AN APPROACH TO MODELING ON FINANCIAL TIME SERIES DATA WITH REGIME SHIFTS

Daisuke Yokouchi*, Takeshi Kato** and Yoshimitsu Aoki***

Abstract

This paper proposes a new method to analyze time series data with regime shifts and makes the following three contributions: (1) it suggests an exponential weighted estimation algorithm for autoregressive model with time varying coefficients, (2) it gives a visualization technique of structural change points and an outlier measure based on the Mahalanobis distance and (3) it illustrates that our method works for hedge fund return data and high frequency FX data.

I. Introduction

Many financial time series often show abrupt breaks in their behavior coming from some events like financial crisis or government policy change. Researchers and practicians regard the time series as ones with several regimes shifts, model the behavior by using some time series models, and have applied the obtained model for its prediction, risk measurement and so on. The Hidden Markov (HM) model as originated by L.E. Baum and T. Petrie (1966) is one of the most famous models used for such a purpose and called regime switching model in financial literatures.

The most common HM model in financial literatures is HM autoregressive model which is enabled to switch among several autoregressive (AR) models with different parameters (Hamilton 1989; 1994). In order to fit this kind of HM model, the number of regimes during data term must be explicitly countable. However, it is not so easy to reveal the correct number through statistical data analysis. In addition, if the obtained number of regimes was large, it is quite difficult in fitting the HM model to the data.

GARCH family models and AR model with time varying coefficients have been also applied to represent a time series with regime shift. This is because their time varying parameters including volatility can absorb the differences among each models in all the regimes. However, by using these models, another problem may arise such as the difficulty in catching the regime shifts when they occur from variation of parameters in the charts. In other words, unlike the HM model, these models are not enough to detect change points of regimes.

Moreover, the research works so far assume normal distribution for error term of the

* Hitotsubashi University, yokouchi@ics.hit-u.ac.jp, Japan
** Sophia Universisty, tkskato@sophia.ac.jp, Japan
*** F-viz Corp., aoki@f-viz.co.jp, Japan

models. Consequently those models are not robust to outliers which are often observed at regime switching points like financial crises, to some greater or lesser degree. Therefore the models often don't work for a while after the regime shift happened.

One of the methods to delete infections by outliers is to utilize weighted methods of moments. Some researchers or practicians occasionally take advantage of Exponentially Weighted Moving Average (EWMA), especially for estimation of historical volatility. This is because the EWMA puts more weights on more recent observations. This technique makes the method more robust to outliers by adjusting a parameter of the weight. But it is available for estimation of moments only although its computational cost is low. Another solution to robustness is to replace the normal distribution of the error term to another fat tailed one. Some researchers and practicians often adopt this solution, although it can not solve the issue about change point detection as we mentioned. In addition, since its calculation cost is generally expensive, it is not cost-effective to fit this kind of model to big amounts of time series data, tick data of foreign exchange, for example.

In summary, the models stated above have both advantages and disadvantages to represent timeseries with regime shifts. In order to resolve the disadvantages and to handle big data which we would have no other choice but to deal with in the near future, we have to take the following points into consideration:

- a time weighted estimation method for time series model with time varying parameters which enables us to do high-speed calculation applicable for real time model estimation of tick data, and
- an index for detecting a time point of regime shift from the fitted model.

In this paper, we propose a new framework consisting of the two components, a real-time estimation method for AR(k) model with time varying parameters and a change point detection index created from the model fitting.

II. *Our Framework*

1. An Estimation Algorithm for AR(k) Model with Time Varying Parameters

We assume that x_t ($t=1, 2, \dots, T$) are observations from AR(k) model

$$x_t - \mu_t = \sum_{i=1}^{k} a_{t,i}(x_{t-i} - \mu_{t-i}) + \varepsilon_t \tag{1}$$

where $\varepsilon_t \sim N(0, \sigma_t^2)$ and k is a positive integer. In order to estimate the time varying parameters μ_t, $a_{t,i}$, σ_t^2 from data up to time t, we develop a parameter estimation algorithm based on Yule-Walker equations and idea of EWMA. Note that in this paper we use n-period EWMA $\bar{x}_n$ at t defined by Robert([6]) as

$$\bar{x}_n = \frac{x_t + rx_{t-1} + r^2 x_{t-2} + \cdots + r^{n-1} x_{t-n-1}}{1 + r + r^2 + \cdots + r^{n-1}} \approx r\bar{x}_{n-1} + (1-r)x_t \quad (n \to \infty) \quad . \tag{2}$$

The basic concept of our algorithm is quite simple. In short, for every t, exponentially

weighted mean and autocovariances are calculated from data up to t, substituted the autocovariances into Yule-Walker equations, and solve the equations in order to obtain an estimation of time varying coefficient $a_{t,i}$.

The following shows the details of our algorithm.

Let $l > k$ be a positive integer. Set initial values to

$$\hat{\mu}_l = \frac{1}{l} \sum_{m=1}^{l} x_m$$

$$C_{l,j} = \frac{1}{l} \sum_{m=j+1}^{l} (x_m - \hat{\mu}_l)(x_{m-j} - \hat{\mu}_l) \quad (j=0, 1, \dots, k)$$

In addition let $r \in (0, 1)$ and $\hat{a}_{l,j}$ $(j=1, 2, \dots, k)$ be the solutions of Yule-Walker equation

$$\sum_{i=1}^{k} \hat{a}_{l,j} C_{l,j-i} = C_{l,j} \quad (j=1, 2, \dots, k)$$

where $C_{l,-m} = C_{l,m}$. And an initival value of variance of the error term ε_t is set with

$$\hat{\sigma}_l^2 = \frac{1}{l-k} \sum_{m=k+1}^{l} \left\{ (x_m - \hat{\mu}_l) - \sum_{i=1}^{k} a_{l,i}(x_{m-i} - \hat{\mu}_l) \right\}^2 .$$

For every $(t=l+1, l+2, \dots, T)$, by using equation (2), calculate time varying mean

$$\hat{\mu}_t = r\hat{\mu}_{t-1} + (1-r)x_t \, , \tag{3}$$

time varying autocovariance

$$C_{t,j} = rC_{t-1,j} + (1-r)(x_t - \hat{\mu}_t)(x_{t-j} - \hat{\mu}_t) \quad (j=0, 1, \dots, k) \tag{4}$$

and solve Yule-Walker equation with $\hat{a}_{t,j}$ $(j=1, 2, \dots, k)$.

Note that the equation (4) represents a kind of heteroskedastic structure. This is because $C_{t,0}$ corresponds to an estimated value of σ_t and the estimate $\hat{\sigma}_t$ is dependent on $\hat{\sigma}_{t-1}$. And $r\hat{\sigma}_{t-1}$ determines the persistence in volatilty similar to that of IGARCH(1,1) model: if volatility was high yesterday, it will be still high today.

2.   An Index of Change Point Detection

Regime shifts in financial time series often occur when markets face with crisis or government abruptly change its policy. At the same time, prices and returns on many securities or bonds take extreme values, and the model estimated before the regime shift can not work in many cases. By observing this fact, we establish a method to detect a signal of such a regime shift by helping hand of outlier detection.

Self information, which is often called Shannon entropy (Shannon 1948), is one of the most famous criterion for outlier detection. It provides a measure of uncertainty in a random

variable and is applied to statistical model selection. However, it is not available for continuous random variable like the error term of the equation (1). For gaussian random variable, Mahalanobis distance is well known as a powerful tool for outlier detection. The malahanobis distance D_m of a realization x of random variable $X \sim N(\mu, \sigma^2)$ is defined as

$$D_m(x) = \sqrt{\frac{(x - \mu)^2}{\sigma^2}} \quad . \tag{5}$$

Then, we add the following procedures to our algorithm.

If $t+1 \leq T$, for every t, we calculate predicted value of x_{t+1} at time t,

$$\hat{x}_{t+1} = \hat{\mu}_t + \sum_{i=1}^{k} a_{t,i}(x_{t-i+1} - \hat{\mu}_t)$$

and the mahalanobis squared distance

$$D_m^2(x_{t+1}) = \frac{(x_{t+1} - \hat{x}_{t+1})^2}{\hat{\sigma}_t^2} \quad . \tag{6}$$

When $D_m^2(x_{t+1})$ is quite large, x_{t+1} becomes a candidate of change point.

In order to lay down a clear criterion to regard x_{t+1} as a change point or not, we can utilize x^2 test since $D_m^2 \sim x^2(1)$. Determining a significant level of the x^2 test, we can get rid of arbitrariness of its threshold adjustment.

III.   *Applications of Our Framework*

1.   **Application to Monthly Return Data of Hedge Fund Index**

We introduce an example of application of our framework to empirical data. In this section, we deal with monthly returns of HFRX global hedge fund index as an example. Hedge fund returns are said to be autocorrelated occasionally (Getmansky 2003; Miura 2009). To check the fact of autoregressiveness, we calculated autoregressive models in a rolling method with 12 months data ending at the month t. Figure 1 illustrates the order of autoregression determined by AIC. It tells us monthly the return data of HFRX global hedge fund index seems to have a kind of time varying autoregressive structure since the order of AR increase and decrease as time goes on. Then, we considered that it is appropriate for fitting of our AR model with time varying parameters.

We use the return data of HFRX global hedge fund index from April 2003 to June 2010, and fit the model (1) for $k=1$ to the data by using statistical computing environment R. Note that we rewrite the model (1) for $k=1$ as

$$x_t = a_t + b_t x_{t-1} + \varepsilon_t \tag{7}$$

for the convenience, and we call a_t "intercept" and b_t "coefficient".
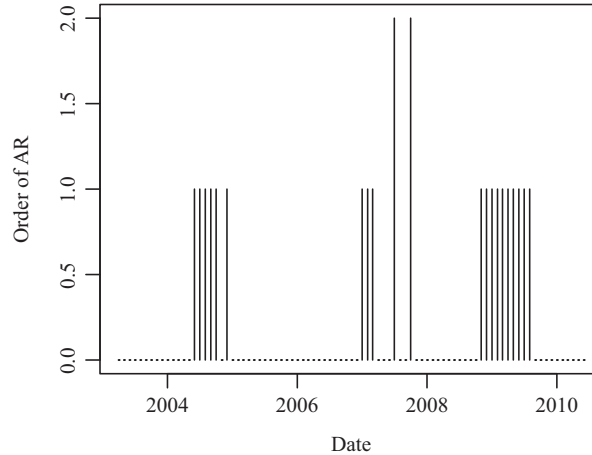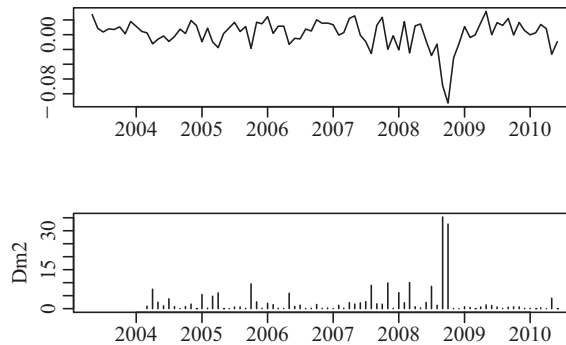
FIGURE 1.   ROLLING ORDER OF AR

FIGURE 2.   HFRX

The upper figure of Figure 5 illustrates the original series and each bar of the lower figure illustrates values of our change point indexes based on Mahalanobis distance for every observations. From the latter half of 2007 to the first half of 2008, somewhat large movements of our indices can be shown. And shortly after, the two large movements can be shown in the lower figure. The time of the two movements clearly corresponds to the time of Lehman's fall which we consider as a regime shift point. Figure 3 illustrates movements of time varying coefficients of the model (7). The upper figure indicates b_t and the lower indicates a_t. After the Lehman's fall, the value of b_t is significantly increased, in fact, it is about double. So we found that the returns of HFRX global hedge fund index become more autocorrelated after the Lehman's fall. If a_t represents so called "hedge fund alpha", we may consider that many hedge funds in the world have difficulty in generating absolute returns after the Lehman's fall because of its abrupt change for investment.

To verify the validity of our change point index, we statistically examine the data before
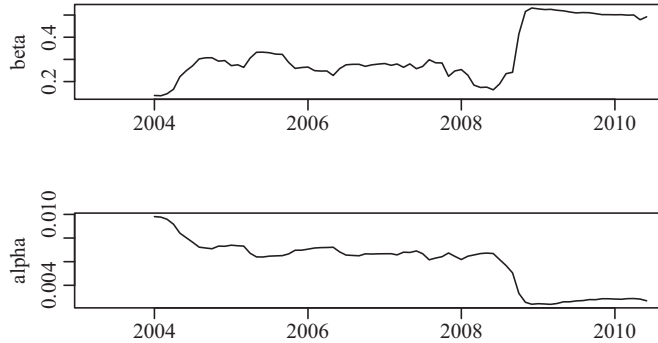
FIGURE 3.   HFRX2

TABLE 1.   DESCRIPTIVE STATISTICS ON HFRX GLOBAL HEDGE FUND INDEX

	period 1	period 2	entire period
Mean	0.003682878	0.003597818	0.001720629
Standard deviation	0.01354429	0.01541439	0.01882657
Shapiro Wilk test	0.09136079	0.485497	3.103738e-08
Ljung Box test	0.1815455	0.271946	1.057707e-05

and after the time when the change point index marked the extremely high values at September 2008 and October 2008. Firstly we eliminated the two records and divide the dataset into the two parts, period 1 from April 2004 to August 2008 and period 2 from November 2009 to June 2010. And we calcluate several descriptive statistics on the datasets and execute both Shapiro Wilk test of normality and Ljung Box test of autocorrelation. Table 1 shows the obtained values of sample mean, sample standard deviation and p values of the two tests. Sample means and standard deviations in the period 1 and 2 are very close to each other. Since the p values in the period 1 and 2 are larger than 0.1, the null hypothesis that returns are not autocorrelated is not rejected respectively. However period 1 has no normality and period 2 has normality since null hypothesis on period 1 is rejected at the significant level of 0.1. That is, each distribution of the two periods are different each other

2.   Application to One Minute Data of Foreign Exchange

As an example of fitting our framework to a large amount of financial timeseries data, we take up one minute's foreign exchange data from US Dollar to EURO. We obtained the data with 7331 observation time points from Bloomberg and fit the model (7) to the data. The data length is equivalent to about 5 days from the morning August 13th to the morning 18th 2012.

As is the case in the example of HFRX global hedge fund index in the previous section, Figure 5 shows a part of original time series and calculation results of our change point index, and Figure 3 shows movements of time varying parameter b_t and a_t within the same time period. From these figures, we found that a_t and b_t move inversely each other, and they abruptly move when a regime shift seems to happen. However, at this time we can not interpret
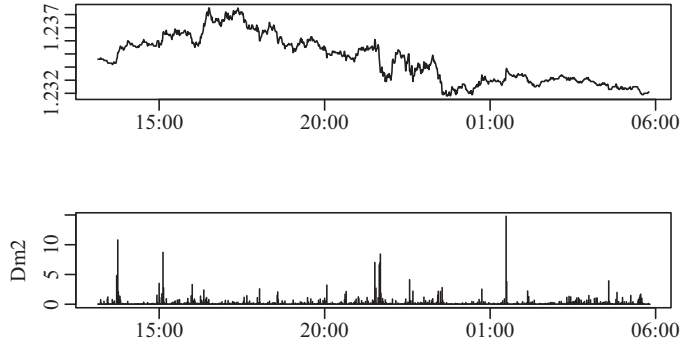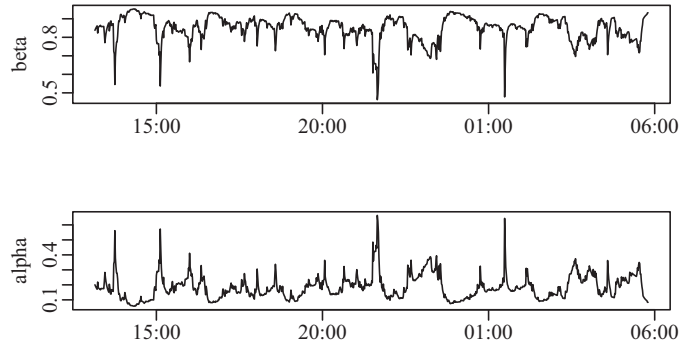
FIGURE 4.   EURO-USD

FIGURE 5.   EURO-USD2

the meaning of these inversely movements since we don't have enough knowledge concerning FX. Calculation cost of our algorithm is so cheap as we mentioned that we could easily fit our framework to the dataset by using R installed in generic laptop PC[1]. In actual, it took only 1.6 CPU seconds to obtain this calculation result although our source code includes several for loop statements which become much computational load for interpreter programming language such as R, MATLAB, and so on.

IV.   *Concluding Remarks*

As was seen in section III, we consider that our framework could work for financial time series which seem to have regime shifts, without using complicated statistical models and their fitting algorithms. In actual, our fitting program in R have about 60 lines in appendix.

Moreover, we also think that it is not difficult to extend our fitting algorithm for univariate timeseries to one for muitivariate timeseries. This is because our algorithm basically work if autocovariance of univariate time series of our algorithm was replaced to autocovariance matrix

---

[1] Lenovo X220, Intel Core i7 CPU 2.80GHz, 8GB RAM

of multivariate time series in section II.1.

APPENDIX

An Implementation of Algorithm in Section II.1 Using R Language

    The following function definition of R is an example of implementation of the algorithm in section II.1. This program is so simple that it works in S-PLUS which is commercial distribution of S language and an ancestor of R. Note that for better code readability we dare to write many for loop statements although it costs more computational time.

```
EWSA=function(x,k=2,l=10,r=0.99){
    mu0=mean(x[1:l])
    C0=acf(x[1:l],plot=F,type="cov",lag.max=k)$acf
    container=NULL
    for(i in 1:k){
        tmp=NULL
        for(j in 1:k){
            tmp[j]=as.numeric(C0)[abs(i-j)+1]
        }
        container=rbind(container,tmp)
    }

    a0=solve(container,C0[2:(k+1)])
    x.hat0=t(a0)%*%(x[l:1][1:k]-mu0)+mu0
    sigma2.0=ar(x[1:l],aic=F,order.max=k)$var
    mu.t=NULL
    C.t=list()
    a.t=list();
    x.hat.t=NULL
    sigma2.t=NULL
    Dm=NULL
    Dl=NULL
    mu.t[l]=mu0
    C.t[[l]]=as.numeric(C0)
    a.t[[l]]=a0
    sigma2.t[l]=sigma2.0

    for(i in 1:(l-1)){
        C.t[[i]]=rep(NA,k+1)
        a.t[[i]]=rep(NA,k)
    }

    for(i in (l+1):length(x)){
        mu.t[i]=r*mu.t[i-1]+(1-r)*x[i]
```

```
        tmpC=NULL
        for(j in 1:(k+1)){
            tmpC[j]=(x[i]-mu.t[i])*(x[i-j+1]-mu.t[i])
        }
        C.t[[i]]=r*C.t[[(i-1)]]+(1-r)*tmpC
        container=NULL
        for(m in 1:k){
            tmp=NULL
            for(n in 1:k){
                tmp[n]=as.numeric(C.t[[i]])[abs(m-n)+1]
            }
            container=rbind(container,tmp)
        }
        a.t[[i]]=solve(container,C.t[[i]][2:(k+1)])
        x.hat.t[i+1]=t(a.t[[i]])%*%(x[i:1][1:k]-mu.t[i])+mu.t[i]
        sigma2.t[i]=r*sigma2.t[i-1]+(1-r)*(x[i]-mu.t[i])^2
        if(i!=length(x)){
            Dm[i+1]=(x[i+1]-x.hat.t[i+1])^2/sigma2.t[i]
        }
    }
    return(list(k=k,mu.t=mu.t,C.t=C.t,a.t=a.t,x.hat.t=x.hat.t,
    sigma2.t=sigma2.t,Dm=Dm,series=x))
}
```

Arguments

- **x** : univariate time series data.
- **k** : order of AR.
- **l** : the number of samples used for computing initial values.
- **r** : weight.

Value

-
- **mu.t** : time varying mean.
- **C.t** : time varying autocovariance.
- **a.t** : time varying coefficients of AR.
- **x.hat.t** : predicted value of AR.
- **sigma2.t** : time varying variance of error term.
- **Dm** : Mahalanobis distance.
- **x** : original time series.

REFERENCES

Baum, L.E. and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics 37 (6), 1554-1563.

Getmansky, M., Lo, A. and Makarov, I. (2003), An econometric model of serial correlation and illiquidity in hedge fund returns. Journal of Financial Economics 74, 529-609.

Hamilton, J.D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica 57, 357-384

Hamilton, J.D. (1994). Time Series Analysis, Princeton, NJ: Princeton University Press

Miura, R., Aoki, Y. and Yokouchi, D. (2009), A Note on Statistical Models for Individual Hedge Fund Returns, Mathematical Methods of Operations Research. Vol.69, 3, 553-577. Springer.

Roberts S.D. (1959) Properties of Control Chart Zone Tests, The Bell System Technical Journal 37, 83-114.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27 (3), 379-423.