Dealing with Zero Elements in Duration Data: A Nonparametric Approach

Alex Lewandowski (20458543)

April 22, 2016

Abstract

In this paper we develop a method for dealing with zeros in duration data. The presence of zeros presents issues during econometric analysis, primarily when logarithms or ratios are involved. Currently, the methods for imputation have strict parametric assumptions, which limit the flexibility of their application. We develop a method that is fully nonparametric and is based on long range dependence, a well established stylized fact of duration data. This method is then employed using data from the foreign exchange rate. To our best knowledge, this is the most flexible imputation method available for data exhibiting long range dependence.

Keywords — Imputation, Zeros, Nonparametric, Long Range Dependence, Duration

1 Introduction

Analysts are frequently faced with undesirable zeros in their dataset. These zeros could be actual values, such as in a counting process where there are no occurrences of interest. However, zeros are often a result of truncation or censoring as pointed out by Humphreys (2013). This is often the case when the measurements are not perfectly precise, and successive differences are considered. The resulting differences are then zero when consecutive measurements are too close together. The question is, why is this a problem? Simply, the transformations available to the analyst become limited. Specifically, logarithmic transformations and ratios are not possible if a zero is present. In this paper, we will investigate a method to impute values for zeros as they occur in financial duration.

Today, we have very wide availability of data in all forms. In particular, financial data that was once aggregated to daily, weekly, monthly and quarterly levels can now be analyzed in its purest form; ultra high frequency. For example, ultra high frequency data can be the arrival of a new currency quote in the foreign exchange market or a transaction on a stock exchange. These observations can be radically, slightly or no different from the previous one. Another aspect that makes ultra high frequency data unique is the fact that quotes arrive at irregularly spaced time intervals. Observations 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 observation will occur. This presents a new set of questions on financial duration, which will give insight on the microstructure of markets.

Alleviating the issue of zeros in ultra high frequency data is a form of imputation. So, we would like to develop a method of substituting values in place of zeros. There are already many general methods to impute values for censored, missing or zero values. However, none are specifically designed with duration data in mind. In this respect, this paper deals with a completely new problem. Previous attempts at imputation were ad-hoc methods; Martín-Fernández et al. (2003) deals with zeros in the context of compositional data. The most simple approaches simply delete entries that were undesirable. Other methods would substitute fixed values that were sensible in the context of the problem. Next, formal single stochastic imputation was developed, simulating data from a sensible distribution in the context of the problem. These simulated points would be then substituted for the undesirable ones before statistical analysis. Finally, multiple imputation

was developed by Rubin (2004), the method applies single imputation numerous times. The econometric analysis would then be conducted on each of the data sets, and the result would be averaged. Both single and multiple imputation is discussed in detail by Pigott (2001). In this paper, we will be augmenting the multiple imputation method to suit the context of duration data. In particular, we utilize long range dependence to allow for fully nonparametric multiple imuptation.

The problem of zeros in duration data is of particular note, since the zeros are not missing at random, which is a central assumption to traditional single and multiple imputation. As Sun et al. (2008) points out, duration clusters similar to volatility. Thus, periods of lower duration cluster together which violates the missing at random assumption. Further, the zeros pose problems because duration data is highly leptokurtic, meaning that it exhibits large outliers very frequently. One way to deal with outliers is to take a log transformation, but this isn't possible if there are zero values. Other instances of logarithmic transforms present themselves in Generalized Method of Moment (GMM) estimation. Specifically, Lux (2008) employs a specific class of logarithmic differenced moments to eliminate issues of long range dependence in the estimation of a fractal model. The GMM procedure requires sample moments to be calculated, but this is not be possible if a single zero value is present. Thus, the practical value of dealing with zeros is clear in the context of duration data.

As discussed earlier, multiple imputation is a popular method to deal with missing or zero values. However, strong parametric assumptions hamper its flexibility as pointed out by Aerts et al. (2002). In addition, there are computational concerns when dealing with big data. Ultra high frequency data, which is the origin of duration data, can be very large. This depends on the popularity of the financial item in question, but from our experience, year long datasets of highly traded financial instruments can be larger than 10 gigabytes. Multiple imputation may be infeasible for certain econometric procedures, such as GMM where large matrix calculations must be made. In our paper we consider computationally lenient calculations on two monthly datasets with different ratios of zero to nonzero elements.

The paper is organized as follows, Section 2 deals with the theoretical analysis of the method we propose. Section 3 applies the method to empirical datasets. Lastly, Section 4 summarizes the findings of the paper, provides limitations and points out avenues for further research.

2 Theoretical Analysis

We investigate the basic problem of estimating the duration that is not recorded due to measurement restrictions. However, durations are measured by a difference in successive times. Thus, we start our theoretical section focusing on time and denote t_i as the true time of day that a financial trade or quote is recorded. This can take any value between 0 and $24 \cdot 60 \cdot 60 = 86400$. The measured time $\hat{t_i}$ represents a censored observation of time that is dependent on the precision of the measurement. Then the relationship between the true time and the censored measurement is,

$$t_i = \hat{t_i} + \epsilon_i$$

Where we have that the measurement error is a random variable with some continuous distribution on a bounded interval. As discussed earlier, the interval of the distribution is dependent on the precision of the measurement which is known up to the nearest u, typically 1, 0.1, 0.01 or 0.001. As a result, we have that $0 < \epsilon_i < u$ for all i. This is an assumption on the way the data is collected, but it is very flexible. Essentially, we are assuming that the measurements are taken with precision but truncated past the threshold point u. Alternatively, we can assume that the value is measured with precision and then rounded to the desired threshold u. The result is the same in the nonparametric setting, since we will construct the denisty from the data alone.

Next, we take the difference between successive times to yield the duration. We note that this new random variable does not necessarily have the same distribution as ϵ_i. More formally defined, we have that

$$d_i = t_{i+1} - t_i = \hat{t}_{i+1} - \hat{t}_i + (\epsilon_{i+1} - \epsilon_i)$$

Since we are only imputing values for which $\hat{t}_{i+1} - \hat{t}_i = 0$, this simplifies to $d_i = \xi_i$ where we call ξ_i the residual and have $\xi_i = \epsilon_{i+1} - \epsilon_i$. If we know the distribution of this random variable, we can produce an estimate. However, the distribution is not known and further, it seems that we have no information on how to construct this distribution empirically.

There is an important stylized fact in durations, one which assists in the estimation of these unobserved residuals. Specifically, we consider the long range dependence of durations. Formally, this is defined as the

persistence in autocorrelations, where the autocorrelations decay extremely slowly. Another way to look at it is through a fractal lens, where local behavior mirrors global behavior. The exact relationship between long range dependence and self-similarity can be found in Sun et al. (2008). What this enables us to do is to impose the same density found for the available data and scale it down to the desired interval that we need. This method is not universal, and can only be applied when long range dependence is present. However, this is particularly interesting since there is no literature on the long range dependence on times, while durations are just a function of successive times. A theoretical treatment of this peculiar fact will be an avenue for further research.

This procedure, which we denote as Long Range Density Scaling (LRDS), allows us to replace zeros with sample values from this empirical density. The density is calculated in the standard way using a Gaussian kernel to avoid zeros.

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

Thus, we sample from this density numerous times and perform multiple imputations. The general idea behind multiple imputation is that multiple samples from a distribution average out the bias and produces consistent estimates. If one wanted to calculate $\mathbb{E}(log(d_i))$ for all $1 \leq i \leq N$, using m-imputations, then one would calculate $\frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{j=1}^{N} log(d_j^{(i)})$. We note that $d_j^{(i)}$ is the j^{th} value of the i^{th} sample. We also note that these values are the same if d_j is originally nonzero, but the values will differ if d_j is originally zero. This could be extended to any econometric analysis done on the data. Simply conduct the econometric analysis $m > 1$ times for each randomly drawn sample. In the next section, we observe how our method works in practice.

3 Empirical Analysis

Since the post Bretton-Woods period, econometricians have placed significant emphasis on modeling exchange rates. The foreign exchange market is the largest by far, with trillions of dollars being traded every day by consumers, banks and corporations. 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 investigate the

market microstructure. Our empirical research examines ultra high frequency recordings of the CAD/USD exchange rate during the month of January 2015 and February 2016. The data encompasses quotes from all hours of the day and contains millions of quote observations with time, bid and ask prices. The data was obtained from TrueFX, who 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, there are zeros present in the data. This exact amount of zeros varies with the dataset, with January having 17% of the duration observations as zero. The February dataset, however, has only 0.5% of their observations as zeros.

Jan	Original	Augmented	Feb	Original	Augmented
$\mathbb{E}(d_i)$	0.8452	0.8245	$\mathbb{E}(d_i)$	0.2981	0.2981
$\mathbb{V}(d_i)$	103.3924	103.3924	$\mathbb{V}(d_i)$	56.6342	56.6352
$\mathbb{S}(d_i)$	658.7213	658.7219	$\mathbb{S}(d_i)$	1259.4150	1259.4150
$\mathbb{K}(d_i)$	466.4000	465467.0000	$\mathbb{K}(d_i)$	1669598.0000	1669598.0000

Table 1: Summary statistics for original and augmented datasets

As discussed in the theoretical section, we proceed by kernel density estimation and scale it to the precision threshold $u = 0.001$. That is, $0 < \xi_i < 0.001$ and we draw n samples from this empirical density equal to the amount of zeros in the dataset. Augmented datasets are then constructed with the same values as the original datasets, but with zeros replaced by samples from the empirical density. We choose to run $m = 20$ imputations and take an average of the result. The estimates seem to converge quickly and, as you can see from Table 1, the standard summary statistics are not changed by the imputed values. This is very good and the reason for this is that the outliers tend to dictate the overall behavior, as observed in the extremely large kurtosis value in each dataset.

Next, we calculate moments that we could not have in the original dataset. As discussed before, we are primarily concerned with logarithmic transformations and ratios, the results can be found in Table 2. The logarithms are evaluated nicely but the ratios are explosive and have very high variances. However, this is still much better that the results in the original dataset. The augmented dataset provides a good imputation

for the zero values, since the summary statistics that exist for the original dataset do not change much. The empirical density used, as well as other graphs can be found in the appendix.

Jan	Original	Augmented	Feb	Original	Augmented
$\mathbb{E}(log(d_i))$	$-\infty$	-3.3342	$\mathbb{E}(log(d_i))$	$-\infty$	-2.6294
$\mathbb{V}(log(d_i))$	∞	1.5881e-07	$\mathbb{V}(log(d_i))$	∞	1.2605e-09
$\mathbb{E}(1/d_i)$	∞	1.794e+7	$\mathbb{E}(1/d_i)$	∞	2.4685e+5
$\mathbb{V}(1/d_i)$	∞	1.0332e+14	$\mathbb{V}(1/d_i)$	∞	1.0332e+14

Table 2: Summary statistics for actual and augmented datasets

4 Conclusion

In this paper, we surveyed the current methods for imputation and utilize nonparametric methods of multiple imputation in the context of duration data. Most methods tend to be ad hoc and developed with a certain type of data in mind, and this paper expands the literature. Our theoretical analysis of the problem shows that there exists some unknown distribution for the residual that is truncated or censored after recording. To retrieve estimates of this, we scale the empirical density down to the threshold that the measurements do not observe. In the context of duration data, this is justified by the long range dependence in the data, which is loosely stated as global and local exhibiting similar behavior. This is done m times, generating $m > 1$ datasets. Statistical analysis is then applied to each dataset and the result is averaged. We apply our method to foreign exchange market data and find that the estimates provided after imputation were quite good.

The largest issue with this approach is computational, since duration datasets tend to be quite large. Indeed, performing a statistical analysis on a large dataset may take very long, and multiple imputation requires the result to be reproduced many times. A way to choose samples based on local information, such as K-Nearest Neighbours in Aerts et al. (2002), may improve this method and should be explored in future research. Another issue is that the zeros are not the only values that are truncated. Every value is subject to the same precision threshold, and estimating the residual for every value is a direction for further research.

Specifically, investigating the small and large sample properties of this imputation should be investigated. However, a direct empirical need for imputation for non-zero values is unclear.

Lastly, more precise data would allow for validation of this technique. Currently, there is no data source available that can provide measurements more precise than what is used in our analysis. If more precise data becomes available, we can artificially censor the data and conduct an equivalent analysis as in this paper. Then, the estimated statistics can be compared to the true values using the full uncensored data. We leave this for future research and for those with access to more accurate data.

Appendix

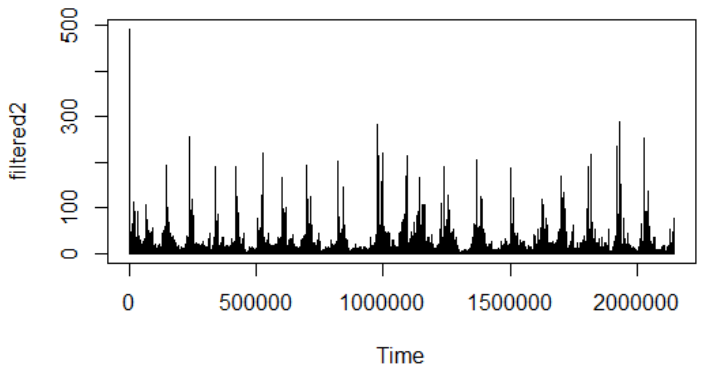

Figure 1: Duration plot for January 2015

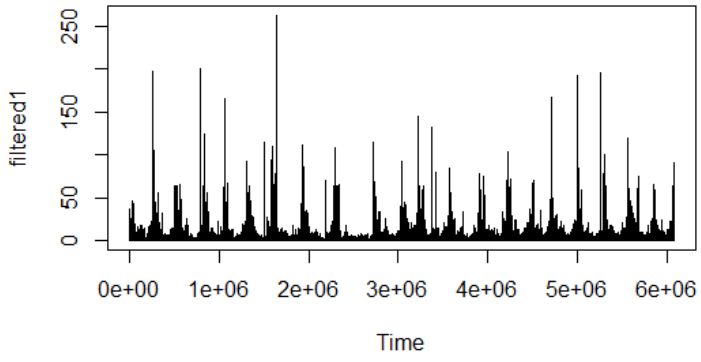

Figure 2: Duration plot for February 2016

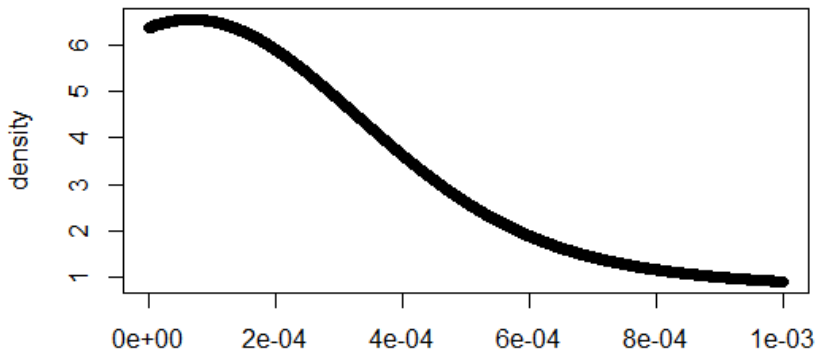

Figure 3: Scaled density for January 2015

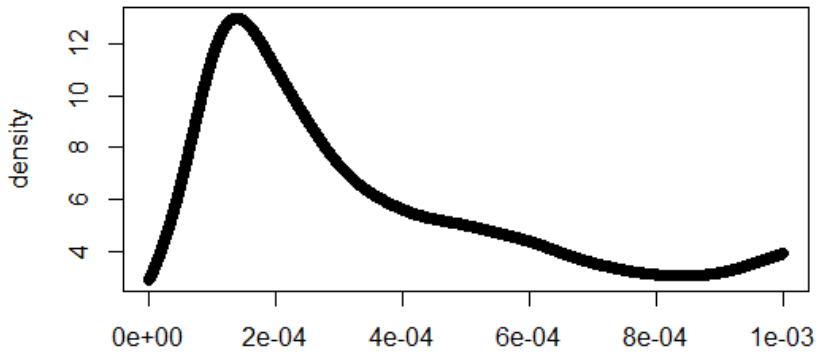

Figure 4: Scaled density for Feburary 2016

References

Aerts, M., G. Claeskens, N. Hens, and G. Molenberghs (2002). Local multiple imputation. *Biometrika 89*(2), 375–388.

Humphreys, B. R. (2013). Dealing with zeros in economic data. *Department of Economics, University of Alberta, Alberta*.

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.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation 7*(4), 353–383.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, Volume 81. John Wiley & Sons.

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.