Thesis Proposal: Marked Point Process Intensity Estimation on Trade Duration Process

Mingyu Tang
Department of Statistics
Carnegie Mellon University

Advisor: Dr. Mark Shervish

Abstract

The research is motivated by a desire to model the stochastic process of trades in a limit-order-book market. We model the trade process as a marked point process. We propose a semiparametric model for the conditional intensity function, attempting to capture the effect of the recent past in a nonparametric way and the effect of the more distant past using a parametric time series model. Our framework provides more flexibility than the most commonly used family of models, known as Autoregressive Conditional Duration (ACD), in terms of the shape of the density of durations and in the form of dependence across time. We also propose a way to incorporate the effect of additional explanatory variables (i.e. marks in the marked point process.) So far, we have shown that the framework works much better than the ACD family in the sense of both prediction log-likelihood and various diagnostic tests using data from the NYSE. We propose two main extensions to the existing framework. One is to model multiple point processes, specifically a bivariate version, to predict both the duration to the next event but also the side (buy or sell) of the next event. The other aspect is to take into account the effect of information that arrives between two consecutive events to update the conditional intensity before the next event. In general, the framework can be used both to estimate intensity functions in self-exciting point process, and to estimate a series of conditional densities for a random process.

1 Introduction

In today's financial world, most markets rely on a *limit order book* (LOB) to match buyers and sellers. High frequency traders and market makers rely on real-time access to the LOB in order to implement their trading algorithms. When a LOB is broadly available, markets become more transparent and efficient. With the development of infrastructure and data collection techniques, researchers, traders and regulators are greatly interested in LOB data nowadays.

There are three major types of orders placed in a LOB, limit orders, market orders and cancellations. A trader places a market order when he wants to buy or sell a certain number of shares at the current best available price. The market order will be matched with existing nonexecuted orders immediately, and this leads to an execution (also called a trade). A limit order is placed when a trader specifies a price and the number of shares he wants to buy or sell at that price or better. Such orders will remain in the LOB unexecuted until either the trader cancels them or they get executed. When a trader changes his mind about an unexecuted limit order, he can place a cancellation order which cancels all or part of the unexecuted portion. All of these orders are recorded chronologically in the LOB. By following the flow of orders over time, we can construct the complete history of executed and unexecuted orders at every time during each trading day. The rich and high quality of LOB data

provides opportunities for a variety of research topics. Gould, Porter, Williams, McDonald, Fenn and Howison (2012) [GPW$^+$12] provide a thorough survey of LOBs. Popular research topics include the construction of optimal order execution strategies in Obizhaeva and Wang (2005) [OAW05], the impact of orders on the rest of the market in Eisler et al. (2010) [EZBK10], and the empirical analysis on various features of the order flow in Cont (2000) [Con00]. Modelling parts of the order flow, such as the trade execution process and the price duration process is another rich research area. Poisson processes are employed to describe the order flows in Cont, Stoikov and Talreja (2010) [CST10]. Traditional methods for optimal order execution derive from econometric approaches, such as the perfect-rationality approach, zero-intelligence approach, agent-based models and so forth. On the other hand, with the development of machine learning in recent years a variety of new techniques have been employed in the analysis of LOBs. Examples include SVM, Lasso and Graphical models. Although people have a better understanding of LOBs than ever before, there are still many things that are not yet well understood. open questions, such as cross stock effects, cross exchange effects, hidden order effects, dark pool effects, etc.

Every single order in LOB carrys varying degrees of market information. For that reason, they can all potentially affect market prices. From the flow of these orders, we can identify those events in which we are interested. The sequences of these events naturally form various processes. For example, the market orders and executions form the trade process. The events that change the price form a price process. Similarly, we can define a volume process, a touch event process(touch event is an event that changes the best available price on either side.The best available price on the buy side is defined as the bid price, while the best available price on the sell side is defined as the ask price) and so forth. The times that elapse between two consecutive events are defined as durations. Thus, each of the above processes is an example of a duration process. Traders, regulators and researchers have shown great interest in various duration processes. The price duration process helps us understand the evolution of price. The volume duration process shows how the market demand changes over time. The trade duration process helps to understand how likely it is that an order will be executed in a certain time. Naturally, a better modeling of these processes can help high frequency traders to make better algorithms.

A common feature among all of these duration processes is self-excitation. This refers to the occurrence of an event increasing the likelihood of future occurrences. Self-exciting processes exhibit a clustering effect. In other words, long durations are likely to be followed by long durations and short durations are likely to be followed by short durations. In order to capture the self-exciting feature and predict how the duration evolves over time, self-exciting point process models have been introduced, including two main types of models. One type of model is the intensity model. The intensity function, which characterizes the instantaneous probability of an occurrence at each time, is modeled as a function of the history of the process. Bauwens and Hautsch (2006) [BH06] give a good survey of current methods of modeling intensity functions. The most common intensity model is the Hawkes process, introduced by Hawkes (1971) [Haw71], in which the conditional intensity function is modeled as a linear combination of the effects of all the past events. And the effect of a certain past event is modeled as a decaying function applied on the time elapsed from that particular event. Zhao's thesis (2010) [Zha10] proposed another intensity model, in which the conditional intensity function depends on the number of events in the most recent past time window of a fixed length. The other main type of model is direct modeling of durations between events. This type of model was first introduced by Engle and Russel (1997, 1998) [ER98]. Their popular autoregressive conditional duration (ACD) model assumes that expectated value of the next duration has an autoregressive structure so as to capture the self-exciting feature. There is rich literature extending the ACD model to other duration models. Pacurar (2006) [Pac06] has given a comprehensive survey on the development of ACD models. In particular, Bauwens and Veredas (2004) [BV04] proposed a stochastic conditional duration (SCD) model, in which they introduce a stochastic noise in the autoregressive formula for the mean duration. Also, a log-ACD model, a threshold ACD model, and a stochastic volatility duration model are all designed to improve the original ACD model in specific ways. Furthermore, Russell (1999) [Rus99] has developed an intensity model based on duration called the autoregressive conditional intensity (ACI) model.

Although the ACD family of models successfully capture the self-exciting features of the duration processes, Bauwens, Giot, Grammig and Veredas (2000) [BGGV00] has shown that none of the ACD models can pass the a model evaluation criterion based on the probability integral transform

theorem proposed by Diebold, Gunther and Tay (1998) [DGT98] in terms of trade duration process. This fact dramatically undermines the credibility of ACD family.

Among all the processes mentioned above, trade duration process, along with its attached information such as execution price, trade volume, trade duration and so forth, carries the most valuable market information and almost all the automatic trading algorithm will react to trades immediately. In particular, the duration of trades has great power in predicting the side of future trades, which the high frequency traders extremely care about. Therefore, In this proposal, we focus on the dynamics of trade flow and propose a semi parametric framework, combining nonparametric conditional density estimation and parametric time series models, to estimate the conditional intensity function of a duration model, especially for but not limited to the trade duration process. Since trades are irregularly time-spaced events, the trade duration process is typically characterized as a marked point process, where the trades are target events and the other associated features, including price, spread, trade side, and other features of the LOB are marks.

Another interesting problem we hope to address is to model a bivariate version of the duration process. High frequency traders are particularly interested in on which side the next trade would occur in order to decide whether to take a long or short position. And the side of next trade is known to be highly correlated with the recent trade durations. In particular, a trade with a short duration tends to be on the same side as the previous trade, while a long duration has nearly 50% chance of being on either side. And a lot of research have extended the intensity models and duration models to bivariate versions. Shek (2011) [She11] developed the bivariate Hawkes process. Bauwens and Giot (2002) [BH03] developed the asymmetric ACD model. And the ACI model by Russell (1999) [Rus99] is itself a multivariate intensity model. Bauwens and Giot (2006) [BG06] extended the ACI model by introducing a latent common intensity factor to the multivariate intensity model. We have already set up a parametric model for predicting the side of a trade after we predict the duration based on our univariate point process. We intend to extend our framework to a bivariate version to model the sides of trades along with durations. Another natural extension is multivariate point process. Such an extension is motivated by modeling the cross-stock effect and cross-exchange effect. We intend to 1) combine the SPIE duration model and Hawkes process model and 2) combine the SPIE duration model and copula to model the multivariate point process.

The rest of the proposal is organized in the following way. In section 2, we give a brief overview of our dataset. In section 3, we give a brief review of ACD models, their extensions and their drawbacks. In section 4, we introduce our framework for modeling the duration process. We propose a systematic way to incorporate additional explanatory variables. In section 5, we present experimental results on LOB data from the New York Stock Exchange and compare our framework with the ACD model. In section 6, model diagnostics are introduced. Section 7 gives discussion and conclusion. Section 8 is future work.

2 Data Description

In this section, we give a detailed description of the dataset that we use. The LOB dataset is extremely huge and complex. Table 1 shows a small piece of the original LOB dataset format. Each row corresponds to one order. The columns from left to right are respectively the time of the order (measured in millisecond from the midnight), the stock name, the price of that particular order, the existing number of shares sitting on the corresponding price after this order occurs, the size of the order, the existing number of active orders at that level after this order occurs, the side of the order (B for buy and S for sell) and the type of the order (O for new order, C for cancellation, E for trade, X for multiple orders and N for others).

Table 1: Limit order book data sample. Time is measured in millisecond after midnight. Depth is number of existing shares after the arrival of the order. Size is the number of shares the coming order carries. Type includes O(new limit order), E(market order), C(cancellation).

time	stock name	price	depth	size	number of existing orders	side	type
43026177	JPM	124.50	2000	300	5	S	C
43026179	JPM	124.42	400	100	3	B	C
43026180	JPM	124.43	100	100	1	B	O

Since we focus on the dynamics of trade flow in this proposal, we filtered out the market orders and produced a list of triples of (T_t, B_t, L_t) as in Table 2, where T_t is log of duration, L_t is side of trade (+1 for sell side and -1 for buy side), and B_t is the book pressure imbalance, defined as the log ratio of the number of sell-side shares at the ask price to number of buy-side shares at the bid price. We observe such a triple whenever a trade occurs. Although book pressure imbalance actually changes much more frequently than trades occur, for the time being we keep track of the book pressure imbalance only when a trade happens. We intend to address the problem of how the trade duration T_t evolves over time by introducing a new class of semi parametric methods to model the durations. Part of the model will be to incorporate the effects of marks (such as book pressure imbalance and side of trade.)

Table 2: Trade Flow Sample.

log duration	book pressure imbalance	label
3.91	-0.588	-1
6.82	-0.134	-1
6.10	-1.946	+1

3 Review of ACD model

In this section, we briefly review the ACD model, which was first proposed by Engle and Russell (1998), along with some of its variants. Let x_i denote the time elapsed between two consecutive events at times t_{i-1} and t_i, i.e. $x_i = t_i - t_{i-1}$. An ACD model attempts to capture the time dependence in the duration process by modeling the conditional expectation of the next duration given the past, i.e. $E(x_i|\mathcal{F}_{i-1})$, where $\mathcal{F}_{i-1}$ denotes the information available before time t_{i-1}. A common ACD model is:

$$x_i = \Psi_i \epsilon_i,$$

where

$$\Psi_i = \omega + \alpha\epsilon_{i-1} + \beta\Psi_{i-1},$$

$\{\epsilon_i\}_{i\geq 1}$ is a process of IID positive random variables with mean 1, $\omega > 0$, $\alpha > 0$ and $\beta > 0$ are parameters with $\alpha + \beta < 1$ (to allow the Ψ_i to have a common mean.) Thus, $E(x_i|\mathcal{F}_{i-1}) = \Psi_i$. The particular model specified above is called ACD(1,1) because of the introduction of one order lag of both ϵ_i and Ψ_i. The distribution of ϵ_i is assumed to be from a parametric family with a long tail. Common choices include, Gamma, Weibull and Burr families.

As an example, consider the family of Weibull(1, γ) distributions. Then the parameters can be estimated by maximum likelihood method. In a point process, the intensity function is equal to the hazard function, and the hazard function of ϵ_i, which is defined as density over survivor function, can be derived as:

$$h(\epsilon_i) = \gamma\epsilon_i^{\gamma-1}$$

The conditional hazard function of x_i can be obtained as

$$h(x_i|\mathcal{F}_{i-1}) = \frac{\gamma}{\Psi_i}(\frac{x_i}{\Psi_i})^{\gamma-1}$$

When $\gamma = 1$, the Weibull distribution is an exponential distribution. If $\gamma > 1$, the hazard function is increasing. If $\gamma < 1$, the hazard function is decreasing. The process will be self-exciting when the hazard is decreasing, that is $\gamma < 1$.

3.1 Additional Explanatory Variables in ACD

In a market microstructure dataset, such as our NYSE dataset, events are usually associated with some additional explanatory variables, such as volume, spread, price and so forth. These variables can be characterized as marks in the point process and they can have an impact on the intensity of events. In the original ACD models and its variants, the effects of additional explanatory variables are incorporated in the autoregressive formula, as in Equation 1:

$$\Psi_i = \omega + \alpha\epsilon_{i-1} + \beta\Psi_{i-1} + \delta^T u_i, \tag{1}$$

where, u_i is a vector of additional explanatory variables and δ^T is the corresponding coefficients. Such a specification indicates that the additional variables only affect durations through their means.

3.2 ACD Variants

Among all of the extensions of ACD models, the Log-ACD by Bauwens and Giot (2000) [BG00] is the most popular. The objective of such an extension is to relax the constraint of positivity of coefficients, especially when additional explanatory variables are incorporated. The model is specified as:

$$x_i = \Psi_i \epsilon_i, \tag{2}$$
$$\Psi_i = e^{\psi_i}, \tag{3}$$
$$\psi_i = \omega + \alpha \epsilon_{i-1} + \beta \psi_{i-1}. \tag{4}$$

The exponential transformation guarantees the positivity of conditional duration Ψ_i.

Another important extension is the threshold ACD (TACD) by Zhang, Russell and Tsay (1999) [ZRT01]. TACD allows more flexibility in the autoregressive part:

$$\psi_i = \begin{cases} \omega_1 + \alpha_1 \epsilon_{i-1} + \beta_1 \psi_{i-1} & , if\ 0 < x_{i-1} <= r_1 \\ \omega_2 + \alpha_2 \epsilon_{i-1} + \beta_2 \psi_{i-1} & , if\ r_1 < x_{i-1} <= r_2 \\ \omega_3 + \alpha_3 \epsilon_{i-1} + \beta_3 \psi_{i-1} & , if\ r_2 < x_{i-1} < \infty \end{cases}$$

Bauwens and Veredas (1999) [BV04] assumes that the durations are driven by a stochastic latent factor and introduced the stochastic conditional duration model (SCD). This stochastic latent variable is interpreted as dynamic and random market information. Thus, the autoregressive formula is modified to:

$$\psi_i = \omega + \beta \psi_{i-1} + u_i$$

3.3 Drawbacks of ACD Models

Bauwens, Giot, Grammig and Veredas (2000) [BGGV00] have shown that the ACD model and its variants do not perform well with trade duration data under the evaluation method proposed by Diebold, Gunther and Tay (1998)[DGT98]. There are two main reasons for this poor performance.

First, in ACD models and their variants, there is a strong parametric assumption on the distribution of the ϵ process, in particular being positive with a long-tailed distribution. Gamma and Weibull distributions are most commonly used, and both of these include exponential distributions as special cases. However, as we will show in the next section, the trade durations do not have such a ideal parametric distribution. Furthermore, the empirical distribution of log-durations appears to be bimodal, which undermines the performance of any parametric ACD model and its variants. Such a bimodal feature is the direct reason that makes the ACD model and its variants fail to pass the Kolmogorov Smirnov test in the model evaluation proposed by Diebold, Gunther and Tay (1998) [DGT98].

Another important restriction on ACD models and their variants is that the time dependency is incorporated only in the expectation of the duration. However, as we show in the rest of the proposal, we can see that the previous trade duration affects the ensuing trade duration in a more general way. In particular, it changes both the locations and the relative sizes of the two modes. Therefore, modeling the time dependency solely in terms of the mean duration cannot capture the more general effect of previous durations. Although some of the extensions are designed to overcome this shortcoming, they are still not flexible enough to capture the effects on the modes of the distribution.

4 Semiparametric Intensity Estimation (SPIE)

In this section, we propose a semi parametric model for the conditional intensity function of a point process or a sequence of conditional densities for durations. Suppose that we observe a series of occurrence times $\{t_i\}$ from a point process. By taking the differences of successive occurrence times, we get a series of durations $\{\Delta t_i\}$. Suppose that the conditional density function (given the past) for Δt_i is $p_i(t)$ with CDF $P_i(t)$. For the point process, the conditional intensity function λ is defined as:

$$\lambda(t|\mathcal{F}_t) = lim_{\Delta t \to 0} \frac{1}{\Delta t} P(\text{One event occurs in the time interval } [t, t+\Delta t] | \mathcal{F}_t)$$

In a duration-based point process, the estimated intensity $\hat{\lambda}_t$ is essentially the hazard function, which can be obtained by $\hat{haz}_i(.) = \hat{p}_i(.)/(1 - \hat{P}_i(.))$.

Since the duration has a long tail, we transform to the log scale in the rest of the proposal and use $\{T_i\}$ to denote the logarithms of the durations. We let $f_i(.)$ denotes the density of T_i, and $F_i(.)$ denotes its CDF. Naturally, $p_i(.)$ and $P_i(.)$ can be directly derived from $f_i(.)$ and $F_i(.)$ and vice-versa. Therefore, the goal of the following report is to estimate the density function $f_i(.)$ of T_i given $\{T_j\}_{j=1}^{i-1}$.

Our semi parametric framework is characterized in the following way:

1. Construct an initial nonparametric estimator $\hat{f}_i(.)$ for log-durations. Its corresponding CDF is $\hat{F}_i(.)$.
2. Calculate the *generalized residuals* $c_i = \hat{F}_i(T_i)$ and construct *transformed generalized residuals* $p_i = \Phi^{-1}(C_i)$.
3. Construct a parametric time series model for the transformed generalized residuals: $TSmodel(p_i|\{p_j\}_{j=1}^{i-1})$.
4. Estimate the predictive distribution of the transformed generalized residuals under the time series model: $P_i|\{p_j\}_{j=1}^{i-1} \sim N(\hat{\mu}_i, \hat{\sigma}_i^2)$.
5. Transform back to the distribution of log-durations and compute the adjusted estimated conditional density $\hat{f}_i^*$ of log-durations. That is, if P_i has the distribution in step 4, then $T_i = \hat{F}_i^{-1}(\Phi(P_i)) \sim \hat{f}_i^*(t)$, where T_i is the random variable of adjusted predictive distribution of log-duration.

In step 1, $\hat{F}_i(.)$ is any nonparametric estimator of the conditional CDF $\hat{F}_i(.|\mathcal{F}_{i-1})$ of T_i. Due to the curse of dimensionality and heavy computational demands, this first estimate will depend on only the most recent past. This attempts to capture the short-term memory nonparametrically. In principle, the marginal distribution of T_i or $\hat{F}_i(T_i|T_{i-1})$ could be a good starting point. In the example in section 5, the nonparametric estimator is a kernel conditional density estimator given the previous duration, $\hat{F}_i(T_i|T_{i-1})$. Thus, the c_i can be regarded as generalized residuals and are supported on $(0, 1)$. In step 2, Φ^{-1} is the standard normal quantile function. The transformed generalized residuals p_i should then be approximately standard normal random variables. In step 3, $TS-model$ is any parametric time series model that we can fit to the series $\{p_i\}_{i=1}^n$. Using the time series model, we compute the conditional predictive distribution of each transformed generalized residual P_i. Typically, an ARIMA model could be used in this step. However, with market microstructure data, especially for trade flow, long memory (slowly decaying autocorrelation function) is present. Thus, ARFIMA models might be a better choice. In the example in section 5, we use an ARFIMA(1,d,1) model, which provides a good fit. According to the time-series model, the predictive distribution of P_i is the normal distribution with mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ in step 4. Step 5 transforms to the predictive distribution of T_i back in the log-duration space. If P_i has the distribution in step 4, then $T_i = \hat{F}_i^{-1}(\Phi(P_i))$ has the estimated predictive distribution of log-durations given the past. The adjusted estimated density function of T_i is $\hat{f}_i^*(t)$.

4.1 Adjusted New PDF and CDF

In this section, we derive the adjusted estimated conditional density function $\hat{f}_i^*(t)$ of T_i. We start with the initial nonparametric density estimate $\hat{f}_i(t)$. The adjusted CDF is:

$$\hat{F}_i^*(t) = P(T_i <= t) = P(\hat{F}_i^{-1} \circ \Phi(P_i) <= t) = P(P_i <= \Phi^{-1} \circ \hat{F}_i(t))$$

$$\hat{f}_i^*(t) = \frac{\partial}{\partial t}\hat{F}_i^*(t) = \hat{h}_i(\Phi^{-1}(\hat{F}_i(t))) * \Phi^{-1\prime}(\hat{F}_i(t)) * \hat{f}_i(t) = \frac{\hat{h}_i(\Phi^{-1}(\hat{F}_i(t)))}{\phi(\Phi^{-1}(\hat{F}_i(t)))} * \hat{f}_i(t)$$

where, $\phi(.)$ is the standard normal density, $P_i \sim N(\hat{\mu}_i, \hat{\sigma}_i^2)$ and $\hat{h}_i^*(.)$ is density of P_i. The new density function is the product of the initial density and $\frac{\hat{h}_i(\Phi^{-1}(\hat{F}_i(t)))}{\phi(\Phi^{-1}(\hat{F}_i(t)))}$. If no time series model

is fit to the P_i in step 3, then $\hat{h}_i(.) = \phi(.)$. The adjusted predictive CDF can also be derived straightforwardly:

$$\begin{aligned}
\hat{F}_i^*(t) &= \int_0^t \hat{f}_i^*(s) = \int_0^t \frac{\hat{h}_i(\Phi^{-1}(\hat{F}_i(s)))}{\phi(\Phi^{-1}(\hat{F}_i(s)))} \hat{f}_i(s) ds \\
&= \int_0^{\hat{F}_i(t)} \frac{\hat{h}_i(\Phi^{-1}(s_1))}{\phi(\Phi^{-1}(s_1))} ds_1 \quad \text{, where } s_1 := \hat{F}_i(s)) \\
&= \int_{-\infty}^{\Phi^{-1}(\hat{F}_i(t))} \frac{\hat{h}_i(s_2)}{\phi(s_2)} \phi(s_2) ds_2 \quad \text{, where } s_2 := \Phi^{-1}(s_1) \\
&= \hat{H}_i(\Phi^{-1}(\hat{F}_i(t)))
\end{aligned}$$

Here $\hat{H}_i(.)$ is CDF corresponding to the density $\hat{h}_i(.)$. Accordingly, the adjusted generalized residual c_i^* is:

$$c_i^* = \hat{F}_i^*(T_i) = \hat{H}_i(\Phi^{-1}(\hat{F}_i(T_i)) = \hat{H}_i(\Phi^{-1}(c_i))$$

If the time series model does improve the estimation, the adjusted generalized residual c_i^* should be closer to having a uniform $U(0,1)$ distribution than does c_i, both marginally and conditional on any past information (prior to the occurrence of T_i.) Empirical results will be shown in the example in section 5.

The difference between the adjusted log-likelihood (after fitting the parametric time-series model) and the initial nonparametric log-likelihood is:

$$\begin{aligned}
ll_{adjusted} - ll_{initial} &= \sum \log(\hat{f}_i^*(T_i)) - \sum \log(\hat{f}_i(T_i)) \\
&= \sum \log(\hat{h}_i(\Phi^{-1}(\hat{F}_i(T_i)))) - \sum \log(\phi(\Phi^{-1}(\hat{F}_i(T_i)))) \\
&= \sum \log(\hat{h}_i(p_i)) - \sum \log(\phi(p_i)).
\end{aligned}$$

This indicates that, as long as the linear time-series model fits the transformed generalized residuals $\{p_i\}_{i=1}^n$ better than the IID standard normal model, it increases the log-likelihood in the original log-duration space. The increase in log-likelihood in the original log-duration space is the same as the increase in the transformed generalized residual space.

4.2 Additional Explanatory Variables

In a marked point process, marks are observed along with the target events. These marks, or additional explanatory variables, may have an impact on the distribution of log-duration. In the ACD models, the marks' information is incorporated in the autoregressive formula for the conditional mean of duration. Analogously, in our SPIE framework, it is natural to incorporate the additional variables' effects in the parametric time-series model (step 3).

In a typical ARMA model, additional variables can be incorporated in two ways. One is called an ARMAX model. The other is called an ARMA model with regression. Their formulas, state space representations, parameter estimation and prediction are described in Appendix. These two methods can both be summarized in the following format:

$$\begin{aligned}
x_{t+1} &= \Phi x_t + \gamma u_{t+1} + \Theta w_t, \text{ where } w_t \sim N(0, Q), \\
y_t &= A x_t + \Gamma u_t + v_t, \text{ where } v_t \sim N(0, R), \\
cov(w_t, v_s) &= S\delta_s^t \\
x_1 &= 0,
\end{aligned}$$

where y_t are observations, x_t are hidden states, w_t and v_t are gaussian noise, and u_t are additional variables, one of which might be a fixed constant to allow for an intercept.

Likewise, with ARFIMA models, common extensions are the ARFIMAX model and ARFIMA with regression model, in which $y_t - \Gamma u_t$ comes from an ARFIMA model as in Equation 5. In the

example of trade flow in section 5, the additional variable book pressure imbalance is incorporated in such a manner.

$$\phi(B)\Delta^d(y_t - \Gamma u_t) = \theta(B) \tag{5}$$

where $\Delta = (1 - B)$ and B is the back-shift operator.

The state space representation of ARFIMA model and its corresponding parameter estimation and prediction are described in Appendix.

5 Experimental Results

In this section, we apply the SPIE framework to the trade flow from the limit order book of J.P. Morgan stock on the New York Stock Exchange (NYSE) for dates between 07/07/2010 and 07/14/2010. The trading procedure on the NYSE is discussed by Schwarts (1993) [Sch93] and Hasbrouck, Sofianos and Sosebee (1993) [HSS93]. The dataset includes a series of pairs (T_i, B_i), where T_i is log-duration and B_i is an additional explanatory variable, book pressure imbalance. In what follows, we will describe the specific methods that we used in each step of the SPIE framework and specify the way we incorporated the additional expanatory variable. The improvement, in the sense of negative log-likelihood, of each step is given afterwards, followed by the comparison between our model and the ACD model.

We follow the SPIE framework as we described in section 3. In particular, in the first step, since it is well-known that all of the duration processes in market microstructure data are self-exciting (especially inter-trade durations,) and since the most recent durations carry the most information, we used nonparametric kernel conditional density estimation in Hall, Racine and Li (2004) [HRL04] for T_i given T_{i-1}. We call the estimated conditional cdf $\hat{F}_i(.|T_{i-1})$. Figure 1 shows the conditional densities estimation conditional on different values of the previous log-duration T_{i-1}. It is clear that the self-exciting feature is captured by the conditional density estimation to a large extent. No matter what the previous duration T_{i-1} is, the distribution of the current duration T_i has two humps. And as the previous duration increases, both the proportion and location of the second hump increase. The bimodal characteristic explains why those parametric conditional duration models with a unimodal residual distribution cannot capture the dynamics of trade flow very well.

In step 1, the duration needs to be preprocessed before applying the kernel conditional density estimation because the duration is measured in millisecond and thus discrete. In order to make kernel density estimation of log-duration eligible, we added a little noise on the original duration to make it continuous. In detail, we transformed all the original durations in training data and test data by adding a constant 0.5 millisecond first. And when it comes to the kernel conditional density estimation, we add an IID uniform noise $U \sim$ Uniform(-0.5, 0.5). Such a transformation is to guarantee that the log duration is still positive and that the transformed durations still keep their original numerical order. Intuitively, the noise is so small that it should not make a significant difference on the estimation.

The noise must be employed in the first step in order to make duration continuous and avoid the bandwidth in the kernel conditional density estimation crash. However, the noise may be incorporated in the time series modeling step or not. Although we still need a theoretical proof that the noise will not undermine the estimation, we have tested multiple versions of combinations of noises for multiple times. And since the noise is small, the parameter estimation result in time series modeling and the model evaluation result are both consistent and stable. In particular, the rest of this section shows the result with noise employed only in nonparametric kernel conditional density estimation. In the appendix, the detailed adjustment of the estimated function with noise $\hat{f}$ in step 1 and $\hat{h}$ in step 3 as well as model evaluation are provided. In the future work, I will deliver a simulation test to justify the estimation with noise.

In step 2 of SPIE, we get the transformed generalized residuals p_i. As in Figure 2 and Figure 3, the autocorrelation function and partial autocorrelation function of the transformed generalized residuals on 6 days are shown. It is clear that p_i has a long memory with slowly decaying ACF. Thus, we employ an ARFIMA model in step 3.

In step 3 of SPIE, we first consider a time series predictor without book pressure imbalance. By minimizing the AIC score among all of the ARFIMA models, ARFIMA(1,d,1) without intercept

is the chosen model, as in Equation 6. Table 3 shows the results of parameter estimation and the improvement of log-likelihood compared to the purely kernel conditional density estimation. on both training data and test data. ("Test data" refers to the next trading day throughout this report.) From table 3, we can see that parameter estimation is generally stable except on July 9th and the improvement of log-likelihood is consistent on both training data and test data.

$$(1-\phi B)(1-B)^d P_i = (1-\theta B)Z_i, \;\; where \;\; Z_i \sim WN(0,\sigma^2) \tag{6}$$

Table 3: Parameter estimation of fractional integrated ARMA model on the transformed generalized residuals. Standard Deviation is attached the following parenthesis. Δll is log likelihood improvement compared to pure kernel conditional density estimation in step 1. Test data is the data on the next trading day.

Date	ϕ	θ	d	σ^2	$\Delta\, ll_{training}$	$\Delta\, ll_{test}$
07/07	0.149(0.039)	0.393(0.044)	0.211(0.015)	0.97	159	195
07/08	0.133(0.036)	0.387(0.042)	0.208(0.015)	0.97	194	113
07/09	0.088(0.047)	0.296(0.049)	0.175(0.013)	0.97	116	96
07/12	0.199(0.049)	0.409(0.054)	0.186(0.017)	0.97	98	149
07/13	0.157(0.040)	0.376(0.046)	0.182(0.015)	0.98	149	99

In order to incorporate the effects of book pressure imbalance, we consider an ARFIMAX model with two lags of book pressure imbalance as in Equation 7. $|B_i|$ is the absolute value of book pressure imbalance. Book pressure imbalance is measuring the degree of imbalance between buying desire and selling desire. Since it is the degree of imbalance that determines the duration to next trade, we take into account the absolute value of book pressure imbalance instead of original book pressure imbalance). It turns out both b_0, b_1 and b_2 make a difference significantly. Table 4 shows the parameter estimation and the log-likelihood improvement compared to pure nonparametric density estimation on both training data and test data. It is clear that both the parameter estimation and the improvement of log-likelihood are generally consistent across days except that on July 9th, there is a clear deviation. And the improvement of incorporating the book pressure imbalance is significant from the difference of Δll_{test} between the Table3 and Table 4, indicating that the incorporation of book pressure imbalance through an ARFIMAX model performs much better than ignoring book pressure imbalance.

$$(1-\phi B)(1-B)^d(P_i - b_0 - b_1|B_i| - b_2|B_{i-1}|) = Z_i, \;\; where \;\; Z_i \sim WN(0,\sigma^2) \tag{7}$$

Table 4: Parameter estimation of fractional integrated ARMA with book pressure imbalance as additional explanatory variables.

Date	ϕ	θ	d	β_0	β_1	β_2
07/07	0.110(0.050)	0.329(0.058)	0.188(0.017)	0.096(0.038)	-0.225(0.010)	0.128(0.010)
07/08	0.088(0.041)	0.314(0.056)	0.185(0.015)	0.096(0.036)	-0.227(0.009)	0.127(0.009)
07/09	0.052(0.060)	0.247(0.067)	0.161(0.016)	0.113(0.035)	-0.232(0.011)	0.122(0.010)
07/12	0.180(0.064)	0.371(0.073)	0.173(0.019)	0.066(0.039)	-0.217(0.012)	0.133(0.012)
07/13	0.118(0.044)	0.328(0.051)	0.175(0.015)	0.091(0.032)	-0.246(0.009)	0.159(0.009)
Date	σ^2	$\Delta\, ll_{training}$	$\Delta\, ll_{test}$			
07/07	0.896	467	528			
07/08	0.897	530	374			
07/09	0.892	372	287			
07/12	0.908	292	601			
07/13	0.893	609	348			

Then in steps 4 and 5 of SPIE, by transforming the estimated density of P_i back to the density of T_i, we obtain the new estimated density

$$\hat{f}_i^*(t|\mathcal{F}_{i-1}) = \frac{\hat{h}_i(\Phi^{-1}(\hat{F}_i(t)))}{\phi(\Phi^{-1}(\hat{F}_i(t)))}\hat{f}_i(t),$$

where $\hat{f}_i(t)$ and $\hat{F}_i(t)$ are respectively nonparametric kernel conditional estimated PDF and CDF and $\hat{h}_i(.)$ is one-step ahead predicted distribution of P_i from the ARFIMA model on the sequence of p_i.

Figure 4 shows an intuitive idea of how much each step improves the model. The pure nonparametric conditional density estimation works as a baseline. The blue bins reflect the amount of improvement of the ARFIMA model as the time series predictor in step 3 in terms of log-likelihood compared to the pure kernel conditional density estimation in step 1. On average, ARFIMA model improves by 0.5 percent. On the other hand, the red bins reflect the amount of improvement of ARFIMAX model as the time series predictor in step 3 in terms of log-likelihood compared to the pure kernel conditional density estimation in step 1. The ARFIMAX model improves by 1.3 percent on average. Both models make a consistent and significant improvement.

5.1 Comparison to ACD Model

In this section, we implemented the most common version of the ACD model, i.e. log-ACD model as in Equations (2), (3), and (4). We then compare it to our SPIE model in terms of negative log-likelihood on the test data. Table 5 shows the parameter estimation results for the log-ACD model by means of maximizing log-likelihood and the negative log-likelihood on both training data and test data. Obviously, the estimated parameters are consistent and the autoregressive feature is significant since $\hat{\beta}$ is close to 1. In addition, γ is significantly smaller than 1, which suggests a decreasing hazard function, i.e. the self-exciting feature. The original negative log-likelihood is derived as in Equation 8, where x_i is the original duration and $\lambda = \frac{1}{\Gamma(1+1/\gamma)}$ to make sure that the mean of ϵ is 1. In order to compare it with our SPIE model, where we model log-duration we derived the negative log-likelihood for log-duration in log-ACD model as in Equation 9.

Table 5: Parameter estimation of log Autoregressive Conditional Duration modle. $ll_{training}$ is negative log likelihood of training data. ll_{test} is negative log likelihood of test data, which is the next trading day.

Date	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\omega}$	$\hat{\gamma}$	$ll_{training}$	ll_{test}
07/07	0.119	0.955	0.242	0.415	27380	31617
07/08	0.134	0.937	0.355	0.416	31604	22516
07/09	0.163	0.922	0.485	0.402	22492	21484
07/12	0.214	0.927	0.420	0.426	21512	34119
07/13	0.144	0.930	0.392	0.406	33925	26767

$$-ll_{original-duration} = -\sum log(f(x_i|\mathcal{F}_{i-1})) = -\sum log[\frac{\gamma}{\lambda}(\frac{x_i}{\lambda\Phi_i})^{\gamma-1}e^{-(x_i/\lambda\Phi_i)^\gamma}\frac{1}{\Phi_i}] \quad (8)$$

$$-ll_{log-duration} = -ll_{original-duration} - \sum log(x_i) \quad (9)$$

To compare the performance of our SPIE framework and ACD model visually, Figure 5 shows the comparison of negative log-likelihood between the two models. The log-likelihoods are calculated on test data with the model trained from the previous trading day's data. Obviously, the SPIE framework outperforms the log-ACD model to a great extent.

6 Model Diagnostics

Diebold, Gunther and Tay (1998) [DGT98], henceforth referred to as DGT, proposed a method of density forecast evaluation based on the probability integral transform. According to DGT, under the null hypothesis that the predicted density is the true density, the probability integral transform $c_i = \int_{-\infty}^{T_i} f_i(t)dt$ is an IID sample of uniform random variables on the interval (0,1). In particular, they don't have any autocorrelation dependency. Furthermore, Bauwens, Giot, Grammig and Veredas (2000) [BGGV00] compared the main existing conditional duration models, including ACD, log-ACD, TACD, SCD and SVD models, by means of the DGT method. Although these models work

well on the price duration process and volume duration process, none of them work on the trade duration process.

In this section, we investigated our specific model in section 4 by the DGT method. As we described in Section 3.1, the new generalized residual is derived as $c_i^* = \int_{-\infty}^{T_i} f_i^*(t)dt$. Under the null hypothesis that $f_i^*(t)$ is the true model, C_i^* should be IID uniform on (0,1) and have no autocorrelation dependency.

6.1 IID Uniform Test

In this section, we test the null hypothesis that $H_0 : C_i^* \sim U(0,1)$ by a Kolmogorov Smirnov test. Table 6 shows the p-values of the KS tests on both training data and test data on multiple days without the effect of book pressure imbalance. Table 7 shows the counterpart with the effect of book pressure imbalance. Obviously, both the ARFIMA and ARFIMAX model on all test days passed the Kolmogorov Smirnov test at the 0.05 confidence level. In other words, the generalized residuals from our model appear to be close to having a uniform (0,1) distribution.

Table 6: P-value of KS test of model without book pressure imbalance

Date	P-value on training data	P-value on test data
07/07	0.82	0.21
07/08	0.23	0.54
07/09	0.67	0.38
07/12	0.33	0.43
07/13	0.46	0.74

Table 7: P-value of KS test of model with book pressure imbalance

Date	P-value on training data	P-value on test data
07/07	0.68	0.92
07/08	0.89	0.85
07/09	0.81	0.64
07/12	0.86	0.24
07/13	0.77	0.15

6.2 Autocorrelation Test

In this section, we test the null hypothesis that $H_0 : C_i^*$ are not autocorrelated by the Ljung-Box test. Table 8 shows the p-values of the Ljung-Box tests on both training data and test data on multiple days with multiple number of lags Q without the effect of book pressure imbalance. Table 9 shows the counterpart with the effect of book pressure imbalance. We can see from the tables that the generalized residuals generally do not have autocorrelation dependency except that the predictor trained on July 8th and tested on July 9th did not pass the Ljung Box test.

Table 8: P-value of Ljung Box test of model without book pressure imbalance. Q represents the number of autocorrelations considered.

Date	P-value on training data			P-value on test data		
	Q=5	Q=10	Q=15	Q=5	Q=10	Q=15
07/07	0.89	0.66	0.84	0.10	0.11	0.14
07/08	0.18	0.15	0.21	0.02	0.01	0.005
07/09	0.17	0.26	0.12	0.16	0.31	0.63
07/12	0.62	0.60	0.85	0.39	0.23	0.37
07/13	0.61	0.25	0.50	0.38	0.04	0.13

Table 9: P-value of Ljung Box test of model with book pressure imbalance. Q represents the number of autocorrelations considered.

Date	P-value on training data			P-value on test data		
	Q=5	Q=10	Q=15	Q=5	Q=10	Q=15
07/07	0.89	0.81	0.89	0.27	0.46	0.27
07/08	0.37	0.52	0.41	0.03	0.03	0.03
07/09	0.11	0.26	0.13	0.36	0.63	0.80
07/12	0.67	0.78	0.91	0.55	0.66	0.69
07/13	0.91	0.81	0.86	0.11	0.03	0.10

6.3 Pearson Correlation Test

In this section, we test whether the density forecast is well explained by the previous information, i.e. H_0 : the generalized residuals have 0 correlation with the previous time gap and other marks at various lags. We used the Pearson correlation test between generalized residuals and any previous information. The p-values of Pearson correlation tests on various lags and multiple days are shown in Table 10. Likewise, the test is also conducted on the ARFIMAX model with book pressure imbalance. Table 11 and Table 12 respectively show the p-values of tests for durations and book pressure imbalance. In general, the both models fit very well because most previous durations and book pressure imbalances have no correlation with generalized residuals.

Table 10: P-value of Pearson Correlation test between generalized residuals and previous durations at different lags under ARFIMA model without book pressure imbalance on test data.

	P-value				
Lag	07/08	07/09	07/12	07/13	07/14
1	0.57	0.62	0.60	0.88	0.44
2	0.27	0.46	0.14	0.18	0.91
3	0.20	0.77	0.49	0.06	0.28
4	0.71	0.03	0.10	0.66	0.12
5	0.13	0.00	0.25	0.30	0.27
6	0.14	0.10	0.87	0.19	0.02
7	0.04	0.08	0.52	0.48	0.50
8	0.24	0.02	0.28	0.16	0.23
9	0.77	0.18	0.34	0.08	0.26
10	0.13	0.07	0.91	0.69	0.82

Table 11: P-value of Pearson Correlation test between generalized residuals and previous durations at different lags under ARFIMAX model with book pressure imbalance on test data

	P-value				
Lag	07/08	07/09	07/12	07/13	07/14
1	0.65	0.48	0.32	0.53	0.01
2	0.74	0.93	0.18	0.55	0.67
3	0.23	0.61	0.54	0.21	0.30
4	0.98	0.06	0.24	0.54	0.03
5	0.16	0.00	0.38	0.12	0.39
6	0.39	0.38	0.80	0.18	0.05
7	0.17	0.17	0.62	0.49	0.67
8	0.64	0.02	0.34	0.20	0.18
9	0.71	0.18	0.84	0.18	0.40
10	0.39	0.23	0.82	0.48	0.61

Table 12: P-value of Pearson Correlation test between generalized residuals and absolute value of previous book pressure imbalance at various lags under ARFIMAX model with book pressure imbalance on test data

	P-value				
Lag	07/08	07/09	07/12	07/13	07/14
1	0.94	0.33	0.11	0.01	0.00
2	0.60	0.38	0.21	0.11	0.21
3	0.00	0.50	0.06	0.00	0.04
4	0.23	0.79	0.68	0.41	0.75
5	0.60	0.27	0.58	0.26	0.58
6	0.69	0.20	0.40	0.07	0.42
7	0.06	0.12	0.79	0.40	0.34
8	0.43	0.74	0.69	0.20	0.99
9	0.10	0.87	0.31	0.32	0.29
10	0.76	0.89	0.86	0.66	0.61

7 Discussion and Conclusion

We proposed a new semi-parametric framework for estimating the conditional intensity function of a marked point process. In particular, we applied our framework to trade-duration processes. By the DGT evaluation method, our model generally passes the Kolmogorov Smirnov test for the uniform distribution, the Ljung-Box test of independence as well as the Pearson correlation test between residuals and auxilliary variables. Bauwens, Giot, Grammig and Veredas (2000) [BGGV00] claimed that the ACD model and all of its variants model fail to pass both the Kolmogorov Smirnov test and the Ljung-Box test. In addition, our model based on the SPIE framework has consistently significantly better performance than the ACD model and its variants in the sense of log-likelihood on test data. Therefore, our framework has great potential for modeling the intensity functions of duration processes, especially the trade duration process.

Furthermore, we have introduced a systematic way to incorporate additional explanatory variables in the SPIE framework. In the example in section 5, we have shown that the incorporation of effects of book pressure imbalance significantly improves our model in terms of log-likelihood and the new model generally continues to pass the residual tests.

The essence of the framework lies in two aspects. On one hand, the framework decomposes the modeling of the shapes of distributions and the modeling of time dependency. Nonparametric density estimation captures the shapea of distribution, while parametric time series model captures the time dependency. On the other hand, in the particular case of trade-duration process, the framework seperates the long memory and short memory estimation. The nonparametric conditional density estimation captures the recent information to a precise extent, while the time series model captures the relative distant information.

The reason that the framework works and outperforms the existing ACD family on the particular trade duration processes is that the nonparametric conditional density estimation is able to recover the bimodal distribution, while all the ACD family models assume unimodal distributions. Furthermore, the ARFIMA model on the transformed generalized residuals is able to capture the autocorrelation dependence in a more flexible way than merely through the expectation.

The SPIE framework is designed for estimating a conditional intensity function. But it can also be used for estimating and predicting any series of densities, in principle.

8 Work in the Future

8.1 Multivariate Point Process

Our method is aimed at modeling the dynamics of trade flow for a single stock. A natural extension is to model multivariate point processes, motivated by capturing cross stock and cross exchange effects, explaining bivariate buy/sell flow as well as explaining flows of limit orders and cancellations. The

popular current existing models include the Hawkes Process ($\lambda(t|\mathcal{F}_t) = \lambda_0 + \sum_j \sum_{i=1}^{n_i} \gamma_j \phi_j(t - t_i)$), GLM-framework($log\lambda(t|\mathcal{F}_t) = \gamma_0 + \sum_j \sum_{i=1}^{Q} \gamma_i^j \Delta N_{t-i}^j$) and some ACD variants. Their essential idea is to sum up the effects of all of the past events. Yet, they suffer from ignoring the interactive effects of previous events and failure in the goodness-of-fit test. Possible future work directions include (1) combining our SPIE framework with a Hawkes process, (2) introducing a latent process as a driving force, and (3) asymmetric SPIE modeling to deal with buy/sell flow. The asymmetric SPIE is discussed in the next subsection.

8.1.1 Asymmetric SPIE

In practice, traders care about not only the duration of a trade but also on which side the next trade would occur. We already know that trade-duration is closely related with the side of a trade and that longer durations increase the probability of changing side from the previous trade to the limit of one half.

In order to estimate the probability $P(L_i = 1|\mathcal{F}_{i-1})$ of the side of a trade, there are two obvious methods. Since it is known that duration has a great impact on label, we can estimate the duration first and then estimate the probability of $L_i = 1$ conditional on the estimated duration, i.e. $P(L_i = 1|\mathcal{F}_{i-1}) = \int_t P(L_i = 1|T_i = t, \mathcal{F}_{i-1})P(T_i = t|\mathcal{F}_{i-1})dt$. Another method is an extension of our SPIE framework to an asymmetric version, in which there are two types of target events. At each time, both of the two types of events are waiting to occur. But we can only observe the one that occurs first, and the other unobserved duration is a latent duration. We can use nonparametric density estimation for both types of events, $\hat{g}_i^{(1)}(.)$ and $\hat{g}_i^{(2)}(.)$ and then model the transformed generalized residuals by a bivariate time series model with missing values.

In our example, we would fit a bivariate ARFIMA model, in which only one value is observed at each period and the other one is censored to lie in an interval. A state space representation is needed in order to estimate parameters or make predictions for ARMA model or ARFIMA models with missing values. Palma and Chan (1997) [PC97] investigated Kalman filtering with infinite dimensional state methods to fit an ARFIMA model with missing values. Shumway and Stoffer (1982) [RD82] modified the EM algorithm to deal with missing values in state space models. Our ARFIMA model can be expressed in the following equations. Here, X_t are hidden states, one of $y_t^{(1)}$ and $y_t^{(2)}$ is the observed value and the other is censored to lie in the interval $(l_t, +\infty)$. The one-sided censoring indicates that the unobserved duration is longer than the observed one. Without loss of generality, suppose that $y_t^{(2)}$ is missing, then l_t can be derived by $\Phi^{-1}(\hat{g}_i^{(1)}(T_i))$.

$$\begin{cases} X_{t+1} = \Phi X_t + \Theta\epsilon_t, \;\; t = 1, 2..n \\ \begin{pmatrix} y_t^{(1)} \\ y_t^{(2)} \end{pmatrix} = \begin{pmatrix} A_t^{(1)} \\ A_t^{(2)} \end{pmatrix} X_t + \begin{pmatrix} \epsilon_t^{(1)} \\ \epsilon_t^{(2)} \end{pmatrix} \end{cases}$$

In order to estimate parameters and make predictions from our model with interval-censoring, the censored data must be included in the log-likelihood. A modified version of Kalman filter or EM algorithm to deal with the problem of restrictions is needed.

8.2 Update Information between Two Consecutive Trades

Our model is based on analysis of durations and the marks at the times when trades occur. However, some marks evolve continuously over time. For example, the book pressure imbalance is changing between two consecutive trades. So far, our model does not take into account that evolution. Modeling the mark process is actually attractive because quite a few trades have long durations. In such cases, how the marks change during the long inter-trade intervals is more informative in predicting the next trade than the values of the marks when the intervals start.

In duration models, the hazard function is essentially equivalent to the intensity function. Thus, modeling the conditional intensity has potential. The Hawkes process [Haw71] models the conditional intensity function in continuous time. The information from past trades decays over clock time rather than over event time as in duration models. Zhao's (2010) thesis [Zha10] models the conditional intensity depending on the number of events in the previous fixed-time-length window.

For modeling multivariate point processes, we intend to start by combining the SPIE framework with a Hawkes Process to keep updating the information between the current trade and the next trade. In particular, using the SPIE framework, whenever a trade occurs, we obtain a new hazard function for next duration $h_i(.)$. Then the events after the current trade are incorporated by Hawkes kernels, that is,

$$\lambda(t) = h_i(t) + \sum_k \sum_{x_i < t^*_{k,j} < t} \phi_k(t - t^*_{k,j}), \;\; where\, t > x_i,$$

where k represents the types of events, x_i is the clock time of ith trade, $t^*_{k,j}$ is the clock time of jth event of type k, and $\phi_k(.)$ is a linear combination of kernels, such as exponential kernel, Weibull kernel, gamma kernel. Such a parameterization helps identify whether the extra information (other events) excites or inhibits the target point process. And this generalization can also be extended to modeling multivariate point processws to detect causality among processes.

8.3 Model Improvement and General Ideas

There are also a few ways to further improve our model, including automating model selection, modeling deterministic intraday patterns as well as employing index model(i.e. the single variable condition can be replaced by a linear combination of several variables). A few general ideas will also be considered in the future work, including consistency analysis, bootstrap simulation, nonlinear correlation test as well as noise justification in theoretical analysis and simulation. In the next two subsections, we provide a brief of bootstrap simulation and nonlinear correlation testing.

8.3.1 Bootstrap Simulation

In this section, we discuss an algorithm for bootstrapping the SPIE models and how we could implement it. With the bootstrapping tool, we can estimate the bootstrap standard error and a confidence interval for the parameters in the time series models. Moreover, we can estimate the confidence interval of the nonparametric conditional density estimation of current durations given any previous information. As in our example, we can estimate a confidence interval for $f(T_i|T_{i-1})$. We could bootstrap the SPIE model as follows:

1. Given log-duration sequences $\{T_i\}_{i=1}^n$, estimate a series of density $\hat{f}^*_i(.|\mathcal{F}_{i-1})$ by the SPIE framework.
2. Construct the new generalized residuals $c_i = \hat{F}^*_i(T_i), i = 1, 2...n$. By our assumptions, the $\{c_i\}_{i=1}^n$ form a sequenc of IID U(0,1) random variables.
3. Resample the new generalized residuals $\{c_i\}_{i=1}^n$ with replacement to get $\{c^*_i\}_{i=1}^n$.
4. Construct the bootstrap dataset $\{T^*_i\}_{i=1}^n$ as follows. Keep the first few durations the same as the original durations, i.e. $T^*_i = T_i, i = 1, 2, ..m$, for small m. Given $\{T^*_i\}_{i=1}^{j-1}$, we can get a prediction of the density of the next duration $\hat{f}^*_j$ using the SPIE model. Then the next duration is constructed as $T^*_j = \hat{F}^{*(-1)}_j(c^*_j)$.
5. Using the bootstrap durations, repeat the parameter estimation process to get new parameter estimates and new nonparametric conditional density estimates.
6. Repeat steps 3–5 a large number, B, of times and get a set of bootstrap parameter estimates and conditional density estimates $\{\hat{\Theta}^*_i, i = 1, 2...B\}$. These, can be used to construct confidence intervals for both parameters in the time series models and nonparametric conditional densities.

Bootstrap simulation is a powerful tool. There are two difficulties in the bootstrap simulation process. First, it is generally time-consuming for nonparametric conditional density estimation. Second, to calculate the inverse function of the predicted distribution $\hat{F}^{*(-1)}_j(.)$ is not trivial.

8.3.2 Nonlinear Correlation Test

In section 6.3, we have shown by the Pearson correlation test that our model's generalized residual c_i are not linear correlated with some of the previous information, including previous durations T_{i-j}

and book pressure imbalance B_{i-j} at any time lag $j > 0$. However, this test is not enough, even if our model perfectly estimates the duration density. Under the ideal model, c_i should be independent of all information in the entire past $\mathcal{F}_{i-1}$.

An interesting test we plan to implement is to divide the dataset into m subgroups conditional on certain past information. For example, the entire dataset could be divided into 30 subgroups by the $i/30$, i=1,2..30 percentiles of the T_{i-1}. Do the KS test and get a p-value on each subgroup. Ideally, the m p-values should come from a uniform (0,1) distribution. Thus, a pp-plot of the m p-values and the theoretical uniform (0,1) distribution would be helpful to identify any patterns.

References

[BG00] L. Bauwens and P. Giot. The logarithmic acd model: An application to the bid/ask quote process of two nyse stocks. *Annales dEconomie et de Statistique, 60:117149.*, 2000.

[BG06] L. Bauwens and P. Giot. Stochastic conditional intensity processes. *Journal of Financial Econometrics, 4:450493.*, 2006.

[BGGV00] Luc BAUWENS, Pierre GIOT, Joachim GRAMMIG, and David VEREDAS. A comparison of financial duration models via density forecasts. *CORE DISCUSSION PAPER 2000/60*, 2000.

[BH03] L. Bauwens and N. Hautsch. Asymmetric acd models: Introducing price infor- mation in acd models with a two state transition model. *Empirical Economics, 28:1.*, 2003.

[BH06] Luc Bauwens and Nikolaus Hautsch. Modelling financial high frequency data using point processes. *Core Discussion Paper*, 2006.

[BV04] L. Bauwens and D. Veredas. The stochastic conditional duration model: A latent factor model for the analysis of financial durations. *Journal of Economet- rics, 119:381412.*, 2004.

[Con00] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *QUANTITATIVE FINANCE VOLUME 1 (2001)223236*, 2000.

[CST10] Rama Cont, Sasha Stoikov, and Rishi Talreja. A stochastic model for order book dynamics. *OPERATIONS RESEARCH, Vol. 58, No. 3, MayJune 2010, pp. 549563*, 2010.

[DGT98] F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating density forecasts, with applications to financial risk management. *International Economic Review, 39:863883*, 1998.

[ER98] R. F. Engle and J. R. Russell. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica, 66:11271162.*, 1998.

[EZBK10] Eisler, Z., J. P. Bouchaud, and J. Kockelkoren. The price impact of order book events: market orders, limit orders and cancellations. *arXiv:0904.0900.*, 2010.

[GPW+12] Martin D. Gould, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. Limit order books. *arXiv:1012.0349v3*, 2012.

[Haw71] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika, 58:8390.*, 1971.

[HRL04] Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association, Vol. 99, No. 468 (Dec., 2004), pp. 1015- 1026*, 2004.

[HSS93] J. Hasbrouck, G. Sofianos, and D. Sosebee. Orders, trades, .reports and quotes and new york stock exchange. *Working Paper, NYSE.*, 1993.

[OAW05] Obizhaeva, A., and J. Wang. Optimal trading strategy and supply/demand dynamics. *Working Paper, SSRN eLibrary*, 2005.

[Pac06] Maria Pacurar. Autoregressive conditional duration (acd) models in finance: A survey of the theoretical and empirical literature. *ISSN: 1707-410X.*, 2006.

[PC97] Wilfredo Palma and Ngai Hang Chan. Estimating and foresting of long-memory processes with missing values. *Journal of Forcasting.*, 1997.

[RD82] R.H.Shumway and D.S.Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series.*, 1982.

[RD11] R.H.Shumway and D.S.Stoffer. Time series analysis and its applications. *Springer.*, 2011.

[Rus99] J. R. Russell. Econometric modeling of multivariate irregularly-spaced high- frequency data. *Working Paper, University of Chicago*, 1999.

[Sch93] R.A. Schwartz. Reshaping equity markets. *Business One Irwin*, 1993.

[She11] Howard Shek. Modeling high frequency market order dynamics using self-excited point process. 2011.

[Zha10] Linqiao Zhao. A model of limit order book dynamics and a consistent estimation procedure. *Doctoral Dissertation*, 2010.

[ZRT01] M. Y. Zhang, J. Russell, and R. S. Tsay. A nonlinear autoregressive con- ditional duration model with applications to financial transaction data. *Journal of Econometrics, 104:179207.*, 2001.

Appendix

A Figures

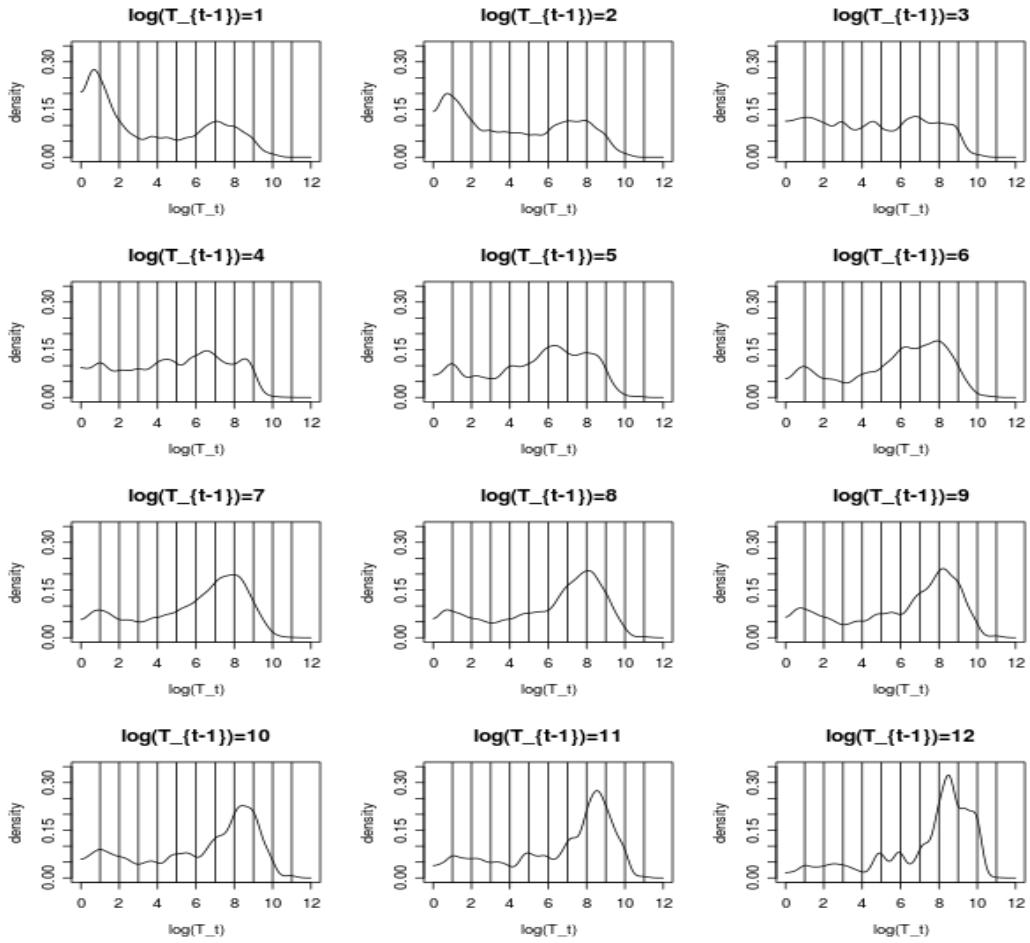

Figure 1: Conditional Density of T_i Given T_{i-1}

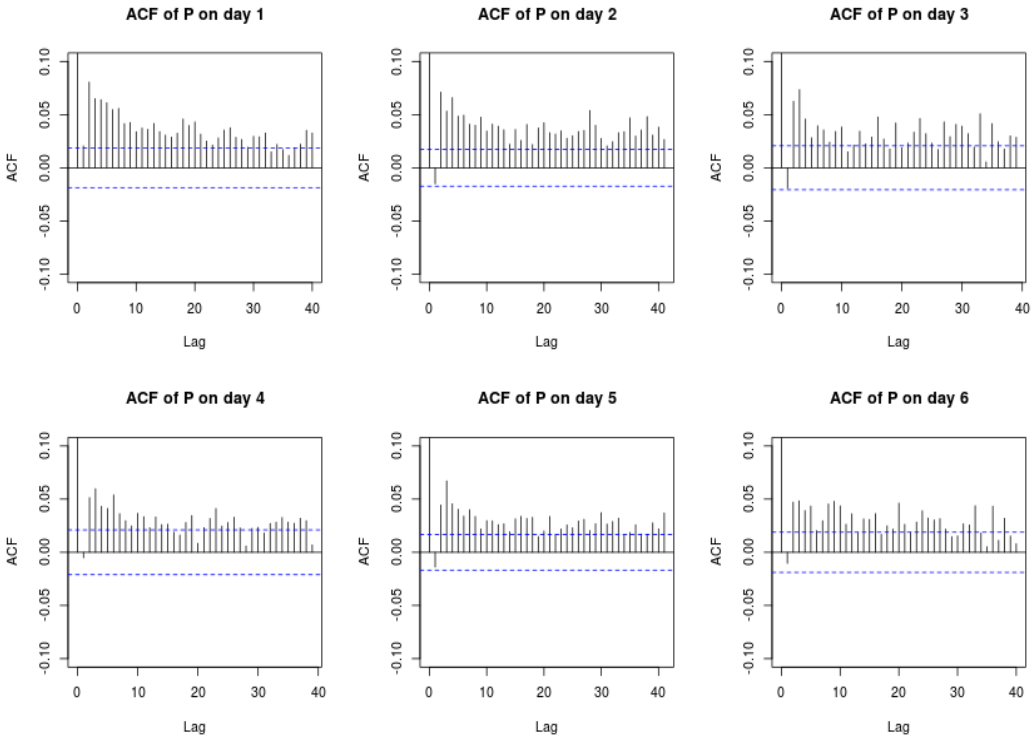

Figure 2: ACF and PACF of Generalized Residual P_i

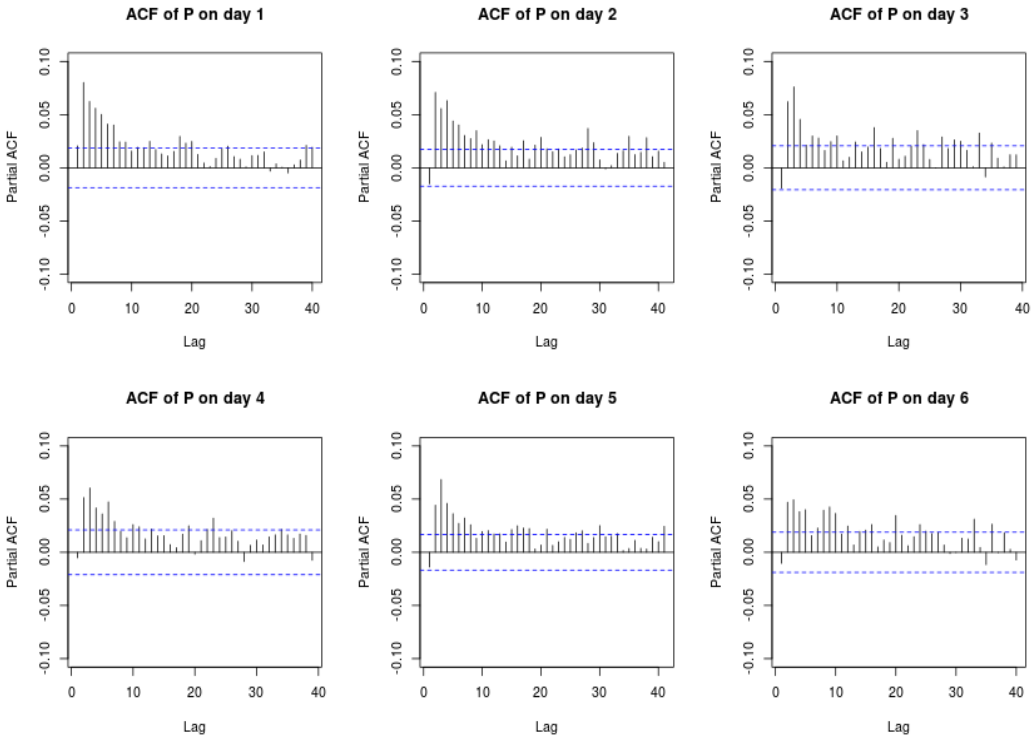

Figure 3: ACF and PACF of Generalized Residual P_i

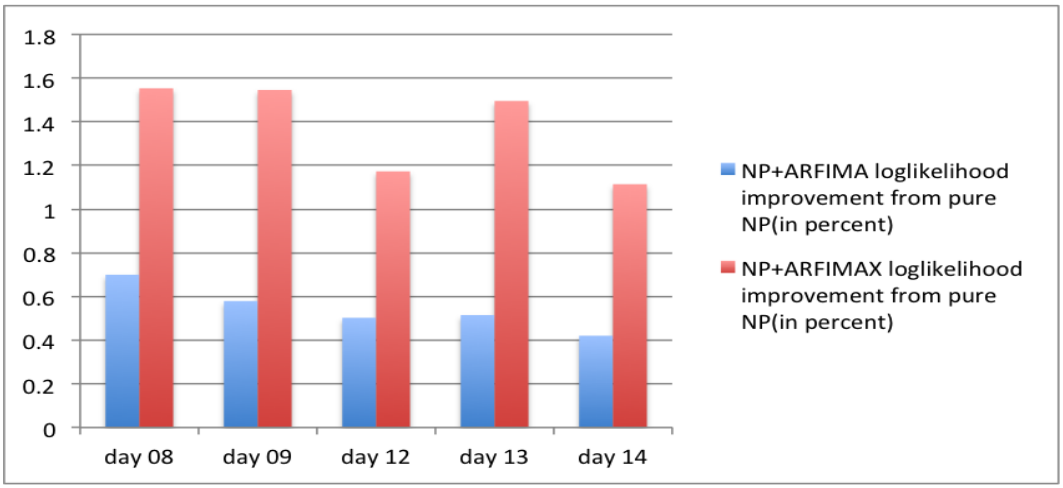

Figure 4: log-likelihood improvement of each step.

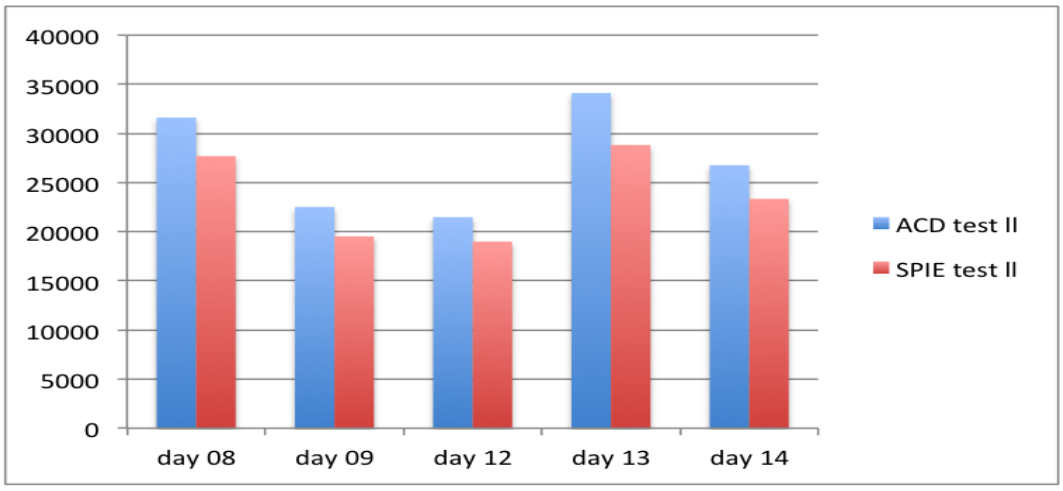

Figure 5: log-ACD -loglikelihood vs. SPIE -loglikelihood

B State Space Representation of ARMA Model [1]

There are multiple ways to represent ARMA models as state space models. In this project, we use the following representation. An ARMA(p,q) model $\phi(B)y_t = \theta(B)Z_t$ can be represented as a state space model:

$$y_t = Ax_t + Z_t$$
$$x_{t+1} = \Phi x_t + \Theta Z_t,$$

where, assuming $p \geq q$,

$$\Phi = \begin{pmatrix} \phi_1 & I & 0 & \cdots & 0 \\ \phi_2 & 0 & I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{p-1} & 0 & 0 & \cdots & I \\ \phi_p & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$\Theta = \begin{pmatrix} \theta_1 + \phi_1 \\ \vdots \\ \theta_q + \phi_q \\ \phi_{q+1} \\ \vdots \\ \phi_p \end{pmatrix}$$

In particular, an ARMA(1,1) model $y_t - \phi t_{t-1} = Z_t + \theta Z_{t-1}$ can be represented as the state space model:

$$y_t = x_t + Z_t$$
$$x_{t+1} = \phi x_t + (\theta + \phi)Z_t$$

B.1 State Space Representation of ARMAX and ARMA with regression Models

Exogenous variables u_t, including fixed constant intercept, may enter into the state equation or into the observation equation, leading to ARMAX or ARMA with regression model respectively.

A typical ARMAX model $\phi(B)y_t = \gamma u_t + \theta(B)Z_t$ can be represented as a state space model:

$$y_t = Ax_t + Z_t$$
$$x_{t+1} = \Phi x_t + \gamma u_t + \Theta Z_t$$

And a typical ARMA model with regression $y_t - \Gamma u_t \sim ARMA(p, q)$ can be represented as a state space model:

$$y_t = Ax_t + \Gamma u_t + Z_t$$
$$x_{t+1} = \Phi x_t + \Theta Z_t$$

In general, the ARMAX model and ARMA model with regression can be combined and represented as:

$$x_{t+1} = \Phi x_t + \gamma u_{t+1} + \Theta w_t$$
$$y_t = Ax_t + \Gamma u_t + v_t$$
$$cov(w_s, v_t) = S\delta_s^t$$
$$x_0 \sim N(\mu_0, \Sigma_0)$$

where μ_0 and Σ_0 are the initial mean and variance, Φ is the transition matrix,Q and R are state and observation covariance matrices, and γ and Γ are input coefficient matrices. In particular, when $\gamma = 0$ and $w_t = v_t$, it reduces to ARMA with regression. And when $\Gamma = 0$ and $w_t = v_t$, it reduces to ARMAX model.

[1] The state space representation of ARMA model, Kalman Filter and log-likelihood are referenced from [RD11]

B.2 State Space Representation of ARFIMA model [2]

A moving average representation of an $ARFIMA(p, d, q)$ model is given by:

$$y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \tag{10}$$

where ψ_j are the coefficients of $\psi(z) = \sum_{j=0}^{\infty} = \theta(z)/\phi(z)(1-z)^{-d}$. The standard ARFIMA model does not have a finite-dimensional state-space representation. However, if we truncate equation (10) at the first m terms, then it can be represented as:

$$x_{t+1} = Fx_t + H\epsilon_t$$
$$y_t = Gx_t + \epsilon_t$$

where

$$F = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$H = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_m \end{pmatrix}$$

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \end{pmatrix}$$

C Kalman Filter for State Space Models

C.1 Kalman Filtering for State Space Models model with independent errors

Consider a state space model with initial conditions $x_0^0 = \mu_0$ and $P_0^0 = \Sigma_0$, and specified as

$$x_t = \Phi x_{t-1} + \gamma u_t + w_t$$
$$y_t = Ax_t + \Gamma u_t + v_t$$
$$w_t \overset{\text{iid}}{\sim} N(0, Q)$$
$$v_t \overset{\text{iid}}{\sim} N(0, R)$$
$$cov(w_s, v_t) = 0.$$

Then for $t = 1, ..., n$, the update is derived as:

$$x_t^{t-1} = \Phi x_{t-1}^{t-1} + \gamma u_t$$
$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q$$

with

$$x_t^t = x_t^{t-1} + K_t(y_t - Ax_t^{t-1} - \Gamma u_t)$$
$$P_t^t = [I - K_t A]P_t^{t-1},$$

where the Kalman gain is

$$K_t = P_t^{t-1} A'[AP_t^{t-1} A' + R]^{-1}$$

[2] Referenced from [PC97]

C.2 Kalman Filtering for State Space Model model with correlated errors

Consider a state space model with initial conditions x_1^0 and P_1^0, and specified as

$$x_{t+1} = \Phi x_t + \gamma u_{t+1} + \Theta w_t$$

$$y_t = Ax_t + \Gamma u_t + v_t$$

$$w_t \overset{\text{iid}}{\sim} N(0, Q)$$

$$v_t \overset{\text{iid}}{\sim} N(0, R)$$

$$cov(w_s, v_t) = S\delta_s^t$$

Then, for $t = 1, ..., n$, the update is derived as:

$$x_{t+1}^t = \Phi x_t^{t-1} + \gamma u_{t+1} + K_t \epsilon_t$$

$$P_{t+1}^t = \Phi P_t^{t-1} \Phi' + \Theta A \Theta' - K_t \Sigma_t K_t',$$

where $\epsilon_t = y_t - Ax_t^{t-1} - \Gamma u_t$, and the gain matrix is given by

$$K_t = [\Phi P_t^{t-1} A_t' + \Theta S][AP_t^{t-1} A' + R]^{-1}.$$

Then the filter values are given by

$$x_t^t = x_t^{t-1} + P_t^{t-1} A'[AP_t^{t-1} A' + R]^{-1} \epsilon_{t+1}$$

$$P_t^t = P_t^{t-1} A'[AP_t^{t-1} A' + R]^{-1} AP_t^{t-1}.$$

And the initial value is:

$$x_1^0 = E(x_1) = \Phi \mu_0 + \gamma u_1$$

$$P_1^0 = var(x_1) = \Phi \Sigma_0 \Phi' + \Theta Q \Theta'$$

D Log likelihood of State Space Model

In the state space model, we use Θ to represent the parameters in the model, $\Theta = \{\mu_0, \Sigma_0, \Phi, Q, R, \gamma, \Gamma\}$. The likelihood is computed based on innovations $\epsilon_t = y_t - Ax_t^{t-1} - \Gamma u_t$ and covariance matrices $\Sigma_t = AP_t^{t-1} A' + R$. Thus, the negative log-likelihood can be written as:

$$-lnL_Y(\Theta) = \frac{1}{2} \sum_{t=1}^{n} log|\Sigma_t(\Theta)| + \frac{1}{2} \sum_{t=1}^{n} \epsilon(\Theta)' \Sigma_t(\Theta)^{-1} \epsilon_t(\Theta)$$

E Durations with Noise

Throughout the section, X_i and T_i are used to represent original duration and log-duration respectively. The duration variable X is measured in millisecond and hence discrete. It ranges from 1 millisecond up to more than 50000 milliseconds and shows a extremely long tail feature. Thus, it is reasonable to fit the duration variable as a continuous variable. Log transformation is employed before modeling is due to the fact that 1) the most recent duration carries the most information and we desire to start with kernel conditional density in the first step. 2) the marginal long tail feature makes the kernel conditional density estimation unreliable. Thus, the log transformation limits the support of T. In fact the nonparametric conditional density on log-duration is equivalent to nonparametric conditional density with adaptive bandwidth. Therefore, a kernel conditional density estimation on the log-duration should perform well. A natural question arises, that is, how to deal with the discreteness of the log-duration before fitting the kernel conditional density estimation. The bandwidth will crash if we apply directly the kernel conditional density estimation on the discrete log-duration. In order to deal with the discreteness of log-duration, we imposed a small noise on the original durations.

E.1 Implementation

We first add 0.5 milliseconds on to all the durations no matter whether it is training data or test data and denote it by Y_i, i.e. $Y_i = X_i + 0.5$. Then, define $Y_i^* = Y_i + U_i$, where $U_i \sim U(-0.5, 0.5)$. Thus, Y_i^* is a continuous variable ranging from 1 to positive infinity theoretically. Define $T_i = log(Y_i)$, $T_i^* = log(Y_i^*)$, $T_i^- = log(Y_i - 0.5)$ and $T_i^+ = log(Y_i + 0.5)$.

Following the SPIE framework, in the first step, we construct the kernel conditional density $\hat{f}(.|T_{i-1}^*)$ by sequence of durations with noise T_i^*. In the second step, calculate the generalized residual by the original discrete log-durations $c_i = \hat{f}(T_i|T_{i-1})$ and $p_i = \Phi^{-1}(c_i)$. And construct $\hat{h}_i(.)$ by sequence of $\{p_i\}$. When predicting the next duration, we discretize the final adjusted CDF $\hat{F}_i^*(.)$ by $P(Y_i = j) = P(T_i = log(j)) = \hat{F}_i^*(log(j + 0.5)|T_{j=1}^{i-1}) - \hat{F}_i^*(log(j - 0.5)|T_{j=1}^{i-1})$.

When we evaluate the model on the test data, the goodness-of-fit test and log- likelihood should be adjusted accordingly.

1) By the discrete version of PITT, the adjusted generalized residual is derived by $c_i^* = \hat{F}_i^*(T_i^-|T_{j=1}^{i-1}) + [\hat{F}_i^*(T_i^+|T_{j=1}^{i-1}) - \hat{F}_i^*(T_i^-|T_{j=1}^{i-1})]V$, where $V \sim U(0, 1)$. And if the model performs well, then $c_i^* \overset{\text{iid}}{\sim} U(0, 1)$. Also note that the generalized residuals can be approximated by $c_i^* \approx \hat{F}_i^*(T_i^*|T_{j=1}^{i-1})$.

2) The likelihood on the test data is derived as:

$$
\begin{aligned}
Likelihood &= \prod_i P(Y_i = y_i|\mathcal{F}_{i-1}) = \prod_i P(T_i = log(y)|\mathcal{F}_{i-1}) \\
&= \prod_i [\hat{F}_i^*(T_i^+|T_{j=1}^{i-1}) - \hat{F}_i^*(T_i^-|T_{j=1}^{i-1})] \\
&= \prod_i [\hat{F}_{Y_i^*}(log(y_i + 0.5)|y_{j=1}^{i-1}) - \hat{F}_{Y_i^*}(log(y_i - 0.5)|y_{j=1}^{i-1})] \\
&\approx \prod_i \hat{f}_{Y_i^*}(y_i^*) \\
&= \prod_i \hat{f}_i^*(log(y_i^*)) \frac{1}{y_i^*} = \prod_i \hat{f}_i^*(T_i^*)/e^{T_i^*}
\end{aligned}
$$

where, $\hat{f}^*(.) = \hat{f}_{T^*}(.)$, $\hat{F}^*(.) = \hat{F}_{T^*}(.)$, and the relationship between the distributions of Y^* and T^* is $\hat{f}_{Y^*}(y) = \hat{f}^*(log(y))\frac{1}{y}$ and $\hat{F}_{Y^*}(y) = \hat{F}^*(log(y))$

And thus, $log-likelihood \approx \sum_i log(\hat{f}_i^*(T^*)) - \sum_i T_i^*$. In principle, we estimated the distributions of continuous sequences of variables, $\{T_i^*\}$ and $\{Y_i^*\}$ and discretize their distributions to get the distributions of original sequences of discrete variables, $\{T_i\}$ and $\{Y_i\}$.

By simulating noises multiple times, we have proved that the construction of $\hat{f}$ in step 1 and $\hat{h}$ in step 3 are stable and consistent. Also, the adjusted generalized residuals pass the goodness-of-fit tests and the log likelihood on the test data improves the ACD model significantly and consistently.