Vol.2010-MPS-81 No.11
Vol.2010-BIO-23 No.11
2010/12/16

ランダム行列の固有値分布との比較による米国株価変動のトレンド抽出

田中美栄子，木戸丈剛
鳥取大学工学研究科
鳥取県鳥取市湖山町南 4-101,680-8552

Trend-extraction of Stock Prices in the American Market by Means of RMT-PCM

Mieko Tanaka-Yamawaki and Takemasa Kido
Graduate School of Engineering, Tottori University,
Tottori, 680-8552 Japan

概要

ランダム行列理論から導かれる固有値分布式を利用する主成分抽出法（RMT-PCM）は，株式市場のように多くの要因に依存する複雑系に対して有効性を発揮する．本手法の特徴は，対象とする系の乱雑性と複雑性を積極的に利用する点にある．特に対象とする系が提供してくれる数値データが，乱雑性の極限で成立する理論式が有効であるようなパラメータ領域にぴったり嵌るような場合，本手法は系の特徴抽出のための大変便利なツールとなりうる．このことを確かめるため，我々は，米国株式市場の価格の解析を広範囲に行った．1994 年,1998 年,2002 年の tick 価格と，1994 年～2009 年の日次終値を使用した解析を行うことにより，過去十数年の年次変化や年内変化を追跡することに成功した．そこで本手法をツール化する上での問題点を整理し，広範囲の対象に適用可能なアルゴリズムの確立に処する．

1. まえがき

株式市場においては常時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．このとき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,それに基づいて各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．

主成分抽出の方法はいろいろあるが,株式市場のように要素数もデータ長も膨大である場合にはむしろその乱雑性を積極的に利用する方法が好ましい．本論文で検討する,ランダム行列理論[1,7]から導かれる同時刻相関行列の固有値分布の公式[2]を利用する方法（RMT-PCM)ランダム部分の従う法則性を理論に任せ,残差部分を主成分として取りだすことにより,従来の主成分分析に比べて主成分数を明確なアルゴリズムに基づいて選出できるのが利点である．

本手法のアイデアは約 10 年前に提案された,多数の時系列間の同時刻相関行列のスペクトルと,ランダム行列から作った相関行列のスペクトルを比較して残差を主成分として扱う手法[3,4,5,6,8]の延長上にあるが,本稿では tick 価格データベースである NYSE-TAQ を用いて一日あたり 6 時点の同時刻相関行列を扱うことで 1 年分のデータのみで手法の有効性を保証するパラメータ領域を確保し,1994 年,1998 年,2002 年と３つの異なる年度の結果を比較できるようにした[9,10,11,12,13]．加えて 1994～2009 の日次終値を用いることにより,2 年分[14],4 年分,8 年分,16 年分という異なる期間を用いた結果を比較できるようにし,結果の整合性と手法の有効性を定量的に検討する．

† 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,

Vol.2010-MPS-81 No.11
Vol.2010-BIO-23 No.11
2010/12/16

2．金融時系列の同時刻相関行列

株価時系列解析では，直接株価を比較するのでなく収益率

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

を使用することが多い．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．もっと便利なのは対数収益

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

であり，対数中の分子は S(t)＋ΔS(t)であるから株価の増分ΔS が株価 S(t)に対して十分小さい時，

$$r(t)=\log\left(1+\frac{\Delta S(t)}{S(t)}\right)\cong\frac{\Delta S(t)}{S(t)} \tag{3}$$

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

$$C_{i,j}=\sum_{t=1}^{T}r_i(t)r_j(t) \tag{4}$$

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$\langle r\rangle$を差引いて分散の平方根σで割っておくことである．

$$x_i(t)=\frac{r_i(t)-\langle r_i\rangle}{\sigma_i} \tag{5}$$

式(5)によって正規化した時系列 $x_i(t)$ の内積を取って式(4)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(4)より，

$$C_{i,j}=C_{j,i} \tag{6}$$

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 $V^{-1}CV$ 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 の固有ベクトルに対応し、次式で表される固有値問題の解となる．

$$\sum_{j=1}^{N}C_{i,j}v_{k,j}=\lambda_k v_{k,i} \tag{7}$$

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，

$$\sum_{i=1}^{N}(v_{k,i})^2=1 \tag{8}$$

異なる列 k と k’に対しては直交する．

$$\sum_{i=1}^{N}v_{k,i}v_{k',i}=0 \tag{9}$$

Vol.2010-MPS-81 No.11
Vol.2010-BIO-23 No.11
2010/12/16

3. ランダム行列スペクトルによる主成分抽出法(RMT_PCM)

相関行列 C の固有値と固有ベクトルを計算し,固有値のうち RMT 理論式と一致する部分はランダム成分として捨て,残差部分を主成分とする．株価時系列の対数収益は乱数に非常に近いために上位数個の固有値を除いて RMT 式で良く近似できる点がこの方法の利点である．相関行列の固有値を大きい方から採用してゆく方法で主成分分析を行うこと自体は従来から知られているが,株価相関のように相関行列の次元が数百以上に及ぶ場合には,RMT 式との比較が意味を持ち,RMT_PCM が有効性を発揮する.RMT 式は $N \to \infty,\ T \to \infty,\ Q \equiv T/N = \text{const.}$ の極限で次式により与えられる[2].

$$P_{RMT}(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{10}$$

ここで固有値λの上限と下限は以下のようである.

$$\lambda_\pm = (1 \pm 1/\sqrt{Q})^2 \tag{11}$$

4. 株式市場の日中変動（1 年データ）

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.

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なければいけない．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.

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.

1994 年の 419 社の 1 時間変動に対する，相関行列の固有値分布は RMT 式に重なるスペクトルとそれより大きな離散固有値に分かれる.1994 年の場合,N=419 社に対する解析結果はランダム理論値の最大値が 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 を有意成分と見なせる．理論式の最大値であるλ$_+$の右の領域に浸み出した連続スペクトル部分は有意成分ではない.この理由として,これらの固有ベクトル成分のランダム性が高いことと,式(2)で対数収益に換算した際に付加わる特徴的な癖[15]が大半であることなどが挙げられる．この点については稿を改めて詳しく論じたい.

1998 年の N=490 社に対する結果は,ランダム部分の最大固有値が,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 が有意成分候補となる.

最後に 2002 年の場合,569 社に対する結果はランダム部分の最大値が,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[5]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.

Vol.2010-MPS-81 No.11
Vol.2010-BIO-23 No.11
2010/12/16

年次変動を見てゆくと,その時代ごとに優勢だった業種がこの解析によってあぶり出されていることが確認できる．1994 年頃までは株式市場をけん引していた,車・鉱業・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．

Table 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)

5　株式市場の日次終値（2 年,4 年,8 年,16 年データ）

以上は，tick データ利用により一日当たり６データを取ることで１年ごとに一つの解析を行い，年次変化を見てきた．これは１年で 252 営業日しかないため，1 ファイル当たりのデータ数が株式数 N＝400～500 以上となる条件を満たすために必要であったためである．しかし 2 年分をつなげたデータを用いれば一日当たり１データしか取れない日次データであっても N<T の条件を満たす．但し，２年では T=504 となるため，Q=T/N 値の境界 (Q=1)近くをとることになり，注意が必要である．我々は 2 つの方法でこの点に対応した．一つは 2 年分のデータを繋げた場合の結果を，4 年分繋げた結果，8 年分繋げた結果と比較し，先の日中データの結果との比較に於いてその中と経過点とみなすことである．いまひとつは，機械乱数を用いて様々な N と T の値に対して本手法の是非をシミュレートし，問題点を見出すこと[12]である，ここでは前者を主体に報告し,後者は稿を改めて論じたい．

日中データの場合と同様の方法により,Table 1 に対応する日次データの結果を Table 2，Table 3,Table 4，Table 5 にそれぞれ 2 年データ,4 年データ、8 年データ、16 年データに対して示す．

各表においては,略称として,エ（エネルギー），材（素材），財サ（資本財およびサービス），生（生活必需品），健（ヘルスケア），金（金融），情（情報技術）,電（電気通信サービス），公（公益事業）を用いた．

Table　2　各固有ベクトルの上位 20 成分の主な業種（2 年データ）

	94-95	96-97	98-99	00-01	02-03	04-05	06-07	08-09
u_1	金 6,公 8	金 8	金 14	財 6 金 10	金 16	金 12	金 17	財 9
u_2	公 17	公 20	公 20	情 20	生 10 公 10	情 19	エ 20	金 16
u_3	エ 16	エ 19	情 11	生 8 金 12	エ 19	金 13 公 7	公 20	生 8 健 6
u_4	偏無	偏無	材 12 財 8	公 20	公 20	偏無	情 9	サ 10 H:10
u_5	生 8 健 8	金 8	健 7 金 7	健 6 金 9	健 20	サ 13	金 9 情 9	サ 8 健 9

各期間ともに u_2 の主要成分が特定の業種に集中し,その業種が期間によって変化しており,これをその期間のトレンドと考えることができる．そこで表 5 から読み取れる各期間の特徴を以下に述べる．

94-95 データでは u_2，u_3，u_5 には業種の偏りがみられるが，u_4 では業種の偏りがみられない．94-95 データでは偏って大きい成分が u_4 以外の固有ベクトルでは 10 成分以上有るのに対し，u_4 は 4 成分しかなかったため，20 成分までみると相関関係がなくなり，業種の偏りがみられないと考えられる．このことから 94-95 データの特徴は u_2，u_3，u_5 から読み取ると公益事業，エネルギー，ヘルスケア，生活必需品関連の銘柄の株がランダムでない動きをしていると考えられる．同様に 96-97 データでは u_2，u_3 には業種の偏りがみられるが,u_4,u_5 では顕著に大きな成分が少なかったため業種の偏りがみられない．また u_5 の固有値 λ_5 が 4.39 と低いため，ランダムな固有値とも考えられる．このことから 96-97 デー

Vol.2010-MPS-81 No.11
Vol.2010-BIO-23 No.11
2010/12/16

タの特徴は u_2，u_3 からのみ読み取ると公益事業，エネルギー関連の銘柄に絞られる．98-99 データでは u_2 の公益事業と u_4 の素材関連の銘柄がランダムでない動きをしていると考えられる．00-01 データでは情報技術，金融，生活必需品，公益事業関連の銘柄が，02-03 データでは u_2～u_5 全ての固有ベクトルで業種の偏りがみられ，公益事業，エネルギー，ヘルスケア，生活必需品がランダムでない変動が起こり特に公益事業について特徴的な変動あったと推測される．04-05 データでは，情報技術，金融，サービス関連の銘柄でランダムでない変動が起ったと推測される．06-07 データでは，エネルギー，公益事業，情報技術，金融関連の銘柄で， 08-09 データでは，金融が目立ち次いでサービス，ヘルスケア，情報技術の銘柄でランダムでない変動が起ったと推測される．

次に業種の偏り以外にも上位 20 成分の相関関係を見る．00-01 データの業種の偏らなかった u_5 の主要な成分は，ランダムな変動をする銘柄の集まりかといえばそうではなかった． λ_5 の値も大きく u_5 の値の大きな成分も多く，相関の値も大きかったため主要成分同士は相関関係にあると思われる．つまり，00-01 データの u_5 の主要成分は業種によらない相関関係にある銘柄だと考えられる．他にも 04-05 データの u_2 の成分では大きな要素はなかったが相関関係にある要素の集まりであった．
各成分の業種内訳の定量的な比率を棒グラフに表したものを Fig1 に示す．左図は日中変動の 1 年データによる結果であり，右図は日次変動の 2 年データの結果である．

Table 3 各固有ベクトルの上位 20 成分の主な業種（4 年データ）

	94-97	98-01	02-05	06-09
u_1	金 6, 公 7, 健 4,	金 11, 財サ 7, 情 2	金 14 財サ 3 情 3	金 8 財サ 7 材 5
u_2	(+)公 20/(−)情 20	(+)公 20/(−)情 20	(+)エ 16 公 4/(−)情 20	(+)エ 15 公 4 材 1/(−)金 17
u_3	(+)エ 19/(−)情 19	(+)エ 20/(−)金 9 生 8	(+)エ 16 情 4/(−)生 16 金 4	(+)公 11 生 6 健 3/(−)材 1 エ 19
u_4	(+)偏無/(−)エ 12 公 8	(+)公 19/(−)材 13	(+)公 20/(−)エ 17 財サ 3	(+)情 15 財サ 5/(−)金 19
u_5	(+)偏無/(−)偏無	偏無	(+)金 14 財サ 4/(−)健 17	(+)健 14 金 2/(−)公 11 金 9

Table 4 各固有ベクトルの上位 20 成分の主な業種（8 年データ）

	94-01	02-09
u_1	金 13, 財サ 7	金 11 材 4 財サ 4 情 1
u_2	(+)公 20/(−)情 20	(+)エ 19 公 1/(−)金 20
u_3	(+)金 10 生 4 公 4 電 2/(−)エ 20	(+)情 19 健 1/(−)エ 8 金 12
u_4	(+)情 8 公 12/(−)材 12 財サ 4 金 3 生 1	(+)公 14 生 5 健 1/(−)情 20
u_5	(+)金 4 公 1 生 1 健 10 金 3/(−)材 12 財サ 8	(+)財サ 17 生 3/(−)公 6 金 14

Table 5 各固有ベクトルの上位 20 成分の主な業種（16 年データ）

	94-09
u_1	金 12, 材 4, 財サ 3
u_2	(+)情 20/(−)公 17 エ 3
u_3	(+)エ 13 情 7/(−)金 20
u_4	(+)公 19 生 1 /(−)エ 19 金 1
u_5	(+)生 8 材 5 財サ 7/(−)金 14 公 3 情 3

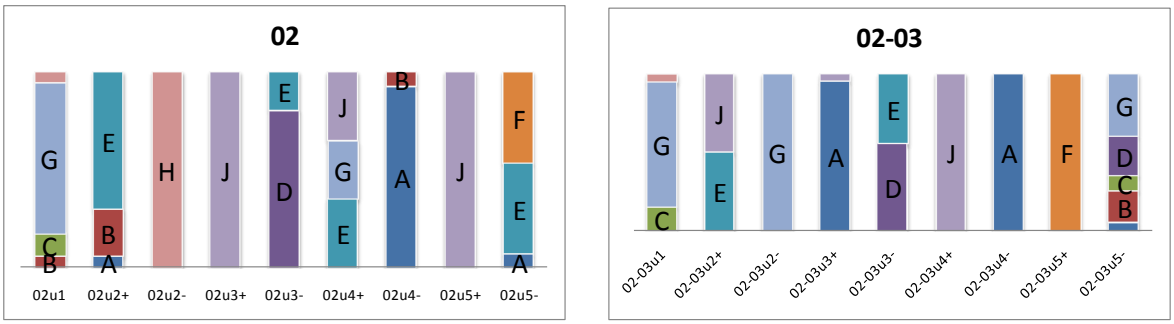

Fig1. 固有ベクトル成分の業種内訳　日中価格 2002 年データ(左) 日次終値 2002-2003 年データ(右)

6 まとめ

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_PCM) 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.日次終値を使った解析では,16年にわたるデータが使える一方で,データ長Tを大きく取るには長い期間をひとまとめにしなければならず,最低2年分が必要である.しかも2年分をまとめただけでは T=504 となって Q=T/N の適用範囲の境界に近く,理論式の信用度が落ちる.4年分をまとめると Q>2 が保障され,適用範囲の問題はなくなる.本稿では2年データから16年データまで長さを替えて実験を行い,それらを比較することで,結果に大きな問題が生じないことを実証した.

参考文献

[1]例えば M. L. Mehta, “Random Matrices”, Academic Press 3rd edition, 2004.
[2] A. M. Sengupta and P. P. Mitra, “Distribution of singular values for some random matrices”, Physical Review E 60, pp. 3389-, 1999.
[3] V. Plerou, et. al,“Random matrix approach to cross correlation in financial data”, Physical Review E 65, 066126, 2002.
[4] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, Physical Review Letters, 83, pp.1471-1474, 1999.
[5] L. Laloux, P. Cizeaux, J. -P. Bouchaud, and M. Potters, Physical Review Letters, 83, pp.1467-1470,1999.
[6] 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.
[7] 永尾太郎,ランダム行列の基礎, 東京大学出版会, 2005.
[8] 青山秀明, 他:経済物理学, 共立出版, 2008.
[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.
[12] 伊藤大哲,″ランダム行列理論の固有値地分布に基づく主成分分析手法の適用条件″, 鳥取大学工学部平成21年度卒業論文.
[13] 田中美栄子,木戸丈剛,″ 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",FIT2010：第9回情報科学技術フォーラム講演論文集(電子情報通信学会・情報処理学会) pp.153-156,2010.
[14] 木戸丈剛,田中美栄子,″ランダム行列の固有値分布との比較による米国株価日次変動のトレンド抽出", FIT2010：第 9 回情報科学技術フォーラム講演論文集(電子情報通信学会・情報処理学会) pp.157-162, 2010.
[15] Mieko Tanaka-Yamawaki, "Extracting Principal Components from Pseudo-Random Data by Using Random Matrix Theory", Econophysics Colloquium 2010 (Taipei, Nov.4-6, 2010).