DISENTANGLING AND ASSESSING UNCERTAINTIES IN MULTIPERIOD CORPORATE DEFAULT RISK PREDICTIONS

By Miao Yuan[*], Cheng Yong Tang[†,1], Yili Hong[*,2] and Jian Yang[‡,3]

Virginia Tech[*], Temple University[†] and University of Colorado Denver[‡]

Measuring the corporate default risk is broadly important in economics and finance. Quantitative methods have been developed to predictively assess future corporate default probabilities. However, as a more difficult yet crucial problem, evaluating the uncertainties associated with the default predictions remains little explored. In this paper, we attempt to fill this blank by developing a procedure for quantifying the level of associated uncertainties upon carefully disentangling multiple contributing sources. Our framework effectively incorporates broad information from historical default data, corporates' financial records, and macroeconomic conditions by (a) characterizing the default mechanism, and (b) capturing the future dynamics of various features contributing to the default mechanism. Our procedure overcomes the major challenges in this large scale statistical inference problem and makes it practically feasible by using parsimonious models, innovative methods, and modern computational facilities. By predicting the marketwide total number of defaults and assessing the associated uncertainties, our method can also be applied for evaluating the aggregated market credit risk level. Upon analyzing a US market data set, we demonstrate that the level of uncertainties associated with default risk assessments is indeed substantial. More informatively, we also find that the level of uncertainties associated with the default risk predictions is correlated with the level of default risks, indicating potential for new scopes in practical applications including improving the accuracy of default risk assessments.

1. Introduction. Measuring the corporate default risk has long been crucial in many business decisions. Examples include loan evaluation where a bank analyzes the credit quality of a borrower over various future potential borrowing periods, internal control considerations where corporate management needs to periodically and accurately assess the firm's present financial condition, investment screening where investors predict financial health of investments under consideration and screen out undesirable investments, and determining the credit ratings by rating agencies.

Also, recent introduction and expansion of credit derivative markets have renewed interests in this topic. According to the survey by the International Swaps and Derivatives Association (ISDA), the credit default swap (CDS) market, the most popular type of credit derivatives, has exploded over the past decade to about $30 trillion in 2010, up from $0.9 trillion in 2001. The default probabilities underlie the pricing of such financial instrument, and CDS reflects the market-based estimate of default probabilities. The Basel II bank regulation has further pushed the topic to the center of the banking regulation. In particular, based on the Basel II accord, banks and bank regulators need to determine the appropriate level of regulatory and economic capital to be held by a bank to be in line with the credit risk represented by its loan portfolio, where borrower default probabilities play an explicit role.

In the finance literature, there are two broad categories of approaches for corporate default modeling—the structural and the reduced form modeling approaches. The classical structural approach of Merton (1974) assumes that a firm defaults when its assets drop to a sufficiently low level relative to its liabilities. A key implication is that a firm's conditional default probability is completely determined by the only key variable, its distance to default, which is closely related to the firm's annual asset growth, accounting for its levels of liabilities and volatilities; see, among others, the review of the structural approaches in Altman, Resti and Sironi (2004). Campbell, Hilscher and Szilagyi (2008) argue that despite the impressive predictive power of the Merton's (1974) structural model, in light of its restrictive functional form, it is better to use a reduced-form model, allowing more covariates entering default predictions; see also Duffie and Lando (2001) on the dependence of defaults with other covariates. The first generation of the reduced-form models, for example, those in Beaver (1966, 1968) and Altman (1968), primarily rely on the multiple discriminant analysis (MDA), classifying a firm into one of the possible categories based on the score and rank computed from its individual characteristics. The second generation approaches, such as Ohlson (1980) and Zmijewski (1984), propose to use the logistic regression analysis, attempting to assess the conditional probability that a firm would go default in the next period of time. Both the first and second generations may be considered as static approaches, as they have been using single-period classification or probability models with the bankruptcy data but have ignored their multiple-period feature. As pointed out by Shumway (2001), static models would produce biased and inconsistent estimates of bankruptcy probabilities due to their ignoring the dynamics over time, and may introduce an unnecessary selection bias into the estimates. The current generation of the reduced form models incorporates the dynamics over time by examining the duration of the default events. Shumway (2001) proposes a hazard function based duration modeling with time-dependent predictors; see also Chava and Jarrow (2004), Campbell, Hilscher and Szilagyi (2008), Bharath and Shumway (2008), and Beaver, Correia and McNichols (2012) among others. In

particular, Chava and Jarrow (2004) demonstrate the superior predicting performance of Shumway's (2001) model over the first [i.e., Altman (1968)] and second [i.e., Zmijewski (1984)] generations of models. The most recent development of the reduced form modeling approaches on default predictions has an emphasis on the corporate defaults over multiple periods; see, for example, Duffie, Saita and Wang (2007) and Duan, Sun and Wang (2012).

While various quantitative procedures have been developed for predicting corporate default probabilities, point predictions are serving as the dominating measures in the current state of knowledge. A blank in the literature is that the point predictions are equipped with no assessment of the associated prediction uncertainties. A main reason behind it is that the task is too challenging due to the huge scale and high complexity of the problem. Clearly, the historical corporate and macroeconomic time series data are of high dimensionality and complexity. Meanwhile, all companies exposed to future default risks would require assessments of their default risks and the associated level of uncertainties. The level of complexity would substantially increase further if one concerns the prediction with multiple future periods. Fundamentally, there are multiple contributing sources to the uncertainties including those from the default mechanism, the future dynamics of the corporates and economic environment, and the model estimation errors; see Section 2 for more details on the source of uncertainties.

Our investigation intends to develop a procedure obtaining prediction intervals for quantifying the level of uncertainties associated with default risk predictions, taking the three aforementioned sources of uncertainties into account. Studies on the prediction intervals with duration modeling have been documented in the literature in areas such as the reliability; see, for example, Hong, Meeker and McCalley (2009) and reference therein. Nevertheless, existing methods do not apply to the scenario of corporate default prediction due to the unique challenging practical aspects of the problem; see Sections 2 and 3.

Given the complexity of the models in this scenario, explicit formulas generally do not exist for constructing valid prediction intervals. Thus, our framework resorts to resampling procedures built upon parametric stochastic models. When the number of companies at default risk is at the order of tens of thousand with history of tens of years, challenges are arising from (1) complicated structure of covariates with large number of unknown parameters, (2) large scale of the data sets, and (3) complicated data structure. For example, the data set in our analysis in Section 4 for the US market over the period 1990–2009 contains more than 10,000 companies, and the number of monthly observations exceeds 1,000,000. However, only few companies have observations during the entire period because many companies either went default or exited the market due to other reasons. Among all companies in the data set, missing data are overwhelming and the time horizons for those observations are highly heterogeneous among the companies. Our Section 3 provides detail on our framework developed for uncertainties assessments with prediction intervals for both point predictions and total defaults

predictions, overcoming those challenges by using parsimonious models, innovative and computationally efficient methods, and powerful computational facilities.

The proposed framework in this study will contribute to the literature from several important aspects. First and foremost, compared with the current practice of default probability prediction which typically yields only the point estimate, the introduction of default prediction intervals dramatically improves our understanding and knowledge especially for model diagnosis and statistical inferences. Furthermore, a distinguished feature of our measure is allowing for not only multiple sources of uncertainties but also the asymmetric nature of default probability prediction intervals so that the lower bound of default probability prediction would not go below zero (which is obviously not sensible). Appropriately quantifying the associated uncertainties is the key to valid statistical inference on the future default probabilities. For example, to assess how well their model of default prediction performs, Campbell, Hilscher and Szilagyi (2008) compare the fitted point estimate of probability of failure (which is the average of such estimates from each company) with the actual default rate in the market and conclude (p. 2916) that their model somewhat overpredicts failures in 1974 to 1975, underpredicts for much of the 1980s, and then overpredicts in the early 1990s. Obviously, additional scope of the problem may be provided if the prediction intervals are taken into consideration. Also, the availability of uncertainty quantification for default probability prediction would enable us to further conduct effective forecasting evaluations, for example, along the line of Diebold and Mariano (1995), where one would examine whether the apparent improvement of forecast accuracy is statistically significant. In our data analysis reported in Section 4, facilitated by the prediction intervals, we are able to show that the out-of-sample aggregated predictions for the total number of defaults work quite well for multiple years.

Moreover, the uncertainties of default probability prediction should be crucial in improving our understanding of default risk pricing on financial markets, and may provide a new venue of exploring distress risk and/or credit risk in asset pricing. For example, Ding et al. (2012) document the puzzling negative relationship between stock returns and default risk as measured by default probability. Giesecke et al. (2011) report a puzzling finding on the US corporate bond market that credit spreads are roughly twice as large as default losses and do not respond to realized default rate. The missing uncertainties of default probability predication could be important. As an illustration, we demonstrate by our data analysis in Section 4 that the assessments of uncertainties associated with predicted default probabilities for individual companies are indeed highly informative. First of all, the level of uncertainties can be high, especially for those companies with high predicted default probabilities. Second, and more interestingly, we found that by incorporating the width of the prediction interval in a logistic regression for the binary variable defined as a company going default or not, significant interaction is found between the width of the prediction interval and the point default probability prediction. This shows that the level of uncertainties associated with the point default

probability prediction can be informative practically for solving problems. Additionally, the uncertainties of default probability prediction should shed light on many important issues in finance where default/credit risk plays a central role. For example, Giesecke and Kim (2011) explore the systemic risk of the financial sector defined as the conditional probability of failure of a sufficiently large proportion of financial institutions. We show in our Section 4 that our procedure is capable of equipping prediction intervals with aggregated predicted total number of defaults, a feature that can benefit various studies of systemic risks.

The rest of this paper is organized as follows. In Section 2, we disentangle the sources of uncertainties contributing to the default predictions. We present in Section 3 our framework for predictions and assessing their associated levels of uncertainties for future default risks of individual companies and the total number of future defaults on the market. Section 4 comprehensively analyzes a large-scale US market data by developing default probability predictions and quantitatively assessing their associated level of uncertainties. A simulation study evaluating the accuracy of our method is presented in Section 5. Section 6 concludes this paper and draws the picture for future research. The Supplementary Material [Yuan et al. (2018)] contains the detail of the EM algorithm in our method.

2. Sources of uncertainties in default risk predictions.

2.1. *Stochastic time-to-event.* We construct our procedure in a setting for multiperiod corporate default risk prediction. For the default mechanism, we focus on the recent reduced form models for the durations of the defaults and other competing risk events. Event duration modeling can be broadly classified into the time-to-event data analyses, a subject that has been intensively studied in areas such as reliability in engineering studies, and survival analysis [see, for example, Meeker and Escobar (1998), and Kalbfleisch and Prentice (2002)]. The key device for the duration modeling is the event intensity function. In the survival or time-to-event analysis, intensity function modeling plays a central role; see the monograph Kalbfleisch and Prentice (2002) and reference therein. For corporate defaults in finance, predictions with the intensity function model also involve the stochastic nature of the explanatory variables, and we refer to Duffie (2011) as an overview.

Modeling the durations with intensity function treats the time when a company defaults in the future as a random variable. Meanwhile, to accommodate the fact that a company may exit the market before going default due to events other than bankruptcy, for example, being acquired by another company, incorporating the competing risks are required in modeling the time-to-exit of the companies. Generally speaking, suppose there are K types of credit events competing with each other so that in only one event that occurs the first is observed. Let T_k be the time to the event of type k $(k = 1, \ldots, K)$. We denote the event intensity function by $\lambda_k(t)$ $[\lambda_k(t) \geq 0, t \geq 0]$ for the kth type of event. The event intensity function connects to the survival function $S_k(t) = P(T_k > t)$ by $S_k(t) = \exp\{-\int_0^t \lambda_k(u)\, du\}$;

see, among others, Kalbfleisch and Prentice (2002) for more detail on the intensity function.

Additional to observing that the intensity $\lambda_k(t)$ is a function of time, it is natural to expect that features including the financial healthiness, profitability, growing perspective,etc., are affecting the future default occurrences. Meanwhile, the macroeconomic conditions also have impact on future defaults. From the predictive perspective, the company specific features and the macroeconomic conditions are also stochastic, a source that also contributes to the uncertainties in the corporate default predictions. Therefore, adequately incorporating the dynamic features is crucial in both predicting the defaults and assessing the associated level of uncertainties.

We describe here how to incorporate the dynamic features as explanatory covariate information in the intensity function. Modeling and estimating the stochastic covariate will be discussed in Section 2.4. We denote by a random vector $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})^{\mathrm{T}}$ indexed by time t containing generic covariates of dimensionality p. The observed covariate process is then $\boldsymbol{x}(t_1, t_2) = \{\boldsymbol{x}_s : t_1 < s \leq t_2\}$ representing available covariate information from t_1 to t_2. Subsequently, the intensity function for event type k ($k = 1, \ldots, K$) for a company at time t with covariate $\boldsymbol{x}_t$ is characterized by $\lambda_k(t; \boldsymbol{x}_t)$. Such an intensity function models the rate (i.e., probability per unit time) that event k will happen instantly after time t, conditioning on the covariate value. The total events intensity (i.e., something happens) for a company at time t is $\lambda(t; \boldsymbol{x}_t) = \sum_{k=1}^{K} \lambda_k(t; \boldsymbol{x}_t)$ by noting that competing risk events are mutually exclusive. We also define the cumulative intensity function for the event type k as $\Lambda_k[t; \boldsymbol{x}(0, t)] = \int_0^t \lambda_k(s; \boldsymbol{x}_s) \, ds$ ($k = 1, \ldots, K$). Then the aggregated cumulative intensity function is $\Lambda[t; \boldsymbol{x}(0, t)] = \sum_{k=1}^{K} \Lambda_k[t; \boldsymbol{x}(0, t)]$.

2.2. *Parametric intensity function and its estimation.* For practical applications, parametric intensity functions $\lambda_k(\cdot)$ ($k = 1, \ldots, K$) are often assumed for effectively analyzing time-to-event data with meaningful practical interpretations. In our work, we consider that the intensity function of event type k at time t has the exponential additive form $\lambda_k(t; \boldsymbol{x}_t) = \exp(\beta_{k0} + \beta_{k1}x_{1t} + \cdots + \beta_{kp}x_{pt})$. Treating $\boldsymbol{x}_t$ as random, the framework is referred to as doubly stochastic in Duffie, Saita and Wang (2007) for corporate default predictions. The parameter $\boldsymbol{\beta} = (\beta_{10}, \ldots, \beta_{1p}, \ldots, \beta_{K0}, \ldots, \beta_{Kp})^{\mathrm{T}}$ is unknown and needs to be estimated from historical defaults data. Therefore, uncertainties associated with the parameter estimation contribute to the uncertainties in the default predictions.

We now describe the maximum likelihood (ML) method for estimating the parameter in the intensity function. For each company, the time-to-event data are denoted by $\{t_i, \boldsymbol{\delta}_i, \boldsymbol{x}_i(0, t_i)\}$ ($i = 1, \ldots, n$), where n is the number of companies. Here t_i is the event time for company i if one of the K events happens, and t_i is the last observation time τ if no event occurred during the data collection period. The event indicator for company i is $\boldsymbol{\delta}_i = (\delta_{1i}, \ldots, \delta_{Ki})^{\mathrm{T}}$, where $\delta_{ki} = 1$

and $\delta_{li} = 0, l \neq k$ if event k happens to company i, and $\delta_{li} = 0, l = 1, \ldots, K$, if no event happens until the last observation time τ. Last, the observed covariate history from the time origin to t_i for company i is denoted as $\boldsymbol{x}_i(0, t_i) = \{\boldsymbol{x}_{i,s} : 0 < s \leq t_i\}$, with $\boldsymbol{x}_{i,s}$ representing the covariates of company i at time s. In this investigation, we consider $K = 2$ types of events hereinafter, that is, a company defaults $(k = 1)$ or exits the market due to other events $(k = 2)$.

We note that the cumulative distribution function (cdf) of the random variable T for the time-to-event of a company, given its covariate history $\boldsymbol{x}(0, t)$, is $F_T(t) = P(T \leq t) = 1 - e^{-\int_0^t \sum_{k=1}^2 \lambda_k(s; \boldsymbol{x}_s)\, ds} = 1 - e^{-\Lambda[t; \boldsymbol{x}(0,t)]}$. The marginal cdf of time to event type k, denoted as T_k, is $F_{T_k}(t) = 1 - e^{-\int_0^t \lambda_k(s; \boldsymbol{x}_s)\, ds} = 1 - e^{-\Lambda_k[t; \boldsymbol{x}(0,t)]}$. The probability density function (pdf) of T_k is $f_{T_k}(t) = \lambda_k(t; \boldsymbol{x}_t)e^{-\Lambda_k[t; \boldsymbol{x}(0,t)]}$. To differentiate different types of observed events, let $\Delta_k, k = 1, 2$ be the event indicators. That is, $\Delta_1 = 1, \Delta_2 = 0$ if it is a default, $\Delta_1 = 0, \Delta_2 = 1$ if it is an exist due to other reason, and $\Delta_1 = \Delta_2 = 0$ if no event occurred by the latest observation time (denoted by τ) in the data set. Due to two types of competing risks, the observed time-to-event of a company is therefore $T = \min(T_1, T_2)$. The fraction of failing probability due to the type k event is

$$F_k(t) = \Pr(T \leq t, \Delta_k = 1) = \Pr(T_k \leq t, T_l > T_k; \text{ for all } l \neq k)$$

$$= \int_0^t f_{T_k}(t_k) \prod_{l \neq k}[1 - F_{T_l}(t_k)]\, dt_k = \int_0^t \lambda_k(s; \boldsymbol{x}_s)e^{-\Lambda[s; \boldsymbol{x}(0,s)]}\, ds.$$

The joint likelihood of the event times or the last observation times t_i's of the n given the covariate processes $\boldsymbol{x}_{i,t_i}$ at t_i and covariate history $\boldsymbol{x}_i(0, t_i)$ $(i = 1, \ldots, n)$ is then given by

$$L_T(\boldsymbol{\beta}|\text{DATA}) = \prod_{i=1}^n \left(\left(\prod_{k=1}^2 \{\lambda_k(t_i; \boldsymbol{x}_{i,t_i})e^{-\Lambda[t_i; \boldsymbol{x}_i(0,t_i)]}\}^{\delta_{ki}} \right) \right.$$

(2.1)

$$\left. \times \{e^{-\Lambda[t_i; \boldsymbol{x}_i(0,t_i)]}\}^{\prod_{k=1}^2 (1-\delta_{ki})} \right),$$

where $\lambda_k(t; \boldsymbol{x}_t)$ is proportional to the probability that a company has an event of type k between time t and $t + dt$, where dt is an infinitesimal amount of time, $e^{-\Lambda[t; \boldsymbol{x}(0,t)]}$ gives the probability of observing a company that survives to time t. The parameters $\boldsymbol{\beta}$ are then estimated by maximizing the joint likelihood of the event times in (2.1). In practice, the covariate history $\boldsymbol{x}_i(0, t_i)$ for company i is only discretely observable. Therefore, integration of the intensity function of event type k [i.e., $\int_0^{t_i} \lambda_k(s; \boldsymbol{x}_{i,s})\, ds$] can be approximated by discretization.

We remark that the uncertainties associated with the parameter estimation can be conventionally quantified using the standard errors by inverting the observed Fisher information matrix. In the literature, this type of standard errors are typically reported as a measure of level of uncertainties. However, considering only

uncertainties in the parameter estimations is not yet adequate for assessing the uncertainties associated with the multiperiod corporate default predictions. The reason is that the parameter estimation procedure is a static one conditioning on the covariate process so that it fails to incorporate any future dynamics in the covariate process.

2.3. *Covariate process and discrete-time observations.* The intensity function, on one hand, by its definition is a function of continuous time. On the other hand, those covariates used for modeling the intensity function can only be observed at a grid of discrete time points. Thus, the survival function, which relates to the integration of the intensity function, is typically approximated by taking the intensity function to be piecewise constant between two adjacent observation times. Such an approximation in fact relates the intensity function modeling to multiperiod binary response variable regression analysis with the logit or other link functions; see, for example, Shumway (2001) and Duffie, Saita and Wang (2007). Moreover, the one-period ahead default predictions using the logistic regression type of approaches can also be viewed as a result of piecewise constant approximation of the intensity function; see, among others, the studies of Hillegeist et al. (2004) and Bharath and Shumway (2008). Clearly, the accuracy of the approximation is worse with a longer time interval between two observations, and so is the accuracy of the prediction using the logistic regression type approaches. Therefore, multiperiod corporate default predictions require a more accurate account for the future dynamics; see Duffie, Saita and Wang (2007) and Duan, Sun and Wang (2012).

In constructing our procedure, the dynamics of the covariates process clearly play an important role in corporate default predictions. For example, the multiperiod approach of Duffie, Saita and Wang (2007) relies on a parsimonious high-dimensional vector autoregressive time series model for the covariate process, and numerical approximation for integrations with respect to the future dynamics is needed for assessing the multiperiod future corporate default probabilities. Our model in this investigation for the high-dimensional vector time series incorporates an autoregressive component for capturing the predictive information in the conditional mean function. The autoregressive structure is a benchmarking one and has broad applications in time series analysis; see the overviews in the monographs Tsay (2010), Durbin and Koopman (2012) and Tsay (2014). To exploit further the systematic/structural dynamics among the covariates, we further equip the innovations of the vector time series with a dynamic factor model that disentangles the contributions from a systematic factor component and an idiosyncratic component.

In statistics and financial econometrics, the dynamic factor models are advantageous for its parsimony and have demonstrated their promising predictive performance in broad areas; see the monograph Durbin and Koopman (2012) for an overview. In the credit risk modeling literature, the dynamic factor models have

been demonstrated to be effective and have many successful applications. Among them, for example, a dynamic factor model is applied in conjunction with the intensity function modeling with parametric baseline hazard in Koopman, Lucas and Monteiro (2008) for credit rating transitions, incorporating dynamic frailty accounting for the dependence between defaults. Koopman, Lucas and Schwaab (2011) investigate the dynamic frailty-correlations between defaults in different segments and groups of firms, and demonstrate the impact from the latent dynamic factors. Equipping the latent factors with attractive practical interpretations such as those effects from the macro, industrial, regional factors, etc., the dynamic models are capable of incorporating multiple sources of contributions to assessing the default risks. Koopman, Lucas and Schwaab (2012) consider the dynamic factors models incorporating multiple effects for the default counts modeling for the 2008 credit crisis. More recently, Schwaab, Koopman and Lucas (2017) investigate global credit risk concerning multiple countries, and demonstrate the impacts from the country and industry factors.

An appealing feature of the dynamic factor models is their convenience in computations. For Gaussian models, the Kalman filter can be conveniently applied, which is the case for our dynamic factor model. The Kalman filter is also capable of handling missing data and mixed frequency data; see Bräuning and Koopman (2014) for a recent investigation on forecasting with mixed frequency data. When additional non-Gaussian observations are incorporated with the dynamic models, the importance sampling based techniques are demonstrated useful for estimations. As examples, the defaults counts data are handled together with other covariate in the dynamic factor models of Koopman and Lucas (2008), Koopman, Lucas and Schwaab (2011), and Schwaab, Koopman and Lucas (2017).

The main objective of our covariate process modeling concerns the dynamic for corporate defaults predictions at the individual firm level, with a combination of the autoregressive and dynamic factor structures. Our approach integrates the covariate model with the dynamic intensity function model outlined in Section 2.1, based on which we conduct the multiperiod corporate default predictions. Here our attempt targets at the corporate default prediction at the firm level, and it is different from predicting the default counts as in existing studies, for example, Koopman, Lucas and Schwaab (2011, 2012) and Schwaab, Koopman and Lucas (2017). Computationally, it is clearly a more demanding objective. Hence exploiting the impacts and benefits from the latent dynamic factors is clearly desirable in this scenario. Our approach with the modeling device for corporate default predictions has a few new methodological features of their own interests. First, the systematic contributions from the latent factors among the firms are exploited in conjunction with the autoregressive structure in a dynamic intensity function for capturing the dynamics in predicting corporate defaults. Second, the separated treatments of the covariate process and the default mechanism allow feasibility and convenience for multiperiod predictions for individual firms with the dynamic factor model—only the computationally more efficient Kalman filter is involved when handling high-dimensional time series from thousands of firms.

2.4. *Parametric stochastic covariate process and its estimation.* We now describe the parametric model for the covariate process considered in our framework. Specifically, we denote by $\mathbb{X}_t$, $t = 1, \ldots, \tau$ the observed covariate process including both firm-specific covariates for all the companies and the macroeconomic covariates at time t. The firm-specific and macroeconomic covariates, serving as effective reflection of the profitability as well as leverage ratio of assets to debts of a company, and indicators for the economic condition of the nation, are used to model the default risks. Inspired by Duffie, Saita and Wang (2007), we focus on two firm-specific variables—the distance to default ($D_{i,t}$) and the trailing one-year stock log-return ($V_{i,t}$) for company i at time t. Here, the distance to default Merton (1974) is a classical measure in corporate credit risk analysis especially from a structural model point of view. Roughly speaking, the distance to default is defined as the number of standard deviations of annual asset growth by which the log asset level exceeds the firm's log liabilities. In the classical model of Merton (1974), a company's conditional default probability is completely determined by its distance to default. In our studies, we use the distance to default calculated by the method proposed in Duan, Sun and Wang (2012). For macroeconomic variables, we choose the trailing one-year return on the S&P 500 index (S_t) and the three month Treasury bill rate (r_t). Hence, we have $\mathbb{X}_t = (\mathbf{D}_t^{\mathrm{T}}, \mathbf{V}_t^{\mathrm{T}}, r_t, S_t)^{\mathrm{T}}$, $t = 1, \ldots, \tau$ where $\mathbf{D}_t = (D_{1,t}, \ldots, D_{n,t})^{\mathrm{T}}$, $\mathbf{V}_t = (V_{1,t}, \ldots, V_{n,t})^{\mathrm{T}}$, and τ is the total number of time points. That is, $\mathbb{X}_t$ is $m \times 1$ vectors where $m = 2n + 2$ and n is the number of firms. In the data set for our studies, the observations are available monthly. To adjust for evident quarterly seasonal effect in the time series, we take a difference of order 3, resulting in a new m-dimensional vector time series $\mathbf{X}_t = \mathbb{X}_{t+3} - \mathbb{X}_t$ ($t = 1, \ldots, \tau'$), where $\tau' = \tau - 3$.

Modeling the dynamics of $\mathbf{X}_t$ is the most challenging task in default predictions and assessing the associated level of uncertainties, because of the fact that $\mathbf{X}_t$ is of very high dimensionality. Take, for example, the US market, the total number of companies has exceeded 10,000 since 1990. Moreover, in an active market, new companies are almost continuously formed while many existing ones exit the market for various reasons, resulting in highly un-balanced observations of the time series, that is, the origin and end times of the components in $\mathbf{X}_t$ are different with possible missing data for some periods of time. Furthermore, the behaviors of components in $\mathbf{X}_t$ are expected to inter-relate with each other in some complicated ways. Thus jointly modeling the tremendously high-dimensional time series becomes daunting, while further dedicated effort is also necessary for developing methods of parameter estimation and assessing the associated level of uncertainties.

We consider a highly parsimonious time series model for $\mathbf{X}_t$ with two key components specifically for default predictions: (a) a mean-reverting autoregressive structure in the conditional mean of $\mathbf{X}_t$ given prior observations, and (b) a dynamic factor model for the innovations to capture the correlations among the components in $\mathbf{X}_t$. We refer to Tsay (2010) as an introduction for modeling vector valued time

series, and Pan and Yao (2008) and Lam and Yao (2012) for recent development of factor models for multivariate time series.

Specifically, the conditional mean model is a modified version of the one considered in Duffie, Saita and Wang (2007):

$$(2.2) \qquad \mathbf{X}_t - \boldsymbol{\mu} = \boldsymbol{\Theta}(\mathbf{X}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t, \qquad t = 2, \dots, \tau'.$$

Model (2.2) is a vector auto-regression model mainly to capture the conditional dependence with the mean reverting effects of all the covariates. The coefficient matrix $\boldsymbol{\Theta}$ is designed in a parsimonious way following Duffie, Saita and Wang (2007) as

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\kappa}_D & \mathbf{0} & \mathbf{b} & 0 \\ \mathbf{0} & \boldsymbol{\kappa}_V & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{0} & \kappa_r & 0 \\ \mathbf{0} & \mathbf{0} & 0 & \kappa_S \end{pmatrix},$$

where $\boldsymbol{\kappa}_D = \kappa_D \mathbf{I}_n$, $\boldsymbol{\kappa}_V = \kappa_V \mathbf{I}_n$, $\mathbf{b} = b\mathbf{1}_n$. $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathbf{1}_n$ is an n-dimensional vector taking value 1 for all of its elements. Here we define the mean reverting vector as $\boldsymbol{\kappa} = (\kappa_D, \kappa_V, \kappa_r, \kappa_s, b)^{\mathrm{T}}$. The first four elements $\kappa_D, \kappa_V, \kappa_r, \kappa_s$ of $\boldsymbol{\kappa}$ capture the mean reverting effects of the four selected covariate processes. The current distance to default is modeled jointly by the mean reverting of the previous value, and the effect of departure of Treasury bill rate r_{t-1} from its mean at the previous month through the parameter b.

To further capture the serial and cross-sectional dependence between components in $\mathbf{X}$, we propose to apply the following dynamic factor model (DFM) for the innovation vector $\boldsymbol{\varepsilon}_t$:

$$(2.3) \qquad \boldsymbol{\varepsilon}_t = \boldsymbol{\Lambda} \mathbf{F}_t + \mathbf{e}_t,$$

$$(2.4) \qquad \mathbf{F}_t = \mathbf{A} \mathbf{F}_{t-1} + \boldsymbol{\eta}_t, \qquad t = 2, \dots, \tau',$$

where the latent factor $\mathbf{F}_t$ is a $q \times 1$ vector following an auto-regression process with order 1 [i.e.,VAR(1)]. The principal component analysis (PCA) is a convenient device for the factor model structure (2.3); see, for example, Stock and Watson (2002). The dynamic factor structure introduced by (2.4) has demonstrated its merits in numerous credit risk modeling; see the discussions in Section 2.3. The state space methods, Durbin and Koopman (2012) are convenient for handling the model setting with (2.3) and (2.4). Here, $\boldsymbol{\eta}_t$'s are assumed to be independently and identically distributed (iid) normal random vectors from $N(\mathbf{0}, \mathbf{Q})$, for some positive definite matrix $\mathbf{Q}$. Here $\boldsymbol{\Lambda}$ is a $m \times q$ matrix of factor loadings, and $\mathbf{A}$ is a $q \times q$ matrix of autoregressive coefficients. The random vectors $\boldsymbol{\eta}_t$ and $\mathbf{e}_t$ are independent normal random vectors. The covariance matrix of $\mathbf{e}_t$ is assumed to be a diagonal matrix $\mathbf{P}$. Here the factor model with loading $\boldsymbol{\Lambda}$ is high parsimonious by the fact that the number of common factor q is usually very small, which drastically reduces the number of parameters in the covariance matrix of $\boldsymbol{\varepsilon}_t$. As in our

data analysis, the number of the factor is chosen as $q = 2$ by using the method of
Bai and Ng (2008). Facilitated by the dynamic factor model, the future dynamics
of the covariate process can be effectively incorporated in default predictions and
uncertainties assessments.

We collectively denote by $\boldsymbol{\theta_x} = \{\boldsymbol{\mu}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \mathbf{A}, \mathbf{P}, \mathbf{Q}\}$ all parameters in the covariate proces. We develop an expectation-maximization (EM) algorithm for estimating parameters in the dynamic factor model specified by (2.3) and (2.4), whose detail is given in the Supplementary Material [Yuan et al. (2018)]. Specifically, our EM algorithm efficiently incorporates the hidden factor $\mathbf{F}_t$ in this tremendously large scale problem with high-dimensional time series and highly un-balanced observations. In our EM algorithm, both the E-step and M-step can be conveniently executed for practical implementations. Most remarkably, the matrices inversions in our EM algorithm only involves those of size $q \times q$, making it most computationally efficient and feasible for this large scale default prediction problems.

3. Predictions and uncertainties assessments.

3.1. *Procedures for future default predictions.* In our study, predicting the future corporate default probabilities given the available current information is the key objective. For different levels of interests such as assessing the overall level of credit risks, one may also need to predict the total number of defaults for the overall market system and certain market sectors.

Let us begin with the method for individual corporate default predictions, and then the method for aggregated default predictions.

With the intensity model and the covariate model described in Sections 2.2 and 2.4, the future default probability of company i within s time units in the future after the last observation time τ is

$$\rho_i(s; \boldsymbol{\theta}) = \mathbf{E}\{\Pr[\tau < T_i \leq \tau + s, \Delta_1 = 1 | T > \tau] | \mathcal{F}_\tau\}$$

$$(3.1) \qquad = \mathbf{E}\Bigg\{\Bigg[\int_\tau^{\tau+s} \lambda_1(t; \boldsymbol{x}_{i,t})$$

$$\times \exp\big(-\{\Lambda[t; \boldsymbol{x}_i(0, t)] - \Lambda[\tau; \boldsymbol{x}_i(0, \tau)]\}\big) dt\Bigg]\Bigg|\mathcal{F}_\tau\Bigg\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_T^{\mathrm{T}}, \boldsymbol{\theta}_{\mathbf{x}}^{\mathrm{T}})^{\mathrm{T}}$ contains the parameters of the covariate model $\boldsymbol{\theta_x}$ and those of the time-to-event model $\boldsymbol{\theta}_T$. Here, the expectation is understood as conditioning on the information up to time τ, denoted by $\mathcal{F}_\tau$, and $\Delta_1 = 1$ indicates that it is the default probability of interest so that the corresponding intensity function $\lambda_1(\cdot)$ and the cumulative hazard function are involved in (3.1). We note that $\rho_i(s; \boldsymbol{\theta})$ is a sub-distribution function because $\rho_i(\infty; \boldsymbol{\theta}) < 1$ due to the existence of other types of competing events.

Because no simple analytical expression for the expectation in (3.1) is available, we use a Monte Carlo simulation approach to evaluate $\rho_i(s; \boldsymbol{\theta})$. The following

algorithm is for computing $\hat{\rho}_i(s; \widehat{\boldsymbol{\theta}})$ with the estimate $\widehat{\boldsymbol{\theta}}$ from the methods described in Sections 2.2 and 2.4.

ALGORITHM 1.

1. Simulate $\mathbf{X}^*(\tau', \tau' + s)$, the future differenced covariate processes, for all the firm-specific and macroeconomic covariates from their distribution $\mathbf{X}(\tau', \tau' + s)|\mathbf{X}_{\tau'}$ as specified in Section 2.4.
2. By inverting the differencing operator, the simulated covariate processes denoted by $\mathbb{X}^*(0, \tau + s)$ is reconstructed from the combined differenced data $\mathbf{X}^*(0, \tau' + s) = \{\mathbf{X}(0, \tau'), \mathbf{X}^*(\tau', \tau' + s)\}$, where $\mathbf{X}(0, \tau')$ and $\mathbf{X}^*(\tau', \tau' + s)$ are respectively the historical and simulated future data.
3. For each company i, numerically compute

$$\rho_i^*(s; \widehat{\boldsymbol{\theta}}) = \int_\tau^{\tau+s} \lambda_1(t; \boldsymbol{x}_{i,t}^*) \exp\left(-\{\Lambda[t; \boldsymbol{x}_i^*(0, t)] - \Lambda[\tau; \boldsymbol{x}_i(0, \tau)]\}\right) dt.$$

4. Repeat steps 1–3 M times to obtain $\rho_i^{*m}(s; \widehat{\boldsymbol{\theta}}), m = 1, \ldots, M$ where M is the prespecified number of the simulation replications.
5. The prediction of $\rho_i(s; \boldsymbol{\theta})$ is obtained by $\hat{\rho}_i(s; \widehat{\boldsymbol{\theta}}) = M^{-1} \sum_{m=1}^M \rho_i^{*m}(s; \widehat{\boldsymbol{\theta}})$.

In the first step of Algorithm 1, to forecast the differenced covariate process, we calculate the following for $s = 1, 2, \ldots$

$$\tilde{\boldsymbol{\varepsilon}}_{\tau'+s} = \widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{F}}_{\tau'+s} + \hat{\mathbf{e}}_{\tau'+s} \quad \text{and} \quad \mathbf{X}_{\tau'+s} = \hat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\Theta}}\mathbf{X}_{\tau'+s-1} + \tilde{\boldsymbol{\varepsilon}}_{\tau'+s},$$

where $\widehat{\mathbf{F}}_{\tau'+s}$ is forecasted from the fitted DFM for $\widehat{\mathbf{F}}_t$ and $\hat{\mathbf{e}}_{\tau'+s}$ is drawn from $N(\mathbf{0}, \widehat{\mathbf{P}})$.

As for the point prediction for the aggregated number of defaults in the market coverage of interest, let N_s be the cumulative number of defaults at s time units after the last observation time τ and $\text{RS}(t)$ be the set collecting companies of interests that are at risk of default at time t. It follows that $N_s = \sum_{i \in \text{RS}(\tau)} I_i(s)$ and $I_i(s) \sim \text{Bernoulli}[\rho_i(s; \boldsymbol{\theta})]$. The point prediction for N_s is $\widehat{N}_s = \sum_{i \in \text{RS}(\tau)} \hat{\rho}_i(s; \widehat{\boldsymbol{\theta}})$.

Though both the point predictions for $\rho_i(s)$ and N_s are informative for measuring future default risks, they do not reflect the uncertain nature of the predictions that we have discussed earlier. In what follows, we describe how to assess the uncertainties associated with the predictions.

3.2. *Assessing uncertainties at the aggregate level.* Prediction intervals (PIs) are used to quantify uncertainty in prediction of future random quantities. Let N_s be the cumulative number of events at a future point. A $100(1 - \alpha)\%$ PI for N_s is defined as $\Pr(\underline{N}_s \leq N_s \leq \widetilde{N}_s) = 1 - \alpha$. To assess the uncertainties associated with the predicted number of defaults, a natural choice is to supply a PI denoted by $[\underline{N}_s, \widetilde{N}_s]$. A naive (plug-in) PI for this purpose is obtained by solving

(3.2) $$F_{N_s}(\underline{N}_s; \widehat{\boldsymbol{\theta}}) = \frac{\alpha}{2} \quad \text{and} \quad F_{N_s}(\widetilde{N}_s; \widehat{\boldsymbol{\theta}}) = 1 - \frac{\alpha}{2}.$$

Here $F_{N_s}(n_s; \boldsymbol{\theta}), n_s = 0, 1, \ldots, n'$ is the cdf of N_s where n' is the number of companies in the RS(τ), $1 - \alpha$ is the desired coverage probability. Note that N_s is a sum of non-identically distributed Bernoulli random variables. An explicit form for $F_{N_s}(n_s; \boldsymbol{\theta})$ is

$$
\begin{aligned}
F_{N_s}&(n_s; \boldsymbol{\theta}) \\
&= \frac{1}{n'+1} \sum_{l=0}^{n'} \left\{ \frac{1 - \exp[-\boldsymbol{i}\omega l(n_s + 1)]}{1 - \exp(-\boldsymbol{i}\omega l)} \right. \\
&\qquad \left. \times \prod_{i \in \mathrm{RS}} [1 - \rho_i(s; \boldsymbol{\theta}) + \rho_i(s; \boldsymbol{\theta}) \exp(\boldsymbol{i}\omega l)] \right\},
\end{aligned}
$$
(3.3)

where $\boldsymbol{i} = \sqrt{-1}$ and $\omega = 2\pi/(n'+1)$. The cdf in (3.3) is obtained from a discrete Fourier transform of the characteristic function of N_s, which can be viewed as a generalization of the binomial distribution for a collection of firms with homogeneous default probability. We refer to Hong (2013) for more details on the derivation and an efficient implementation for computing $F_{N_s}(n_s; \boldsymbol{\theta})$. Alternatively, one can use some approximation methods such as the ordinary normal approximation or normal approximation with second order correction [e.g., Volkova (1996)].

The PI in (3.2) ignores the uncertainties in $\widehat{\boldsymbol{\theta}}$. Thus the coverage probability is generally smaller than the nominal $1 - \alpha$ level. These PIs can be calibrated to improve the coverage probability. We will use resampling method by parametric bootstrap to do the calibration.

Using the predictive distribution in Lawless and Fredette (2005), a $100(1 - \alpha)\%$ PI for N_s, denoted by $[\underset{\sim}{N_s}, \widetilde{N}_s]$, is obtained by

$$
\underset{\sim}{N_s} = v_{\alpha/2} \quad \text{and} \quad \widetilde{N}_s = v_{1-\alpha/2}.
$$
(3.4)

Here v_α is the α lower quantile of random variable N_s^* specified by distribution $F_{N_s}(\cdot; \widehat{\boldsymbol{\theta}})$ in which $\widehat{\boldsymbol{\theta}}$ is also treated as a random variable. In practice, v_α can be computed by simulations. That is, v_α is approximated by the α sample quantile of $N_s^{*b}, b = 1, \ldots, B$. Specifically, we obtain N_s^{*b} from $F_{N_s}(n_s; \widehat{\boldsymbol{\theta}}^{*b})$ given $\widehat{\boldsymbol{\theta}}^{*b} = (\widehat{\boldsymbol{\theta}}_T^{*b\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{x}}^{*b\mathrm{T}})^{\mathrm{T}}$, in which $\widehat{\boldsymbol{\theta}}_T^{*b}$ was simulated from $\mathrm{N}(\widehat{\boldsymbol{\theta}}_T, \Sigma_{\widehat{\boldsymbol{\theta}}_T})$ and $\widehat{\boldsymbol{\theta}}_{\mathbf{x}}^{*b}$ was estimated from the simulated covariate processes.

The simulation procedure based on parametric bootstrap is as follows.

ALGORITHM 2.

1. Simulate the differenced covariate processes $\mathbf{X}^*(1, \tau')$ from the model (2.2), (2.3), and (2.4) with $\widehat{\boldsymbol{\theta}}$. For each company, the differenced observations at the first month are kept and we do not extrapolate any period with no observations in the original data set.

2. Re-estimate parameters in the covariate model $\widehat{\boldsymbol{\theta}}_{\mathbf{x}}^*$ based on the simulated processes through the EM algorithm in the Supplementary Material [Yuan et al. (2018)].

3. Take a random sample of $\widehat{\boldsymbol{\theta}}_T^*$ from its asymptotic distribution $\mathrm{N}(\widehat{\boldsymbol{\theta}}_T, \Sigma_{\widehat{\boldsymbol{\theta}}_T})$, where $\widehat{\boldsymbol{\theta}}_T$ and $\Sigma_{\widehat{\boldsymbol{\theta}}_T}$ are estimated from the observed data by the methods in Section 2.2.

4. With the simulated data $\mathbf{X}^*(1, \tau')$ and the new parameter estimates $\widehat{\boldsymbol{\theta}}^* = (\widehat{\boldsymbol{\theta}}_T^{*\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{x}}^{*\mathrm{T}})^\mathrm{T}$, Algorithm 1 is implemented to predict the default probabilities $\rho_i^*(s; \widehat{\boldsymbol{\theta}}^*)$, $i = 1, \ldots, n$.

5. Take a random sample N_s^* from its distribution (3.3) with parameter values $\widehat{\boldsymbol{\theta}}^*$.

6. Repeat steps 1 to 5 B times to obtain N_s^{*b}, $b = 1, \ldots, B$.

7. The $100(1 - \alpha)\%$ calibrated PI for N_s is $\{N_s^{*([(\alpha/2)B])}, N_s^{*([(1-\alpha/2)B])}\}$, where $N_s^{*(\cdot)}$ is the ordered version of N_s^{*b} and $[\cdot]$ is the round function.

The fundamental rationale of the above algorithm and the one in the next section is to incorporate all sources of uncertainties as discussed earlier, that is, those from the stochastic default mechanism, the stochastic covariate process, and the parameter estimation procedures.

3.3. *Assessing uncertainties for corporate default probabilities.* By applying Algorithm 1, one may predict the multiperiod ahead of the default probabilities for individual corporations for evaluating the future default risk. Clearly, all sources of uncertainties are contributing in the point default probability estimations. For example, the final point prediction is the average of a range of possible default probabilities, and the different level of variations among those re-generated default probabilities are reflecting different levels of uncertainties associated with point default probability predictions. For reflecting the level of uncertainties associated with the default probability predictions, we propose to construct calibrated prediction intervals based on historical data for companies that are at risk incorporating all contributing sources of uncertainties.

The procedure for constructing the prediction interval is based on a large scale parametric bootstrap, similar to the one for the aggregated defaults prediction. Specifically, to incorporate the uncertainties in parameter estimation, in each iteration of resampling, we first simulate the differenced processes from the fitted covariate model and estimate parameters using the simulated data. To incorporate uncertainties associated with the parameter estimation of the time-to-event model, we re-generate model parameters from the estimated joint asymptotic distributions. Finally, we simulate multiperiod ahead values of the covariate process given the last observation in the historical data with the re-estimated parameter. Then, for each replication of the resampling procedure and for each at risk company, one multiperiod ahead default probability can be obtained. By repeating the procedure a number of times, we obtain the distribution of the predicted default probabilities

and construct the prediction interval correspondingly. More specifically, we have the following algorithm.

ALGORITHM 3.

1. Simulate the differenced covariate processes $\mathbf{X}^*(1, \tau')$ from the model (2.2), (2.3), and (2.4) with $\widehat{\boldsymbol{\theta}}$. For each company, the differenced observations at the first month are kept and we do not extrapolate any period with no observations in the original data set.
2. Re-estimate parameters in the covariate model $\widehat{\boldsymbol{\theta}}_{\mathbf{x}}^*$ based on the simulated processes through the EM algorithm in the Supplementary Material [Yuan et al. (2018)].
3. Take a random sample of $\widehat{\boldsymbol{\theta}}_T^*$ from its asymptotic distribution $\mathrm{N}(\widehat{\boldsymbol{\theta}}_T, \Sigma_{\widehat{\boldsymbol{\theta}}_T})$, where $\widehat{\boldsymbol{\theta}}_T$ and $\Sigma_{\widehat{\boldsymbol{\theta}}_T}$ are estimated from the observed data by the methods in Section 2.2.
4. With the simulated data $\mathbf{X}^*(1, \tau')$ and the new ML estimates $\widehat{\boldsymbol{\theta}}^* = (\widehat{\boldsymbol{\theta}}_T^{*\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{x}}^{*\mathrm{T}})^{\mathrm{T}}$, Algorithm 1 is implemented to predict the default probabilities $\rho_i^*(s; \widehat{\boldsymbol{\theta}}^*)$, $i = 1, \ldots, n$.
5. Repeat steps 1 to 4 B times to obtain $\rho_i^{*b}(s)$, $b = 1, \ldots, B$.
6. The $100(1 - \alpha)\%$ PI of default probability for the ith company at s time units after the last observation time τ is $\{\rho_i^{*([(\alpha/2)B])}(s), \rho_i^{*([(1-\alpha/2)B])}(s)\}$, where $\rho_i^{*(\cdot)}(s)$ is the ordered version of $\rho_i^{*b}(s)$ and $[\cdot]$ is the round function.

4. US corporate default data analysis.

4.1. *Data overview.* We now illustrate an application of our prediction framework on a US Corporate data set containing observations from January 1990 to November 2009. The data set contains defaults and other credit events information of the United States (US) public firms together with their stock market data from the CRSP (i.e., The Center for Research in Security Prices) database and accounting data from the Compustat database. The entire data set has around 12,000 US companies and more than 1,000,000 firm-specific monthly observations. Handling the entire data set is difficult given our limited resources, and it could take a very long time because we need to perform a reasonable number of replicated studies for validation and assessment purposes. So we choose a subset of the data with three industrial sectors—electronic product manufacturers, holding and investment offices, and business services. These three industrial sectors contain 3271 firms of the US market and have experienced a majority number of defaults. Specifically, among the 3271 companies on the market from January 1990 up to November 2009, 164 defaulted and 2049 exited due to other reasons, leaving 1058 companies at risk at the end of November 2009. Time-to-event information is available in the data set in terms of the occurrence time of an event and its type which is used to

determine if it is a default or an exit due to other reasons. For the firm specific covariate, we consider the distance to default (D_t) and the trailing one year stock return (V_t) of each company following our discussion in Section 2.4. To incorporate macroeconomic conditions, we consider monthly data of the trailing one-year S&P 500 return (S_t) and the three-month Treasury bill rate (r_t) as covariates as well.

4.2. *Model estimations.* By applying the proportional hazard model introduced in Section 2.2 using the covariate described in Section 2.4, the default and other exit intensity functions are modeled by $\lambda_k(t; \boldsymbol{x}_t) = \exp(\beta_{k0} + \beta_{k1} D_t + \beta_{k2} V_t + \beta_{k3} r_t + \beta_{k4} S_t)$ where $k = 1, 2$ are respectively corresponding to defaults and other exits. We estimate parameters in the time-to-event model by maximizing the log likelihood function as described in Section 2.2. Standard errors for the ML estimates are calculated by inverting the observed information matrix. The point estimates and 95% confidence intervals based on asymptotic normality are given in Table 1.

From Table 1, we see negative $\hat{\beta}_{11}$ but positive $\hat{\beta}_{21}$, indicating that lower risk is associated with larger value of distance to default, but in that case the firm is associated with higher chance of exit the market due to other reasons. Negative $\hat{\beta}_{12}$ and $\hat{\beta}_{22}$ show that higher stock return implies lower risk for both default and other forms of exits which may be due to that the trailing one-year stock return is an important indicator for a company's profitability. As for the effect of macroeconomic variables, negative $\hat{\beta}_{13}$ and $\hat{\beta}_{23}$ indicate that an increase in three-month Treasury bill rate manifests lower risks for both default and other exits, demonstrating the impacts on credit events from the overall economic condition of the environment. As for the effect from the trailing one-year return of the S&P 500 index upon controlling the level of other covariates, an increase in its value is associated with higher default risk, but its impact on other exits is not statistically significant. The negative impact might be due to the correlations between the individual stock returns and the S&P 500 index; see also the discussion in Duffie, Saita and Wang (2007).

To predict future dynamics of the covariates, we apply the dynamic factor model specified by (2.2), (2.3), and (2.4) described in Section 2.4. To estimate the parameters, we apply our EM algorithm discussed in Section 2.4 whose detail is given in the Supplementary Material [Yuan et al. (2018)].

The mean reverting parameters in (2.2) are estimated as $\hat{\kappa} = (0.63766, 0.63551, 0.89208, 0.63546, -0.00714)^{\mathrm{T}}$ with estimated standard errors $(0.0031, 0.0049, 0.0343, 0.0333, 0.0176)^{\mathrm{T}}$. The parameter $\mathbf{A}$ in the vector autoregressive model (2.4) for the hidden factors $\mathbf{F}_t$ is estimated as

$$\widehat{\mathbf{A}} = \begin{pmatrix} 0.3734 & 0.2144 \\ -0.0599 & 0.4803 \end{pmatrix},$$

	Default					**Other Exits**			
Para.	**Est.**	**SE**	**95% CI**		**Para.**	**Est.**	**SE**	**95% CI**	
			Lower	**Upper**				**Lower**	**Upper**
β_{10}	−6.9126	0.2018	−7.3081	−6.5171	β_{20}	−5.2646	0.0666	−5.3950	−5.1341
β_{11}	−0.6803	0.0867	−0.8502	−0.5105	β_{21}	0.0504	0.0084	0.0339	0.0669
β_{12}	−1.1467	0.0646	−1.2734	−1.0200	β_{22}	−0.3295	0.0401	−0.4081	−0.2509
β_{13}	−0.3091	0.0542	−0.4153	−0.2028	β_{23}	−0.0450	0.0160	−0.0763	−0.0137
β_{14}	1.9431	0.3974	1.1642	2.7219	β_{24}	−0.0839	0.1404	−0.3590	0.1913

whose entry wise estimated standard errors are 0.0955, 0.1368, 0.0364, 0.0598 for $\hat{A}_{11}$, $\hat{A}_{12}$, $\hat{A}_{21}$, and $\hat{A}_{22}$ where $\hat{A}_{ij}$ is the ijth component of $\widehat{\mathbf{A}}$. The EM algorithm also returns estimations of the loading matrix $\mathbf{\Lambda}$ and covariance matrix $\mathbf{P}$ of $\mathbf{e}_t$ in (2.3) containing many components that are omitted here.

4.3. *Total defaults predictions and uncertainties.* In what follows, when we are conducting multiple-period predictions, only data up to the origin of predictions are used when applying our prediction framework. For each prediction period, the actual observed defaults and covariate processes after the origin of predictions were held out and only used for validation and assessments afterward. For example, observed data from January 1990 to December 2008 were used to predict the default risks one year ahead in 2009 and so forth.

We first consider four respective one-year periods during 2006–2009, and for each period we conduct a one-year ahead prediction of the total number of defaults. We apply Algorithm 2 in Section 3.2 to obtain the prediction intervals for the total numbers. Figure 1 summarizes the results. Specifically for each one-year period, Figure 1 shows the predicted cumulative number of defaults and the associated 90% two-sided PI. In each panel of Figure 1, the solid step plot with dots indicates the actual cumulative numbers of defaults by month, and the solid straight line represents the predicted average number of defaults.

From Figure 1, we find that the predictive assessments of the overall credit risk levels are different between these four years by observing that, for example, the total predicted number of defaults in 2006 is much smaller than that of 2007. Such an observation well matches the situations that actually happened. Hence, for assessing the overall level of future credit risks, historical data are informative, and it is promising evidence for using quantitative methods to incorporate historical data information for future predictions. We also note that Figure 1 is reporting the cumulative number of predicted defaults, and the widening trend of the PIs reflect an expected fact that the level of uncertainties associated with predictions is
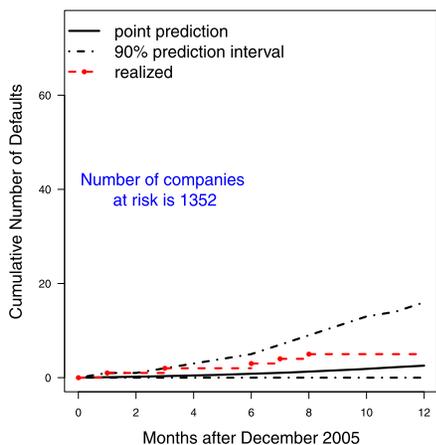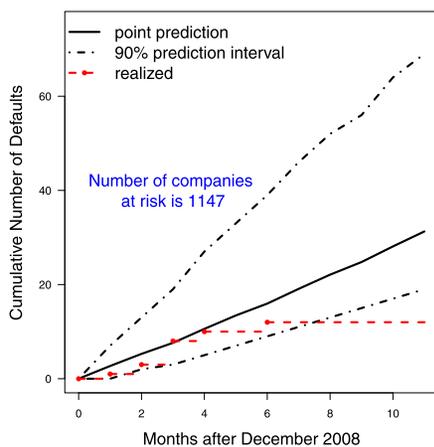
(a) Defaults Predictions for 2006
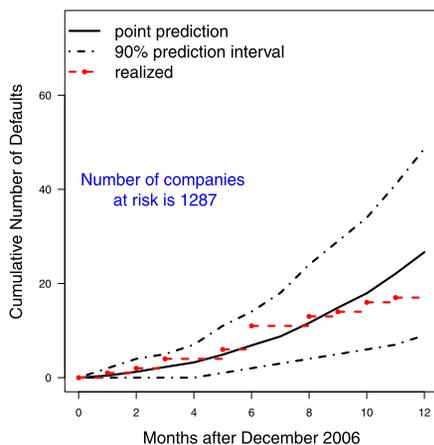
(b) Defaults Predictions for 2007
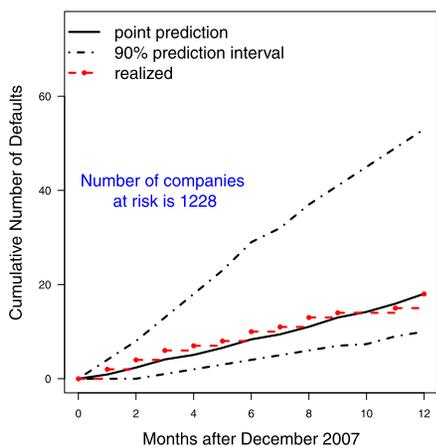
(c) Defaults Predictions for 2008

Defaults Predictions for 2009

FIG. 1. *Cumulative number of defaults in the one-year periods and the associated PI for all the units at risk.*

increasing over time. That is, one always needs to take more caution when applying predictions over a longer term due to higher level of associated uncertainties.

Additionally, we also see that predictions with the same model are not working equally well for all four one-year periods. For 2007 and 2008, the mean predictions agrees well with the actual cumulative numbers of defaults, showing that the model works very well for these periods of time. For 2006, the actual number of defaults is above the mean prediction but still falls between the mean prediction and the upper bounds of the PI (i.e., the 95% percentile). Such an observation may indicate that the actual situation in 2006 was somewhat different from what happened

before so that the model used was not able to perfectly reflect the future situation. Nevertheless, the approach still performs reasonably well by observing the fact that the PI is narrow for 2006 and the actual number of defaults still fall within the PI. In 2009, the mean predictions well matches the actual cumulative number of defaults in the first six months. However, as we can see from the figure, there is an abrupt change in the defaults occurrence by observing that no default was recorded in the second half of 2009, a phenomenon that may be due to the governmental interventions. Our data used for prediction are only up to the beginning of 2009, thus predicting such an abruption is hard for a quantitative method. Further investigation on modifying the parametric modeling may be needed to incorporate sudden change of the market conditions.

4.4. *Individual default risk predictions and uncertainties.* We now present the performance of the point predictions and PIs for future default probabilities of individual firms. We find from our studies that the level of uncertainties associated with the point predictions, quantified by the width of PIs, can be highly informative in analyzing and predicting default risks.

For the same four one-year periods (2006, 2007, 2008, and 2009) as in Section 4.3, we assess each individual firm's default risks with monthly point predictions for its future default probabilities and the associated PIs using Algorithm 3 in Section 3.3. The results for some selected firms are reported in Figure 2. We choose four companies that actually went default during the periods of time. For comparisons, we also present side by side another four companies from the same industrial sector but did not go default. By presenting the results in the same scale, we clearly see striking differences between the profiles of the default predictions. Specifically, those companies who actually went default are predicted to have substantially higher probabilities of going default. Additionally, the associated PIs are also much wider than those companies who actually did not go default.

From the predictions for individual firms, it is promising to observe that the predicted default probabilities for those companies who actually went default are among the highest, providing us a crucial device for differentiating companies based on quantitative credit risk assessments. Emerge Interactive Inc., defaulted in 2006, was a technology company providing food-safety, individual-animal tracking, and supply-management services. According to the point default probability predictions at the end of 2006, Emerge Interactive Inc. has the highest default risk among all the five companies that actually defaulted in 2006. Overall, its predicted default probability is ranked the 8th out of all 1352 companies at risk. Lehman Brothers Holdings Inc., defaulted in 2007, was the fourth largest investment bank in the US. The predicted default probability of Lehman Brothers in 2007 is ranked the 141st out of all 1287 companies at risks. Bankunited Financial Corp, defaulted in 2008, was a savings and loan association. Its predicted default probability is the 4th out of all 1228 companies at risk. TierOne Corporation, defaulted in 2009, was
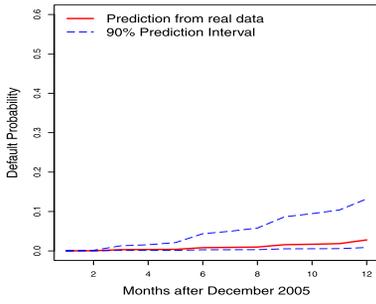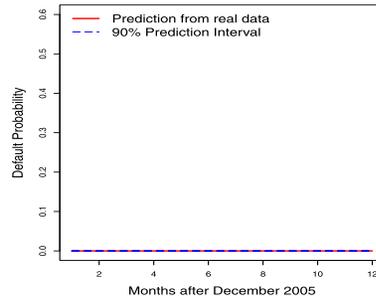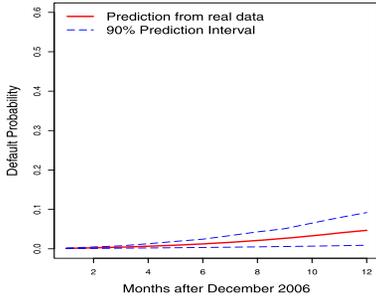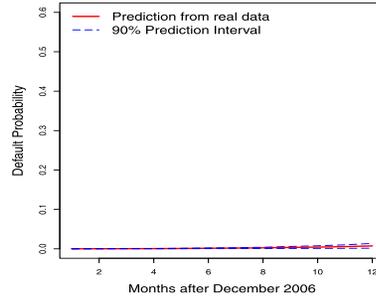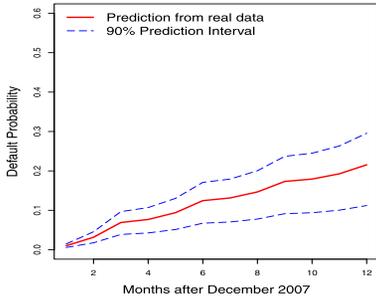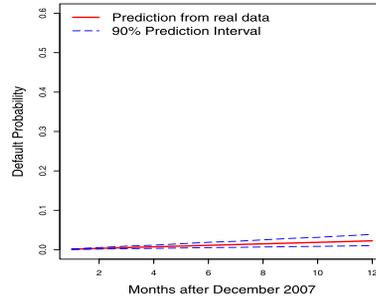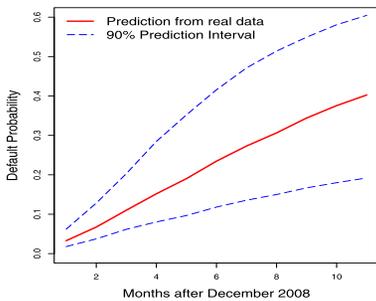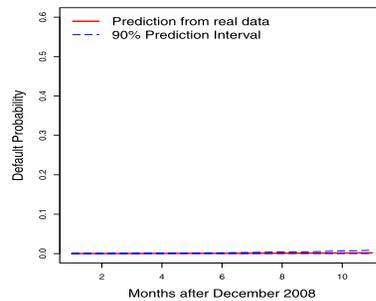
FIG. 2. *Predictions for individual default probabilities and the associated* 90% *PI.*

the holding company for TierOne Bank. Its predicted default probability in 2009 is ranked the 9th out of all 1228 companies at risk.

Besides the level of default risks assessed by the point predictions, the associated PIs are providing information from a new dimension. Visually, it is clear to see that the PIs are wider for those actually defaulted companies. Numerically, for example, the point prediction for the probability of Emerge Interactive Inc going default in 2016 is 0.0347 and the associated 90% PI is (0.0087, 0.1319). In contrast, the counterparts for Microsoft are 0.00002 and (0, 0.00005), indicating striking differences between high risk and low risk companies. As shown in the coming Section 4.6, we actually find that the level of uncertainties measured by the length of the PIs can provide extra information additional to the point predictions that can be potentially used for improving the accuracy of default predictions.

4.5. *Power curves and prediction performances.* We now evaluate the out-of-sample default prediction performances for the four one-year periods. For such a purpose, we plot the receiver operating characteristic (ROC) curves in Figure 3, which are also referred to as power curves in the literature, for example, Duffie, Saita and Wang (2007). A power curve is constructed by plotting the cumulative fraction of actual defaults versus the corresponding percentile the quantitative measure used to predictively rank all firms at risk. That is, a steeply increasing curve is the evidence of good performance using the corresponding ranking measure. Equivalently, larger area under the curve (AUC) means better predictive performance. Here, we consider two quantities—the predicted default probabilities and the lengths of the associated PIs—for ranking all firms at risk to differentiate the defaulted firms. Figure 3(a) and 3(b) respectively show the power curves corresponding to these two quantities.

From Figure 3, we can see that both predictive quantities have reasonable prediction performances, achieving AUCs near 0.9 out of the maximum 1. This again demonstrates the promising applications of quantitative methods for predicatively assessing corporate default risks. The predicted point default probability overall performs slightly better than the width of the PI. Since the width of PI is not intended for predicting future defaults, such an observation itself is interesting and informative and further research on credit risks and their evaluations are needed for understanding such a phenomenon. Moreover, we find that the width of PI is complementary to the point prediction of default probability; see Section 4.6.

4.6. *Default predictions and associated uncertainties.* We see from Section 4.5 that point default probability predictions and the width of the associated PIs perform similarly effective for differentiating defaulted firms. Then a natural question of interests is can the level of uncertainties measured by the width of PIs provide extra information for enhancing the prediction performance? As an attempt to explore the answer for that question, we conduct a logistic regression with

the default status as the response variable and both the predicted default probability and the width of the associated PIs as predictor. Summary of the model fitting is reported in Table 2.

As for the adequacy of the logistic regression model fitting, the deviance of the model is 578.61, while the deviance for the null model is 400.79. A chi-square test yields a p-value less than 0.001, showing that the model is highly significant. For the overall goodness of fit, we also did the Hosmer–Lemeshow test whose p-value is 0.5335, indicating that the model provides a good fit to the data. For comparison, we fit another logistic regression model dropping the length of the prediction interval from the model. As a result, the Hosmer–Lemshow test Hosmer, Lemeshow and Sturdivant (2013) of the reduced model has a p-value less than 0.001, a significant evidence that the reduced model without the PI length is not adequate.

Table 2 confirms that both larger predicted default probability and wider PI indicate higher default risk. However, given the point prediction and the interaction between the two variables, the width of PI is no longer significantly associated with the default risk. Most interestingly, however, a highly significant interaction between the two predictors is detected by the logistic regression, telling that using the PI width besides the point default predictions is statistically informative. Such an observation is quite reasonable from the perspective that one could be more confident in predicting a company's future default with smaller range of the prediction intervals. That is, the width of PI has the potential of providing extra information for assessing the corporate default risks, suggesting an interesting topic for further investigations.

4.7. *Model diagnostics.* For assessing the adequacy of the model fitting to the data set, we conduct some model diagnostics. For assessing the effect of the
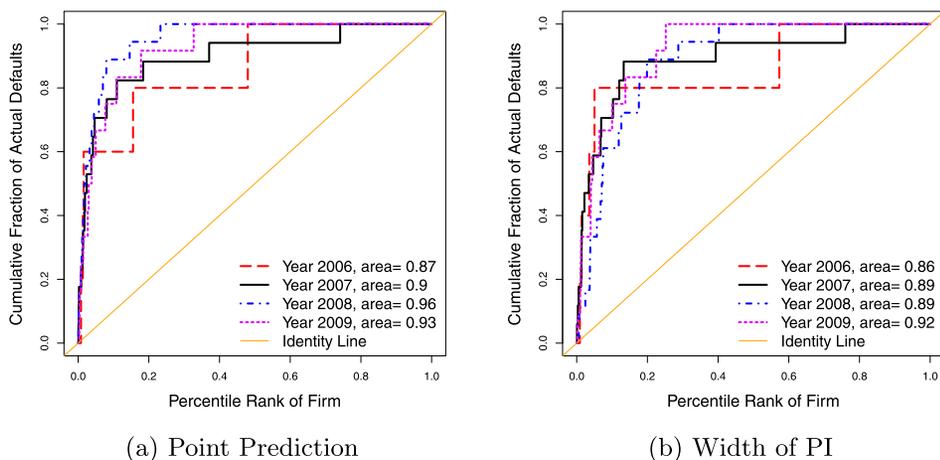

(a) Point Prediction   (b) Width of PI

FIG. 3. *Out of sample prediction power curves by the point prediction and width of PI.*

TABLE 2
Summary of the Logistic regression model

	Estimate	**Std. Error**	*z* value	**Pr(> |z|)**
Intercept	−6.1340	0.2898	−21.1691	< 0.0001
PI width	2.0742	3.3577	0.6177	0.5367
Point prediction	49.6683	6.9704	7.1256	< 0.0001
PI width × Point prediction	−99.8712	14.4081	−6.9316	< 0.0001

(a) Defaults

(b) Other Exits
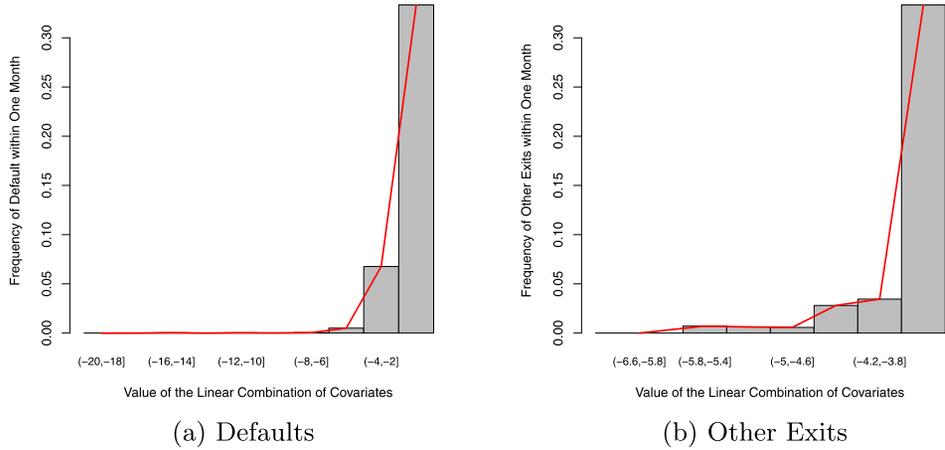
FIG. 4. *Empirical frequency of monthly defaults vs values of the estimated linear combinations broken down into ten intervals. The red solid line is the estimated exponential function.*

dynamic factor model specified by (2.3) and (2.4), we attempted a fitting of a two-factor model but with no dynamic structure (2.4). As a result, we found that the dynamic model improves the fitting in the sense of reducing the mean residual sum of squares by 10%. To check the exponential linear form of the intensity functions, we calculate the estimated values of the specified linear functions and break down the range of the values into 10 intervals. Then we aggregate the companies according to 10 intervals of the values of the estimated linear functions, and then obtain the respective total numbers of the companies. The empirical frequencies of the defaults respectively on each interval are reported by bars in Figure 4(a), overlayed by the values of the exponential linear functions. The shape of the red line in Figure 4(a) satisfactorily validates exponential linear form of the intensity function for defaults. Similarly, we obtain Figure 4(b) and validates the exponential form of the intensity function for other types of exits.

For checking the adequacy of the model for the covariate process, we conducted some numerical and graphical model diagnostics for checking the goodness of fit. We compute the fitted covariate values according to the model and compare
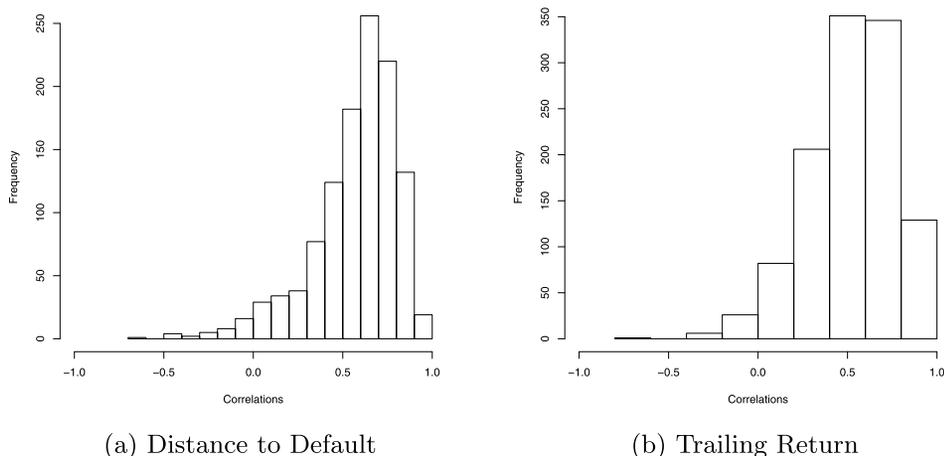
(a) Distance to Default   (b) Trailing Return

FIG. 5. *Correlations between the fitted covariates and observed covariates for the two firm specific variables.*

them with the actual observed values. In an ideal situation the fitted values and the observed values should display strong positive correlations. Figure 5 shows the histograms of the correlations between the fitted covariates and the observed values for the two firm specific covariate. From Figure 5, we can see that a majority of the correlations are reasonably high, indicating an overall good fit of the models. The average correlations for the distance to default and the trailing returns are 0.57 and 0.52 respectively. Given the large number of individual companies, some lack of fit is inevitable, and some dedicated further adjustment for the models can also be possible. The fitting of the covariate model for the two macro economic variables is also reasonably good. The correlations between the fitted values and the observed values are 0.56 and 0.82 respectively for the returns of the Treasury bill and the SP 500 index. Overall, the model for the covariates modeling is flexible and fit the data reasonably well considering that we use one model for covariates of so many companies.

5. A simulation study. The simulation setting is constructed based on the data set of Section 4 on US corporate defaults from 1990 to 2009. Specifically, we first take a random sample of size n from the set of the companies that are at risk as of the end of year 2008. We vary $n \in \{400, 600, 800, 1000\}$ to assess the impact from the scale of the problem on the accuracy of the framework. Intuitively with the same period of time, the larger the n is, the more difficult it is to predict future defaults.

Upon selecting the n companies, we generate both the events of default and other types of exist based on the parametric intensity models outlined in Sections 2.3 and 4.2. To ensure reasonable numbers of events in the simulations and for

simplicity, we set the same intensity functions for both events of defaults and other exists, particularly, with the same parameters $\beta = (-5.26, 0.1, -1.2, -0.045, -0.084)^{\mathrm{T}}$. These values closely mimic those estimated from the real data set as reported in Table 1.

We then estimate the parameters of the covariate model outlined in Section 2.4 based on the n companies, and fixed the values of the parameters throughout the simulations with size n. With the estimated parameters, we simulate the monthly differenced values of the covariate from the same model as outlined in Section 2.4. Then we generate the covariate process from the simulated differenced values, with the first observations of the companies taken from that of the random sample of size n. Then, with the simulated covariate process, we generate both events of defaults and other exits with the parametric intensity functions specified as described earlier.

Upon generating the simulated data set with both the covariate process and the time-to-events, we apply our method to produce prediction intervals of the cumulative number of defaults in each month of 2008. We also compare the two types of prediction intervals—calibrated and uncalibrated ones—as outlined in Section 3.2. The simulations for each sample size n are repeated 240 times. The results of the accuracy in terms of the percentage that the prediction intervals cover the true cumulative default numbers are reported in Figure 6.

From Figure 6, we have a few observations. First and foremost on the empirical accuracy of the uncertainties assessment, we observe that the empirical frequencies of the prediction coverages are close to the nominal levels for all multiple period predictions when n is smaller. When n is larger, the coverage of the PI is also very accurate for the cumulative default counts within a closer time horizon from the origin of predictions. Second, we see that the coverages generally get worse for larger time horizon cumulative predictions. Since we are examining the cumulative predictions, the main reason should be the error aggregations in the cumulative counts predictions. That is, even the coverages of the prediction intervals for the number of events at each individual month are close to the nominal level, the performance of the PI for cumulative counts will still be more off the target due to that all errors are aggregating together. Given the same amount of information, this also reflects the practical difficulty in obtaining accurate predictions for longer time horizons. Third, the calibrated intervals perform substantially better than the uncalibrated ones, indicating the merits of applying calibrated procedures for prediction intervals. Since the naive prediction intervals (without calibration) only capture the intrinsic randomness in the random variables and ignores the uncertainty from parameter estimations, they tend to be narrow so that the coverage tends to be smaller than the nominal level. In contrast, the advantage of calibration relies on the fact that it takes the additional source of uncertainties into account. Hence the coverage of the calibrated prediction intervals tend to be closer to the nominal level.

6. Discussions and future work. We consider the challenging problem of assessing uncertainties associated with corporate default predictions by carefully disentangling and quantifying the contributing sources for the point predictions. An application of our framework to a large-scale US Corporate data set shows that our point predictions have good out-of-sample performance, and is promising in quantifying the uncertainties in predictions. Our framework also helps for a better understanding of the default mechanism by providing an additional dimension of insights from assessing the level of uncertainties associated with point predictions. With limited access to a powerful modern computational facility (160 hours in parallel on 80 CPUs), we accomplish the tasks for assessing the uncertainties associated with corporate default predictions, demonstrate the feasibility for solving
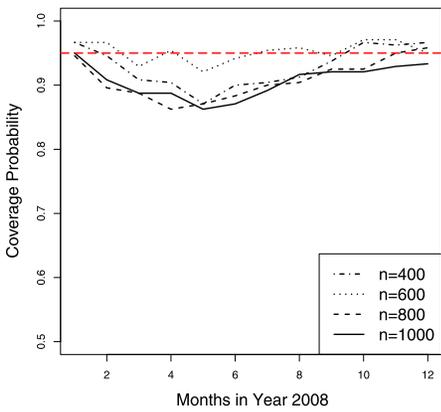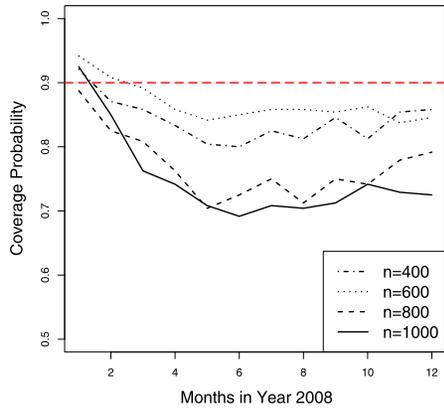

(a) 95% calibrated PI
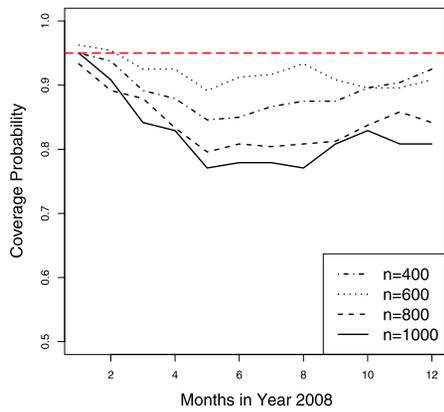
(b) 95% uncalibrated PI
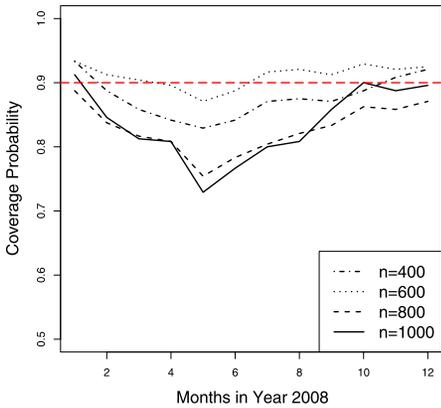
(c) 90% calibrated PI

(d) 90% uncalibrated PI

FIG. 6. *Coverage probability of calibrated and uncalibrated PI from simulation study.*

this long overdue important large scale practical problem with many challenging practical features. Our study reveals high level of uncertainties associated default predictions especially when conducting longer term predictions. We believe that is mainly due to the nature of the prediction problem involving unknown future dynamics of those factors affecting the default mechanism. Hence more cautions are necessary when using quantitative tools for mid- and long-term default predictions. We also note that the rationale of the quantitative modeling is to predict future events based on historical information. Thus for all model based quantitative predictions, they should be interpreted as if the market conditions are consistent with the historical scenarios, then the predictions are valid. Otherwise, one has to take serious cautions.

Further investigations for assessing the uncertainties associated with corporate default risk predictions are clearly desirable. Our framework needs parametric models for the default mechanism and the covariate processes. Extending the scope of the framework and evaluating its robustness are clearly important. For example, other methods dealing with the default mechanism may be considered. Additional considerations on features such as the cycling effect Koopman and Lucas (2005) and systemic risk Giesecke and Kim (2011) can also be investigated. Interesting questions also include how to efficiently incorporate more variables, and how a violation of the parametric models may affect the accuracy of the assessed level of uncertainties. There is another important consideration on the correlations between defaults. Recent investigations reveal the correlations and even clustering effect of the occurrences of corporate defaults; see, among others, the frailty modeling approaches with random effects in Duffie et al. (2009), the dynamic frailty modeling approaches of Koopman, Lucas and Monteiro (2008) and Koopman, Lucas and Schwaab (2011), and the jump cumulative default intensity function approach of Peng and Kou (2009). It will also be interesting to explore the impact from the random and even clustering effects in constructing the prediction intervals. We hope to conduct further investigations on these problems in future. Both theoretical and practical investigations are also needed for exploring the high-dimensional covariate process for large number of companies with the global market data. Instead of resorting to the simulation based approach we take, a potential direction for assessing uncertainties associated with default predictions could be applying the Bayesian approaches that are capable of producing the posterior distribution of the quantities of interests. How to develop parsimonious and effective Bayesian modeling, and how to design and implement efficient practical computational framework are interesting and open challenging problems.

SUPPLEMENTARY MATERIAL

Supplement to "Disentangling and assessing uncertainties in multiperiod corporate default risk predictions" (DOI: 10.1214/18-AOAS1170SUPP; .pdf). Detail of the EM algorithm.

REFERENCES

ALTMAN, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *J. Finance* **23** 589–609.

ALTMAN, E., RESTI, A. and SIRONI, A. (2004). Default recovery rates in credit risk modelling: A review of the literature and empirical evidence. *Economic Notes* **33** 183–208.

BAI, J. and NG, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics* **3** 89–163.

BAI, J. and WANG, P. (2015). Identification and Bayesian estimation of dynamic factor models. *J. Bus. Econom. Statist.* **33** 221–240. MR3337059

BAŃBURA, M. and MODUGNO, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *J. Appl. Econometrics* **29** 133–160. MR3233736

BEAVER, W. H. (1966). Financial ratios as predictors of failure. *J. Acc. Res.* **4** 71.

BEAVER, W. H. (1968). Market prices, financial ratios, and the prediction of failure. *J. Acc. Res.* **6** 179.

BEAVER, W. H., CORREIA, M. and MCNICHOLS, M. F. (2012). Do differences in financial reporting attributes impair the predictive ability of financial ratios for bankruptcy? *Rev. Acc. Stud.* **17** 969–1010.

BHARATH, S. T. and SHUMWAY, T. (2008). Forecasting default with the merton distance to default model. *Rev. Financ. Stud.* **21** 1339–1369.

BÖHNING, D., DIETZ, E., SCHAUB, R., SCHLATTMANN, P. and LINDSAY, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math.* **46** 373–388.

BRÄUNING, F. and KOOPMAN, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *Int. J. Forecast.* **30** 572–584.

CAMPBELL, J. Y., HILSCHER, J. and SZILAGYI, J. (2008). In search of distress risk. *J. Finance* **63** 2899–2939.

CHAVA, S. and JARROW, R. A. (2004). Bankruptcy prediction with industry effects. *Rev. Finance* **8** 537–569.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.

DING, A. A., TIAN, S., YU, Y. and GUO, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *J. Amer. Statist. Assoc.* **107** 990–1003. MR3010885

DUAN, J.-C., SUN, J. and WANG, T. (2012). Multiperiod corporate default prediction—A forward intensity approach. *J. Econometrics* **170** 191–209. MR2955949

DUFFIE, D. (2011). *Measuring Corporate Default Risk*. Oxford Univ. Press, Oxford.

DUFFIE, D. and LANDO, D. (2001). Term structures of credit spreads with incomplete accounting information. *Econometrica* **69** 633–664. MR1828538

DUFFIE, D., SAITA, L. and WANG, K. (2007). Multi-period corporate default prediction with stochastic covariates. *J. Financ. Econ.* **83** 635–665.

DUFFIE, D., ECKNER, A., HOREL, G. and SAITA, L. (2009). Frailty correlated default. *J. Finance* **64** 2089–2123.

DURBIN, J. and KOOPMAN, S. J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed. *Oxford Statistical Science Series* **38**. Oxford Univ. Press, Oxford. MR3014996

GIESECKE, K. and KIM, B. (2011). Systemic risk: What defaults are telling us. *Manage. Sci.* **57** 1387–1405.

GIESECKE, K., LONGSTAFF, F., SCHAFER, S. and STREBULAEV, I. (2011). Corporate bond default risk: A 150-year perspective. *J. Financ. Econ.* **102** 233–250.

HILLEGEIST, S. A., KEATING, E. K., CRAM, D. P. and LUNDSTEDT, K. G. (2004). Assessing the probability of bankruptcy. *Rev. Acc. Stud.* **9** 5–34.

HONG, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Comput. Statist. Data Anal.* **59** 41–51. MR3000040

HONG, Y., MEEKER, W. Q. and MCCALLEY, J. D. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *Ann. Appl. Stat.* **3** 857–879. MR2750685

HOSMER, D. W., LEMESHOW, S. and STURDIVANT, R. X. (2013). *Applied Logistic Regression*. John Wiley and Sons Ltd.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR1924807

KOOPMAN, S. J. and LUCAS, A. (2005). Business and default cycles for credit risk. *J. Appl. Econometrics* **20** 311–323. MR2221901

KOOPMAN, S. J. and LUCAS, A. (2008). A non-Gaussian panel time series model for estimating and decomposing default risk. *J. Bus. Econom. Statist.* **26** 510–525. MR2459348

KOOPMAN, S. J., LUCAS, A. and MONTEIRO, A. (2008). The multi-state latent factor intensity model for credit rating transitions. *J. Econometrics* **142** 399–424. MR2408742

KOOPMAN, S. J., LUCAS, A. and SCHWAAB, B. (2011). Modeling frailty-correlated defaults using many macroeconomic covariates. *J. Econometrics* **162** 312–325. MR2795620

KOOPMAN, S. J., LUCAS, A. and SCHWAAB, B. (2012). Dynamic factor models with macro, frailty, and industry effects for U.S. default counts: The credit crisis of 2008. *J. Bus. Econom. Statist.* **30** 521–532. MR2990495

LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663

LAWLESS, J. F. and FREDETTE, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92** 529–542. MR2202644

MEEKER, W. Q. and ESCOBAR, L. A. (1998). *Statistical Methods for Reliability Data*. Wiley, New York.

MERTON, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *J. Finance* **29** 449–470.

OHLSON, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *J. Acc. Res.* **18** 109.

PAN, J. and YAO, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95** 365–379. MR2521589

PENG, X. and KOU, S. (2009). Default clustering and valuation of collateralized debt obligations. Working paper.

SCHWAAB, B., KOOPMAN, S. J. and LUCAS, A. (2017). Global credit risk: World, country and industry factors. *J. Appl. Econometrics* **32** 296–317. MR3631944

SHUMWAY, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *J. Bus.* **74** 101–124.

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271

TSAY, R. S. (2010). *Analysis of Financial Time Series*, 3rd ed. Wiley, Hoboken, NJ. MR2778591

TSAY, R. S. (2014). *Multivariate Time Series Analysis*: *With R and Financial Applications*. Wiley, Hoboken, NJ. MR3236787

VOLKOVA, A. Y. (1996). A refinement of the central limit theorem for sums of independent random indicators. *Theory Probab. Appl.* **40** 791–794. MR1405154

YUAN, M., TANG, C. Y., HONG, Y. and YANG, J. (2018). Supplement to "Disentangling and assessing uncertainties in multiperiod corporate default risk predictions." DOI:10.1214/18-AOAS1170SUPP.

ZMIJEWSKI, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *J. Acc. Res.* **22** 59.

M. YUAN
Y. HONG
DEPARTMENT OF STATISTICS
VIRGINIA TECH UNIVERSITY
213 HUTCHESON HALL
BLACKSBURG, VIRGINIA 24060
USA
E-MAIL: miaoy89@vt.edu
        yilihong@vt.edu

C. Y. TANG
DEPARTMENT OF STATISTICAL SCIENCE
TEMPLE UNIVERSITY
1810 NORTH 13TH ST
PHILADELPHIA, PENNSYLVANIA 19122
USA
E-MAIL: yongtang@temple.edu

J. YANG
BUSINESS SCHOOL
UNIVERSITY OF COLORADO DENVER
1475 LAWRENCE ST
DENVER, COLORADO 80202
USA
E-MAIL: Jian.Yang@ucdenver.edu