Nonlinear Correlation Tracking Technique in Data Mining of Financial Markets*

YI Dong-yun, ZHANG Wei-ming, DU Xiao-yong

(Department of Mathematics and System Science, National University of Defense Technology, Changsha 410073, China)

E-mail: dyyi@nudt.edu.cn

Received November 25, 1998; accepted October 18, 1999

Abstract: Financial data mining is one of the most challenging research directions in information society. Financial data with random characteristics make it difficult to find out the rule hidden in data. In this paper, it is pointed out that correlation coefficient can not capture nonlinear information, which is the serious defect of classic correlation analysis. Furthermore, the properties of the high-order correlation coefficient are discussed, and it is proved that high-order correlation can not only describe the hidden nonlinear correlation, but also fill up the space between classic correlation and independence. The computational simplicity makes the high-order correlation coefficient be an effective technique to track nonlinear relation between variables. Finally, the above results are applied to the correlative analysis between stock price and stock trading volume, and the computing results show that the high-order correlation coefficient can track the time-varying nonlinear characteristics.

Key words: nonlinear analysis; data mining; financial data

Financial data mining is one of the most challenging research directions in information society[1,2]. Financial data are of random characteristics, which makes it difficult to find out the rule hidden in data. The traditional assumption was founded upon the theory of market efficiency, which stated simply the so-called "random walk" model in statistical terms. Yet this posed a serious dilemma between theory and practice as trader did continue to make profits in short term[3].

Econometric tests specified a more general model for the time sequence behavior of asset returns, including auto-regressive and other terms, but it only provided a linear structure. Generally speaking, it is difficult to describe the nonlinear structure of capital markets. Based on the correlation dimension, Brock W. put forward a method that can test for independence in 1987[4], and then presented a further general test for nonlinear Granger Causality in 1992[5]. N. Refenes conducted the study of neural network method in 1997[6]. It turned out that under those model-based test procedures, it is possible to reject the special hypothesis of random walk. Brock's

* This project is supported by the National Natural Science Foundation of China under Grant Nos. 60003013 and 69872039 (国家自然科学基金). **YI Dong-yun** was born in 1965. He is an associate professor in Department of Mathematics and System Science of National University of Defense Technology. His research interests are in dynamic data mining, network data mining, multiagent system and financial engineering. **ZHANG Wei-ming** was born in 1962. He is a professor in Department of Information System Engineering and Management Science of National University of Defense Technology. His research interests are in information system engineering and intelligent decision system. **DU Xiao-yong** was born in 1976. He is an M. S. student in Department of Mathematics and System Science of National University of Defense Technology. His research interests are in nonlinear system analysis and information fusion.

model needs the normal distribution hypothesis and the above models have the defect that they can not illustrate the time-varying dynamic characteristics of nonlinear structure of capital markets.

In this paper, high-order correlation coefficient is discussed, and for the first time we have proved that high-order correlation can fill up the space between correlation and independence. The high-order correlation coefficient is easy to compute and the computing results show that it can capture the time-varying characteristics of financial data, which is very useful to build dynamic financial prediction model.

The description of relationship between random variables involves two concepts: one is covariance $\mathrm{Cov}(X,Y)$, or correlation coefficient $\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)^{1/2}\mathrm{Var}(Y)^{1/2}}$, where $\mathrm{Var}(X)$ is the variance of variable X; the other is independence, or $F(x,y)=F(x)F(y)$, where $F(x,y)$ is a joint distribution function and $F(\cdot)$ is a distribution function. The correlation coefficient is easy to compute whereas the independence is difficult to justify.

It is all known that if X and Y are independent, then they are not correlative, but the opposite is not true. From view of statistics, if X and Y are not correlative, it only shows that there is no obviously linear correlation between X and Y, but it is unknown whether nonlinear relation exists or not.

Example 1. $|\rho_{XY}|=1\Leftrightarrow Y=aX+b$, a.s..

Example 2. Suppose that $Y=X^2$ and $X\sim N(0,1)$, where $N(0,1)$ is a normal distribution with zero mean and unit variance, then $\mathrm{Cov}(X,Y)=0$, and $\rho_{XY}=0$.

Example 1 clearly shows the linear nature of correlation coefficient, whereas Example 2 shows the serious defect, i.e., correlation coefficient can not capture nonlinear information. Hence, it is the key problem to determine the space between correlation and independence.

In next section, this problem will be completely solved with the definition of high order correlation coefficient.

1 Definition and Theorem

We first illustrate an example.

Example 3. Let $Y=\cos(X)$, $X\sim U(0,2\pi)$, where $U(0,2\pi)$ is a uniform distribution. Then we can get the following results

$$\rho_{XY}=0.0491,\ \rho_{X^2Y}=0.234,\ \rho_{X^3Y}=0.345,\ \rho_{X^4Y}=0.463,\ \rho_{X^5Y}=0.517.$$

The computed results show that $\mathrm{Cov}(X^k,Y)$ or ρ_{X^kY}, $k=1,2,3\ldots$, can describe the hidden nonlinear correlation, so we introduce high-order correlation.

Definition 1. If there exist two positive integers k,l such that $\mathrm{Cov}(X^k,Y^l)\neq 0$, then we say there is (k,l)-order correlation between X and Y. $\mathrm{Cov}(X^k,Y^l)$ is called (k,l)-order covariance and $\rho_{X^kY^l}$ is called (k,l)-order correlation coefficient.

It is obvious that ρ_{XY} is the simplest situation while $k=l=1$. If $k>1$ or $l>1$, then we say there is high-order correlation between X and Y. In Example 2, we can get $\rho_{X^2Y}=1$, i.e., X and Y are of $(2,1)$-order correlation.

Theorem 1. Suppose that X and Y are two random variables with $|EX^k|<\infty$, $|EY^l|<\infty$, $|EX^kY^l|<\infty$, then X and Y are independent if and only if $\rho_{X^kY^l}=0$, for $k,l=1,2,3,\ldots$.

Proof. First note that $E(X^k-EX^k)(Y^l-EY^l)=EX^kY^l-EX^kEY^l$, so $\rho_{X^kY^l}=0$ is equivalent to $EX^kY^l-EX^lEY^k=0$. The necessity is obvious from the property of independent random variables. The sufficiency is proved as follows.

Suppose $EX^kY^l-EX^kEY^l=0$ for $k,l=1,2,\ldots$. Let $F_1(x),F_2(y),F(x,y)$ be the distributed functions of random variables X,Y and random vector (X,Y) respectively. Thus the characteristic functions of X,Y and

(X,Y) can be respectively denoted as $f_1(s),f_2(t),f(s,t)$, where

$$f_1(s)=Ee^{isX}=\int_{R^1}e^{isx}dF_1(x),\ f_2(t)=Ee^{itY}=\int_{R^1}e^{ity}dF_2(y),$$

$$f(s,t)=Ee^{i(sX+tY)}=\iint_{R^2}e^{i(sx+ty)}dF(x,y).$$

Considering the existence of EX^k, EY^l and EX^kY^l, it can be obtained that $EX^k=i^{-k}f_1^{(k)}(0)$, $EY^l=i^{-l}f_2^{(l)}(0)$, $EX^kY^l=i^{-(k+l)}\left.\frac{\partial^{k+l}f(s,t)}{\partial s^k\partial t^l}\right|_{(s,t)=(0,0)}$. From the given condition $EX^kY^l-EX^kEY^l=0$, we get

$$\left.\frac{\partial^{k+l}f(s,t)}{\partial s^k\partial t^l}\right|_{(s,t)=(0,0)}=f_1^{(k)}(0)\cdot f_2^{(l)}(0). \tag{1}$$

In terms of Taylor's expansion formula, the following expressions can be derived

$$f_1(s)=f_1(0)+f'_1(0)\cdot s+\ldots+\frac{1}{n!}f_1^{(n)}(0)\cdot s^n+\ldots \tag{2}$$

$$f_2(t)=f_2(0)+f'_2(0)\cdot t+\ldots-\frac{1}{n!}f_2^{(n)}(0)\cdot t^n+\ldots \tag{3}$$

$$f(s,t)=f(0,0)+\left.\frac{\partial f(s,t)}{\partial s}\right|_{(s,t)=(0,0)}\cdot s+\left.\frac{\partial f(s,t)}{\partial t}\right|_{(s,t)=(0,0)}\cdot t+\ldots+\frac{1}{n!}\sum_{k=0}^{n}C_n^k\left.\frac{\partial^n f(s,t)}{\partial s^k\partial t^{n-k}}\right|_{(s,t)=(0,0)}\cdot s^k\cdot t^{n-k}+\ldots \tag{4}$$

With Eq. (1) and Eq. (4), we have

$$f(s,t)=\sum_{n=1}^{\infty}\sum_{k=0}^{n}\left(\frac{1}{k!}f_1^{(k)}(0)\cdot s^k\right)\left(\frac{1}{(n-k)!}f_2^{(n-k)}(0)\cdot t^{n-k}\right). \tag{5}$$

For any real numbers s and t, there exist two real numbers $\bar{s}$ and $\bar{t}$ with $|s|<|\bar{s}|$, $|t|<|\bar{t}|$. Considering Eqs. (2) and (3), power series $\sum_{n=0}^{\infty}\frac{1}{n!}f_1^{(n)}(0)\cdot\bar{s}^n$ converges to $f_1(\bar{s})$ and $\sum_{n=0}^{\infty}\frac{1}{n!}f_2^{(n)}(0)\cdot\bar{t}^n$ to $f_2(\bar{t})$ respectively. Then, the two series, $\sum_{n=0}^{\infty}\frac{1}{n!}f_1^{(n)}(0)\cdot s^n$ and $\sum_{n=0}^{\infty}\frac{1}{n!}f_1^{(n)}(0)\cdot t^n$, are absolutely convergent by the property of power series. With the product property of absolutely convergent series, we have

$$f_1(s)\cdot f_2(t)=\left(\sum_{n=0}^{\infty}\frac{1}{n!}f_1^n(0)\cdot s^n\right)\cdot\left(\sum_{n=0}^{\infty}\frac{1}{n!}f_2^n(0)\cdot t^n\right)=$$

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n}\left(\frac{1}{k!}f^{(k)}(0)\cdot s^k\right)\left(\frac{1}{(n-k)!}f^{(n-k)}(0)\cdot t^{n-k}\right)=f(s,t).$$

This expression indicates that the characteristic function of random vector (X,Y) can be expressed as the product of those of random variables X and Y, i.e., random variable X is independent of Y. So we prove the theorem.

Theorem 1 firstly gives an essential description of the relation between correlation and independence, i.e., if there is no linear relation between X^k and Y^l for every $k,l=1,2,3,\ldots$, then X and Y are independent. Theorem 1 shows that high-order correlation fills up the space between classic correlation and independence. The following corollary gives an equivalent condition.

Corollary 1. Under the condition of Theorem 1, X and Y are independent if and only if $\mathrm{Cov}(E(X/Y),E(Y/X))=0$, where $E(X/Y)$ is the conditional expectation of X to Y.

Proof is omitted.

Moreover, considering the difference between variable and dependent variable, we give the following utilitarian definition.

Definition 2. If there exists a positive integer k such that $\mathrm{Cov}(X^k,Y)\neq 0$, then we say there is k-order correlation between variable X and dependent variable Y. If $k>1$, then we say there is high-order correlation between variable X and dependent variable Y.

Similarly we have the following results.

Theorem 2. Suppose that X and Y are two random variables with $|EX^k|<\infty$, $|EY|<\infty$, $|EX^kY|<\infty$, then that variable X and variable Y are independent is equivalent to one of the following conditions:

(1) $\rho_{X^kY}=0$, for $k=1,2,3,\ldots$;

(2) $\mathrm{Cov}(Y,E(Y/X))=0$.

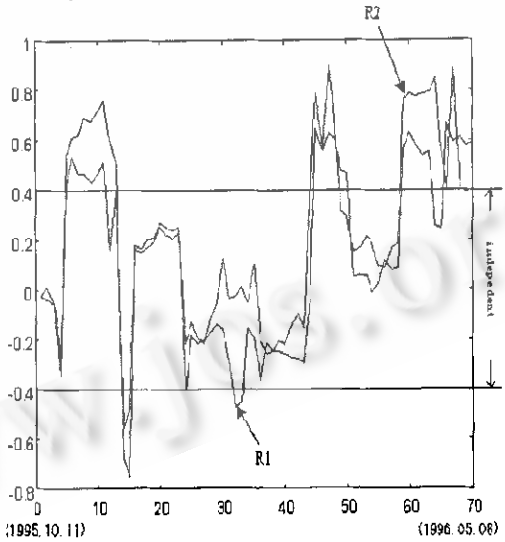
Proof is omitted.

It is obvious that $\rho_{XY}\approx 0$ in Example 3, but $\rho_{X^kY}>0$, $k>1$, so we know there is high-order correlation between variable X and dependent variable Y.

2 Applications

The analyses in above Examples 1~3 are done under the condition of having known function relation between X and Y. Next, we will analyze an actual example. The price curve in Fig. 2 gives actual stock price of Shenzhen Development Bank CO., LTD. from October 11 in 1995 to May 8 in 1996, denoted by $S(t)$. Let $X(t)=(S(t)-S(t-1))/S(t-1)$ denote stock price ratio, $Y(t)=(A(t)-A(t-1))/A(t-1)$ denote stock trading volume ratio and $Z(t)=A(t)/A0$ denote stock trading volume relative position, where $A(t)$ denotes trading stock volume and $A0$ denotes total stock circulation volume.

Figure 1 gives R1-value curve and R2-value curve where R1-value is the estimate of (2,1)-order correlation coefficient between trading volume ratio and price ratio, and R2-value is the estimate of (2,1)-order correlation coefficient between trading volume relative position and price ratio according to actual transaction data and sliding windows data processing technique[7].



R1: (2,1)-order correlation between trading volume ratio and price ratio

R2: (2,1) order correlation between trading volume relative position and price ratio

Fig. 1

Moreover, let $R1^*=2*R1+\mathrm{Mean}(\{S(t)\})$ and $R2^*=2*R2+\mathrm{Mean}(\{S(t)\})$. Price curve and (2,1)-order correlation coefficient curve $R1^*$ between trading volume ratio and price ratio are described in Fig. 2, and price curve and (2,1)-order correlation curve $R2^*$ between trading volume relative position and price ratio are shown in Fig. 3.

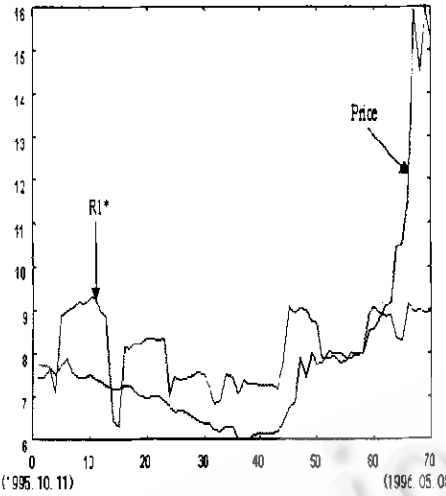


Fig. 2 Price and (2,1)-order correlation between trading volume ratio and price ratio

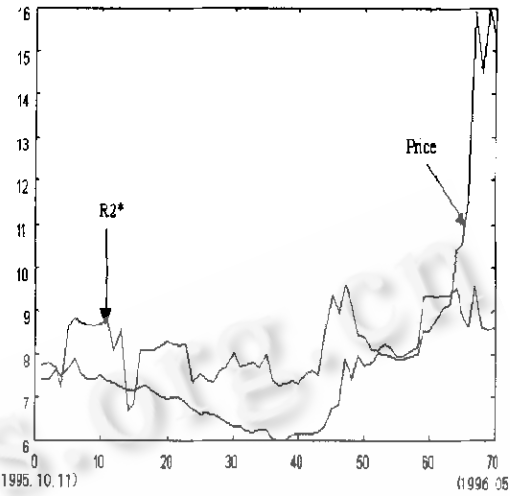


Fig. 3 Price and (2,1)-order correlation between trading volume relative position and price ratio

From the computed results, we can draw the following conclusions:

(1) The stock price ratio has nonlinear correlation with stock trading volume ratio and the nonlinear correlation is time-varying;

(2) The stock price ratio has nonlinear correlation with stock trading volume relative position and the nonlinear correlation is time-varying;

(3) The nonlinear correlation between stock price ratio and stock trading volume ratio is in accordance with the nonlinear correlation between stock price ratio and stock trading volume relative position;

(4) Whether stock market prices follow random walks or not[6] is time-varying, so analytical results between price and volume are valid only when there exists the intensive correlation. The result is important for us to do research, for example, association discovery in finance markets.

(5) Time-varying nonlinear correlation between price and volume gives not only intensive support and implication but also complication for applying neural network and chaos to study finance markets and to predict stock price[8].

3 Conclusions

How to track the nonlinear time-varying information is an important issue in financial data mining. This article deals with the problem by the discussion of independence and classic correlation. The paper gives some examples and then proves that the high-order correlation coefficients can fill up the space between independence and classic correlation. Furthermore, with the simplicity of computation, it is possible to capture all nonlinear information with high-order correlation coefficients. Finally, we apply the results to analyze an actual problem, the relation between stock price ratio and stock trading volume. The computational results show that high-order correlation coefficients can effectively tracking the time-varying linear and nonlinear characteristics hidden in financial data, and some interesting and useful results are obtained.

References:

[1] John, G. H., Miller, P. Building long/short portfolios using rule induction. In: Computational Intelligence for Financial Engineering. Piscataway NJ: IEEE Press, 1996.

[2] John, G.H. Stock selection using rule induction. IEEE Expert, 1996, 52~58.

[3] Lo, A., Mackinlay, A.C. Stock market prices do not follow random walks. Review Financial Studies, 1998, 1:203~238.

[4] Brock, W., Dechert, W.D., Scheinkman, J. A test for independence based on the correlation dimension. Working Paper, Department of Economics, University of Wisconsin, Madison, 1987.

[5] Beak, E., Brock, W. A general test for nonlinear granger causality: bi-variable model. Working Paper, Iowa State University and University of Wisconsin, Madison, 1992.

[6] Refenes, N., Burgess, A.N., Bentz, Y. Neural networks in financial engineering: a study in methodology. IEEE Transcations on Neural Networks, 1997, 8(6):1222~1267.

[7] Wang, Z.M., Yi. Dong-yong. Measure data modeling and parameter estimating. Changsha: National University of Defense Technology Press, 1996.

[8] Tong Fu, Fei Liang-jun. Computational intelligence approach for discovering the prediction model of financial market. Journal of Software, 1999, 10(4):395~399 (in Chinese).

附中文参考文献:

[8] 童頫,费良俊.发现金融市场预测模型的计算智能方法.软件学报,1999,10(4):395~399.

金融数据挖掘中的非线性相关跟踪技术

易东云, 张维明, 杜小勇

(国防科学技术大学 数学与系统科学系,湖南 长沙 410073)

摘要: 金融数据挖掘是信息社会中一个极具挑战性的研究方向.金融数据的随机特性使得隐藏在数据中的内在规则难以被发现.指出了经典相关分析的缺陷,进一步讨论了高阶相关系数的性质,证明了高阶相关不仅能描述隐藏的非线性相关信息,而且正好刻画了线性相关与独立之间的空白.因此,完全可以利用高阶相关性的计算简单性对金融数据中的时变非线性相关特性进行实时跟踪,克服了 Brock W. 等人于1987年和1992年提出的 Granger-Causality 独立性检验方法中需要正态假设和非实时性的缺点.最后,将上述结果应用于股票价格与成交量之间的相关分析.数值结果显示高阶相关能跟踪隐藏在数据中的时变非线性相关特性.

关键词: 非线性分析;数据挖掘;金融数据

中图法分类号: TP18 **文献标识码**: A