経済研究
Vol. 41, No. 3, Jul. 1990

特集　計量経済学の方法と応用

金融時系列分析と逐次分析法

高　橋　一

1. はじめに

近年金融時系列分析は多くの統計学者の興味を喚起しつつある．それは今まで多くの計量経済モデルでは統計分析に必要な十分多くのデータを得ることが難しかったこともあり現代統計学の経済学への応用が非常に限られていた．そしてそれ故そこから新しい統計学的問題も残念ながら殆ど生まれてこなかったという事に対する反動とも言えよう．一方癌研究等に代表される医学への応用，そして品質管理への応用は過去30年の間にアメリカを中心として急速な進歩を遂げてきている．そしてそれらの中から逐次分析法，比例ハザードモデルを始めとする数多くの統計学手法が生まれてきている．しかしながら平均株価指数，為替レート等の金融時系列分析に於いては近年比較的容易に大きな標本を取ることが可能となってきているのみならず，金融理論の発展，特にモダンポートフォリオの理論の発展に伴い，金融時系列の統計分析の重要性は年々高まりつつある．

さて過去40年コンピュータを利用した数多くの統計手法が利用されつつある．そのうちの幾つかは交差確認法の様にもともと存在してはいたが，想像を絶する計算量のため殆ど使われる事の無かったものも有る．又逐次分析法の様に一応の理論は有るが計算が面倒である事より広く応用される事の無かった方法もある．一方ブーツトラップ法，射影追跡法，そしてCART (Classification And Regression Tree)等の様にコンピュータ無しでは全く考えられなかった新しい方法も考案されて来ている．本稿ではそれらの新しい統計手法の中から逐次分析法(その中でもCusum検定と呼ばれる品質管理に於ける手法)とCARTに注目しそれらの金融時系列分析への応用について考えて行く．

本研究は，日本証券奨学財団研究調査助成の援助のもとに行なわれた．その資金援助に謝意を表したい．

2. 逐次分析法

本節では逐次分析法と呼ばれる統計手法の金融時系列分析への応用を考える．逐次分析法とは元々第二次世界大戦中にアメリカで，A. Wald等によって開発された手法で，統計的決定を行なう際予め必要とされる標本数を固定せずに行なう方法である．換言すれば，なんらかの結論を出すのに十分な情報を得るまで標本を取り続ける方法である．これは主に統計学的な品質管理法等に於いて製品が極端にout of controlの状態にある時などは出来る限り早く製造工程に問題が有ると言う結論を出したい．このような場合決められた数の標本を全て取り終わる前に製造工程は問題なく動いていると言う仮説が誤りである事が明かになる事は珍しくない．その時全ての標本を取り終わるまで決定を待つ事は時間と資源の無駄で有るが，もしも途中でサンプリングを止めてしまったならば統計学的に意味のある結論を引き出すことは出来ない．逐次分析法はこの様なときサンプリングを途中で停止する事を認めつつ，かつ統計学的に意味のある結論を引き出す方法である．以下我々は逐次分析法を分布の構造変化時点の発見問題へ応用する．その為に先ず問題を次の様に定式化する．

$x_1, x_2, \cdots$ を独立な確率変数列で x_i は確率密度関数 $f_i(x)$ に従うものとする．そして我々は帰無仮説

(2.1)　H_0:　$f_i(x)=f(x)$
for all $i=1,2,\cdots$

を対立仮説

(2.2)　H_v:　$f_i(x)=f(x)$
for all $i=1,2,\cdots,v-1$,
$f_i(x)=g(x)$
for all $i=v,v+1,\cdots$

に対して検定する．しかし変化時点 v は一般には未知であり対立仮説として

(2.3)　A：少なくとも一つの v に対し H_v が成立する．

を考え実際に構造変化が起きたか否かを検定する．さらにもしも変化が起きていたのならば出来る限り早くその事実を発見する事が目的である．(勿論帰無仮説の下では $v=\infty$ である)．このような問題には品質管理の場面に於いてしばしば出会うであろう．一方金融時系列分析に於いても例えば日経平均の収益率 x_i の系列が今まで正の平均を持っていたが，何等かのマクロ的な条件の変化から負の平均を持つ様に成る事があろう．この時我々の関心事は出来るだけ早くその変化を見つける事である．又その変化は系列の分散に現われる事もあるし，平均，分散の両方に現われてくる事も有り得る．何れにしても問題は如何にして精度の高い v の推定値，又は変化時点を知らせる指標 τ を求めるかである．即ち帰無仮説の下で τ は出来る限り大きな値を取る(変化が起こったと言うことは誤りであるから)，一方対立仮説の下では変化が起きた後，それをでき得る限り早く発見する指標を求めたい．形式的には，ある与えられた正定数 B に対し

(2.4)　$E_0(\tau)\geq B$

という条件の下で

(2.5)　$\sup E_v\{\tau-v+1\mid\tau\geq v\}\qquad v\geq 1$

を最小にする τ を求めることである．

この問題に対するアドホックではあるが一般的な解答は Siegmund(1985, Ch 2)で論じられている．以下簡単にそれを要約しておこう．$x_1,\cdots,x_n$ が観測されたとき H_0 を H_v に対し検定するときの対数尤度比は

(2.6)　$\sum_{k=v}^{n}\log\{g(x_k)/f(x_k)\}$

で与えられる．従って少なくとも一つの v に対して H_v が成立しているという対立仮説 A に対する対数尤度比は

(2.7)　$\lambda_n=\max_{0\leq k\leq n}(S_n-S_k)=S_n-\min_{0\leq k\leq n}S_k$

但し，

(2.8)　$S_n=\sum_{k=1}^{n}X_k=\sum_{k=1}^{n}\log\{g(x_k)/f(x_k)\}$

で与えられる．帰無仮説の下では X_k の期待値は負，一方 $k\geq v$ であれば X_k の期待値は正となる．従って S_n は初めは負のドリフトを持つが変化時点 v 以降は正のドリフトが支配し始める．その正のドリフトを如何に早くピックアップするかが問題である．標本採集の途中でその構造変化を知らせる指標を停止時刻(Stopping Time)τ と書くとき我々はそれを次のように定義できるであろう．(但しここで停止時刻と言う呼び方は品質管理や通常の仮説検定問題に於ける用語法に従っている．それらの分野ではなんらかの決定が行なわれたならば，それ以上の標本採集を行なわない．即ち標本採集をそこで停止するからである．これは金融時系列分析等の場合とは若干異なる．そこでは構造変化が認識された後もデータを集め続ける事が普通である)．

もしも尤度比が大きければ，それは対立仮説が正しいという確信が増すと言うことであるから，ある定数 c に対して

(2.9)　$\tau=\tau(c)=\inf\{n:S_n-\min_{0\leq k\leq n}S_k\geq c\}$
$=\inf\{n:\lambda_n\geq c\}$

が直感的には分かりやすい停止時刻であろう．さて

$$(2.10)\quad \sup_{v\geq 1} E_v\{\tau-v+1 \mid \tau\geq v\}=E_1\{\tau\}$$

が成立する事は $E_v\{\tau-v+1 \mid \tau\geq v\}$ に於ける最悪の場合のケースが実は $v-1$ で S_n が最小値を取る時であると言うことより簡単に証明できる．そこで我々は結局 $E_0\{\tau\}$ 及び $E_1\{\tau\}$ を c の関数として表現する事により与えられた密度関数と B より c を決定できる．所で Wald (1947)の方法を若干変更することにより

$$(2.11)\quad E_0\{\tau\}\fallingdotseq|e^c-c-1|/|\rho_0|$$
$$(2.12)\quad E_1\{\tau\}\fallingdotseq(e^c+c-1)/\rho_1$$

となる事が証明できる．但し，

$$(2.13)\quad \rho_0=\int_{-\infty}^{\infty}\{\log g(x)/f(x)\}f(x)\,dx,$$
$$(2.14)\quad \rho_1=\int_{-\infty}^{\infty}\{\log g(x)/f(x)\}g(x)\,dx$$

である．以下これを各種のケースに適用して行く．そのために $x_1, x_2, \cdots$ を基本的には互いに独立な正規分布 $N(\mu_i, \sigma_i^2)$ に従う確率変数列とし，やや行き過ぎた単純化ではあるが構造変化をその平均の変化，分散の変化，又は両方の変化という様に定式化する事とする．変数列 x_i を為替レートの対数変換値，ないしは日経平均の収益率とするならば，平均値の変化はその収益率の変化となる．一方分散の変化はそのリスクの変化に対応する．

2.1. 平均の変化

通常の仮説検定問題と同様，分散未知の場合と既知の場合とに分けて考える必要があるが，ここでは簡単の為分散は既知と仮定する．従って一般性を失う事なく $\sigma_i^2=1$ と仮定する．帰無仮説のもとでの平均を 0，対立仮説 $A^{(1)}$ のもとでの平均を $+1$($A^{(2)}$ のもとでは -1)とするならば，

$$(2.15)\quad X_i^{(1)}=x_i-1/2 \qquad (X_i^{(2)}=-x_i-1/2)$$

となり，

$$(2.16)\quad S_n^{(j)}=\sum_{i=1}^{n}(-1)^{(j)}x_i-n/2, \qquad (j=1,2)$$

である．従って停止時刻 $\tau^{(j)}j=1,2$ は(2.9)で S_n を上の $S_n^{(j)}$ で置き換えたもので定義される．$\tau^{(1)}, \tau^{(2)}$ はそれぞれ均衡状態にあった系列が均衡状態からプラスの方向，またはマイナスの方向へ乖離したか否かを判定する．そこで

$$(2.17)\quad \tau=\min\{\tau^{(1)}, \tau^{(2)}\}$$

を定義すれば τ は均衡状態からの乖離を判定する．この時

$$(2.18)\quad (E\{\tau\})^{-1}=(E\{\tau^{(1)}\})^{-1}+(E\{\tau^{(2)}\})^{-1}$$

が成立している(Siegmund, 1985 p 28)．一方株価の収益率等の動きを分析するときには例えば今現在収益率がマイナスであるのならばそれがプラスに転ずるときが何時かかを知る必要がある．形式的にはこれは帰無仮説のもとでは $\mu=-\Delta$ そして対立仮説のもとでは $\mu=\Delta$ となる正規分布に対する変化時点検出問題である．従って，ルーティーン計算より

$$(2.19)\quad S_n=\sum_{i=1}^{n}2\Delta x_i$$

となり

$$(2.20)\quad \tau=\tau(c)=\inf\{n: S_n-\min_{0\leq k\leq n} S_k\geq c\}$$

が求める停止時刻となる．即ち $\tau=m$ が仮に観測されたとすると，これは第 m 時点で収益率が既にプラスに転じたと決めるのに十分な証拠(情報)が集まった事を意味する．τ の定義より，これは一種のフィルタールールと成っていることは云うまでもないであろう．今までのフ

ィルタールールが単に勘と経験とからのみ成り立っていたのとは異なりτは確率モデルに基づいている．それ故定量的な性質を求める事が可能となっている．実際(2.11)よりτの期待値は

$$(2.21)\quad E_{\Delta}\{\tau\}\fallingdotseq|\exp[-2\Delta c]+2\Delta c-1|/(2\Delta^2)$$

で与えられる．ただ残念ながら(2.21)の数値的な精度はあまりよくない事が知られている．そこで(2.11)を更新理論(Renewal Theory)からの結果と組み合わせることにより，より優れた精度を持つ近似公式が求められた(Siegmund, 1985)．これは一言でいえば(2.11)は(2.9)でτを定義するとき$S_n-\min S_k=c$が成立していると仮定して求められている．しかしながら一般には$S_n-\min S_k-c\geqq 0$であり，over-shootが存在する．Siegmundはそのover-shootの大きさの極限分布の平均値を求め(正規性の仮定下では概ね0.584)それの定数倍をcに加えるより，より精度の高い近似式を求めることに成功した．それによれば，例えば$c=6$, $\Delta=0.4$の時$E_{-0.4}\{\tau\}\fallingdotseq 944(940)$, $E_{0.4}\{\tau\}\fallingdotseq 14.8(14.9)$と成る．但しここで括弧内の数字はVan Dobben de Bruyn(1968)が与えた数値計算に基ずくものである．勿論cを小さくすれば変化時点の発見をもっと早く行なう事が可能である．

一方分散未知のケースは基本的にはt-分布を用いれば良いわけだが問題は若干複雑になって来る．このことは次の分散の変化時点の発見問題の所で一緒に論ずる．

2.2. 分散の変化

株価等の日次収益率の平均値は概ねゼロと成っているが，その標準偏差値はゼロに近いながらも平均値より一桁大きい事が普通である．又Black-Schole(1973)により代表される株式オプションの価格決定公式は当該株式の収益率の分散には依存するがその平均値とは独立である．このように応用の多くの側面において分散の測定及びその変化時点の発見が最近重要な問題となってきている．分散変化モデルの中で最も簡単なものは，「帰無仮説H_oの下では確率変数列$x_1, x_2, \cdots$は既知の平均μ分散σ_o^2を持つ正規分布に従う．一方対立仮説Aではある未知の時点v以降は平均μ分散σ_1^2の正規分布に従う」というものであろう．この時(2.7), (2.8)の対数尤度関数は

$$(2.22)\quad \lambda_n=S_n-\min_{0\leqq k\leqq n} S_k$$

但し

$$(2.23)\quad S_n=(1/2)\{n\log(\sigma_0/\sigma_1)^2-(\sigma_1^{-2}-\sigma_0^{-2})\sum_{i=1}^{n}(x_i-\mu)^2\}$$

と成る．上記λ_nを用いて(2.9)で停止時間を定義することにより，これも又平均の変化時点発見問題と同様に議論できる．実用上はσ_0^2, μは過去のデータより推定されたものを用いればよいであろう．又σ_1^2に関してはσ_o^2の定数倍等が考えられる．

さて平均が未知であると仮定すると話はかなりややこしくなってくる．分散の値に関する通常の仮説検定問題に於いては，問題が位置に関して不変であることより$x_1, x_2, \cdots, x_n$が観測された時，それらを

$$(2.24)\quad y_{k-1}=[x_1+\cdots+x_{k-1}-(k-1)x_k]/[k(k-1)]^{1/2},\ k=2,\cdots,n.$$

と変換する．仮説H_0の下でy_kは平均0分散σ_o^2の独立な正規分布に従う．又対立仮説H_1では平均0分散σ_1^2のやはり独立な正規分布に従う．しかし$n>v$の時，対立仮説H_vの下では$y_1, \cdots, y_{v-1}$は$N(0, \sigma_0^2)$, $y_v, \cdots, y_{n-1}$は$N(0, \sigma_1^2)$に従う独立な確率変数となり(2.8)で定義される対数尤度比の形が非常に複雑となり理論的にはともかく実用的な検定とは成り難い．これは収益率の系列に正規分布を仮定すること自体にかなり問題を含んでいる事にも関連している．Taylor(1986)を始め多くの実証研究によ

り，アメリカ，イギリスに於ける主な金融時系列が正規性の仮定を満たしていない事が指摘されてきている．日本に於いても例えば刈屋，佃，丸(1989)により同様の結果が得られている．そこで便宜的にせよ正規分布を仮定し分析を行なうことの唯一のメリットはそれが理論的にも又応用面でも簡単な結果，形式を提供するからである．従って変換(2.24)を用いる理由は余り無い．そこで以下では少々異なったモデルを考えて行く．

実証分析の結果から多くの金融時系列は平均値に対し対称ではあるが正規分布と比べ裾野が重い分布(尖度が3より有意に大きい)となっている事が指摘されてきている．この様な分布を説明する為にここではいわゆる混合正規分布を考える．このようなモデルについては純粋に統計学的見地から論じられたものから(DeGroot, 1970)金融時系列分析に関連したものまで(Praetz, 1972, Blattberg-Gonedes, 1974, Kon, 1984)数多く存在する．しかしモデル自体は次の通りである．

$$(2.25)\quad x_i \overset{i.i.d.}{\sim} N(\mu, 1/R_i),\qquad R_i \overset{i.i.d.}{\sim} h(r)\qquad i=1,2,\cdots$$

即ち，第i日目の収益率は二つの確率変数(x_i, R_i)により決定される．先ず何らかの社会経済的理由からその日の収益率の分散$1/R_i$が決定される．そしてそれをgivenとしてx_iの条件付き分布が平均μ分散$1/R_i$の正規分布となる．R_iを分布$N(\mu, 1/R_i)$の精度(precision)と呼ぶことがある．さて重い裾を与え，かつ数学的に展開が容易なRの分布として代表的なものが次に挙げるパラメータ(a, b)を持つガンマ分布である．

$$(2.26)\quad R\sim h(r|a,b)=[b^a/\Gamma(a)]r^{a-1}e^{-br},\qquad r>0$$

ここで$\Gamma(a)=\int_0^\infty u^{a-1}e^{-u}du$はガンマ関数である．又$E\{R\}=a/b$，$Var\{R\}=a/b^2$である．分散の逆数にガンマ分布を仮定する事は主にベイズ統計学の中で行なわれてきている．これはデータが与えられたときのRの条件付き分布が再びガンマ分布となるという性質を持っているからである．又この時(x, R)の同時分布は

$$\begin{aligned}(2.27)\quad f(x,r|\mu,a,b)&=n(x|\mu,1/r)h(r|a,b)\\&=[b^a/(2\pi)^{1/2}\Gamma(a)]r^{a-1/2}\\&\quad\exp\{-r[(x-\mu)^2/2+b]\}\end{aligned}$$

となる．

さてここでxの周辺分布(無条件確率分布)を計算すると，それは

$$\begin{aligned}(2.28)\quad f(x)&=f(x|\mu,a,b)\\&=b^a\Gamma(a+(1/2))/[\Gamma(a)(2\pi)^{1-2}\\&\quad\{b+(1/2)(x-\mu)^2\}^{a+(1/2)}]\end{aligned}$$

と成る．ここでもしも$a=b$, $\mu=0$とすると，これは自由度$a/2$のt-分布の密度関数に外ならない．一般にt-分布は正規分布と比べ大きな尖度を持つことはよく知られている．さて$x_1,\cdots,x_n$が観測された時パラメータa, b μの最尤推定量はデータの尤度

$$\begin{aligned}(2.29)\quad \Lambda_n&=\prod_{i=1}^n b^a\Gamma(a+(1/2))\\&\quad/[\Gamma(a)(2\pi)^{1/2}\{b+(1/2)(x_i-\mu)^2\}^{a+(1/2)}]\end{aligned}$$

をa, b, μに関し最大化を計る事により得られる．この一種の経験ベイズ的手法は微分等を使いその解をClosed Formで求めることは勿論出来ないが，少なくとも数値解析的には実行可能である(cf. Blattberg-Gonedes, 1974)．

一方構造変化をa, bの変化で表現する事によりCusum検定を行なうことも可能である．その為にある与えられた定数の組(μ, a, b)と(μ, a', b')とに対し次の様なモデルを考えよう

$$(2.30)\quad H_0:(x_i, R_i)\sim f(x, r|\mu, a, b)\quad \text{for all } i=1,2,\cdots$$

(2.31) $H_v: (x_i, R_i) \sim f(x, r|\mu, a, b)$ for all $i=1, 2, \cdots, v-1$
$(x_i, R_i) \sim f(x, r|\mu, a', b')$ for all $i=v, v+1, \cdots$

少なくとも一つの v に対して H_v が成り立つという対立仮説を上と同様 A と書くことにすれば，H_0 を A に対し検定するときの対数尤度関数は(2.7)の λ_n で与えられる．但しここでは

(2.32) $S_n=\sum_{k=1}^{n} X_k$

又

$X_k=\log\{b'^{a'}\Gamma(a'+(1/2))/[\Gamma(a')(2\pi)^{1/2}\{b'+(1/2)(x_i-\mu)^2\}^{a'+(1/2)}]\}-\log\{b^a\Gamma(a+(1/2))/[\Gamma(a)(2\pi)^{1/2}\{b+(1/2)(x_i-\mu)^2\}^{a+(1/2)}]\}$

と成っている．ガンマ分布との混合正規分布を考えている為(2.13), (2.14)の計算を含む，τ の期待値の導出は正規分布の場合と較べ難しいが，少なくとも数値的に求めることは可能である．

実際のデータにこの方法を当てはめる際にはパラメータの値を幾らに設定したら良いかが問題となる．ここでは一応平均は既知としてあるから，便宜上 $\mu=0$ としておこう．(日次データの場合にはほぼ問題なく言えると思われる．)一方より重要なパラメータ (a, b) に付いては過去のデータより異なった社会経済状態に対応する幾つかのパターンを選び各々の場合に於て (a, b) の値を予め最尤法で推定しておく．過去何十日(60日ぐらい)かのデータをもとに現時点の (a, b) 値を推定しそれを帰無仮説下の分布のパラメータ値とし，上で求めた幾つかのパラメータ値へ変化したか否かをテストする．ここでは同時に幾つかの停止時刻を観測する訳で，我々はその中で一番小さなものに注目する．即ち対応する状態へと分散の分布が変化したと云う決定を下す．現時点ではここで述べた方法の統計学的諸性質はまだ完全には研究されてはいない，が分散構造の変化時点を発見する数少ない実用的な方法の一つである．

3. CART 法

CART はもともとノンパラメトリックな判別分析法の一種として1970年代後半に Richard Olshen, Charles Stone, Jerome Friedman そして Leo Breiman 等の人々により考案・発展されてきた手法である．正規性の仮定下に於ける判別分析では Fisher や Mahalanobis の判別関数が有名である．そしてその有効性は多くの実例により示されてきている．実際 Fisher の線形判別関数がかなりロバストであることは(理論的にはまだ不明の所が多いが)多くの統計学者の一致した意見と言えよう．しかしながら(説明変数の)次元が極端に大きくなったり，質的な変量を含む場合や欠損値が存在する時には，それは必ずしもうまく行くとは限らない．もちろん概念的には与えられたデータから多変量密度関数を推定し(例えば単に多次元のヒストグラムを描いて見る事もまたは Breiman による Nearest Neighborhood 法の様なよりソフィスティケイトされた方法をとるにせよ)それを用いて Bayes 解を求める等いくつかの方法が考えられるが，問題は次元の呪いと呼ばれる現象である．これは一口で云えばヒストグラム(最もナイーブな密度関数の推定としての)を作るとき十単位あたり 1000 個のデータは一次元の分布に対しては一単位当り平均して 100 個のデータが利用可能という事より常識的には充分な数のデータと言える．しかし二次元分布ではそれは平均して一単位当り 3.12 個，三次元では僅か1個になってしまう．そして十次元分布では約 0.2 個となり殆どゼロとなってしまう．換言すれば常識的にそれほど高い次元でなくとも多次元分布の密度関数の推定には莫大な量のデータが必要となって来る．しかし分析の目的がそれらの変数により説明されるであろうある変数の値ないしはそれが属するクラスの予測や説明にあるのならば密度関数を完全に知る必要は無い．目的に即した情報のみを選び分析を進めれば良いであろう．そこで提案された一つの方法が CART である．CART の使用

図 1

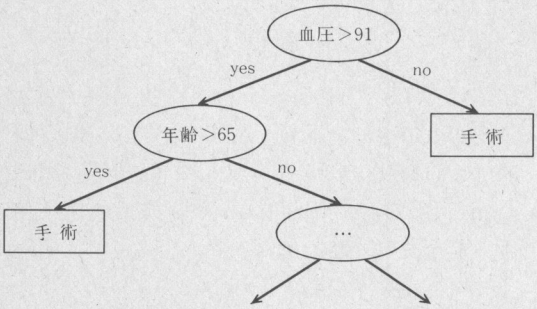
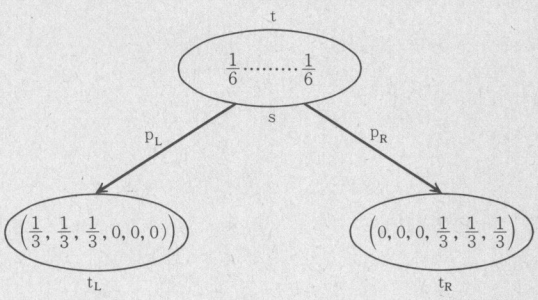


図 2



法は至って簡単である．それは週刊誌等によく見られるyes-noテスト，即ちdecision-treeである．具体的には次の例の様なものである．

例．心臓マヒの患者が急救病院に運び込まれたとき先ず最初に行なわなければならないことは緊急手術が必要か否かの判断であろう．もちろんそれは熟練した医師が患者の状態やいくつかの検査結果をもとに決めるべきことであるが，過去の経験からそれらの判断はかなりの部分においてマニュアル化可能であろう．例えば先ず真っ先に運ばれてきた患者の血圧をはかってみて，もしもそれが91以下であれば直ちに手術を行なう．それ以上であれば，次に年齢が65才以上か否かを調べそうでなければ内科的な治療を行なう，等々の事が考え得る．これを図式化したものが以下に示した“決定樹”(decision-tree)ないしは簡単にツリーである．

さてここでこれ以後用いる用語を定義しておこう．ツリーにおける各分岐点(上の図中丸ないし四角で囲われたもの)を“ノード”(又は節)と呼ぶ．ノードの中でも最終的な決定を行なうもの(四角で囲われたもの)を特にターミナル・ノードと呼び他のものと区別する．ツリーを作るためのもととなるデータ$(\underset{\sim}{x}, y)$を学習標本Ξ(learning sample)と呼ぶ．$\underset{\sim}{x}$はp次元の説明変数ヴェクター，yは非説明変数である．但し通常の回帰モデルとは異なり$\underset{\sim}{x}$の次元pは各標本毎に異なっても構わないし$\underset{\sim}{x}$の各要素は量的(numeric)であっても質的(categorical)であっても，又それらの混合型であっても構わないものとする．yについていえば，もしも量的であれば回帰モデルとなるし，質的であればクラス分け問題となる．pの値が動きうる事と，混合型の説明変数を認める事が通常のモデルとは大きく異なる点である．ツリーを作るための基本的な考え方とその使い方はL. Breiman, J. Friedman, R. Olshen and C. Stone(1984)に詳しい．又は高橋(1990)に簡単な説明がある．

3.1. ツリーの構築

ここではクラス分けを目的としたツリー(Classification tree)を学習標本から作る時の基本的な考え方を述べよう．今仮にあるノードtに6種の異なったクラスに属するデータ(例えば$y=1, 2, 3, 4, 5, 6$)がそれぞれ同じ割合で(1/6ずつ)存在しているものとする．クラス分けと云う目的からはこれは最悪の状態といえる．ここでもしも$\underset{\sim}{x}$の値によってこのノードを左右二つのノードt_L, t_Rに分ける時上図で示される分割規則(splitting rule)sがもしも存在すればそれはより純粋なノードが作られたと云う意味に於いて，よりベターな状態と言えよう．但しp_R, p_Lはそれぞれtの要素で左側のノードへ振り分けられるものと右側のノードへ振り分けられるものとの比率である．分割規則sは具体的には，例えば，$\underset{\sim}{x}$の要素x_mがある数以上であれば(またはAというカテゴリーに属せば)右のノードへそのサンプルを送り出す．又はより一般的に$\underset{\sim}{x}$の要素の線形結合を考えその値の大小でクラス分けを行なう．前者を変数x_mに於ける分割，後者を線形結合による分割と呼ぶ．よりベターな状態はそのノードの持つ“純粋度”の様なもので表現されうると考えられる．このプロセスを何回か繰り返す事により

図 3 (13 日)

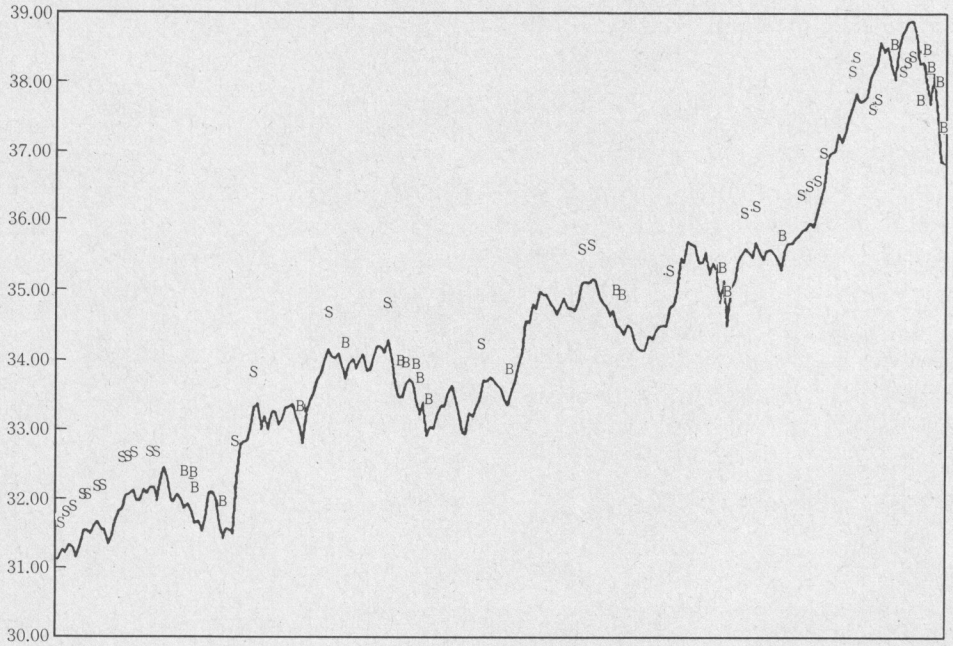
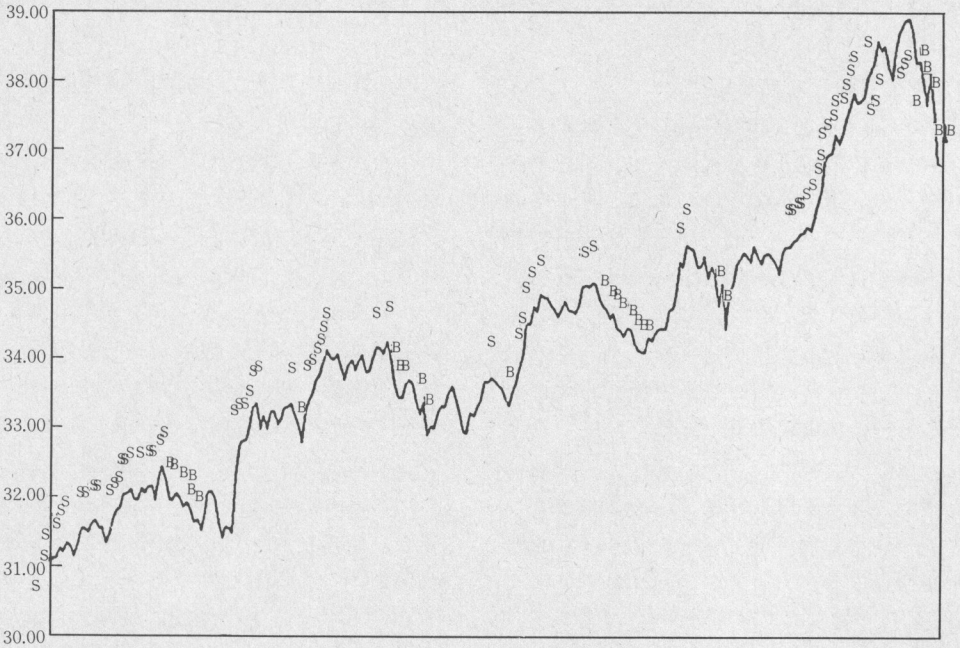
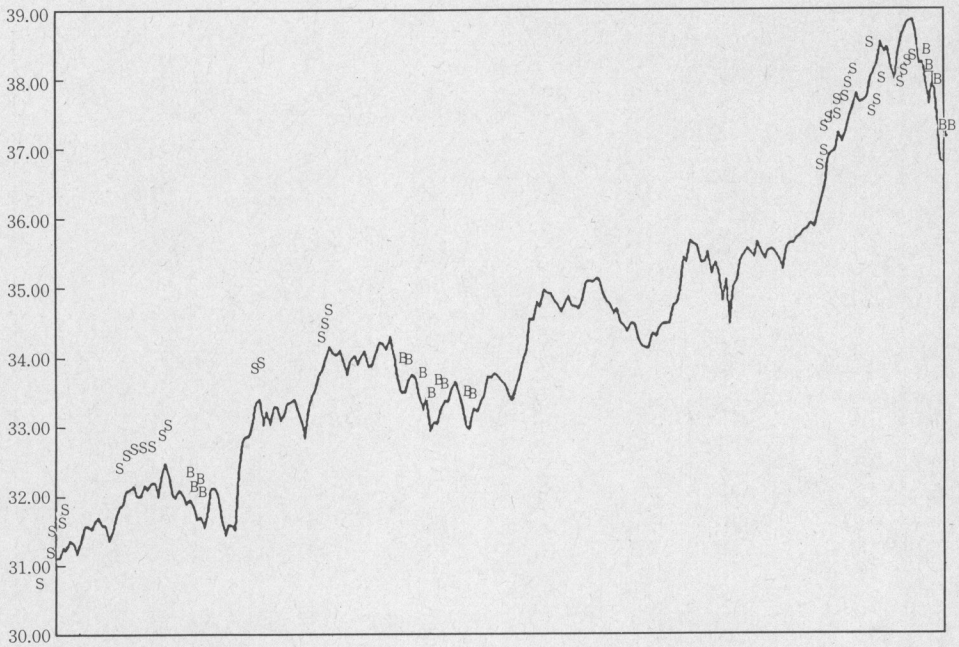


図 4 (26 日)



我々は十分“純粋”なノードに到達出来る．これをターミナルノードとし，そこでクラスの決定を行う．実際にはこの様な操作は全てプログラム化されている為，我々がやらなければならないことは基本的には“純粋さ”を計る指標を何にするか，クラス分けにおける“予測の誤判別率”をどの程度とするか等を予め指定しておく事のみである．そして学習標本よりCARTはツリーを構築する．我々はこのツリーを用いて新たに与えられたデータ(説明変数)から，そ

図 5 (52日)



のクラスを予測(推定)する.

3.2　東証指数に対する応用

CARTに関する理論的研究は今の所おもに独立なデータに限定されてはいるが，時系列データに対しても十分適用可能であると考えられている．実際その一つの応用例として東証株価平均の価格変化予測に応用してみたものがある．以下それを簡単に見て行く．ここで扱っているデータは東京証券取引所で上場されている日経225種の日次の平均株価である．株式市場で利益を得るためには株価の安いところでそれを買い，高いところで売りに回れば良いことは誰の目にも明かであるが問題は値下がりと値上がり各々のピークを如何に発見するかである．前節で考えた様なモデルも一つの方法であるがここではCARTを使ってその予測を行ってみる．先ず説明変数として日経平均指数の順位相関係数，RSI，株価水準，ボリウムレィシオ，そして暴落レィシオの13日，26日，そして52日の移動平均及びそれらの標準偏差を考えた．一方ツリーを作るための学習標本の非説明変数は過去数年間に渡る最適売買タイミングを事後的に調べたものを用いた．その結果得られたものが次のグラフである．(これは1980年から1988年迄の上記説明変数と事後的に決定された最適タイミングを学習標本としたものから作られたツリーを用いた1989年の最適売買タイミングを求めたものである)．図中 S, B はそれぞれCARTによる最適売り，買いタイミングの予測である．この結果が満足すべきものか否かは判断の分かれるところで有ろうが，フィルタールールを始めとする多くの株価予測法と比較して殆ど遜色の無い結果がある意味で非常にナイーブな方法(基本的にはCARTがやっていることは過去のデータの整理である)により得られたと言うことは，注目に値するであろう．

その他にもCARTの性格上例えば景気指標の構築等にも応用可能であろう．又その欠損値に対するロバスト性や説明変数の持つ(標本毎に次元が異なっても良い事に加え量的，質的を問わずあらゆる種類のデータを含み得ると言う)高い自由度故，所得統計等の改善に大きく役立つと考えられる．

(CARTによる株式モデルの計算は大和証券システム開発部の宮本氏，同システム企画部の

田中氏にお願いした．両氏のご協力に感謝する．）

（一橋大学経済学部）

References

［1］ Black, F. and M. Scholes, "The pricing of options and coporate liabilities," *Journal of Political Economy*, 81,(1973), 637-659.

［2］ Blattberg, R. C. and N. J. Gonedes, "A comparison of the stable and Student distributions as a statistical models for stock prices," *Journal of Business*, 47(1974), 244-280.

［3］ Breiman, L., Friedman, J., Olshen, R. A. and C. J. Stone, *Classification and Regression Trees*, Wadworth, 1984.

［4］ van Dobben de Bruyn, C. S. *Cumulative Sum Tests*, London, Griffin, 1968.

［5］ De Groot, M., *Optimal Statistical Decisions*, McGraw-Hill New York.

［6］ 刈屋武昭，佃良彦，丸淳子編著『日本の株価変動』東洋経済新報社，1989年

［7］ Kon, S. J., "Models of stock returns-a comparison," *Journal of Finance*, 39,(1984), 147-165.

［8］ Praetz, P. D., "The distribution of share price changes," *Journal of Business*, 45,(1972), 49-55.

［9］ Siegmund, D., *Sequential Analysis, Tests and Confidence Intervals*, Springer-Verlag New York Inc., 1985.

［10］ 高橋 一「Classification and Regression Trees とその応用」官庁統計に於ける解析方法の改善に関する調査研究(II)，統計研究会，1990年

［11］ Taylor, S.,*Modeling Financial Time Series*, John Wiley and Sons Ltd., 1986.

［12］ Wald, A., *Sequential Analysis*, New York : Jhon Wiley and Sons, Inc., 1947.