Corporate Probability of Default: A Single-Index Hazard Model Approach

Shaobo Li, Shaonan Tian, Yan Yu, Xiaorui Zhu, and Heng Lian*

August 30, 2022

*Shaobo Li is Assistant Professor (shaobo.li@ku.edu), School of Business, University of Kansas; Shaonan Tian is Associate Professor (shaonan.tian@sjsu.edu), Department of Marketing and Business Analytics, Lucas College and Graduate School of Business, San Jose State University; Yan Yu (corresponding author) is Joseph S. Stern Professor of Business Analytics (Yan.Yu@uc.edu), Department of Operations, Business Analytics, and Information Systems, Carl H. Lindner College of Business, University of Cincinnati; Xiaorui Zhu is Assistant professor (xzhu@towson.edu), Department of Business Analytics and Technology Management, College of Business and Economics, Towson University; and Heng Lian is Associate Professor (henglian@cityu.edu.hk), Department of Mathematics, City University of Hong Kong.

Corporate Probability of Default: A Single-Index Hazard Model Approach

Abstract

Corporate probability of default (PD) prediction is vitally important for risk management and asset pricing. In search of accurate PD prediction, we propose a flexible yet easy-to-interpret default-prediction single-index hazard model (DSI). By applying it to a comprehensive U.S. corporate bankruptcy database we constructed, we discover an interesting V-shaped relationship, indicating a violation of the common linear hazard specification. Most importantly, the single-index hazard model passes the Hosmer-Lemeshow goodness-of-fit calibration test while neither does a state-of-the-art linear hazard model in finance nor a parametric class of Box-Cox transformation survival models. In an economic value analysis, we find that this may translate to as much as three times of profit compared to the linear hazard model. In model estimation, we adopt a penalized-spline approximation for the unknown function and propose an efficient algorithm. With a diverging number of spline knots, we establish consistency and asymptotic theories for the penalized-spline likelihood estimators. Furthermore, we re-examine the distress risk anomaly, that is, higher financially distressed stocks deliver anomalously lower excess returns. Based on the PDs from the proposed single-index hazard model, we find that the distress risk anomaly has weakened or even disappeared during the extended period.

Keywords: Asset Pricing; Bankruptcy Prediction; Nonparametric; Penalized Splines; Survival.

1 Introduction

The corporate probability of default (PD) plays a crucial role in risk management and asset pricing (Altman, 1968; Shumway, 2001; Vassalou and Xing, 2004; Campbell et al., 2008; Ding et al., 2012; Tian et al., 2015). It is often a key metric for credit rating agencies, such as Moody's and Standard & Poor's, to deliver risk assessment for investors. The PDs are also tied to credit spread and loan rate calculations for corporate bonds. In response to 2008 financial crisis, the Basel Committee on Banking Supervision developed an international regulatory framework for banks, Basel III, under which the PD is an essential input to calculate banking's capital and liquidity. Indeed, the PD is a core risk parameter to conduct stress testing and comprehensive capital analysis and review (CCAR), overseen by the Federal Reserve regularly. A dramatic increase in corporate bankruptcies has already been witnessed since the beginning of the global COVID-19 pandemic in 2020. In the United States alone, there are more than 340 companies filed for bankruptcy including big names such as Hertz and J.C. Penney, citing financial distresses due to the COVID-19 (Scigliuzzo et al., 2020). The magnitude of bankruptcies, in terms of assets, has far surpassed the year of 2008, suggesting an unprecedented financial distress (Shen, 2020). In the face of such large levels of financial stress, closely-monitored stress testing and CCAR attain a vital importance that goes beyond the needs to meet regulatory requirements. Undoubtedly, an accurate prediction of the PD is critically important. In search of accurate PD prediction, an accurate, flexible, yet easily-interpretable statistical model is warranted.

In this paper, we build a comprehensive bankruptcy database of U.S. publicly traded firms and develop a flexible default-prediction single-index hazard model (DSI) for corporate PD prediction. One of the most important findings is shown in Figure 1, which depicts the predicted PDs from the proposed DSI model, benchmarking with a state-of-the-art bankruptcy prediction model in finance (diagonal line), termed as CHS (Campbell et al., 2008),[1] or equiv-

[1]CHS is short for Campbell, Hilscher and Szilagyi who are the authors of Campbell et al. (2008). They

alently the Cox discrete-time model (Cox, 1972). Another benchmark model is an "optimal" parametric discrete transformation survival model (DTM) that adopts a class of inverse of Box-Cox and logarithm transformations (Ding et al., 2012).

We observe from Figure 1 that the predicted one-year-ahead PDs[2] from the three models virtually overlay each other in the area to the left of the shaded window, which corresponds to the observations whose PDs fall below the 85 percentile based on CHS model. As the percentiles get larger, the three models deliver noticeably different PDs. In the shaded window, which corresponds to 85 to 99 percentiles in the PDs predicted by CHS, the predicted PDs from our proposed DSI model are the largest, followed by those from the DTM and then from CHS. To the right of the shaded window is the most interesting top one percentile of predicted PDs, where the order is completely reversed and the discrepancy becomes even more substantial. These interesting findings may suggest that for a majority of small predicted PDs at a given time point, the cash reserves based on the predicted PDs from CHS are similar to those from DSI and DTM. But for the highest PD estimates from the top one percentile, the cash reserves from the CHS model would potentially be calculated overly conservatively compared to those from the other two models.

For corporate bankruptcy prediction, a desirable model is usually evaluated in two dimensions: discrimination and calibration performance. While discrimination assesses models' ability to discriminate two dichotomous events, calibration evaluates the agreement between the predicted probabilities and the actual proportions of the event occurrence. Given the crucial role of PD estimates in capital requirement calculation under BASEL III, a model with good calibration performance is essential. In this work, we adopt Hosmer-Lemeshow (HL) goodness-of-fit test (Hosmer Jr et al., 2013) to assess models' calibration power. To

adopted the reduced form of Shumway's linear hazard model (Shumway, 2001), or equivalently the Cox discrete-time model, and proposed market-based firm-specific variables as bankruptcy predictors.

[2]The plotted one-year-ahead PDs are the average of bins that are taken based on percentiles of the predicted PDs from the benchmark CHS model.

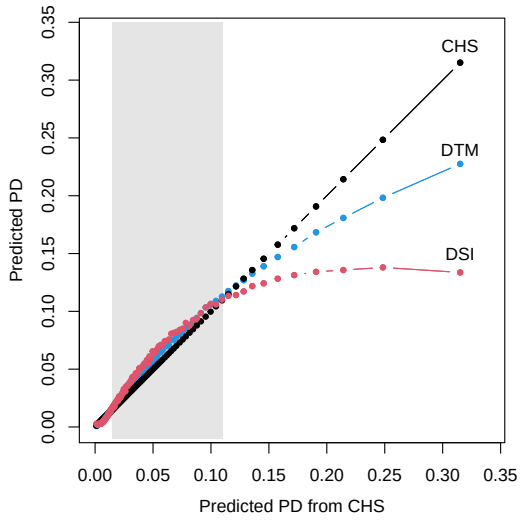

Figure 1: Predicted PDs from a state-of-the-art bankruptcy prediction model in finance, CHS of Campbell et al. (2008) (diagonal line); an optimal discrete transformation survival model (DTM) of Ding et al. (2012); and the proposed default-prediction single-index hazard model (DSI) for corporate bankruptcy prediction.

the best of our knowledge, from our limited empirical analysis, we find that our proposed default-prediction single-index hazard model is the only model in the literature that passes the HL test for both in-sample and out-of-sample predictions across multiple time periods regardless of whether the years of 2008 financial crisis are included or not.[3] Our empirical results also demonstrate a superior discriminatory power of the proposed DSI model.

Additionally, we assess the economic value of different PD prediction models through a business lending practice, where different lenders use different PD prediction models for credit scoring and calculate credit spread for each borrower. Companies with the highest default risk are rejected. For the remaining companies, lenders offer competitive price based

[3]Hosmer-Lemeshow calibration test has been rarely conducted in the corporate bankruptcy literature. It mostly fails based on our limited replication of existing works. To the best of knowledge, Ding et al. (2012) may be the only work documented in the previous literature reporting a satisfactory calibration performance using Hosmer-Lemeshow goodness-of-fit test. Although the "optimal" discrete transformation survival model (DTM) passed the Hosmer-Lemeshow test according to Ding et al. (2012) for a sample period 1980-2006, for the extended sample period 1980-2016 including 2008 financial crisis, we find that DTM with $c = 10$ remains optimal among a class of Box-Cox and logarithm transformation survival models but no longer passes Hosmer-Lemeshow calibration test.

on their PD prediction model, and borrowers choose the lender with the best offer. Under a fixed loan market, profit of each lender can be calculated given prespecified values of loss given default (LGD) and credit spread of the highest quality loan, hence the economic value of different PD prediction models. Our empirical results show that the lender who adopts the proposed DSI consistently generates the highest profit across out-of-sample periods, where the profit may be as much as three times comparing to the lender using CHS model.

The discrepancy observed from Figure 1 may be partly due to the potential violation of the linear assumption in the widely-regarded state-of-the-art CHS or Shumway's linear hazard model in finance. Nielsen et al. (1998) noted that the covariate effect is rarely specified precisely by a parametric model in many applications. In fact, based on the annual data of 17,862 publicly traded firms in the United States from 1980-2016, our empirical analysis finds strong evidence suggesting that the linear specification may be severely violated. More specifically, a V-shaped functional relationship (see Figure 2 and more details in empirical results) is unveiled by applying our proposed DSI model that has the following form,

$$\log\left(\frac{h(t_j|\boldsymbol{x}_{i,t_{j-1}})}{1-h(t_j|\boldsymbol{x}_{i,t_{j-1}})}\right) = \alpha_j + \eta(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{i,t_{j-1}}), \tag{1}$$

where α_j is the baseline hazard at time point t_j and $\eta(\cdot)$ is a flexible univariate function. With a penalized spline estimation, this flexible function turns out to be the aforementioned V shape for the corporate bankruptcy database we construct. The linear projection $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{i,t_{j-1}}$ is the so-called single index, which maps $\boldsymbol{x}_{i,t_{j-1}}$, company i's specific characteristics in previous year t_{j-1}, to a univariate index. The hazard function, $h(t_j|\boldsymbol{x}_{i,t_{j-1}})$ is interpreted as the probability of default of company i in year t_j given its financial characteristics observed in the previous year, i.e., $\Pr(T = t_j|T \geq t_j, \boldsymbol{X} = \boldsymbol{x}_{i,t_{j-1}})$. If $\eta(\cdot)$ is specified as a linear functional form, model (1) reduces to CHS model, or more generally, Cox discrete-time hazard model.

One of the most appealing features of model (1) is that the flexible univariate function $\eta(\cdot)$

can capture potential nonlinear relationship, while the single-index coefficients $\boldsymbol{\beta}$ preserves some model interpretability. A nice by-product is that the single index may be itself of interest. In the bankruptcy application, it may yield an interesting "financial default index" for practitioners. To estimate model (1), we adopt a penalized-spline approximation for its computational stability (Yu and Ruppert (2002); Ruppert et al. (2003); and references therein). Penalized splines can be viewed as a generalization of smoothing splines, and the penalty function helps to prevent over-fitting. Based on a recent theoretical development by Huang and Su (2021), we establish a stochastic bound for the estimated function $\hat{\eta}$ together with the parametric components with a diverging number of knots. Asymptotic normality is shown for the single-index coefficients with the optimal $\sqrt{n}$ order. To the best of our knowledge, this is the first work that establishes the asymptotic results for penalized-spline likelihood estimators with a diverging number of knots under the semiparametric single-index hazard model framework.

The proposed default-prediction single-index hazard model for corporate PD prediction can be viewed as a discrete counterpart of the well studied continuous single-index hazard models in survival analysis. To mention a few, Huang and Liu (2006) developed the single-index proportional hazards model using polynomial splines at pre-specified fixed knots. Lu et al. (2006) studied the partially-linear single-index proportional hazard model using local linear fit adopting an algorithm similar to Carroll et al. (1997). Wang (2004) allowed some covariates to be time-dependent with local partial likelihood estimation along with some missing data imputation.

However, models studied in aforementioned works and other continuous survival literature cannot be directly applied to model corporate default probability. This is mainly because the use of calendar time in our study, while a common time origin is applied to all individuals under the framework of continuous survival analysis. In corporate default prediction problem, it is necessary to use actual calendar time because companies with the same

financial characteristics still have different probability of default at different calendar time due to varying macroeconomic conditions (Ding et al., 2012). Therefore, companies entering the database at different time (depending on their initial public offering (IPO) schedule) should not share the common starting point. With the calendar time, the discrete-time hazards model enjoys the "memoryless" feature that the conditional PD only depends on the latest observation, rather than the entire trajectory of covariate vectors in continuous survival models. In fact, the likelihood function of continuous survival models would be ill-defined for corporate bankruptcy studies as the cumulative conditional hazard integrates the entire trajectory of the time-varying covariates $\boldsymbol{x}(t)$, many of which are clearly not available for the bankruptcy data.

We summarize our contributions as following. First, our work contributes to the literature by unveiling an interesting V-shaped functional relationship (see Figure 2). Such a non-monotone V shape contradicts to the common linearity assumption in the widely-used CHS model or Shumway's model. We propose a default-prediction single-index hazard model with a flexible function $\eta(\cdot)$ along with a nice by-product of "financial default index" and apply it to a comprehensive corporate bankruptcy database we construct.

Second, we empirically show that the proposed DSI model achieves superior prediction performance in both calibration and discriminatory power, compared to benchmark models. To the best of our knowledge, for the U.S. publicly traded companies the proposed DSI model is the only model that passes the Hosmer-Lemeshow goodness-of-fit test (Hosmer Jr et al., 2013) for both in-sample and out-of-sample prediction with various prediction periods, including 2008 financial crises. This suggests that the proposed model is able to recover the PDs more accurately, which may possibly lead to important implications in economical capital reserve calculation in risk management. In addition to the calibration test, the proposed default-prediction single-index hazard model also yields superior prediction accuracy with respect to discrimination. We further assess the economic value of different PD prediction

model through a lending practice under a fixed loan market. The lender who adopts the proposed DSI model for credit scoring consistently generates the highest profit throughout the entire out-of-sample prediction period. The highest economic value generated by DSI model again provides evidence of its superior performance in PD prediction.

Third, we revisit a vital asset pricing implication by applying our proposed default-prediction single-index hazard model to an extended sample period including 2008 financial crisis. The probability of default has also been directly linked to the asset pricing literature. Fama and French (1996) conjecture that the investors demand a positive premium to bear the distress risk. Instead, we find that the return premium associated with higher distress risk is negative. Specifically, the firms with higher distress risk earn lower excess return after controlling the common three factors of Fama and French (1996) or five-factors of Fama and French (2015) for asset pricing. This finding is inconsistent with the original conjecture by Fama and French (1996) but consistent with Campbell et al. (2008), Ding et al. (2012), and Gao et al. (2018). However, unlike documented in these studies, our study shows that the negative distress return anomaly, that is, higher financially distressed stocks deliver anomalously significant lower excess returns, has weakened and even disappeared in this extended sample period including 2008 financial crisis.

Last but not least, our work contributes to the statistical and econometric literature by establishing the asymptotic results for the penalized-spline likelihood estimators of the semiparametric single-index hazard models. The rest of this paper is organized as following: Section 2 describes the corporate bankruptcy database we have constructed and used for our empirical study. Section 3 introduces more details of the default-prediction single-index hazard model along with an estimation algorithm and large sample property. The empirical results are elaborated in Section 4, followed by a simulation study mimicking the real data in Section 5. An asset pricing implication is investigated in Section 6. We conclude the paper in Section 7. Online supplementary materials include additional tables and empirical

results, technical details, and proofs of theorems.

2 Data

In this paper, we construct a bankruptcy database consisting of all the U.S. publicly traded firms listed on the New York Stock Exchange, American Stock Exchange, and NASDAQ from 1980 to 2016.[4] The bankruptcy indicator is coded as "1" for the year that a company filed for bankruptcy protection under either Chapter 7 (liquidation) or Chapter 11 (reorganization), and "0" if the company is healthy, deleted, or delisted due to other reasons such as merger and acquisition. The firm-specific covariates include both accounting and market-based financial data that are collected from the Standard & Poor's COMPUSTAT database and the Center for Research in Security Prices (CRSP) database maintained by the Wharton Research Data Services (WRDS). In particular, we merge a company's quarterly updated accounting information from the COMPUSTAT database with its monthly trading data from the CRSP database. We carefully align the company's fiscal year to the calendar year and also lag all the annual accounting information by four months to ensure that the accounting information is available to the market at the time of prediction (Shumway, 2001; Chava and Jarrow, 2004). As a result, our constructed database consists of 189,037 firm-year observations and 1,589 bankruptcy events during the sample period. A detailed frequency table is shown in Table S1 of supplementary materials.

For the firm-specific covariates, we follow the formation of Campbell et al. (2008) and use the market-valued total asset (MTA) rather than the book value to construct eight exploratory variables, LTMTA, NIMTA, CASHMTA, MBE, RSIZE, EXRET, SIGMA and PRICE. In specific, LTMTA is defined as total liability over market-valued total assets.

[4]Due to substantial delays of bankruptcy disputes as well as some delays in default status updates in COMPUSTAT database, following the literature, we end the sampling period of this study in the year 2016 to avoid inaccurate records.

NIMTA is the ratio of net income to market valued total assets. CASHMTA is constructed by dividing cash and short-term assets by the market value of total assets. We obtain RSIZE as the logarithm of firm's relative size to the S&P 500 index value and EXRET as the annual log excess return against the S&P 500 index return. SIGMA is the volatility of firm stock returns in the past 3 months. Following Campbell et al. (2008), we take the log of stock price as the PRICE variable and also adopt the 10% of the difference between market and book equity to the book value of total assets to construct the market-to-book ratio, MBE. In addition, we also winsorize all predictors at the 1st and 99th percentile in order to reduce the affects by extremely values. Table 1 provides the detailed summary statistics. One observation from Table 1 is that the bankruptcy group has quite different financial characteristics from the non-bankruptcy group. The former tends to have higher debt and liabilities relative to their assets, smaller size in terms of their asset values and market capitalization, weaker profitability, and lower realized stock returns than that of the latter. It is also clear that bankruptcy firms are usually more volatile. The average market return volatility for the bankruptcy group is 1.170, while the volatility is only 0.607 for the non-bankruptcy group. The bankruptcy group also has a lower average trading log-price of 0.467, comparing to 2.268 for the non-bankruptcy group.

Table 1: Summary statistics for bankruptcy predictors.

	Bankruptcy firms (No. Firm-year = 1,589)					*Non-bankruptcy firms (No. Firm-year = 187,448)*				
Variable	Mean	Std.	Min	Med	Max	Mean	Std.	Min	Med	Max
LTMTA	0.638	0.296	0.730	0.014	0.970	0.437	0.283	0.404	0.014	0.970
NIMTA	-0.221	0.248	-0.149	-0.771	0.159	-0.020	0.135	0.016	-0.771	0.159
CASHMTA	0.109	0.180	0.037	0.000	0.747	0.102	0.133	0.053	0.000	0.747
MBE	5.964	12.359	1.582	0.225	59.495	2.882	6.574	1.620	0.225	59.495
RSIZE	-12.539	1.580	-12.670	-14.839	-5.308	-10.508	2.078	-10.621	-14.839	-5.308
EXRET	-0.738	0.765	-0.654	-1.943	1.178	-0.123	0.518	-0.072	-1.943	1.178
SIGMA	1.170	0.650	1.047	0.120	2.419	0.607	0.437	0.478	0.120	2.419
PRICE	0.467	1.323	0.454	-1.520	4.437	2.268	1.309	2.516	-1.520	4.676

3 Semiparametric Single-Index Hazard Model

3.1 Default-Prediction Single-index Hazard Model

Let n be the total number of firms and $\boldsymbol{x}_{i,t_j}$ be the d-dimensional time-dependent covariate vector denoting company i's specific financial characteristics observed at time t, where $t = t_1, ..., t_j, ..., t_J$ are the J fixed discrete observation time in the whole sample period. For our annual data, these are end of year observations. Let A_i denote the starting time and D_i denote the end time that company i, $i = 1, 2, ..., n$, is first and last observed in the database respectively during the sample period. Denote δ_i the censoring indicator, where $\delta_i = 1$ if the ith company files bankruptcy at $t = D_i$ during the sample period; and $\delta_i = 0$ otherwise.

A company may enter the database at different time depending on their initial public offering (IPO) schedule. A company may also exit the database at different time due to its bankruptcy status as well as other delisting reasons. In particular, for healthy companies whose IPO dates are prior to the sample period, their starting date $A_i = t_1$ will be the same as our starting year 1980 in the sample period. For companies with IPO dates later than 1981, then their starting date A_i is their first public trading year. For example, the initial public offering year of Amazon is 1997 at a price of \$18 per share. The starting date for "Amazon" is the year of 1997, where $A_i > t_1$. This kind of different starting time in bankruptcy prediction is very different from the traditional survival analysis. On the other hand, the end time D_i is subject to right censoring at the end of the sample period. If a company files for bankruptcy after the end of sampling period, then $D_i = t_J$. A healthy company may also exit the database through other delisting reasons such as merger and acquisition. For example, Bank One corporation merged with JPMorgan Chase & Co. in 2004. Here for Bank One, the end year $D_i = 2004$ but the censoring indicator $\delta_i = 0$. A healthy company such as Bank One can exit from the database earlier than the end of sample period, through, e.g., merger or acquisition, where $D_i < t_J$ but $\delta_i = 0$.

Let T be the random variable of the calendar year when bankruptcy is filed. As companies do not share the same starting point, T is different from the survival time in the traditional survival analysis. Let

$$p_{i,t_j} = \Pr(T = t_j | T \geq t_j, \boldsymbol{X} = \boldsymbol{x}_{i,t_{j-1}})$$

or simply $p_{i,t_j} = h(t_j|\boldsymbol{x}_{i,t_{j-1}})$ be the conditional probability that company i files for bankruptcy at time t_j given it has survived past time t_{j-1}. Our single-index hazard model (1) for corporate PD prediction can be rewritten as

$$p_{i,t_j} = \frac{1}{1 + \exp\{-\alpha_j - \eta(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{i,t_{j-1}})\}}. \tag{2}$$

For model identifiability, we follow Yu and Ruppert (2002) to impose the constraints that the single-index parameter $\|\boldsymbol{\beta}\|_2 = 1$ and the first element $\beta_1 > 0$.

The log-likelihood function of model (2) takes form

$$l_n(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{A_i < t_j \leq D_i} \left[\delta_{i,j} \log p_{i,t_j} + (1 - \delta_{i,j}) \log(1 - p_{i,t_j})\right], \tag{3}$$

where $\delta_{i,j} = \delta_i I\{D_i = t_j\}$. The mathematical derivation of (3) can be found in supplementary materials (Section B). The memoryless property of model (1) or (2) can also be seen from such mathematical derivation. One of the keys in estimating model (2) is to obtain a consistent estimate of the flexible function $\eta(\cdot)$. In what follows, we discuss penalized spline estimation in detail and establish its asymptotic properties.

3.2 Penalized Spline Estimation

Penalized splines or P-splines can be considered as a generalization of smoothing splines allowing a flexible choice of knots and penalty (see Ruppert et al. (2003) for a review). Yu and Ruppert (2002) showed that P-splines approach to single-index models is advantageous over

other approaches such as local methods (Carroll et al., 1997). Unlike careful knot-placement in regression splines, an appealing feature of P-spines estimation is that the smoothness is tuned by a single penalty parameter λ as in smoothing splines (e.g. Huang and Liu (2006)).

We illustrate P-splines using the truncated power basis for simplicity. Other bases such as B-splines can be easily adopted. The truncated power basis is defined as $\mathbf{B}(u) = (u, u^2, \ldots, u^q, (u - v_1)_+^q, \ldots, (u - v_K)_+^q)$, where q is the polynomial degree, and $v_1, \ldots, v_K$ are K interior knots. The truncation function $(u - v_k)_+^q$ equals $(u - v_k)^q$ if $u \geq v_k$, and 0 otherwise. Popular ways to place spline knots are equally-spaced or equally at sample quantiles. Any function $\eta(u)$ with $q - 1$ continuous derivatives can be approximated by

$$\gamma_1 u + \gamma_2 u^2 + \ldots + \gamma_q u^q + \gamma_{q+1}(u - v_1)_+^q + \ldots + \gamma_{q+k}(u - v_K)_+^q = \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{B}(u),$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{q+K})$ is the $q + K$-dimensional spline coefficient column vector. Note that the intercept term 1 in the spline basis $\mathbf{B}(u)$ as well as the corresponding constant spline coefficient term γ_0 are omitted due to the baseline constant $\boldsymbol{\alpha}$.

Assume that $\eta(u)$ is defined on the interval $[a, b]$, then the penalized-spline log-likelihood can be written as

$$Q_{n,\lambda}(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{n} l_n(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) - \lambda \int_a^b [\eta^{(m)}(u)]^2 du, \tag{4}$$

where $\lambda \int_a^b [\eta^{(m)}(u)]^2 du$ is a general form of the penalty term, $\eta^{(m)}$ is the m-th derivative of η for $m \geq 2$, and $\lambda \geq 0$ is a roughness penalty parameter that can be chosen by generalized cross validation. For $m = 2$, the penalty can be expressed as $\int [\eta''(u)]^2 du = \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{P} \boldsymbol{\gamma}$, where $\mathbf{P}$ is symmetric and positive semidefinite.

3.3 Asymptotic Theories

To present asymptotic theories, we first introduce a few notations. Let $\boldsymbol{\zeta}$ be the unconstrained parameter vector after reparameterization of $\boldsymbol{\beta}$ and $\mathbf{J}(\boldsymbol{\zeta}) = \partial\boldsymbol{\beta}/\partial\boldsymbol{\zeta}$ the Jacobian transformation matrix. Denote $\mathbf{J}(\boldsymbol{\zeta}_0)$ by $\mathbf{J}_0$, where $\boldsymbol{\zeta}_0$ is the true value of $\boldsymbol{\zeta}$. For ease of notation, we replace t_j with j and let $\boldsymbol{x}_{ij} = \boldsymbol{x}_{i,t_{j-1}}$. Define $m_{ij}(a) = -\kappa_{ij}(\delta_{ij}a - \log(1+e^a))$, where $\kappa_{ij} = I\{A_i < t_j \leq D_i\}$. Thus $m_{ij}(\alpha_{0j} + \eta_0(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0))$ is the negative log-likelihood for individual i at time j. We write the number of splines basis $q+K$ simply as K for simplicity since K diverges with n. Without loss of generality, we consider interval $[0,1]$ for $[a,b]$. Further, we define the "projection"

$$(\widetilde{\boldsymbol{\eta}}_0 = (\widetilde{\eta}_{01}, \ldots, \widetilde{\eta}_{0d})^{\mathrm{T}}, \widetilde{\boldsymbol{\alpha}}_0) = \arg\min_{\boldsymbol{\eta},\boldsymbol{\alpha}} E[\sum_{j=1}^{J} m''_{ij}(a_{0j})\|\eta'_0(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{x}_{ij} - \alpha_j - \boldsymbol{\eta}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0)\|^2]. \quad (5)$$

We also need the following regularity conditions to establish our results.

(A1) $P(\kappa_{ij} = 1|\boldsymbol{x}_{ij}) > 0$ for all $j \in \{1, \ldots, J\}$. $P(\delta_{ij} = 1|A_i \leq t_j \leq D_i, \boldsymbol{x}_{ij}) = p_{ij}$.

(A2) $\boldsymbol{x}_{ij}$ are bounded and $\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0$ takes value in $[0,1]$.

(A3) $\eta_0 \in W^p([0,1])$ (Sobolev space of order p) for some integer $p \geq 2$. We also assume $\widetilde{\eta}_{0j} \in W^p([0,1]), j = 1, \ldots, J$. Dimension K of the spline space satisfies $K \to \infty$ and $K\log(n)/n \to 0$.

(A4)

$$E[\mathbf{J}_0^{\mathrm{T}} \sum_{j=1}^{J} m''_{ij}(a_{0j})\{\eta'_0(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0)\}^{\otimes 2}\mathbf{J}_0]$$

and

$$E[\sum_{j=1}^{J} m''_{ij}(a_{0j})(\eta'_0(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0))^2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{T}}]$$

are positive-definite matrices, where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ for any (column) vector $\mathbf{a}$.

(A5) The splines degree q satisfies $q \geq \max\{p-1, m\}$.

Remark 1. *Assumption (A1) specifies the true model. For assumption (A2), boundedness of predictors is usually assumed for spline-based estimation, since spline functions are typically constructed on a finite closed interval. (A3) assumes the smoothness of the true function. The matrices in (A4) are related to the asymptotic covariance matrix of the index parameter. Some projection similar to that defined in* (5) *is often used in separametric models, which represents the effect of the nonparametric part on the parametric part. If $J = 1$, $\widetilde{\eta}$ takes the form of conditional expectation as in Liang et al. (2010). For the general case, there seems to be no simple characterization of $\widetilde{\eta}$.*

The following result shows consistency of penalized-spline likelihood estimators of the unknown function η, the baselines $\boldsymbol{\alpha}$, and single-index coefficients $\boldsymbol{\beta}$.

Theorem 1. *Under assumptions (A1)-(A5), and that $K(K^{-2p} + \lambda K^{2(m-p)_+} + \frac{1}{n\lambda^{1/(2m)}} \wedge \frac{K}{n}) = o(1)$, there exists a local maximizer of the penalized-spline (log-)likelihood (4) that satisfies*

$$\|\widehat{\eta} - \eta_0\|^2 + \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|^2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = O_p\left(K^{-2p} + \lambda K^{2(m-p)_+} + \frac{1}{n\lambda^{1/(2m)}} \wedge \frac{K}{n}\right).$$

Remark 2. *The first two terms in the rate correspond to squared bias and the third is the variance term. To see that the rate obtained here can produce the optimal rate, we can choose $K \asymp n^{\frac{1}{2p+1}}$ and $\lambda \lesssim n^{-\frac{2m}{2p+1}}$, and we will get the optimal rate $n^{-\frac{2p}{2p+1}}$. This situation is often referred to as "light penalization" where the complexity is mainly controlled by choosing the optimal number of knots. On the other hand, assuming $m = p$, choosing $\lambda \asymp n^{-\frac{2p}{2p+1}}$ and K can be much larger than $n^{\frac{1}{2p+1}}$, we still can get the optimal rate $n^{-\frac{2p}{2p+1}}$, which is referred to as "heavy penalization".*

The result below establishes asymptotic normality for the single-index coefficients $\boldsymbol{\beta}$.

Theorem 2. *Under assumptions for Theorem 1, there exists a local maximizer of the penalized-spline (log-)likelihood (4) that satisfies*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{J}_0 \boldsymbol{\Sigma}^{-1} \mathbf{J}_0^{\mathrm{T}}),$$

where $\boldsymbol{\Sigma} = \mathbf{J}_0^{\mathrm{T}} \left\{ \sum_{j=1}^{J} m''_{ij}(a_{0j}) \left(\eta'_0(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \boldsymbol{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \right)^{\otimes 2} \right\} \mathbf{J}_0$, *and* $\widetilde{\alpha}_{0j}$, $\widetilde{\boldsymbol{\eta}}_0$ *are as defined in* (5).

Proofs of both Theorems 1 and 2 are given in online supplementary materials.

3.4 Algorithm

We propose an efficient iterative algorithm to maximize the penalized-spline log-likelihood function (4). The solution can be decomposed to two components: the single-index coefficients $\boldsymbol{\beta}$ and the rest coefficients $\boldsymbol{\gamma}_\alpha = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$. They can be estimated iteratively, or through the profile likelihood estimation (Liang et al., 2010; Yu et al., 2017) using standard nonlinear optimization software such as `nlm`. However, large nonlinear optimization may be computationally expensive and unstable. Instead, we advocate an iterative algorithm, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_\alpha$ are estimated iteratively. Specifically, given the single-index coefficients $\hat{\boldsymbol{\beta}}$, the P-spline estimates, $\hat{\boldsymbol{\gamma}}_{\alpha,\lambda}$, can be obtained straightforwardly as $\boldsymbol{\gamma}_\alpha \mathbf{B}_\alpha(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{i,t_{j-1}})$ is linear with respect to coefficients $\boldsymbol{\gamma}_\alpha$. Existing tools such as `gam` function in the R package `mgcv` can be readily used for the P-spline estimation. To estimate the single-index coefficient $\boldsymbol{\beta}$, we apply the following linear approximation to the unknown function η, so that

$$\eta(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{i,t_{j-1}}) \approx \eta(\boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{x}_{i,t_{j-1}}) + [\nabla \eta(\boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{x}_{i,t_{j-1}})]^{\mathrm{T}} (\boldsymbol{\beta} - \boldsymbol{\beta}_0). \tag{6}$$

We further approximate $\eta(\boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{x}_{i,t_{j-1}})$ by $\hat{\boldsymbol{\gamma}}_{\alpha,\lambda}^{\mathrm{T}} \mathbf{B}_\alpha(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{i,t_{j-1}})$ using an estimated $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ along with the spline approximation. Consequently, the highly nonlinear problem of maximizing

(4) turns into a linear problem facilitating an efficient algorithm summarized below.

Step 0 Obtain an initial estimate for the single-index coefficient $\hat{\boldsymbol{\beta}}^{(0)}$ using the Cox discrete hazard model. Reparameterize $\hat{\boldsymbol{\beta}}^{(0)}$, such that $\|\hat{\boldsymbol{\beta}}^{(0)}\|_2 = 1$ and $\text{sgn}(\hat{\beta}_1^{(0)}) = 1$.

Step 1 Set $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(0)}$, so that the problem reduces to a univariate smoothing problem. We maximize the penalized-spline log-likelihood objective function (4) with respect to $\boldsymbol{\gamma}_\alpha$, and obtain the P-spline estimates $\hat{\boldsymbol{\gamma}}_{\alpha,\lambda}$.

Step 2 With the linear approximation (6), we update the single-index coefficient $\hat{\boldsymbol{\beta}}$ by maximizing the log-likelihood function (4) with respect to $\boldsymbol{\beta}$. Reparameterize $\hat{\boldsymbol{\beta}}$ such that $\|\hat{\boldsymbol{\beta}}\|_2 = 1$ and $\text{sgn}(\hat{\beta}_1) = 1$.

Step 3 Repeat steps 1 and 2 until all parameter estimates converge.

3.5 Connection to Existing Bankruptcy Prediction Models

Our default-prediction single-index models (DSI) shows great promises especially in calibration performance as demonstrated in the next section. It is the first model passing the Hosmer-Lemeshow test including 2008 financial crisis period on a comprehensive US corporate bankruptcy database, to the best of out knowledge. Furthermore, it is flexible and semiparametric, encompassing the state-of-the-art CHS model for bankruptcy prediction.

Our approach is also partly motivated by the class of discrete transformation models (DTM, Ding et al. (2012)), which is a parametric class of transformation survival models that yield better performance comparing to CHS model. In particular, DTM is derived by applying a class of monotonic transformation functions $G(\cdot)$, inverses of logarithm and Box-Cox transformations, to $-\log[S(t_j|\boldsymbol{x}_{t_j})/S(t_{j-1}|\boldsymbol{x}_{t_{j-1}})]$, where $S(t_j|\boldsymbol{x}_{t_j})$ is the conditional

survival function. That is,

$$G\left(-\log\frac{S(t_j|\boldsymbol{x}_{t_j})}{S(t_{j-1}|\boldsymbol{x}_{t_{j-1}})}\right) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{t_j})G\left(-\log\frac{S_0(t_j)}{S_0(t_{j-1})}\right), \tag{7}$$

where $S_0(t_j)$ is the baseline survival function. The DTM model is motivated by its continuous counterpart (Zeng and Lin, 2006), which has the following transformation relationship,

$$\frac{d}{dt}G(\Lambda(t|\boldsymbol{x}_t)) = \exp\{\beta^T\boldsymbol{x}_t\}\frac{d}{dt}G(\Lambda_0(t)), \tag{8}$$

where $\Lambda(t|\boldsymbol{x}_t) = \int_0^t \lambda_0(s)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(s))ds$ is the cumulative conditional hazard function. It is important to note that (7) and (8) are fundamentally different as (7) essentially applies the transformation function $G(\cdot)$ to the difference of cumulative hazard, while (8) applies $G(\cdot)$ to the cumulative hazard and then takes derivative. The continuous survival model would not be applicable for corporate default prediction due to the use of calendar time, that companies do not share the common origin time, and that the conditional cumulative hazard function $\Lambda(t|\boldsymbol{x}_t)$ would be ill-defined. Similarly, the continuous counterpart of our single-index hazard model (Wang, 2004; Huang and Liu, 2006; Lu et al., 2006) is not suitable for the bankruptcy prediction application.

We can rewrite the DTM model of transformation relationship (7) as

$$\frac{p_{i,t_j}}{1-p_{i,t_j}} = G^*\left(\exp(\alpha_j^* + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{t_j})\right), \tag{9}$$

where $G^*(u) = G^{-1}(\log(1+u))$ and α_j^* is the transformed baseline that only depends on time. Interestingly, we further find that if $G(\cdot)$ in equation (7) is chosen to be the inverse of a family of logarithm transformation considered by Ding et al. (2012) and Zeng and Lin (2006), which is defined as $G_c(u) = (\exp(cu)-1)/c$ for $c > 0$ and $G_c(u) = u$ if $c = 0$, then $G_c^*(u) = (cu+1)^{1/c} - 1$ for $c > 0$ and $G_c^*(u) = \exp(u) - 1$ if $c = 0$, which is the same Box-Cox

transformation considered by Zeng and Lin (2006) and Chen et al. (2002).

Our proposed DSI model (1) can be rewritten as

$$\frac{p_{i,t_j}}{1-p_{i,t_j}} = \exp\left(\alpha_j^* + \eta(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{t_j})\right). \tag{10}$$

Note that both DTM (model (9)) and our proposed DSI (model (10)) encompass the state-of-the-art CHS model for corporate bankruptcy prediction. However, DTM adopts a class of monotonic parametric transformations, Box-Cox transformations, $G^*(\cdot)$, on the exponential term $\exp(\alpha_j^* + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{t_j})$, while DSI applies a flexible nonparametric function $\eta(\cdot)$ to the linear projection or single index $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{t_j}$. Indeed, we discover an interesting V-shaped relationship, which is discussed next.

4 Empirical Results

We report empirical results of our proposed default-prediction single-index hazard model (DSI) on the bankruptcy data described in Section 2. In particular, we compare our DSI with popular benchmark models, namely, the state-of-the-art CHS (Campbell et al., 2008) model in finance and the optimal discrete transformation survival model (DTM) (Ding et al., 2012). We document the estimation results based on the full sample data from 1980 to 2016. We then examine models' out-of-sample prediction performance through an expanding window. A robustness check over various prediction periods is conducted and included in online supplementary materials.

4.1 Estimation Results and Assessment of PD Prediction Accuracy

After fitting the default-prediction single-index hazard model on the bankruptcy database we built, an interesting nonlinear relationship is unveiled from the estimated unknown flexible function $\hat{\eta}$. We visualize a V shape in Figure 2, where the shaded interval is the 95% confidence band. It is clear that the relationship depicted in Figure 2 is nonlinear, indicating a severe violation to the common assumption of linearity adopted by popular models such as CHS (Campbell et al., 2008). This interesting finding may also provide some initial empirical evidence to the so-called "financial frictions", where a firm with an extremely low index value may suffer from a similar risk of bankruptcy as a firm with an extremely high index value due to reasons like a restricted credit supply (Giordani et al., 2014).

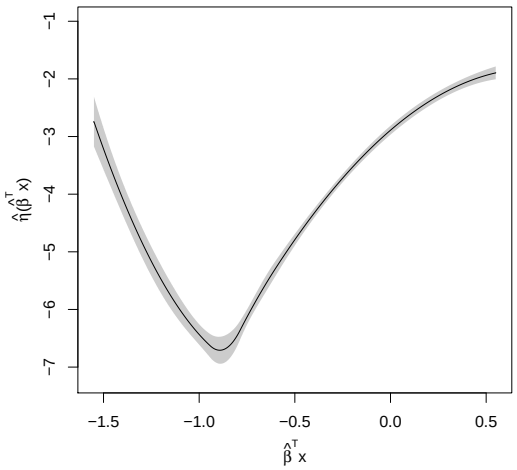

Figure 2: Estimated unknown flexible function $\hat{\eta}(\cdot)$ from the proposed default-prediction single-index hazard model. The range of horizontal axis is the estimated "single-index" based on the full sample period.

In the binary prediction problem, the probabilistic prediction accuracy is commonly assessed in two dimensions: calibration and discrimination. A model with good calibration performance is crucial for PD estimation, especially in economic environments, like the one

the world is currently in, where closely-monitored stress testing and CCA are vitally important. However, in bankruptcy literature, formal calibration testing is rarely performed. Ding et al. (2012) employed the popular Hosmer-Lemeshow (HL) goodness-of-fit test (Hosmer Jr et al., 2013) to evaluate the calibration performance for corporate bankruptcy prediction. Hosmer-Lemeshow (HL) goodness-of-fit test assesses how close it is between the expected and observed event rates. Its null hypothesis can be interpreted as the model has good calibration. Hence a larger p-value would indicate better calibration performance.

Discrimination, on the other hand, evaluates a model's ability to differentiate the observations in different classes, i.e., bankruptcy vs. non-bankruptcy. One of the most commonly-used measures of discrimination is the area under the receiver operating characteristic (ROC) curve, known as AUC, where AUC = 0.5 for a completely random guess and AUC = 1 for a perfect discrimination. A higher AUC value indicates better discrimination.

In addition, decile tables and pseudo-R^2 are commonly-used metrics for assessing model performance in the corporate bankruptcy prediction literature. The decile table summarizes the proportion of observed events in cumulative bins that are constructed by categorizing the sorted predicted probabilities from the largest to smallest. For example, the cumulative bins can be 90-100%, 80-100%,..., 0-100% of predicted PDs, and for each bin, the proportion of the bankruptcies out of total number of bankruptcies is calculated. Therefore, the calculated proportions monotonically increase across the cumulative bins, and it equals to 1 for the bin of 0-100%. A larger number in the top bins is desirable. Pseudo-R^2 is defined as $1-\log L_1/\log L_0$, where L_1 and L_0 are the likelihood from fitted model and null model respectively.

In Panel A of Table 2, we summarize the model estimation results on the full sample period. Importantly, we observe that our proposed default-prediction single-index hazard model is able to pass the Hosmer-Lemeshow goodness-of-fit test with a large p-value of 0.665. On the other hand, the Hosmer-Lemeshow test rejects both benchmark models, implying that their estimated PDs are clearly not well-calibrated. Furthermore, in the cumulative decile

ranking table, DSI consistently shows a higher proportion of bankruptcy firms in the top bins than the other two models. For example, in the top 10% bin, 65.2% bankruptcy firms are captured by DSI, while CHS and DTM capture 63.6% and 64.0%, respectively. The reported pseudo-R^2 and AUC deliver a similar message, showing an enhanced in-sample performance of the DSI model over the CHS and DTM models. Taken together, we conclude that the proposed DSI model has substantially improved the accuracy for corporate bankruptcy prediction, especially in terms of the calibration performance for model fitting.

Table 2: In-sample and out-of-sample prediction assessment based on the p-value of Hosmer-Lemeshow (H-L) goodness-of-fit χ^2-test, AUC, Pseudo-R^2, and the decile ranking table for the proposed default-prediction single-index hazard model (DSI) for corporate bankruptcy prediction; a state-of-the-art bankruptcy prediction model in finance, CHS of Campbell et al. (2008); and an optimal discrete transformation survival model (DTM) of Ding et al. (2012). The in-sample model fitting is based on the full sample period from 1980-2016, and the out-of-sample prediction is based on expanding window predictions.

	Panel A. In-sample (1980-2016)			Panel B. Out-of-sample (2006-2016)		
	DSI	CHS	DTM	DSI	CHS	DTM
p-value of H-L χ^2-test	0.665	0.000	0.000	0.281	0.000	0.000
AUC	0.858	0.844	0.845	0.881	0.823	0.818
Pseudo-R^2	0.196	0.182	0.186	0.275	0.235	0.233
Decile rankings						
90-100%	0.652	0.636	0.640	0.730	0.645	0.652
80-100%	0.795	0.781	0.783	0.833	0.759	0.752
70-100%	0.852	0.840	0.846	0.872	0.794	0.801
60-100%	0.887	0.866	0.868	0.901	0.826	0.823
50-100%	0.910	0.894	0.895	0.926	0.840	0.844
0-100%	1.000	1.000	1.000	1.000	1.000	1.000

We further cross-compare the DSI model with the CHS model by investigating the two tail ends from each model. It is interesting to note that there are 14 bankrupted companies with their predicted PDs ranked in the top 10 percentile from DSI but ranked in the lowest or safest 10 percentile from CHS. However, we find none if the opposite filter is applied, i.e., the lowest 10 percentile from DSI but highest 10 percentile from CHS. This fact again echoes

with the non-linear V-shaped relationship discovered in Figure 2.

4.2 Out-of-Sample Prediction

We now shift our attention to the out-of-sample performance. To evaluate models' out-of-sample performance, we adopt an expanding window approach. Specifically, we start with training data from 1980 to 2005 and make predictions for the following year of 2006. We then expand our training data by one year (i.e. from 1980 to 2006) and make predictions for the year of 2007. We keep expanding this training sample window until the last year in our sampling period is used as the testing data.

We summarize the out-of-sample prediction results in Panel B of Table 2. Consistent with our discussion in the previous section, our proposed default-prediction single-index hazard model tends to outperform the other two benchmark models in all the commonly-adopted metrics in the bankruptcy prediction literature. For example, the DSI passed the Hosmer-Lemeshow calibration test with a p-value of 0.281 whereas neither CHS (Campbell et al., 2008) nor the DTM (Ding et al., 2012) passed the test. Furthermore, in the top 10% bin, we observe nearly 10% more (73.0% vs. 64.5% vs. 65.2%) bankruptcy firms captured correctly by DSI, comparing to that of CHS model and DTM model. Consistent improvements in pseudo-R^2 and AUC demonstrate the advantage of our proposed DSI in providing better prediction performance. A robustness check across different periods are conducted, and the results are summarized in Table S2 of supplementary.

In summary, from our limited empirical analysis, an interesting non-linear, non-monotonic, V-shaped relationship is unveiled by our proposed DSI hazard model, while current most popular models in finance, based on the linearity assumption, would not have the ability to capture. Most importantly, to the best of knowledge, DSI is perhaps the only model that can pass the Hosmer-Lemeshow goodness-of-fit test and provide substantially better-calibrated PD estimates both in-sample and out-of-sample. This may have potentially important im-

plications in practice. As shown in Figure 1, the DSI, CHS, and DTM models yield similar predicted PDs for the majority of the lower percentiles in the CHS predicted PDs. However, the three predicted PD curves are noticeably different as the predicted PDs get larger. In the top one percentile of CHS predicted PDs, instead of lining up along the 45-degree CHS line, the predicted PDs from the DSI model fall considerably under the CHS prediction curve with the DTM predictions in between. For the CHS predicted PDs between 85 to 99 percentiles in the shaded window, the order is completely reversed where the predicted PDs from the DSI model generally falls slightly above the 45-degree CHS line. These interesting findings may suggest that the cash reserves based on majority small predicted PDs from CHS are similar as those from the default-prediction single-index hazard model and the transformation survival model. But for higher predicted PDs, the cash reserves based on the CHS model may be possibly optimistic or overly conservative, especially for exceptionally high PDs.

4.3 Economic Value

A direct application of a PD prediction model is that it can be used as a credit scoring model for pricing during the lending practice. Specifically, in the loan market, different banks may adopt different credit scoring models to assess loans to individual firms, after which the banks make decisions of either reject or lend money with certain price. If not rejected, the borrowers then choose the lender who offers the lowest price (credit spread), which is mainly determined by the predicted value of borrower's default probability. Here we adopt the following expression of credit spread derived by Blöchlinger and Leippold (2006).

$$R = \frac{p_{i,t_j}}{1 - p_{i,t_j}} \text{LGD} + k, \tag{11}$$

where LGD is loss in loan value given default, which is often prespecified; k is also a prespecified value that is the credit spread for the highest quality loan.

Following a similar setup of Agarwal and Taffler (2008), we assume a 100 billion (USD) loan market and three lenders with each one using a different PD prediction model, namely DSI, CHS, and DTM. We refit all three models by excluding financial firms as potential lenders and use the expanding window approach to predict PD for each year from 2006 to 2016. All three lenders reject the loans in the highest 5% PD category. If at least two lenders offer the same price to a loan according to (11), the borrower will randomly choose one.

Table 3 shows the empirical results of economic value for each lender under the described lending practice. The results are averaged over years, where the left panel aggregates the entire prediction period (2006-2016) which includes 2008 financial crisis, while the aggregation period of the right panel (2011-2016) excludes the financial crisis. It is clear to see that for both scenarios, the lender who adopts the proposed default-prediction single-index model receives the highest profit and largest market share. The profits based on DSI and DTM are substantially larger than CHS, providing another possible evidence that the linearity assumption in CHS may be inadequate to characterize the actual relationship. The average credit spread is 44 base points for DSI in both panels, slightly higher than DTM but much lower than CHS. Credit spread relies on the predicted PD, and a higher credit spread leads to a higher revenue but lower market share. Thus, the predicted PD by DSI well balances such a tradeoff, leading to the highest profit. The results shown in Table 3 provide another empirical evidence that the proposed DSI model is more preferred than the other two benchmarking models in a lending practice. Additionally, we notice that the share of defaulters for DSI is higher than the other two. This does not contradict to the decile rankings in Table 2 as the latter is evaluated for each model independently while the share of defaulters is based on a shared market where each company is obligated only to one lender.

Table 3: Comparison of economic value between different PD prediction models. Left panel shows the results averaged over 2006 to 2016 while the right panel is aggregated from 2011 to 2016. Market share is the proportion of borrowers obligated to a lender out of all firms that are not rejected by all three lenders. Share of defaulters is the percentage of defaulted firms to which the loan is granted out of the total number of default. Average credit spread is the credit spread averaged over all firms obligated to a lender. Revenue = Market size ($100B) × Market share × Average credit spread, Loss = Market size ($100B) × Share of defaulters × Prior probability of default × LGD, where the prior probability of default is the sample failure rate in the same prediction year, and Profit = Revenue − Loss.

	2006-2016			2011-2016		
	DSI	CHS	DTM	DSI	CHS	DTM
Market share (%)	48.25	6.58	45.17	49.26	5.08	45.67
Share of defaulters (%)	23.22	13.29	11.39	22.67	16.19	7.14
Average credit spread (%)	0.44	0.78	0.37	0.44	0.75	0.37
Revenue ($mm)	212.72	52.47	167.50	217.03	38.04	169.66
Loss ($mm)	29.76	1.86	22.26	27.79	2.03	8.78
Profit ($mm)	182.96	50.61	145.23	189.24	36.01	160.88

5 Simulation Study

We further conduct a simulation study, mimicking the real bankruptcy process. We simulate bankruptcy data largely based on the distribution of the real bankruptcy data. Specifically, at each time point $j = 1, ..., 36$, we generate N_j firms, where $N_j \sim \text{Poisson}(N_j^*)$, and N_j^* is a proportion of the number of new firms entered at time j in the real data.[5] For each simulated firm that enters at time j, the covariates (bankruptcy predictors) are simulated from multivariate normal distribution, using the same mean and variance-covariance matrix as the real bankruptcy data. The probability of default of company i is calculated as $1/(1 + \exp(-\hat{\alpha}_j - \eta(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{ij})))$, where $\eta(u) = 5.5u + 1.3u^2 - 1.8u^3$ is adopted to mimic the V shape shown in Figure 2. The time-varying baseline parameters are obtained from $\hat{\alpha}_j = \log(\frac{\hat{p}_j}{1-\hat{p}_j})$, where $\hat{p}_j$ is the overall default rate at time j observed from our real bankruptcy data. The binary bankruptcy indicator is simulated based on the Bernoulli distribution with calculated probability of default in previous step. Firms entered at time j stays until default happens.

[5]For the results presented, we use a proportion (0.1) to lower the sample size to reduce the computational cost of simulation. Qualitatively similar results are observed otherwise.

We generate 500 bankruptcy datasets and apply the three modeling approaches, the default-prediction single-index hazard model (DSI), the state-of-the-art discrete linear hazard model CHS, and the "optimal" discrete transformation survival model (DTM), on each simulated data. Table 4 compares the mean squared error, bias and standard error of coefficient estimates across the three models based on the 500 simulated datasets. It is clear that the DSI model provides the least MSE and bias in coefficient estimates. Table 5 further shows several metrics in models' performance assessment. Specifically, based on the 500 replicates, we calculate (i) the mean of absolute deviance of the estimated PD, i.e., $|\hat{\mathbf{p}} - \mathbf{p}_0|$, where $\mathbf{p}_0$ is the true PD and $\hat{\mathbf{p}}$ is the estimated PD; (ii) the mean of AUC; (iii) the mean of pseudo-R^2; and (iv) the Hosmer-Lemeshow calibration test rejection rate. The advantage of adopting the proposed DSI model is noticeable. Specifically, the PD estimates from DSI model provide the minimum deviations from the true PDs. While neither CHS nor DTM model passes the Hosmer-Lemeshow test among any of the 500 simulation runs, only about 4.6% of the simulation runs reject our proposed DSI model, suggesting a strong calibration power of the DSI model in this simulation setting. In addition, we note that DSI model consistently provides the highest AUC values and pseudo-R^2 among the three models. Overall, through this limited simulation study mimicking the real bankruptcy process, we show that the proposed DSI model for corporate bankruptcy is able to provide superior calibration and discrimination performance in predicting the probability of default.

6 Asset Pricing

An important application of predicted PDs is to be utilized as the default risk for constructing investment portfolios. As the conjecture in Fama and French (1996), investors may expect a positive association between the expected return and default risk when holding stocks. Such a positive relationship has been confirmed by a number of studies (Vassalou and Xing, 2004;

Table 4: Mean squared error (MSE), bias, and standard error (S.E.) of coefficient estimates based on the simulated data, mimicking the real bankruptcy data, for the proposed default-prediction single-index hazard model (DSI); a state-of-the-art bankruptcy prediction model in finance, CHS of Campbell et al. (2008); and an optimal discrete transformation survival model (DTM) of Ding et al. (2012).

	DSI			CHS			DTM		
	MSE	Bias	S.E.	MSE	Bias	S.E.	MSE	Bias	S.E.
LTMTA	0.000	0.006	0.021	0.160	-0.399	0.031	0.197	-0.443	0.017
NIMTA	0.000	-0.010	0.020	0.094	0.296	0.082	0.253	0.469	0.182
CASHMTA	0.000	-0.012	0.018	0.076	0.256	0.103	0.325	0.496	0.281
MBE	0.006	0.077	0.007	0.088	0.296	0.030	0.098	0.094	0.298
RSIZE	0.022	-0.149	0.004	0.369	-0.604	0.068	0.361	-0.236	0.553
EXRET	0.000	-0.008	0.010	0.002	0.037	0.020	0.015	0.098	0.072
SIGMA	0.001	-0.020	0.025	0.068	-0.259	0.019	0.049	-0.167	0.146
PRICE	0.031	0.176	0.007	0.629	0.786	0.105	0.522	0.408	0.597

Table 5: Assessment of model fitting and prediction accuracy based on the simulated data, mimicking the real bankruptcy process, for the proposed default-prediction single-index hazard model (DSI); a state-of-the-art bankruptcy prediction model in finance, CHS of Campbell et al. (2008); and an optimal discrete transformation survival model (DTM) of Ding et al. (2012).

	DSI		CHS		DTM	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
$\lvert\hat{\mathbf{p}} - \mathbf{p}_0\rvert$	0.0030	0.0001	0.0136	0.0007	0.0145	0.0006
AUC	0.9511	0.0064	0.7864	0.0172	0.7878	0.0167
Psuedo-R^2	0.5422	0.0170	0.2202	0.0240	0.1824	0.0199
H-L test rejection rate	0.0460		1.0000		1.0000	

Chava and Jarrow, 2004; Aretz et al., 2018). In contrast, some empirical studies document a negative relationship that holding stocks with high default probabilities harvests anomalous low returns (Dichev, 1998; Griffin and Lemmon, 2002; Campbell et al., 2008; Da and Gao, 2010; George and Hwang, 2010; Gao et al., 2018).

This puzzling default risk anomaly has attracted vital attention due to its challenges in both practice and theory. We revisit the puzzle by using the predicted PDs from our proposed DSI model in the extended sample period including 2008 financial crisis. In particular, we obtain firms' one-year-ahead predicted PDs as described in Section 4 using the expanding

window approach starting from 1980-1982, and construct 10 decile-portfolios by sorting the predicted PDs. These 10 portfolios are updated each year accordingly. We also construct three long-short (buy-sell) portfolios, LS9010, LS9505, and LS9901, to investigate the risk-return anomaly. That is, we suppose investor to hold the stocks associated with highest 10 (5, or 1) percent default risk in long position and those with lowest 10 (5, or 1) percent in short position for a number of years (1983 to 2016 for our study). If the positive relation between risk and return holds, such long-short portfolios would be expected to gain positive excess returns, which are the difference between stock and the S&P 500 index return.

Table 6 reports the asset pricing results. In Panel A, we show that the long-short portfolios LS9010, LS9505, and LS9901 yield monthly average excess returns of -0.17%, -0.27%, and -1.39% respectively. The most extreme portfolio has significant excess returns with t-statistics of -2.12. This observation implies a weak anomaly, i.e., the distressed stocks with high predicted PDs from our DSI model yield negative returns.

Table 6: Returns on Bankruptcy Risk-Sorted Portfolios

This table reports the average value-weighted excess returns and its regression of two models: Fama-French three-factor model (Market, SMB, HML), and Fama-French five-factor model (Market, SMB, HML, RMW, CMA). These factors are market factor (Market), size factor (SMB), value factor (HML), profitability factor (RMW), and investment factor (CMA). We sort all stocks based on the one-year ahead expanding window prediction of bankruptcy from our default-prediction single-index model and divide the stocks into 10 portfolios based on deciles. For example, 0 to 10th percentile is denoted as "0010" and 90th to 100th percentile is "9000". The long-short portfolios LS9010 or LS9505 go long with the 10% or 5% riskiest stocks and short the 10% or 5% safest stocks. The results for mean excess returns, alphas of two models are reported in Panel A. In Panel B, we show the Fama-French three-factor regression coefficients. We also report the corresponding values of t-statistics in parentheses.

Portfolios Risk	0010 Low	1020	2030	3040	4050	5060	6070	7080	8090	9000 High	LS9010	LS9505	LS9901
Panel A. Portfolio Excess Return & Alphas of Factor Models													
Mean excess return	0.71	0.75	0.72	0.78	0.63	0.61	0.78	0.70	0.86	0.54	−0.17	−0.27	−1.39
	(2.91)	(3.32)	(3.19)	(3.46)	(2.8)	(2.73)	(3.21)	(2.69)	(2.43)	(1.22)	(−0.50)	(−0.65)	(−2.12)
3-factor alpha	0.13	0.17	0.14	0.15	−0.03	−0.04	0.01	−0.15	−0.31	−0.65	−0.78	−0.86	−1.77
	(1.59)	(2.42)	(1.49)	(1.99)	(−0.29)	(−0.47)	(0.09)	(−1.17)	(−2.07)	(−2.65)	(−3.03)	(−2.72)	(−2.86)
5-factor alpha	0.12	0.10	0.03	0.15	−0.04	−0.08	−0.12	−0.17	−0.08	−0.31	−0.43	−0.49	−1.15
	(1.24)	(1.35)	(0.28)	(1.79)	(−0.37)	(−0.8)	(−1.21)	(−1.38)	(−0.49)	(−1.14)	(−1.55)	(−1.46)	(−1.63)
Panel B. Three-Factor Regression Coefficients													
Market	0.98	0.94	0.92	0.95	0.97	0.94	1.06	1.11	1.43	1.50	0.53	0.52	0.35
	(41.73)	(46.1)	(31.71)	(51.99)	(26.6)	(37.43)	(25.08)	(24.46)	(15.20)	(17.58)	(6.13)	(4.30)	(2.23)
SMB	−0.03	−0.07	0.01	0.03	−0.01	−0.03	−0.08	−0.11	−0.01	0.65	0.68	0.89	1.15
	(−0.83)	(−2.66)	(0.17)	(0.84)	(−0.21)	(−0.48)	(−0.95)	(−1.68)	(−0.13)	(6.13)	(6.80)	(6.41)	(4.43)
HML	−0.27	−0.16	−0.12	−0.02	0.08	0.11	0.27	0.46	0.86	0.63	0.90	0.76	0.36
	(−6.88)	(−4.8)	(−1.97)	(−0.51)	(2.05)	(2.40)	(3.93)	(6.49)	(6.03)	(4.44)	(6.34)	(3.24)	(0.87)
Panel C. Portfolio Characteristics													
RSIZE	−9.74	−9.78	−9.83	−9.90	−10.04	−10.17	−10.26	−10.42	−11.02	−11.69			
MBE	2.15	2.15	2.13	2.12	2.09	2.00	1.92	1.84	1.88	2.60			

In addition to the mean excess return, we also attempt to offer some insights to an important theoretical asset-pricing question: "How can factor models explain the stock return?" Specifically, we regress each portfolio's monthly excess return on the factors and report the estimated intercepts (alphas). If a factor model can explain stocks' excess return, the estimated alpha is expected to be zero. We apply both Fama-French three-factor model (Fama and French, 1996) and the five-factor model (Fama and French, 2015), which are among the most agreed factor models. The three-factor model includes the market factor (Market), the size factor (Small Minus Big, or SMB), and the value factor (High Minus Low, or HML), while the five-factor model has two additional factors: profitability factor (Robust Minus Weak, or RMW), and investment factor (Conservative Minus Aggressive, or CMA).

As shown in Panel A, the three-factor model cannot fully explain the abnormal negative excess return, as the estimated alphas are significant for the long-short portfolios LS9010 (-0.78%), LS9505 (-0.86%), and LS9901 (-1.77%). However, the five-factor model can explain the excess return with insignificant alphas. These findings imply that the distress anomaly as evidenced by negative alphas may still exist but has clearly weakened, especially under the five-factor model, in the extended sample period including 2008 financial crisis comparing to the earlier sample periods considered in Campbell et al. (2008) or Ding et al. (2012). Panel B shows the factor loadings (estimated coefficients) for the three-factor models. The values are qualitatively consistent with those in the existing literature (Campbell et al., 2008; Ding et al., 2012; Hou et al., 2015).

We also report the portfolio characteristics in Panel C. The average relative size (RSIZE) of portfolios implies that the size decreases along with the increase of default risk. On the other hand, the firms in the two tails of predicted PDs are associated with a high market-to-book equity ratio (MBE). These two pieces of evidence are consistent with Campbell et al. (2008) and Ding et al. (2012). In addition, one referee made an interesting suggestion that constructing a new distress factor using our calibrated default probability may also have

important asset pricing implications. This is an interesting question to explore further in the future.[6]

In summary, our findings based on the period of 1983 to 2016 imply that the default risk anomaly has weakened or even disappeared using the predicted PDs from the proposed DSI model. The stocks with higher default risk no longer earn strong anomalously lower excess returns than those firms with lower default risk after adjusting for the risk factors.

7 Conclusion

In this paper, we develop a flexible default-prediction single-index hazard model (DSI) for corporate PD prediction, and asymptotic properties have been established for a penalized-spline likelihood estimation. Applying the proposed DSI model to a comprehensive corporate bankruptcy database we build, we have a number of interesting findings. First, we discover a V-shaped relationship of the systematic component with the "financial default" single index. This is in stark contrast to the popular linear assumption in the state-of-the-art bankruptcy prediction model CHS (Campbell et al., 2008), indeed the Cox discrete hazard model (Cox, 1972). The uncovered V shape suggests that the linearity assumption may be severely violated. Second and most importantly, the proposed DSI model passes the Hosmer-Lemeshow goodness-of-fit calibration tests while neither does CHS nor an optimal transformation survival model (DTM). In our empirical study, we observe that majority small predicted PDs from the three models are close to each other. However, for higher predicted PDs, the three models yield noticeably different predictions. These findings may have important implications in practice. For example, the cash or capital reserve calculations

[6]We have attempted to build a new factor called distress-minus-healthy (DMH). Preliminary results have shown that this factor can explain the same distress anomaly we investigated here. Furthermore, we re-examined a profitability anomaly that is associated with the gross profits-to-assets (GP/A) ratio (Novy-Marx, 2013). Our preliminary results show that the new DMH factor helps but still cannot fully explain the profitability anomalies. A more in-depth study is needed to explore the new distress factor and potential implications.

based on the CHS predicted PDs may be optimistic or conservative, especially for extremely high PDs. Third, we examine the important asset pricing implication based on the predicted PDs from the proposed model. We find that the negative distress anomaly, that the higher is the distress risk the lower excess return even after controlling important factors, has weakened and even disappeared during the extended period including 2008 financial crisis.

Probability of default prediction has many applications in a variety of fields beyond corporate bankruptcy, such as credit card, commercial and residential loans, corporate bonds, mortgage borrowing, and foreclosure process. It is our hope that the developed flexible default-prediction single-index hazard model may be potentially adopted in these fields to deliver accurate prediction and high impact in practice.

References

Agarwal, V. and Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8):1541–1551.

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

Aretz, K., Florackis, C., and Kostakis, A. (2018). Do stock returns really decrease with default risk? new international evidence. *Management Science*, 64(8):3821–3842.

Blöchlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, 30(3):851–873.

Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, 63(6):2899–2939.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489.

Chava, S. and Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, 8(4):537–569.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Da, Z. and Gao, P. (2010). Clientele change, liquidity shock, and the return on financially distressed stocks. *Journal of Financial and Quantitative Analysis*, 45(1):27–48.

Dichev, I. D. (1998). Is the risk of bankruptcy a systematic risk? *The Journal of Finance*, 53(3):1131–1147.

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

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, 51(1):55–84.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Gao, P., Parsons, C. A., and Shen, J. (2018). Global relation between financial distress and equity returns. *The Review of Financial Studies*, 31(1):239–277.

George, T. J. and Hwang, C.-Y. (2010). A resolution of the distress risk and leverage puzzles in the cross section of stock returns. *Journal of Financial Economics*, 96(1):56–79.

Giordani, P., Jacobson, T., Von Schedvin, E., and Villani, M. (2014). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, 49(4):1071–1099.

Griffin, J. M. and Lemmon, M. L. (2002). Book-to-market equity, distress risk, and stock returns. *The Journal of Finance*, 57(5):2317–2336.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

Huang, J. Z. and Liu, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, 62(3):793–802.

Huang, J. Z. and Su, Y. (2021). Asymptotic properties of penalized spline estimators in concave extended linear models: Rates of convergence. *The Annals of Statistics*, 49(6):3383–3407.

Liang, H., Liu, X., Li, R., and Tsai, C.-L. (2010). Estimation and testing for partially linear single-index models. *The Annals of Statistics*, 38(6):3811–3836.

Lu, X., Chen, G., Singh, R. S., and K. Song, P. X. (2006). A class of partially linear single-index survival models. *Canadian Journal of Statistics*, 34(1):97–112.

Nielsen, J. P., Linton, O., Bickel, P. J., et al. (1998). On a semiparametric survival model with flexible covariate effect. *The Annals of Statistics*, 26(1):215–241.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108(1):1–28.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Scigliuzzo, D., Saul, J., Harrington, S. D., and Pogkas, D. (2020). The covid bankruptcies: California pizza to prom-dress retailer. https://www.bloomberg.com/graphics/2020-us-bankruptcies-coronavirus/.

Shen, L. (2020). The 20 biggest companies that have filed for bankruptcy because of the coronavirus pandemic. https://fortune.com/2020/06/29/companies-filing-bankruptcy-2020-during-coronavirus-pandemic-covid-19-economy-industries/.

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

Tian, S., Yu, Y., and Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100.

Vassalou, M. and Xing, Y. (2004). Default risk in equity returns. *The Journal of Finance*, 59(2):831–868.

Wang, W. (2004). Proportional hazards regression models with unknown link function and time-dependent covariates. *Statistica Sinica*, pages 885–905.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054.

Yu, Y., Wu, C., and Zhang, Y. (2017). Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing*, 27(2):571–582.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.

Supplemental Materials for "Corporate Probability of Default: A Single-Index Hazard Model Approach"

A Additional Tables and Empirical Results

A.1 Bankruptcy Frequency Table

In Table S1, we summarize the bankruptcy events and active firms in our database. The second column shows the number of bankruptcies reported, and the third column is the total number of active firms each year. The last column summarizes the corresponding bankruptcy percentage. We observe more bankruptcy events following the recession in early 1990s, the dot-com bubble burst in 2001, and the 2008 financial crisis.

A.2 Robustness Check with Different Prediction Periods

In order to check the robustness of the proposed default-prediction single-index hazard model (DSI), we assess the prediction accuracy based on various prediction periods. Table S2 shows the p-values of Hosmer-Lemeshow test, the AUC, and pseudo-R^2. We find that our proposed DSI passes the Hosmer-Lemeshow test over all prediction periods while delivering the highest AUC values and pseudo-R^2. It is clear that the proposed DSI consistently outperforms the other two benchmark models. The improvement is even stronger when the testing periods include 2008 financial crisis. Based on our limited empirical study, we find that the proposed default-prediction single-index hazard model can offer robust prediction accuracy even when 2008 financial crisis is included.

Table S1: Count of bankruptcy firms and total number of firms over year.

Year	Bankruptcy firms	Total firms	Proportion (%)
1980	16	2477	0.565
1981	17	2588	0.580
1982	24	4209	0.546
1983	27	4757	0.547
1984	46	5188	0.867
1985	55	5209	0.979
1986	76	5312	1.318
1987	40	5508	0.635
1988	55	5537	0.795
1989	64	5296	1.190
1990	64	5227	1.358
1991	113	5152	2.038
1992	62	5343	1.048
1993	37	5523	0.598
1994	42	6694	0.553
1995	46	6898	0.565
1996	39	7260	0.468
1997	61	7477	0.669
1998	85	7459	0.965
1999	56	7010	0.685
2000	61	6824	0.806
2001	68	6304	0.984
2002	62	5621	1.050
2003	40	5175	0.696
2004	22	4946	0.344
2005	18	4877	0.328
2006	10	4822	0.166
2007	13	4773	0.210
2008	40	4608	0.825
2009	88	4284	1.914
2010	33	4099	0.781
2011	18	3940	0.406
2012	20	3785	0.476
2013	17	3670	0.409
2014	15	3721	0.403
2015	21	3774	0.450
2016	18	3690	0.461

Table S2: Robustness check on out-of-sample prediction performance over different periods based on the p-value of Hosmer-Lemeshow (H-L) goodness-of-fit χ^2-test, AUC, and Pseudo-R^2 for the proposed default-prediction single-index hazard model (DSI) for corporate bankruptcy prediction; a state-of-the-art bankruptcy prediction model in finance, CHS of Campbell et al. (2008); and an optimal discrete transformation survival model (DTM) of Ding et al. (2012).

	p-value of H-L χ^2-test			AUC			Pseudo-R^2		
Period	DSI	CHS	DTM	DSI	CHS	DTM	DSI	CHS	DTM
2006-2016	0.281	0.000	0.000	0.881	0.823	0.818	0.275	0.235	0.233
2007-2016	0.139	0.000	0.000	0.875	0.814	0.809	0.268	0.227	0.223
2008-2016	0.094	0.000	0.000	0.873	0.811	0.805	0.268	0.226	0.222
2009-2016	0.120	0.000	0.000	0.876	0.830	0.828	0.294	0.266	0.264

B Derivation of Likelihood Function (3)

For bankruptcy prediction, discrete-time hazard model framework is most popularly adopted (Shumway, 2001; Campbell et al., 2008; Tian et al., 2015). An appealing feature of the discrete-time hazard model is the memoryless property, which implies that a company's default probability at time t is *conditionally* independent from that of time $t-1$.

In particular, the companies in the data consist of two types: (i) companies that have bankrupted at or before t_J ($\delta_i = 1$) , and (ii) companies that have been censored at or before t_J ($\delta_i = 0$), where δ_i is the censoring indicator. For (i), the probability that firm i experiences

bankruptcy at time D_i can be expressed as

$$\begin{aligned}
\Pr(T = D_i) &= \Pr(T = D_i | T \geq D_i) \Pr(T \neq D_i - 1 | T \geq D_i - 1) \Pr(T \neq D_i - 2 | T \geq D_i - 2) ... \\
&\quad ... \Pr(T \neq A_i + 1 | T \geq A_i + 1) \Pr(T \neq A_i | T \geq A_i) \\
&= h(D_i | \boldsymbol{x}_{i,D_i-1})[1 - h(D_i - 1 | \boldsymbol{x}_{i,D_i-2})][1 - h(D_i - 2 | \boldsymbol{x}_{i,D_i-3})] ... \\
&\quad ...[1 - h(A_i + 1 | \boldsymbol{x}_{i,A_i})][1 - h(A_i | \boldsymbol{x}_{i,A_i-1})] \\
&= p_{i,D_i}(1 - p_{i,D_i-1})(1 - p_{i,D_i-2}) ... (1 - p_{i,A_i+1})(1 - p_{i,A_i}) \\
&= p_{i,D_i} \prod_{A_i \leq t < D_i} (1 - p_{i,t}).
\end{aligned}$$

Note that the i.i.d. assumption is *not* imposed. Similarly, for (ii), the censored companies, the bankruptcy may happen after D_i. So the corresponding probability is

$$\begin{aligned}
\Pr(T > D_i) &= \Pr(T \neq D_i | T \geq D_i) \Pr(T \neq D_i - 1 | T \geq D_i - 1) \Pr(T \neq D_i - 2 | T \geq D_i - 2) ... \\
&\quad ... \Pr(T \neq A_i + 1 | T \geq A_i + 1) \Pr(T \neq A_i | T \geq A_i) \\
&= [1 - h(D_i | \boldsymbol{x}_{i,D_i-1})][1 - h(D_i - 1 | \boldsymbol{x}_{i,D_i-2})][1 - h(D_i - 2 | \boldsymbol{x}_{i,D_i-3})] ... \\
&\quad ...[1 - h(A_i + 1 | \boldsymbol{x}_{i,A_i})][1 - h(A_i | \boldsymbol{x}_{i,A_i-1})] \\
&= (1 - p_{i,D_i})(1 - p_{i,D_i-1})(1 - p_{i,D_i-2}) ... (1 - p_{i,A_i+1})(1 - p_{i,A_i}) \\
&= \prod_{j \in \{j : A_i \leq t_j \leq D_i\}} (1 - p_{i,t_j}).
\end{aligned}$$

Let us denote $T_i = \{j : A_i \leq t_j \leq D_i\}$ and $T_i^{(-1)} = \{j : A_i \leq t_j < D_i\}$. Then the likelihood

for all companies $(i = 1, ..., n)$ is

$$\begin{aligned} L &= \prod_{i=1}^{n} [\Pr(T = D_i)]^{\delta_i} \times [\Pr(T > D_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[p_{i,D_i} \prod_{j \in T_i^{(-1)}} (1 - p_{i,t_j}) \right]^{\delta_i} \left[\prod_{j \in T_i} (1 - p_{i,t_j}) \right]^{1-\delta_i}, \end{aligned}$$

where δ_i is the censoring indicator defined at the beginning.

By taking log on both sides, we obtain the log-likelihood function

$$l = \log L = \sum_{i=1}^{n} \left[\delta_i \log p_{i,D_i} + \delta_i \sum_{j \in T_i^{(-1)}} \log(1 - p_{i,t_j}) + (1 - \delta_i) \sum_{j \in T_i} \log(1 - p_{i,t_j}) \right].$$

It can be simplified by following algebra

$$\begin{aligned} l &= \sum_{i=1}^{n} \left[\delta_i \log p_{i,D_i} + \delta_i \sum_{j \in T_i^{(-1)}} \log(1 - p_{i,t_j}) + (1 - \delta_i) \sum_{j \in T_i} \log(1 - p_{i,t_j}) \right] \\ &= \sum_{i=1}^{n} \left[\delta_i \log p_{i,D_i} - \delta_i \log(1 - p_{i,D_i}) + (\delta_i) \sum_{j \in T_i} \log(1 - p_{i,t_j}) + (1 - \delta_i) \sum_{j \in T_i} \log(1 - p_{i,t_j}) \right] \\ &= \sum_{i=1}^{n} \left[\delta_i \log \left(\frac{p_{i,D_i}}{1 - p_{i,D_i}} \right) + \sum_{j \in T_i} \log(1 - p_{i,t_j}) \right]. \end{aligned}$$

Now in the square bracket of last line, the first term can be rewritten as

$$\delta_i \log \left(\frac{p_{i,D_i}}{1 - p_{i,D_i}} \right) = \sum_{j \in T_i} \delta_{i,j} \log \left(\frac{p_{i,t_j}}{1 - p_{i,t_j}} \right),$$

where $\delta_{i,j} = \delta_i I\{D_i = t_j\}$. The above equation holds because if $\delta_i = 0$, then $\delta_{i,j} = 0$ for all $j \in T_i$; if $\delta_i = 1$, then $\delta_{i,j} = 0$ for all $j \in T_i^{(-1)}$ but $\delta_{i,j} = 1$ for $D_i = t_j$.

Finally, the log-likelihood has the following form

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \left[\sum_{j \in T_i} \delta_{i,j} \log \left(\frac{p_{i,t_j}}{1 - p_{i,t_j}} \right) + \sum_{j \in T_i} \log(1 - p_{i,t_j}) \right] \\
&= \sum_{i=1}^{n} \sum_{j \in T_i} \left[\delta_{i,j} \log \left(\frac{p_{i,t_j}}{1 - p_{i,t_j}} \right) + \log(1 - p_{i,t_j}) \right] \\
&= \sum_{i=1}^{n} \sum_{j \in T_i} \left[\delta_{i,j} \log p_{i,t_j} + (1 - \delta_{i,j}) \log(1 - p_{i,t_j}) \right],
\end{aligned}
$$

which takes an equivalent form as the log-likelihood of the classical logistic regression.

C Technical Details and Proof of Theorems

C.1 Identifiability and Reparametrization

For model identifiability, the single-index parameter is constrained such that $||\boldsymbol{\beta}|| = 1$ and the first element $\beta_1 > 0$. We reparameterize $\boldsymbol{\beta}$ to handle this constraint. Let column vector $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_{d-1})^{\mathrm{T}}$ and define $\boldsymbol{\beta}^{\mathrm{T}} = (1, \boldsymbol{\zeta}^{\mathrm{T}})/\sqrt{1 + ||\boldsymbol{\zeta}||^2}$. Now the reparametrized parameter $\boldsymbol{\zeta}$ is unconstrained. The Jacobian matrix of $\partial\boldsymbol{\beta}/\partial\boldsymbol{\zeta}$ is

$$
\mathbf{J}(\boldsymbol{\zeta}) = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\zeta}} = -\left(1 + \|\boldsymbol{\zeta}\|^2\right)^{-\frac{3}{2}} \begin{bmatrix}
\zeta_1 & \zeta_2 & \cdots & \zeta_{d-1} \\
-(1 + \|\boldsymbol{\zeta}\|^2) + \zeta_1^2 & \zeta_2 \zeta_1 & \cdots & \zeta_{d-1} \zeta_1 \\
\zeta_1 \zeta_2 & -(1 + \|\boldsymbol{\zeta}\|^2) + \zeta_2^2 & \cdots & \zeta_{d-1} \zeta_2 \\
\vdots & \vdots & \ddots & \vdots \\
\zeta_1 \zeta_{d-1} & \zeta_2 \zeta_{d-1} & \cdots & -(1 + \|\boldsymbol{\zeta}\|^2) + \zeta_{d-1}^2
\end{bmatrix}.
$$

C.2 Proof of Theorems 1 and 2

We provide proofs of Theorems 1 and 2. Lemmas and their proofs are presented in the end of this section.

First, we introduce some additional notations used in the proof. Denote the true parameters by $\mathbf{g}_0 = (\eta_0, \boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0J})^{\mathrm{T}}, \boldsymbol{\beta}_0)$. Denote $\ell_{ij}(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ the *negative* log-likelihood for individual i at time t_j, i.e., $\ell_{ij}(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) = -\kappa_{ij}(\delta_{ij}\log p_{ij} + (1-\delta_{ij})\log(1-p_{ij}))$, where $\kappa_{ij} = I\{A_i < t_j \leq D_i\}$. Denote $\|\eta\|^2_{(m)}$ the penalty $\int (\eta^{(m)}(u))^2 du$ for some integer $m \geq 2$, where $\eta^{(m)}$ is the m-th derivative of η. The corresponding inner product is denoted by $\langle \eta_1, \eta_2 \rangle_{(m)} = \int \eta_1^{(m)}(u)\eta_2^{(m)}(u)du$. Denote $E(\cdot)$ the expectation and $E_n(\cdot)$ the sample average. Let C be a generic positive constant whose value can change at different appearances.

C.2.1 Proof of Theorem 1

Theorem 1 follows immediately from Propositions 1 and 2 below, which bound the approximation and estimation error, respectively.

Proposition 1. *Under the assumptions of Theorem 1, there is a local minimizer* $\mathbf{g}_\lambda = (\eta_\lambda, \boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda) = \arg\min_{\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}} E\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda\|\eta\|^2_{(m)}$ *that satisfies*

$$\|\eta_\lambda - \eta_0\|^2 + \lambda\|\eta_\lambda\|^2_{(m)} + \|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\alpha}_\lambda - \boldsymbol{\alpha}_0\|^2 = O(K^{-2p} + \lambda K^{2(m-p)_+}),$$

where the minimization with respect to η *is over* $\eta \in G_n$ *with* G_n *being the space of splines with the given degree and knots sequence,*

Proof of Proposition 1. By Proposition 2.1 of Huang and Su (2021), there exists some $\eta^* \in G_n$ with

$$\|\eta^* - \eta_0\| = O(K^{-p}), \|\eta^*\|_{(m)} = O(K^{(m-p)_+}), \text{ and } \|\eta^* - \eta_0\|_\infty = O(K^{-p+1/2}). \tag{S1}$$

Denote $s^2 = K^{-2p} + \lambda K^{2(m-p)_+}$. We only need to show that for any $\|\eta - \eta^*\|^2 + \lambda\|\eta -$

$\eta^*\|_{(m)}^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2 = Ls^2$, with $L > 0$ sufficiently large,

$$E[\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})] + \lambda\|\eta\|_{(m)}^2 > E[\ell(\eta^*, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)] + \lambda\|\eta^*\|_{(m)}^2, \tag{S2}$$

which implies there exists a local minimizer of $E[\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})] + \lambda\|\eta\|_{(m)}^2$ that satisfies $\|\eta_\lambda - \eta^*\|^2 + \lambda\|\eta_\lambda - \eta^*\|_{(m)}^2 + \|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\alpha}_\lambda - \boldsymbol{\alpha}_0\|^2 = O(s^2)$. Then, combining this with (S1) and using triangle inequality, the results hold.

Now we show (S2). We note that

$$\begin{aligned}
& E[\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})] + \lambda\|\eta\|_{(m)}^2 - E[\ell(\eta_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)] \\
\geq\ & C(\|\eta - \eta_0\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2) + \lambda\|\eta\|_{(m)}^2 \\
\geq\ & C(\|\eta - \eta^*\|^2 - \|\eta_0 - \eta^*\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2) + \frac{\lambda}{2}\|\eta - \eta^*\|_{(m)}^2 - \lambda\|\eta^*\|_{(m)}^2 \\
\geq\ & CLs^2 - Cs^2.
\end{aligned} \tag{S3}$$

The first inequality holds by Lemma 1. This is because $\|\eta - \eta_0\|_\infty \leq \|\eta - \eta^*\|_\infty + \|\eta^* - \eta_0\|_\infty \leq C\sqrt{KL}s = o(1)$. (Note we have the property $\|\eta\|_\infty \leq C\sqrt{K}\|\eta\|, \forall \eta \in G_n$. See proposition 2.2 of Huang and Su (2021).) The second inequality holds by triangle inequality and that $\frac{1}{2}\|\eta_1 - \eta_2\|_{(m)}^2 \leq \|\eta_1\|_{(m)}^2 + \|\eta_2\|_{(m)}^2$ for functions η_1, η_2.

Furthermore, by Lemma 1 again,

$$\begin{aligned}
& E[\ell(\eta^*, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)] + \lambda\|\eta^*\|_{(m)}^2 - E[\ell(\eta_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)] \\
\leq\ & C\|\eta^* - \eta_0\|^2 + \lambda\|\eta^*\|_{(m)}^2 \\
\leq\ & Cs^2.
\end{aligned} \tag{S4}$$

Since L is sufficiently large, combining (S3) and (S4), we establish (S2). □

Proposition 2. *Under the assumptions of Theorem 1, there exists a local minimizer* $\widehat{\mathbf{g}} =$

$(\widehat{\eta}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ *of* $E_n\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda\|\eta\|^2_{(m)}$ *with* $\eta \in G_n$ *such that*

$$\|\widehat{\eta} - \eta_\lambda\|^2 + \lambda\|\widehat{\eta} - \eta_\lambda\|^2_{(m)} + \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_\lambda\|^2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|^2 = O_p\left(\frac{1}{n\lambda^{1/(2m)}} \wedge \frac{K}{n}\right),$$

where $a \wedge b = \min\{a, b\}$.

Proof of Proposition 2. Denote $u^2 = \frac{1}{n\lambda^{1/(2m)}} \wedge \frac{K}{n}$. We only need to show that for all $(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with $\|\eta - \eta_\lambda\|^2 + \lambda\|\eta - \eta_\lambda\|^2_{(m)} + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\lambda\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda\|^2 = Lu^2$ with L sufficiently large, we have with probability approaching one,

$$E_n\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda\|\eta\|^2_{(m)} > E_n\ell(\eta_\lambda, \boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda) + \lambda\|\eta_\lambda\|^2_{(m)}. \tag{S5}$$

Let $h(\xi) = E\ell(\mathbf{g}_\lambda + \xi(\mathbf{g} - \mathbf{g}_\lambda)) + \lambda\|\eta_\lambda + \xi(\eta - \eta_\lambda)\|^2_{(m)}$, where $\mathbf{g} = (\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\mathbf{g}_\lambda = (\eta_\lambda, \boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda)$. Then the difference of the two sides of (S5) is

$$\begin{aligned} & E_n\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda\|\eta\|^2_{(m)} - E_n\ell(\eta_\lambda, \boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda) - \lambda\|\eta_\lambda\|^2_{(m)} \\ = \;& h(1) - h(0) + (E_n - E)\{\ell(\mathbf{g}) - \ell(\mathbf{g}_\lambda)\}. \end{aligned}$$

Since $\xi = 0$ is a local minimizer of $h(\xi)$ (by the definition of $(\eta_\lambda, \boldsymbol{\alpha}_\lambda, \boldsymbol{\beta}_\lambda)$), we have $h'(0) = 0$ and the above is bounded below by

$$h''(\xi^*) + (E_n - E)\{\ell(\mathbf{g}) - \ell(\mathbf{g}_\lambda)\},$$

for some $\xi^* \in [0, 1]$. Using Lemmas 2 and 3, $(E_n - E)\{\ell(\mathbf{g}) - \ell(\mathbf{g}_\lambda)\} = O_p(\sqrt{L}u^2)$ while $h''(\xi^*) \geq CLu^2$, and thus the above is positive with probability approaching one, which established the result. □

C.2.2 Proof of Theorem 2

For asymptotic normality of $\boldsymbol{\beta}$, we need to rely more on the specific form of the likelihood to get an explicit asymptotic covariance matrix.

Since the estimator $(\widehat{\eta}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ locally minimizes the penalized loss, $\xi = 0$ locally minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} m_{ij}\left(\widehat{\alpha}_j + \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) + \xi\{\alpha_j + \eta(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \widehat{\alpha}_j - \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\}\right) + \lambda\|\widehat{\eta} + \xi(\eta - \widehat{\eta})\|_{(m)}^2,$$

for any $\boldsymbol{\alpha}, \boldsymbol{\beta}, \eta \in G_n$.

By the first-order optimality condition, we have

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} m_{ij}'(\widehat{a}_j)\{\widehat{\alpha}_j + \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \alpha_j - \eta(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\} + 2\lambda\langle\widehat{\eta}, \eta - \widehat{\eta}\rangle_{(m)} = 0, \tag{S6}$$

where $\widehat{a}_j = \widehat{\alpha}_j + \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ (note that previously we have also defined $a_{0j} = \alpha_{0j} + \eta_0(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0)$).

Then, we set $\alpha_j = \widehat{\alpha}_j + \widetilde{\alpha}_{0j}\mathbf{1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\eta = \widehat{\eta} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\widetilde{\boldsymbol{\eta}}^*$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in (S6), where $\widetilde{\boldsymbol{\eta}}^*$ is the spline approximation of $\widetilde{\boldsymbol{\eta}}_0$ with $\|\widetilde{\eta}_j^* - \widetilde{\eta}_{0j}\| = O(K^{-p})$, $\|\widetilde{\eta}_j^*\|_{(m)} = O(K^{(m-p)_+})$ (Proposition 2.1 of Huang and Su (2021)). Then (S6) becomes

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} m_{ij}'(\widehat{a}_j)\left\{\widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0) - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\widetilde{\boldsymbol{\eta}}^*(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0) - \widetilde{\alpha}_{0j}\mathbf{1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right\} \\
&+2\lambda\langle\widehat{\eta}, (\widetilde{\boldsymbol{\eta}}^*)^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\rangle_{(m)} = 0.
\end{aligned} \tag{S7}$$

We will now show that the left-hand side in the above is equal to

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\left\{m_{ij}'(a_{0j}) + m_{ij}''(a_{0j})\left(\widehat{\alpha}_j - \alpha_{0j} + (\widehat{\eta} - \eta_0)(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0) + \eta_0'(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}_{ij}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right)\right\} \\
&\quad\cdot\left\{(\eta_0'(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0))^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right\} + o_p\left(\frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|\right).
\end{aligned} \tag{S8}$$

To see this, we note by Taylor's expansion,

$$
\begin{aligned}
& m'_{ij}(\widehat{a}_j) - m'_{ij}(a_{0j}) - m''_{ij}(a_{0j}) \left(\widehat{\alpha}_j - \alpha_{0j} + (\widehat{\eta} - \eta_0)(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) + \eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \\
= \; & m''_{ij}(a_j^*)(\widehat{a}_j - a_{0j}) - m''_{ij}(a_{0j}) \left(\widehat{\alpha}_j - \alpha_{0j} + (\widehat{\eta} - \eta_0)(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) + \eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \\
= \; & (m''_{ij}(a_j^*) - m''_{ij}(a_{0j}))(\widehat{a}_j - a_{0j}) \\
& + m''_{ij}(a_{0j}) \left(\widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}) - \eta_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - (\widehat{\eta} - \eta_0)(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - \eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \\
= \; & (m''_{ij}(a_j^*) - m''_{ij}(a_{0j}))(\widehat{a}_j - a_{0j}) + m''_{ij}(a_{0j})(\widehat{\eta}'(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}^*) - \eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)) \mathbf{x}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),
\end{aligned}
$$

where a_j^* lies between $\widehat{a}_j$ and a_{0j} and $\boldsymbol{\beta}^*$ lies between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. The above implies

$$
\begin{aligned}
& \frac{1}{n} \sum_{i,j} m'_{ij}(\widehat{a}_j) - m'_{ij}(a_{0j}) - m''_{ij}(a_{0j}) \left(\widehat{\alpha}_j - \alpha_{0j} + (\widehat{\eta} - \eta_0)(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) + \eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \\
= \; & o_p(n^{-1/2}),
\end{aligned} \tag{S9}
$$

since $\|\widehat{\eta}' - \eta'_0\| = o_p(n^{-1/4})$. We also have

$$
\begin{aligned}
& \left\{ \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}) - \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}} \widetilde{\boldsymbol{\eta}}^*(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - \widetilde{\alpha}_{0j} \mathbf{1}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} \\
& - \left\{ (\eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} \\
= \; & (\widehat{\eta}'(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}^*) - \eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)) \mathbf{x}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\widetilde{\boldsymbol{\eta}}_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - \widetilde{\boldsymbol{\eta}}^*(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),
\end{aligned}
$$

which means

$$
\begin{aligned}
& \frac{1}{n} \sum_{i,j} \left\{ \left\{ \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}) - \widehat{\eta}(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}} \widetilde{\boldsymbol{\eta}}^*(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - \widetilde{\alpha}_{0j} \mathbf{1}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} \right. \\
& \left. - \left\{ (\eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} \right\} \\
= \; & o_p(n^{-1/4} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|).
\end{aligned} \tag{S10}
$$

Furthermore, the penality term in (S7) is

$$\begin{aligned} & 2\lambda \left| \langle \widehat{\eta}, (\widetilde{\boldsymbol{\eta}}^*)^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rangle_{(m)} \right| \\ \leq \; & 2\sqrt{\lambda \|\widehat{\eta}\|_{(m)}^2} \sqrt{\lambda \sum_{j=1}^{J} \|\widetilde{\eta}_j^*\|_{(m)}^2} \cdot \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\|. \end{aligned}$$

We have $\lambda \|\widehat{\eta}\|_{(m)}^2 = O_p(s^2 + u^2) = o_p(n^{-1/2})$ (u^2 and s^2 as defined in the proof of Propositions 1 and 2) and similarly $\lambda \|\widetilde{\eta}_j^*\|_{(m)}^2 = o_p(n^{1/2})$. Thus we have

$$2\lambda \left| \langle \widehat{\eta}, (\widetilde{\boldsymbol{\eta}}^*)^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rangle_{(m)} \right| = o_p(n^{-1/2} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\|). \tag{S11}$$

Combining (S9)–(S11) proves (S8).

Then, using the definition of projection (5), we have

$$\sum_{j=1}^{J} m_{ij}''(a_{0j})(\alpha_j + \eta(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)) \left\{ \eta_0'(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)(\mathbf{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\}$$

has mean zero for any $\boldsymbol{\alpha}, \eta$, and thus (S8) implies

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \left\{ m_{ij}'(a_{0j}) + m_{ij}''(a_{0j}) \left(\eta_0'(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right) \right\} \\ & \cdot \left\{ (\eta_0'(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0))^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} = o_p \left(n^{-1/2} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \right). \end{aligned}$$

Noting that $\sum_j m_{ij}'(a_{0j}) = -\sum_j \kappa_{ij} (\delta_{ij} - \frac{\exp\{\alpha_{0j} + \eta_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)\}}{1 + \exp\{\alpha_{0j} + \eta_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)\}})$ has mean zero, and further using $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathbf{J}_0(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) + O_p(\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\|^2)$, the central limit theorem immediately implies the asymptotic normality.

$$\sqrt{n}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^{-1})$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{J}_0 \boldsymbol{\Sigma}^{-1} \mathbf{J}_0^{\mathrm{T}}),$$

where $\boldsymbol{\Sigma} = \mathbf{J}_0^{\mathrm{T}} \sum_{j=1}^{J} \left\{ \sum_{j=1}^{J} m''_{ij}(a_{0j}) \left(\eta'_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \mathbf{x}_{ij} - \widetilde{\alpha}_{0j} - \widetilde{\boldsymbol{\eta}}_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \right)^{\otimes 2} \right\} \mathbf{J}_0$. □

C.2.3 Lemmas

Lemma 1. *There exist positive constants C_1, C_2, C_3 such that whenever $\|\eta - \eta_0\|_\infty + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le C_1$, we have*

$$\begin{aligned} C_2(\|\eta - \eta_0\|^2 + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2) &\le E[\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})] - E[\ell(\eta_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)] \\ &\le C_3(\|\eta - \eta_0\|^2 + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2). \end{aligned}$$

Proof of Lemma 1. Since $\mathbf{g}_0 = (\eta_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ minimizes $E[\ell(\mathbf{g})]$, Taylor's expansion shows

$$E[\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})] - E[\ell(\eta_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)] = \int_0^1 \frac{d^2 E[\ell(\mathbf{g}_0 + \xi(\mathbf{g} - \mathbf{g}_0))]}{d\xi^2} d\xi.$$

We have

$$\begin{aligned} \frac{d^2 E\ell(\mathbf{g}_0 + \xi(\mathbf{g} - \mathbf{g}_0))}{d\xi^2} &= E[\textstyle\sum_{j=1}^{J} m''_{ij}(\alpha_{0j} + \eta_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) + \xi(\alpha_j - \alpha_{0j} + \eta(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) - \eta_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0))) \\ &\qquad \cdot \{\alpha_j - \alpha_{0j} + \eta(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) - \eta_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)\}^2]. \end{aligned}$$

Since m''_{ij} is bounded and bounded away from zero on bounded interval, and using Taylor's expansion and assumption (A4) it is easy to see $\{\alpha_j - \alpha_{0j} + \eta(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) - \eta_0(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)\}^2 \asymp \|\eta - \eta_0\|^2 + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$. □

Lemma 2. *If* $\|\eta\|_\infty + \|\eta_\lambda\|_\infty \le C$ *for some constant* $C > 0$*, then*

$$\begin{aligned} &\frac{\partial^2\{E[\ell(\mathbf{g}_\lambda + \xi(\mathbf{g} - \mathbf{g}_\lambda))] + \lambda\|\eta_\lambda + \xi(\eta - \eta_\lambda)\|^2_{(m)}\}}{\partial \xi^2} \\ \asymp\ & C(\|\eta - \eta_\lambda\|^2 + \lambda\|\eta - \eta_\lambda\|^2_{(m)} + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\lambda\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda\|^2), \forall \xi \in [0, 1]. \end{aligned}$$

Proof of Lemma 2. As in the proof of Lemma 1, we have

$$\begin{aligned} \frac{d^2 E\ell(\mathbf{g}_\lambda + \xi(\mathbf{g} - \mathbf{g}_\lambda))}{d\xi^2} &= E[\textstyle\sum_{j=1}^J m''_{ij}(\alpha_{\lambda j} + \eta_\lambda(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\lambda) + \xi(\alpha_j - \alpha_{\lambda j} + \eta(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \eta_\lambda(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\lambda))) \\ &\quad \cdot\{\alpha_j - \alpha_{\lambda j} + \eta(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \eta_\lambda(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\lambda)\}^2], \end{aligned}$$

and $\{\alpha_j - \alpha_{\lambda j} + \eta(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \eta_\lambda(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\lambda)\}^2 \asymp \|\eta - \eta_\lambda\|^2 + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\lambda\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda\|^2$. Furthermore, it is easy to see $d^2\|\eta_\lambda + \xi(\eta - \eta_\lambda)\|^2_{(m)}/d\xi^2 = \|\eta - \eta_\lambda\|^2_{(m)}$. □.

Lemma 3. *Let* $\mathcal{F} := \{(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta}) : \eta \in G_n, \boldsymbol{\alpha} \in \mathbb{R}^J, \boldsymbol{\beta} \in \mathbb{R}^d, \|\eta'\|_\infty \le C, \|\eta - \eta_\lambda\| + \sqrt{\lambda}\|\eta - \eta_\lambda\|_{(m)} + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\lambda\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda\| \le \sqrt{L}u\}$*, where* $u^2 = \frac{1}{n\lambda^{1/(2m)}} \wedge \frac{K}{n}$ *is as defined in Proposition 2, we have*

$$\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}} (E_n - E)\{\ell(\mathbf{g}) - \ell(\mathbf{g}_\lambda)\} = O_p(\sqrt{L}u^2).$$

Proof of Lemma 3. It is easy to see that $\ell(\eta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is Lipschitz continuous on each of the three variables on $\mathcal{F}$. Then, using the symmetrization inequality and the contraction inequality for the Rademacher process (Theorem 2.1 and Theorem 2.3 in Koltchinskii (2011)),

we have

$$
\begin{aligned}
& E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}(E_n-E)\{\ell(\mathbf{g})-\ell(\mathbf{g}_\lambda)\}\right)^2\right] \\
\leq\ & CE\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}\frac{1}{n}\sum_i\sigma_i\{\ell_i(\mathbf{g})-\ell_i(\mathbf{g}_\lambda)\}\right)^2\right] \\
\leq\ & C\Bigg(E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}\frac{1}{n}\sum_i\sigma_i\{\ell_i(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})-\ell_i(\eta,\boldsymbol{\alpha},\boldsymbol{\beta}_\lambda)\}\right)^2\right] \\
& +E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}\frac{1}{n}\sum_i\sigma_i\{\ell_i(\eta,\boldsymbol{\alpha},\boldsymbol{\beta}_\lambda)-\ell_i(\eta,\boldsymbol{\alpha}_\lambda,\boldsymbol{\beta}_\lambda)\}\right)^2\right] \\
& +E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}\frac{1}{n}\sum_i\sigma_i\{\ell_i(\eta,\boldsymbol{\alpha}_\lambda,\boldsymbol{\beta}_\lambda)-\ell_i(\eta_\lambda,\boldsymbol{\alpha}_\lambda,\boldsymbol{\beta}_\lambda)\}\right)^2\right]\Bigg) \\
\leq\ & C\Bigg(\sum_{j=1}^J E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}\frac{1}{n}\sum_i\sum_j\sigma_i\{\eta(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\lambda)-\eta_\lambda(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\lambda)\}\right)^2\right] \\
& +E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}\frac{1}{n}\sum_i\sigma_i\|\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda\|\right)^2\right] \\
& +E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}\frac{1}{n}\sum_i\sigma_i\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_\lambda\|\right)^2\right]\Bigg),
\end{aligned}
\tag{S12}
$$

where σ_i are i.i.d. Rademacher variables independent of all other variables.

Define $\mathcal{F}_\eta = \{\eta \in G_n : \|\eta'\|_\infty \leq C, \|\eta-\eta_\lambda\| + \sqrt{\lambda}\|\eta-\eta_\lambda\|_{(m)} \leq \sqrt{L}u\}$. For a fixed $j \in \{1,\ldots,J\}$, denoting $z_i = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\lambda$ for simplicity of notation, we will now show that

$$
E\left[\sup_{\eta\in\mathcal{F}_\eta}\left(\frac{(1/n)\sum_{i=1}^n\sigma_i\eta(z_i)}{\|\eta\|+\sqrt{\lambda}\|\eta\|_{(m)}}\right)^2\right] = O(u^2). \tag{S13}
$$

Assume $\psi_k, k=1,\ldots,K$ is an orthonormal basis of G_n (that is, $E[\psi_k(z)\psi_{k'}(z)] = I\{k=k'\}$). Write $\eta = \sum_{k=1}^K \eta_k\psi_k$ with $\eta_k = E[\eta(z)\psi_k(z)]$. By Proposition 2.4 of Huang and Su (2021),

we have $\|\eta\|^2 = \sum_{k=1}^K \eta_k^2$ and $\|\eta\|_{(m)}^2 \asymp \sum_{k=1}^K k^{2m}\eta_k^2$. Thus

$$\begin{aligned} & E\left[\sup_{\eta\in\mathcal{F}_\eta}\left(\frac{(1/n)\sum_{i=1}^n \sigma_i\eta(z_i)}{\|\eta\|+\sqrt{\lambda}\|\eta\|_{(m)}}\right)^2\right] \\ \leq\ & CE\left[\sup_{\eta\in\mathcal{F}_\eta}\frac{1}{n^2}\frac{\left(\sum_{i=1}^n\sigma_i\sum_{j=1}^K\eta_k\psi_k(z_i)\right)^2}{\sum_{k=1}^K(1+\lambda k^{2m})\eta_k^2}\right] \\ =\ & CE\left[\sup_{\eta\in\mathcal{F}_\eta}\frac{1}{n^2}\frac{\left(\sum_{k=1}^K\eta_k(\sum_{i=1}^n\sigma_i\psi_k(z_i))\right)^2}{\sum_{k=1}^K(1+\lambda k^{2m})\eta_k^2}\right]. \end{aligned} \tag{S14}$$

Using the Cauchy-Schwarz inequality, we have

$$\left(\sum_{k=1}^K\eta_k(\sum_{i=1}^n\sigma_i\psi_k(z_i))\right)^2 \leq \left(\sum_{k=1}^K(1+\lambda k^{2m})\eta_k^2\right)\left(\sum_{k=1}^K\frac{(\sum_{i=1}^n\sigma_i\psi_k(z_i))^2}{1+\lambda k^{2m}}\right).$$

Plugging the above into (S14), we get

$$\begin{aligned} & E\left[\sup_{\eta\in\mathcal{F}_\eta}\left(\frac{(1/n)\sum_{i=1}^n \sigma_i\eta(z_i)}{\|\eta\|+\sqrt{\lambda}\|\eta\|_{(m)}}\right)^2\right] \\ \leq\ & \frac{C}{n}E\left[\sum_{k=1}^K\frac{\psi_k^2(z)}{1+\lambda k^{2m}}\right] \\ =\ & \frac{C}{n}\sum_{k=1}^K\frac{1}{1+\lambda k^{2m}}. \end{aligned}$$

Using Proposition 2.5 of Huang and Su (2021) which stated that $\sum_{k=1}^K\frac{1}{1+\lambda k^{2m}} = O(\lambda^{-1/(2m)})$ and the trivial bound $\sum_{k=1}^K\frac{1}{1+\lambda k^{2m}} \leq K$, we obtain (S13).

Using (S13) for the first term on the right side of (S12), while the second and the third

terms are easily bounded by $O(n^{-1}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_\lambda\|^2+\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_\lambda\|^2))$, we see that

$$E\left[\left(\sup_{\mathbf{g}=(\eta,\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}(E_n-E)\{\ell(\mathbf{g})-\ell(\mathbf{g}_\lambda)\}\right)^2\right] \;=\; O(Lu^4).$$

An application of Markov's inequality then completes the proof of the Lemma. □.

References

Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, 63(6):2899–2939.

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

Huang, J. Z. and Su, Y. (2021). Asymptotic properties of penalized spline estimators in concave extended linear models: Rates of convergence. *The Annals of Statistics*, 49(6):3383–3407.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033. Springer Science & Business Media.

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

Tian, S., Yu, Y., and Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100.