Understanding Credit Risk for Chinese Companies using Machine Learning: A Default-Based Approach

Edward I. Altman, Xiaolu Hu, and Jing Yu*

March 2021

* Altman is from NYU Stern School of Business, New York University, Hu is from RMIT University, and Yu is from University of Sydney. Authors' contact information: Altman: eia1@stern.nyu.edu; Hu: xiaolu.hu@rmit.edu.au; and Yu: jing.yu@sydney.edu.au. The authors would like to thank the participants at the 1st Credit Scoring and Credit Rating Conference organized by Southwestern University of Finance and Economics in Chengdu China in October 2020, and the research seminars of the University of Sydney, Tianjin University and RMIT University for their helpful comments.

Understanding Credit Risk for Chinese Companies using Machine Learning: A Default-Based Approach

Abstract

In response to the recent elevated corporate credit risk environment in China's credit market, we develop a probability of default (*PD*) measure for Chinese companies using actual corporate bond defaults by applying the Least Absolute Shrinkage and Selection Operator (LASSO) machine learning model. Our *PD* measure is applicable to publicly listed and also, importantly, to unlisted companies. Our measure's bond default prediction accuracy outperforms models generated by alternative machine learning techniques and other prominent credit risk measures. Further analysis documents a large pricing effect of corporate default risk using our *PD* measure in primary and secondary bond markets. The pricing effect of default risk became more pronounced following two crucial market events in 2014 that raised market awareness of credit risk and is stronger for bonds likely traded by retail and foreign investors. In the cross section of bond and stock returns, we observe a positive distress risk premium after controlling for common risk factors. Finally, stocks of low *PD* firms outperformed those of high *PD* firms during the COVID-19 pandemic.

Keywords: Credit risk; Chinese bond market; Z-score; Distress risk premium, COVID-19 Pandemic

JEL codes: G12; G15; G23; G24

In recent years, China has moved slowly toward a more Western-style corporate debt market with more defaults, and in which investors are forced to be more discriminating about credit risks. *- Wall Street Journal (2020)*[1]

1. Introduction

The Chinese onshore bond market has grown rapidly in the past two decades with a market capitalization of USD 16 trillion as of August 2020 according to ChinaBond, making it the second largest fixed income market only to the US. As a percentage of a nation's total credit market, China's corporate bond market has now surpassed the US and is the largest in the world (Cherian, Mo and Subramanyam, 2020). Panel A of Figure 1 illustrates the rapid expansion of China's credit bond market from nearly non-existence in 2007 to a massive market at present. While slow in opening its doors to foreign investors, the Chinese credit market is gradually winning global recognition[2], accelerated by global investors' hunt for yield opportunities in the recent low interest rate environments and their search for a semi-safe haven asset given the country's early recovery from the COVID-19 outbreak. Panel B of Figure 1 displays the still low but potent growth of foreign holdings in China's credit market. However, as expressed in the opening quote, along with the unprecedented credit boom comes an increasing number of corporate bond defaults since the first bond default in 2014. Not long after the first default, on April 21, 2015, the first state-owned enterprise (SOE) bond default by Baoding Tianwei Group sent a shockwave across the market and challenged the long-held view on the implicit government guarantees for SOE borrowing. Recently, S&P Global Ratings (2020) warned of record high corporate defaults in China's credit bond market toward the end of 2020 as USD 0.93 trillion (CNY 6.5 trillion)[3] worth of corporate bonds were scheduled to reach maturities.[4] The warning was soon realized by the surge of credit defaults during the year, with a few salient SOE defaults rippling through the onshore credit market.[5] As the Chinese large and rising credit

---

[1] Yu, Xie, 20 November 2020, Missed Payments Rattle Confidence in China Corporate Bonds, *Wall Street Journal*.

[2] Although foreign restrictions prevented foreign investors to tap into the onshore Chinese bond market in the past, two recent moves greatly contributed the rapid expansion of foreign ownership in the local bonds in China. In 2017, the launch of *Bond Connect* program allowed foreign investors to trade Chinese bonds via Hong Kong stock exchange. Starting from April 1 2019, Bloomberg Barclays Global Aggregate Index has added the Chinese government and policy bank bonds. In February 2020, JP Morgan EMBI global index also added Chinese bonds, making it second-largest weighted country following Mexico. The inclusion into major global fixed income indices will usher in global fund managers who adopt index-based trading strategies and further improve the liquidity of China's credit market among foreign investors.

[3] The exchange rate used throughout the paper is 1 USD = 6.9618 CNY, the spot rate as of the end of 2019.

[4] Source: China Corporate Outlook 2020: Steep Walls, Few Catapults, S&P Global Ratings (2020).

[5] See, for e.g., the following Bloomberg articles for anecdotal evidence: "China Investors Brace for Record Defaults in Risky End to 2020" on August 23 2020 and "China State Banks Cut Corporate Bond Exposure Amid Rout" on November 13 2020.

market becomes riskier, it is pivotal to forecast corporate default risk and understand the pricing implications of such risk for enhanced security selection and risk management.

[Insert Figure 1 here]

To date, there are not many desirable candidate measures of default risk available to investors. Local credit ratings are policy-dependent[6] and highly skewed toward AA and higher. As of 2019, 90% of local bonds issued by nonfinancial companies are rated AA and above. A large proportion of high credit ratings indicates the lack of information content in the credit ratings in the onshore bond market.[7] While there is a rich literature on measuring financial distress risk, most of the distress risk models are calibrated using US data, and thus may not be suitable for emerging markets like China. Few notable exceptions that develop financial distress risk models for emerging market firms include Altman (2005) for emerging markets, Zhang, Altman and Yen (2010) for Chinese firms and very recently, Asis, Chari and Haas (2020) for emerging markets. In particular, Zhang, Altman and Yen (2010) attempt to build a modified Z score for Chinese firms. However, due to the absence of actual corporate bond defaults at the time, the authors then develop their default prediction model based on stocks under special treatment (ST), a warning for stocks with potential exchange delisting risk. Exploiting the rising number of corporate bond defaults since 2014, this study builds the corporate default risk model for Chinese companies based on actual bond defaults using both financial and nonfinancial information by employing a machine learning algorithm (MLA). Compared with other distress risk metrics, our probability of default (PD) measure has three important advantages: (1) Instead of benchmarking the US studies, we recognize the considerable differences between the US and China and employ a unique local set of variables to select the most predictive variables; (2) our measure does not require stock market inputs, and thus works for private bond issuers which accounted for 82% of corporate bond new issues from 2007 to 2019; and (3) we adopt a trained MLA and generate a parsimonious model which can be easily applied at low cost.

[6] In China, the credit ratings from accredited providers are used to calibrate capital requirements for banks and as investment guidance for insurance funds and money market funds. Banks and insurance firms are required to invest only in bonds rated A and above. In addition, only debt issues rated AA and above are eligible for bond repurchases in the exchange-based market (China Securities Depository and Clearing Corporation Limited (CSDC hereafter), 2017). Money market funds can only invest in bonds with credit ratings of AA+ and above (China Securities Regulatory Commission (CSRC hereafter), 2015), and insurance funds must hold debt instruments with ratings above A (see Amstad and He, 2018).

[7] As we complete this study, the lack of credibility of local credit ratings was once again reinforced by an AAA-rated Yongcheng Coal and Electricity Holding Group's bond default in early November 2020, calling for investors' caution over how idiosyncratic credit risk is in China and how reliable ratings are, as remarked by the *Wall Street Journal* article titled "Bond Ratings Get Lost in Translation" on November 20 2020.

As a nascent trend, MLA is gaining popularity in the recent finance literature in areas of asset pricing[8] and corporate default prediction[9]. In particular, these algorithms are argued to have advantages in selecting estimation variables for better bankruptcy predictions (Barboza, Kimura and Altman, 2017). In this study, we primarily use one particular MLA, the least absolute shrinkage and selection operator (LASSO) model, to determine the predictive variables for corporate bond defaults for Chinese bond issuers. It avoids overfitting problem by using a penalty function that removes all but the strongest predictors with cross validation (Hastie, Tibshirani and Friedman, 2001; Tibshirani, 1996). Further comparisons between the default model generated by LASSO with others derived from alternative MLAs as well as conventional default prediction models not based on machine learning techniques bolster our confidence in our choice of the MLA.

Our development of the corporate credit risk metric for Chinese companies proceeds in two steps. Based on a sample of 90 companies that defaulted on their corporate bonds from 2014 to 2018, we employ a LASSO technique to select financial and nonfinancial variables from the candidate variable list and derive a linear logit model on basis of quarterly financial information. We then examine the relative performance of our model in a secondary hold-out sample from 2019 to May 2020. Our quarterly *PD* measure estimated from the LASSO exhibits superior overall accuracy over other MLA methods including the support vector machines (SVMs), artificial neural networks (ANNs) and Random Forest (RF). It outperforms other frequently used credit risk models including credit ratings, Zhang, Altman and Yen's (2010) Z_{China} model, Merton's (1974) expected default probability (EDP), and Asis, Chari, and Haas's (2020) emerging market distress risk measure in the hold-out sample. This is evidenced by low Type I and Type II errors (approximately 15% and 21%) and high AUC ratio (90.11%). Distinguished from other models that often require stock market inputs and thus are limited to publicly listed companies, our *PD* measure shows reasonably high predictive accuracy for public and private companies, with an accuracy ratio above 80% for both groups.

[8] See Bali, Goyal, Huang, Jiang and Wen (2020), Chinco, Clark-Joseph and Ye (2019), Feng, Giglio, and Xiu (2020), Freyberger, Neuhierl, and Weber (2020) Lettau and Pelger (2020) among others.
[9] See, for example, Azayite and Achchab (2016), Kim and Sohn (2010), Olson (2012) and Pike, Sapriza and Zimmermann (2019).

Next, we assess the pricing effects of corporate credit risk in the primary and secondary corporate bond markets. After controlling for bond credit ratings and other bond/firm-specific characteristics, we observe a robust positive effect of corporate credit risk proxied by our *PD* measure in bond prices, suggestive of the incremental explanatory power of our *PD* measure in corporate bond pricing. Financially, the yield differential between bonds predicted to default and those not to default is 61 basis points (bps) in the primary market and 72 bps in the secondary market. While trivial in the past, the pricing effects of corporate credit risk become much more pronounced following two impactful market events that signal the government's waning interests in bad debt bailouts: the first corporate bond default in March 2014 and the ineligibility of repo financing for bonds rated below AAA, a policy change announced in December 2014. Further investigation analyzes investor heterogeneity by exploiting a distinct segmented bond market structure in China and the recent liberalization policy that significantly enhanced foreign holdings in China's corporate bond market. Without specific holding-level data, our preliminary analysis provides suggestive evidence that corporate credit risk is priced to a larger extent for more risk-averse retail and foreign investors than domestic institutional investors.

To demonstrate the investment value of our corporate credit risk measure, we also perform portfolio-level analysis to show the return predictivity of our *PD* measure in the cross section of corporate bonds and stocks. The universe of corporate bonds and stocks are sorted into quintile portfolios based on quarterly *PD* measure. We document a significant and positive credit risk premium in both markets. For corporate bonds, the monthly abnormal return on a zero-investment hedge portfolio (equal- or value-weighted) which buys bonds from issuers with the highest *PD* and sells short bonds from issuers with the lowest *PD* is in the range of 0.032%-0.115%, varying with the choice of benchmark asset pricing model for bonds. In line with the cross-sectional regression evidence, the asset pricing effects of credit risk enlarge after the first bond default and asset pledgeability reform that raise the investor awareness of credit risk and after the introduction of *Bond Connect* that effectively opens the market to foreign investors. We observe similar findings in the stock market. The monthly return differential between two extreme credit risk stock portfolios (i.e., High - Low) is in the range of 0.243%-0.320%. The collective evidence suggests that investors demand higher returns for securities of companies with greater financial distress risk, consistent with a positive distress risk premium. In the final step, we test the stock return predictivity of our credit risk model surrounding the outbreak of the COVID-19 pandemic, which invokes a sudden and unprecedented stress test on firms' financial strength. Our results show that stocks of low *PD* companies significantly outperform those of high *PD* companies, further substantiating the validity of our measure in capturing a firm's financial resilience amidst large-scale economic crises.

We make several contributions to the existing literature. First, our study contributes to a large stream of the credit risk literature mostly devoted to the US by developing a novel *PD* measure for an emerging market, China, based on actual corporate bond defaults. Our simple *PD* estimation only requires inputs from publicly available information. Hence, it is applicable at low cost and suits not only public but also private firms, which dominate China's onshore bond market. Second, despite the fast emerging literature on the massive but still mysterious Chinese onshore bond market, the extant studies on this market mostly control for credit risk using either established credit risk models built in more developed markets or simply the policy-dependent, but less informative, credit ratings. We fill the gap in the literature by designing an up-to-date credit risk measure using newly available actual corporate bond defaults. Developed based on the typical characteristics of defaulted Chinese bond issuers, our *PD* measure exhibits superior predictive accuracy and is priced in bond prices. Third, a long-standing asset pricing anomaly exists in the asset pricing literature (e.g., Dichev, 1998; Griffin and Lemmon, 2002; Johnson, 2004 and Penman Richardson and Tuna, 2007), which finds lower expected stock returns associated with higher financial distress risk, contrary to the conventional wisdom of risk and return tradeoff. On the basis of our *PD* measure, we document a positive credit risk premium, that is, a positive link between corporate credit risk and cross-sectional equity returns in China, addressing the "financial distress puzzle" in a new emerging market setting.

The remainder of the paper is organized as follows. Section 2 provides an overview of China's onshore bond market and conducts a literature survey on the contemporary research in the intersection of the two streams of studies that investigate the nascent emerging debt markets in China and those that attempt to develop credit risk models for emerging markets. Our primary *PD* measure is developed using a multivariate logit regression model using LASSO and is compared with other MLA models and commonly used credit risk metrics in Section 3. Section 4 examines the pricing effect of corporate default risk in the primary and secondary bond markets. Section 5 performs portfolio-level analyses and explores the investment implications of our default risk measure in the cross section of bond/stock markets over the longest possible timespan and Section 6 concludes.

2. Institutional background and literature review

2.1 Onshore Chinese bond market

There is no doubt that the Chinese onshore bond market experienced exponential growth since 2007. As shown in Figure 1 Panel A, the overall bond market expanded dramatically from less than CNY 20 trillion (~ USD 2.87 trillion) in 2007 to nearly CNY 110 trillion (~ USD 15.80 trillion) as of August 2020. In parallel, corporate bond market grew from merely CNY 0.66 trillion (~ USD 0.09 trillion) in 2007 to slightly over CNY 23 trillion (~ USD 3.30 trillion) a decade later, ranking first in the world. To explain the rapid market expansion, Chen, He and Liu (2020) attribute the accelerated development of corporate bond market to the hangover effect of the four-trillion-yuan stimulus package in 2009. Chinese local governments financed the stimulus through bank loans in 2009 and then resorted to non-bank debt financing after 2012 under the rollover pressure from maturing bank loans.

As the primary interest of this study, our discussion focuses on publicly issued debt instruments from nonfinancial companies. There are four main categories of local corporate bonds: short-term commercial papers, medium-term notes (MTNs), corporate bonds and enterprise bonds. Considering the short-term nature of commercial papers (less than one year), our analysis only covers MTNs, corporate bonds and enterprise bonds and terms them broadly as corporate bonds. The divergence of these debt securities reflects in their trading venues and governing authorities. Short-term commercial papers and MTNs are traded on the interbank market and governed by the People's Bank of China (PBoC). Corporate bonds are traded on the exchange-based market (i.e., Shanghai or Shenzhen Stock Exchange) and governed by the CSRC. Enterprise bonds, mainly issued by SOEs, are regulated by the National Development and Reform Commission (NDRC)[10] and traded on either the interbank market or cross-traded on both the interbank and the exchange-based markets. More than half of the corporate bonds are issued by publicly listed companies, whereas less than 1% of enterprise bonds are issued by public firms (Livingston, Poon and Zhou, 2018).

[10] More detail about the regulatory structure can be found from Amstad and He (2018).

In the past, the onshore bond market had been off-limits to foreign investors due to the policy restrictions. While they are granted access to the exchange market, this market is much smaller in terms of market value and trading volume relative to the interbank market. The liberalization of the interbank market gradually occurred from 2002 but picked up speed after 2010 driven by the central government's will to promote direct financing through financial markets. Primarily guided by the publication of the Regulation on Domestic Investment by Qualified Foreign Institutional Investor (QFII) in 2006, foreign investors qualified as a QFII candidate are given access to the interbank market. Over the period from 2010-2015, the PBoC launched several reforms to further liberalize the market including substantially expanding the list of QFII institutions and increasing the foreign investment quota from USD 30 billion to USD 80 billion. By July 2015, almost all types of foreign institutional investors can participate in the interbank markets with no approval requirements under quota limits. Despite a series of significant policy reforms, Cherian, Mo and Subramanyam (2020) observe a mere 1.1% market share of foreign investment in the interbank market by June 2017. They, however, highlight an important event that considerably boosted the foreign investment activities in the local bond market. The *Bond Connect* program in July 2017, which established a trading platform in Hong Kong with an intention to allow offshore investors and mainland investors to invest in each other's bond markets, triggered a 70% increase in foreign holdings in the interbank market one year later.

In line with the accelerated liberalization view, graphic evidence as shown in Figure 1 Panel B suggests that global investors' appetite for onshore China bond market is growing rapidly in recent years in face of the economic tension between China and the US and the COVID-19 pandemic. Compared with only CNY 0.10 trillion (~ USD 0.01 trillion) in 2007, the total foreign holdings tally up to CNY 2.46 trillion (~ USD 0.35 trillion) as of August 2020 according to ChinaBond. As a prominent example, China bonds were added to two major global fixed income indices, Bloomberg Barclays Global Aggregate Index in April 2019 and JP Morgan EMBI in February 2020. As a further push to the opening-up of the interbank market, China bond officials announced an extension of trading hours from the closing time of 5:00 pm to 8:00 pm to accommodate the steadily increasing global investors from the US and Europe from September 2020 onwards.

2.2 Corporate bond credit and default risk in China

Historically, there were literally no bond defaults in the Chinese credit market as the Chinese government, central or local, bailed out all financially distressed bond issuers like in many socialist and transition economies (Faccio, Masulis and McConnell, 2006; Zhu, 2016). As the bond market (corporate bond market) grew from only 47.54% (2.43%) in proportion to GDP in 2007 to nearly 100% (20.05%) in 2019, implicit government support can no longer sustain the long-term development of China's debt markets (see Figure 2 Panel A). Unsurprisingly, the government's tolerance for corporate defaults shifted in early 2014, marked by the State Council's releasing of a policy document, the Guiding Principles for the Healthy Development of Capital Markets. This document called for boosting direct financing through capital markets and easing out government support in financial markets. Afterwards, the rate of corporate bankruptcies in the overall economy grew from 0.03% in 2015 to 0.10% in 2018. The rate of bankruptcies in the US stayed around 0.40% throughout the period (see Figure 2 Panel B).[11] As for the debt markets, the first bond default of "11 Chaori Bond" rocked the local credit market, followed by another hit from the first SOE bond default of Baodian Tianwei Group in 2015.[12] Ever since, the number of corporate bond defaults is ramping up. Compared with six bond defaults with a total value of CNY 1.34 billion (~ USD 0.19 billion) in 2014, the number of bond defaults exceeded 142 in 2019, a total value of CNY 107.4 billion (~ USD 15.43 billion) in 2019 (see Figure 2 Panel C).[13] The rising bond defaults could be attributable to three interrelated factors: (1) the central policy stance changes to boost direct financing in the economy for a healthy financial system;[14] (2) the local financial markets' gradual integration into global financial system restricts governmental intervention; and (3) the government's ability to rescue distressed firms diminishes in wake of a phenomenal wave of defaults and bankruptcies post 2014. Nevertheless, the upsurge of corporate bond defaults poses pressing challenges for investors to assess and price the default risk for Chinese companies.

[11] Bankruptcy rate is the number of bankruptcies over total number of firms. The number of new accepted bankruptcy cases in China is sources from Supreme People's Court annual report and the total number of firms in China is from National Bureau of Statistics of China. The number of business bankruptcy filing in the US is sourced from American Bankruptcy Institute and the number of firms in the US is from United States Census Bureau.

[12] Jin, Wang, and Zhang (2018) provide an excellent summary of the first ever SOE bond default of Baodian Tianwei group. They find that the reduction of implicit government guarantees, manifested by this event, has real effects on SOEs. Following the SOE bond default, there is a decline in investment and net debt issuance and an increase in cash holdings in SOEs compared to non-SOEs.

[13] We only consider the first-time default of each bond.

[14] As part of an effort to limit the government's intervention in economic activities, Li and Ponticelli (2020) note two major changes of China's bankruptcy system. In 2006, a new bankruptcy law which drew on the judicial experiences of the US and Europe was approved and took effect in June 2007, with an intent to strengthen creditor rights protection. Second, between 2007 and 2017, 97 specialized courts with better trained judges were introduced in Chinese cities to expedite the processing of local bankruptcy cases.

[Insert Figure 2 Here]

Credit ratings are commonly used in developed markets as a market barometer for corporate default risk. But this seems not the case in China. The lack of rating differentiation and inflated credit ratings are well known (Jiang and Packer, 2017; Livingston, Poon and Zhou, 2018; Hu, Shi, Wang, and Yu, 2020). Roughly 99% of the newly issued corporate bonds are rated AA or above in the period of 2014-2019. As shown in Panel C of Figure 2, among the 392 defaulted bonds in the period of 2014-2019, 27% of their issuers were rated AA+ or higher at issuance. Part of the reason is the overweight of state ownership in their rating algorithms (see, for e.g., Moody's (2014)). Yet, the waning of government support prompts fresh concerns about using policy-dependent credit ratings to capture credit risk.

2.3 Literature review

The rapid development of China's onshore credit market, coupled with the growing global investors' attention, have stimulated a burgeoning literature examining the economics of this vast but still nascent fixed income market. Some prominent examples are in order. Amstad and He (2018) provide an introductory overview of onshore Chinese bond market, and Cherian, Mo and Subramanyam (2020) conduct an extensive survey on policy reforms focusing on foreign liberalization of the local bond market and its liquidity effects. Hu, Pan and Wang (2020) review the characteristics of Chinese capital markets including stocks, government bonds, corporate credit bonds and financial bonds and empirically analyze the historical returns of stock and bond markets in China.

With the improved data availability in Chinese corporate bonds, the recent literature has witnessed an emergence of studies exploring typical new features of China's bond market. For example, Chen, He and Liu (2020) adopt a macroeconomic perspective and identify an underlying economic driver for the recent proliferation of Chinese credit bond market through the hangover effect of the four-trillion-yuan stimulus plan imposed in 2009. Several studies explore the bond pricing effects. For example, Ding, Xiong and Zhang (2020) uncover an issuance overpricing, rather than underpricing, in the primary Chinese credit markets. Jiang and Packer (2017) and Livingston, Poon and Zhao (2018) show evidence of a significant correlation between credit ratings and bond offering yields, although the local rating scales are very broad, pooling bonds of different default risks into few high rating categories. Turning to the secondary bond market, Cherian, Mo and Subramanyam (2020) and Mo and Subramanyam (2020) investigate whether and how the liquidity effect is priced and to what extent a sequence of market liberalization policies has affected the liquidity effects in China's bond market. Geng and Pan (2019) demonstrate the improved price discovery among non-SOE issuers following the first bond default. Other contemporary studies take advantage of the special settings in China's bond market and explore topics wherein there remain knowledge gaps in developed markets such as implicit government guarantees (Jin, Wang, and Zhang, 2019), the role of retail investors (Liu, Wang, Wei and Zhong, 2019; Mo and Subramanyam, 2020), asset pledgeability (Chen, Chen, He, Liu and Xie, 2019), and credit enhancements (Gao, Huang and Mo, 2020).

Despite its first-order importance for bond investors, the literature that attempts to gauge credit risk among Chinese companies is surprisingly scarce. The scholarly endeavor to modelling business credit risk can be dated back to Altman (1968) who develops a Z score for large US companies using the multivariate discriminant analysis (MDA) technique. Over the past 50 years, the growing importance of credit risk in capital markets has spurred a rich literature. In search of credit risk metrics, there are three dominant methodological approaches used in academic research: the discriminant analysis (Altman, 1968), the logit regression (Ohlson, 1980; Shumway, 2001; Campbell, Hilscher and Szilagyi, 2008), and the structural model inspired by Merton's (1974) option pricing model (Vassalou and Xing, 2004; Bharath and Shumway, 2008). These credit risk models are estimated in the more developed US market and thus may not be valid in emerging markets. In sharp contrast, not much literature exists for modelling credit risk for companies from emerging markets due in part to the dearth of reliable corporate default data in these marketplaces. A few exceptions include Altman (2005) and Asis, Chari and Hass (2020). Given the tremendous differences in institutional structures and accounting standards between China and other emerging markets, a generalized business failure model for emerging markets may not be suitable for Chinese firms. Zhang, Altman and Yen (2010) make an early attempt to build a Z_{China} score in the original Z score framework. However, in absence of actual corporate defaults, the model is developed based off a sample of ST stocks. Recognizing actual default probabilities as an important input in default risk prediction models, we aim at improving the credit risk model for Chinese firms by exploring a sample of firms that defaulted on their bonds.

3. Credit risk prediction model development

3.1 Sample formation

At start, it is important to note that firms considered as defaulted are those that have defaulted on their bonds in this study. Our analysis focuses on nonfinancial Chinese credit bonds issued by both private and publicly listed companies. The bond default information is retrieved from the Wind database for the period from 2014 to 2018. To avoid unnecessary duplicates in our sample, we only analyze the first-time bond default for each issuer. At each quarter end, we match the default information for that date with the most recently available financial statement information. We assume that financial information becomes available three months after fiscal quarter end and allow up to one year gap for the accounting information to be matched with respective bond defaults. After excluding firms with missing financial information, there remain 90 companies that defaulted on their credit bonds, among which 11 are publicly listed companies trading on either Shanghai or Shenzhen stock exchanges. In terms of state ownership, our default sample includes 17 SOEs and 73 non-SOEs. Over the same time period, there are 5,233 corporate bond issuers with available financial information, suggesting an average default rate of 1.72%. Each quarter, we form our estimation sample by randomly selecting non-defaulted issuing companies to maintain an average default rate around the actual default rate of 1.72%. For the validity test, we form a secondary hold-out sample of 47 defaulted firms between 2019 and May 2020[15]. Table 1 describes our estimation and hold-out sample in each quarter.

[Insert Table 1 Here]

3.2 Variable selection

Drawn from a rich literature relating to the US and China's credit risk prediction models[16], we compile an array of financial variables in each of the following six accounting ratio categories: Leverage, Liquidity, Profitability, Coverage, Activity and Structure. The specific accounting variables under each category are outlined in columns 1-2 of Table 2. Apart from the conventionally used financial variables, we also consider two nonfinancial variables, namely,

[15] To ensure that our primary results are not driven by a particular random selection outcome, we bootstrap the random selection 1,000 times and find that our LASSO model provides an average accuracy ratio of 89.69% in the estimation period and 89.62% during the holdout period. The evidence confirms that our primary findings are insensitive to the random selection procedure.
[16] See, among others, Altman (1968), Altman, Fargher and Kalotay (2010), Beaver (1966), Law and Roache (2015), Mselmi, Lahiani and Hamza (2017), Tian, Yu and Guo (2015) and Zhang, Altman and Yen (2010).

SOE (a dummy variable which equals one if the firm is owned by the government, and zero otherwise) and *Local GDP* (the GDP growth rate of the province where the company is headquartered). The selection of *SOE* tends to capture a firm's government backing, explicit or implicit, following the notion that state-owned firms are more likely to receive preferential government treatments. We also consider the economic conditions of local governments given the important role of local governments to both SOE and non-SOE companies in China. Huang, Li, Ma, and Xu (2017) pinpoint the decentralization of SOEs from central to local governments since 1998. After decentralization, the rights regarding the SOE's finance, labor, wage, social welfare, and personnel are transferred to local government; all income taxes are turned to local public finance, providing a clear incentive for local officials to support local SOEs. Not only for SOEs, Allen, Qian, and Qian (2005) highlight the role of local governments in the growth of nonstate firms, particularly in areas of successful economic growth. In the survey of non-SOE firms from economically prosperous regions in China, they find that more than 40% state that local governments support their growth without demanding profit sharing. We anticipate that provinces with more advanced economic development are more capable of providing financial support for local firms. While there is potential to extend the list of nonfinancial variables, we aim at focusing on publicly available information that can be easily acquired at low cost.

[Insert Table 2 Here]

Primarily, we consider models generated by a LASSO-based machine learning technique as well as two conventional corporate default models constructed using stepwise selection process and multivariate discriminant analysis (MDA). Compared with other MLAs, an apparent advantage of LASSO is that it outputs a linear logit model which is more suitable for identifying a business failure problem because its dependent variable takes a value of either zero (non-defaulted firm) or one (defaulted firm) and for conveniently giving the estimated probability of default. Specifically, we develop our default model using LASSO based on Bayes Information criterion (BIC) following the recommendation of Freyberger, Neuhierl and Weber (2020) and Yuan and Lin (2006).[17] Table 2 displays the default models estimated based

[17] In unreported results, we explore the other two LASSO techniques, adaptive LASSO and LASSO based on the cross-validation (CV) criterion. We find that the logit model from adaptive LASSO is rather similar to the one derived from LASSO BIC. LASSO CV, however, yields a logit model that is counterintuitive. Hence, in our paper we report our model on basis of LASSO BIC technique throughout the paper unless otherwise specified.

on the three statistical approaches. In selecting the predictive model, we must trade off the statistical significance against economic intuition. The predictive model from LASSO recruits five financial variables, that is, *Total liabilities/Total assets*, *Interest bearing debt/Total assets*, *Cash and short-term investment/Total assets*, *Net Income/Total assets* and *Current liabilities/Total liabilities* and two non-financial variables as shown in column 3. All signs of these variables are consistent with our expected economic interpretations. The stepwise forward selection process[18], however, generates counterintuitive predictions on *Total liabilities/Total assets*, (*Current assets – Inventory)/Current liabilities*, *Current assets/Total assets* and *Current liabilities/Total liabilities*; Likewise, MDA also gives counterintuitive predictions on *Book equity/Total liabilities* and Interest *expenses/ EBITDA* (column 5). The lack of economic justifications motivates us to focus on the LASSO based predictive model for the subsequent analysis. We perform further comparison tests to demonstrate the forecasting accuracy of our selected model in later sections.

In terms of the selected financial variables, it is necessary to draw some initial comparisons with the Z_{China} score developed by Zhang, Altman and Yen (2010). The previous Z_{China} model closely resembles $Z^{''}$ score developed for emerging markets in Altman (2005) and involves two profitability variables (*Net profit/Total assets* and *Retained earnings/Total assets*), one liability variable (*Total liabilities/Total assets*) and one activity variable (*Working capital/Total assets*). Our new *PD* model, however, weighs on two liability variables, one profitability variable, one liquidity variable and one structure variable. The distinct model inputs highlight the fact that the previous Z_{China} is estimated using ST-prefixed stocks which, by definition, are assigned based on their financial profitability. Our *PD* model is derived from corporate bond defaults which emphasizes more on the liability-side of firm financials and manifests a firm's debt capacity.

3.3 Multivariate logit regression model

After choosing the LASSO based predictive model, we take a granular look at the model in Table 3. Panel A of Table 3 compares the mean and median differences in the predictive variables across default and control groups. The statistical differences are mostly consistent

[18] We start from estimating the logit model and then eliminate the least helpful covariates one by one, until all the remaining input variables are efficient, that is, their significance level is below a pre-specified critical level (5%).

with our priors. The mean and median of *Total liabilities/Total assets* and *Interest bearing debt/Total assets* are statistically greater in the default group than the control group. The cash holding, *Cash and short-term investment/Total assets* and the financial profitability variable, *Net income/Total assets*, are statistically lower in the default group. The structure variable, *Current liabilities/Total liabilities*, is significantly higher in the default group. Defaulted bond issuers are less likely to be an SOE or located in less developed regions.

In Panel B, we run the logit regression model using the LASSO selected variables. Before analyzing the coefficient estimates of the multivariate logit model, we perform a series of goodness-of-fit tests suited for binary logistic regression models. The Chi-squared statistic of the Wald test for whether the coefficients for predictors are simultaneously equal to zero is 173.61, significant at 1%, suggesting that each selected predictor is statistically important for the risk prediction model. This suggests that including these predicting variables altogether improves the fit of the model. Further, the Hosmer-Lemeshow test is employed to assess the discrepancy between predicted probabilities and observed probabilities, with a null hypothesis that predicted probabilities deviate from observed probabilities in a way predicted by binomial distribution. The *p*-value of this test is 0.290, suggesting that our logit model fits the data well[19]. Overall, the diagnostic tests indicate that we are fitting an appropriate prediction model. We further quantify the predictive accuracy of the model by computing its Area Under the Curve (AUC) ratio defined by Keenan and Sobehart (1999) (also known as the Receiver Operating Characteristics (ROC) curve). The ROC curve shows the proportion of true positives (a defaulter is correctly classified as a defaulter) versus the proportion of false positives (a non-defaulter is wrongly classified as a defaulter), whereas the AUC ratio measures the entire area underneath the ROC curve and reflects the aggregate performance of a predictive model. Hence, a higher value of AUC indicates a better discriminatory accuracy of the model. As shown in Table 3, our logit model's in-sample AUC ratio is 89.65%, with a 95% confidence interval of [86.45%, 92.86%], indicating a superior predictive power. As for the concern about the limited number of defaults, we perform a powerful *Leave-one-out* validation test, whereby we build the model based on N-1 companies, and hold out one firm to check its accuracy. Then we repeat this test N times. Eventually the test generates an average accuracy ratio based on N samples. We continue to observe a reasonably high accuracy ratio of 89.70% in this test, confirming the

[19] We apply 10 groups when conducting the Hosmer-Lemeshow test. The result does not qualitatively change when we change the group number to 3, 5 or 7.

predictive power of our *PD* model. All the input variables exhibit expected signs. Our estimated *PD* increases with the leverage variables, *Total liabilities/Total assets* and *Interest bearing debt/Total assets*, increases with the short-term liabilities, *Current liabilities/Total liabilities* and decreases with the profitability variable, *Net income/Total assets* and the liquidity variable, *Cash and short-term investment/Total assets*. In terms of two nonfinancial variables, *SOE* and *Local GDP* both show negative correlations with the probability of default.

[Insert Table 3 Here]

3.4 Validity test

As an important next step, we perform the sanity check on a hold-out sample, consisting of 47 defaulted issuers observed between 2019 and May 2020. Following the same matching procedure used to form the estimation sample, we are able to locate 2,686 non-defaulted companies which have outstanding bonds and available financial information over the same time period.

On basis of this hold-out sample, we compare the prediction accuracy of our primary model, the LASSO-based logit model, against other three machine learning models including the SVMs, the ANNs and the RF, and other four default prediction models including the Z_{China} score model in Zhang, Altman and Yen (2010)[20], the logit model estimated using numerical credit ratings (*Credit Rating*), the EDP (*EDP*) measure following Bharath and Shumway (2008) for publicly listed companies, and the logit models using the distress risk measure developed by Asis, Chari and Haas (2020) for publicly listed companies in emerging markets. The construction of the EDP measure and Asis, Chari, and Haas's model are detailed in Sections OA1 and OA2, respectively.

In Panel A of Table 4, we attempt to compare the predictive power of various MLA models and existing default risk prediction models. All the MLA models apply a ten-fold cross validation. The additional MLA methods use the candidate variables listed in Table 2 as the model inputs. We train a SVM with Gaussian radial basis function (RBF) kernel with

[20] The Z_{China} score in Zhang, Altman, and Yen (2010) is computed as follows: Z-score=-0.460×(Total liabilities/Total assets)+9.32×ROA+0.388×(Working capital/Total assets)+1.158×(Retained earnings/Total assets)+0.517.

parameters C equals to 10 and γ equals to 0.1.[21] We introduce 20 hidden layers in the ANNs with 500 times of training[22]. For the random forest model, we also apply 500 times of iteration and require a minimum of five observations on each ending branch. The MLA models together with the existing credit risk metric, we derive the probability of default using a logit model framework. Subsequently, the Youden index is used to determine the cutoff points for the logit model estimations. As for MDA-based models, the cutoff value for the previous Z_{China} metric is 0.5 following Zhang, Altman and Yen (2010).

[Insert Table 4 Here]

To demonstrate the accuracy of these risk prediction models, we rely on the following five indicators: In-sample AUC (the AUC of estimation period), Hold-out sample AUC (the AUC of testing period), Type I error rate (the probability of classifying a defaulter as a non-defaulter, failed warning), Type II error rate (the probability of predicting a non-defaulter as a defaulter, false warning) and Accuracy (1- the weighted average of Type I and II error rates). Panel B of Table 4 reports the misclassification and accuracy rates for each model. All the prediction models yield an accuracy ratio above 70% except the Z_{China} for the hold-out sample. Our primary model has a second highest prediction accuracy with a hold-out sample AUC of 90.11%, only 0.6% lower than the random forest model. Figure 3 clearly depicts the high accuracy of our models in both estimation and hold out samples. By observing the AUC in estimation sample and hold-out sample, we can also find the overfitting problem in SVMs and RF models that their hold-out sample AUC values are substantially reduced although they produce a near 100% in-sample AUC.

Turning to the misclassification, the LASSO logit model exhibits the lowest Type I error rate of 14.89% and a reasonable level of Type II error rates of 21%. Comparably, although the ANNs, RF and credit ratings provide a lower Type II error and thus a higher weighted average accuracy, their failed warning rates (Type I error) is much higher than the LASSO model, which are 31.91%, 100% and 36.96%, respectively. This indicates 30%-100% of the defaults cannot be detected if employing these three models. Except that, the results from ANNs and RF are hard to interpret, which increase the application cost of these two methods. The SVMs, Z_{China}, EDP and Asis et al.'s models have higher Type I and Type II error rates than our primary

[21] When we change the parameters for other combinations, the results are qualitatively unchanged.
[22] The results are qualitatively unchanged when using 10 hidden layers.

model. Trading off between the misclassification rate, accuracy rate and application costs, our LASSO logit model has superior prediction performance among all the considered prediction models with a low cost. It is therefore applied as the primary default risk prediction model hereafter.

[Insert Figure 3 Here]

To substantiate the reliability of our model, we present further validity tests on subsamples of listed and unlisted firms in Panel B of Table 4. Irrespective of the estimation sample and hold-out sample, we find that our model exhibits a high accuracy of predicting bond defaults for both publicly listed and private companies. The realized prediction accuracy is above 83% and the AUC is greater than 81% for listed and unlisted companies on both samples. This evidence confirms the advantage of our model in forecasting financial distress risk for the prevalent unlisted bond issuers.

Next, we expand the prediction horizon up to 28 quarters and track the quarterly changes in our *PD* measure in this enlarged window. To be specific, for each company that defaulted its corporate bond from 2014 to May 2020, we find a comparable non-defaulted company that is in the same industry and closest in sales. In Figure 4, we plot the cross-sectional averages of the *PD* measures for defaulters and non-defaulters 28 quarters leading up to the default events. It is clear that our *PD* measure starts to rise 8 quarters prior to the actual bond defaults, suggesting the effectiveness of this measure.

[Insert Figure 4 Here]

4. The pricing effect of credit risk in the primary and secondary bond markets

Credit risk is of the first-order importance to bond investments. In the event of corporate bond default, bond investors lose a significant portion, if not all, of their investments. To compensate for the potential default risk exposure, it is plausible that bond investors demand a higher return for bonds issued by companies of high default risk. If it truly captures the underlying credit risk, we expect our *PD* measure to be reflected in the bond pricing process. In this section, we link our *PD* prediction to corporate bond prices in the primary and secondary markets. Further investigation explores market episodes that carry implications for corporate credit risk and takes into consideration investor heterogeneity in segmented bond markets.

4.1 Main empirical analyses

We test the pricing effect of our *PD* measure using the following regression specification:

$$Yield\ spread_{i,t} = \alpha_{i,t} + \beta_1 PD_{i,t-1} + \beta_2 X_{i,t-1} + \varepsilon_{i,t} \tag{1}$$

The dependent variable is alternately at-issue bond yield spread in the primary bond market and bond market yield spread in the secondary market where a Treasury note of comparable maturity is used as a benchmark bond. The list of bond- and issuer-specific control variables are assembled based on prior literature (e.g., Ziebart and Reiter (1992), Blume, Lim and Mackinlay (1998), Campbell and Taksler (2003), and Livingston, Poon and Zhou (2018)). We list the control variables in Table 5 and define them in Appendix 1. The summary statistics of the two dependent variables as well as control variables can be found in Online Appendix Table OA2. We estimate Equation (1) at the bond issue level in the primary market analysis and at the bond-month intervals in the secondary market analysis. Our *PD* measure is estimated for each bond issue in the primary market analysis and for each bond-month observation in the secondary market analysis under the assumption that quarterly financial information becomes available three months after the financial quarter end (CSRC, 2007; NAFMII, 2012).

The corporate bond sample used in these analyses is formed in the following manners. We first extract all the median and long-term corporate bonds in form of nonfinancial medium-term notes (MTN), enterprise bonds and corporate bonds from 2007 through May 2020 from the Wind database.[23] The initial selection leads to 17,826 corporate bond new issues. Further filtering criteria are imposed to form the final sample. Specifically, bonds with missing identifiers, more than one coupon payment per year and missing information on control variables are removed from our analysis. Our final sample of primary bond market is comprised of 17,450 bonds from 4,491 issuers. To construct the secondary market sample, we merge this primary bond market sample with bond trading information including the monthly close price,

[23] We begin our corporate bond sample in 2007 because no such bonds exist prior to 2007 after filtering.

monthly weighted average trading price and the yield to maturity (YTM).[24] To further mitigate the bond market peculiarities, we drop the following bond-month observations: (1) have a maturity of less than one month, (2) an average daily trading price falls below CNY 5 or above CNY 130, (3) have less than three-month trading history during our sample period, or (4) have missing information on dependent or control variables. After the screening, there remain 228,248 bond-month observations for 13,963 bonds issued by 4,084 issuers between 2007 and May 2020.

Before a formal empirical analysis on the relation between our *PD* measure and bond prices, we first display the time series variation in the *PD* measure. Figure 5 illustrates the mean, 90th percentile and maximum value of the *PD* measure over time in the primary and secondary bond markets as well as stock markets. Regardless of the three statistics, we consistently observe a heightened corporate default risk after 2014 following the first corporate bond default in these markets. All of the three statistics of the *PD* measure dipped in the first quarter of 2020 after the COVID-19 outbreak thanks to the government's efforts to encourage creditors to refinance debt and accept payment delays. The upward trend in the *PD* measure resonates with mounting anecdotal evidence on the rising default risk in the fast growing China corporate bond markets.

To explore the information content of Chinese credit ratings, we also graphically compare our *PD* measures with the issuer's ratings. In Figure 6, we plot the mean, 90th percentile and max value of *PD* in each of the following rating categories: AAA, AA+, AA, AA- and below AA-. If ratings are informative, we expect to observe these statistics trending up as we move across ratings from the highest to the lowest. However, for the primary bond market in Panel A, none of the three statistics exhibit a clear upward sloping pattern; for the secondary market, only mean and 90th value increase as credit ratings deteriorate. Moreover, there is substantial variability in the range of *PD* values in each rating category for both bond markets. For instance, the max value of *PD* is 54.88%, but the mean is only 1.10% for AAA-rated bonds. The overall graphic evidence is suggestive of an overly broad, if not uninformative, credit rating system in Chinese corporate bond market.

[24] In Wind database, there is much missing YTM information in the early years. We replace the missing YTM by applying the following equation: $\mathrm{PV}_t = \frac{FV}{(1+YTM_t)^{N_t-1+d_t/365}} + \int_{k=0}^{N-1} \frac{C_t}{(1+YTM_t)^{k+d_t/365}}$ where *PV* is the monthly close dirty price, *FV* is the notional face value of the bond, *N* is the total number of coupon payments left at time *t*, *d* is the number of days between time *t* and the next coupon payment date and *C* is the coupon payment per period.

[Insert Figures 5 and 6 Here]

Next, we examine the pricing effects of our *PD* measure by estimating Equation (1) specified above, controlling for a host of bond- and firm-specific characteristics as well as industry and year fixed effects.[25] To mitigate the impact of outliers, we winsorize all of the firm-level financial variables at the top and bottom 1% of the sample distribution. We begin with estimating the baseline regression model in the primary bond market. In column 1, we first exclude credit rating dummies, *Leverage*, *ROA* and *Cash* from the list of controls in the base regressions for the concern that there could be high correlations between these variables and our *PD* measure.[26] The preliminary result shows the coefficient of *PD* is 0.108, significantly at 1% level. We then respectively include *Leverage*, *ROA* and *Cash* in column 2 and include credit rating dummies in column 3. It is reassuring to note that the *PD* coefficient is 0.105 in column 2 and 0.092 in column 3, both significant at 1% level. The regression model shown in column 4 incorporates all of the bond and firm characteristic variables, and our *PD* measure remains statistically significant. We interpret the economic significance of our findings in two ways. First, based on the full model estimation in column 4, a one-standard-deviation increase in *PD* is associated with a yield increase of 24.25 bps.[27] Second, we introduce a *PD (dummy)* which equals one if *PD* is above its cutoff point (0.981%) and zero otherwise[28]. In column 5, we observe that bonds issued by firms are predicted to default experience an additional yield of 61.4 bps compared to those that are predicted not to default. Turning to the secondary bond market, the pricing effect of *PD* is of a similar magnitude. Take the model estimation in column 10 with *PD (dummy)* for instance. The yield differential between predicted defaulters and non-defaulters is as large as 71.8 bps in the secondary market. In Panel B of Table 5, we further split our full sample into bonds issued by publicly listed and private companies in columns 1-2 and columns 5-6, and into SOE and non-SOE issuers in columns 3-4 and columns 7-8. We find that the *PD* coefficient remains positive and statistically significantly at 1% level across these subsamples. The overall evidence suggests the statistically and economically significant effect of our *PD* measure in corporate bond pricing, even after controlling for credit ratings.

[25] Notably, our results are insensitive to the inclusion of firm fixed effects.

[26] Financial statements are common information source for the construction of our *PD* measure and for third-party credit ratings. *Leverage*, *ROA* and *Cash* also show high correlations with the predictors of our logit model (i.e., *Total liabilities/Total assets*, *Net income/Total assets* and *Cash and short-term investment/Total assets*).

[27] As reported in Table OA2, the sample standard deviation of *PD* is 2.694%.

[28] This cutoff threshold is obtained using the Youden index.

The results pertaining to credit rating dummy variables are also noteworthy here. Using AAA-rated bonds as the base case, a sequence of credit rating dummies is included in our baseline model. They are AA+, AA, AA- and below AA-. Across all the model specifications, we find consistent evidence of positive coefficients on all of the rating dummy variables. This suggests a price discount for bonds rated below AAA, relative to AAA rated bonds. Additionally, there is a monotonic increase in the estimated coefficient of the rating dummy moving from AA+ to below AA-. It means that bond investors recognize the value of bond ratings and incorporate it into the pricing process. From this perspective, the credit ratings in China's bond market are not absolutely uninformative, albeit high. In Table OA3 of the Online Appendix, we inspect the sensitivity of bond prices to the *PD* measure across credit ratings. In both primary and secondary bond markets, we note a larger coefficient for *PD* among corporate bonds rated AA+ and AA relative to those rated AAA or below AA, indicating the lack of informativeness for AA+ and AA ratings.[29]

[Insert Table 5 Here]

We conduct a number of robustness tests. First, instead of using the Treasury note yield as the benchmark rate, we use an alternative benchmark rate – yields on comparable bonds issued by China Development Bank (CDB). The results presented in columns 1 and 3 in Table OA4 of the Online Appendix suggest that our primary results are insensitive to this benchmark interest rate. Second, we follow Ang, Bai, and Zhou (2018) and use a synthetic yield as the benchmark interest rate. Specifically, we isolate the bond yield spread using a synthetic Treasury bond in a two-step procedure. In the first step, we fit the zero-coupon Treasury bond yield curve using Svensson (1994)'s exponential functional form. Next, we back out the implied price of the synthetic Treasury bond with the same coupon rate, coupon frequency and maturity date as the corresponding bond issue by discounting each cash flow using the Treasury bond zero-coupon rates on the yield curve. The matching Treasury bond yield is then calculated using its implied price. Compared with our primary bond yield spread measure that only considers the maturity in matching, this spread measure controls for all the cash flow effects unique to each bond issue. The results presented in columns 2 and 4 of Table OA4 confirm the robustness of our primary findings. Third, we replace the *YTM spread* with an alternative

[29] Due to the limited number of observations for the "Below AA-" rating category, we classify bonds with AA- or below AA- ratings as one group, referred to as "Below AA".

measure, *YTM spread (Avg)* where the YTM of each bond-month observation is calculated from its trading volume-weighted average price. As can be seen in column 5 in Table OA4, our results continue to hold. We further control bond and year-month fixed effects for the secondary bond market analysis in columns 6 and 7 respectively in Table OA4, and our results are qualitatively the same. Last, we exclude bonds with puttable, callable or sinking fund arrangements from the analysis in columns 1 and 3 of Table OA5, and exclude Chengtou bonds, a type of enterprise bonds issued by local government financial vehicles (LGFVs) and presumably backed by local governments, in columns 2 and 4 of the table. Our results continue to hold on these alternative samples.

4.2. Exogenous shocks to corporate credit risk in China

Two salient market events have considerably shifted investor credit risk perception and elevated the awareness of corporate credit risk in recent years. One, the first corporate bond default by Chaori group in 2014 and the first SOE bond default in 2015 have important market implications by signifying the government's increased tolerance of corporate failures and the gradual exit of government bailouts in Chinese credit bond market. [30] We anticipate an increased credit risk premium after the first bond default. In parallel with the inception of bond defaults, the second event pertains to the repo pledgeability reform that was released on December 8, 2014. The exchange bond market announced that enterprise bonds with credit ratings below AAA are no longer eligible as repo collateral. This sudden policy move is rooted in the 2009 four-trillion stimulus package (Chen, He and Liu, 2020). In response to the stimulus policy, local governments funded infrastructure investments through LGFVs. In face of the rollover risk about five years later, LGFVs resort to the bond market and refinance bank loans and continuing infrastructure investments by issuing Chengtou bonds. To rein in the systemic risk implied by these bonds, the State Council explicitly banned the local government backing of these bonds in October 2014 and subsequently financial regulators removed the repo pledgeability of bonds rated below AAA.[31] The loss of collateral value of a large proportion of

[30] In keeping with the view of the exit of government bailouts in corporate bond defaults, Geng and Pan (2019) find that the link between credit spreads and the default risk measured by EDP only becomes statistically significant following the first bond default in 2014. Using the first SOE bond default in 2015, Jin, Wang, and Zhang (2018) suggest that implicit government guarantees are worth of at least 1.75% of bond value. Moreover, the reduction of implicit government support has significant real effects including reduced corporate investment and increased cash holdings in SOEs compared with non-SOEs. The two studies provide support for the increased credit risk and concerns over such risk in China's onshore bond market.

[31] Chen, Chen, He, Liu and Xie (2019) provide extensive discussions and in-depth analysis on the causal effect of this asset pledgeability policy in the exchange bond market.

corporate bonds rated below AAA could cut investors' risk tolerance, further increasing investors' demand for credit risk premium in bond prices.

To test the marginal effects of the two exogenous events on corporate credit risk, we augment our baseline model (1) by incorporating a *Post 2014* dummy variable capturing the first bond default event and a *Post Pledgeability* dummy variable reflecting the asset pledgeability policy reform. Specifically, *Post 2014* (*Post Pledgeability*) takes the value of one if the observation is after March 2014 (December 2014) and zero if otherwise. To differentiate the event effects, our focus lies on the interaction term between our *PD* measure and the two event dummies. We limit the sample to exchange-traded bonds when analyzing the effect of the pledgeability policy as it is only applicable in the exchange bond market. Although targeted at enterprise bonds, we believe that the policy has a widespread impact in raising investor awareness of credit risk for all bonds traded on the exchange market. Therefore, we report our further analysis on the sample of all the exchange-traded bonds. After including the interaction term, we find a statistically insignificant coefficient on *PD* but significant coefficient on *PD×Post 2014* for both the primary and secondary bond markets in columns 1 and 3 of Table 6, consistent with the elevated credit risk in China's bond market post the first bond default. Accordingly, investors' credit risk appetite shifts following the removal of the collateral eligibility of below-AAA rated bonds in the repo transactions. Across the primary and secondary markets as shown in columns 2 and 4 of Table 6, the interactions between *PD* and *Post Pledgeability* are robustly positive and significant. This evidence is again in line with the declined investor risk tolerance after the policy reform.

[Insert Table 6 Here]

4.3 Investor heterogeneity in segmented bond markets

In this section, we take advantage of the special market structures and reforms in China and explore the varying pricing effect of credit risk proxied by our *PD* measure across retail, institutional and foreign investor groups. *Ex ante*, we expect a stronger pricing effect of credit risk for retail and foreign investors than domestic institutional investors due to their differentiated access to firm information. In general, domestic institutional investors have privileged access to information. Prior studies provide ample evidence of large creditors participating in direct conversations with the borrowing company's management and gaining access to confidential information such as asset growth updates and covenant renegotiations.[32] The information advantage of this nature is largely limited for small retail investors and foreign investors who are still new in this market and trading from overseas. The unlevelled information access reduces retail and foreign investors' risk tolerance, aggravating their credit risk aversion. To test these conjectures, we distinguish between retail and institutional investors by exploiting a segmented market structure in Chinese corporate bond market and between domestic and foreign investors invoked by a recent bond market reform, *Bond Connect*, which opens the bond market to foreign investors.

As discussed in Section 2, a distinctive feature of China's corporate bond market is the co-existence of two bond trading venues: the interbank and exchange markets. While the interbank markets are dominated by large institutional investors, the exchange is characterized by small transactions and open for both retail and institutional investors (Hu, Pan, and Wang, 2020; Mo and Subramanyam, 2020). Without actual retail investor holdings information, we attempt to explore the credit risk perception of retail investors by comparing the pricing effect of credit risk between the interbank and exchange markets. A dummy variable, *Exchange*, which equals one if the bond is traded on exchanges and zero if otherwise, is introduced into the baseline regression model. As shown in columns 1 and 3 of Table 7, we observe positive and statistically significant coefficients on *PD*× *Exchange* across the primary and secondary bond markets, suggesting that retail investors attach more importance to credit risk derived from publicly available information.

Further, following the introduction of the *Bond Connect* program in July 2017, foreign investors are given access to onshore bond market. The northbound link of this program allows foreign investors to access the interbank bond market. Mo and Subramanyam (2020) document a significant liquidity improvement of this liberalization policy. Taking a different perspective, we investigate the policy's credit risk implications by estimating a regression model analogous to the above market segmentation test. We incorporate into the base model an interaction term between *PD* and a *Post Connect* dummy variable, which equals one if the observation is after July 2017 and zero otherwise. As expected, the coefficient of *PD*×*Post Connect* is consistently positive and statistically significant in the yield spread regressions on both the primary and secondary markets, as illustrated in columns 2 and 4 of Table 7. The evidence supports the

[32] See, for e.g., Andersen (2006), Ivanshina and Sun (2010), and Han and Zhou (2014).

notion that the involvement of risk averse foreign investors improves the extent to which credit risk is priced in bond markets. Extending Mo and Subramanyam (2020), our findings demonstrate that foreign investors not only increase the demand of liquidity but also the awareness of credit risk in China's bond market, improving the price discovery in this nascent bond market.

[Insert Table 7 Here]

5. Credit risk premium in bond and stock markets: Portfolio-level analysis

We have thus far examined the pricing effects of corporate credit risk, measured by our *PD* variable, in the cross section of corporate bond returns in the bond-level analysis. To explore the investment values of our *PD* measure, we perform portfolio-level analysis on the relation between corporate credit risk and bond/stock returns in this section. With this approach, we can look into whether corporate credit risk is priced in the cross section of bonds and stocks in excess of traditional risk factors widely documented in the asset pricing literature. In the last step, we extend the stock market analysis to the COVID-19 period and inspect the role of credit risk in driving stock performance during the crisis.

5.1. Univariate portfolio analysis on corporate bond returns

The asset pricing implication of credit risk is carried out based on the bond transactions data from the Wind database. In addition to the filtering criteria imposed to form the secondary bond market data for the cross-sectional regression analysis (see Section 4.1), we also require a bond has at least three months of trading price data over the 12-month period following the monthly bond portfolio formation. As a result, the final sample used in our bond portfolio analysis consists of 186,139 bond-month observations of 11,975 bonds from 3,758 issuers from October 2008 to May 2020.

The construction of corporate bond portfolios closely follows Bai, Bali and Wen (2019). Every month, we generate a *PD* estimate for a corporate bond issuer. To prevent the look-ahead bias, our one-quarter ahead *PD* measure is updated using previous quarter financial information, assuming new financial information is released three months after the fiscal quarter end. Then, for every month, we sort corporate bonds into quintiles in accordance with the bond issuer's

PD measure from low (Quintile 1) to high (Quintile 5). The bond portfolios are rebalanced using newly updated monthly *PD* estimates. Last, we derive a monthly time series of expected excess returns on portfolios from the expected excess returns on individual bonds, equal- or value-weighted (by outstanding debt amount). To proxy for a bond's expected excess return in a given month, we measure the average of monthly realized returns in excess of risk-free interest rate (i.e., one-year fixed term deposit rate) in the next 12-month period after the monthly one-quarter *PD* is updated. The following equation illustrates the derivation of monthly realized bond returns:

$$r_{i,t} = \frac{P_{i,t} + AI_{i,t} + C_{i,t}}{P_{i,t-1} + AI_{i,t-1}} - 1 \tag{2}$$

where $P_{i,t}$ is the monthly volume-weighted average clean trading price of bond *i* in month *t*, $AI_{i,t}$ is the accrued interest, and $C_{i,t}$ is the coupon paid by bond *i* in month *t*.

Table 8 displays the values of *PD* estimates and abnormal portfolio returns (i.e., alphas) of quintile bond portfolios by accounting for common market risk factors in a multitude of asset pricing models. From left to right, we consider raw excess portfolio returns, alphas estimated based on CAPM, three-factor, five-factor, seven-factor and ten-factor bond asset pricing models. In the CAPM model, the only risk factor is stock market excess return, MKT^{Stock} which is the difference between A-share stock market return and one-year fixed term deposit rate. The three-factor model expands the CAPM model by adding stock market size (*SML*) and value (*HML*) factors following Fama and French (1992)'s method. Considering the typical characterization of bond market, we further augment the stock market three-factor model by including the default risk (*DEF*) and term risk (*TERM*) factors in the five-factor pricing model. *DEF* is the difference between the monthly long-term credit bond return (measured by the CSI Credit Bond Index return) and the monthly long-term Treasury bond return (measured by CSI Treasury Bond Index return), while *TERM* is the difference between the monthly long-term Treasury bond return (measured by CSI Treasury Bond Index return) and the one-year deposit rate. The seven-factor pricing model accounts for additional bond risk factors including bond market excess return (MKT^{Bond}), downside risk factor (*DRF*), the liquidity risk factor (*LRF*), the return reversal factor (*REV*) and the credit risk factor (*CRF*), motivated by Bai, Bali and Wen (2019). The ten-factor pricing model comprehensively controls for all the above-mentioned risk factors. For brevity, we provide a detailed discussion about these factor pricing models in Appendix 2 and present the descriptive statistics of these risk factors used for bond portfolio analysis in Table OA6 of the Online Appendix.[33]

In Panel A of Table 8, we examine the equally weighted excess returns and risk-adjusted alphas on portfolios sorted on quarterly *PD* measure. The bottom row, "High-Low", reports the abnormal returns on a zero investment strategy of buying bonds in Quintile 5 with the highest *PD* and selling bonds in Quintile 1 with the lowest *PD*. The *t*-tests for the statistical significance of portfolio returns are based on Newey-West standard errors, corrected for heteroskedasticity and serial correlation of three lags[34]. In column 1, we note a large spread of the *PD* values across quintile portfolios. The average *PD* is only 0.077% in the low-risk bond portfolio, but it ramps up considerably to 4.969% in the high-risk portfolio. Without controlling for any other risk factors, column 2 shows that raw excess returns increase monotonically from the low-risk to high-risk portfolios. The average monthly excess return on the low-risk portfolio is 0.216%, whereas is 0.294% in the high-risk portfolio, implying an annualized excess return of 0.924% (0.077% on a monthly basis) on the long-short credit risk hedge portfolio.

It is possible that our *PD* measure simultaneously captures other common risk factors known to the literature. To rule out this possibility, we estimate portfolio alphas in various asset pricing models well established in the asset pricing literature. Consistent with Bai, Bali and Wen (2019), stock market risk factors have very limited explanatory power in cross-sectional bond returns. The alpha differential between low- and high-risk bond portfolios hardly changes after adjusting for CAPM or stock market three-factor asset pricing model. The alpha differential slightly enlarges when we estimate five-, seven, and ten-factor models that explicitly consider bond market risk factors. Taking the ten-factor model for example, the high-risk portfolio alpha exceeds that of low-risk portfolio by 0.115%, amounting to an annualized difference of 1.380%. As shown in Panel B, the portfolio analysis based on value-weighted portfolio returns yield qualitatively the same results. Taken together, the evidence here is in line with the conventional risk-return trade-off that bond investors demand a price discount (i.e., high bond returns) for bonds that have large credit risk exposure.

[33] In Table OA7 of the Online Appendix, we use an alternative method to construct the DRF and CRF factors by defining the 5% VaR as the second lowest monthly return observation over the past 36 months. Our previous results continue to hold.
[34] The number of lags is determined based on the rule-of-thumb: $0.75 \times N^{(1/3)}$ where N is the number of months in the sample.

[Insert Table 8 Here]

We conduct further investigation on the dynamic pricing effect of corporate credit risk following major market events and policy updates. As mentioned above, we anticipate an increased credit risk premium following the first corporate bond default in 2014 and the loss of collateral benefits of corporate bonds rated below AAA in 2017. In Panel C, we split the full sample into two subperiods before and after the first bond default in 2014. As expected, while the *PD* measure is significantly larger in the high-risk portfolio than in the low-risk counterpart, the return differential between the two portfolios is statistically insignificant prior to 2014. After 2014, the first default event serves as a wake-up call for bond investors and markedly improves investors' awareness of corporate credit risk. Consistently, the excess return of high-risk portfolio is significantly greater than that of low-risk portfolio post 2014, suggesting that at-risk bonds compensate investors for a higher return due to raised credit risk. The asset pledgeability policy reform signifies the officials' concern about the contagion effect of credit risk from bond market to other financial sectors. In addition, the significant drop of collateral value of bonds rated below AAA further weakens investors' risk tolerance for financially troubled bond issuers. Consistent with the decline in credit risk tolerance among market participants, we observe the vastly different pricing effects of corporate credit risk before and after the 2017 pledgeability reform in Panel D. Note that we only include exchange-traded bonds in this set of tests because this policy is only applicable to this market. Despite the significant *PD* difference between high and low default risk portfolios, their return differential is insignificant prior to the policy reform but rather statistically and economically significant post the policy change (0.410; t-value = 5.382). The above results remain robust using value-weighted portfolio excess returns, substantiating the cross-sectional regression analysis on the important bond pricing implications of the two credit risk events documented in Table 8.

In Panel E, we examine the asset pricing impact of the establishment of the *Bond Connect* program, which considerably promoted the foreign investors' participation in onshore bond market. Interestingly, we find that the return differential between high- and low-risk portfolios is only statistically significant after the introduction of the *Bond Connect* program. The portfolio analysis once again confirms the findings of cross-sectional regression analysis that bond market liberalization improves the extent to which credit risk is priced in bond prices.

5.2 Univariate portfolio analysis on stock returns

We further our portfolio analysis to stock markets, motivated by the long-standing asset pricing puzzle in the credit risk literature. It is conventional wisdom that stocks of financially distressed firms exhibit greater expected returns as a compensation for investors to assume default risk. However, extant empirical studies generate much mixed results in this regard depending on the measurement of distress risk and expected returns. Some studies including Vassalou and Xing (2004), Chava and Purnanandam (2010) and Friewald, Wagner and Zechner (2014) document a positive relationship between distress risk and expected stock returns. Yet, several others find the opposite. Prominent examples include Dichev (1998), Griffin and Lemmon (2002), Johnson (2004) and Penman, Richardson and Tuna (2007). Some evidence exists to attribute the distress risk puzzle to arbitrage limits and slow market reaction to financial distress risk. We complement these studies by examining the asset pricing implications of corporate credit risk in an emerging market.

To form the stock sample, we include all A-share stocks listed on Shanghai and Shenzhen stock exchanges between April 1994 and May 2020. We then purge out the following stocks: financial stocks, stocks of companies with missing information on book equity and market capitalization and those with negative book equity value as well as ST stocks given that the prices of ST stocks are capped by regulation. We also require a stock has at least three months of trading price data over the 12-month period following the monthly stock portfolio formation. The final sample covers 333,806 firm-month observations for 3,095 unique publicly listed companies spanning 312 months.

Similar to the bond portfolio analysis, every month between April 1994 and May 2020, we estimate a quarterly *PD* measure using previous quarter financial information to prevent the look-ahead bias. Then, the universe of stocks is sorted into stock quintiles based solely on the monthly *PD* measure from lowest (Quintile 1) to highest (Quintile 5). Our quintile portfolios are updated monthly. To compute the expected portfolio return, we aggregate the expected individual stock excess returns in each portfolio, each month using the equal-weighted or value-weighted (by market value) approach. The expected stock return is calculated as the average of monthly excess returns in the next 12 months after *PD* is estimated, where individual stock excess return in a given month is the difference between realized monthly stock return and risk-free rate.

Along with the mean value of *PD*, Panels A and B of Table 9, respectively, report the equal-weighted and value-weighted portfolio returns across five stock portfolios. The spread of the *PD* measure is narrower compared with that in bond portfolio analysis. The average *PD* is 0.026% in the low-risk group, and is 3.948% in the high-risk group, reflecting the better capital access and thus lower default risk of public companies relative to private peers. We sequentially demonstrate the return differential between high-risk and low-risk portfolios based on raw excess returns, CAPM-adjusted alphas, and Fama-French three-factor alphas. Regardless of the asset pricing model, we consistently observe a positive relation between credit risk and expected stock portfolio return. In terms of the raw stock returns, the equally weighted excess return on the lowest-risk portfolio is 1.356%, and it increases to 1.677% in the highest-risk portfolio. The return differential between the two portfolios is 0.320% (equivalent to 3.840% annualized excess return), statistically significant at 1% level. We examine whether common risk factors could explain away the significance of credit risk using CAPM and Fama-French three-factor model as the benchmark asset pricing models in columns 3 and 4. Even after controlling for market risk factors, the abnormal return between low- and high-risk portfolios remains statistically and economically unchanged. Similar results are obtained using value-weighted portfolio returns in Panel B. Collectively, our findings in regard to a positive credit risk premium in the cross section of stock returns run counter to the financial distress risk puzzle, observed in the US literature.

[Insert Table 9 Here]

5.3 Credit risk and stock performance during the COVID-19 pandemic

The sudden outbreak of COVID-19 pandemic and the subsequent lockdowns posed an unprecedented economic hardship to companies in China and around the world. Among others, firms that are in financial distress prior to this crisis are expected to take the hardest hit due to their already weak financials. In this section, we are interested to find out the link between credit risk and stock performance pre- and post-COVID outbreak. Empirically, we conduct both stock-level and portfolio-level analysis using our *PD* measure as a proxy for corporate credit risk. To execute the empirical plans, we retrieve daily stock returns over the period from September to May 2020, with 23rd January 2020 as the cutoff date to define the start of the crisis following the government's official announcement. Similar to Section 5.2, we exclude financial stocks and ST stocks. The sample used in this analysis covers 574,339 firm-day observations from 3,665 public companies. For each company, we calculate its quarterly *PD* measure. Next, we estimate the following equation,

$$R_{i,t} = \alpha_{i,t} + \beta_1 Post\ COVID_{i,t} + \beta_2 Post\ COVID_{i,t} \times PD_i + \varepsilon_{i,t} \tag{3}$$

where *Post COVID* is a time dummy variable which equals one if the observation is on or after January 23, 2020 and zero otherwise. $R_{i,t}$ denotes the daily return of stock *i* on day *t*. In columns 1-2 in Panel A of Table 10, we regress daily stock returns on *Post COVID* and its interaction with *PD*, without or with the control of day and firm fixed effects. Our interest lies on the coefficient estimate of the interaction term, which tells us about the differential stock performance of companies with varying credit risk. Without including day fixed effects, column 1 gives us the result of a negative and significant coefficient on *Post COVID*, indicating an average within-firm decline in stock return after the onset of COVID-19. More importantly, the coefficient of *PD*×*Post COVID* is statistically negative, consistent with an even worse stock performance for high-risk firms. The estimated coefficient of the interaction term continues to be significant after the inclusion of day fixed effects in column 2.

A concern that naturally arises from the daily stock return regressions is the mechanical relation between large sample size and statistical significance. To mitigate this concern, we follow Albuquerque, Koskinen, Yang and Zhang (2020) and use pre- and during-crisis cumulative stock returns to measure stock performance and condition cumulative returns on corporate credit risk, captured by our *PD* measure. In other words, for each stock, we respectively compute its cumulative stock return over the pre-crisis period (i.e., cumulative daily returns from 1st September 2019 through 22nd January 2020) and over the during-crisis period (i.e., cumulative daily returns from 23rd January 2020 through 30th May 2020). The modified regression model is specified below:

$$CR_{i,T} = \alpha + \beta_1 Post_{i,T} + \beta_2 Post\ COVID_{i,T} \times PD_i + \varepsilon_{i,T} \tag{4}$$

where $CR_{i,T}$ is the cumulative daily return for stock *i* in time period *T* which alternatively represents two periods before and after the COVID-19 outbreak. The stock-level regression model substantially reduces the sample from 574,339 stock-days to 7,330 observations. It is reassuring that the coefficient of *Post×PD COVID* continues to be negative and statistically significant in the new regression framework with the control of firm fixed effects.

[Insert Table 10 Here]

Table 10 present the numerical comparison of portfolio returns between high- and low-credit risk groups before and after the crisis outbreak based on daily portfolio returns in Panel B and cumulative portfolio returns in Panel C. In Panel C, we find that the average daily portfolio return decreases from 0.118% pre-crisis to 0.003% after, a net significant difference of -0.115% for the group of stocks with lowest *PD* values. The high *PD* stocks experience an even bigger price decline during the crisis. The average return declines from 0.133% to -0.087%, a net decline of -0.220% per day for the stocks with the highest default probability, significant at 1%. The same results hold for cumulative portfolio returns computed over the pre- and post-crisis periods. The magnitude of cumulative return declines due to the crisis in general increases with corporate credit risk. Overall, the stocks with low credit risk prior to the crisis outperforms those that already expose to credit risk prior to the economic hit, resonating with a contemporary US study by Glossner, Matos, Ramelli, and Wagner (2020) that institutional investors prioritized corporate financials, the “hard” measures of firm resilience, during the COVID-19 crisis.

6. Conclusions

In response to the mounting evidence on the rising default risk in the ever expanding but still nascent onshore Chinese corporate bond market, this study is devoted to develop a simple credit risk scoring system that is suited for not only publicly listed companies but also private companies that account for a vast majority of bond issuances in China’s credit bond market using the machine learning technique. Exploiting the emergence of corporate bond defaults since 2014, our *PD* measure is estimated from a multivariate logit regression model derived by LASSO BIC that involves five financial variables and two nonfinancial variables including a firm’s state ownership status and the economic development of local government. Our new credit risk model exhibits high accuracy and superior to other MLA techniques including SVMs, ANNs and RF, and third-party credit ratings as well as credit risk models developed based on the US or other emerging market data.

After controlling for credit ratings as well as other bond and issuing company characteristics, we observe a significant and robust pricing effect of corporate credit risk measured by our *PD* measure in both primary and secondary bond markets. The value relevance of our *PD* measure becomes more pronounced following two impactful market events that raise the investor awareness of credit risk in corporate bond market: the first corporate default in March 2014 and the removal of bonds rated below AAA for repo financing in December 2014. Additionally, the enhanced market liberalization after the establishment of the *Bond Connect* program also improves the extent to which credit risk is priced in bond markets. We conclude by exploring the investment implications of our new credit risk measure. Employing portfolio-level analysis, we document a positive credit risk premium in the cross section of corporate bond and stock returns, robust to other common risk factors documented in the asset pricing literature. That is, investors are compensated with higher returns for taking financial distress risk. During the COVID-19 pandemic, our credit risk measure identifies financially troubled companies whose stocks considerably underperformed others, suggestive of a valid link between corporate credit risk and firm resilience to unexpected disasters. Summing up, our credit risk carries important practical implications for bond and stock market investors. Against the backdrop of the overly broad credit ratings, growing credit risk and heightened global interest in China's corporate bond markets, it is essential for investors to correctly gauge credit risk in order to preclude credit defaults and investment losses.

References

Allen, F., Qian, J. and Qian, M., 2005. Law, finance, and economic growth in China. *Journal of Financial Eeconomics* 77, 57-116.

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23, 589-609.

Altman, E.I., 2005. An emerging market credit scoring system for corporate bonds. *Emerging markets review* 6, 311-323.

Altman, E.I, Fargher, N. and Kalotay, E., 2010. A Simple Empirical Model of Equity-Implied Probabilities of Default. *The Journal of Fixed Income* 20, 71-85.

Albuquerque, R., Koskinen, Y., Yang, S. and Zhang, C., 2020. Resiliency of environmental and social stocks: An analysis of the exogenous COVID-19 market crash. *The Review of Corporate Finance* Studies 9, 593-621.

Amstad, M., and Z. He., 2018. Chapter 6: Chinese bond market and interbank market. In: Amstad M., G. Sun and X. Wei (eds.), Handbook of China's Financial Systems.

Anderson, J., 2006. As lenders, hedge funds draw insider scrutiny. New York Times (October 16).

Ang, A., Bai, J., Zhou, H., 2018. The great wall of debt: Real estate, political risk, and Chinese local government credit spreads. Working paper, BlackRock, Georgetown University, and Tsinghua University.

Asis, G., Chari, A. and Haas, A., 2020. In search of distress risk in Emerging markets. Working paper, National Bureau of Economic Research.

Azayite, F.Z. and Achchab, S., 2016. Hybrid discriminant neural networks for bankruptcy prediction and risk scoring. *Procedia Computer Science* 83, 670-674.

Bai, J., Bali, T.G. and Wen, Q., 2019. Common risk factors in the cross-section of corporate bond returns. *Journal of Financial Economics* 131, 619-642.

Bali, T.G., Goyal, A., Huang, D., Jiang, F. and Wen, Q., 2020. The cross-sectional pricing of corporate bonds using big data and machine learning. Working paper.

Bao, J., Pan, J. and Wang, J., 2011. The illiquidity of corporate bonds. *The Journal of Finance* 66, 911-946.

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

Beaver, W.H., 1966. Financial ratios as predictors of failure. *Journal of Accounting Research* 4, 71-111.

Bharath, S.T., and Shumway, T., 2008. Forecasting default with the Merton distance to default model. The *Review of Financial Studies* 21, 1339-1369.

Blume, M.E., Lim, F., MacKinlay, A.C., 1998. The declining credit quality of US corporate debt: Myth or reality? *The Journal of Finance* 53, 1389–1413.

Breiman, L., 1996. Bagging predictors. *Machine learning* 24, 123-140.

Breiman, L., 2001. Random forests. *Machine learning* 45, 5-32.

Campbell, J.Y., Taksler, G.B., 2003. Equity volatility and corporate bond yields. *The Journal of Finance* 58, 2321–2350.

Campbell, J.Y., Hilscher, J. and Szilagyi, J., 2008. In search of distress risk. *The Journal of Finance* 63, 2899-2939.

Chari, A., Stedman, K.D. and Lundblad, C., 2020. Taper Tantrums: Quantitative Easing, Its Aftermath, and Emerging Market Capital Flows. *The Review of Financial Studies. Forthcoming*.

Chava, S., and Purnanandam, A., 2010. Is default risk negatively related to stock returns?. *The Review of Financial Studies* 23, 2523-2559.

Chen, H., Chen, Z., He, Z., Liu, J. and Xie, R., 2019. Pledgeability and asset prices: Evidence from the Chinese corporate bond markets (No. w26520). National Bureau of Economic Research.

Chen, Z., He, Z. and Liu, C., 2020. The financing of local government in China: Stimulus loan wanes and shadow banking waxes. *Journal of Financial Economics* 137, 42-71.

Cherian J., Mo, J and Subrahmanyam, M., 2020. China's Corporate Credit Bond Market: Development and Policy Issues. Working Paper.

Chinco, A., Clark‐Joseph, A.D. and Ye, M., 2019. Sparse signals in the cross‐section of returns. *The Journal of Finance* 74, 449-492.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning* 20, 273-297.

CSDC, 2017. No 47, Notice on matters related to the release of the "business guidelines on access standards for pledge-type repurchase eligibility and the value of discount coefficient of standard bonds (2017 revision). Retrieved from http://www.chinaclear.cn/zdjs/editor_file/20170407171435538.pdf

CSRC, 2007. Measures for the administration of information Disclosure of listed companies. Retrieved from http://www.csrc.gov.cn/pub/zjhpublic/zjh/200804/t20080418_14481.htm.

CSRC, 2015. No 120, Measures for the supervision and administration of money market funds. Retrieved from http://www.csrc.gov.cn/pub/zjhpublic/G00306201/201512/t20151218_288348.htm

Dichev, I.D., 1998. Is the risk of bankruptcy a systematic risk?. *The Journal of Finance* 53, 1131-1147.

Ding, Y., Xiong, W. and Zhang, J., 2020. Overpricing in China's corporate bond market (No. w26815). National Bureau of Economic Research.

Djankov, S., McLiesh, C and Shleifer, A., 2007. Private credit in 129 countries. *Journal of Financial Economics* 84, 299-329.

Faccio, M., Masulis, R.W. and McConnell, J.J., 2006. Political connections and corporate bailouts. *The Journal of Finance* 61, 2597-2635.

Fama, E.F., and French, K.R., 1992. The cross-section of expected stock returns. *The Journal of Finance* 47, 427-465.

Feng, G., Giglio, S. and Xiu, D., 2020. Taming the factor zoo: A test of new factors. *The Journal of Finance* 75, 1327-1370.

Freyberger, J., Neuhierl, A. and Weber, M., 2020. Dissecting characteristics nonparametrically. *The Review of Financial Studies* 33, 2326-2377.

Friewald, N., Wagner, C. and Zechner, J., 2014. The cross-section of credit risk premia and equity returns. *The Journal of Finance* 69, 2419-2469.

Gao, H., Huang, Y. and Mo, J., 2020. Boosted Credit Ratings in China: The Effects of Credit Enhancement on Bond Pricing. Available at SSRN 3549270.

Ge, Y., and Qiu, J., 2007. Financial development, bank discrimination and trade credit. *Journal of Banking & Finance* 31, 513-530.

Geng, Z., and Pan, J., 2019. Price Discovery and Market Segmentation in China's Credit Market (No. w26575). National Bureau of Economic Research.

Glossner, S., Matos, P., Ramelli, S. and Wagner, A.F., 2020. Where do institutional investors seek shelter when disaster strikes? Evidence from COVID-19. Working paper.

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

Han, S., and Zhou, X., 2014. Informed Bond Trading, Corporate Yield Spreads, and Corporate Default Prediction. *Management Science*, 60, 675-694.

Hu, G.X., Pan, J. and Wang, J., 2020. Chinese capital market: An empirical overview. *Critical Finance Review. Forthcoming*.

Hu, X., Shi, J., Wang, L. and Yu, J., 2020. Foreign ownership in Chinese credit ratings Industry: Information revelation or certification?. *Journal of Banking & Finance*, in press.

Huang, Z., Li, L. Ma, G. and Xu, L., 2017. Hayek, local information and commanding heights: Decentralizing state-owned enterprises in China. *American Economic Review* 107, 2455-2478.

Ivashina, V., and Sun, V., 2010. Institutional stock trading on loan market information. *Journal of Financial Economics* 100, 284-303.

Jiang, X., and Packer, F., 2019. Credit ratings of Chinese firms by domestic and global agencies: Assessing the determinants and impact. *Journal of Banking & Finance* 105, 178-193.

Jin, S., Wang, W. and Zhang, Z., 2018. The Value and Real Effects of Implicit Government Guarantees. Working paper.

Johnson, T.C., 2004. Forecast dispersion and the cross section of expected returns. *The Journal of Finance* 59, 1957-1978.

Keenan, S., and Sobehart, J.R., 1999. Performance measures for credit risk models. Moody's Risk Management Services, 10.

Kim, H.S. and Sohn, S.Y., 2010. Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research* 201, 838-846.

Klapper, L., Laeven, L. and Rajan, R., 2012. Trade credit contracts. *The Review of Financial Studies* 25, 838-867.

Law, D., and Roache, S.K., 2015. Assessing Default Risks for Chinese Firms: A Lost Cause? (No. 15-140). International Monetary Fund.

Lettau, M. and Pelger, M., 2020. Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies* 33, 2274-2325.

Li, B., and Ponticelli, J., 2020. Going bankrupt in China (No. w27501). National Bureau of Economic Research.

Liu, C., Wang, S., Wei, K.J. and Zhong, N., 2019. The demand effect of yield-chasing retail investors: Evidence from the Chinese enterprise bond market. *Journal of Empirical Finance* 50, 57-77.

Livingston, M., Poon, W.P., Zhou, L., 2018. Are Chinese credit ratings relevant? A study of the Chinese bond market and credit rating industry. *Journal of Banking and Finance* 87, 216-232.

McMillan, J. and Woodruff, C., 1999. Interfirm relationships and informal credit in Vietnam. *The Quarterly Journal of Economics* 114, 1285-1320.

Merton, R.C., 1974. On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *The Journal of Finance* 29, 449-470.

Min, J.H. and Lee, Y.C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications* 28, 603-614.

Mselmi, N., Lahiani, A. and Hamza, T., 2017. Financial distress prediction: The case of French small and medium-sized firms. *International Review of Financial Analysis* 50, 67-80.

Mo, J., and Subrahmanyam, M.G., 2020. Policy Interventions and Liquidity in Segmented Chinese Credit Bond Markets. Available at SSRN 3522891.

Moody's, 2014. Government-related issuers, Moody's Investors Service.

NAFMII, 2012. Information disclosure rules for debt financing instruments of non-financial enterprises in the inter-bank bond market. Retrieved from http://114.255.185.134/ggtz/gg/201205/t20120516_15010.html

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18, 109-131.

Olson, D.L., Delen, D. and Meng, Y., 2012. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems* 52, 464-473.

Penman, S.H., Richardson, S.A. and Tuna, I., 2007. The book-to-price effect in stock returns: accounting for leverage. *Journal of Accounting Research* 45, 427-467.

Petersen, M.A., and Rajan, R.G., 1997. Trade credit: theories and evidence. *The Review of Financial Studies* 10, 661-691.

Pike, T., Sapriza, H. and Zimmermann, T., 2019. Bottom-up leading macroeconomic indicators: An application to non-financial corporate defaults using machine learning. Working paper.

Shumway, T., 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business* 74, 101-24.

S&P Global Ratings, 2020. China Corporate Outlook 2020: Steep Walls, Few Catapults (January 16).

Svensson, L. E., 1994. Estimating and interpreting forward interest rates: Sweden 1992-1994. Working paper, National Bureau of Economic Research.

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, 831-868.

Wang, G., Hao, J., Ma, J. and Jiang, H., 2011. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications* 38, 223-230.

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

Zhang, L., Altman, E.I. and Yen, J., 2010. Corporate financial distress diagnosis model and application in credit rating for listing firms in China. *Frontiers of Computer Science in China* 4, 220-236.

Zhu, N., 2016. China's Guaranteed Bubble: How implicit government support has propelled China's economy while creating systemic risk. McGraw-Hill Education.

Ziebart, D.A., and Reiter, S.A., 1992. Bond ratings, bond yields and financial information. *Contemporary Accounting Research* 9, 252–282.

Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101, 1418-1429.

Appendix 1 Variable definitions

Variable	Definition (Data Source)
Yield Spread measures	
Offering spread	Percentage difference between the bond offering yield and yield on a Treasury note of comparable maturity. (Wind)
Offering spread (CDB)	Percentage difference between the bond offering yield and yield on a bond issued by China Development Bank of comparable maturity. (Wind)
Offering spread (Synthetic)	Percentage difference between the bond offering yield and the yield on a synthetic central government bond yield estimated using Ang, Bai and Zhou (2018)'s approach. (Wind)
YTM spread	Percentage difference between the bond yield to maturity at each month end and yield on a Treasury note of comparable maturity. (Wind)
YTM spread (Avg)	Percentage difference between daily trading volume-weighted average yield to maturity in a given month and yield on a Treasury note of comparable maturity. (Wind)
YTM spread (CDB)	Percentage difference between the bond yield to maturity at each month end and yield on a bond issued by China Development Bank of comparable maturity. (Wind)
YTM spread (Synthetic)	Percentage difference between the bond yield to maturity at each month end and the yield on a synthetic central government bond yield estimated using Ang, Bai and Zhou (2018)'s approach. (Wind)
Credit risk measures	
PD	Probability of default estimated from the logit regression model derived by LASSO BIC in Table 3. (Wind)
PD (dummy)	A dummy variable which equals one if *PD* is equal to or higher than the cutoff 0.981% and zero otherwise.
EDP	Expected default probability estimated from Merton's bond pricing model following Bharath and Shumway (2008) as elaborated in OA1. (CSMAR)
Z_{China}	Modified Z score for Chinese firms specified in Zhang, Altman and Yan (2010). (Wind)
Distress risk	Financial distress risk estimated using the prediction model in Asis, Chari and Haas (2020). (CSMAR, World Bank, Bloomberg, BIS, FRED, Fed, COBE)
Bond control variables	
AAA	A dummy variable which equals one if the bond rating is AAA and zero otherwise. (Wind)
AA+	A dummy variable which equals one if the bond rating is AA+ and zero otherwise. (Wind)
AA	A dummy variable which equals one if the bond rating is AA and zero otherwise. (Wind)
AA-	A dummy variable which equals one if the bond rating is AA- and zero otherwise. (Wind)
Below AA-	A dummy variable which equals one if the bond rating is below AA- and zero otherwise. (Wind)
Maturity	Number of years to the maturity date of a particular debt issue. (Wind)
ISize	Outstanding debt amount in CNY billions. (Wind)
Puttable	A dummy variable which equals one for a puttable bond and zero otherwise. (Wind)
Callable	A dummy variable which equals one for a callable bond and zero otherwise. (Wind)
Sinking fund	A dummy variable which equals one if the bond has sinking fund and zero otherwise. (Wind)
Cross	A dummy variable which equals one if the bond trades in both interbank and exchange-based market, and zero otherwise. (Wind)
MTN	A dummy variable which equals one for an MTN bond and zero otherwise. (Wind)
Enterprise	A dummy variable which equals one for an enterprise bond and zero otherwise. (Wind)
Corporate	A dummy variable which equals one for a corporate bond and zero otherwise. (Wind)
Issuer control variables	
Public	A dummy variable which equals one for a publicly listed issuer and zero otherwise. (Wind)
Age	Age of the issuer in years. (Wind)
Leverage	Total liabilities scaled by total assets. (Wind)
Tangibility	Property, plant and equipment scaled by average total assets. (Wind)
Sales	Natural log of sales in CNY billions. (Wind)
ROA	Operating income scaled by total assets. (Wind)
Growth	Change in operating revenues from the previous year. (Wind)
Cash	Cash and cash equivalents scaled by current liabilities. (Wind)
Other variables	
Exchange	A dummy variable which equals one if the bond is traded on the exchange market and zero otherwise. (Wind)
Post 2014	A dummy variable which equals one if the observation is after March 2014 and zero otherwise. (Wind)
Post Pledgeability	A dummy variable which equals one if the observation is after December 2014 and zero otherwise. (Wind)
Post Connect	A dummy variable which equals one if the bond is issued after July 2017 and zero otherwise. (Wind)
Post COVID	A dummy variable which equals one if the observation is after 23rd January 2020, and zero otherwise

Appendix 2 Asset pricing models for bonds and stocks

Panel A: Common risk factors for bonds and stocks

Risk Factor	Factor notation	Definition	Data source
Stock market excess return	MKT^{Stock}	Difference between A-share stock market return and one-year fixed term deposit rate.	CSMAR
Size factor	SMB	Value-weighted average return difference between small stock portfolio and large stock portfolio across the book-to-market portfolios.	CSMAR
Value factor	HML	Value-weighted average return difference between high book-to-market stock portfolio and low book-to-market stock portfolio across the size portfolios.	CSMAR
Default risk factor	DEF	Difference between monthly long-term credit bond return (CSI Credit Bond Index) and long-term Treasury bond return (CSI Treasury Bond Index).	Wind
Term risk factor	TERM	Difference between monthly long-term Treasury bond return (CSI Treasury Bond Index) and one-year deposit rate.	Wind
Bond market excess return	MKT^{Bond}	Value-weighted average return of all corporate bonds in our sample in excess of one-year deposit rate.	Wind
Downside risk factor	DRF	We first form bivariate portfolios by independently sorting bonds into three rating categories (i.e., AAA, AA+ and others) and five quintiles based on their 5% VaR in the last 24 months with a minimum of six-month data. DRF is the value-weighted average return difference between the highest-VaR portfolio and the lowest-VaR portfolio across the rating portfolios.	Wind
Liquidity risk factor	LRF	We form bivariate portfolios by independently sorting bonds into three rating categories (i.e., AAA, AA+ and others) and five quintiles based on their illiquidity measure. LRF is the value-weighted average return difference between the highest-illiquidity and the lowest-illiquidity portfolios across the rating portfolios. The monthly illiquidity measure (*ILLIQ*) is constructed following Bao, Pan and Wang (2011) using daily bond trading data; it is computed as $-Cov_t(\Delta p_{it}, \Delta p_{it+1})$ for each month, where Δp_{it} is the change in transaction price for day *t* of bond *i*.	Wind
Return reversal factor	REV	We form bivariate portfolios by independently sorting bonds into three rating categories (i.e., AAA, AA+ and others) and five quintiles based on their previous month bond returns. REV is the value-weighted average return difference between short-term loser (i.e., lowest past bond returns) and the short-term winner (i.e.., highest past bond returns) portfolio across the rating portfolios.	Wind
Credit risk factor	CRF	Average of CRF_{VaR}, CRF_{ILLIQ} and $CRF_{REV,}$ where CRF_{VaR}, $CRF_{ILLIQ,}$ and CRF_{REV} are the value-weighted average return difference between the lowest-rating portfolio and the highest-rating portfolio across the VaR portfolios, bond illiquidity portfolios, and short-term return reversal portfolios, respectively.	Wind

Panel B: Asset pricing models for corporate bonds

CAPM	$Ret_t = \alpha_t + \beta_1 MKT_t^{Stock} + \varepsilon_t$
Three factor model	$Ret_t = \alpha_t + \beta_1 MKT_t^{Stock} + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_t$
Five-factor model	$Ret_t = \alpha_t + \beta_1 MKT_t^{Stock} + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 DEF_t + \beta_5 TERM_t + \varepsilon_t$
Seven-factor model	$Ret_t = \alpha_t + \beta_1 MKT_t^{Bond} + \beta_2 DEF_t + \beta_3 TERM_t + \beta_4 DRF_t + \beta_5 LRF_t + \beta_6 REV_t + \beta_7 CRF_t + \varepsilon_t$
Ten-factor model	$Ret_t = \alpha_t + \beta_1 MKT_t^{Stock} + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 MKT_t^{Bond} + \beta_5 DEF_t + \beta_6 TERM_t + \beta_7 DRF_t + \beta_8 LRF_t + \beta_9 REV_t + \beta_{10} CRF_t + \varepsilon_t$

Panel C: Asset pricing models for stocks

CAPM	$Ret_t = \alpha_t + \beta_1 MKT_t^{Stock} + \varepsilon_t$
Three-factor model	$Ret_t = \alpha_t + \beta_1 MKT_t^{Stock} + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_t$

Appendix 3 Machine learning methodology

In this Appendix, we briefly introduce the four MLA techniques we used in this study.

We assume that the probability of a firm default event between time t and $t + h$ takes the functional form:

$$P\left(D_{i,t,t+h} = 1\right) = f(X_{i,t}; \theta) \tag{A1}$$

where $D_{i,t,t+h}$ is the indicator function that is 1 if firm i defaults between time t and $t + h$, f is a flexible function of firm i's M-dimensional characteristics $X_{i,t}$, and $\theta = (\theta_1, \cdots, \theta_M)'$ is a vector of parameters that needs to be estimated. We index firms by $i = 1, \cdots, N$ and quarters by $t = 1, \cdots, T$, where N is the number of firms at time t.

A.1 Linear Regression

The linear prediction regression assumes that f can be estimated by a linear function as:

$$f(X_{i,t}; \theta) = X'_{i,t}\theta \tag{A2}$$

where θ can be estimated by the ordinary least squares (OLS) via the optimization problem:

$$\min_{\theta} L(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{N} (D_{i,t,t+h} - f(X_{i,t}; \theta))^2 \tag{A3}$$

Based on Wooldridge (2001), the estimate of θ in Eq. A3 is unbiased and efficient if the number of predictors (M) is relatively small, while T is relatively large.

A.2 Penalized Linear Regression: LASSO

The OLS model has the issue of overfitting by including all candidate variables when the number of predictors is large. LASSO is one of the most widely used methods to reduce this overfitting issue by adding a penalty term to the objective function. The penalty is imposed to tradeoff between in-sample performance reduction and out-of-sample stability improvement. The θ in LASSO is estimated via:

$$\min_{\theta} L(\theta; .) \equiv L(\theta) + \phi(\theta; .) \tag{A4}$$

where $\phi(\theta;.)$ is the penalty function. The estimates of some variables of θ can be regularized and shrunk towards zero based on the penalty function. A general penalty function as follows is widely used in machine learning literature:

$$\phi(\theta;\lambda,\rho) = \lambda(1-\rho)\sum_{j=1}^{M}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{M}\theta_j^2 \tag{A5}$$

where $\lambda > 0$ is a hyperparameter controlling for the amount of shrinkage; the larger the value of λ, the greater the amount of shrinkage. The estimation model in Eq. A5 reduces to the standard OLS if $\lambda = 0$. When $\rho = 0$, Eq. A5 corresponds to LASSO, which sets a subset of θ to exactly zero. In this sense, the LASSO is a sparsity modelling technique and can be used for a variable selection. To decide the value of λ, cross-validation, AIC and BIC criteria are used. Adaptive LASSO (Zou, 2006) adds weights to $\phi(\theta;.)$.

A.3 Support vector machines (SVMs)

The support vector machines (SVMs) are classification techniques based on statistical learning theory. SVMs produces a binary classifier, so-called optimal separating hyperplanes, through an extremely non-linear mapping of the input vectors into the high-dimensional feature space. SVMs constructs a linear model to estimate the decision function using non-linear class boundaries based on support vectors. If the data are linearly separated, SVM trains linear machines for an optimal hyperplane that separates the data without error and into the maximum distance between the hyperplane and the closest training points. The training points that are closest to the optimal separating hyperplane are called support vectors.

The optimisation problem of SVMs is to minimize the following:

$$\frac{1}{2}w'w + C\sum_{i=1}^{N}\xi_i \tag{A6}$$

subject to

$$d_i[w'\phi(X_i) + b] \geq 1 - \xi_i \tag{A7}$$

where $\xi_i \geq 0$ are the margins of error related to classification cost C, d_i are the classifications in the training set and $\phi(X)$ transforms space $\mathbb{R}^N$. $\phi(X)$ does not need to be known, as a kernel function is applied so that $K(X) = \phi(X_i)'\phi(X_j)$. The kernel function is predetermined. The traditional kernel functions are:

$$K(X_i, X_j) =< X_i, X_j > \quad \text{(A8)}$$

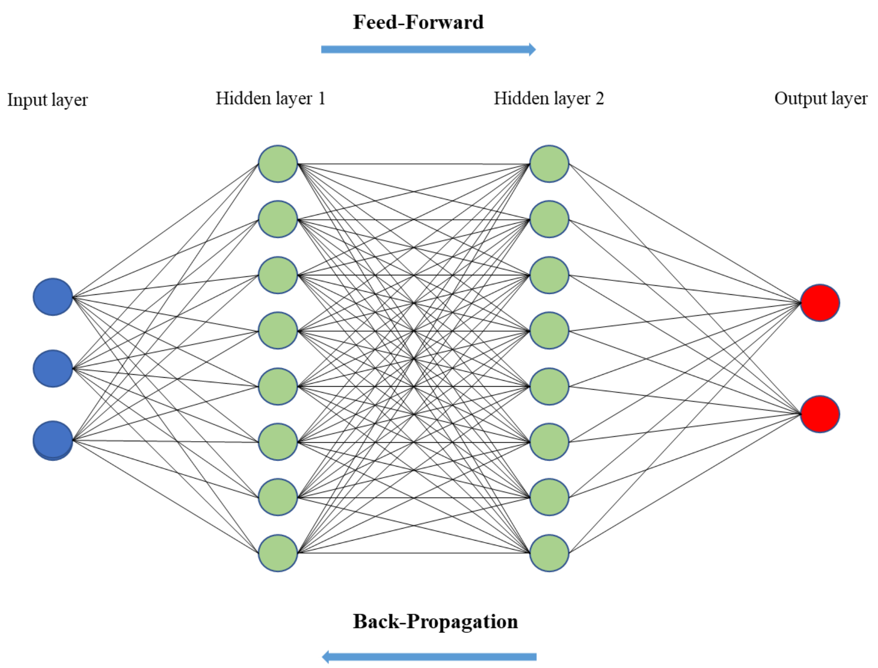
$$K(X_i, X_j) = \exp(-\gamma\|X_i - X_j\|^2) \quad \text{(A9)}$$

where γ is a positive constant. Eq. A8 is called the linear kernel and Eq. A9 is the radial basis function (RBF). Based on this method, our study defines the default problem as a non-linear problem and uses the RBF kernel to optimize the hyperplane.

A.4 Artificial neural networks: back-propagation networks (BPNs)

Back-propagation is the essence of neural net training. It is the method of fine-tuning the weights of a neural net based on the error rate obtained in the previous iteration. Proper tuning of the weights allows users to reduce error rates and to make the model reliable by increasing its generalization. The network, as shown in Figure A1, includes an input layer of raw candidate predictors, one or more hidden layers that interact and nonlinearly transform the candidate predictors, and an output layer that generates outcome prediction.

Figure A1: Back-propagation networks with two hidden layers



The back-propagation process trains the model in the following steps:

- Inputs arrive through the preconnected path.
- Inputs are modelled using real weights. The weights are usually randomly selected.
- Calculate the output for every neuron from the input layer to the hidden layers, to the output layer with transformations.
- Calculate the error in the outputs.
- Travel back from the output layer to the hidden layers to adjust the weights such that the error is decreased.
- Keep repeating the process until the desired output is achieved.

The model becomes more flexible by adding hidden layers between the inputs and outputs. In this paper, we apply 20 hidden layers for BPNs. The results are qualitatively unchanged by applying 10 layers instead.

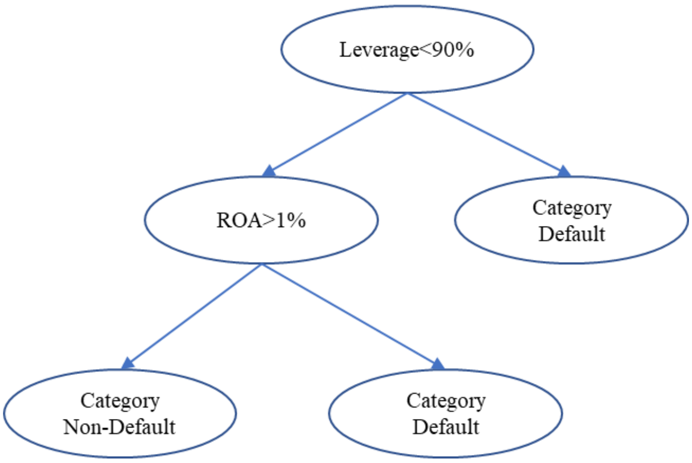
A.4 Random Forests

Random forests method is augmentations of the simple decision tree. Unlike linear models, decision trees are fully nonparametric. A single tree "grows" in a sequence of steps. At each step, a new "branch" separates the data based on one of the default predictors. A single tree starts with all observations, the method finds the variable and associated threshold value that best splits the observations in two groups, as measured by some objective loss function. Each group is then split again into two groups by the same processes. However, the splitting predictor variable does not have to be the same. This process continues until each group has limited number of observations in it, or until some predefined number of splits is achieved. Random forests apply bagging which repeatedly selects a random sample with replacement of the training set and fits trees to these samples. Through averaging the bagging outcomes, random forests can reduce the overfit in individual bootstrap samples and make the predictive performance more stable.

Panel A of Figure A2 shows an example of a single tree with two predictors, *Leverage* and *ROA*. The tree separates to a partition based on characteristic values. First, observations are sorted by *Leverage*. Those above the breakpoint of 90% are assigned to Category Default. Those with lower *Leverage* then further sorted by *ROA*. Firms with lower than 1% *ROA* are assigned to Category Default, while the rest go into Category Non-Default. Panel B of Figure A2 demonstrate the framework of random forests.

Figure A2: Decision tree example

Panel A: A single classification tree



Panel B: Random forests

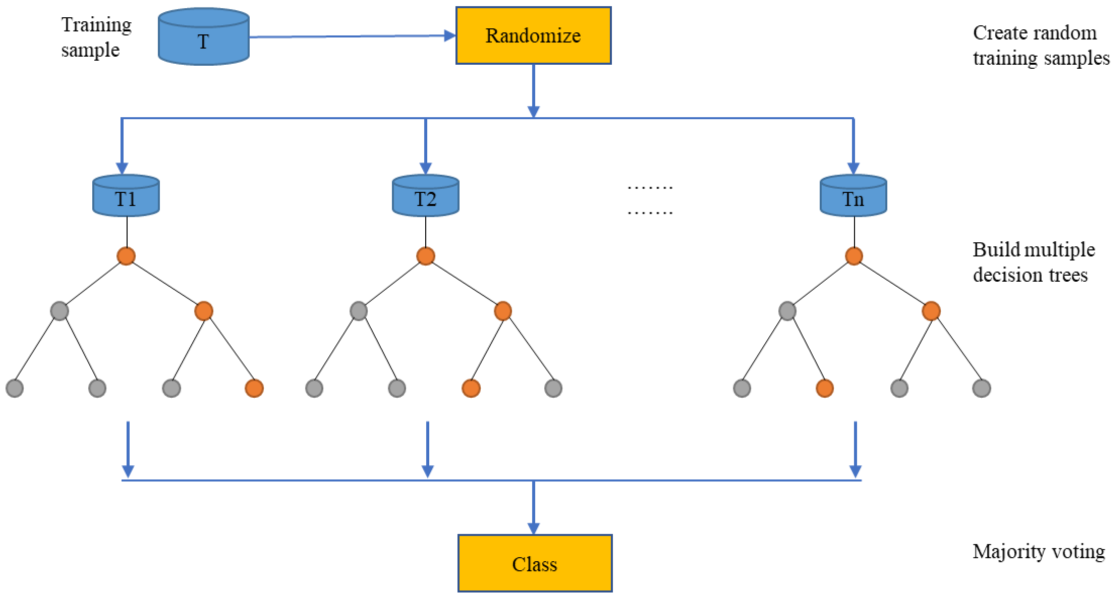


Table 1. Distribution of corporate defaults over time

This table summarizes the number of defaulted (Defaulted) and non-defaulted (Non-defaulted) corporate bond issuers and the total number of corporate bond issuers every quarter from 2014 to May 2020.

Panel A: Number of default firms			
Year of default	**Defaulted**	**Non-defaulted**	**Total**
2014Q1	1	57	58
2014Q2	1	57	58
2015Q2	5	286	291
2015Q3	2	114	116
2015Q4	7	400	407
2016Q1	8	457	465
2016Q2	6	343	349
2016Q3	2	114	116
2016Q4	6	343	349
2017Q1	1	57	58
2017Q2	2	114	116
2017Q3	3	171	174
2017Q4	5	286	291
2018Q1	1	57	58
2018Q2	8	457	465
2018Q3	16	915	931
2018Q4	16	915	931
2019Q1	11	629	640
2019Q2	7	400	407
2019Q3	9	514	523
2019Q4	11	629	640
2020Q1	6	343	349
2020Q2	3	171	174
Total	**137**	**7,829**	**7,966**

Table 2. Variable selection

This table lists out the candidate financial and nonfinancial variables considered to construct the corporate default prediction model for Chinese companies by accounting ratio category. Column 1 indicates the variable notations. Column 2 describes the variables. Columns 3-5 report the variables selected and their estimated signs in predicting default event by difference methodologies, including LASSO using BIC criteria, a forward stepwise selection procedure and MDA. The sample period is from 2014 to 2018.

Variables	Description	LASSO BIC	Stepwise forward	MDA
(1)	(2)	(3)	(4)	(5)
Financial variables				
Leverage				
X11	Book equity/Total liabilities			√ (+)
X12	Total liabilities/Total assets	√ (+)	√ (-)	√ (+)
X13	Interest bearing debt/Total assets	√ (+)	√ (+)	
Liquidity				
X21	Cash/Total assets			√ (-)
X22	Cash and short-term investment/Total assets	√ (-)	√ (-)	
X23	Current assets/Current liabilities		√ (-)	
X24	(Current assets – Inventory)/Current liabilities		√ (+)	
X25	Cash and short-term investment/Current liabilities			
Profitability				
X31	EBIT/Sales			
X32	EBITDA/Sales			
X33	Net income/Total assets	√ (-)	√ (-)	√ (-)
X34	Net income/Sales			
X35	Retained earnings/Total assets			
Coverage				
X41	Interest expenses/ EBITDA			√ (-)
X42	Interest expenses/ EBIT			
X43	Debt/EBITDA			
X44	Debt/EBIT			
Activity				
X51	Working capital/Total assets		√ (-)	
X52	Accounts payable/Sales			
X53	Accounts receivable/Sales			
X54	Inventory/Operating expense			
Structure				
X61	Current assets/Total assets		√ (+)	
X62	Fixed assets/Total assets			
X63	Current liabilities/Total liabilities	√ (+)	√ (-)	
Nonfinancial variables				
SOE	A dummy variable which equals one if the firm is owned by the government, and zero otherwise	√ (-)	√ (-)	√ (-)
Local GDP	GDP growth rate of the province where company is headquartered	√ (-)	√ (-)	√ (-)

Table 3. Corporate default prediction model

This table reports the summary statistics for the selected variables through LASSO BIC in our model in Panel A, final multivariate logit regression model designed for predicting corporate default in Panel B. ***, ** and * indicate statistical significance of the *t*-tests for mean differences and the Wilcoxon tests for median differences at the 1%, 5%, and 10% levels, respectively. The estimation sample period is from 2014 to 2018.

Panel A: Summary statistics for selected variables

Variable	**Mean**		**Median**		**Mean diff**	**Median diff**
	Default	Non-default	Default	Non-default		
(1)	(2)	(3)	(4)	(5)	(2)-(3)	(4)-(5)
PD	14.302	1.500	7.942	0.315	12.802***	7.627***
Total liabilities/Total assets	0.690	0.575	0.646	0.585	0.115***	0.061***
Interest bearing debt/Total assets	0.407	0.336	0.377	0.326	0.071	0.051***
Cash and short-term investment/Total assets	0.099	0.111	0.075	0.098	-0.012***	-0.023***
Net income/Total assets	-0.016	0.017	0.011	0.011	-0.033***	0.000
Current liabilities/Total liabilities	0.653	0.517	0.663	0.513	0.136***	0.150***
SOE	0.189	0.799	0.000	1.000	-0.610***	-1.000***
Local GDP	6.949	7.569	7.300	7.700	-0.620***	-0.400***

Panel B: Logit model with raw value predictors

Log (PD/1-PD) =		
	-2.617	
	0.818	Total liabilities/Total assets
	3.373	Interest bearing debt/Total assets
	-5.275	Cash and short-term investment/Total assets
	-11.986	Net income/Total assets
	1.210	Current liabilities/Total liabilities
	-3.053	SOE
	-0.237	Local GDP
Pseudo R-squared	27.70%	
Wald test	Chi2=173.61***	
Hosmer-Lemeshow test	*p*-value=29.02%	
In sample AUC	89.65%	
95% Confidence Interval	[86.45%, 92.86%]	
Leave-one-out AUC	89.70%	

Table 4 *PD* model prediction accuracy

Panel A reports the in-sample AUC and hold-out sample AUC values, Types I and II error rates as well as accuracy rate across different default risk prediction models. Note that EDP and Asis, Chari and Haas (2020) default risk measures can only be estimated for the publicly listed companies. The sample period for estimation sample is from 2014 to 2018 and from 2019 to May 2020 for hold-out sample. Panel B demonstrates our *PD* model's prediction accuracy for public and private companies on the estimation sample and the hold-out sample. Our *PD* prediction model is shown in Table 3 Panel B.

Panel A: Misclassification and accuracy rates across models

Model number	**Model name**	**In-sample AUC**	**Hold-out sample AUC**	**Type I error**	**Type II error**	**Accuracy**
1	Logit model derived from LASSO BIC	89.65%	90.11%	14.89%	21.41%	78.70%
2	SVMs	99.35%	72.06%	31.91%	37.30%	62.79%
3	ANNs	89.62%	88.51%	29.79%	14.97%	84.78%
4	Random Forest	100%	90.71%	100.00%	0.00%	98.28%
5	Z_{China} in Zhang, Altman and Yen (2010)	52.00%	63.93%	51.06%	23.86%	75.67%
6	Credit Rating	75.95%	78.00%	36.96%	9.15%	90.37%
7	EDP	65.53%	83.15%	22.22%	27.19%	72.99%
8	Asis, Chari and Haas (2020) distress risk	83.84%	77.48%	38.89%	23.13%	76.29%

Panel B: Model predictive accuracy on various subsamples

	# of Defaults (1)	**Correct prediction** (2)	**Incorrect prediction** (3)	**Prediction accuracy** (4)	**AUC** (5)
Estimation sample					
Public companies	11	11	0	100.00%	95.82%
Private companies	79	71	8	89.87%	90.55%
Hold-out sample					
Public companies	18	15	3	83.33%	81.79%
Private companies	29	25	4	86.21%	91.16%

Table 5. Yield spread and corporate default risk

This table reports the OLS regression results of *Offering spread* and *YTM spread* on *PD*, along with other bond and issuer characteristic controls for full sample and subsamples in Panel A and Panel B, respectively. The dependent variable is bond offering yield spread (*Offering spread*) for columns 1-5 in Panel A and columns 1-4 in Panel B, which is the percentage difference between the issue's offering yield and the yield on a Treasury bond of comparable maturity. The dependent variable in columns 6-10 in Panel A and columns 5-8 in Panel B is the trading yield spread in the secondary market, *YTM spread*, which is the percentage difference between the bond yield to maturity at month end and yield on a Treasury bond of comparable maturity. *PD* is the probability of default measure estimated from our logit regression model derived using LASSO BIC (i.e., Panel B Table 3). The bond control variables include bond credit rating dummies (*AA+*, *AA*, *AA-* and *Below AA-*), *Maturity, ISize, Puttable, Callable, Sinking fund, Cross, Corporate* and *Enterprise*, while the issuing firm control variables are *Public, Age, Leverage, Tangibility, Sales, Growth, ROA*, and *Cash*. Industry and year fixed effects are also included. In columns 5 and 10, we replace *PD* with a dummy variable, *PD (dummy)*, which equals one if *PD* is equal to or higher than 1.822% and zero otherwise. In Panel B, columns 1-2 and 5-6 report the results for public and private firms, respectively. Columns 3-4 and columns 7-8 report the results for SOEs and non-SOEs, respectively. All variables are defined in Appendix 1. *YTM spread* and the continuous issuer-specific financial variables are winsorized at the top and bottom 1% of the sample distribution. The robust standard errors (reported in parentheses) are clustered at the firm level in columns 1-5 in Panel A and columns 1-4 in Panel B, and at the bond level in columns 6-10 in Panel A and columns 5-8 in Panel B. ***, ** and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A: Full sample regressions

	Primary Market Dep.var= *Offering spread*					**Secondary Market Dep.var= *YTM spread***				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
PD	0.108***	0.105***	0.092***	0.090***		0.112***	0.111***	0.091***	0.091***	
	(7.876)	(7.683)	(8.662)	(8.508)		(21.328)	(21.019)	(19.516)	(19.334)	
PD (dummy)					0.614***					0.718***
					(13.630)					(27.832)
AA+			0.659***	0.655***	0.673***			0.611***	0.615***	0.626***
			(20.595)	(20.375)	(20.862)			(26.686)	(26.875)	(27.489)
AA			1.202***	1.197***	1.180***			1.060***	1.066***	1.040***
			(36.318)	(36.092)	(36.116)			(39.836)	(40.194)	(39.400)
AA-			2.000***	1.989***	1.951***			2.139***	2.134***	2.125***
			(26.671)	(26.401)	(27.099)			(23.208)	(23.408)	(23.801)
Below AA-			2.307***	2.266***	2.178***			3.738***	3.721***	4.564***
			(27.277)	(26.020)	(25.006)			(21.360)	(21.201)	(21.412)
Maturity	-0.075***	-0.073***	-0.034***	-0.033***	-0.034***	0.003	0.003	0.027***	0.027***	0.029***
	(-10.775)	(-10.576)	(-6.070)	(-6.001)	(-6.152)	(0.661)	(0.804)	(7.231)	(7.162)	(7.640)
ISize	-0.090***	-0.086***	-0.043***	-0.041***	-0.033***	-0.043***	-0.045***	-0.026***	-0.029***	-0.021***
	(-4.946)	(-4.860)	(-4.603)	(-4.504)	(-3.683)	(-7.105)	(-7.364)	(-6.081)	(-6.597)	(-4.656)
Puttable	0.195***	0.192***	0.038	0.038	0.044	0.084**	0.087**	-0.027	-0.027	-0.048
	(5.846)	(5.797)	(1.362)	(1.376)	(1.616)	(2.428)	(2.529)	(-0.865)	(-0.862)	(-1.524)
Callable	0.459***	0.458***	0.534***	0.531***	0.514***	2.472***	2.468***	2.514***	2.514***	2.479***
	(11.599)	(11.707)	(14.828)	(14.928)	(14.632)	(36.576)	(36.482)	(37.102)	(37.047)	(37.322)
Sinking fund	0.574***	0.574***	0.457***	0.462***	0.464***	0.098**	0.086*	-0.023	-0.032	-0.035
	(11.781)	(11.841)	(11.148)	(11.199)	(10.827)	(2.158)	(1.909)	(-0.528)	(-0.739)	(-0.816)
Cross	-0.075*	-0.074*	-0.020	-0.018	-0.035	-0.036	-0.031	-0.035	-0.034	-0.051
	(-1.708)	(-1.677)	(-0.502)	(-0.456)	(-0.892)	(-1.006)	(-0.867)	(-1.059)	(-1.009)	(-1.556)

Enterprise	0.262***	0.267***	0.323***	0.325***	0.338***	0.226***	0.212***	0.273***	0.266***	0.289***
	(4.947)	(5.034)	(7.035)	(7.072)	(7.286)	(5.233)	(4.928)	(6.938)	(6.758)	(7.443)
Corporate	-0.071**	-0.076**	0.064**	0.062**	0.042	0.171***	0.180***	0.195***	0.205***	0.185***
	(-2.260)	(-2.434)	(2.373)	(2.326)	(1.561)	(3.976)	(4.176)	(5.084)	(5.346)	(4.731)
Public	-0.130***	-0.146***	-0.079**	-0.087***	-0.001	-0.157***	-0.150***	-0.119***	-0.106***	-0.030
	(-3.190)	(-3.516)	(-2.419)	(-2.588)	(-0.023)	(-4.552)	(-4.225)	(-3.808)	(-3.336)	(-0.943)
Age	-0.004**	-0.005***	-0.004***	-0.004***	-0.004**	-0.007***	-0.007***	-0.005***	-0.005***	-0.005***
	(-2.486)	(-2.582)	(-2.784)	(-2.769)	(-2.454)	(-6.647)	(-6.625)	(-5.640)	(-5.610)	(-5.047)
Leverage		0.121		-0.003	-0.038		-0.012		-0.187**	-0.280***
		(1.087)		(-0.036)	(-0.408)		(-0.119)		(-1.995)	(-3.030)
Tangibility	-0.086	-0.164***	-0.142**	-0.037***	-0.047***	-0.254***	-0.182**	-0.243***	-0.289***	-0.346***
	(-1.136)	(-15.309)	(-2.270)	(-3.973)	(-4.785)	(-4.423)	(-2.386)	(-4.796)	(-4.057)	(-5.044)
Sales	-0.153***	0.019	-0.032***	0.027*	0.037***	-0.168***	-0.167***	-0.037***	-0.031***	-0.047***
	(-14.832)	(1.066)	(-3.499)	(1.900)	(2.768)	(-22.526)	(-21.918)	(-5.247)	(-4.286)	(-6.317)
Growth	0.017	0.349**	0.026*	0.246**	0.144	0.047***	0.056***	0.035***	0.041***	0.053***
	(0.951)	(2.528)	(1.858)	(2.082)	(1.232)	(4.190)	(4.977)	(3.343)	(3.990)	(5.196)
ROA		0.944**		0.384	-0.977**		-0.964***		-1.151***	-2.770***
		(2.106)		(1.037)	(-2.473)		(-2.658)		(-3.607)	(-8.718)
Cash		-0.094***		-0.032	-0.057***		-0.101***		-0.041**	-0.093***
		(-4.019)		(-1.585)	(-2.796)		(-5.748)		(-2.573)	(-5.968)
Constant	2.615***	2.281***	1.692***	1.485***	1.660***	1.249***	1.361***	0.770***	1.030***	1.279***
	(9.096)	(6.922)	(7.969)	(6.234)	(7.183)	(6.718)	(6.678)	(5.151)	(6.321)	(7.499)
Observations	17,450	17,450	17,450	17,450	17,450	228,248	228,248	228,248	228,248	228,248
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.416	0.418	0.535	0.535	0.535	0.358	0.359	0.435	0.436	0.431

Table 5 (Cont.)

Panel B: Subsample regressions								
	Primary Market Dep.var= ***Offering spread***				**Secondary Market Dep.var=** ***YTM spread***			
	Public	Private	SOE	Non-SOE	Public	Private	SOE	Non-SOE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PD	0.078***	0.094***	0.098**	0.024***	0.069***	0.096***	0.033***	0.042***
	(5.649)	(6.317)	(2.286)	(3.483)	(12.631)	(13.704)	(4.999)	(8.926)
Constant	1.148***	1.646***	1.757***	1.451***	0.788**	2.156***	1.442***	2.029***
	(3.586)	(12.200)	(8.272)	(3.031)	(2.490)	(17.683)	(8.930)	(4.318)
Observations	3,110	14,340	14,602	2,848	48,461	179,787	187,995	40,253
Bond controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.571	0.534	0.571	0.460	0.432	0.467	0.502	0.334

Table 6. Bond market shocks: Implicit bailout and asset pledgeability policy

This table reports the OLS regression results of *Offering spread* and *YTM spread* on *PD* around two bond market events, along with other bond and issuer characteristic controls. The dependent variable is bond offering yield spread (*Offering spread*) in columns 1-2, which is the percentage difference between the bond offering yield and the yield on a Treasury bond of comparable maturity. The dependent variable in columns 3-4 is the bond yield spread in the secondary market, *YTM spread,* which is the percentage difference between the bond yield to maturity at month end and the yield on a Treasury bond of comparable maturity. *PD* is the probability of default measure estimated from our logit regression model derived using LASSO BIC (i.e., Panel B in Table 3). The bond control variables include bond credit rating dummies (*AA+*, *AA*, *AA-* and *Below AA-*), *Maturity*, *ISize*, *Puttable*, *Callable*, *Sinking fund*, *Cross*, *Corporate* and *Enterprise*, while the issuing firm control variables are *Public*, *Age*, *Leverage*, *Tangibility*, *Sales*, *Growth*, *ROA*, and *Cash*. *Post 2014* is a dummy variable which equals one if the observation is after March 2014 and zero otherwise. *Post Pledgeability* is a dummy variable which equals one if the observation is after December 2014 and zero otherwise. Industry fixed effects are also included. All variables are defined in Appendix 1. *YTM spread* and the continuous issuer-specific financial variables are winsorized at the top and bottom 1% of the sample distribution. The robust standard errors (reported in parentheses) are clustered at the firm level in columns 1-2 and at the bond level in columns 3-4. ***, ** and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Primary Market Dep.var = *Offering spread*		**Secondary Market Dep.var = *YTM spread***	
	First Default	Pledgeability Reform	First Default	Pledgeability Reform
	(1)	(2)	(3)	(4)
PD	0.015	0.063***	0.009*	0.027***
	(1.213)	(4.043)	(1.660)	(3.888)
Post 2014	-0.380***		-0.344***	
	(-17.243)		(-19.636)	
Post 2014 × PD	0.074***		0.091***	
	(4.755)		(11.794)	
Post Pledgeability		-0.519***		-0.590***
		(-17.693)		(-23.169)
Post Pledgeability × PD		0.032**		0.073***
		(2.005)		(8.095)
Constant	1.756***	2.167***	1.805***	3.034***
	(15.812)	(19.138)	(20.469)	(20.313)
Observations	17,450	7,651	228,248	100,235
Bond controls	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	No	No	No	No
Adjusted R-squared	0.455	0.488	0.401	0.311

Table 7. Investor heterogeneity

This table reports the OLS regression results of *Offering spread* and *YTM spread* on *PD* taking into account investor heterogeneity, along with other bond and issuer characteristic controls. The dependent variable is bond offering yield spread (*Offering spread*) in columns 1-2, which is the percentage difference between the bond offering yield and the yield on a Treasury bond of comparable maturity. The dependent variable in columns 3-4 is the bond yield spread in the secondary market, *YTM spread,* which is the percentage difference between the bond yield to maturity at month end and yield on a Treasury bond of comparable maturity. *PD* is the probability of default measure estimated from the logit regression model derived using LASSO BIC (i.e., Panel B in Table 3). The bond control variables include bond credit rating dummies (*AA+*, *AA*, *AA-* and *Below AA-*), *Maturity*, *ISize, Puttable, Callable, Sinking fund, Cross*, *Corporate* and *Enterprise*, while the issuing firm control variables are *Public*, *Age*, *Leverage*, *Tangibility*, *Sales*, *Growth*, *ROA*, and *Cash*. *Exchange* is a dummy variable which equals one if the bond is traded on an exchange market and zero otherwise. *Post Connect* is a dummy variable which equals one if the observation is after July 2017 and zero otherwise. Industry and year fixed effects are also included. All variables are defined in Appendix 1. *YTM spread* and the continuous issuer-specific financial variables are winsorized at the top and bottom 1% of the sample distribution. The robust standard errors (reported in parentheses) are clustered at the firm level in columns 1-2 and at the bond level in columns 3-4. ***, ** and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Primary Market Dep.var = *Offering spread*		**Secondary Market Dep.var =*YTM spread***	
	Market Seg	Bond Connect	Market Seg	Bond Connect
	(1)	(2)	(3)	(4)
PD	0.080***	0.054***	0.086***	0.051***
	(5.790)	(5.978)	(80.340)	(13.613)
Exchange	0.035		0.187***	
	(1.245)		(20.712)	
Exchange × PD	0.018**		0.008***	
	(2.184)		(5.770)	
Post Connect		0.189***		0.116***
		(7.459)		(5.997)
Post Connect × PD		0.055***		0.083***
		(3.761)		(9.121)
Constant	1.504***	1.244***	1.039**	1.434***
	(7.869)	(11.189)	(2.268)	(16.264)
Observations	17,450	17,450	228,248	228,248
Bond controls	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	Yes	No	Yes	No
Adjusted R-squared	0.536	0.455	0.436	0.406

Table 8. Bond portfolios sorted by corporate default risk

Quintile portfolios are formed every month from October 2008 to May 2020 by sorting bonds based on one-quarter ahead *PD* measure, which is estimated from our logit regression model derived using LASSO BIC (i.e.., Panel B in Table 3). Quintile 1 is the portfolio of bonds with the lowest *PD* and Quintile 5 consists of bonds with the highest *PD*. The portfolios in Panel A are equal weighted while the portfolios in Panel B are value weighted using debt outstanding as weights. We present the portfolio returns (equally weighted and value weighted) surrounding the first bond default in March 2014 in Panel C, the asset pledgeability reform for exchange-traded bonds in Panel D, and the establishment of the *Bond Connect* program in Panel E. Panels A and B report the average *PD*, average monthly excess return in the next 12 months over one-year fixed term deposit rate, CAPM alpha, three-factor alpha, five-factor alpha, seven-factor alpha and ten-factor alpha for each quintile portfolio. The last row of the two panels shows the differences in *PD*, excess return and alphas. All the asset pricing models for corporate bonds are described in detail in Appendix 2 Panel B. Monthly excess returns and alphas are expressed in percentage. In Panels C-D, we only report the average *PD* value and excess return for each quintile portfolio. Newey-West adjusted *t*-statistics are given in parentheses.

Panel A: Equal-weighted portfolio returns

	PD	**Excess return**	**CAPM alpha**	**Three-factor alpha**	**Five-factor alpha**	**Seven-factor alpha**	**Ten-factor alpha**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Low	0.077	0.216	0.216	0.203	0.195	0.170	0.153
		(5.039)	(5.038)	(4.861)	(5.215)	(3.373)	(3.268)
2	0.162	0.194	0.194	0.184	0.175	0.149	0.134
		(4.945)	(4.929)	(4.695)	(5.130)	(3.472)	(3.222)
3	0.287	0.208	0.209	0.203	0.196	0.170	0.163
		(5.689)	(5.705)	(5.633)	(6.023)	(3.892)	(3.805)
4	0.695	0.241	0.242	0.239	0.232	0.200	0.195
		(8.152)	(8.211)	(8.032)	(8.947)	(5.921)	(5.635)
High	4.969	0.294	0.295	0.291	0.285	0.272	0.268
		(9.486)	(9.576)	(8.905)	(9.512)	(8.389)	(7.656)
High-Low	4.892	0.077	0.079	0.088	0.090	0.102	0.115
		(2.175)	(2.221)	(2.510)	(2.533)	(2.141)	(2.595)

Panel B: Value-wiehgted portfolio returns

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Low	0.078	0.211	0.211	0.203	0.195	0.166	0.154
		(5.451)	(5.469)	(5.387)	(5.960)	(3.584)	(3.421)
2	0.163	0.200	0.201	0.197	0.189	0.150	0.144
		(5.970)	(6.001)	(5.895)	(6.463)	(4.226)	(3.962)
3	0.287	0.206	0.207	0.203	0.196	0.168	0.162
		(6.177)	(6.222)	(6.051)	(6.468)	(4.259)	(4.013)
4	0.682	0.220	0.222	0.219	0.212	0.177	0.172
		(7.404)	(7.529)	(7.247)	(7.954)	(5.268)	(4.800)
High	5.167	0.243	0.244	0.241	0.235	0.227	0.225
		(8.323)	(8.456)	(8.059)	(8.620)	(7.441)	(6.769)
High-Low	5.089	0.032	0.034	0.038	0.040	0.062	0.071
		(2.111)	(2.120)	(1.988)	(2.002)	(2.014)	(2.218)

Table 8 (Cont.)

Panel C: Portfolio returns surrounding first bond default

	Equal weighted returns				**Value weighted returns**			
	Before 2014		**After 2014**		**Before 2014**		**After 2014**	
	PD	Excess return	*PD*	Excess return	*PD*	Excess return	*PD*	Excess return
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Low	0.039	0.254	0.110	0.183	0.041	0.179	0.110	0.230
		(5.476)		(2.691)		(4.533)		(3.737)
2	0.093	0.214	0.222	0.176	0.094	0.179	0.224	0.219
		(4.791)		(2.862)		(4.517)		(4.238)
3	0.188	0.213	0.375	0.204	0.188	0.162	0.375	0.246
		(4.598)		(3.722)		(3.731)		(5.184)
4	0.473	0.197	0.894	0.280	0.459	0.138	0.881	0.294
		(4.766)		(7.198)		(3.673)		(8.041)
High	3.550	0.221	6.233	0.385	3.742	0.149	6.436	0.335
		(4.740)		(11.409)		(3.861)		(10.841)
High-Low	3.511	-0.023	6.123	0.202	3.701	-0.030	6.326	0.105
		(-1.656)		(4.017)		(-1.257)		(2.797)

Panel D: Portfolio returns surrounding asset pledgeability reform

	Equal weighted returns				**Value weighted returns**			
	Before pledgeability reform		**After pledgeability reform**		**Before pledgeability reform**		**After pledgeability reform**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Low	0.042	0.346	0.107	-0.018	0.042	0.304	0.108	0.030
		(6.682)		(-0.278)		(6.483)		(0.472)
2	0.089	0.309	0.217	-0.128	0.090	0.281	0.220	-0.004
		(5.720)		(-1.883)		(5.283)		(-0.072)
3	0.158	0.269	0.486	-0.047	0.157	0.243	0.474	0.078
		(4.904)		(-0.753)		(4.670)		(1.396)
4	0.392	0.256	2.215	0.244	0.391	0.235	2.223	0.226
		(5.081)		(4.596)		(4.406)		(5.064)
High	3.233	0.325	8.378	0.391	3.396	0.270	8.860	0.310
		(5.288)		(7.772)		(5.046)		(8.818)
High-Low	3.191	-0.021	8.271	0.410	3.354	-0.034	8.752	0.281
		(-1.965)		(5.382)		(-1.655)		(5.130)

Panel E: Portfolio returns surrounding *Bond Connect* program

	Equal weighted returns				**Value weighted returns**			
	Before *Bond Connect* program		**After *Bond Connect* progrma**		**Before *Bond Connect* program**		**After *Bond Connect* program**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Low	0.058	0.226	0.134	0.186	0.062	0.194	0.129	0.262
		(4.140)		(4.955)		(3.978)		(7.283)
2	0.126	0.196	0.276	0.187	0.127	0.176	0.274	0.275
		(3.936)		(5.150)		(4.176)		(11.124)
3	0.234	0.188	0.456	0.273	0.234	0.170	0.457	0.323
		(4.079)		(8.184)		(4.185)		(12.280)
4	0.575	0.216	1.079	0.321	0.563	0.184	1.060	0.335
		(5.929)		(10.904)		(5.198)		(13.022)
High	4.391	0.259	6.810	0.404	4.636	0.206	6.855	0.359
		(7.657)		(7.767)		(6.223)		(9.739)
High-Low	4.333	0.033	6.676	0.218	4.574	0.012	6.726	0.097
		(0.819)		(4.957)		(0.391)		(3.405)

Table 9. Stock portfolios sorted by corporate default risk

Quintile portfolios are formed every month from April 1994 to May 2020 by sorting stocks based on one-quarter ahead *PD* measure, which is derived from our logit regression model derived using LASSO BIC (i.e., Panel B in Table 3). Quintile 1 is the portfolio with the lowest *PD* and Quintile 5 is the portfolio with the highest *PD*. The portfolios in Panel A are equal weighted while the portfolios in Panel B are value weighted using market capitalization as weights. For each quintile portfolio, we report its average value of *PD*, average monthly excess return in the next 12 months over one-year fixed term deposit rate, the CAPM alpha, the three-factor alpha estimated from the Fama-French three-factor model. The last row in each panel shows the differences in *PD* value, excess return, and alphas between the highest and lowest quintile portfolios. The asset pricing models for stocks are described in detail in Appendix 2. Excess returns and alphas are expressed in monthly percentage terms. Newey-West adjusted *t*-statistics are given in parentheses.

Panel A: Equal-weighted portfolio returns				
	PD	**Excess return**	**CAPM alpha**	**Three-factor alpha**
	(1)	(2)	(3)	(4)
Low	0.026	1.356	1.268	1.312
		(3.298)	(3.288)	(3.309)
2	0.108	1.494	1.399	1.435
		(3.490)	(3.479)	(3.467)
3	0.292	1.614	1.511	1.545
		(3.544)	(3.529)	(3.493)
4	0.755	1.625	1.524	1.551
		(3.536)	(3.525)	(3.473)
High	3.948	1.677	1.574	1.591
		(3.682)	(3.678)	(3.604)
High-Low	3.922	0.320	0.306	0.279
		(3.289)	(3.225)	(2.953)

Panel B: Value-weighted portfolio returns				
	(1)	(2)	(3)	(4)
Low	0.023	1.070	0.991	1.067
		(2.677)	(2.641)	(2.800)
2	0.106	1.142	1.051	1.115
		(2.718)	(2.667)	(2.773)
3	0.287	1.286	1.187	1.248
		(2.859)	(2.816)	(2.882)
4	0.747	1.234	1.143	1.203
		(2.787)	(2.753)	(2.814)
High	3.772	1.367	1.268	1.311
		(3.012)	(2.981)	(2.986)
High-Low	3.749	0.297	0.277	0.243
		(2.441)	(2.336)	(2.040)

Table 10. Corporate default risk and stock performance during the COVID-19 pandemic

Panel A of this table reports the OLS regression results of stock returns on *PD* between October 2019 and May 2020 where the dependent variable is daily stock return in columns 1-2 and cumulative stock return over the pre- and post-COVID outbreak in columns 3. *Post COVID* is a dummy variable which equals one if the observation is after 23rd January 2020, and zero otherwise. *PD* is the probability of default measure estimated from our logit regression model derived using LASSO BIC (i.e., Panel B in Table 3). Panel B reports the average daily returns on quintile stock portfolios sorted on the *PD* measure as of the end of 2018 for the pre- and post-COVID periods. Panel C reports the cumulative returns on quintile portfolios sorted on the *PD* measure as of the end of Q1 2019 for pre- and post-COVID periods. Robust standard errors (reported in parentheses) are clustered at the firm level. ***, ** and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A: OLS regressions

	Daily stock returns		**Cumulative stock returns**
	(1)	(2)	(3)
Post COVID	-0.138***		-13.266***
	(-18.943)		(-13.781)
PD	0.004***		0.217
	(4.022)		(1.592)
Post COVID×PD	-0.005***	-0.003***	-0.240**
	(-4.970)	(-5.167)	(-2.317)
Constant	0.123***	0.151***	10.376***
	(23.254)	(3.597)	(15.482)
Observations	574,339	574,339	7,330
Firm fixed effects	Yes	Yes	Yes
Day fixed effects	No	Yes	No
Adjusted R-squared	0.006	0.273	0.115

Panel B: Average daily portfolio returns

	Low *PD*	**2**	**3**	**4**	**High *PD***
Pre-period	0.118	0.124	0.165	0.147	0.133
Post-period	0.003	0.010	0.009	-0.005	-0.087
Diff	-0.115***	-0.114***	-0.156***	-0.152***	-0.220***

Panel C: Cumulative portfolio returns

	Low *PD*	**2**	**3**	**4**	**High *PD***
Pre-period	9.130	9.825	13.600	12.388	10.996
Post-period	-1.732	-0.855	-1.415	-2.799	-8.302
Diff	10.862***	10.680***	15.015***	15.187***	19.298***

Figure 1. China's onshore bond market development between 2007 and August 2020
This figure plots the size of Chinese onshore bond market by total outstanding in CNY billions between 2007 and August 2020 in Panel A and illustrates foreign investor holdings in China's onshore bond market during the same period in Panel B. Data is sourced from ChinaBond.

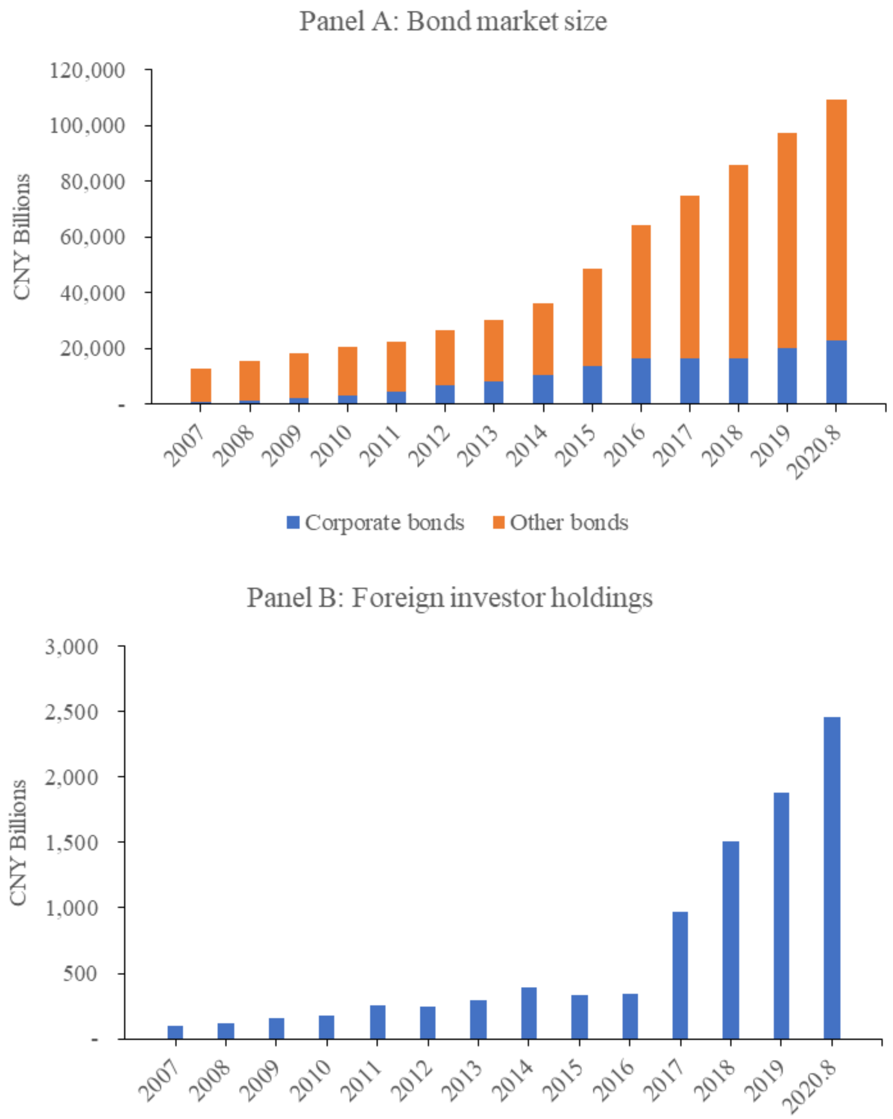


Figure 2. Potential credit risk in China's onshore bond market

Panel A plots Chinese bond marekt size in proportion to its GDP between 2007 and 2019. Panel B draws comparisons in business bankruptcy cases between the US and China from 2015 to 2018. Panel C displays the total number of corporate bond defaults and number of bond defaults in the following issuer credit rating categories: AA+ and above, AA+ below and unrated from 2014 to 2019. Data sources: Bond market information, economic variables and bond defaults data are retrieved from Wind database; The Chinese bankruptcy information is taken from the Supreme People's Court and the US bankruptcy information is from American Bankruptcy Institute; the total number of firms in China and the US is from National Bureau of Statistics of China and United States Census Bureau, respectively. The sample periods vary depending on data availability.

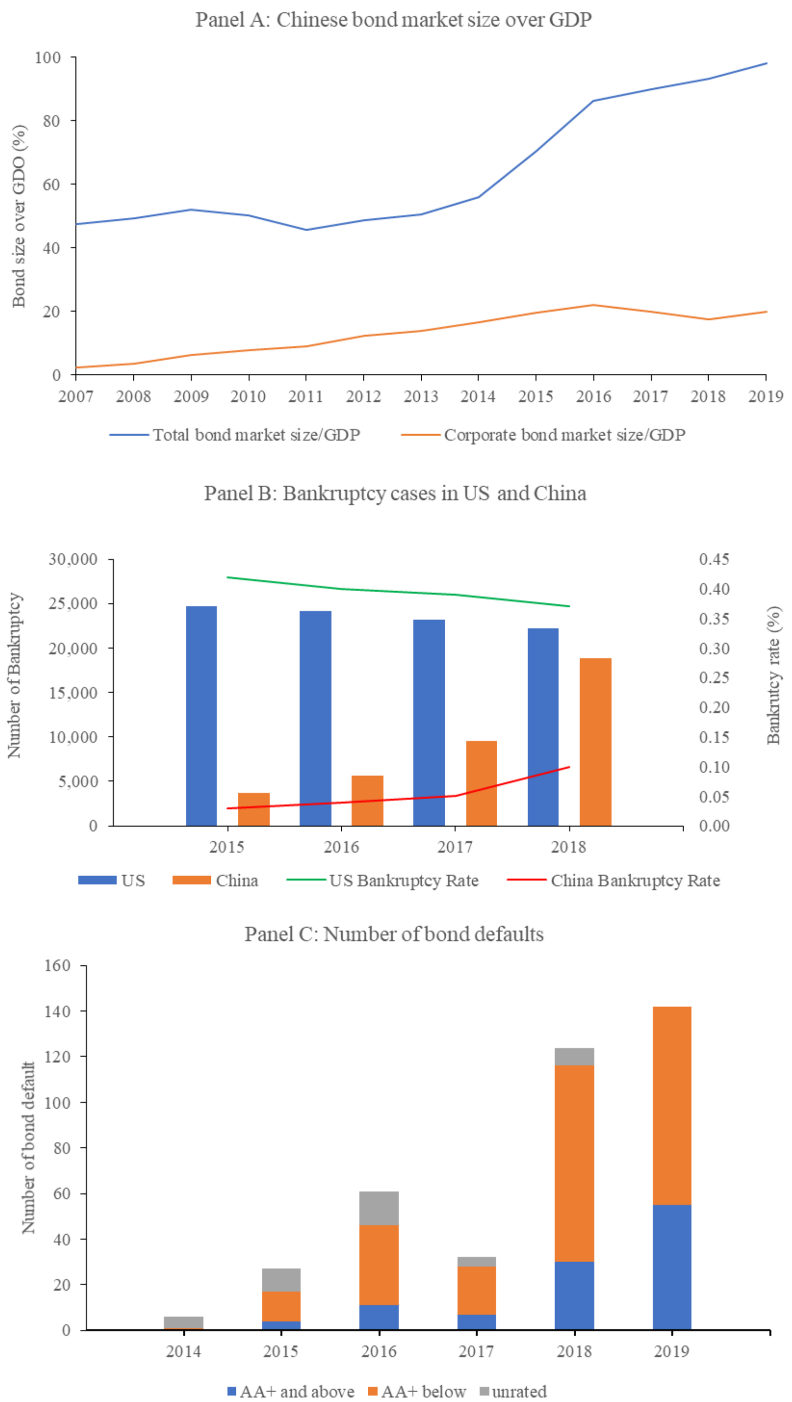
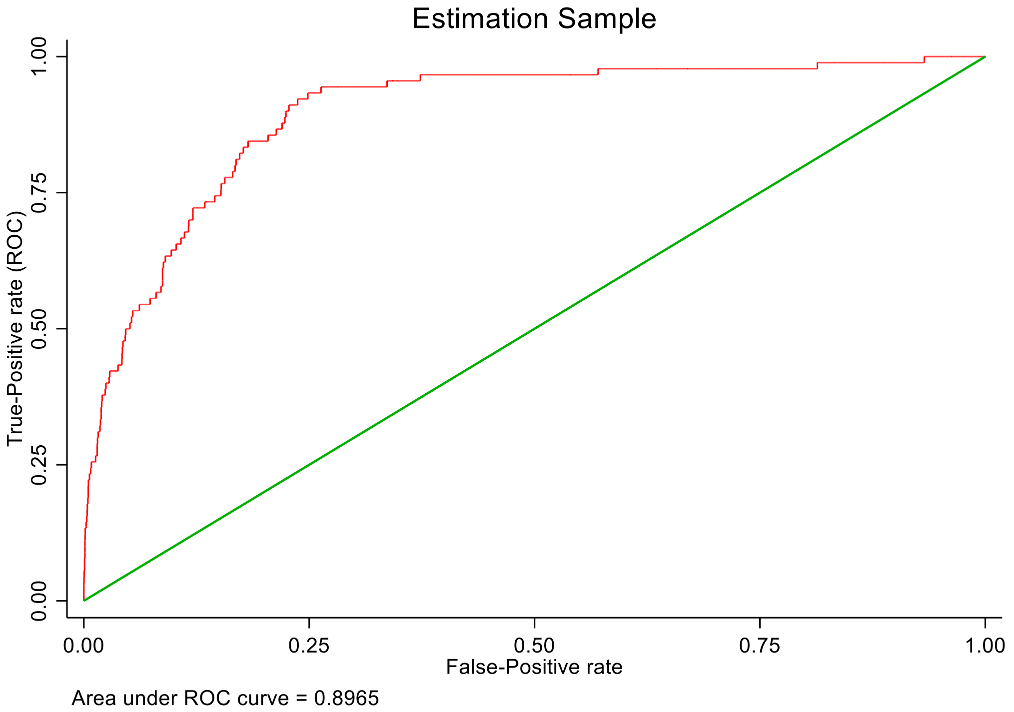


Figure 3. ROCs of the default risk models using LASSO Model

This figure plots out the ROCs for our default risk models derived from LASSO BIC, for the estimation and hold-out sample, respectively.



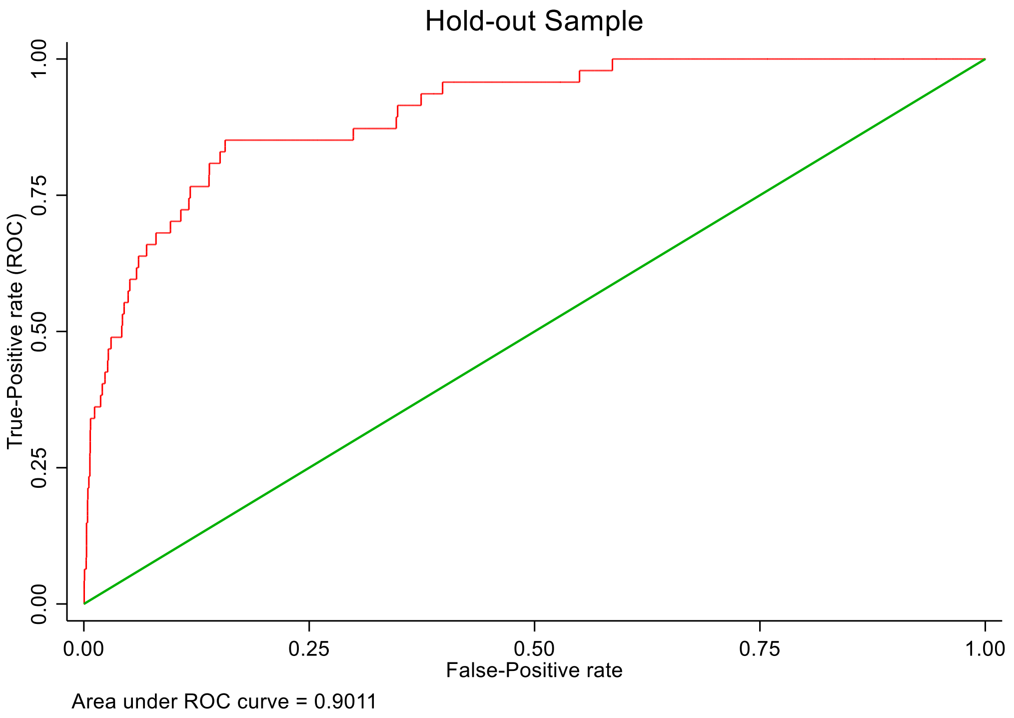


Figure 4. Average default risk of defaulted and otherwise similar non-defaulted firms

This figure plots the time series of the cross-sectional averages of the yearly *PD* measures of defaulted and otherwise similar non-defaulted firms over the 28-quarter period before bond defaults. We locate one non-defaulted firm similar in characteristics to each defaulted firm in the same industry and with the closest sales value. Numbers in X-axis denote the number of quarters prior to the default.

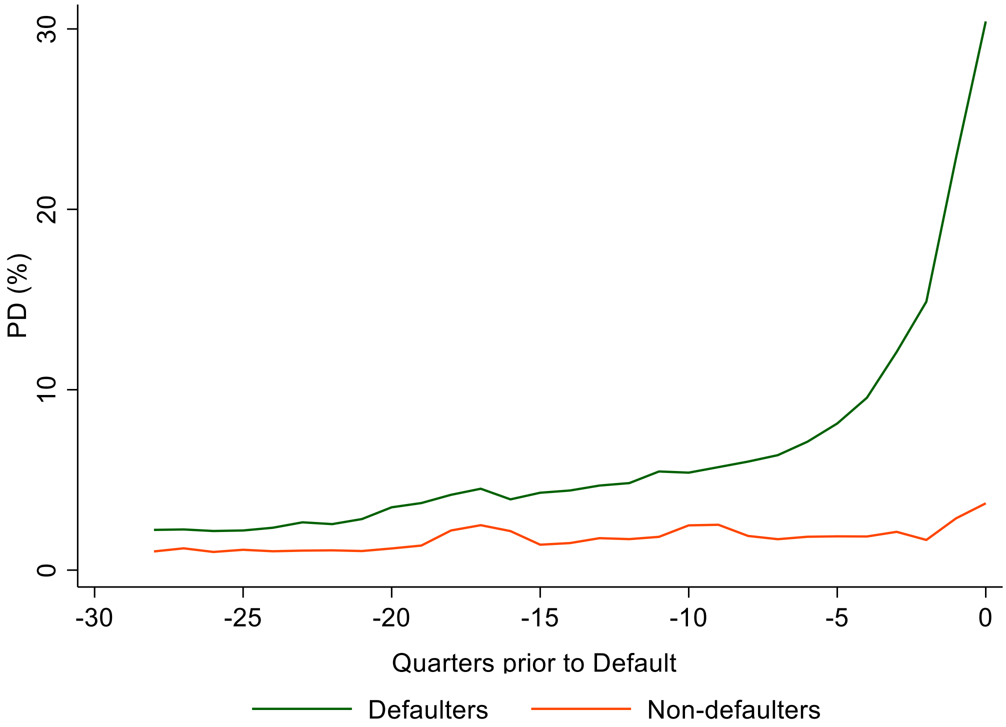


Figure 5. Corporate default risk across time

This figure plots the yearly time series of mean, 90 percentile (P90) and maximum (Max) value of the *PD* measures for primary bond market (Panel A), secondary bond market (Panel B) and stock market (Panel C).

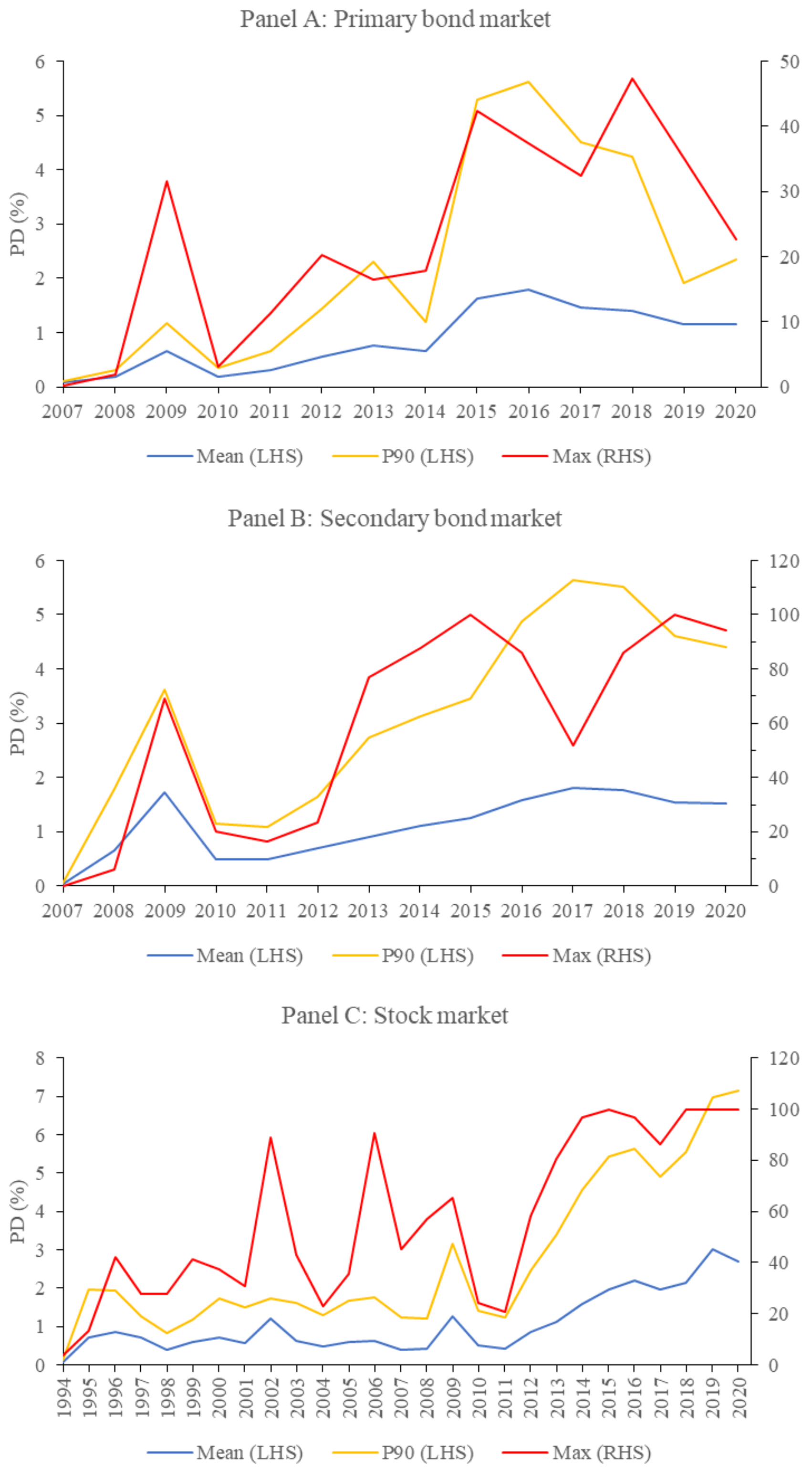
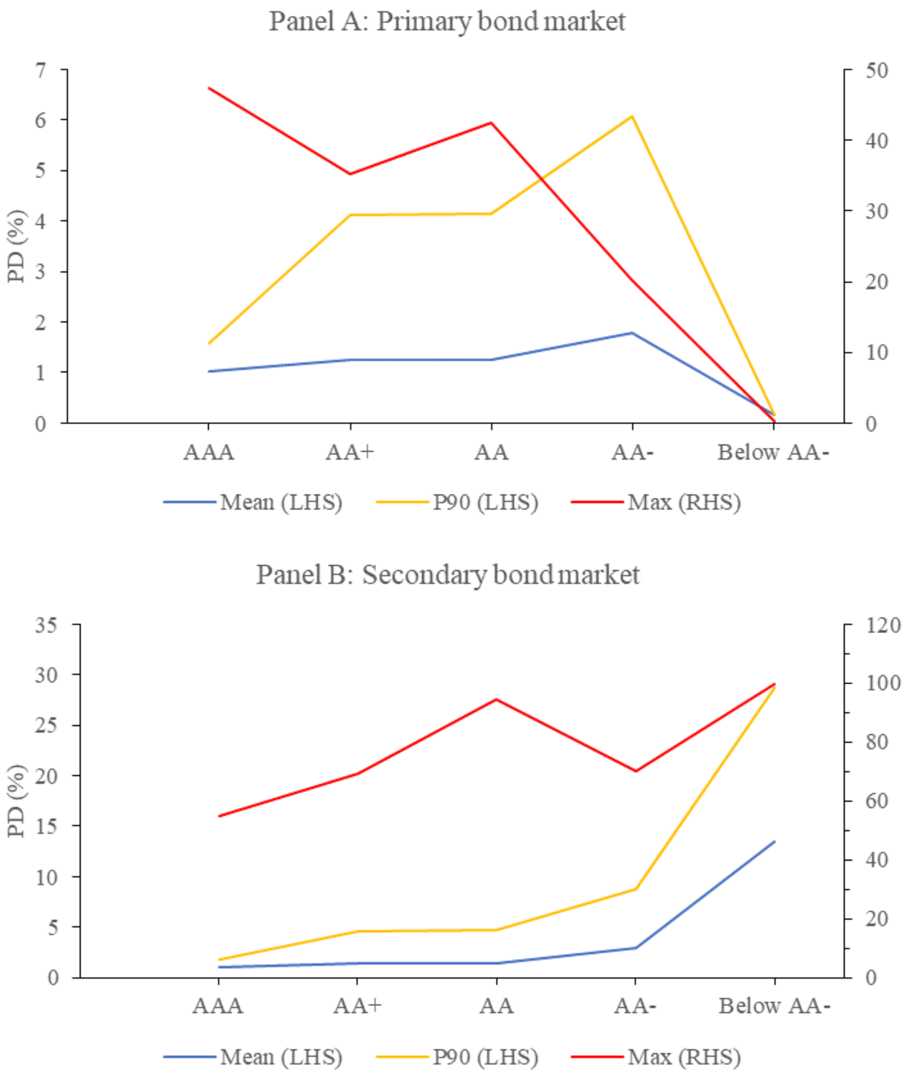


Figure 6. Probability of default by credit ratings

This figure plots the mean, 90th percentile (P90) and maximum (Max) value of the yearly *PD* measures in each credit rating category for primary bond market (Panel A) and secondary bond market (Panel B).



Online Appendix

to Accompany

Understanding Credit Risk for Chinese Companies using Machine Learning: A Default-Based Approach

OA1. Expected default probability

We construct a simplified EDP measure following Bharath and Shumway (2008) on basis of the classic Merton (1974) bond pricing model. The Merton's model stipulates that the equity of the firm is a call option on the underlying value of the firm with a strike price equal to the face value of the firm's debt. On this premise, the value of the firm's debt can be derived by the put-call parity. The Merton's EDP measure in Bharath and Shumway involves the following key inputs: the firm's total debt, market capitalization, previous year historical stock returns and volatility. Specifically, EDP is calculated from $N(-DD)$, where $N(\cdot)$ refers to the cumulative standard normal distribution and *DD* refers to the distance to default calibrated as below:

$$DD = \frac{\ln[(MC+D)/D]+\left(r_{t-1}-0.5\sigma_V^2\right)T}{\sigma_v\sqrt{T}}, \quad \text{(A1.1)}$$

where *MC* is market capitalization, *D* is the sum of current liabilities plus one-half of long-term debt, *T* is the presumed forecast horizon which is assumed to be one year in our analysis, r_{t-1} is the previous year stock return, and σ_v is the approximate volatility of the firm's value over the past year, which can be estimated using the volatility of the firm's equity using the following equation:

$$\sigma_v = \frac{MC}{MC+D}\sigma_{t-1} + \frac{D}{MC+D}(0.05 + 0.25 \times \sigma_{t-1}), \quad \text{(A1.2)}$$

where σ_{t-1} is the volatility of daily stock returns over the past year. We only retain EDP estimates with a minimum of 50-day daily stock returns in a given year. We first predict the monthly default probability using the above method and then derive a quarterly default risk estimate by taking a simple average of the monthly forecasts.

OA2. Emerging market distress risk measure

This section describes the construction of a financial distress risk measure developed in Asis, Chari and Haas (2020), calibrated for companies in emerging market economies. In a nutshell, the measure is derived by estimating a logit model of forecasting corporate default probabilities based on a set of firm fundamentals and stock market variables, along with local and global economic indicators.

The detailed list of firm-level accounting and stock market variables is summarized in Table OA1 of Online Appendix, along with variable definitions and data sources. The sample used to estimate Asis, Chari and Haas's model covers 49,580 firm-month observations from December 2013 to June 2020. To draw comparison with our *PD* measure, we first predict the monthly default probability using Model 7 in Table 3 of Asis, Chari and Haas (2020) and then derive a quarterly default risk estimate by taking a simple average of the monthly forecasts. The only modification we apply to their prediction model is that our estimation excludes the *Prior Default* dummy variable given that our sample only retains first-time default events and thus this dummy variable assumes the same value of zero.

Table OA1. Variables in Asis, Chari and Haas (2020)

Variable	Definition	Data source
Firm and stock characteristics		
Excess returns	Log (1 + stock return) - log (1 + home market index return).	CSMAR
Stock price	Log of share price at month end.	CSMAR
Return volatility	Standard deviation of daily returns over the previous month.	CSMAR
Relative size	Log (market cap) - log (home country stock market cap).	CSMAR
Profitability	Ratio of net income to the market value of total assets, where the market value of assets is equal to the sum of the firm's market capitalization and total liabilities.	CSMAR
Leverage	Ratio of total liabilities to the market value of total assets.	CSMAR
Cash	Ratio of cash and cash equivalents to the market value of total assets.	CSMAR
Market-to-book ratio	Ratio of market capitalization to book value of equity, where book value of equity is equal to total assets minus total liabilities.	CSMAR
Local economic characteristics		
Unemployment rate	Unemployment rate in a given year.	World Bank
Inflation rate	Monthly change in Consumer Price Index (CPI).	Bank for International Settlements
Real interest rate	Real interest rate in a given year.	World Bank
Sovereign spread	JP Morgan Emerging Markets Bonds Spread.	World Bank
ΔFX 12m avg	The 12-month average of the monthly changes in a country's bilateral exchange rate against the US dollar.	Bank for International Settlements
Global economic characteristics		
Fed funds rate	Fed Funds Rate.	FRED
Yield slope (5y-FF)	Difference between the US 5-year Treasury rate and the Fed funds rate.	FRED
Fed MP surprise	US monetary policy surprise measured as the daily change in 5-year US Treasury futures on FOMC announcement days following Chair, Stedman and Lundblad (2020).	FRED, Bloomberg and Fed
VIX	CBOE Implied Volatility Index.	CBOE
TED spread	Component of the TED spread orthogonal to VIX, where the TED spread is the difference between 3-month LIBOR rate and 3-month T-bill rate.	CBOE and FRED
ΔBroad US$	Monthly percentage change in the broad US dollar index.	FRED

Table OA2. Summary statistics of bond and issuer characteristic variables used in the analysis of primary and secondary bond markets

This table presents the mean (Mean), standard deviation (Std.Dev), median (Median), minimum (Min), maximum (Max), the 1st percentile (p1), 25th percentile (p25), 75th percentile (p75) and 99th percentile (p99) values of various bond and issuer characteristics for the sample used in the primary bond market analysis in Panel A and in the secondary market in Panel B. Panel A includes 17,450 bonds issued by 4,491 unique issuers between 2007 and May 2020 in the primary bond market. Panel B consists of 228,248 bond-month observations for 13,963 bonds issued by 4,084 unique issuers trading in the secondary bond market between 2007 and May 2020. All variables are defined in Appendix 1.

Panel A: Summary statistics of key variables used in the primary market analysis

	N	Mean	Std.Dev	Median	Min	Max	p1	p25	p75	p99
Offering spread	17,450	2.247	1.111	2.060	0.110	6.540	0.490	1.360	2.990	4.980
Offering spread (CDB)	17,450	1.669	1.063	1.447	-0.411	5.958	0.091	0.844	2.292	4.518
Offering spread (Synthetic)	17,450	2.282	1.122	2.089	0.127	6.575	0.516	1.388	3.038	5.069
PD	17,450	1.169	2.694	0.312	0.000	47.373	0.015	0.148	0.727	12.059
AAA	17,450	0.411	0.492	0.000	0.000	1.000	0.000	0.000	1.000	1.000
AA+	17,450	0.277	0.447	0.000	0.000	1.000	0.000	0.000	1.000	1.000
AA	17,450	0.299	0.458	0.000	0.000	1.000	0.000	0.000	1.000	1.000
AA-	17,450	0.013	0.112	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Below AA-	17,450	0.000	0.008	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Maturity	17,450	4.962	1.983	5.000	2.000	20.000	2.000	3.000	6.000	10.000
ISize	17,450	1.303	1.524	1.000	0.030	30.000	0.170	0.500	1.500	7.500
Puttable	17,450	0.258	0.437	0.000	0.000	1.000	0.000	0.000	1.000	1.000
Callable	17,450	0.092	0.288	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Sinking fund	17,450	0.161	0.367	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Cross	17,450	0.189	0.392	0.000	0.000	1.000	0.000	0.000	0.000	1.000
MTN	17,450	0.530	0.499	1.000	0.000	1.000	0.000	0.000	1.000	1.000
Enterprise	17,450	0.221	0.415	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Corporate	17,450	0.249	0.432	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Public	17,450	0.178	0.383	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Age	17,450	17.234	8.292	16.000	0.000	72.000	4.000	11.000	22.000	39.000
Leverage	17,450	0.585	0.164	0.612	0.007	0.962	0.143	0.481	0.698	0.866
Tangibility	17,450	0.286	0.214	0.269	-0.555	0.992	-0.153	0.132	0.420	0.830
Sales	17,450	2.012	2.072	1.936	-9.762	7.995	-2.019	0.369	3.521	6.655
Growth	17,450	0.957	48.445	0.128	-0.988	6144.235	-0.518	0.017	0.305	3.759
ROA	17,450	0.043	0.037	0.034	-0.176	0.501	0.002	0.020	0.055	0.183
Cash	17,450	0.729	18.686	0.352	0.000	1921.684	0.043	0.226	0.559	2.947

Table OA2 (Cont.)

Panel B: Summary statistic of key variables used in the secondary market analysis										
	N	**Mean**	**Std.Dev**	**Median**	**Min**	**Max**	**p1**	**p25**	**p75**	**p99**
YTM spread	228,248	2.530	6.108	2.049	-9.927	973.304	0.280	1.376	3.001	7.953
YTM spread (Avg)	228,248	2.466	3.490	2.128	-4.219	98.854	-0.829	1.435	2.994	8.118
YTM spread (CDB)	228,248	1.947	6.118	1.413	-10.615	973.172	-0.278	0.829	2.318	7.579
YTM spread (Synthetic)	228,248	2.244	6.117	1.754	-9.111	973.643	0.065	1.129	2.612	7.864
PD	228,248	1.358	3.302	0.316	0.000	99.921	0.019	0.154	0.806	14.670
AAA	228,248	0.345	0.475	0.000	0.000	1.000	0.000	0.000	1.000	1.000
AA+	228,248	0.297	0.457	0.000	0.000	1.000	0.000	0.000	1.000	1.000
AA	228,248	0.339	0.473	0.000	0.000	1.000	0.000	0.000	1.000	1.000
AA-	228,248	0.017	0.129	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Below AA-	228,248	0.002	0.045	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Maturity	228,248	3.510	2.055	3.236	0.082	19.997	0.186	1.970	4.811	9.501
ISize	228,248	1.729	2.260	1.000	0.000	30.000	0.200	0.750	2.000	11.000
Putable	228,248	0.257	0.437	0.000	0.000	1.000	0.000	0.000	1.000	1.000
Callable	228,248	0.053	0.225	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Sinking fund	228,248	0.219	0.414	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Cross	228,248	0.250	0.433	0.000	0.000	1.000	0.000	0.000	0.000	1.000
MTN	228,248	0.497	0.500	0.000	0.000	1.000	0.000	0.000	1.000	1.000
Enterprise	228,248	0.313	0.464	0.000	0.000	1.000	0.000	0.000	1.000	1.000
Corporate	228,248	0.190	0.392	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Public	228,248	0.212	0.409	0.000	0.000	1.000	0.000	0.000	0.000	1.000
Age	228,248	15.952	7.765	15.000	0.000	72.000	3.000	10.000	21.000	36.000
Leverage	228,248	0.585	0.156	0.605	0.003	1.337	0.189	0.479	0.697	0.867
Tangibility	228,248	0.289	0.207	0.281	-0.630	0.997	-0.170	0.141	0.428	0.776
Sales	228,248	2.046	2.111	1.915	-13.085	8.007	-2.027	0.369	3.545	7.603
Growth	228,248	0.398	8.286	0.106	-0.999	1028.908	-0.590	-0.010	0.278	3.076
ROA	228,248	0.039	0.047	0.031	-0.583	8.611	-0.019	0.017	0.051	0.159
Cash	228,248	0.574	9.070	0.365	0.000	1921.684	0.046	0.235	0.582	2.800

Table OA3. Yield spread and *PD* by rating categories

This table reports the OLS regression results of *Offering spread* and *YTM spread* on *PD* by rating categories and foreign ownership in rating agencies, along with other bond and issuer characteristic controls. In columns 1-4, the dependent variable is bond offering yield spread (*Offering spread),* which is the percentage difference between bond offering yield and yield on a Treasury bond of comparable maturity. In columns 5-8, the dependent variable is trading yield spread, *YTM spread,* which is the percentage difference between bond yield to maturity at month end and yield on a Treasury bond of comparable maturity. *PD* is the probability of default measure estimated by the logit regression model derived by LASSO BIC (i.e., Panel B in Table 3). The bond control variables include *Maturity, ISize, Puttable, Callable, Sinking fund, Cross, Corporate* and *Enterprise,* while the issuing firm control variables are *Public, Age, Leverage, Tangibility, Sales, Growth, ROA*, and *Cash*. Industry and year fixed effects are also included. Columns 1 and 5 report the results for AAA rated bonds. Columns 2 and 6 report the results for AA+ rated bonds. Columns 3 and 7 report the results for AA rated bonds. Columns 4 and 8 report the results for bonds with below AA ratings. All variables are defined in Appendix 1. *YTM spread* and the continuous issuer-specific financial variables are winsorized at the top and bottom 1% of the sample distribution. The robust standard errors (reported in parentheses) are clustered at the firm level in columns 1-4 and at the bond level in columns 5-8. ***, ** and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Primary Market Dep.var = *Offering spread*				**Secondary Market Dep.var = *YTM spread***			
	AAA	AA+	AA	Below AA	AAA	AA+	AA	Below AA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PD	0.057***	0.119***	0.094***	0.054**	0.073***	0.116***	0.100***	0.039***
	(3.640)	(9.999)	(8.854)	(2.388)	(8.971)	(11.602)	(10.776)	(6.535)
Constant	1.296***	2.954***	3.067***	1.737**	1.429***	3.246***	2.740***	2.849***
	(4.340)	(12.069)	(11.923)	(1.995)	(6.956)	(12.821)	(8.515)	(4.244)
Observations	7,178	4,829	5,219	224	78,786	67,710	77,422	4,330
Bond controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.385	0.397	0.354	0.398	0.525	0.342	0.276	0.460

Table OA4. Alternative yield spread measures and inclusion of other fixed effects

This table reports the OLS regression results of alternative yield spread measures on *PD*, along with other bond and issuer characteristic controls. Columns 1 and 3 use the yield on a bond of comparable maturity issued by CDB as the benchmark rate for the primary and secondary market bonds, respectively. Columns 2 and 4 use the yield on a synthetic bond constructed following Ang, Bai and Zhou (2018) method as the benchmark rates for primary and secondary market bonds, respectively. In column 5, the dependent variable is an average yield spread, which is computed as the percentage difference between the trading volume-weighted average of daily bond yields in a given month and yield on a Treasury note of comparable maturity for secondary market bonds. Bond fixed effects and Year-month fixed effects are added in the baseline regression model in columns 6 and 7, respectively. *PD* is the probability of default measure estimated from the logit regression model derived by LASSO BIC (i.e., Panel B in Table 3). The bond control variables include bond credit rating dummies (*AA+*, *AA*, *AA-* and *Below AA-*), *Maturity, ISize, Puttable, Callable, Sinking fund, Cross*, *Corporate* and *Enterprise,* while the issuing firm control variables are *Public*, *Age*, *Leverage, Tangibility, Sales, Growth, ROA,* and *Cash*. Industry and year fixed effects are also included for all the regression models. All variables are defined in Appendix 1. *YTM spread (CDB)*, *YTM spread (Synthetic)*, *YTM spread (Avg)*, *YTM spread* and the continuous issuer-specific financial variables are winsorized at the top and bottom 1% of the sample distribution. The robust standard errors (reported in parentheses) are clustered at the firm level in columns 1-2 and at the bond level in columns 3-7. ***, ** and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Primary Market		**Secondary Market**				
	Offering spread (CDB)	*Offering spread (Synthetic)*	*YTM spread (CDB)*	*YTM spread (Synthetic)*	*YTM spread (Avg)*	Bond fixed effects	Year-month fixed effects
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PD	0.092***	0.090***	0.091***	0.090***	0.090***	0.034***	0.092***
	(8.599)	(8.482)	(19.328)	(19.365)	(18.754)	(8.772)	(19.326)
Constant	1.066***	1.541***	0.662***	0.991***	1.188***	2.172***	1.461***
	(6.121)	(6.635)	(3.408)	(4.388)	(8.854)	(8.901)	(16.951)
Observations	17,450	17,450	228,248	228,248	228,248	228,248	228,248
Bond controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	No
Bond fixed effects	No	No	No	No	No	Yes	No
Year-month fixed effects	No	No	No	No	No	No	Yes
Adjusted R-squared	0.494	0.543	0.425	0.425	0.350	0.747	0.456

Table OA5 Alternative corporate bond samples

This table reports the OLS regression results of *Offering spread* in the primary market and *YTM spread* in the secondary market on *PD*, along with other bond and issuer characteristic controls over various subsamples. Columns 1 and 3 report the baseline results for the sample excluding bonds with puttable, callable and sinking fund options. Columns 2 and 4 report the results for the sample excluding Chengtou bonds. The dependent variable is bond offering yield spread (*Offering spread*) in columns 1-2, which is the percentage difference between bond offering yield and yield on a Treasury bond of comparable maturity. The dependent variable in columns 3-4 is the trading yield spread in the secondary market, *YTM spread,* which is the percentage difference between the bond yield to maturity at the month end and yield on a Treasury bond of comparable maturity. *PD* is the probability of default measure estimated from the logit regression model derived by LASSO BIC (i.e., Panel B in Table 3). The bond control variables include bond credit rating dummies (*AA+, AA, AA-* and *Below AA-*), *Maturity, ISize, Puttable, Callable, Sinking fund, Cross, Corporate* and *Enterprise,* while the issuing firm control variables are *Public, Age, Leverage, Tangibility, Sales, Growth, ROA*, and *Cash*. Industry and year fixed effects are also included. All variables are defined in Appendix 1. *YTM spread* and the continuous issuer-specific financial variables are winsorized at the top and bottom 1% of the sample distribution. The robust standard errors (reported in parentheses) are clustered at the firm level in columns 1-2 and at the bond level in columns 3-4. ***, ** and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Primary Market Dep.var = *Offering spread*		**Secondary Market Dep.var = *YTM spread***	
	Non-PCS	Non-Chengtou	Non-PCS	Non-Chengtou
	(1)	(2)	(3)	(4)
PD	0.074***	0.074***	0.085***	0.074***
	(7.266)	(7.819)	(13.381)	(17.163)
Constant	1.076***	1.117***	0.354***	0.773***
	(4.398)	(4.259)	(2.809)	(3.550)
Observations	8,914	9,154	111,184	116,638
Bond controls	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
Adjusted R-squared	0.550	0.543	0.431	0.451

Table OA6. Summary statistics of risk factors in bond and equity markets

This table reports the summary statistics including mean (Mean), standard deviation (Std.Dev), median (Median), min (Min), max (Max), 1st (p1), 25th (p25), 75th (p75) and 99th (p99) percentiles of common risk factors used in the asset pricing models for determining the cross section of corporate bond returns for the period October 2008 - May 2020 in Panel A and of stock returns for the period April 1994 - May 2020 in Panel B. All the risk factors are described in detail in Appendix 2.

Panel A: Common risk factors in corporate bond market										
	N	**Mean**	**Std.Dev**	**Median**	**Min**	**Max**	**p1**	**p25**	**p75**	**p99**
MKT^{Stock}	138	0.719	7.840	1.235	-27.434	19.927	-26.559	-3.401	4.269	18.748
SMB	138	1.058	4.814	1.020	-22.118	22.838	-8.903	-1.789	3.550	11.932
HML	138	-0.174	3.555	0.031	-14.369	15.951	-8.601	-1.978	1.485	9.222
MKT^{Bond}	138	0.228	0.514	0.246	-1.668	2.233	-1.269	-0.055	0.507	1.588
REV	138	1.212	0.863	1.193	-1.650	5.991	-1.334	0.846	1.516	4.980
DRF	138	0.080	0.774	0.080	-4.110	2.599	-2.406	-0.270	0.433	2.591
DRF36	138	0.040	0.762	0.019	-3.638	3.113	-1.849	-0.280	0.286	2.786
LRF	138	0.294	1.891	-0.065	-3.247	12.355	-3.045	-0.390	0.395	11.823
CRF	138	0.144	0.644	0.191	-3.218	2.431	-1.494	-0.155	0.492	1.762
CRF36	138	0.146	0.622	0.168	-2.868	1.953	-1.306	-0.207	0.484	1.810
DEF	138	-0.040	1.306	-0.055	-3.790	4.510	-3.360	-0.700	0.550	3.150
TERM	138	0.147	1.818	0.258	-4.781	6.757	-4.712	-1.021	1.297	4.368
Panel B: Common risk factors in equity market										
MKT^{Stock}	312	0.996	11.113	0.561	-27.434	112.636	-24.118	-4.765	5.050	30.561
SMB	312	0.851	4.732	0.812	-22.118	22.838	-11.376	-1.752	3.549	10.558
HML	312	0.243	3.311	0.406	-14.369	15.951	-8.965	-1.527	2.025	9.222

Table OA7. Alternative downside risk factor for bond returns

This table replicates the bond portfolio tests in columns 6-7 of Panels A and B in Table 8 using an alternative downside risk factor constructed following Bai, Bali, and Wen (2019), which is the 5% VaR calculated as the second lowest monthly return observation over the past 36 months. We require a minimum of six monthly observations over the past 36 months for this downside risk factor construction. Newey-West adjusted *t*-statistics are given in parentheses.

Panel A: Equal-weighted portfolio returns			
	PD	**Seven-factor bond alpha (36-month VaR)**	**Ten-factor bond alpha (36-month VaR)**
	(1)	(2)	(3)
Low	0.077	0.173	0.156
		(3.528)	(3.399)
2	0.162	0.151	0.136
		(3.574)	(3.311)
3	0.287	0.173	0.166
		(4.135)	(4.014)
4	0.695	0.201	0.196
		(5.950)	(5.631)
High	4.969	0.273	0.269
		(8.342)	(7.545)
High-Low	4.892	0.100	0.113
		(2.147)	(2.603)
Panel B: Value-weighted portfolio returns			
	(1)	(2)	(3)
Low	0.078	0.168	0.156
		(3.693)	(3.530)
2	0.163	0.152	0.146
		(4.327)	(4.043)
3	0.287	0.169	0.164
		(4.365)	(4.082)
4	0.682	0.178	0.174
		(5.262)	(4.765)
High	5.167	0.230	0.227
		(7.483)	(6.744)
High-Low	5.089	0.062	0.071
		(1.982)	(2.023)