複数時系列データのソフトクラスタリングとその金融市場への応用

桐畑　誠† 馬　強†

† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †kirihta@db.soc.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　金融市場は様々な要因が絡み合って形成されており，投資者の認識も異なる場合が多い．市場の状況を理解することが難しいため，投資家，とりわけ初心者が効率よく取引するのが困難である．市場の変化は，突然起こる大暴落や緩やかな上昇・下落など様々な形態があり，このような市場の状況を分析することが，投資家の意思決定支援に重要である．本研究では，多様な市場の変化を捉えるため，部分時系列における異常値の割合の変化に着目したクラスタリング手法を提案する．

キーワード　ソフトクラスタリング，時系列データマイニング，Gaussian graphical models

1 はじめに

近年，NISA(少額投資非課税) 制度の効果もあって，ますます多くの人が投資を行うようになってきている．さらに日本政府は積立 NISA やジュニア NISA といった少額からの投資も援助する施策を打ち出し，投資を奨励している．金融庁による NISA 制度の効果検証結果 [1] によると，NISA 口座の開設者の約 3 割は投資未経験者であり，NISA の導入によって投資未経験者への投資の裾野拡大の効果があったと見られるとともに，その効果は若い世代ほど大きかった．

平成 30 年 3 月時点での NISA 口座は 1167 万，累計買付金額は 13 兆円であり，制度の開始以来，順調に推移している．しかし，日本証券協会の調査 [2] によると，株式や投資信託を含む有価証券の保有率は約 18%とかなり少ない．また，金融商品に興味を持っている人の割合は金融商品の保有率よりも高くなっている．すなわち，金融商品には興味があるが，まだ金融商品の購入に踏み切れていない人がいるということである．株式・投資信託・公社債の非購入理由のアンケート結果では,「興味がない」以外では「十分な知識をまだ持っていない」「ギャンブルのようなもの」や「値動きに神経を使うのが嫌」といった金融商品に対する苦手意識が目立った．これは，金融市場の状況を理解するのが難しいことが 1 つの原因であると考えられる．

金融市場は，緩やかな上昇・下落や突然起こる大暴落など様々に変化していくが，この市場の状況を理解していないと大きな損失を出すことがある．しかし，金融市場というのは様々な要因が絡み合って形成されており，投資者の認識も異なる場合が多い．市場の状況を理解することが難しいため，投資家，とりわけ初心者が効率よく取引するのが困難である．したがって，金融市場の局面を推定することで，市場理解を支援し，初心者の投資活動をサポートする．

近年，自動車や人の動作推定など時系列データから状態を推定する研究が盛んであり，一定の成果を上げている．そこで，自動車に取り付けた各種のセンサーデータからカーブや減速などの運転動作を推定する TICC(Toeplitz Inverse Covariance-Based Clustering) [3] の金融市場分析への応用を試みる．動作推定の研究では，動作を識別することに重きを置いているため動作が急激に変化しているものを対象にしていることが多い．しかし，金融市場は，緩やかな上昇・下落など時間がかかる変化があるため，既存手法ではこれらの変化に対応出来ない．また，2008 年 9 月に起こったリーマンショックのように過去に類を見ない局面が発生することがあるため，クラスタ数の設定やクラスタの定義が必要な既存手法では処理できない場合が多い．

そこで，本研究では，データ点の割り当てに、異常検知を応用することで想定外の局面に対応したクラスタリング手法である NEO-TICC を提案し，金融市場においてその有用性を確かめる．本研究での貢献は，時系列の状態推定において，想定外の局面に対応したクラスタリング手法を提案し，実験を通じてその有効性を検証したことである．

本稿の構成は次の通りである．2 節で関連研究について整理し，3 節で我々が提案するクラスタリング手法について説明する．4 節では，予備実験を行っている．5 節で結論および今後の課題を述べる．

2 関連研究

市場に影響を及ぼす情報の全てに目を通し，正確な意思決定を行うのは極めて困難である．そのため情報技術を用いた市場分析は注目を集めており，一定の成果が報告されている [4]... [8]. 粟納らや大西らの研究 [4] [5] では，運用報告書やニュースなどのテキストデータと基準価額などの数量データを併用して，投資信託商品を状態空間モデルでモデリングすることで，価格に影響を与えている要因を明らかにした．さらに，大西らはトレンドと呼ばれる上昇局面と下降局面に時系列を分割して分析することで，要因分析の精度を高めると共に，各局面で要因の影響度に違いがあることを確かめた．また，我々の先行研究 [6] では，柔軟なトレンド検知手法をモデルと組み合わせることで予測性能が向上させ，トレンドの重要性を示した．渡部 [7] はマルコフスイッチングモデルを用いて日本の景気循環を分析した．景気一致指数という日本の景気を表した指数に対して，好

景気と不景気を表現した 2 つの状態変数を追加して，景気循環の分析を行い，景気転換点の分析やリーマンショックにおける構造変化を発見した．

これらの研究は単一時系列に対してトレンドや景気を考えているが，実際の金融市場は多数の時系列が関わり合って成り立っているので，複数時系列で考える必要がある．Birch ら [8] は DAX30 の日次リターンに対する相関関係を利用した企業ネットワークを推定した．ネットワークに対してクラスタリングを行ったところ，同じクラスタに同じ業種の企業が含まれるように分類することが出来た．さらに，リーマンショックとその回復時のネットワークを比較すると，不況時には企業間の繋がりが弱くなり，好況時には企業間の繋がりが強くなることを示した．

自動車や人の動作推定など時系列データから状態を推定する研究が近年盛んであり，本研究では，この技術の金融市場分析への応用を試みる．松原ら [9] は手足につけたモーションセンサーのデータに対して，隠れマルコフモデルを拡張した MLCM（multi-level chaing model）を用いることで，拍手や羽ばたきなど人間のダンス動作をセンサーデータから推定した．松原らは変数の数などもコストと考えることで，ハイパーパラメータを設定することなく推定可能な手法を提案している．さらに，Google Trend のデータを用いたモーションセンサー以外の時系列データへの応用例も示している．

David ら [3] は Gaussian inverse covariance matrix を推定することで変数間のネットワーク構造を求め，このネットワーク構造の変化を用いて時系列のクラスタリングを行う TICC(Toeplitz Inverse Covariance-Based Clustering) を提案している．自動車のセンサーデータに対してこの手法を適用することで，方向転換や加速，減速などの動作を推定することが出来る．またグラフ構造を用いることで，各動作の時にどのセンサーが重要かを分析することが可能になる．

Whang ら [10] は K-Means 法によるクラスタリングにおいて，各データ点が必ず一つのクラスタに所属しなければいけないという制約を無くし，オーバーラップや外れ値を許したソフトクラスタリング手法 NEO-K-Means を提案している．この手法では，制約を上手く与えることで，ソフトクラスタリングを実現しているため，幅広いクラスタリングに応用できる可能性がある．本論文では，異常検知を応用した時系列のソフトクラスタリング手法を提案することで，より柔軟な金融市場の分析と共に，市場変化の予兆発見を試みる．

3 提 案 手 法

本論文では，異常検知を応用し，想定外の局面に対応した NEO(Non Exhaustive, Overlap)-TICC を提案する. 各データ点がどのクラスタに所属しているかを決める際に，どのクラスタからも異常と判定された点は，現在考えているクラスタの想定外の局面として未知と判定する．提案手法の処理概要を図 1 に示す．

入力として n 次元で長さ T の時系列データ x を受け取り，各

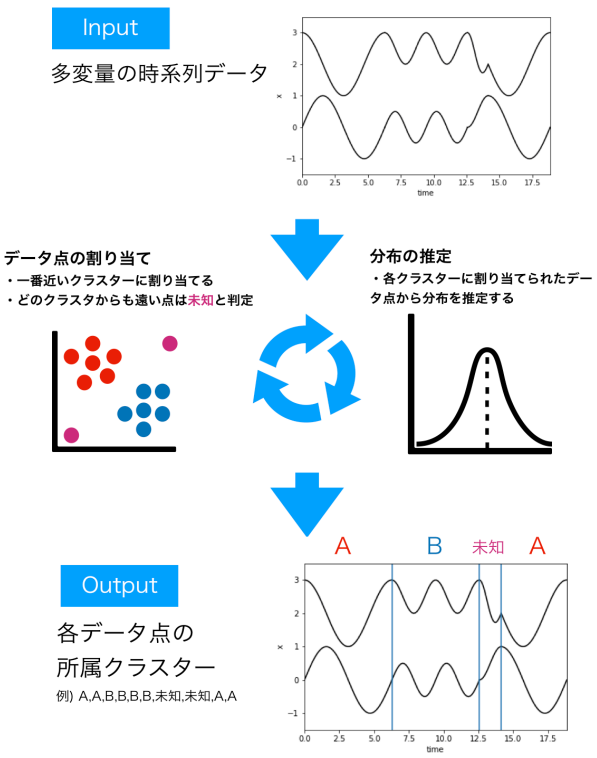

図 1: 提案手法の処理概要図

データ点 x_t が所属するクラスタの系列を出力する．クラスタの数は設定した K 個に，そのどれにも所属しないことを表す未知クラスタを加えた K+1 個である．

未知以外の各クラスタは正規分布 $N(\mu_i, \Sigma_i)$ で表現し，各時点 x_t がどの分布から生成されやすいかでセグメンテーションおよびクラスタリングを行う．以下の二つのステップを変化しなくなるまで繰り返すことで，クラスタリングを実行する．

- クラスタへの割り当て

後述する異常度を用いることで，未知の局面を考慮した各クラスタへのデータ点の割り当てを行う．

- 各クラスタの分布の推定

割り当てられたデータ点を用いて，クラスタ毎の分布を推定する．

3.1 データ点の割り当て

3.1.1 異 常 度

データ点 x のあるクラスタ i に対する異常度 $a_i(x)$ を考える．n 次元正規分布 $N(\mu, \Sigma)$ からの m 個のデータで求めた標本平均 $\hat{\mu}_i$, 標本共分散 $\hat{\Sigma}_i$ を用いて，新たなデータ x の異常度 $a_i(x)$ を以下の式で定義することが出来る．

$$a_i(x) = (x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (x - \hat{\mu}_i) \qquad (1)$$

この式の右辺はマハラノビス距離であり，平均からの離れ具合を共分散によって補正したものになっている．この値が大きいほど，想定している分布の中心から離れた点になり，異常な点と判断される．

ここで，ホテリングの T^2 法 [12] より，$T^2 = \frac{m-n}{(m+1)n} a(x)$ に

より定義される統計量 T^2 は自由度 (n,m-n) の F 分布に従う．したがって，F 分布の性質を用いて，異常と判断するための閾値 a_{th} を求めることが出来る．以下で定義される正答率を α として与えると，正常であるものを異常だと判定する率である誤報率は $1-\alpha$ になる．

$$\text{正答率} = \frac{\text{正しく正常と判断したデータ数}}{\text{正常であるデータ数}}$$

F 分布の確率密度関数を $F(x|n, M-n)$ とすると誤報率 $1-\alpha$ から閾値 a_{th} は以下の式で求められる．

$$1-\alpha = \int_0^{a_{th}} F(x|N, T-N)dx \qquad (2)$$

新たなデータ x に対して計算される統計量 T^2 が a_{th} を超えていればそのデータは異常と判断する．

1 時点のみではなく，部分時系列における異常度を考えることで，局面の変化を捉えることが可能になる．ウィンドウ幅 $w < T$ の部分時系列 X を考えると，X_t は以下のように定義できる．

$$X_t = \begin{pmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-w+1} \end{pmatrix} \qquad (3)$$

X_t における w 個の時点に対して上記の各クラスタからの統計量 T^2 を計算し，閾値 a_{th} を超える時点の数が多いデータ点 X_t はクラスタ i に対して異常と判断する．異常と判定をする個数の閾値を $r = ceil(w \times (1-\alpha))$ とする．ceil() は小数点第一位で切り上げを行う関数である．

図 2 に異常判定の例を示す．異常度が閾値を超える点が幅 w の部分時系列のなかに 2 個以上あればクラスタから異常と判断する．閾値を超えていないデータ点は赤色に，閾値を超えているデータ点は紫色をしている．X_{t-1} の時は閾値を超えている点が 1 個であるので，正常と判断され，X_t では，閾値を超えている点が 2 個になったので，異常と判断される．

3.1.2 割り当てアルゴリズム

Algorithm1 にしたがって，逐次的にデータ点 x_t の各クラスタへの割り当てを実行する．部分時系列に変換した X_t における異常度を用いて各クラスタに対して X_t が異常かどうかを判定する．k 個のクラスタのうち，X_t を正常と判断したクラスタは N_t に，X_t を異常と判断したクラスタは A_t に入れる．N_t と A_t を基に 3 つの場合に分けて割り当てを行う．

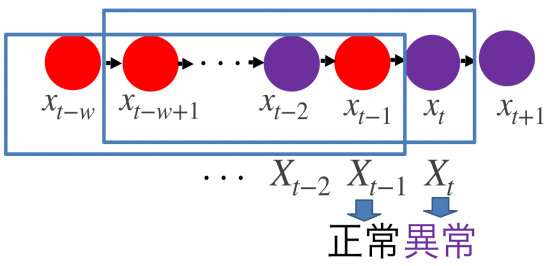

図 2: 異常判定の例

Algorithm 1 クラスタの割り当てアルゴリズム

記号:
$X_t(\in R^{n\times w})$:部分時系列の t 時点における値.
K:k 個のクラスタ集合 $\{1, 2, \ldots, k\}$
$N_t(\subseteq K)$:X_t を正常と判断したクラスタ集合
$A_t(\subseteq K)$:X_t を異常と判断したクラスタ集合
$f(X_t, i)$:X_t の各次元に対して統計量 T^2 を計算し，和を取る関数
p_t:X_t に割り当てられたクラスタ

Input: p_{t-1}, N_t, A_t
Output: p_t

```
if A_t = K then
    p_t = −1
end if
if N_t ≠ φ then
    if p_{t−1} ∈ N_t then
        p_t = p_{t−1}
    else
        p_t = arg min_{i∈N_t} f(X_t, i)
    end if
end if
return p_t
```

1. 全てのクラスタが X_t を異常と判断 (行番号 1)
   未知クラスタに x_t を割り当てる (行番号 2)
2. t-1 時点と同じクラスタが X_t を正常と判断 (行番号 5)
   x_t を t-1 時点と同じクラスタに割り当てる (行番号 6)
3. t-1 時点と異なるクラスタが X_t を正常と判断 (行番号 7)
   x_t を異常度の合計が一番小さいクラスタに割り当てる (行番号 8)

t=0 の割り当ては 3 番の場合として考え，異常度の合計が一番小さいクラスタに割り当てる．t=1 以降は t-1 時点の割り当てを用いて，逐次的に割り当てを求める．

3.2 分布の推定

各クラスタに割り当てられたデータ点を用いて，分布を推定する．ただし，割り当てのうち，未知クラスタに割り当てられたデータ点は，分布の推定には用いない．

グラフィカル lasso [11] と呼ばれる手法を用いて各クラスタの精度行列 Λ_i をスパース推定する．精度行列は分散共分散行列の逆行列なので，正規分布を求めていることになる．簡単のため，標本平均 $\hat{\mu}_i$ で中心化したデータ $\hat{x}_t$ を考えると，クラスタ i における対数尤度は以下になり，これを最大化するような精度行列を求める．

$$\begin{aligned} &\ln \textstyle\prod_{\hat{x}_t \in C_i} N(x_t|0, \Sigma_i) \\ &= \textstyle\sum_{x_t \in C_i} \{-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln det(\Sigma_i) - \frac{1}{2}\hat{x}_t^T \Sigma_i^{-1} \hat{x}_t\} \\ &= -\frac{|C_i|}{2}\{tr(\Lambda_i S_i) - \ln det\Lambda_i\} + const. \end{aligned}$$

S_i はクラスタ i での標本共分散行列，$|C_i|$ はクラスタ i に割り当てられているデータ数，const. は定数項を表している．この対数尤度の式の {} の中身に正則化項を追加した以下の目的関数を最小化することで，スパースな精度行列を求める．

$$minimize\ tr(\Lambda_i S_i) - \ln det(\Lambda_i) + \lambda\|\Lambda_i\|_1 \qquad (4)$$

λは正則化項であり，この値が大きければ大きいほどスパースな精度行列が推定される．これを各クラスタごとに解くことで，割り当てられたデータ点に応じて分布を推定することができる．

データ点の各クラスタへの割り当てが変わらなくなるまで，割り当てと分布の推定を繰り返すことでクラスタリングを実行する．

4 実験及び評価

4.1 概　　要

2011 年 8 月から 2012 年 10 月までの TOPIXCore30 の日次データを用い，提案手法の確認を行った．実験では，人工的に作成した想定外の局面に対して提案手法が有効であるかどうかを確認し，ケーススタディで株価データから景気の局面を分類出来るか検証している．

4.2 データセットの作成

2011 年 8 月から 2012 年 10 月までの TOPIXCore30 のうち株式会社リクルートホールディングスとアステラス製薬株式会社を除いた 28 社の対数収益率を用いて人工データを作成した．対数収益率は株価を対数に変換し，前日との差分を取ったものである．人工データは好況，不況に加えて，想定外の局面として，時間をかけて分布が変化していく移行局面を含む．2012 年 3 月が景気の転換点であるため，2011 年 11 月から 2012 年 2 月を局面 1（好況）、2012 年 4 月から 2012 年 7 月を局面 2（不況）として、それぞれの分布 $N_1(\mu_1, \Sigma_1), N_2(\mu_2, \Sigma_2)$ を推定する．移行局面では 2 つの分布の合成比率が線形に変化することを想定したデータを生成する．すなわち，局面 1 から局面 2 への移行局面が始まってから t 時点のデータは $N((1-\frac{t}{50})\mu_1 + \frac{t}{50}\mu_2, (1-\frac{t}{50})\Sigma_1 + \frac{t}{50}\Sigma_2)$ の分布から生成される．

[局面 1，移行局面，局面 2，移行局面，局面 1，移行局面，局面 2] の順に分布に従うデータを生成する．局面 1，2 は 150 個ずつ，移行期間は 50 個ずつ計 750 個のデータを生成した．

4.2.1 評　　価

上記の方法で 20 セットのデータを作成し，移行期間は未知を正解ラベルとした三つのクラスへの分類問題として評価を行う．推定に用いたパラメータは，クラスタ数 $k = 2$，ウィンドウ幅 $w = 4$，正答率 $\alpha = 0.9$，$\lambda = 1.0 \times 10^{-10}$ である．クラスタリングの精度は表 1 のようになった．未知クラスタはクラスタ 3，移行局面を局面 3 とし，Num(i,j) = 局面 j のデータをクラスタ i と判断した個数とする時，局面 1 の適合率 P_1, 再現率 R_1，F1 値 $F1_1$ の計算は以下のようになる．

$$P_1 = \frac{Num(1,1)}{Num(1,1) + Num(1,2) + Num(1,3)} \qquad (5)$$

$$R_1 = \frac{Num(1,1)}{Num(1,1) + Num(2,1) + Num(3,1)} \qquad (6)$$

$$F1_1 = \frac{2}{\frac{1}{P1} + \frac{1}{R1}} \qquad (7)$$

局面 2，移行局面についても同様の計算を行い，20 セットの算術平均を取る．

表 1 より，局面 1，局面 2 に関しては適合率も再現率も高く分類できている．移行局面では再現率が低いが適合率は低くはない．すなわち，移行局面のデータを未知と判定できている数は少ないが，他局面のデータを未知と判定している数も多くはないということであり，局面 1 と 2 の分類に大きな影響を与えることなく異常なデータを検知出来ている．

また，実際にデータセットに対して提案手法を適用した各局面ごとの分類結果は表 2 のようになった．移行局面における中盤のデータを未知と判定している．移行局面が線形に前の分布から後の分布に変化することを考えると，中盤はどちらの分布からも離れた分布になっているはずなので，これを上手く捉えることが出来ている．しかし，初期や終盤は局面 1 や 2 からそれほど変化したものではないので，ほとんど 1 か 2 に分類されてしまっている．想定外の局面をどのような局面にするかは今後考えていく必要がある．

4.3 ケーススタディ

2011 年 8 月から 2012 年 10 月のデータに対して，提案手法を適用し，2012 年 3 月で局面が変化するかどうかを確かめる．人工データに対する実験と同じパラメータを用いて推定を行ったところ，図 3 のようになった．図 3 より局面は 2012 年 2 月の初旬に変化しており，2012 年の 9 月頃にまた変化していると推定された．2012 年 3 月が景気の転換点であるので，提案手法で景気の転換点を推定することは出来なかったが，分布が変化していくことは確認できた．未知と判定されている局面がいくつかあるが，何か特別な出来事があったのかなどの原因は調査出来ておらず，今後の課題としたい．

表 1: クラスタリングの精度 (20 セットの平均)

	適合率	再現率	F 値
局面 1	0.738	0.991	0.846
移行局面	0.632	0.0377	0.0698
局面 2	0.882	0.994	0.934

表 2: 局面ごとの混同行列

	クラスタ 1	未知	クラスタ 2
局面 1	147	0	0
移行局面	30	2	18
局面 2	0	0	150
移行局面	26	3	21
局面 1	147	3	0
移行局面	28	2	19
局面 2	0	3	147

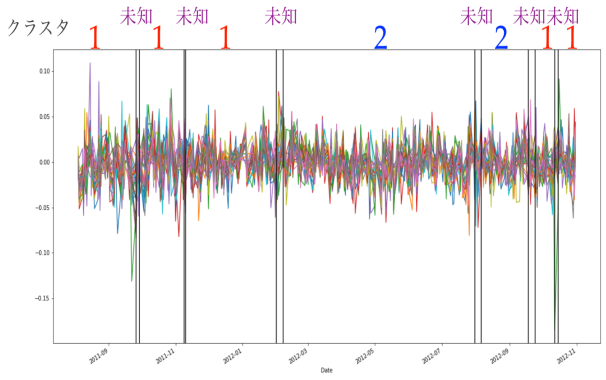

図 3: 実データに対するクラスタリング結果

5 終わりに

本論文では，異常度の変化を用いることで想定外の局面を考慮したクラスタリングを行う NEO-TICC を提案している．クラスタごとに分布を推定し，各時点のデータがどのクラスタの分布に近いかでクラスタリングを行いつつ，どのクラスタからも離れている異常なデータ点を未知と判断することで，想定外の局面に対処している．

TOPIXCore30 の日次データを用いて，移行局面を含んだ人工データを作成し，実験を行ったところ，移行局面を未知と判断することが可能とわかった．さらに，2011 年 8 月から 2012 年 10 月のデータに対して，提案手法を適用したところ，景気の転換点とはずれていたが，金融市場における分布の変化を確認することは出来た．

今回は異常な局面は全て未知と判断したが，実際に金融市場を分析するには未知の局面がどういう局面なのかを詳しくしていく必要がある．実データに対する実験で未知と検知された局面がいくつか存在したため，今後この局面を詳しく調べていくことで，未知の局面をさらに詳しく表現することが可能になると考えられる．

6 謝　　辞

本研究の一部は科研費（16K12532）と 京都大学教育研究振興財団の助成による．

文　　献

[1] 金融庁，NISA 制度の効果検証結果， http://www.fsa.go.jp/policy/nisa/20161021-1.html，(2016).

[2] 日本証券業協会，証券投資に関する全国調査（調査結果概要），http://www.jsda.or.jp/shiryo/chousa/data/files/h30/H30gaiyou20181219.pdf，2018.

[3] Hallac, David and Vare, Sagar and Boyd, Stephen and Leskovec, Jure, Toeplitz inverse covariance-based clustering of multivariate time series data, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 215–223, 2017.

[4] Awano, Yuki and Ma, Qiang and Yoshikawa, Masatoshi, Causal Analysis for Supporting Users' Understanding of Investment Trusts, In Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, pp. 524–528, 2014.

[5] Nobuaki Onishi, Qiang Ma, Factor analysis of investment trust products by using monthly reports and news articles, Twelfth International Conference on Digital Information Management (ICDIM), pp. 32–37, 2017.

[6] Makoto Kirihata, Qiang Ma, Global Analysis of Factors by Considering Trends to Investment Support, International Conference on Database and Expert Systems Applications, pp. 119–133 2018.

[7] 渡部敏明，マルコフ・スイッチング・モデルを用いた日本の景気循環の計量分析，経済研究，Vol. 60，No. 3，pp. 253–265，2009.

[8] Birch, Jenna and Pantelous, Athanasios A and Soramäki, Kimmo，Analysis of correlation based networks representing DAX 30 stock price returns，Computational Economics，Vol. 47，No. 4，pp. 501–525，2016.

[9] Yasuko Matsubara, Yasushi Sakurai, Christos Faloutsos, AutoPlait: Automatic Mining of Co-evolving Time Sequences, ACM SIGMOD Conference, pp. 193–204, 2014.

[10] Whang, Joyce Jiyoung and Hou, Yangyang and Gleich, David and Dhillon, Inderjit S.， Non-exhaustive, Overlapping Clustering. IEEE transactions on pattern analysis and machine intelligence, 2018.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, Sparse inverse covariance estimation with the graphical lasso. Biostatistics, Vol. 9, No. 3, pp. 432—441, 2008.

[12] 井手剛，杉山将．異常検知と変化検知．機械学習プロフェッショナルシリーズ．講談社サイエンティフィク，2015.