社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2009-ICS-154 (22)
2009 / 3 / 4

Identification of Stochastic Process based on Genetic Algorithm

KangrongTan† ShozoTokinaga††

† Faculty of Economics, Kurume University
†† Faculty of Economics, Kyushu University
E-mail: †tankouyuu@kurume-u.ac.jp, ††tokinaga@en.kyushu-u.ac.jp

Identification of Stochastic Process based on Genetic Algorithm

Kangrong TAN† and Shozo TOKINAGA††

† Faculty of Economics, Kurume University 1635 Miimachi, Kurume City, Japan
†† Faculty of Kyushu University 6-19-1 Hakozaki, Higashi-ku, Fukuoka City, Japan
E-mail: †tankouyuu@kurume-u.ac.jp, ††tokinaga@en.kyushu-u.ac.jp

Abstract This study deals with the problem on how to simultaneously estimate the parameters in a stochastic process, especially under some complicated circumstances. Some previous research suggest using χ^2-fitting to estimate these parameters. But, it is certainly difficult to carry a χ^2-fitting with several unknown distributional parameters. Here in this study, we suggest estimating these paramters simultaneously by using Genetic Algorithm (GA). At first we explain Tsallis distribution and entropy model related to the Fokker-Planck equation, which is usually used to describe time-space evolution of particles. Since Tsallis distribution can provide dynamical traces of probability density functions (p.d.f) which evolve over different time spans. Different from conventional Brownian motion, Tsallis distribution is evolving as an anomalous diffusion process, and it includes two types distributions, namely, one is a distribution with finite moments, the other is a distribution with infinite moments. Actually there are several parameters to be optimized simultaneously, it is not easy for some simple χ^2-fitting to estimate. Thus, we propose to use the GA-based procedure to simultaneously optimize parameters of Tsallis anomalous diffusion process. In our numerical studies, we find that our proposed method works well on tracing the whole evolving picture of returns distribution of the High Frequency Data(HFD) in the stock market.

Key words Stochastic process, Tsallis diffusion process, Identification, Genetic Algorithm, High Frequency Data

1. Introduction

Recent worldwide financial crisis, begun with the bankruptcy of the Lehman Brothers Holdings Inc., has caused hefty losses in many financial markets. Many banks have been or prepared to be injected with governmental capital, since the dysfunctional markets liquidity among the banks, security companies, and insurance corporates etc. One of the problems of this crisis is wheather financial institutions have evaluated the right risks related to their daily businesses.

Mathematical models have been built for them to evaluate various financial risks. But, most of them are based on normal assumption. The problems to trace continuous evolution of probability density functions over different time spans are crucial to detect systematic changes in markets. This study deals with Genetic Algorithm (GA) based optimization of Tsallis anomalous diffusion process and its applications to evolution analysis of returns distributions. So far, well-defined models, such as Brownian motion or Geometric Brownian motion are utilized to analyze price changes in markets due to their own simplicity. However, these assumptions are not accurate enough to detect characteristics of the evolution of pdfs over different time spans, and remain serious biases in some cases. As alternatives, modifications of original p.d.fs are proposed by using mixture distributions,

and quantitative analyses have shown that these alternatives can represent the characteristics of returns distribution, such as kurtosis and heavy-tailed behavior of p.d.fs[1]. However, it is well observed that a returns distribution usually evolves over different time spans, and then methods are needed to figure out whole pictures of continuous evolution process of p.d.fs over different time spans[2][3][4][5].

The problem to approximate the probability density function (called p.d.f) or distribution function is still basic and crucial task in various fields such as engineering, economics and finance as well as statistics. Especially, for the analysis of probability corresponding to rare events we must focus on the distribution tails, and sometimes we are puzzled by so-called heavy-tailed or long-tailed distributions. Heavy tails or long tails lead us to overestimation or underestimation of rare events, so that serious accidents such as large packet losses in network traffic and hefty damages to financial assets occur. Thus, more accurate approaches to approximate p.d.fs are necessitated. This study deals with Genetic Algorithm (GA) based optimization of Tsallis anomalous diffusion process and its applications to evolution analysis of returns distributions.

In previous works for the approximation of p.d.f, we find several successful results using the GA [1][6][7][8][9]. By using two typical distributions (Gamma and log-normal distributions), the approximation of p.d.f for natural phenomena such as water flow is shown. But, in these cases the complexity of the mixture distribution is limited. As alternatives, modifications of original p.d.fs are proposed by using mixture distributions, and quantitative analyses have shown that these alternatives can represent the characteristics of returns distributions, such as kurtosis and heavy-tailed behavior of pdfs. For the prediction of error distribution or the generation of random numbers, the combination of multiple p.d.fs is used to generate variables by optimizing the weight among p.d.fs. But, the application of the mixture distribution is oriented only for the error estimation and random number generation and the basic p.d.f is limited. Moreover, it is well observed that a returns distribution usually evolves over different time spans, and thus methods are needed to figure out the whole picture of continuous evolution process of p.d.fs over different time spans.

As to grasp the whole picture of how a returns distribution evolves dynamically, we propose to model it as a Tsallis distribution as shown in previous work [4]. But, the authors haven't shown that, how to optimize several parameters in Tsallis distribution simultaneously, instead, it just states that the parameters are obtained from a simple χ^2-fitting without shown readers any algorithm used in this fitting. Usually, it is difficult to estimate several paramters by only using χ^2-fitting. Thus, we propose to use Genetic Algorithm (GA) to optimize these statistical parameters simultaneously, since GA has the ability to reach a global optimal solution without stuck in local ones [5][6][7][8][9].

Firstly we explain Tsallis distribution and entropy model related to the Fokker-Planck equation which is usually used to describe time-space evoluation of particles[2][3]. It is assumed that p.d.fs are usually time variant (time dependent), and are described by time t as well as stochastic variable x. The process (evolution) of the p.d.fs is described by the Fokker-Planck equation. Different form conventional Brownian motion, Tsallis distribution is related to an anomalous diffusion process, and it includes two types distributions, namely, one is a distribution with finite moments, the other is a distribution with infinite moments.

Secondly we show how to use the GA-based procedure to optimize parameters of Tsallis anomalous diffusion process. Since Tsallis distribution can provide dynamical traces of p.d.fs which evolve over different time spans. In our numerical studies, we apply our proposed method to identifying the evolution of real stock returns and find that our method works well on tracing the whole evolving picture of returns distributions.

The rest of this study is organized as follows. Section 2 summarizes the basic properties of Tsallis distribution and Fokker-Planck equation, and shows how to optimize the parameters by using GA. Section 3 briefly reviews and summarizes the evolution evidence of returns distributions over varying time spans, and presents some applications and their numerical results with real market data sets.

2. Tsallis anomalous diffusion process

2.1 Tsallis entropy and Fokker-Planck equation

It is assumed that p.d.fs are usually time variant, and are described by time t as well as stochastic variable x. The evolution process of p.d.fs can be described by the Fokker-Planck equation. Tsallis entropy is defined as follows.

$$S_q = -\frac{1}{1-q}(1 - \int P(x,t)^q dx \qquad (1)$$

It is clear that S_q will converge into a usual entropy when q takes limit to 1, namely, $S = -\int PInP$. Here, $P(x,t)$ is probability density function at time t, and parameter q is independent of time t. So as to insure the consistency of a p.d.f, the following equations are imposed as constraints. Equation (2) works as a constraint to make $P(x,t)$ as a p.d.f in common sense. Equation (3)(4) and (5)(6) are so called q-mean, and q-variance. They are different from usual mean and variance unless $q = 1$.

$$\int P(x,t)dx = 1 \tag{2}$$

$$E(x-\bar{x}(t))_q = \int (x-\bar{x}(t))P(x,t)^q dx \tag{3}$$

$$= 0 \tag{4}$$

$$E(x-\bar{x}(t))_q^2 = \int (x-\bar{x}(t))^2 P(x,t)^q dx \tag{5}$$

$$= \sigma_q(t)^2 \tag{6}$$

By maximizing the Tsallis entropy constrained by above equation (2)-(6) for some fixed q, it yields,

$$P(x,t) = \frac{1}{Z(t)}(1+\beta(t)(q-1)(x-\bar{x}(t))^2)^{\frac{1}{1-q}} \tag{7}$$

Where $Z(t)$ and $\beta(t)$ are Lagrange multipliers corresponding to equation (2) and (5)(6).

$$Z(t) = \frac{B(\frac{1}{2}, \frac{1}{q-1}-\frac{1}{2})}{\sqrt{(q-1)\beta(t)}} \tag{8}$$

$$\beta(t) = \frac{1}{2\sigma_q(t)^2 Z(t)^{q-1}} \tag{9}$$

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{10}$$

Since

$$\sigma^2(t) = E(x-\bar{x}(t))^2 \tag{11}$$

then

$$\sigma^2(t) = \begin{cases} \frac{1}{(5-3q)\beta(t)}, & q < \frac{5}{3} \\ \infty, & q \geq \frac{5}{3} \end{cases} \tag{12}$$

Where q should be less than $\frac{5}{3}$ if it has a distribution with finite variance, otherwise it has a distribution with infinite variance. The value of q in practice is usually fitted by the real market data sets.

For the nonlinear Fokker-Planck equation

$$\frac{\partial P(x,t)^\mu}{\partial t} = -\frac{\partial}{\partial x}(F(x)P(x,t)^\mu) + \frac{D\partial^2 P(x,t)^\nu}{2\partial x^2} \tag{13}$$

It can be solved by Equation (7) when $q = 1+\mu-\nu$. Where F(x) is supposed to be a linear drift term, namely, $F(x) = a - bx$. And here if

$$\frac{dx}{dt} = F(x) + \sqrt{DP(x,t)^{1-q}}\xi(t) \tag{14}$$

where $\xi(t)$ is a Gaussian noise, namely,

$$< \xi(t)\xi(t') >= \delta(t-t') \tag{15}$$

Where the diffusion coefficient term is $DP(x,t)^{1-q}$, and it is called as subdiffusion in the case $q < 1$, and called as superdiffusion in the case $q > 1$. It is clearly different from the normal Brownian motion, since the diffusion coefficient term is only D in the normal Brownian motion. Therefore, this model can be used as to fit nonlinear diffusion process. The application details of the scheme are discussed in the numerical experiments in Section 3.

2.2 GA-based parameter optimization of Tsallis distribution

2.2.1 Why GA-based method is adopted

Let us simply explain why we propose to use GA-based (Genetic Algorithm:GA) method here. As mentioned above, previous work [4] suggests to use a χ^2-fitting to estimate the statistical parameters in Tsallis distribution. But, the authors haven't shown that, how to optimize several parameters in Tsallis distribution simultaneously, instead, it just states that the parameters are obtained from a simple χ^2-fitting without shown readers any algorithm used in this fitting. Usually, it is difficult to estimate several paramters by using χ^2-fitting. Thus, we propose to use Genetic Algorithm (GA) to optimize these statistical parameters simultaneously, since GA has the ability to reach a global optimal solution without stuck in local ones [8].

Namely, it is hard for one to estimate parameters q, and $\beta(t)$ or $Z(t)$ simultaneously. It becomes more complicated when there are several data sets available for several different time spans. It is necessary to consider each fitting result of each different time span. Therefore, it turns out to be a multiobjective optimization problem. Usually it is not easy to get optimal solution in dealing with such a multiobjective optimization problem. Usual optimization methods probably converge into some local optimal solutions.

So far, GA, as one of the most efficient optimization methods, which converges rather into a global solution than a local one in search of optimal solution, has been widely applied in many research fields ranging from scientific researches to social studies [8].

2.2.2 Our proposed GA scheme

Here, suppose that we have several data sets for different time spans, then we can get several likelihood functions for the data sets, say, $L_0, L_1, ..L_{m-1}$ which share the same parameter q, with different parameters $\beta(t_0)$, $\beta(t_1)$, ..., $\beta(t_{m-1})$. Let $V = \Sigma_{i=0}^{m-1} L_i$, we consider that, the optimal solution is a set of q, $\beta(t_0)$, $\beta(t_1)$, ..., $\beta(t_{m-1})$ which makes V reach the maximum.

Our GA scheme is designed as follows.

Step 1: Initial population

Generate random numbers as individuals of the first generation with certain population. Here, each individual represents a set of parameters in P(x,t) (equation (7)), namely, q, and $\beta(t)$ or $Z(t)$.

Step 2: Evaluation of fitness

Evaluate the fitness of each individual based on predetermined fitness function, then to sort all individuals of the generation according to their fitness values.

Step 3: Selection of individuals

Select two individuals with higher fitness values from the

generation at a certain probability. The selection strategy has a great deal of variations, Roulette strategy is adopted in our applications.

Step 4: Genetic operations

Carry out genetic operations, namely, crossover operation, and mutation operation to two selected individuals to reproduce their offsprings and put them into the pool of next generation.

Here, a crossover operation means to randomly decide crossover positions on the two selected individuals at first, then to exchange parts of two individuals each other. Basically, there are two methods to do this, one is one-point crossover, the other is multipoints crossover. The later one is applied to our application. A mutation operation means to randomly decide mutation positions under a certain probability, and then to change those position values of a selected individual. It also has two ways to do this. One is one-point mutation, the other is multipoints mutation. The later one is adopted in our application.

Step 5: Replacement of individuals

Reevaluate the fitness of each individual of the new generation, to see if the results meet the terminal conditions, such as repeating times, or error range etc. If it does, then GA terminates, else it goes back to Step 3.

And fitness function for evaluating jth individual is defined as

$$Fitness_j = \frac{1}{(n-1)} \frac{V_j}{\sum_j V_j} \tag{16}$$

where V_j is a sum of likelihood values corresponding to L_is.

Overall, the differences between our GA-based method and the previous work [4] are as follows.

1)Our GA-based method optimizes the parameters simultaneously. In previous work [4], it suggested a χ^2-fitting to do it, which is not easy to be done under the circumstance of several parameters unknown.

2)Our GA-based parameter-fitting is based upon multiple evolving distributions over different time spans, not depending on only one of the evolving distributions as shown in [4]. Our method is to estimate the parameters of the evolving family, it is a multiobjective optimization approach.

3. Applications

3.1 Evolution of returns distribution

Prior to simulation studies for the evolution analysis of time series of real stock returns, we briefly summarize the significance to trace the changes of statistical properties of returns. It is seen that returns obtained from stock prices are evaluated over varying time spans, such as, one hour, one day, one week, one month, one year, and so on. Usually their statistics such as kurtosis, standard deviation tend to bear different distributions over different time spans. Seemingly, the evolving distribution is getting closer to normal distribution, as the time span is getting longer. However, in fact, it can be shown that most of p.d.fs are not normals. Normality will be rejected by statistical tests, such as Jarque-Bera test.

Here, we examine two different stock returns. One of them consists of the daily returns of Standard & Poors Index. The other comes from high frequency transaction data of IBM stock. Here, return r_t at time t is computed by $r_t = logp_t - logp_{t-1}$.

Firstly, descriptive statistics kurtosis and standard deviation of returns of S&P Index over different time spans are summarized in Table-1.

Stock A	Kurtosis	S.D.
1-day	35.15	0.00389
7-day	8.067	0.00984
14-day	5.844	0.01423
21-day	3.854	0.01772
28-day	3.214	0.02048
56-day	2.550	0.02896
112-day	0.839	0.04148
224-day	0.257	0.06028

Table-1: Evolution of kurtosis and standard deviation of daily returns of S&P Index over different time spans

As is seen from Table-1, kurtosis and standard deviation are getting nearer and nearer to normal as time span evolves longer. But, it is seen that their statistical properties never become to be identical to the normal distribution.

Secondly, kurtosis and standard deviation of returns of IBM stock over different time spans are summarized in Table-2. Figure-1 shows the plot of the time series of returns.

IBM	Kurtosis	S.D.
5-sec	5804.676	0.000284
10-sec	2452.795	0.000419
20-sec	461.969	0.000547
40-sec	294.132	0.000717
80-sec	195.777	0.000938
160-sec	156.321	0.001181
320-sec	85.974	0.00164
640-sec	60.107	0.00225
1280-sec	36.856	0.00317
2560-sec	59.566	0.00432

Table-2: Evolution of kurtosis and standard deviation of tick returns of IBM stock over different time spans

It is also seen from Table-2, just as in the Table-1, kurtosis and standard deviation of returns of IBM stock are getting nearer and nearer to normal as time span evolves longer. But, they never reveal as the statistical properties of usual normal distribution.

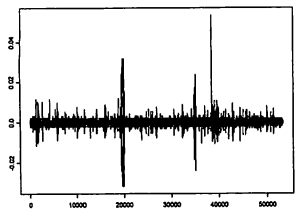

Figure-1: Plots of returns of IBM Stock

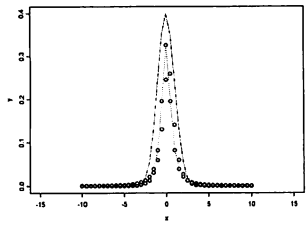

Figure-2: Plot of p.d.f of S&P day-1

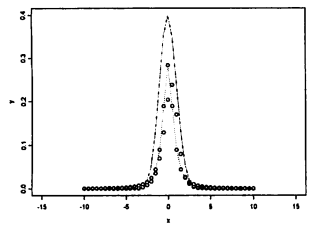

Figure-3: Plot of p.d.f of S&P day-14

3.2 Applications for evolution analysis

In this section, we apply our proposed method to real market data set A (S & P Index) and B (IBM Stock Prices) obtained from real stock prices. And all the returns of these two stocks are standardized to have mean of 0 and standard deviation of 1. The data set A is consisting of daily stock prices. The procedure of this research is applied to estimating Tsallis distribution with 1-day, 14-day, 28-day returns. Parameters of the GA are selected as follows.

Population size : 200

Crossover and mutation probabilities: 0.42 and 0.31 respectively, $q \in (1, 5]$, and $\beta(t) \in [0.001, 50]$.

Furthermore, we employ elite-keeping policy in GA. An elite-keeping policy is understood to copy an individual with higher or highest fitness into next generation automatically. We repeat GA procedure for sufficient times and get the same global optimal solution. The estimated values of q and $\beta(t_0)$ are 2.23, and 1.68 respectively. It is seen that it is a superdiffusion process with infinite variance since $q = 2.23 > \frac{5}{3} > 1$ holds.

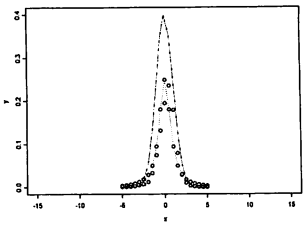

Figure-4: Plot of p.d.f of S&P day-28

We show the results of the estimated distributions and the empirical distributions from Figure-2 to 4. In all figures, the outside dot-dash line denotes Standard Normal distribution, the inner doted line denotes the estimated Tsallis distribution, and the circls denotes returns samples. It is seen from these figures, the numerical results that the estimated $P(x,t)$ dynamically traces the evolution of returns distribution over varying time spans.

On the other hand, both in academic research and business practice, returns are often modeled as a process of Brownian motion, or Geometric Brownian motion, such as, $r_T = logp_T - logp_0 \sim \phi((\mu - \frac{\sigma^2}{2})T, \sigma\sqrt{T})$, where $\phi(m,s)$ denotes a normal distribution with mean m and standard deviation s. Compared to the Standard Normal distribution in these figures, it is clear that Tsallis distribution catches the evolution behavior of returns distribution much better than the normal distribution does.

So as to show the capability of the proposed method, we apply the same procedure to data set B, returns of high frequency transaction data (High Frequency Data) of IBM. The evolution is characterized by Tsallis anomalous diffusion process shown from Figure-5 to 7, for the time spans (interval) with 20 seconds, 80 seconds and 320 seconds, respectively. By applying GA procedure to it, we get the estimated values of q and $\beta(t_0)$ are 1.01, and 0.04 respectively. It is seen that it is a superdiffusion process with finite variance since q =1.01< $\frac{5}{3}$ holds.

Seen from Figure-5 to 7, the tallest outside dot-dash line denotes Standard Normal distribution, the inner doted line denotes the estimated Tsallis distribution, and the circls denotes returns samples, respectively. It is seen from these figures, the numerical results show that the estimated $P(x,t)$ almost exactly overlaps where the returns samples lies, and dynamically traces the evolution of returns distribution over varying time spans under the circumstances of tick data as well as the daily data above. Meanwhile, compared the estimated Tsallis distribution to the Standard Normal distribution in these figures, it is clear that Tsallis anomalous diffusion process catches the evolution behavior of returns

distribution much better than Brownian motion or Geometric Brownian motion does.

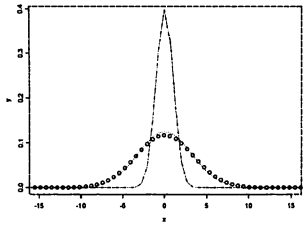

Figure-5: Plot of pdf of IBM sec-20

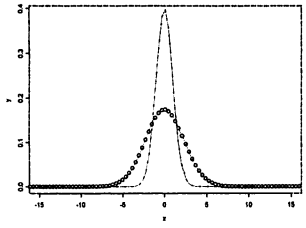

Figure-6: Plot of pdf of IBM sec-80

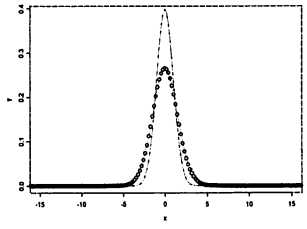

Figure-7: Plot of pdf of IBM sec-320

4. Conclusion

In this research, we showed GA based optimization of Tsallis anomalous diffusion process and its applications to evolution analysis of returns distribution. We explained how to model returns distribution using Tsallis distribution and entropy model in relation to the Fokker-Planck equation, and then we showed how to use the GA-based procedure to optimize parameters of Tsallis anomalous diffusion process using a multiobjective approach. In numerical studies, we found that our proposed method works well on tracing the whole evolving picture of returns distribution, even under the circumstance of High Frequency Data. It is important for us to do the further researches to see how the Value at Risk (VaR) or Conditional Value at Risk (CVaR) changes over the different time spans in risk management or risk measurement.

The authors would like to thank Professor Joe Gani for his precious comments that significantly improved this research. And Professor Chris Heyde for his valuable discussion, unfortunately he passed away last March. This research was partly supported by the Japan Society for the promption of Science under the grant number (B)19310099 and (C)19510164, and the authors would like to thank the organization.

Reference

[1]Tan, K., and Tokinaga, S. (2006). Identifying returns distribution by using mixture distribution optimized by Genetic Algorithm, *Proc. of NOLTA2006*, pp.119-122, 2006.

[2]Tsallis, C., and Bukman, D. J. (1996). Anomalous diffusion in the presence of external forces: exact time-dependent solutions and entropy, Physical Review E **54**, pp.2197.

[3]Heyde, C., C., and Leonenko, N., N. (2005). Student processes, *Appl. Prob*, Vol. **37**, pp.342-365.

[4]Micheal, F., and Johnson, M., D. (2003). Financial market dynamics, *Physica A: Statistical mechanics and its applications*, vol. **320**.

[5]Tan, K., and Tokinaga, S. (2007). Genetic Algorithm-based parameter optimization of Tsallis distribution and its application to financial market, *Proc. NOLTA2007*.

[6]Bhattacharjya, R.K. (2004). Optimal design of unit hydrographs using probability distribution and genetic algorithm, *Sadhana*, vol. **29**, Part **5**, pp.499-508.

[7]Dutre P., Suykens F., and Wilems Y. (1998). Optimized Monte Carlo path generation using Genetic Algorithm, *Report CW267*, Department of computer science, Katholieke Universiteit Leuven.

[8]Goldberg D.(1989). Genetic Algorithm: in Search, Optimization, and Machine Learning, Addison-Wesley Press.

[9]Tan, K., and Tokinaga, S. (1999). Optimization of fuzzy inference rules by using the genetic algorithm and its application to bond rating, *Journal of Operations Research Japan*, vol. **42**, **3**, pp.302-315.