A-021

東証 tick 価格の RMT 主成分分析による 3 箇月単位の主要セクタ抽出

Extraction of Major Industrial Sectors per Quarter Year Based on the Application of the RMT-PCA on the tick-wise stock Prices in the Tokyo Market

山本 敦史† 田中 美栄子†
Atushi Yamamoto Mleko Tanaka-Yamamwaki

1．はじめに

ランダム行列の固有値分布がデータ長とデータ数の比のみで表される簡単な関数となることを利用し，多量な時系列データの変動から相関の強い変動をするデータとランダムな変動をするデータに分離することができる．これは，多量の時系列データが測定されている事象ならば適用可能なため，金融工学や地震学等，生態学など様々な分野で研究されている[1-6]．我々はこの手法に RMT-PCA（Random Matrix Theory oriented Principal Component Analysis）と名付け，株式データへの適用を試みている．これにより，大量の株価の数値データから株価が連動する銘柄を抽出できれば銘柄選択の指標になるのではないかと考えた．

我々の研究室では S&P500 リストの 500 社の日次データから 2 年単位の主要業種の分析を行っている[7-9]．しかし，近年の投資の IT 化により短期間の投資を行う人が増えており，短期間の動向を知る需要が高まっている．そこで本稿では，TOPIX500 の構成銘柄の 30 分毎の株価を用い，1 年単位と 3 箇月単位での主要セクタの抽出を行った．さらに,抽出された銘柄を経済に関連する業種をまとめた関連株に分類し，実際の経済動向との比較を行った．

2．相関行列の固有値分布

ランダムデータに対しては相関行列の固有値分布の理論式は

$$Q = \frac{L}{N} > 1 \quad , \quad \mathrm{N} \to \infty \quad , \quad \mathrm{L} \to \infty \qquad (1)$$

$$\lambda_{max} = \left(1 + \frac{1}{\sqrt{Q}}\right)^2 , \lambda_{min} = \left(1 - \frac{1}{\sqrt{Q}}\right)^2 \qquad (2)$$

$$P_{rmt}(\lambda) = \frac{Q}{2\pi\lambda}\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{\min})} \qquad (3)$$

で与えられる．いまN 行L 列のランダム行列があるとすると，L とN の比率Q = L/N より最大固有値：λ_{max}と最小固有値：λ_{min}を求めることができる．つまり，固有値分布はQ にのみ依存する．実データはランダムではないので式 (3) に一致せず，実際に求めた相関行列の固有値はこの式の範囲から飛び出してしまう．飛び出した固有値に相当する自由度が，ランダム成分ではない主成分ということになる．

†鳥取大学大学院工学研究科

3. 主要業種の抽出方法

3.1 株価データの編集

今回，2007 年 1 月 4 日~ 2009 年 12 月 31 日までの期間の 2009 年の TOPIX500 構成銘柄（平成 21 年 10 月 30 日現在）から各銘柄 30 分毎の株価を用いる．ランダム行列理論を用いるには同時刻にすべての銘柄の価格データが必要となる．本研究では過去研究同様，取引が行われなかった時刻を現状維持と考え，一つ前の時系列の株価と同じ値を与えることで補正を行っている．

3.2 データの正規化

まず，各銘柄 30 分毎の株価の変動を調べる．銘柄により株価に大きな差が存在するためここでは対数差を算出し変動の割合に着目する．さらに

$$g_{i,j} = \frac{X_i(t) - \langle X_i \rangle}{\sqrt{\langle X_i^2 \rangle - \langle X_i \rangle^2}} \qquad (4)$$

より求めた対数差の平均値が 0，分散が 1 となるように正規化を行う．

3.3 相関行列の算出

正規化した値より *N* 行 *L* 列の行列 *G* を作成する．

$$G = \begin{pmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,L} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ g_{N,1} & g_{N,2} & \cdots & g_{N,L} \end{pmatrix} \qquad (5)$$

*N*は銘柄数，L は時系列長を示す．作成した行列の各銘柄との同時刻の内積を取り相互相関行列を算出する．

$$C = \frac{1}{L} G^T G \qquad (6)$$

この行列 C の(i, j)要素は規格化された銘柄 i と銘柄 j の時系列変動間の内積として[−1:1]の範囲の値を取り，二つの時系列の類似度に相当する．この C の固有値を求め，ランダム行列理論の理論式による固有値分布と比較する．

3.4 固有値の比較による主成分の分離

理論式は N → ∞, L → ∞に於いて成立するが､これと比較するデータは N も L も有限である．本稿で使用する東証データは N=486，L= 642 であり，これと同じパラメータを持

つ擬似乱数を用いて，ランダム理論により導かれた式(1)で近似できるのかどうかを確認する必要がある．結果は図 1 に示すように理論式と擬似乱数に依る結果が良く一致しており，ここで使う N と L に対してランダム行列理論を用いることの妥当性が示された．

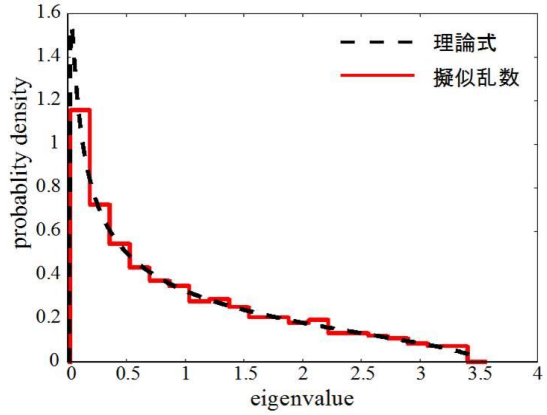

図 1. 擬似乱数の固有値分布

一方，2009 年 1~3 月の株価データより作成した相互相関行列の固有値分布と同じパラメータの理論式の比較結果は図 2 に示すように，最大固有値 3.5 より以上の固有値が現れた．時系列間の変動に相関が無い場合大きな固有値がたくさん現れた，これらの大きな固有値を主成分としてランダム成分から分離する．

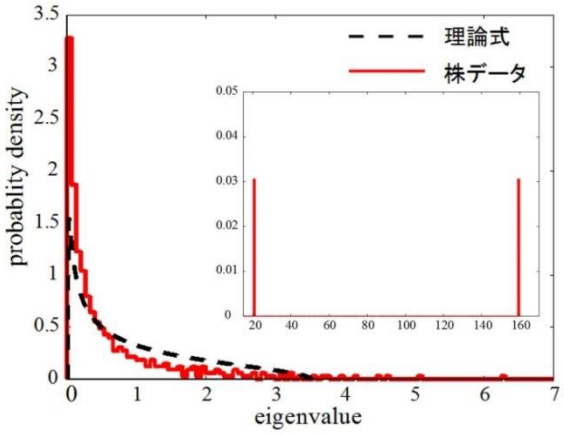

図 2. 株データの固有値分布

3.5　固有ベクトルの業種内連動

固有値を大きい順に並べ，λ_1，λ_2，…，λ_N とし，それぞれに対応する固有ベクトルをu_1，u_2，…，u_N と呼ぶこととする．固有値が大きいほどその固有値が相関の強い成分を持つことを示すため，まず第 1 固有値：λ_1に着目する．第 1 固有値の固有ベクトル成分：u_1 を図 3 に示す．

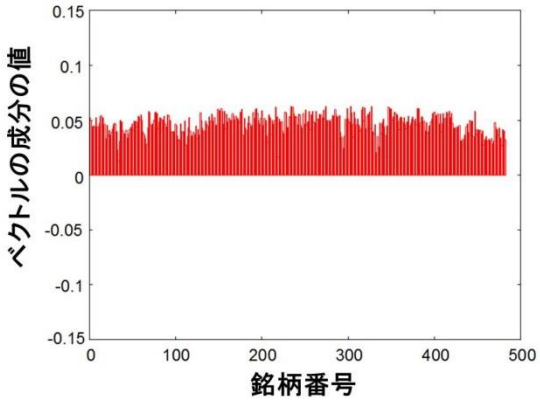

図 3. 第一固有ベクトル成分

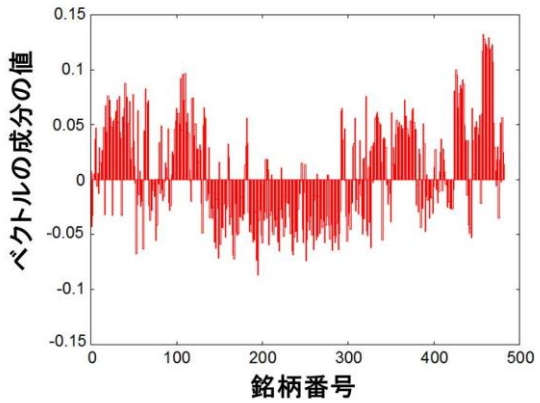

図 4. 第二固有ベクトル成分

図 3 では全ての銘柄のベクトルの成分が正値で平均的であり，銘柄間の差異が見られない．ここで例として 2009 年 1~ 3 月を挙げると，この期間のλ_1の値は 157 を取り，全ての固有値の和(=N) は 33 %である．これは固有値 476 個の内，1 つの固有値が約 3 割を占めているということであり，非常に大きい値であることが分かる．文献[10] にはu_1の固有状態は TOPIX 指数のような市場全体の変動を示すと解釈できることが記述されている．u_1から有効な情報が得られないため次点に大きい固有値である第 2 固有値の固有ベクトル成分：u_2に着目し，これを図 4 に示す．

図 4 を見るとベクトルの成分は正負に突出しており銘柄の偏りを顕著に示している．東証株価は業種によって証券コードが割り振られているため，図 4 を見ると同業種と強い相関を示していることが分かる．さらにこの突出した要素は同業種が同じ符号に固まる傾向を持つため，u_2 成分が正に大きい 10 銘柄，負に大きい 10 銘柄をそれぞれ相関の高い銘柄として抽出し，これらの銘柄の属する業種を表 1 の分類表に従い同定した結果を次の 4 章に示す．図にするにあたり，17 個の業種それぞれに銘柄コードを参考とした番号を付与した．

表 1. 17 業種による分類

「17：建設・資材」	「80：商社・卸売」
「20：食品」	「81：小売」
「30：素材・科学」	「83：銀行」
「45：医薬品」	「85：金融」
「50：エネルギー資源」	「88：不動産」
「54：鉄鋼・非鉄」	「90：運輸・物流」
「60：機械」	「94：情報通信・
「65：電機・精密」	サービスその他」
「70：自動車・輸送機」	「95：電力・ガス」

4. 主要業種の分析

4.1 TOPIX17 による業種分類

図 5，図 6 に TOPIX17 による各期間の主要業種の抽出結果を示す．抽出した銘柄を業種に分類すると正と負の 2 つの成分の内，一方には資源や機械等の業種，一方には金融やガス等の業種が固まることが分かった．これを資源，機械と金融，ガスで別々に時系列順に並べると，それぞれに時系列による主要な業種の移り変わりを見ることができた．

まず，先行研究[7-9] で行われていた 1 年間単位の主要業種の抽出結果を図 5 に示す．図 5（左）を見ると 2007 年は「54：鉄鋼・非鉄」の割合が突出しており主要業種であったことが分かる．また図 5（右）を見ると，2007 年は「83：銀行」，2008 年，2009 年では「95：電力・ガス」がそれぞれ主成分を占めており，相関の高い変動があったことが分かる．

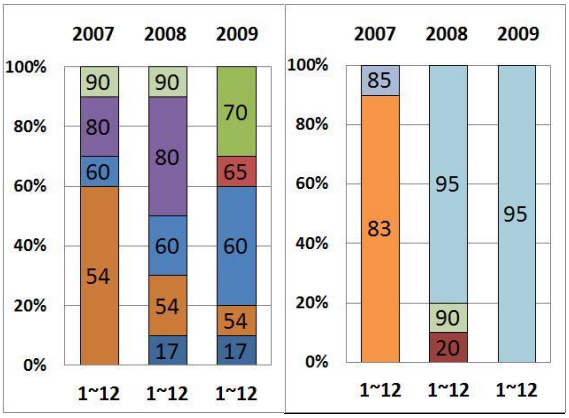

図 5. 1 年間単位の主要業種の抽出結果
（左）機械等，（右）金融等

次に，3 箇月単位の主要業種の抽出結果を図 6 に示す．図 6（上）を見ると，2007 年の 1 月~3 月から 2008 年の 4 月~6 月まで「54：鉄鋼・非鉄」が常に主要業種を占めており，さらに割合の増減の流れを見ることができる．また 2008 年の 7 月~9 月，10 月~12 月を境に，「65：電機・精密」が徐々に増えている事もわかる．図 6（下）でも 2008 年の 10 月~12 月を境に，それまで「83：銀行」が主成分を占めることが多いが，「95：電力・ガス」に移り変わっている様子を見ることができる．なお，図 6 の 2007 年の 7 月~9 月，2008 年の 10 月~12 月は共に複数の業種が集まっており，主要と言える業種が見られない．これは 2007 年 8 月にサブプライムローン問題を発端とした株価急落，2008 年 10 月にはリーマンショックが起こり，多くの銘柄が一斉に下落したため主成分に複数の業種が集まったと考えることができる．このような変動は 1 年単位の分析では見ることができなかった．3 箇月単位で分析を行うことで 1 年単位では見られなかった主要業種の細かい移り変わりを見ることができるようになった．

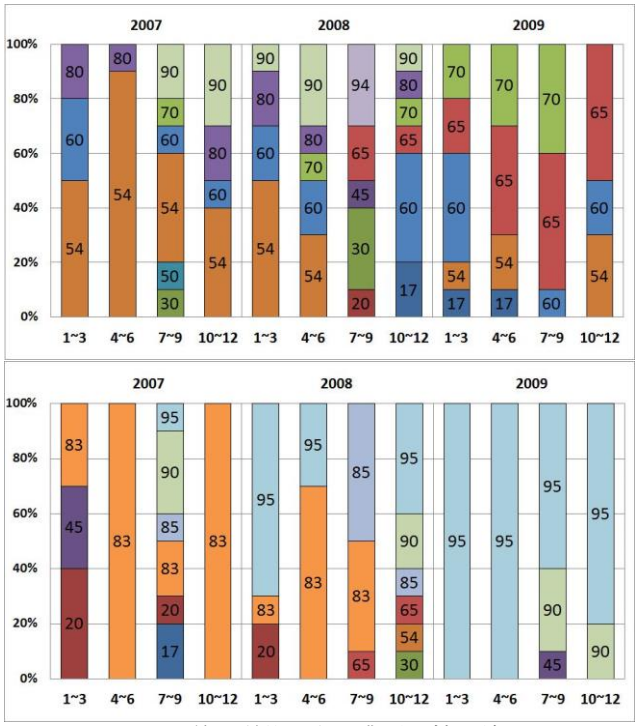

図 6. 3 箇月単位の主要業種の抽出結果
（上）機械等，（下）金融等

4.2 関連株による分類

ここまでは TOPIX17 による業種の分類を行ってきた．この節では，以下の表 2 の 7 つの関連株による分類を行い，上記で行った 17 業種の分類を元に関連株としてまとめ，実際の出来事と対比することで結果の評価を行う．7 つの関連株に分けるにあたり，17 業種に分類する時と同様に業務内容が複数の関連株に属している銘柄もある．ここでは [11] を参考に関連株による分類を行った．各棒グラフに 3 箇月単位で区切った業種の比率とその頭文字を示した．

図 7 に 3 箇月単位の主要業種の関連株による分類結果を示す．本研究で扱った時期は 2008 年 11 月のリーマンショックの影響が大きかった為に 2009 年の 1 月~3 月以降，ディフェンシブ株が図 3 に示す業種関連株に大きく表れる．これは，2008 年 11 月のリーマンショック後に市場が混乱する一方，その影響が少なく変動が安定していた「電力」株が主要関連株に現れたと考えられる．また図 9 では，2008 年の 4 月 6 月を境に主要関連株が市況関連株から外需関連株へ徐々に移行していく様子を見ることができる．市況関連株は原料の取引価格，外需関連株は国外での業績に影響する銘柄であり，実際に 2007 年~2008 年にかけて原油価格が急高騰している．2008 年の 11 月にはトヨタショック，その後もクライスラーや GM の破綻など外需に影響する出来事が起きており，図 7 と一致する．このように第 2

固有値の変遷を追うことで株の大まかな動向を知ることができる．

表 2. 関連株の分類

分類名	内訳-TOPIX17 による業種分類
外需関連株	「65:電機・精密」， 「70:自動車・輸送機」
内需関連株	「17:建設・資材」， 「88:不動産」
景気循環株	「30:素材・化学」，「60:機械」
ディフェンシブ株	「20:食品」，「45:医薬品」， 「95:電力・ガス」
消費関連株	「81:小売」， 「90:運輸・物流」， 「94:情報通信・サービスその他」
金利敏感株	「83:銀行」，「85:金融」
市況関連株	「50:エネルギー資源」， 「54:鉄鋼・非鉄」， 「80:商社・卸売」

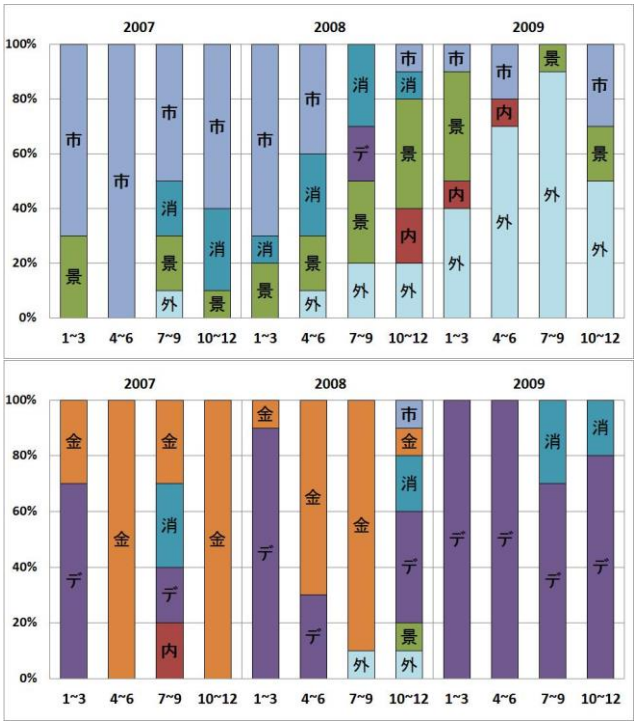

図 7. 関連株による分類（3 箇月），
（上）機械等，（下）金融等

5. まとめ

本研究では 2007 年~2009 年の東証株価データを 3 箇月単位に区切り各期間で主要セクタの抽出を行った．各期間で 2 番目に大きい固有値の固有ベクトルに着目し分析をおこなった．今回ベクトルの値が正，負に突出したそれぞれ 10 銘柄を主要セクタとして抽出することで，2 つの相関の高い銘柄群とその変遷を見ることができた．さらに評価法として関連株に分類し実際の経済動向との対比を行ったところ，実験結果と実際の経済動向に高い相関を見ることができた．しかし，現在，株の取引は秒単位で行われており今回の結果から短期投資戦略の指標とするのは難しい．今回ランダム行列理論を用いるにあたり，$Q = L/N > 1$ という条件の下，L を増やすことで短期化が可能になった． N を減らして実験を行えば更なる短期化を行うことができると考えられる．また，本稿では述べていない第 3 以降の固有値の固有ベクトルの分析を行うことが今後の課題である．

参考文献

[1] V.Plerou, P.Gopikrishnan, B.Rosenow, L.A.N.Amaral and H.E.Stanley:Random matrix approach to cross correlations in nancial data, Physical Review E, Vol. 65,pp. 066-126 (2002).

[2] M.L.Mehta: Random Matrices, Academic Press, 3rd edition (2004).

[3] D.Wang, et.al.,:Quantifying and modeling long-range cross correlations in multiple time series with applications to word stock indices : Physical Review E, Vol. 83,pp. 046-121 (2011).

[4] B.Podobnik, et. al.,:Time-lag cross-correlations in collective phenomena，EPL, Vol. 90，pp. 68001 (2010).

[5] R.N.Mantegna, et. al.,:Hierarchical structure in financial markets，EUR.Phys.J.B, Vol. 11，pp. 193-197 (1999).

[6] A.M.Sengupta, P.P.Mitra:Distribution of singular values for some random matrices, Physical Review E, Vol. 60, pp. 33-89 (1999).

[7] 木戸丈剛，楊欣，田中美栄子，高石哲弥：RMT 公式を用いた主成分抽出法による日本及び米国株価の年次トレンドの比較，情報処理学会論文誌，数理モデル化と応用，Vol. 4，pp. 104-110，(2011).

[8] 木戸丈剛，田中美栄子：ランダム行列理論との比較による米国日時変動のトレンド抽出, FIT2010, 第 9 回情報化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).

[9] M.Tanaka-Yamawaki, T.Kido, R.Itoi:Trend Extraction of Stock Prices in the American Market by Means of RMT-PCA, Inteligent Decision Technologies, SIST 10, pp. 637-646 (2011).

[10] 青山秀明，家富宏，池田洋一，相馬亘，藤原義久：経済物理学，第 5 章(2008).

[11] NSJ 証券：初めての株式投資，
http://www.nsjournal.jp/web contents/stock hajimete/3-2.html.