Predicting Corporate Bond Ratings via a Semiparametric Index Model

Yixiao Jiang*

Abstract

This paper proposes a semiparametric model with multiple-index to predict credit ratings on corporate bonds. The model explicitly takes conflicts of interest into account and allows ratings to depend flexibly on risk predictors through a semiparametric index structure. Asymptotic normality for the proposed estimator is derived after reducing the bias. Using Moody’s rating data from 2001-2016, the proposed semiparametric model outperforms the ordered-probit model at prediction, with the largest predictive gain of 18% in 2007. Relative to its rating standard in 2001, Moody’s ratings become more stringent after the 2007-2009 financial crisis, which may reflect the regulatory effects of the Dodd-Frank Act.

1 Introduction

The development of corporate bond rating prediction models has attracted lots of research interest in the academic and business communities. Several studies in the bond rating literature have used statistical models, including Linear Probability Model, multivariate discriminant analysis (MDA), and ordered-choice models to predict bond ratings (Blume et al., 1998, Horrigan, 1966, Kaplan and Urwitz, 1979, Pinches and Mingo, 1973, West, 1970). Bond rating is, in a way, a classification problem. A growing strand of papers leverage machine learning methods, such as variables selection, support vector machines, and neural networks (Bellotti et al., 2011, D’Rosario and Hsieh, 2018, Huang et al., 2004, Lee, 2007, Sermpinis et al., 2018) to predict ratings. Machine learning can result

*Assistant Professor, Department of Economics, Christopher Newport University. Email: yixiao.jiang@cnu.edu. This paper is based on the first chapter of my PhD dissertation titled “Semiparametric Estimation of Financial Risk: Corporate Default, Credit Ratings, and Implied Volatility”. I thank my advisor, Roger Klein, for his advice and comments.
in a boost in predictive power, but often at a cost of reduced economic interpretability because of the opacity and complex computation structures. As such, such methods may not be optimal when the focal interest is not purely forecasting.

The goal of this paper is to propose a more interpretable predictive model that can also shed light on the Credit Rating Agencies' (CRA) rating methodology. Given the public concern that the ratings were contaminated by various conflicts of interest in the 2007-2009 financial crisis, I include a measure of conflicts of interest in the rating model along with other firm and bond level risk predictors, such as firm asset and leverage ratio. Specifically, I formulate a semiparametric ordered-response model where the rating probability is fully nonparametric in the proposed measure of conflict of interest and depends on other firm/bond characteristics through an index structure. The proposed semiparametric model maintains the same level of economic interpretability as the ordered probit model (OPM) and displays a better predictive power. Unlike machine learning methods in which parameters generally do not carry any economic meanings, parameters in the proposed semiparametric models quantify the relative importance of risk attributes in determining credit ratings. Structural objects, like the Average Structural Function, can also be recovered if endogeneity in the data is taken care of.

While there are numerous studies on single-index models in the extant literature\(^1\), only a handful of studies are available on the estimation of multiple-index regression models. Identification results of multiple-index models are first established by Ichimura and Lee (1991) and Horowitz (1998), where the authors use semiparametric least square (SLS) to estimate the index parameters. Donkers and Schafgans (2008) proposed a derivative-based method, which extends the average derivative estimator of Powell et al. (1989), to estimate such models without the need to pre-specify the number of indices. A recent study, Ahn et al. (2017), proposed to estimate the index coefficients (up to scale) based on an eigenvalue approach. Studies in the statistical literature, such as Xia et al. (2002) and Xia (2008), also employ different variants of multiple-index models to flexibly reduce the data dimension. All aforementioned studies, with the exception of Ichimura and Lee (1991), require estimating nonparametric objects and/or derivatives and therefore may be difficult to implement when sample size is small. This paper, in contrast, extends the maximum likelihood approach of

Klein and Sherman (2002) to estimate the index parameters directly without any nonparametric intermediate steps.

Specifically, I employ multivariate kernel density estimators to estimate the semiparametric rating probability function entering a quasi-likelihood. Bias-reducing kernels are often used in the literature to ensure such estimators having an appropriately low order of bias, which is required to establish asymptotic results. However, when the object of interest is a probability, these kernels can deliver estimated probability outside of [0,1] and render estimation results difficult to interpret. To circumvent this challenge, I obtain alternative bias reduction by using a “recursive differencing” strategy proposed by Shen and Klein (2019) and then establish $\sqrt{N}$ normality for the proposed estimator. Through simulation evidence, I show the proposed semiparametric model behaves well in finite sample.

Based on Moody’s rating data from 2001-2016, the proposed semiparametric model (SIM) outperforms the ordered-probit model (OPM) at prediction. I show SIM boosts predictive power in a measurable and identifiable way by capturing genuine aspects of the underlying rating mechanism. In years when estimated parameters in OPM and SIM diverge substantially, SIM consistently delivers superior predictive power. The improvement is especially pronounced during the financial crisis. For example, SIM correctly predicts 18.1% more ratings than the OPM in 2007. During the crisis, the error term in the model usually has a “fat” left-tail to reflect the elevated default risk (Kozlowski et al., 2015, Meine et al., 2016), which violates the normality assumptions under OPM. As such, it is no surprise that SIM turned out to be a better statistical depiction of the data and hence had a better predictive power.

I also use the developed predictive model to investigate structural changes in Moody’s rating standards. Suppose Moody’s assigns ratings in later period according to its rating standard in 2001, I ask whether a bond would have received a higher or lower rating while displaying the same risk characteristics. To be specific, I use the 2001 data to train a SIM and then predict ratings in later years based on the estimated parameters. Due to its semiparametric nature, this approach reduces the likelihood of misspecification errors. Empirically I find ratings standards started to tighten after the 2007-2009 crisis, which may reflect the regulatory effects of the Dodd-Frank Act. Relative to its standard in 2001, Moody’s ratings have dropped by more than half-a-notch for bonds with identical risk attributes.

The rest of this paper is organized as follow. In Section 2, I introduce the proposed measure of
conflicts of interest and other firm/bond characteristics that are employed in the predictive model. Section 3 outlines the model and estimation strategy. Section 4, I provide the asymptotic results for the proposed estimator and discuss the assumptions required to establish them. In so doing, I sketch out the proofs, and provide complete technical details in the Appendix. In Section 5, I compare the forecasting performance of the semiparametric model with several benchmarks and report the time-series variation in Moody’s rating standard. Section 6 concludes.

2 Data and Variable Construction

As described below, I formulate a model to predict Moody’s ratings on corporate bonds at issuance (e.g., initial ratings) from 2001, when Moody’s went public, to 2016. From Mergent’s Fixed Income Securities Database (FISD), I first obtain initial ratings on corporate bonds issued by public firms covered by either the Center for Research in Security Prices (CRSP) or Compustat. In light of the prior literature, I select several firm and bond characteristics as predictor variables. Description and summary statistics of these characteristics are provided in Table 1.

Credit rating agencies consider two risk metric when assigning ratings: probability of default (PD) and loss given defaults (LGD). Moody’s rating reflects both the probability of default and loss given default, whereas S&P’s ratings reflect only the default risk (Alp, 2013). PD is primarily affected by characteristics of the issuer firm, such as the level of asset and leverage ratio. LGD, on the other hand, depends on characteristics of a particular bond issue such as the subordination status (senior or junior debt) and the issue amount. As such, I include both firm-specific as well as bond-specific characteristics when modeling Moody’s rating procedure. After combining data from multiple sources, the final sample is composed of 11196 bonds issued by 1492 firms.

One lesson that investors and policy makers learned in the last financial crisis is credit ratings are contaminated with conflicts of interest. Credit Rating Agencies (CRAs) that have large market shares, such as Moody’s and Standard & Poor’s, are increasingly owned by large institutional investors, making their role as an unbiased financial market “gatekeeper” ever more suspect. As

---

2 I choose not to include macro variable because the model will be estimated in a short rolling window recursively (e.g., year-by-year). Therefore I include only firm and bond specific variables rather than macro variables that represent common shock to all firms.

3 Moody’s was founded as a private company in 1900, acquired by Dun&Bradstreet (D&B) in 1962, and remained one of its divisions until October 4, 2000, when it was spun off and listed on the NYSE. S&P has been a fully owned division of McGraw-Hill, a publicly traded company, since 1966. Going public makes CRAs more vulnerable to conflicts of interest. For example, Kedia et al. (2016) finds that Moody’s assigns favorable ratings towards issuers that Moody’s shareholders have invested in.
Table 1: Firm and Bond Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSET</td>
<td>logarithm of the issuer asset</td>
<td>9.643</td>
<td>2.280</td>
<td>4.360</td>
<td>14.324</td>
</tr>
<tr>
<td>ASSET VOLATILITY</td>
<td>variance of asset</td>
<td>0.230</td>
<td>0.169</td>
<td>0.003</td>
<td>1.416</td>
</tr>
<tr>
<td>LEVERAGE</td>
<td>firm leverage ratio</td>
<td>0.264</td>
<td>0.178</td>
<td>0.002</td>
<td>1.212</td>
</tr>
<tr>
<td>PROFIT</td>
<td>operating performance</td>
<td>0.026</td>
<td>0.058</td>
<td>-0.739</td>
<td>0.436</td>
</tr>
<tr>
<td>AMT</td>
<td>logarithm of the issuing amount</td>
<td>12.224</td>
<td>1.681</td>
<td>2.708</td>
<td>19.337</td>
</tr>
<tr>
<td>SENIORITY</td>
<td>a bond’s subordination status</td>
<td>0.809</td>
<td>0.393</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>MFOI</td>
<td>conflicts of interest</td>
<td>0.002</td>
<td>0.004</td>
<td>0.000</td>
<td>0.051</td>
</tr>
</tbody>
</table>

Note: A number of firm and bond characteristics are selected to predict credit ratings based on the bond rating literature (Blume et al., 1998, Horrigan, 1966, Jiang et al., 2012, Kaplan and Urwitz, 1979, West, 1970). These variables are: (1) Issuer size, defined as the value of the firm’s total assets (ASSET) (2) Stability variable (ASSET VOLATILITY), defined as the variance of the firm’s total assets in the last 16 quarters. (3) Firm leverage, defined as the ratio of long-term debt to total assets (LEVERAGE). (4) Operating performance, defined as operating income before depreciation divided by sales (PROFIT). (5) Issue size, defined as the par value of the bond issue (AMT). (6) Subordination status, a 0-1 dummy variable that is equal to one if the bond is a senior bond (SENIORITY). Firm-level variables are computed using a five-year arithmetic average of the annual ratios because Kaplan and Urwitz (1979) note that bond raters might look beyond a single year’s data to avoid temporary anomalies. The conflicts of interest measure MFOI is defined in the main text based on institutional shareholding data.

noted by Kedia et al. (2016), the credit rating decisions may be influenced by the CRA’s shareholders. Therefore incorporating a bond issuer’s liaison with Moody’s shareholders may have marginal predictive power on ratings. To construct such a measure, I first obtain the list of Moody’s shareholders and approximate each shareholder j’s “influence” on Moody’s by the percentage of Moody’s stock that they hold, λj. Next, I obtain the portfolio weight of bond issuer i in shareholder j’s investment portfolio, termed pij. The data on institutional shareholding is obtained from Thomson Reuters (13F).

I use the following variable, termed Moody-Firm-Ownership-Index (MFOI), to summarily characterize a bond issuer i’s liaison with Moody’s shareholders,

\[
MFOI_i = \sum_{j=1}^{M} p_{ij} \lambda_j, \quad \text{Moody’s has } j = 1, 2, \cdots, M \text{ shareholders}
\]  

By construction, a bond issuer i has a larger MFOI in two cases. First, when Moody’s shareholders invest more in the i’s stock, which suggests issuer i is more attractive to institutional investors. In the expression above, this happens when pij increases for some shareholder j. Alternatively, MFOI tends to be large when the common shareholders of issuer i and Moody’s have strong influence over Moody’s. When common shareholders have a higher stake in Moody’s (e.g., λj is large), Moody’s might have a greater incentive to assign favorable ratings. In both cases, a larger MFOI intuitively
leads to a higher credit ratings.\footnote{Kedia et al. (2016) employed a binary version of the above measure to proxy conflicts of interest. They focus exclusively on firms that are important to Moody’s two stable large shareholders: Berkshire Hathaway and Davis Selected Advisors. Specifically, if bond issuer i’s stock accounts for at least 0.25% in either Berkshire Hathaway’s or Davis Selected Advisors’ investment portfolio, then the binary variable takes value one, indicating that issuer i has a liaison with Moody’s. They further show that the average rating for a Moody-related firm is 0.213 notch better than a comparable non-Moody-related firm.}

3 Empirical Model

3.1 Model and Motivation for the Estimator

The observed bond rating, $Y_i$, is discrete, taking one of the values from 1 to 7, and is related to a bond’s latent credit quality $y_i^*$ as follow

$$Y_i = \begin{cases} 1 & \text{if } y_i^* < T_1 \\ 2 & \text{if } T_1 \leq y_i^* < T_2 \\ \vdots & \text{and so on} \\ 7 & \text{if } T_6 \leq y_i^* \end{cases}$$

(2)

with the $T_j$ being parameters such that $T_1 < T_2 < \cdots < T_6$. Thus, the range of credit quality $y_i^*$ is partitioned into 7 mutually exclusive and exhaustive intervals, and the numerical rating $Y_i$ indicates the interval into which a particular bond’s credit quality falls. Consider the following model for $y^*$,

$$y_i^* = g(X_i, S_i)$$

(3)

where $X_i = [F_i, B_i, MFOI_i]$\footnote{The vector $X$ is assumed to be exogenous throughout. Intuitively, and as one might have expected, some information contained in $S$, e.g., manager’s ability, may also drive institutional investors’ investment decision, implying that $MFOI$ is endogenous. The problem of endogeneity can be handled, for example, using the control function approach proposed by Blundell and Powell (2004) provided with a valid exclusion restriction. However, since the main focus of this paper is to predict rating rather than making causal inference, I abstract away from the problem of endogenous regressors.}. For each bond $i$, assume $F_i = (F_{i1}, F_{i2}, \cdots) \in \mathcal{R}^F$ is a vector of firm characteristics, $B_i = (B_{i1}, B_{i2}, \cdots) \in \mathcal{R}^B$ is a vector of bond characteristics, and $MFOI_i$ is the aforementioned measure of conflicts of interest. The model allows for both continuous as well as discrete characteristics. $S_i$ is a scalar error term summarizing private information which the rating agency takes into account but is not observed to the econometrician. Instead of making distributional assumption about $S$, I make an index assumption that will be provided below.

\footnote{The numerical rating matches the seven ordinal rating categories: Aaa = 7, Aa = 6, A = 5, Baa = 4, Ba = 3, B = 2, and C = 1 (from the highest credit quality to the lowest).}
With $Y_{ik} = 1\{Y_i = k\}$, a key object of estimation interest is $P_{ik} \equiv Pr(Y_{ik} = 1|X_i)$, which is
the probability that a bond will be rated in category $k$ conditional on the vector of explanatory
variables. In terms of prediction, the model would naturally predict the rating to be the category
that associated with the highest probability. In the extant literature, one prevailing parametric
approach to estimate the rating probability is the ordered-probit model (OPM), which assumes

$$y_i^* = X_i\beta_0 + S_i, \quad S \sim N(0, 1)$$

(4)

Given the normality of $S$, the rating probabilities can be written as

$$P_{ik} \equiv Pr(Y_{ik} = 1|X_i) = \begin{cases} 
\Phi(T_k - X_i\beta_0) & \text{if } k = 1 \\
\Phi(T_k - X_i\beta_0) - \Phi(T_{k-1} - X_i\beta_0) & \text{if } k = 2, 3, \ldots, 6 \\
1 - \Phi(T_{k-1} - X_i\beta_0) & \text{if } k = 7 
\end{cases}$$

where $\Phi$ is the cdf of a standard normal random variable. The slope coefficients $\beta_0$ and the cutoff
points $T_k, k = 1, \cdots, 6$, can be estimated through maximizing the following log-likelihood function,$^7$

$$Q = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{7} Y_{ik} \ln(P_{ik})$$

(5)

However, if the assumption in (4) is incorrectly imposed, the resulting estimator is typically
inconsistent and will not perform well at prediction. Since the true rating process is unobserved
by the public, it is unclear how different sources of information are used by the CRA to assign
ratings. Putting linear and additive restriction as in (4) seems much too strong because firm and
bond characteristics might jointly affect a bond’s credit quality $y^*$ in various ways.$^8$ Apart from
constraining the form of the rating function, the normality assumption on the error term $S$ is also
hard to justify in practice. For example, during the financial crisis, the error term is likely to have a
“fat” left-tail. In both regards, researchers want to switch to a model that relax these assumptions.

$^7$In the case when $y^* = X_i\beta_0 + c_0 + S$ in (4) and the variance of the homoskedastic error term is $\sigma^2$, identification
is up to location and scale: one can at most identify $\beta^* \equiv \beta_0/\sigma$, $T_k^* = (T_k - T_1)/\sigma$ (the “pseudo cutoff points”),
and $c^* = (c_0 - T_1)/\sigma$ (the “pseudo intercept”). See Amemiya (1981) for a discussion.

$^8$For example, an increase in the firm leverage ratio may have a much stronger impact on credit quality for senior
bond than for junior bonds. However, this interactive effect is precluded by (4).
In a more general nonparametric formulation,

\[ Pr(Y_{ik} = 1|X_i) = F_k(X_i), \text{ for } k = 1, 2, \cdots, 7 \]  

(6)

This specification imposes few restrictions on the form of the joint distribution of the data. Therefore, there is little room for misspecification\(^9\). However, when the dimension of \( X \) is large, the resulting estimator will have considerable variance due to the “curse of dimensionality.”

Recall from before that the vector of explanatory variable represent information from three distinct sources: firm characteristics (\( F \)), bond characteristics (\( B \)), and conflicts of interest (\( MFOI \)). In order to estimate the above probability with a moderately sized sample, I define the following two indices,

Firm Index: \( W_{Fi} \equiv F_1 \beta_{10}^F + F_2 \beta_{20}^F + \cdots \)  

(7)

Bond Index: \( W_{Bi} \equiv B_1 \beta_{10}^B + B_2 \beta_{20}^B + \cdots \)  

(8)

and make the model *semiparametric*:

\[ Pr(Y_{ik} = 1|X_i) = Pr(Y_{ik} = 1|W_{Fi}, W_{Bi}, MFOI_i), \text{ for } k = 1, 2, \cdots, 7 \]  

\[ \equiv P_k(W_{Fi}, W_{Bi}, MFOI_i) \]  

(9)

The “index assumption” in (9) implies that firm and bond characteristics enter the rating function as two separate indices, whereas our main variable of interest, MFOI, enter the rating function nonparametrically by itself. This index assumption is employed to reduce the estimation burden from dimensionality: under a purely nonparametric specification in (6), the rating probability depends on a potentially high-dimensional vector \( X_i \) whereas in the semiparametric model it depends on a 3-dimensional vector \( W_i = [W_{Fi}, W_{Bi}, MFOI_i] \).

The link function \( P_k(\cdot, \cdot, \cdot) \) in the proposed semiparametric model (SIM) will be estimated *nonparametrically*. Comparing to OPM, this specification permits ratings to depend more flexibly on firm characteristics, bond characteristics, and conflict of interest without constraining the functional form. Moreover, \( P_k(\cdot, \cdot, \cdot) \) is category-specific and permits the rating agency to have a different rating

---

\(^9\)The threshold parameters \( T_k \) are subsumed in the category-specific nonparametric rating function \( F_k \). Unless additional parametric assumptions are made on \( F_k \), these thresholds parameters cannot be identified.
function for each rating category.

3.2 Estimation strategy

Normalization  In the SIM described by equations (7)-(9), the slope coefficients in the firm and bond indices, \( \beta^F \) and \( \beta^B \), can only be identified up to location and scale\(^{10}\). In this regard, I normalize the coefficient of a firm characteristics \( F_1 \) and a bond characteristics \( B_1 \) to one and define the normalized indices as:

\[
\text{Normalized Firm Index: } V_{Fi} \equiv W_{Fi}/\beta_{10}^F = F_{1i} + F_i\theta_0^F \\
\text{Normalized Bond Index: } V_{Bi} \equiv W_{Bi}/\beta_{10}^B = B_{1i} + B_i\theta_0^B
\]

where \( \theta_0^F = (\beta_{20}^F/\beta_{10}^F, \beta_{30}^F/\beta_{10}^F, \cdots) \) and \( \theta_0^B = (\beta_{20}^B/\beta_{10}^B, \beta_{30}^B/\beta_{10}^B, \cdots) \). Note that the identifiable parameter \( \theta_0 = (\theta_0^F, \theta_0^B)' \) captures the relative impact of other firm and bond characteristics to \( F_1 \) and \( B_1 \), respectively. I refer \( \theta_0 \), the parameters of interest in the SIM, as the “index coefficients”.

In below, I develop an \( \sqrt{N} \)-normal estimator for \( \theta \).

Estimation  In the parametric case, the form of the log-likelihood is known and the model is estimated by maximizing the likelihood in (5). In contrast, the form of likelihood for SIM is unknown since I do not make distributional assumptions. Nevertheless, the index structure will help to develop an estimator for the likelihood. Specifically, I employ an estimator based on an extension of the approach in Klein and Sherman (2002) which maximizes the following estimated log-likelihood function:

\[
\hat{Q}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \tau_i \{ \sum_{k=1}^{7} Y_{ik} \text{Ln}(\hat{P}_k(V_{Fi}, V_{Bi}, MFOI_i)) \} \]

(11)

where \( \hat{P}_k(V_{Fi}, V_{Bi}, MFOI_i) \) is a semiparametric estimator of the rating probability defined in (9). Note that both \( V_F \) and \( V_B \), as shown in (10), are functions of \( \theta \). \( \tau_i \) is a trimming function that removes observations for which densities become too small\(^{11}\).

\(^{10}\) Since the functional form of \( P_k(\cdot) \) in (9) is not specified, conditioning on the original index \( W_{Fi}, W_{Bi}, MFOI_i \) and a linear transformation of them deliver the same amount information on ratings. Therefore, without some normalization, the limiting log-likelihood function cannot be uniquely maximized at the true parameters, which is necessary for identification.

\(^{11}\) This trimming function will be defined formally in the next section.
The rating probability \( P_k(V_{Fi}, V_{Bi}, MFOI_i) \) describes the conditional probability that bond \( i \) is assigned rating \( k \) given the three indices. Intuitively, this probability can be approximated by the percentage of “similar” bonds that are assigned the same rating. In practice, econometricians use kernels, which is a function \( K_h(\cdot) \) that is symmetric and integrated to one, to take the average around a small neighborhood \( h \) around bond \( i \). Formally, with \( V_i = [V_{Fi}, V_{Bi}, MFOI_i] \), this probability is estimated as,

\[
\hat{P}_k(V_{Fi}, V_{Bi}, MFOI_i) = \frac{1}{N-1} \sum_{j \neq i} Y_{jk} K_h(V_i - V_j) = \hat{f}(V_i)/\hat{g}(V_i)
\]

(12)

where \( \hat{g}(V_i) \) is the estimated density of the three-dimensional index \( V_i \) and

\[
K_h(V_i - V_j) = \frac{1}{h_1 h_2 h_3} K \left( \frac{V_{Fj} - V_{Fi}}{h_1} \right) K \left( \frac{V_{Bj} - V_{Bi}}{h_2} \right) K \left( \frac{MFOI_j - MFOI_i}{h_3} \right)
\]

(13)

In this paper I use a second-order Gaussian kernel: \( K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2) \) and \( h \equiv (h_1, h_2, h_3)' \) is a bandwidth parameter that goes to zero as sample size approaches infinity. Intuitively \( h \) controls the number of similar bonds that are included in the kernel-weighted average; a smaller bandwidth \( h \) will lead to a smaller bias (only the \( j \)'s that are very similar to \( i \) are included) but larger variance (fewer observations are included).

4 Large Sample Results

To establish asymptotic results it is necessary to control the bias in the underlying density estimators defined in (12). Compared to the single-index ordered model considered by Klein and Sherman (2002), the estimator developed here has an even slower convergence rate because the underlying density depends on a three-dimensional index vector (e.g., curse of dimensionality). I obtain bias reduction first by employing a “recursive differencing” strategy proposed by Shen and Klein (2019). As a second source of bias reduction, I exploit a “residual” property of expected semiparametric probability functions. To describe these bias controls rigorously, I require definitions and assumptions which is introduced below.
4.1 Definitions and Assumptions

The aforementioned “recursive differencing” estimator is formally defined in D.4, with the underlying index and density estimators defined in D.1-D.3. This estimator differs from the estimator in (12) by an adjustment term $\hat{\Delta}_{ji}^k$. As described in equations (18)-(19), this term is constructed by first estimating the rating probabilities for bond $i$ and bond $j$ based on the standard Nadaraya-Watson estimator, and then take the difference.

The intuition for this bias correction is the following: the probability that bond $i$ receives rating $k$ is approximated by a (weighted) local average of $Y_{jk}$ within the bandwidth $h$, where the kernel function $K_h(V_j - V_i)$ assigns a higher weight to bond $j$ that are “similar” to $i$. However, in finite sample, the estimator has a bias because each $j$, though similar, is still different from $i$. From standard results$^{12}$, the bias of the estimator goes to zero at the rate of $h^2$ as $h \to 0$. After removing an estimate of the differences between $j$ and $i$, the kernel-weighted average of $Y_{jk} - \hat{\Delta}_{ji}^k$ is a better estimate of $P_k(V_{Fi}, V_{Bi}, MFOI_i)$. Specifically, Shen and Klein (2019) proves the bias of this “recursive differencing” estimator $\hat{P}_k^*(V_{Fi}, V_{Bi}, MFOI_i)$ goes to zero at the rate of $h^4$, whereas the variance stays the same order at $1/Nh^3$.

D.1 Firm and Bond Index: Let $F_1$ denote a continuous firm characteristics and $F'$ denotes the vector of firm characteristics other than $F_1$ such that $F_i \equiv [F_{1i}, F'_i]$ for. $B_1$ and $B'$ are defined similarly for the vector of bond characteristics. I define the firm and bond index as

$$
\text{Firm Index: } W_{Fi} \equiv F_{1i} \beta_{F10}^F + F'_i \beta_{F0}^F \\
\text{Bond Index: } W_{Bi} \equiv B_{1i} \beta_{B10}^B + B'_i \beta_{B0}^B
$$

Under A.1-A.3 below, I define the normalized indices as:

$$
\text{Normalized Firm Index: } V_{Fi} \equiv W_{Fi}/\beta_{F10}^F = F_{1i} + F'_i \theta_{0}^F \\
\text{Normalized Bond Index: } V_{Bi} \equiv W_{Bi}/\beta_{B10}^B = B_{1i} + B'_i \theta_{0}^B
$$

where $\theta_{0}^F \equiv (\beta_{F20}/\beta_{F10}, \beta_{F30}/\beta_{F10}, \cdots)$ and $\theta_{0}^B \equiv (\beta_{B20}/\beta_{B10}, \beta_{B30}/\beta_{B10}, \cdots)$ are referred as the index coefficients.

$^{12}$See, for example, the lecture notes of Hansen (2009) for a discussion.
D.2 Kernels: Let \( V_j \equiv (V_{Fj}, V_{Bj}, MFOI_j) \) denotes the value of the normalized index for observation \( j \) and \( v \equiv (v_F, v_B, MFOI) \) denote a fixed point of interest. Define a multivariate kernel function

\[
K_h(V_j - v) = \frac{1}{h_1 h_2 h_3} K\left(\frac{V_{Fj} - v_F}{h_1}\right) K\left(\frac{V_{Bj} - v_B}{h_2}\right) K\left(\frac{MFOI_j - MFOI}{h_3}\right)
\]

where \( K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right) \) and \( h \equiv (h_1, h_2, h_3)' \) is a bandwidth parameter that goes to zero as sample size approaches infinity.

D.3 Density Estimator: Let \( g(v) \) denotes the joint density of the three-dimensional index \( V_j \) at a fixed point \( v \), a (leave-one-out) kernel-weighted density estimator \( \hat{g}(v) \) is defined as

\[
\hat{g}(v) = \frac{1}{N-1} \sum_{j=1}^{N-1} K_h(V_j - v)
\]

D.4 Estimated Probability Referring to D.2-D.3, an initial estimator for the probability that a bond with characteristics \( v = (V_F, V_B, MFOI) \) will be rated in category \( k \), denoted \( P_k(v) \), is defined as

\[
\hat{P}_k(v) \equiv \frac{1}{N-1} \sum_{j=1}^{N-1} Y^k_j K_h(V_j - v) \hat{g}(v)
\]

Based on this initial estimator at \( v \), a recursive-differencing estimator for \( P_k(v) \), in the spirit of Shen and Klein (2019), is defined as

\[
\hat{P}^*_k(v) = \frac{1}{N-1} \sum_{j=1}^{N-1} [Y^k_j - \hat{\Delta}^k_j(v)] K_h(V_j - v) \frac{\hat{g}(v)}{\hat{g}(v)} = \hat{f}(v) / \hat{g}(v)
\]

where \( \hat{\Delta}^k_j(v) = \hat{P}^k(V_j) - \hat{P}^k(v) \) is a bias-correction adjustment.

D.5 Trimming Functions: Let \( W_i^d \) denote the \( d \)-th column of a \( D \)-dimensional continuous vector \( W_i \). Define \( \hat{\tau}_{id} \equiv 1\{ a_d < \hat{W}_i^d < \hat{b}_d \} \) and \( \hat{\tau}_i = \prod_{d=1}^D \hat{\tau}_{id} \), where \( a_d, \hat{b}_d \) are, respectively, lower and upper sample quantiles for \( W_i^d \). When \( W_i^d = X_i \), I refer to \( \tau_{ix} \) as \( X \)-trimming; With \( \hat{V}_i \) as the estimated index, when \( W_i^d = \hat{V}_i \), I refer to \( \tau_{iv} \) as index trimming.
D.6 First- and Second-stage estimator: Based on D.1-D.5, I define:

\[ \hat{\theta}_1 = \arg\max_{\theta} Q_1(\theta), \quad Q_1(\theta) = N^{-1} \sum_{i=1}^{N} \tau_{i1}(\sum_{k=1}^{7} Y_i^k \ln(\hat{P}^*_k(V_i))), \] (20)

\[ \hat{\theta}_2 = \arg\max_{\theta} Q_2(\theta), \quad Q_2(\theta) = N^{-1} \sum_{i=1}^{N} \tau_{i2}(\sum_{k=1}^{7} Y_i^k \ln(\hat{P}^*_k(V_i))), \] (21)

where \( V_i \equiv (V_{Fi}(\theta), V_{Bi}(\theta), MFOI_i) \) as defined in D.1. The second stage estimator \( \theta_2 \) differs from the first stage in that the trimming function \( \tau_{i\theta} \) is based on the estimated index: use the definition in D.5 with \( W_i = (V_{Fi}(\hat{\theta}_1), V_{Bi}(\hat{\theta}_1), MFOI_i) \).

The proofs for asymptotic properties of the recursive differencing estimator also exploit a residual-like property of the derivative (with respect to the parameters \( \theta \)) of the expected semiparametric probability function, with this derivative having conditional expectation of zero when evaluated at the true parameter values \( \theta_0 \). By using this property, I can further control for the bias in the gradient to the log-likelihood function in (11), which is essential to establishing asymptotic normality. To this end, a two-stage estimation procedure described in D.6 is needed: I first estimate the model under \( X\)-trimming defined in D.5. The resulting estimates, \( \hat{\theta}_1 \), are employed to obtain estimated indices. A second-stage estimator, \( \hat{\theta}_2 \), is then obtained from re-estimating the model with trimming based on estimated indices. This residual-like property has been employed by Klein and Shen (2010) as bias controls.

How does the trimming function affect the convergence property of the proposed estimator? Trimming impact the variance of the estimator because fewer observations are used, though it does not effect the variance order. For the bias, note that trimming gives protection against small density denominators and hence “stabilize” the probability estimator. In addition, the order of the bias increases (i.e, the bias of the underlying kernel-type density estimator is \( O(h^2) \) at an interior point and \( O(h) \) at the boundary) near the boundary of the support for variables on which we trim. Trimming resolves that as well. The trimming function restricts estimation in all stages to employ only the middle 99 percentile (\( \hat{a}_d = 0.005, \hat{b}_d = 0.995 \) for all \( d \)) of the data. Such trimming thresholds, as explained below, are selected based on simulation evidence.

To obtain convergence properties for the proposed estimator in D.4 and asymptotic normality for the second-stage estimator of the index parameters defined in D.6, I make the following assumptions.

A.1 Data: For each \( k \) in 1, \( \cdots, 7 \), the vector \( (Y_i^k, X_i) \) is \( iid \) over \( i \), and takes on values in a
compact and finite support\textsuperscript{13}. The columns of \( X_i = [F_i, B_i, MFOI_i] \) are linearly independent with probability one.

**A.2 The Error Term:** The error term \( S_i \) is conditionally independent of \( X_i \); \( E[S_i|X_i] = 0 \), and independent across \( i \).

**A.3 Continuous firm and bond characteristics** Referring to the firm and bond index defined in D.3, I require the index coefficient of \( F_1 \) and \( B_1 \) are nonzero: \( \beta_{F10} \neq 0, \beta_{B10} \neq 0. \)

**A.4 Index Assumption** Referring to the normalized index defined in D.1, let \( V_i(\theta_0) \equiv [V_{F1_i}, V_{B1_i}, MFOI_i] \), the following index assumption is assumed to hold for all \( i \) and \( k \):

\[
E[Y^k_{i} = 1|X_i] = E[Y^k_{i} = 1|V_i(\theta_0)] \tag{22}
\]

**A.5 Parameter Space** The vector of true parameters values \( \theta_0 \equiv [\theta_{F0}, \theta_{B0}] \) for the model in lies in the interior of a compact parameter space, \( \Theta \).

**A.6 Conditional Densities** Let \( g(v|y,x) \) denote the density of the index \( V_i(\theta) \) defined in A.4 conditioning on \( Y_i = y \) and \( X_i = x \). Denote \( \nabla^d g(t|y,x) \) as the partial or cross partial derivatives up to order \( d \). With \( g \) defined on a compact support, I assume \( g > 0 \) and \( \nabla^d g(t|y,x) \) to be uniformly bounded for \( d = 0, 1, 2, 3 \) on the interior of its support.

**A.7 Bandwidth Parameter** Referring to the kernel estimator defined in D.2, the bandwidth parameter \( h \to 0 \) as \( N \to \infty \). Specifically, with \( h = (h_1, h_2, h_3) \), I choose \( h_z = 0.97\sigma_zN^{-r} \) according to Silverman (1982)\textsuperscript{14} where \( \sigma_z \) is the standard deviation of the three indices respectively \( (z = 1, 2, 3) \) and \( r \) is a parameter that affects the rate that \( h \) goes to zero. In this paper, \( r = 1/10.01 \).

The choice of \( r = 1/10.01 \) in A.7 requires a bit explanation. To ensure asymptotic normality of the proposed estimator, we need to show (i) the estimated Hessian converges uniformly to the expected Hessian at the true parameter values and (ii) the gradient is centered at zero (e.g no asymptotic bias) and also converges to the true gradient. As will be shown below, the uniform convergence of the Hessian matrix will generate a upper bound \( (r > 1/10) \), whereas the bias reduction mechanism

\textsuperscript{13}X \textsuperscript{*} denotes variables like firm’s asset, leverage ratio, etc, which are naturally bounded from the above

\textsuperscript{14}To be specific, let \( h_z = c_zN^{-r} \), Silverman (1982) shows that the MSE-optimal \( c = (\frac{1}{\sqrt{\pi}})^{(4+\theta)}\sigma_z \). In the case with three indices, the constant is roughly 0.97.
will generate a lower bound \((r > 1/12)\). Within these bounds, I choose \(r = 1/10.01\) which is close to the optimal bandwidth \((r_o = 1/7\) in the case of a three-index model, which minimizes MSE of the kernel estimates of conditional probability), according to Hansen (2009).

Assumptions A.1-A.5 define the index model that I propose to estimate. An index formulation of low dimension is important for obtaining reasonable results in finite samples. In A.3, I assume that the firm and bond indices satisfies the identification conditions in Ichimura and Lee (1991). Specifically, each index must contain at least one continuous variable that belongs to the model in the statistical sense. In addition, I require densities for continuous variables and the indices must be sufficiently smooth, as implied by A.6. The smoothness conditions are standard in the literature and have been discussed in Klein and Spady (1993).

### 4.2 Asymptotic Theorems

Denote \(\hat{\theta}_1\) and \(\hat{\theta}_2\) as the first and second stage estimators respectively defined in D.6. In this section I provide and discuss the asymptotic properties for these two estimators. The Appendix contains formal proofs for all required intermediate lemmas and the main theorems given below. In what follows, I first establish consistency for estimators in both stages using standard uniform convergence arguments. Then, I turn to the proofs for asymptotic normality of \(\hat{\theta}_2\).

For \(m = 1, 2\), recall that the estimator \(\hat{\theta}_m\) maximizes the estimated log-likelihood function \(\hat{Q}_m(\theta)\) defined in D.6 for both stages. From Lemma C.3, \(\sup_{\theta}|\hat{Q}_m(\theta) - \hat{Q}(\theta)| \overset{p}{\to} 0\), where \(\hat{Q}(\theta)\) is obtained from \(\hat{Q}_m(\theta)\) by replacing all estimated functions with their probability limits. From standard argument, \(\hat{Q}(\theta)\) converges uniformly to its expectation \(E[\hat{Q}(\theta)]\). As I show in the Appendix that \(E[\hat{Q}(\theta)]\) is uniquely maximized at \(\theta_0\), so the estimators in both stages are consistent.

**Theorem 4.1 (Consistency).** Under A.1-A.7 and with \(1/12 < r < 1/10\),

\[
|\hat{\theta}_1 - \theta_0| = o_p(1); \quad |\hat{\theta}_2 - \theta_0| = o_p(1)
\]

**Proof.** See Appendix \(\square\)

To derive the asymptotic distribution of the second stage estimator \(\hat{\theta}_2\), recall that \(\hat{\theta}_2\) is obtained by maximizing the log-likelihood function \(Q_2\) defined in D.6 with the *index-trimming*. Let \(G(\theta)\) and \(H(\theta)\) denote the Gradient and Hessian matrices, which are the first and second derivative of
the log-likelihood function in (21):

\[ \hat{G}(\theta) = \nabla_{\theta'} \hat{Q}_2(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{7} \frac{Y_k}{P_k^*} \nabla_{\theta} \hat{P}_k^* \]  \hspace{1cm} (23)

\[ \hat{H}(\theta) = \nabla_{\theta'} \hat{Q}_2(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{7} \left( \frac{Y_k}{P_k^*} \nabla_{\theta} \hat{P}_k^* - \frac{Y_k}{P_k^*} \nabla_{\theta} \hat{P}_k^* \right) \]  \hspace{1cm} (24)

To simplify presentation, I suppress the trimming function \( \tau_i^\alpha \) and denote the double summation \( \sum_{i=1}^{N} \sum_{k=1}^{7} \) as \( \sum_{i,k} \). From a Taylor expansion of the estimated Gradient at \( \hat{\theta}_2 \),

\[ \hat{G}(\hat{\theta}_2) = \hat{G}(\theta_0) + \hat{H}(\theta^+)(\hat{\theta}_2 - \theta_0) \quad \theta^+ \in (\theta_0, \hat{\theta}_2) \]  \hspace{1cm} (25)

Since the estimated gradient must be zero at \( \theta_0 \), the above expression simplifies to

\[ \sqrt{N}(\hat{\theta}_2 - \theta_0) = -\hat{H}(\theta^+)^{-1} \sqrt{N} \hat{G}(\theta_0) \quad \theta^+ \in (\theta_0, \hat{\theta}_2) \]  \hspace{1cm} (26)

By Lemma 5 in the Appendix, the estimated Hessian \( \hat{H}(\theta^+) \) will uniformly converge to \( H_0 \equiv E[H(\theta_0)] \) when the bandwidth parameter \( r < 1/10 \). The estimated gradient \( \hat{G}(\theta_0) \) has an asymptotic expansion as follow:

\[ \sqrt{N}\hat{G}(\theta_0) = N^{-1/2} \sum_{i,k} \frac{Y_k - \hat{P}_k^*}{P_k^*} \nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0} \]  \hspace{1cm} (27)

\[ = N^{-1/2} \sum_{i,k} \frac{Y_k - P_k}{P_k^*} \nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0} + N^{-1/2} \sum_{i,k} \frac{P_k - \hat{P}_k^*}{P_k^*} \nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0} \]

\[ = \underbrace{N^{-1/2} \sum_{i,k} \frac{Y_k - P_k}{P_k^*} \nabla_{\theta} P_k |_{\theta = \theta_0}}_{A} - \underbrace{N^{-1/2} \sum_{i,k} (Y_k - P_k) \left( \frac{\nabla_{\theta} P_k |_{\theta = \theta_0}}{P_k} - \frac{\nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0}}{P_k^*} \right)}_{B} \]

\[ = \underbrace{N^{-1/2} \sum_{i,k} \frac{P_k - \hat{P}_k^*}{P_k} \nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0}}_{A_1} - \underbrace{N^{-1/2} \sum_{i,k} (P_k - \hat{P}_k^*) \left( \frac{\nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0}}{P_k} - \frac{\nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0}}{P_k^*} \right)}_{B_1} \]

\[ - \underbrace{N^{-1/2} \sum_{i,k} (P_k - \hat{P}_k^*) \left( \frac{\nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0}}{P_k} - \frac{\nabla_{\theta} \hat{P}_k^* |_{\theta = \theta_0}}{P_k^*} \right)}_{B_2} \]

\[ = A_1 + B_1 + o_p(1) \]
In the Appendix, I formally show that each term in the asymptotic expansion, except $A_1$, is $o_p(1)$. The aforementioned residual-like property plays a critical role in showing $B_1 = o_p(1)$. Combining (26) and (27), $\sqrt{N}(\hat{\theta}_2 - \theta)$ has the following asymptotic linear representation

$$\sqrt{N}(\hat{\theta}_2 - \theta) = -H_0^{-1}N^{-1/2}\sum_{i,k} \frac{Y_k - P_k}{P_k} \nabla_{\theta} P_k|_{\theta = \theta_0} + o_p(1)$$

(28)

where $H_0 \equiv E[H(\theta_0)]$ is the probability limit of the Hessian matrix evaluated at $\theta_0$. After applying the Lindberg-Levy CLT, the second stage estimator $\hat{\theta}_2$ defined in D.6 is $\sqrt{N}$–normal,

**Theorem 4.2 (Normality).** Assume $A.1$-$A.7$ and with the bandwidth parameter $1/12 < r < 1/10$,

$$\sqrt{N}(\hat{\theta}_2 - \theta_0) \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = H_0^{-1}E[G'(\theta_0)G(\theta_0)]H_0^{-1}$ with $H_0 = E[\sum_{k=1}^7 \left( \frac{Y_k}{P_k} \nabla_{\theta} P_k|_{\theta = \theta_0} - \frac{Y_k}{P_k} \nabla_{\theta} P_k|_{\theta = \theta_0} \right)]$, $G(\theta_0) = \sum_{k=1}^7 \frac{Y_k}{P_k} \nabla_{\theta} P_k|_{\theta = \theta_0}$, and $P_k = Pr(Y_{ik} = 1|V_i(\theta_0))$.

*Proof.* See Appendix

5 Results

5.1 Simulation evidence

Before applying the proposed estimator to predict corporate bond ratings, Monte Carlo experiments are used to investigate the finite sample performance of different estimators. As discussed earlier, the main advantage of the proposed multi-index semiparametric model is the ability to flexibly capture the interactive effects among covariates. As such, I consider the following data-generating process.

The latent variable model in the simulations has the form

$$y^* = \beta_0(\delta_0 + X_1 + \theta_0 X_2)e^{X_3^2} + U$$

(29)

where $y^*$ is the latent response variable, $X_1, X_2, X_3$ are explanatory variables, $\delta_0, \beta_0, \text{ and } \theta_0$ are unknown parameters, and $U$ is an error term, independent of $X = (X_1, X_2, X_3)^{15}$. The observed

\footnote{More specifically, $U$ is generated from a $\chi^2(1)$ distribution, standardized to have mean zero and unit variance. $X_1, X_3 \sim N(0, 1)$ and $X_2$ is a standardized $\chi^2(1)$.}
data consists of \((Y, X)\) where

\[
Y = 0\{-\infty < y^* \leq c_0\} + 1\{c_0 < y^* \leq c_1\} + 2\{c_1 < y^*\} \tag{30}
\]

In each simulation, \(\delta_0 = \beta_0 = \theta_0 = 2\). The threshold points \(c_0\) and \(c_1\) are the 33th and 66th percentiles of the \(y^*\) distribution. Of estimation interest is the slope parameter \(\theta_0\).

The object of the simulations is two folds. First, I compare the proposed multiple-index semi-parametric estimator (SIM-M) with three other estimators in terms of precision. Second, I evaluate the performance of these estimators under different bandwidth \((r \text{ defined in A.7})\) and trimming parameters \((t_i \text{ defined in D.5})\). The three other estimators are (1) ordered probit model (OPM), (2) Semiparametric single index model (SIM-1) considered by Klein and Sherman (2002), and (3) Semiparametric Sieve estimator (SIM-Sieve) considered by Chen (2007). The Sieve estimator also has a multi-index structure as SIM-M.

More specifically, I compare the RMSE and predictive accuracy of the four aforementioned estimators in six simulation designs, with the results presented in Table 2. These six simulations are generated by the permutation of three different trimming thresholds (90 percentile, 95 percentile, and 99 percentile) and two bandwidth parameters: a “small” bandwidth with \(r = 1/11.99\) and a “large” bandwidth with \(r = 1/10.01\). The sample size is 2000 with a replication of 50.

Across different designs, I find the proposed SIM-M model generates the smallest RMSE, followed by the Sieve estimator with the same model. The magnitude of bias is comparable between SIM-M and SIM-Sieve, yet the variance of SIM-M is much smaller. Given the DGP in (29), apparently both the single-index model (SIM-1) and ordered probit model (OPM) suffers from misspecification bias. Therefore it is not surprising to find these two models underperform in terms of RMSE and prediction accuracy. As for the choice of bandwidth and trimming parameters, I find a large bandwidth \(r = 1/10.01\) and a high trimming threshold, 99 percentile, perform the best. As discussed in the last section, \(1/12 < r < 1/10\) ensures the proposed SIM-M estimator has a asymptotic normal distribution. Within the range of such permissible values, \(r = 1/10.01\) would minimize the MSE of the estimated conditional probability. However, there is no theoretical explanation on why a higher trimming threshold performs better.
Figure 1: Classification accuracy between SIM and OPMs

(a) In-sample

Note: The unshaded (blank), horizontally-lined, and dark-solid bars in each graph correspond to the percentage of correct prediction (e.g., 0-1 loss) from (1) OPM/MFOI: ordered probit model without MFOI, (2) OPM: ordered probit model with MFOI, and (3) SIM: the 3-index semiparametric model with MFOI. In the upper graph, I report the PCP for in-sample prediction: e.g., I estimate the three models using the rating data in year $t$ and then predict the rating in the same year. In the lower graph, I report the PCP for out-of-sample prediction: e.g., I estimate the three models using the rating data in year $t$ and then predict the rating in $t+1$ year.
### Table 2: Simulation evidence

<table>
<thead>
<tr>
<th></th>
<th>Large Bandwidth $r = 1/10$</th>
<th>Small Bandwidth $r = 1/11.99$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trimming</strong></td>
<td><strong>TRUE</strong></td>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td>SIM-1</td>
<td>0.90</td>
<td>2</td>
</tr>
<tr>
<td>SIM-M</td>
<td>2</td>
<td>1.938</td>
</tr>
<tr>
<td>SIM-Sieve</td>
<td>2</td>
<td>2.471</td>
</tr>
<tr>
<td>OPM</td>
<td>2</td>
<td>2.612</td>
</tr>
<tr>
<td>SIM-1</td>
<td>0.95</td>
<td>2</td>
</tr>
<tr>
<td>SIM-M</td>
<td>2</td>
<td>1.947</td>
</tr>
<tr>
<td>SIM-Sieve</td>
<td>2</td>
<td>2.471</td>
</tr>
<tr>
<td>OPM</td>
<td>2</td>
<td>2.547</td>
</tr>
</tbody>
</table>

DGP: $Y = \begin{cases} 0 & \{ -\infty < y^* \leq c_0 \} \\ 1 & \{ c_0 < y^* \leq c_1 \} \\ 2 & \{ c_1 < y^* \} \end{cases} + \beta_0 (X_1 + \theta_0 X_2 - \alpha_0) e^{X_3^2} + U$. The parameter being estimated is $\theta$, which is set to be $2$ in all designs. The two bandwidths are chosen to be the two endpoints of the permissible values $1/12 < r < 1/10$. Such a range of $r$, as shown in Theorem 4.2 assures asymptotic normality. SIM-1 is the semiparametric single-index model considered by Klein and Sherman (2002). SIM-sieve estimates the same model as SIM-M, except that the conditional probability defined in (12) is estimated via 3rd-order series approximation: $\hat{P}_{\text{sieve}}(v) = \sum_{j=0}^{3} v^j \beta_j$, with $v = (x_1 + \theta x_2, x_3)'$.

### 5.2 Empirical illustration: predicting bond ratings

Despite the generality with which the proposed semiparametric framework accounts for the influence of explanatory variables, the complexity of the estimation procedure raises the question of whether, and to what extent, these features can be addressed by a simpler model. To answer this question, I first compare the proposed semiparametric model (SIM-M) with two parametric alternatives: (1) a benchmark ordered probit model (OPM) and (2) an ordered probit model excluding MFOI (OPM/MFOI). A comparison between OPM and OPM/MFOI unveils the marginal predictive power of incorporating conflicts of interest, whereas a comparison between OPM and SIM-M investigates the predictive improvement from relaxing functional/distributional form assumptions.

I evaluate predictive accuracy according to a 0-1 loss function\footnote{Let $y_i$ be the true rating and $\hat{y}_i$ be the predicted rating for a bond $i$, the 0-1 loss function $L_{0-1} = \frac{1}{N} \sum_{i} 1(y_i = \hat{y}_i)$. The loss function essentially computes the percentage of ratings that are correctly predicted.} in 16 training samples (2001-2016) and 15 holdout samples (2002-2016) are reported in Figure 1 by year. Parameters are estimated using the rating data in year $t$ (the training sample, with $t = 2001, \ldots, 2015$), whereas out-of-sample prediction is performed on ratings in year $t + 1$ (the holdout sample). First of all, I find incorporating...
MFOI boosts the predictive powers uniformly as OPM outperforms OPM/MFOI statistically in all 31 samples by significant margins. Secondly, SIM-M also outperforms OPM in all 16 training samples. For the 15 holdout samples, SIM performs better in 9, with the largest 18.1% prediction improvement in 2007 (from 54.7% to 64.6%).

The second finding implies that relaxing functional-form and distributional assumptions improves prediction, which should not be surprising. For example, during the financial crisis, the distribution of the error term $S$ may deviate from a normal distribution significantly and display a “fat” left-tail, capturing the elevated default risk. SIM-M takes the unknown distribution of error term into account in a way that OPM cannot. Consequently, the misspecification error should result in a considerable difference in the OPM and SIM-M estimates and therefore harm OPM’s predictive power.

To support this argument empirically, in Table 3, I report parameter estimates from the two models, termed $\theta_{SIM-M}$ and $\theta_{OPM}$, for each year $t$ and examine the relationship between predictive power and the “divergence” of the two estimates over time. Specifically, I use a revised version of the $L_1$-norm of $\theta_{SIM-M} - \theta_{OPM}$, termed $L_1$, to measure the degree of divergence\(^{17}\). As shown in the last two columns of Table 3, when $\theta_{OPM}$ and $\theta_{SIM-M}$ diverge substantially (e.g., $L_1$ increases), the out-of-sample predictive improvement in SIM-M, termed $\Delta\%$, tends to increase. The pearson correlation coefficient between $L_1$ and $\Delta\%$ is 0.688\(^{18}\). This strong, positive co-movement between the two metric intuitively suggests that SIM-M boosts the predictive power by correctly specifying the underlying rating model.

### 5.3 Additional predictive results

The previous section examines the overall predictive performance based on the 0-1 loss function and the comparison is limited to two parametric models. In this section, I expand the forecasting horse-race in three important ways. First, two other semiparametric methods studied in the Monte Carlo, a semiparametric single index model (SIM-1) considered by Klein and Sherman (2002) and a Semiparametric Sieve estimator (SIM-Sieve) considered by Chen (2007), are considered. Second, predictive accuracy is evaluated for each rating category because in practice, forecasting the ratings

\(^{17}\)The $L_1$-norm, or the taxicab/Manhattan norm, of a $d$-vector $x$ is: $\left| \sum_{k=1}^{d} [x^k_{SIM-M} - x^k_{OPM}] \right| = \sum_{k=1}^{d} |x^k_{SIM-M} - x^k_{OPM}|$, where $x^k$ is the $k$-th argument of the vector $x$. In our case, I report the norm as a percentage: $L_1 = \frac{\sum_{k=1}^{d} |x^k_{SIM-M} - x^k_{OPM}|}{\sum_{k=1}^{d} x^k_{OPM}}$ to take the different scale in each dimension of $\theta$ into account.

\(^{18}\)Based on a univariate linear regression, I find that for a 1% increase in $L_1$ leads to a 0.2% increase in the out-of-sample predictive improvement of SIM. This effect is statistically significant at 1% significance level.
### Table 3: Relationship between Estimation Difference and Predictive Improvement

<table>
<thead>
<tr>
<th>Year</th>
<th>PROFIT</th>
<th>ASSET VOLATILITY</th>
<th>LEVERAGE</th>
<th>(x)</th>
<th>(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_{SIM}$</td>
<td>$\theta_{OPM}$</td>
<td>Diff1*</td>
<td>$\theta_{SIM}$</td>
<td>$\theta_{OPM}$</td>
</tr>
<tr>
<td>2002</td>
<td>35.2</td>
<td>22.7</td>
<td>43.2%</td>
<td>-4.2</td>
<td>-2.2</td>
</tr>
<tr>
<td>2003</td>
<td>26.4</td>
<td>14.3</td>
<td>59.3%</td>
<td>-1.8</td>
<td>-1.6</td>
</tr>
<tr>
<td>2004</td>
<td>35.7</td>
<td>19.6</td>
<td>57.9%</td>
<td>-3.2</td>
<td>1.0</td>
</tr>
<tr>
<td>2005</td>
<td>17.6</td>
<td>16.6</td>
<td>5.9%</td>
<td>-1.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>2006</td>
<td>22.1</td>
<td>18.2</td>
<td>19.2%</td>
<td>-4.2</td>
<td>-1.9</td>
</tr>
<tr>
<td>2007</td>
<td>15.5</td>
<td>12.9</td>
<td>18.8%</td>
<td>-0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>2008</td>
<td>20.3</td>
<td>18.7</td>
<td>8.1%</td>
<td>-2.5</td>
<td>-2.0</td>
</tr>
<tr>
<td>2009</td>
<td>16.5</td>
<td>10.4</td>
<td>45.3%</td>
<td>-4.7</td>
<td>-1.5</td>
</tr>
<tr>
<td>2010</td>
<td>34.5</td>
<td>29.9</td>
<td>14.1%</td>
<td>-0.8</td>
<td>-3.0</td>
</tr>
<tr>
<td>2011</td>
<td>23.0</td>
<td>14.3</td>
<td>46.9%</td>
<td>0.6</td>
<td>1.5</td>
</tr>
<tr>
<td>2012</td>
<td>20.4</td>
<td>20.0</td>
<td>1.7%</td>
<td>-0.7</td>
<td>-0.5</td>
</tr>
<tr>
<td>2013</td>
<td>26.9</td>
<td>27.3</td>
<td>-1.8%</td>
<td>-1.6</td>
<td>-0.6</td>
</tr>
<tr>
<td>2014</td>
<td>35.9</td>
<td>35.2</td>
<td>2.1%</td>
<td>-2.7</td>
<td>-2.1</td>
</tr>
<tr>
<td>2015</td>
<td>36.6</td>
<td>34.6</td>
<td>5.6%</td>
<td>-1.2</td>
<td>-0.8</td>
</tr>
<tr>
<td>2016</td>
<td>34.4</td>
<td>39.2</td>
<td>-12.8%</td>
<td>-0.7</td>
<td>-0.9</td>
</tr>
</tbody>
</table>

* Diff is the percentage difference in SIM and OPM estimates := \( \frac{\theta_{SIM} - \theta_{OPM}}{0.5(\theta_{SIM} + \theta_{OPM})} \)

** \( L_1^{**} \) is the \( L_1 \)-norm between \( \theta_{SIM} \) and \( \theta_{OPM} \) as a percentage: \( \frac{|\theta_{SIM} - \theta_{OPM}|}{3(\text{Diff1} + \text{Diff2} + \text{Diff3})} \)

*** \( \Delta_t\%^{***} \): the percentage predictive improvement in SIM over OPM in year \( t \). For example, in 2007, OPT correctly predicts 54.67% ratings, whereas SIM correctly predicts 64.57%. This is a \( \frac{64.57 - 54.67}{54.67} = 18.1\% \) improvement.

Note: In this table I report the year-by-year parameter estimates from SIM and OPM on three firm characteristics: PROFIT (Column 1-3), ASSET VOLATILITY (Column 4-6), and LEVERAGE (Column 7-9). Definitions of these variables can be found in Table 1. For each variable, I report the parameter estimate from the SIM, OPM, and their percentage difference. Recall that in SIM, \( \theta_{SIM} \) represents a variable’s relative impact to ASSET (identification up to location and scale). To make a fair comparison, I normalize the OPM estimates in the same way: \( \theta_{OPM} = \frac{\beta_{OPM}}{\beta_{ASSET}} \). In Column 10, I calculate average estimation difference in the three parameters. In Column 11, I report the predictive improvement as a percentage. The pearson correlation coefficient between the last two columns is 0.688.
Table 4: Out-of-sample prediction results from OPM, SIM1, SIM-M, and Sieve

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OPM</td>
<td>SIM1</td>
<td>SIM-M</td>
</tr>
<tr>
<td>L_{0-1}</td>
<td>0.547</td>
<td>0.541</td>
<td>0.646</td>
</tr>
<tr>
<td>L_{MSE,\alpha} &amp; 0.2</td>
<td>0.500</td>
<td>0.123</td>
<td>0.181</td>
</tr>
<tr>
<td>&amp; 0.4</td>
<td>0.537</td>
<td>0.196</td>
<td>0.235</td>
</tr>
<tr>
<td>&amp; 0.6</td>
<td>0.574</td>
<td>0.316</td>
<td>0.324</td>
</tr>
<tr>
<td>&amp; 0.8</td>
<td>0.611</td>
<td>0.486</td>
<td>0.449</td>
</tr>
</tbody>
</table>

Note: Within each panel, SIM-1 is the semiparametric single-index model considered by Klein and Sherman (2002). Sieve estimates the same model as SIM-M, except that the conditional probability defined in (12) is estimated via 3rd-order series approximation: \( \hat{P}_{sieve}(v) = \sum_{j=0}^{3} v^j \beta_j \). The first row \( L_{0-1} \) reports the percentage of correction prediction (the 0-1 loss). The second to the fifth row report the asymmetric MSE with different \( \alpha \) where a smaller number corresponds to a better prediction. Specifically, \( L_{MSE,\alpha} = \frac{1}{N} \sum_i \left[ \mathbb{1}(y_i < \hat{y}_i)(y_i - \hat{y}_i)^2 + (y_i \geq \hat{y}_i)\alpha(y_i - \hat{y}_i)^2 \right] \), \( \alpha = 0.2, 0.4, 0.6, 0.8, 1 \). This loss function is also employed by Christoffersen and Diebold (1996). Compared to the regular MSE loss, which sets equal penalties for over- and under-prediction, the asymmetric MSE sets a greater penalty for predictions that are exceedingly optimistic. The degree of asymmetry is governed by the parameter \( \alpha \) in that the smaller \( \alpha \) is, the greater the penalty for optimistic prediction.

Results on the out-of-sample prediction are reported in Table 4 wherein a variety of loss functions are employed. Within each year panel, the first four columns correspond to the four estimators being compared. The first row reports the percentage of correction prediction (the 0-1 loss). The second to the fifth row report the asymmetric MSE with different \( \alpha = 0.2, 0.4, 0.6, 0.8 \) where a smaller number corresponds to a better prediction. The last column reports the rank of SIM-M according to different loss functions, with 1 being the best. During the 2007-2009 financial crisis, SIM-M and SIM-Sieve perform significantly better than SIM-1 and OPM. This finding is consistent with the Monte Carlo study.

In Table 5 I evaluate the predictive accuracy of the four competing models by rating category.
To keep the presentation concise, I report only the 0-1 loss. The last column reports the average predictive accuracy during 2007-2009, with the bold case numbers highlighting the winner(s). It seems if predicting investment-grade bonds (with the exception of Aa) is most important, then SIM-M and SIM sieve are much better than OPM, but if predicting high-yield bonds (especially C) is most important, then actually OPM is better.

5.4 Time-series variation in rating standards

In this section, I employ the proposed SIM to investigate to what extent did Moody’s inflate ratings during the 2007-2009 financial crisis. I use a narrow definition of rating inflation similar to Alp (2013): If Moody’s use the standard they use in 2001 to assign ratings in later periods, then I ask whether a firm holding the same risk characteristics receives a higher or lower rating. A higher (predicted) rating implies deflated rating, whereas a lower (predicted) rating implies inflated rating.

To be specific, I estimate the SIM using the 2001 data only and “store” the index coefficient \( \hat{\theta}_{2001} \) along with the nonparametric rating probability function \( \hat{P}_{k,2001} \) for each rating category \( k \in \{1, \ldots, 7\} \). Then I predict what the earlier ratings would have been based on \( \hat{\theta}_{2001} \) and \( \hat{P}_{k,2001} \). Given the predicted rating \( \hat{Y}_i \) and the actual rating \( Y_i \), I claim the ratings in a particular year \( T \) are inflated if the mean discrepancy (\( D_T \)),

\[
D_T = \frac{1}{N_T} \sum_{i \in T} \hat{Y}_i - Y_i > 0
\]

Conversely, the ratings are deflated if \( D_T < 0 \), implying the standard for the year \( T \) is more stringent than the 2001 standard.

Tracing the mean discrepancy by year identifies time-series variation in Moody’s corporate bond rating standards. In Figure 2, I report the time series of \( D_t \) for each year following 2001 until 2016. The red reference line represents the 2001 standards, so data points above (below) the reference line imply the ratings on that year are inflated (deflated) relative to the 2001 standard. The blue band around the point estimates represent the 95% confidence interval.

There is a clear structural break in rating standards in 2009 where ratings started to become more stringent afterwards. Holding all risk attributes constant, the average ratings in 2009 become 0.4 notch lower compared to those in 2008. Moody’s rating standard became even more strict from 2012 to 2014 relative to its standard in 2009, which may reflects regulatory effects of the Dodd-Frank
Table 5: Out-of-sample prediction results from OPM, SIM1, SIM-M, and Sieve (by rating category)

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OPM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aaa</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Aa</td>
<td>87.82%</td>
<td>90.00%</td>
<td>48.57%</td>
<td><strong>75.46%</strong></td>
</tr>
<tr>
<td>A</td>
<td>31.63%</td>
<td>51.28%</td>
<td>63.57%</td>
<td>48.82%</td>
</tr>
<tr>
<td>Baa</td>
<td>79.64%</td>
<td>45.45%</td>
<td>62.61%</td>
<td>62.57%</td>
</tr>
<tr>
<td>Ba</td>
<td>0.00%</td>
<td>0.00%</td>
<td>31.46%</td>
<td><strong>10.49%</strong></td>
</tr>
<tr>
<td>B</td>
<td>42.67%</td>
<td>50.00%</td>
<td>0.00%</td>
<td>30.89%</td>
</tr>
<tr>
<td>C</td>
<td>53.85%</td>
<td>0.00%</td>
<td>38.46%</td>
<td><strong>30.77%</strong></td>
</tr>
<tr>
<td><strong>SIM1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aaa</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Aa</td>
<td>92.02%</td>
<td>65.45%</td>
<td>54.29%</td>
<td>70.59%</td>
</tr>
<tr>
<td>A</td>
<td>31.93%</td>
<td>75.00%</td>
<td>57.36%</td>
<td>54.76%</td>
</tr>
<tr>
<td>Baa</td>
<td>82.63%</td>
<td>38.46%</td>
<td>85.22%</td>
<td>68.77%</td>
</tr>
<tr>
<td>Ba</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>B</td>
<td>24.00%</td>
<td>66.67%</td>
<td>25.00%</td>
<td>38.56%</td>
</tr>
<tr>
<td>C</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td><strong>SIM-M</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aaa</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td><strong>33.33%</strong></td>
</tr>
<tr>
<td>Aa</td>
<td>86.97%</td>
<td>78.18%</td>
<td>45.71%</td>
<td>70.29%</td>
</tr>
<tr>
<td>A</td>
<td>60.24%</td>
<td>51.92%</td>
<td>52.71%</td>
<td><strong>54.96%</strong></td>
</tr>
<tr>
<td>Baa</td>
<td>88.02%</td>
<td>60.14%</td>
<td>87.39%</td>
<td>78.52%</td>
</tr>
<tr>
<td>Ba</td>
<td>0.00%</td>
<td>0.00%</td>
<td>12.36%</td>
<td>4.12%</td>
</tr>
<tr>
<td>B</td>
<td>26.67%</td>
<td>83.33%</td>
<td>19.23%</td>
<td><strong>43.08%</strong></td>
</tr>
<tr>
<td>C</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td><strong>Sieve</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aaa</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td><strong>33.33%</strong></td>
</tr>
<tr>
<td>Aa</td>
<td>90.76%</td>
<td>75.45%</td>
<td>31.43%</td>
<td>65.88%</td>
</tr>
<tr>
<td>A</td>
<td>59.34%</td>
<td>50.64%</td>
<td>38.76%</td>
<td>49.58%</td>
</tr>
<tr>
<td>Baa</td>
<td>88.62%</td>
<td>66.43%</td>
<td>92.17%</td>
<td><strong>82.41%</strong></td>
</tr>
<tr>
<td>Ba</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>B</td>
<td>26.67%</td>
<td>83.33%</td>
<td>2.88%</td>
<td>37.63%</td>
</tr>
<tr>
<td>C</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Note: this table summarizes the percentage of correct prediction (0-1 loss) for each rating category. The last column reports each of the four competing model’s the average performance during 2007-2009.
Figure 2: Moody’s actual ratings relative to the 2001 standards, full sample, 2001-2013

Act passed in 2010. Dimitrov et al. (2015) also finds that following the Dodd-Frank Act, CRAs issue lower ratings, give more false warnings, and issuing downgrades that are less informative. Before the structural break in 2009, Moody’s rating standard did not change much, as the times series of $D_T$ remains flat at the zero-line until 2008. The actual ratings in 2007 are, on average, significantly better than the predicted ratings using the 2001 standard. However, the economic magnitude of rating inflation is rather small (e.g., less than one-tenth of a notch).

Our findings is consistent with studies in the credit rating literature. Blume et al. (1998) is the first to note that rating standards have become more conservative for the period 1978 to 1995. They show the coefficients of the year dummies in an ordered-probit model have decreased sharply in this 18-year period. Baghai et al. (2014) shows that this trend has continued until at least 2009 and suggests the increased conservatism is not fully warranted because defaults have declined over the same period. Alp (2013) finds S&P’s rating standard did not change much before 2002 but has became more stringent since 2002, especially for the investment-grade bonds. The rating standard in a semiparametric model is summarized not only by the coefficients of risk characteristics, but also the (nonparametric) link function that flexibly allows for interaction and nonlinearities.
6 Conclusion

During the 2007-2009 financial crisis, credit rating agencies (CRAs) have been criticized for assigning inflated ratings to mortgage-backed securities and other structured products. Similar phenomenon have been documented on other asset classes such as corporate bonds (Kedia et al., 2016, Strobl and Xia) and CDOs (Griffin and Tang, 2012). The relaxed rating standard was largely attributable to conflicts of interest arising out of the CRA’s business model. In this paper, I propose a economically-interpretable model to predict credit ratings and investigate the time series variation in CRA’s rating standards.

Compared with extant bond rating models, the proposed model has two key features: (i) I explicitly consider the impact of conflicts of interest on ratings through common shareholders, (ii) the model imposes few distributional and functional form restriction on the underlying rating process. Asymptotic results of the proposed estimator is established after several bias corrections. While the focus of this paper is on predicting corporate bond ratings, the estimation and inference methods developed in here can be readily applied to other semiparametric model with multiple indices.
References


Simi Kedia, Shivaram Rajgopal, and Xing Zhou. Does it matter who owns moody’s, 2016.


Günter Strobl and Han Xia. The issuer-pays rating model and ratings inflation: Evidence from corporate credit ratings.

