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

December 3, 2025

Abstract

This paper addresses the challenge of forecasting corporate bankruptcy, 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 model. Our approach handles censoring through inverse probability weighting and achieves accurate estimation with numerous mixed-frequency predictors by employing a sparse-group penalty. We derive finite sample bounds for the estimation error and demonstrate the superior performance of our method through Monte Carlo simulations. Finally, we apply our model to predict the financial distress of Chinese-listed firms.

Keywords: High-dimensional mixed-frequency censored data; Time-varying logit model; Sparse-group LASSO; Oracle inequality; Corporate survival analysis;

1 Introduction

Regulators and investors are increasingly focused on identifying vulnerable firms and developing accurate models to predict firm failures[1] well in advance, as the ability to correctly predict such failures could result in substantial financial savings. This has led to an extensive body of literature 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. These contemporary models have shown superior accuracy in predicting firm failures (Barboza et al., 2017).

While much of the existing literature focuses on predicting firm failure at a fixed time point, treating it as a binary outcome prediction problem, some studies have explored the application of survival analysis to this issue (Li et al., 2023). In contrast, this paper examines a different aspect by focusing on survival time T, specifically investigating whether firms can survive for t years, expressed as $\mathbb{1}\{T \leq t\}$, given that they have already survived s years.

Logistic regression, a powerful and well-established method for binary classification, serves as the foundation for our investigation into whether the logit model can be extended to handle high-dimensional censored bankruptcy data. This extension not only promises significant statistical improvements but also offers concrete economic advantages.

First, to address right-censored data when fitting logistic regression models, previous research has suggested weighting the estimating equations (Zheng et al., 2006) or the outcomes (Scheike et al., 2008). Building on these approaches, we develop a time-varying logit model. However, these traditional models often struggle to select a small set of relevant predictors from a vast pool of candidate variables, a challenge that becomes even more pronounced in high-dimensional financial data, where bankruptcy is often associated with numerous flow variables. Flow variables, by definition, are measured over time rather than at a specific point in time, adding complexity to the modeling process.

Second, financial flow variables are typically measured quarterly (high-frequency), whereas the bankruptcy status of a firm is a low-frequency event, occurring only once during the firm's lifetime. In many studies on distress prediction, annual financial variables are used with equal weighting across different quarters (e.g., Annual Return on Assets (ROA)). We introduce a Mixed-Data Sample (MIDAS) weighting scheme developed by Ghysels et al. (2007), which assigns different weights to each quarter, thus improving the fit and predictive power of the model (Audrino et al., 2019; Babii et al., 2022; Beyhum and Striaukas, 2023). The individual weights derived from the MIDAS scheme capture the relationship between bankruptcy status

[1] Other similar expressions include firm bankruptcy, financial distress, company insolvency, etc.

and past values of an independent covariate.

Third, high-dimensional mixed-frequency data regressions often involve specific data structures that, when properly accounted for, can enhance the performance of unrestricted estimators. These structures are represented by groups of lags for a single (high-frequency) covariate. To exploit this, we apply the sparse-group LASSO regularization technique, which effectively handles such structures (Simon et al., 2013). The sparse-group LASSO estimator's key advantage is its ability to combine both sparse and dense signals among covariates, offering improved model accuracy (Chen et al., 2020; Babii et al., 2023). Additionally, we establish a non-sharp oracle inequality for the sparse-group LASSO estimator in the context of a censored dataset.

Finally, to validate our methodology, we conduct a series of simulations designed to closely mimic the bankruptcy data of Chinese firms. We evaluate the performance of different models across various scenarios with differing survived time s and prediction horizons t. The simulation results strongly support the use of the proposed methodology, which integrates the MIDAS weighting scheme and the sparse-group LASSO penalty. In empirical analysis, our model outperforms traditional logistic regression models that do not incorporate MIDAS or data structure information, yielding superior predictions for both short-term and long-term horizons.

Outline The paper is organized as follows. Section 2 introduces the necessary notations used throughout the study. In Section 3, we first present the time-varying logistic regression model, followed by a discussion on employing the MIDAS weighting technique and incorporating group structure information among variables. This section culminates in introducing a new estimator, termed the sparse-group LASSO MIDAS estimator. Section 4 is dedicated to the analysis of the oracle inequality of the proposed estimator within the context of censored data. In Section 5, we present the results of the simulation studies. Finally, in Section 6, we construct a dataset on Chinese firm distress and assess the prediction performance of our model in the real dataset, a comparison with several other models is included.

2 Notation

For $p \in \mathbf{N}$, put $[p] = \{1, 2, \ldots, p\}$. 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 $q \in [1, \infty)$ and $|\boldsymbol{b}|_\infty = \max_{j\in[p]} |b_j|$ if $q = \infty$. For a group structure $\mathcal{G}$, the $\ell_{2,1}$ group norm of $\boldsymbol{b} \in \mathbb{R}^p$ is defined as $\|\boldsymbol{b}\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$. For a symmetric $p \times p$ matrix $\mathbf{A} = (a_{i,j})$, let $\operatorname{vech}(\mathbf{A}) \in \mathbb{R}^{p(p+1)/2}$ be its vectorization consisting of the lower triangular and diagonal elements and we denote its entry wise max norm $|\mathbf{A}|_\infty = |\operatorname{vech}(\mathbf{A})|_\infty = \max_{i,j} |a_{ij}|$ and smallest eigenvalue $\lambda_{\min}(\mathbf{A})$. 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 $-J$. Let $\mathcal{G}$ be a partition of $[p]$ defining the group structure, for a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, the sparse-group structure is described by a pair $(S_{\boldsymbol{\beta}}, \mathcal{G}_{\boldsymbol{\beta}})$, where $S_{\boldsymbol{\beta}} = \{j \in [p] : \beta_j \neq 0\}$ and $\mathcal{G}_{\boldsymbol{\beta}} = \{G \in \mathcal{G}_{\boldsymbol{\beta}} : \boldsymbol{\beta}_G \neq 0\}$ are the support and respectively the group support of $\boldsymbol{\beta}$. We use $|S|$ to denote the cardinality of a set S. Define $\sqrt{s_{\boldsymbol{\beta}}} = \alpha\sqrt{|S_{\boldsymbol{\beta}}|} + (1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|}$ as the cardinality of this pair, where α is a control parameter. 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 an (sufficiently large) absolute constant v such that $a_N \leq v b_N$ for all $N \geq 1$ and $a_N \sim b_N$ if $a_N \lesssim b_N$ and $b_N \lesssim a_N$.

3 High-dimensional Censored MIDAS Logistic Regression

3.1 Time-Varying Logistic Regression Model

In corporate survival analysis, we define the random variable T as the survival time of a firm, measured from the firm's Initial Public Offering date (IPO) to the occurrence of financial distress day.[2] The variable T is right censored by the censoring time C denoting the time difference between the IPO date and the censoring event, which occurs at the end of the follow-up period.[3,4] Our primary interest 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. Specifically, we assume that the survival time T follows the logistic regression model:

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

where $\boldsymbol{Z}$ is the covariate vector and $\tilde{\boldsymbol{\beta}}_0(t,s)$ is the true parametric parameter which is related to t and s. It's worth mentioning that in practice we are interested in several fixed s and t. For simplicity, we use $\tilde{\boldsymbol{\beta}}_0$ as shorthand for $\tilde{\boldsymbol{\beta}}_0(t,s)$.

Before presenting the model assumptions, we first introduce the following notation: $\widetilde{T} = T \wedge C$, $\delta(t) = \mathbb{1}\{C \geq t \wedge T\}$ and $H(t) = P(C \geq t)$ which is the survival function of the censoring time C. We aim to estimate the true parameter $\tilde{\boldsymbol{\beta}}_{\mathbf{0}}$ of the distribution defined in (1). To do this, we then impose the following assumptions on T and C:

Assumption 3.1. *(Independent censoring) C is independent of T and X, that is $C \perp\!\!\!\perp (T, \boldsymbol{Z})$. Furthermore, we have data $T_i, C_i, \boldsymbol{Z}_i \overset{IID}{\sim} T, C, \boldsymbol{Z}, \forall i \in [N]$ and the covariate $\boldsymbol{Z}_i$ has a finite variance.*

[2] Initial Public Offering Date: The date on which the firm's stock is first offered to the public.

[3] Since companies are not listed as soon as they are created, T in our context is slightly different from the survival time typically considered in traditional survival analysis (Li et al., 2023).

[4] In the firm distress prediction context, there are no firms that drop out during the observation period.

We argue Assumption 3.1 is reasonable in the context of corporate survival analysis since the censoring time of a firm only depends on the observation period and no firms drop out during that period.

Assumption 3.2. *(Sufficient follow-up) Given the prediction horizon t, we have $P\left(\widetilde{T} \geq t\right) \geq C_r$, where C_r is a positive constant. Furthermore, we have $\widetilde{T} \geq s$.*

Remark 1. *Through Assumption 3.2, we find $H(t \wedge T) = P(C \geq t \wedge T) \geq P(C \geq t) \geq P\left(\widetilde{T} \geq t\right) \geq C_r$. Then we see that $H(t \wedge T) > 0$, which ensures that the formula in (15) is well defined since the denominator $H(t \wedge T)$ cannot be 0. $\widetilde{T} \geq s$ ensures the duration of the observation period should be larger than s, which means we have at least s years' information of the firm.*

Based on Assumptions 3.1 and 3.2, the true parameter $\tilde{\boldsymbol{\beta}}_0$ of the distribution (1) could be achieved by solving the population conditional maximum likelihood problem

$$\tilde{\boldsymbol{\beta}}_0 = \operatorname{argmax}_{\tilde{\boldsymbol{\beta}}} \mathbb{E}\left[\mathbb{1}\{T \leq t\} \log p\left(\boldsymbol{Z}\right) + \left(1 - \mathbb{1}\{T \leq t\}\right) \log\left(1 - p\left(\boldsymbol{Z}\right)\right) \mid T \geq s\right], \qquad (2)$$

where $p\left(\boldsymbol{Z}\right) = \frac{\exp\left(\boldsymbol{Z}^\top \tilde{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{Z}^\top \tilde{\boldsymbol{\beta}}\right)}$ and the MLE (Maximum likelihood estimation) estimator can be obtained through maximum likelihood estimation. However, we cannot fully observe the indicator $\mathbb{1}\{T \leq t\}$ in (2) for all firms in the sample due to the censoring problem. To address this issue, we rely on Corollaries A.1, A.2 and A.3 (see Appendix A), allowing us to obtain $\tilde{\boldsymbol{\beta}}_0$ by solving

$$\operatorname{argmax}_{\tilde{\boldsymbol{\beta}}} \mathbb{E}\left[\mathbb{1}\{\widetilde{T} \geq s\}\left(\frac{\delta(t)\mathbb{1}\{\widetilde{T} \leq t\}}{H(t \wedge \widetilde{T})}\boldsymbol{Z}^\top \tilde{\boldsymbol{\beta}} - \log(1 + \exp(\boldsymbol{Z}^\top \tilde{\boldsymbol{\beta}}))\right)\right],$$

where $\delta(t) = \mathbb{1}\{C \geq t \wedge T\} = 1 - \mathbb{1}\{\widetilde{T} \leq t\}\left(1 - \mathbb{1}\{T \leq C\}\right)$. Then the estimator is obtained from:

$$\operatorname{argmin}_{\tilde{\boldsymbol{\beta}}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\widetilde{T}_i \geq s\}\left(-\frac{\delta_i(t)\mathbb{1}\{\widetilde{T}_i \leq t\}}{\widehat{H}\left(t \wedge \widetilde{T}_i\right)}\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}} + \log\left(1 + \exp(\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}})\right)\right), \qquad (3)$$

where $\widehat{H}\left(t \wedge \widetilde{T}\right)$ serves as a consistent estimator of $H\left(t \wedge \widetilde{T}\right)$. The method we employ here is Outcome Weighted Inverse Probability of Censoring Weighting (OIPCW). A similar approach for addressing the censoring problem is called Inverse Probability Weighting (IPW) (Horvitz and Thompson, 1952; Zheng et al., 2006; Beyhum et al., 2024). Further details regarding both OIPCW and IPW can be found in Blanche et al. (2023).

3.2 MIDAS: Mixed-Frequency Data Sampling

A substantial body of literature (Zheng et al., 2006; Scheike et al., 2008) examined both the estimator (3) and the IPW-type estimator, both of which have demonstrated considerable effectiveness in traditional survival analysis. However, applying these methods to firm distress prediction reveals several limitations:

- In firm distress prediction, the outcome (i.e., financial distress or bankruptcy) and covariates are often sampled at differing frequencies. Distress events are low-frequency occurrences, typically happening only once in a firm's lifetime, whereas financial covariates that inform the prediction are often available at higher frequencies, such as annually or quarterly. This mismatch in sampling frequencies introduces challenges for accurate modeling.

- Traditional survival analysis generally incorporates a limited number of covariates. Prior research Zheng et al. (2006) highlighted that using the IPW-logit estimator, similar to (3), presents challenges in selecting a small, relevant subset of markers from a large pool of candidate variables, especially in medical studies. In distress prediction, this issue is intensified by the extensive range of available financial variables. Estimating models with high-dimensional, censored data thus remains a complex problem.

- Relationships among covariates are often disregarded in estimation. For instance, a financial variable measured at time s and its lagged values may show similar predictive patterns for firm distress. Additionally, certain financial variables, such as Return on Assets (ROA) and Return on Equity (ROE) (Audrino et al., 2019), share common properties and could be grouped accordingly. Nevertheless, existing literature has not adequately considered such group structures when using the estimator (3) or IPW-type estimator.

This section addresses the challenge of handling high-dimensional mixed-frequency censoring data in firm distress prediction. Specifically, we examine a sample in which all N firms have already survived at least s years. The covariate x is observed m times in the period between year $s-1$ and s. We assume that the financial distress status $\mathbb{1}\{T \leq t\}$, is affected by s years of lagged $\left\{x_{s-\frac{j-1}{m}}, j \in [d]\right\}$, where $d = s \times m$ represents the total number of lags. More generally, let there be K groups of covariates $\left\{x_{s-\frac{j-1}{m},k}, j \in [d]\right\}_{k\in[K]}$. Each group covariate is defined as $\boldsymbol{Z}_{i,k} = \left(x_{i,s,k}, x_{i,s-\frac{1}{m},k}, \ldots, x_{i,s-\frac{d}{m},k}\right)$, $i \in [N], k \in [K]$, which is measured at some higher frequency with d observations. We then reconsider the term $\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}}_0$

in the distribution function (1) as

$$\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}}_0 = \sum_{k=1}^{K} \psi\left(L^{1/d}; \tilde{\boldsymbol{\beta}}_{0,k}\right) x_{i,s,k}, i \in [N],$$

where $\psi\left(L^{1/d}; \tilde{\boldsymbol{\beta}}_{0,k}\right) x_{i,s,k} = \frac{1}{d}\sum_{j=1}^{d} \tilde{\beta}_{0,j,k} x_{i,s-\frac{j-1}{m},k}$ is a high-frequency lag polynomial. Note that the lag polynomial $\psi\left(L^{1/d}; \tilde{\boldsymbol{\beta}}_{0,k}\right) x_{i,s,k}$ involves the same d number of high-frequency lags for each covariate $k \in [K]$, which is done for the sake of simplicity and can easily be relaxed.

Undoubtedly, when dealing with numerous covariates sampled at high frequencies, the total number of parameters to be estimated can become significantly larger than the effective sample size, a common issue in financial distress prediction. This imbalance often results in poor estimation and reduced out-of-sample prediction accuracy in finite samples. For instance, as the real dataset demonstrated in Section 6.1, all the firms in the dataset have survived $s = 6$ years, and the $K = 95$ covariates are measured $m = 4$ times per year (e.g., in a yearly or quarterly setting). This configuration necessitates the estimation of $6 \times 4 \times 95 + 1 = 2280$ parameters, including the intercept. However, if the sample contains fewer than 2280 observations, the curse of dimensionality arises, not only complicating computations but also making them time-consuming and costly.

To mitigate this challenge, the LASSO estimator (Tibshirani, 1996) provides an effective method for parameter estimation in high-dimensional settings. By enforcing sparsity, LASSO selects a relevant subset of covariates, leading to simpler, more interpretable models and improved prediction accuracy. In this paper, we extend this approach by applying the sparse-group LASSO (Simon et al., 2013), which incorporates the group structure among covariates into the estimation process.

Additionally, to handle mixed-frequency data, we begin by parameterizing the high-frequency lag polynomial using the Mixed-Data Sampling (MIDAS) regression framework (Ghysels et al., 2006) as follows:

$$\psi\left(L^{1/d}; \tilde{\boldsymbol{\beta}}_{0,k}\right) x_{i,s,k} = \sum_{l=1}^{L} \beta_{0,k,l} w_l \left(\frac{j-1}{d}\right) x_{i,s-\frac{j-1}{m},k} + E_i, i \in [N], k \in [K], j \in [d], \quad (4)$$

where $\boldsymbol{\beta}_{0,k} = (\beta_{0,k,1}, \beta_{0,k,2}, \cdots, \beta_{0,k,L})^\top$ is L-dimensional vector of coefficients with $L \leq d$, $\{w_l : l = 1, \ldots, L\}$ constitutes a collection of functions known as the dictionary W and E_i is the approximation error. $\omega : [0,1] \times \mathbb{R}^L \to \mathbb{R}$ represents a weight function. Finally, this approach reduces the number of parameters to be estimated from $K \times d + 1$ to $K \times L + 1$, offering a significant dimensionality reduction.

The simplest form of the dictionary W consists of algebraic power polynomials, also known as Almon polynomials (Almon, 1965). More broadly, the dictionary may include arbitrary approximating functions, such as the classical orthogonal bases of $L_2[0,1]$.[5] Using orthogonal polynomials, in particular, tends to reduce multicollinearity and improve performance in finite samples. Notably, specifying dictionaries deviates from standard MIDAS regressions and results in a computationally appealing convex optimization problem (Babii et al., 2022).

Furthermore, all high-frequency lags (or the approximating functions used to parameterize the lag polynomial) of a single covariate are treated as a group. While other grouping structures, such as combining related covariates like Return on Assets (ROA) and Return on Equity (ROE)—could be explored, this paper adopts the simplest group structure as outlined.

We specifically emphasize the use of the sparse-group LASSO (Simon et al., 2013) to incorporate the group structure into the estimation procedure. Unlike the group LASSO (Yuan and Lin, 2006), which enforces sparsity solely between groups, the sparse-group LASSO encourages sparsity both within and between groups. This dual sparsity allows us to efficiently capture predictive information from each group, enhancing the model's ability to identify relevant covariates.

3.3 Estimator with High-dimensional Censored Mixed-Frequency Data

Let $x_{i,s,k}$ represent the k-th covariate of firm i measured at time s, and we use $\boldsymbol{Z}_{i,k} = \left(x_{i,s-\frac{j-1}{m},k}\right)_{j\in[d]} \in \mathbb{R}^{1\times d}$ to store the k-th covariate measured at some higher frequency with d observations before time s. Let $\boldsymbol{X}_i = (\boldsymbol{Z}_{i,1}W, \boldsymbol{Z}_{i,2}W, \ldots, \boldsymbol{Z}_{i,K}W)^\top \in \mathbb{R}^{KL\times 1}, i \in [N]$ be the design matrix, where $W = \left(w_l\left(\frac{j-1}{d}\right)/d\right)_{j\in[d],l\in[L]}$ is a $d \times L$ weight matrix. We then see the distribution (1) becomes

$$P(T \leq t \mid \boldsymbol{X}, T \geq s) = \frac{\exp\left(\boldsymbol{X}^\top \boldsymbol{\beta}_0 + \boldsymbol{E}\right)}{1 + \exp\left(\boldsymbol{X}^\top \boldsymbol{\beta}_0 + \boldsymbol{E}\right)}, \tag{5}$$

where $\boldsymbol{\beta}_0 = \left(\boldsymbol{\beta}_{0,1}, \boldsymbol{\beta}_{0,2}, \cdots, \boldsymbol{\beta}_{0,K}\right)^\top$ and $\boldsymbol{E} = (E_1, E_2, \cdots, E_N)^\top$. Furthermore, let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_K^\top\right)^\top$, where $\boldsymbol{\beta}_k \in \mathbb{R}^L, k \in [K]$ denotes the parameters of the high-frequency lag polynomial $\boldsymbol{Z}_{i,k}W$. Notice that the high-frequency parameters $\boldsymbol{\beta}_k, k \in [K]$ are related to t and s. The proposed sparse-group LASSO estimator $\hat{\boldsymbol{\beta}}$ solves the following problem

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\widetilde{T}_i \geq s\} \left(- \frac{\delta_i(t) \mathbb{1}\{\widetilde{T}_i \leq t\}}{\widehat{H}\left(t \wedge \widetilde{T}_i\right)} \boldsymbol{X}_i^\top \boldsymbol{\beta} + \log\left(1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})\right) \right) + \lambda \Omega(\boldsymbol{\beta}), \tag{6}$$

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

where the sparse group LASSO penalty is: $\Omega(\boldsymbol{\beta}) = \alpha|\boldsymbol{\beta}|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_{2,1}$, $\|\boldsymbol{\beta}\|_{2,1} = \sum_{G\in\mathcal{G}} |\boldsymbol{\beta}_G|_2$ is the group LASSO norm and $\mathcal{G}$ is a group structure (partition of $[g]$) among all the covariates. Notice that we have the intercept term in the simulation and empirical application, but we don't penalize it. The amount of penalization in (6) is controlled by the regularization parameter $\lambda \geq 0$ while $\alpha \in [0,1]$ is a weight parameter that determines the relative importance of the sparsity and the group structure. Setting $\alpha = 1$, we obtain the LASSO estimator while setting $\alpha = 0$, leads to the group LASSO estimator, which is reminiscent of the elastic net.

The survival function $H(\cdot)$ of the censroing time C which does not depend on covariates (Assumption 3.1), can be estimated by the Kaplan Meier estimator (Kaplan and Meier, 1958):

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

with $N(j) = \sum_{i=1}^{N} \mathbb{1}\left\{\widetilde{T}_i \leq j, \delta_i(j) = 0\right\}, dN(j) = N(j) - \lim_{j'\to j, j'<j} N(j')$ which denotes the jump of the process N at the time j or the number of events (in risk) that happened at time j, and $\widetilde{T}(j) = \sum_{i=1}^{N} \mathbb{1}\left(\widetilde{T}_i \geq j\right)$ which means firms known to have been not in risk up to time j.

4 Oracle Inequality

In this section, we outline the main assumptions for the proposed estimator (6) and analyze the estimation properties of the sparse-group LASSO estimator within the context of censoring data. Both the LASSO and the group LASSO estimators are covered. For simplicity, we consider a sample in which all N firms have already survived s years here, and we focus on the estimator $\hat{\boldsymbol{\beta}}$ in our theoretical analysis, omitting the intercept term, as the intercept is not penalized in our model.

Definition 1. *(Risk Function)*

$$\begin{aligned} R(\boldsymbol{\beta}) &= \mathbb{E}\left[\mathbb{1}\{\widetilde{T} \geq s\}\left(-\frac{\delta(t)\mathbb{1}\{\widetilde{T} \leq t\}}{H(t\wedge\widetilde{T})}\boldsymbol{Z}^\top\tilde{\boldsymbol{\beta}} + \log(1+\exp(\boldsymbol{Z}^\top\tilde{\boldsymbol{\beta}}))\right)\right] \\ &= \mathbb{E}\left[\mathbb{1}\{\widetilde{T} \geq s\}\left(-\frac{\delta(t)\mathbb{1}\{\widetilde{T} \leq t\}}{H(t\wedge\widetilde{T})}\left(\boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{E}\right) + \log(1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{E}))\right)\right], \end{aligned}$$

and

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

where $\boldsymbol{E}$ is the approximation error.

Assumption 4.1. *(Data) We have i.i.d data T_i, C_i, and*

$$\boldsymbol{X}_i = (\boldsymbol{Z}_{i,1}W, \boldsymbol{Z}_{i,2}W, \ldots, \boldsymbol{Z}_{i,K}W)^\top = (X_{i,1}, \ldots, X_{i,j}, \ldots, X_{i,p})^\top \in \mathbb{R}^p, i \in [N]$$

satisfies $\max\limits_{|u|_2=1} \mathbb{E}\left(\left|\boldsymbol{X}_i^\top u\right|^q\right) \leq K_0 < \infty$, where $p = KL$ and $q \geq 4$. Additionally, W is orthogonal which means $WW^T = I$, where I is an identify matrix.

Remark 2. *The q-th moment of a linear combination of random variables will control the moments of the individual variables, including their pairwise products. It's clear to see that we both have $\max\limits_{1\leq j\leq p} \mathbb{E}\left(|X_{i,j}|^q\right) \leq K_0$ and $\max\limits_{1\leq j,l\leq p} \mathbb{E}\left(\left|X_{i,j}X_{i,l}\right|^{\frac{q}{2}}\right) \leq \max\limits_{1\leq j,l\leq p} \mathbb{E}\left(\left|\frac{X_{i,j}+X_{i,l}}{\sqrt{2}}\right|^q\right) \leq K_0$.*

Compared to Van De Geer (2016), they only allowed for the data with $\max\limits_{1\leq j\leq p} \mathbb{E}\left(\exp\left(|X_{i,j}|^2\right)\right) < \infty$, which excludes the heavy-tailed distributions. Assumption 4.1 may allow for heavy-tailed distributions commonly encountered in financial and economic time series, for example, asset returns and volatilities.

Assumption 4.2. *(Restricted Eigenvalue) There exists a constant $\gamma_{\mathrm{H}} > 0$ such that the minimum eigenvalue*

$$\lambda_{\min}\left(\mathbb{E}\left[\frac{\exp(\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}}_0)}{\left(1+\exp(\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}}_{\boldsymbol{0}})\right)^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right]\right) \geq \gamma_{\mathrm{H}}.$$

Remark 3. *Since W is orthogonal and $\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}}_0 = \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i$, we have $\frac{\exp(\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}}_0)}{\left(1+\exp(\boldsymbol{Z}_i^\top \tilde{\boldsymbol{\beta}}_{\boldsymbol{0}})\right)^2}\boldsymbol{Z}_i\boldsymbol{Z}_i^\top = \frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top$. It follows*

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

Assumption 4.2 is similar to the compatibility condition in Van De Geer (2008, 2016) and was used in Caner (2023); Han et al. (2023). In linear regression, the restricted eigenvalue condition implies the restricted strong convexity (RSC) of the loss function which underpins the statistical guarantee of the M-estimator. However, in the logistic regression model, the Hessian matrix of the loss function depends on $\boldsymbol{\beta}$, creating some technical difficulty verifying the RSC. In the context of LASSO, Han et al. (2023) developed a local RSC (LRSC) within a neighborhood of $\boldsymbol{\beta}_0$, which is closely related to the quadratic margin condition of $R(\boldsymbol{\beta})$ in Caner (2023); Van De Geer (2016). It's worth mentioning that when there is no approximation error E, Van De Geer (2008); Caner (2023); Beyhum and Portier (2024) also additionally assumed that

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

is satisfied for $\forall\tilde{\boldsymbol{\beta}}$ within a neighborhood of $\tilde{\boldsymbol{\beta}}_0$, which can directly imply the quadratic margin condition. Here, we relax such assumption and provide another way that proves the quadratic margin condition is satisfied on the conditional risk function $R(\boldsymbol{\beta}|\boldsymbol{X})$.

Assumption 4.3. *(Regularization) $\forall\epsilon_1\geq 0$, $\exists U_{\epsilon_1}>0$, suppose that*

$$\lambda\sim K_C\left(\frac{\sqrt{\log p}}{\sqrt{N}}\vee\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}}\vee\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}}\vee U_{\epsilon_1}\frac{p^{\frac{2}{q}}}{\sqrt{N}}\right),$$

where K_C is a constant related to K_0 and C_r. C_r and K_0 are as in Assumption 3.2 and 4.1.

The regularization parameter λ is determined by the first term $\frac{\sqrt{\log p}}{\sqrt{N}}$ that resembles the well-known sub-Gaussian rate, the second term and third term with a polynomial term of p represents the possible heavy tail of the data and the last term $\frac{p^{\frac{2}{q}}}{\sqrt{N}}$ shows the effect of the censoring issue together with the heaviness of the tail. It is worth mentioning that under Assumption 3.1 and 3.2, the convergence rate of the Kaplan-Meier estimator $\widehat{H}\left(t\wedge\widetilde{T}\right)$ is $O_P\left(\frac{1}{\sqrt{N}}\right)$.

Assumption 4.4. *(Candidate Oracle) The candidate oracle $\boldsymbol{\beta}$ that satisfies*

$$\Omega\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\leq 36\frac{s_{\boldsymbol{\beta}}\lambda}{\gamma_{\mathrm{H}}}+16\lambda^{-1}\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right),$$

and

$$\frac{2s_{\boldsymbol{\beta}}\lambda}{\gamma_{\mathrm{H}}}+(\frac{144}{\gamma_{\mathrm{H}}}+\frac{288\bar{K}_0}{\gamma_{\mathrm{H}}^2})s_{\boldsymbol{\beta}}^2p^{\frac{2}{q}}\lambda+(\frac{64}{\gamma_{\mathrm{H}}}+\frac{128\bar{K}_0}{\gamma_{\mathrm{H}}})\frac{s_{\boldsymbol{\beta}}p^{\frac{2}{q}}}{\lambda}\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)\leq 1,$$

where $\sqrt{s_{\boldsymbol{\beta}}}=\alpha\sqrt{|S_{\boldsymbol{\beta}}|}+(1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|}$ and $(S_{\boldsymbol{\beta}},\mathcal{G}_{\boldsymbol{\beta}})$ are the support and the group support of the candidate oracle $\boldsymbol{\beta}$. $\bar{K}_0$ is a constant related to K_0, γ_{H} and K_0 are as in Assumption 4.2 and 4.1.

Assumption 4.4 allows us not to impose strict sparsity on the true parameter $\boldsymbol{\beta}_0$, and the candidate oracle $\boldsymbol{\beta}$ is trading off approximation error between $R(\boldsymbol{\beta}|\boldsymbol{X})$ and $R(\boldsymbol{\beta}_0|\boldsymbol{X})$. Notice that the choice of $\boldsymbol{\beta}$ can be the true parameter $\boldsymbol{\beta}_0$ which will substantially simplify the expression.

Theorem 4.1. *Let $\boldsymbol{\beta}_0$ be the true parameter and $\hat{\boldsymbol{\beta}}$ is the proposed estimator (6). Under Assumption 3.1, 3.2, 4.2 ,4.1, 4.3 and 4.4, we have*

$$\Omega(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \lesssim \frac{s_{\boldsymbol{\beta}}\lambda}{\gamma_{\mathrm{H}}} + \lambda^{-1}\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right) + \lambda^{-1}\frac{1}{N}\sum_{i=1}^{N}|E_i|, \tag{7}$$

and

$$R(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) - R(\boldsymbol{\beta}|\boldsymbol{X}) \lesssim \frac{s_{\boldsymbol{\beta}}\lambda^2}{\gamma_{\mathrm{H}}} + R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right) + \frac{1}{N}\sum_{i=1}^{N}|E_i|$$

hold with probability at least $1 - \epsilon_1 - \frac{2}{p} - \frac{c}{\log p}$. ϵ_1 is as in Assumption 4.3 which is related to the censoring issue in our model and c is a universal constant.

Remark 4. *(Effects of censoring data) The constant K_C in the regularization parameter λ is negatively related to C_r, where C_r is related to the censoring issue, see Assumption 3.2 and 4.3. When the data has a higher censoring rate, it means a lower C_r, and then leads to a higher K_C. Consequently, the convergence rate of the proposed estimator becomes slow.*

Remark 5. *It is important to note that, although we focus on the logistic regression model with a specific sparse-group LASSO penalty in this analysis, the results can be readily extended to the Generalized Linear Model (GLM) with a structured sparsity penalty. Moreover, to the best of our knowledge, the assumptions we impose on the covariates and the risk function are the weakest, even in the absence of the censoring problem.*

Theorem 4.1 provides the oracle inequality with the proposed sg-LASSO estimator (6). When the tuning parameter $\alpha = 1$ is in the sparse group LASSO penalty, Theorem 4.1 can be reduced to Proposition 3.1 of Han et al. (2023) without considering the censoring issue. $\sqrt{s_{\boldsymbol{\beta}}} = \alpha\sqrt{|S_{\boldsymbol{\beta}}|} + (1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|}$ is the effective sparsity constant (Babii et al., 2022) and the constant reflects the benefits of the sparse-group structure for the sg-LASSO estimator that comes from $\alpha\sqrt{|S_{\boldsymbol{\beta}}|} + (1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|} \leq \sqrt{|S_{\boldsymbol{\beta}}|}$.

Remark 6. *The main challenges in proving Theorem 4.1 stems from establishing an empirical process result that links $R_N(\boldsymbol{\beta})$ and $R(\boldsymbol{\beta}|\boldsymbol{X})$ while accounting for the censoring issue. By imposing Assumption 4.2 solely on $\boldsymbol{\beta}_0$, we construct a lower bound for the conditional risk function $R(\boldsymbol{\beta}|\boldsymbol{X})$ in the neighborhood of $\boldsymbol{\beta}_0$. This bound is related to the sample effective sparsity, rather than the population effective sparsity typically assumed in the literature. This distinction complicates the proof of the oracle inequality, as it requires demonstrating that under certain assumptions, the sample effective sparsity approximates the population effective sparsity with high probability. Additionally, our framework accommodates the choice of a candidate oracle. Further details of the proof are provided in Appendix B.*

Assumption 4.5. *we have $\frac{1}{N}\sum_{i=1}^{N}|E_i| = O_P(s_{\boldsymbol{\beta}_0}\lambda^2)$.*

Since the $|\cdot|_1$-norm is equivalent to the Ω-norm whenever groups have fixed size, we deduce from Theorem 4.1 that

Corollary 4.1. *Let the candidate oracle $\boldsymbol{\beta}$ be the true parameter. Suppose that Assumption 3.1, 3.2, 4.2 ,4.1, 4.3, 4.4 and 4.5 hold, when $N, p \to \infty$, then*

$$\left|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right|_1 = O_P\left(\frac{s_{\boldsymbol{\beta}_0}\sqrt{\log p}}{\sqrt{N}} \vee \frac{s_{\boldsymbol{\beta}_0} p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}} \vee \frac{s_{\boldsymbol{\beta}_0} p^{\frac{2}{q}} \log p}{N^{1-\frac{2}{q}}}\right),$$

and

$$R(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X}) = O_P\left(\frac{s_{\boldsymbol{\beta}_0}\log p}{N} \vee \frac{s_{\boldsymbol{\beta}_0} p^{\frac{8}{q}}\log p}{N} \vee \frac{s_{\boldsymbol{\beta}_0} p^{\frac{8}{q}} \log^2 p}{N^{2-\frac{4}{q}}}\right).$$

Remark 7. *Suppose the data is sub-Gaussian with finite moments of all orders, which means $q \to \infty$. When p increases slower than N, we then see the well-known sub-Gaussian rate $\frac{s_{\boldsymbol{\beta}_0}\sqrt{\log p}}{\sqrt{N}}$ of the sg-LASSO estimator.*

5 Simulation

In this section, we evaluate the predictive performance of three methods through simulations:

- LASSO-UMIDAS: An unstructured LASSO estimator without unrestricted MIDAS weights.
- LASSO-MIDAS: An unstructured LASSO estimator using MIDAS weights.
- 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. Simulation results underscore the method's strengths, particularly in achieving superior prediction accuracy with finite sample data.

5.1 Simulation Design

Given a time of interest t, our goal now is to generate the survival time T that follows the distribution

$$P(T \leq t \mid \widetilde{\boldsymbol{X}}, T \geq s) = \frac{\exp\left(\widetilde{\boldsymbol{X}}\boldsymbol{\theta}(\boldsymbol{t},\boldsymbol{s})\right)}{1+\exp\left(\widetilde{\boldsymbol{X}}\boldsymbol{\theta}(\boldsymbol{t},\boldsymbol{s})\right)}. \tag{8}$$

To do this, we first generate T using the following formula

$$T = s + \exp\left(\frac{\log(\frac{\zeta}{1-\zeta}) - \widetilde{\boldsymbol{X}}\boldsymbol{\theta}}{\widetilde{\boldsymbol{X}}\boldsymbol{l}}\right), \tag{9}$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_K)^\top$, $\boldsymbol{l} = (1, 1, \ldots, 1)^\top$ and $\zeta \sim \text{Uniform}(0, 1)$. θ_0 is the intercept and $(\theta_1, \ldots, \theta_K)$ represents the high-frequency lag polynomial coefficients. $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2, \ldots, \widetilde{\boldsymbol{X}}_N)^\top$ and $\widetilde{\boldsymbol{X}}_i = (1, \widetilde{Z}_{i,1}, \ldots, \widetilde{Z}_{i,K})^\top$, we add K high-frequency lag polynomials $\widetilde{Z}_{i,k}$ which are generated by

$$\widetilde{Z}_{i,k} = \frac{1}{d}\sum_{j=1}^{d} \omega\left(\frac{j-1}{d}; \boldsymbol{\beta}_k\right) x_{i,s-\frac{j-1}{m},k}, i \in [N], k \in [K]. \tag{10}$$

All the observations in the simulated dataset have survived at least s years, and we are interested in yearly/quarterly frequency $m = 4$, then for each high-frequency lag polynomial $\widetilde{Z}_{i,k}$, we have $d = s \times m$ quarter lags' information to use. In (10), only $\widetilde{Z}_{i,1}$ and $\widetilde{Z}_{i,2}$ have nonzero coefficients which rely on a commonly used weighting scheme: $\omega(u; \boldsymbol{\beta}_k)$ for $k = 1, 2$ are determined by beta densities, respectively, equal to $Beta(1, 3)$, $Beta(2, 3)$; see Ghysels et al. (2007); Ghysels and Qian (2019); Babii et al. (2022), for further details. The left $K - 2$ lag polynomials $\widetilde{Z}_{i,k}, k \in [3, \ldots, K]$ are all noisy high-frequency lag polynomials.[6]

Notice that the reverse of times series $\{x_{i,s-\frac{j-1}{m},k}\}$ is $\{x_{i,\frac{j}{m},k}\}, j \in [d]$. For the generation of $x_{i,\frac{j}{m},k}$, we first initiate the processes $(x_{i,\frac{1}{m},1}, \ldots, x_{i,\frac{1}{m},K}) \sim N\left(\mathbf{0}, \Sigma(1-\rho^2)\right)$ with the covariance matrix $\Sigma_{u,v} = \rho_0^{|u-v|}, u, v \in [K]$. Then, the high-frequency covariates $x_{i,\frac{j}{m},k}, j \in [d], k \in [K]$ are generated as the following 3 scenarios:

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

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

In addition, we consider one more scenario that allows the covariates to have heavy tails. As same as before, we first initiate the processes $(x_{i,\frac{1}{m},1}, \ldots, x_{i,\frac{1}{m},K}) \sim$ student-$t(10)$ with the covariance matrix $\Sigma_{u,v} = \rho_0^{|u-v|}, u, v \in [K]$. The high-frequency covariates are following

- Scenario 3: $x_{i,\frac{j}{m},k} = \rho x_{i,\frac{j-1}{m},k} + \nu_{i,k}, k \in [K], j \in \{2, 3, \ldots, d\}$, and $(\nu_{i,1}, \ldots, \nu_{i,K}) \sim_{\text{i.i.d}}$ student-t with degree 10 and its covariance matrix $\Sigma(1-\rho^2)\frac{\text{degree}}{\text{degree}-2}$. $\rho = 0.1$ and $\rho_0 = 0.1$.

Before estimation, we transform these covariates to their absolute values to ensure that $\widetilde{\boldsymbol{X}}\boldsymbol{l}$ in the second row of (12) remains strictly positive. For the k-th variable, ρ regulates the degree

[6] The true coefficients of these noisy high-frequency lag polynomials are all zero.

of time series dependence among its lags, while ρ_0 represents the level of cross-sectional dependence across all variables.

Finally, the probability function of the generating survival time T in (9) is

$$P(T \leq t \mid \widetilde{\boldsymbol{X}}, T \geq s) = \frac{\exp\left(\widetilde{\boldsymbol{X}}\boldsymbol{\theta}(\boldsymbol{t},\boldsymbol{s})\right)}{1+\exp\left(\widetilde{\boldsymbol{X}}\boldsymbol{\theta}(\boldsymbol{t},\boldsymbol{s})\right)}, \tag{11}$$

where $\boldsymbol{\theta}(\boldsymbol{t},\boldsymbol{s}) = (\theta_0 + \log(t-s), \theta_1 + \log(t-s), \theta_2 + \log(t-s), 0, \ldots, 0)^\top$. This probability function is followed by (9) and $\zeta \sim U(0,1)$

$$\begin{aligned} T \leq t &\Leftrightarrow s + \exp\left(\frac{\log(\frac{\zeta}{1-\zeta}) - \widetilde{\boldsymbol{X}}\boldsymbol{\theta}}{\widetilde{\boldsymbol{X}}\boldsymbol{l}}\right) \leq t \\ &\Leftrightarrow \frac{\log(\frac{\zeta}{1-\zeta}) - \widetilde{\boldsymbol{X}}\boldsymbol{\theta}}{\widetilde{\boldsymbol{X}}\boldsymbol{l}} \leq \log(t-s) \\ &\Leftrightarrow \log\left(\frac{\zeta}{1-\zeta}\right) \leq \widetilde{\boldsymbol{X}}\left(\log(t-s)\boldsymbol{l} + \boldsymbol{\theta}\right) = \widetilde{\boldsymbol{X}}\boldsymbol{\theta}(\boldsymbol{t},\boldsymbol{s}) \\ &\Leftrightarrow (11). \end{aligned} \tag{12}$$

The censoring time C is generated by the shifted exponential distribution $C \sim s + \exp(\gamma)$. We select γ to maintain a censoring rate of approximately 81% in the simulated dataset, matching the rate observed in the real dataset (see Section 6.1).

To illustrate this generation process, we provide the following specific example:

Example 5.1. *Consider the case where* $K = 50$*, and only two relevant lag polynomials contribute to the true* $\boldsymbol{\theta}$ *in* (9) *for generating* T*. The elements of the true* $\boldsymbol{\theta}$ *are:* $\theta_0 = 1$*,* $\theta_1 = 1$*,* $\theta_2 = -1$*, and* $\theta_k = 0$ *for* $k \in [3, \ldots, K]$*. Thus, the true vector* $\boldsymbol{\theta}(\boldsymbol{t},\boldsymbol{s})$ *in* (11) *is* $(1 + \log(t-s), 1 + \log(t-s), -1 + \log(t-s), 0, \ldots, 0)^\top$*. The two true weighting schemes* $\omega(u; \boldsymbol{\beta}_k) \in \mathbb{R}^d$ *in* (10) *are beta densities, say* $\omega(u; \boldsymbol{\beta}_1) = Beta(1,3)$ *and* $\omega(u; \boldsymbol{\beta}_2) = Beta(2,3)$.

Given that the dictionary W *is included in the sg-LASSO-MIDAS* (6)*, our focus is on estimating* $\hat{\boldsymbol{\beta}}_1$ *and* $\hat{\boldsymbol{\beta}}_2 \in \mathbb{R}^L$*, which are the coefficients for the two relevant lag polynomials. It is crucial to note the differences in dimensionality:* $\omega(u; \boldsymbol{\beta}_k) \in \mathbb{R}^d$*,* $\hat{\boldsymbol{\beta}}_k \in \mathbb{R}^L$*, and* $W \in \mathbb{R}^{d \times L}$*. The relationship between the estimator (left side) and the true parameters (right side) is expressed as follows*

$$W\hat{\boldsymbol{\beta}}_1 \to Beta(1,3) \times (1 + \log(t-s)),$$

$$W\hat{\boldsymbol{\beta}}_2 \to Beta(2,3) \times (-1 + \log(t-s)).$$

All settings for $\boldsymbol{\theta}$*, along with the weighting schemes* $\omega(u; \boldsymbol{\beta}_1) = Beta(1,3)$ *and* $\omega(u; \boldsymbol{\beta}_2) =$

$Beta(2, 3)$ are applied in our simulation settings.

For the choice of the MIDAS weight function W in the sg-LASSO-MIDAS, we employ 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 (6).[7] 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)

To finalize the simulation design, we discuss the choices for s and t. Consistent with prior research (Audrino et al., 2019), we incorporate up to $s = 6$ years of data for predictive modeling in our simulations. For t, we use the percentiles $t = \{t_1 = 10\%, t_2 = 30\%, t_3 = 50\%\}$ of the set $\{T_i : T_i \text{ is uncensored}, i \in [N]\}$. Table 1 summarizes all the settings used in this simulation.

Table 1: Summary of simulation settings.

s	6 years
K	2 relevant and 48 noisy high-frequency lag polynomials
m	4, yearly/quarterly setting
d	$s \times m$ lags
T	See (9)
N	$\{800, 1200\}$
t	$\{10\%, 30\%, 50\%\}$ percentile of the set $\{T_i : T_i \text{ is uncensored }, i \in [N]\}$
C	Shifted Exponential distribution $s + exp(\gamma)$
Censoring Rate	Around 81%
W	Gegenbauer polynomials with $\alpha_{\text{poly}} = -\frac{1}{2}$ and up to size $L = 3$ shifted to $[0, 1]$
$\widetilde{\boldsymbol{X}}$	See (10)
$\boldsymbol{\theta(t, s)}$	$(1 + \log(t - s), 1 + \log(t - s), -1 + \log(t - s), 0, \ldots, 0)^\top$

A key remaining question is how to evaluate the classification results from our model. Although Receiver Operating Characteristic (ROC) curves are widely used for assessing classification performance, the traditional ROC curve approach is not fully appropriate in this context due to the presence of censoring, which means the bankruptcy status of some firms is only partially observed. To overcome this limitation, we introduce an ROC curve estimator that can effectively measure classification performance in both the simulated and real datasets.

5.2 Evaluation Metric: Time-Dependent ROC Curves

Recalling the definitions of sensitivity and specificity in the ROC curves, we see that in our model, the distress status $\mathbb{1}\{T \leq t\}$ is time-dependent of t. Consequently, both sensitivity,

[7] 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.

or the "true positive rate" (TPR), and specificity, or the "false positive rate" (FPR), are also functions that depend on t

$$
\begin{aligned}
Se(c,t) &= P[M > c \mid D(t) = 1], \\
Sp(c,t) &= P[M \geq c \mid D(t) = 0].
\end{aligned}
$$

We define the estimated probability $M_i := p(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i) = \frac{\exp\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)}$ to discriminate a firm's financial distress. The threshold c is used to classify a firm as distressed if $M_i \geq c$, or as non-distressed if $M_i < c$, with $D(t) = \mathbb{1}\{T \leq t\}$ 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 higher ROC curve signifies better performance in distinguishing between firms that have failed and those that have not. However, in practice, some firms are censored, meaning that their distress status $D(t)$ cannot be observed. 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).

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

$$
\hat{S}_{\lambda_N}(c,t) = \frac{1}{N} \sum_i \hat{S}_{\lambda_N}(t \mid M_i)\, \mathbb{1}\{M_i > c\},
$$

where $\hat{S}_{\lambda_N}(t \mid M_i)$ is a suitable estimator of the conditional survival function characterized by a parameter λ_N

$$
\hat{S}_{\lambda_N}(t \mid M_i) = \prod_{a \in \mathcal{T}_N, a \leq t} \left\{ 1 - \frac{\sum_j K_{\lambda_N}(M_j, M_i)\, \mathbb{1}\{\widetilde{T}_j = a\}\delta_j}{\sum_j K_{\lambda_N}(M_j, M_i)\, \mathbb{1}\{\widetilde{T}_j \geq a\}} \right\},
$$

where $\mathcal{T}_N$ is a set of the unique values of $\widetilde{T}_i$ for observed events, $\delta_i = \mathbb{1}\{T_i \leq C_i\}$ and $K_{\lambda_N}(M_j, M_i)$ is a kernel function that depends on a smoothing parameter λ_N. Akritas (1994) used a $0/1$ nearest neighbor kernel, $K_{\lambda_N}(M_j, M_i) = \mathbb{1}\{-\lambda_N < \hat{F}_M(M_i) - \hat{F}_M(M_j) < \lambda_N\}$ where $F_M(x)$ is the distribution function of M and $2\lambda_N \in (0,1)$ represents the percentage of individuals that are included in each neighborhood (boundaries). The resulting sensitivity and specificity are defined by

$$
\begin{aligned}
\widehat{Se}(c,t) &= \frac{\left(1 - \hat{F}_M(c)\right) - \hat{S}_{\lambda_N}(c,t)}{1 - \hat{S}_{\lambda_N}(t)}, \\
\widehat{Sp}(c,t) &= 1 - \frac{\hat{S}_{\lambda_N}(c,t)}{\hat{S}_{\lambda_N}(t)},
\end{aligned}
$$

where $\hat{S}_{\lambda_N}(t) = \hat{S}_{\lambda_N}(-\infty, t)$. 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. They established the usual $\sqrt{N}$-consistency and asymptotic normality 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 λ_N is chosen to be $\mathcal{O}\left(N^{-\frac{1}{3}}\right)$. In our paper, we use the default value of the λ_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).

5.3 Simulation Results

First, we estimate three different LASSO-type regression methods. In the last approach sg-LASSO-MIDAS, each covariate and its high-frequency lags share the same group, therefore, we have K groups. Table 2 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.

Table 2: 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	$1 + K \times L$

We start by comparing the prediction results for sample sizes $N \in \{800, 1200\}$ across three simulation scenarios, followed by an examination of the recovery of the MIDAS weight function. To assess 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 on the test dataset, with the average estimated AUC obtained from 100 simulated datasets for each sample size. The tuning parameters λ and α in the sg-LASSO-MIDAS are selected using 5-fold stratified cross-validation to maximize the AUC, with the same procedure applied to the LASSO-MIDAS.

Table 3 reports the estimated average AUCs for 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 nonlinear 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. This simulation evidence strongly supports the advantage of using MIDAS weighting and incorporating the internal structure of covariates in high-dimensional settings.

Table 3: Estimated average AUCs (standard error) on the test dataset of the three different methods: LASSO-UMIDAS (LASSO-U), LASSO-MIDAS (LASSO-M), sg-LASSO-MIDAS (sg-LASSO-M). $t = \{t_1 = 10\%, t_2 = 30\%, t_3 = 50\%\}$ percentile of the set $\{T_i : T_i \text{ is uncensored }, i \in [N]\}$, and $N = \{800, 1200\}$.

	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	0.541 (0.016)	0.541 (0.009)	0.509 (0.005)	0.563 (0.026)	0.603 (0.011)	0.564 (0.007)
LASSO-M	0.760 (0.035)	0.730 (0.016)	0.665 (0.012)	0.765 (0.028)	0.784 (0.011)	0.777 (0.006)
sg-LASSO-M	0.803 (0.030)	0.769 (0.015)	0.719 (0.011)	0.807 (0.021)	0.811 (0.007)	0.791 (0.005)
	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	0.610 (0.029)	0.642 (0.014)	0.608 (0.012)	0.673 (0.026)	0.716 (0.009)	0.720 (0.009)
LASSO-M	0.684 (0.037)	0.789 (0.013)	0.765 (0.012)	0.770 (0.026)	0.862 (0.005)	0.876 (0.003)
sg-LASSO-M	0.729 (0.033)	0.809 (0.010)	0.795 (0.008)	0.800 (0.021)	0.867 (0.004)	0.878 (0.003)
	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	0.525 (0.021)	0.526 (0.007)	0.505 (0.001)	0.545 (0.014)	0.519 (0.007)	0.502 (0.001)
LASSO-M	0.686 (0.048)	0.609 (0.017)	0.538 (0.009)	0.741 (0.044)	0.625 (0.020)	0.562 (0.014)
sg-LASSO-M	0.711 (0.051)	0.630 (0.021)	0.551 (0.012)	0.768 (0.044)	0.651 (0.014)	0.594 (0.019)

Table 13, 14 and 15 show the recovery of the two true weight functions $Beta(1,3) \times (1 + \log(t-s))$, and $Beta(2,3) \times (-1 + \log(t-s))$, as described in Example 5.1. Entries in odd rows are the average mean integrated squared error and the simulation standard error in even rows. The recovery performance of the LASSO-UMIDAS is typically the worst compared with the other two methods. sg-LASSO-MIDAS seems to improve even more than LASSO-MIDAS. Entries in odd rows show the average mean integrated squared error (MISE), while the even rows display the simulation standard error. The recovery performance of LASSO-UMIDAS is generally the poorest compared to the other two methods, with sg-LASSO-MIDAS showing the most improvement over LASSO-MIDAS.

It is important to note that, due to the censoring issue in the data, the convergence rate of the proposed estimator does not increase as rapidly as it would under no-censoring conditions as the sample size N grows. This finding is consistent with our theoretical results. Additionally, as t changes, the two true weight functions, $Beta(1,3) \times (1 + \log(t - s))$, and $Beta(2,3) \times (-1 + \log(t - s))$ also change accordingly.

6 Empirical Analysis: Predicting Chinese Firms Bankruptcy with Mixed-Frequency Censored Data

6.1 Database and Measurement of Financial Distress in China

We construct a dataset comprising Chinese publicly traded firms from the manufacturing sector on the Shanghai and Shenzhen Stock Exchanges, with their financial statuses categorized as either Special Treatment (ST) or No-ST.[8] A firm is classified as ST firm if it meets any of the following criteria:

- two consecutive years of earnings are negative;
- one recent year of earnings is negative and the most recent year of equity is negative;
- the most recent year's audited financial statements conclude with substantial doubt;
- other situations identified by the stock exchange as abnormal activities or a high risk of delisting.

According to Li et al. (2021), ST firms serve as a reliable indicator of financial distress in China. Therefore, we use the ST indicator to represent a firm's financial distress in this paper.

The dataset is sourced from the IFIND database https://www.hithink.com/ifind.html, one of China's leading financial data providers. Additionally, we have developed an R package, "Survivalml," which includes this dataset and is publicly available at https://github.com/Wei-M-Wei/Survivalml. Detailed information about the package and dataset can be found in Appendix C.

The raw dataset consists of 1614 companies, of which 299 were classified as ST and 1315 as No-ST. 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 11 provides detailed information on these financial variables; see Appendix C for further details.

[8] The initial public dates of these firms were between January 1, 1985, and December 31, 2015.

We then 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 1 illustrates the prediction procedure applied to the real dataset. The observation period spans from $1985/01/01$ to $2020/12/31$, and 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 its 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.

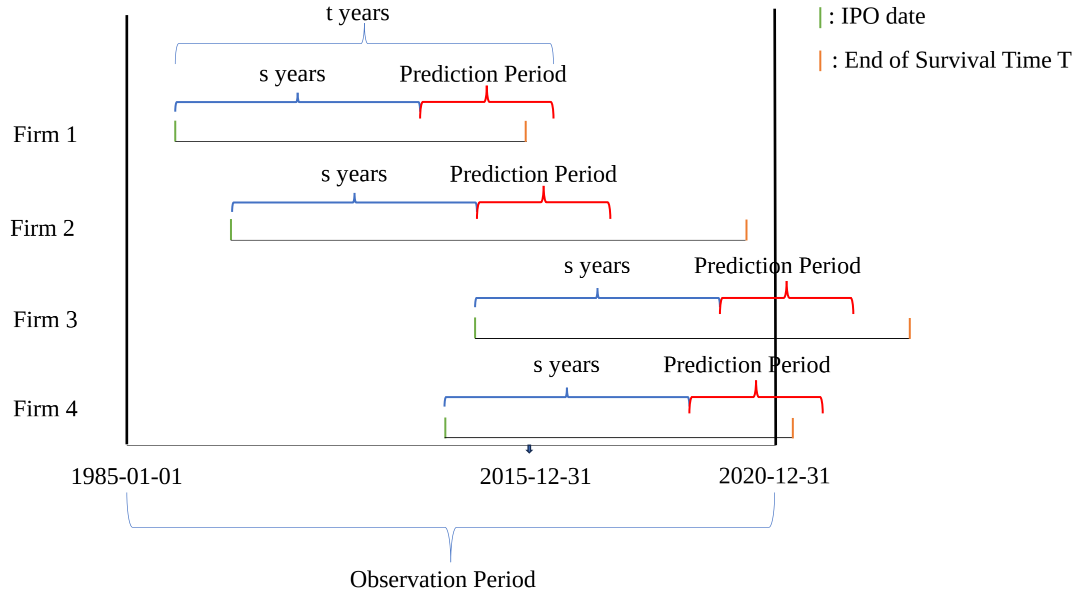

Figure 1: Prediction procedure in the empirical application.

6.2 Estimation Procedure

We now describe the estimation procedure in the empirical application. In practice, 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 $j = 1, 2, \ldots, d$, and $d = s \times m$. We organize all the lags of the covariate into a group vector $\boldsymbol{Z}_{i,k}$:

$$\boldsymbol{Z}_{i,k} = \left(x_{i,s-\frac{1}{m},k}, x_{i,s-\frac{1}{m},k}, \ldots, x_{i,\frac{1}{m},k} \right), i \in [N], k \in [K],$$

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

Next, we aggregate the lagged covariate vector $\boldsymbol{Z}_{i,k}$ 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 $L = 3$.[9]

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

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

where each $\boldsymbol{X}_i = (\boldsymbol{Z}_{i,1}W, \boldsymbol{Z}_{i,2}W, \ldots, \boldsymbol{Z}_{i,K}W)^\top, i \in [N]$. This matrix $\boldsymbol{X}$ is then used in sg-LASSO-MIDAS and LASSO-MIDAS. Notice that we include the intercept term in the estimation procedure.

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

- **Simple Logistic Regression**: We benchmark a simple low-dimensional logistic regression with the latest lags of all financial covariates. 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 $d = s \times m - 1$ coefficients per group covariate $\boldsymbol{Z}_{i,k}, k \in [K]$, 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 $\boldsymbol{Z}_{i,k} \in \mathbb{R}^{1 \times d}, k \in [K]$. We aggregate the group covariate $\boldsymbol{Z}_{i,k}$ using Gegenbauer polynomials $W \in \mathbb{R}^{d \times L}$. 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 $\boldsymbol{Z}_{i,k} \in \mathbb{R}^{1 \times d}, k \in [K]$, which is aggregated using Gegenbauer polynomials $W \in \mathbb{R}^{d \times L}$. 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 $1 + K \times L$.

[9] These polynomials are also known as the second type of Chebyshev polynomials.

The choice of s dictates the historical information captured in the covariate matrix $\boldsymbol{X}$, while t denotes the prediction horizon. Existing literature on firm distress prediction, particularly in the United States, often examines prediction horizons 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 this empirical application, we select $s = 6$ years as a reference, based on the firm classification criteria for Special Treatment outlined in subsection 6.1. We establish prediction horizons of $t = 8, 8.5, 9$ years, with the goal of predicting firm distress within these timeframes. Additionally, to explore longer forecasting periods, we also consider another example with $s = 10$ years and prediction horizons $t = 13, 13.5, 14$ years. These extended prediction horizons allow us to capture longer-term financial distress risks. As highlighted by Li et al. (2021), such prediction horizons are crucial for predicting firm financial distress in China, and they provide meaningful and practical benchmarks for firm failure prediction.

In practice, missing data in financial variables can arise due to various factors, including inconsistent reporting practices across firms, differing regulatory requirements, incomplete disclosures, and delays in data availability following IPOs. Given the significant 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 available, these methods often result in too few firms or variables being retained for meaningful analysis. Furthermore, the censoring rate in the sub-dataset plays an important 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.1 in Appendix C.1. Figure 2 illustrates the distribution of listing, financial distress, and censored firms in different years for the complete dataset with $s = 6$ years. The figure reveals that a large proportion of firms were listed in 2010, and the majority of financial distress occurred between 2016 and 2020, highlighting critical time frames for prediction.

To evaluate the prediction performance of the different methods, we randomly split the dataset into in-sample (80%) and out-of-sample (20%), ensuring that both sets maintain the same proportion of the event indicator $\delta_i(t)\mathbb{1}\{\widetilde{T}_i \leq t\}$. Tuning parameters for the sg-LASSO-MIDAS and LASSO-MIDAS are selected using 5-fold stratified cross-validation, where the optimal parameters are those that maximize the AUC on the out-of-sample. The AUC estimator used in this procedure follows the method described in Section 5.2. Specifically, we perform a grid search over the regularization parameter α in the sparse group LASSO penalty, with values in the set $[0.1, 0.3, 0.5, 0.7, 0.9, 1]$. This process is repeated 10 times, each time using a different random split of the data. All the models are trained in the same dataset and evaluated in the same test dataset.

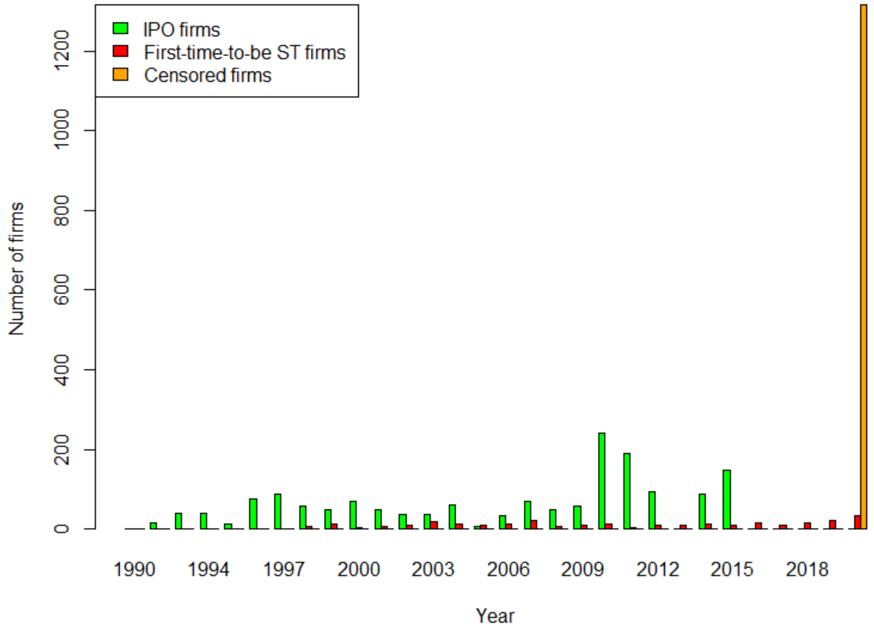

Figure 2: Number of IPO, first-time-to-be ST, and censored firms across different years in the raw dataset.

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

6.3 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 Chinese economy. The dataset contains 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. Given that the sub-datasets differ by the value of s, the set of macroeconomic variables selected will vary accordingly for each sub-dataset. Additionally, we use the same MIDAS dictionary W for the macroeconomic covariates as for the financial covariates, ensuring consistency in how we aggregate the high-frequency data over time.

Table 4 summarizes the details of the two sub-datasets categorized by different values of s. For the sub-dataset where $s = 6$ years, we use all available information across each firm's entire survival period. This allows us to utilize the maximum historical data available for firms with 6 years of survival. However, for the sub-dataset where $s = 10$ years, we limit the covariates to those from the last 4 years of each firm's survival period. This adjustment is made because firms that have survived for 10 years were generally listed in the 1990s, and significant missing data tends to be present in the early years after their IPOs. Therefore, restricting the covariates to the more recent 4 years ensures better data quality and a more robust analysis.

Table 4: Summary information of the dataset with $s = 6, 10$ years

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

In addition, we also apply an oversampling technique to address the imbalance in the dataset due to the high censoring rate, which leads to an unequal proportion of firms experiencing distress versus those that are censored. This imbalance could negatively impact the performance of distress prediction models, as the minority class (distressed firms) may be underrepresented:

- Oversampling: Since the 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.

Table 5: (Distress prediction performance) Estimated average AUCs (95% confidence interval) on the out-of-sample 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
	Bench Mark		
Simple-Logit	0.714 [0.666, 0.760]	0.698 [0.664, 0.736]	0.782 [0.754, 0.816]
	Kaplan-Meier Weights (6)		
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]
	Kaplan-Meier Weights (6) (Cross-Validation for 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]
	Pair-Wise Difference Test between sg-LASSO-M and LASSO-U		
p-value	0.098	0.000	0.000

The last row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS compared to LASSO-UMIDAS with cross-validation for maximizing AUC

As shown in Tables 5 and 6, the LASSO-MIDAS and sg-LASSO-MIDAS consistently outperform the LASSO-UMIDAS, which aligns with our expectations. 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 AUC generally yields better results than cross-validation based on likelihood scores. This supports the idea that maximizing the AUC

Table 6: (Distress prediction performance) Estimated average AUCs (95% confidence interval) on the out-of-sample 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
	Bench Mark		
Simple-Logit	0.621 [0.566, 0.682]	0.598 [0.555, 0.647]	0.635 [0.600, 0.675]
	Kaplan-Meier Weights (6)		
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]
	Kaplan-Meier Weights (6) (Cross-Validation for 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]
	Pair-Wise Difference Test between sg-LASSO-M and LASSO-U		
p-value	0.000	0.010	0.065

The last row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS compared to LASSO-UMIDAS with cross-validation for maximizing AUC is more effective for evaluating prediction performance in this context, as AUC captures the model's ability to discriminate between distressed and non-distressed firms more effectively than likelihood-based measures.

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.

To further assess the performance difference, we conduct a bootstrap pairwise test (James A. Hanley, 1983; Robin et al., 2011) comparing the prediction ability of sg-LASSO-MIDAS and LASSO-UMIDAS, which is widely used in comparing two AUCs. The results indicate that the improvement of sg-LASSO-MIDAS over 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. These findings strongly support the advantages of using MIDAS weights and considering the group structure of covariates in practice. Overall, the empirical results highlight the superiority of the sg-LASSO-MIDAS across different scenarios.

Notably, the set of selected variables tends to vary across different prediction horizons t. Specifically, the models are generally more effective at identifying firms that fail within the recent few years and their discriminative power seems to decrease as the prediction horizon extends to the long term.

We also examine the financial types selected by the sg-LASSO-MIDAS, as shown in Figure 3. The Z-score-related financial variables appear to play a significant role across all prediction horizons in the prediction of firm distress. This finding aligns with previous literature (Altman, 1968), as the Z-score model has been extensively used in both academic research and industry for predicting corporate defaults (Altman et al., 2017). Further details on the selected financial covariates are provided in Figure 4 in Appendix D.

6.4 Does the Censoring Problem Matter?

What happens if we neglect the effect of censoring data? To explore this, we consider an estimator that ignores the impact of censored observations. The procedure is as follows:

- **Comparison Model**: We consider the following model:

$$R_{C,N}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{C_i \geq t\} \left(\mathbb{1}\{\widetilde{T}_i \leq t\}\boldsymbol{X}_i^{\top}\boldsymbol{\beta} + \log\left(1 + \exp(\boldsymbol{X}_i^{\top}\boldsymbol{\beta})\right)\right),$$

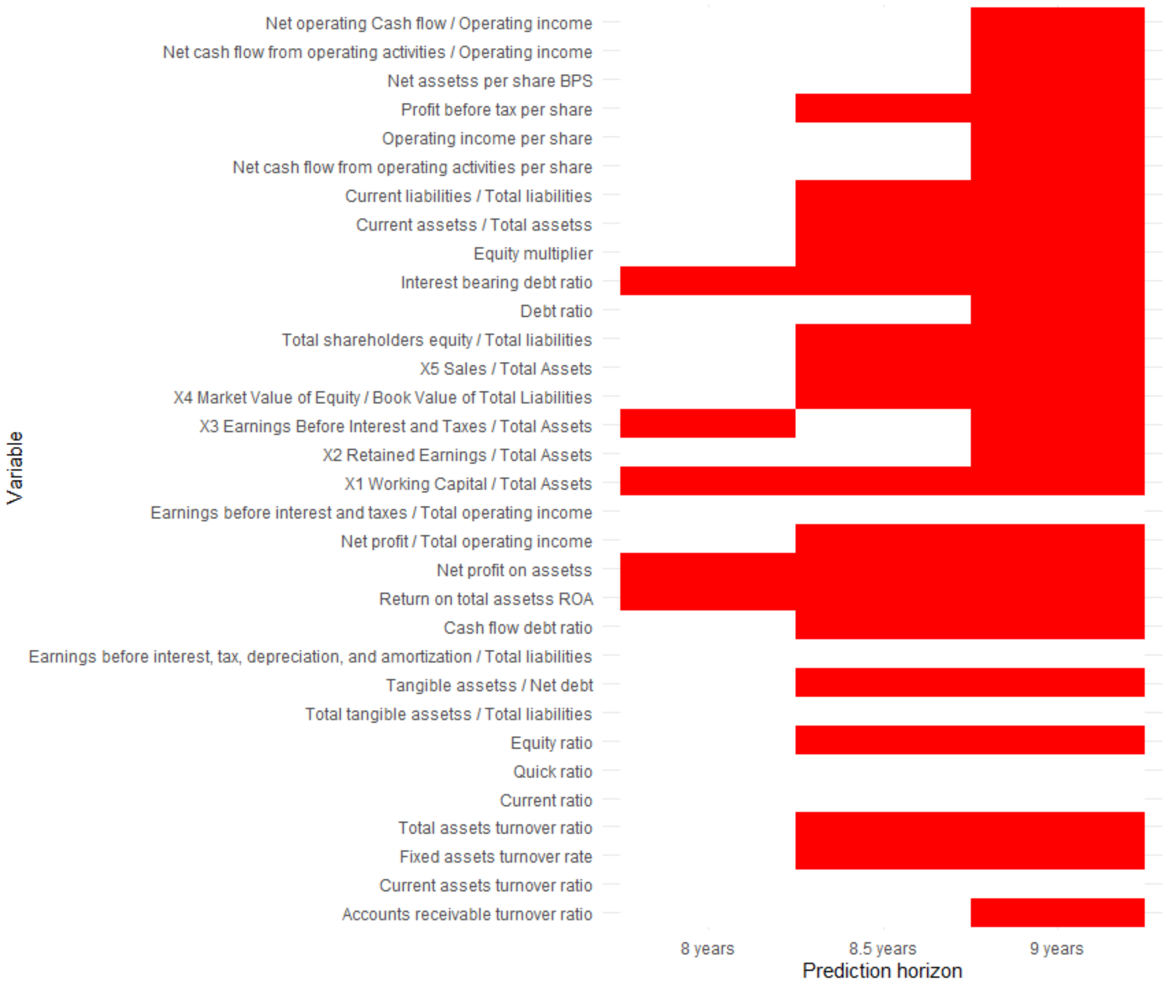

Figure 3: The proportion of selected financial variables in each type when $s = 6$ years.

and the comparison estimator $\hat{\boldsymbol{\beta}}_C$ is obtained by:

$$\hat{\boldsymbol{\beta}}_C = \text{argmin}_{\boldsymbol{\beta}} R_{C,N}(\boldsymbol{\beta}) + \lambda\Omega(\boldsymbol{\beta}). \tag{13}$$

In this comparison model, we exclude firms for which $\mathbb{1}\{C_i \geq t\}$, that are censored beyond the prediction horizon, from the dataset. However, it is important to note that we still estimate the AUC using the same test dataset for all models.

Table 7: (Distress prediction performance) Estimated average AUCs (95% confidence interval) on the out-of-sample 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
	Kaplan-Meier Weights (6)		
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]
	Comparison Model (13) (only firms observed $\geq t$ years)		
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]
	Pair-Wise Difference Test between sg-LASSO-M and Comparison Model		
p-value	0.001	0.495	0.021

Notes: The last row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS with Kaplan-Meier Weights compared to the comparison model that doesn't consider the censoring problem.

When censoring is ignored, the performance of the comparison model deteriorates across all scenarios, emphasizing the importance of properly accounting for censoring in predictive modeling. We also conduct a pairwise comparison between the sg-LASSO-MIDAS and the comparison model. For the scenarios where $s = 6$, $t = 8.5$ years, and $s = 10$, $t = 13$ years, sg-LASSO-MIDAS shows a numerical improvement over the comparison model, though the difference is not statistically significant. However, in other scenarios, including censoring significantly enhances model performance, with results being statistically significant at least at the 5% level.

6.5 One More distress prediction Application

In this section, we present another application of distress prediction using the same dataset, but with a different method for dividing the in-sample and out-of-sample sets compared to the previous subsection.

Table 8: (Distress prediction performance) Estimated average AUCs (95% confidence interval) on the out-of-sample 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
	Kaplan-Meier Weights (6)		
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]
	Comparison Model (13) (only firms observed $\geq t$ years)		
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]
	Pair-Wise Difference Test between sg-LASSO-M and Comparison Model		
p-value	0.238	0.022	0.001

Notes: The last row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS with Kaplan-Meier Weights compared to the comparison model that doesn't consider the censoring problem.

Recall that the financial dataset spans from January 1, 1985, to December 31, 2020. To better align with practical applications, when $s = 6$ years, we first select the time point of $2016/12/31$. Firms that had already survived 6 years prior to this date are used as the in-sample (544 firms), while the remaining firms that had not yet survived 6 years by $2016/12/31$, are placed in the out-of-sample (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. We tune the regularization parameters λ and α using 5-fold stratified cross-validation for AUC. Specifically, we use a grid of $[0.9, 0.91, 0.92, \ldots, 1]$ to search for the optimal regularization parameter α in the sparse group LASSO penalty. All the other settings are consistent with those in the previous section, except we use a dictionary W composed of Gegenbauer polynomials shifted to $[0, 1]$ with parameter $\alpha_{\text{poly}} = \frac{1}{2}$ and size $L = 3$.

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

Tables 9 and 10 report the estimated AUCs on the out-of-sample. The second-to-last row presents the pairwise test between sg-LASSO-MIDAS and LASSO-UMIDAS, and the last row shows the test between the sg-LASSO-MIDAS and the comparison model. When $s = 6$ years, sg-LASSO-MIDAS significantly outperforms LASSO-UMIDAS, while LASSO-MIDAS performs similarly to sg-LASSO-MIDAS. Furthermore, macroeconomic data augmented prediction seems to be similar to the purely financial model in most scenarios, however, we observe the prediction performance is more stable compared with only using financial data. Additionally, The model with Kaplan-Meier weights performs statistically better than the model without censoring at the 5% level, except for the $t = 9$ years prediction horizon, further highlighting the advantage of accounting for censoring in the prediction model. When $s = 10$ years, sg-LASSO-MIDAS is numerically superior to both LASSO-UMIDAS and the comparison model, though the difference is statistically significant only for $t = 14$ years.

7 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. By integrating advanced machine learning techniques with Mixed-Data Sampling (MIDAS), the study offers significant contributions to the field, with implications that extend across diverse domains.

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

Secondly, we address the complexities introduced by high-dimensional censored data in logistic regression. To extend the existing literature, with minimal assumptions on both covariates and risk function, we develop a robust theoretical framework that establishes the non-asymptotic properties of the sg-LASSO estimator for censored heavy-tailed data. This framework is readily extendable to generalized linear models with structured sparsity estimators. To overcome the limitations of traditional performance measures in the presence of censoring, we introduce a time-dependent ROC curve estimator, offering a reliable metric for classification accuracy across simulated and empirical datasets.

A key practical contribution is the construction of a comprehensive dataset of Chinese publicly traded manufacturing firms, incorporating survival and censoring time information. This dataset enables a variety of empirical analyses, particularly in the context of firm distress prediction. To address the challenges of mixed-frequency data, we introduce a class of MIDAS regressions with sg-LASSO estimator, incorporating orthogonal polynomial-based lag selection. Empirical findings demonstrate that sg-LASSO-MIDAS consistently outperforms unstructured LASSO models across multiple scenarios. Notably, the inclusion of censoring information significantly enhances model performance, providing robust insights for predict-

Table 9: (Application 2) Estimated AUCs (95% confidence interval) on the out-of-sample 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		
Simple Logit	0.730 [0.565, 0.901]	0.725 [0.576, 0.860]	0.671 [0.543, 0.790]
	Kaplan-Meier Weights (6)		
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]
	Kaplan-Meier Weights (6) (Cross-Validation for 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]
	Comparison Model (13) (only firms observed $\geq t$ years)		
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]
	Pair-Wise Difference Test between sg-LASSO-M and LASSO-U		
p-value	0.001	0.000	0.000
	Pair-Wise Difference Test between sg-LASSO-M and Comparison Model		
p-value	0.009	0.042	0.107

Notes: The last second row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS compared to LASSO-UMIDAS with cross-validation for maximizing AUC. The last row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS with Kaplan-Meier Weights compared to the comparison model.

Table 10: (Application 2) Estimated AUCs (95% confidence interval) on the out-of-sample 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		
Simple Logit	0.619 [0.420, 0.777]	0.598 [0.431, 0.752]	0.647 [0.502, 0.796]
	Kaplan-Meier Weights (6)		
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]
	Kaplan-Meier Weights (6) (Cross-Validation for 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]
	Comparison Model (13) (only firms observed $\geq t$ years)		
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]
	Pair-Wise Difference Test between sg-LASSO-M and LASSO-U		
p-value	0.318	0.255	0.097
	Pair-Wise Difference Test between sg-LASSO-M and Comparison Model		
p-value	0.459	0.290	0.072

Notes: The last second row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS compared to LASSO-UMIDAS with cross-validation for maximizing AUC. The last row reports the p-value of the pair-wise difference test across the three prediction horizons of the sg-LASSO-MIDAS with Kaplan-Meier Weights compared to the comparison model.

ing firm distress under real-world conditions.

Finally, the methodologies developed in this paper have broad applicability beyond corporate distress prediction. The integration of time-varying logit 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.

References

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

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 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.

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

Beyhum, J. and Portier, F. (2024). High-dimensional nonconvex lasso-type m-estimators. *Journal of Multivariate Analysis*, 202:105303.

Beyhum, J. and Striaukas, J. (2023). Sparse plus dense midas regressions and nowcasting during the covid pandemic.

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

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.

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.

Chen, C.-C., Chen, C.-D., and Lien, D. (2020). Financial distress prediction model: The effects of corporate governance indicators. *Journal of Forecasting*, 39(8):1238–1252.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, 86(5):1867–1900.

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.

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. 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.

Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Annals of Statistics*, pages 49–58.

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.

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.

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.

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

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.

Scheike, T. H., Zhang, M.-J., and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220.

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*, volume 2159 of *Lecture Notes in Mathematics*. Springer International Publishing, Cham.

Van De Geer, S. and Muro, A. (2014). On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electronic Journal of Statistics*, 8(2):3031 – 3061.

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.

Appendices

The appendix provides supplementary materials supporting the main contributions of this paper. It includes the proof of Theorem 4.1, detailed information on the real dataset, and additional simulation and empirical results. Section A explores the relationship between the true parameter $\boldsymbol{\beta}_0$ and the estimator (3), accounting for the effects of censoring. Section B outlines the derivation of the oracle inequality for the sg-LASSO estimator (6) in the context of censored data. Section C provides an in-depth description of the real dataset used in this study, while Section D presents further simulation and empirical results. Finally, Section E introduces the dictionaries used in constructing the MIDAS weight function.

A Process on dealing with the censoring problem

Based on Assumption 3.1 and 3.2, by the maximum likelihood estimation, the estimator $\hat{\boldsymbol{\beta}}$ is to solve

$$\operatorname{argmax}_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}\{T_i \le t\}\mathbb{1}\left\{T_i \ge s\right\}\log\left(p\left(\boldsymbol{Z}_i\right)\right) + \left(1-\mathbb{1}\{T_i \le t\}\right)\mathbb{1}\left\{T_i \ge s\right\}\log\left(1-p\left(\boldsymbol{Z}_i\right)\right)\right), \quad (14)$$

where $p\left(\boldsymbol{Z}_i\right) = \frac{\exp\left(\boldsymbol{Z}_i^{\top}\tilde{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{Z}_i^{\top}\tilde{\boldsymbol{\beta}}\right)}$. We first present the following corollaries to show the equivalence between the estimator (3) and (14).

Corollary A.1. *Based on the Assumption 3.2, the Law of large numbers shows:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left\{T_i \le t\right\}\mathbb{1}\left\{T_i \ge s\right\} \xrightarrow{P} \mathbb{E}_{T,\boldsymbol{Z}}[\mathbb{1}\left\{T \le t\right\}\mathbb{1}\left\{T \ge s\right\}] = \mathbb{E}_{T,\boldsymbol{Z}}[\mathbb{1}\left\{T \le t\right\} \mid T \ge s\}]P(T \ge s),$$

then we have the following convergence in probability:

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{T_i \le t\}\mathbb{1}\left\{T_i \ge s\right\}\log\left(p\left(\boldsymbol{Z}_i\right)\right) + \left(1-\mathbb{1}\{T_i \le t\}\right)\mathbb{1}\left\{T_i \ge s\right\}\log\left(1-p\left(\boldsymbol{Z}_i\right)\right) \xrightarrow{P} \\ &\mathbb{E}_{T,\boldsymbol{Z}}\left[\mathbb{1}\{T \le t\}\log\left(p\left(\boldsymbol{Z}\right)\right) + \left(1-\mathbb{1}\{T \le t\}\right)\log\left(1-p\left(\boldsymbol{Z}\right)\right) \mid T \ge s\right]P(T \ge s).\end{aligned}$$

Corollary A.2. *Based on Assumption 3.1 and 3.2, we have:*

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}\{T \le t\}\log(p\left(\boldsymbol{Z}\right)) + \left(1-\mathbb{1}\{T \le t\}\right)\log(1-p\left(\boldsymbol{Z}\right)) \mid T \ge s\right]P(T \ge s) \\ &= \mathbb{E}\left[\mathbb{1}\{\widetilde{T} \ge s\}\left(\frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \le t\}\log(p\left(\boldsymbol{Z}\right)) + \left(1-\frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \le t\}\right)\log(1-p\left(\boldsymbol{Z}\right))\right)\right]\end{aligned}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{T \leq t\}\log(p(\boldsymbol{Z})) + (1-\mathbb{1}\{T \leq t\})\log(1-p(\boldsymbol{Z}))\middle| T \geq s\right] \\
&= \mathbb{E}\left[\frac{\mathbb{1}\{T \leq t\}}{H(t \wedge T)}\mathbb{E}[\mathbb{1}\{C \geq t \wedge T\} \mid T, \boldsymbol{Z}]\log(p(\boldsymbol{Z})) + \left(1 - \frac{\mathbb{1}\{T \leq t\}}{H(t \wedge T)}\mathbb{E}[\mathbb{1}\{C \geq t \wedge T\} \mid T, \boldsymbol{Z}]\right)\log(1-p(\boldsymbol{Z}))\middle| T \geq s\right] \\
&= \mathbb{E}\left[\mathbb{E}\big[\log(p(\boldsymbol{Z}))\frac{\delta(t)}{H(t \wedge T)}\mathbb{1}\{T \leq t\}\big| T, \boldsymbol{Z}\big] + \log(1-p(\boldsymbol{Z})) - \mathbb{E}\big[\log(1-p(\boldsymbol{Z}))\frac{\delta(t)}{H(t \wedge T)}\mathbb{1}\{T \leq t\}\big| T, \boldsymbol{Z}\big]\middle| T \geq s\right] \\
&= \mathbb{E}\left[\frac{\delta(t)}{H(t \wedge T)}\mathbb{1}\{T \leq t\}\log(p(\boldsymbol{Z})) + \left(1 - \frac{\delta(t)}{H(t \wedge T)}\mathbb{1}\{T \leq t\}\right)\log(1-p(\boldsymbol{Z}))\middle| T \geq s\right] \\
&= \mathbb{E}\left[\frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\log(p(\boldsymbol{Z})) + \left(1 - \frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\right)\log(1-p(\boldsymbol{Z}))\middle| T \geq s\right] \\
&= \frac{1}{P(T \geq s)}\mathbb{E}\left[\mathbb{1}\{T \geq s\}\left(\frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\log(p(\boldsymbol{Z})) + \left(1 - \frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\right)\log(1-p(\boldsymbol{Z}))\right)\right] \\
&= \frac{1}{P(T \geq s)}\mathbb{E}\left[\mathbb{1}\{\widetilde{T} \geq s\}\left(\frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\log(p(\boldsymbol{Z})) + \left(1 - \frac{\delta(t)}{H\left(t \wedge \widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\right)\log(1-p(\boldsymbol{Z}))\right)\right].
\end{aligned}$$

The first, second, and third equality rely on the Assumption 3.1 and the law of iterated expectations. The forth equality relies on that if $\mathbb{1}\{C \geq t \wedge T\}\mathbb{1}\left\{T \leq t\right\} = 1$, we must have $t \wedge T \leq C$ and $T \leq t$, which means $T \leq C$, then we can see $T = \widetilde{T}$. On the other hand, if $\mathbb{1}\{C \geq t \wedge T\}\mathbb{1}\left\{\widetilde{T} \leq t\right\} = 1$, we have following cases

- When $T \leq t$, it just means $\mathbb{1}\{C \geq t \wedge T\}\mathbb{1}\left\{T \leq t\right\} = 1$,
- When $T \geq t$, because $\mathbb{1}\{C \geq t \wedge T\} = 1$, we have $C \geq t$, then $\widetilde{T} \geq t$ which contradicts to $\mathbb{1}\left\{\widetilde{T} \leq t\right\} = 1$,

then we can conclude that $\mathbb{1}\{C \geq t \wedge T\}\mathbb{1}\left\{T \leq t\right\} = 1 \Leftrightarrow \mathbb{1}\{C \geq t \wedge T\}\mathbb{1}\left\{\widetilde{T} \leq t\right\} = 1$. The fifth equality uses the following equation when given any function of $f(\boldsymbol{Z}, T, C)$

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{Z}, T, C)\mathbb{1}\{T \geq s\}] \\
&= \mathbb{E}[f(\boldsymbol{Z}, T, C) \cdot 1 \mid T \geq s]P(T \geq s) + \mathbb{E}[f(\boldsymbol{Z}, T, C) \cdot 0 \mid T < s]P(T < s) \\
&= \mathbb{E}[f(\boldsymbol{Z}, T, C) \mid T \geq s]P(T \geq s).
\end{aligned}$$

For the last equality, we consider the following two cases

- if $\mathbb{1}\{\widetilde{T} \geq s\} = 1$, it means $\mathbb{1}\{T \geq s\} = 1$.
- if $\mathbb{1}\{T \geq s\} = 1$ and based on Assumption 3.2, we always have $\widetilde{T} \geq s$.

Overall, we finish the proof. □

Corollary A.3. *Rely on Assumption 3.2, the Law of Large Numbers implies the following convergence in probability:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\widetilde{T}_i \geq s\}\left(\frac{\delta_i(t)}{H\left(t\wedge\widetilde{T}_i\right)}\mathbb{1}\{\widetilde{T}_i \leq t\}\log\left(p\left(\boldsymbol{Z}_i\right)\right)+\left(1-\frac{\delta_i(t)}{H\left(t\wedge\widetilde{T}_i\right)}\mathbb{1}\{\widetilde{T}_i \leq t\}\right)\log\left(1-p\left(\boldsymbol{Z}_i\right)\right)\right)$$
$$\xrightarrow{P}\mathbb{E}_{T,C,\boldsymbol{Z}}\left[\mathbb{1}\{\widetilde{T} \geq s\}\left(\frac{\delta(t)}{H\left(t\wedge\widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\log\left(p\left(\boldsymbol{Z}\right)\right)+\left(1-\frac{\delta(t)}{H\left(t\wedge\widetilde{T}\right)}\mathbb{1}\{\widetilde{T} \leq t\}\right)\log\left(1-p\left(\boldsymbol{Z}\right)\right)\right)\right]. \tag{15}$$

By Using Corollary A.1, A.2 and A.3 sequentially, and we use a consistent estimator $\widehat{H}\left(t\wedge\widetilde{T}\right)$ to replace $H\left(t\wedge\widetilde{T}\right)$ in (15), then we see the equivalence between the estimator (3) and (14).

B Proofs of Theorem 4.1

Outline: In Section B.1, we present a Lemma about the properties of the sparse group LASSO norm. In Section B.2, we define different types of effective sparsity. Section B.3 introduces the one point margin condition for the conditional risk function. Section B.4 presents technical Lemmas which are related to concentration inequalities. The subsequent two sections B.5 and B.6 are dedicated to addressing two key questions:

- Section B.5: We establish probability inequalities for the empirical process $\forall\boldsymbol{\beta}', \boldsymbol{\beta}\in\mathbb{R}^p$

  $$\sup_{\boldsymbol{\beta}':\Omega(\boldsymbol{\beta}'-\boldsymbol{\beta})\leq M_{\boldsymbol{\beta}}}\left|\left[R_N(\boldsymbol{\beta}')-R(\boldsymbol{\beta}'|\boldsymbol{X})\right]-\left[R_N(\boldsymbol{\beta})-R\left(\boldsymbol{\beta}|\boldsymbol{X}\right)\right]\right|,$$

  where $M_{\boldsymbol{\beta}}>0$ that is related to $\boldsymbol{\beta}$, $R_N(\cdot)$ is the sample risk function and $R(\cdot)$ is the population risk function. The presence of censoring further complicates this empirical process.

- Section B.6: We show that the sample effective sparsity can approach to the population effective sparsity. Several foundational results were established by Van De Geer and Muro (2014), and these findings were later extended to encompass more general sparsity-inducing norms, as discussed in Van De Geer (2016). It is important to note that these works primarily consider the isotropic condition of the covariates. In contrast, our study advances these results by accommodating scenarios involving heavy-tailed data.

Notably, all the results we obtain can be readily extended to generalized linear models with structured sparsity estimators.

B.1 Sparse Group LASSO Norm

We consider the sparse-group LASSO penalty $\Omega(\cdot)$ in the whole proof. Note that $\Omega(\cdot)$ can be decomposed as a sum of two seminorms $\Omega(\boldsymbol{b}) = \Omega^+(\boldsymbol{b}) + \Omega^-(\boldsymbol{b}), \forall \boldsymbol{b} \in \mathbb{R}^p$ with

$$\Omega^+(\boldsymbol{b}) = \alpha \left|\boldsymbol{b}_{S_{\boldsymbol{\beta}}}\right|_1 + (1-\alpha) \sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |\boldsymbol{b}_G|_2 \, , \quad \Omega^-(\boldsymbol{b}) = \alpha \left|\boldsymbol{b}_{S^c_{\boldsymbol{\beta}}}\right|_1 + (1-\alpha) \sum_{G \in \mathcal{G}^c_{\boldsymbol{\beta}}} |\boldsymbol{b}_G|_2 \, ,$$

where $(S_{\boldsymbol{\beta}}, \mathcal{G}_{\boldsymbol{\beta}})$ are the support and the group support of a candidate oracle $\boldsymbol{\beta}$ and α is a tuning parameter. It is clear to see that $\Omega(\cdot)$ depends on the choice of $\boldsymbol{\beta}$.

Lemma B.1. *Denote $\Omega_*(\cdot)$ as the dual norm of $\Omega(\cdot)$, it satisfies*

1. *For any vectors $x, y \in \mathbb{R}^p$, $\left|x^\top y\right| \leq \Omega_*(x)\Omega(y)$.*
2. *$\forall z \in \mathbb{R}^p$, we have $\Omega_*(z) \leq \alpha |z|_1^* + (1-\alpha)|z|_{2,1}^*$, where $|\cdot|_1^*$ is the dual norm of $|\cdot|_1$ and $|\cdot|_{2,1}^*$ is the dual norm of $|\cdot|_{2,1}$. Furthermore, let $G^* = \max_{G \in \mathcal{G}} |G|$ be the size of the largest group in $\mathcal{G}_{\boldsymbol{\beta}}$, we also have $\Omega_*(z) \leq G^* |z|_\infty$.*
3. *$\forall X \in \mathbb{R}^{N \times p}$ and $\forall z \in \mathbb{R}^p$, we have $\Omega_*(Xz) \leq G^* \left|X\right|_\infty \Omega(z)$ (see definition of $|X|_\infty$ in Section 2).*

Proof. The first statement is trivial. For the second statement, $\Omega(\cdot)$ is a norm, and by the convexity of $x \mapsto x^{-1}$ on $(0, \infty)$, we have

$$\begin{aligned} \Omega_*(z) &= \sup_{b \neq 0} \frac{|\langle z, b\rangle|}{\Omega(b)} \leq \sup_{b \neq 0} \left\{ \alpha \frac{|\langle z, b\rangle|}{|b|_1} + (1-\alpha) \frac{|\langle z, b\rangle|}{|b|_{2,1}} \right\} \\ &\leq \alpha \sup_{b \neq 0} \frac{|\langle z, b\rangle|}{|b|_1} + (1-\alpha) \sup_{b \neq 0} \frac{|\langle z, b\rangle|}{|b|_{2,1}} \\ &= \alpha |z|_1^* + (1-\alpha)|z|_{2,1}^*. \end{aligned}$$

We also know that $|z|_1^* = |z|_\infty$ and $|z|_{2,1}^* = \left(\sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |z_G|_2\right)^* = \max_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |z_G|_2 \leq \sqrt{G^*}|z|_\infty$, also see Appendix of Babii et al. (2023). Then

$$\Omega_*(z) \leq \alpha |z|_1^* + (1-\alpha)|z|_{2,1}^* \leq \alpha |z|_\infty + (1-\alpha)\sqrt{G^*}|z|_\infty \leq G^* |z|_\infty$$

since $G^* \geq 1$. The third statement is argued by:

$$\begin{aligned}
\Omega^*(Xz) &\leq \alpha |Xz|_\infty + (1-\alpha) \max_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |[Xz]_G|_2 \\
&\leq \alpha |X|_\infty |z|_1 + (1-\alpha)\sqrt{G^*} |X|_\infty \, |z|_2 \\
&\leq |X|_\infty \left(\alpha |z|_1 + (1-\alpha)\sqrt{G^*} \, |z|_2 \right) \\
&\leq |X|_\infty \left(\alpha |z|_1 + (1-\alpha)\sqrt{G^*} \, |z|_{2,1} \right) \\
&\leq G^* |X|_\infty \Omega(z)
\end{aligned}$$

since $|[Xz]_G|_2 \leq |X|_\infty |z_G|_2$, $|z|_2 \leq |z|_{2,1} = \sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |z_G|_2$ and $G^* \geq \sqrt{G^*} \geq 1$. Notice that for the last two arguments, we do not try to obtain sharp inequalities here. □

B.2 Effective Sparsity

In the whole proof, we have i.i.d data $T_i, C_i, \boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,j}, \ldots, X_{i,p})^\top \in \mathbb{R}^p, i \in [N]$ that satisfies Assumption 3.1, 3.2 and 4.1. Especially, we allow for the heavy-tailed data in our proof, which satisfies $\max_{|u|_2=1} \mathbb{E}\left(\left|\boldsymbol{X}_i^\top u\right|^q\right) \leq K_0 < \infty$, where $q \geq 4$.

First, we present several definitions that are inspired by Van De Geer (2016). $\forall \boldsymbol{\beta}', \Delta \in \mathbb{R}^p$, we define the pseudo-norm $\hat{\tau}(\cdot)$ and its population version $\tau(\cdot)$

$$\hat{\tau}^2_{\boldsymbol{\beta}'}(\Delta) := \frac{1}{N} \sum_{i=1}^N \frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}' + E_i)}{\left(1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}' + E_i)\right)^2} \left|\boldsymbol{X}_i^\top \Delta\right|_2^2, \quad \tau^2_{\boldsymbol{\beta}'}(\Delta) := \mathbb{E}\left[\hat{\tau}^2_{\boldsymbol{\beta}'}(\Delta)\right].$$

Furthermore, let

$$\frac{1}{\mathrm{C}_M^2(\boldsymbol{X}_i)} = \left(\frac{1}{1 + \exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i + M\Omega_*(\boldsymbol{X}_i)\right)} \right) \left(1 - \frac{1}{1 + \exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i - M\Omega_*(\boldsymbol{X}_i)\right)} \right).$$

It follows that $\forall \boldsymbol{\beta}'$ that satisfies $\Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) \leq M$

$$\begin{aligned}
&\frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}' + E_i)}{\left(1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}' + E_i)\right)^2} \\
&= \left(\frac{1}{1 + \exp\left(\boldsymbol{X}_i^\top \left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) + \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)} \right) \left(1 - \frac{1}{1 + \exp\left(\boldsymbol{X}_i^\top \left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) + \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)} \right) \\
&\geq \left(\frac{1}{1 + \exp\left(\left|\boldsymbol{X}_i^\top \left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right)\right| + \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)} \right) \left(1 - \frac{1}{1 + \exp\left(-\left|\boldsymbol{X}_i^\top \left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right)\right| + \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)} \right) \\
&\geq \frac{1}{\mathrm{C}_M^2(\boldsymbol{X}_i)}
\end{aligned}$$

since $\left|\boldsymbol{X}_i^\top\left(\boldsymbol{\beta}'-\boldsymbol{\beta}_0\right)\right| \leq \Omega\left(\boldsymbol{\beta}'-\boldsymbol{\beta}_0\right)\Omega_*(\boldsymbol{X}_i) \leq M\Omega_*(\boldsymbol{X}_i)$. We also define

$$\hat{\tau}_M^2(\Delta) := \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\boldsymbol{X}_i^\top\Delta\right|_2^2}{\mathrm{C}_M^2(\boldsymbol{X}_i)}.$$

Notice that, $\forall\boldsymbol{\beta}'$ that satisfies $\Omega\left(\boldsymbol{\beta}'-\boldsymbol{\beta}_0\right) \leq M$, we have $\hat{\tau}_{\boldsymbol{\beta}'}^2(\Delta) \geq \hat{\tau}_M^2(\Delta)$.

For a stretching factor $D>0$, a pair set $S=(S_{\boldsymbol{\beta}},\mathcal{G}_{\boldsymbol{\beta}})$ and its cardinality $s_{\boldsymbol{\beta}}=\alpha\sqrt{|S_{\boldsymbol{\beta}}|}+(1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|}$ of a candidate oracle $\boldsymbol{\beta}$, we define the population version of effective sparsity:

$$\Gamma_\Omega^2\left(D,\mathcal{S}_{\boldsymbol{\beta}},\tau_{\boldsymbol{\beta}'}(\Delta)\right) := \left(\min\left\{\tau_{\boldsymbol{\beta}'}^2(\Delta) : \Delta\in\mathbb{R}^p, \Omega^+(\Delta)=1, \Omega^-(\Delta)\leq D\right\}\right)^{-1}, \tag{16}$$

and its compatibility constant

$$\phi^2(D,\mathcal{S}_{\boldsymbol{\beta}},\tau_{\boldsymbol{\beta}'}(\Delta)) := \min\left\{s_{\boldsymbol{\beta}}\tau_{\boldsymbol{\beta}'}^2(\Delta) : \Delta\in\mathbb{R}^p, \Omega^+(\Delta)=1, \Omega^-(\Delta)\leq D\right\}. \tag{17}$$

The sample version is

$$\Gamma_\Omega^2\left(D,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_{\boldsymbol{\beta}'}(\Delta)\right) := \left(\min\left\{\hat{\tau}_{\boldsymbol{\beta}'}^2(\Delta) : \Delta\in\mathbb{R}^p, \Omega^+(\Delta)=1, \Omega^-(\Delta)\leq D\right\}\right)^{-1}, \tag{18}$$

and its compatibility constant

$$\phi^2(D,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_{\boldsymbol{\beta}'}(\Delta)) := \min\left\{s_{\boldsymbol{\beta}}\hat{\tau}_{\boldsymbol{\beta}'}^2(\Delta) : \Delta\in\mathbb{R}^p, \Omega^+(\Delta)=1, \Omega^-(\Delta)\leq D\right\}. \tag{19}$$

Next, we can clearly see that:

$$\Gamma_\Omega^2(D,\mathcal{S}_{\boldsymbol{\beta}},\tau_{\boldsymbol{\beta}'}(\Delta)) = \frac{s_{\boldsymbol{\beta}}}{\phi^2(D,\mathcal{S}_{\boldsymbol{\beta}},\tau_{\boldsymbol{\beta}'}(\Delta))}, \tag{20}$$

$$\Gamma_\Omega^2(D,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_{\boldsymbol{\beta}'}(\Delta)) = \frac{s_{\boldsymbol{\beta}}}{\phi^2(D,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_{\boldsymbol{\beta}'}(\Delta))}. \tag{21}$$

By the definition of the sample effective sparsity (18), it follows $\forall\boldsymbol{\beta}',\Delta\in\mathbb{R}^p$, we have

$$\Omega^+(\Delta) \leq \hat{\tau}_{\boldsymbol{\beta}'}(\Delta)\,\Gamma_\Omega\left(D,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_{\boldsymbol{\beta}'}(\Delta)\right).$$

It's worth mentioning that all $\hat{\tau}_{\boldsymbol{\beta}'}(\Delta)$ above can be replaced by $\hat{\tau}_M(\Delta)$. Notice that these notations are slightly modified version compared with Van De Geer (2016) and the choice of $\boldsymbol{\beta}'$ can be arbitrary, e.g. $\boldsymbol{\beta}'=\boldsymbol{\beta}_0$. The following is a useful Lemma to tie the population compatibility constant to Assumption 4.2.

Lemma B.2. *Suppose that Assumption 4.2 holds, we have* $\phi^2\left(D,\mathcal{S}_{\boldsymbol{\beta}},\tau_{\boldsymbol{\beta}_0}(\Delta)\right) \geq \gamma_{\mathrm{H}}$.

Proof. Recall the definition of $\phi^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta))$, see (17), we work on the event

$$\left\{\Omega^+(\Delta) = 1, \Omega^-(\Delta) \leq D\right\}.$$

By the property of the smallest eigenvalue of a symmetric matrix and Assumption 4.2, we have

$$\gamma_{\mathrm{H}} = \min_{|u|_2=1} u^\top \mathbb{E}\left[\frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)\right)^2} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right] u.$$

By Jensen's inequality, we then see

$$\sqrt{|S_{\boldsymbol{\beta}}|\Delta^\top \Delta} \geq \sqrt{|S_{\boldsymbol{\beta}}|\Delta_{S_{\boldsymbol{\beta}}}^\top \Delta_{S_{\boldsymbol{\beta}}}} \geq |\Delta_{S_{\boldsymbol{\beta}}}|_1,$$

and

$$\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|\Delta^\top \Delta} \geq \sqrt{|\mathcal{G}_{\boldsymbol{\beta}}| \sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} \Delta_G^\top \Delta_G} \geq \sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |\Delta_G|_2 .$$

The last inequality comes from: Assume $\mathcal{G}_{\boldsymbol{\beta}} = \{G_1, G_2, \ldots, G_k\}$ and let $\Delta_{G_i}^\top \Delta_{G_i} = g_i^2 \in \mathbb{R}^+$, again by Jensen's inequality

$$\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}| \sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} \Delta_G^\top \Delta_G} = \sqrt{k \sum_{i=1}^{k} \Delta_{G_i}^\top \Delta_{G_i}} = \sqrt{k \sum_{i=1}^{k} g_i^2} \geq \sum_{i=1}^{k} |g_i| = \sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |\Delta_G|_2 .$$

Overall, since $\sqrt{s_{\boldsymbol{\beta}}} = \alpha\sqrt{|S_{\boldsymbol{\beta}}|} + (1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|}$, it is clear to see that

$$\begin{aligned}\sqrt{s_{\boldsymbol{\beta}}\Delta^\top \Delta} &= \alpha\sqrt{|S_{\boldsymbol{\beta}}|}\sqrt{\Delta^\top \Delta} + (1-\alpha)\sqrt{|\mathcal{G}_{\boldsymbol{\beta}}|}\sqrt{\Delta^\top \Delta} \\ &\geq \alpha|\Delta_{S_{\boldsymbol{\beta}}}|_1 + (1-\alpha)\sum_{G \in \mathcal{G}_{\boldsymbol{\beta}}} |\Delta_G|_2 = \Omega^+(\Delta) = 1.\end{aligned}$$

Then we have

$$s_{\boldsymbol{\beta}} \tau_{\boldsymbol{\beta}_0}^2(\Delta) = \sqrt{s_{\boldsymbol{\beta}}}\Delta^\top \mathbb{E}\left[\frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)\right)^2} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right] \Delta\sqrt{s_{\boldsymbol{\beta}}} \geq \gamma_{\mathrm{H}}$$

since $\sqrt{s_{\boldsymbol{\beta}}\Delta^\top \Delta} \geq 1$. □

B.3 One Point Margin Condition for the Conditional Risk Function

Let $\mathbb{B}_{local} := \{\boldsymbol{\beta}' : \Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) \leq M\}, \exists M > 0$, which is a convex neighborhood of the true parameter $\boldsymbol{\beta}_0$. We consider the relationship between the sample risk function $R_N(\boldsymbol{\beta})$ and the population risk function $R\left(\boldsymbol{\beta}\right)$. To simplify, we omit the intercept term and $\mathbb{1}\{\widetilde{T}_i \geq s\}$ here, since it does not have an impact on the proof. We define

$$\widehat{f}_i(t)=\frac{\delta_i(t)}{\widehat{H}\left(t\wedge\widetilde{T}_i\right)}\mathbb{1}\{\widetilde{T}_i\leq t\}, f_i(t)=\frac{\delta_i(t)}{H\left(t\wedge\widetilde{T}_i\right)}\mathbb{1}\{\widetilde{T}_i\leq t\}.$$

Then we see that

$$R_N(\boldsymbol{\beta})=\frac{1}{N}\sum_{i=1}^{N}\Bigg(-\widehat{f}_i(t)\boldsymbol{X}_i^\top\boldsymbol{\beta}+\log\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta})\right)\Bigg),$$

and

$$R(\boldsymbol{\beta})=\mathbb{E}\left[-f_i(t)\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)+\log\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)\right].$$

Their gradient and Hessian matrix functions are as follows

$$\dot{R_N}(\boldsymbol{\beta})=\frac{1}{N}\sum_{i=1}^{N}\left(-\widehat{f}_i(t)+\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta})}\right)\boldsymbol{X}_i^\top, \dot{R}(\boldsymbol{\beta})=\mathbb{E}\left[\left(-f_i(t)+\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}\right)\boldsymbol{X}_i^\top\right],$$

and

$$\ddot{R_N}(\boldsymbol{\beta})=\frac{1}{N}\sum_{i=1}^{N}\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta})}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta})\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top, \ddot{R}(\boldsymbol{\beta})=\mathbb{E}\left[\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right].$$

Lemma B.3. *(One point margin condition for the conditional risk) We consider the conditional theoretical risk function $R(\boldsymbol{\beta}|\boldsymbol{X})$ here. For $\tilde{\boldsymbol{\beta}}\in\mathbb{B}_{local}$, one has the one point margin condition:*

$$R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\geq\frac{\hat{\tau}_M^2\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)}{2}.$$

Proof. $\boldsymbol{\beta}_0$ is the true parameter which means $\dot{R}\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)=0$, then using the Taylor expansion at $\boldsymbol{\beta}_0$, we can easily see:

$$\begin{aligned}
&R\left(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}\right)-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\\
&=\dot{R}\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\frac{1}{2}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^\top\ddot{R}\left(\boldsymbol{\beta}'|\boldsymbol{X}\right)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\
&=\dot{R}\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\frac{1}{2}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^\top\left(\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}'+E_i\right)}{\left(1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}'+E_i\right)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right)(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\
&\geq\frac{\hat{\tau}_M^2\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)}{2},
\end{aligned}$$

where $\boldsymbol{\beta}' \in \mathbb{B}_{local}$ since $\mathbb{B}_{local}$ is convex. $\dot{R}\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right) = 0$ is followed by

$$\begin{aligned}\dot{R}\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right) &= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}\left[-f_i(t)\boldsymbol{X}_{\boldsymbol{i}}|\boldsymbol{X}_{\boldsymbol{i}}\right] + \frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}{1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}\boldsymbol{X}_i\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}\left[-f_i(t)|\boldsymbol{X}_{\boldsymbol{i}}\right]\boldsymbol{X}_{\boldsymbol{i}} + \frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}{1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}\boldsymbol{X}_i\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(-\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}{1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}\boldsymbol{X}_i + \frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}{1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i)}\boldsymbol{X}_i\right) = 0,\end{aligned}$$

where we see $\mathbb{E}[f_i(t)|\boldsymbol{X}_i]$ is obtained by

$$\begin{aligned}\mathbb{E}[f_i(t)|\boldsymbol{X}_i] &= \mathbb{E}\left[\frac{\delta_i(t)}{H\left(t\wedge\widetilde{T}_i\right)}\mathbb{1}\{\widetilde{T}_i \le t\}|\boldsymbol{X}_i\right] = \mathbb{E}\left[\frac{\delta_i(t)}{G\left(t\wedge T_i\right)}\mathbb{1}\{T_i \le t\}|\boldsymbol{X}_i\right]\\ &= \mathbb{E}\left[\frac{\mathbb{1}\{T_i \le t\}}{G\left(t\wedge T_i\right)}\mathbb{E}\left[\mathbb{1}\{C_i \ge t\wedge T_i\}|T_i,\boldsymbol{X}_i\right]|\boldsymbol{X}_i\right] = \mathbb{E}\left[\mathbb{1}\{T_i \le t\}|\boldsymbol{X}_i\right]\\ &= \frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i\right)}\end{aligned}$$

since $\delta_i(t)\mathbb{1}\{\widetilde{T}_i \le t\} = \mathbb{1}\{C_i \ge t\wedge T_i\}\mathbb{1}\{T_i\wedge C_i \le t\} = 1$ means $\widetilde{T}_i = T_i$, and $G\left(t\wedge T_i\right) = P(C \ge t\wedge T_i) = \mathbb{E}\left[\mathbb{1}\{C_i \ge t\wedge T_i\}|T_i,\boldsymbol{X}_i\right]$. □

The following Section B.4 presents some useful Lemmas for the proof of Theorem 4.1.

B.4 A Fuk-Nagaev Concentration Inequality and some technical Lemmas

Theorem B.1. *Define i.i.d random vectors across $i \in [N]$, $\boldsymbol{Z}_i = (Z_{i,1},\cdots,Z_{i,j},\ldots,Z_{i,p})^\top$. Let $\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} Z_{i,j}, \quad \mu_j = \mathbb{E}[Z_{i,j}]$, we have the following*

$$P\left(\max_{1\le j\le p}|\hat{\mu}_j - \mu_j| \ge K_2\frac{\sqrt{\log p}}{\sqrt{N}} + K_{U,1}\frac{\log p}{N^{1-\frac{2}{q}}} + K_{U,2}\frac{\sqrt{\log p}}{\sqrt{N}}\right) \le \frac{C_1}{\log p}, \tag{22}$$

where $q \ge 4$, $K_2 = 2\tilde{K}_2\max_{1\le j\le p}\mathbb{E}\left[|Z_{i,j}|^2\right]$, $K_{U,1} = 2\tilde{K}_2\mathbb{E}\left[\max_{1\le j\le p}|Z_{i,j}|^{\frac{q}{2}}\right]^{\frac{2}{q}}$, $K_{U,2} = \sqrt{\mathbb{E}\left[\max_{1\le j\le p}|Z_{i,j}|^2\right]}$, and $C_1, \tilde{K}_2$ are universal constants.

Proof. First, let

$$\sigma^2 = \max_{1\le j\le p}\sum_{i=1}^{N}\mathbb{E}[|Z_{i,j}|^2], K_U = \mathbb{E}\left[\max_{1\le j\le p}|Z_{i,j}|^2\right], \tilde{K}_U = \mathbb{E}\left[\max_{1\le j\le p}|Z_{i,j}|^{\frac{q}{2}}\right].$$

Then we use the following maximal inequality: By Lemma D.2 of Chernozhukov et al. (2019) is: (set $s = 2$ in their Lemma, for any $a > 0$)

$$P\left[\max_{1\le j\le p} |\hat{\mu}_j - \mu_j| \ge 2\mathbb{E}\left(\max_{1\le j\le p} |\hat{\mu}_j - \mu_j|\right) + \frac{a}{N}\right] \le \exp\left(-a^2/3\sigma^2\right) + K_1 \frac{NK_U}{a^2}, \quad (23)$$

where constant $K_1 > 0$ depends on η and s. When η and s are decided, we can see K_1 is a universal positive constant.

Lemma D.3 of Chernozhukov et al. (2019) shows that with $\tilde{K}_2 > 0$ a universal positive constant, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{1\le j\le p} |\hat{\mu}_j - \mu_j|\right] &\le \tilde{K}_2 \left[\frac{\sigma\sqrt{\log p}}{N} + \frac{\sqrt{\mathbb{E}[\max_{1\le i\le N}\max_{1\le j\le p} |Z_{i,j}|^2]}\log p}{N}\right] \\
&\le \tilde{K}_2 \left[\frac{\sigma\sqrt{\log p}}{N} + \frac{N^{\frac{2}{q}}(\tilde{K}_U)^{\frac{2}{q}}\log p}{N}\right]
\end{aligned}$$

since

$$\begin{aligned}
\mathbb{E}[\max_{1\le i\le N}\max_{1\le j\le p} |Z_{i,j}|^2] &= \mathbb{E}\left[\max_{1\le i\le N}\left(\max_{1\le j\le p} |Z_{i,j}|^2\right)\right] \\
&\le \mathbb{E}\left[\max_{1\le i\le N}\left(\max_{1\le j\le p} |Z_{i,j}|^2\right)^{\frac{q}{4}}\right]^{\frac{4}{q}} \\
&\le \mathbb{E}\left[N\left(\max_{1\le j\le p} |Z_{i,j}|^2\right)^{\frac{q}{4}}\right]^{\frac{4}{q}} \\
&= N^{\frac{4}{q}}\mathbb{E}\left[\left(\max_{1\le j\le p} |Z_{i,j}|^{\frac{q}{2}}\right)\right]^{\frac{4}{q}}.
\end{aligned}$$

Let $a = \sqrt{K_U N \log p}$ in (23), then we have

$$\begin{aligned}
&P\left(\max_{1\le j\le p} |\hat{\mu}_j - \mu_j| \ge 2\tilde{K}_2\left[\frac{\sigma\sqrt{\log p}}{N} + \frac{(\tilde{K}_U)^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}}\right] + \frac{\sqrt{K_U}\sqrt{\log p}}{\sqrt{N}}\right) \\
&\le \exp\left(\frac{-K_U N \log p}{3\sigma^2}\right) + \frac{K_1}{\log p}.
\end{aligned}$$

Since $\sigma^2 = N \max\limits_{1\le j\le p} \mathbb{E}\left[|Z_{i,j}|^2\right] \le NK_U$, we have $\frac{\sigma^2}{N} = \max\limits_{1\le j\le p} \mathbb{E}\left[|Z_{i,j}|^2\right] \le K_U$, so that

$$
\begin{aligned}
&P\left(\max_{1\le j\le p} |\hat{\mu}_j - \mu_j| \ge K_2 \frac{\sqrt{\log p}}{\sqrt{N}} + K_{U,1}\frac{\log p}{N^{1-\frac{2}{q}}} + K_{U,2}\frac{\sqrt{\log p}}{\sqrt{N}}\right) \\
&= P\left(\max_{1\le j\le p} |\hat{\mu}_j - \mu_j| \ge 2\tilde{K}_2 \left[\max_{1\le j\le p} \mathbb{E}\left[|Z_{i,j}|^2\right] \frac{\sqrt{\log p}}{\sqrt{N}} + \frac{(\tilde{K}_U)^{\frac{1}{q}} \log p}{N^{1-\frac{2}{q}}}\right] + \frac{\sqrt{K_U}\sqrt{\log p}}{\sqrt{N}}\right) \\
&\le \exp\left(\frac{-K_U N \log p}{3\sigma^2}\right) + \frac{K_1}{\log p} \le \exp\left(-\frac{\log p}{3}\right) + \frac{K_1}{\log p} \le \frac{3}{\log p} + \frac{K_1}{\log p} = \frac{C_1}{\log p},
\end{aligned}
$$

where $K_1, \tilde{K}_2$ are universal constants, $C_1 = K_1 + 3$, $K_2 = 2\tilde{K}_2 \max\limits_{1\le j\le p} \mathbb{E}\left[|Z_{i,j}|^2\right]$, $K_{U,1} = 2\tilde{K}_2 \mathbb{E}\left[\max\limits_{1\le j\le p} |Z_{i,j}|^q\right]^{\frac{1}{q}}$, and $K_{U,2} = \sqrt{\mathbb{E}\left[\max\limits_{1\le j\le p} |Z_{i,j}|^2\right]}$. □

Lemma B.4. *Recall that* $f_i(t) = \frac{\delta_i(t)}{H\left(t\wedge\widetilde{T}_i\right)} \mathbb{1}\{\widetilde{T}_i \le t\}, i \in [N]$, *under Assumption 3.1, 3.2 and 4.1, we have the following*

$$
\left|\frac{1}{N}\sum_{i=1}^{N} f_i(t)\boldsymbol{X}_i - \mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right|_\infty \ge A_1 \frac{\sqrt{\log p}}{\sqrt{N}} + A_2 \frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}} + A_3 \frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}} \tag{24}
$$

holds with probability at most $\frac{C_2}{\log p}$, *and* C_2 *is a universal constant and* A_1, A_2, A_3 *are all positively related to* K_0 *and negatively related to* C_r.

Proof. Let $\boldsymbol{Z}_i$ in Theorem B.1 be $f_i(t)\boldsymbol{X}_i$ and by Theorem B.1, we know that

$$
\left|\frac{1}{N}\sum_{i=1}^{N} f_i(t)\boldsymbol{X}_i - \mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right|_\infty \ge K_4 \frac{\sqrt{\log p}}{\sqrt{N}} + K_{U,3}\frac{\log p}{N^{1-\frac{2}{q}}} + K_{U,4}\frac{\sqrt{\log p}}{\sqrt{N}}
$$

holds with probability at most $\frac{C_2}{\log p}$. $C_2, \tilde{K}_4$ are universal constants, $K_4 = 2\tilde{K}_4 \max\limits_{1\le j\le p} \mathbb{E}\left[|f_i(t)X_{i,j}|^2\right]$, $K_{U,3} = 2\tilde{K}_4 \mathbb{E}\left[\max\limits_{1\le j\le p} |f_i(t)X_{i,j}|^{\frac{q}{2}}\right]^{\frac{2}{q}}$, and $K_{U,4} = \sqrt{\mathbb{E}\left[\max\limits_{1\le j\le p} |f_i(t)X_{i,j}|^2\right]}$.

First, we see that

$$
\begin{aligned}
\mathbb{E}[f_i^{\frac{q}{2}}(t)|\boldsymbol{X}_i] &= \mathbb{E}\left[\frac{\delta_i^{\frac{q}{2}}(t)}{H^{\frac{q}{2}}\left(t\wedge\widetilde{T}_i\right)} \mathbb{1}^{\frac{q}{2}}\{\widetilde{T}_i \le t\}\middle|\boldsymbol{X}_i\right] = \mathbb{E}\left[\frac{\delta_i(t)}{H^{\frac{q}{2}}\left(t\wedge T_i\right)}\mathbb{1}\{T_i \le t\}\middle|\boldsymbol{X}_i\right] \\
&= \mathbb{E}\left[\frac{\mathbb{1}\{T_i \le t\}}{H^{\frac{q}{2}}\left(t\wedge T_i\right)}\mathbb{E}\left[\mathbb{1}\{C_i \ge t\wedge T_i\}|T,\boldsymbol{X}_i\right]|\boldsymbol{X}_i\right] = \mathbb{E}\left[\frac{\mathbb{1}\{T_i \le t\}}{H^{\frac{q}{2}-1}\left(t\wedge T_i\right)}\middle|\boldsymbol{X}_i\right] \\
&\le \frac{1}{C_r^{\frac{q}{2}-1}} \frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0 + E_i\right)}
\end{aligned}
$$

since $\delta_i(t)\mathbb{1}\{\widetilde{T}_i \le t\} = \mathbb{1}\{C_i \ge t\wedge T_i\}\mathbb{1}\{T_i \wedge C_i \le t\} = 1$ means $\widetilde{T}_i = T_i$, and $H\left(t\wedge T_i\right) =$

$P(C \geq t \wedge T_i) = \mathbb{E}\left[\mathbb{1}\{C_i \geq t \wedge T_i\}|T_i, \boldsymbol{X}_i\right]$, and by Assumption 3.2. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{1\leq j\leq p}|f_i(t)X_{i,j}|^{\frac{q}{2}}\right] \\
&= \mathbb{E}\left[f_i^{\frac{q}{2}}(t)\left(\max_{1\leq j\leq p}|X_{i,j}|^{\frac{q}{2}}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[f_i^{q}(t)\left(\max_{1\leq j\leq p}|X_{i,j}|^{\frac{q}{2}}\right)\middle|\boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\left(\max_{1\leq j\leq p}|X_{i,j}|^{\frac{q}{2}}\right)\mathbb{E}\left[f_i^{\frac{q}{2}}(t)|\boldsymbol{X}_i\right]\right] \\
&\leq \mathbb{E}\left[\left(\frac{1}{C_r^{\frac{q}{2}-1}}\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0+E_i\right)}\right)\left(\max_{1\leq j\leq p}|X_{i,j}|^{\frac{q}{2}}\right)\right] \\
&\leq \frac{\mathbb{E}\left[\left(\max_{1\leq j\leq p}|X_{i,j}|^{\frac{q}{2}}\right)\right]}{C_r^{\frac{q}{2}-1}}
\end{aligned}$$

since $\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0\right)} \leq 1$. By Holder's inequality and Assumption 4.1, we then see

$$\mathbb{E}\left[\max_{1\leq j\leq p}|X_{i,j}|^2\right]^{\frac{1}{2}} \leq \mathbb{E}\left[\max_{1\leq j\leq p}|X_{i,j}|^{\frac{q}{2}}\right]^{\frac{2}{q}} \leq \mathbb{E}\left[\max_{1\leq j\leq p}|X_{i,j}|^{q}\right]^{\frac{1}{q}} \leq \left(p\max_{1\leq j\leq p}\mathbb{E}\left[|X_{i,j}|^q\right]\right)^{\frac{1}{q}} \leq (pK_0)^{\frac{1}{q}}, \tag{25}$$

which means $\mathbb{E}\left[\max_{1\leq j\leq p}|f_iX_{i,j}|^{\frac{q}{2}}\right]^{\frac{2}{q}} \leq \frac{(pK_0)^{\frac{1}{q}}}{C_r^{1-\frac{2}{q}}}$ that is $K_{U,3} \leq 2\tilde{K}_4\frac{(pK_0)^{\frac{1}{q}}}{C_r^{1-\frac{2}{q}}} \sim p^{\frac{1}{q}}$ and $K_{U,4} \leq \frac{(pK_0)^{\frac{1}{q}}}{\sqrt{C_r}} \sim p^{\frac{1}{q}}$.

Also for $\max_{1\leq j\leq p}\mathbb{E}\left[|f_iX_{i,j}|^2\right]$, we can similarly prove that

$$\max_{1\leq j\leq p}\mathbb{E}\left[|f_iX_{i,j}|^2\right] \leq \max_{1\leq j\leq p}\frac{\mathbb{E}\left[|X_{i,j}|^2\right]}{C_r} \leq \frac{1\wedge\max_{1\leq j\leq p}\mathbb{E}\left[|X_{i,j}|^q\right]}{C_r}, q \geq 4,$$

which means $K_4 \leq 2\tilde{K}_4\left(\frac{1\wedge\max_{1\leq j\leq p}\mathbb{E}\left[|X_{i,j}|^q\right]}{C_r}\right)$. Notice that K_4, $K_{U,3}$ and $K_{U,4}$ are all positively related to K_0 and negatively related to C_r.

Overall, under Assumption 3.1, 3.2, and 4.1, there exist C_2 which is a universal constant and $A_1, A_2, A_3 \sim K_0$, we have

$$
\begin{aligned}
&P\left(\left|\frac{1}{N}\sum_{i=1}^{N} f_i(t)\boldsymbol{X}_i - \mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right|_\infty \geq A_1\frac{\sqrt{\log p}}{\sqrt{N}} + A_2\frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}} + A_3\frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}}\right)\\
&\leq P\left(\left|\frac{1}{N}\sum_{i=1}^{N} f_i(t)\boldsymbol{X}_i - \mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right|_\infty \geq K_4\frac{\sqrt{\log p}}{\sqrt{N}} + K_{U,3}\frac{\log p}{N^{1-\frac{2}{q}}} + K_{U,4}\frac{\sqrt{\log p}}{\sqrt{N}}\right)\\
&\leq \frac{C_2}{\log p}.
\end{aligned}
$$

□

Lemma B.5. $\forall\boldsymbol{\beta}\in\mathbb{R}^p$, *under Assumption 4.1, we have the following*

$$
\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}\boldsymbol{X}_i - \mathbb{E}\left[\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}\boldsymbol{X}_i\right]\right|_\infty \geq A_4\frac{\sqrt{\log p}}{\sqrt{N}} + A_5\frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}} + A_6\frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}} \tag{26}
$$

holds with probability at most $\frac{C_3}{\log p}$*, and* C_3 *is a universal constant and* $A_4, A_5, A_6 \sim K_0$.

Proof. $\forall i$, let $\boldsymbol{Z}_i$ in Theorem B.1 be $\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}\boldsymbol{X}_i$ and then we know that

$$
\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}\boldsymbol{X}_i - \mathbb{E}\left[\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}\boldsymbol{X}_i\right]\right|_\infty \geq K_6\frac{\sqrt{\log p}}{\sqrt{N}} + K_{U,5}\frac{\log p}{N^{1-\frac{2}{q}}} + K_{U,6}\frac{\sqrt{\log p}}{\sqrt{N}}
$$

holds with probability at most $\frac{C_3}{\log p}$. $C_3, \tilde{K}_6$ are universal constants. It follows from (25) that

$$
K_6 = 2\tilde{K}_6\max_{1\leq j\leq p}\mathbb{E}\left[\left|\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}X_{i,j}\right|^2\right] \leq 2\tilde{K}_6\max_{1\leq j\leq p}\mathbb{E}\left[\left|X_{i,j}\right|^2\right] \leq 2\tilde{K}_6\left(1\wedge\max_{1\leq j\leq p}\mathbb{E}\left[\left|X_{i,j}\right|^q\right]\right) \sim K_0
$$

$$
K_{U,5} = 2\tilde{K}_6\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}X_{i,j}\right|^{\frac{q}{2}}\right]^{\frac{2}{q}} \leq 2\tilde{K}_6\mathbb{E}\left[\max_{1\leq j\leq p}\left|X_{i,j}\right|^{\frac{q}{2}}\right]^{\frac{2}{q}} \leq 2\tilde{K}_6(pK_0)^{\frac{1}{q}},
$$

and

$$
K_{U,6} = \sqrt{\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}X_{i,j}\right|^2\right]} \leq \sqrt{\mathbb{E}\left[\max_{1\leq j\leq p}\left|X_{i,j}\right|^2\right]} \leq (pK_0)^{\frac{1}{q}}.
$$

The order of $K_{U,5}$ and $K_{U,6}$ is similarly obtained as $K_{U,3}$ in Lemma B.4. Then we can say that there exist C_3 which is a universal constant and $A_4, A_5, A_6 \sim K_0$, we have the following

$$\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}\boldsymbol{X}_i-\mathbb{E}\left[\frac{\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}{1+\exp\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i\right)}\boldsymbol{X}_i\right]\right|_\infty \geq A_4\frac{\sqrt{\log p}}{\sqrt{N}}+A_5\frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}}+A_6\frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}} \tag{27}$$

holds with probability at most $\frac{C_3}{\log p}$. □

Lemma B.6. *Under Assumption 4.1, $\forall\boldsymbol{\beta}\in\mathbb{R}^p$, the following*

$$\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top-\mathbb{E}\left[\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right|_\infty \geq B_1\frac{\sqrt{\log p}}{\sqrt{N}}+B_2\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}}+B_3\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}} \tag{28}$$

holds with probability at most $\frac{C_4}{\log p}$, and C_4 is a universal constant, $B_1, B_2, B_3 \sim K_0$.

Proof. Let the $\boldsymbol{Z}_i$ in Theorem B.1 becomes $\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}vech(\boldsymbol{X}_i\boldsymbol{X}_i^\top)$, and by Theorem B.1, we know that

$$\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top-\mathbb{E}\left[\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right|_\infty \geq K_8\frac{\sqrt{2\log p}}{\sqrt{N}}+K_{U,7}\frac{2\log p}{N^{1-\frac{2}{q}}}+K_{U,8}\frac{\sqrt{2\log}}{\sqrt{N}} \tag{29}$$

holds with probability at most $\frac{C_4'}{2\log p}$. C_4', $\tilde{K}_8$ are universal constants. It follows from Assumption 4.1 that

$$K_8=2\tilde{K}_8\max_{1\leq j,l\leq p}\mathbb{E}\left[\left(\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\right)^2|X_{i,j}X_{i,l}|^2\right]\leq 2\tilde{K}_8\max_{1\leq j\leq p}\mathbb{E}\left[|X_{i,j}X_{i,l}|^2\right]\lesssim 2\tilde{K}_8K_0,$$

$$K_{U,7}=2\tilde{K}_8\mathbb{E}\left[\max_{1\leq j,l\leq p}\left(\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\right)^{\frac{q}{2}}|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right]^{\frac{2}{q}}\leq 2\tilde{K}_8\mathbb{E}\left[\max_{1\leq j,l\leq p}|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right]^{\frac{2}{q}}\leq 2\tilde{K}_8p^{\frac{2}{q}}K$$

and

$$K_{U,8}=\sqrt{\mathbb{E}\left[\max_{1\leq j,l\leq p}\left(\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\right)^2|X_{i,j}X_{i,l}|^2\right]}\leq\sqrt{\mathbb{E}\left[\max_{1\leq j,l\leq p}|X_{i,j}X_{i,l}|^2\right]}\leq p^{\frac{2}{q}}K_0^{\frac{2}{q}},$$

since

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq j,l\leq p}|X_{i,j}X_{i,l}|^2\right]^{\frac{1}{2}}\leq\mathbb{E}\left[\max_{1\leq j,l\leq p}|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right]^{\frac{2}{q}}&\leq\mathbb{E}\left[\sum_{j=1}^{p}\max_{1\leq l\leq p}|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right]^{\frac{2}{q}}\\&\leq\left(p\max_{1\leq j,l\leq p}\mathbb{E}\left[|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right]\right)^{\frac{2}{q}}\leq p^{\frac{2}{q}}K_0^{\frac{2}{q}}.\end{aligned} \tag{30}$$

The order of $K_{U,7}$ and $K_{U,8}$ is similarly obtained as $K_{U,3}$ in Lemma B.4. Then we can say that there exist C_4 which is a universal constant and $B_1, B_2, B_3 \sim K_0$, we have the following

$$\left|\frac{1}{N}\sum_{i=1}^{N}\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top - \mathbb{E}\left[\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right|_\infty \geq B_1\frac{\sqrt{\log p}}{\sqrt{N}} + B_2\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} + B_3\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}} \tag{31}$$

holds with probability at most $\frac{C_4}{\log p}$. □

Lemma B.7. *Under Assumption 4.1, the following*

$$\left|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^\top - \mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right|_\infty \geq B_4\frac{\sqrt{\log p}}{\sqrt{N}} + B_5\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} + B_6\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}} \tag{32}$$

holds with probability at most $\frac{C_5}{\log p}$*, and* C_5 *is a universal constant,* $B_4, B_5, B_6 \sim K_0$*.*

Proof. Let the $\boldsymbol{Z}_i$ in Theorem B.1 becomes $vech(\boldsymbol{X}_i\boldsymbol{X}_i^\top)$, and by Theorem B.1, we know that

$$\left|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^\top - \mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right|_\infty \geq K_{10}\frac{\sqrt{2\log p}}{\sqrt{N}} + K_{U,9}\frac{2\log p}{N^{1-\frac{2}{q}}} + K_{U,10}\frac{\sqrt{2\log p}}{\sqrt{N}} \tag{33}$$

holds with probability at most $\frac{C_5'}{2\log p}$. $C_5', \tilde{K}_8$ are universal constants, and from (30)

$$K_{10} = 2\tilde{K}_{10}\max_{1\leq j,l\leq p}\mathbb{E}\left[|X_{i,j}X_{i,l}|^2\right] \lesssim 2\tilde{K}_{10}K_0,$$

$$K_{U,9} = 2\tilde{K}_{10}\mathbb{E}\left[\max_{1\leq j,l\leq p}|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right]^{\frac{2}{q}} \leq 2\tilde{K}_{10}p^{\frac{2}{q}}K_0^{\frac{2}{q}},$$

and

$$K_{U,10} = \sqrt{\mathbb{E}\left[\max_{1\leq j,l\leq p}|X_{i,j}X_{i,l}|^2\right]} \leq p^{\frac{2}{q}}K_0^{\frac{2}{q}}.$$

The order of $K_{U,9}$ and $K_{U,10}$ is similarly obtained as $K_{U,3}$ in Lemma B.4. Then we can say that there exist C_5 which is a universal constant and $B_4, B_5, B_6 \sim K_0$, we have the following

$$\left|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^\top - \mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right|_\infty \geq B_4\frac{\sqrt{\log p}}{\sqrt{N}} + B_5\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} + B_6\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}} \tag{34}$$

holds with probability at most $\frac{C_5}{\log p}$. □

B.5 Probability Inequalities for the Empirical Process

Theorem B.2. *Under Assumptions 3.1, 3.2 and 4.1, $\forall \epsilon_1 > 0$, $\exists U_{\epsilon_1} > 0$, define*

$$\lambda_\epsilon \sim a_1 \frac{\sqrt{\log p}}{\sqrt{N}} + a_2 \frac{p^{\frac{1}{q}} \log p}{N^{1-\frac{2}{q}}} + a_3 \frac{p^{\frac{1}{q}} \sqrt{\log p}}{\sqrt{N}}, \lambda_C \sim \frac{p^{\frac{2}{q}}}{\sqrt{N}},$$

and

$$\lambda' \sim \lambda_\epsilon + U_{\epsilon_1} \lambda_C.$$

We have

$$\begin{aligned} &P\left(\sup_{\boldsymbol{\beta}' : \Omega(\boldsymbol{\beta}' - \boldsymbol{\beta}) \le M_{\boldsymbol{\beta}}} |[R_N(\boldsymbol{\beta}') - R(\boldsymbol{\beta}'|\boldsymbol{X})] - [R_N(\boldsymbol{\beta}) - R(\boldsymbol{\beta}|\boldsymbol{X})]| \le \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \right) \\ &\ge 1 - \epsilon_1 - \frac{1}{p} - \frac{c_1}{\log p}, \end{aligned} \tag{35}$$

where $\lambda' \sim \lambda_\epsilon + U_{\epsilon_1}\lambda_C$. c_1 is a universal constant, a_1, a_2, a_3 are all constants positively related to K_0 and negatively related to C_r.

Proof. Recall the empirical and theoretical risk function, by Lemma B.1 we see that

$$\begin{aligned} &\sup_{\boldsymbol{\beta}' : \Omega(\boldsymbol{\beta}' - \boldsymbol{\beta}) \le M_{\boldsymbol{\beta}}} |[R_N(\boldsymbol{\beta}') - R(\boldsymbol{\beta}'|\boldsymbol{X})] - [R_N(\boldsymbol{\beta}) - R(\boldsymbol{\beta}|\boldsymbol{X})]| \\ &= \sup_{\boldsymbol{\beta}' : \Omega(\boldsymbol{\beta}' - \boldsymbol{\beta}) \le M_{\boldsymbol{\beta}}} \left| \frac{1}{N} \sum_{i=1}^{N} \left(-\widehat{f}_i(t) + \mathbb{E}\left[f_i(t) | \boldsymbol{X}_i \right] \right) \boldsymbol{X}_i^\top (\boldsymbol{\beta}' - \boldsymbol{\beta}) \right| + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \\ &\le \sup_{\boldsymbol{\beta}' : \Omega(\boldsymbol{\beta}' - \boldsymbol{\beta}) \le M_{\boldsymbol{\beta}}} \Omega_* \left(\frac{1}{N} \sum_{i=1}^{N} \left(-\widehat{f}_i(t) + \mathbb{E}\left[f_i(t) | \boldsymbol{X}_i \right] \right) \boldsymbol{X}_i \right) \Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta} \right) + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \\ &\le \Omega_* \left(\frac{1}{N} \sum_{i=1}^{N} \left(-\widehat{f}_i(t) + \mathbb{E}\left[f_i(t) | \boldsymbol{X}_i \right] \right) \boldsymbol{X}_i \right) M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \\ &\le G^* \left| \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{f}_i(t) - \mathbb{E}[f_i(t) | \boldsymbol{X}_i] \right) \boldsymbol{X}_i \right|_\infty M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i|, \end{aligned} \tag{36}$$

where G^* represents the size of the largest group in $\boldsymbol{X}_i$. Then we have a look at the following

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{f}_i(t)-\mathbb{E}[f_i(t)|\boldsymbol{X}_i]\right)\boldsymbol{X}_i\right|_\infty \\
&=\left|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{f}_i(t)-f_i(t)+f_i(t)-\mathbb{E}[f_i(t)|\boldsymbol{X}_i]\right)\boldsymbol{X}_i\right|_\infty \\
&\leq\left|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{f}_i(t)-f_i(t)\right)\boldsymbol{X}_i\right|_\infty+\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_i(t)\boldsymbol{X}_i-\mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]+\mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]-\mathbb{E}[f_i(t)|\boldsymbol{X}_i]\boldsymbol{X}_i\right)\right|_\infty \\
&\leq\left|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{f}_i(t)-f_i(t)\right)\boldsymbol{X}_i\right|_\infty+\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_i(t)\boldsymbol{X}_i-\mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right)\right|_\infty \\
&+\left|\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}[f_i(t)|\boldsymbol{X}_i]\boldsymbol{X}_i-\mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right)\right|_\infty \\
&:=(a1)+(a2)+(a3).
\end{aligned}
\tag{37}
$$

Firstly, we consider the first part of the last row of (37). Recall the definition of $\widehat{f}_i(t)$ and $f_i(t)$, we have

$$
\begin{aligned}
\left|\widehat{f}_i(t)-f_i(t)\right| &= \left|\frac{\delta_i(t)}{\widehat{H}\left(t\wedge\widetilde{T}_i\right)}\mathbb{1}\{\widetilde{T}_i\leq t\}-\frac{\delta_i(t)}{H\left(t\wedge\widetilde{T}_i\right)}\mathbb{1}\{\widetilde{T}_i\leq t\}\right| \\
&\leq\left|\frac{1}{\widehat{H}\left(t\wedge\widetilde{T}_i\right)}-\frac{1}{H\left(t\wedge\widetilde{T}_i\right)}\right| \\
&=\left|\frac{\widehat{H}\left(t\wedge\widetilde{T}_i\right)-H\left(t\wedge\widetilde{T}_i\right)}{\widehat{H}\left(t\wedge\widetilde{T}_i\right)H\left(t\wedge\widetilde{T}_i\right)}\right|.
\end{aligned}
$$

By standard properties of the Kaplan-Meier estimator $\widehat{H}$, (see Theorem 1.1 in Gill (1983))), it holds

$$
\sup_{x\in[0,\tau]}|H(x)-\widehat{H}(x)|=O_P\left(N^{-1/2}\right),
$$

where τ is defined in Assumption 3.2. In Assumption 3.2, we assume $H\left(t\wedge\widetilde{T}\right)>0$ and $t\in[0,\tau]$, which means $t\wedge\widetilde{T}\in[0,\tau]$. It follows that:

$$
\max_i\left|\frac{1}{\widehat{H}\left(t\wedge\widetilde{T}_i\right)}-\frac{1}{H\left(t\wedge\widetilde{T}_i\right)}\right|=O_P\left(N^{-1/2}\right),
\tag{38}
$$

which means

$$\max_i \left|\widehat{f}_i(t) - f_i(t)\right| = O_P\left(N^{-1/2}\right). \tag{39}$$

Then $\forall \epsilon_1 > 0$, $\exists U_{\epsilon_1} > 0$, we have

$$P\left(\max_i \left|\widehat{f}_i(t) - f_i(t)\right| \le \frac{U_{\epsilon_1}}{\sqrt{N}}\right) \ge 1 - \epsilon_1. \tag{40}$$

For any $k > 0$, we see that

$$P\left(\max_{1\le j\le p} |X_{i,j}| > k\right) \le \sum_{1\le j\le p} P\left(|X_{i,j}| > k\right) \le p \max_{1\le j\le p} P\left(|X_{i,j}| > k\right) \le p\frac{\max\limits_{1\le j\le p} \mathbb{E}\left(|X_{i,j}|^q\right)}{k^q}, i \in [N],$$

since we have $\max\limits_{1\le j\le p} \mathbb{E}\left(|X_{i,j}|^q\right) \le K_0$ from Assumption 4.1. Let $k = K_0^{\frac{1}{q}} p^{\frac{2}{q}}$, then we have

$$P\left(\max_{1\le j\le p} |X_{i,j}| > K_0^{\frac{1}{q}} p^{\frac{2}{q}}\right) \le p\frac{\max\limits_{1\le j\le p} \mathbb{E}\left(|X_{i,j}|^q\right)}{K_0 p^2} = \frac{1}{p}, i \in [N].$$

By Lemma B.1, we can see

$$\max_i \Omega_*\left(\boldsymbol{X}_i\right) \le \max_i G^* \left|\boldsymbol{X}_i\right|_\infty, \tag{41}$$

Combine (40) and (41), we have $\forall \epsilon_1 > 0$, $\exists U_{\epsilon_1} > 0$

$$\max_i \left|\widehat{f}_i(t) - f_i(t)\right| \Omega_*\left(\boldsymbol{X}_i\right) \le \frac{U_{\epsilon_1} G^* K_0^{\frac{1}{q}} p^{\frac{2}{q}}}{\sqrt{N}} \tag{42}$$

holds with probability $1 - \epsilon_1 - \frac{1}{p}$.

Secondly, we consider the $(a2)$. By Lemma B.4, we have

$$P\left(\left|\frac{1}{N}\sum_{i=1}^N \left(f_i(t)\boldsymbol{X}_i - \mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right)\right|_\infty \ge A_1\frac{\sqrt{\log p}}{\sqrt{N}} + A_2\frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}} + A_3\frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}}\right) \le \frac{C_2}{\log p}. \tag{43}$$

Finanlly, as for $(a3)$, we know that

$$\mathbb{E}[f_i(t)|\boldsymbol{X}_i] = \frac{\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}{1 + \exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)},$$

and

$$\mathbb{E}[f_i(t)\boldsymbol{X}_i] = \mathbb{E}\left[\mathbb{E}[f_i(t)\boldsymbol{X}_i|\boldsymbol{X}_i]\right] = \mathbb{E}\left[\boldsymbol{X}_i\mathbb{E}[f_i(t)|\boldsymbol{X}_i]\right] = \mathbb{E}\left[\frac{\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}{1 + \exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}\boldsymbol{X}_i\right].$$

Applying Lemma B.5, we have

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}[f_i(t)|\boldsymbol{X}_i]\boldsymbol{X}_i-\mathbb{E}\left[f_i(t)\boldsymbol{X}_i\right]\right)\right|_{\infty}\geq A_4\frac{\sqrt{\log p}}{\sqrt{N}}+A_5\frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}}+A_6\frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}}\right)\leq\frac{C_3}{\log p}, \tag{44}$$

where C_2, C_3 are universal constants, A_1, A_2, A_3 are all constants positively related to K_0 and negatively related to C_r and $A_4, A_5, A_6 \sim K_0$.

To conclude, let $a_1 = A_1 + A_4$, $a_2 = A_2 + A_5$, $a_3 = A_3 + A_6$, $c_1 = C_2 + C_3$, combine (42), (43) and (44), it follows that $\forall \epsilon_1 > 0$, $\exists U_{\epsilon_1} > 0$, the following

$$\sup_{\boldsymbol{\beta}':\Omega(\boldsymbol{\beta}'-\boldsymbol{\beta})\leq M_{\boldsymbol{\beta}}}\left|\left[R_N(\boldsymbol{\beta}')-R(\boldsymbol{\beta}'|\boldsymbol{X})\right]-\left[R_N(\boldsymbol{\beta})-R\left(\boldsymbol{\beta}|\boldsymbol{X}\right)\right]\right|\leq\lambda' M_{\boldsymbol{\beta}}+\frac{1}{N}\sum_{i=1}^{N}2|E_i| \tag{45}$$

holds with probability at least $1-\epsilon_1-\frac{1}{p}-\frac{c_1}{\log p}$, where $\lambda_\epsilon \sim a_1\frac{\sqrt{\log p}}{\sqrt{N}}+a_2\frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}}+a_3\frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}}$, $\lambda_C \sim \frac{p^{\frac{2}{q}}}{\sqrt{N}}$, $\lambda' \sim \lambda_\epsilon + U_{\epsilon_1}\lambda_C$, c_1 is a universal constant, a_1, a_2, a_3 are all constants positively related to K_0 and negatively related to C_r. □

B.6 Relationship between Population and Sample Effective Sparsity

Now we tie our population (16) and sample effective sparsity (18).

Lemma B.8. *Given a stretching factor D and let $U_{H_1} = B_1\frac{\sqrt{\log p}}{\sqrt{N}}+B_2\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}}+B_3\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}}$, which is defined in (28). Under Assumption 4.1, we have*

$$\Gamma_{\Omega}^{-2}\left(D,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_{\boldsymbol{\beta}_0}(\Delta)\right)\geq\Gamma_{\Omega}^{-2}\left(D,\mathcal{S}_{\boldsymbol{\beta}},\tau_{\boldsymbol{\beta}_0}(\Delta)\right)-G^*(D+1)^2U_H$$

holds with probability at least $1-\frac{C_4}{\log p}$, where D is a stretching factor, C_4 is a universal constant, $B_1, B_2, B_3 \sim K_0$, and G^* represents the size of the largest group in $\mathcal{G}_{\boldsymbol{\beta}}$.

Proof. Let $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\frac{\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0+E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top\boldsymbol{\beta}_0+E_i)\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^\top$and $\Sigma = \mathbb{E}\left[\hat{\Sigma}\right]$, we have

$$\begin{aligned}\left|\Delta^\top\hat{\Sigma}\Delta\right| &\geq \left|\Delta^\top\mathbb{E}\left[\hat{\Sigma}\right]\Delta\right|-\left|\Delta^\top\left(\hat{\Sigma}-\mathbb{E}\left[\hat{\Sigma}\right]\right)\Delta\right|\\ &\geq \left|\Delta^\top\mathbb{E}\left[\hat{\Sigma}\right]\Delta\right|-G^*\Omega^2\left(\Delta\right)\left|\hat{\Sigma}-\mathbb{E}\left[\hat{\Sigma}\right]\right|_{\infty}.\end{aligned}$$

Since by Lemma B.1, we have

$$\begin{aligned}\left|\Delta^\top\left(\hat{\Sigma}-\mathbb{E}\left[\hat{\Sigma}\right]\right)\Delta\right| &\leq \Omega\left(\Delta\right)\Omega_*\left(\left(\hat{\Sigma}-\mathbb{E}\left[\hat{\Sigma}\right]\right)\Delta\right)\\ &\leq \Omega\left(\Delta\right)G^*\left|\left(\hat{\Sigma}-\mathbb{E}\left[\hat{\Sigma}\right]\right)\right|_{\infty}\Omega\left(\Delta\right)\\ &= G^*\Omega^2\left(\Delta\right)\left|\hat{\Sigma}-\mathbb{E}\left[\hat{\Sigma}\right]\right|_{\infty}.\end{aligned}$$

When $\Omega^{+}(\Delta) = 1$ and $\Omega^{-}(\Delta) \leq D$, we can see $\Omega\left(\Delta\right) = \Omega^{+}(\Delta) + \Omega^{-}(\Delta) \leq D + 1$. Then

$$\left|\Delta^{\top}\hat{\Sigma}\Delta\right| \geq \left|\Delta^{\top}\mathbb{E}\left[\hat{\Sigma}\right]\Delta\right| - G^{*}(D+1)^{2}\left|\hat{\Sigma} - \mathbb{E}\left[\hat{\Sigma}\right]\right|_{\infty}.$$

Applying Lemma B.6, it follows

$$P\left(\left|\hat{\Sigma} - \mathbb{E}\left[\hat{\Sigma}\right]\right|_{\infty} \leq U_{H}\right) \geq 1 - \frac{C_4}{\log p}.$$

Then recall the definition of $\hat{\tau}^{2}_{\boldsymbol{\beta}_0}\left(\Delta\right)$ and $\tau^{2}_{\boldsymbol{\beta}_0}\left(\Delta\right) = \mathbb{E}\left(\hat{\tau}^{2}_{\boldsymbol{\beta}_0}\left(\Delta\right)\right)$, we have

$$\hat{\tau}^{2}_{\boldsymbol{\beta}_0}\left(\Delta\right) \geq \tau^{2}_{\boldsymbol{\beta}_0}\left(\Delta\right) - G^{*}(D+1)^{2}U_{H} \tag{46}$$

holds with probability at least $1 - \frac{C_4}{\log p}$. Take into account (16), (17), (18), (19) and minimize both left and right sides of (46) with respect to Δ, it follows

$$\Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{\boldsymbol{\beta}_0}(\Delta)) \geq \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta)) - G^{*}(D+1)^{2}U_{H}$$

holds with probability $1 - \frac{C_4}{\log p}$. □

Lemma B.9. $\forall \boldsymbol{\beta}' \in \mathbb{R}^{p}$ *which satisfies* $\Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) \leq M$, $\exists M \geq 0$. *Let*

$$U_{H_2} := 2\left(G^{*}(D+1)^{2}\left(B_4\frac{\sqrt{\log p}}{\sqrt{N}} + B_5\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} + B_6\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}}\right) + G^{*}\tilde{K}_0(D+1)^{2}\right)MG^{*}K_0^{\frac{1}{q}}p^{\frac{2}{q}}.$$

Under Assumption 4.1, we have

$$\Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{\boldsymbol{\beta}'}(\Delta)) \geq \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{M}(\Delta)) \geq \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{\boldsymbol{\beta}_0}(\Delta)) - U_{H_2}$$

holds with probability $1 - \frac{C_5}{\log p} - \frac{1}{p}$. *$D$ is a stretching factor, C_5 is a universal constant, $\tilde{K}_0, B_4, B_5, B_6 \sim K_0$, and G^{*} represents the size of the largest group in $\mathcal{G}_{\boldsymbol{\beta}}$.*

Proof. Let $\hat{\Sigma}_M = \frac{1}{N}\sum_{i=1}^{N}\frac{\boldsymbol{X}_i\boldsymbol{X}_i^{\top}}{\mathrm{C}^{2}_{M}(\boldsymbol{X}_i)}$ and $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\frac{\exp(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}_0+E_i)}{\left(1+\exp(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}_0+E_i)\right)^{2}}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}$, for any Δ, we have

$$\left|\Delta^{\top}\hat{\Sigma}_M\Delta\right| \geq \left|\Delta^{\top}\hat{\Sigma}\Delta\right| - \left|\Delta^{\top}\left(\hat{\Sigma}_M - \hat{\Sigma}\right)\Delta\right|.$$

Since

$$\begin{aligned}\left|\Delta^{\top}\left(\hat{\Sigma}_M - \hat{\Sigma}\right)\Delta\right| &= \left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\mathrm{C}^{2}_{M}(\boldsymbol{X}_i)} - \frac{\exp(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}_0 + E_i)}{\left(1 + \exp(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}_0 + E_i)\right)^{2}}\right)\Delta^{\top}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\Delta\right| \\ &\leq \max_{i}\left|\frac{1}{\mathrm{C}^{2}_{M}(\boldsymbol{X}_i)} - \frac{\exp(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}_0 + E_i)}{\left(1 + \exp(\boldsymbol{X}_i^{\top}\boldsymbol{\beta}_0 + E_i)\right)^{2}}\right|\left|\frac{1}{N}\sum_{i=1}^{N}\Delta^{\top}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\Delta\right|.\end{aligned}$$

For any $k > 0$, we have

$$P\left(\max_{1\le j\le p} |X_{i,j}| > k\right) \le \sum_{1\le j\le p} P\left(|X_{i,j}| > k\right) \le p \max_{1\le j\le p} P\left(|X_{i,j}| > k\right) \le p\frac{\max\limits_{1\le j\le p} \mathbb{E}\left(|X_{i,j}|^q\right)}{k^q}, i \in [N],$$

since we have $\max_{1\le j\le p} \mathbb{E}\left(|X_{i,j}|^q\right) \le K_0$ from Assumption 4.1, let $k = K_0^{\frac{1}{q}} p^{\frac{2}{q}}$, then we have

$$P\left(\max_{1\le j\le p} |X_{i,j}| > K_0^{\frac{1}{q}} p^{\frac{2}{q}}\right) \le p\frac{\max\limits_{1\le j\le p} \mathbb{E}\left(|X_{i,j}|^q\right)}{K_0 p^2} \le \frac{1}{p}, i \in [N],$$

By Lemma B.1, we can see

$$\max_i \Omega_*\left(\boldsymbol{X}_i\right) \le \max_i G^* \left|\boldsymbol{X}_i\right|_\infty \le G^* K_0^{\frac{1}{q}} p^{\frac{2}{q}} \tag{47}$$

holds with probability at least $1 - \frac{1}{p}$.

As $\Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) \le M$, we have

$$\begin{aligned}
&\left|\frac{1}{\mathrm{C}_M^2(\boldsymbol{X}_i)} - \frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)\right)^2}\right| \\
&\le \left|\frac{1}{\mathrm{C}_M^2(\boldsymbol{X}_i)} - \left(\frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i + M\Omega_*(\boldsymbol{X}_i)\right)}\right)\left(1 - \frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}\right)\right| \\
&+ \left|\left(\frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i + M\Omega_*(\boldsymbol{X}_i)\right)}\right)\left(1 - \frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}\right) - \frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)}{\left(1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i)\right)^2}\right| \\
&= \left|\left(\frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i + M\Omega_*(\boldsymbol{X}_i)\right)}\right)\left(\frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)} - \frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i - M\Omega_*(\boldsymbol{X}_i)\right)}\right)\right| \\
&+ \left|\left(\frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + M\Omega_*(\boldsymbol{X}_i)\right)} - \frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}\right)\left(1 - \frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}\right)\right| \\
&\le \left|\frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)} - \frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i - M\Omega_*(\boldsymbol{X}_i)\right)}\right| \\
&+ \left|\frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i + M\Omega_*(\boldsymbol{X}_i)\right)} - \frac{1}{1+\exp\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + E_i\right)}\right| \\
&\le 2\left|M\Omega_*(\boldsymbol{X}_i)\right|.
\end{aligned}$$

It follows

$$\begin{aligned}
\max_i \left|\frac{1}{\mathrm{C}_M^2(\boldsymbol{X}_i)} - \frac{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0)}{\left(1+\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0)\right)^2}\right| &\le 2\max_i M\Omega_*\left(\boldsymbol{X}_i\right) \\
&\le 2MG^* K_0^{\frac{1}{q}} p^{\frac{2}{q}}
\end{aligned} \tag{48}$$

holds with probability at least $1 - \frac{1}{p}$.

As for $\left|\frac{1}{N}\sum_{i=1}^{N} \Delta^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \Delta\right|$, when $\Omega^+(\Delta) = 1$ and $\Omega^-(\Delta) \le D$, we can see $\Omega\left(\Delta\right) = \Omega^+(\Delta) + \Omega^-(\Delta) \le D + 1$, then we have

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^{N} \Delta^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \Delta\right| &= \left|\frac{1}{N}\sum_{i=1}^{N} \Delta^\top \left(\boldsymbol{X}_i \boldsymbol{X}_i^\top - \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\right)\Delta + \Delta^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\Delta\right| \\ &\le G^*\Omega^2(\Delta)\left|\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^\top - \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\right|_\infty + \max_i G^*\Omega^2(\Delta)\left|\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\right|_\infty \\ &\le G^*(D+1)^2\left(B_4\frac{\sqrt{\log p}}{\sqrt{N}} + B_5\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} + B_6\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}}\right) + G^*\tilde{K}_0(D+1)^2, \end{aligned} \tag{49}$$

holds with probability at least $1 - \frac{C_5}{\log p}$. The first inequality comes from Lemma B.1. The first part of the last inequality is obtained from Lemma B.7 and the second part is from $\max_{1\le j,l\le p}\mathbb{E}\left(|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right) \le K_0$ from Assumption 4.1, which implies $\tilde{K}_0 = \max_{1\le j,l\le p}\mathbb{E}\left(|X_{i,j}X_{i,l}|\right) \le 1 \wedge \max_{1\le j,l\le p}\mathbb{E}\left(|X_{i,j}X_{i,l}|^{\frac{q}{2}}\right) \sim K_0$ when $q \ge 4$.

Combine (48) and (49), we conclude that

$$\left|\Delta^\top\left(\hat{\Sigma}_M - \hat{\Sigma}\right)\Delta\right| \le U_{H_2}$$

holds with probability at least $1 - \frac{C_5}{\log p} - \frac{1}{p}$. By the same argument in Lemma B.8, we then have

$$\hat{\tau}^2_{\boldsymbol{\beta}'}\left(\Delta\right) \ge \hat{\tau}^2_M\left(\Delta\right) \ge \hat{\tau}^2_{\boldsymbol{\beta}_0}\left(\Delta\right) - U_{H_2} \tag{50}$$

holds with probability at least $1 - \frac{C_5}{\log p} - \frac{1}{p}$. Minimize both left and right sides of (50) with respect to Δ, we have

$$\Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{\boldsymbol{\beta}'}(\Delta)) \ge \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\Delta)) \ge \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{\boldsymbol{\beta}_0}(\Delta)) - U_{H_2}$$

holds with probability $1 - \frac{C_5}{\log p} - \frac{1}{p}$. C_5 is a universal constant, $\tilde{K}_0, B_4, B_5, B_6 \sim K_0$, and G^* represents the size of the largest group in $\mathcal{G}_{\boldsymbol{\beta}}$. □

Theorem B.3. *$\forall \boldsymbol{\beta}' \in \mathbb{R}^p$ which satisfies $\Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) \le M$, $\exists M \ge 0$. We define*

$$U_{H_1} := G^*(D+1)^2\left(B_1\frac{\sqrt{\log p}}{\sqrt{N}} + B_2\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} + B_3\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}}\right),$$

$$U_{H_2} := 2(G^*D+G^*)^2\left(B_4\frac{\sqrt{\log p}}{\sqrt{N}} + B_5\frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} + B_6\frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}}\right) M K_0^{\frac{1}{q}} p^{\frac{2}{q}} + 2\tilde{K}_0(G^*D+G^*)^2 M K_0^{\frac{1}{q}} p^{\frac{2}{q}}.$$

Under Assumption 4.1, the following

$$\Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{\boldsymbol{\beta}'}(\Delta)) \geq \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\Delta)) \geq \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta)) - U_{H_1} - U_{H_2},$$

holds with probability at least $1 - \frac{C_4}{\log p} - \frac{C_5}{\log p} - \frac{1}{p}$*.* D *is a stretching factor,* C_4, C_5 *are universal constants,* $\tilde{K}_0, B_1, B_2, B_3, B_4, B_5, B_6 \sim K_0$*, and* G^* *represents the size of the largest group in* $\mathcal{G}_{\boldsymbol{\beta}}$*.*

B.7 Proof of Theorem 4.1

Proof. First, we assume a candidate oracle $\boldsymbol{\beta}$ that satisfies $\Omega\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right) \leq M_{\boldsymbol{\beta}}$ and let $M = 2M_{\boldsymbol{\beta}}$, we have $\mathbb{B}_{local} = \{\boldsymbol{\beta}' : \Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) \leq 2M_{\boldsymbol{\beta}}\}$ is a convex neighborhood of the true parameter $\boldsymbol{\beta}_0$. Then we see that $\{\boldsymbol{\beta}' : \Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}\right) \leq M_{\boldsymbol{\beta}}\} \subset \mathbb{B}_{local}$. Note that $\Omega(\cdot)$ depends on the candidate oracle $\boldsymbol{\beta}$. The whole proof is divided into two parts and mainly answers the following questions

- Since one point margin condition is only satisfied for the conditional risk $R(\cdot|\boldsymbol{X})$ with the pseudo-norm $\hat{\tau}_M(\cdot)$, we show that under the certain condition that the sample effective sparsity for $\hat{\tau}_M(\cdot)$ can be replaced by the population effective sparsity for $\tau_{\boldsymbol{\beta}_0}(\cdot)$.

- We then work on the event

$$\left\{\left|\left[R_N(\tilde{\boldsymbol{\beta}}) - R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X})\right] - \left[R_N(\boldsymbol{\beta}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right)\right]\right| \leq \lambda' M_{\boldsymbol{\beta}}\right\},$$

  where $\tilde{\boldsymbol{\beta}} \in \mathbb{B}_{local}$. For simplicity, we assume the candidate oracle $\boldsymbol{\beta}$ is near the true parameter $\boldsymbol{\beta}_0$. The goal is to prove $\Omega\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq M_{\boldsymbol{\beta}}$, where $M_{\boldsymbol{\beta}}$ depends on how well this candidate oracle is trading off approximation error $(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)$, estimation error which is related to the regularization parameter λ, and smallish coefficients $s_{\boldsymbol{\beta}}$.

B.7.1 Part 1

Define $b = \left(G^*D + G^*\right)^2 \left(B_1 \vee B_2 \vee B_3 \vee B_4 \vee B_5 \vee B_6\right)$. By Theorem B.3, let

$$\lambda_\Gamma \sim b\left(\frac{\sqrt{\log p}}{\sqrt{N}} \vee \frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}} \vee \frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}}\right).$$

such that

$$U_{H_1} \leq \lambda_\Gamma, U_{H_2} \leq 4\lambda_\Gamma M_{\boldsymbol{\beta}} p^{\frac{2}{q}} + 4\bar{K}_0 M_{\boldsymbol{\beta}} p^{\frac{2}{q}},$$

where $\bar{K}_0 = \tilde{K}_0(G^*D + G^*)^2 K_0^{\frac{1}{q}} \sim K_0$. Then we see that $\forall \boldsymbol{\beta}'$ which satisfies $\Omega\left(\boldsymbol{\beta}' - \boldsymbol{\beta}_0\right) \leq 2M_{\boldsymbol{\beta}}$

$$\Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_{\boldsymbol{\beta}'}(\Delta)) \geq \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\Delta)) \geq \Gamma^{-2}(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta)) - U_{H_1} - U_{H_2}$$

holds with probability at least $1 - \frac{C_4}{\log p} - \frac{C_5}{\log p} - \frac{1}{p}$, where C_4, C_5 are universal constants. It follows

$$\frac{\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta))}{1 - (U_{H_1} + U_{H_2})\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta))} \geq \Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\Delta))$$

when $(U_{H_1} + U_{H_2})\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta)) < 1$.

What is interesting is to understand under what condition $\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\Delta))$ can be bounded by $\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta))$. By Lemma B.2 and (20), we also know $\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta)) \leq \frac{s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}}$. Then as long as

$$(U_{H_1} + U_{H_2})\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta)) \leq \frac{s_{\boldsymbol{\beta}}\lambda_\Gamma + 4s_{\boldsymbol{\beta}}\lambda_\Gamma M_{\boldsymbol{\beta}} p^{\frac{2}{q}} + 4\bar{K}_0 s_{\boldsymbol{\beta}} M_{\boldsymbol{\beta}} p^{\frac{2}{q}}}{\gamma_{\mathrm{H}}} \leq \frac{1}{2}, \tag{51}$$

we can see that

$$\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\Delta)) \leq 2\Gamma^2(D, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\Delta)) \leq \frac{2s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}}$$

holds with probability at least $1 - \frac{C_4}{\log p} - \frac{C_5}{\log p} - \frac{1}{p}$. It is worth mentioning that this result is a new development based on Lemma A.3 of Caner (2023) and Lemma 12.3 of Van De Geer (2016) since we impose the least assumptions on the covariates $\boldsymbol{X}$.

B.7.2 Part 2

We define $\tilde{\boldsymbol{\beta}} := d\hat{\boldsymbol{\beta}} + (1-d)\boldsymbol{\beta}$, where $d := \frac{M_{\boldsymbol{\beta}}}{M_{\boldsymbol{\beta}} + \Omega\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}$. It is clear to see

$$\begin{aligned}\Omega\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \Omega\left(d\hat{\boldsymbol{\beta}} + (1-d)\boldsymbol{\beta} - \boldsymbol{\beta}\right) &= d\Omega\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &= \frac{M_{\boldsymbol{\beta}}\Omega\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{M_{\boldsymbol{\beta}} + \Omega\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)} \leq M_{\boldsymbol{\beta}},\end{aligned} \tag{52}$$

which shows that $\Omega\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq M_{\boldsymbol{\beta}}$ and therefore $\tilde{\boldsymbol{\beta}} \in \mathbb{B}_{local}$. The remaining proof is to show that $\Omega\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq M_{\boldsymbol{\beta}}$.

Note that $R_N(\cdot) + \lambda\Omega(\cdot)$ is convex and $\hat{\boldsymbol{\beta}}$ is the minimizer of (6), we get

$$\begin{aligned}R_N(\tilde{\boldsymbol{\beta}}) + \lambda\Omega(\tilde{\boldsymbol{\beta}}) &\leq dR_N(\hat{\boldsymbol{\beta}}) + d\lambda\Omega(\hat{\boldsymbol{\beta}}) + (1-d)R_N(\boldsymbol{\beta}) + (1-d)\lambda\Omega\left(\boldsymbol{\beta}\right) \\ &\leq R_N(\boldsymbol{\beta}) + \lambda\Omega\left(\boldsymbol{\beta}\right).\end{aligned} \tag{53}$$

Add the term $\left[R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right)\right]$ to both sides of (53), we have:

$$R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right) \leq -\left[\left(R_N(\tilde{\boldsymbol{\beta}}) - R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X})\right) - \left(R_N(\boldsymbol{\beta}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right)\right)\right] + \lambda\Omega\left(\boldsymbol{\beta}\right) - \lambda\Omega(\tilde{\boldsymbol{\beta}}). \tag{54}$$

By the definition of $\Omega(\cdot)$, since $\Omega^-(\boldsymbol{\beta}) = 0$, we have

$$\begin{aligned} \Omega(\boldsymbol{\beta}) - \Omega(\tilde{\boldsymbol{\beta}}) &= \Omega^+(\boldsymbol{\beta}) + \Omega^-(\boldsymbol{\beta}) - \Omega^+(\tilde{\boldsymbol{\beta}}) - \Omega^-(\tilde{\boldsymbol{\beta}}) \\ &= \Omega^+(\boldsymbol{\beta}) - \Omega^+(\tilde{\boldsymbol{\beta}}) - \Omega^-(\tilde{\boldsymbol{\beta}}) \\ &= \Omega^+(\boldsymbol{\beta}) - \Omega^+(\tilde{\boldsymbol{\beta}}) - \Omega^-(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\leq \Omega^+(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \Omega^-(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}). \end{aligned} \tag{55}$$

From (52), we know $\Omega\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq M_{\boldsymbol{\beta}}$ and look back to Theorem B.2, we let

$$\lambda_\epsilon \sim a\left(\frac{\sqrt{\log p}}{\sqrt{N}} \vee \frac{p^{\frac{1}{q}}\sqrt{\log p}}{\sqrt{N}} \vee \frac{p^{\frac{1}{q}}\log p}{N^{1-\frac{2}{q}}}\right), \lambda_C \sim \frac{p^{\frac{2}{q}}}{\sqrt{N}},$$

where a_1, a_2, a_3 are all constants positively related to K_0 and negatively related to C_r and $a = a_1 \vee a_2 \vee a_3$. Applying Theorem B.2, together with (54) and (55), we can see for any $\epsilon_1 > 0$, $\exists U_{\epsilon_1} > 0$, let $\lambda' \sim \lambda_\epsilon + U_{\epsilon_1}\lambda_C$,

$$\begin{aligned} R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right) &\leq \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^N 2|E_i| + \lambda\Omega\left(\boldsymbol{\beta}\right) - \lambda\Omega(\tilde{\boldsymbol{\beta}}) \\ &\leq \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^N 2|E_i| + \lambda\Omega^+(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \lambda\Omega^-(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \end{aligned} \tag{56}$$

holds with the probability at least $1 - \epsilon_1 - \frac{1}{p} - \frac{c_1}{\log p}$ where c_1 is a universal constant.

Then, we divide the rest proof into two cases:

(1) Case 1: $\lambda\Omega^+(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^N 2|E_i| + R(\boldsymbol{\beta}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X})$. In (56), move the $\lambda\Omega^-(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ to the left side, we have:

$$\lambda\Omega^-(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R(\boldsymbol{\beta}|\boldsymbol{X}) \leq 2\lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^N 4|E_i| + R(\boldsymbol{\beta}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X}).$$

Since we have $R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right) \geq 0$ by Lemma B.3, then

$$\lambda\Omega^-(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq 2\lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^N 4|E_i| + 2\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X})\right).$$

Combine $\lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \leq \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| + R(\boldsymbol{\beta}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X})$, we have

$$\lambda\Omega(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq 3\lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 6|E_i| + 3\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X})\right). \tag{57}$$

(2) Case 2 : $\lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \geq \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| + R(\boldsymbol{\beta}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X})$. By using $\lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \leq \lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) + R(\boldsymbol{\beta}_0|\boldsymbol{X}) - R(\boldsymbol{\beta}|\boldsymbol{X})$ and plugging into (56), we have :

$$\begin{aligned} R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right) + \lambda\Omega^{-}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) &\leq \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| + \lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &\leq 2\lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) + R(\boldsymbol{\beta}_0|\boldsymbol{X}) - R(\boldsymbol{\beta}|\boldsymbol{X}). \end{aligned} \tag{58}$$

Since we always have $R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right) \geq 0$ by Lemma B.3, we can see $\Omega^{-}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq 2\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$.

Based on the definition of effective sparsity and let $D = 2$, we have

$$\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq \hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\Gamma_{\Omega}\left(2, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right),$$

then plug $\frac{1}{2}\lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$ into both sides (58), and by Jensen's inequality

$$\begin{aligned} & R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right) + \lambda\Omega^{-}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \frac{1}{2}\lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ & \leq \frac{3}{2}\lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \\ & \leq \frac{3}{2}\lambda\hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\Gamma_{\Omega}\left(2, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right) + \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \\ & \leq \frac{3}{2}\lambda\hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\Gamma_{\Omega}\left(2, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right) + \frac{3}{2}\lambda\hat{\tau}_M\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\Gamma_{\Omega}\left(2, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right) + \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i| \\ & \leq \frac{\hat{\tau}_M^2(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}{2} + \frac{9}{8}\lambda^2\Gamma_{\Omega}^2(2, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})) + \frac{\hat{\tau}_M^2\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)}{2} + \frac{9}{8}\lambda^2\Gamma_{\Omega}^2\left(2, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M\left(\boldsymbol{\beta}-\boldsymbol{\beta}\right)\right) + \lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 2|E_i|. \end{aligned} \tag{59}$$

Since $\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \mathbb{B}_{local}$, by Lemma B.3, we know that

$$R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right) \geq \frac{\hat{\tau}_M^2\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)}{2},$$

and

$$R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right) \geq \frac{\hat{\tau}_M^2\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)}{2}.$$

Then we have

$$\begin{aligned}
&\lambda\Omega^{-}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})+\frac{1}{2}\lambda\Omega^{+}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\\
&\leq \frac{9}{8}\lambda^2\Gamma^2\left(2,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_M\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right)+\frac{9}{8}\lambda^2\Gamma^2\left(2,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_M\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right)+\lambda' M_{\boldsymbol{\beta}}+\frac{1}{N}\sum_{i=1}^{N}2|E_i|+2\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right).
\end{aligned} \tag{60}$$

Let $\lambda = 12\lambda' \vee \lambda_\Gamma$ which has the same order as

$$\lambda \sim K_C\left(\frac{\sqrt{\log p}}{\sqrt{N}} \vee \frac{p^{\frac{2}{q}}\sqrt{\log p}}{\sqrt{N}} \vee \frac{p^{\frac{2}{q}}\log p}{N^{1-\frac{2}{q}}} \vee U_{\epsilon_1}\frac{p^{\frac{2}{q}}}{\sqrt{N}}\right),$$

where K_C is a constant related to K_0 and C_r. Notice that since we choose the stretching factor $D=2$, then K_C does not depend on D. Let $\lambda M_{\boldsymbol{\beta}} = 36\frac{s_{\boldsymbol{\beta}}\lambda^2}{\gamma_{\mathrm{H}}}+24\frac{1}{N}\sum_{i=1}^{N}|E_i|+16\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)$, we see that as long as

$$\gamma_{\mathrm{H}}s_{\boldsymbol{\beta}}\lambda_\Gamma+144s_{\boldsymbol{\beta}}^2p^{\frac{2}{q}}\lambda(\lambda_\Gamma+\bar{K}_0)+\frac{32\gamma_{\mathrm{H}}s_{\boldsymbol{\beta}}p^{\frac{2}{q}}(\lambda_\Gamma+\bar{K}_0)\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda}\leq\frac{\gamma_{\mathrm{H}}^2}{2}, \tag{61}$$

which implies

$$\begin{aligned}
&\frac{s_{\boldsymbol{\beta}}\lambda_\Gamma+4s_{\boldsymbol{\beta}}\lambda_\Gamma M_{\boldsymbol{\beta}}p^{\frac{2}{q}}+4\bar{K}_0s_{\boldsymbol{\beta}}M_{\boldsymbol{\beta}}p^{\frac{2}{q}}}{\gamma_{\mathrm{H}}}=\frac{s_{\boldsymbol{\beta}}\lambda_\Gamma+4s_{\boldsymbol{\beta}}p^{\frac{2}{q}}(\lambda_\Gamma+\bar{K}_0)M_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}}\\
&=\frac{s_{\boldsymbol{\beta}}\lambda_\Gamma+144\frac{s_{\boldsymbol{\beta}}^2p^{\frac{2}{q}}\lambda}{\gamma_{\mathrm{H}}}(\lambda_\Gamma+\bar{K}_0)+\frac{64\gamma_{\mathrm{H}}s_{\boldsymbol{\beta}}p^{\frac{2}{q}}(\lambda_\Gamma+\bar{K}_0)(R(\boldsymbol{\beta}|\boldsymbol{X})-R(\boldsymbol{\beta}_0|\boldsymbol{X}))}{\lambda}}{\gamma_{\mathrm{H}}}\\
&\leq\frac{1}{2},
\end{aligned} \tag{62}$$

then based on (51), we can have

$$\Gamma^2(2,\mathcal{S}_{\boldsymbol{\beta}},\hat{\tau}_M(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0))\leq 2\Gamma^2(2,\mathcal{S}_{\boldsymbol{\beta}},\tau_{\boldsymbol{\beta}_0}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0))\leq\frac{2s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}}$$

holds with probability at least $1-\frac{C_4}{\log p}-\frac{C_5}{\log p}-\frac{1}{p}$.

Here we give a loose version of (61). If we have

$$\gamma_{\mathrm{H}}s_{\boldsymbol{\beta}}\lambda+144(\frac{\gamma_{\mathrm{H}}}{2}+\bar{K}_0)s_{\boldsymbol{\beta}}^2p^{\frac{2}{q}}\lambda+\frac{64(\frac{\gamma_{\mathrm{H}}^2}{2}+\gamma_{\mathrm{H}}\bar{K}_0)s_{\boldsymbol{\beta}}p^{\frac{2}{q}}\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda}\leq\frac{\gamma_{\mathrm{H}}^2}{2},$$

which implies $s_{\boldsymbol{\beta}}\lambda_{\Gamma} \leq s_{\boldsymbol{\beta}}\lambda \leq \frac{\gamma_{\mathrm{H}}}{2}$, it follows that

$$
\begin{aligned}
&\gamma_{\mathrm{H}} s_{\boldsymbol{\beta}}\lambda_{\Gamma} + 144 s_{\boldsymbol{\beta}}^{2} p^{\frac{2}{q}}\lambda(\lambda_{\Gamma} + \bar{K}_0) + \frac{64\gamma_{\mathrm{H}} s_{\boldsymbol{\beta}} p^{\frac{2}{q}}(\lambda_{\Gamma} + \bar{K}_0)\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda} \\
&= \gamma_{\mathrm{H}} s_{\boldsymbol{\beta}}\lambda_{\Gamma} + 144\lambda(s_{\boldsymbol{\beta}}\lambda_{\Gamma} + s_{\boldsymbol{\beta}}\bar{K}_0)s_{\boldsymbol{\beta}} p^{\frac{2}{q}} + \frac{64\gamma_{\mathrm{H}} p^{\frac{2}{q}}(s_{\boldsymbol{\beta}}\lambda_{\Gamma} + s_{\boldsymbol{\beta}}\bar{K}_0)\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda} \\
&\leq \gamma_{\mathrm{H}} s_{\boldsymbol{\beta}}\lambda + 144(\frac{\gamma_{\mathrm{H}}}{2} + s_{\boldsymbol{\beta}}\bar{K}_0)s_{\boldsymbol{\beta}} p^{\frac{2}{q}}\lambda + \frac{64\gamma_{\mathrm{H}} p^{\frac{2}{q}}(\frac{\gamma_{\mathrm{H}}}{2} + s_{\boldsymbol{\beta}}\bar{K}_0)\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda} \\
&\leq \gamma_{\mathrm{H}} s_{\boldsymbol{\beta}}\lambda + 144(\frac{\gamma_{\mathrm{H}}}{2} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\beta}}\bar{K}_0)s_{\boldsymbol{\beta}} p^{\frac{2}{q}}\lambda + \frac{64\gamma_{\mathrm{H}} p^{\frac{2}{q}}(\frac{\gamma_{\mathrm{H}}}{2} s_{\boldsymbol{\beta}} + s_{\boldsymbol{\beta}}\bar{K}_0)\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda} \\
&= \gamma_{\mathrm{H}} s_{\boldsymbol{\beta}}\lambda + 144(\frac{\gamma_{\mathrm{H}}}{2} + \bar{K}_0)s_{\boldsymbol{\beta}}^{2} p^{\frac{2}{q}}\lambda + \frac{64(\frac{\gamma_{\mathrm{H}}^{2}}{2} + \gamma_{\mathrm{H}}\bar{K}_0)s_{\boldsymbol{\beta}} p^{\frac{2}{q}}\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda} \leq \frac{\gamma_{\mathrm{H}}^{2}}{2},
\end{aligned}
$$

since $s_{\boldsymbol{\beta}} \geq 1$.

We can conclude that if $\gamma_{\mathrm{H}} s_{\boldsymbol{\beta}}\lambda + 144(\frac{\gamma_{\mathrm{H}}}{2} + \bar{K}_0)s_{\boldsymbol{\beta}}^{2} p^{\frac{2}{q}}\lambda + \frac{64(\frac{\gamma_{\mathrm{H}}^{2}}{2} + \gamma_{\mathrm{H}}\bar{K}_0)s_{\boldsymbol{\beta}} p^{\frac{2}{q}}(R(\boldsymbol{\beta}|\boldsymbol{X}) - R(\boldsymbol{\beta}_0|\boldsymbol{X}))}{\lambda} \leq \frac{\gamma_{\mathrm{H}}^{2}}{2}$, then

$$
\Gamma^{2}(2, \mathcal{S}_{\boldsymbol{\beta}}, \hat{\tau}_M(\boldsymbol{\beta} - \boldsymbol{\beta}_0)) \leq 2\Gamma^{2}(2, \mathcal{S}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}_0}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)) \leq \frac{2 s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}}
$$

holds with probability at least $1 - \frac{C_4}{\log p} - \frac{C_5}{\log p} - \frac{1}{p}$. Look back to (60), we then have

$$
\lambda\Omega^{-}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \lambda\Omega^{+}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \frac{9\lambda^{2} s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}} + 2\lambda' M_{\boldsymbol{\beta}} + 4\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right). \tag{63}
$$

Combine the results (57) in Case 1 and (63) in Case 2, we conclude here that under the event

$$
\left\{\left|\left[R_N(\tilde{\boldsymbol{\beta}}) - R(\tilde{\boldsymbol{\beta}}|\boldsymbol{X})\right] - \left[R_N(\boldsymbol{\beta}) - R\left(\boldsymbol{\beta}|\boldsymbol{X}\right)\right]\right| \leq \lambda' M_{\boldsymbol{\beta}}\right\},
$$

we have

$$
\lambda\Omega(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \frac{9\lambda^{2} s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}} + 3\lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 6|E_i| + 4\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)
$$

holds with probability at least at least $1 - \frac{C_4}{\log p} - \frac{C_5}{\log p} - \frac{1}{p}$. Then by Theorem B.2, we have

$$
\lambda\Omega(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \frac{9\lambda^{2} s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}} + 3\lambda' M_{\boldsymbol{\beta}} + \frac{1}{N}\sum_{i=1}^{N} 6|E_i| + 4\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)
$$

holds with probability at least at least $1 - \epsilon_1 - \frac{2}{p} - \frac{c}{\log p}$, where $c = c_1 + C_4 + C_5$ is a universal constant.

Recall the value of $\lambda = 12\lambda' \vee \lambda_{\Gamma}$ and $M_{\boldsymbol{\beta}} = 36\frac{s_{\boldsymbol{\beta}}\lambda}{\gamma_{\mathrm{H}}} + 24\lambda^{-1}\frac{1}{N}\sum_{i=1}^{N}|E_i| + 16\lambda^{-1}\left(R(\boldsymbol{\beta}|\boldsymbol{X}) - R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)$, it follows

$$\begin{aligned}\frac{M_{\boldsymbol{\beta}}}{M_{\boldsymbol{\beta}}+\Omega\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}\Omega(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &= \Omega(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &\leq \frac{9\lambda s_{\boldsymbol{\beta}}}{\gamma_{\mathrm{H}}}+\frac{3\lambda'}{\lambda}M_{\boldsymbol{\beta}}+\frac{4\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda}+\frac{6\frac{1}{N}\sum_{i=1}^{N}|E_i|}{\lambda} \\ &\leq \frac{1}{4}M_{\boldsymbol{\beta}}+\frac{1}{4}M_{\boldsymbol{\beta}}=\frac{1}{2}M_{\boldsymbol{\beta}}.\end{aligned} \tag{64}$$

Then it's clear to see that

$$\frac{M_{\boldsymbol{\beta}}}{M_{\boldsymbol{\beta}}+\Omega\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}\Omega(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq \frac{M_{\boldsymbol{\beta}}}{2} \Leftrightarrow 2\Omega(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq M_{\boldsymbol{\beta}}+\Omega(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}),$$

hence we have $\Omega(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq M_{\boldsymbol{\beta}}$.

Overall, under Assumption 3.1, 3.2, 4.2 and 4.1, $\forall\epsilon_1$, $\exists U_{\epsilon_1}>0$, as long as

$$\gamma_{\mathrm{H}}s_{\boldsymbol{\beta}}\lambda+144(\frac{\gamma_{\mathrm{H}}}{2}+\bar{K}_0)s_{\boldsymbol{\beta}}^2p^{\frac{2}{q}}\lambda+\frac{64(\frac{\gamma_{\mathrm{H}}^2}{2}+\gamma_{\mathrm{H}}\bar{K}_0)s_{\boldsymbol{\beta}}p^{\frac{2}{q}}\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)}{\lambda} \leq \frac{\gamma_{\mathrm{H}}^2}{2},$$

we have

$$\Omega(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq M_{\boldsymbol{\beta}} \lesssim \frac{s_{\boldsymbol{\beta}}\lambda}{\gamma_{\mathrm{H}}}+\lambda^{-1}\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)+\lambda^{-1}\frac{1}{N}\sum_{i=1}^{N}|E_i| \tag{65}$$

holds with probability at least $1-\epsilon_1-\frac{2}{p}-\frac{c}{\log p}$. c is a universal constant and $\bar{K}_0$ is a constant depends on K_0. Furthermore, we see that

$$\begin{aligned}R(\hat{\boldsymbol{\beta}}|\boldsymbol{X})-R(\boldsymbol{\beta}|\boldsymbol{X}) &\leq -\left[\left[R_n(\hat{\boldsymbol{\beta}})-R(\hat{\boldsymbol{\beta}}|\boldsymbol{X})\right]-\left[R_n(\boldsymbol{\beta})-R(\boldsymbol{\beta}|\boldsymbol{X})\right]\right]+\lambda\Omega(\boldsymbol{\beta})-\lambda\Omega(\hat{\boldsymbol{\beta}}) \\ &\leq \lambda' M_{\boldsymbol{\beta}}+\lambda\Omega^{+}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &\leq \left(\lambda'+\lambda\right)M_{\boldsymbol{\beta}} \\ &\lesssim \frac{s_{\boldsymbol{\beta}}\lambda^2}{\gamma_{\mathrm{H}}}+\left(R(\boldsymbol{\beta}|\boldsymbol{X})-R\left(\boldsymbol{\beta}_0|\boldsymbol{X}\right)\right)+\frac{1}{N}\sum_{i=1}^{N}|E_i|.\end{aligned}$$

□

C Description of Real Dataset and Additional Simulation Results

In this paper, we build an R package called 'Survivalml' for the methodology and real dataset. This dataset contains 1614 Chinese publicly traded firms in the manufacturing industry with observed financial status. The first listing date, risk date, and industry code of each firm are contained in the dataset. All the financial variables are quarterly measured from $1985/01/01$ to $2015-12-31$. The following table shows the names and units of the variables. We attach a more detailed table, which includes the source and formula definition of these variables, in the Supplementary Materials.

Financial Variables	Unit
Operation ability Related	
Inventory Turnover	times
Accounts Receivable Turnover Ratio	%
Accounts Payable Turnover Ratio	%
Current Assets Turnover Ratio	%
Fixed Assets Turnover Ratio	%
Total Assets Turnover Ratio	%
Debt Related	
Current Ratio	%
Quick Ratio	%
Equity Ratio	%
Total Tangible Assets / Total Liabilities	%
Total Tangible Assets / Interest-Bearing Debt	%
Total Tangible Assets / Net Debt	%
Earnings Before Interest, Tax, Depreciation, and Amortization / Total Liabilities	%
Cash Flow Debt Ratio	%
Time Interest Earned Ratio	%
Long-Term Debt to Capitalization Ratio	%
Profit Related	
Weighted Return on Equity (ROE)	%
Deducted Return on Equity (ROE)	%
Return on Assets (ROA)	%
Net Profit on Assets	%
Return on Invested Capital (ROIC)	%

Sales Margin	%
Gross Profit Margin	%
Net Profit / Total Operating Income	%
Earnings Before Interest and Taxes / Total Operating Income	%
Basic Earnings Per Share (year-on-year growth rate)	%
Diluted Earnings Per Share (year-on-year growth rate)	%
Total Operating Income (year-on-year growth rate)	%
Gross Profit (year-on-year growth rate)	%
Operating Profit (year-on-year growth rate)	%
Total Profit (year-on-year growth rate)	%
Net Profit (year-on-year growth rate)	%
Potential Related	
Net Cash Flow From Operating Activities	%
Cash in Net Profit (year-on-year growth rate)	%
Net Assets (year-on-year growth rate)	%
Total Debt (year-on-year growth rate)	%
Total Assets (year-on-year growth rate)	%
Net Cash Flow (year-on-year growth rate)	%
Z-score Related	
X1 - Working Capital / Total Assets	%
X2 - Retained Earnings / Total Assets	%
X3 - Earnings Before Interest and Taxes / Total Assets	%
X4 - Market Value of Equity / Book Value of Total Liabilities	%
X5 - Sales / Total Assets	%
Capital Related	
Total Shareholders’ Equity / Total Liabilities	%
Debt Ratio	%
Interest-Bearing Debt Ratio	%
Equity Multiplier	%
Current Assets / Total Assets	%
Current Liabilities / Total Liabilities	%
Stock Related	
Earnings Per Share EPS - basic	Yuan
Net Cash Flow from Operating Activities Per Share	Yuan
Operating Income Per Share	Yuan

Profit before Tax Per Share	Yuan
Net Assets Per Share BPS	Yuan
Cash Related	
Net Cash Flow from Operating Activities / Operating Income	%
Net Cash Flow from Operating Activities / Net Income from Operating Activities	%
Net Operating Cash Flow / Operating Income	%

C.1 Data Pre-Process

Since the raw dataset has many missing values in different variables, we propose the following algorithm to extract a complete sub-dataset in which all the firms have survived at least s years. This algorithm balances the number of firms, the number of variables, and the censoring rate of the sub-dataset.

Algorithm 1 Data Process Algorithm

```
Require: Raw dataset, $s$, two initial values $c_1 = 25$ and $c_2 = 25$, step $l = 50$
 1: In the raw dataset, we only select firms that satisfy $\mathbb{1}\{\widetilde{T} \geq s\}$ to form a new dataset. Then
    define the dimension of this dataset as $(N, p)$
 2: For each firm $i$, calculate the number of missing values in variables, say $M_{1,i}$
 3: For each variable $k$, calculate the number of missing values in firms, say $M_{2,k}$
 4: for $a \in (c_1, c_1 + 1 \times l, c_1 + 2 \times l, \ldots, N)$ do
 5:     for $b \in (c_2, c_2 + 1 \times l, c_2 + 2 \times l, \ldots, p)$ do
 6:         Delete firm $i$ with $M_{1,i} \geq b$
 7:         Find variable $k$ with $M_{2,k} \geq a$, then delete all the lags of this variable.
 8:         Delete firms that still have missing values in variables and then calculate the dimension of this sub-dataset, say $N_{a,b}$ and $p_{a,b}$
 9:         if $N_{a,b}/N \geq 0.5$ and $p_{a,b}/p \geq 0.5$, calculate the number of uncensored firms $C_{a,b}$
10:     end for
11: end for
12: We finally picked up the sub-dataset which has the most uncensored firms
```

This algorithm balances the number of firms, the number of variables, and the number of uncensored firms in the sub-dataset. We will get no firms if we delete the firms with missing variable values in the raw dataset. The situation is almost the same if we delete the variables that have missing values. If we want to control the number of uncensored firms to be the largest, without considering the dimension of the processed dataset, we will still see only a few variables left. Step 6 and 7 in Algorithm C.1 is to delete the firms and variables that have extremely large numbers of missing values. Step 9 balances the dimension and the number of uncensored firms of the processed dataset. Table C.1 shows the dimension of a sub-dataset with $s = 6$ chosen by:

- Method 1: Delete all firms that have missing values.
- Method 2: Delete all variables that have missing values.
- Method 3: Using Algorithm C.1, without considering the dimension of processed dataset in Step 9.
- Method 4: Using Algorithm C.1.

Table 12: Dimensions of sub-datasets with $s = 6$ by using different data-process methods (to be revised)

Method	Method 1	Method 2	Method 3	Method 4
Number of firms N	0	1403	951	901
Number of variables K (without lags)	57	0	3	32

D Additional Results

Table 13: Shape of weights estimation accuracy of the three different models: LASSO-UMIDAS (LASSO-U), LASSO-MIDAS (LASSO-M), sg-LASSO-MIDAS (sg-LASSO-M). Entries in odd rows are the average mean integrated squared error, and in even rows the simulation standard error. Results are based on 100 simulated datasets for each sample size.

	$s = 6$, Scenario 1					
	$N = 800$			$N = 1200$		
	$t = t_1$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1, 3) \times (1 + log(t - s))$	0.689	0.633	0.620	0.680	0.609	0.591
	0.000	0.003	0.002	0.001	0.001	0.001
$Beta(2, 3) \times (-1 + log(t - s))$	2.482	2.481	2.481	2.482	2.477	2.472
	0.000	0.000	0.000	0.000	0.000	0.000
	$t = t_2$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1, 3) \times (1 + log(t - s))$	1.092	1.022	0.989	1.076	0.948	0.933
	0.000	0.001	0.001	0.001	0.002	0.001
$Beta(2, 3) \times (-1 + log(t - s))$	1.893	1.878	1.873	1.893	1.807	1.811
	0.000	0.000	0.000	0.000	0.001	0.001
	$t = t_3$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1, 3) \times (1 + log(t - s))$	1.543	1.500	1.467	1.538	1.419	1.395
	0.000	0.001	0.001	0.000	0.001	0.001
$Beta(2, 3) \times (-1 + log(t - s))$	1.433	1.426	1.422	1.433	1.379	1.381
	0.000	0.000	0.000	0.000	0.001	0.001

Figure 4 shows the selected financial covariates obtained by sg-LASSO-MIDAS when $s = 6$ years. These selected covariates are marked in red.

Table 14: Shape of weights estimation accuracy of the three different models: LASSO-UMIDAS (LASSO-U), LASSO-MIDAS (LASSO-M), sg-LASSO-MIDAS (sg-LASSO-M). Entries in odd rows are the average mean integrated squared error, and in even rows the simulation standard error. Results are based on 100 simulated datasets for each sample size.

	$s = 6$, Scenario 2					
	$N = 800$			$N = 1200$		
	$t = t_1$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1,3) \times (1 + log(t - s))$	0.674	0.625	0.616	0.660	0.596	0.590
	0.001	0.001	0.001	0.002	0.001	0.001
$Beta(2,3) \times (-1 + log(t - s))$	2.482	2.437	2.438	2.479	2.345	2.346
	0.000	0.000	0.000	0.000	0.000	0.001
	$t = t_2$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1,3) \times (1 + log(t - s))$	1.079	1.012	1.001	1.049	0.974	0.966
	0.001	0.001	0.001	0.002	0.001	0.001
$Beta(2,3) \times (-1 + log(t - s))$	1.892	1.827	1.827	1.875	1.765	1.765
	0.000	0.000	0.000	0.000	0.000	0.001
	$t = t_3$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1,3) \times (1 + log(t - s))$	1.536	1.492	1.472	1.514	1.435	1.422
	0.001	0.001	0.001	0.001	0.001	0.001
$Beta(2,3) \times (-1 + log(t - s))$	1.432	1.401	1.398	1.422	1.348	1.349
	0.000	0.000	0.000	0.000	0.000	0.000

Table 15: Shape of weights estimation accuracy of the three different models: LASSO-UMIDAS (LASSO-U), LASSO-MIDAS (LASSO-M), sg-LASSO-MIDAS (sg-LASSO-M). Entries in odd rows are the average mean integrated squared error, and in even rows the simulation standard error. Results are based on 100 simulated datasets for each sample size.

	$s = 6$, Scenario 3					
	$N = 800$			$N = 1200$		
	$t = t_1$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1,3) \times (1 + log(t - s))$	0.530	0.521	0.517	0.529	0.515	0.511
	0.000	0.000	0.000	0.000	0.000	0.000
$Beta(2,3) \times (-1 + log(t - s))$	2.790	2.790	2.790	2.790	2.790	2.790
	0.000	0.000	0.000	0.000	0.000	0.000
	$t = t_2$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1,3) \times (1 + log(t - s))$	0.988	0.979	0.977	0.988	0.978	0.974
	0.000	0.000	0.000	0.000	0.000	0.000
$Beta(2,3) \times (-1 + log(t - s))$	2.034	2.034	2.034	2.034	2.033	2.033
	0.000	0.000	0.000	0.000	0.000	0.000
	$t = t_3$					
	LASSO-U	LASSO-M	sg-LASSO-M	LASSO-U	LASSO-M	sg-LASSO-M
$Beta(1,3) \times (1 + log(t - s))$	1.393	1.389	1.387	1.392	1.387	1.385
	0.000	0.000	0.000	0.000	0.000	0.000
$Beta(2,3) \times (-1 + log(t - s))$	1.576	1.576	1.576	1.576	1.575	1.575
	0.000	0.000	0.000	0.000	0.000	0.000

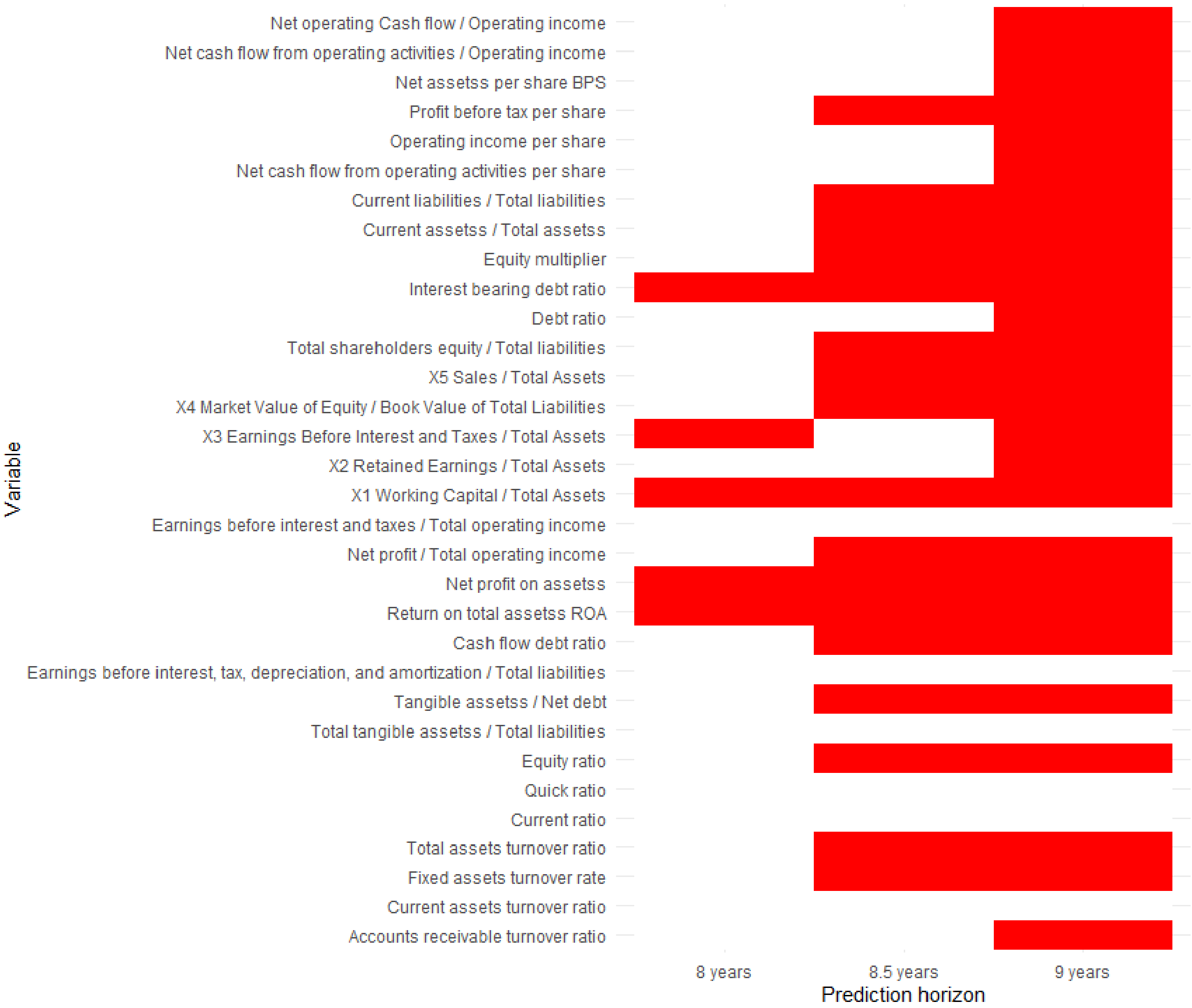

Figure 4: Selected financial variables by sg-LASSO-MIDAS when $s = 6$ years.

E Dictionaries

In this section, we review the choice of dictionaries for the MIDAS weight function. It is possible to construct dictionaries using arbitrary sets of functions, including a mix of algebraic polynomials, trigonometric polynomials, B-splines, Haar basis, or wavelets. In this paper, we mostly focus on dictionaries generated by orthogonalized algebraic polynomials, though it might be interesting to tailor the dictionary for each particular application. The attractiveness of algebraic polynomials comes from their ability to generate a variety of shapes with a relatively low number of parameters, which is especially desirable in low signal-to-noise environments. The general family of appropriate orthogonal algebraic polynomials is given by Jacobi polynomials that nest Legendre, Gegenbauer, and Chebychev's polynomials as a special case.

Example A.1.1 (Jacobi polynomials). Applying the Gram-Schmidt orthogonalization to the power polynomials $\{1, x, x^2, x^3, \ldots\}$ with respect to the measure

$$\mathrm{d}\mu(x) = (1-x)^{\alpha_{\text{poly}}}(1+x)^{\beta_{\text{poly}}}\mathrm{d}x, \quad \alpha_{\text{poly}}, \beta_{\text{poly}} > -1,$$

on $[-1, 1]$, we obtain Jacobi polynomials. In practice, Jacobi polynomials can be computed through the well-known tree-term recurrence relation for $n \geq 0$

$$P_{n+1}^{(\alpha_{\text{poly}},\beta_{\text{poly}})}(x) = axP_{n}^{(\alpha_{\text{poly}},\beta_{\text{poly}})}(x) + bP_{n}^{(\alpha_{\text{poly}},\beta_{\text{poly}})}(x) - cP_{n-1}^{(\alpha_{\text{poly}},\beta_{\text{poly}})}(x)$$

with $a = (2n+\alpha_{\text{poly}}+\beta_{\text{poly}}+1)(2n+\alpha_{\text{poly}}+\beta_{\text{poly}}+2)/2(n+1)(n+\alpha_{\text{poly}}+\beta_{\text{poly}}+1)$, $b = (2n+\alpha_{\text{poly}}+\beta_{\text{poly}}+1)\left(\alpha_{\text{poly}}^2-\beta_{\text{poly}}^2\right)/2(n+1)\,(n+\alpha_{\text{poly}}+\beta_{\text{poly}}+1)(2n+\alpha_{\text{poly}}+\beta_{\text{poly}})$, and $c = (\alpha_{\text{poly}}+n)(\beta_{\text{poly}}+n)(2n+\alpha_{\text{poly}}+\beta_{\text{poly}}+2)/(n+1)(n+\alpha_{\text{poly}}+\beta_{\text{poly}}+1)(2n+\alpha_{\text{poly}}+\beta_{\text{poly}})$. To obtain the orthogonal basis on $[0, 1]$, we shift Jacobi polynomials with affine bijection $x \mapsto 2x - 1$.

For $\alpha_{\text{poly}} = \beta_{\text{poly}}$, we obtain Gegenbauer polynomials, for $\alpha_{\text{poly}} = \beta_{\text{poly}} = 0$, we obtain Legendre polynomials, while for $\alpha_{\text{poly}} = \beta_{\text{poly}} = -1/2$ or $\alpha_{\text{poly}} = \beta_{\text{poly}} = 1/2$, we obtain Chebychev's polynomials of two kinds.

In the mixed frequency setting, non-orthogonalized polynomials, $\{1, x, x^2, x^3, \ldots\}$, are also called Almon polynomials. It is preferable to use orthogonal polynomials in practice due to reduced multicollinearity and better numerical properties. At the same time, orthogonal polynomials are available in Matlab, R, Python, and Julia packages, see more details in the R package 'midasml'.

Gegenbauer polynomials with $\alpha_{\text{poly}} = -\frac{1}{2}$, which is also known as Chebychev's polynomial, are our default recommendation, while other choices of α_{poly} and β_{poly} are preferable if we want to accommodate MIDAS weights with other integrability/tail properties. For example, Babii et al. (2022); Beyhum and Striaukas (2023) recommend Legendre polynomials when nowcasting GDP.