A-009

ランダム行列との比較による株価日中変動の相関行列解析

Correlation Analysis of Intra-day Stock Prices in Comparison to Random Matrices

田中美栄子† 木戸丈剛†

Mieko Tanaka-Yamawaki Takemasa Kido

1. まえがき

株式市場においては多数の株価が互いに連動しながら一見ランダムに動いている．主要株の多くが連動して動くと市場全体に影響を及ぼす目立った動きとなるが,連動のネットワークは非定常であり,仮に或る時刻にその詳細を知ったとしてもその知識を有効に使う手立てを迅速に見つけ出すことは難しい．しかし主要な株価が連動して動いているときにその動きを牽引する大きな主成分を迅速に抽出する計算方法があれば,それに基づいてその時刻における市場の特徴抽出を行い,時間変化を追うことが可能になる．

そのような方法の候補として,多数の時系列間の同時刻相関行列のスペクトルを,ランダム行列から作った相関行列のスペクトルからの外れ成分として扱う方法があり,近年広範囲な分野で興味が持たれている．

金融時系列に対しては,NYSE 株価のうち S&P500 に入っているものを選んでその日次変動を扱った Plerou 等の研究[1,2]やそれとほぼ同時に発表された Bouchaud 等の研究[3,4]が良く知られており,日本株に対しても青山等[5]が同様の方法で解析を行っている．本研究は上記の手法の適用範囲を調べた上で,米国株価の日中変動に対して有意な主成分を抽出し,これに基づいて株価相関の年次変動を追跡できることを示す.

2. 金融時系列の同時刻相関行列

株式市場では数百から数千の異なる銘柄の株式が常時取引されており，そのうち連動して動いている銘柄の組を抽出することは，市場の動きの解析に重要な意味を持つ．通常は同業種に属する銘柄どうしが連動することにより相関が見られるが，異業種間にも相関や反相関が見られることがある．銘柄ごとに株価の大きさはバラバラなので，直接株価を比較するのでなく，収益率という量を使用することが多い．収益率は株価の変化をもとの時刻 t における株価 S(t)に対して時刻 Δt 後の株価 S(t+Δt)が何割上がったか，または下がったかという比率

$$\frac{S(t+\Delta t)-S(t)}{S(t)}=\frac{\Delta S(t)}{S(t)} \tag{1}$$

で表現する．この量は単位に依存しないため，平均数万円の株価の増減も平均数百円の株価の増減も同様に扱うことができる．しかしもっと便利なのは対数収益と呼ばれるもので，株価を対数で表したうえでその差を取って定義する．

† 鳥取大学大学院工学研究科情報エレクトロニクス専攻,
Tottori University, Graduate School of Engineering, Department of Information and Electronics,

時刻 t の対数収益 r(t)は両時刻の株価の対数の差で表される．

$$r(t)=\log(S(t+\Delta t))-\log(S(t)) \tag{2}$$

対数の性質からこれは

$$r(t)=\log\left(\frac{S(t+\Delta t)}{S(t)}\right) \tag{3}$$

と書けるが，対数の中の分子のほうは S(t)＋ΔS(t)であるから株価の増分ΔS が元の株価に対して十分小さい時,

$$r(t)=\log\left(1+\frac{\Delta S(t)}{S(t)}\right)\cong\frac{\Delta S(t)}{S(t)} \tag{4}$$

となって事実上，式(1)の収益率に等しい．式(2)で定義しておけば割算を使わずに計算できるので便利であり，今後は株価の変化といえばこの対数収益で表すことにする．本論文では複数の銘柄を扱うため,i 番目の銘柄の収益率の時系列を $r_i(t)$ と添え字 i を付けて表す．全銘柄数が N のとき,この添え字 i は 1 から N までの整数となる．

二つの銘柄 i と j の相関 $C_{i,j}$ は各時刻 t におけるそれぞれの対数収益 $r_i(t)$ と $r_j(t)$ の時系列ベクトルの内積

$$C_{i,j}=\sum_{t=1}^{T} r_i(t)r_j(t) \tag{5}$$

で表される．定義からこれは行 i と列 j の入れ替えに対して対称である．

後で便利なようにそれぞれの時系列の値を正規化しておく．これは t=1 から t=T の期間における r の平均値が 0 で分散が 1 になるように，r から平均値<r >を差引いて分散の平方根 σ で割っておくことである．

$$x_i(t)=\frac{r_i(t)-<r_i>}{\sigma_i} \tag{6}$$

式(6)によって正規化した時系列 $x_i(t)$の内積を取って式(5)のように計算した相関 $C_{i,j}$ を行列の形に並べると，当然これは正方行列であり，対角成分は全て 1 となる．また式(5)より，

$$C_{i,j}=C_{j,i} \tag{7}$$

となるので相関行列は対称行列でもある．対称行列は直交行列 V，すなわち $V^t=V^{-1}$ を満たす行列，を使った相似変換

V^{-1} C V により対角行列に変換できる．このような V の各列は正方行列 C の固有ベクトル

$$v_k = \begin{pmatrix} v_{k,1} \\ v_{k,2} \\ . \\ v_{k,N} \end{pmatrix} \tag{8}$$

に対応し，C を掛けると元のベクトル v_k に比例する．

$$Cv_k = \lambda_k v_k \tag{9}$$

全部で N 個の k の値，それぞれにに対して常識が成立し，比例係数 λ_k は固有値となる．成分を顕わに書くと

$$\sum_{j=1}^{N} C_{i,j} v_{k,j} = \lambda_k v_{k,i} \tag{10}$$

となる．このような固有ベクトル v_k は正規直交系を形成する．つまり，各ベクトル v_k は長さが 1 に規格化され，

$$v_k \cdot v_k = \sum_{n=1}^{N} (v_{k,n})^2 = 1 \tag{11}$$

異なる列 k と l に対しては直交する．

$$v_k \cdot v_l = \sum_{n=1}^{N} v_{k,n} v_{l,n} = 0 \tag{12}$$

これはクロネッカーのデルタを使って次のように書ける．

$$v_k \cdot v_l = \delta_{k,l} \tag{13}$$

この右辺は k=l ならば 1，そうでなければ 0 であることを示す．同時刻相関行列 C は対称行列なので Jacobi 回転アルゴリズムを繰り返すことにより,固有値と固有ベクトルを求めることができる[6].

3. ランダム行列スペクトルによる主成分抽出法 (RMT_PCM)

Jacobi 回転を繰り返して C の対角化を行うことは，N 個の式(8)からなる時系列セットを回転によって固有ベクトルのセットに変換することに対応する．

$$y(t) = Vx(t) \tag{14}$$

成分で書くと

$$y_i(t) = \sum_{j=1}^{N} v_{i,j} x_j(t) \tag{15}$$

となるが，添え字 i の異なる内積はゼロとなり，同じ i どうしの内積は次式により i 番目の固有値に一致する．

$$\begin{aligned} &\frac{1}{T}\sum_{j=1}^{N} y_i(t) y_i(t) \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{N} v_{i,l} x_l(t) \sum_{m=1}^{N} v_{i,m} x_m(t) \\ &= \sum_{l=1}^{N}\sum_{m=1}^{N} v_{i,l} v_{i,m} C_{l,m} \\ &= \lambda_i \end{aligned} \tag{16}$$

最大固有値の値が他の固有値と比べて抜きん出て大きいことは，その固有ベクトル成分の分散が他方向の成分の分散より抜きん出て大きいことを意味し，その方向を主成分とすることによって多変量のデータが一変量で近似できることになる．すなわち多次元のデータであってもデータの分布の顕著な方向に軸を取れば事実上 1 次元とみなして次元圧縮ができることになる．

このようにして最大固有値に対応する固有ベクトルを第 1 主成分とみなすことができ，同様にして 2 番目に大きい固有値に対応する固有ベクトルを第 2 主成分，その次に大きい固有値に対応するものを第 3 主成分，等と対応させて行くことができる．

もともとの相関行列 C の対角成分の和（trace）は対角成分の全てが 1 に規格化されていることにより独立な時系列の数 N に等しく，これは相似変換によって変化しないから全固有値の和は N となる．すなわち対角化によって生じた元の 1 より大きな成分は主成分として残し，1 より小さい成分はノイズとして捨てる．しかしこれでは N が大きい場合は主成分が多く残りすぎて状況が見えにくい．

固有値分布が非常に偏っている場合には，顕著に大きな固有値のみを或る基準に従って残し，それ以外はノイズとする慣習も存在する．例えば固有値の大きい方から累積和を取って行き，それが固有値全体の和である N の 8 割に達するまでを主成分とみなし，それより小さい成分をノイズとして捨てることにより有意な主成分とノイズを分離する．

しかし上記の 2 基準は数百から数千の銘柄が入り組んで上下する株式市場の価格時系列に応用するには次元数 N が大きすぎて不適当である．例えば 1 を超える固有値が数個以内に収まることは滅多になく，また，累積固有値が N の 8 割に達するまで主成分を取っていくと，これも数個以内に収まることはない．これは問題の次元 N が数百以上でランダム性の強い問題に共通した欠点である．

文献[1-5]で採用された方法はランダム行列理論[7]から導かれる固有値スペクトルの式を利用する方法であった．ランダム行列理論によれば，同時刻相関行列の固有値分布は，時系列の長さを T とすると，

$$N \to \infty,\ T \to \infty,\ Q \equiv T/N = \text{const.}$$

の極限で

$$P_{RMT}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{17}$$

となることが知られている．すなわち正規化されたランダム時系列の相関行列の固有値 λ は

$$\lambda_- < \lambda < \lambda_+ \tag{18}$$

の範囲に式(17)に従って分布し，その上限と下限は

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \tag{19}$$

で与えられる[8].

そこで得られた固有値のうちで，ランダム行列理論式に一致する，またはその範囲に存在するものをランダム成分とみなし，

$$\lambda \gg \lambda_{+} \qquad (20)$$

となる部分を，大きい方から順に主成分とする方法である．ランダム成分の上限である　　が有意成分の下限値であるかというとそうではないことが実データの分析結果から分かる．対応する固有ベクトルの成分分布にランダム性が高いと，主成分として特徴抽出に寄与することができないからである．この意味において本手法は固有値分布だけでなく固有ベクトルの成分分布をも併用した基準を採用する．この方法を従来の主成分分析法と区別して,ランダム行列理論式のスペクトルとの比較による主成分抽出法(Random Matrix Theory oriented Principal Component Method, RMT_PCA と略）と呼ぶことにしよう. RMT_PCA は従来法とは異なり,株式市場における非常に多くの株式の相関を扱う場合のように,数百から数千におよぶサイズの次元からたった数個の主成分を分離する際の理論的基礎付けを用意するものである．

4. 株式市場の日中変動の解析結果

以下では前述の RMT_PCM の方法を株価の日中データに適用した結果を述べる．使用したデータは米国株価の tick データ（NYSE-TAQ）の 1994 年－2002 年の期間であり，各年の trade 価格のセットを 1 データとして解析した結果をもとに，主成分の時間変化を追跡し，比較する[9].

同時刻相関行列を計算するためには使用する N 個全ての銘柄に対して T 個の全時刻で価格が必要となる．全ての tick 時刻に対してこれを満たす株価は存在しない．しかし我々の目的である，当該年の市場を牽引する主成分の抽出という目的に対しては，取引の十分活発な人気株のみを対象にしても良いと考えられる．そこで NYSE の営業時間である 9 時半から 3 時半の間で，定時の 10 時から 1 時間毎に 15 時までの 6 時刻の近辺（誤差 30 分以内とした）に取引のあった銘柄のみを選んでその trade 値（実際に約定した価格の記録）を式(1)－(4)の株価 S(t)として解析を行った[9].

このようにすると 1994 年，1998 年，2002 年はいずれも各々252 日の営業日があり，1 日あたり 6 データとして年間のデータ数が T=1512 となる．このすべてに trade 値の存在する銘柄 N は 1994 年で N=419 銘柄，1998 年で N=490 銘柄，2002 年で N=569 銘柄となった．

このような手間をかけずに直近の過去に約定した値を使用すれば T をもっと大きくできる．これは文献で before-tick などと呼ばれている方法である．または各 tick 時刻における ask（売り気配）や bid（買い気配）等の気配値を使用しても T を大きくできる．これらに対して我々の方法は定時の前後 30 分以内に実際に取引された価格を使用するもので，定時の周りに幅を持たせた block-tick 法とでも呼ぶべきものである．どれが最適であるかは今後の研究に待つところが大きい．本稿では後者の block-tick 法を一貫して用いる．

1994 年の 419 社の 1 時間変動に対する，式(5)の相関行列の固有値分布を Fig.1 に角グラフで示す.点線で示すランダム行列理論式(17)に重なるスペクトルとランダム行列の最大固有値より大きな離散固有値に分かれる．

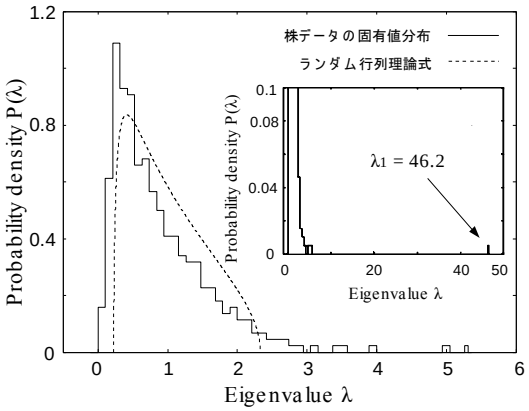

Fig.1　米国株価(1994)419 銘柄相関行列の結果

計算式から導かれる最大固有値は N=419,T=1512 の場合,Q=T/N=3.6 より λ_{+}=2.3 となるが，ランダム部分でも乱数度が低ければ λ_{+}より大きな領域にも固有値が分布するので，連続スペクトルの途切れる 3 以上の固有値：λ_1 = 46.2, λ_2 = 5.25, λ_3 = 5.04, λ_4 = 3.90, λ_5 = 3.51, λ_6 = 3.41, λ_7 = 3.11 を有意成分と見なすことにする．しかしこの選択には任意性が残る．

同様に 1998 年の 490 社に対する相関行列の固有値分布を求め,上述のアルゴリズムでノイズ成分を分離する．計算式から導かれる最大固有値は N=490, T=1512 の場合,Q=T/N=3.09 より λ_{+}=2.5 となり,そのうち 3.5 を越える 7 固有値：λ_1 = 81.1, λ_2 = 10.3, λ_3 = 6.9, λ_4 = 5.7, λ_5 = 4.8, λ_6 = 3.9, λ_7 = 3.5 が有意成分の候補となる．

最後に 2002 年の 569 社に対する相関行列の固有値分布については，計算式から導かれる最大固有値は N=569, T=1512 の場合,Q=T/N=2.66 より λ_{+}=2.6 となり,その内の 10 固有値：λ_1 = 166.4, λ_2 = 20.6, λ_3 = 11.3, λ_4 = 8.6, λ_5 = 7.7, λ_6 = 6.5, λ_7 = 5.8, λ_8 = 5.3, λ_9 = 4.1, λ_{10} = 4.0 が有意成分の候補となる．

上記固有値に対応する固有ベクトル成分のうち正値上位 10 個を Table 1 に示す. U_1 の成分は大企業が同符号で多数並び，文献[1]の 1990～1996 年の日次データと定性的に同じ結果となるものの,その銘柄は同じではない.,第 4 固有ベクトル U_3～U_4 に半導体関連企業が多い点も日次データに類似であるが，1 時間変動の場合は U_5に石油関連が集中する．

固有ベクトルの成分構成については，1994 年では，U_1 は大企業が多数並び,1998 年から 2002 年にかけて金融業と銀行が並んだ.この意味としては，1998 年以降に金融業と銀行が巨大化して株式市場の中心となった事が反映していると解釈できる．

また,1994 年には優勢だった鉱物や半導体に代わって 1998 年以降は食品や電気・エネルギー関連株が上位に出ているのが特徴的であるが,これは産業構造が 1994 年から 1998 年を経て 2002 年に至る間に変化して,鉱物・半導体が下火になり,代わって金融,食品,電気・エネルギー株などが米国株価の主力となったことを示唆すると考えられる．

NYSE 株価 1 時間変動と文献[1]の日次変動の結果との比較と理論式の適用範囲の明確化を本稿の主目的とした.1994 年の N=419 社の 1 時間変動を取った T=1512 の時系列に対する結果[9]は,第一固有成分に大企業多数が集まり S&P500

指標に近い組合せとなり,第 3, 4 成分に半導体関連が集まるなど,定性的には余り違わない結果となった.一方,2002 年の N=569 社の 1 時間変動を取った T=1512 の時系列に対する結果は, 第一固有成分に金融業と銀行が多く集まり,第 2-第 10 固有ベクトルの成分に食品や電気・エネルギー関連株が上位に現れた.

このことは 1994 年から 2002 年の間に半導体産業が下火になる一方,金融,食品,電気・エネルギー株などが NYSE の主力となる方向に産業構造が変化してきたことを反映していると考えられる.中間の 1998 年の N=490 社に対する同様の結果は,1994 年と 2002 年に至る変化の過渡期の状況を表しており，半導体関連が下火になる一方で，銀行・金融，環境・エネルギー関連が浮上する様子が観察される.

表 1 固有ベクトルの構成要素上位 10 成分の業種分布

U_k	1994 年	1998 年	2002 年
U_1	銀行(2), 車(2)	銀行(5), 金融(3)	金融(5), 銀行(3)
U_2	鉱業(7)	電気・エネルギー(10)	食品(6)
U_3	半導体(8), 集積回路(2)	銀行(2)	電気・エネルギー(10)
U_4	半導体(3) PC(3)，薬(2)	半導体・集積回路(10)	食品(4), 電気・エネルギー(4)
U_5	石油(9)	鉱業(6)	電気・エネルギー(9)
U_6	産業機械(2)	梱包(2)	電気・エネルギー(4)，食品(2)
U_7	特徴なし	特徴なし	小売(4), 鉱業(2)
U_8	通信(2), 車(2)	通信(2), 投資(2), 石油(2)	小売(9)
U_9	通信(4), 銀行(2)	小売(4)	鉱業(5), 通信(3)
U_{10}	通信(2)	医療(5)	通信(8)

5. まとめと今後の展望

株式市場における非常に多くの株式の相関を扱う場合,数百から数千におよぶサイズの次元をもつ,非常にランダム性の強いデータからたった数個の主成分を分離する必要がある.本論文で検討した,ランダム行列理論式を使った主成分抽出法(RMT_PCA）は,次元数が数百以上の大きな場合に適し,時系列長が次元数に比べてはるかに大きく取れる tick 時系列に向く方法であること,アルゴリズムがはっきりしていること,ランダム部分を RMT との照合することにより明確な方法で分離できること,等の利点を持っている．tick 時系列への適用は我々以前にはなく,新規な試みであることなどから株式市場のみならず,広範囲のデータ・マイニングに対して有効であると予想される．今回の米国株価 tick 時系列の解析では,同時刻相関行列を計算するために朝 10 時から午後 3 時までの定時に最も近い tick 値を代表値として取り出すことで, 1 日あたり 6 データを用い,1 年毎に十分な長さの時系列を確保することで年次変化を追う事が出来た．1 年あたりの証券取引の営業日が 250 日程度であることから,1 日 6 データ取れば 1 年で T=1500 データとなり,N=約 500 の株に対して Q=T/N=3 程度を確保できるからである．しかし定時に実取引が行われたかどうかに拘らなければ,blocktick 法でなく,直近の過去値を取引の有無に拘わらず用いる,before-tick 法を用いてもっと短い期間で同等の Q 値を確保できる事になり,年次でなく 4 半期毎の変化や,月次変化をも追うことが可能になる.

また,Q>1 の条件の下限に近い領域で起き得る誤差を機械乱数を使って見積もった結果,Q が 1 に近づくにつれて誤差は増大するものの,Q<1.3 の領域を除けば実用的には本手法が有効であることが示唆される[10,11]．本手法を広範囲のデータに適用した結果を蓄積することで,現時点ではまだ結論に至らない知見を明確化できるものと考えられる.

参考文献

[1] V. Plerou,et.al.,"Random matrix approach to cross correlation in financial data", Physical Review E 65, 066126, 2002.
[2] V. Plerou, P. Gopikrishnan, B.Rosenow, L.A.N.Amaral, and H.E. Stanley, Physical Review Letters, 83, 1471, 1999.
[3] L.Laloux, P. Cizeaux, J.-P. Bouchaud, and M.Potters, 83, 1467, 1999.
[4] J.-P. Bouchaud and M. Potters, " Theory of Financial Risks", Cambridge University Press, 2000: "金融リスクの理論"(森平監訳)朝倉書店,2003.
[5] 青山秀明,他 4 名, "経済物理学"(共立出版,2008)
[6] 小国力,"Fortran 95, C&JAVA による新数値計算法- -数値計算とデータ分析"(サイエンス社,1997)
[7] M.L.Mehta, "Random Matrices", Academic Press 3rd edition, 2004.
[8] A.M.Sengupta and P.P.Mitra, "Distribution of singular values for some random matrices", Physical Review E vol.60, pp.3389-, 1999.
[9] 田中美栄子、田中瑶子、伊藤大哲、中村元紀、木戸丈剛、川村綾、佐藤彰洋：ランダム行列との比較による NYSE 株価 1 時間変動の相関行列分析(1), 素粒子論研究(京都大学基礎物理学研究所)117 巻 5 号, E85-E86, 2009 年 12 月.
[10] 田中美栄子、伊藤大哲、田中瑶子、木戸丈剛：ランダム行列理論との比較による NYSE 株価 1 時間変動の解析(2), 素粒子論研究(京都大学基礎物理学研究所)117 巻 5 号,E87-E88, 2009 年 12 月.
[11] 田中美栄子、田中瑶子、伊藤大哲,"ランダム行列との比較による NYSE 株価 1 時間変動の相関行列解析,統計数理研究所共同研究リポート(2010 年 3 月)第 241 巻, 「経済物理とその周辺(6)」(統計数理研究所), 27-31, 2010 年 3 月.