High-dimensional censored MIDAS logistic regression for corporate survival forecasting

Wei Miao[†], Jad Beyhum[‡], Jonas Striaukas[*] and Ingrid Van Keilegom[†]

[†]ORSTAT, KU Leuven

[‡]Department of Economics, KU Leuven

[*]Department of Finance, Copenhagen Business School

Abstract

This paper addresses the challenge of forecasting corporate distress, a problem marked by three key statistical hurdles: (i) right censoring, (ii) high-dimensional predictors, and (iii) mixed-frequency data. To overcome these complexities, we introduce a novel high-dimensional censored MIDAS (Mixed Data Sampling) logistic regression. Our approach handles censoring through inverse probability weighting and achieves accurate estimation with numerous mixed-frequency predictors by employing a sparse-group penalty. We establish finite-sample bounds for the estimation error, accounting for censoring, the MIDAS approximation error, and heavy tails. The superior performance of the method is demonstrated through Monte Carlo simulations. Finally, we present an extensive application of our methodology to predict the financial distress of Chinese-listed firms. Our novel procedure is implemented in the R package `Survivalml`.

Keywords: Corporate survival analysis; high-dimensional censored data; mixed-frequency data; logistic regression; sparse-group LASSO

1 Introduction

Regulators, lenders, and investors are increasingly focused on identifying vulnerable firms and developing accurate models to predict firm failures well in advance, as the ability to correctly predict such failures could result in a more resilient financial stability policy and better financial outcomes for market participants. As a result, an extensive body of literature is dedicated to understanding the determinants of firm failures. Traditional statistical models, such as discriminant analysis (Almon, 1965), logistic regression (Ohlson, 1980), and hazards models (Shumway, 2001), along with other time-sensitive approaches (Duffie et al., 2007), have historically been the main focus of study. However, with the advent of more extensive datasets in recent years, the focus has increasingly shifted toward machine learning methods, which are better equipped to handle high-dimensional data. Such models have shown superior accuracy in predicting firm failures (Barboza et al., 2017) due to their efficient handling of rich data sources. Over time, the task of forecasting corporate survival has gained significant attention due to its critical economic implications and its close connection to other challenges, such as predicting household loan defaults.

In this paper, we focus on the task of predicting the probability that a firm will fail within the first t years after its initial listing, conditional on its survival for the first s years, where $s < t$. This problem presents three significant statistical challenges. First, data are often right-censored, meaning that for some firms, we only know that they have survived up to a time s' where $s < s' < t$. Second, the high dimensionality of the predictors adds complexity. Modern data sets provide a wealth of variables for each listed firm, increasing the analytical burden. Third, the mixed-frequency nature of the data compounds the difficulty. For each potential predictor, we observe numerous lags, exacerbating the challenge of managing the proliferation of parameters.

As highlighted in our review of the literature below, in our view, the existing methods for this prediction task do not adequately address all three challenges. To bridge this gap, we propose a novel high-dimensional censored MIDAS logistic regression method that addresses these complexities. Our approach is based on a high-dimensional logistic regression framework to estimate survival probabilities. To address right-censoring, we make use of a tool from the survival analysis literature called outcome-weighted inverse probability of censoring weighting, as described in Blanche et al. (2023). The mixed-frequency nature of the data is managed using mixed data sampling (MIDAS), an approach developed and popularized by Ghysels et al. (2007). This method approximates the coefficients of the lags of each variable using a finite-dimensional series basis, known as the dictionary. Finally, to handle the high dimensionality of the predictors, we apply a sparse-group LASSO penalty. This penalty not only manages the dimensionality of the regressors but also accounts for the group structure of the predictors, which corresponds to the lags of the original variables, as discussed in Babii et al. (2022).

We derive finite-sample bounds on the estimation error of our estimator. Notably, these bounds allow for heavy-tailed variables, and account for both the approximation error and right-censoring, which are novel contributions to the literature on high-dimensional logistic regression models. The finite-sample performance of our method is evaluated through simulations, demonstrating its robustness against natural alternatives. Furthermore, we showcase the practical advantages of our approach through an application to forecasting the financial distress of Chinese listed firms. In this context, our method significantly outperforms the standard logistic regression benchmark and other competing methods over several horizons, underscoring its empirical effectiveness. Several practical augmented prediction methods, including oversampling and incorporating macro data into the model, are utilized. To further demonstrate the effectiveness of the proposed method that includes censoring information, a comparison is conducted with a method that excludes censored firms. Finally, our novel approach is implemented in the R package `Survivalml` to make it readily accessible for practitioners.[1]

Literature review. Let us first review how the existing methods address the three challenges we described, which are inherent in corporate survival forecasting. This review will stress the advantages of our methodology over popular alternatives. Given the extensive literature on this topic, we do not aim to provide an exhaustive review. Instead, we focus on surveying key approaches to the problems at hand. We also cite papers on the related problem of forecasting loan default.

To address the right-censoring of data, many studies restrict their analysis to firms that were first listed more than t years before the end of the follow-up period (see, for instance, Audrino et al., 2019; Petropoulos et al., 2020). Under the classical assumption of independent censoring, this approach avoids selection bias. However, it discards data on firms listed less than t years ago, leading to a loss of efficiency. Another common strategy is to directly model the hazard rate of firm failure, using methods such as Cox models or single-index models (e.g., Ding et al., 2012; Lee, 2014; Kim et al., 2016; Zhou et al., 2022; Li et al., 2023). Although effective in some contexts, this approach has limitations. Typically, the primary interest lies in estimating the probability of failure, not the hazard rate, and, therefore, modeling the survival probability as we do is more natural to solve the problem at hand. Furthermore, none of the aforementioned hazard-based approaches explicitly account for the challenges posed by high-dimensional mixed-frequency data. Instead, they applied their methods to pre-selected low-dimensional sets of predictors and lags, bypassing the complexity of high-dimensional data structures.

Let us now address the challenge of parameter proliferation, which arises from both the

[1] The package is publicly available at `https://github.com/Wei-M-Wei/Survivalml`.

high-dimensionality and the mixed-frequency nature of the data. Several studies have used LASSO as a selection tool to predict corporate bankruptcy; see, for example, Petropoulos et al. (2020); Barbaglia et al. (2023). However, these studies do not address censoring, lack theoretical results, and do not utilize the MIDAS framework. The application of MIDAS in a logistic regression framework for corporate bankruptcy prediction was explored by Audrino et al. (2019). While their work incorporates the MIDAS approach, it is limited to a low-dimensional set of predictors and does not consider censoring. More closely related to our study, Jiang et al. (2021) examined a penalized logistic regression with the norm ℓ_1. However, their approach does not account for censoring, lacks theoretical underpinnings, and employs what is referred to as unrestricted MIDAS. Unlike our approach, which uses a restricted MIDAS procedure, unrestricted MIDAS includes all lags as predictors, effectively bypassing the dimension reduction benefits of the MIDAS framework.[2]

Finally, we compare our theoretical results to the existing literature. The theory of penalized estimators of the high-dimensional logistic regression model has been extensively studied under various situations. As already mentioned, no existing study allows for censoring or approximation error. Van De Geer (2008); Meier et al. (2008); Van De Geer (2008); Bühlmann and Van De Geer (2011); Van De Geer (2016) analyzed these models using fixed design or isotropy conditions of covariates, which are often unsuitable for financial econometric data. More recently, Caner (2023) relaxed these assumptions, allowing for random covariates designs with non-normal covariates in the context of penalized Generalized Linear Models (GLM). However, compared to the present paper, this work imposes additional assumptions on the shape of the second-order partial derivatives of the loss function. Similarly, Han et al. (2023) developed the theory for GLM with LASSO by establishing local restricted strong convexity of the loss function, which is related to the quadratic margin condition in the present paper; see Appendix B.3. In the context of mixed-frequency data, Babii et al. (2022, 2023) developed the theoretical foundation for high-dimensional time series and panel data linear regression models while accounting for the MIDAS approximation error. However, the theory for logistic regression models incorporating such approximation errors remains unexplored and none of the aforementioned studies have addressed the challenges posed by censored data.

Outline. The paper is organized as follows. In Section 2, we first present the model, followed by a discussion on employing the MIDAS weighting technique and incorporating group structure information among variables. Section 3 is dedicated to the analysis of the estimation properties of the proposed estimator. Section 4 presents the results of the simulation

[2] It is worth noting that the term "unrestricted MIDAS" is somewhat misleading, as this approach directly incorporates all lags as independent variables. Consequently, it does not take advantage of the dimension-reduction capabilities inherent in the MIDAS methodology.

studies. In Section 5, we construct a dataset on Chinese firm distress and assess the prediction performance of the proposed methods in the real dataset, a comparison with several other approaches is included.

Notation. For $\ell \in \mathbb{N}$, we define $[\ell] = \{1, 2, \ldots, \ell\}$. For a vector $\boldsymbol{b} \in \mathbb{R}^p$, its ℓ_q norm is denoted as $|\boldsymbol{b}|_q = \left(\sum_{j\in[p]} |b_j|^q\right)^{1/q}$ if $q \in [1, \infty)$. For a matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ be its smallest eigenvalue. For a vector $\Delta \in \mathbb{R}^p$ and a subset $J \subset [p]$, let Δ_J be a vector in $\mathbb{R}^p$ with the same coordinates as Δ on J and zero coordinates on J^c, where J^c is the complement of the subset J. The cardinality of a set S is $|S|$. For $a, b \in \mathbb{R}$, we put $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Lastly, we write $a_N \lesssim b_N$ if there exists a (sufficiently large) absolute constant v such that $a_N \leq v b_N$ for all $N \geq 1$. The indicator function is denoted by $\mathbb{1}\{\cdot\}$.

2 High-dimensional censored MIDAS logistic regression

2.1 Logistic regression model

In corporate survival analysis, we focus on the survival time T of a firm. The random variable T represents the duration from the firm's Initial Public Offering (IPO) date to the occurrence of financial distress. Specifically, the IPO date refers to the first day the firm's stock is publicly traded. Since companies are not listed immediately after their creation, the survival time T in our context differs slightly from the typical survival time considered in traditional survival analysis (Li et al., 2023).

Our main objective is to predict the probability that a firm will survive up to t years, given that it has already been publicly listed for s years. In practice, the survival time T is right-censored by the censoring time C, which denotes the duration between the IPO date and the censoring event, occurring at the end of the follow-up period. Hence, we do not directly observe T, but rather the censored value $\widetilde{T} = T \wedge C$, with the indicator $\delta = \mathbb{1}\{T \leq C\}$.

The financial distress status, indicated by $\mathbb{1}\{T \leq t\}$, is influenced by covariates $\boldsymbol{Z} \in \mathbb{R}^{K_z}$. We assume, for the moment, that $\boldsymbol{Z}$ has finite variance to ensure well-defined expectations and model the survival indicator $\mathbb{1}\{T \leq t\}$ using a logistic regression model:

$$P(T \leq t \mid \boldsymbol{Z}, T \geq s) = \frac{\exp\left(\boldsymbol{Z}^\top \boldsymbol{\theta}_0(t,s)\right)}{1 + \exp\left(\boldsymbol{Z}^\top \boldsymbol{\theta}_0(t,s)\right)}, \tag{1}$$

where $\boldsymbol{Z}$ is the covariate vector, and $\boldsymbol{\theta}_0(t,s) \in \mathbb{R}^{K_z}$ is the vector of true parameters specific to t and s. For convenience, we use $\boldsymbol{\theta}_0$ as shorthand for $\boldsymbol{\theta}_0(t,s)$. Note that in practice, the model includes an intercept term, which enters the variable $\boldsymbol{Z}$.

Model (1) is typically estimated via maximum likelihood estimation. This method is based on the characterization of $\boldsymbol{\theta}_0$ as the solution to the population conditional maximum likelihood problem:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\mathbb{R}^{K_z}}{\arg\max}\, \mathbb{E}\left[\mathbb{1}\{T \leq t\}\boldsymbol{Z}^\top\boldsymbol{\theta} - \log\left(1+\exp(\boldsymbol{Z}^\top\boldsymbol{\theta})\right)\middle| T \geq s\right]. \tag{2}$$

In (2), we have rewritten the classical logistic model's likelihood in a simplified form; see Lemma A.1 for a proof.[3]

However, equation (2) cannot be directly used for estimation due to the fact that the survival time T is not always observed. To address this issue, we apply the outcome-weighted inverse probability of censoring weighting (OIPCW) method, as outlined by Blanche et al. (2023).[4] This method relies on two standard assumptions about the censoring mechanism, which we describe below. The first assumption is the assumption of independent censoring:

Assumption 2.1. *C is independent of T and $\boldsymbol{Z}$.*

This is a standard assumption in survival analysis. We argue that Assumption 2.1 is reasonable in corporate survival analysis because the censoring time for a firm is solely determined by the observation period, with no firms censored before. The second assumption concerns sufficient follow-up:

Assumption 2.2. $P\left(\widetilde{T} \geq t\right) > 0$.

This assumption implies that some firms have been observed for more than t years without experiencing financial distress, which is necessary for model estimation.

Under Assumptions 2.1 and 2.2, we obtain an alternative characterization of $\boldsymbol{\theta}_0$, relying only on observed or estimable quantities:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\mathbb{R}^{K_z}}{\arg\max}\, \mathbb{E}\left[\frac{\delta(t)\mathbb{1}\{\widetilde{T} \leq t\}}{H(t\wedge\widetilde{T})}\boldsymbol{Z}^\top\boldsymbol{\theta} - \log\left(1+\exp(\boldsymbol{Z}^\top\boldsymbol{\theta})\right)\middle| \widetilde{T} \geq s\right], \tag{3}$$

where $H(u) = P(C \geq u | C \geq s)$ is the survival probability of C at time u conditional on $C \geq s$ and $\delta(t) = \mathbb{1}\{C \geq t \wedge T\} = 1 - \mathbb{1}\{\widetilde{T} \leq t\}\delta$ is the observation indicator. Equation (3) is proven in Appendix A, see Lemma A.2. Essentially, the expectation in (3) weighs the uncensored observations that fail between s and t by the weights $1/H(t\wedge\widetilde{T})$ to ensure they are representative of firms with survival times between s and t.[5]

The function H is not directly observed but can be estimated under Assumption 2.1 using the classical Kaplan-Meier estimator (Kaplan and Meier, 1958), as described below.

[3] The characterization (2) is valid under a full-rank condition stated in Lemma A.1.

[4] An alternative approach for addressing censoring is Inverse Probability Weighting (IPW) (Horvitz and Thompson, 1952; Zheng et al., 2006; Beyhum et al., 2024b,a). Further details on both OIPCW and IPW can be found in Blanche et al. (2023).

[5] The expectation in (3) is well-defined since $H(t\wedge\widetilde{T}) \geq H(t) = P(C \geq t | C \geq s) = P\left(C \geq t\right)/P(C \geq s) \geq P(\widetilde{T} \geq t)/P(C \geq s) > 0$ almost surely by Assumption 2.2.

2.2 Estimation with mixed-frequency data

Consider an i.i.d. sample of firms $(\widetilde{T}_i, \delta_i, \boldsymbol{Z}_i), i \in [N]$, such that for all $i \in [N]$, $\widetilde{T}_i \geq s$, i.e., all firms in the sample are observed for at least s years.[6]

For prediction, we use s years of lagged covariates $\left\{x_{i,s-\frac{j-1}{m}} \in \mathbb{R}^K, j \in [d]\right\}$, where $d = s \times m$ represents the total number of lags, and m is the frequency of observation. The covariates can be observed at varying frequencies, and although not all lags may enter the regression, we omit such cases for simplicity. The k^{th} covariate and its lags are represented as $\widetilde{\boldsymbol{Z}}_{i,k} = \left(x_{i,s,k}, x_{i,s-\frac{1}{m},k}, \ldots, x_{i,s-\frac{d-1}{m},k}\right)^\top$, $k \in [K]$. The complete vector of lagged covariates and intercept is denoted as $\boldsymbol{Z}_i = \left(1, \widetilde{\boldsymbol{Z}}_{i,1}^\top, \widetilde{\boldsymbol{Z}}_{i,2}^\top, \ldots, \widetilde{\boldsymbol{Z}}_{i,K}^\top\right)^\top \in \mathbb{R}^{K_z}$ with $K_z = K \times d + 1$.

The function H can be estimated using the Kaplan-Meier estimator:

$$\widehat{H}(u) = \prod_{j \leq u} \left(1 - \frac{dN(j)}{\widetilde{T}(j)}\right),$$

where $N(j) = \sum_{i=1}^N \mathbb{1}\{\widetilde{T}_i \leq j, \delta_i(j) = 0\}$ is the number of units at risk at time j, and $dN(j) = N(j) - \lim_{j' \to j, j' < j} N(j')$ denotes the jump of the process N at time j. Additionally, $\widetilde{T}(j) = \sum_{i=1}^N \mathbb{1}(\widetilde{T}_i \geq j)$ is the number of firms known to be at risk at time j.

We consider datasets that are high-dimensional. For instance, in our empirical application, as summarized in Table 3, if we consider firms that have survived $s = 6$ years, with $K = 95$ covariates measured $m = 4$ times per year, the total number of parameters to estimate is $6 \times 4 \times 95 + 1 = 2,281$, including the intercept. When the sample size is not much larger than the number of parameters, the curse of dimensionality arises, complicating computations and reducing estimation precision.

To address this, dimension-reduction techniques are necessary. A common approach is to directly apply the LASSO (Tibshirani, 1996) to the original predictors. For an i.i.d. sample $\{(\widetilde{T}_i, \delta_i, \boldsymbol{Z}_i), i \in [N]\}$, the ℓ_1-norm penalized estimator minimizes:

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{K_z}}{\arg\min} \frac{1}{N} \sum_{i=1}^N \left(-\frac{\delta_i(t)\mathbb{1}\{\widetilde{T}_i \leq t\}}{\widehat{H}(t \wedge \widetilde{T}_i)} \boldsymbol{Z}_i^\top \boldsymbol{\theta} + \log\left(1 + \exp(\boldsymbol{Z}_i^\top \boldsymbol{\theta})\right)\right) + \mu |\boldsymbol{\theta}|_1, \qquad (4)$$

where μ is a regularization parameter.[7]

Here, we follow a different approach to reducing the dimension based on Mixed-Data Sampling (Ghysels et al., 2006, MIDAS), which is designed to address parameter proliferation in mixed-frequency data. MIDAS approximates the coefficients of high-frequency lag

[6] As shown in the previous section, considering firms with at least s years of observation does not introduce selection bias under the independent censoring assumption.

[7] In our practical implementation, we never penalize the intercept coefficient.

polynomials using a finite dictionary of functions. Specifically, let us write

$$\boldsymbol{Z}_i^\top \boldsymbol{\theta}_0 = \boldsymbol{\theta}_{0,1} + \frac{1}{d}\sum_{k=1}^{K}\sum_{j=1}^{d} \omega_k\left(\frac{j-1}{d}\right) x_{i,s-\frac{j-1}{m},k}, \quad i \in [N], \tag{5}$$

where $\omega_k : [0,1] \mapsto \mathbb{R}$, $k \in [K]$, are weight functions for the lag polynomials such that $\omega_k\left(\frac{j-1}{d}\right)/d = \boldsymbol{\theta}_{0,1+d(k-1)+j}$. Let $\{w_l : l = 1, \dots, L\}$ be the dictionary of functions. For each $k \in [K]$, we assume there exist coefficients $\boldsymbol{\beta}^*_{0,k} = (\boldsymbol{\beta}^*_{0,k,1}, \boldsymbol{\beta}^*_{0,k,2}, \dots, \boldsymbol{\beta}^*_{0,k,L})^\top \in \mathbb{R}^L$ such that:

$$\omega_k(u) \approx \sum_{l=1}^{L} \boldsymbol{\beta}^*_{0,k,l} w_l(u), \quad u \in [0,1].$$

This reduces the number of parameters from $K \times d + 1$ to $K \times L + 1$. MIDAS approaches perform well in various contexts (Ghysels et al., 2020) and often outperform the LASSO estimator in (4). The simplest dictionary consists of algebraic power polynomials (e.g., Almon polynomials (Almon, 1965)), but other orthogonal bases of $L_2[0,1]$ can be used to improve performance with correlated covariates.[8]

To further enhance dimension reduction, we apply sparse-group LASSO (Simon et al., 2013), which incorporates group structures among covariates. Unlike group LASSO (Yuan and Lin, 2006), which enforces sparsity between groups, sparse-group LASSO encourages sparsity both within and between groups. Let:

$$\boldsymbol{X}_i = \left(1, \widetilde{\boldsymbol{Z}}_{i,1}^\top W, \widetilde{\boldsymbol{Z}}_{i,2}^\top W, \dots, \widetilde{\boldsymbol{Z}}_{i,K}^\top W\right)^\top \in \mathbb{R}^{KL+1},$$

where $W = \left(w_l\left(\frac{j-1}{d}\right)/d\right)_{j\in[d], l\in[L]}$ is a $d \times L$ weighting matrix. Define $\boldsymbol{\beta} \in \mathbb{R}^{KL+1}$, and let the penalty be

$$\Omega(\boldsymbol{\beta}) = \alpha|\boldsymbol{\beta}|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_{2,1}, \quad \|\boldsymbol{\beta}\|_{2,1} = \sum_{G\in\mathcal{G}} |\boldsymbol{\beta}_G|_2,$$

with $\mathcal{G}$ representing the group structure. The sparse-group LASSO estimator minimizes

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{KL+1}}{\arg\min}\, R_N(\boldsymbol{\beta}) + \lambda\Omega(\boldsymbol{\beta}), \tag{6}$$

where

$$R_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} -\frac{\delta_i(t)\mathbb{1}\{\widetilde{T}_i \leq t\}}{H(t \wedge \widetilde{T}_i)} \boldsymbol{X}_i^\top \boldsymbol{\beta} + \log\left(1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})\right).$$

Here, $\lambda \geq 0$ controls the regularization, and $\alpha \in [0,1]$ balances the sparsity and group

[8] $L_2[0,1]$ denotes the space of square-integrable functions $f : [0,1] \to \mathbb{R}$.

penalties. Choosing $\alpha = 1$ recovers LASSO, while $\alpha = 0$ corresponds to group LASSO.[9] In our analysis, we adopt the simplest group structure $\mathcal{G} = \{G_k : k \in [K+1]\}$, where $G_1 = \{1\}$ and $G_k = \{(1 + (k-2)L) + 1, \ldots, 1 + (k-1)L\}$ corresponds to parameters for the same high-frequency covariates. As shown by Babii et al. (2022), this structure enhances prediction performance. Note that alternative groupings, such as pairing related covariates like Return on Assets (ROA) and Return on Equity (ROE), could also be considered. We call **sg-LASSO-MIDAS** the approach embodied by equation (6).

In the empirical sections, we also examine two alternative methods as benchmarks. The first method **LASSO-UMIDAS**, from equation (4) employs unrestricted lag polynomials combined with LASSO. Unlike the MIDAS approach, LASSO-UMIDAS does not impose restrictions on the polynomials, resulting in the need to estimate a significantly larger number of parameters. Moreover, no structural constraints are applied to the coefficients of the lags. As a second alternative, we consider **LASSO-MIDAS**, which adopts the MIDAS approximation employed in the sg-LASSO-MIDAS method but avoids the group penalty. It corresponds to equation (6) with a fixed mixing parameter of $\alpha = 1$.

3 Theoretical results

In this section, we outline the main assumptions for the proposed estimator (6) and analyze theoretically the finite sample properties of the sparse-group LASSO estimator within the context of censored data. Both the LASSO and the group LASSO estimators are covered as special cases.[10] Recall that we have an i.i.d. sample $\{(\widetilde{T}_i, \delta_i, \boldsymbol{Z}_i),\ i \in [N]\}$ such that $\widetilde{T}_i \geq s$ for all $i \in [N]$. We focus on the estimator $\widehat{\boldsymbol{\beta}}$ in our theoretical analysis. We consider an asymptotic regime where N goes to infinity and $p = KL + 1$ goes to infinity as a function of N. High-dimensional ℓ_1-norm penalized logistic regression has been studied in the literature, see, for instance, Van De Geer (2008, 2016); Caner (2023) and Han et al. (2023). However, none of these studies account for censoring nor allow for approximation errors. Instead, we explicitly take into account the approximation error stemming from the MIDAS approximation defined as

$$E_i = \boldsymbol{Z}_i^\top \boldsymbol{\theta}_0 - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0, \quad i \in [N],$$

where $\boldsymbol{\beta}_0 = \left(\boldsymbol{\theta}_{0,1}, (\boldsymbol{\beta}^*_{0,1})^\top, (\boldsymbol{\beta}^*_{0,2})^\top, \ldots, (\boldsymbol{\beta}^*_{0,K})^\top\right)^\top \in \mathbb{R}^p$ is the true parameter of interest. Let $\boldsymbol{E} = (E_1, E_2, \ldots, E_N)^\top$ collect all approximation errors.

We start by introducing the following assumptions.

[9] In our practical implementation, we do not penalize the intercept coefficient, however, for simplicity, we do not write this in the equations.

[10] We treat α as constant for the theory but optimize it through cross-validation in practice.

Assumption 3.1. *(Data) We have i.i.d. data* $\{(\widetilde{T}_i, \delta_i, \boldsymbol{Z}_i),\ i \in [N]\}$, *and there exists* $q \geq 4$ *and* $K_0 > 0$ *such that* $\max\limits_{|\boldsymbol{u}|_2=1} \mathbb{E}\left(\left|\boldsymbol{X}_i^\top \boldsymbol{u}\right|^q\right) \leq K_0$.

This condition just requires that the variables have more than 4 finite moments, allowing for polynomial tails commonly observed with financial variables.

Assumption 3.2. *There exists a constant* $\gamma_{\mathrm{H}} > 0$ *such that the minimum eigenvalue*

$$\lambda_{\min}\left(\mathbb{E}\left[\frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right) \geq \gamma_{\mathrm{H}}.$$

Assumption 3.2 is similar to the compatibility condition discussed in Van De Geer (2008, 2016); Caner (2023), as well as the restricted Fisher-information matrix eigenvalue condition described in Han et al. (2023). This is a high-dimensional version of the full-rank condition guaranteeing the asymptotic properties of the maximum likelihood estimator in the (low-dimensional) logistic regression with misspecification error.

Next, we need to introduce additional definitions. Let $S_{\boldsymbol{\beta}_0} = \{j \in [p] :\ \boldsymbol{\beta}_{0,j} \neq 0\}$ and $\mathcal{G}_{\boldsymbol{\beta}_0} = \{j \in [K+1] :\ (\boldsymbol{\beta}_0)_{G_j} \neq \boldsymbol{0}\}$ be the support and the group support of the target parameter $\boldsymbol{\beta}_0$. Let $s_{\boldsymbol{\beta}_0} = \alpha\sqrt{|S_{\boldsymbol{\beta}_0}|} + (1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}_0}|}$ be the sparsity level and $G^* = \max_{G \in \mathcal{G}_{\boldsymbol{\beta}_0}} |G|$ be the size of the largest group in $\mathcal{G}_{\boldsymbol{\beta}_0}$. For simplicity, we suppose that $s_{\boldsymbol{\beta}_0} \geq 1$ (otherwise, it suffices to replace $s_{\boldsymbol{\beta}_0}$ by $s_{\boldsymbol{\beta}_0} \vee 1$ in all assumptions and bounds involving $s_{\boldsymbol{\beta}_0}$). We impose the following assumption on $s_{\boldsymbol{\beta}_0}$.

Assumption 3.3. *It holds that*

$$s_{\boldsymbol{\beta}_0} G^* \left(\frac{p^{\frac{2}{q}} \log p}{N^{1-\frac{2}{q}}} \vee \frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}}\right) = o(1),$$

and

$$s_{\boldsymbol{\beta}_0} (G^*)^{\frac{3}{2}} \left(\frac{\lambda s_{\boldsymbol{\beta}_0}}{\gamma_{\mathrm{H}}} + \frac{\lambda^{-1}}{N}|\boldsymbol{E}|_1\right)(Np\log p)^{\frac{1}{q}} = o_P(1).$$

This is a condition on the degree of sparsity $s_{\boldsymbol{\beta}_0}$, the size of the largest group G^*, the ℓ_1-norm of the approximation error $|\boldsymbol{E}|_1 = \sum_{i=1}^N |E_i|$, and the relative growth rate of N and p. The condition is more likely to hold when $s_{\boldsymbol{\beta}_0}$, G^*, $|\boldsymbol{E}|_1$ or $1/q$ are smaller and p does not grow too quickly with N. Note also that if λ is too low or too large, the condition might fail to hold. This condition allows establishing a connection between empirical and population effective sparsity, enabling the extension of the quartic margin condition to its sampled version. Van De Geer and Bühlmann (2009) briefly discussed it specifically for data with Gaussian tails in the case of the LASSO for the linear model. We extend this framework to accommodate heavy-tailed data and approximation error in the logistic regression model. When there is no such approximation error, that is $E_i = 0$, for all $i \in [N]$, a similar assumption imposed on λ and $s_{\boldsymbol{\beta}_0}$ is used in Van De Geer (2016) and Han et al. (2023).

We now establish bounds on the estimation error, presenting two distinct types. The first type pertains to the parameter estimation error $\Omega\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)$, while the second focuses on prediction accuracy. Consider a scenario where, for some $\boldsymbol{z}=(1,\tilde{\boldsymbol{z}})^\top \in \mathbb{R}^{K_z}$, we aim to estimate $P(\boldsymbol{z})=P(T\leq t \mid \boldsymbol{Z}=\boldsymbol{z}, T\geq s)$, representing the probability that a firm with covariates $\boldsymbol{z}$, having survived at least s years, fails before t. We estimate $P(\boldsymbol{z})$ using $\widehat{P}(\boldsymbol{z})=\frac{\exp\left(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}\right)}$ where $\boldsymbol{x}=\left(1,\tilde{\boldsymbol{z}}^\top W\right)^\top$. Our goal is to provide a bound for the error $\widehat{P}(\boldsymbol{z})-P(\boldsymbol{z})$. This bound will depend on the term $e=\boldsymbol{z}^\top\boldsymbol{\theta}_0-\boldsymbol{x}^\top\boldsymbol{\beta}_0$, which represents the MIDAS approximation error at the covariate $\boldsymbol{z}$. The following theorem formally states this result.

Theorem 3.1. *Let Assumptions 2.1, 2.2, 3.1, 3.2 and 3.3 hold. If* $p^{\frac{1}{q}}\sqrt{\log p}/N^{\frac{1}{2}-\frac{1}{q}}=o(\lambda)$, *then, with probability going to* 1, *we have*

$$\Omega\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\lesssim\frac{\lambda s_{\boldsymbol{\beta}_0}}{\gamma_{\mathrm{H}}}+\lambda^{-1}\frac{1}{N}|\boldsymbol{E}|_1,$$

and

$$\widehat{P}(\boldsymbol{z})-P(\boldsymbol{z})\lesssim\frac{\lambda s_{\boldsymbol{\beta}_0}|\boldsymbol{x}|_\infty}{\gamma_{\mathrm{H}}}+\lambda^{-1}\frac{1}{N}|\boldsymbol{E}|_1\,|\boldsymbol{x}|_\infty+|e|.$$

Let us now discuss the theorem. First, we require that $p^{\frac{1}{q}}\sqrt{\log p}/N^{\frac{1}{2}-\frac{1}{q}}$ is negligible with respect to λ. For bounds on the LASSO under sub-Gaussian errors, it suffices that λ is of the order of $\sqrt{\log p/N}$. Our condition is stricter due to the presence of heavy-tailed variables. However, as $q\to\infty$, the variables are no longer heavy-tailed, and we recover the order $\sqrt{\log p/N}$ for λ. As is standard in the literature, in practice, we select λ in practice via cross-validation (see Sections 4 and 5). The dependence of our rates on λ aligns with those for the standard LASSO estimator in high-dimensional regression. Specifically, the estimation error of $\widehat{\boldsymbol{\beta}}$ is of the order $\lambda s_{\boldsymbol{\beta}_0}$.

Our bounds also depend on the ℓ_1-norm $|\boldsymbol{E}|_1$ of the approximation error. To the best of our knowledge, this work is the first to establish results for the high-dimensional logistic regression model with an approximation error. Such results, however, are well-established for the LASSO in linear regression (see Bickel et al., 2009). To achieve this result, we bound the difference between the empirical and population loss functions not only by terms related to the empirical process but also by a term dependent on the approximation error $\boldsymbol{E}$. Addressing this challenge is particularly difficult due to the nonlinearity of the problem. Interested readers are referred to Appendix B for the detailed proof of Theorem 3.1. Regarding the estimated prediction probability, its error bound matches the order of the parameter estimation error, with an additional term which is a function of the MIDAS approximation error e.

Finally, as a concluding remark, note that since the Kaplan-Meier estimator converges at the $\sqrt{N}$-rate, the fact that the weights $\delta_i(t)/H(t \wedge \widetilde{T}_i)$ are estimated does not affect the rate of convergence of our estimator.

4 Simulations

We evaluate the predictive performance of three methods through simulations: i) LASSO-UMIDAS, an unstructured LASSO estimator without unrestricted MIDAS weights, ii) LASSO-MIDAS, an unstructured LASSO estimator using MIDAS weights, and iii) sg-LASSO-MIDAS, a structured sparse-group LASSO estimator with MIDAS weights. The sg-LASSO-MIDAS approach, specifically, highlights the advantages of leveraging group structures and dictionaries within a high-dimensional framework, offering a compelling comparison to LASSO-MIDAS and LASSO-UMIDAS (see, e.g., Babii et al., 2022). Simulation results showcase the method's strengths, particularly in achieving superior prediction accuracy with finite sample data.

4.1 Simulation design

Let us describe the data-generating process. There are $K = 50$ high-frequency covariates, but only the first two enter the model. All the observations in the simulated dataset have survived at least s years, and we are interested in a yearly/quarterly frequency $m = 4$. We consider $s = 6$ years of lagged data.

For the generation of $x_{i,\frac{j}{m},k}, j \in [d]$, we first initiate the processes by letting $\left(x_{i,\frac{1}{m},1}, \dots, x_{i,\frac{1}{m},K}\right)^\top$ follow a $\mathcal{N}\left(\mathbf{0}, \Sigma(1-\rho^2)\right)$ distribution with $\Sigma_{u,v} = \rho_0^{|u-v|}, u, v \in [K]$. Then, the high-frequency covariates $x_{i,\frac{j}{m},k}, j \in [d], k \in [K]$ are generated according to the following scenarios:

Scenario 1: $x_{i,\frac{j}{m},k} = \rho x_{i,\frac{j-1}{m},k} + \nu_{i,k}$, $k \in [K], j \in \{2, 3, \dots, d\}$, and $(\nu_{i,1}, \dots, \nu_{i,K})^\top \sim_{\text{i.i.d}} \mathcal{N}\left(\mathbf{0}, \Sigma(1-\rho^2)\right)$ with $\rho = 0.1$ and $\rho_0 = 0.1$.

Scenario 2: $x_{i,\frac{j}{m},k} = \rho x_{i,\frac{j-1}{m},k} + \nu_{i,k}$, $k \in [K], j \in \{2, 3, \dots, d\}$, and $(\nu_{i,1}, \dots, \nu_{i,K})^\top \sim_{\text{i.i.d}} \mathcal{N}\left(\mathbf{0}, \Sigma(1-\rho^2)\right)$ with $\rho = 0.6$ and $\rho_0 = 0.1$.

In addition, we consider one more scenario that allows the covariates to have heavy tails. Similarly as before, we first initiate the processes with $\left(x_{i,\frac{1}{m},1}, \dots, x_{i,\frac{1}{m},K}\right)^\top \sim$ student-$t(2)$ with the covariance matrix $\Sigma_{u,v} = \rho_0^{|u-v|}, u, v \in [K]$. Then, the third scenario is as follows.

Scenario 3: $x_{i,\frac{j}{m},k} = \rho x_{i,\frac{j-1}{m},k} + \nu_{i,k}$, $k \in [K], j \in \{2, 3, \dots, d\}$, and $(\nu_{i,1}, \dots, \nu_{i,K})^\top \sim_{\text{i.i.d}}$ student-t with degree 2 and its covariance matrix $\Sigma(1-\rho^2)$, with $\rho = 0.1$ and $\rho_0 = 0.1$.

It is clear that ρ regulates the degree of time series dependence among lagged covariates, while ρ_0 represents the level of cross-sectional dependence across all K covariates. Finally, before generating T_i, we transform the covariates to their absolute values which ensures that the distribution functions will be increasing in t for all $\boldsymbol{Z}$.

The last step is to generate T_i. To do so, we let

$$T_i = s + \exp\left(\frac{\log(\frac{\zeta}{1-\zeta}) - \left(1 + \sum\limits_{j\in[d]} \widetilde{\omega}_1\left(\frac{j-1}{d}\right) x_{i,s-\frac{j-1}{m},1} - \sum\limits_{j\in[d]} \widetilde{\omega}_2\left(\frac{j-1}{d}\right) x_{i,s-\frac{j-1}{m},2}\right)}{1 + \sum_{k=1}^{2} \sum\limits_{j\in[d]} \widetilde{\omega}_k\left(\frac{j-1}{d}\right) x_{i,s-\frac{j-1}{m},k}}\right),$$

for all $i \in [N]$, where $\zeta \sim \text{Uniform}(0,1)$. The weighting schemes $\widetilde{\omega}_k(u), u \in [0,1]$ for $k = 1,2$ correspond to beta densities, respectively, equal to $\mathbf{Beta}(1,3)$, $\mathbf{Beta}(2,3)$, see Ghysels et al. (2007); Ghysels and Qian (2019); Babii et al. (2022), for further details. This generation scheme guarantees that the survival function of T satisfies (1), where $\boldsymbol{\theta}_0(t,s)$ is such that $\boldsymbol{\theta}_{0,1}(t,s) = 1 + \log(t-s)$, $\boldsymbol{\theta}_{0,1+j}(t,s) = (1 + \log(t-s))\widetilde{\omega}_1\left(\frac{j-1}{d}\right)$, $j \in [d]$, $\boldsymbol{\theta}_{0,1+d+j}(t,s) = (\log(t-s) - 1)\widetilde{\omega}_2\left(\frac{j-1}{d}\right)$, $j \in [d]$ and $\boldsymbol{\theta}_{0,k}(t,s) = 0$ for all $k \in \{2d+2,\ldots,K_z\}$. Remark that only the first two high-frequency covariates are relevant.

The censoring time $C_i, i \in [N]$ is generated by the shifted exponential distribution $C_i \sim s + \exp(\gamma)$, where we select γ to maintain a censoring rate $\left(\sum_{i=1}^{N} \mathbb{1}\{T_i > C_i\}\right)/N$ of approximately 81% in the simulated dataset, matching the rate observed in the real dataset (see Section 5.1).

For the choice of the MIDAS weight function W in the sg-LASSO-MIDAS, we use a dictionary comprising Gegenbauer polynomials shifted to the interval $[0,1]$, with the parameter $\alpha_{\text{poly}} = -\frac{1}{2}$, and size of $L = 3$ as specified in (5).[11] The use of such orthogonal polynomials is advantageous in practice, as they help to reduce multicollinearity and improve numerical stability; for further details on dictionaries, see Appendix A in Babii et al. (2022). Regarding t, we set it to the following percentiles $t = \{t_1 = 10\%, t_2 = 30\%, t_3 = 50\%\}$ of the set $\{T_i : T_i \text{ is uncensored}, i \in [N]\}$.

Concerning the evaluation of classification performance, Receiver Operating Characteristic (ROC) curves are widely used in the literature. However, traditional ROC curves are not fully suitable in this context due to censoring, where the status of firms is only partially observed. To address this limitation, we use the ROC curve estimator developed by Heagerty et al. (2000), which was specifically designed to evaluate classification performance effectively in the presence of censoring.

[11] Notice that we shift the basis of the Gegenbauer polynomials to the interval $[0,1]$. The parameter α_{poly} defines the type of Gegenbauer polynomials. When $\alpha_{\text{poly}} = 1$, they correspond to Legendre polynomials, and when $\alpha_{\text{poly}} = \frac{1}{2}$, they correspond to Chebyshev polynomials.

4.2 Evaluation metric: ROC curves with censoring

Recalling the definitions of sensitivity and specificity in the ROC curves, we see that in our model, both sensitivity, or the "true positive rate" (TPR), and specificity, or the "false positive rate" (FPR), are also functions that depend on t:

$$\begin{aligned} Se(c,t) &= P[\Upsilon_i > c \mid T_i \le t], \\ Sp(c,t) &= P[\Upsilon_i \le c \mid T_i > t], \end{aligned} \tag{7}$$

where $\Upsilon_i := p\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{X}_i\right) = \frac{\exp\left(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}\right)}$ is the estimated probability.[12] The threshold c is used to classify a firm as distressed if $\Upsilon_i > c$, or as non-distressed if $\Upsilon_i \le c$, with $\mathbb{1}\{T \le t\}$ indicating whether the firm has failed by time t.

A ROC curve illustrates the full range of True Positive Rates (TPR) and False Positive Rates (FPR) across all possible threshold values c. A larger area under the ROC curve (AUC) signifies better performance in distinguishing between firms that have failed and those that have not. In practice, the status $\mathbb{1}\{T \le t\}$ in (7) cannot be fully observed due to censoring. To address this issue, various ROC curve estimators have been proposed in Heagerty et al. (2000); Cai et al. (2006); Heagerty and Zheng (2005); Amico et al. (2020).

Here, we employ the Nearest Neighbor estimator (Heagerty et al., 2000) to account for the censored data and evaluate the ROC curves. Let

$$\widehat{S}_{\kappa_N}(c,t) = \frac{1}{N}\sum_{i=1}^{N} \widehat{S}_{\kappa_N}(t \mid \Upsilon_i)\, \mathbb{1}\{\Upsilon_i > c\},$$

where $\widehat{S}_{\kappa_N}(t \mid \Upsilon_i)$ is a suitable estimator of the conditional survival function characterized by a parameter κ_N:

$$\widehat{S}_{\kappa_N}(t \mid \Upsilon_i) = \prod_{a \in \mathcal{T}_N, a \le t} \left\{ 1 - \frac{\sum_j \Psi_{\kappa_N}(\Upsilon_j, \Upsilon_i)\, \mathbb{1}\{\widetilde{T}_j = a\}\delta_j}{\sum_j \Psi_{\kappa_N}(\Upsilon_j, \Upsilon_i)\, \mathbb{1}\{\widetilde{T}_j \ge a\}} \right\},$$

where $\mathcal{T}_N$ is a set of the unique values of $\widetilde{T}_i$ for observed events, $\delta_i = \mathbb{1}\{T_i \le C_i\}$ and $\Psi_{\kappa_N}(\Upsilon_j, \Upsilon_i)$ is a kernel function that depends on a smoothing parameter κ_N. Following the approach in Heagerty et al. (2000), we used a $0/1$ nearest neighbor kernel (Akritas, 1994), $\Psi_{\kappa_N}(\Upsilon_j, \Upsilon_i) = \mathbb{1}\{-\kappa_N < \widehat{F}_\Upsilon(\Upsilon_i) - \widehat{F}_\Upsilon(\Upsilon_j) < \kappa_N\}$ where $\widehat{F}_\Upsilon(\cdot)$ is the empirical distribution function of Υ and $2\kappa_N \in (0,1)$ represents the percentage of individuals that are included in each neighborhood (boundaries). The resulting sensitivity and specificity are

[12] When defining Υ_i, we treat $\widehat{\beta}$ as fixed because it is estimated on the train set. The probabilities in $Se(c,t)$ and $Sp(c,t)$ are over the distribution of the test set.

defined by:

$$\widehat{Se}(c,t) = \frac{\left(1-\widehat{F}_{\Upsilon}(c)\right) - \widehat{S}_{\kappa_N}(c,t)}{1-\widehat{S}_{\kappa_N}(t)},$$
$$\widehat{Sp}(c,t) = 1 - \frac{\widehat{S}_{\kappa_N}(c,t)}{\widehat{S}_{\kappa_N}(t)},$$

where $\widehat{S}_{\kappa_N}(t) = \widehat{S}_{\kappa_N}(-\infty,t)$. Both sensitivity and specificity above are monotone and bounded in $[0,1]$.

Heagerty et al. (2000) used bootstrap resampling to estimate the confidence intervals for this ROC curve estimator. Motivated by the results of Akritas (1994) and Cai et al. (2011), Hung and Chiang (2010) discussed the asymptotic properties of the estimator and concluded that bootstrap resampling techniques can be used to estimate the variances of the proposed ROC curve. In practice, Heagerty et al. (2000) suggested that the value for κ_N is chosen to be $O\left(N^{-\frac{1}{3}}\right)$. In the present paper, we use the default value of the κ_N produced in the documentation of the R package 'SurvivalROC', which is consistent with the choice found in Blanche et al. (2013). For further details on other ROC curve estimators in the survival analysis, we refer to Kamarudin et al. (2017).

4.3 Simulation results

We compute results for the three different LASSO-type regression methods. In the structured approach, sg-LASSO-MIDAS, each covariate and its high-frequency lags share the same group, therefore, we have $K+1$ groups (one group corresponding to the intercept). Table 1 presents the number of parameters (including the intercept) to be estimated in each of the three methods. It is evident that the two methods using MIDAS weights help mitigate the high-dimensional problem when $s \times m$ exceeds L. We start by comparing the prediction results for sample sizes $N \in \{800, 1200\}$ across three simulation scenarios, followed by examining the recovery of the MIDAS weight function. To assess the prediction performance, we randomly split the simulated dataset into a training dataset (80%) and a test dataset (20%), ensuring that both sets maintain the same proportion of the event indicator $\delta_i(t)\mathbb{1}\{\widetilde{T}_i \leq t\}$. We then calculate the estimated AUC in the test dataset, with the average estimated AUC obtained from 100 simulated datasets for each sample size. The tuning parameters in the sg-LASSO-MIDAS

Table 1: Number of parameters to be estimated in different methods.

Methods	Number of estimated parameters
LASSO-UMIDAS	$1 + K \times s \times m$
LASSO-MIDAS	$1 + K \times L$
sg-LASSO-MIDAS	$1 + K \times L$

and LASSO-MIDAS models are selected using 5-fold stratified cross-validation to maximize the AUC, and the same procedure is applied to the LASSO-UMIDAS model. Specifically, we perform a grid search over the regularization parameter α in the sparse group LASSO penalty, with values in the set $\{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$ and, as standard, λ is chosen in a grid which follows Liang et al. (2024). Table 2 reports the estimated average AUCs in the test dataset. As shown, sg-LASSO-MIDAS achieves the highest AUC values across different simulation scenarios. Both sg-LASSO-MIDAS and LASSO-MIDAS, using weight function approximations, outperform LASSO-UMIDAS. LASSO without MIDAS weighting generally demonstrates the poorest predictive performance. As expected, the predictive performance improves with an increase in sample size N. These results remain robust as the persistence parameter ρ of covariates increases from 0.1 to 0.6. Although all three methods perform less effectively with heavy-tailed covariates, sg-LASSO-MIDAS continues to outperform the others. In Tables 2, 3 and 4 of Appendix D, we report additional results for the estimation accuracy of the true parameters. It is worth noting that the increase of the parameter estimation accuracy with the sample size is not particularly large, as the high censoring rate in the simulated datasets limits the increase in the number of uncensored firms ($\mathbb{1}\{T_i \leq C_i\} = 1$)

Table 2: Estimated average AUCs (standard deviation) in the test dataset of the three different methods: LASSO-UMIDAS (LASSO-U), LASSO-MIDAS (LASSO-M), sg-LASSO-MIDAS (sg-LASSO-M). $s = 6$ and $t = \{t_1 = 10\%, t_2 = 30\%, t_3 = 50\%\}$ percentile of the set $\{T_i : T_i \text{ is uncensored }, i \in [N]\}$.

	Scenario 1					
	$N = 800$			$N = 1200$		
	$t = t_1$	$t = t_2$	$t = t_3$	$t = t_1$	$t = t_2$	$t = t_3$
LASSO-U	0.584 (0.160)	0.591 (0.125)	0.569 (0.098)	0.652 (0.122)	0.646 (0.094)	0.611 (0.092)
LASSO-M	0.870 (0.094)	0.847 (0.083)	0.793 (0.088)	0.911 (0.057)	0.884 (0.056)	0.843 (0.051)
sg-LASSO-M	0.903 (0.087)	0.884 (0.059)	0.825 (0.076)	0.928 (0.048)	0.913 (0.041)	0.867 (0.054)
	Scenario 2					
	$N = 800$			$N = 1200$		
	$t = t_1$	$t = t_2$	$t = t_3$	$t = t_1$	$t = t_2$	$t = t_3$
LASSO-U	0.628 (0.196)	0.673 (0.105)	0.636 (0.110)	0.680 (0.179)	0.752 (0.084)	0.718 (0.079)
LASSO-M	0.859 (0.121)	0.871 (0.087)	0.823 (0.087)	0.908 (0.074)	0.911 (0.045)	0.887 (0.048)
sg-LASSO-M	0.884 (0.113)	0.905 (0.065)	0.862 (0.076)	0.935 (0.052)	0.936 (0.032)	0.912 (0.037)
	Scenario 3					
	$N = 800$			$N = 1200$		
	$t = t_1$	$t = t_2$	$t = t_3$	$t = t_1$	$t = t_2$	$t = t_3$
LASSO-U	0.611 (0.180)	0.606 (0.124)	0.553 (0.108)	0.620 (0.159)	0.637 (0.097)	0.584 (0.092)
LASSO-M	0.769 (0.188)	0.793 (0.116)	0.754 (0.120)	0.820 (0.150)	0.852 (0.083)	0.831 (0.080)
sg-LASSO-M	0.774 (0.193)	0.822 (0.102)	0.783 (0.111)	0.848 (0.129)	0.878 (0.081)	0.849 (0.085)

to only about 76 as the sample size N grows from 800 to 1200.[13] This simulation evidence strongly supports the advantage of using MIDAS weighting and incorporating the internal structure of covariates in high-dimensional settings.

5 Empirical application

5.1 Data

We construct a dataset of all publicly traded Chinese manufacturing firms listed on the Shanghai and Shenzhen Stock Exchanges. These firms' financial statuses are classified as either Special Treatment (ST) or No-ST.[14] A firm is designated as an ST firm if it meets any of the following criteria: i) two consecutive years of earnings are negative; ii) one recent year of earnings is negative and the most recent year of equity is negative; iii) the most recent year's audited financial statements conclude with substantial doubt; and iv) other situations identified by the stock exchange as abnormal activities or a high risk of delisting. According to Li et al. (2021), ST status is a reliable indicator of financial distress in China. Therefore, we use the ST indicator as a proxy for a firm's financial distress.

The dataset is sourced from the IFIND database https://www.hithink.com/ifind.html, one of China's leading financial data providers. The database contains mostly manually extracted, covering financial data such as stocks, bonds, funds, futures, and indexes. Additionally, we have developed an R package, `Survivalml`, which is publicly available at https://github.com/Wei-M-Wei/Survivalml. Detailed information about the package and dataset can be found in Appendix F.

The raw dataset consists of 1614 companies, of which 299 were classified as ST and 1315 as No-ST.[15] The dataset exhibits a censoring rate of approximately 81%. We collect 57 quarterly measured financial variables, categorized into 8 types (number of covariates in each type), as follows: Operation-Related (6), Debt-Related (10), Profit-Related (16), Potential-Related (6), Z-score Related (5) (Altman, 1968), Capital-Related (6), Stock-Related (5), and Cash-Related (3). Table 5 provides detailed information on these financial variables; see Appendix F for further details. Figure 1 presents the distribution of IPO, first-time-to-be ST, and censored firms across different years of the raw dataset. Many of them were listed in 2010 and 2011, several firms were publicly listed in 2013 and the financial distress firms seem to be distributed evenly between 1999 and 2020.[16]

[13] In simulation results not shown for brevity, when the dataset's censoring rate is approximately 30%, notable improvements both in parameter estimation and estimated AUC are observed as the sample size N increases from 800 to 1200.

[14] The initial public offering (IPO) dates of these firms fall between 1985, January 1st and 2015, December 31st.

[15] There are no mergers in the dataset.

[16] Since China put froze IPOs in 2013, there were only a limited number of IPO firms in this year.

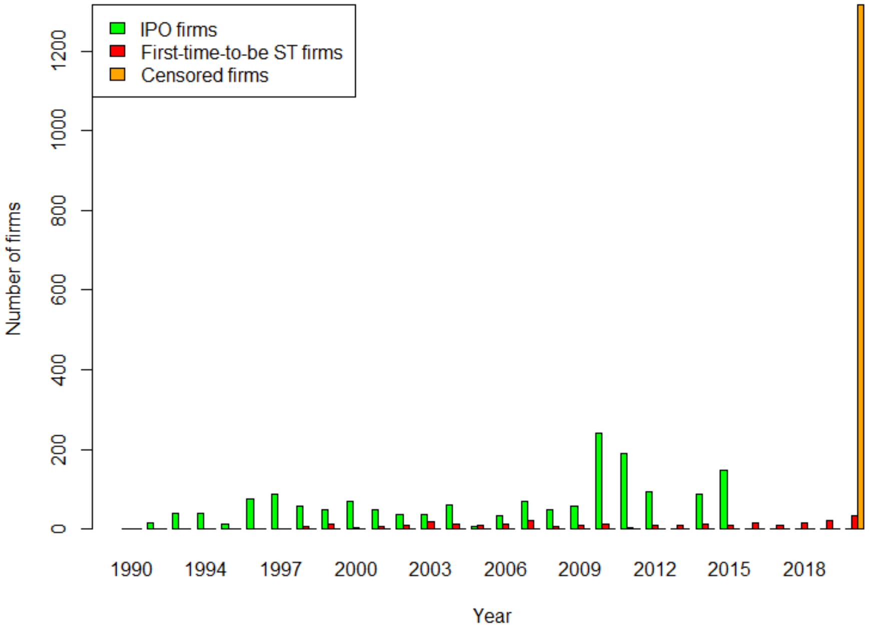

Figure 1: Number of IPO, first-time-to-be ST, and censored firms across different years in the raw dataset.

In addition, we construct a sub-dataset in which all firms have survived at least s years. The goal is to use these s years of information to predict whether a firm will fail within t years. Figure 2 shows the prediction procedure applied to the real dataset. The observation period spans from 1985, January 1^{st}, to 2020, December 31^{st}. The survival time T of each firm i is defined as the interval between the firm's IPO date and the first instance when the company was classified as ST. If a firm was never classified as ST, we only observe the censoring time C, which is the interval from the IPO date to the end of the observation period. Both T and C are measured in years. Firms 1 and 2 represent uncensored firms, so their survival time can be fully observed within the observation period. Firms 3 and 4 are censored, and we can only observe their censoring time C. Thus, for all firms, only $\widetilde{T} = T \wedge C$ is observable.

5.2 Estimation procedure

We now describe the estimation procedure in the empirical application. First, we note that all public firms report their financial information with a one-quarter delay. Consequently, if a firm has survived for s years, only $s \times 4 - 1$ quarters' worth of financial covariate information will be available for analysis.

Let $x_{i,s-\frac{j-1}{m},k}$ represent the k-th financial covariate of firm i, measured at time $s - \frac{j-1}{m}$, where $j = 2, \ldots, d$, and $d = s \times m$. We organize all the lags of the covariate into a group

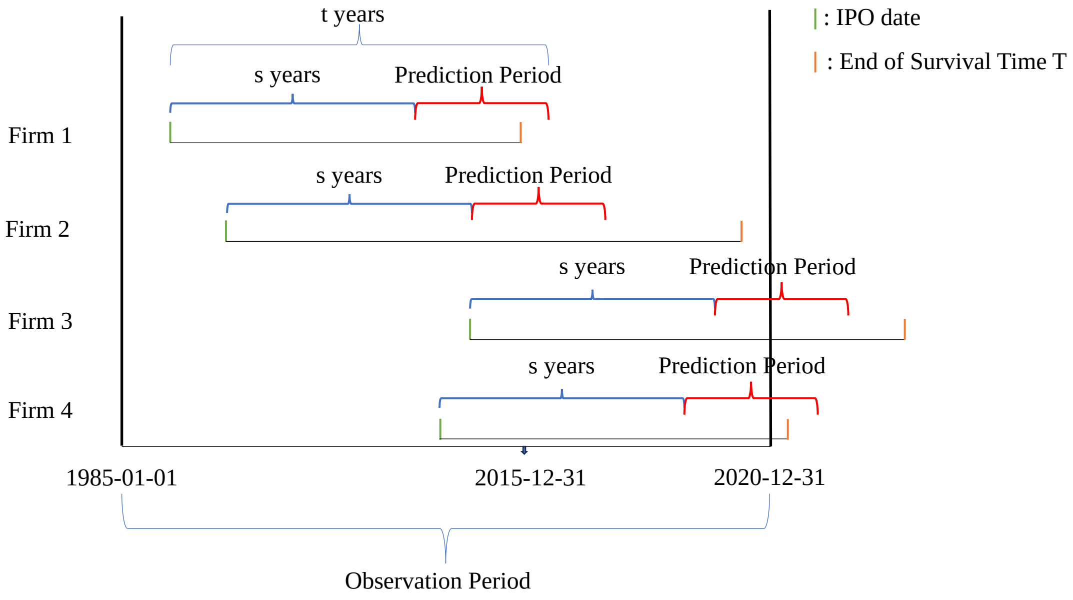

Figure 2: Prediction procedure in the empirical application.

vector $\widetilde{\boldsymbol{Z}}_{i,k}$:

$$\widetilde{\boldsymbol{Z}}_{i,k} = \left(x_{i,s-\frac{1}{m},k}, x_{i,s-\frac{2}{m},k}, \dots, x_{i,\frac{1}{m},k}\right)^\top, \quad i \in [N], \quad k \in [K],$$

where $x_{i,\frac{1}{m},k}$ refers to the k-th covariate measured in the next quarter following the firm's IPO date, and $m = 4$ denotes the quarterly frequency of the financial covariates.

Next, we aggregate the lagged covariate vector $\widetilde{\boldsymbol{Z}}_{i,k}$ using a dictionary W, which consists of Gegenbauer polynomials shifted to the interval $[0, 1]$ with parameter $\alpha_{\text{poly}} = -\frac{1}{2}$ and size $L = 3$.[17]

Finally, we construct the covariate matrix $\boldsymbol{X}$ as follows:

$$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_N)^\top,$$

where each $\boldsymbol{X}_i = \left(1, \widetilde{\boldsymbol{Z}}_{i,1}^\top W, \widetilde{\boldsymbol{Z}}_{i,2}^\top W, \dots, \widetilde{\boldsymbol{Z}}_{i,K}^\top W\right)^\top, i \in [N]$. This matrix $\boldsymbol{X}$ is then used in sg-LASSO-MIDAS and LASSO-MIDAS. Notice that we include the intercept term but do not penalize it in the estimation procedure.

We compare the performance of firm distress predictions using the following methods.

Logistic regression. As a benchmark, we consider a simple unpenalized logistic regression with the latest lag of all financial covariates. We solve the empirical version of (3), in which

[17] These polynomials are also known as the second type of Chebyshev polynomials on the interval $[0, 1]$.

the function H is estimated using the Kaplan-Meier estimator. This is considered a reasonable starting point for distress predictions. Only the latest lag for each covariate is used, and the total number of parameters we need to estimate is $1 + K$, where K is the number of covariates without their lags.

LASSO-U (LASSO-UMIDAS). We estimate $d - 1 = s \times m - 1$ coefficients per group covariate $\widetilde{\boldsymbol{Z}}_{i,k} \in \mathbb{R}^{d-1}, k \in [K]$, using the unstructured LASSO estimator. The total number of parameters to estimate is $1 + K \times (s \times m - 1)$, where s is the number of years survived by the firm, and m represents the number of lags for each covariate.

LASSO-M (LASSO-MIDAS). Each high-frequency covariate and its d lags are grouped into $\widetilde{\boldsymbol{Z}}_{i,k} \in \mathbb{R}^{d-1}, k \in [K]$. We aggregate the group covariate $\widetilde{\boldsymbol{Z}}_{i,k}$ using Gegenbauer polynomials $W \in \mathbb{R}^{(d-1)\times L}$. We apply a Lasso penalty to induce sparsity. The total number of parameters, including the intercept, to estimate is $1 + K \times L$, where L is the size of the Gegenbauer polynomial dictionary.

sg-LASSO-M (sg-LASSO-MIDAS). Similarly to LASSO-MIDAS, each high-frequency covariate and its d lags form a group $\widetilde{\boldsymbol{Z}}_{i,k} \in \mathbb{R}^{d-1}, k \in [K]$, which is aggregated using Gegenbauer polynomials $W \in \mathbb{R}^{(d-1)\times L}$. Instead of using a Lasso penalty, we use the sparse-group Lasso penalty to induce sparsity in the group covariates. The total number of parameters to estimate is $1 + K \times L$.

The choice of s dictates the historical information captured in the covariate matrix $\boldsymbol{X}$, while t denotes the prediction horizon. The existing literature on firm distress prediction, particularly in the United States, often examines prediction horizon t ranging from 1 quarter to 2 years (Cole and White, 2012). In practical applications, such as for bank regulators, models need to identify potential failures well in advance. For example, Audrino et al. (2019) developed a MIDAS-type method with prediction horizons of 1 and 2 years.

For our empirical application, we select a reference period of $s = 6$ years. Using the firm classification criteria for Special Treatment outlined in Section 5.1, we establish prediction horizons of $t = 8, 8.5, 9$ years to forecast firm distress within these intervals. To investigate longer forecast periods, we also consider an additional case with $s = 10$ years and prediction horizons of $t = 13, 13.5, 14$ years. Longer prediction horizons provide information on the risks of long-term financial distress. As highlighted by Li et al. (2021), these prediction horizons are critical for accurately forecasting firm financial distress in China. They also offer meaningful and practical benchmarks for evaluating firm failure prediction models.

In practice, missing data in financial variables can arise due to various factors, including inconsistent reporting practices between firms, differing regulatory requirements, incomplete disclosures, and delays in data availability after IPOs. Given the substantial amount of missing data in the raw dataset, we construct a complete sub-dataset for each s by selecting firms with consistent s-year observations. While common approaches for handling missing data, such as removing variables or firms with missing values, are widely used, these methods often result in retaining too few firms or variables for meaningful analysis. Furthermore, the firms with observable status in the sub-dataset play a critical role in the predictive modeling process. To address these challenges, we propose an algorithm that balances dimensionality and the number of uncensored firms in the selected sub-dataset, as outlined in Algorithm C in Appendix C.

To evaluate the prediction performance of the different methods, we randomly split the dataset into in-sample (80%) and out-of-sample (20%) sets, ensuring that both sets maintain the same proportion of the event indicator $\delta_i(t)\mathbb{1}\{\widetilde{T}_i \leq t\}$. The tuning parameters for sg-LASSO-MIDAS, LASSO-MIDAS, and LASSO-UMIDAS are selected using stratified 5-fold cross-validation, where the optimal parameters are those that maximize the AUC in the out-of-sample set.[18] Additionally, as an alternative, cross-validation to maximize the likelihood score is also investigated. The AUC estimator employed in this procedure follows the method described in Section 4.2. Specifically, we perform a grid search over the regularization parameter α in the sparse group LASSO penalty, with values in the set $\{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$ and, as standard, λ is chosen in a grid which follows Liang et al. (2024). This process is repeated 10 times as a robustness check, each time using a different random split of the data. All models are trained on the same training set and evaluated on the same test set.

For each split, the AUC is computed on the out-of-sample data, and the out-of-sample data is then bootstrapped 1000 times to calculate the AUC for each bootstrap sample. The AUC values for each bootstrap sample are subsequently averaged across the 10 different splits, resulting in 1000 averaged AUC values. The final performance is reported as the overall average AUC, along with a two-side 95% confidence interval, which is calculated based on these 1000 bootstrapped averages. This approach ensures a robust performance evaluation by accounting for variability in the data and model performance.[19]

On top of the simple logistic regression and the LASS0-UMIDAS, LASS0-MIDAS, and sg-LASSO-MIDAS with cross-validation for the AUC or the likelihood score, we consider other alternative approaches.

Macro data augmented prediction. We first assess whether incorporating macroeconomic data can enhance the accuracy of distress prediction models. The macroeconomic dataset for China is sourced from the Federal Reserve Bank of Atlanta's China Macroeconomy Project (`https://www.atlantafed.org/cqer/research/china-macroeconomy#Tab2`), which provides a comprehensive set of macroeconomic variables relevant to the Chi-

[18] In this paper, unless specified otherwise, the default choice for cross-validation is to maximize the AUC.

[19] We note that the bootstrap approach is not theoretically validated for the regularized estimators we consider in this paper.

nese economy. The dataset includes 98 macroeconomic variables, measured quarterly, and spans the same time period as the financial data collected for the firms in our study.

To merge the macroeconomic data with the financial dataset, we select only those macroeconomic variables that do not have missing values across all firms within each financial sub-dataset. Since the sub-datasets differ based on the value of s, the set of macroeconomic variables selected will vary accordingly for each sub-dataset. Furthermore, we use the same MIDAS dictionary W for the macroeconomic covariates as for the financial covariates, ensuring consistency in the aggregation of high-frequency data over time.

Table 3 summarizes the details of the two sub-datasets categorized by different values of s. For the sub-dataset with $s = 6$ years, we use all available information across each firm's entire survival period, allowing us to leverage the maximum historical data available for firms with 6 years of survival. In contrast, for the sub-dataset with $s = 10$ years, we restrict the covariates to those from the last 4 years of each firm's survival period. This adjustment is necessary because firms that have survived for 10 years were generally listed in the 1990s, and significant missing data is often observed in the early years following their IPOs. By focusing on the most recent 4 years, we ensure better data quality and a more robust analysis.

Table 3: Summary information of the dataset with $s = 6$ and $s = 10$ years.

	$s = 6$ years	$s = 10$ years
Number of firms N	901	784
Number of uncensored firms	67	80
Number of financial covariates $K_{\text{financial}}$ (including lags)	32 (736)	36 (540)
Number of macro covariates K_{macro} (including lags)	63 (1449)	63 (945)
30% percentile of $\widetilde{T}$ (years)	9.512	13.369
50% percentile of $\widetilde{T}$ (years)	10.285	15.411
30% percentile of T among uncensored firms (years)	7.789	11.032
50% percentile of T among uncensored firms (years)	8.934	13.844

Oversampling. In addition, we apply an oversampling technique to address the imbalance in the dataset caused by the high censoring rate, which results in an unequal proportion of firms experiencing distress versus those that are not. This imbalance could adversely affect the performance of distress prediction models, as the minority class (distressed firms) may be underrepresented. Since the empirical dataset has a high censoring rate, we face a class imbalance between those firms that eventually experience distress $\mathbb{1}\{T_i \leq C_i\}\mathbb{1}\{\widetilde{T}_i \leq t\} = 1$ and those that do not or we do not observe $\mathbb{1}\{T_i \leq C_i\}\mathbb{1}\{\widetilde{T}_i \leq t\} = 0$. To balance this, for the training dataset, we randomly duplicate the observations from the minority class (firms that experience distress) until the proportion of distressed firms reaches 15% of the training dataset. This step helps mitigate the imbalance and ensures that the model is exposed to a sufficient number of distressed firms during training. Tuning parameters are selected using 5-fold stratified cross-validation, where the optimal parameters maximize the likelihood score.

Does censoring matter for prediction? We compare with an approach that applies LASS0-UMIDAS, LASS0-MIDAS, and sg-LASSO-MIDAS to the sub-dataset where censored firms with censoring time smaller than t ($\mathbb{1}\{C_i < T_i\}\mathbb{1}\{C_i < t\} = 1$) have been removed.[20] This is the approach usually taken in the literature, since it allows to ignore censoring, see the discussion in the literature review of the introduction. The limitation of this procedure is that it does not use all observations, resulting in a loss of precision.

5.3 Application results

The results are presented in Tables 4 and 5. The LASSO-MIDAS and sg-LASSO-MIDAS consistently outperform the LASSO-UMIDAS, which aligns with our expectations and the logistic regression benchmark. Among the two, the sg-LASSO-MIDAS provides a slight performance advantage over LASSO-MIDAS, particularly when $s = 10$ years, indicating that the sparse-group Lasso regularization is beneficial for the prediction task, especially when incorporating a larger historical window of data.

When we compare the performance of models based on cross-validation using different metrics, we observe that cross-validation based on the AUC generally yields better results than cross-validation based on likelihood scores. This is logical since our target measure is the AUC itself.

Additionally, while integrating macroeconomic data does not improve prediction performance over the purely financial model when $s = 6$ years, it enhances performance when $s = 10$ years. This suggests that macroeconomic variables become more relevant with a larger historical window, offering supplementary information that helps improve prediction accuracy, especially for firms with longer survival periods. However, oversampling does not seem to provide any additional benefit in improving prediction performance.

When we remove censored firms with $C_i < t$, the performance of our methodologies deteriorates across all scenarios, emphasizing the importance of properly accounting for censoring in predictive modeling.

To further assess the performance difference, we conduct a pairwise comparison test, a widely used method for comparing two AUCs. Slightly modifying the approach in James A. Hanley (1983); Robin et al. (2011), we specifically test whether the estimated AUC of

[20] Given the prediction horizon t, the distress status of censored firms cannot be observed if their censoring time is shorter than t and, clearly, censored firms with $\mathbb{1}\{C_i \geq t\}$ are not distressed.

Table 4: (Distress prediction performance) Estimated average AUCs (95% confidence interval) in the out-of-sample set with $s = 6$ years and prediction horizons $t = 8, 8.5, 9$ years.

	$s = 6$ years		
	$t = 8$ years	$t = 8.5$ years	$t = 9$ years
	Benchmark		
Logistic reg.	0.714 [0.666, 0.760]	0.698 [0.664, 0.736]	0.782 [0.754, 0.816]
	Cross-validation for the AUC		
LASSO-U	0.797 [0.755, 0.844]	0.756 [0.717, 0.801]	0.765 [0.734, 0.802]
LASSO-M	0.838 [0.793, 0.872]	0.817 [0.774, 0.864]	0.811 [0.778, 0.843]
sg-LASSO-M	0.823 [0.789, 0.865]	0.821 [0.778, 0.861]	0.806 [0.776, 0.840]
	Cross-Validation for the likelihood score		
LASSO-U	0.710 [0.671, 0.753]	0.644 [0.587, 0.708]	0.761 [0.718, 0.804]
LASSO-M	0.701 [0.665, 0.738]	0.851 [0.817, 0.894]	0.808 [0.774, 0.843]
sg-LASSO-M	0.782 [0.747, 0.820]	0.813 [0.773, 0.862]	0.795 [0.760, 0.835]
	Macro Data Augmented		
LASSO-U	0.790 [0.767, 0.819]	0.740 [0.718, 0.772]	0.721 [0.698, 0.748]
LASSO-M	0.823 [0.797, 0.848]	0.810 [0.786, 0.836]	0.782 [0.761, 0.804]
sg-LASSO-M	0.820 [0.800, 0.846]	0.806 [0.783, 0.830]	0.798 [0.778, 0.822]
	Oversampling with Financial Data		
LASSO-U	0.833 [0.800, 0.863]	0.760 [0.716, 0.800]	0.773 [0.746, 0.808]
LASSO-M	0.801 [0.760, 0.838]	0.832 [0.806, 0.864]	0.825 [0.799, 0.855]
sg-LASSO-M	0.810 [0.768, 0.851]	0.834 [0.808, 0.861]	0.822 [0.800, 0.851]
	Data without censored firms satisfying $C_i < t$		
LASSO-U	0.707 [0.659, 0.768]	0.792 [0.759, 0.829]	0.731 [0.695, 0.776]
LASSO-M	0.787 [0.741, 0.829]	0.817 [0.777, 0.854]	0.744 [0.701, 0.791]
sg-LASSO-M	0.753 [0.702, 0.816]	0.820 [0.778, 0.856]	0.776 [0.733, 0.821]

Table 5: (Distress prediction performance) Estimated average AUCs (95% confidence interval) in the out-of-sample set with $s = 10$ years and prediction horizons $t = 13, 13.5, 14$ years.

	$s = 10$ years		
	$t = 13$ years	$t = 13.5$ years	$t = 14$ years
	Benchmark		
Logistic reg.	0.621 [0.566, 0.682]	0.598 [0.555, 0.647]	0.635 [0.600, 0.675]
	Cross-Validation for the AUC		
LASSO-U	0.566 [0.514, 0.633]	0.628 [0.591, 0.674]	0.669 [0.637, 0.709]
LASSO-M	0.773 [0.738, 0.818]	0.653 [0.615, 0.700]	0.688 [0.654, 0.726]
sg-LASSO-M	0.818 [0.781, 0.847]	0.671 [0.635, 0.718]	0.702 [0.667, 0.737]
	Cross-Validation for the likelihood score		
LASSO-U	0.572 [0.537, 0.625]	0.555 [0.520, 0.603]	0.622 [0.593, 0.663]
LASSO-M	0.787 [0.754, 0.831]	0.609 [0.579, 0.668]	0.701 [0.671, 0.739]
sg-LASSO-M	0.815 [0.779, 0.853]	0.657 [0.630, 0.710]	0.701 [0.677, 0.739]
	Macro Data Augmented		
LASSO-U	0.573 [0.542, 0.611]	0.702 [0.677, 0.727]	0.678 [0.659, 0.701]
LASSO-M	0.747 [0.728, 0.779]	0.670 [0.647, 0.703]	0.691 [0.671, 0.716]
sg-LASSO-M	0.773 [0.750, 0.795]	0.707 [0.687, 0.738]	0.725 [0.706, 0.749]
	Oversampling with Financial Data		
LASSO-U	0.655 [0.602, 0.711]	0.636 [0.590, 0.675]	0.697 [0.668, 0.731]
LASSO-M	0.704 [0.649, 0.753]	0.627 [0.590, 0.677]	0.679 [0.640, 0.714]
sg-LASSO-M	0.685 [0.637, 0.738]	0.615 [0.578, 0.664]	0.669 [0.634, 0.706]
	Data without censored firms satisfying $C_i < t$		
LASSO-U	0.764 [0.726, 0.809]	0.619 [0.577, 0.665]	0.695 [0.667, 0.733]
LASSO-M	0.759 [0.726, 0.815]	0.624 [0.588, 0.669]	0.616 [0.575, 0.658]
sg-LASSO-M	0.802 [0.764, 0.845]	0.625 [0.592, 0.676]	0.642 [0.610, 0.680]

sg-LASSO-MIDAS is superior to that of LASSO-UMIDAS.[21] The results indicate that the improvement of the sg-LASSO-MIDAS over the LASSO-UMIDAS is statistically significant at least at the 10% significance level across all scenarios, with the largest gap observed when $s = 10$ and $t = 13$ years. We also conduct a pairwise comparison between the sg-LASSO-MIDAS applied to the dataset with and without censored firms satisfying $C_i < t$. For the scenarios where $s = 6$, $t = 8.5$ years, and $s = 10$, $t = 13$ years, sg-LASSO-MIDAS performs better on the full dataset than on the dataset without censored firms satisfying $C_i < t$, though the difference is not statistically significant. However, in other scenarios, including censoring significantly improves model performance, with results being statistically significant, at least at the 5% level. These findings strongly support the advantages of using MIDAS weights, considering the group structure of covariates, and incorporating the censoring information in practice. Overall, the empirical results highlight the superiority of the sg-LASSO-MIDAS across different scenarios.

To better understand which covariates are useful for prediction, we examine the financial types selected by the sg-LASSO-MIDAS, as illustrated in Figure 3 in Appendix D. Financial variables related to the Z-score appear to play a pivotal role across all prediction horizons in forecasting firm distress. This observation aligns with prior research (Altman, 1968), as the Z-score model has been widely employed in both academic studies and industry to predict corporate defaults (Altman et al., 2017). Further details on the selected financial covariates are presented in Figures 1 and 2 in Appendix D.

Table 6: Pairwise difference test across different scenarios: Null hypothesis H_0: estimated AUC of the first prediction approach is larger. We use $*$ and $**$ to indicate 10% and 5% significance, respectively.

$s = 6$ years			$s = 10$ years		
$t = 8$ years	$t = 8.5$ years	$t = 9$ years	$t = 13$ years	$t = 13.5$ years	$t = 14$ years
sg-LASSO-M $vs.$ LASSO-U			sg-LASSO-M $vs.$ LASSO-U		
0.098^{*}	0.000^{**}	0.000^{**}	0.000^{**}	0.010^{**}	0.065^{*}
sg-LASSO-M with $vs.$ without censored firms satisfying $C_i < t$			sg-LASSO-M with $vs.$ without censored firms satisfying $C_i < t$		
0.001^{**}	0.495	0.021^{**}	0.238	0.022^{**}	0.001^{**}

5.4 Additional results

To further show the performance of the proposed method, we present another application of distress prediction using the same dataset, but with a different method for dividing the

[21] We have 1000 bootstrapped average AUCs for each method as described before. The p-value is calculated as the proportion of sg-LASSO-MIDAS's AUC values that are smaller than those of another method.

in-sample and out-of-sample sets compared to the previous subsection, which is closer to real-time prediction.

Recall that the financial dataset spans from 1985, January 1^{st}, to 2020, December 31^{st}. To better align with practical applications, when $s = 6$ years, we first select the time point $2016/12/31$. Firms that had already survived 6 years prior to this date are used as the in-sample set (544 firms), while the remaining firms that had not yet survived 6 years by $2016/12/31$ are placed in the out-of-sample set (357 firms). Thus, the actual observation period ranges from $1985/01/01$ to $2016/12/31$. The prediction horizons are set as $t = 8, 8.5, 9$ years, as in previous analyses. The regularization parameters λ and α using 5-fold stratified cross-validation for AUC. Specifically, we use a grid of $\{0.9, 0.91, 0.92, \dots, 1\}$ to search for the optimal regularization parameter α in the sparse group LASSO penalty. As before, λ is chosen in a grid which follows Liang et al. (2024). All other settings are consistent with those in the previous section, except that we use a dictionary W composed of Gegenbauer polynomials shifted to $[0, 1]$ with parameter $\alpha_{\text{poly}} = \frac{1}{2}$ and size $L = 3$.

For $s = 10$ years, which is relatively large, we select a new time point of $2013/12/31$ to allow for more years of prediction after this date. The prediction horizons are set to $t = 13, 13.5, 14$ years. Firms that had survived for $s = 10$ years before $2013/12/31$ are used as the in-sample set (311 firms), while those that had not survived $s = 10$ years by this time are treated as the out-of-sample set (473 firms).

Tables 7 and 8 report the estimated AUCs in the out-of-sample set. The second-to-last row presents the pairwise test between sg-LASSO-MIDAS and LASSO-UMIDAS, while the last row presents the comparison between sg-LASSO-MIDAS applied to data with and without censored firms satisfying $C_i < t$.

For $s = 6$ years, sg-LASSO-MIDAS significantly outperforms LASSO-UMIDAS, whereas LASSO-MIDAS performs similarly to sg-LASSO-MIDAS. Furthermore, the macroeconomic data-augmented prediction appears comparable to the purely financial model in most scenarios. However, the prediction performance is observed to be more stable when macroeconomic data is included compared to using only financial data. Additionally, sg-LASSO-MIDAS performs statistically better at the 5% level when the dataset includes censored firms with $C_i < t$, except for the $t = 9$ years prediction horizon, highlighting the advantage of accounting for censoring in the prediction model. For $s = 10$ years, sg-LASSO-MIDAS applied to the full dataset is numerically superior to both sg-LASSO-MIDAS applied to the dataset without censored firms satisfying $C_i < t$ and LASSO-UMIDAS. However, both of the differences are statistically significant at 10% level only for $t = 14$ years.

Table 7: (Additional application) Estimated AUCs (95% confidence interval) in the out-of-sample set with $s = 6$ years and prediction horizons $t = 8, 8.5, 9$ years.

	$s = 6$ years		
	$t = 8$ years	$t = 8.5$ years	$t = 9$ years
	Benchmark		
Logistic reg.	0.730 [0.565, 0.901]	0.725 [0.576, 0.860]	0.671 [0.543, 0.790]
	Cross-Validation for the AUC		
LASSO-U	0.734 [0.565, 0.880]	0.688 [0.558, 0.824]	0.666 [0.539, 0.798]
LASSO-M	0.898 [0.829, 0.952]	0.866 [0.774, 0.940]	0.812 [0.728, 0.910]
sg-LASSO-M	0.898 [0.829, 0.952]	0.845 [0.746 ,0.932]	0.812 [0.728, 0.910]
	Cross-Validation for the likelihood score		
LASSO-U	0.736 [0.566, 0.881]	0.714 [0.588, 0.838]	0.552 [0.445, 0.669]
LASSO-M	0.898 [0.828, 0.956]	0.841 [0.753, 0.938]	0.789 [0.680, 0.908]
sg-LASSO-M	0.898 [0.828, 0.956]	0.835 [0.748, 0.932]	0.760 [0.648, 0.898]
	Macro Data Augmented		
LASSO-U	0.734 [0.645, 0.812]	0.688 [0.616, 0.764]	0.666 [0.583, 0.751]
LASSO-M	0.898 [0.868, 0.933]	0.809 [0.745, 0.877]	0.769 [0.739, 0.813]
sg-LASSO-M	0.899 [0.869, 0.933]	0.874 [0.838, 0.915]	0.759 [0.731, 0.811]
	Data without censored firms satisfying $C_i < t$		
LASSO-U	0.680 [0.543, 0.822]	0.745 [0.648, 0.829]	0.628 [0.512, 0.787]
LASSO-M	0.807 [0.679, 0.926]	0.767 [0.671, 0.884]	0.778 [0.682, 0.885]
sg-LASSO-M	0.807 [0.679, 0.926]	0.767 [0.671, 0.884]	0.773 [0.658, 0.888]
	sg-LASSO-M $vs.$ LASSO-U		
p-value	0.001^{**}	0.000^{**}	0.000^{**}
	sg-LASSO-M with $vs.$ without censored firms satisfying $C_i < t$		
p-value	0.009^{**}	0.042^{**}	0.107

Notes: The second-to-last row reports the p-value from the pairwise difference test across the three prediction horizons, with the null hypothesis that the estimated AUC of sg-LASSO-MIDAS is superior to LASSO-UMIDAS's. The last row presents the p-value from the pairwise test across the three prediction horizons, comparing sg-LASSO-MIDAS applied to data with and without censored firms satisfying $C_i < t$. We use $*$ and $**$ to indicate 10% and 5% significance, respectively.

Table 8: (Additional application) Estimated AUCs (95% confidence interval) in the out-of-sample set with $s = 10$ years and prediction horizons $t = 13, 13.5, 14$ years.

	$s = 10$ years		
	$t = 13$ years	$t = 13.5$ years	$t = 14$ years
		Benchmark	
Logistic reg.	0.619 [0.420, 0.777]	0.598 [0.431, 0.752]	0.647 [0.502, 0.796]
		Cross-Validation for the AUC	
LASSO-U	0.685 [0.483, 0.879]	0.753 [0.529, 0.895]	0.681 [0.541, 0.847]
LASSO-M	0.775 [0.528, 0.943]	0.784 [0.685, 0.878]	0.801 [0.704, 0.880]
sg-LASSO-M	0.758 [0.499, 0.946]	0.803 [0.706, 0.879]	0.801 [0.704, 0.880]
		Cross-Validation for the likelihood score	
LASSO-U	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]	0.690 [0.547, 0.846]
LASSO-M	0.789 [0.687, 0.902]	0.806 [0.694, 0.907]	0.696 [0.577, 0.865]
sg-LASSO-M	0.789 [0.687, 0.902]	0.806 [0.694, 0.907]	0.696 [0.577, 0.865]
		Macro Data Augmented	
LASSO-U	0.685 [0.574, 0.806]	0.643 [0.559, 0.756]	0.738 [0.650, 0.827]
LASSO-M	0.775 [0.718, 0.842]	0.787 [0.741, 0.845]	0.788 [0.715, 0.848]
sg-LASSO-M	0.788 [0.724, 0.837]	0.797 [0.750, 0.855]	0.796 [0.716, 0.853]
		Data without censored firms satisfying $C_i < t$	
LASSO-U	0.662 [0.491, 0.893]	0.507 [0.288, 0.701]	0.662 [0.474, 0.806]
LASSO-M	0.737 [0.609, 0.900]	0.653 [0.453, 0.859]	0.801 [0.704, 0.880]
sg-LASSO-M	0.745 [0.605, 0.902]	0.731 [0.544, 0.894]	0.678 [0.521, 0.802]
		sg-LASSO-M $vs.$ LASSO-U	
p-value	0.318	0.255	0.097*
		sg-LASSO-M with $vs.$ without censored firms satisfying $C_i < t$	
p-value	0.459	0.290	0.072*

Notes: The second-to-last row reports the p-value from the pairwise difference test across the three prediction horizons, with the null hypothesis that the estimated AUC of sg-LASSO-MIDAS is superior to LASSO-UMIDAS's. The last row presents the p-value from the pairwise test across the three prediction horizons, comparing sg-LASSO-MIDAS applied to data with and without censored firms satisfying $C_i < t$. We use $*$ and $**$ to indicate 10% and 5% significance, respectively.

6 Conclusion

This paper presents a novel approach to corporate survival analysis, addressing the challenges of high-dimensional censored data sampled at both consistent and mixed frequencies.

The first major contribution is the introduction of the sparse-group LASSO estimator for high-dimensional censored MIDAS logistic regressions. This estimator effectively accommodates hierarchical data structures and facilitates model selection both within and across groups, unifying classical LASSO and group LASSO under a broader, more flexible framework.

Secondly, we develop the theory for logistic regression with high-dimensional censored data sampled at different frequencies. To extend the existing literature with assumptions on fixed design or isotropic conditions of the covariates, we develop the non-asymptotic properties of the proposed sparse-group LASSO estimator for censored, heavy-tailed data. This framework is readily extendable to generalized linear models with structured sparsity estimators. Furthermore, we consider the approximation error, which, to the best of our knowledge, is a novel contribution in the context of logistic regression. This error may arise from various sources, including approximations in the MIDAS weight function and/or deviations from exact sparsity.

A key practical contribution is an application to a comprehensive dataset of publicly traded Chinese manufacturing firms, integrating survival and censoring time information alongside numerous high-frequency financial covariates. Empirical findings indicate that sg-LASSO-MIDAS consistently outperforms unstructured LASSO approaches across various scenarios. Notably, the inclusion of censoring information significantly enhances prediction performance, providing valuable insights for predicting firm distress under real-world conditions.

Overall, the methodologies developed in this paper have broad applicability beyond corporate distress prediction. The integration of logistic models, MIDAS, and regularized machine learning techniques holds promise for applications in areas such as disease diagnosis, solvency evaluation, fraud detection, customer churn analysis, and labor market studies.

Acknowledgments

The authors thank seminar participants at Tilburg University for useful comments.

Funding sources

Wei Miao gratefully acknowledges financial support from the China Scholarship Council through the grant 202306130036. Jad Beyhum gratefully acknowledges financial support from the Research Fund KU Leuven through the grant STG/23/014. Jonas Striaukas gratefully acknowledges the financial support from the European Commission, MSCA-2022-PF Individual Fellowship, Project 101103508. Project Acronym: MACROML. Ingrid Van Keilegom acknowledges support from the FWO (Research Foundation Flanders) through the projects G0I3422N and G047524N.

Supplementary material

Online Appendix: Proof of Theorem 3.1, empirical dataset pre-processing algorithm, additional empirical and simulation results, MIDAS dictionaries, and details on the empirical dataset (.pdf file).

R package: R package 'Survivalml' that implements our method is available on https://github.com/Wei-M-Wei/Survivalml.

References

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, 22(3):1299–1327.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1):178–196.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., and Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management & Accounting*, 28(2):131–171.

Amico, M., Van Keilegom, I., and Han, B. (2020). Assessing cure status prediction from survival data using receiver operating characteristic curves. *Biometrika*, 108(3):727–740.

Audrino, F., Kostrov, A., and Ortega, J.-P. (2019). Predicting US bank failures with MIDAS logit models. *Journal of Financial and Quantitative Analysis*, 54(6):2575–2603.

Babii, A., Ball, R. T., Ghysels, E., and Striaukas, J. (2023). Machine learning panel data regressions with heavy-tailed dependent data: Theory and application. *Journal of Econometrics*, 237(2):105315.

Babii, A., Ghysels, E., and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Barbaglia, L., Manzan, S., and Tosetti, E. (2023). Forecasting loan default in europe with machine learning. *Journal of Financial Econometrics*, 21(2):569–596.

Barboza, F., Kimura, H., and Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417.

Beyhum, J., Centorrino, S., Florens, J.-P., and Van Keilegom, I. (2024a). Instrumental variable estimation of dynamic treatment effects on a duration outcome. *Journal of Business & Economic Statistics*, 42(2):732–742.

Beyhum, J., Tedesco, L., and Van Keilegom, I. (2024b). Instrumental variable quantile regression under random right censoring. *The Econometrics Journal*, 27(1):21–36.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397.

Blanche, P. F., Holt, A., and Scheike, T. (2023). On logistic regression with right censored data, with or without competing risks, and its use for estimating treatment effects. *Lifetime Data Analysis*, 29(2):441–482.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

Cai, T., Gerds, T. A., Zheng, Y., and Chen, J. (2011). Robust prediction of t-year survival with data from multiple studies. *Biometrics*, 67(2):436–444.

Cai, T., Pepe, M. S., Zheng, Y., Lumley, T., and Jenny, N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics*, 7(2):182–197.

Caner, M. (2023). Generalized linear models with structured sparsity estimators. *Journal of Econometrics*, 236(2):105478.

Cole, R. A. and White, L. J. (2012). Déjà vu all over again: The causes of us commercial bank failures this time around. *Journal of Financial Services Research*, 42:5–29.

Ding, A. A., Tian, S., Yu, Y., and Guo, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *Journal of the American Statistical Association*, 107(499):990–1003.

Duffie, D., Saita, L., and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3):635–665.

Ghysels, E., Kvedaras, V., and Zemlys-Balevičius, V. (2020). Mixed data sampling (MIDAS) regression models. In *Handbook of Statistics*, volume 42, pages 117–153. Elsevier.

Ghysels, E. and Qian, H. (2019). Estimating MIDAS regressions via OLS with polynomial parameter profiling. *Econometrics and Statistics*, 9:1–16.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2):59–95.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90.

Han, Y., Tsay, R. S., and Wu, W. B. (2023). High dimensional generalized linear models for temporal dependent data. *Bernoulli*, 29(1):105–131.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hung, H. and Chiang, C.-T. (2010). Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian Journal of Statistics*, 37(4):664–679.

James A. Hanley, B. J. M. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.

Jiang, C., Xiong, W., Xu, Q., and Liu, Y. (2021). Predicting default of listed companies in mainland china via U-MIDAS logit model with group lasso penalty. *Finance Research Letters*, 38:101487.

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: Current methods and applications. *BMC Medical Research Methodology*, 17:1–19.

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

Kim, M. H.-Y., Ma, S., and Zhou, Y. A. (2016). Survival prediction of distressed firms: Evidence from the chinese special treatment firms. *Journal of the Asia Pacific Economy*, 21(3):418–443.

Lee, M.-C. (2014). Business bankruptcy prediction based on survival analysis approach. *International Journal of Computer Science & Information Technology*, 6(2):103.

Li, C., Lou, C., Luo, D., and Xing, K. (2021). Chinese corporate distress prediction using lasso: The role of earnings management. *International Review of Financial Analysis*, 76:101776.

Li, S., Tian, S., Yu, Y., Zhu, X., and Lian, H. (2023). Corporate probability of default: A single-index hazard model approach. *Journal of Business & Economic Statistics*, 41(4):1288–1299.

Liang, X., Cohen, A., Heinsfeld, A. S., Pestilli, F., and McDonald, D. J. (2024). sparsegl: An *R* package for estimating sparse group lasso. *Journal of Statistical Software*, 110(6).

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):53–71.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., and Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, 36(3):1092–1113.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1):101–124.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Van De Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645.

Van De Geer, S. (2016). *Estimation and Testing Under Sparsity*. Springer International Publishing, Cham.

Van De Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zheng, Y., Cai, T., and Feng, Z. (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics*, 62(1):279–287.

Zhou, F., Fu, L., Li, Z., and Xu, J. (2022). The recurrence of financial distress: A survival analysis. *International Journal of Forecasting*, 38(3):1100–1115.