Corporate Default Predictions and Methods for Uncertainty Quantifications

Miao Yuan

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Yili Hong, Chair
Inyoung Kim
Laura P. Sands
Hongxiao Zhu

17th June, 2016
Blacksburg, Virginia

Key Words: Default Risk; Dynamic Factor Model; High-Dimensional Time Series; (Log)-Location-Scale Distributions; Tolerance Interval.

Corporate Default Predictions and Methods for Uncertainty Quantifications

Miao Yuan

Abstract

Regarding quantifying uncertainties in prediction, two projects with different perspectives and application backgrounds are presented in this dissertation.

The goal of the first project is to predict the corporate default risks based on large-scale time-to-event and covariate data in the context of controlling credit risks. Specifically, we propose a competing risks model to incorporate exits of companies due to default and other reasons. Because of the stochastic and dynamic nature of the corporate risks, we incorporate both company-level and market-level covariate processes into the event intensities. We propose a parsimonious Markovian time series model and a dynamic factor model (DFM) to efficiently capture the mean and correlation structure of the high-dimensional covariate dynamics. For estimating parameters in the DFM, we derive an expectation maximization (EM) algorithm in explicit forms under necessary constraints. For multi-period default risks, we consider both the corporate-level and the market-level predictions. We also develop prediction interval (PI) procedures that synthetically take uncertainties in the future observation, parameter estimation, and the future covariate processes into account.

In the second project, to quantify the uncertainties in the maximum likelihood (ML) estimators and compute the exact tolerance interval (TI) factors regarding the nominal confidence level, we propose algorithms for two-sided control-the-center and control-both-tails TI for complete or Type II censored data following the (log)-location-scale family of distributions. Our approaches are based on pivotal properties of ML estimators of parameters for the (log)-location-scale family and utilize the Monte-Carlo simulations. While for Type I censored data, only approximate pivotal quantities exist. An adjusted procedure is developed to compute the approximate factors. The observed CP is shown to be asymptotically accurate by our simulation study. Our proposed methods are illustrated using real-data examples.

Corporate Default Predictions and Methods for Uncertainty Quantifications

Miao Yuan

General Audience Abstract

This dissertation focuses on developing new prediction methods with quantifications of uncertainties within the predictions. The first project in Chapter 2 contributes to the corporate default risk predictions based on event time data as well as high-dimensional corporate and macroeconomic indexes recorded over time. The second project in Chapter 3 proposes a general algorithm to compute an interval (i.e., a tolerance interval or TI) that can cover any specified content (e.g., 90%) of all the future observations based on existing samples following (log)-location-scale family of distributions.

In Chapter 2, we develop innovative methods to predict the multi-period default risks for individual companies and the number of defaults in the market. Our model is especially useful in debt rating and pricing, also controlling risks in the financial system. Specifically, there are several highlights in our methods. First, we incorporate both company-level and market-level financial indexes into the risks for exits of companies due to default and other reasons. The co-movement of these indexes over time (i.e., thousands of time series totally for these indexes) drives the stochastic and dynamic nature of the corporate risks. Second, we propose a parsimonious Markovian time series model to remove the long-time trend and the grand mean of these indexes, then achieve dimension reduction by finding only a few latent factors evolving over time can capture the co-movement of thousands of de-trended indexes for individual companies and the macroeconomics. We derive an efficient parameter estimation procedure all in explicit mathematical forms. Third, based on the predicted values of these indexes, we predict the multi-period default risks at the corporate level and market level. We also assess the uncertainties within the predictions. The application of our methods on a large-scale dataset of the US market shows our predictions are accurate and the quantification of uncertainties can effectively reveal how confident we are with these predictions.

In Chapter 3, we compute the commonly used types of two-sided TI for samples (probably including incomplete observations) following the (log)-location-scale family of distributions. The specified confidence level is guaranteed through accurately quantifying the randomness in the existing observations. For which, we develop a general algorithm to do re-sampling and simulate the distribution of pivotal quantities (for (log)-location-scale family of distributions, the probability distribution of pivotal quantities do not depend on the true parameter values, which makes the simulation procedure possible. For some types of incomplete observations, this property holds only approximately). We then obtain a curve of eligible TI factors from a large grid of values. In this way, our algorithm successfully overcomes the challenges caused by the fact that there exist no explicit forms for parameter estimators of most distributions in these two families, and controls the uncertainties in both endpoints of the TI simultaneously. Our proposed methods are illustrated using real-data examples.

Dedication

To my family.

Acknowledgments

I owe my deepest gratitude to my advisor - Dr. Yili Hong. With his patient guidance, I become a qualified PhD candidate with rewarding research experience. I do not only obtain deeper understanding of statistics and enlightening research insights, but also acquire scientific and generally applicable research methods. I also appreciate my committee members Dr. Inyoung Kim, Dr. Runlong Tang, Dr. Hongxiao Zhu, and Dr. Laura Sands for their supports, time and insightful questions to make my research work better.

I would like to further acknowledge Dr. Laura Sands to support me with a research assistant position. I feel so lucky to work on several projects with her, from which I learn valuable knowledge and research insights in the health-care and clinical fields, as well as how to apply statistical methods to solve research questions in these fields.

I am also thankful to the Department of Statistics which provides me the opportunity to know smart and kind professors, to learn comprehensive statistical knowledge, to study with outstanding students, and to gain valuable working experience from teaching and statistical collaboration. I want to acknowledge the advanced research computing of Virginia Tech to provide me with the powerful computing resources. Without them, none of my research projects can proceed.

I am also sincerely grateful to my dear friends for their warm accompany and help both inside and outside of school. I appreciate competent and supportive friends in our research group: Yuanyuan Duan, Zhibing Xu, Khaled Bedair, Caleb King, Yimeng Xie, I-Chen Lee and Zhongnan Jin.

I want to give special thanks to my fiancé Di Hu. Last but not least, I deeply appreciate my family for their tremendous and constant love, patience, encouragement and supports both mentally and financially through my life and graduate study. I cannot thank enough to my parents, grandparents and my twin sister for their inestimable contributions to my achievements.

Contents

1 General Introduction **1**

1.1 Background 1

1.1.1 Default Prediction 1

1.1.2 Tolerance Intervals 5

1.2 Motivation 6

1.2.1 Correlation Among Individual Defaults 6

1.2.2 Estimation of Parameters in the Covariate Dynamics 8

1.2.3 Tolerance Interval Under Right-Censored Data Following (Log)-Location-Scale Distributions 8

1.2.4 Quantifying Uncertainties in Predictions 9

1.3 Outline of this Dissertation 10

2 Disentangling and Assessing Uncertainties in Multiperiod Corporate Default Risk Predictions **14**

2.1 Introduction 16

2.1.1 Related Literature 20

2.1.2 Overview 21

2.2 Sources of Uncertainties in Default Risk Predictions 21

2.2.1 Stochastic Time-to-Event 21

2.2.2 Stochastic Covariates in Default Predictions 22

2.2.3 Parametric Intensity Function and Its Estimation 23

2.2.4 Parametric Stochastic Covariate Process and Its Estimation 27

2.3 Predictions and Uncertainties Assessments 31

2.3.1 Procedures for Future Default Predictions 31

2.3.2 Assessing Uncertainties at the Aggregate Level 33

2.3.3 Assessing Uncertainties for Corporate Default Probabilities 35

2.4 US Cooperate Default Data Analysis . . . 37
2.4.1 Parameter Estimates for Time-to-Event Models . . . 38
2.4.2 Parameter Estimates for Covariate Model . . . 39
2.4.3 Aggregate Default Predictions and Uncertainties . . . 39
2.4.4 Individual Risk Prediction and Uncertainties . . . 40
2.4.5 Prediction Performance: Power Curve . . . 43
2.4.6 Default Predictions and Associated Uncertainties . . . 45
2.5 Discussions and Future Research . . . 47
Appendices . . . 48
2.A Details of the EM Algorithm for Estimating Parameters in the Covariate Model . . . 48
2.B Calibration . . . 57

3 Two-sided Tolerance Intervals for Members of the (Log)-Location-Scale Family of Distributions 62

3.1 Introduction . . . 63
3.1.1 Motivation . . . 63
3.1.2 Literature Review . . . 63
3.1.3 Overview . . . 65
3.2 Data and Model . . . 65
3.2.1 Data . . . 65
3.2.2 Model . . . 67
3.2.3 Parameter Estimation . . . 68
3.3 Two-sided Tolerance Interval Procedures . . . 69
3.3.1 Control-the-Center TIs . . . 69
3.3.2 Control-Both-Tails TIs . . . 71
3.3.3 TIs with Equal Error Probabilities . . . 72
3.4 Computation of Tolerance Intervals . . . 73
3.4.1 Control-the-Center TIs . . . 73
3.4.2 Control-Both-Tails TIs . . . 76

3.4.3 Computation of TIs for the Log-Location-Scale Family of Distributions 76
3.5 Applications 76
3.5.1 Air Lead Level Data 76
3.5.2 Pressure Vessel Failure Data 79
3.5.3 Locomotive Control Failure Data 80
3.6 Simulations for Type I Censoring 82
3.7 Conclusions and Areas for Future Research 84
Appendices 85
3.A Proof of Result 1 85
3.B Proof of Result 2 86

4 Conclusions and Future Work **91**
4.1 Conclusions 91
4.2 Future Directions 93

List of Figures

2.1 Estimated frequency of events versus the linear combination values to check the functional form of intensities. 24

2.2 Cumulative number of defaults in the one-year periods and the associated PI for all the units at risk. 41

2.3 Predictions for individual default probabilities and the associated 90% PI. . . 44

2.4 Out of sample prediction power curves by the point prediction and width of PI. 45

3.1 Lognormal probability plot for the air lead level data. 78

3.2 Contour plots of factors for control-the-center TI. 78

3.3 Contour plots of factors for control-both-tails TI. 79

3.4 Probability plots for the pressure vessel failure data. 80

3.5 Probability plots for the locomotive control failure data. 82

3.6 Plots of the CP vs nominal confidence level for TIs with content 0.9, under Type I censored data. 84

List of Tables

2.1 ML estimates of parameters and their asymptotic standard errors based on data over January 1990 to December 2008. 38

2.2 Summary of the Logistic regression model. 46

3.1 Levels of Lead in Air ($\mu g/m^3$) 77

3.2 Tolerance intervals for the levels of lead in air data 77

3.3 Failure times of pressure vessels in hours. 79

3.4 Tolerance intervals for the pressure vessel failure data. 80

3.5 Miles to failure of locomotive controls in units of thousands of miles. 81

3.6 Tolerance intervals for the locomotive control failure data. 81

Chapter 1 General Introduction

1.1 Background

1.1.1 Default Prediction

Default prediction is to forecast the risk that a company will fail to fulfill its debt or obligations based on the performance of the firms and the macroeconomic condition. Default prediction for individual companies and the market is essential to credit risk management, business, financial and regulatory decision making. It also provides the critical basis for rating and pricing the corporate debt as well as credit portfolios.

During the past 150 years, the United States (US) market has gone through a number of major shocks of large-scale corporate defaults, in which 20-50% of all corporate bonds defaulted, one can see Giesecke et al. (2013) for more details. From Duan (2010) and Peng and Kou (2009), the effectiveness of existing credit rating models has been seriously doubted since the outburst of 2007-2009 financial crisis. As a result, it is urgent to discover the mechanism of default and develop effective models to assess the risks for corporate defaults.

Existing models for default prediction can be roughly divided into two categories: indirect assessment based on Poisson process of events that requires forecasting the future covariate dynamics or direct computation without modeling the covariate process. The former includes Duffie et al. (2007), Duffie et al. (2009), Peng and Kou (2009), etc, and the latter includes Duan et al. (2012).

Most literature regarding default prediction focus on the economic sectors, e.g., the number of defaults in a credit portfolio, especially since the financial crisis. Two general categories of such models are "bottom-up" and "top-down", one can refer to Peng and Kou (2009) and Duan (2010) for extensive introduction of these two types of models.

1.1.1.1 Time-to-Event Data

Time-to-event data often arise in reliability and survival analysis. They refer to the time to occurrence of the event of interest. Such event can be failure of a product in a life test, the first exacerbation or recurrence of some disease, resolution of pain, and death, etc. It is common for time-to-event data to have censoring or truncated observations. In this dissertation, we mainly deal with right-censored data, which means the time point when the event occurs is beyond the duration of study time. Thus we only know the event happens after the end of the study.

1.1.1.2 Intensity Functions

To model the probabilistic distribution of time-to-event, an important function to define is the hazard or intensity function $\lambda(t)$, which gives the instantaneous rate of occurrence of the corresponding event at time t (Lawless, 2003). The intensity function can be specified as a parametric, semiparametric or nonparametric function (perhaps incorporating covariates). Intensity function of an event determines the form of distribution functions of the time-to-event. The commonly used parametric distributions for time-to-event data include the (log)-location-scale family of distributions (e.g., (log)-normal, smallest extreme value, Weibull, (log)-logistic distributions, see Meeker and Escobar 1998 and Lawless 2003). Most of them (except (log)-normal) have the corresponding intensity functions in explicit forms. One example of the famous semiparametric models is the Cox proportional hazard model, which can use nonparametric functions as the baseline hazard. One may see details about Cox proportional hazard model in Klein and Moeschberger (2003). Another example for semiparametric methods uses a linear combination of B-splines to model the event intensity, as proposed by Rosenberg (1995). For nonparametric methods, we can use the Nelson-Aalen estimator of the cumulative hazard function (see Lawless 2003).

1.1.1.3 Competing Risks

Competing risks often arise in modeling time-to-event data, since there are always more than one possible causes for failures/successes. The multiple causes are considered as mutually exclusive events. That is, the occurrence of one type of event makes the time to other types of events censored/infinite. In this way, the sub-distribution (here it is called a sub-distribution because the function is still less than 1 as time goes to infinity) of time to failure due to the specific event of interest is conditional on the fact that other types of events do not happen before the event of interest. Ignoring the risks of other types of events biases the estimated distribution of time to the event of interest due to failure in identifying the systematic causes of random censoring. The competing risks are often assumed to be conditionally independent given covariates and thus are additive. That is, the instantaneous rate of failure is equal to the sum of rates of different events. In our data, companies can exit the market due to default and other types of events (e.g., acquisition and merge).

1.1.1.4 Covariate Process and Dynamic Factor Model

Since default risk can intrinsically be revealed by firm-specific and macroeconomic factors, many literature regarding default prediction or the collateralized debt obligations (CDO) pricing use stochastic intensity functions incorporating covariate dynamics for events like default. Especially, the "bottom up" models generally predict the number of defaults within a credit portfolio by modeling intensity functions of individual companies. Majority of the studies assume defaults of different companies are conditionally independent given their covariate processes, which means the correlation among individual defaults can only be captured through their covariate processes, or only through some macroeconomic-factor dynamics.

A DFM is an extended factor model used in a multivariate time-series context. That is, co-movement of the observed time series $\mathbf{Y}_t$ can be driven by a low-dimensional latent factor process $\mathbf{F}_t$ (more generally, several lags of factors influence $\mathbf{Y}_t$ together, see Bai and Wang

2012) plus an idiosyncratic component $\mathbf{e}_t$. The latent factor $\mathbf{F}_t$ is assumed to follow a vector autoregressive process with order h (VAR(h)). Since the economic variables are usually driven by a few common factors, the benefit to use a low-dimensional factor representation is considerable when the economic or financial time series has a high dimension and the data are only available during relative short period of time. In this situation, modeling the co-movement of factors instead of the economic variables shows much more efficiency, and makes the estimation of parameters easier and more stable. Several studies about the DFM show that only few dynamic factors can explain satisfactory proportion of variability in many economic variables, one may refer to Stock and Watson (2011) for details.

1.1.1.5 Kalman Filtering and Smoothing

Kalman filtering and smoothing are recursive procedures to obtain inferences on a state-space model. A basic state-space model generally consists of a model for the observed measurements explained by a possibly latent state vector at time t, and a Markovian time series (or VAR) model that describes the dynamics of the state vectors over time. The state-space model can be nonlinear and non-Gaussian, but here we only introduce the Kalman filtering and smoothing procedures for a linear and Gaussian state-space model. In Chapter 2, the dynamic factor model (DFM) has a state-space model representation, with the residual vector $\boldsymbol{\varepsilon}_t$ be the observed measurements and the dynamic factors $\mathbf{F}_t$ be the state vector that is driven by a VAR(1) process.

Kalman filter is a sequential updating inference procedure for the state vector and was first derived by Kalman (1960). It requires specification of an initial state distribution and a linear state-space model with Gaussian errors. As new observations are received over time, the Kalman filter calculates the normal distribution of each new state vector at time t using the observations at time t and the estimated distribution of the previous state vector. Because both the initial state and the error terms of the state-space model are Gaussian, and state-space model shows linear relationship, the derived distributions of each observed

response vector and state vector are Gaussian. When obtain a new observation, Kalman filter updates the mean and covariance matrix of the corresponding new state vector through a set of recursion equations, which can be derived in a Bayesian manner.

After applying Kalman filter, we have the derived normal distribution of each state vector at time t given information up to the same time t. While Kalman smoother can use all the information up to the last observation time (LOT, denoted as T here) to inform the estimation of all the previous state vectors. That is why the smoother is called a retrospective update procedure. Starting from the estimation of the latest state vector obtained from the filter, it smooths the estimation of each previous state vector in a backward manner recursively.

Kalman filter and smoother are often used in time series model to estimate the likelihood. In Chapter 2, we apply the filter to update the conditional moments of the dynamic factors sequentially over time, and then use the smoother to incorporate all the available data into the estimation of these moments. Then we plug in the estimated moments of the dynamic factors at all the time points from the smoother into the conditional log likelihood in the E-step to replace the latent factors.

1.1.2 Tolerance Intervals

For univariate distributions, a tolerance interval (TI) is constructed so that it can cover at least a specified proportion of the population with some required confidence level, based on an observed random sample from that distribution. TIs have applications in many fields when the data are from a single univariate or multivariate distribution. For example, given service life of a sample of a certain LED product, it is of interest to calculate a TI that covers the lifetime of at least 90% of this kind of LED products with certain level of confidence. Thus, a one-sided tolerance bound (TB) is the confidence limit for a specified distribution quantile, a two-sided TI is expected to cover a pair of quantiles with the space between them to be at least a specified content of the distribution.

TI is similar to a simultaneous prediction interval (SPI), which is expected to give the bounds of one or several future observations based on a random sample from that distribution. Essentially SPI has the same function as TI, except that it is supposed to bound m out of $n, (n > m)$ future observations rather than a continuous proportion of the distribution as the TI does. One can see Xie et al. (2014) for a general algorithm to compute the SPI for data from the (log)-location-scale family of distributions. For multivariate distributions, TI and SPI can be extended to tolerance and prediction regions, respectively.

TI can also be used in the regression setting including the mixed effect model, to bound a proportion of the distribution of the response variable. Besides, nonparametric TI and Bayesian TI have also been derived. For a comprehensive introduction of TI, see Krishnamoorthy and Mathew (2009).

1.2 Motivation

1.2.1 Correlation Among Individual Defaults

We use the covariate time series to explain the underlying changes in the intensities of default and other exits over time. Treating December 2008 as the last observation time (LOT), our covariate data contain two firm-specific variables of 1,147 companies at risk and two macroeconomic variables, all spanning 228 months. That makes a covariate matrix with 2,296 rows and 228 columns. All the firms included in the covariate process model survived to the LOT, but they were formed at different time points. The challenges in modeling covariate process lie in the high-dimensionality of the covariate data, considerable amount of missing data and the complicated correlation structure among individual corporates across time. These challenges also make the estimation of parameters difficult.

Some studies generated correlation among individual default intensities only through some common market-level factors. For example, Duffie and Gârleanu (2001) and Mortensen (2006) modeled the common factors and firm-specific factors of individual companies as

independent affine-jump diffusion processes. For more details about derivation of affine-jump diffusion process, see Duffie et al. (2000). Peng and Kou (2009) only modeled the process of market factors. They used a Pólya process to model the macroeconomic factor dynamics to generate rapid increase in the cumulative intensity simultaneously for all the companies during financial crisis, while they used the trapezoidal approximation to the integral of a mild mean-reverting process under normal economic conditions.

Our covariate model is a generalization of that in Duffie et al. (2007). We both assume the conditional independence among intensities given the covariate processes, and specify the grand mean vector of the covariate processes in a Markovian time series model (i.e., the first order difference). One slight difference is that we propose to remove the quarterly trend in the covariate data by differencing before applying the Markovian time series model. More importantly, we develop a more flexible, yet still efficient method to model the serial and cross-sectional correlation structure within the covariate processes. Duffie et al. (2007) modeled the first-order differenced time series of firm-specific covariates and trailing one-year return on S&P 500 index as linear combinations of independent normal distributed vectors. The correlations among these covariate processes are modeled by specifying a common normal vector multiplying different volatility parameters. They also assume the Treasury bill rates are independent from the firm-specific covariates (except the distance to default) and the trailing one-year return on S&P 500 index.

While, we model the correlations among de-trended individual covariate processes and the macroeconomic dynamics through a dynamic factor model (DFM). In particular, we apply a DFM on the residuals $\boldsymbol{\varepsilon}_t$ of the Markovian time series model (equation (2.3) in Section 2.2.4 of Chapter 2), to capture the co-movement of the covariate time series without quarterly or general trend. In this way, we try to capture the correlation structure among all the individual and macroeconomic covariate processes by a few latent dynamic factors, also acquire better model interpretations.

1.2.2 Estimation of Parameters in the Covariate Dynamics

Parameter estimation in the DFM is not trivial due to the latency of factors, complexity of the model and missing observations. The maximum likelihood (ML) estimates of a DFM generally do not have closed-forms, and a direct numerical optimization of the log likelihood is very computationally intensive if possible. The missing observations make estimation more complicated.

There are literature focusing on the estimation of parameters in a DFM using the EM algorithm. In particular, one can see Banbura and Modugno (2012) for details about the derivation of EM algorithm for a general DFM with missing data. However, Banbura and Modugno (2012) does not fully address the identification problems in estimation of a DFM. We add the necessary constraints to ensure identification of all the parameters.

Moreover, our covariate model contains a Markovian time series model for the observed covariate processes with quarterly effect removed, which adds one more layer of complexity in the parameter estimation. Our study has addressed all these problems.

1.2.3 Tolerance Interval Under Right-Censored Data Following (Log)-Location-Scale Distributions

When censored observations are present, maximum likelihood (ML) estimators of parameters of the (log)-location-scale family of distributions generally do not have closed forms. Even normal distribution does not have explicit-form ML estimators when data are censored. Thus numerical method is needed to obtain the ML estimates. Because the one-sided TI is essentially a confidence limit for the distribution quantile, it does not have a closed-form considering the distribution quantile is an expression of the location, scale parameters and the quantile of the corresponding standard distribution in the location-scale family. In literature, one-sided TI is usually calculated by the Monte-carlo simulations using the pivotal quantities, see Krishnamoorthy and Mathew (2009) for details. Note that when data are type II censored, ML estimators of parameters of the (log)-location-scale family of distributions

have exact pivotal properties. When data are Type I censored, ML estimators of parameters of these families only have approximate pivotal properties.

Because for a two-sided TI, both the lower and upper bounds have uncertainties associated with the ML estimates, the two endpoints are not simple confidence limits of certain distribution quantiles. We need a TI algorithm to control the randomness in the two endpoints simultaneously. Krishnamoorthy and Xie (2011) developed algorithms to calculate two-sided TI for symmetric distributions in the location-scale family using the Monte-Carlo simulations based on the pivotal properties of ML estimators and the symmetric TI factors. When data are Type I censored, they also provided adjusted TI factors. While, there is no study about computing two-sided TI for the non-symmetric distributions in the location-scale family. Thus, we develop a general method that can be applied for all the distributions in the location-scale family to calculate the exact two-sided TI when data are complete or Type II censored. Also an adjusted procedure is developed for Type I censored data. These algorithms can be applied on log transformed data from a log-location-scale family based on the one-to-one relationship between a location-scale distribution and the corresponding log-location-scale distribution. Using our methods, for example, when a sample of possibly censored data follow the Weibull distribution, TIs can be obtained to estimate a proportion of the distribution.

1.2.4 Quantifying Uncertainties in Predictions

When predicting based on a random sample, it is important to quantify the variability in the prediction. For default prediction, we use the time-to-event data of 3,271 companies, covariate process data of 1,147 companies and two macroeconomic variables, all spanning a period of 228 months. Thus the prediction contains variabilities from the ML estimates of time-to-event model, ML estimates of the covariate model and the future covariate process. However, point predictions of the default probability and number of defaults only take account of the randomness in the future covariate process. It is therefore desirable to develop

a procedure that can quantify the uncertainties in predictions resulting from all the three sources. In the default prediction project in Chapter 2, we compute PIs incorporating all the uncertainties stated above for the individual default probabilities and the number of defaults. The PIs are computed by bootstrapping the covariate processes conditional on the observations in the first month.

To calculate the two-sided TI based on random samples from a distribution in the (log)-location-scale family, the challenge is also to quantify variabilities in the ML estimates. Because the whole procedure does not have closed forms, we use the Monte-Carlo simulations to estimate the uncertainties and search for the factors of TI that yield the desired confidence level. To incorporate the variabilities in the ML estimates, factors generally have larger magnitude than the corresponding distribution quantiles.

1.3 Outline of this Dissertation

Following the introduction, Chapter 2 is based on Yuan et al. (2015), in which we present a competing risks model for exits due to default and other reasons incorporating covariate information. The covariate processes without quarterly trend are modeled by Markovian time series with a DFM representation for the error vector. We predict the multiperiod default probabilities for individual companies and number of defaults in the market based on the simulated future covariate processes. To quantify the uncertainties from multiple sources, we also propose procedures for computing PIs for both individual default probabilities and the number of defaults in the sectors considered. Chapter 3 is based on Yuan et al. (2016), in which we develop a general algorithm to calculate the two-sided TI for complete and right-censored data from the (log)-location-scale family of distributions. To validate the asymptotic accuracy of the TI under Type I censored data, we conduct a simulation study with different settings of the expected number of exact observations. We also derive the equal-tailed constraint to obtain the unique TI. Chapter 4 gives some general conclusions and plans for future work.

Bibliography

J. Bai and P. Wang. Identification and estimation of dynamic factor models. Department of Economics Discussion Papers. Columbia University, New York. Available at http://academiccommons.columbia.edu/catalog/ac%3A146472, 2012.

M. Banbura and M. Modugno. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journel of Applied Econometrics*, 29:133–160, 2012.

J. Duan, J. Sun, and T. Wang. Multiperiod corporate default prediction – a forward intensity approach. *Journal of Econometrics*, 170:191–209, 2012.

J. C. Duan. Clustered defaults. *National University of Singapore working paper, available at Social Science Research Network 1511397*, 2010.

D. Duffie and N. Gârleanu. Risk and valuation of collateralized debt obligations. *Financial Analysts Journal*, 57:41–59, 2001.

D. Duffie, J. Pan, and K. Singleton. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68:1343–1376, 2000.

D. Duffie, L. Saita, and K. Wang. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83:635–665, 2007.

D. Duffie, A. Eckner, G. Horel, and L. Saita. Frailty correlated default. *The Journal of Finance*, 64:2089–2123, 2009.

K. Giesecke, F. A. Longstaff, S. Schaefer, and I. A. Strebulaev. Macroeconomic effects of corporate default crisis: A long-term perspective. *Journal of Financial Economics*, 111: 297–310, 2013.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(Series D):35–45, 1960.

J. P. Klein and M. L. Moeschberger. *Survival Analysis - Techniques for Censored and Truncated Data*. Springer Verlag, New York, 2003. ISBN 978-0-387-95399-1.

K. Krishnamoorthy and T. Mathew. *Statistical Tolerance Regions - Theory, Applications, and Computation*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2009. ISBN 978-0-470-38026-0.

K. Krishnamoorthy and F. Xie. Tolerance intervals for symmetric location-scale families based on uncensored or censored samples. *Journal of Statistical Planning and Inference*, 141:1170–1182, 2011.

J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, 2003. ISBN 0-471-37215-3.

W. Q. Meeker and L. A. Escobar. *Statistical Methods for Reliability Data*. John Wiley & Sons, 1998. ISBN 0-471-14328-6.

A. Mortensen. Semi-analytical valuation of basket credit derivatives in intensity-based models. *Journal of Derivatives*, 13:8–26, 2006.

X. Peng and S. Kou. Default clustering and valuation of collateralized debt obligations. unpublished manuscript, 2009.

P. S. Rosenberg. Hazard function estimation using B-splines. *Biometrics*, 51:874–887, 1995.

J. H. Stock and M. W. Watson. Dynamic factor models. *Oxford Handbook of Economic Forecasting*, 1:35–59, 2011.

Y. Xie, Y. Hong, W. Q. Meeker, and L. A. Escobar. Simultaneous prediction intervals for the (log)-location-scale family of distributions. *Statistics Preprints*, 2014. Paper 128.

M. Yuan, C. Tang, Y. Hong, and J. Yang. Disentangling and assessing uncertainties in multiperiod corporate default risk predictions. Submitted, 2015.

M. Yuan, Y. Hong, W. Q. Meeker, and L. A. Escobar. Two-sided tolerance intervals for the (log)-location-scale family of distributions. *Quality Technology and Quantitative Management*, 2016. In press.

Chapter 2 Disentangling and Assessing Uncertainties in Multi-period Corporate Default Risk Predictions

Abstract

Measuring credit risks for individual companies, industrial segments, and market systems is fundamentally and broadly important in economics and finance. For such a purpose, various quantitative methods have been developed to predictively assess the probabilities of companies going default in future. However, as a more difficult and crucial problem, evaluating the uncertainties associated with the default predictions remains little explored. In this chapter, for the first time in the scenario of default predictions, we develop a procedure for quantifying the level of associated uncertainties by carefully disentangling multiple contributing sources. Our framework effectively incorporates broad information from historical default data, 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 development of the framework overcomes major challenges in this tremendously large-scale statistical inference problem and makes it practically feasible by using parsimonious models, innovative methods, and modern computational facilities. By appropriately predicting the marketwide total number of defaults and assessing the associated uncertainties, our method can effectively evaluate the aggregated market credit risk level. Upon analyzing a US market data set with our method, we demonstrate that the level of uncertainties associated with default risk assessments is indeed substantial. More importantly and 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 benefiting practical applications

including improving the accuracy of default risk assessments.

Key Words: Competing Risks; Default Probability; Dynamic Factor Model; EM Algorithm; High-Dimensional Time Series; Prediction Interval.

2.1 Introduction

Default risk prediction 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 would predict financial health of investments under consideration and screen out undesirable investments, and determination of credit ratings by rating agencies (e.g. Altman 1968; Duffie et al. 2007).

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.

While various quantitative procedures have been developed for predicting default probabilities in the literature, little progress has been made to quantify the intrinsic uncertainties in default predictions. The main reason behind the lack of such an investigation is that the task is too challenging due to the extremely large-scale of the problem with multiple contributing sources to the uncertainties associated with predictions, which include those due to the future default mechanism, the future dynamics of the targets at default risk, and model parameter estimations. In Section 2.2, we carefully elaborate on those three sources

of uncertainties contributing to the default risk predictions.

The objective of our study is to develop a procedure obtaining prediction intervals (PIs) for quantitatively assessing the level of uncertainties associated with default risk prediction, taking the three aforementioned sources of uncertainties into account. To the best of our knowledge, no such or similar measure on default prediction uncertainties has been considered in the literature. Though assessing uncertainties associated with predictions has been studied in the literature especially in areas such as reliability (e.g. Hong et al. 2009), existing methods do not apply in the default predictions due to the unique challenging practical aspects of the problem, particularly related to developing PIs with unprecedentedly large-scale of the problem, as detailed in Sections 2.2 and 2.3. Technically speaking, all companies on the market at a given time are at risk of going default in the future, and each of them needs an assessment of its predicted default probability as well as the associated level of uncertainties.

Synthetically speaking, one needs stochastic modeling device(s) with multiple levels of considerations for fulfilling the objective of future default predictions. The first level is for the stochastic nature of the future default mechanism, upon given the features of the company and the status of the macroeconomic environment, which are collectively called covariates in our study hereinafter. Because the future covariates themselves are not deterministic, a second level consideration is required to incorporate the stochastic nature of the covariate processes; see, for example, the doubly stochastic framework of Duffie et al. (2007), and the forward intensity approach of Duan et al. (2012). Clearly, both levels of stochastics are contributing to the uncertain nature of the default predictions.

Our framework is designed to incorporate the aforementioned two levels of stochastic features in appropriately constructing predictions intervals. Given the complexity of the modeling devices, generally applicable explicit forms do not exist for constructing valid PIs. Thus, our framework resorts to resampling procedures designed based on 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, we remark drastically practical challenges from 1) complicated

covariates structure with large number of parameters to be estimated, 2) the large-scale of the data sets, and 3) the fact that observations for the subjects are highly un-balanced in the sense that some companies are on the market for many years while some are there for only a few months. For an example, the data set in our analysis in Section 2.4 for the US market over the period 1990 – 2009 contains more than 10,000 companies, and the number of monthly observations of the covariates exceed 1,000,000. Thousands of parameters are involved by using parametric models for the default mechanism and the covariate processes. Among all companies in the data set, missing data are overwhelming and the time horizons for those observations are highly heterogeneous among companies. Our Section 2.3 provides details on our dedicatedly developed framework for uncertainties assessments with PIs for both point predictions for the individual default probabilities and total number of defaults, overcoming those challenges by using parsimonious models, innovative computationally efficient method, and powerful computational facilities.

The proposed framework for measuring default prediction uncertainty 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 uncertainty dramatically improves our understanding and knowledge especially for model diagnosis and statistical inference on default probability prediction. Furthermore, in contrast with recent few studies which also use the naive measure of standard deviations of default probability prediction, a distinguished feature of our measure is to adequately allow for not only the multiple sources of uncertainties but also the asymmetric nature of default probability prediction uncertainty so that the lower bound of default probability prediction would not go below zero (which is obviously not sensible). Through appropriate quantification of default prediction uncertainties, we can produce PIs around the point predictions on the future default probability, which is key to sound statistical inference on the default probability prediction. For example, to assess how well their model of default prediction performs, Campbell et al. (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 PIs are taken into consideration. Also, in addition to the in-sample comparison, the out-of-sample default probability prediction is typically used in this line of the literature. The availability of default probability prediction uncertainty would easily enable us to further conduct the forecasting evaluation test, e.g., 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 2.4, facilitated by the PIs, we are able to show that the out-of-sample aggregated predictions for the total number of defaults work reasonably 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, which is sensible both theoretically and practically. As an illustration, we demonstrate by our data analysis in Section 2.4 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, more interestingly, we found that by incorporating the width of the prediction interval (PI) 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 PI 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 2.4 that our procedure is capable of equipping PIs with aggregated predicted total number of defaults, a feature that can benefit various studies of systemic risks. Clearly, additional insights can be added to these issues where uncertainties associated default probability predictions should be informative.

2.1.1 Related Literature

For recent studies on the topics of default predictions,

Duffie et al. (2007) modeled default and other types of exit for each individual company as independent Poisson processes conditional on the covariate processes. They used a proportional hazard model and built a parametric time series model on the selected firm-specific and market-related covariates. They also evaluated their model performance regarding corporate default prediction for multiple periods using real data. Duan et al. (2012) developed a forward intensity procedure to predict corporate default probabilities for different prediction horizons without modeling the covariate process. They modeled the default intensity and exit intensity as functions of possibly different sets of covariates with time-varying coefficients. Duffie et al. (2009) estimated the distribution of losses for debt portfolios of US companies by modeling the default intensity that incorporates a common dynamic latent covariate and firm-specific latent covariates. Bai and Wang (2012) proposed a set of identification restrictions for a dynamic factor model with factors modeled by a VAR(h) process. To allow for the prediction of concurrent defaults under common shocks, Duan (2010) developed a hierarchical intensity structure which models the default risks at common, group and individual levels. While other types of exit is modeled as an independent Poisson process.

Peng and Kou (2009) proposed a prediction method for the individual conditional default probabilities given dynamics of the market factors, which were modeled by Pólya process to produce strong default clustering effect during financial crises. For quantification of prediction uncertainties based on datasets with complied structures, Hong et al. (2009) developed a prediction procedure and its associated calibration method, to account for uncertainties in future random quantities and parameter estimates. Hong and Meeker (2010) described a model for using the dynamic covariate information for failure probability predictions in a reliability setting.

2.1.2 Overview

The rest of this chapter is organized as follows. In Section 2.2, we disentangle the sources of uncertainties contributing to the default predictions. We present in Section 2.3 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 2.4 comprehensively analyzes a large-scale US market data by developing default probability predictions and quantitatively assessing their associated level of uncertainties. Section 2.5 concludes this chapter and draws the picture for future research.

2.2 Sources of Uncertainties in Default Risk Predictions

2.2.1 Stochastic Time-to-Event

The foremost source of uncertainties in default risk predictions roots in the nature of the key objective of interest – stochastic occurrence of future defaults. That is, evaluating the corporate default risk concerns the unknown future event of a company going default or not, attributing the stochastic nature requiring characterizing the default mechanism. Investigating and modeling based on historical records of corporate default data can be broadly classified into the so-called time-to-event data analyses, which are intensively studied

and documented in areas including reliability in engineering studies, and survival analysis in biostatistics (see, for example, Meeker and Escobar 1998, and Kalbfleisch and Prentice 2002). Nevertheless, unique practical features associated with corporate default data bring up new interests and challenges. For example, due to broad marketwide interests on default risks of so many companies, the collected data set is inevitably huge in its size and contains tremendous volume of broad information. Moreover, defaults among companies are relatively rare events so that the marginal default probability among companies is low, whose prediction is known to associate with a higher level of uncertainties. While at the same time, a high level of precision is desirable when predicting the corporate default risk.

To incorporate the stochastic time-to-event in default risks assessments, we consider the framework of stochastic event time. That is, the time when a company defaults in the future is a random variable. Meanwhile, to accommodate the fact that a company may exit the market before going default due to reasons other than bankruptcy, e.g. being acquired by another company, the so-called competing risks are required in modeling the time-to-exit of the companies. Generally speaking, suppose there are K types of events, competing with each other so that not all of their occurrences can be practically observed, even when they are associated with the same company. Let T_k be the time to the event of type k. Following the convention of time-to-event data analysis (e.g., Kalbfleisch and Prentice 2002), the event intensity function $\lambda_k(t)$ $(\lambda_k(t) \geq 0, t \geq 0)$ for the kth type of event is defined as $S_k(t) = \exp\left[-\int_0^t \lambda_k(u)du\right]$, where $S_k(t) = P(T_k > t)$ is the survival probability, i.e. the probability that the kth event happens after time t. Parametric intensity function and its estimation will be discussed in Section 2.2.3 after the introduction of stochastic covariates.

2.2.2 Stochastic Covariates in Default Predictions

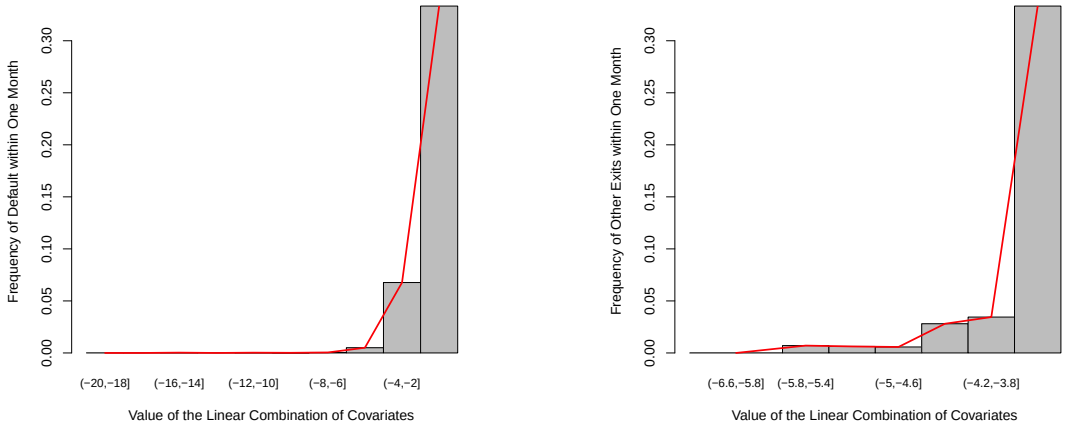
Additional to observing that the intensity $\lambda_k(t)$ is a function of time, it is natural to expect that features including the financial healthiness, profitability, growing perspective, etc are affecting the future default occurrences. Meanwhile, the macroeconomic conditions also

have impact on the future defaults. From the predictive perspective, the company specific features and the macroeconomic conditions are also stochastic, a source that will inevitably contribute 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, where the latter is clearly more challenging and is a main concern in our work.

Let us first describe how to incorporate the dynamic features as explanatory covariate information in the intensity function, and modeling and estimation for the stochastic covariate will be discussed in Section 2.2.4. We denote a random vector $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})^{\mathrm{T}}$ indexed by time t for the time series containing features, where p is the total number of firm-specific and macroeconomic covariates for a company. The observed covariate process is then denoted by $\boldsymbol{x}(t_1, t_2) = \{\boldsymbol{x}_s : t_1 < s \leq t_2\}$, which records available covariate information from time t_1 to time 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 given the covariate data. The total intensity of events (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 the assuming conditional independence between competing risk events. We also define the cumulative intensity by t from the time origin for 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 cumulative intensity of exit for a company up to time t is $\Lambda[t; \boldsymbol{x}(0, t)] = \sum_{k=1}^{K} \Lambda_k[t; \boldsymbol{x}(0, t)]$.

2.2.3 Parametric Intensity Function and Its Estimation

For practical applications, parametric forms of the intensity function $\lambda_k(\cdot)$ are often imposed for effectively analyzing time-to-event data with meaningful practical interpretations. In our work, we consider the parametric intensity function of event type k at time t with the



(a) Default (b) Other Exits

Figure 2.1: Estimated frequency of events versus the linear combination values to check the functional form of intensities.

convenient exponential additive form as

$$\lambda_k(t; \boldsymbol{x}_t) = \exp\left(\beta_{k0} + \beta_{k1} x_{1t} + \cdots + \beta_{kp} x_{pt}\right). \tag{2.1}$$

To check the model fitting of both intensity functions as exponential of the linear predictor, we plotted the actual frequency of defaults within one month on the y-axis for each value of the linear combination of covariates. The exponential shape of the line in Figure 2.1a validates (2.1) as the intensity function for default. Specifically, we use the estimated coefficient values in the default intensity based on all the available data to calculate the linear combination values for each company at risk in each month. The time period included are year 2005 - 2008. Because the linear combination values are continuous and default is a rare event, we aggregate the number of defaults in each interval of the linear combination values. The ten intervals have the same width on the range of the linear combination values. Similarly, we obtained Figure 2.1b and validated the exponential form of the intensity function for other exits.

With $\boldsymbol{x}_t$ in the event intensities modeled by some stochastic process, the framework is

referred as doubly stochastic in Duffie et al. (2007) for default probability predictions. In practice, the parameter $\boldsymbol{\beta} = (\beta_{10}, \ldots, \beta_{1p}, \ldots, \beta_{K0}, \ldots, \beta_{Kp})^{\mathrm{T}}$ is unknown and needs to be estimated using historic corporate default data. Therefore, uncertainties associated with the parameter estimation also 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 specific company, the time-to-event data are denoted by $\{t_i, \boldsymbol{\delta}_i, \boldsymbol{x}_i(0, t_i)\}$ for $i = 1, \ldots, n$, 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 studying the historical corporate default data, we follow the approach in existing studies in Duffie et al. (2007) and consider $K = 2$ types of events, i.e., a company defaults ($k = 1$) or exits the market due to other reasons ($k = 2$).

We note that the cumulative distribution function (cdf) of time to event T for a company, given its covariate history $\boldsymbol{x}(0, t)$, is $F_T(t) = P(T \leq t) = 1 - e^{-\int_0^t \sum_{k=1}^K \lambda_k(s; \boldsymbol{x}_s) ds} = 1 - e^{-\Lambda[t; \boldsymbol{x}(0,t)]}$. 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 $\Delta_k, k = 1, \ldots, K$ be the event indicators. That is, $\Delta_k = 1, \Delta_l = 0, l \neq k$ if the event that happened is of type k and $\Delta_l = 0, l = 1, \ldots, K$ if no event occurred by the latest observation time (denoted by τ) in the data set. Due to K types of competing risks, the observed time-to-event of a company is therefore $T = \min(T_1, \ldots, T_K)$. The sub-distribution function of T_k, which gives

the fraction failing due to event of type k, is

$$\begin{aligned} F_k(t) =& \Pr(T \le t, \Delta_k = 1) = \Pr(T_k \le t, T_l > T_k; \text{ for all } l \neq k) \\ &= \int_0^t f_{T_k}(t_k) \prod_{l \neq k} [1 - F_{T_l}(t_k)] \, dt_k = \int_0^t \lambda_k(s; \boldsymbol{x}_s) e^{-\Lambda[s; \boldsymbol{x}(0,s)]} ds. \end{aligned}$$

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 $\boldsymbol{x}_i(0,t_i)$ $(i = 1, \cdots, n)$ is then given by

$$L_T(\boldsymbol{\beta}|\text{DATA}) = \prod_{i=1}^{n} \left(\left(\prod_{k=1}^{K} \left\{ \lambda_k(t_i; \boldsymbol{x}_{i,t_i}) e^{-\Lambda[t_i; \boldsymbol{x}_i(0,t_i)]} \right\}^{\delta_{ki}} \right) \times \left\{ e^{-\Lambda[t_i; \boldsymbol{x}_i(0,t_i)]} \right\}^{\prod_{k=1}^{K}(1-\delta_{ki})} \right), \tag{2.2}$$

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 survives to time t. The parameters $\boldsymbol{\beta}$ are then estimated by maximizing the joint likelihood of the event times in (2.2).

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 reasonably approximated by $\sum_{t=1}^{t_i} \lambda_k(t, \boldsymbol{x}_{i,t})$. One can also see such approximations in Duffie et al. (2007) and Duan et al. (2012).

To remark, the level of uncertainties associated with the parameter estimation can be routinely quantified using the standard errors from the inverse of the local information matrix. In the literature, this type of standard errors are usually reported as a measure of level of uncertainties. Though it is one step forward for the ultimate goal of quantifying the uncertainties in default predictions, incorporating only uncertainties in the parameter estimations is not adequate for assessing the level of uncertainties in the default predictions. The clear 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 feature

processes as discussed in Section 2.2.2.

2.2.4 Parametric Stochastic Covariate Process and Its Estimation

For predicting future default probability, one needs to appropriately incorporate the aforementioned stochastic covariate process. Given the scale of the number of companies of interest in credit risk assessment, highly parsimonious model for the covariate process is necessary. Upon specifying the parametric model for the covariate process, the parameter estimation procedure together with the dynamic features introduced by the model are contributing in the uncertainties associated with the default prediction.

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 nationwide market, are used to model the default risks. Following existing studies, e.g., Duffie et al. (2007) and Duan et al. (2012), we narrow our scope to 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. 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 the only key variable, its distance to default. In our studies, we use the distance to default calculated by the method proposed in Duan et al. (2012). Though Duan et al. (2012) found that individual company's trailing one-year stock return was insignificant given other variables, we included it because it is highly significant in both the intensity functions of default and other types of exits given the other three variables using our data, please see Table 2.1. 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), as Duffie et al. (2007) and Duan et al. (2012). Hence, we have $\mathbb{X}_t = (\mathbf{D}_t^{\mathrm{T}}, \mathbf{V}_t^{\mathrm{T}}, r_t, S_t)^{\mathrm{T}}$, $t = 1, \cdots, \tau$ 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 remove quarterly seasonal effect in the time series, we take a difference of order 3, and resulting in a new m-dimensional vector time series $\mathbf{X}_t$ ($t = 1, \ldots, \tau'$), where $\tau' = \tau - 3$.

Modeling the dynamic features 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 tremendous 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 highly un-balanced observations of the time series, i.e., the origin and end times of the components in $\mathbf{X}_t$ are different with possible missing data for some period of time. Furthermore, the behaviors of components in $\mathbf{X}_t$ are expected to inter-related 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 propose a highly parsimonious time series model for $\mathbf{X}_t$ with two key components specifically for default predictions: a) a mean-reverting 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), Lam and Yao (2011), 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 et al. (2007):

$$\mathbf{X}_t - \boldsymbol{\mu} = \boldsymbol{\Theta}(\mathbf{X}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t, \quad t = 2, \ldots, \tau'. \tag{2.3}$$

Model (2.3) is essentially 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 et al. (2007) as follows,

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

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. While 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 one month.

To further capture the serial and cross-sectional dependence between components in $\mathbf{X}$, we propose to apply the following dynamic factor model (DFM) in Stock and Watson (2002) for the innovation vector $\boldsymbol{\varepsilon}_t$:

$$\boldsymbol{\varepsilon}_t = \boldsymbol{\Lambda}\mathbf{F}_t + \mathbf{e}_t, \tag{2.4}$$

$$\mathbf{F}_t = \mathbf{A}\mathbf{F}_{t-1} + \boldsymbol{\eta}_t, \quad t = 2, \ldots, \tau', \tag{2.5}$$

where the latent factor $\mathbf{F}_t$ is a $s \times 1$ vector following an auto-regression process with order 1 (i.e.,VAR(1)). Here, $\boldsymbol{\eta}_t' s$ are assumed to be independently and identically distributed (iid)

normal random vectors from NOR($\mathbf{0}, \mathbf{Q}$), for some positive definite matrix $\mathbf{Q}$. Here $\mathbf{\Lambda}$ is a $m \times s$ matrix of factor loadings, and $\mathbf{A}$ is a $s \times s$ 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 $\mathbf{\Lambda}$ is highly parsimonious by the fact that the number of common factor s can be very small, which drastically reducing the number of parameters in the covariance matrix of $\boldsymbol{\varepsilon}_t$.

Comparing our covariate model with Duffie et al. (2007) can reveal the advantages of applying DFM on modeling the covariate processes. First we model the first-order time dependence in the innovation vectors $\boldsymbol{\varepsilon}_t$ by specifying a VAR(1) model for the factors, while they assume serial independence among the innovation vectors. Second, our model has a much more flexible variance and cross-sectional correlation structure than theirs. For example, in Duffie et al. (2007), the variances of all the innovation terms of the firm-specific and macroeconomic covariates cannot change across time and different companies; the cross-sectional correlation between the innovation terms of the firm-specific covariates for company i and j at time t stay the same across different i, j and t; the cross-sectional correlations between the innovation terms of firm-specific covariates and the one-year S&P 500 return at time t are the same across different companies and time points; the innovation terms of the treasury bill rates are independent of other innovations. While our model has the flexibility to fit all of the variances and cross-sectional correlations without any equality or independence constraints. Moreover, using two factors to fit the estimated innovation vectors $\widetilde{\boldsymbol{\varepsilon}}_t$ of the differenced covariate data up to the end of 2008, the DFM reduces mean residual sum of squares in fitting $\widetilde{\boldsymbol{\varepsilon}}_t$ by about 10% compared to the mean.

As in our data analysis, the number of factor is chosen as $s = 2$ 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 develop an expectation-maximization (EM) algorithm for estimating parameters in

dynamic factor model specified by (2.4) and (2.5), whose detail is given in the Supplementary Material. 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, computationally intensive matrices inversion in our EM algorithm only involves those of size $s \times s$, making it most computationally efficient and feasible for this large-scale default prediction problems.

2.3 Predictions and Uncertainties Assessments

2.3.1 Procedures for Future Default Predictions

Clearly, predicting future corporate default probabilities given 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 aggregately predict the total defaults for the overall market system and certain market sectors. Let us begin with describing the method for individual corporate default predictions, and then the method for aggregating default predictions. Under the time-to-event and covariate process models discussed earlier, the conditional default probability of company i within s time units in future after the last observation time τ is,

$$\begin{aligned}\rho_i(s;\boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{x}_i(\tau,\tau+s)|\boldsymbol{x}_{i,\tau}}\left\{\Pr[\tau < T_i \le \tau+s, \Delta_1 = 1 | T > \tau]\right\} \\ &= \mathrm{E}_{\boldsymbol{x}_i(\tau,\tau+s)|\boldsymbol{x}_{i,\tau}}\left\{\frac{\Pr[\tau < T_1 \le \tau+s, T_1 < T_l, T_1 > \tau; l \ne 1]}{\Pr[T > \tau]}\right\} \\ &= \mathrm{E}_{\boldsymbol{x}_i(\tau,\tau+s)|\boldsymbol{x}_{i,\tau}}\left\{\int_\tau^{\tau+s} \lambda_1(t;\boldsymbol{x}_{i,t}) \exp\left(-\left\{\Lambda[t;\boldsymbol{x}_i(0,t)] - \Lambda[\tau;\boldsymbol{x}_i(0,\tau)]\right\}\right) dt\right\},\end{aligned} \tag{2.6}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_T^{\mathrm{T}}, \boldsymbol{\theta}_{\mathbf{X}}^{\mathrm{T}})^{\mathrm{T}}$ contains the parameters of the covariate model $\boldsymbol{\theta}_{\mathbf{X}}$ and those of the time-to-event model $\boldsymbol{\theta}_T$. Here, the conditional expectation is taken with respect to the probability distribution of the covariate process given information at time τ. The $\lambda_1(\cdot)$ is the default intensity function, and the covariate processes $\boldsymbol{x}_i(t_1,t_2) = \{\boldsymbol{x}_{i,u} : t_1 < u \le t_2\}$

are for company i during the period of t_1 to t_2. Note that $\rho_i(s;\boldsymbol{\theta})$ is a conditional sub-distribution function rather than a cumulative distribution function because $\rho_i(\infty;\boldsymbol{\theta}) < 1$ due to the existence of other types of exits.

Because there are no simple analytical expressions for the expectation in (2.6), 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 estimated parameter $\widehat{\boldsymbol{\theta}}$.

Algorithm 1:

1. Simulate the differenced covariate processes $\mathbf{X}^*(\tau', \tau'+s)$ containing all the firm-specific and macroeconomic covariates from its conditional distribution $\mathbf{X}(\tau', \tau'+s)|\mathbf{X}_{\tau'}$ given the information up to τ', with $\widehat{\boldsymbol{\theta}}_{\mathbf{X}} = \{\widehat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Theta}}\}$ estimated from the observed covariate data.

2. Obtain the covariate processes, which are denoted by $\mathbb{X}^*(0, \tau+s)$ from the differenced data $\mathbf{X}^*(0, \tau'+s) = \{\mathbf{X}(0,\tau'), \mathbf{X}^*(\tau', \tau'+s)\}$ and the originally observed covariate values at the first three months $\mathbb{X}(0,3)$, where $\mathbf{X}(0,\tau')$ and $\mathbf{X}^*(\tau', \tau'+s)$ are differenced covariate values calculated respectively from the observed and simulated covariate processes.

3. For each company i, compute

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

4. Repeat steps 1-3 M times to obtain $\rho_i^{*m}(s;\widehat{\boldsymbol{\theta}}), m = 1, \cdots, M$.

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 fit VAR(1) model on $\hat{\mathbf{F}}_t$, and calculate the following for $s = 1, 2, \cdots$

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

where $\widehat{\mathbf{F}}_{\tau'+s}$ is foretasted from the fitted VAR(1) model for $\widehat{\mathbf{F}}_t$ and $\hat{\mathbf{e}}_{\tau'+s}$ can be drawn from NOR$(\mathbf{0}, \widehat{\mathbf{P}})$.

As for the point prediction for the aggregate 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 $RS(t)$ be the set collecting companies of interests that are at risk of default at time t. Clearly, $N(s) = \sum_{i \in RS(\tau)} I_i(s)$ and $I_i(s) \sim$ Bernoulli$[\rho_i(s; \boldsymbol{\theta})]$. The point prediction for $N(s)$ is $\widehat{N}(s) = \sum_{i \in RS(\tau)} \hat{\rho}_i(s; \widehat{\boldsymbol{\theta}})$.

Though both the point predictions for $\rho(s)$ and $N(s)$ are informative for measuring future default risks, they did 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.

2.3.2 Assessing Uncertainties at the Aggregate Level

To assess the uncertainties associated with the predicted number of defaults, a natural choice is to supply a PI denoted by $\left[\underset{\sim}{N},\ \widetilde{N}\right]$. A naive (plug-in) PI for this purpose is obtained by solving

$$F_N(\underset{\sim}{N}; \widehat{\boldsymbol{\theta}}) = \frac{\alpha}{2}, \quad \text{and} \quad F_N(\widetilde{N}; \widehat{\boldsymbol{\theta}}) = 1 - \frac{\alpha}{2}. \tag{2.7}$$

Here $F_N(n_k; \boldsymbol{\theta}), n_k = 0, 1, \cdots, n^*$ is the cdf of $N(s)$ where n^* is the number of companies in the $RS(\tau)$, $1 - \alpha$ is the desired coverage probability. Note that $N(s)$ is a sum of non-identically distributed Bernoulli random variables. An explicit form for $F_N(n_k; \boldsymbol{\theta})$ is

$$F_N(n_k; \boldsymbol{\theta}) = \frac{1}{n^* + 1} \sum_{l=0}^{n^*} \left\{ \frac{1 - \exp[-\boldsymbol{i}\omega l(n_k + 1)]}{1 - \exp(-\boldsymbol{i}\omega l)} \prod_{i \in RS} [1 - \rho_i(s; \boldsymbol{\theta}) + \rho_i(s; \boldsymbol{\theta}) \exp(\boldsymbol{i}\omega l)] \right\}, \tag{2.8}$$

where $\boldsymbol{i} = \sqrt{-1}$ and $\omega = 2\pi/(n^* + 1)$. The cdf in (2.8) is obtained from a discrete Fourier transform of the characteristic function of $N(s)$. See Hong (2013) for more details on the

derivation and an efficient implementation for computing $F_N(n_k;\boldsymbol{\theta})$. Alternatively, one can use approximation methods such as the ordinary normal approximation or normal approximation with second order correction (e.g., Volkova 1996).

The PI in (2.7) 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 property. 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 $\left[\underset{\sim}{N},\ \widetilde{N}\right]$, is obtained by

$$\underset{\sim}{N} = v_{\alpha/2} \quad \text{and} \quad \widetilde{N} = v_{1-\alpha/2}. \tag{2.9}$$

Here v_α is the α lower quantile of $N(s)$ in the distribution $F_N(n_k;\widehat{\boldsymbol{\theta}}^*)$, where both N and $\widehat{\boldsymbol{\theta}}^*$ are treated as random variables. The α quantile v_α in (2.9) can be approximated by simulations. In particular, v_α is approximated by the α sample quantile of $F_N(n_k;\widehat{\boldsymbol{\theta}}^{*b}), b = 1,\cdots,B$. Here we sample N^{*b} from $F_N(n_k;\widehat{\boldsymbol{\theta}}^{*b})$ given the ML estimates $\widehat{\boldsymbol{\theta}}^{*b} = (\widehat{\boldsymbol{\theta}}_T^{*b\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{X}}^{*b\mathrm{T}})^{\mathrm{T}}$. To obtain samples of $\widehat{\boldsymbol{\theta}}^{*b}$, $\widehat{\boldsymbol{\theta}}_T^{*b}$ was simulated from $\mathrm{NOR}(\widehat{\boldsymbol{\theta}}_T, \Sigma_{\widehat{\boldsymbol{\theta}}_T})$ and $\widehat{\boldsymbol{\theta}}_{\mathbf{X}}^{*b}$ was estimated from the simulated covariate processes.

Specifically, 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 (2.3), (2.4) and (2.5), with parameters estimated from the real data before the prediction period. For each company, the differenced observations at the first month as well as the missing pattern of the differenced data are fixed. Note that each company entered the dataset and exited from the market at different times due to defaults or other reasons.

2. Estimate parameters in the covariate model $\widehat{\boldsymbol{\theta}}_{\mathbf{X}}^* = (\widehat{\boldsymbol{\Omega}}^*, \hat{\boldsymbol{\mu}}^*, \widehat{\boldsymbol{\Theta}}^*)$ based on the simulated processes through the EM algorithm in the Appendix.

3. Take a random sample of $\widehat{\boldsymbol{\theta}}_T^*$ from its asymptotic distribution $\text{NOR}(\widehat{\boldsymbol{\theta}}_T, \Sigma_{\widehat{\boldsymbol{\theta}}_T})$, where $\widehat{\boldsymbol{\theta}}_T$ and $\Sigma_{\widehat{\boldsymbol{\theta}}_T}$ are estimated from the observed data by the methods in Section 2.2.3.

4. With the originally observed covariate data in the first three months $\mathbb{X}(0,3)$, the differenced covariate data $\mathbf{X}(1,\tau')$ and the new parameter estimates $\widehat{\boldsymbol{\theta}}^* = (\widehat{\boldsymbol{\theta}}_T^{*\text{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{X}}^{*\text{T}})^{\text{T}}$, **Algorithm 1** is implemented to predict the default probabilities $\rho_i^*(s; \widehat{\boldsymbol{\theta}}^*), i = 1, \cdots, n$.

5. Take a random sample $N^*(s)$ from its distribution (2.8) with parameter values $\widehat{\boldsymbol{\theta}}^*$.

6. Repeat steps 1 to 5 B times to obtain $N^{*b}(s), b = 1, \cdots, B$.

7. The $100(1-\alpha)\%$ calibrated PI for $N(s)$ is $\left\{N^{*[(\alpha/2)B]}(s),\ N^{*[(1-\alpha/2)B]}(s)\right\}$, where $N^{*[b]}(s)$ is the ordered version of $N^{*b}(s)$ 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, i.e., those from the stochastic default mechanism, the stochastic covariate process, and the parameter estimation procedures.

2.3.3 Assessing Uncertainties for Corporate Default Probabilities

By applying **Algorithm 1**, one may predict the multiperiod ahead default probabilities for individual cooperate for evaluating the future default risk. Clearly, all sources of uncertainties are contributing in the point default probability estimations. For example, as intuitive in the Monte-Carlo simulation procedure in **Algorithm 1**, 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 PIs based on historical data for at risk companies that incorporate all contributing sources of uncertainties.

The procedure for constructing the PI is based on a large-scale parametric bootstrap, similar as 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 parameters. 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 PI correspondingly. More specifically, we have the following algorithm.

Algorithm 3:

1. Simulate $\mathbf{X}^*(1, \tau')$ from the covariate model (2.3), (2.4) and (2.5) using the parameter estimates from real data before the prediction period. The differenced observations of each company at the first month and the missing pattern of the differenced processes are fixed.

2. Estimate parameters in the covariate model and obtain $\widehat{\boldsymbol{\theta}}^*_{\mathbf{X}} = (\widehat{\boldsymbol{\Omega}}^*, \hat{\boldsymbol{\mu}}^*, \widehat{\boldsymbol{\Theta}}^*)$ based on the simulated processes, through the procedures in the Appendix.

3. Take a random sample of the ML estimates of the time-to-event model $\widehat{\boldsymbol{\theta}}^*_T$ from $\text{NOR}(\widehat{\boldsymbol{\theta}}_T, \Sigma_{\widehat{\boldsymbol{\theta}}_T})$, where $\widehat{\boldsymbol{\theta}}_T$ and $\Sigma_{\widehat{\boldsymbol{\theta}}_T}$ are estimated from the observed data by the methods in Section 2.2.3.

4. With the originally observed covariate data in the first three months $\mathbb{X}(0, 3)$, the differenced covariate data $\mathbf{X}(1, \tau')$ and the new ML estimates $\widehat{\boldsymbol{\theta}}^* = (\widehat{\boldsymbol{\theta}}^{*\mathrm{T}}_T, \widehat{\boldsymbol{\theta}}^{*\mathrm{T}}_{\mathbf{X}})^{\mathrm{T}}$, **Algorithm 1** is implemented to predict the default probabilities $\rho^*_i(s; \widehat{\boldsymbol{\theta}}^*), i = 1, \cdots, n$.

5. Repeat steps 1 to 4 B times to obtain $\rho^{*b}_i(s), b = 1, \cdots, B$.

6. The $100(1-\alpha)\%$ PI of default probability for the ith company at s time units after the Last observation time τ is $\left\{\rho_i^{*[(\alpha/2)B]}(s),\ \rho_i^{*[(1-\alpha/2)B]}(s)\right\}$, where $\rho_i^{*[b]}(s)$ is the order statistics of $\rho_i^{*b}(s)$ and $[\,\cdot\,]$ is the round function.

2.4 US Cooperate Default Data Analysis

To demonstrate our methods for assessing the level of uncertainties associated with default predictions, we analyze a sample of the data set studied in Duan et al. (2012). The data set contains defaults and other information of the United States (US) public firms 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 companies in the US and more than 1,000,000 firm-specific monthly observations. Our method is applied on a subset of data containing 3,271 firms from three industrial sectors (i.e., electronic product manufacturers, holding and investment offices, and business services) over the period from January 1990 to November 2009. To reflect the macroeconomic condition, we use monthly data of the trailing one-year S&P 500 return and the three-month Treasury bill rate as covariates as well. Besides the covariate data, we also have the time-to-event information such as the beginning and exit times, and exit type for each company. Specifically, among all the 3,271 companies from January 1990 up to November 2009, 164 defaulted and 2,049 exited due to other reasons, leaving 1,058 companies at risk at the end of November 2009.

In Section 2.4.1, we first apply the estimation procedures on the time-to-event and covariate data introduced earlier. To evaluate the one-year out-of-sample prediction, we only use the data before each prediction period. For example, to predict the default risks in 2009, data spanning January 1990 to December 2008 are used. We will report parameter estimates in the time-to-event model based on the original data, and estimates of the covariate model based on the differenced covariate data.

Table 2.1: ML estimates of parameters and their asymptotic standard errors based on data over January 1990 to December 2008.

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

2.4.1 Parameter Estimates for Time-to-Event Models

Parameter estimates in the time-to-event model are obtained by maximizing the log likelihood as introduced in Section 2.2.3. For the ML estimates, standard errors are calculated by the inverse of the local information matrix. The point estimation as well as 95% normal approximated confidence intervals are displayed in Table 2.1. Here β_{10} and β_{20} stand for the intercept terms in the intensity functions of default and other exits, respectively. For the coefficients of firm-specific and macroeconomic covariates, $\beta_{11}, \cdots, \beta_{14}$ are the coefficients of companies' distance to default, trailing one-year stock log return, the three-month Treasury bill rate and the trailing one-year return on the S&P 500 index in the default intensity function. $\beta_{21}, \cdots, \beta_{24}$ are the coefficients of corresponding terms in the intensity function for other exits.

From Table 2.1, larger distance to default indicates significantly lower risk for default but higher chance for exit due to other reasons. The trailing one-year stock return is an important indicator for a company's profitability, such that higher stock return implies lower risk for both default and other forms of exits. Increasing three-month Treasury bill rate manifests lower risk for default and other exits. Rising trailing one-year return on the S&P 500 index leads to higher risk for default, while its impact on other exits is not significant based on our results.

2.4.2 Parameter Estimates for Covariate Model

To estimate parameters $\widehat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Theta}}$ in the covariate model, we only include companies in the risk set at the beginning of the prediction period. To predict for defaults in 2009, the differenced covariate data of the 1,147 companies at risk are used to fit the covariate model.

The covariance matrix $\mathbf{P}$ of $\mathbf{e}_t$ is a large diagonal matrix with 2,296 rows. The estimated variances of elements in $\mathbf{e}_t$ range from 0.0013 to 21.5736. The estimated matrix $\mathbf{A}$ of autoregressive coefficients in the VAR(1) model (2.5) for the factors $\mathbf{F}_t$ is equal to:

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

which makes sure that (2.5) is a stationary process. The mean reverting parameters $\hat{\boldsymbol{\kappa}} = (0.6377, 0.6355, 0.8921, 0.6355, -0.0071)^{\mathrm{T}}$. The factor loading matrix $\boldsymbol{\Lambda}$ of $\boldsymbol{\varepsilon}_t$ is a 2296×2 matrix and the estimates of its elements range between -0.3532 and 0.6068. The estimated mean vector $\hat{\boldsymbol{\mu}}$ of the differenced data $\mathbf{X}_t, t = 1, \cdots, \tau'$ has 2296 elements. Specifically, the mean vector of the differenced distances to default over 19 years range from -0.5925 to 0.1952. The mean values of the differenced one-year stock log return of the 1147 companies fall between -0.0501 and 0.1156. The mean of the differenced three-month Treasury bill rate and the S&P 500 trailing one-year return are -0.009 and 0.0011, respectively.

2.4.3 Aggregate Default Predictions and Uncertainties

For predicting the aggregated total number of defaults, we consider the one-year periods: 2006, 2007, 2008 and 2009. Each time the number of defaults during one year is predicted, including both the expected number based on parameters estimated from real data and the PI based on parameters estimated from parametrically bootstrapped processes. To check the performance of out-of-sample predictions, only data before each prediction period is used for parameter estimation.

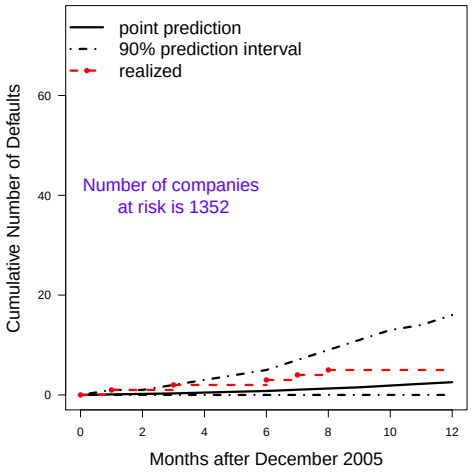
Specifically, Figure 2.2 shows the cumulative number of defaults during each one-year period, and the associated 90% two-sided PI for all the companies at risk at the beginning of the corresponding period. The PIs are calculated through **Algorithm 2** in Section 2.3.2. The solid step plot with dots shows the real number of defaults, and the solid straight line represents the predicted mean number of defaults.

For 2007 and 2008, the mean prediction agrees well with the actually realized cumulative number of events, which provides a validation of the proposed model. For 2006, the realized number of defaults falls between the mean prediction and the upper bounds of the PI (i.e., the 95% percentile), which shows the benefits of quantifying the uncertainties in the point prediction for the number of corporate defaults. In 2009, the mean predictions are good for the realized cumulative number of defaults in the first six months. However, it is very difficult to forecast the abrupt change in the trend of the realized number of defaults. Our predictions follow the previous trend and do not predict that there was no default during July to November 2009. Such an observation facilitated with PIs suggests a change in the occurrence of the default, and further investigation on the parametric modeling may be needed to incorporate the time-varying effect.

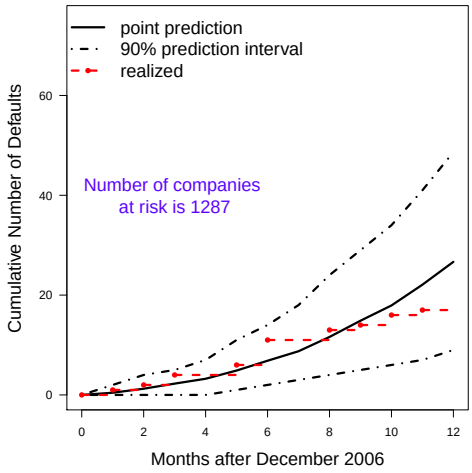
2.4.4 Individual Risk Prediction and Uncertainties

In this section, we will present the performance of our point predictions and PIs for the default probabilities of individual companies. Besides, we find that the uncertainty within the point prediction, quantified by the width of PI, also can be highly informative in analyzing default predictions for the individual corporate default risks. And moreover, the width of PI can interact with the point prediction to shed a light on the default risk.

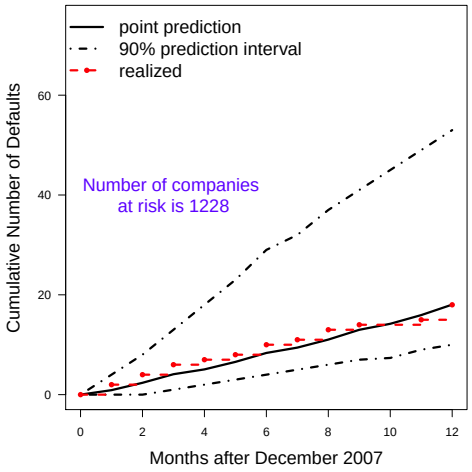
In this section, we present the out-of-sample predictions for the individual corporate default risks over each one-year period (2006, 2007, 2008 and 2009). The solid curve shows the predicted default probabilities using parameter estimates based on the real data before the prediction period, which can be viewed as the mean predictions for the probabilities



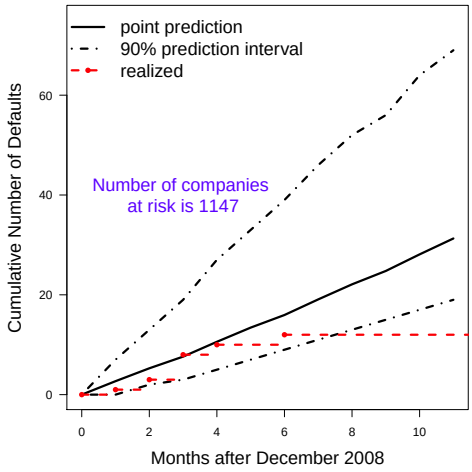
(a) Number of defaults during 2006



(b) Number of defaults during 2007



(c) Number of defaults during 2008



(d) Number of defaults during 2009

Figure 2.2: Cumulative number of defaults in the one-year periods and the associated PI for all the units at risk.

that a company at risk will default no later than a specific month. While the dashed lines represent the 90% PIs for the default probabilities. Individual PIs are calculated based on the **Algorithm 3** in Section 2.3.3.

The y-axes of each pair of plots in the same year in Figure 2.3 have the same scale. Plots on the left panel show predictions for companies that defaulted in the prediction period. For comparisons, the right panel presents companies that did not default in the same one-year prediction period. In the plots on the left side, the mean predictions at each month and the variabilities in the point predictions based on simulations are substantially higher than those in the corresponding plots on the right side.

Emerge Interactive Inc. was a technology company providing food-safety, individual-animal tracking and supply-management services and it defaulted in 2006. Let the mean prediction for the probability that default will occur by the last month of the prediction period represent the default risk. Then, Emerge Interactive Inc had the highest risk among the five companies that actually defaulted in 2006. And it ranked the 8th in the default risk among all the 1352 companies at risk. Lehman Brothers Holdings Inc was the fourth largest investment bank in the US and it defaulted in 2007. Lehman Brothers ranked the 14th and the 141th in terms of the default risk among the 17 companies that defaulted and among all the 1287 companies at risk, respectively. Bankunited Financial Corp was a savings and loan association that defaulted in 2008. In terms of default risk predictions, it had the 3rd highest risk among the 18 companies that defaulted, and ranked the 4th among all the 1228 companies at risk. TierOne Corporation was the holding company for TierOne Bank, it defaulted in 2009 with the 3rd and 9th highest default risk among the companies that defaulted in the same prediction period and among all the 1147 companies at risk, respectively.

The PIs are computed to quantify the uncertainties in the mean predictions. In 2006, the mean prediction for the cumulative default risk of Emerge Interactive Inc by December is 0.0347 and the associated 90% PI is (0.0087, 0.1319). For Microsoft, the mean prediction

is 0.00002 and the 90% PI is (0, 0.00005). During 2009, for TierOne Corporation, the mean prediction for the cumulative default risk by November is 0.4025 and the associated 90% PI is (0.1919, 0.6046). For Tellabs, the mean prediction is 0.0020 with 90% PI (0.0003, 0.0085).

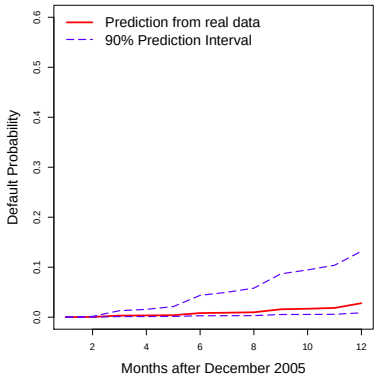
The out-of-sample prediction power curves in Section 2.4.5 will further evaluate the effectiveness of our model towards the individual default risk predictions considering the whole risk sets of each prediction period.

2.4.5 Prediction Performance: Power Curve

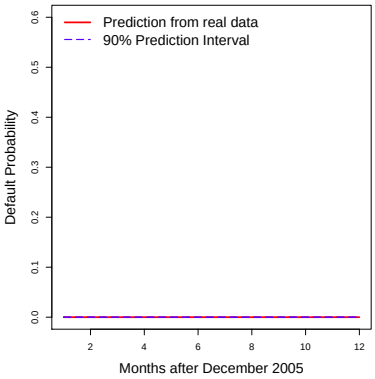
We now evaluate the out-of-sample prediction of the default risks for each company within one-year periods (i.e., 2006, 2007, 2008, and 2009) with the receiver operating characteristic (ROC) curve, also known as power curve as in other studies, e.g., Duffie et al. (2007). We found that the default risk within one year can be quantified in two ways: the predicted mean default probability or the amount of uncertainty within that point prediction. That is, greater variability with the point prediction for the default probability also reveals higher risk for default. Specifically, for the point prediction, we use the mean predicted probability that a specific company will default by the last month of the target period. And the uncertainty is quantified by the width of the PI for the cumulative default probability of the last month.

Figures 2.4a and 2.4b show the prediction power curves of ranking the companies by the point prediction and the width of PI, respectively. In the two figures, each step indicates an actual default. And the left-most point on each step represents the cumulative fraction of actual defaults and the corresponding percentile ranks of the predicted default risks that are needed to capture those defaults. From the areas under the curves, the width of PI has slightly worse performance than the point prediction in terms of the accuracy of ranks, although they provide different ranks for a specific company. By comparing the areas under the four curves, the predictions are more accurate for 2009 and 2008 than for 2007 and 2006.

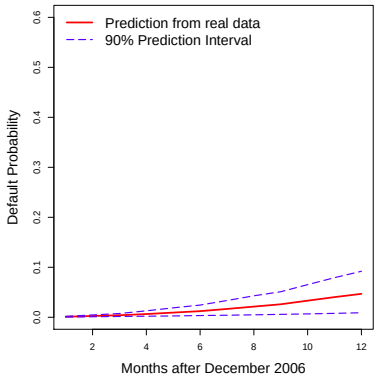
Let's look more closer at the story that the prediction power curves tell us, taking the curve drawn by the point predictions as an example. In 2006, five out of 1352 companies



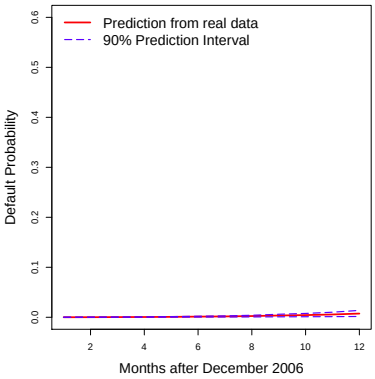
(a) Emerge Interactive Inc



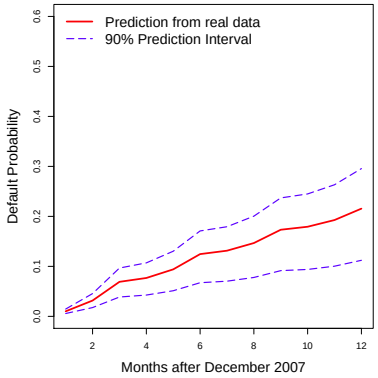
(b) Microsoft Corp



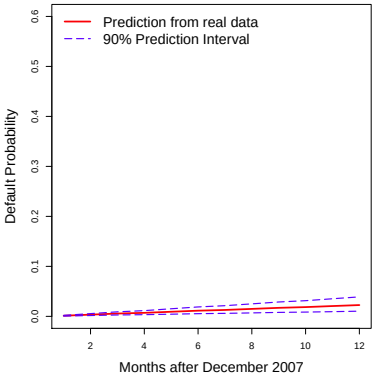
(c) Lehman Brothers



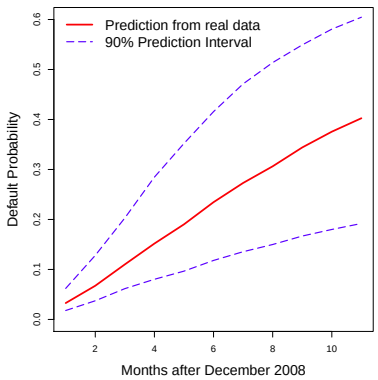
(d) City National Corp



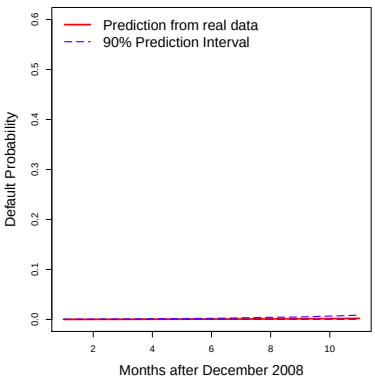
(e) Bankunited Financial



(f) Bank of America Corp

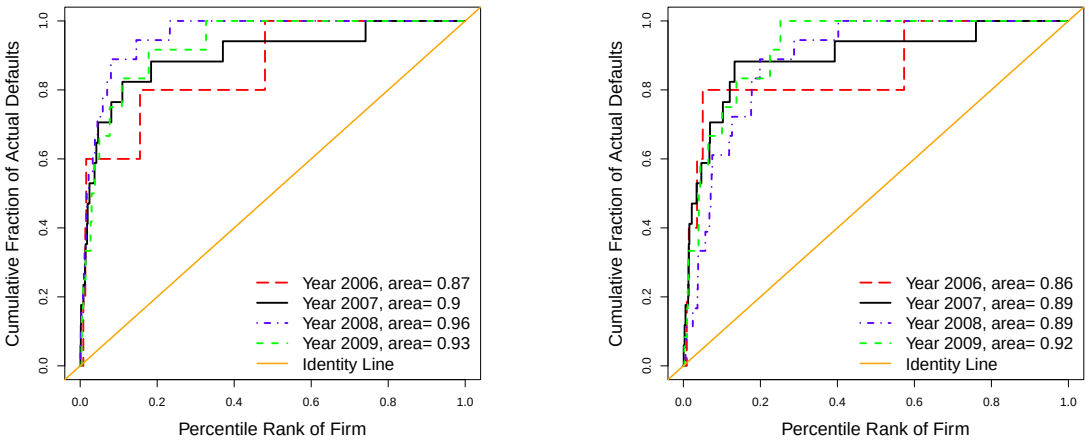


(g) TierOne Corp



(h) Tellabs Inc

Figure 2.3: Predictions for individual default probabilities and the associated 90% PI.



(a) Point prediction (b) Width of PI

Figure 2.4: Out of sample prediction power curves by the point prediction and width of PI.

defaulted. The first default is captured by the 0.59% highest default probability. In addition, 60% of the actual defaults can be predicted using the highest 1.41% default risks, and 80% actual defaults can be predicted within the highest 22.19% default risks. During the year 2007, totally 17 among the 1287 companies at risk actually defaulted. The first actual default ranks at the highest 0.08% in default risks. The default risks for 70.59% of the actual defaults are within the highest 4.66%. And 88.24% actual defaults can be captured by the highest 18.41% default risks. In 2008, 18 of the 1228 companies at risk defaulted, the one with the highest default risk ranks at 0.08%. 72.22% and 88.89% defaults occur on the companies within the highest 4.48% and 7.98% default probabilities, respectively. During 2009, 12 out of the 1147 companies at risk defaulted. The actual default with the highest predicted risk ranks at 0.35% among all the companies at risk. 66.67% and 91.67% of the defaults can be predicted using the highest 4.97% and 17.79% default risks, respectively.

2.4.6 Default Predictions and Associated Uncertainties

In Section 2.4.5, we saw both higher prediction for the mean cumulative default probability and its uncertainty imply higher default risk. Here we attempt to see how those two quantities work together to predict defaults.

Table 2.2: Summary of the Logistic regression model.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-6.1340	0.2898	-21.1691	0.0000
PI width	2.0742	3.3577	0.6177	0.5367
Point prediction	49.6683	6.9704	7.1256	0.0000
PI width × Point prediction	-99.8712	14.4081	-6.9316	0.0000

Specifically, we use the predicted mean default probability and the width of PI as in Section 2.4.5 as covariates, and whether each company defaulted in the prediction period as the response to fit logistic regression models. To have more defaults in the data set, we combine the data of all the companies at risk in each of the four prediction periods (2006, 2007, 2008 and 2009) together, resulting in 5014 observations in total. The result of logistic regression model is displayed in Table 2.2.

Table 2.2 confirms that higher predicted cumulative default probability within a year means an increase in the default risk. However, given the point prediction and the interaction between the two, main effect of the width of PI is not significant any more. Among the companies with predicted mean default probability lower than 0.05, only the point prediction is significantly associated with the default risk. Among companies with high predicted mean default probability, the PI width has a negative association with the default risk. For example, a logistic regression model for a subgroup (sample size = 189) with the point prediction higher than 0.08 shows that the higher predicted mean default probability is associated with higher default risk (p - value = 0.049), while wider PI is associated with lower default risk (p - value = 0.058). Interaction between the two variables is not significant any more (p - value = 0.479). Therefore, it is important to consider the amount of uncertainties along with the high default probabilities from predictions to make a thorough decision.

In addition, digging into the causes for high uncertainties will shed more light on how to properly use the uncertainty to make a decision. From a simple comparison of the time series plots of the distance to default before the prediction period between companies with the highest uncertainties and lowest uncertainties, we see (1) companies with high uncertainties

generally have low and wavy distance to default approaching the prediction period, while companies with low uncertainties usually have high, stable or straightly increasing distance to default near the prediction period; (2) companies with high uncertainties usually enter the market later (have more missing covariate data) than companies with low uncertainties; (3) companies with high uncertainties usually have decreasing distance to default toward 0 or negative values near the prediction period. (4) during 2007 and 2008, companies with relatively low uncertainties also have decreasing and unstable distance to default, but they usually have longer distance to default than the companies with high uncertainties.

Similarly, comparing the trailing one-year stock log return, companies with the highest uncertainties typically have (1) more frequently and drastically fluctuations, (2) decreasing trend near the prediction period. Besides, they have lower stock return and more missing values than the companies with low uncertainties.

2.5 Discussions and Future Research

We consider the challenging problem of assessing uncertainties associated default predictions by carefully disentangling and quantifying the contributing sources for the point predictions. For such a purpose, we first build a competing risks model for default and other types of exits with simple intensity functions, and then apply a highly parsimonious Markovian time series and dynamic factor model for the covariate process. For parameter estimations, we dedicatedly develop an EM algorithm for estimating the parameters in the dynamic factor models that can handle tremendously high-dimensional time series and highly un-balanced observations with substantial amount of missing data. Based on the time-to-event and covariate process models, we consider point predictions for individual default probabilities and the aggregated number of defaults in the market based on appropriately simulating future covariate processes. For assessing the level of associated uncertainties, we develop the calibrated PIs for the individual companies and the population by incorporating uncertainties in parameter estimation besides the future processes. An application of our methods on 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.

With limited access to a powerful modern computation facility (160 hours in parallel on 80 CPUs), we are able to accomplish the tasks for assessing the uncertainties associated with cooperate default predictions. We clearly demonstrate the feasibility for solving this long overdue important practical problem with the considerable size of the data and many challenging practical features. Assessing the uncertainties associated with cooperate default predictions is particularly meaningful with respect to a better understanding of the point predictions and additional dimension of insights on the default risks.

However, formal variable selection and model selection for the intensity functions need to be implemented to achieve the best predictive performance. More investigations for assessing the uncertainties associated with corporate default risk predictions are clearly desirable. For example, other methods dealing with the default mechanism can be considered, and practical features such that defaults may not occur independently, suggesting that frailty models could provide additional insights to the problem. Both theoretical and practical investigations are also needed for exploring the tremendously high-dimensional covariate process for large number of companies with the whole marketwide data.

Appendices

2.A Details of the EM Algorithm for Estimating Parameters in the Covariate Model

The mean vector of $\mathbf{X}_t, t = 1, \cdots, \tau'$ is denoted by $\boldsymbol{\mu}$. Let $\mathbf{M}$ be an $m \times \tau'$ matrix, with element 0 indicating missing in the differenced covariates and 1 otherwise. Let $\mathbf{W}_t$ be an $m \times m$ diagonal matrix whose ith diagonal component takes value 1 if the ith element of $\mathbf{X}_t$, denoted by X_{it}, is observed and 0 otherwise. Let $\mathbf{B}_t$ be a matrix obtained by removing those rows in $\mathbf{I}_m$, if those corresponding X_{it}'s are not observed. Note that $\mathbf{W}_t = \mathbf{B}_t^{\mathrm{T}}\mathbf{B}_t$. Here, $\mathbf{W}_t$

and $\mathbf{B}_t$ are defined to remove the missing values in the differenced covariate processes from the estimation procedure.

We now derive the ML estimators of the time series model (2.3) and EM algorithm for the DFM in (2.4) and (2.5) in explicit forms when substantial amount of missing data are present. We also address the identification problems in estimating the DFM by imposing necessary constraints (for details, one can refer to Bai and Wang 2012).

We derive the ML estimators of the time series model given parameter estimates of the DFM by directly maximizing the log likelihood. While parameter estimation for the DFM is not trivial because $\mathbf{F}_t$ is not observable and the ML estimators do not have closed forms.

Note that computing the ML estimates of the time series model and those of the DFM are not separate procedures. Because ML estimates of the former (mean vector $\hat{\boldsymbol{\mu}}$ and coefficient matrix $\widehat{\boldsymbol{\Theta}}$) depend on the estimates of the latter ($\widehat{\boldsymbol{\Omega}}$) through the likelihood function, and reversely ML estimates of the latter also depend on those of the former through the estimated set of residuals $\widetilde{\boldsymbol{\varepsilon}}_t$. Thus, we iteratively update ML estimates of the DFM and those of the time series model.

Before implementing the EM algorithm, the first step is to obtain the initial values $\boldsymbol{\mu}^{(0)}, \boldsymbol{\Theta}^{(0)}, \boldsymbol{\Sigma}^{(0)}$ and $\boldsymbol{\Omega}^{(0)} = \{\boldsymbol{\Lambda}^{(0)}, \mathbf{A}^{(0)}, \mathbf{P}^{(0)}\}$, which will be explained at the end of this appendix. Then we iterate between applying the EM algorithm to estimate parameters $\boldsymbol{\Omega} = \{\boldsymbol{\Lambda}, \mathbf{A}, \mathbf{P}\}$ in (2.4) and (2.5), and updating the ML estimates $\hat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Theta}}$ of model (2.3) until convergence. Let us introduce the derivation of the EM algorithm first.

Let $\boldsymbol{\theta}_{\mathbf{X}}$ denote all the parameters in the covariate models (2.3), (2.4) and (2.5). Particularly, in the $(j+1)^{\text{th}}$ E-step, the expectation of the joint log likelihood $\mathbb{E}_{\mathbf{F}}\left\{l\left[\boldsymbol{\Omega} \mid \mathbf{X}(0,\tau'), \widehat{\boldsymbol{\theta}}_{\mathbf{X}}^{(j)}\right]\right\}$ is computed conditional on all the differenced covariate data $\mathbf{X}(0,\tau')$ through $\widetilde{\boldsymbol{\varepsilon}}$, and given the parameter estimates $\widehat{\boldsymbol{\theta}}_{\mathbf{X}}^{(j)}$ from the j^{th} iteration of the estimation procedure. Here l stands for the log likelihood, $\widetilde{\boldsymbol{\varepsilon}} = (\widetilde{\boldsymbol{\varepsilon}}_2, \cdots, \widetilde{\boldsymbol{\varepsilon}}_{\tau'})$ and $\mathbf{F} = (\mathbf{F}_1, \cdots, \mathbf{F}_{\tau'})$. In the $(j+1)^{\text{th}}$ M-step, ML estimates $\widehat{\boldsymbol{\Omega}}^{(j+1)}$ are obtained by maximizing the conditional log likelihood function calculated in the $(j+1)^{\text{th}}$ E-step. Referring to Banbura and Modugno (2012), the EM procedure

is derived as the following steps given initial estimates $\boldsymbol{\Omega}^{(0)}$ and $\boldsymbol{\mu}^{(0)}, \boldsymbol{\Theta}^{(0)}, \boldsymbol{\Sigma}^{(0)}$.

For the conditional expectation step, Kalman filter and smoother procedures are applied to obtain explicit forms of the first and second order moments of the latent factors $\mathbf{F}$ given all the available covariate data. They will be used in the maximization step as available data to compute $\widehat{\boldsymbol{\Omega}}$.

Let $\widehat{\mathbf{A}}^{(j)}$ denote the estimated matrix $\mathbf{A}$ from the j^{th} iteration of EM algorithm, and similarly define the notation for other parameters. Also let $\mathbf{F}_{i|i'}$ and $\boldsymbol{\Xi}_{i|i'}$ denote the conditional expectation and covariance matrix of $\mathbf{F}_i$ given the differenced covariate data up to time i', respectively. In the Kalman Filter, all the conditional means and covariance matrices of factors $\mathbf{F}_t's$ given the differenced covariate data up to the same time point t are updated. That is, to incorporate information forward in time and obtain $\mathbf{F}_{t|t}$ and $\boldsymbol{\Xi}_{t|t}$, for $t = 2, \ldots, \tau'$. Given initial values $\mathbf{F}_{1|1} = \boldsymbol{\mu}_1 = \underline{0}$ and $\boldsymbol{\Xi}_{1|1} = \boldsymbol{\Xi}_1 = \mathbf{I}_2$, the Kalman filtering updates at the $(j+1)^{\text{th}}$ expectation step are as follows,

$$
\begin{aligned}
\mathbf{F}_{t|t-1}^{(j+1)} &= \widehat{\mathbf{A}}^{(j)} \mathbf{F}_{t-1|t-1}^{(j)}, \\
\boldsymbol{\Xi}_{t|t-1}^{(j+1)} &= \widehat{\mathbf{A}}^{(j)} \boldsymbol{\Xi}_{t-1|t-1}^{(j)} (\widehat{\mathbf{A}}^{(j)})^{\mathrm{T}} + \widehat{\mathbf{Q}}^{(j)}, \\
\mathbf{v}_t &= \mathbf{B}_t \widetilde{\boldsymbol{\varepsilon}}_t - \mathbf{B}_t \widehat{\boldsymbol{\Lambda}}^{(j)} \mathbf{F}_{t|t-1}^{(j+1)}, \\
\mathbf{U}_t &= \mathbf{B}_t \widehat{\boldsymbol{\Lambda}}^{(j)} \boldsymbol{\Xi}_{t|t-1}^{(j+1)} (\widehat{\boldsymbol{\Lambda}}^{(j)})^{\mathrm{T}} \mathbf{B}_t^{\mathrm{T}} + \mathbf{B}_t \hat{\mathbf{P}}^{(j)} \mathbf{B}_t^{\mathrm{T}}, \\
\mathbf{F}_{t|t}^{(j+1)} &= \mathbf{F}_{t|t-1}^{(j+1)} + \boldsymbol{\Xi}_{t|t-1}^{(j+1)} (\widehat{\boldsymbol{\Lambda}}^{(j)})^{\mathrm{T}} \mathbf{B}^{\mathrm{T}} \mathbf{U}_t^{-1} \mathbf{v}_t, \\
\boldsymbol{\Xi}_{t|t}^{(j+1)} &= \boldsymbol{\Xi}_{t|t-1}^{(j+1)} - \boldsymbol{\Xi}_{t|t-1}^{(j+1)} (\widehat{\boldsymbol{\Lambda}}^{(j)})^{\mathrm{T}} \mathbf{B}_t^{\mathrm{T}} \mathbf{U}_t^{-1} \mathbf{B}_t \widehat{\boldsymbol{\Lambda}}^{(j)} \boldsymbol{\Xi}_{t|t-1}^{(j+1)}.
\end{aligned}
$$

The inverse matrix $\mathbf{U}_t^{-1}$ takes the explicit form involving inverting a diagonal matrix and an

$s \times s$ matrix,

$$\begin{aligned}\mathbf{U}_t^{-1} =&(\mathbf{B}_t\widehat{\mathbf{P}}^{(j)}\mathbf{B}_t^{\mathrm{T}})^{-1} - (\mathbf{B}_t\widehat{\mathbf{P}}^{(j)}\mathbf{B}_t^{\mathrm{T}})^{-1}\mathbf{B}_t\widehat{\mathbf{\Lambda}}^{(j)}(\mathbf{\Xi}_{t|t-1}^{(j+1)})^{1/2} \\ &\left[\mathbf{I} + (\mathbf{\Xi}_{t|t-1}^{(j+1)})^{1/2}(\widehat{\mathbf{\Lambda}}^{(j)})^{\mathrm{T}}\mathbf{B}_t^{\mathrm{T}}(\mathbf{B}_t\widehat{\mathbf{P}}^{(j)}\mathbf{B}_t^{\mathrm{T}})^{-1}\mathbf{B}_t\widehat{\mathbf{\Lambda}}^{(j)}(\mathbf{\Xi}_{t|t-1}^{(j+1)})^{1/2}\right]^{-1}(\mathbf{\Xi}_{t|t-1}^{(j+1)})^{1/2} \\ &(\widehat{\mathbf{\Lambda}}^{(j)})^{\mathrm{T}}\mathbf{B}_t^{\mathrm{T}}(\mathbf{B}_t\widehat{\mathbf{P}}^{(j)}\mathbf{B}_t^{\mathrm{T}})^{-1}.\end{aligned}$$

The conditional expectations and covariance matrices given all the differenced data up to time τ' are obtained by Kalman smoothing and backward recursive calculation. Here $\mathbb{E}$ means the conditional expectation, cov means the conditional covariance matrix between two variables, and var means the covariance matrix of a single variable given all the observed covariate data. In particular,

$$\begin{aligned}\mathbb{E}^{(j+1)}(\mathbf{F}_{\tau'}) &= \mathbf{F}_{\tau'|\tau'}^{(j+1)}, \\ \mathrm{var}^{(j+1)}(\mathbf{F}_{\tau'}) &= \mathbf{\Xi}_{\tau'|\tau'}^{(j+1)}, \\ \mathrm{cov}^{(j+1)}(\mathbf{F}_{\tau'}, \mathbf{F}_{\tau'-1}) &= \left[\mathbf{I} - \mathbf{\Xi}_{\tau'|\tau'-1}^{(j+1)}(\widehat{\mathbf{\Lambda}}^{(j)})^{\mathrm{T}}\mathbf{B}_t^{\mathrm{T}}\mathbf{U}_t^{-1}\mathbf{B}_t\widehat{\mathbf{\Lambda}}^{(j)}\right]\widehat{\mathbf{A}}^{(j)}\mathbf{\Xi}_{\tau'-1|\tau'-1}^{(j+1)}.\end{aligned}$$

For $t = \tau', \ldots, 3$, calculate

$$\begin{aligned}\mathbf{J}_{t-1} &= \mathbf{\Xi}_{t-1|t-1}^{(j+1)}(\widehat{\mathbf{A}}^{(j)})^{\mathrm{T}}(\mathbf{\Xi}_{t|t-1}^{(j+1)})^{-1}, \\ \mathbb{E}^{(j+1)}(\mathbf{F}_{t-1}) &= \mathbf{F}_{t-1|t-1}^{(j+1)} + \mathbf{J}_{t-1}\left[\mathbb{E}^{(j+1)}(\mathbf{F}_t) - \mathbf{F}_{t|t-1}^{(j+1)}\right], \\ \mathrm{var}^{(j+1)}(\mathbf{F}_{t-1}) &= \mathbf{\Xi}_{t-1|t-1}^{(j+1)} + \mathbf{J}_{t-1}\left[\mathrm{var}^{(j+1)}(\mathbf{F}_t) - \mathbf{\Xi}_{t|t-1}^{(j+1)}\right]\mathbf{J}_{t-1}^{\mathrm{T}}.\end{aligned}$$

For $t = \tau', \ldots, 3$, we have,

$$\mathrm{cov}^{(j+1)}(\mathbf{F}_{t-1}, \mathbf{F}_{t-2}) = \mathbf{\Xi}_{t-1|t-1}^{(j+1)}\mathbf{J}_{t-2}^{\mathrm{T}} + \mathbf{J}_{t-1}\left[\mathrm{cov}^{(j+1)}(\mathbf{F}_t, \mathbf{F}_{t-1}) - \widehat{\mathbf{A}}^{(j)}\mathbf{\Xi}_{t-1|t-1}^{(j+1)}\right]\mathbf{J}_{t-2}^{\mathrm{T}}.$$

To compute the second moments using the computed expectations and covariance matrices,

we use the following formulas,

$$\mathbb{E}^{(j+1)}(\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}) = \mathbb{E}^{(j+1)}(\mathbf{F}_t)\mathbb{E}^{(j+1)}(\mathbf{F}_t^{\mathrm{T}}) + \mathrm{var}^{(j+1)}(\mathbf{F}_t), \quad t = 1, \ldots, \tau',$$
$$\mathbb{E}^{(j+1)}(\mathbf{F}_t\mathbf{F}_{t-1}^{\mathrm{T}}) = \mathbb{E}^{(j+1)}(\mathbf{F}_t)\mathbb{E}^{(j+1)}(\mathbf{F}_{t-1}^{\mathrm{T}}) + \mathrm{cov}^{(j+1)}(\mathbf{F}_t, \mathbf{F}_{t-1}), \quad t = 2, \ldots, \tau'.$$

At the $(j+1)^{\text{th}}$ maximization step, due to the identification problem of $\mathbf{\Lambda}\mathbf{F}_t$, we calculate the updated estimates of $\mathbf{\Lambda}$ using Lagrange multiplier as,

$$\mathrm{vec}(\widehat{\mathbf{\Lambda}}^{(j+1)}) = \left[\sum_{t=1}^{\tau'} \mathbb{E}^{(j+1)}(\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}) \otimes \mathbf{W}_t + \begin{pmatrix} k & 0 \\ 0 & 0 \end{pmatrix} \otimes \widehat{\mathbf{P}}^{(j)}\right]^{-1} \mathrm{vec}\left[\sum_{t=1}^{\tau'} \mathbf{W}_t\widetilde{\boldsymbol{\varepsilon}}_t\mathbb{E}^{(j+1)}(\mathbf{F}_t^{\mathrm{T}})\right]. \tag{10}$$

Here k is the Lagrange multiplier that used to constrain the length of the first column of $\widehat{\mathbf{\Lambda}}^{(j+1)}$ to be 1, and $\otimes$ representing the Kronecker product. The value of k can be easily obtained because it is a one-dimensional root finding problem.

Because the factors $\mathbf{F}_t$'s are unobservable, to identify their directions, we also need to constrain each column of $\widehat{\mathbf{\Lambda}}$ to have a positive inner product with the identity vector $\mathbf{1}_m$. Equation (10) is actually a block-wise low dimensional matrix product. In particular, the inverse matrix on the right side of (10) can be simplified to

$$\begin{aligned}
&\left[\sum_{t=1}^{\tau'} \mathbb{E}^{(j+1)}(\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}) \otimes \mathbf{W}_t + \begin{pmatrix} k & 0 \\ 0 & 0 \end{pmatrix} \otimes \widehat{\mathbf{P}}^{(j)}\right]^{-1} \\
=&\sum_{i=1}^{m} \left\{ \left[\sum_{t=1}^{\tau'} \mathbb{E}^{(j+1)}(\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}})\mathbf{W}_{t,ii} + \begin{pmatrix} k & 0 \\ 0 & 0 \end{pmatrix} \widehat{\mathbf{P}}_{ii}^{(j)}\right]^{-1} \otimes \mathbf{E}_i \right\},
\end{aligned}$$

where $\mathbf{W}_{t,ii}$ is the (i,i)th element of $\mathbf{W}_t$, $\mathbf{E}_i$ is a $m \times m$ matrix with value one for the (i,i)th

element and zero otherwise, and $\mathbf{P}_{ii}$ is the (i,i)th element of $\mathbf{P}$. In addition,

$$\widehat{\mathbf{A}}^{(j+1)} = \left[\sum_{t=1}^{\tau'} \mathbb{E}^{(j+1)}(\mathbf{F}_t \mathbf{F}_{t-1}^{\mathrm{T}})\right] \left[\sum_{t=1}^{\tau'} \mathbb{E}^{(j+1)}(\mathbf{F}_{t-1} \mathbf{F}_{t-1}^{\mathrm{T}})\right]^{-1}.$$

The matrix $\mathbf{P}$ is estimated by

$$\begin{aligned}\widehat{\mathbf{P}}^{(j+1)} =& \tau'^{-1} \mathrm{diag}\left\{\sum_{t=1}^{\tau'} \mathbf{W}_t \left[\widetilde{\boldsymbol{\varepsilon}}_t \widetilde{\boldsymbol{\varepsilon}}_t^{\mathrm{T}} - \widetilde{\boldsymbol{\varepsilon}}_t \mathbb{E}^{(j+1)}(\mathbf{F}_t^{\mathrm{T}})(\widehat{\boldsymbol{\Lambda}}^{(j+1)})^{\mathrm{T}} - \widehat{\boldsymbol{\Lambda}}^{(j+1)} \mathbb{E}^{(j+1)}(\mathbf{F}_t) \widetilde{\boldsymbol{\varepsilon}}_t^{\mathrm{T}}\right.\right. \\ &\left.\left. + \widehat{\boldsymbol{\Lambda}}^{(j+1)} \mathbb{E}^{(j+1)}(\mathbf{F}_t \mathbf{F}_t^{\mathrm{T}})(\widehat{\boldsymbol{\Lambda}}^{(j+1)})^{\mathrm{T}}\right] \mathbf{W}_t + (\mathbf{I} - \mathbf{W}_t) \widehat{\mathbf{P}}^{(j)} (\mathbf{I} - \mathbf{W}_t)\right\}.\end{aligned}$$

The estimation procedure is completed by iteratively implementing the E-step, M-step and updating ML estimates of the time series model. In practice, one can use the stopping rules such as fixed number of iterations, relative element-wise changes in parameter estimates below some small threshold or the Aitken acceleration-based stopping criterion exploited in Böhning et al. (1994). Next, we present ML estimates of the time series model in explicit forms.

The covariate time series specified by (2.3) is parsimonious involving few number of parameters. With parameter estimates obtained in the M-step, we update the mean vector $\hat{\boldsymbol{\mu}}$ and the mean reverting parameters $\hat{\boldsymbol{\kappa}} = (\hat{\kappa}_D, \hat{\kappa}_V, \hat{\kappa}_r, \hat{\kappa}_S, \hat{b})^{\mathrm{T}}$ using the ML estimation. First, the covariance matrix $\widehat{\boldsymbol{\Sigma}}$ of innovation vectors $\boldsymbol{\varepsilon}_t$ can be approximated by:

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Lambda}} \left(\widehat{\mathbf{Q}} \widehat{\mathbf{Q}}^{\mathrm{T}} + \widehat{\mathbf{A}} \widehat{\mathbf{Q}} \widehat{\mathbf{A}}^{\mathrm{T}} + \widehat{\mathbf{A}}^2 \widehat{\mathbf{Q}} \widehat{\mathbf{A}}^{2\mathrm{T}} + \widehat{\mathbf{A}}^3 \widehat{\mathbf{Q}} \widehat{\mathbf{A}}^{3\mathrm{T}} + \widehat{\mathbf{A}}^4 \widehat{\mathbf{Q}} \widehat{\mathbf{A}}^{4\mathrm{T}} + \widehat{\mathbf{A}}^5 \widehat{\mathbf{Q}} \widehat{\mathbf{A}}^{5\mathrm{T}}\right) \widehat{\boldsymbol{\Lambda}}^{\mathrm{T}} + \widehat{\mathbf{P}}.$$

Define matrices

$$\mathbf{X}_{\boldsymbol{\Theta},t} = \begin{pmatrix} \mathbf{D}_t & \mathbf{0} & \mathbf{0} & \mathbf{0} & r_t\mathbf{1}_n \\ \mathbf{0} & \mathbf{V}_t & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & 0 & r_t & 0 & 0 \\ 0 & 0 & 0 & S_t & 0 \end{pmatrix},$$

$$\mathbf{A}_{11} = \frac{c}{\tau'}\widehat{\boldsymbol{\Sigma}}^{-1},$$

$$\mathbf{A}_{12} = \frac{1}{\tau'}\widehat{\boldsymbol{\Sigma}}^{-1}\left(\sum_{t=1}^{\tau'-1}\mathbf{W}_t\mathbf{X}_{\boldsymbol{\Theta},t}\right),$$

$$\mathbf{A}_{22} = \frac{1}{\tau'}\sum_{t=1}^{\tau'-1}\left(\mathbf{W}_t\mathbf{X}_{\boldsymbol{\Theta},t}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}_t\mathbf{X}_{\boldsymbol{\Theta},t}\right),$$

$$\mathbf{b}_1 = \frac{1}{\tau'}\widehat{\boldsymbol{\Sigma}}^{-1}\left(\sum_{t=1}^{\tau'-1}\mathbf{W}_t\mathbf{X}_{t+1}\right),$$

$$\mathbf{b}_2 = \frac{1}{\tau'}\left[\sum_{t=1}^{\tau'-1}(\mathbf{W}_t\mathbf{X}_{\boldsymbol{\Theta},t})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}_t\mathbf{X}_{t+1}\right],$$

where constant c is the total number of non-missing values in $\widetilde{\boldsymbol{\varepsilon}}_t$, $t = 2, \ldots, \tau'$. Then, the mean vector $\hat{\boldsymbol{\mu}}$ and the mean reverting parameters $\hat{\boldsymbol{\kappa}}$ are updated by

$$\begin{pmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{\kappa}} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}{}^{\mathrm{T}} & \mathbf{A}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}.$$

Because $\widetilde{\boldsymbol{\varepsilon}}_t = (\mathbf{X}_t - \hat{\boldsymbol{\mu}}) - \widehat{\boldsymbol{\Theta}}(\mathbf{X}_{t-1} - \hat{\boldsymbol{\mu}})$, the updated $\hat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Theta}}$ will be used to compute a new set of $\widetilde{\boldsymbol{\varepsilon}}_t$, $t = 2, \ldots, \tau'$, which are involved in a new implementation of the expectation step (E-step) and maximization step (M-step). First, the initial values of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Theta}$ are calculated. The initial estimate of $\boldsymbol{\mu}$ is $\boldsymbol{\mu}^{(0)} = \bar{\mathbf{X}} = (\sum_{t=1}^{\tau'}\mathbf{X}_t)/\tau'$. The initial estimate of $\boldsymbol{\Theta}$, denoted by $\boldsymbol{\Theta}^{(0)}$, can be obtained by minimizing the sum of squares of $\widetilde{\boldsymbol{\varepsilon}}_t$ as following,

$$L(\boldsymbol{\Theta}) = \sum_{t=1}^{\tau'}\left[(\mathbf{X}_t - \boldsymbol{\mu}^{(0)}) - \boldsymbol{\Theta}(\mathbf{X}_{t-1} - \boldsymbol{\mu}^{(0)})\right]^{\mathrm{T}}\mathbf{B}_{t,t-1}^{\mathrm{T}}\mathbf{B}_{t,t-1}\left[(\mathbf{X}_t - \boldsymbol{\mu}^{(0)}) - \boldsymbol{\Theta}(\mathbf{X}_{t-1} - \boldsymbol{\mu}^{(0)})\right],$$

with respect to $\mathbf{\Theta}$. Here $\mathbf{B}_{t,t-1}$ is the matrix obtained by removing those rows in $\mathbf{I}_m$, if those corresponding X_{it}'s or $X_{i,t-1}$'s are not observed.

Initial values of the residual vectors $\boldsymbol{\varepsilon}_t, t = 2, \cdots, \tau'$ are obtained by

$$\widetilde{\boldsymbol{\varepsilon}}_t^{(0)} = (\mathbf{X}_t - \boldsymbol{\mu}^{(0)}) - \mathbf{\Theta}^{(0)}(\mathbf{X}_{t-1} - \boldsymbol{\mu}^{(0)}).$$

The initial estimate $\mathbf{\Sigma}^{(0)}$ of the covariance matrix $\mathbf{\Sigma}$ can be calculated element-wise by:

$$r_{ij} = \frac{\sum_{t=2}^{\tau'} \widetilde{\varepsilon}_{it}^{(0)} \widetilde{\varepsilon}_{jt}^{(0)} I(X_{it} \& X_{i,t-1} \text{ are observed}) I(X_{jt} \& X_{j,t-1} \text{ are observed})}{\sum_{t=2}^{\tau'} I(X_{it} \& X_{i,t-1} \text{ are observed}) I(X_{jt} \& X_{j,t-1} \text{ are observed})}.$$

The resulting sample covariance matrix $\mathbf{\Sigma}^{(0)}$ may not be positive definite, but the largest few eigenvalues should be positive. Obtain $\mathbf{\Lambda}^{(0)}$ by the eigenvectors of the s largest eigenvalues of $\mathbf{\Sigma}^{(0)}$, and calculate $\mathbf{F}_t^{(0)} = \mathbf{\Lambda}^{(0)\mathrm{T}} \mathbf{W}_t \widetilde{\boldsymbol{\varepsilon}}_t^{(0)}$.

The initial estimates $\mathbf{A}^{(0)}, \mathbf{P}^{(0)}$ are obtained using $\mathbf{\Lambda}^{(0)}, \mathbf{F}_t^{(0)}, \widetilde{\boldsymbol{\varepsilon}}_t^{(0)}$ above. Initial value $\mathbf{A}^{(0)}$ can be obtained by regressing $\mathbf{F}_t^{(0)}$ against $\mathbf{F}_{t-1}^{(0)}$. That is,

$$\mathbf{A}^{(0)} = (\sum_{t=2}^{\tau'} \mathbf{F}_t^{(0)} \mathbf{F}_{t-1}^{(0)\mathrm{T}})(\sum_{t=2}^{\tau'} \mathbf{F}_{t-1}^{(0)} \mathbf{F}_{t-1}^{(0)\mathrm{T}})^{-1}.$$

Besides, $\mathbf{P}^{(0)}$ is a diagonal matrix calculated based on (2.4). The (i,i)th element of $\mathbf{P}^{(0)}$ is the variance of the ith row of $\widetilde{\boldsymbol{\varepsilon}}_t^{(0)} - \mathbf{\Lambda}^{(0)} \mathbf{F}_t^{(0)}$.

$$X_i(t) = [r(t), s(t), D_i(t), V_i(t)]^{\mathrm{T}}$$

$$\begin{bmatrix} r(t_{j+1}) \\ s(t_{j+1}) \\ D_i(t_{j+1}) \\ V_i(t_{j+1}) \end{bmatrix} = \begin{bmatrix} r(t_j) \\ s(t_j) \\ D_i(t_j) \\ V_i(t_j) \end{bmatrix} + \begin{bmatrix} k_r[\theta_r - r(t_j)] \\ k_s[\theta_s - s(t_j)] \\ k_D[\theta_{iD} - D_i(t_j)] + r_D[\theta_r - r(t_j)] \\ k_V[\theta_{iV} - V_i(t_j)] \end{bmatrix} + \begin{bmatrix} 0 \\ r_s \\ I \end{bmatrix} w(t_{j+1}) + \begin{bmatrix} \epsilon_r(t_{j+1}) \\ \epsilon_s(t_{j+1}) \\ \epsilon_i(t_{j+1}) \end{bmatrix} \tag{11}$$

Here $\epsilon_r(t) \sim N(0, \sigma_r^2)$, $\epsilon_s(t) \sim N(0, \sigma_s^2)$, $w(t) \sim N(0, \Sigma_w)$, and $\epsilon_i(t) \sim N(0, \Sigma_\epsilon)$.

$$X_i(t_{j+1}) = X_i(t_j) + \mu[\boldsymbol{\theta}_X, X_i(t_j)] + cw(t_{j+1}) + \boldsymbol{\epsilon}_i(t_{j+1})$$

$c = [0, r_s, I]^{\mathrm{T}}$.

$$L_X(\boldsymbol{\theta}_X | DATA) = \prod_{t_{j+1} \leq \tau} A_{j+1}\left[X_i(t_{j+1}) | X_i(t_j), \boldsymbol{\theta}_X; i \in RS(t_{j+1})\right].$$

Here $RS(t)$ is the risk set at time t

$$\begin{aligned} &A_{j+1}\left[x_i(t_{j+1}) | X_i(t_j), \boldsymbol{\theta}_X; i \in RS(t_{j+1})\right] \\ &= \int_w \left(\prod_{i \in RS(t_{j+1})} f_{\boldsymbol{\epsilon}}\left[x_i(t_{j+1}) - x_i(t_j) - \mu[\boldsymbol{\theta}_X, x_i(t_j)] - cw; \boldsymbol{\theta}_X\right] \right) f_w(w; \boldsymbol{\theta}_X) dw. \end{aligned}$$

Here $f_{\boldsymbol{\epsilon}}$ and f_w are multivariate normal pdfs.

2.B Calibration

To account for the uncertainty in $\widehat{\boldsymbol{\theta}}$, we use a parametric bootstrap simulation to approximate the distribution of $\widehat{\boldsymbol{\theta}}$. The calibration has two parts: first, we use bootstrap to generate the bootstrap version of the covariate process $\boldsymbol{x}_i^*(t_i), i = 1, 2, \cdots, n$. Because we assume a parametric model for the covariate process, parametric simulation methods can be used here to generate $\boldsymbol{x}_i^*(t_i)$. Repeating the ML estimation procedure, one obtains the bootstrap version of estimates of the parameters for the covariate process.

The second part of the calibration process is to obtain the bootstrap version estimates of parameters for the failure-time distribution. The traditional bootstrap method that uses simple random sampling with replacement can be problematic with heavy censoring, as it can result in bootstrap samples without enough failures for the estimation of the parameters. Here we use the random weighted bootstrap method (e.g., Newton and Raftery 1994, Jin et al. 2001) to obtain the bootstrap version estimates of the parameters. See Hong et al. 2009 for another application of random weighted bootstrap in calibration PIs. In particular, with a set of random weights Z_i generated from any positive continuous distribution with $\mathrm{E}(Z_i) = \sqrt{\mathrm{Var}(Z_i)}$, the random weighted likelihood is

$$L_T^*(\boldsymbol{\theta}_T|DATA) = \prod_{i=1}^{n} \left\{ \left[\prod_{k=1}^{K} \left(\lambda_k[t_i; x^*(t_i)] e^{-\Lambda[t_i;\boldsymbol{x}_i^*(t_i)]}\right)^{Z_i\delta_{ki}} \right] \times \left(e^{-\Lambda[t_i;\boldsymbol{x}_i^*(t_i)]}\right)^{Z_i \prod_{k=1}^{K}(1-\delta_{ki})} \right\}$$

Here $\boldsymbol{x}_i^*(t_i)$ is the bootstrap sample generated in the first part. The bootstrap versions of the parameter estimates for the failure-time distribution can be obtained by maximizing the random weighted likelihood. Combining with the bootstrap version estimates of the parameter for the covariates, we obtain the bootstrap version of $\widehat{\boldsymbol{\theta}}$, which is denoted by $\widehat{\boldsymbol{\theta}}^*$.

With B bootstrap samples of $\widehat{\boldsymbol{\theta}}^*$, the calibration of PIs for the population can be done by using a procedure similar to the procedure described in Section 6.2 of Hong et al. (2009). Here B is usually chosen to be a large number (e.g., $B = 10,000$). The calibration of PIs for individuals can be done by using a procedure similar to the procedure described in

Section 5.4 of Hong et al. (2009).

Alternatively, one can use a random sample of ML estimates of parameters in the intensity functions from their asymptotic distributions, if computational time is limited.

$$
\begin{aligned}
\mathbb{E}^{(j)}(\mathbf{F}_t) &= \mathbb{E}^{(j)}(\mathbf{F}_t|\mathcal{F}_{t-1}, \mathbf{v}_t, \ldots, \mathbf{v}_\tau) = \mathbf{F}_{t|t-1} + \boldsymbol{\Xi}_{t|t-1}\mathbf{q}_{t-1}, \\
\mathbf{q}_{t-1} &= \{\boldsymbol{\Lambda}^{(j)}\}^{\mathrm{T}}\mathbf{U}_t^{-1}\mathbf{v}_t + \mathbf{L}_t^{\mathrm{T}}\mathbf{q}_t, \\
\mathbf{L}_t &= \mathbf{A}^{(j)} - \mathbf{A}^{(j)}\boldsymbol{\Xi}_{t|t-1}\{\boldsymbol{\Lambda}^{(j)}\}^{\mathrm{T}}\mathbf{U}_t^{-1}\{\boldsymbol{\Lambda}^{(j)}\}, \\
\mathbf{q}_\tau &= \mathbf{0}, \\
\mathbb{E}^{(j)}(\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}) &= \mathbb{E}^{(j)}(\mathbf{F}_t)\mathbb{E}^{(j)}(\mathbf{F}_t^{\mathrm{T}}) + \mathrm{var}^{(j)}(\mathbf{F}_t), \\
\mathrm{var}^{(j)}(\mathbf{F}_t) &= \mathrm{var}^{(j)}(\mathbf{F}_t|\mathcal{F}_{t-1}, \mathbf{v}_t, \ldots, \mathbf{v}_\tau) = \boldsymbol{\Xi}_{t|t-1} - \boldsymbol{\Xi}_{t|t-1}\mathbf{G}_{t-1}\boldsymbol{\Xi}_{t|t-1} \\
\mathbf{G}_{t-1} &= \{\boldsymbol{\Lambda}^{(j)}\}^{\mathrm{T}}\mathbf{U}_t^{-1}\{\boldsymbol{\Lambda}^{(j)}\} + \mathbf{L}_t^{\mathrm{T}}\mathbf{G}_t\mathbf{L}_t, \\
\mathbf{G}_\tau &= \mathbf{0}, \\
\mathbb{E}(\mathbf{F}_t\mathbf{F}_{t-1}^{\mathrm{T}}) &= \mathbb{E}^{(j)}(\mathbf{F}_t)\mathbb{E}^{(j)}(\mathbf{F}_{t-1}^{\mathrm{T}}) + \mathrm{cov}^{(j)}(\mathbf{F}_t, \mathbf{F}_{t-1}), \\
\mathrm{cov}^{(j)}(\mathbf{F}_t, \mathbf{F}_{t-1}) &= \boldsymbol{\Xi}_{t|t}\mathbf{J}_{t-1}^{\mathrm{T}} + \mathbf{J}_t\{\mathrm{cov}^{(j)}(\mathbf{F}_{t+1}, \mathbf{F}_t) - \mathbf{A}^{(j)}\boldsymbol{\Xi}_{t|t}\}\mathbf{J}_{t-1}^{\mathrm{T}}, \\
\mathrm{cov}^{(j)}(\mathbf{F}_\tau, \mathbf{F}_{\tau-1}) &= \{\mathbf{I} - \boldsymbol{\Xi}_{\tau|\tau-1}(\boldsymbol{\Lambda}^{(j)})^{\mathrm{T}}\mathbf{U}_t^{-1}\boldsymbol{\Lambda}^{(j)}\}\mathbf{A}^{(j)}\boldsymbol{\Xi}_{\tau-1|\tau-1}, \\
\mathbf{J}_t &= \boldsymbol{\Xi}_{t|t}(\mathbf{A}^{(j)})^{\mathrm{T}}\boldsymbol{\Xi}_{t+1|t}^{-1}.
\end{aligned}
$$

Bibliography

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

J. Bai and S. Ng. Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3:89–163, 2008.

J. Bai and P. Wang. Identification and estimation of dynamic factor models. Department of Economics Discussion Papers. Columbia University, New York. Available at http://academiccommons.columbia.edu/catalog/ac%3A146472, 2012.

M. Banbura and M. Modugno. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journel of Applied Econometrics*, 29:133–160, 2012.

D. Böhning, E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46:373–388, 1994.

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

F. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economics Statistics*, 13:253–263, 1995.

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

J. Duan, J. Sun, and T. Wang. Multiperiod corporate default prediction – a forward intensity approach. *Journal of Econometrics*, 170:191–209, 2012.

J. C. Duan. Clustered defaults. *National University of Singapore working paper, available at Social Science Research Network 1511397*, 2010.

D. Duffie, L. Saita, and K. Wang. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83:635–665, 2007.

D. Duffie, A. Eckner, G. Horel, and L. Saita. Frailty correlated default. *The Journal of Finance*, 64:2089–2123, 2009.

K. Giesecke and B. Kim. Systemic risk: what defaults are telling us. *Management Science*, 57:1387–1405, 2011.

K. Giesecke, F. Longstaff, S. Schafer, and I. Strebulaev. Corporate bond default risk: a 150-year perspective. *Journal of Financial Economics*, 102:233–250, 2011.

Y. Hong. On computing the distribution function for the poisson binomial distribution. *Computational Statistics and Data Analysis*, 59:41–51, 2013.

Y. Hong and W. Q. Meeker. Field-failure and warranty prediction based on auxiliary use-rate information. *Technometrics*, 52:148–159, 2010.

Y. Hong, W. Q. Meeker, and J. D. McCalley. Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *The Annals of Applied Statistics*, pages 857–879, 2009.

Z. Jin, Z. Ying, and L. Wei. A simple resampling method by perturbing the minimand. *Biometrika*, 88(2):381–390, 2001.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons Inc, 2nd edition, 2002.

C. Lam and Q. Yao. Estimation for latent factors for high-dimensional time series. *Biometrika*, 98:901–918, 2011.

C. Lam and Q. Yao. Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40:694–726, 2012.

J. F. Lawless and M. Fredette. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92:529–542, 2005.

W. Q. Meeker and L. A. Escobar. *Statistical Methods for Reliability Data*. John Wiley & Sons, 1998. ISBN 0-471-14328-6.

R. C. Merton. On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, 29:449–470, 1974.

M. A. Newton and A. E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48, 1994.

J. Pan and Q. Yao. Modelling multiple time series via common factors. *Biometrika*, 95: 365–379, 2008.

X. Peng and S. Kou. Default clustering and valuation of collateralized debt obligations. unpublished manuscript, 2009.

J. H. Stock and M. Watson. Has the business cycle changed and why? *NBER Macroeconomics Annual*, 17:159–230, 2002.

R. Tsay. *Analysis of Financial Time Series*. New York: Wiley, 3rd edition, 2010.

A. Y. Volkova. A refinement of the central limit theorem for sums of independent random indicators. *Theory of Probability and Its Applications*, 40:791–794, 1996.

Chapter 3 Two-sided Tolerance Intervals for Members of the (Log)-Location-Scale Family of Distributions

Abstract

In this chapter, we propose methods to calculate exact factors for two-sided control-the-center and control-both-tails tolerance intervals (TI) for the (log)-location-scale family of distributions, based on complete or Type II censored data. With Type I censored data, exact factors do not exist. For this case we developed an algorithm to compute approximate factors. Our approaches are based on Monte-Carlo simulations. We also provide algorithms for computing TIs that control the probability in both tails of a distribution. A simulation study for Type I censored data shows that the estimated coverage probability (CP) is close to the nominal confidence level when the expected number of uncensored observations is moderate to large. We illustrate the methods with applications using different combinations of distributions and types of censoring.

Key Words: Censored Data; Lognormal; Maximum Likelihood; Monte-Carlo Simulation; Pivotal Quantity; Weibull.

3.1 Introduction

3.1.1 Motivation

In many applications such as quality control and reliability engineering, one is often interested in estimating the content of a distribution with a high degree of confidence, based on an observed random sample (perhaps censored) from that distribution. Such an estimation procedure is different from obtaining a confidence interval to contain a scalar characteristic, such as a distribution mean or quantile. For example, one often needs an interval that can cover at least a proportion β of a distribution. Tolerance intervals (TIs) are frequently used to describe the distribution of a process output, a particular random quantity characteristic, or the lifetime distribution of a product (e.g., Krishnamoorthy and Mathew 2009). More applications of TIs are also described in Hahn and Meeker (1991). A TI to contain a proportion of the distribution is used for this purpose. In this chapter, we use $(\beta, 1-\alpha)$ to denote a TI that has $100(1-\alpha)\%$ confidence to cover at least a content β of a population.

Members of the (log)-location-scale family of distributions (e.g., the normal, the Weibull and the lognormal) are commonly used to model data in the physical and engineering sciences. Applications involving life testing often result in right-censored data. Thus, the objective of this chapter is to develop a general procedure to compute two-sided TI for the location-scale and log-location-scale families of distributions based on complete and right censored data.

3.1.2 Literature Review

In this chapter, we consider TIs for continuous distributions. For the normal distribution, TIs based on both complete and censored data have been well developed. Specifically, Odeh and Owen (1980) provided exact factors of one-sided tolerance bounds (TBs) and two-sided TIs with various combinations of confidence levels, contents, and sample sizes for a normal distribution. The tables of Odeh and Owen (1980), however, only apply to the case where

the degrees of freedom is equal to the sample size minus one. Weissberg and Beatty (1960) and Howe (1969) proposed approximate factors for normal TIs with sample variance having arbitrary degrees of freedom. Eberhardt et al. (1989) developed a FORTRAN program to compute the exact factors of normal TIs for a wide ranges of degrees of freedom, effective sample sizes, and confidence levels. Their methods show advantages especially when the effective sample size is much smaller than the degree of freedom, for example, in a regression setting.

In the situation where censored observations are present, Krishnamoorthy and Mathew (2009) discussed the calculation of one-sided TBs for normal and the Weibull distributions. One-sided TBs for lognormal and the smallest extreme value (SEV) distribution can be obtained based on their one-to-one relationship with normal and the Weibull distributions, respectively. For the Laplace distribution, computation of one-sided TBs and two-sided TIs were studied, for example, in Shyu and Owen (1986a) and Shyu and Owen (1986b), using Monte-Carlo simulations. For symmetric distributions in the location-scale family, Krishnamoorthy and Xie (2011) derived algorithms for computing control-the-center TIs and equal-tail TIs using pivotal quantities based on ML estimators under complete and Type II censored data. They also considered an adjusted method for Type I censored data. Our objective is to provide a generic algorithm that can compute TIs for both symmetric and non-symmetric distributions in the (log)-location-scale family. For the (log)-location-scale family of distributions, Xie et al. (2014) developed a general method to calculate the exact one-sided prediction bounds and two-sided prediction intervals under complete and Type II censored data, and an adjusted procedure for Type I censored data.

Bergquist (2006) developed a parametric bootstrap procedure to calculate the one-sided TBs and two-sided TIs for response of a (non)-linear-mixed-effect model in the pharmaceutical application setting. They concluded that the bootstrap TIs usually have considerably lower actual CP than the nominal confidence level when the sample size is small to moderate, in the situations involving a simple random sample and a (non)-linear-mixed-effect model.

The performance of their parametric bootstrap procedures largely depends on the accuracy of the model parameter estimates. Our methods, however, are based on the pivotal properties of the location-scale family of distributions. Thus, the computed TI factors do not depend on the parameter estimates, when data are complete or Type II censored. Besides, we evaluate and control the CP directly in the Monte-Carlo simulations to calculate the factors. Hence our methods do not have the problems pointed out in Bergquist (2006).

One-sided TBs have been well studied in literature. The lower (upper) TB is one-sided lower (upper) confidence bound with confidence level $100(1-\alpha)\%$ on the β quantile [$(1-\beta)$ quantile]. Our focus will therefore be on two-sided TIs. There are two types of commonly-used TIs: control-the-center TIs and control-both-tails TIs. One can also see description of these two types of TI in Odeh and Owen (1980), Hahn and Meeker (1991, Chapter 4) and Krishnamoorthy and Mathew (2009, Chapter 1).

3.1.3 Overview

The rest of this chapter is organized as follows. Section 3.2 introduces the data and distributions used in this chapter. Section 3.3 introduces definitions of two-sided TI procedures for the (log)-location-scale family of distributions. Section 3.4 describes algorithms for computing TIs for the (log)-location-scale family of distributions. Section 3.5 focuses on applications of the developed procedure for three examples. Section 3.6 discusses the results of a simulation study for Type I censored data. Section 3.7 contains conclusions and some areas for future research.

3.2 Data and Model

3.2.1 Data

We consider TIs for complete, Type I (time), and Type II (failure) censored data. For complete and Type II censored data, our TI procedure is exact. For Type I censored data,

we derive an approximate procedure.

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ denote n independent random variables corresponding to the response of interest. The censoring indicator δ_i, $i = 1, 2, \ldots, n$ is defined as $\delta_i = 1$ when X_i is uncensored and $\delta_i = 0$ when X_i is right censored. The observed data are denoted by $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$. For simplicity, and without loss of generality, we assume that $x_1 < \cdots < x_n$.

We consider three types of data:

1. Complete data. In this case the $x_1, x_2, \ldots, x_n$ are the realizations of the X_i. Because there is no censoring, $\delta_i = 1$ for all i.

2. Type II censored data. In this case the data consist of the r smallest observations $x_1 < \cdots < x_r$ in the sample of size n and the additional information that the remaining $(n - r)$ sample values are censored at x_r. Then $\delta_i = 1$ for $i = 1, \ldots, r$ and $\delta_i = 0$ for $i = r+1, \ldots, n$. Note that r is a pre-specified integer between 1 and n but x_r is random. To ensure estimability of the parameters we consider situations in which $r \geq 2$. Type II censoring sometimes arises in life testing.

3. Type I censored data. The data consist of the r smallest observations $(x_1, \ldots, x_r)$ that satisfy $x_i \leq x_c$, where the fixed bound x_c is pre-specified, and the additional information that the remaining $(n - r)$ observations are censored at x_c. Then $\delta_i = 1$ for $i = 1, \ldots, r$ and $\delta_i = 0$ for $i = r + 1, \ldots, n$. Note that, in this case, the number of uncensored observations r is random and x_c is fixed. We exclude the case $r = 0$ because the maximum likelihood (ML) estimator of the parameters does not exist. Type I censored data arise in life testings and also in other applications such as when a measurement system has an upper bound for its readings.

3.2.2 Model

Because the (log)-location-scale family of distributions has many applications for modeling lifetime data or time to event data in reliability and survival applications, and many other areas of science and engineering, our study focuses on computing TIs for members in this family of distributions. The location-scale family of distributions is characterized by a location parameter $-\infty < \mu < \infty$ and a scale parameter $\sigma > 0$, which are typically unknown and need to be estimated from data. The cumulative distribution function (cdf) and the probability density function (pdf) of the location-scale distributions are

$$F(x;\boldsymbol{\theta}) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad \text{and} \quad f(x;\boldsymbol{\theta}) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), \tag{3.1}$$

respectively, where $\boldsymbol{\theta} = (\mu, \sigma)$. A random variable Y belongs to the log-location-scale family if $X = \log(Y)$ follows a location-scale distribution. For the log-location-scale family, $\exp(\mu)$ is a scale parameter and $\sigma > 0$ is a shape parameter. The cdf and pdf of the log-location-scale family of distributions are

$$F(y;\boldsymbol{\theta}) = \Phi\left[\frac{\log(y)-\mu}{\sigma}\right] \quad \text{and} \quad f(y;\boldsymbol{\theta}) = \frac{1}{\sigma y}\phi\left[\frac{\log(y)-\mu}{\sigma}\right],$$

respectively. Here $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf for a standard distribution in the location-scale family ($\mu = 0$ and $\sigma = 1$), respectively.

Some commonly-used distributions in the location-scale family, their standard pdf $\phi(x)$ and cdf $\Phi(x)$, along with the corresponding log-location-scale distributions are summarized

as follows:

$$\begin{aligned}
\text{Normal (Lognormal)}: &\quad \phi_{\text{norm}}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), &\quad \Phi_{\text{norm}}(z) = \int_{-\infty}^{z} \phi_{\text{norm}}(v)dv \\
\text{Logistic (Loglogistic)}: &\quad \phi_{\text{logis}}(z) = \frac{\exp(z)}{[1+\exp(z)]^2}, &\quad \Phi_{\text{logis}}(z) = \frac{\exp(z)}{1+\exp(z)} \\
\text{LEV (Fréchet)}: &\quad \phi_{\text{sev}}(z) = \exp[-z - \exp(-z)], &\quad \Phi_{\text{sev}}(z) = \exp[-\exp(-z)] \\
\text{SEV (Weibull)}: &\quad \phi_{\text{lev}}(z) = \exp[z - \exp(z)], &\quad \Phi_{\text{lev}}(z) = 1 - \exp[-\exp(z)].
\end{aligned}$$

where SEV and LEV are the smallest extreme value and largest extreme value distributions, respectively.

3.2.3 Parameter Estimation

Because parametric distributions in the (log)-location-scale family are used to model the (possibly censored) data, it is desirable to obtain parameter estimates by the maximum likelihood (ML) method. The ML method is easy to implement given a parametric model and the data censoring type.

Assuming the right censored samples are independent and identically distributed (iid), the likelihood function of the parameters $\boldsymbol{\theta} = (\mu, \sigma)$ given the data is,

$$L(\boldsymbol{\theta}|\text{DATA}) = \prod_{i=1}^{n} [f(x_i)]^{\delta_i} [1 - F(x_c)]^{1-\delta_i},$$

where $f(\cdot)$ and $F(\cdot)$ are the pdf and cdf given in (3.1), respectively. Because ML estimates of the (log)-location-scale family generally do not have explicit forms for censored data, numerical methods are applied to find the values of $\boldsymbol{\theta}$ that maximize the log likelihood function. The ML estimates are denoted by $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma})$.

3.3 Two-sided Tolerance Interval Procedures

In this chapter, we develop methods to construct two-sided TIs for the members in the (log)-location-scale family of distributions. In particular, we construct a $(\beta, 1-\alpha)$ TI,

$$\left[\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha), \quad \widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha)\right] \tag{3.2}$$

for given data $\boldsymbol{x}$ from a particular member of this family of distributions. Here, β is the content of the distribution that the TI should cover with $100(1-\alpha)\%$ confidence. We focus on the TIs for the location-scale family of distributions. For a distribution in the log-location-scale family, we first compute, using the log transformed data, the TI for the corresponding location-scale distribution as $\left[\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha), \quad \widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha)\right]$. Then, the desired TI is obtained as

$$\left\{\exp\left[\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha)\right], \quad \exp\left[\widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha)\right]\right\}.$$

3.3.1 Control-the-Center TIs

Let $F(x; \boldsymbol{\theta})$ be the cdf of the random variable X from which the not-yet-observed data $\boldsymbol{X}$ will be taken. For notational simplicity, let

$$\Delta F_L(\boldsymbol{x}) = F\left[\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha); \boldsymbol{\theta}\right] \quad \text{and} \quad \Delta F_U(\boldsymbol{x}) = 1 - F\left[\widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha); \boldsymbol{\theta}\right] \tag{3.3}$$

be the population contents below $\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha)$ and above $\widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha)$ for any realization of $\boldsymbol{X} = \boldsymbol{x}$, respectively. Then, the population content of the TI in (3.2) is

$$\Delta F(\boldsymbol{x}) = 1 - \Delta F_U(\boldsymbol{x}) - \Delta F_L(\boldsymbol{x}). \tag{3.4}$$

In applications, one is interested in the probability with which the random quantity $\Delta F(\boldsymbol{X})$ exceeds β. The TI in (3.2) is said to be an exact $(\beta, 1-\alpha)$ control-the-center TI if

$$\Pr\left\{\Delta F(\boldsymbol{X}) > \beta\right\} = 1 - \alpha. \tag{3.5}$$

The coverage probability (CP) of a control-the-center TI is

$$\mathrm{CP}(\boldsymbol{\theta}) = \Pr\left\{\Delta F(\boldsymbol{X}) > \beta\right\} = \mathrm{E}_{\boldsymbol{X}}\left(\mathrm{I}\left\{\Delta F(\boldsymbol{X}) > \beta\right\}\right), \tag{3.6}$$

where $\mathrm{E}_{\boldsymbol{X}}$ is the expectation with respect to the distribution of $\boldsymbol{X}$, and the indicator function $\mathrm{I}[A]$ is equal to 1 when the statement A is true and is equal to 0 otherwise. Note that in some cases, the CP will depend on the parameter $\boldsymbol{\theta}$ through the probability distribution of the data $\boldsymbol{X}$.

For members of the location-scale family of distributions, the TI is constructed using the following forms,

$$\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha) = \widehat{\mu} + g_{L(1-\alpha,\beta)}\widehat{\sigma} \quad \text{and} \quad \widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha) = \widehat{\mu} + g_{U(1-\alpha,\beta)}\widehat{\sigma}, \tag{3.7}$$

where $g_L = g_{L(1-\alpha,\beta)}$ and $g_U = g_{U(1-\alpha,\beta)}$ are the factors that provide the desired CP.

Result 1 *Let* $\boldsymbol{Z} = (Z_1, Z_2)$ *where* $Z_1 = (\widehat{\mu} - \mu)/\sigma$ *and* $Z_2 = \widehat{\sigma}/\sigma$. *The CP in* (3.6) *only depends on the distribution of* $\boldsymbol{Z}$. *That is,*

$$\mathrm{CP}(\boldsymbol{\theta}) = E_{\boldsymbol{Z}}\left(\mathrm{I}\left\{\Phi(Z_1 + g_U Z_2) - \Phi(Z_1 + g_L Z_2) > \beta\right\}\right), \tag{3.8}$$

where the factors g_L *and* g_U *are chosen to satisfy*

$$1 - \alpha = E_{\boldsymbol{Z}}\left(\mathrm{I}\left\{\Phi(Z_1 + g_U Z_2) - \Phi(Z_1 + g_L Z_2) > \beta\right\}\right), \tag{3.9}$$

when data are complete or Type II censored.

The proof of **Result 1** is given in Appendix 3.A. When data are complete or Type II censored, the distributions of pivotal quantities $\boldsymbol{Z}$ for a location-scale distribution do not depend on any unknown parameters (e.g., Krishnamoorthy and Mathew 2009, pages 16–17). When data are Type I censored, the quantities $\boldsymbol{Z}$ are approximately pivotal and thus their distributions depend on the unknown parameters. In particular, we have the following result for Type I censored data.

Result 2 *For any given censoring time x_c, the quantities $\boldsymbol{Z}$ are approximately pivotal under Type I censoring. Their distributions only depend on the expected fraction of uncensored observations $p_f = \Phi[(x_c - \mu)/\sigma]$.*

A proof of this result is given in Appendix 3.B.

3.3.2 Control-Both-Tails TIs

The control-the-center TI only guaranties that the population content is at least β but the procedure is not specific on the amount of content in each tail. In some applications, in contrast to the control-the-center TI, one needs to construct a more stringent TI that leaves no more than a specified amount of content in each tail of the distribution. The control-both-tails TI is used for this purpose. See Hahn and Meeker (1991, Chapter 4) for a description of the two alternative types of TIs. Although it is possible to specify different probabilities in each tail, it is more common that the specified probabilities are equal and we will follow that convention. The TI in (3.2) is said to be an exact $(\beta, 1-\alpha)$ control-both-tails TI if

$$\Pr\{\Delta F_L(\boldsymbol{X}) \leq (1-\beta)/2, \Delta F_U(\boldsymbol{X}) \leq (1-\beta)/2\} = 1-\alpha. \tag{3.10}$$

Note that the population content in-between the interval endpoints is at least β for the TI in (3.10). The TI in (3.10), however, is more stringent than the TI in (3.5) because it requires the same content in each tail.

The CP of a control-both-tails TI is

$$\begin{aligned} \mathrm{CP}(\boldsymbol{\theta}) &= \Pr\left\{\Delta F_L(\boldsymbol{X}) \le (1-\beta)/2, \Delta F_U(\boldsymbol{X}) \le (1-\beta)/2\right\} \\ &= \mathrm{E}_{\boldsymbol{X}}\left(\mathrm{I}\left\{\Delta F_L(\boldsymbol{X}) \le (1-\beta)/2, \Delta F_U(\boldsymbol{X}) \le (1-\beta)/2\right\}\right). \end{aligned} \tag{3.11}$$

For members of the location-scale family of distributions, the TI is constructed using the following forms

$$\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha) = \widehat{\mu} + g'_{L(1-\alpha,\beta)}\widehat{\sigma} \quad \text{and} \quad \widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha) = \widehat{\mu} + g'_{U(1-\alpha,\beta)}\widehat{\sigma},$$

where $g'_L = g'_{L(1-\alpha,\beta)}$ and $g'_U = g'_{U(1-\alpha,\beta)}$ are the factors that provide the desired CP.

Result 3 *The CP in* (3.11) *only depends on the distribution of* $\boldsymbol{Z}$*. That is,*

$$\mathrm{CP}(\boldsymbol{\theta}) = E_{\boldsymbol{Z}}\left(\mathrm{I}\left\{\Phi(Z_1 + g'_L Z_2) \le (1-\beta)/2, \Phi(Z_1 + g'_U Z_2) \ge (1+\beta)/2\right\}\right) \tag{3.12}$$

and the factors g'_L *and* g'_U *are chosen to satisfy*

$$1-\alpha = E_{\boldsymbol{Z}}\left(\mathrm{I}\left\{\Phi(Z_1 + g'_L Z_2) \le (1-\beta)/2, \Phi(Z_1 + g'_U Z_2) \ge (1+\beta)/2\right\}\right), \tag{3.13}$$

when data are complete or Type II censored.

The proof of **Result 3** is similar to that of **Result 1**. Thus it is omitted.

3.3.3 TIs with Equal Error Probabilities

The pairs of (g_L, g_U) for a $(\beta, 1-\alpha)$ TI are not uniquely determined. It is appropriate to calculate a "balanced" TI, such that the two one-sided TBs constructed by its lower and upper bounds have equal error probabilities. That is, if $[\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha),\ \widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha)]$ is a balanced TI, then using the lower end point (the upper end point) of the TI as a lower (upper) tolerance bound to contain a fraction $(1+\beta)/2$ of the population has the desirable

property that the coverage probabilities of these two bounds are the same. Specifically, we define the CP for each of these bounds as

$$\mathrm{CP}_L(\boldsymbol{\theta}) = \Pr\left\{F\left[\underline{T}(\boldsymbol{x}, \beta, 1-\alpha); \boldsymbol{\theta}\right] \leq (1-\beta)/2\right\} \tag{3.14}$$

$$\mathrm{CP}_U(\boldsymbol{\theta}) = \Pr\left\{1 - F\left[\widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha); \boldsymbol{\theta}\right] \leq (1-\beta)/2\right\}. \tag{3.15}$$

A $(\beta, 1-\alpha)$ TI with equal error probabilities is the TI in (3.5) or (3.10) with an additional constraint that

$$\mathrm{CP}_U(\boldsymbol{\theta}) = \mathrm{CP}_L(\boldsymbol{\theta}). \tag{3.16}$$

3.4 Computation of Tolerance Intervals

In this section, we develop algorithms to calculate factors of control-the-center and control-both-tails TI with CP equal to the nominal confidence level. To obtain the unique pair of factors, equation (3.9) and (3.16) need to be solved for control-the-center and (3.13) and (3.16) for control-both-tails TI.

Because the joint distribution of $\boldsymbol{Z}$ is complicated and does not have an explicit form, especially with censored observations, and there are usually no simplifications for (3.8) and (3.12) when the distribution in the location-scale family is non-symmetric, the equations need to be solved numerically. In our algorithm, the CP in (3.8) or (3.12) of a given procedure is evaluated directly through Monte-Carlo simulations from the joint distribution of $\boldsymbol{Z}$.

3.4.1 Control-the-Center TIs

For complete, Type II, or Type I censored data from the location-scale family of distribution, we develop **Algorithm 1** to calculate a control-the-center TI.

3.4.1.1 Algorithm 1

Denote by $\widehat{\mu}$ and $\widehat{\sigma}$ the ML estimators for μ and σ obtained from the available data.

1. Use the corresponding location-scale distribution with parameters $(\widehat{\mu}, \widehat{\sigma})$ to simulate $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ with the same sample size n and censoring pattern as the available data. For notation simplicity, the entries in $\boldsymbol{x}^*$ are sorted such that $x_1^* < x_2^* < \cdots < x_n^*$.

   - If the original data are complete, the data for the this iteration of the simulation are $x_1^*, \ldots, x_n^*$.
   - If the original data are Type II censored, then the data for the simulation are $x_1^*, \ldots, x_r^*$ and the additional information that the remaining $(n-r)$ data points are censored at x_r^*.
   - If the original data are Type I censored, retain as exact observations the r realizations $x_1^* < \cdots < x_r^*$ values that satisfy $x_i^* \leq x_c$ where x_c is the same as the censoring bound in the data. The remaining $(n-r)$ simulated values are censored at x_c. In this censoring scheme r is random. Note that the expected proportion of non-censored observations, in the simulation, equal to the ML $\widehat{p}_f = \Phi[(x_c - \widehat{\mu})/\widehat{\sigma}]$ for the fraction of non-censored observations for the original data. Samples with $r = 0$ are discarded.

2. Compute the ML estimates $(\widehat{\mu}^*, \widehat{\sigma}^*)$ using the simulated data $\boldsymbol{x}^*$.

3. Repeat steps 1-2 B times. The simulated data in each iteration and the corresponding ML estimates are denoted as $\boldsymbol{x}_j^*$ and $(\widehat{\mu}_j^*, \widehat{\sigma}_j^*)$, respectively. Then for any pair of (g_L, g_U) in the two-dimensional domain, we calculate the CP using

$$\mathrm{CP}(g_L, g_U) \approx \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}\Big[\Phi\left(\frac{\widehat{\mu}_j^* + g_U \widehat{\sigma}_j^* - \widehat{\mu}}{\widehat{\sigma}}\right) - \Phi\left(\frac{\widehat{\mu}_j^* + g_L \widehat{\sigma}_j^* - \widehat{\mu}}{\widehat{\sigma}}\right) > \beta\Big].$$

In step 3, one specifies ranges for g_L and g_U that are typically wide enough. Then one computes the CP for a grid of g_L and g_U. The factors (g_L, g_U) that provide $\mathrm{CP}(g_L, g_U) = 1-\alpha$ are chosen subject to the equal error probability in each tail constraint introduced in Section 3.3.3. The intersection point of the two curves (3.17) and (3.18) are calculated using linear interpolation.

$$\mathrm{CP}(g_L, g_U) = 1 - \alpha, \tag{3.17}$$

$$\mathrm{CP}_L(g_L) = \mathrm{CP}_U(g_U). \tag{3.18}$$

Computing the line corresponding to the constraint (3.18) requires calculation of the CP for one-sided TBs as follows

$$\mathrm{CP}_L(g_L) = \mathrm{CP}_L\left[\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha)\right] \approx \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left[\Phi\left(\frac{\widehat{\mu}_j^* + g_L\widehat{\sigma}_j^* - \widehat{\mu}}{\widehat{\sigma}}\right) \leq (1-\beta)/2\right],$$

$$\mathrm{CP}_U(g_U) = \mathrm{CP}_U\left[\widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha)\right] \approx \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left[1 - \Phi\left(\frac{\widehat{\mu}_j^* + g_U\widehat{\sigma}_j^* - \widehat{\mu}}{\widehat{\sigma}}\right) \leq (1-\beta)/2\right],$$

where $\widehat{\mu}_j^*, \widehat{\sigma}_j^*$ are the ML estimates used to compute $\mathrm{CP}(g_L, g_U)$.

When data are complete or Type II censored, the quantities Z_1 and Z_2 have the exact pivotal properties. Thus, $\widehat{\mu} = 0$ and $\widehat{\sigma} = 1$ can also be used in **Algorithm 1**, providing the advantage that for a specified distribution in the location-scale family, the factors (g_L, g_U) for given n and $r \leq n$ could be computed once and for all. Similarly, $\widehat{\mu} = 0$ and $\widehat{\sigma} = 1$ can be specified for the Monte-Carlo simulations when calculating CP of the one-sided TBs for the equal-tail constraint (3.18).

When data are Type I censored, however, Z_1 and Z_2 are only approximately pivotal. Asymptotically, the joint distribution of $\boldsymbol{Z}$ only depends on the expected fraction of uncensored observations p_f, which can be estimated using the ML method as $\widehat{p}_f = (x_c - \widehat{\mu})/\widehat{\sigma}$. Hence, $\widehat{\mu}$ and $\widehat{\sigma}$ are specified in **Algorithm 1**, providing the advantage that the censoring point x_c in the Monte-Carlo simulations is the same as that of the real data.

3.4.2 Control-Both-Tails TIs

For distributions in the location-scale family, we can use a procedure that is similar to the control-the-center TI procedure to construct control-both-tails TI with equal tail-probabilities. In this case, CP in **Algorithm 1** should be evaluated as:

$$\mathrm{CP}(g'_L, g'_U) \approx \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left[\Phi\left(\frac{\widehat{\mu}_j^{**} + g'_U\widehat{\sigma}_j^{**} - \widehat{\mu}}{\widehat{\sigma}}\right) \geq \frac{1+\beta}{2} \text{ and } \Phi\left(\frac{\widehat{\mu}_j^{**} + g'_L\widehat{\sigma}_j^{**} - \widehat{\mu}}{\widehat{\sigma}}\right) \leq \frac{1-\beta}{2}\right].$$

The contour line $\mathrm{CP}_L(g'_L) = \mathrm{CP}_U(g'_U)$ of factors for equal-tail TIs is the same, regardless of the types of TI (i.e., control-the-center or control-both-tails).

3.4.3 Computation of TIs for the Log-Location-Scale Family of Distributions

If data follow a distribution in the log-location-scale family, we first take the log transformation of the data. Then for the transformed data from the corresponding location-scale distribution, **Algorithm 1** can be implemented to compute the factors of the desired TI. Last, take anti-logs (exponential) of the TI endpoints for the log transformed data and we get the TI for the original data as $\Big[\exp(\widehat{\mu} + g_L\widehat{\sigma}), \quad \exp(\widehat{\mu} + g_U\widehat{\sigma})\Big]$.

3.5 Applications

In this section, our method is illustrated in three different applications to compute TIs for the Weibull, the lognormal, and the loglogistic distributions.

3.5.1 Air Lead Level Data

Levels of lead-in-air data shown in Table 3.1 were studied in Krishnamoorthy et al. (2006). This dataset, collected by the National Institute of Occupational Safety and Health (NIOSH) at the Alma American Labs on February 23, 1989, contains 15 air lead levels from different spots within the facility. To evaluate the health risk for the staff, researchers needed to

Table 3.1: Levels of Lead in Air ($\mu g/m^3$)

200	120	15	7	8	6	48	61
380	80	29	1000	350	1400	110	

Table 3.2: Tolerance intervals for the levels of lead in air data

Model	ML Estimates $\widehat{\mu}$	ML Estimates $\widehat{\sigma}$	TI Control-Center	TI Control-Tails
Lognormal	4.3329	1.6805	$g_L = -2.37,\ g_U = 2.37$ TI = (1.42, 4087.48)	$g'_L = -2.61,\ g'_U = 2.61$ TI = (0.95 6118.09)

compute the ($\beta = 0.9, 1 - \alpha = 0.9$) two-sided TIs to describe the distribution of the air lead levels, based on for the complete data. Krishnamoorthy et al. (2006) suggested that the data can be described well by a lognormal distribution.

The ML estimates of lognormal distribution parameters are $\widehat{\mu} = 4.3329$ and $\widehat{\sigma} = 1.6805$. Following **Algorithm 1**, we simulate samples of size 15 from the standard normal distribution. The number of Monte-Carlo samples was chosen to be 100,000 to make simulation error negligible. Figure 3.2a is a contour plots of pairs of (g_L, g_U) that yield control-the-center TIs with content $\beta = 0.9$ and eight different levels of CP between 0.90 and 0.97. The equal-tail TI is illustrated at the confidence level 90%. The factors of an equal-tail TI to control the center are presented as the coordinates of the intersection point in Figure 3.2b. We calculate the coordinates of the intersection point of $\mathrm{CP}(g_L, g_U) = 0.9$ with the contour line $\mathrm{CP}_L(g_L) = \mathrm{CP}_U(g_U)$ as described in Section 3.4.1. Figure 3.3a and 3.3b are similar contour plots for control-both-tails TIs. The results of the computed factors and TIs are summarized in Table 3.2.

That is, by the control-the-center TI, we have 90% confidence that a proportion 0.9 of the air lead levels will fall between 1.42 and 4087.48 $\mu g/m^3$. The corresponding control-both-tails TI indicates that there is 90% chance that no more than a proportion 0.1 of the air lead levels will exceed 6118.09 or fall below 0.95 $\mu g/m^3$.

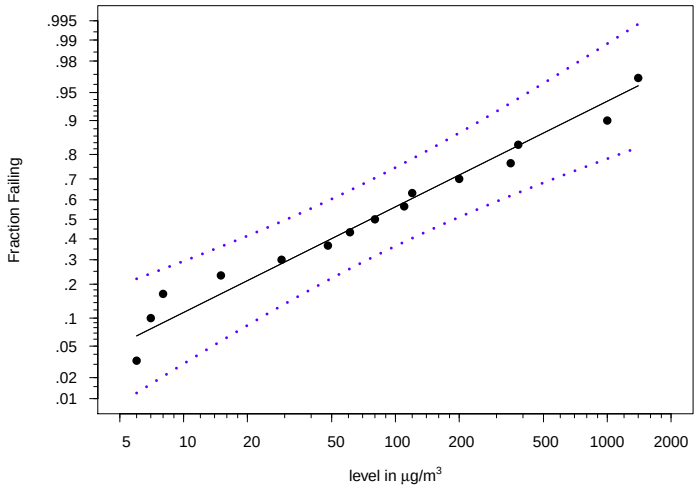
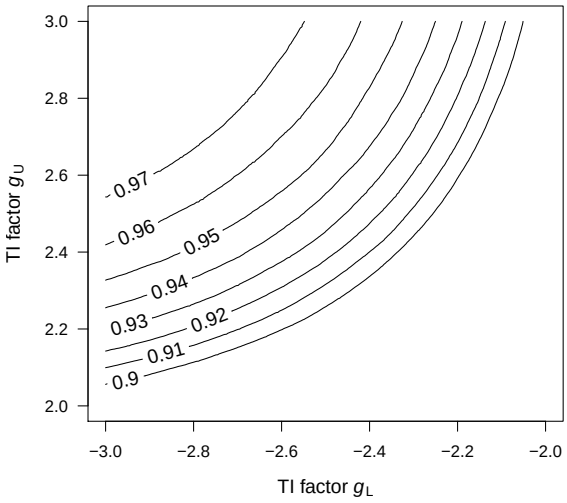
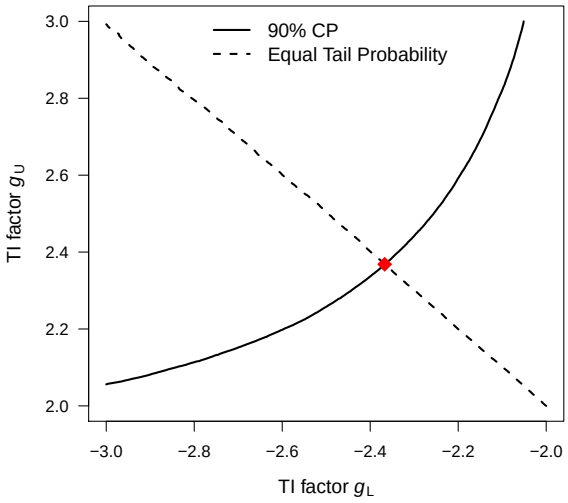


Figure 3.1: Lognormal probability plot for the air lead level data.

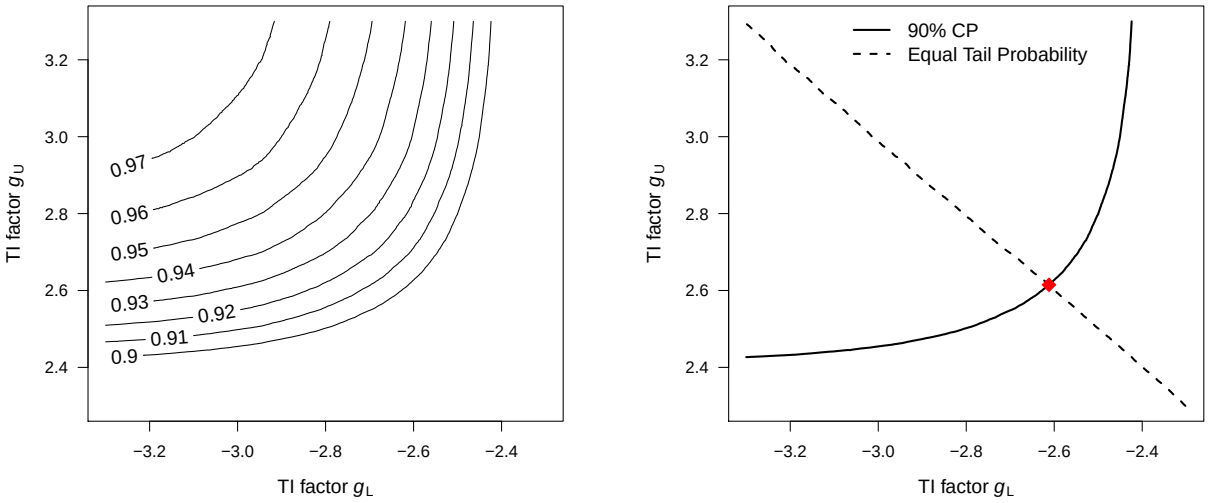


(a) Factors for TI with $\beta = 0.9$ and different CP



(b) Factors for $(\beta = 0.9, 1 - \alpha = 0.9)$ TI with equal tails

Figure 3.2: Contour plots of factors for control-the-center TI.



(a) Factors for TI with $\beta = 0.9$ and different CP

(b) Factors for $(\beta = 0.9, 1 - \alpha = 0.9)$ TI with equal tails

Figure 3.3: Contour plots of factors for control-both-tails TI.

Table 3.3: Failure times of pressure vessels in hours.

2.2	4.0	4.0	4.6	6.1	6.7	7.9	8.3
8.5	9.1	10.2	12.5	13.3	14.0	14.6	15.0

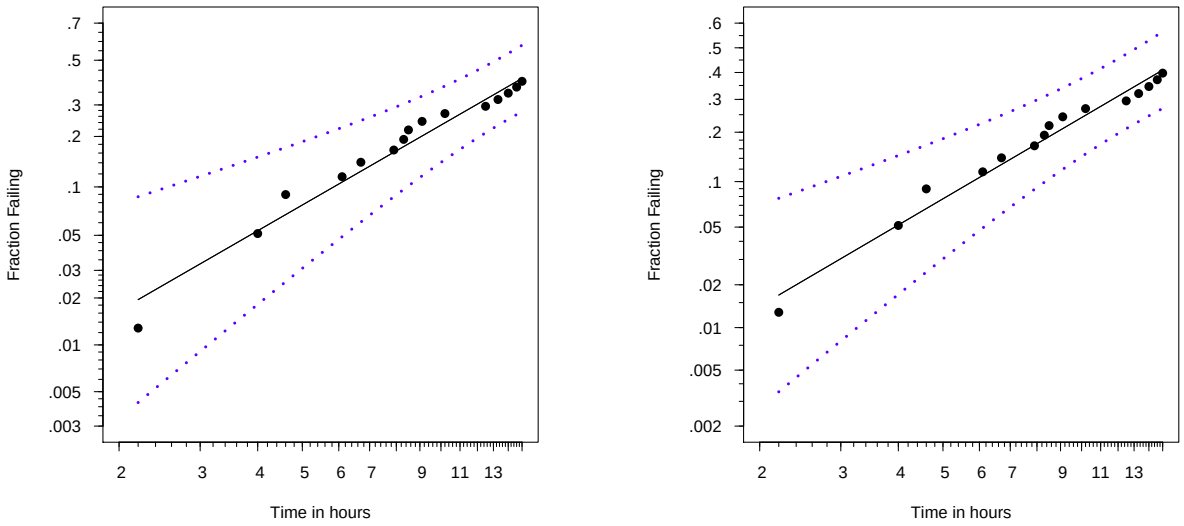
3.5.2 Pressure Vessel Failure Data

Failure times of 39 pressure vessels were given in Martz and Waller (1982) and used in Howlader and Weiss (1992). The data are shown in Table 3.3. We treat the largest $n-r = 23$ observations as right censored (i.e., we assume that the 39 pressure vessels were observed until there were $r = 16$ failures). Therefore the data are Type II censored. From Figure 3.4, we see that both the Weibull and the loglogistic distributions provide good description for the data. TIs under both distributions are computed to characterize the lifetime distribution of the pressure vessels so that one can see the sensitivity with respect to the specification of the distribution. For the content and confidence level $(\beta = 0.9, 1 - \alpha = 0.9)$, the computed factors and TIs are shown in Table 3.4.

Based on the TIs using the Weibull distribution, we have 90% confidence that at least

Table 3.4: Tolerance intervals for the pressure vessel failure data.

Model	ML Estimates $\widehat{\mu}$	$\widehat{\sigma}$	TI Control-Center	TI Control-Tails
Weibull	3.0796	0.5835	$g_L = -4.09,\ g_U = 2.19$ TI = (2.00, 77.98)	$g'_L = -4.38,\ g'_U = 2.45$ TI = (1.69 90.77)
Loglogistic	2.8979	0.5195	$g_L = -4.06,\ g_U = 4.78$ TI = (2.20, 217.44)	$g'_L = -4.33,\ g'_U = 5.21$ TI = (1.91 272.00)



(a) Weibull probability plot

(b) Loglogistic probability plot

Figure 3.4: Probability plots for the pressure vessel failure data.

a proportion 0.9 of the pressure vessels will have life times between 2 and 77.98 hours, and no more than a proportion 0.1 of the pressure vessels will fail within 1.69 hours or after 90.77 hours. The upper endpoints of the loglogistic distribution TIs are much larger due to extrapolation and the heavier tail of that distribution.

3.5.3 Locomotive Control Failure Data

Nelson (1982, page 324) gives the miles to failure (in units of 1000 miles) of 96 different locomotive controls. The data are shown in Table 3.5. Krishnamoorthy and Xie (2011) constructed two-sided TIs using both lognormal and loglogistic distributions to describe the

Table 3.5: Miles to failure of locomotive controls in units of thousands of miles.

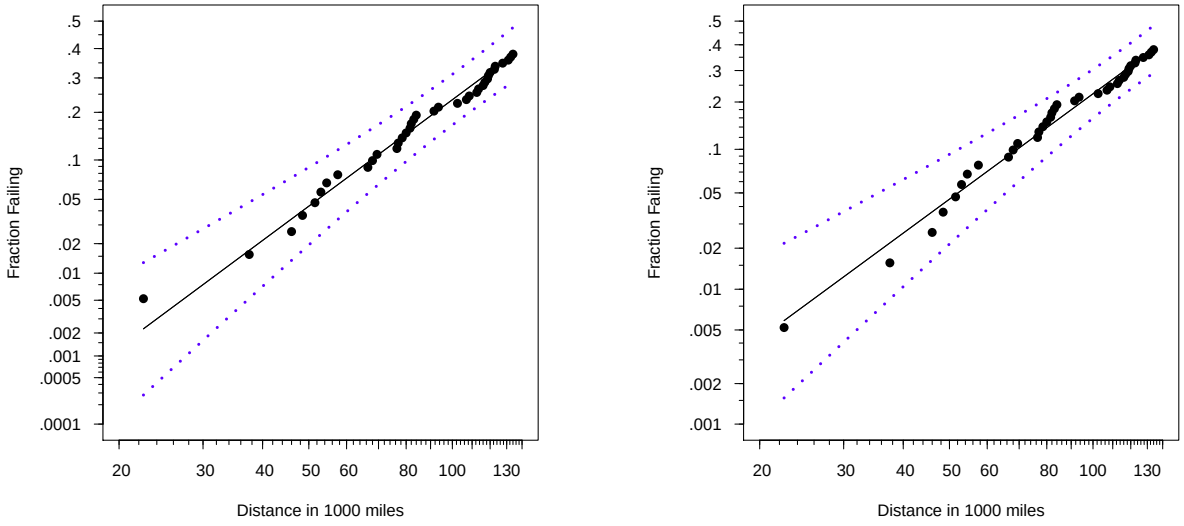
22.5	37.5	46.0	48.5	51.5	53.0	54.5	57.5	66.5	68.0
69.5	76.5	77.0	78.5	80.0	81.5	82.0	83.0	84.0	91.5
93.5	102.5	107.0	108.5	112.5	113.5	116.0	117.0	118.5	119.0
120.0	122.5	123.0	127.5	131.0	132.5	134.0			

Table 3.6: Tolerance intervals for the locomotive control failure data.

Model	ML Estimates		TI	TI
	$\widehat{\mu}$	$\widehat{\sigma}$	Control-Center	Control-Tails
Lognormal	5.1169	0.7055	$g_L = -1.90,\ g_U = 2.10$ TI = (43.67, 733.08)	$g'_L = -1.99,\ g'_U = 2.23$ TI = (41.05 804.38)
Loglogistic	5.0830	0.3837	$g_L = -3.50,\ g_U = 3.78$ TI = (42.02, 687.72)	$g'_L = -3.65,\ g'_U = 3.98$ TI = (39.72 743.84)

failure time distribution of these locomotive controls. The observation period ended at $x_c =$ 135 thousand miles, by which 37 locomotive controls had failed. Thus the miles to failure of the locomotive controls make a Type I censored dataset. A $(\beta = 0.9, 1-\alpha = 0.9)$ TI is needed to estimate the lifetime and assess the reliability of the controls. Because both lognormal and loglogistic distributions fit the data well, we compare TIs for these two distributions so that one can see the sensitivity with respect to the specification of the distribution. Results of the computed factors and TIs are presented in Table 3.6. The TIs based on the lognormal distribution are more conservative.

From the calculated TI using the lognormal distribution, we have 90% confidence that at least a proportion 0.9 of these locomotive controls will have life times between 43.67 and 733.08 thousand miles. Also we can be 90% confident that no more than a proportion 0.1 of these locomotive controls will run less than 41.05 or more than 804.38 thousand miles.



(a) Lognormal probability plot

(b) Loglogistic probability plot

Figure 3.5: Probability plots for the locomotive control failure data.

3.6 Simulations for Type I Censoring

Because the TIs with Type I censored data are not exact, this section presents the results of a simulation to evaluate the actual CP. The actual CP depends on the unknown parameters through the expected number of uncensored observations. As the expected number of uncensored observations increases to infinity, the ML estimates approach the true parameter values and the actual CP will approach the nominal confidence level. Thus, to evaluate finite sample properties of the TI procedure, we examine the actual CP of TIs as a function of the expected number of uncensored observations.

We consider both control-the-center and control-both-tails TI for the Weibull distribution with $(\mu = 0, \sigma = 1)$. Denote the expected number and expected proportion of uncensored observations as k_f and p_f, respectively. Then, the procedure of the simulation study for the control-the-center TI with equal tails is as follows:

1. Generate $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ from the standard SEV distribution with pre-specified censoring point x_c, which is the p_f lower quantile of the standard SEV distribution.

That is $x_c = \log[-\log(1-p_f)]$. The sample size n is determined by $n = k_f/p_f$. Then calculate ML estimates $(\widehat{\mu}, \widehat{\sigma})$.

2. For simulated data $\boldsymbol{x}$, use **Algorithm 1** to find pairs of (g_L, g_U) with CP equal to the nominal confidence level based on B_2 Monte-Carlo samples. Then identify the unique pair of factors (g_{L0}, g_{U0}) for equal-tail TI by solving (3.17). Details are given in Section 3.4.1.

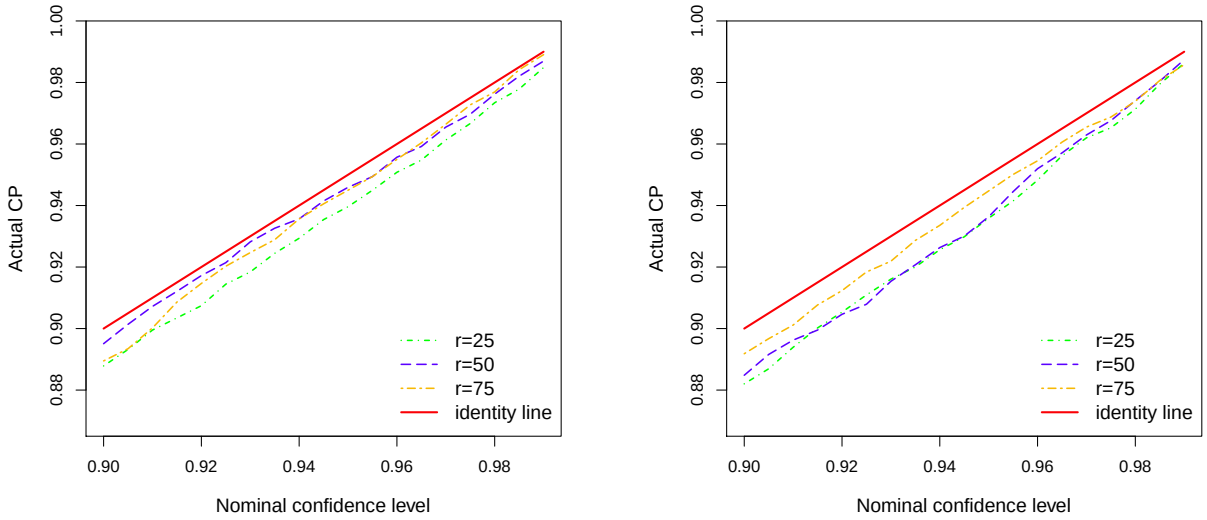
3. To calculate the CP, repeat steps 1 to 2 B_1 times. Denote the simulated data by $\boldsymbol{x}_j^*$, the corresponding ML estimates by $(\widehat{\mu}_j^*, \widehat{\sigma}_j^*)$ and the equal-tail factors by (g_{Lj}, g_{Uj}), $j = 1, \cdots, B_1$. Then calculate TI conditional on each $\boldsymbol{x}_j^*$ as $(\widehat{\mu}_j^* + g_{Lj}\widehat{\sigma}_j^*, \quad \widehat{\mu}_j^* + g_{Uj}\widehat{\sigma}_j^*)$. The CP based on B_1 samples from the standard SEV distribution can be obtained as,

$$\mathrm{CP}(\boldsymbol{\theta}) \approx \frac{1}{B_1}\sum_{j=1}^{B_1} \mathrm{I}\Big[\Phi\left(\widehat{\mu}_j^* + g_{Uj}\widehat{\sigma}_j^*\right) - \Phi\left(\widehat{\mu}_j^* + g_{Lj}\widehat{\sigma}_j^*\right) > \beta\Big].$$

The simulation procedure to study the control-both-tails TI is similar, it only differs in the calculation of the CP.

To see the performance **Algorithm 1** under type I censoring, we consider values of k_f equal to $25, 50, 75$. The expected proportion of uncensored observations p_f is specified as 0.25. Correspondingly, the sample size n takes the values $100, 200, 300$ for both types of TI. The number of type I simulations B_1 is 4000.

The CP compared to the nominal confidence levels are summarized in Figure 3.6. The lines of control-the-center and control-both-tails TI with $k_f = 75$ are close to the identity line. The actual CP is slightly off the nominal confidence level. Also the CP is closer to the nominal one when the nominal confidence level is larger for all of the values of the expected number of uncensored observations considered. In addition, as the expected number of uncensored observations increases, the CP tends to approach the nominal confidence level for both types of TI. Overall, Figure 3.6 displays satisfactory results when k_f is moderate



(a) Control-the-center TI with equal tails

(b) Control-both-tails TI with equal tails

Figure 3.6: Plots of the CP vs nominal confidence level for TIs with content 0.9, under Type I censored data.

to large. The plots on the two panels also show that the CP of control-both-tails TI tends to be lower than those of control-the-center TI when they have the same k_f and nominal confidence level.

3.7 Conclusions and Areas for Future Research

In this study, we developed a general method to compute control-the-center TIs and control-both-tails TIs for the (log)-location-scale family of distributions. Previous published work did not provide a procedure to calculate two-sided TIs for Weibull (SEV) distribution. When data are complete or Type II censored, our method provides exact factors. For Type I censoring, our method provides factors that give approximate TIs. In this case, the factors are asymptotically accurate. Our simulation study showed that under Type I censored data, the CP is close to the nominal confidence level when the expected number of uncensored observations is moderate to large.

The selection of an appropriate lifetime distribution for the data is an important step.

Knowledge of the data-generation mechanism can be useful for some applications. If the failure is due to fracture from fatigue in ductile materials, the lognormal distribution can be an appropriate model (see, for example, Chapter 11 of Meeker and Escobar 1998). If the failure is due to fracture from fatigue in brittle materials, the Weibull distribution usually provides a good model. In other situations, probability plots are widely used to select an appropriate distribution.

For the approximate TIs, the CP approaches its nominal confidence level as the expected number of uncensored observations increases. The CP might not be satisfactory when the expected number of uncensored observations is small. It might be possible to investigate alternative methods like bias correction or other methods with another layer of simulation that might provide improved CP.

Sometimes it is also of interest to find the shortest TI. For this purpose, we only need to find the factors on the curve $\mathrm{CP}(g_L, g_U) = 1 - \alpha$ with the absolute difference between g_U and g_L reaching the minimum. One could also develop approaches to compute TIs for distributions outside the (log)-location-scale family, for applications involving regression models, and tolerance regions for multivariate distributions that are often used to characterize the multidimensional output of a process.

Appendices

3.A Proof of Result 1

When a TI is to be computed for a univariate continuous distribution, such as the normal distribution or the Weibull distribution, the CP can be expressed as

$$\mathrm{CP}(\boldsymbol{\theta}) = \int \mathrm{I}\left\{F[\widetilde{T}(\boldsymbol{x}, \beta, 1-\alpha); \boldsymbol{\theta}] - F[\underset{\sim}{T}(\boldsymbol{x}, \beta, 1-\alpha); \boldsymbol{\theta}] > \beta\right\} f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x},$$

where $f(\boldsymbol{x}; \boldsymbol{\theta})$ is the joint density function of the data $\boldsymbol{X}$ and the integration is over the region of $\boldsymbol{x}$ values for which $f(\boldsymbol{x}; \boldsymbol{\theta}) > 0$. When the TI endpoints can be expressed as

functions of model parameter estimates $\widehat{\boldsymbol{\theta}}$, the CP can be computed as

$$\mathrm{CP}(\boldsymbol{\theta}) = \int \mathrm{I}\left\{F[\widetilde{T}(\widehat{\boldsymbol{\theta}}, \beta, 1-\alpha); \boldsymbol{\theta}] - F[\underset{\sim}{T}(\widehat{\boldsymbol{\theta}}, \beta, 1-\alpha); \boldsymbol{\theta}] > \beta\right\} h(\widehat{\boldsymbol{\theta}}; \boldsymbol{\theta}) d\widehat{\boldsymbol{\theta}},$$

where $h(\widehat{\boldsymbol{\theta}}; \boldsymbol{\theta})$ is the joint density function (sampling distribution) of the parameter estimator $\widehat{\boldsymbol{\theta}}$ and the integration is over the region of $\widehat{\boldsymbol{\theta}}$ values for which $h(\widehat{\boldsymbol{\theta}}; \boldsymbol{\theta}) > 0$.

Following from (3.7), the CP of a two-sided TI to control the center for the location-scale family of distributions can be expressed as

$$\begin{aligned}\mathrm{CP}(\boldsymbol{\theta}) &= \mathrm{E}_{\widehat{\mu},\widehat{\sigma}}\left(\mathrm{I}\left\{\Phi\left[\frac{\widetilde{T}(\widehat{\mu},\widehat{\sigma},\beta,1-\alpha)-\mu}{\sigma}\right] - \Phi\left[\frac{\underset{\sim}{T}(\widehat{\mu},\widehat{\sigma},\beta,1-\alpha)-\mu}{\sigma}\right] > \beta\right\}\right) \\ &= \mathrm{E}_{\widehat{\mu},\widehat{\sigma}}\left\{\mathrm{I}\left[\Phi\left(\frac{\widehat{\mu}+g_{U(1-\alpha;\beta,n)}\widehat{\sigma}-\mu}{\sigma}\right) - \Phi\left(\frac{\widehat{\mu}+g_{L(1-\alpha;\beta,n)}\widehat{\sigma}-\mu}{\sigma}\right) > \beta\right]\right\}, \qquad (19)\end{aligned}$$

where $\Phi(z)$ is the cdf of a particular standard location-scale distribution and the expectation is with respect to the joint distribution of the estimators $(\widehat{\mu}, \widehat{\sigma})$. Following (19),

$$\begin{aligned}\mathrm{CP}(\boldsymbol{\theta}) &= \mathrm{E}_{\widehat{\mu},\widehat{\sigma}}\left\{\mathrm{I}\left[\Phi\left(\frac{\widehat{\mu}-\mu}{\sigma}+g_{U(1-\alpha;\beta,n)}\frac{\widehat{\sigma}}{\sigma}\right) - \Phi\left(\frac{\widehat{\mu}-\mu}{\sigma}+g_{L(1-\alpha;\beta,n)}\frac{\widehat{\sigma}}{\sigma}\right) > \beta\right]\right\} \\ &= \mathrm{E}_{Z_1,Z_2}\left\{\mathrm{I}\left[\Phi\left(Z_1+g_{U(1-\alpha;\beta,n)}Z_2\right) - \Phi\left(Z_1+g_{L(1-\alpha;\beta,n)}Z_2\right) > \beta\right]\right\},\end{aligned}$$

where the expectation is with respect to the joint distribution of $Z_1 = (\widehat{\mu}-\mu)/\sigma$ and $Z_2 = \widehat{\sigma}/\sigma$.

3.B Proof of Result 2

From Lawless (2003, pp. 562-563), the quantities Z_1 and Z_2 for a location-scale family of distribution are pivotal because the number of uncensored observations r is fixed from sample to sample. When data are Type I censored, the exact pivotal properties do not hold. It can be proved, however, that when the expected proportion of uncensored observations p_f is fixed, Z_1 and Z_2 are asymptotically pivotal.

Let $F(x)$ be the cdf of the population and $F_n(x)$ denote its empirical distribution function. By the Glivenko-Cantelli Theorem,

$$\sup_{x\in\mathbb{R}}|F_n(x)-F(x)| \xrightarrow{\text{a.s.}} 0. \tag{20}$$

That is, the actual proportion of uncensored observations

$$\hat{p}_f=\sum_{i=1}^{n}[I(X_i\le x_c)]/n \xrightarrow{\text{a.s.}} p_f=F(x_c), \text{for all } x_c.$$

Thus, the actual number of uncensored observations $r=n\hat{p}_f$ is approximately equal to np_f, which is a fixed number.

In particular, for $\boldsymbol{x}=(x_1,\cdots,x_n)$ independently and identically distributed (i.i.d.) as $F(x;\mu,\sigma)$ and $\boldsymbol{x'}=(x'_1,\cdots,x'_n)$ i.i.d. as $F(x;\mu',\sigma')$ with $x'_i=ax''_i+b$, $\mu'=a\mu+b$ and $\sigma'=a\sigma$ (i.e., F is a member of the location-scale family of distributions), when they have the same expected proportion of uncensored observations p_f,

$$\begin{aligned}
L_{\boldsymbol{x}}(\mu,\sigma)=&\frac{1}{\sigma^r}\left[\prod_{i=1}^{r_1}\phi\left(\frac{x_i-\mu}{\sigma}\right)\right]\left[1-\Phi\left(\frac{x_c-\mu}{\sigma}\right)\right]^{n-r_1},\\
L_{\boldsymbol{x}'}(\mu',\sigma')=&\frac{1}{(a\sigma)^r}\left[\prod_{i=1}^{r_2}\phi\left(\frac{x'_i-\mu'}{\sigma'}\right)\right]\left[1-\Phi\left(\frac{ax_c+b-\mu'}{\sigma'}\right)\right]^{n-r_2}\\
=&\frac{1}{(a\sigma)^r}\left[\prod_{i=1}^{r_2}\phi\left(\frac{x''_i-\mu}{\sigma}\right)\right]\left[1-\Phi\left(\frac{x_c-\mu}{\sigma}\right)\right]^{n-r_2}\\
=&\frac{1}{(a\sigma)^r}\left[\prod_{i=1}^{r_1}\phi\left(\frac{x''_i-\mu}{\sigma}\right)\right]\left[1-\Phi\left(\frac{x_c-\mu}{\sigma}\right)\right]^{n-r_1}\left[\prod_{i=r_1+1}^{r_1+\delta}\phi\left(\frac{x''_i-\mu}{\sigma}\right)\right]\\
&\Big/\left[1-\Phi\left(\frac{x_c-\mu}{\sigma}\right)\right]^{\delta},
\end{aligned}$$

where $r_1=\sum_{i=1}^{n}I(x_i\le x_c)$, $r_2=\sum_{i=1}^{n}I(x'_i\le ax_c+b)$ are the number of exact observations in $\boldsymbol{x}$ and $\boldsymbol{x'}$, respectively. Set $\mid r_2-r_1\mid=\delta$, and ϕ, Φ to be the pdf and cdf of the corresponding standard distribution.

Based on (20), as $n \to \infty$, for a fixed p_f, we have

$$\sup_{x_c \in \mathbb{R}} |r_1/n - p_f| \xrightarrow{\text{a.s.}} 0, \sup_{x_c \in \mathbb{R}} |r_2/n - p_f| \xrightarrow{\text{a.s.}} 0. \text{ Therefore, } \sup_{x_c \in \mathbb{R}} \delta/n \xrightarrow{\text{a.s.}} 0.$$

Hence,

$$L_{\boldsymbol{x}'}(a\mu + b, a\sigma) \xrightarrow{\text{a.s.}} \frac{1}{a^r} \left\{ \frac{1}{\sigma^r} \left[\prod_{i=1}^{r_1} \phi\left(\frac{x_i'' - \mu}{\sigma} \right) \right] \left[1 - \Phi\left(\frac{x_c - \mu}{\sigma} \right) \right]^{n - r_1} \right\} \text{ for all } x_c, \quad (21)$$

where $\boldsymbol{x''} = (x_1'', \cdots, x_n'')$ is an independent sample from the same distribution as $\boldsymbol{x}$. Based on (21), $\hat{\mu}' \approx a\hat{\mu} + b$ and $\hat{\sigma}' \approx a\hat{\sigma}$ (i.e., ML estimators $\hat{\mu}$ and $\hat{\sigma}$ from Type I censored data are approximately equivariant) when n is large enough. Therefore, quantities Z_1 and Z_2 in **Result 2** are approximately pivotal under Type I censored samples of large size.

Bibliography

M. L. Bergquist. *Caution Using Bootstrap Tolerance Limits with Application to Dissolution Specification Limits*. PhD thesis, North Carolina State University, 2006.

K. R. Eberhardt, R. W. Mee, and C. P. Reeve. Computing factors for exact two-sided tolerance limits for a normal distribution. *Communications in Statistics - Simulation and Computation*, 18:397–413, 1989.

G. J. Hahn and W. Q. Meeker. *Statistical Intervals - A Guide for Practitioners*. John Wiley & Sons, Inc., New York, 1991. ISBN 0-471-88769-2.

W. G. Howe. Two-sided tolerance limits for normal populations - some improvements. *Journal of the American Statistical Association*, 64:610–620, 1969.

H. A. Howlader and G. Weiss. Log-logistic survival estimation based on failure-censored data. *Journal of Applied Statistics*, 19:231240, 1992.

K. Krishnamoorthy and T. Mathew. *Statistical Tolerance Regions - Theory, Applications, and Computation*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2009. ISBN 978-0-470-38026-0.

K. Krishnamoorthy and F. Xie. Tolerance intervals for symmetric location-scale families based on uncensored or censored samples. *Journal of Statistical Planning and Inference*, 141:1170–1182, 2011.

K. Krishnamoorthy, T. Mathew, and G. Ramachandran. Generalized p-values and confidence intervals: a novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene*, 3:642–650, 2006.

J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, 2003. ISBN 0-471-37215-3.

H. F. Martz and R. A. Waller. *Bayesian Reliability Analysis*. John Wiley & Sons, Inc., New York, 1982. ISBN 0-471-86425-0.

W. Q. Meeker and L. A. Escobar. *Statistical Methods for Reliability Data*. John Wiley & Sons, 1998. ISBN 0-471-14328-6.

W. Nelson. *Applied Life Data Analysis*. John Wiley& Sons, Inc., New York, 1982.

R. E. Odeh and D. B. Owen. *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. Marcel Dekker, Inc., New York, 1980. ISBN 0-8247-6944-9.

J. Shyu and D. B. Owen. One-sided tolerance intervals for the two-parameter double exponential distribution. *Communications in Statistics - Simulation and Computation*, 15: 101–119, 1986a.

J. Shyu and D. B. Owen. Two-sided tolerance intervals for the two-parameter double exponential distribution. *Communications in Statistics - Simulation and Computation*, 15: 479–495, 1986b.

A. Weissberg and G. H. Beatty. Tables of tolerance-limit factors for normal distributions. *Technometrics*, pages 483–500, 1960.

Y. Xie, Y. Hong, W. Q. Meeker, and L. A. Escobar. Simultaneous prediction intervals for the (log)-location-scale family of distributions. *Statistics Preprints*, 2014. Paper 128.

Chapter 4 Conclusions and Future Work

4.1 Conclusions

In this dissertation, we develop methods of predicting the default probabilities for individual companies and number of corporate defaults for the market based on large-scale event-time data and covariate data. We also propose a calibrated PI procedure to evaluate the uncertainties within predictions. Besides, we extend the computation of two-sided TIs for only symmetric distributions to any distribution in the location-scale and log-location-scale families.

In Chapter 2, we have three main achievements. The first is to build and estimate the time-to-event model. We apply a competing risks model for default and other types of exits, model the intensities of default and other exits with selected covariates, and obtain the ML estimates by numerically optimizing the log likelihood function. The second is modeling simultaneously two macroeconomic and 2,294 individual-level covariate processes spanning 228 months. In particular, we build a Markovian time series for the differenced covariate processes, apply a DFM with the optimal number of factors on the residuals of the Markovian time series to capture their correlation structure. We derive the EM algorithm to estimate parameters of the DFM in explicit forms. Specifically, we apply Kalman fitering and smoothing to estimate the conditional moments of the latent factors, and use them to replace the latent parts of the loglikelihood in the E-step. Our estimation algorithm is applicable when large amount of missing data of arbitrary pattern exist in the covariate processes. We also successfully address the identification problems in estimating the DFM by adding necessary constraints, and iteratively update ML estimates of the Markovian time series and the DFM until convergence. For our data, this "two-parts" iterative procedure converges very well in 100 iterations. Third, regarding prediction for the individual default probabilities

and the number of defaults in the market, we do not only provide point predictions based on large amount of simulated future covariate processes, but also develop a calibrated PI procedure to synthetically incorporate the variabilities in the ML estimates of time-to-event model and the covariate model, as well as randomness in the future covariate dynamics. The widths of such PIs are validated to successfully reveal the amount of uncertainties associated with predictions.

Generally, our covariate model and parameter estimation procedure can be applied to financial data of high dimensions collected over time. And our calibrated PI procedure can be used in other situations where uncertainties of some random quantities come from multiple sources.

In Chapter 3, we develop a general algorithm to calculate the exact factors of two-sided TIs when complete or type II censored data follow a distribution in the (log)-location-scale family. The algorithms for both control-the-center and control-both-tails TIs are provided. To obtain a unique pair of factors, we can compute the two types of TI with equal-tail probabilities. For type I censored data, we propose an adjusted procedure to calculate the approximate factors of the two types of TIs. The simulation study for Type I censored data shows that our algorithm can give TIs with the observed CP close to the nominal ones when the expected number of failures is moderate to large. Essentially, our methods are based on the pivotal properties of ML estimators of parameters for the location-scale family of distributions and the Monte-Carlo simulations. Our methods can be directly applied on log transformed data from the log-location-scale family of distributions.

The method for complete and Type II censored data can be applied to other types of data as long as the exact pivotal properties hold. Also, the method for Type I censored data can be applied when the pivotal properties are only approximate.

4.2 Future Directions

For the default prediction project, we can improve our work in three directions corresponding to the three achievements. First, in the time-to-event model, we can apply a more flexible class of intensity functions (e.g., nonparametric intensity functions) or conduct formal selection for the intensity functions and the variables they include. It may also be beneficial to apply a frailty model to better capture the clustered defaults of individual companies. Second, in the covariate model, we can try more flexible model specification, e.g., to allow parameters in the DFM to change over time or select the order of time dependence, to make abrupt changes in the default trend possible. Third, for quantifying uncertainties in predictions, it is promising to study the theoretical properties of the calibrated PI procedure with the complicated covariate model.

For the TI paper, our method is essentially intended to quantify the variabilities in ML estimates through the Monte-Carlo simulations. Since only approximate pivotal properties hold for type I censored data, the performance of our methods in this case depends on the accuracy of ML estimates through the expected number of failures. When the expected number of exact observations is small, it is desirable to develop effective bias correction or other methods. In the future, it is also of interest to develop general methods to compute TIs for other univariate distributions, applications involving regression models, and tolerance regions for the commonly used multivariate distributions.