Summary of Doctoral Thesis: Bayesian Analysis of Duration Data

Tomoki Toyabe

February 19, 2025

Contents

1 Background

1.1 Tick Data Analysis

Before the 1980s, the shortest interval data available in finance was daily data. With advances in computing, exchanges began recording all trades with timestamps, prices, and volumes, leading to tick data analysis. The former half of this doctral thesis studies intraday seasonality, where trading activity typically follows a U-shaped pattern—high after market open, decreasing midday, and rising again before close. To model this, the dissertation extends the SCD model and integrates a Markov Chain Monte Carlo (MCMC) approach for estimation.

The SCD model, a non-Gaussian state-space model, treats transaction intervals as random variables influenced by market factors. The study applies a novel estimation technique using a hybrid Gibbs/Metropolis-Hastings sampling scheme.

1.2 Survival Analysis

In ACD and SCD models, duration refers to the time interval between consecutive trades, whereas in survival analysis, it represents the time until an event occurs. Survival analysis is widely used in economics to study unemployment periods and firm survival durations.

Censoring is a common issue in survival analysis, occurring when observed data falls below a threshold, leading to estimation bias. Right censoring is particularly problematic as it results in missing true event times when they exceed the observation period. Various methods address this issue, including parametric models that assume distributions such as Weibull, exponential, and normal, as well as semi-parametric models like the Cox proportional hazards model.

Positive-unlabeled (PU) learning is a machine learning approach that learns from labeled positive data and unlabeled data. Unlike traditional classification, which relies on both positive and negative examples, PU learning enables the use of previously unutilized unlabeled data. There are two main approaches to PU learning. The case-control scenario involves separately extracting positive and unlabeled samples, similar to non-random sampling in statistics. In contrast, the single-training-set scenario generates a dataset once, where only a portion of the positive examples is labeled while the rest remain unlabeled. This distinction plays a crucial role in defining the labeling mechanism in PU learning models.

The latter part of this paper proposes a method for integrating PU learning into survival analysis, a relatively unexplored area in existing research. Few studies have applied PU structures to survival analysis, but some alternative approaches have been explored. One study attempted to compensate for the lack of labeled survival data by assigning pseudo-labels to unlabeled patient data. Another study framed time-to-event prediction as a multi-objective regression problem, treating censored observations as partially labeled data and using a random forest model for estimation.

Despite their relevance, these approaches differ from traditional PU learning estimation frameworks, which rely on specific assumptions about the relationship between positive and unlabeled data. This paper aims to bridge this gap by developing a new methodology for PU survival analysis based on these foundational principles.

2 Bayesian Simultaneous Estimation of the SCD Model with B-Splines for Intraday Seasonality

This chapter presents a Bayesian simultaneous estimation method for capturing intraday seasonality in stock tick data. The method extends the Stochastic Conditional Duration (SCD) model by incorporating B-splines, allowing for a more flexible estimation of intraday patterns.

2.1 Stochastic Conditional Duration Model

The SCD model, introduced by Bauwens (2004), is a non-Gaussian state-space model that represents observed durations as a product of a latent variable and a positive random component. The latent variable follows a stationary AR(1) process, and both the latent variable and the error term are mutually independent. The error term in the SCD model is commonly assumed to follow either a Gamma or Weibull distribution. This assumption allows the model to capture variations in financial transaction durations effectively.

To enhance the flexibility of the SCD model, B-splines are introduced for modeling intraday variations. B-splines provide a smooth and data-driven approach, enabling the model to better capture time-dependent fluctuations in trading activity.

2.2 Estimation Method

The Stochastic Conditional Duration (SCD) model is a non-linear, non-Gaussian state-space model consisting of a measurement equation and a transition equation with state variables.

The transition equation follows a stationary AR(1) process, where the joint probability distribution is determined by a tridiagonal positive definite matrix under specific conditions.

In Bayesian inference, prior distributions are assigned to key parameters, including the latent variables and model coefficients. The joint posterior density function incorporates the likelihood function based on the Weibull or Gamma distribution, the state-space model, and the prior distributions. However, due to the intractability of the joint posterior distribution, an MCMC (Markov Chain Monte Carlo) method is applied. The MCMC method generates random samples from the posterior distribution, enabling the numerical evaluation of posterior statistics necessary for Bayesian inference through Monte Carlo integration.

2.3 Empirical Applications

The Stochastic Conditional Duration (SCD) model is estimated using tick data from Toyota Motor Corporation (TMC) in Aichi, Asahi Group Holdings, Ltd. (AGH) in Tokyo, and Japan Airlines Co., Ltd. (JAL) in Tokyo. The dataset is sourced from Nikkei NEEDS and includes three specific trading dates: January 4, 2016, June 23, 2016, and July 27, 2016. These dates are selected for their economic significance, as January 4 marks the first trading day of the year, June 23 is the day before the Brexit vote, and July 27 is the day before the Bank of Japan's policy meeting.

The companies are listed on the Tokyo Stock Exchange (TSE), where the daily trading session is divided into a morning session from 9:00 to 11:30 and an afternoon session from 12:30 to 15:00. For this study, only the morning session is analyzed, with the time period normalized so that 9:00 corresponds to zero and 11:30 marks the end of the observed timeframe.

The empirical analysis examines the duration between trades for these stocks during significant market events to assess the performance of the SCD model in capturing intraday seasonality and market microstructure effects. By focusing on high-frequency trading data, the study evaluates the model's capability to explain variations in trade durations and its potential application in financial market analysis.

3 Bayesian Simultaneous Estimation Capturing Intraday and Intra-Deferred Future Seasonalities

In this section, we focus on the tendency for deferred futures to have higher trading volume and the problems with connecting futures data of different contract months as deferred futures. Japanese gold futures contracts are listed simultaneously in six contract months with different maturities up to about one year, and when the nearer contract expires about every two months, a new contract month is listed. Naturally, the same "deferred" futures contract will have a due date about one year in the future immediately after listing, but just before the next contract month's futures contract is listed, the due date will be about 10 months away. Since the reason for the higher trading volume in the deferred futures is thought to be due to the length of time to the due date, it is natural to assume that changes in the due date within the deferred futures affect the frequency of transactions. This pattern, or in other words, the seasonality observed in deferred futures, is described using the Bernstein polynomial, similar to the previously mentioned intraday seasonality, and is estimated alongside the aforementioned autocorrelation within the SCD model.

3.1 Stochastic Conditional Duration Model For Commodity Tick Data

To enhance the model's ability to capture market dynamics, the number of regression coefficients from the previous model has been increased. This modification allows the model to separately represent two key seasonal patterns: intraday seasonality, which reflects systematic variations in trading activity within a single trading session, and contract-month seasonality, which accounts for fluctuations in trading behavior across different contract maturities. By treating these seasonal effects independently, the model can more accurately describe how trade durations evolve over time.

In addition to incorporating these seasonal effects, the model has been extended to include a broader set of explanatory variables. Along with transaction-specific data such as bid-ask spreads and trading volume, the influence of news events has been introduced as a new factor affecting trade durations. The inclusion of news effects acknowledges the role of external information in shaping market activity, as traders adjust their behavior in response to new developments.

By integrating these enhancements, the stochastic conditional duration model provides a more comprehensive framework for understanding transaction timing in commodity markets. This refined approach allows for a deeper analysis of how different factors—ranging from intraday trading patterns and contract-month variations to liquidity measures and external information—interact to influence market microstructure.

3.2 Estimation Method

A key aspect of the estimation process in this chapter is the use of Ancillarity-Sufficiency Interweaving Strategy (ASIS), which improves the efficiency of MCMC sampling. ASIS is designed to address the slow mixing of MCMC chains that can occur in Bayesian models by strategically interweaving two parameterizations of the latent states. This approach enhances the convergence speed and reduces autocorrelation between samples, leading to more reliable inference.

In the context of the SCD model, ASIS is particularly useful for handling the dependence structure within the state variables. By alternating between different reparameterizations, the method ensures that the posterior distribution is explored more effectively.

3.3 Empirical Results

The empirical analysis in this study is based on tick-by-tick data for gold futures, covering the period from September 2022 to June 2023. The data was obtained from the Tokyo Commodity Exchange (TOCOM), and the analysis focuses specifically on transactions that took place during the daytime trading session, which runs from 8:45 AM to 3:15 PM. A key reason for selecting this period is that it includes holiday trading data, which was introduced in September 2022. The study examines the trading volume across different contract months, revealing an important contrast between Japanese and U.S. futures markets. While U.S. futures typically exhibit higher trading volumes for near-term contracts, the results indicate that, in Japan, trading volumes tend to be greater for deferred futures. This finding is consistent with previous research on Japanese commodity futures.

To analyze the duration between trades and the underlying market dynamics, the SCD model is applied to the dataset. The model incorporates both intraday seasonality and contract-month variations, allowing for a detailed examination of the factors influencing trade durations. By including additional explanatory variables such as bid-ask spreads, trading volume, and external influences like news events, the analysis aims to provide a comprehensive understanding of transaction timing.

The results reveal several important insights. The study confirms that trade durations exhibit clear intraday seasonal patterns, with shorter intervals during peak trading hours and longer durations at the market's opening and closing times. The contract-month effect is also found to be a significant factor, as different contract maturities display distinct trading behaviors. The inclusion of news effects in the model suggests that unexpected announcements tend to shorten trade durations, indicating that traders react quickly to new information, leading to increased market activity. Moreover, the findings show that bid-ask spreads and trading volume have a substantial impact on transaction durations, reinforcing the idea that liquidity conditions influence the frequency of trades.

4 Positive-Unlabeled Survival Analysis (PU Learning)

4.1 Model

Traditional survival analysis assumes that when censoring occurs, the event time is always greater than the censoring time. However, this study takes a different approach by assuming that the event time remains unobservable in the case of censoring, regardless of whether it is greater or smaller than the censoring time.

This assumption aligns with real-world situations where the exact event time is unknown, leading to what the study refers to as unlabeled data. To illustrate this concept, a scenario is considered in which a company attempts to make sales to customers from a pre-existing customer list. Each customer is represented by an identifier, and the objective is to determine whether a sale was successfully made. However, the outcome may remain partially unobservable, making it difficult to infer precise patterns.

By introducing this alternative framework, the study challenges conventional survival analysis models and provides a more realistic way to handle censored data in practical applications.

4.2 Likelihood

This section introduces the likelihood function for the model described previously. It begins by considering the case where the censoring time is observed and explains how the likelihood function is constructed based on this information.

The likelihood function is formulated using probability density functions that represent the distributions of both the event time and the censoring time. These distributions are parameterized, and the likelihood is expressed as a function of these parameters. The model distinguishes between two cases: when the event is observed and when it is censored. The likelihood function incorporates both of these situations by weighting the probability of each case accordingly.

To refine the estimation process, assumptions about the relationship between the event time and the censoring time are introduced. The likelihood function is then rewritten to reflect these assumptions, leading to a formulation that considers a truncated distribution in cases where the event time is censored. This approach ensures that the model accurately represents the underlying data-generating process while incorporating the uncertainty introduced by censoring.

By constructing the likelihood function in this way, the study provides a framework for estimating model parameters effectively, even when event times are only partially observed.

This formulation is essential for applying the proposed survival analysis model to real-world datasets, where incomplete observations are common.

4.3 Extension to Nonparametric Bayesian Model

To allow for greater flexibility in modeling the distribution of survival times, an extension of the survival model to a nonparametric Bayesian framework is introduced. Instead of assuming a fixed parametric form for the probability density function of survival time, the model employs a Dirichlet Process Gamma Mixture Model, which can accommodate more general and complex distributions.

The Dirichlet process is used as a prior distribution, allowing the survival time distribution to be represented as an infinite mixture of components. Each component in the mixture is governed by a set of parameters drawn from a baseline distribution. The mixture components are weighted according to a stick-breaking process, where the weights follow a Beta distribution. This hierarchical structure ensures that the model can adaptively determine the number of mixture components needed to best fit the data.

By using this nonparametric Bayesian approach, the model gains the flexibility to capture more complex survival time distributions that may not be well-described by traditional parametric models. This extension is particularly useful when dealing with heterogeneous data or unobserved latent structures in survival analysis. The ability to estimate a flexible distribution makes this approach more robust and applicable to real-world data where assumptions about a specific parametric form may not hold.

4.4 Simulation

This section conducts simulations to verify the correctness of the likelihood functions proposed. The simulations involve generating synthetic data, estimating model parameters using different methods, and analyzing the results to assess the model's performance.

The data generation process is first described, where the dataset consists of variables representing survival times, censoring times, censoring indicators, and covariates. The survival and censoring times are drawn from an exponential distribution, while the covariates are generated from a multivariate normal distribution with two dimensions.

Once the data is generated, the study applies parametric estimation methods to fit the model and evaluate its effectiveness. The estimation process involves determining the parameters of the survival and censoring distributions, along with their relationships to covariates. The results are then analyzed to interpret how well the model captures the true underlying patterns in the simulated data.

By comparing the estimated parameters with the true values used in data generation, the study assesses the model's accuracy. The findings confirm that the proposed likelihood functions provide valid estimations, supporting the theoretical framework developed earlier. This simulation study strengthens the reliability of the model and demonstrates its practical applicability in survival analysis scenarios where censoring plays a significant role.

5 Conclusions

5.1 SCD Model: Intraday Sesasonality

Chapter 2 proposed an extension of the SCD model for estimating the intraday seasonality and duration clustering simultaneously, while previous studies remove this intraday seasonality externally. The intraday seasonality is represented by a B-Spline, and its coefficients are simulated along with the model parameters in the MCMC loop. This has the advantage not only of allowing simultaneous estimation of the intraday seasonality, but also of facilitating the extension to the estimation of the effect of external information such as the limit order information. One limitation of the method in this chapter is that the number of knots and smoothing parameters must be selected based on an information criterion, which is computationally expensive. Future research should address these issues and reduce the computational cost. We also use the block-sampling for state variable and change a setting for the parameter in the Gamma distribution for a stable sampling.

In the empirical analysis, the proposed SCD model is applied to three days of microsecond tick-data for three stocks listed on the TSE stock market. The proposed method outperforms the existing method in the WAIC sense in many of the estimation results for the three stocks with different trading frequencies, consistent with previous studies. Since very few studies have incorporated the limit order book information into the SCD models, the use of the limit order book information and discussion of the corresponding parameter estimation results are also contributions of this paper. We find that the spread has a positive impact on the trading time interval for all patterns, while order size has a negative impact for about half of the patterns. Future studies will confirm the superiority of simultaneous estimation in forecasting with a wide range of data.

5.2 SCD Model: Intra-Deferred Future Seasonality

In Chapter 3, we conducted a Bayesian statistical analysis using the SCD model that considers patterns in the deferred futures of gold in Japan, incorporating intraday seasonality, limit order book information, and news effects. The results showed that the intraday seasonality exhibits an inverted U-shape, consistent with prior research, and the impacts of order book information and news effects are also naturally interpretable. However, the differences in the WAIC between models including and excluding seasonality and patterns within the deferred futures varied across expiration months, suggesting the need for further empirical experiments and research using more extensive data.

5.3 PU Survival Analysis

In Chapter 4, we summarize the results of the simulations performed in the previous section. It is clear from Table 5 that the model proposed here for the data containing s_i, which is shown to be uncertain, results in results that are closer to the true value than the existing models. The estimated value of θ_c in the Conventional PM when all c is observed is close to the true value, but it is not even higher than that of PUSA when c is partially absent in the mean square error. This is due to the fact that the likelihood function assumed in our model is for positive unlabeled data.

As a future prospect, we proposed a method based on maximum likelihood estimation for the likelihood function, but it is expected to be used to estimate the model by the MCMC method, which is a Bayesian estimation method. Although we have assumed exponential distributions for the models of survival time and censoring time, it is expected that we can estimate more complicated models such as Weibull distribution by integrating numerically.