Received: 6 July 2022 | Revised: 7 October 2022 | Accepted: 10 October 2022

DOI: 10.1002/sta4.515

ORIGINAL ARTICLE

WILEY

Regularized Buckley–James method for right-censored outcomes with block-missing multimodal covariates

Haodong Wang[1] | Quefeng Li[2] | Yufeng Liu[1,2,3,4,5]

[1]Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, 27599, North Carolina, USA

[2]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, 27516, North Carolina, USA

[3]Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, 27599-7264, North Carolina, USA

[4]Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, 27514, North Carolina, USA

[5]Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, 27514, North Carolina, USA

Correspondence
Yufeng Liu, Department of Statistics & Operations Research, The University of North Carolina at Chapel Hill, 354 Hanes Hall, CB3260, Chapel Hill, NC 27599, USA.
Email: yfliu@email.unc.edu

Funding information
National Institute of General Medical Sciences, Grant/Award Number: R01GM126550; National Science Foundation, Grant/Award Number: DMS2100729; National Institute on Aging, Grant/Award Number: R01AG073259

High-dimensional data with censored outcomes of interest are prevalent in medical research. To analyze such data, the regularized Buckley–James estimator has been successfully applied to build accurate predictive models and conduct variable selection. In this paper, we consider the problem of parameter estimation and variable selection for the semiparametric accelerated failure time model for high-dimensional block-missing multimodal neuroimaging data with censored outcomes. We propose a penalized Buckley–James method that can simultaneously handle block-wise missing covariates and censored outcomes. This method can also perform variable selection. The proposed method is evaluated by simulations and applied to a multimodal neuroimaging dataset and obtains meaningful results.

KEYWORDS
accelerated failure time model, censored data, high dimensional data, missing covariates, moment estimation, survival analysis

1 | INTRODUCTION

Measures of neural activity such as magnetic resonance imaging (MRI) and positron emission tomography (PET) yield thousands of predictor variables for diagnosis and prognosis in patients with diseases such as Alzheimer's disease (AD). Since not all variables contain helpful information for the model, selecting a parsimonious subset of variables with good prediction accuracy can be very important. While linear regression with a scalar response and complete data has been well studied (Tibshirani, 1996), data with censored outcomes and incomplete covariates present new challenges.

AD is a progressive neurodegenerative disease characterized by overall cognitive decline as well as behavioral and functional changes that eventually impair an individual's ability to perform the basic daily activities. People diagnosed with mild cognitive impairment (MCI), which is generally considered as a transitional stage between healthy cognitive aging and dementia, are at significantly increased risk of clinical AD (Gauthier et al., 2006; Knopman et al., 2003). Thus, MCI is a critical prognostic and therapeutic component in AD study, and it is helpful to develop reliable methods to analyze the conversion time from MCI to AD. Although up to 60% of MCI patients convert to AD within 10 years, many return to the normal cognitive function (Manly et al., 2008; Mitchell & Shiri-Feshki, 2009). The AD conversion time of those participants who did not progress to AD during their follow-up period was censored at their last visit time.

Increasing efforts have focused on building predictive models of the AD conversion based on the proportional hazard (PH) model or the accelerated failure time (AFT) model. For example, to examine the usage of MRI and cerebrospinal fluid (CSF) biomarkers to predict the conversion from MCI to AD, Vemuri et al. (2009) used a single-predictor Cox PH model to predict the hazard ratio of the conversion from MCI to AD. They showed that MRI and CSF provide complimentary predictive information about the conversion from MCI to AD. They also showed that combining MRI and CSF can predict better than using either source alone. Liu et al. (2017) used independent component analysis (ICA) and the multivariate Cox PH regression model to identify promising risk factors associated with MCI conversion.

In the literature, many papers also used the AFT model (Cox & Oakes, 2018; Kalbfleisch & Prentice, 2011) to analyze the conversion time of AD, where the response refers to the logarithm of a failure time. The AFT model is based on the linear model and the estimated regression coefficients can help provide useful interpretation (Reid, 1994). It is well known that the linear model and the PH model cannot hold simultaneously except in the case of the extreme value error distribution. Two general estimation strategies to handle censored responses in the AFT model include extensions of least-squares estimators through missing data techniques (Buckley & James, 1979; Koul et al., 1981; Lai & Ying, 1991; Miller & Halpern, 1982) and rank-based methods (Lai & Ying, 1991; Prentice, 1978; Tsiatis, 1990). For example, Oulhaj et al. (2009) used the smoothing AFT procedure with G-splines to predict the period of time before cognitive impairment occurs in community-dwelling elderly. Ning et al. (2011) proposed a generalized Buckley–James type of estimator using right-censored and length-biased data under semiparametric transformation and AFT models. Their proposed method was applied to assess the effect of different diagnostic categories of AD using survival data.

Several authors have also extended the PH and AFT models for variable selection and explored their properties. Tibshirani (1997) and Gui and Li (2005) developed regularized Cox regression methods by adding an ℓ_1 penalization term to the partial likelihood function of the Cox model. Similarly, Datta et al. (2007) and Johnson (2009) added an ℓ_1 penalization term to the Buckley–James estimators for the AFT model. Wang et al. (2008) added the elastic-net penalty in the Buckley–James method for the AFT model to relate high-dimensional genomic data to censored survival outcomes. Johnson (2009) proved that, under suitable regularity conditions, an ℓ_1-penalized Buckley–James estimator with only one iteration yields a root-n consistent solution. Wang and Wang (2010) proposed the Buckley–James boosting method for the semiparametric AFT models with right-censored survival data, which can be used for prediction and variable selection.

In the past few years, there has been extensive research on using neuroimaging data for MCI and AD prediction (Eskildsen et al., 2013; Park & Moon, 2016). However, data in Alzheimer's Disease Neuroimaging Initiative (ADNI) study were collected from different sources, which include MRI, PET, and CSF. Data from a specific modality can be entirely missing due to patient dropouts or other practical issues. This leads to a block-wise missing data structure. Due to the block-wise missing structure with high dimensionality and censored response, it is challenging to identify the patients likely to convert from MCI to AD. It is also interesting to further predict the conversion time for an effective risk estimate, which could lead to an efficient intervention of pharmacological treatments for early AD (Jack Jr, 2012).

Most of the AFT and PH models can only work with complete covariates. To handle incomplete multimodal data in the ADNI study, one may use traditional AFT or PH models by simply removing those observations with missing entries. However, such a procedure may greatly reduce the number of observations and lead to loss of information. Another approach is to perform data imputation, where missing data are replaced by data generated from an imputation model. Imputation methods have been used in both AFT models (Qi et al., 2018) and PH models (Hsu & Yu, 2019; Paik & Tsai, 1997; White & Royston, 2009) to deal with incomplete covariates. Another approach is to use weighted estimating equations for AFT models (Nan et al., 2009; Steingrimsson & Strawderman, 2017) and PH models (Luo et al., 2009; Qi et al., 2005; Steingrimsson & Strawderman, 2017; Wang & Chen, 2001; Xu et al., 2009). They applied the inverse probability weighted (IPW) technique to the existing estimation procedures for the complete covariate cases. In particular, Yu (2011) proposed a revised Buckley–James estimator for data missing by design. In order to deal with multimodal block-wise missing data, Yu et al. (2020) proposed a new direct sparse regression procedure using the estimated covariance matrix from block-missing multimodal data (DISCOM). They first used all available information to estimate the covariance matrix of the predictors and the cross-covariance vector between the predictors and the response variable. Then they used an extended LASSO-type estimator to estimate the coefficients based on the estimated covariance matrix and the cross-covariance vector. Despite its usefulness, however, the DISCOM only considers the linear regression model for uncensored data.

In this paper, we propose a regularized Buckley–James method for variable selection, parameter estimation, and prediction for right-censored outcomes with block-wise missing data. It extends the DISCOM method (Yu et al., 2020) to right-censored survival data. Our proposed method has several attractive properties. First, our approach can handle high-dimensional data and perform variable selection. Second, it works with data with block-wise missing covariates and censored outcomes. Third, our method can still deliver reliable results even if our training data have no observation with complete covariates. Our proposed method includes two steps. The first step is to estimate each element of the covariance and cross-covariance matrices using all available observations. The second step is to use a penalized approach to estimate the sparse regression coefficient vector by the Buckley–James method. Numerical studies and the ADNI data application confirm that the proposed method performs competitively for block-wise missing data.

The remainder of this paper is organized as follows. In Section 2, we introduce the problem background and our model. Simulation studies and a multimodal ADNI data example are presented in Sections 3 and 4. A brief summary of the paper is provided in Section 5.

2 | METHODOLOGY

2.1 | Problem setup and notations

Consider the following semiparametric AFT model,

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta}^* + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}^* = (b_1, ..., b_p)^\top \in \mathbb{R}^p$ is an unknown p-dimensional vector, $\mathbf{T} = (t_1, ..., t_n)^\top \in \mathbb{R}^n$ is the response vector, $\mathbf{X} = (\boldsymbol{x}_1,...,\boldsymbol{x}_n)^\top$ is the $n \times p$ design matrix, and $\epsilon = (\epsilon_1,...,\epsilon_n)^\top$ is the the error vector. Assume that $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. realizations of a random vector $\boldsymbol{X} = (X_1, ..., X_p)^\top$ with zero mean and a covariance matrix $\boldsymbol{\Sigma}_{XX} = (\sigma_{ij}^{XX}) \in \mathbb{R}^{p\times p}$. Denote $\boldsymbol{\Sigma}_{XT} = (\sigma_i^{XT}) \in \mathbb{R}^p$ as the cross-covariance vector between $\mathbf{x}_i$ and t_i for $1 \le i \le n$. Assume that the predictors come from multiple modalities and there are p_k predictors in the kth modality. In addition, assume that $\mathbf{X}$ has block-wise missing values. That is, for each sample, its measurements in one modality can be entirely missing. Let $\tilde{\mathbf{X}} = (\tilde{\boldsymbol{x}}_1,...,\tilde{\boldsymbol{x}}_n)^\top$ be the imputed design matrix, where the missing values in $\mathbf{X}$ are imputed by some imputation methods such as multiple imputation (Rubin, 2004) or the soft-impute algorithm (Mazumder et al., 2010). For simplicity, we use the soft-impute algorithm to calculate $\tilde{\mathbf{X}}$ in our numerical and case studies. The errors ϵ_i for $1 \le i \le n$ are i.i.d. realizations from a random variable ϵ with zero mean and covariance σ_ϵ. Moreover, we further assume that $\mathbf{x}_i$ and ϵ_i are uncorrelated for $1 \le i \le n$.

Let $\mathbf{T}$ denote the transformed failure time, for example, the logarithm of the conversion time from MCI to AD. Suppose that $\mathbf{C} = (c_1,...,c_n)^\top \in \mathbb{R}^n$ is the transformed censoring time which is transformed in the same way as $\mathbf{T}$, with c_i being independent of t_i given $\mathbf{x}_i$. When $\mathbf{T}$ is right censored, we can only observe $(y_i, \delta_i, \mathbf{x}_i)$ for $1 \le i \le n$, where $y_i = \min(t_i, c_i)$, and $\delta_i = 1_{\{t_i \le c_i\}}$ is the censoring indicator for the ith observation.

We employ the following notation throughout this article. For a square matrix $\mathbf{C} = (c_{ii'}) \in \mathbb{R}^{p\times p}$, we denote its diagonal matrix as diag($\mathbf{C}$). For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, we define the largest and smallest eigenvalues of $\mathbf{A}$ as $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively. For a vector $\mathbf{v} \in \mathbb{R}^p$, let $\|\mathbf{v}\|_1 = \sum_i |v_i|$, and $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$.

2.2 | Regularized Buckley–James regression for complete observations

If there is no response censored and no covariate missing, then $t_i = y_i$ for $1 \le i \le n$ and $\mathbf{X}$ are fully observed. Then the least-squares method can be applied to estimate the parameters in model (1) by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{Y} = (y_1,...,y_n)^\top \in \mathbb{R}^n$. For a censored response with complete covariates, the key idea of the Buckley–James method is to replace the censored t_i by its expectation conditional on δ_i and $\boldsymbol{x}_i$. Define the pseudo failure time y_i^* as

$$y_i^* = \begin{cases} y_i & \delta_i = 1; \\ \mathbb{E}(t_i | t_i > y_i,\ \boldsymbol{x}_i) & \delta_i = 0. \end{cases}$$

It can be shown that $\mathbb{E}(y_i^*) = \mathbb{E}(t_i)$ for $1 \le i \le n$; for details, see Smith (2017). With the true $\boldsymbol{\beta}^*$, $\mathbb{E}(t_i | t_i > y_i, \mathbf{x}_i)$ has the form of

$$\begin{aligned} \mathbb{E}(t_i | t_i > y_i, \mathbf{x}_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}^* + \mathbb{E}\left(\epsilon_i | \epsilon_i > y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^*\right) \\ &= \mathbf{x}_i^\top \boldsymbol{\beta}^* + \int_{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^*}^{\infty} \frac{t dF(t)}{1 - F\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^*\right)}, \end{aligned} \tag{2}$$

where F is the distribution function of residual $\epsilon_i(\boldsymbol{\beta}^*) = t_i - \mathbf{x}_i^\top \boldsymbol{\beta}^*$ for $1 \le i \le n$. The distribution of $\epsilon_i(\boldsymbol{\beta}^*)$ can be estimated nonparametrically by the Kaplan–Meier estimator (Kaplan & Meier, 1958)

$$\hat{F}(t) = 1 - \prod_{i:\epsilon_i < t} \left(1 - \frac{d_i}{n_i}\right), \tag{3}$$

where $d_i = \sum_{j=1}^n I(\epsilon_j = \epsilon_i \text{and } \delta_j = 1)$ and $n_i = \sum_{j=1}^n I(\epsilon_j > \epsilon_i)$. After substituting F with $\hat{F}$ in (2), the $\tilde{y}_i^*$ can be simplified as

$$\tilde{y}_i^* = \delta_i y_i + (1-\delta_i)\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \int_{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^*}^{\infty} \frac{t d\hat{F}(t)}{1-\hat{F}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^*)}\right). \tag{4}$$

Then the least-squares method can be applied to the following regression model:

$$\tilde{y}_i^* = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i^*, \tag{5}$$

where $\epsilon_i^* = \tilde{y}_i^* - \boldsymbol{x}_i^\top \boldsymbol{\beta}^*$. In (5), we replace the censored y_i by an estimate of $\mathbb{E}(t_i | t_i > y_i, \boldsymbol{x}_i)$ and treat $\tilde{y}_i^*$ as a pseudo response. Then, estimating $\boldsymbol{\beta}^*$ in (5) becomes a standard least squares problem. Let $\tilde{\mathbf{Y}}^* = (\tilde{y}_1^*, ..., \tilde{y}_n^*)^\top$. The least-squares estimator of $\boldsymbol{\beta}^*$ in model (5) is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \frac{1}{2}(\tilde{\mathbf{Y}}^* - \mathbf{X}\boldsymbol{\beta})^\top (\tilde{\mathbf{Y}}^* - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Y}}^*.$$

The final estimate of $\boldsymbol{\beta}^*$ requires an iterative procedure since values of $\tilde{y}_i^*$ defined in (4) contain $\boldsymbol{\beta}$.

In many areas such as genomic, medicine, and bioinformatics, the number of features p is usually much larger than the sample size n and the classical Buckley–James method fails. Regularization is needed to obtain a stable estimator of $\boldsymbol{\beta}$ with small prediction error. In this case, a modified Buckley–James approach by using penalized least-squares with the penalty term $P_\lambda(\boldsymbol{\beta})$ can be used, where λ is the tuning parameter. To be specific, we consider the following minimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \tag{6}$$

where λ is the tuning parameters and can be determined by cross validation. Given an initial value $\boldsymbol{\beta}^{(0)}$, the final estimator of $\boldsymbol{\beta}$ can be calculated (3), (4), and (6) iteratively.

2.3 | Regularized Buckley–James regression for block-wise missing multimodal observations

Next, we extend the regularized Buckley–James regression to block-wise missing multimodal observations. We assume that the predictors are collected from K modalities, and the kth modality has p_k predictors for $1 \le k \le K$.

Recall that the regularized Buckley–James regression for complete observations iteratively estimates y_i^* by (4) and then solves the minimization problem (6). In order to handle block-wise missing data, given $\tilde{y}_i^*$, we consider the population version of the ℓ_1 penalized least-square estimator

$$\begin{aligned}\boldsymbol{\beta}^0 &= \left(\beta_1^0, \beta_2^0, ..., \beta_p^0\right)^T \\ &= \arg\min_{\beta} \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^n (\tilde{y}_i^* - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2\right] + \lambda\|\boldsymbol{\beta}\|_1.\end{aligned}$$

If both $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{X\tilde{Y}^*}$ are known, $\boldsymbol{\beta}^0$ can be equivalently obtained by solving the following optimization problem:

$$\boldsymbol{\beta}^0 = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{XX}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{X\tilde{Y}^*}^\top \boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1.$$

Therefore, we can obtain the estimator $\hat{\boldsymbol{\beta}}$ if estimators for $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{X\tilde{Y}^*}$ are available. Denote $\hat{\boldsymbol{\Sigma}}_{XX}$ as the estimator of $\boldsymbol{\Sigma}_{XX}$. Next, we explain how to calculate $\hat{\boldsymbol{\Sigma}}_{XX}$ when data are block-wise missing. Define $S_{jk}^{XX} = \{i : x_{ij} \text{ and } x_{ik} \text{ are not missing}\}$, and n_{jt}^{XX} the cardinality of S_{jk}^{XX}. Let $\tilde{\boldsymbol{\Sigma}}_{XX}$ be the sample covariance matrix derived from all observed data, that is, $\tilde{\boldsymbol{\Sigma}}_{XX} = (\tilde{\sigma}_{jt}^{XX})$, where $\tilde{\sigma}_{jt}^{XX} = \sum_{i \in S_{jt}^{XX}} (x_{ij}x_{it}/n_{jt}^{XX})$. Note that $\tilde{\boldsymbol{\Sigma}}_{XX}$ is required to be an unbiased estimator of $\boldsymbol{\Sigma}_{XX}$. When the elements in $\mathbf{X}$ are missing completely at random, the unbiasedness assumption is satisfied. However, the unbiasedness assumption can also hold under some other missing mechanisms.

Since the data $\mathbf{X}$ are block-wise missing, the estimator $\tilde{\boldsymbol{\Sigma}}_{XX}$ defined above can be ill-conditioned. As a result, $\tilde{\boldsymbol{\Sigma}}_{XX}$ is not a good estimator of $\boldsymbol{\Sigma}_{XX}$. Thus, it cannot be used directly in our optimization problem. To resolve this problem, we partition $\tilde{\boldsymbol{\Sigma}}_{XX}$ into K^2 blocks, denoted as $\tilde{\boldsymbol{\Sigma}}^{k_1k_2} \in \mathbb{R}^{p_{k_1} \times p_{k_2}}$ for $1 \le k_1, k_2 \le K$. We let

$$\tilde{\boldsymbol{\Sigma}}_I = \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}^{11} & & & \\ & \tilde{\boldsymbol{\Sigma}}^{22} & & \\ & & \ddots & \\ & & & \tilde{\boldsymbol{\Sigma}}^{KK} \end{pmatrix} \text{ and } \tilde{\boldsymbol{\Sigma}}_C = \begin{pmatrix} \boldsymbol{0} & \tilde{\boldsymbol{\Sigma}}^{12} & \dots & \tilde{\boldsymbol{\Sigma}}^{1K} \\ \tilde{\boldsymbol{\Sigma}}^{21} & \boldsymbol{0} & \dots & \tilde{\boldsymbol{\Sigma}}^{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\boldsymbol{\Sigma}}^{K1} & \tilde{\boldsymbol{\Sigma}}^{K2} & \dots & \boldsymbol{0} \end{pmatrix},$$

where $\tilde{\boldsymbol{\Sigma}}_I$ is a $p \times p$ block-diagonal matrix containing K diagonal blocks of $\tilde{\boldsymbol{\Sigma}}_{XX}$, and $\tilde{\boldsymbol{\Sigma}}_C = \tilde{\boldsymbol{\Sigma}}_{XX} - \tilde{\boldsymbol{\Sigma}}_I$ is a $p \times p$ matrix containing all off-diagonal blocks of $\tilde{\boldsymbol{\Sigma}}_{XX}$. Here, $\tilde{\boldsymbol{\Sigma}}_I$ and $\tilde{\boldsymbol{\Sigma}}_C$ are called the intramodality and cross-modality sample covariance matrices, respectively. Since data are block-wise missing, we use more data to estimate the entries in $\tilde{\boldsymbol{\Sigma}}_I$ than those in the $\tilde{\boldsymbol{\Sigma}}_C$. Thus, the estimator $\tilde{\boldsymbol{\Sigma}}_I$ can be relatively more accurate than $\tilde{\boldsymbol{\Sigma}}_C$. We linearly combine $\tilde{\boldsymbol{\Sigma}}_I$ and $\tilde{\boldsymbol{\Sigma}}_C$ with different weights to estimate $\boldsymbol{\Sigma}_{XX}$. In addition, as in Yu et al. (2020), we adopt the idea of shrinkage estimation of the covariance matrix (Fisher & Sun, 2011) and add the diagonal matrix $\text{diag}(\tilde{\boldsymbol{\Sigma}}_I)$ to our estimator to ensure the resulting estimator to be positive definite. We let

$$\hat{\boldsymbol{\Sigma}}_{XX} = \alpha_1 \tilde{\boldsymbol{\Sigma}}_I + (1-\alpha_1)\text{diag}(\tilde{\boldsymbol{\Sigma}}_I) + \alpha_2 \tilde{\boldsymbol{\Sigma}}_C, \tag{7}$$

where $\alpha_1, \alpha_2 \in [0,1]$ are two shrinkage weights. The diagonal matrix $(1-\alpha_1)\text{diag}(\tilde{\boldsymbol{\Sigma}}_I)$ in (7) ensures that the diagonal entries of our estimator are not shrunk. The eigenvalues of $\hat{\boldsymbol{\Sigma}}_{XX}$ are larger than or equal to $\alpha_1 \lambda_{\min}(\tilde{\boldsymbol{\Sigma}}_I) + (1-\alpha_1)\lambda_{\min}(\text{diag}(\tilde{\boldsymbol{\Sigma}}_I)) + \alpha_2 \lambda_{\min}(\tilde{\boldsymbol{\Sigma}}_C)$ by Weyl's theorem, where $(1-\alpha_1)\lambda_{\min}(\text{diag}(\tilde{\boldsymbol{\Sigma}}_I)) > 0$ since $\text{diag}(\tilde{\boldsymbol{\Sigma}}_I)$ is a positive-definite matrix. Thus, $\hat{\boldsymbol{\Sigma}}_{XX}$ is guaranteed to be positive definite by carefully selecting the tuning parameters α_1 and α_2. In practice, α_1 and α_2 can be chosen from the set $\{(\alpha_1,\alpha_2) : \alpha_1 \in [0,1], \alpha_2 \in [0,1], \hat{\boldsymbol{\Sigma}}_{XX} \text{ is positive semidefinite}\}$ by cross-validation or using an additional tuning dataset.

Let $\tilde{y}_i^{*(m)}$ be the ith failure time calculated in the mth step of the Buckley–James method, $\tilde{\mathbf{Y}}_i^{*(m)} = (\tilde{y}_1^{*(m)}, \dots, \tilde{y}_n^{*(m)})^\top$, $\boldsymbol{\Sigma}_{X\tilde{Y}^*}^{(m)}$ be the covariance vector between $\mathbf{X}$ and $\tilde{\mathbf{Y}}^{*(m)}$, and $\hat{\boldsymbol{\Sigma}}_{X\tilde{Y}^*}^{(m)}$ be an estimator of $\boldsymbol{\Sigma}_{X\tilde{Y}^*}^{(m)}$. Next, we discuss how to calculate $\hat{\boldsymbol{\Sigma}}_{X\tilde{Y}^*}^{(m)}$ when $\mathbf{X}$ is block-wise missing. Let $\boldsymbol{\beta}^{(m-1)}$ be the coefficient vector derived in the $(m-1)$th step. In the mth step, $\tilde{y}^{*(m)}$ is defined as

$$\tilde{y}_i^{*(m)} = \delta_i y_i + (1-\delta_i)\left(\mathbf{x}_i^\top \boldsymbol{\beta}^{(m-1)} + \int_{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(m-1)}}^{\infty} \frac{t d\tilde{F}^{(m)}(t)}{1 - \tilde{F}^{(m)}\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(m-1)}\right)} \right),$$

where $\tilde{F}^{(m)}$ is the estimated distribution function of $t_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(m-1)}$. However, since $\mathbf{X}$ is block-wise missing, $\tilde{y}_i^{*(m)}$ cannot be calculated directly. In order to estimate $\boldsymbol{\Sigma}_{X\tilde{Y}^*}^{(m-1)}$, we decompose it as

$$\begin{aligned} \boldsymbol{\Sigma}_{X\tilde{Y}^*}^{(m-1)} &= \mathbb{E}(\mathbf{X}^\top \tilde{\mathbf{Y}}^{*(m)}) \\ &= \mathbb{E}(\mathbf{X}^\top (\mathbf{X}(\boldsymbol{\beta}^{(m-1)}) + \tilde{\mathbf{E}}^{*(m)})) \\ &= \mathbb{E}(\mathbf{X}^\top \mathbf{X})\boldsymbol{\beta}^{(m-1)} + \mathbb{E}(\mathbf{X}\tilde{\mathbf{E}}^{*(m)}), \end{aligned}$$

where $\tilde{\mathbf{E}}^{*(m)} = (\tilde{e}_1^*(\boldsymbol{\beta}^{(m-1)}), \dots, \tilde{e}_n^*(\boldsymbol{\beta}^{(m-1)}))^\top$ and

$$\tilde{e}_i^*(\boldsymbol{\beta}^{(m-1)}) = \begin{cases} y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(m-1)} & \delta_i = 1; \\ \displaystyle\int_{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(m-1)}}^{\infty} \frac{t d\tilde{F}^{(m)}(t)}{1 - \tilde{F}^{(m)}\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(m-1)}\right)} & \delta_i = 0. \end{cases}$$

Let $\boldsymbol{\Sigma}_{X\tilde{E}^*}^{(m)}$ be the covariance vector between $\mathbf{X}$ and $\tilde{\mathbf{E}}^{*(m)}$, and $\hat{\boldsymbol{\Sigma}}_{X\tilde{E}^*}^{(m)}$ be an estimator of $\boldsymbol{\Sigma}_{X\tilde{E}^*}^{(m)}$. Then we can estimate $\boldsymbol{\Sigma}_{X\tilde{Y}^*}^{(m)}$ as

$$\hat{\boldsymbol{\Sigma}}_{X\tilde{Y}^*}^{(m)} = \hat{\boldsymbol{\Sigma}}_{XX}\boldsymbol{\beta}^{(m-1)} + \hat{\boldsymbol{\Sigma}}_{X\tilde{E}^*}^{(m)}. \tag{8}$$

Define $S_j^X = \{i : x_{ij} \text{ is not missing}\}$ and let n_j^X as the cardinality of S_j^X. In order to estimate $\boldsymbol{\Sigma}_{X\tilde{E}^*}^{(m)}$, let $\hat{\mathbf{E}}^{*(m)} = (\hat{e}_1^*(\boldsymbol{\beta}^{(m-1)}), \dots, \hat{e}_n^*(\boldsymbol{\beta}^{(m-1)}))$ and

$$\hat{e}_i^*(\boldsymbol{\beta}^{(m-1)}) = \begin{cases} y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^{(m-1)} & \delta_i = 1; \\ \displaystyle\int_{y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^{(m-1)}}^{\infty} \frac{t d\hat{F}^{(m)}(t)}{1 - \hat{F}^{(m)}\left(y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}^{(m-1)}\right)} & \delta_i = 0. \end{cases}$$

Here, $\tilde{\boldsymbol{x}}_i$ are the imputed predictors and $\hat{F}^{(m)}$ is the estimated distribution function of $t_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(m-1)}$. Define $\tilde{\boldsymbol{\Sigma}}^{(m)}_{X\tilde{E}^*}$ as the sample covariance matrix using all available data, that is, $\tilde{\boldsymbol{\Sigma}}^{(m)}_{X\tilde{E}^*} = (\tilde{\sigma}_j^{X\tilde{E}^*,(m)})$, where $\tilde{\sigma}_j^{X\tilde{E}^*,(m)} = \sum_{i \in S_j^X} x_{ij}\hat{e}_i^* / n_j^X$. Since our estimator $\hat{\boldsymbol{\Sigma}}_{XX}$ in (7) is a shrinkage estimator, we also use a shrinkage estimator to estimate $\boldsymbol{\Sigma}^{(m)}_{X\tilde{E}^*}$ by

$$\hat{\boldsymbol{\Sigma}}^{(m)}_{X\tilde{E}^*} = \alpha_3 \tilde{\boldsymbol{\Sigma}}^{(m)}_{X\tilde{E}^*}, \tag{9}$$

where $\alpha_3 \in [0,1]$ is the shrinkage weight. In practice, α_3 can also be chosen by cross-validation or using an additional tuning dataset.

In summary, given $\hat{\boldsymbol{\Sigma}}_{XX}$ and $\hat{\boldsymbol{\Sigma}}^{(m)}_{X\tilde{E}^*}$ as defined in (7) and (9), in the mth iteration of the regularized Buckley–James method, we solve the optimization problem

$$\boldsymbol{\beta}^{(m)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \hat{\boldsymbol{\Sigma}}_{XX}\boldsymbol{\beta} - \boldsymbol{\beta}^{(m-1)\top}\hat{\boldsymbol{\Sigma}}_{XX}^\top\boldsymbol{\beta} + \hat{\boldsymbol{\Sigma}}^{(m)\top}_{X\tilde{E}^*}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \tag{10}$$

by the proximal gradient descent algorithm (Parikh et al., 2014).

In Algorithm 1, we summarized the major steps for our proposed method, DISCOM-BJ, given a set of tuning parameters $(\alpha_1,\alpha_2,\alpha_3,\lambda)$.

Algorithm 1 Regularized Buckley–James method by using covariance from multimodality data

Let $\boldsymbol{\beta}^{(0)}$ be the initial value of $\boldsymbol{\beta}$.

while $\left|\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}\right| > d$ **do** ▷ here d is a prespecified constant.

Compute $\epsilon_i(\boldsymbol{\beta}^{(m-1)})$ for $1 \le i \le n$ by

$$\epsilon_i\left(\boldsymbol{\beta}^{(m-1)}\right) = y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(m-1)},$$

where $\tilde{\boldsymbol{x}}_i$ is the imputed predictors of the ith observation.

Compute $\hat{F}^{(m)}(t)$ by

$$\hat{F}^{(m)}(t) = 1 - \prod_{i:\epsilon_i(\boldsymbol{\beta}^{(m-1)})<t}\left(1 - \frac{d_i}{n_i}\right),$$

where $d_i = \sum_{j=1}^n I[\epsilon_j(\boldsymbol{\beta}^{(m-1)}) = \epsilon_i(\boldsymbol{\beta}^{(m-1)})\text{and } \delta_j = 1]$ and $n_i = \sum_{j=1}^n I(\epsilon_j(\boldsymbol{\beta}^{(m-1)}) > \epsilon_i(\boldsymbol{\beta}^{(m-1)}))$.

Compute $\tilde{\mathbf{E}}^{*(m)} = (e_1^*(\boldsymbol{\beta}^{(m-1)}), \ldots, e_n^*(\boldsymbol{\beta}^{(m-1)}))^\top$ by

$$e_i^*(\boldsymbol{\beta}^{(m-1)}) = \begin{cases} y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(m-1)} & \delta_i = 1; \\ \int_{y_i - \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}^{(m-1)}}^{\infty} \frac{t d\hat{F}^{(m)}(t)}{1-\hat{F}^{(m)}(y_i - \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}^{(m-1)})} & \delta_i = 0. \end{cases}$$

Compute $\hat{\boldsymbol{\Sigma}}_{XX}$ and $\hat{\boldsymbol{\Sigma}}^{(m)}_{X\tilde{E}^*}$ by (7) and (9), respectively.

Update $\boldsymbol{\beta}^{(m)}$ by

$$\boldsymbol{\beta}^{(m)} = \min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top\hat{\boldsymbol{\Sigma}}_{XX}\boldsymbol{\beta} - \boldsymbol{\beta}^{(m-1)\top}\hat{\boldsymbol{\Sigma}}_{XX}^\top\boldsymbol{\beta} + \hat{\boldsymbol{\Sigma}}^{(m)\top}_{X\tilde{E}^*}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1.$$

end while

We make two important remarks about the proposed procedure. First, our method applies to any penalty for linear models, including LASSO and elastic net (Zou & Hastie, 2005). Second, to be numerically effective, the starting values $\boldsymbol{\beta}^{(0)}$ may be obtained by using the least-squares estimator treating all observations as uncensored (Buckley & James, 1979). Other choices, for example, using only uncensored observations, are also feasible.

3 | NUMERICAL STUDY

We perform some numerical studies to compare our proposed method (DISCOM-BJ) with some other methods, which include

1. ℓ_2-BJ, which applies the regularized Buckley–James regression to samples with complete observations and uses $P_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2$;
2. Imputed-ℓ_2-BJ, which applies the regularized Buckley–James regression to all samples with missing values being imputed by the soft-thresholded SVD method and uses $P_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2$;
3. ℓ_1-BJ, which applies the regularized Buckley–James regression to samples with complete observations and uses $P_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$;

4. Imputed-ℓ_1-BJ, which applies the regularized Buckley–James regression to all samples with missing values being imputed by the soft-thresholded SVD method and uses $P_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$;
5. Boosting-BJ, which applies the Buckley–James boosting method with linear least-squares (Wang & Wang, 2010) to samples with complete observations;
6. Imputed-Boosting-BJ, which applies the Buckley–James boosting method with linear-least squares (Wang & Wang, 2010) to all samples with missing values being imputed by the soft-thresholded SVD method.

For all examples, we generate the natural logarithm of the true survival time by

$$T = \boldsymbol{x}^\top \boldsymbol{\beta} + \epsilon, \text{ where } \epsilon \sim N(0,1)$$

and set $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{jt})$, where $\sigma_{jt} = 0.6^{|j-t|}$. The data are generated from three modalities whose dimensions p_1,p_2, and p_3 are specified in each example. The true coefficient vector is

$$\boldsymbol{\beta} = \left(b, b, b, \underbrace{0, \ldots, 0}_{p_1-3}, b, b, b, \underbrace{0, \ldots, 0}_{p_2-3}, b, b, b, \underbrace{0, \ldots, 0}_{p_3-3} \right),$$

where b is a constant. We generate ϵ_1,ϵ_2,...,$\epsilon_n \overset{\text{iid}}{\sim} N(0,1)$. The censoring time $\mathbf{C}$ is generated from unif(τ_l,τ_u), where τ_l,τ_u are tuned to achieve the desired censoring rate. The censoring rates are specified in each example.

The training dataset contains 25 samples with complete observations, 25 samples with observations from the third modality, 25 samples with observations from the first and the third modalities, and 25 samples with observations from the first modality. In other words, the missing values in the training data are missing completely at random. The tuning dataset contains 100 samples with complete observations without censoring response and the testing dataset includes 400 samples with complete observations without censoring response. For each method, we train our model with different tuning parameters on the training dataset. Then we choose the optimal tuning parameters minimizing the mean squared error on the tuning dataset.

For each example, the experiment is repeated 50 times. To evaluate the selection performance of the algorithm, we use false-positive rate (FPR) and false-negative rate (FNR) defined as FPR = FP/(FP + TN)and FNR = FN/(FN + TP), where FN is the number of coefficients wrongly estimated as zero, TN is the number coefficients rightfully estimated as zero, TP is the number of coefficients rightfully estimated as nonzero, and FP is the number of coefficients wrongly estimated as nonzero. Furthermore, to evaluate the accuracy of our estimators, the mean squared error MSE $= \|\mathbf{T}_{\text{test}} - \hat{\mathbf{T}}_{\text{test}}\|_2$ and the estimation error $EST = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2$ in the test data are used as the criteria, where T_{test} is the logarithm of the survival time vector in the test dataset, $\hat{T}_{\text{test}}$ is the logarithm of the predicted survival time vector in the test dataset, and $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector.

In Example 1, we examine how our method performs with various signal-to-noise ratios. We set $p = 90$, $p_1 = p_2 = p_3 = 30$ and the censoring rate equal to 50%. In Example 1(a) and 1(b), we set b to be 0.5 and 2, respectively.

In Example 2, we examine how our method performs with various p. We set $b = 1$ and the censoring rate equal to 50%. In Example 2(a), we set p to be 60, where $p_1 = p_2 = p_3 = 20$. In Example 2(b), we set p to be 120, where $p_1 = p_2 = p_3 = 40$.

TABLE 1 Performance comparison of different methods for Example 1 with different signal to noise ratios. The values in the parentheses are the standard errors of the measures.

	Example 1(a) [low signal to noise ratio]				Example 1(b) [high signal to noise ratio]			
	MSE	EST	FPR	FNR	MSE	EST	FPR	FNR
ℓ_2-BJ	4.14 (0.09)	1.34 (0.01)	1.00 (0.00)	0.00 (0.00)	47.19 (1.22)	5.31 (0.05)	1.00 (0.00)	0.00 (0.00)
Imputed-ℓ_2-BJ	2.91 (0.07)	1.21 (0.01)	1.00 (0.00)	0.00 (0.00)	26.26 (0.80)	4.62 (0.05)	1.00 (0.00)	0.00 (0.00)
ℓ_1-BJ	3.64 (0.12)	1.40 (0.02)	0.17 (0.02)	0.49 (0.03)	29.93 (1.49)	4.54 (0.09)	0.23 (0.02)	0.24 (0.02)
Imputed-ℓ_1-BJ	**2.56**(0.09)	1.27 (0.03)	0.16 (0.01)	0.31 (0.02)	16.84 (0.80)	4.22 (0.09)	0.20 (0.01)	0.14 (0.01)
Boosting-BJ	4.37 (0.15)	1.54 (0.03)	0.07 (0.00)	0.58 (0.02)	41.40 (1.70)	5.32 (0.09)	0.06 (0.00)	0.42 (0.03)
Imputed-Boosting-BJ	2.85 (0.09)	1.22 (0.02)	**0.06** (0.00)	0.34 (0.02)	25.67 (0.98)	4.37 (0.08)	**0.04** (0.00)	0.22 (0.02)
DISCOM-BJ	**2.51** (0.09)	**1.21** (0.03)	0.18 (0.02)	**0.26** (0.02)	**15.16** (0.63)	**3.87** (0.08)	0.19 (0.01)	**0.09**(0.01)

In Example 3, we examine how our method performs with various censoring rates. We set $p = 90$, $p_1 = p_2 = p_3 = 30$, and $b = 1$. In Example 3 (a) to 3(f), we respectively let $(\tau_l, \tau_u) \in \{(1, 6.8950), (1, 3.64), (1, 1.21), (-5, 5), (-5, 2.515), (-5, 0.16)\}$ such that the yielding censoring rate $\mathbb{P}(T > C)$ ranges from 0.2 to 0.7 with an increment of 0.1.

We report the simulation results in Tables 1–3. Bold numbers indicate the best results in the corresponding numerical study. Table 1 shows the results of Example 1 with two different signal to noise ratios. Table 2 shows the results of Example 2 with two different dimensions. Table 3 shows the results of Example 3 with different censoring rates. Based on the results, we can see that imputed versions of ℓ_2-BJ, ℓ_1-BJ, and Boosting-BJ perform better than the unimputed version of these methods in terms of the parameter estimation and variable selection. Compared with other existing methods, our proposed DISCOM-BJ delivers the best performance in all these three examples.

TABLE 2 Performance comparison of different methods for Example 2 with different dimensions. The values in the parentheses are the standard errors of the measures.

	Example 2(a) [$p = 60$]				Example 2(b) [$p = 120$]			
	MSE	EST	FPR	FNR	MSE	EST	FPR	FNR
ℓ_2-BJ	10.87 (0.34)	2.49 (0.03)	1.00 (0.00)	0.00 (0.00)	13.73 (0.31)	2.73 (0.02)	1.00 (0.00)	0.00 (0.00)
Imputed-ℓ_2-BJ	5.81 (0.19)	2.16 (0.03)	1.00 (0.00)	0.00 (0.00)	8.84 (0.24)	2.43 (0.02)	1.00 (0.00)	0.00 (0.00)
ℓ_1-BJ	7.58 (0.40)	2.38 (0.04)	0.24 (0.02)	0.28 (0.03)	9.91 (0.45)	2.52 (0.05)	0.14 (0.01)	0.37 (0.03)
Imputed-ℓ_1-BJ	4.77 (0.17)	2.14 (0.04)	0.25 (0.01)	0.14 (0.02)	5.83 (0.22)	2.26 (0.05)	0.16 (0.01)	0.19 (0.01)
Boosting-BJ	10.28 (0.39)	2.65 (0.04)	0.08 (0.00)	0.43 (0.02)	12.65 (0.51)	2.86 (0.05)	0.06 (0.00)	0.48 (0.02)
Imputed-Boosting-BJ	6.91 (0.22)	2.16 (0.04)	**0.06** (0.00)	0.23 (0.02)	7.23 (0.26)	2.20 (0.04)	**0.04** (0.00)	0.24 (0.02)
DISCOM-BJ	**4.46** (0.18)	**1.97** (0.04)	0.29 (0.02)	**0.09** (0.02)	**5.52** (0.24)	**2.08** (0.05)	0.16 (0.01)	**0.13** (0.02)

TABLE 3 Performance comparison of different methods for Example 3 with different censoring rates. The values in the parentheses are the standard errors of the measures.

	Example 3(a) [$\mathbb{P}(T > C) = 0.2$]				Example 3(b) [$\mathbb{P}(T > C) = 0.3$]			
	MSE	EST	FPR	FNR	MSE	EST	FPR	FNR
ℓ_2-BJ	9.39 (0.25)	2.43 (0.02)	1.00 (0.00)	0.00 (0.00)	10.40 (0.25)	2.52 (0.03)	1.00 (0.00)	0.00 (0.00)
Imputed-ℓ_2-BJ	5.88 (0.13)	2.15 (0.02)	1.00 (0.00)	0.00 (0.00)	6.31 (0.14)	2.20 (0.02)	1.00 (0.00)	0.00 (0.00)
ℓ_1-BJ	5.85 (0.31)	2.07 (0.05)	0.26 (0.02)	0.15 (0.02)	6.82 (0.31)	2.16 (0.04)	0.26 (0.02)	0.19 (0.02)
Imputed-ℓ_1-BJ	4.18 (0.16)	1.99 (0.04)	0.18 (0.01)	0.10 (0.01)	4.40 (0.16)	2.01 (0.04)	0.19 (0.01)	0.10 (0.01)
Boosting-BJ	8.10 (0.34)	2.35 (0.05)	0.06 (0.00)	0.31 (0.02)	9.08 (0.35)	2.43 (0.04)	0.06 (0.00)	0.35 (0.02)
Imputed-Boosting-BJ	5.18 (0.19)	1.93 (0.03)	**0.04** (0.00)	0.13 (0.01)	5.71 (0.20)	1.98 (0.03)	**0.04** (0.00)	0.16 (0.02)
DISCOM-BJ	**3.81** (0.14)	**1.81** (0.04)	0.16 (0.01)	**0.07** (0.01)	**4.14** (0.13)	**1.85** (0.03)	0.19 (0.01)	**0.09** (0.01)
	Example 3(c) [$\mathbb{P}(T > C) = 0.4$]				**Example 3(d)** [$\mathbb{P}(T > C) = 0.5$]			
ℓ_2-BJ	11.56 (0.28)	2.57 (0.02)	1.00 (0.00)	0.00 (0.00)	12.65 (0.31)	2.64 (0.03)	1.00 (0.00)	0.00 (0.00)
Imputed-ℓ_2-BJ	6.92 (0.16)	2.27 (0.02)	1.00 (0.00)	0.00 (0.00)	7.46 (0.22)	2.32 (0.03)	1.00 (0.00)	0.00 (0.00)
ℓ_1-BJ	7.76 (0.32)	2.27 (0.04)	0.24 (0.02)	0.24 (0.02)	8.79 (0.39)	2.42 (0.05)	0.17 (0.01)	0.34 (0.03)
Imputed-ℓ_1-BJ	4.78 (0.16)	2.08 (0.04)	0.21 (0.01)	0.11 (0.01)	5.41 (0.28)	2.19 (0.05)	0.19 (0.01)	0.18 (0.02)
Boosting-BJ	10.36 (0.36)	2.56 (0.04)	0.06 (0.00)	0.40 (0.02)	11.46 (0.53)	2.72 (0.06)	0.06 (0.00)	0.43 (0.02)
Imputed-Boosting-BJ	6.47 (0.19)	2.07 (0.03)	**0.05**(0.00)	0.17 (0.02)	7.14 (0.29)	2.17 (0.05)	**0.04**(0.00)	0.24 (0.02)
DISCOM-BJ	**4.48** (0.14)	**1.90** (0.04)	0.22 (0.02)	**0.08** (0.01)	**5.05** (0.23)	**2.03** (0.05)	0.21 (0.02)	**0.12**(0.01)
	Example 3(e) [$\mathbb{P}(T > C) = 0.6$]				**Example 3(f)** [$\mathbb{P}(T > C) = 0.7$]			
ℓ_2-BJ	14.36 (0.34)	2.75 (0.03)	1.00 (0.00)	0.00 (0.00)	15.76 (0.36)	2.83 (0.03)	1.00 (0.00)	0.00 (0.00)
Imputed-ℓ_2-BJ	8.95 (0.26)	2.44 (0.03)	1.00 (0.00)	0.00 (0.00)	10.82 (0.30)	2.61 (0.03)	1.00 (0.00)	0.00 (0.00)
ℓ_1-BJ	10.62 (0.39)	2.54 (0.04)	0.20 (0.02)	0.37 (0.03)	12.09 (0.40)	2.61 (0.03)	0.21 (0.03)	0.44 (0.02)
Imputed-ℓ_1-BJ	6.36 (0.31)	2.31 (0.05)	0.18 (0.01)	0.22 (0.02)	7.42 (0.31)	2.39 (0.05)	0.18 (0.01)	0.28 (0.02)
Boosting-BJ	13.95 (0.55)	2.99 (0.05)	0.06 (0.00)	0.52 (0.02)	17.29 (0.61)	3.36 (0.05)	0.06 (0.00)	0.59 (0.02)
Imputed-Boosting-BJ	8.81 (0.32)	2.39 (0.05)	**0.04** (0.00)	0.30 (0.02)	11.39 (0.33)	2.74 (0.04)	**0.04** (0.00)	0.36 (0.02)
DISCOM-BJ	**5.84** (0.29)	**2.16** (0.05)	0.22 (0.02)	**0.17**(0.02)	**6.83** (0.28)	**2.24** (0.04)	0.36 (0.04)	**0.12**(0.02)

4 | APPLICATION TO THE ADNI STUDY

We apply the DISCOM-BJ to the ADNI study (Mueller et al., 2005) and compare it with several other approaches. A primary goal of this analysis is to identify biological markers and neuropsychological assessments to measure the progression of MCI and early AD. We are interested in predicting the time to convert to state AD of patients who was initially diagnosed as MCI in the ADNI study. We extract biomarkers from three complementary data sources: MRI, PET, and CSF. Note that, as Xue and Qu (2021) stated, our sparsity assumption of the proposed method may not be suitable for raw imaging data or imaging data at small scales since images have to show some visible atrophy for AD. However, the sparsity assumption can still be reasonable for the region of interest (ROI) level data. Thus, we apply the DISCOM-BJ to the ROI level data in ADNI.

We process the image data following the similar procedure as in Yu et al. (2020). For the MRI, after correction, spatial segmentation and registration steps, we obtain the image for each subject based on the Jacob template with 93 manually labeled ROIs. For each of the 93 ROIs in the labeled MRI, we compute the volume of gray matter as a feature. For each PET image, we first align the PET image to its respective MRI image using affine registration. Then, we calculate the average intensity of every ROI in the PET image as a feature. For the CSF modality, five biomarkers are used in this study, namely, amyloid $\beta(A\beta 42)$, CSF total tau (t-tau), tau hyperphosphorylated at threonine 181 (p-tau), and two tau ratios with respect to $A\beta 42$ (i.e., t-tau/$A\beta 42$ and p-tau/$A\beta 42$)

After data processing, we have 93 features from MRI, 93 features from PET, and five features from CSF. There are 376 subjects in total, including 56 subjects with complete MRI, PET, and CSF features and uncensored response, 38 subjects with complete MRI, PET, and CSF features and censored response, 101 subjects with MRI and PET features only, 89 subjects with MRI and CSF features only, and 92 subjects with MRI features only.

In our analysis, we divide the data into training, tuning, and testing sets. The training set consists of all subjects with incomplete observations and 40 randomly selected subjects with complete features. The tuning set consists of another 18 randomly selected subjects with complete observations. The testing set contains the remaining 36 subjects with complete observations. We train our model with different tuning parameters on the training set. Then we choose the tuning parameters that minimize the mean squared error on the tuning set. The testing set is used to evaluate different methods. We used all methods shown in the simulation study to predict the conversion time from MCI to AD. For each method, the analysis is repeated 50 times using different partitions of the data. In addition to compare the sum of MSE of all three responses, we also examine the top features selected by our method.

The results in Table 4 show that our proposed DISCOM-BJ method acquires the best prediction performance with smaller MSE than ℓ_1-BJ, Imputed-ℓ_1-BJ, ℓ_2-BJ, and Imputed-ℓ_2-BJ. To further understand our results, since each MRI and PET features correspond to one ROI, we can examine whether the selected features are meaningful by studying their corresponding brain regions. Table 5 shows the names of top eight features selected by our method, where the first five features are ROIs, and the last three features correspond to the CSF modality. Figure 1 shows these five ROIs of the brain. Among these five brain regions, some regions such as uncus left, middle temporal gyrus left, and hippocampus formation left are known to be highly correlated with AD and MCI by many studies using group comparison methods (Misra et al., 2009; Zhang et al., 2012). It would be interesting to study whether the other two brain regions (middle temporal gyrus right and angular gyrus left) are truly related to the conversion from MCI to AD.

TABLE 4 Performance comparison for the ADNI data. The values in the parentheses are the standard errors of the measures

method	ℓ_2-BJ	Imputed-ℓ_2-BJ	ℓ_1-BJ	Imputed-ℓ_1-BJ	DISCOM-BJ
MSE	0.99(0.04)	1.01(0.04)	0.88(0.03)	0.88(0.04)	**0.84(0.02)**

Note: Bold numbers indicate the best results in this ADNI study.

TABLE 5 Top 8 features selected by DISCOM-BJ

Top features selected by DISCOM-BJ
Uncus left
Hippocampal formation left
Middle temporal gyrus right;
Precuneus left;
Angular gyrus left;
amyloid β (Aβ 42)
CSF total tau(t-tau);
tau hyperphosphorylated at threonine 181

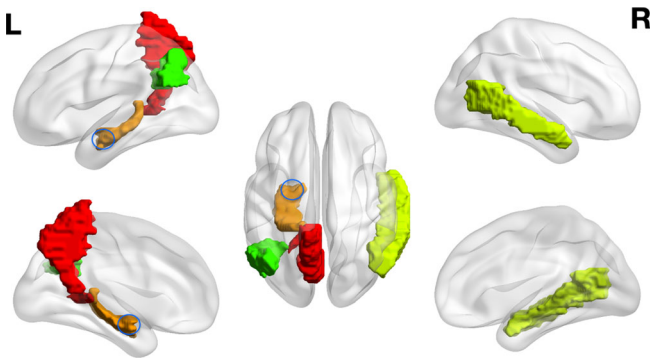

FIGURE 1 Top 5 brain regions selected by DISCOM-BJ, where the uncus left region is highlighted by the blue circle

5 | CONCLUSION

In this paper, we propose an ℓ_1-penalized Buckley–James method using block-missing multimodal predictors and censored responses. In each iteration of Buckley–James method, with pseudo responses, we first estimate the covariance matrix of the predictors using a linear combination of the estimates of the variance of each predictor, the intramodality covariance matrix, and the cross-modality covariance matrix. The proposed estimator of the covariance matrix can be positive semidefinite and more accurate than the sample covariance matrix. In the second step of each iteration, based on the estimated covariance matrix, a penalized estimator is used to deliver a sparse estimate of the coefficients. Extensive simulation studies also indicate that our method has promising performance in estimation, prediction, and model selection for the block-missing multimodal data. Finally, we apply the DISCOM-BJ method to the ADNI dataset to predict the conversion time of the patients from MCI to AD. We demonstrate that our model has accurate prediction and meaningful interpretation.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Yufeng Liu https://orcid.org/0000-0002-1686-0545

REFERENCES

Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika*, *66*(3), 429–436.

Cox, D.R., & Oakes, D. (2018). *Analysis of survival data*: Chapman and Hall/CRC.

Datta, S., Le-Rademacher, J., & Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics*, *63*(1), 259–271.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., & Collins, D.L. (2013). Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *Neuroimage*, *65*, 511–521.

Fisher, T.J., & Sun, X. (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics and Data Analysis*, *55*(5), 1909–1918.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R.C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., & Chertkow, H. (2006). Mild cognitive impairment. *The Lancet*, *367*(9518), 1262–1270.

Gui, J., & Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, *21*(13), 3001–3008.

Hsu, C.-H., & Yu, M. (2019). Cox regression analysis with missing covariates via nonparametric multiple imputation. *Statistical Methods in Medical Research*, *28*(6), 1676–1688.

Jack Jr, C.R. (2012). Alzheimer disease: New concepts on its neurobiology and the clinical role imaging will play. *Radiology*, *263*(2), 344–361.

Johnson, B.A. (2009). On lasso for censored data. *Electronic Journal of Statistics*, *3*, 485–506.

Kalbfleisch, J.D., & Prentice, R.L. (2011). *The statistical analysis of failure time data*: John Wiley & Sons.

Kaplan, E.L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Knopman, D.S., Boeve, B.F., & Petersen, R.C. (2003). Essentials of the proper diagnoses of mild cognitive impairment, dementia, and major subtypes of dementia. In *Mayo Clinic Proceedings*, *78*, Elsevier, pp. 1290–1308.

Koul, H., Susarla, V., & Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, *9*(6), 1276–1288.

Lai, T.L., & Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics*, *19*(2), 531–556.

Liu, K., Chen, K., Yao, L., & Guo, X. (2017). Prediction of mild cognitive impairment conversion using a combination of independent component analysis and the cox model. *Frontiers in Human Neuroscience*, *11*, 33.

Luo, X., Tsai, W.Y., & Xu, Q. (2009). Pseudo-partial likelihood estimators for the Cox regression model with missing covariates. *Biometrika*, *96*(3), 617–633.
Manly, J.J., Tang, M.-X., Schupf, N., Stern, Y., Vonsattel, J.-P.G., & Mayeux, R. (2008). Frequency and course of mild cognitive impairment in a multiethnic community. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *63*(4), 494–506.
Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, *11*, 2287–2322.
Miller, R., & Halpern, J. (1982). Regression with censored data. *Biometrika*, *69*(3), 521–531.
Misra, C., Fan, Y., & Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to ad. *Neuroimage*, *44*(4), 1415–1422.
Mitchell, A.J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica*, *119*(4), 252–265.
Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., & Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, *15*(4), 869–877.
Nan, B., Kalbfleisch, J.D., & Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *The Annals of Statistics*, *37*(5a), 2351–2376.
Ning, J., Qin, J., & Shen, Y. (2011). Buckley–james-type estimator with right-censored and length-biased data. *Biometrics*, *67*(4), 1369–1378.
Oulhaj, A., Wilcock, G.K., Smith, A.D., & de Jager, C.A. (2009). Predicting the time of conversion to MCI in the elderly: Role of verbal expression and learning. *Neurology*, *73*(18), 1436–1442.
Paik, M.C., & Tsai, W.-Y. (1997). On using the Cox proportional hazards model with missing covariates. *Biometrika*, *84*(3), 579–593.
Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and trends® in Optimization*, *1*(3), 127–239.
Park, M., & Moon, W.-J. (2016). Structural mr imaging in the diagnosis of Alzheimer's disease and other neurodegenerative dementia: Current imaging approach and future perspectives. *Korean Journal of Radiology*, *17*(6), 827–845.
Prentice, R.L. (1978). Linear rank tests with right censored data. *Biometrika*, *65*(1), 167–179.
Qi, L., Wang, C.Y., & Prentice, R.L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, *100*(472), 1250–1263.
Qi, L., Wang, Y.-F., Chen, R., Siddique, J., Robbins, J., & He, Y. (2018). Strategies for imputing missing covariates in accelerated failure time models. *Statistics in Medicine*, *37*(24), 3417–3436.
Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, *9*(3), 439–455.
Rubin, D.B. (2004). *Multiple imputation for nonresponse in surveys*, Vol. 81: John Wiley & Sons.
Smith, P.J. (2017). *Analysis of failure and survival data*: Chapman and Hall/CRC.
Steingrimsson, J.A., & Strawderman, R.L. (2017). Estimation in the semiparametric accelerated failure time model with missing covariates: Improving efficiency through augmentation. *Journal of the American Statistical Association*, *112*(519), 1221–1235.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.
Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*(4), 385–395.
Tsiatis, A.A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, *18*(1), 354–372.
Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., & Jack, C.R. (2009). MRI and CSF biomarkers in normal, MCI, and AD subjects: Predicting future clinical change. *Neurology*, *73*(4), 294–301.
Wang, C.Y., & Chen, H.Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics*, *57*(2), 414–419.
Wang, S., Nan, B., Zhu, J., & Beer, D.G. (2008). Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics*, *64*(1), 132–140.
Wang, Z., & Wang, C.Y. (2010). Buckley–James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*, *9*(1), 1–33.
White, I.R., & Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, *28*(15), 1982–1998.
Xu, Q., Paik, M.C., Luo, X., & Tsai, W.-Y. (2009). Reweighting estimators for Cox regression with missing covariates. *Journal of the American Statistical Association*, *104*(487), 1155–1167.
Xue, F., & Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, *116*(536), 1914–1927.
Yu, G., Li, Q., Shen, D., & Liu, Y. (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association*, *115*(531), 1406–1419.
Yu, M. (2011). Buckley–James type estimator for censored data with covariates missing by design. *Scandinavian Journal of Statistics*, *38*(2), 252–267.
Zhang, D., Shen, D., & Initiative, A. D. N. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PloS ONE*, *7*(3), e33182.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, *67*(2), 301–320.

How to cite this article: Wang, H., Li, Q., & Liu, Y. (2022). Regularized Buckley–James method for right-censored outcomes with block-missing multimodal covariates. *Stat*, *11*(1), e515. https://doi.org/10.1002/sta4.515