Exchange Rate Duration under a Markov-Switching Multifractal: A GMM Approach

Alex Lewandowski

Economics 472 Winter 2016

Honours Mathematical Economics

Department of Economics

University of Waterloo

Waterloo, Ontario, Canada

Supervisor: Professor Xu

Reference Style: Chicago 16

Date Submitted: April 8, 2016

Acknowledgements

First and foremost, I would like to thank Professor Xu for his patience, guidance and support in the development of this paper. This research topic was heavily influenced by the interest sparked in his financial econometrics class. Professor Busch was equally helpful in my development as a researcher, posing interesting problems and potential solutions in my economic pursuits. Lastly, I thank my peers for their helpful suggestions, keen insights and company.

Abstract

This paper utilizes the Generalized Method of Moments (GMM) to estimate the Markov-Switching Multifractal Duration (MSMD) model of quote duration. In particular, we utilize a 2-Step GMM to obtain efficient and consistent estimates which are applied to Monte Carlo simulations to demonstrate how the estimation performs. After data cleaning, the model is applied to the foreign exchange market using USD/CAD deseasonalized ultra high frequency data. Best linear forecasts are obtained out-of-sample and compared to the classic Autoregressive Conditional Duration (ACD) model. We find that the MSMD model forecasts better than the estimated ACD model at every horizon.

JEL Classification: C51, C53, C58

Keywords — Markov-switching, Multifractal, Duration, GMM, Forecasting

1 Introduction

Since the beginning of the post Bretton-Woods period, modeling exchange rates has been a central practice of econometrics. The foreign exchange market is by far the largest market, with trillions of dollars worth of assets being traded every day. This will only continue to grow, and with the continual rise in computing power and the staggering amount of data available, there are new opportunities to model the market. Price data that was once aggregated to daily, weekly, monthly and quarterly levels can now be analyzed in its purest form; ultra high frequency or tick data. In the foreign exchange market, tick data is the arrival of a new currency quote. This new quote can be radically, slightly or no different from the previous quote. Another aspect that makes tick data unique is the fact that quotes arrive at irregularly spaced time intervals. Quotes on bid and ask prices may come in at minutes, seconds or even milliseconds at a time, with no way to discern when the next quote will appear. This presents a new set of questions for econometricians to analyze. For example, when is the quote going to update next? How does the wait for a new quote influence the bid and ask prices? These types of questions are of quote duration, which simply refer to the time between two successive quotes. In what follows, we analyze the behavior of quote duration.

Certain stylized facts of exchange rate data persist in the ultra high frequency setting: heavy tails, clustering, regimes, and persistence. Some of these effects are more prominent than others at higher frequencies. Long-memory, for instance, is commonly believed to be captured in higher frequency data, when it is ephemeral in longer frequencies datasets. Others, like regimes, are more prominent at lower frequencies. However, these two stylized facts, regimes and long-memory, are underutilized in the quote duration setting in favor of the ACD model and its variants. In our study, we use a regime-switching model for duration that incorporates long-memory through fractal behavior Namely, the model that we use is an MSMD model. We demonstrate that the continuous-state MSMD model is able to forecast quote durations with a greater accuracy than standard duration models such as ACD. The stylized facts serve as important empirical benchmarks when evaluating our model, and we demonstrate that our model fits these empirical regularities.

One way to think of long-memory is persistence in autocorrelations. Specifically, long-memory is defined as the hyperbolic decay in autocorrelation across a time series. Although there is evidence supporting and opposing long-memory in the context of exchange rates, we proceed with this assumption nonetheless. In our empirical work, we see that long-memory is present in our dataset. This is important in the context of this paper since the model is heavily motivated by previous work in deterministic long-memory processes, such as fractal processes. Ideally, we would want this model to be built on a strong economic foundation. The literature already suggests that forecasts from an MSMD model outperform the current standard, the ACD model, in the stock market. Thus, there is statistical evidence to proceed with the MSMD model. What remains is economic motivation through stylized facts. Fundamentally, we hope to show that the results in the stock market are consistent in the foreign exchange market. If not, we explore reasons for why the exchange rate market may be different than the stock market.

The purpose of this paper is to analyze the merits of a simultaneous regime-switching and long-memory approach to quote duration modeling. Empirically identifying regimes from a time series is exceptionally difficult since the switch depends on exogenous variables that are not available. For example, a regime could be good economic times, bad economic times, high volatility, low volatility or even new economic policy. These regimes can change according to other exogenous variables like new information and investor sentiment, which are particularly hard to quantify. Alternatively, we could model these regime changes using other variables as proxies, but we have to make a priori assumptions on which variables are causing the change. Thus, the economic meaning behind regimes is not a focus of this paper. The solution to this problem is to assume that regimes change randomly, and approximate the changes with a stochastic process. Specifically, we use a Markov-switching approach to regimes, and assume that the regimes change according to a Markov chain. Intuitively, this means that the probability of changing from one regime to another depends only on the current regime. This is consistent with what is found in empirical literature, and we formalize the notion of Markov-switching in section 3.

In order to conduct an empirical study, we need a method to estimate parameters given a dataset. Many methods exist to achieve this end, but we use GMM in this paper. There are problems with this method of estimation; there is no clear-cut algorithm for choosing moments, or deciding how many to use. Maximum Likelihood Estimates (MLE) would be ideal since the estimates are asymptotically consistent and straight forward to calculate. In addition, optimal forecasting techniques have been developed using conditional probabilitiy density functions derived from MLE estimates. However, due to computational concerns, we choose to use GMM for our paper. Specifically, the computational complexity of MLE estimation grows exponentially with respect to the number of components in our volatility process. Moreover, our specific model does not benefit from the optimal forecasting techniques discussed earlier. As such, we focus on minimizing bias and maximizing efficiency of our GMM estimates. We achieve this through best linear prediction, which is done pseudo out-of-sample.

The remainder of this paper is divided into five sections. In Section 2 we present historical developments in the areas of fractal time series, high frequency finance and the trade duration models therein. Section 3 provides a treatment of the theoretical preliminaries, including the derivation of our model and a discussion on estimation and forecasting. Section 4 conducts Monte Carlo simulations to verify the estimation techniques and demonstrate properties of the model. Section 5 explores the data and estimates the parameters of the model and provides out-of-sample forecasts. Finally, we analyze the economic implications, limitations and possible extensions of our research.

2 Literature Review

To begin, we review the seminal paper by Engle and Russell (1998) which developed the theory of stochastic financial duration. As discussed earlier, information arrives at irregular intervals in the ultra high frequency setting. Thus, the author incorporates point processes to model the arrival of price changes. A point process is intuitively described as the number of points falling in some space. In this context, we consider points in times and we define a sequence of random variables $\{T_i\}_{i=1}^{N}$, where $T_1 \leq T_2 \leq \cdots T_N$. T_i denotes the time that the i^{th} tick of data arrives. For irregularly spaced time intervals, we don't necessarily have that $T_{i+1} - T_i$ is a constant for all i. Thus, the variable of interest is duration and it is defined as $d_i = T_{i+1} - T_i$. Then financial duration is nothing more than the time between information arrivals. If the arrival of information carries some change in price, as it usually does, then duration is closely related to volatility. Explicitly stated, shorter durations imply that prices change quickly which further implies high volatility. The key insight by Engle and Russell (1998) is that duration can follow dynamics similar to that of volatility. In particular, the model borrows dynamics from the GARCH family. Denote d_i as the duration, ϵ_i as the innovation and ψ_i is the conditional expectation for the i^{th} tick. Then the ACD(m,q) model is defined as follows

$$
\begin{aligned}
d_i &= \epsilon_i \psi_i \\
\mathbb{E}(d_i \mid d_{i-1}, \ldots d_1) = \psi_i(d_{i-1}, \ldots d_1) &= \psi_i \\
\psi_i &= \omega + \sum_{j=1}^{m} \alpha_j d_{i-j} + \sum_{j=1}^{q} \beta_j \psi_{i-j}
\end{aligned}
$$

We have the additional restriction that ϵ_i are independently, identically distributed and parameterized so that $\mathbb{E}(\epsilon_i) = 1$. The exact specification of the error term, ϵ_i, is left undefined but popular choices include log-normal, exponential, generalized gamma and Weibull. As stated earlier, the ACD model is specified similarly to that of the GARCH family. As a result, the two models share some properties in common: persistence, clustering and heavy tailedness. These are pivotal stylized facts in the literature of price durations, but not all encompassing. The ACD model does not produce the long-memory that is present in the data. Although there exist some augmentations to permit long-memory, they neglect yet another stylized fact: regimes. As Morana and Beltratti (2004) point out, models that incorporate both long-memory and regime switching generate superior forecasts of volatility. This should then naturally extend to forecasts of duration, since duration is a proxy for volatility. Before we direct our attention to a model that is intrinsically long-memory and regime switching, we review some preliminary theory of long-memory.

Long-memory, or long range dependency, can be characterized by persistence at longer time horizons. A common way to analyze the dependency is through fractal time series. This subfield of fractal analysis studies self-similar processes and has seen considerable application in financial modeling because of its ability to model long-memory behavior. In fact, Sun et al. (2008) points out that long-memory processes are asymptotically self-similar. Mandelbrot et al. (1997) was the first to apply the theory of fractals to financial markets. His seminal work on multifractal asset returns laid the theoretical ground work for the Markov-switching multifractal model proposed by Calvet and Fisher (2004).

To motivate the development of the model, let us define what it means for a process to be a multifractal. Informally, if a process exhibits fractal behavior, then the time series exhibits similar behavior on shorter and longer time horizons. That is, the process "remembers" its behavior at different time horizons, thus having long memory. More formally, Segnon and Lux (2013) assert that if $X(t)$ is a fractal process, then it must obey the power scaling law as follows,

$$X(at) \overset{d}{=} a^{H} X(t)$$

Where $\overset{d}{=}$ means equivalent in distribution. Additionally, a process is said to be a multifractal when the following, more general, scaling law holds,

$$X(at) \overset{d}{=} a^{H(a)} X(t)$$

Where we note that $H(\cdot)$ is a random function. A process that obeys a scaling law between different time horizons is said to be self-similar or self-affine. One prominent example of a fractal stochastic process is fractional Brownian motion. The benefit of this property is the fact that local behavior can be extrapolated to global behavior. This is naturally relevant when working with ultra high frequency data, where we have extensive local information.

Now, following Calvet and Fisher (2004), we specify the Markov-switching multifractal for returns. We define r_t and σ_t to be the return and volatility at time t and impose the following structure on the volatility,

$$\begin{aligned} r_t &= \sigma_t \epsilon_t \\ \sigma_t &= \bar{\sigma} \left(M_{1,t} \cdot M_{2,t} \cdots M_{k,t} \right)^{\frac{1}{2}} \\ \epsilon_t &\overset{i.i.d.}{\sim} \mathcal{N}(0,1) \end{aligned}$$

Where $\bar{\sigma}$ is a constant to be estimated and, for all i, $M_{i,t}$ is drawn from the same known distribution with probability $\gamma_i = 1 - (1 - \gamma_k)^{(b^{i-k})}$. Note, we must initialize $\gamma_k \in (0,1)$, making it a parameter of the model. In addition, b is the rate at which probabilities transition and k is the number of multiplicative components in the volatility process. Since the innovation term ϵ_t has only fixed parameters, we fully characterize the Markov-switching Multifractal by the following parameter vector $\boldsymbol{\theta} = (\bar{\sigma}, \gamma_k, b, k)^T$.

The allure of such a model comes from its ability to explain phenomena at every frequency. Firstly, it captures the regime switching behavior at longer frequencies or, equivalently, the long run. Secondly, it incorporates the smooth autocorrelation of medium length frequencies, which is the monthly to quarterly level, sometimes this extends to a business cycle. Lastly, it can generate the large outliers in the short frequencies, such as the jumps in intradaily data. We would like to apply this type of model to quote duration, since quote duration is closely related to volatility. Chen et al. (2013) do this using a mixture of exponential approach. Although his derivation differs slightly from Calvet and Fisher (2004), the MSMD model closely resembles the one of returns. In the following section, we derive this model and discuss its properties.

3 Theoretical Results

3.1 Model Derivation

In this paper, we analyze quote duration in the foreign exchange market. Our efforts are to model quote duration with a fractal time series model. In doing so, we may forecast future values of quote durations. In particular, the model we consider borrows the volatility dynamics of the Markov-switching Multifractal and applies them to the duration setting. Chen et al. (2013) derives the MSMD model in terms of trade durations. We follows an equivalent derivation in the context of quote durations, and we present his derivation here with added exposition.

As before, we begin with a point process that is defined on $\mathbb{R}^+$. That is, we define the arrival of new currency quotes as a sequence, $\{t_i\}_{i=1}^N$, with $0 < t_1 < t_2 < \ldots, t_N$. For every such point process, there exists a counting process, N(t), that counts the number of points or events occurring before a specific time. Mathematically, $N(t)$ counts the number of t_i's, such that $t_i < t$ and more formally, this can be written as $N(t) = \sum_{i\geq 1} \mathbb{1}_{t_i \leq t}$ where

$$\mathbb{1}_{X\leq t} = \begin{cases} 1 & \text{if } X \leq t \\ \\ 0 & \text{otherwise} \end{cases}$$

Then we can define the intensity of the sequence, $\{t_i\}_{i=1}^N$, conditional on all the information before time t (which is denoted as $\mathcal{F}_{t-}$), as follows

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} E(N(t+\Delta t) - N(t) \mid \mathcal{F}_{t-})$$

This expression looks like the definition of a derivative with respect to the counting process. Intuitively, it behaves very similarly to the derivative, giving the instantaneous probability of quote arrival at time t. Then, we specify [0,T] as the interval in time that we are interested in. In essence, we are generalizing the idea of Bernoulli trials to continuous time. Recall that a Bernoulli distribution is characterized by two parameters: n and p. Then, $\lambda(t)$ would be the probability of success on a given trial, or the p parameter, while the interval $[0, T]$ is our "number of trials", or n parameter.

We would like to have an expression for the probability that some set of points, $0 < t_1 < t_2, \dots , t_n < T$, are the only points in [0,T] that have a quote arriving. From the definition of the intensity, we know that the probability that a quote arrives at t_i is $\lambda(t_i)$. Then, the probability that quotes arrive at a sequence of times, $\{t_i\}_{i=1}^{n}$ is $\prod_{j=1}^{n} \lambda(t_j)$. However, the probability that no quote arrives in the open interval (t_i, t_{i+1}) is not as obvious.

Here, we have to make an assumption that the quote durations follow a Poisson point process on each open interval $(t_1, t_2), \dots , (t_i, t_{i+1}), \dots , (t_{n-1}, t_n)$. That is, if we define the random variable Y as the number of new quotes, then $Y \overset{iid}{\sim} POI(\lambda_i)$ and we let $\Lambda_i = \lambda(t_i)(t_{i+1} - t_i)$.

$$f(y) = \frac{\Lambda_i^y e^{-\Lambda_i}}{y!}$$

Note that we are implicitly assuming that Λ_i is constant on each time interval $(t_i, t_i + 1)$. Since $\Lambda_i = \lambda(t_i)(t_{i+1} - t_i)$ and $(t_{i+1} - t_i)$ is already a constant on a fixed interval, we deduce that $\lambda(t_i)$ must also be fixed on each interval. So we denote $\lambda(t_i) = \lambda_i$ to emphasize the fact that it is a constant on the interval. It should be possible to relax this assumption, and future work may consider using an inhomogeneous Poisson point process to model this behavior. In the inhomogeneous case, we have that $\Lambda_i = \int_{t_i}^{t_{i+1}} \lambda(t)dt$. For now, since we are interested in the probability of no new quotes in the interval, we need to evaluate the probability mass function at 0.

$$P(Y = 0) = e^{-\Lambda_i}$$

We combine the two results and find that the probability of new quotes at some set of points $\{t_1, \dots , t_n\}$ and no new quotes in the intervals between $\{t_1, \dots , t_n\}$ is as follows

$$\begin{aligned} p(t_1, \dots, t_n \mid \lambda(\cdot)) &= \prod_{j=1}^{n} \lambda(t_j) e^{-\Lambda_j} \\ &= \prod_{j=1}^{n} \lambda_j e^{-\lambda_j (t_{j+1} - t_j)} \\ &= \prod_{j=1}^{n} \lambda_j e^{-\lambda_j d_j} \end{aligned}$$

In the above formulation, note that the assumption that $\lambda(t_j) = \lambda_j$ is the assumption that λ_j is a constant on each given interval, as stated earlier. Then we have the resulting mixture of exponentials that we set out to derive. For clarity, we rewrite this in terms of duration alone. Recall that $d_i = t_{i+1} - t_i$, then we can write the joint probability of some set of durations $(d_1, \dots, d_{n-1})$ as the following mixture of exponential form,

$$\begin{aligned} p(d_1, \dots, d_{n-1} \mid \lambda(\cdot)) &= \prod_{j=1}^{n} \lambda_j e^{-\lambda_j d_j} \\ &= exp(-\sum_{j=1}^{n} \lambda_j d_j) \prod_{j=1}^{n} \lambda_j \end{aligned}$$

Now note that, conditional on λ_i, each individual $d_i \overset{iid}{\sim} Exp(\lambda_i)$. To make this more clear, consider the example where two quotes arrive at times t_1, t_2. Then we can derive the probability of this occurring, with respect to the duration $d_1 = t_2 - t_1$ as follows

$$\begin{aligned} p(d_1 \mid \lambda(\cdot)) &= \lambda_1 e^{-\lambda_1 d_1} \\ &\sim Exp(\lambda_1) = \frac{Exp(1)}{\lambda_1} \end{aligned}$$

With this, we conclude the mixture of exponential derivation for the MSMD model by Chen et al. (2013). We now need to specify a form for λ_i, which is done similar to the model by Calvet and Fisher (2004).

$$d_i = \frac{\epsilon_i}{\lambda_i} \text{ Where } \epsilon_i \overset{i.i.d.}{\sim} Exp(1)$$

$$\lambda_i = \bar{\lambda} \prod_{j=1}^{k} M_{i,j}$$

Thus, $\bar{\lambda}$ is a constant to be estimated. We also have that, for all i and j, $M_{i,j}$ is a random variable that is drawn from a pre-specified distribution with probability $\gamma_j = 1 - (1-\gamma_k)^{(b^{k-1})}$. As before, we must initialize $\gamma_k \in (0,1)$, which makes it a parameter of the model. In addition, b is the rate at which probabilities transition and k is the number of multiplicative components in the volatility process. Thus, we can characterize the MSMD model by the parameter vector $\boldsymbol{\theta} = (\bar{\lambda}, \gamma_k, b, k)^T$. Further parameters are included depending on the distribution of the innovation term. In our paper, we do not follow a discrete distribution as is common in the literature. Instead, we explore the possibility of a log-normal distribution, $M \sim ln\mathcal{N}(\mu, \sigma^2)$, and thus have,

$$M_{i+1,j} = \begin{cases} \text{Drawn from } ln\mathcal{N}(\mu,\sigma^2) & \text{w.p. } \gamma_j \\ M_{i,j} & \text{w.p. } 1-\gamma_j \end{cases}$$

However, as stated earlier, we wish to normalize $\mathbb{E}(M) = 1$. This implies that $e^{\mu + \frac{\sigma^2}{2}} = 1$ or simply, $\mu = -v, \sigma = \sqrt{2v}$. Then, to fully characterize the lognormal MSMD, we need to estimate a vector of parameters, $\boldsymbol{\theta} = (\bar{\lambda}, \gamma_1, b, k, \mu)^T$.

Next, we go over some properties of this model. First and foremost, Chen et al. (2013) shows that the duration process $\{d_i\}$ is strictly stationary and ergodic. This holds true regardless of any distributional assumptions. In fact, strict stationarity only requires that ϵ_i and, for all j, $M_{i,j}$ are independently distributed. The result follows from the fact that d_i is a measurable function of the strictly stationary vector $(\epsilon_i, M_{i,1}, \ldots, M_{i,k})^T$. Secondly, the process is regime switching because of the multiplicative structure of $M_{i,j}$. The larger the value of j, the lower the probability of drawing a new value for $M_{i,j}$, and so $M_{i,j}$ can remain constant for a long period of time. When $M_{i,j}$ finally does switch, we can think of the next set of durations as under a new regime. A third property is long memory, which we demonstrate in section 4 and 5.

Another important property of this process is the fact that it is over-dispersed. This property is related to fat-tailedness of distributions and simply means that the variance is greater than the mean. The proof of this in the discrete case can be found in Chen et al. (2013), however it turns out that it holds true regardless of the distribution. This property is the main generator of the large outliers in the process, which are also commonly found in duration datasets as noted by Sun et al. (2008).

3.2 Estimation Methodology

This paper uses GMM for the estimation of parameters. This is not the only way to proceed; Calvet and Fisher (2004) use simulated likelihood estimation (via particle filters) with Bayesian updating. It is important to note that a traditional approach to Maximum Likelihood Estimation (MLE) is not possible, since the transition matrix is not well defined for a process with infinite states. Furthermore, the Simulated MLE approach becomes computationally unfeasible very quickly with the number of states growing exponentially with the parameter k. In the binomial case, the states grow by 2^k with allows for estimation for moderate values of k. However, this limits the scope of the model, since it is possible that a higher value of k may be ideal for more volatile assets. Although we could use particle filters, as done in the literature, we explore the effectiveness of GMM for this particular class of duration models.

For the traditional Markov-Switching Multifractal volatility model, Lux (2008) developed moment conditions for efficient GMM estimation. We begin by laying out the GMM methodology for clarity. Recall that if θ_{GMM} is the vector valued estimate resulting from GMM, then it minimizes the function $Q(\theta, W)$.

$$Q(\theta, W) = \min_{\theta} \left\{ \boldsymbol{g}^T(\theta; x) \boldsymbol{W} \boldsymbol{g}(\theta; x) \right\}$$

The estimate is then defined as $\theta_{GMM} = argmin(Q(\theta, W))$. We also note that $\boldsymbol{g}(\cdot)$ is vector valued and is defined to be the difference between an analytic moment and a sample moment. We then have that $\mathbb{E}[g(\cdot)] = 0$, which is a necessary condition for estimation. Solving the minimization problem yields the vector-valued estimate, which depends on $\boldsymbol{W}$, the reweighting matrix. A well known result regarding GMM is that the estimates are consistent and asymptotically normal, but not necessarily efficient. The efficiency depends on how the matrix $\boldsymbol{W}$ is chosen and the optimization technique. In this paper, we use a 2-step GMM and a non-linear optimization algorithm in the programming language R. To compute estimates using a 2-step GMM, we set $\boldsymbol{W} = \boldsymbol{I}$ and then redefine $\boldsymbol{W^*}$ to be a function of the estimates. Lastly, the parameters are computed once more with the new weighting matrix $\boldsymbol{W^*}$.

$$\boldsymbol{W^*} = \frac{1}{N}\sum_{i=1}^{N}\Big(\boldsymbol{g}(\hat{\theta}; d_i)\boldsymbol{g}(\hat{\theta}; d_i)^T\Big)^{-1}$$

Then, the estimate for a 2-step GMM, $\hat{\theta}_{2SGMM} = \underset{\theta}{argmin}(Q(\theta, \boldsymbol{W^*}))$. There are alternative GMM methods that we can explore such as: principal component GMM and continuously updating GMM. Further research should explore the suitability of these other methods, especially for larger datasets and more moments, since computational concerns become an even larger issue. Principal component GMM is especially useful when our weighing matrix $\boldsymbol{W^*}$ is of very high dimension, since it checks the contribution of each moment condition to the estimate through an eigenvalue decomposition. In this context, the eigenvalues act as weights for the moment conditions and can be removed if the values are sufficiently low. This essentially acts as a dimension reduction technique for our matrix without having to go through the entire estimation procedure.

The variance-covariance matrix, $\boldsymbol{V}$, of the estimates is then easily computed. We assume that the 2-step GMM yields efficient estimates and define $\mathcal{J}$ as the Jacobian of a function, which is calculated numerically. Then our estimated variance-covariance matrix is defined as follows,

$$\boldsymbol{\hat{V}} = \Big(\mathcal{J}_\theta \mathbb{E}[\boldsymbol{g}^T(\hat{\theta}; d_t)]\, \boldsymbol{W}^{-1}\, \mathcal{J}_\theta \mathbb{E}[\boldsymbol{g}(\hat{\theta}; d_t)]\Big)^{-1}$$

The GMM methodology, at first glance, seems quite powerful and straightforward. The nuance is in the choice of $g(\cdot)$, the moment conditions. For nonlinear models, such as our multiplicative model, the choice is even less clear. As a preliminary example, we may consider the moment,

$$\begin{aligned}\mathbb{E}(log(d_i)) &= \mathbb{E}(log\left(\frac{e_i}{\lambda_i}\right)) = \mathbb{E}(log(\epsilon_i)) - \mathbb{E}(log(\lambda_i)) \\ &= \mathbb{E}(log(\epsilon_i) - log(\bar{\lambda}) - \mathbb{E}\left(log\left(\prod_{j=1}^{k} M_{i,j}\right)\right) \\ &= \mathbb{E}(log(\epsilon_i)) - log(\bar{\lambda}) - k \cdot \mathbb{E}(log(M))\end{aligned}$$

Where the last line follows form the fact that each $M_{i,j}$ is independent of j and identically distributed. While not immediately clear, $\mathbb{E}(log(\epsilon_i))$ can be evaluated trivially through either simulation or symbolic/numerical integration. Moments of this form, as well as different powers of $log(d_i)$, do not use the transition probability parameters and as a result, cannot fully estimate the model. A naive solution is to fix or assume the transition probability parameters, but some knowledge about the system is required to do so. Instead, we want to look at moments of the autocovariance form. That is, we may choose to look at moments of the form $\mathbb{E}(f(d_i)f(d_{i+T}))$, where $f(\cdot)$ is some transformation of the variable. We consider the specific transformation used by Lux (2008) which is defined as follows,

$$f(d_i, T) = \xi_{i,T} = ln(d_i) - ln(d_{i-T})$$

The motivation for this approach is clear; the model is multiplicative and moreover, we use lognormal M-components. This formulation tackles both these issues and allows for relatively simple moment conditions. However, we choose to use conventional moments, without the transformation. For example, we may use moments of the form $\mathbb{E}(d_i^r), \mathbb{E}(d_i d_{i+T}), \mathbb{E}(d_i^2 d_{i+T}), \mathbb{E}(d_i d_{i+T}^2), \mathbb{E}(d_i^2 d_{i+T}^2)$ for various values of T and r. The moments are derived as follows,

$$\begin{aligned}\mathbb{E}\left(\frac{1}{M^r}\right) &= \int_0^\infty \frac{1}{2x^{r+1}\sqrt{\pi\mu}} exp\left(-\frac{(log(x)+\mu)^2}{4\mu^2}\right) \\ &= \left[\frac{1}{2} erf\left(\frac{log(x)+2r+1}{2\sqrt{\mu}}\right) e^{r(r+1)\mu}\right]_0^\infty\end{aligned}$$

Since the error function is either asymptotically 1 or -1, we have that $\mathbb{E}(\frac{1}{M^r}) = e^{r(r+1)\mu}$. Other moments can be derived equivalently and a moment-matching procedure can be used to choose parameters so that moments are evaluated to desirable values.

$$\begin{aligned}\mathbb{E}(d_i^r) &= \mathbb{E}\left(\frac{e_i^r}{\bar{\lambda}^r}\left(\frac{1}{\prod_{j=1}^k M_{i,j}}\right)^r\right) \\ &= \frac{1}{\bar{\lambda}^r}\int_0^\infty x^r e^{-x} dx \left(\mathbb{E}\left(\frac{1}{M^r}\right)\right)^k \\ &= \frac{r!}{\bar{\lambda}^r} e^{rk(r+1)\mu}\end{aligned}$$

From this result, it is clear that μ must be sufficiently small, as well as $\bar{\lambda}$ sufficiently large, so that the mean is not too large. The other pertinent moments for our consideration are cross moments of varying powers. In particular, we will focus on $\mathbb{E}(d_i d_{i+T})$. The key idea behind the derivation of these moments is that the probability of the j^{th} lognormal volatility component not switching in an interval $[t, t+T]$ is

$(1-\gamma_j)^T$. In the case that the component does not switch, it can be factored into $E(\frac{1}{M})^2$. Thus the moment can be derived as follows,

$$\begin{aligned}
\mathbb{E}(d_i d_{i+T}) &= \mathbb{E}\left(\frac{e_i e_{i+T}}{\bar{\lambda}^2}\frac{1}{\prod_{j=1}^k M_{i,j}\prod_{j=1}^k M_{i+T,j}}\right) \\
&= \frac{1}{\bar{\lambda}^2}\mathbb{E}(e_i)\mathbb{E}(e_{i+T})\mathbb{E}\left(\frac{1}{\prod_{j=1}^k M_{i,j}\prod_{j=1}^k M_{i+T,j}}\right) \\
&= \frac{1}{\bar{\lambda}^2}\prod_{j=1}^k \mathbb{E}\left(\frac{1}{M_{i+T,j}M_{i,j}}\right) \\
&= \frac{1}{\bar{\lambda}^2}\prod_{j=1}^k \left((1-\gamma_j)^T\mathbb{E}\left(\frac{1}{M^2}\right) + (1-(1-\gamma_j)^T)\mathbb{E}\left(\frac{1}{M}\right)^2\right) \\
&= \frac{1}{\bar{\lambda}^2}\prod_{j=1}^k \left((1-\gamma_j)^T e^{6\mu} + (1-(1-\gamma_j)^T)e^{4\mu}\right)
\end{aligned}$$

The cross moments with different powers follow fairly easily from the above derivation, changing the exponents as needed. An important assumption in this derivation is that the e_i is independently and identically distributed. We use a combination of moment conditions to estimate the parameter vector entirely. Using too many moments, as well as too many lags per moment, can lead to computational issues. The most computationally expensive step in the 2-step GMM algorithm is the computation of the reweighting matrix. This is because it involves as many matrix multiplications, which are approximately $\mathcal{O}(n^3)$ in complexity, as there are observations.

For the ACD model, we use the "ACDm" library in R. Similar to our MSMD model, we assume an exponential structure on the innovations, and the package performs automatic model selection. In addition, the parameters are estimated via Quasi-Maximum Likelihood (QMLE). The QMLE is evaluated quite quickly, but other computational concerns also present themselves in the forecasting section, which we elaborate on next.

3.3 Forecasting Methodology

The typical forecasting methodology used in the literature is one of Bayesian updating through Bayes' rule. This requires conditional probability density function to be estimated via MLE or simulated MLE, but since MLE is not feasible for the log-normal MSMD model, we are restricted to other forecasting methods. As a result, we use best linear forecasts in our empirical work. A comparison between optimal forecasts and best linear forecasts of volatility using an MSM model can be found in Lux (2008). Firstly, Monte Carlo simulations concluded that best linear forecasts were less efficient than optimal forecasts, but by a negligible amount. Empirical evidence showed little difference between best linear and optimal forecasts on a variety of currencies and model specifications. In Section 5, we investigate how effective linear forecasts are in an empirical setting. Let us formally define the best h-step linear forecast , $\widehat{d_{i+h}}$, for a stationary process with zero mean as follows,

$$\widehat{d_{i+h}} = \boldsymbol{\phi_n^{(h)}}\, \boldsymbol{d_i}$$

$$\text{Where } \boldsymbol{\Gamma_n}\, \boldsymbol{\phi_n^{(h)}} = \boldsymbol{\gamma_n}$$

$$\boldsymbol{d_i} = (d_1, \dots , d_n)^T$$

That is, $\boldsymbol{\Phi_n^{(h)}}$ is any solution to the above linear system. We also have that $\boldsymbol{\Gamma_n}$ is the variance-covariance matrix, $\boldsymbol{\gamma_n} = (\gamma(h), \dots , \gamma(n + h - 1))$ where $\gamma(\cdot)$ is the autocovariance function. The theory of best linear prediction is a well understood result from time series analysis and it can be shown that it is "best" in the sense that it minimizes Mean Squared Error (MSE).

A direct computation of $\boldsymbol{\Phi_n^{(h)}}$ is impossible, given that the dimension of $\boldsymbol{\Gamma_n}$ is $n \times n$. Even for moderately sized datasets, the inverse of such a matrix is completely infeasible. However, Brockwell and Dahlhaus (2004) develops an extension of the Durbin-Levinson algorithm for computing the inverse of a matrix whose diagonals are constants. Such a matrix is called toeplitz, and the variance-covariance matrix is toeplitz for a stationary process. In particular, algorithm 1 is the classic Durbin-Levinson algorithm for computing the 1-step forecast. That is, we set $h = 1$ from the preceding definition and we have $\widehat{d_{i+1}} = \boldsymbol{\phi_n}\, \boldsymbol{d_i}$. Then the coefficients, $\boldsymbol{\phi_n} = (\phi_n(1), \dots , \phi_n(n))$, are obtained using the following recursive relationship. We begin with $m = 1$, $v_0 = \gamma(0)$ and $\phi_1(1) = \gamma(1)/\gamma(0)$, iterating until $m = n$.

$$\phi_{m+1}(m + 1) = [\gamma(m + 1) - \sum_{j=1}^{m} \phi_m(j)\gamma(m + 1 - j)]v_m^{-1},$$

$$\phi_{m+1}(j) = \phi_m(j) - \phi_{m+1}(m + 1)\phi_m(m + 1 - j),\ j = 1, \dots , m$$

$$v_{m+1} = (1 - \phi^2_{m+1}(m + 1))v_m$$

To calculate an h-step forecast, the coefficients $\boldsymbol{\phi_n^{(h)}} = (\phi_n^{(h)}(1), \dots, \phi_n^{(h)}(n))$ follow the following recursive relationship, this time iterating over h. Note that algorithm 1 is necessary to obtain the initial conditions for algorithm 2. That is, $\phi_n^{(1)}(j) = \phi_n(j)$ and $v_m^{(1)} = v_m$.

$$\phi_m^{(h+1)}(m) = [\gamma(m+h) - \sum_{j=1}^{m-1} \phi_{m-1}(j)\gamma(m+h-j)]v_{m-1}^{-1},$$

$$\phi_m^{(h+1)}(j) = \phi_m^{(h)}(j+1) + \phi_m^{(h)}(1)\phi_m(j) - \phi_m^{(h+1)}(m)\phi_m(m-j),\ j = 1, \dots, m-1,$$

$$v_{m+1}^{(h+1)} = v_m^{(h)} + (\phi_m^{(h)}(1)^2 - \phi_m^{(h+1)}(m)^2)v_{m-1}$$

We then compare our model against the classic ACD model. In addition to MSE, we also use Mean Absolute Error (MAE) as a measure of forecasting accuracy. The main difference between Mean Squared Error (MSE) and Mean Absolute Error (MAE) is that MSE puts a heavier penalty on large errors. The average of M different forecasts can be calculated in one of the following ways.

$$MSE = \frac{\sum_{j=1}^{M}(d_j - \hat{d}_j)^2}{M}$$
$$MAE = \frac{\sum_{j=1}^{M}|d_j - \hat{d}_j|}{M}$$

We use both MSE and MAE to give perspective on which forecasts generate outliers, but we must use our own judgment when deciding whether large outliers are acceptable in our forecasts. We note that both measures suffer from not revealing the direction of the error. The importance of the direction of error depends on the context of application. For example, if we are interested in applying our forecasts to a trading system that is computationally intensive, underestimating the duration for a quote arrival can destroy any possibility of turning a profit when trying to take advantage of ultra high-frequency data. However, we leave these issues for further application.

4 Monte Carlo Simulations

To illustrate some properties of the model, as well as the GMM estimation, we employ Monte Carlo simulations. Specifically, we generate samples from our distribution with fixed parameters. Afterwards, we keep the generated data and estimate the parameters of the model, taking the mean and standard error of the

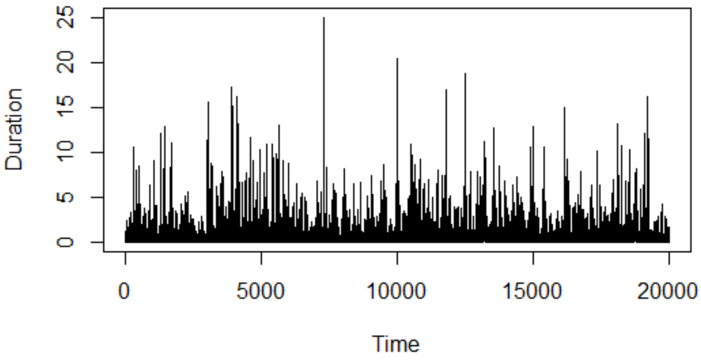

Figure 1: Sample Duration for $\theta = (2, 0.5, 3, 0.05)$ and $\bar{k} = 7$

parameters that we estimate. However, a problem arises in the estimation of k, the number of switching components. This value has to be a positive integer number, and integer optimization problems in general are very difficult. Thus, we fix $\bar{k} = 7$, which Chen et al. (2013) assert is a common finding in the empirical literature. This leaves us with the unknown parameter vector $\boldsymbol{\theta} = (b, \gamma_k, \bar{\lambda}, \mu)^T$, which we estimate using GMM. First, we fix $\theta = (2, 0.5, 3, 0.05)$ and generate 20,000 observations as discussed earlier. We now explore some properties of the data that we generated, which can be examined through the summary statistics and plots.

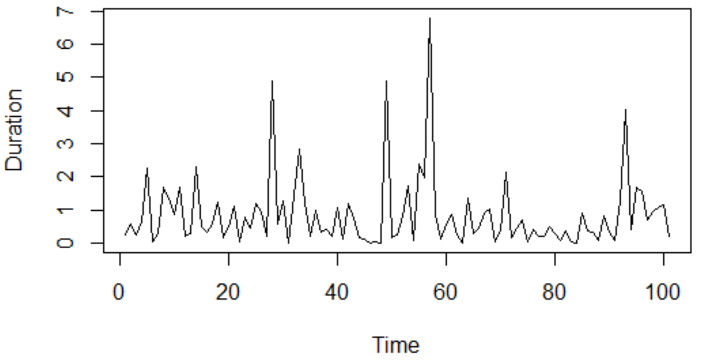

Figure 2: Sample Duration for $i \in [10000, 10100]$

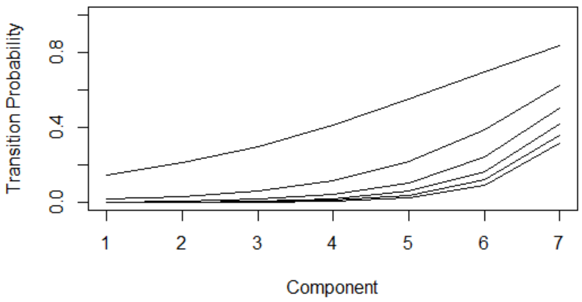

Figure 3: Transition Probabilities for $b = (1.5, 2, 2.5, 3, 3.5, 4), \gamma_k = (5/6, 5/8, 1/2, 5/12, 5/14, 5/16)$

As we can see from Figure 1 and 2, the behavior is quite sporadic in that the model occasionally generates a very large outlier. We also observe periods of very low duration, followed by periods with excessive duration, which demonstrates the regime switching nature of the MSMD model. These are all properties that we expect to see, since this follows exactly from the theoretical derivation in the previous section. We next consider the presence of long-memory in the data, which is revealed in the plot of the autocorrelation function. From Figure 4, we see that the correlation decays hyperbolically with respect to the lag parameter. In addition, these autocorrelations are statistically significant, as indicated by the dashed line which is the line $y = \frac{1}{n} = \frac{1}{20000}$. Before the estimation results, we discuss the choice of parameters for the Monte Carlo simulation. We set $b = 2$ and $\gamma_k = 0.5$ because these parameters dictate the magnitude of the outliers that are generated and also the persistence of regimes. Namely, if γ_k is sufficiently high, then all volatility components are sampled again with high probability. The other parameter controls the rate at which the transition probabilities go from the largest transition probability, γ_k, to the smallest one, λ_1. We simulate sample paths of the transition probabilities for fixed $\bar{k} = 7$ in Figure 3. In addition, we choose values that generate plausible values for duration. Specifically, we chose parameters to avoid outliers that were too gross and to obtain reasonable mean. From an earlier derivation, we have that $\mathbb{E}(d_i) = \frac{\bar{k}}{\lambda} e^{2\mu\bar{k}}$. Thus, μ must be sufficiently small

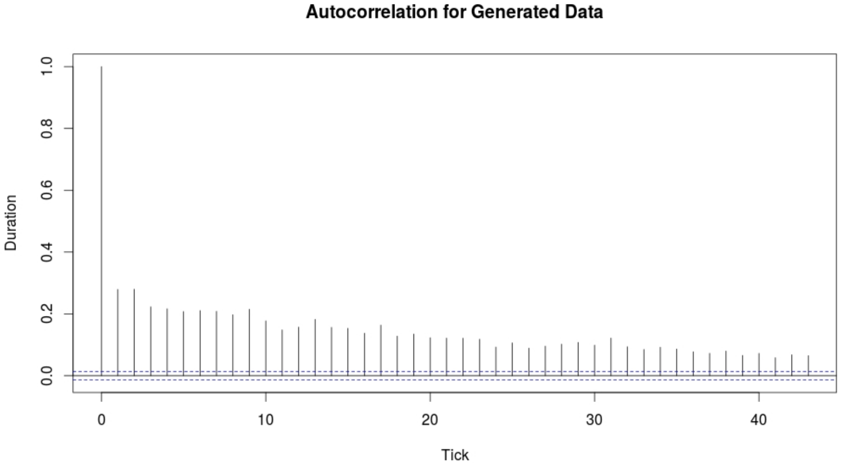

Figure 4: Sample Autocorrelation for $\theta = (2, 0.5, 3, 0.05)$, $\bar{k} = 7$

Next, we wish to examine the properties of the GMM estimates. We use a subset of the moments from the previous section: $\mathbb{E}(d_i), \mathbb{E}(d_i^2), \mathbb{E}(d_i\, d_{i+1}), \mathbb{E}(d_i\, d_{i+10}), \mathbb{E}(d_i^2\, d_{i+5}^2)$. Note that this is not an exhaustive list of moments; there are an infinite amount of moments that we could consider. In the future, it may help to develop an automated method to select moments for this type of model. Nonetheless, with these few moment conditions, we are able to estimate reasonably well using a 2-Step GMM. From Table 1, the mean estimate does converge to the true value, with a reasonable standard error. These estimates may be improved by incorporating more moments, or an alternative GMM estimation procedure, such as principal component GMM. For example, b is the parameter estimated most poorly. One reason for this is b and γ_k drive the regime switching process, which is difficult to fully observe because of its latent nature. Histogram of parameter estimates can be found in the appendix.

	Actual Value	Mean Value	Standard Error
b	2.0000	2.0407	(0.3577)
γ_k	0.5000	0.5488	(0.1509)
λ	3.0000	2.9283	(0.2186)
μ	0.0500	0.0477	(0.0049)

Table 1: Parameter estimates for Monte Carlo simulations, $\bar{k} = 7$

For future Monte Carlo studies, we try different starting parameters to see how robust the GMM-estimates are, and if we can eliminate the edge cases where our algorithm fails to reach a global minimum. This is especially pertinent to this problem because it is a constrained minimization problem, $b > 1, 0 < \gamma_k < 1, \mu > 0, \lambda > 0$. This will further demonstrate the dynamics of an MSMD model, and give insight on how to estimate the parameters more efficiently and robustly.

5 Empirical Results

5.1 Data Analysis

Our empirical research examines ultra high frequency recordings of the CAD/USD exchange rate during the month of January 2015. The data encompasses quotes from all hours of the day and contains more than 2 million quote observations with time, bid and ask prices. Typically, this data is quite hard to come by, and you're expected to pay for it. However, TrueFX provides data for any month between 2008 and 2016. There are other providers of ultra high frequency forex data, but this dataset measured time up to the nearest millisecond. Even with such precision, we have some errors in the data. Particularly, sometimes trades occur at the "same time," and the duration between them appears to be zero. This introduces bias in our estimation and prevents some transformations, such as logarithms. To circumvent this, we substitute the value 0.0009 in lieu of knowledge of how the values are truncated. In a sense, this is a source of measurement error since the value could actually be less than 0.0009, we simply chose the largest value that could not be observed given the data. However, the alternative of throwing away observations also introduces bias since about 4% of the observations do occur at the same time. These errors are not equally dispersed throughout the dataset, instead they cluster together during periods of higher volatility.

On the other hand, both bid and ask prices are up to two decimal places, which is standard for prices. However, we are only concerned with quote duration, and so we do not use the bid, ask or spread in our analysis. By the very definition of quote duration, we have that some of the observations are redundant. The reason for this is that quotes are occasionally updated without any change in price. This would still count as an observation of quote duration since the market acquired new information and decided that the

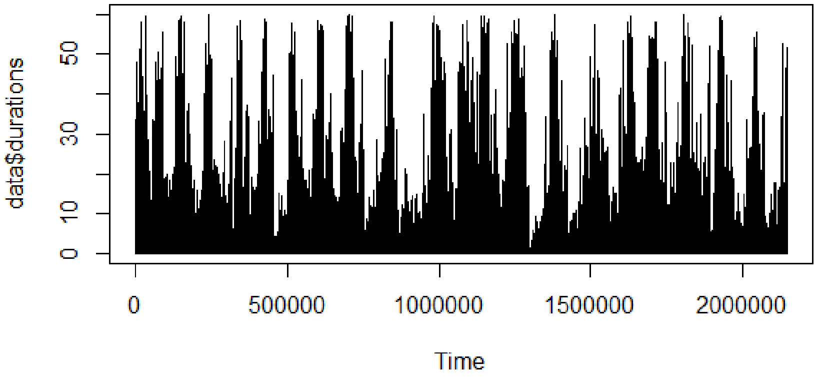

Figure 5: Duration data before seasonal adjustment

exchange rate should be priced the same. In other markets, the data records actual transactions, instead of the currency quotes of exchange rates. An example of this would be stock markets, where we can analyze inter-trade duration, which would involve every observation. However, there is some value in analyzing a contaminated dataset, with the redundant observations. Engle and Russell (1998) found that including the redundant observations did little to change the analysis with his ACD model, and so we compare estimates with both the contaminated data and the pure data. If there is a difference, it would beg the question if the difference is economical or statistical in nature.

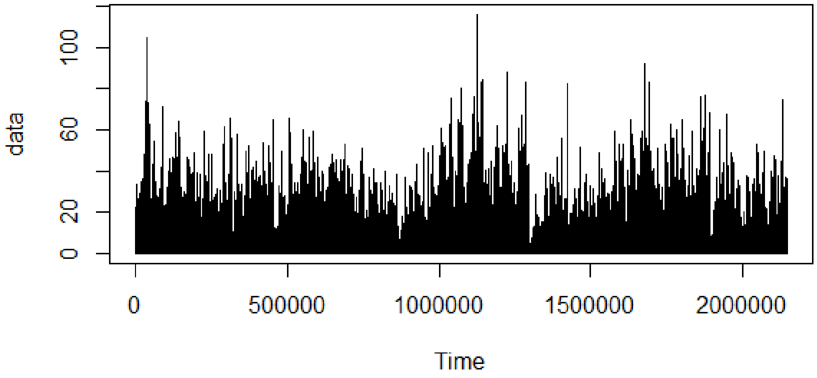

Figure 6: Duration data after seasonal adjustment

Moments	$\mathbb{E}(d_i)$	$\mathbb{V}(d_i)$	$\mathbb{K}(d_i)$	$\mathbb{S}(d_i)$
Estimate	1.0000	5.9825	6.8983	91.6720

Table 2: Summary Statistics for the deseasonazlied USD/CAD dataset

There are other preliminary issues with this dataset, as well as all ultra high-frequency datasets. As noted by Sun et al. (2008), there is seasonality within intradaily data. People need to eat, go home, and sleep; daily human activity is seasonal in that respect. As a result of these factors, trading activity is consistently lower during some periods. For forecasting and estimation purposes, it is required to remove seasonality since it results in nonstationarity. We achieve this through a weight-transformation in the data, using kernel regression to estimate the weights given at each point. Put more formally, we redefine duration as $d_i^* = \frac{d_i}{\phi(t_i)}$ where $\phi(t_i)$ is the Nadarya-Watson estimator evaluated at t_i and $d_i = t_i - t_{i-1}$.

$$\phi(t) = \frac{\sum_{j=1}^{n} K_h(t - t_j) d_i}{\sum_{i=1}^{n} K_h(t - t_j)}$$

For the purposes of calculating this estimator, it is important to measure the times in a consistent manner. Since we are concerned with intradaily seasonality, the time is measured in seconds from midnight of that particular day. After the transformation, we use the set of d_i^* as our dataset, for which we estimate the relevant parameters. This has no effect on the actual inference and prediction, since we can easily undo the transformation after we retrieve the relevant forecasts.

This differs from the approach taken by Chen et al. (2013), where the author specified distinct intervals of activity and used a regular OLS to estimate the weight given in these regions. Instead, we follow the approach by Huptas (2009). Unlike the stock market, this paper deals with the foreign exchange market and so it is not possible to make assumptions on the seasonality of the data. Even if we were to identify which periods should be separated, the choice would be completely arbitrary when the dataset changes to a different currency. For these reasons, we choose to use the Nadarya-Watson estimates so that the results found can be replicated using other datasets.

5.2 Model Fit

In our analysis, we found that the MSMD model fit reasonably well to the in-sample data. For the moment conditions, we use the moments outlined in the previous section, $\mathbb{E}(d_i), \mathbb{E}(d_i^2), \mathbb{E}(d_i\, d_{i+1}), \mathbb{E}(d_i\, d_{i+10}), \mathbb{E}(d_i^2\, d_{i+5}^2)$. We begin by drawing inference on our data based on the parameters. The ACD and MSMD models provide very different perspectives on exchange rate dynamics. For example, from the b and γ_k parameters we can infer that the transition rate between the multiplication components is relatively slow. If b is unusually high, it may suggest a significant change in the data generating process somewhere in the dataset. This sort of data mining could indicate the presence of change points within datasets, although identifying one would be much more difficult. On the other hand, the ACD model provides insight for the conditional heteroskedastic nature of duration. In addition, the autoregressive component makes the model not completely stochastic. This is in contrast to MSMD, where each simulation may be wildly different than the one before it.

Estimation of the ACD model must be done carefully, since $\alpha + \beta < 1$ is a necessary condition for stationarity and it is occasionally violated in our estimation. This is in contrast to the literature, durations are typically mean and covariance stationary, as demonstrated by Chen et al. (2013). Estimation of the MSMD model is also difficult, possibly as a result of the complex nature of the model and its moments. Common problems include converging to local minima in the optimization problem, or not converging at all. The fundamental issue is that the parameter space is not compact. In particular, the estimation would converge to $\gamma_k = 1$, which is not valid. To circumvent this, we bound $\gamma_k \leq 0.95$. Future work may explore better ways to approach this minimization problem, or how to transform the data/model so that current optimization algorithms struggle less in the optimization step.

MSMD	b	γ_k	$\bar{\lambda}$	μ
Estimate	1.9456	0.9500	3.3788	0.0862
S.E.	(0.0285)	(0.0071)	(0.0138)	(0.0003)

ACD	ω	α_1	β_1
Estimate	0.0741	0.1874	0.7560
S.E.	(0.0003)	(0.0006)	(0.0008)

Table 3: Parameter estimates based on deseasonalized CAD/USD dataset

Since we fixed $\bar{k} = 7$ for this estimation, it makes sense for the b parameter to be quite small. This means that the transition probability grows to γ_k relatively slowly. That means that most of the multiplicative components are being re-sampled when a new quote is observed. This makes sense in our relatively small dataset. If we could collect data on a number of years, it would be likely that b would be much higher. That is, more multiplicative components would be "stuck" at some value for extended periods of time. As discussed earlier, if we could find a dataset with known changes in the data generating process (such as some economic shock), then we would expect b to be higher.

Recall that the data was treated so that if $d_i = 0$, then we redefine $d_i = 0.0009$. In the case that $d_i = 0$ is thrown away instead of substituting, we have that b is much higher. In fact, we estimate that $b = 22.423$ when the data is not treated as we did earlier. This actually is consistent with the hypothesis from before. We know that volatility and duration clusters, and that d_i is truncated to 0 because of measurement restrictions. That means that the values that we throw out in periods of high volatility, and they tend to come in consecutive sequences. If these values are omitted, then the points outside of a sequence of zeros will have inflated durations. This is exactly like a sudden regime change, which is reflected in the high b estimate.

5.3 Forecasting Results

The goal of this paper is to demonstrate that the MSMD model is better at forecasting than the current standard, the ACD model. This study focuses on the foreign exchange market, which is new ground for the MSMD model. To that end, we employ cross validation across many subsets and compute pseudo out-of-sample forecasts. That is, we partition the data so that we only fit a portion, and then forecast the data that was left out. The idea is to do this many times, on different subsets, and to take some function of this average. In our case, we use MSE and MAE, but there are many other measures that one could use. Since we are using a single dataset, these forecast measures are sufficient for comparative purposes. However, the actual values do not say anything outside of this comparative setting.

Aggregated values can be found in Table 4, they are represented as ratios of the model's MSE divided by the MSE by a naive forecasting strategy. In our case, the naive strategy is $\hat{d_{i+h}} = d_i$. We find that the MSMD model outperforms the ACD model at every forecasting horizon. In fact, at horizons greater than 1, we find that ACD is outperformed by the naive forecasting strategy. This supports Calvet and Fisher (2004) findings on volatility forecasting, as well as Chen et al. (2013) findings in duration forecasting in the stock market. Referring to Figure 7 for a sample forecast, we find that naive forecasts grossly overestimates. On the other hand, the linear forecasts from the MSMD model stay close to the mean. It is important to note that there is some odd behavior in the MSE and MAE values. Usually, these values increase with higher values of h. In our case, the naive forecasts perform much worse at larger values of h compared to both ACD and MSMD which makes the ratio actually decrease in some cases. Thus, we add to the existing literature by demonstrating that the MSMD model provides superior forecasts within the quote duration setting.

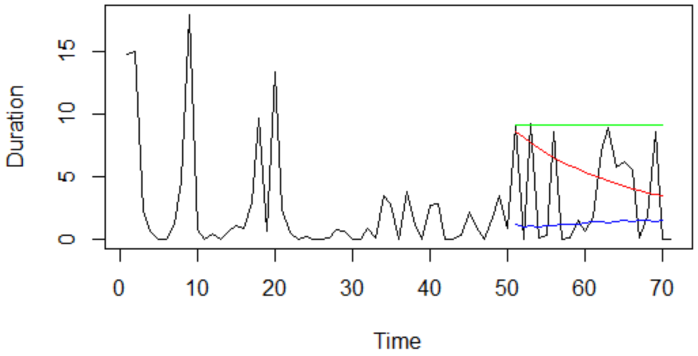

Figure 7: Sample Forecast, green is Naive, red is ACD and blue is MSMD

MSE	ACD	MSMD	MAE	ACD	MSMD
1	0.9911	0.4653	1	1.0077	0.7981
5	0.9688	0.4471	5	1.0783	0.7630
10	0.9664	0.4411	10	1.2068	0.7424
15	0.9888	0.4414	15	1.3516	0.7353
20	1.0340	0.4472	20	1.5100	0.7444

Table 4: MSE and MAE ratios for h-step forecasts, h = (1,5,10,15,20)

6 Conclusion

In this paper, we derive the MSMD model by Chen et al. (2013) and apply it a novel setting. We demonstrate stylized facts through theoretical properties of the model and empirical properties of the data. Then, we develop an estimation methodology by deriving moments for the estimation of a 2-step GMM. In addition, best linear, but not optimal, forecasts are defined and discussed within the context of the MSMD model. Using this theoretical groundwork, we fit the model to ultra high frequency data of the CAD/USD exchange rate during the month of January 2015. We find that the MSMD model outperforms the ACD model at every forecast horizon.

Throughout the paper, numerous comments are left detailing potential pitfalls and problems. These issues are generally computational in nature. One exception is the fundamental choice in moments. There is no clear way to decide which moments to use, but we found that the moments used in this paper provide similar results to Lux (2008). Another problem presents itself in the estimation step, since we cannot estimate k. The literature maintains that $\bar{k} = 7$ is a reasonable assumption for most applications. However, but a method to estimate k with the unknown parameter vector would improve forecasts. The largest problem of all is computational. Specifically, the issue lies in the forecasting step. The matrix that we must invert is $N \times N$, where N is the number of observations. Constructing a matrix for large N (in our case $N \approx 2.1 \cdot 10^7$) is difficult as-is, taking the inverse is impossible. We implement recursive algorithms that are more efficient, but we are still not able to iterate for all the observations due to computational restrictions. A method other than best linear forecasts will be explored for further research.

Other extensions are also available in terms of model specification. For example, the rate parameter λ was assumed to be constant on any open interval (t_i, t_{i+1}). This could be generalized to be an inhomogeneous Poisson point process. This may be more effective as algorithmic trading becomes more prevalent and time intervals become smaller. Lastly, investigating the effect of an increase in trading volume on the predictive capabilities of this model would be very interesting. This could be done by taking a particular month of data over many years, since trading volume has increased greatly in the last 10 years.

Appendix

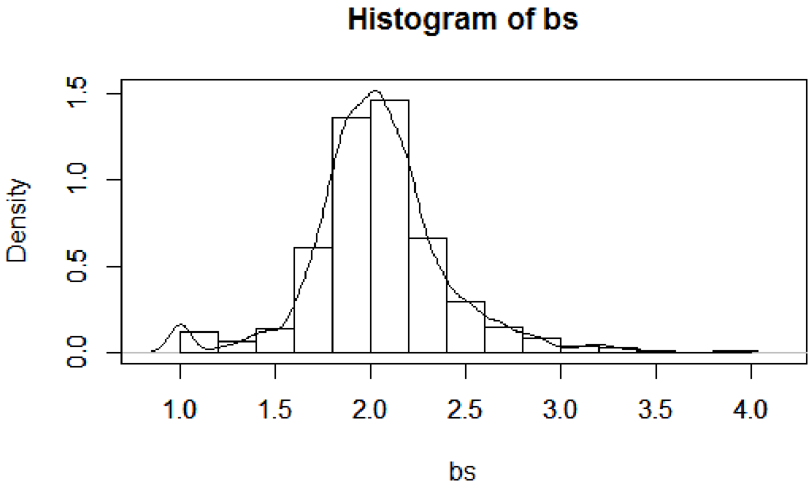

Figure 8: Histogram and Density Estimation for b

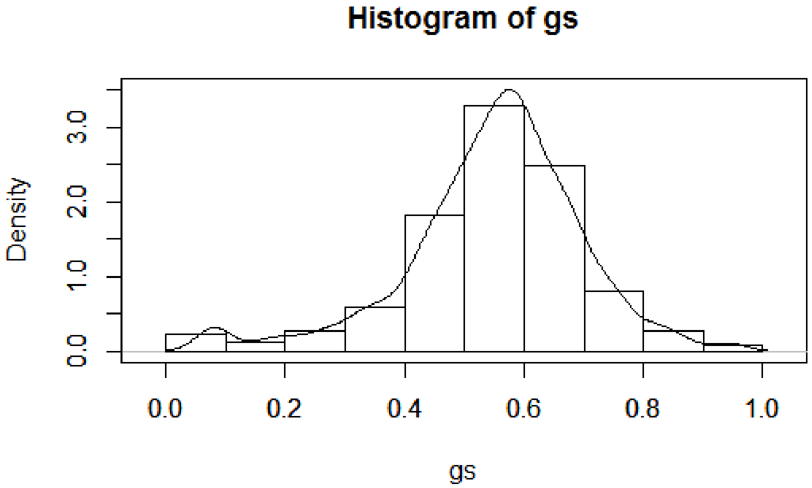

Figure 9: Histogram and Density Estimation for γ_k

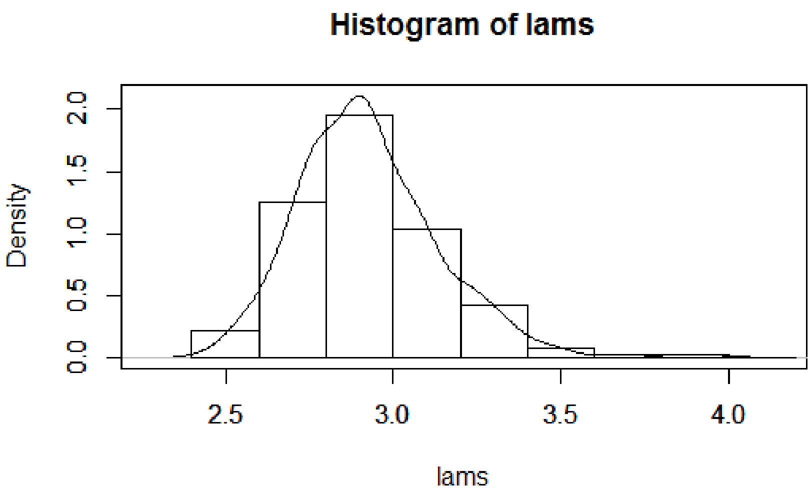

Figure 10: Histogram and Density Estimation for $\bar{\lambda}$

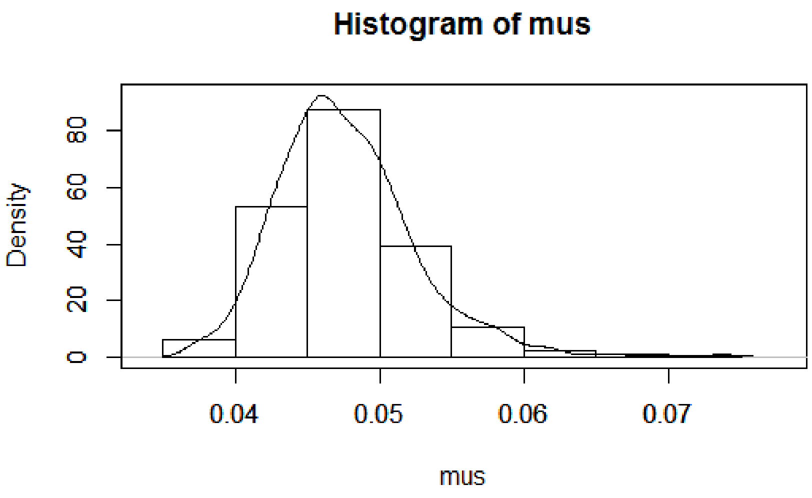

Figure 11: Histogram and Density Estimation for μ

References

Brockwell, P. and R. Dahlhaus (2004). Generalized levinson–durbin and burg algorithms. *Journal of Econometrics 118*(1), 129–149.

Calvet, L. E. and A. J. Fisher (2004). How to forecast long-run volatility: regime switching and the estimation of multifractal processes. *Journal of Financial Econometrics 2*(1), 49–83.

Chen, F., F. X. Diebold, and F. Schorfheide (2013). A markov-switching multifractal inter-trade duration model, with application to us equities. *Journal of Econometrics 177*(2), 320–342.

Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127–1162.

Hall, A. (2005). *Generalized Method of Moments.* Oxford University Press.

Huptas, R. (2009). Intraday seasonality in analysis of uhf financial data: Models and their empirical verification. *Dynamic Econometric Models 9*, 129–138.

Lux, T. (2008). The markov-switching multifractal model of asset returns: Gmm estimation and linear forecasting of volatility. *Journal of business & economic statistics 26*(2), 194–210.

Mandelbrot, B. B., A. J. Fisher, and L. E. Calvet (1997). A multifractal model of asset returns.

Morana, C. and A. Beltratti (2004). Structural change and long-range dependence in volatility of exchange rates: either, neither or both? *Journal of Empirical Finance 11*(5), 629–658.

Segnon, M. and T. Lux (2013). Multifractal models in finance: Their origin, properties, and applications. Technical report, Kiel working paper.

Sun, W., S. Z. Rachev, and F. Fabozzi (2008). Long-range dependence, fractal processes, and intra-daily data. In *Handbook on Information Technology in Finance*, pp. 543–585. Springer.

Žikeš, F., J. Baruník, and N. Shenai (2015). Modeling and forecasting persistent financial durations. *Econometric Reviews*, 1–39.