JPX
JAPAN EXCHANGE GROUP

日本取引所グループ
JAPAN EXCHANGE GROUP

JPX WORKING PAPER

JPXワーキング・ペーパー

混合ガウスモデルを用いた市場注文状況の変化の検出

宮崎 文吾
和泉 潔
鳥海 不二夫
髙橋 諒

2013年3月19日

Vol. 03

備考

JPX ワーキング・ペーパーは，株式会社日本取引所グループ及びその子会社・関連会社（以下「日本取引所グループ等」という．）の役職員並びに外部研究者による調査・研究の成果を取りまとめたものであり，学会，研究機関，市場関係者他，関連する方々から幅広くコメントを頂戴することを意図しております．なお，掲載されているペーパーの内容や意見は筆者ら個人に属し，日本取引所グループ等及び筆者らが所属する組織の公式見解を示すものではありません．

はじめに

近年、情報通信技術の発達に伴う金融取引システムの進展により、大量の注文が取引所市場に集約される中で、多数の有価証券の売買とそれに伴う情報提供が行われている。このため、取引所市場では、株価や出来高を始めとする高頻度データ（ティック・データ）が日々膨大に発生しており、その効率的な活用が期待される。

今回は、東京大学大学院工学系研究科 和泉 潔 准教授、鳥海 不二夫 准教授の両研究室より、一つの可能性をご提示頂いたので紹介する。本研究は、ティック・データの活用により、より効率的に板状況の把握ができないか試みたものである。詳細は後述する本文に譲ることにするが、本研究では、市場参加者が各気配値段に対して提示する発注株数のティック・データに注目し、統計的見地から板の特性を示す近似的な確率分布（混合ガウス分布）を導出、個別の注文と確率分布との適合度を測ることにより、注文状況の異常性を確認できないか検討している。

グローバリゼーションが進み、企業や投資者が世界のマーケットの中で最も投資環境の良い取引市場を選択することが可能となった現在、市場から発せられる情報を効率的に把握し、市場としての透明性を高めることが、国際的な競争力向上に資すると考えている。今後も、このような研究が発端となり、最新の数理科学の知見が、資本市場に新しい示唆を与えてくれることを望む。

株式会社日本取引所グループ　調査グループ　一同

混合ガウスモデルを用いた市場注文状況の変化の検出*

Change detection of orders in stock markets using Gaussian mixture model

宮崎 文吾†, 和泉 潔‡§, 鳥海 不二夫‡, 髙橋 諒**

2013 年 3 月 19 日

概要

本研究では、株式の板情報から抽出した特徴ベクトルに対し、確率モデルを適用することによって、市場状態の変化を検出する手法について検討した。本研究ではそのために、梅岡らの提唱した手法を基にして、板情報から抽出される特徴ベクトルを学習期間と入力期間に分け、学習期間の特徴ベクトル列から混合ガウスモデルを作成し、モデルに対する入力期間の特徴ベクトル列の適合度を判別することで分析を実施した。パラメータの最適化や定量的な閾値決定の方法、本研究では取り扱っていない直近取引価格や注文間隔、キャンセル注文等の情報活用といった改善の余地は残されるものの、特異な取引を効果的に検出する枠組みの構築に向けた基礎研究となり得るものと考える。

---

* 本稿に示されている内容は、筆者ら個人に属し、株式会社日本取引所グループ及びその子会社・関連会社、及び著者らが所属する組織の公式見解を示すものではありません。また、ありうべき誤りは、すべて筆者個人に属します。

† 東京大学工学部

‡ 東京大学大学院工学系研究科

§ 独立行政法人科学技術振興機構 CREST，さきがけ

** 株式会社日本取引所グループ

１．序論

IT 技術と金融取引システムの発展に伴い、金融市場においては、日々膨大な情報が発生しており、効率的かつ的確な情報把握が求められている。このような現状を踏まえ、本研究では、株式の板情報から抽出される特徴ベクトルから確率モデルを求め、確率モデルと観測データ間の関係性を考察する中で、市場の注文状況における特異性を確認する手法を検討した。特異性を確認する研究対象としては、過去増資に際して内部者取引が報告された銘柄を選択した。これは、通常増資情報が予想される場合、1 株あたりの株式価値の希薄化を懸念した売り注文の増大により、板形状が特異性を帯びやすくなることを想定したためである。数理科学的な見地から対象銘柄から求められる確率分布と観測データとの間にある関係性を検討する中で、板形状が把握できれば、より効率的な市場運営に寄与する可能性があり、本研究はその基礎研究として位置づけられる。

本稿の構成を以下に述べる。第 2 章では、株式市場の注文が反映される板情報を用いて、板の形状把握を試みた研究群や、本研究の提案手法で用いる混合ガウスモデルを応用した研究例を紹介する。続く第 3 章では、先行研究を参考にして構築した分析手法について説明する。具体的には、板情報から得られる特徴ベクトルから混合ガウスモデルを作成し、モデルに対して入力期間の特徴ベクトルの点をプロットすることで、不適合度を算出、特異な取引の検出を試みた。第 4 章では、実際に増資に係る内部者取引が報告された銘柄について、同業種の銘柄との比較分析及び分析対象銘柄における期間別分析等を実施し、考察を行った。最後に、最終的なまとめを第 5 章の結論にて行う。

２．先行研究

２．１　板情報の分析

本研究では、板情報から抽出された特徴ベクトルに対して確率モデルを適用する。そのような研究例として西岡らの研究[1]がある。西岡らは板情報から一定期間に寄せられる注文を抽出し、それらを売り注文、買い注文毎に注文株数により大規模･中規模･小規模と分類し、その回数の比率で 6 次元の特徴ベクトルを作成した。注文の規模は市場参加者（機関投資家、個人投資家など）毎に異なる特徴であるとしている。それらの特徴ベクトルを並行 k-means 法によりクラスタリングし、前場あるいは後場の取引開始から 30 分間の特殊時間帯と取引開始後 1 時間からの 30 分間の平常時間帯の隠れマルコフモデル（HMM：Hidden Markov Model）をそれぞれ作成した。その HMM と入力系列の類似度を計算することにより、入力系列を高い精度で特殊時間帯・平常時間帯に識別することに成功している。また、2008 年 9 月のサブプライムローン問題を端緒とした世界的不況の前と後を同様の手法で分析することにより、不況前と不況後では市場の動きが異なっていることを確認した。

続く西岡らの研究[2]では、同様の手法を用いることにより、新日本製鐵の板情報を分析し、2006 年 12 月に新日本製鐵がウジミナス社を持分法適用会社化した際に相転移を起こしたことを検出し、提案手法では GARCH モデルを用いた分析よりも早い時期に市場の変化を捉えることが可能となることを示した。

本研究と同様に板情報に対して混合ガウスモデルを用いて分析し、市場の状態変化を捉えようとした研究には梅岡らの研究[3]がある。梅岡らは、板情報から抽出した特徴ベクトルに対して混合ガウスモデルを作成し、入力系列をそれぞれクラスタリングした。そこから不適合度やクラスタ集中度という指標を求め、2011 年 3 月の東日本大震災前後において市場の状態が変化していることを示した。また、同様の手法により、2008 年 9 月のリーマン・ショック時に市場の状態変化が観察されたことを、ヒストリカル・ボラティリティによる検出法より早く検出でき得ることを示した。

以上、西岡らや梅岡らの研究は、板情報から抽出した特徴ベクトルに対して、確率モデルを適用することによって、市場状態の変化を検出することができる可能性を示している。

２．２　混合ガウスモデル

本研究では板情報から抽出される特徴ベクトルに対して混合ガウスモデルを適応する。本節ではその混合ガウスモデルの一般的性質と、各分野での応用例を紹介する。

d 次元ガウス分布の確率密度関数は式(2.2.1)で表される。

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \tag{2.2.1}$$

ここで、$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^T$ は各変数の平均を並べた平均ベクトルであり、式(2.2.2)で表される $\boldsymbol{\Sigma}$ は分散共分散行列である。ただしσ_{ij}は変数 i と変数 j の共分散（$i = j$ の場合は分散）である。

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{pmatrix} \tag{2.2.2}$$

クラスタ数Kの混合ガウス分布は、式(2.2.3)のように複数のガウス分布の単純な線形重ね合わせで表される。

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{2.2.3}$$

ここで、$N(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ はクラスタk を表現するガウス分布であり、その混合係数である π_k は

$$0 \leq \pi_k \leq 1 \tag{2.2.4}$$

と

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.2.5}$$

を満たさなくてはならない。

このようにして表される混合ガウス分布は、図 2.1 のように複数の峰を持つような分布をうまく表現することができる。図 2.1 はクラスタ数 2 の 2 次元混合ガウス分布の例である。

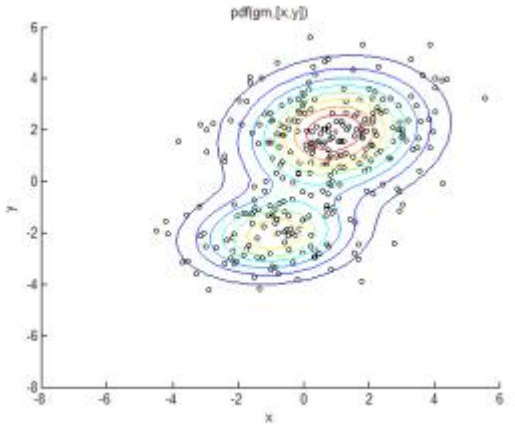

図 2.1　混合ガウス分布の例

（出典　梅岡ら[3]）

混合ガウスモデルは、画像認識や音声識別など様々な分野に利用されている。例えば、村井らの研究[4]は、作動するエレベーターの動画から抽出した 6 次元の学習特徴ベクトル列に対し混合ガウス分布を作成した。そしてそれを動的背景モデルとして、入力特徴ベクトル列が式(2.2.6)を満たした場合に、背景ではない物体（例えば人など）が現われたと判断する。

$$\min_{i \in C} |\boldsymbol{x} - \boldsymbol{\mu}_i| > \theta |\sigma_i| \tag{2.2.6}$$

ここで、θ は判定のための閾値である。θ を大きくすれば判定が厳しくなると言える。このように混合ガウスモデルでは入力ベクトルごとに尤度（起こりやすさ）を判定することができ、その性質を用いて画像認識分野では物体/背景の識別によく用いられる。

また、混合ガウスモデルではクラスタごとに尤度を求めることができるので、入力特徴ベクトルがどのクラスタに属する可能性が高いかを判定することができる。この性質を用いて、混合ガウスモデルは音声情報から各音声に対応する話者を識別するためのモデル化の一般的なアプローチとなっている[5]。

３．分析手法

本研究の分析の流れを図 3.1 に示す。

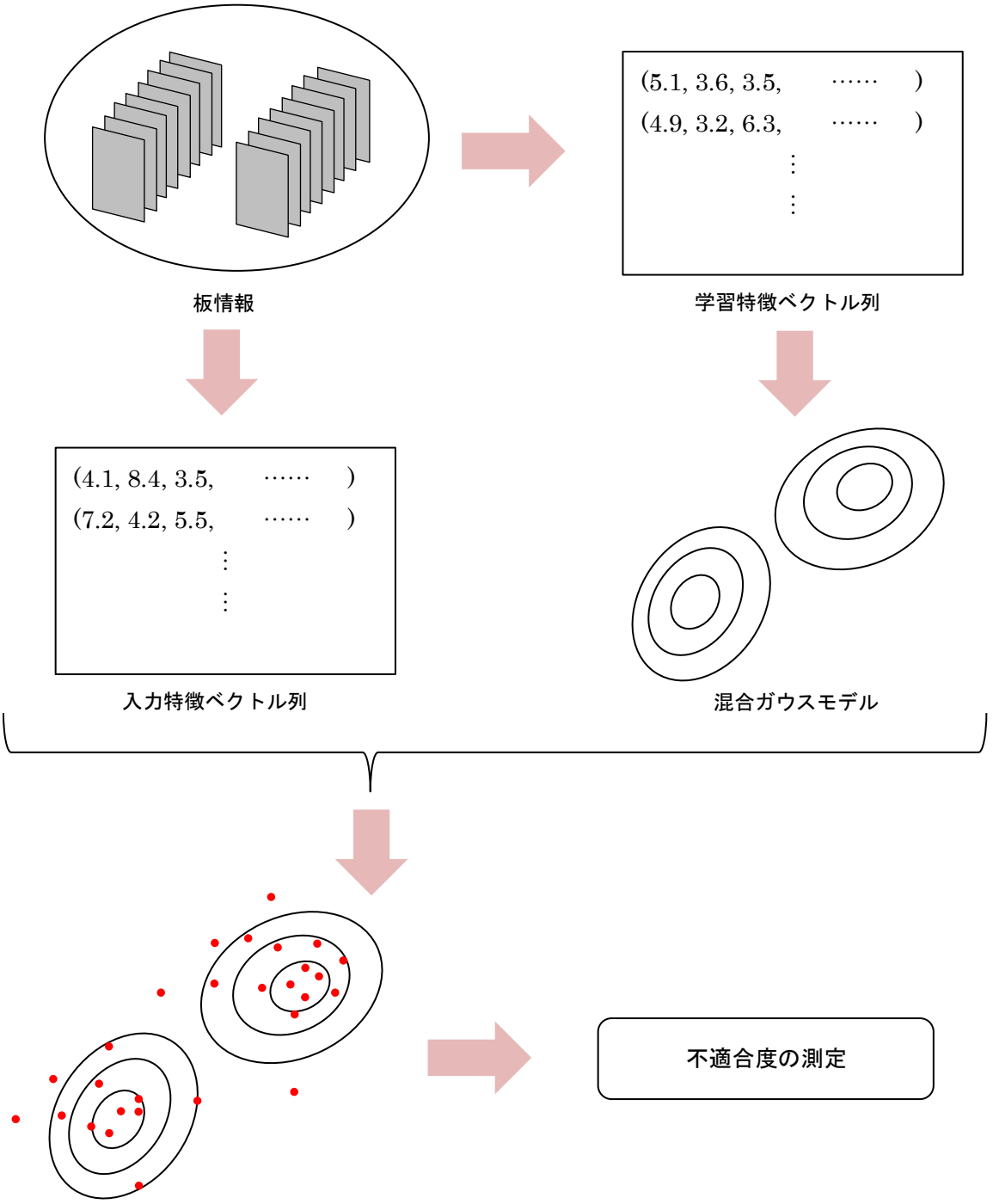

図 3.1　本研究の分析の流れ

本研究の提案する手法は以下のようなステップで行われる。

1．板情報から学習・入力期間ごとに特徴ベクトル列を作成する（3.1 節）

2．学習期間の特徴ベクトル列から混合ガウスモデルを作成する（3.2 節）

3．入力期間の特徴ベクトル列と作成した混合ガウスモデルを比較し、不適合度を計算する（3.3 節）

３．１　特徴ベクトルの作成

本節では板情報からの特徴ベクトルの作成方法について説明する。梅岡ら[3]は、1 日の板情報を 1 分ごとに分割し、板の最良気配値を中心とした各位置に入る注文量を合計し対数を取ったものを、1 分間にその位置に入った回数で割ることにより、8 次元のベクトルを 1 分間に 1 つ得た。そして、その 8 次元ベクトルの各要素に対し、前場の 120 ベクトル、後場の 150 ベクトル*で平均、標準偏差、最大値といった 3 つの特徴量をそれぞれ求め、それをまとめて 1 日一つの 48 次元の特徴ベクトルを作った。本研究では、一定期間に板の最良気配値を中心とした各位置に入る注文量に着目して特徴ベクトルを抽出する。具体的には以下の通りである。

３．１．１　注文量の抽出とラベリング

まず、売り買いそれぞれについて、最良気配値から呼び値の刻みがいくつ分離れているかを、板の最良気配値を中心とした各位置を定義する。具体的には図 3.2 のように定義する。

図 3.2　板の価格位置の例

まず、売り注文に関しては最良売り気配値（図 3.2 の 100 円）を ASK 1 と呼び、A_1 と表記する。ASK 1 から呼び値の刻みが一つ高い価格（図 3.2 の 101 円）を ASK 2 と呼び A_1 と表記する。同様に ASK 3, ASK 4 を定義していくが、ASK 3 以上の位置に入るものは ASK +としてまとめ、A_+と表記する。ASK 1 以下の価格に入った売り注文や成行売り注文（価格を指定せず、即座に執行される売り注文）は便宜上 ASK 0 と呼び、A_0

* 2013 年 1 月現在は前場 150 分、後場 150 分なので、前場後場ともに 150 個のベクトルができる。

と表記する。同様に B_0, B_1, B_2, B_- を定める。

板情報には未執行の累積注文量が記録されているので、そこから特徴ベクトルを抽出するにはまず板情報からある時刻に入った注文量を求めなくてはならない。そこで、連続する 2 枚の板を比較することによって板の変化量を求める。具体的には式(3.1.1)のように各価格における累積注文量の差分を取ればよい。i 番目の板情報スナップショットにおける価格 p の数量を$Q'(i,p)$とすると、その変化量$D'(i,p)$ は

$$D'(i,p) = Q'(i,p) - Q'(i-1,p) \tag{3.1.1}$$

と表される。本研究では価格ではなく最良気配値からの距離に注目しているので、$D'(i,p)$ を価格p ではなくラベル$l(i-1) = [A_0 = p(ASK\ 0),\ A_1 = p(ASK\ 1), \dots, B_- = p(BID-)]$ という対応関係のリストにより対応させなくてはならない。ただし$A_1 = p(ASK\ 1)$ とは、ASK 1 の価格をA_1 というラベルに対応させるという意味である。そのために、$i-1$ 番目の板情報スナップショットにおいて価格とラベルを対応させる関数を用意し、式(3.1.1)をラベルによる表現である式(3.1.2)に変形する。

$$D(i, l(i-1)) = Q(i, l(i-1)) - Q(i-1, l(i-1)) \tag{3.1.2}$$

このとき、注文のキャンセルによりこの値が負になることがあるが、本研究ではキャンセルは考慮せず、そのような$D(i,l)$ は無視する。

なお、基本的な注文抽出の考え方は以上のようであるが、取引が成立したときや新規ラベルに注文が入りラベルに変化があった時（$l(i-1) \neq l(i)$となった時）などは式(3.1.1)や式(3.1.2)だけでは対応できないため、以下のような特殊な操作を施さなければならない。

・$l(i-1) = l(i)$ かつ取引が成立した場合

ラベルが$i, i-1$ 番目の板情報スナップショットで変化しておらず、取引が成立したということは、ASK 1、あるいは BID 1 に入っている注文の一部が執行されたということであるため、注文量は式(3.1.3)のように表すことができる。なお、各時点で取引が成立したか否かは板情報に記載されている。

$$\begin{aligned} D(i, A_0) &= -Q(i, B_1) + Q(i-1, B_1) \\ D(i, B_0) &= -Q(i, A_1) + Q(i-1, A_1) \end{aligned} \tag{3.1.3}$$

これは、B_1/A_1 の減少分が、A_0/B_0に入った注文量と言うことを表す。なお、この場合には、本研究で用いる情報からは成行注文か指値注文化は判断できない。

・$l(i-1) \neq l(i)$ かつ取引が不成立の場合

キャンセルによって注文が消えることは除外してあるので、これは今までに注文が入っていない価格に注文が入ったと考えられる。これはそのためこのときの注文量は式(3.1.4)のように表すことができる。ただし、Pを新しく注文が入った価格とし、L をi 番目の板情報スナップショットにおける価格P の位置とする。

$$D(i,L) = Q'(i,P) \tag{3.1.4}$$

これは、新しい価格に注文が入ったので、「その価格における累積注文量が即ち$i-1$番目の板情報スナップショットからi番目の板情報スナップショットの間に入った注文量である」ということである。

・$l(i-1) \neq l(i)$ かつ取引が成立した場合

取引が成立し、かつラベルに変化があるのは次の 2 通りが考えられる。

（１）買い（売り）の成行注文が大量に入り、BID 1（ASK 1）に入った累積注文量を超えたため、価格が上がった（下がった）場合

この場合の注文量は式(3.1.5), あるいは式(3.1.6)で表される。ただし、Q_{EX} は執行された注文量である。成行注文は、取引が行われている時間帯では即座に全数量執行されるので、注文量は単純に取引執行量と等しくなる。

・買いの成行注文が発注された場合

$$D(i,B_0) = Q_{EX} \tag{3.1.5}$$

・売りの成行注文が発注された場合

$$D(i,A_0) = Q_{EX} \tag{3.1.6}$$

（２）買い（売り）の指値注文が大量に入り、価格が上がった（下がった）場合

この場合の注文量は式(3.1.7), あるいは式(3.1.8)で表される。なお、$B_1(i)$ などはi 番目の板情報スナップショットにおける B_1 の位置という意味である。

・買いの指値注文が発注された場合

$$D(i,B_0) = Q_{EX} + Q(i,B_1(i)) \tag{3.1.7}$$

・売りの指値注文が発注された場合

$$D(i,A_0) = Q_{EX} + Q(i,A_1(i)) \tag{3.1.8}$$

これは、指値買い（売り）注文では対応する価格の売り（買い）数量を超えた場合には、全注文量が即座に執行されずに指値買い（売り）価格の位置$B_1(i)$（$A_1(i)$）に注文量が残るため、その残った注文量と執行された注文量を足すことで出された注文量が求まるということである。

３．１．２　特徴ベクトルの作成

3.1.1 項までで、i 番目の板情報スナップショットに入った注文量$D(i,l)$ が求められ、それはラベル $l=[A_0,A_1,\dots,B\_]$ によってラベル付されている。本項では$D(i,l)$ から特徴ベクトルを作成する方法を説明する。

まず、板情報を T 分ごとに区切り、その T 分間にそれぞれのラベルに入った注文量を合計する。ある T における各ラベルの注文総数 $S(l)$ は式(3.1.9)で表される。

$$S'(l)=\sum_{j=1}^{M}D(j,l) \tag{3.1.9}$$

ただし、M はある T 分間における板情報スナップショット数である。

4 章で紹介する東京証券取引所 FLEX Historical データでは、8:00 ~ 11:00, 12:05 ~ 15:00 計 355 分間の板情報を得ることができるので、T 分ごとに注文総量を求めれば 1 日に$\left[\frac{355-1}{T}+1\right]$回の区間でこれが繰り返される。ただし割り切れない部分の処理については第 4 章で述べる。

また、梅岡ら[3]などは立会時間（実際に取引が行われる時間帯）のみの注文から特徴ベクトルを抽出していた。しかし、事前に未公開情報を持った人の注文は立会時間外に入れられることも十分に考えられるので、本研究では立会時間外の 8:00 ~ 8:59, 12:05 ~ 12:29 の間に入れられる注文も分析対象とする。

以上までの操作で、T 分に 1 つ、$l=[A_0,A_1,\dots,B\_]$ それぞれの注文総量が入った 8 次元のベクトル $\boldsymbol{v'}=\left(S'(A_0),S'(A_1),S'(A_2)S'(A_+),S'(B_0),S'(B_1),S'(B_2),S'(B\_)\right)$ ができる。この v' の各要素に対し式(3.1.10)のように対数を取ることにより正規化する。

$$S(l)=\log\{S'(l)+U\} \tag{3.1.10}$$

ただし、U は対象株式の単元株数である。

３．２　混合ガウスモデル

本節では、前節で作られた学習特徴ベクトル列、学習期間(日)×1 日のベクトル数 $=N$ 個の 8 次元ベクトルに対して混合ガウスモデルを作成する方法を述べる。

なお、学習特徴ベクトル列の各要素はあらかじめ式(3.2.1)に従い正規化しておく。正規化された学習特徴ベクトルの一つ一つを$\boldsymbol{x}_n=\left(x_n(A_0),\ x_n(A_1),\dots,x_n(B_)\right)\ (n=1,2,\dots,N)$ のように書き、ベクトル列全体を$\mathbf{X}$ と書く。

$$x_n(l)=\frac{S_n(l)-\overline{S(l)}}{\sigma_{S(l)}} \tag{3.2.1}$$

ただし、$\overline{S(l)}$ は全特徴ベクトルの要素 l の平均、$\sigma_{S(l)}$ はその標準偏差である。

３．２．１　k-means++クラスタリング

混合ガウスモデルの作成には次項に述べる EM アルゴリズムを用いるが、EM アルゴリズムは収束するまでに必要とする計算回数や、1 度の計算量が多い。そのため、混合ガウスモデルの適切な初期値を見出すために計算量の少なさや単純さで優る k-means アルゴリズムを用いクラスタリングを行い、その後に EM アルゴリズムを適用するという方法がよくとられる[6]。本研究では、Arthur らによる k-means++アルゴリズム[7]を用いて EM アルゴリズムに用いる初期値を求める。k-means++アルゴリズムは、オリジナルの k-means アルゴリズムと比べ初期値依存性の問題が低く、Katsavounidis ら[8]の提案した KKZ 法と比べ外れ値にも耐性があるという特徴を持つ[9]。

k-means++アルゴリズムの k-means アルゴリズムから改良された点は初期のクラスタ中心の選び方である。k-means アルゴリズムでは初期のクラスタ中心を完全にランダムに選んでいたのに対し、k-means++アルゴリズムでは以下のようなステップで初期のクラスタ中心を選ぶ。ここで、クラスタ数はKと固定して考える。クラスタ数については 3.2.3 項で述べる。

(1a) $\mathbf{X}$ からランダムに中心点$\boldsymbol{\mu_1}$ を一つ選ぶ。

(1b) 全点に対し、$D(\boldsymbol{x})$ を計算する。ここで$D(\boldsymbol{x})$ は$\boldsymbol{x}$ からすでに定められた中心点の中で最も近いものとの距離であり、式(3.2.2)で表される。

$$D(\boldsymbol{x}_n) = \min_i |\boldsymbol{x}_n - \boldsymbol{\mu}_i| \tag{3.2.2}$$

この $D(\boldsymbol{x})$ に対し、すべての点において式(3.2.3)に従い確率$\Phi(\boldsymbol{x}_n)$を計算する。

$$\Phi(\boldsymbol{x}_n) = \frac{D(\boldsymbol{x}_n)^2}{\sum_{i=1}^{N} D(\boldsymbol{x}_i)^2} \tag{3.2.3}$$

そして次の中心点$\boldsymbol{\mu}_{i+1}$を、確率$\Phi(\boldsymbol{x}_n)$ に従い一つ選び出す。式(3.2.3)から、$D(\boldsymbol{x}_n)$ が大きい点、つまり最も近い中心への距離が大きい点が次の中心点$\boldsymbol{c}_{i+1}$ に選ばれる確率が高いことがわかる。また、式(3.2.3)の表現から、すでに中心点に選ばれている点が再び中心点に選ばれる確率は 0 であることがわかる。

(1c) 初期の中心点がK 個になるまで、(1b)を繰り返す。

以上(1a ~ c)までが k-means++アルゴリズムで改良された点である。以下からはオリジナルの k-means アルゴリズムと同様に進展する。そのアルゴリズムを参考文献[6]に従って説明する。

まず、k-means アルゴリズムでの目的関数として J を式(3.2.4)で定め、それを最小にする$\{r_{nk}\}$と$\{\boldsymbol{\mu}_k\}$を求めることを目標とする。J は歪み尺度と呼ばれ、各データ点から、それらが割り当てられたクラスタの中心$\boldsymbol{\mu}_k$ までの二乗距離の総和を表している。

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left||\boldsymbol{x}_n - \boldsymbol{\mu}_k\right||^2 \tag{3.2.4}$$

ただし、$r_{nk} \in \{0,1\}$はデータ点のクラスタへの割り当てを表す記号である。具体的には式(3.2.5)で表されるように、$\boldsymbol{x}_n$がk番目のクラスタに割り当てられるとき、つまり$\boldsymbol{x}_n$からk番目のクラスタの中心点$\boldsymbol{\mu}_k$への距離が、ほかのどのクラスタの中心点への距離より近いとき$r_{nk} = 1$とし、そうでない場合には$r_{nk} = 0$とする。

$$r_{nk} = \begin{cases} 1 & k = \mathrm{argmin}_j \left\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\right\|^2 \text{のとき} \\ 0 & \text{それ以外} \end{cases} \tag{3.2.5}$$

(1a～c) で$\{\boldsymbol{\mu}_k\}$が求まり、式(3.2.5)で$\{r_{nk}\}$が求まったので、次に$\{r_{nk}\}$を固定した下での$\boldsymbol{\mu}_k$の最適化を考える。目的関数Jは$\boldsymbol{\mu}_k$の2次関数であり、$\boldsymbol{\mu}_k$で偏微分して0と置くことで最小化できる。

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 2\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{3.2.6}$$

これを解いて、新しい$\boldsymbol{\mu}_k$を得る。

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\boldsymbol{x}_n}{\sum_n r_{nk}} \tag{3.2.7}$$

式(3.2.7)の分母を見ると、それはクラスタkに属すデータ点の個数であることがわかる。その意味から、式(3.2.7)による$\boldsymbol{\mu}_k$の更新式は、k番目のクラスタに属すデータ点の平均を求めていることになる。

以上、式(3.2.7)によるクラスタ平均$\boldsymbol{\mu}_k$の更新と、式(3.2.5)で表されるそれに伴うデータ点の再割り当てを、再割り当てが起こらなくなるまで、あるいは既定の繰り返し回数（本研究では1,000回とした）を超えるまで繰り返す。その結果、すべての特徴ベクトル$\boldsymbol{x}_n$はK個のうちいずれかのクラスタに分類されることになる。

３．２．２　EMアルゴリズム

以上k-means++アルゴリズムによってクラスタリングされたデータ点から、次はEMアルゴリズムを用いて混合ガウスモデルを作成する。EMアルゴリズムは最尤法に基づいて確率モデルのパラメータを推定する手法の一つである。

第2章で説明したように、混合ガウス分布は式(3.2.8)のようにガウス分布の単純な線形重ね合わせで表される。

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.2.8}$$

ただし、d 次元ガウス分布の確率密度関数 $N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は式(3.2.9)で表される。

$$N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \tag{3.2.9}$$

EM アルゴリズムに用いる初期値として、$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ はそれぞれk 番目のクラスタにクラスタリングされたデータ点の平均、分散共分散行列を用いる。混合率 π_k は、式(3.2.10)で表されるように、全データ点のうちk 番目のクラスタにクラスタリングされたものの割合を用いた。

$$\pi_k = \frac{\sum_n r_{nk}}{N} \tag{3.2.10}$$

以降、EM アルゴリズムの流れについて、参考文献[6]に従って説明する。

$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ の初期値がそれぞれのk 番目のクラスタについて与えられたとき、データ点の集合$\mathbf{X}$ の尤度関数 $p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は式(3.2.11)で与えられる。

$$p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{3.2.11}$$

このままでは扱いにくいので、$p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ の対数をとり式(3.2.12)で表される対数尤度関数$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ を考える。

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{3.2.12}$$

ここで、以下の説明のために K 次元の 2 値確率変数 $\boldsymbol{z}$ を導入する。これは 1-of-K 表現を取るとする。つまり、$\boldsymbol{z}$ の成分のうちどれか一つだけが 1 で、他の成分はすべて 0 ということである。また$\boldsymbol{z}$ は式(3.2.13)を満たすものとする。つまり$\boldsymbol{z}$ の周辺分布は混合係数 π_k によって定まる。

$$p(z_k = 1) = \pi_k \tag{3.2.13}$$

この $\boldsymbol{z}$ を用いて、 $\boldsymbol{x}$ が与えられた下での $\boldsymbol{z}$ の条件付き確率 $\gamma(z_k)$を式(3.2.14)で定める。 $\gamma(z_k)$ は負担率と呼ばれ、k 番目のクラスタが $\boldsymbol{x}$ の観測を説明する度合いと考えることができる。なお、1 行目の式変形にはベイズの定理を用いている。

$$\gamma(z_k) = p(z_k = 1|\boldsymbol{x}) = \frac{p(z_k = 1)p(\boldsymbol{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\boldsymbol{x}|z_j = 1)} = \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{3.2.14}$$

さて、EM アルゴリズムの目的は式(3.2.12)で表される対数尤度を最大とするような$\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$を求めることである。そのために、式(3.2.12)を$\boldsymbol{\mu}_k$に関して偏微分して 0 とおくと次式が得られる。

$$\frac{\partial \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial\, \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j N(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{3.2.16}$$

$\boldsymbol{\Sigma}_k$をかけて整理すると、次式を得る。

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n \tag{3.2.17}$$

ただし、N_kは

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{3.2.18}$$

とおいた。N_kはk番目のクラスタに割り当てられる点の実効的な数と考えることができる。式(3.2.17)は$\boldsymbol{\pi}, \boldsymbol{\Sigma}$を固定した下での$\boldsymbol{\mu}$の最大化の条件である。式(3.2.17)を見ると、k番目のガウス要素の平均$\boldsymbol{\mu}_k$は、データ集合の各点の重み付き平均で得られることがわかる。

同様に、式(3.2.12)で表される対数尤度を$\boldsymbol{\Sigma}_k$に関して偏微分して 0 とおき整理すると次式が得られる。

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \tag{3.2.19}$$

最後に、式(3.2.12)で表される対数尤度を混合係数π_kについて最大化する。π_kについては式(3.2.20)で表される制約条件を満たさなくてはならない。

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3.2.20}$$

そのため、$\boldsymbol{\mu}, \boldsymbol{\Sigma}$ に関する最大化で用いたような、単純に偏微分して 0 とおく手法はとれない。そこで、ラグランジュ未定係数法を用いる。つまり、次の量を π_k で微分して 0 とおく。

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right) \tag{3.2.21}$$

すると次式が得られる。

$$\sum_{n=1}^{N} \frac{N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j N(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda = 0 \tag{3.2.22}$$

式(3.2.22)に対して、両辺に π_k を掛け k について和を取り、制約条件(3.2.20)を用いると $\lambda = -N$ を得る。これを用いて λ を消去して変形すると、次の π_k に関する最大化条件を得る。

$$\pi_k^{new} = \frac{N_k}{N} \tag{3.2.23}$$

すなわち、k 番目のガウス要素に関する混合係数 π_k は、全データ点の数のうち k 番目のガウス要素に実効的に属しているデータ点の数の割合と考えることができる。

以上をまとめると、本研究で用いる EM アルゴリズムは次のように書ける。

(1) 初期化

k-means++アルゴリズムを用いて、$\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ の適当な初期値を求める。

(2) E ステップ

現在の $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ を用いて、式(3.2.14)で表される負担率 $\gamma(z_{nk})$ をそれぞれの観測点において計算する。

(3) M ステップ

式(3.2.17), (3.2.19), (3.2.23)を用いて、$\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ を再計算する。再計算された $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ はいずれも対数尤度を大きくすることが保障されている。

(4) 収束判定

次式が成立すれば収束したとして EM アルゴリズムを終了させる。

$$\left(\text{更新後の対数尤度}\right) - \left(\text{更新前の対数尤度}\right) < \delta \times \left(\text{更新前の対数尤度}\right) \tag{3.2.24}$$

ただし、δ は収束判定の厳しさを表す値であり、本研究では $\delta = 0.000001$ に固定した。この収束判定条件を満足するまで(2)E ステップと(3)M ステップを繰り返す。

以上が EM アルゴリズムの流れである。なお、本研究ではプログラム実行のために以下のようなルールを設けた。

・クラスタ数 K は以下の式(3.2.25)を満たさなくてはならない。

$$K < \ln N \tag{3.2.25}$$

・$|\boldsymbol{\Sigma}_k| = 0$ となった場合にはそのクラスタは棄却し、$|\boldsymbol{\Sigma}_k|$ を式(3.2.26)のように十分大きく再配置し、$\boldsymbol{\mu}_k$ は学習特徴ベクトルの中からランダムに一つ選び新たなクラスタを作り EM アルゴリズムを続ける。

$$\boldsymbol{\Sigma}_{k\,ij} = \begin{cases} 1 & i = j \\ 0.1 & i \neq j \end{cases} \tag{3.2.26}$$

・収束しないまま規定最大繰り返し回数 10,000 回に達した場合、そのクラスタ数は無効とし、クラスタ数を一つ少なくして再び初めから k-means++アルゴリズムを始める。これは上の分散共分散行列の配置に伴い、収束しなくなることを避けるためである。

・混合率 π_k は以下の条件式(3.2.27)を満たすものとする。EM アルゴリズムの中でそれを満たさない分布が一つでも現れた場合はその分布は棄却し、クラスタ数を一つ減らして k-means++アルゴリズムからやり直す。これは混合係数が小さいものが存在することによって、クラスタ数が不必要に多くなることを防ぐためである。$\pi_{\min}$ はその厳しさを決める定数である。

$$\min_k \pi_k > \pi_{min} \tag{3.2.27}$$

３．２．３　クラスタ数の決定

以上までの、k-means++アルゴリズムや EM アルゴリズムはクラスタ数 K を所与としてきた。さまざまなクラスタ数でできた混合ガウスモデルの中から、最適なクラスタ数のものを選択する必要がある。今回のように最尤法を用いて推定されたモデルのよさを評価する指標として式(3.2.28)で表される赤池情報量規準(AIC：Akaike's Information Criterion)や、式(3.2.29)で表されるベイズ情報量規準(BIC：Bayesian Information Criterion)などがある。ただし、L は尤度関数、k は自由パラメータの数、N は観測データの数である。

$$\mathrm{AIC} = -2\ln L + 2k \tag{3.2.28}$$

$$\mathrm{BIC} = -2\ln L + k \cdot \ln N \tag{3.2.29}$$

これらの式で与えられる AIC や BIC が最小のモデルが良いモデルとされる。実際、AIC や BIC の右辺第 1 項は尤度に関する項で、負号がついていることから尤度が高くなればそれらの値は小さくなることがわかる。

また、右辺第 2 項は自由パラメータに関する項である。自由パラメータ数を増やせば（混合ガウスモデルの場合それはクラスタ数 K を増やすことと同義である）一般に観測点に対するモデルの当てはまりは良くなり、

対数尤度も高くなる。しかし自由パラメータ数を必要以上に増やすことはノイズにもモデルをフィットさせてしまう、つまりオーバーフィッティング（過剰適合）になってしまう問題がある。そこで増えすぎた自由パラメータ数に対してペナルティを課す必要があるが、情報量規準ではそれが右辺第 2 項に表れている。実際、自由パラメータ数 k が増えると右辺第 2 項は大きくなり、情報量規準は大きくなる（つまりモデルを悪いと評価する方向になる）ことがわかる。右辺第 2 項はこのためバイアス補正項と呼ばれる。また、AIC と BIC でのバイアス補正項を比較すると、$\ln N > 2$ の時、つまり $N \geq 8$ の時は BIC の方がバイアス補正項は大きくなり、自由パラメータ数の増加へのペナルティを大きく課していることがわかる。

AIC と BIC の両者を比較検証している研究には、金田らの研究[10]がある。金田らはクラスタ数の推定において両情報量規準を比較し、BIC の方が良好なパフォーマンスであったとしている。また、水戸のシミュレーション[11]では AIC はクラスタ数を多めに推定する場合があることが示されている。

以上を踏まえ、本研究ではクラスタ数の推定に BIC を用いる。一つの d 次元混合ガウス分布の自由パラメータは、平均 $\boldsymbol{\mu}$ に対して d 個、分散共分散行列 $\boldsymbol{\Sigma}$ に対しては対称行列であることに注意して $d(d+1)/2$ 個である。また、混合ガウス分布を構成するそれぞれのガウス分布に対して混合率 π_k が一つずつ存在するので、クラスタ数が K の混合ガウスモデルの BIC は式(3.2.29)より次のようになる。

$$\mathrm{BIC} = -2\,ln\,L + K\left\{d + \frac{d(d+1)}{2} + 1\right\} \cdot ln\,N \tag{3.2.30}$$

ただし対数尤度 L は式(3.2.12)で書ける。

本研究では式(3.2.25)で表されるクラスタ数 K の条件を満たすもので、その最大値から $K = 1$ までにおいてそれぞれ EM アルゴリズムにより混合ガウスモデルを作成する。作成したモデルに対し式(3.2.30)を適用し BIC をそれぞれ計算し、BIC を最小とするモデルを採用する。ただし EM アルゴリズムの中で、式(3.2.27)で表される最小混合率の条件を満たさないものが一つでも現れた場合は BIC を計算せずに、そのクラスタ数 K は棄却する。

３．３　不適合度

以上のような方法で学習期間の特徴ベクトル列に対して混合ガウスモデルを作成し、できたモデルと入力期間の特徴ベクトル列を比較する。本研究ではその指標として以下で定義する不適合度を用いる。

$$\text{不適合度} = \frac{\text{モデルに不適合とされた入力ベクトルの数}}{\text{全入力ベクトルの数}} \tag{3.3.1}$$

不適合度が高い場合、学習期間で作成した混合ガウスモデルへの入力特徴ベクトルの当てはまりが悪い、すなわち学習期間と入力期間の注文状況が異なっていることを意味する。市場で特異な取引が多く発生しているとすれば、学習期間で作成した混合ガウスモデルではあまり説明されないような入力特徴ベクトルが増え、不適合度は高くなると予想される。

ただし、できた混合ガウスモデルにある入力特徴ベクトルが不適合とされるか否かの基準については、以下の二つの判定式を使った 2 つの場合をそれぞれ実験した。ある入力特徴ベクトルがそれぞれにおける判定式を満たした場合に、そのベクトルを不適合と判断する。

（１）尤度による基準

$$p(\boldsymbol{x}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) < \theta_L \tag{3.3.2}$$

左辺は式(3.2.8)で表される尤度（全クラスタの尤度の重み付き総和）である。これがある閾値 θ_L より小さいとその入力特徴ベクトルは不適合とみなす。

（２）マハラノビス距離による基準

$$\min_k MD_k(\boldsymbol{x}) > \theta_\sigma \tag{3.3.3}$$

ただし、$MD_k(\boldsymbol{x})$ はマハラノビス距離と呼ばれる一種の距離で、以下の式(3.3.4)で表される。

$$MD_k(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)} \tag{3.3.4}$$

これは k 番目のクラスタ中心 $\boldsymbol{\mu}_k$ と $\boldsymbol{x}$ の距離を、分散共分散行列により正規化したものと解釈される。式(3.3.3)を満たすということは、どのクラスタ中心 $\boldsymbol{\mu}_k$ からもマハラノビス距離において閾値 θ_σ より遠いということで、不適合と判断される。

４．実験

本研究では、混合ガウスモデルを用いた板の異常性を効率的に把握する手法を確認するため、以下の表 4.1 に示した 4 銘柄を分析の対象とする。分析対象の銘柄AからDはいずれも、2010 年の増資に際して内部者取引が報告された銘柄である。分析対象とする板情報については、東京証券取引所の提供する FLEX Standard[12]を使用した。FLEX Standard では東京証券取引所に上場する現物株式及び転換社債の板情報などの情報を、2010 年 1 月 4 日から取得できる（なおサービス開始日である 2011 年 1 月 11 日より前のデータは参考情報として取得可能だが一部データの欠損があるとしている)。データは 1 日 1 ファイルの中に全銘柄についての情報が記載され、FLEX Standard の場合板情報は売り買いそれぞれ第 8 気配値まで分類され、それより高い、あるいはより低い価格での注文は Over, Under としてまとめられている。板情報が記録されている時間帯は営業日の 8:00 ~ 11:30, 12:05 ~ 15:00 である。

表 4.1　観察対象とする銘柄

銘柄名	業種	増資発表日
銘柄 A	a	2010/6/25
銘柄 B	b	2010/7/8
銘柄 C	c	2010/8/24
銘柄 D	d	2010/9/29

４．１　実験１：業種内比較

４．１．１　分析銘柄と期間

実験 1 では、表 4.1 で示した 4 銘柄と、それぞれの同業種銘柄とで同時期の不適合度を比較した。比較対象銘柄は日経株価指数 300 採用銘柄を用いた[13]。ただし銘柄 B（業種 b）については自身以外に日経株価指数 300 採用銘柄がないので、日経 500 種平均株価採用銘柄 [14]で代用した。

一般に同じ業種の銘柄の株価は、個別企業のニュースを除くと、市場の動向やマクロ変数に対して同じような動きをすると考えられる。例えば円安というマクロ変数の変化に対しては、輸出企業である自動車メーカーなどにはプラスの材料として株価の上昇をもたらすことが多いが、輸入企業である電力会社などはマイナスの材料として株価の下落をもたらすことが多い。つまり、あるマクロ的なニュースによって特定の業種の市場状態が変化したとき、その期間においてその業種内では一様に不適合度が上下することが考えられるため、比較対象として同業種銘柄を採用することは有効だと考えられる。なお、分析期間については、すべて増資公表が公表日の大引け後であったことから、入力期間を発表当日から 10 営業日前まで、学習期間をさらにそこからさかのぼり 100 営業日前までとした。（いずれの銘柄も 2010 年 5 月 6 日はデータ欠損のため除いている。）

実験 1 で使用したパラメータを以下に示す。

・一つの特徴ベクトルを作る期間 $T = 30$ (分)
・最小混合率 $\pi_{min} = 0.1$
・閾値による尤度 $\theta_L = 0.0001, 0.00001, 0.000001$
・マハラノビス距離による閾値 $\theta_\sigma = 3, 4, 5$

なお、ほかの期間と整合性を保つため 12:05 ~ 12:30 の時間帯に入った注文量は $\frac{30}{25}$ 倍した。

４．１．２　結果

各パラメータにおいて 10 回実験を行い、それぞれの不適合率の平均値を求めた（図 4.1~4.6）。なお、業種平均の上のバーは同一業種内のそれぞれの閾値における不適合度の 0.5 標準偏差を表す。この意味については 4.2 節で詳しく述べる。

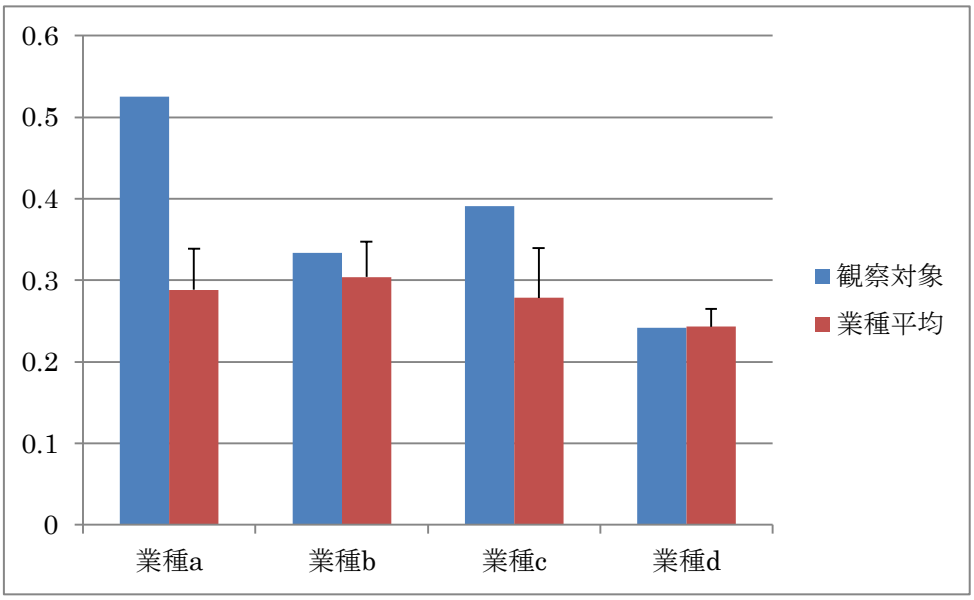

図 4.1　尤度により閾値を定めた場合の結果（$\theta_L = 0.0001$）

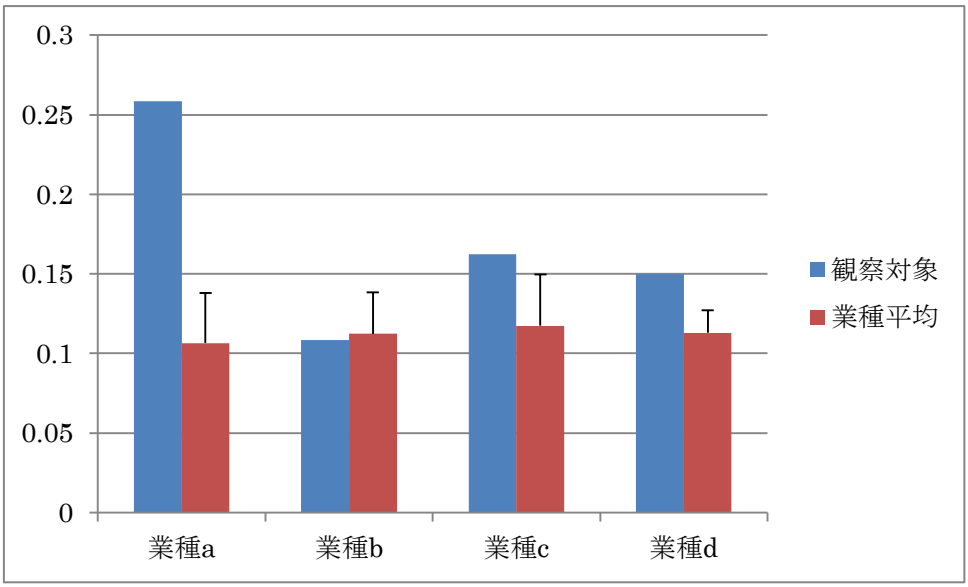

図 4.2　尤度により閾値を定めた場合の結果（$\theta_{\mathrm{L}} = 0.00001$）

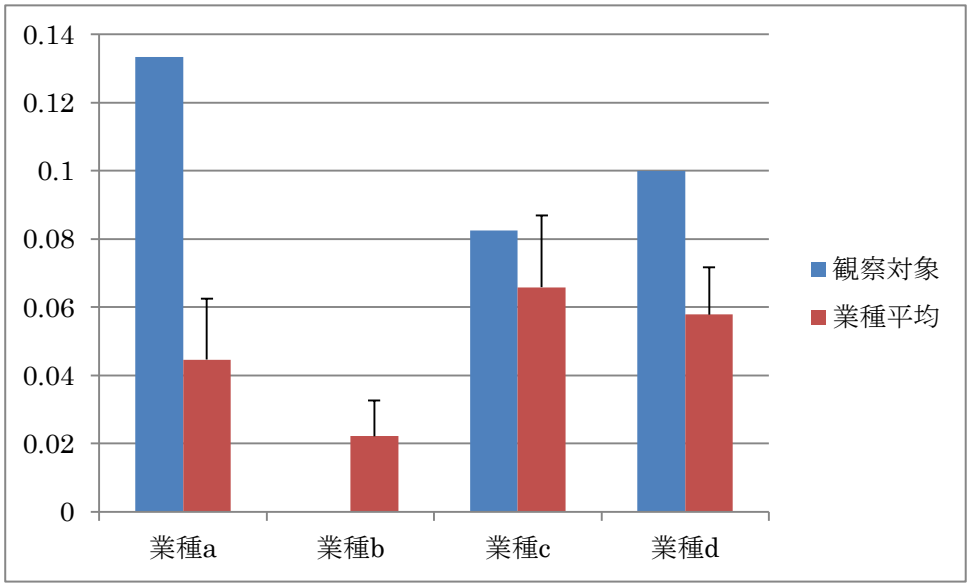

図 4.3　尤度により閾値を定めた場合の結果（$\theta_{\mathrm{L}} = 0.000001$）

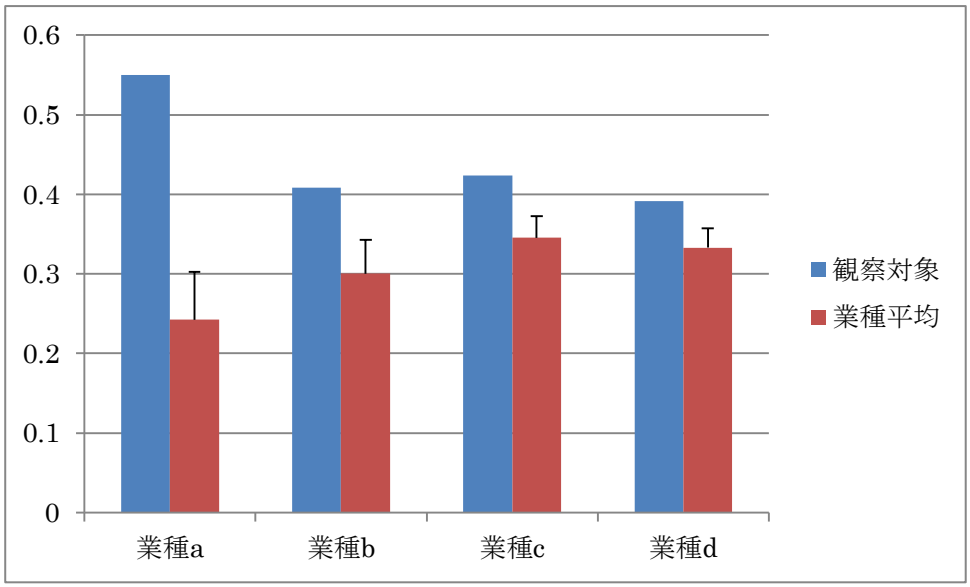

図 4.4　マハラノビス距離により閾値を定めた場合の結果（$\theta_\sigma = 3$）

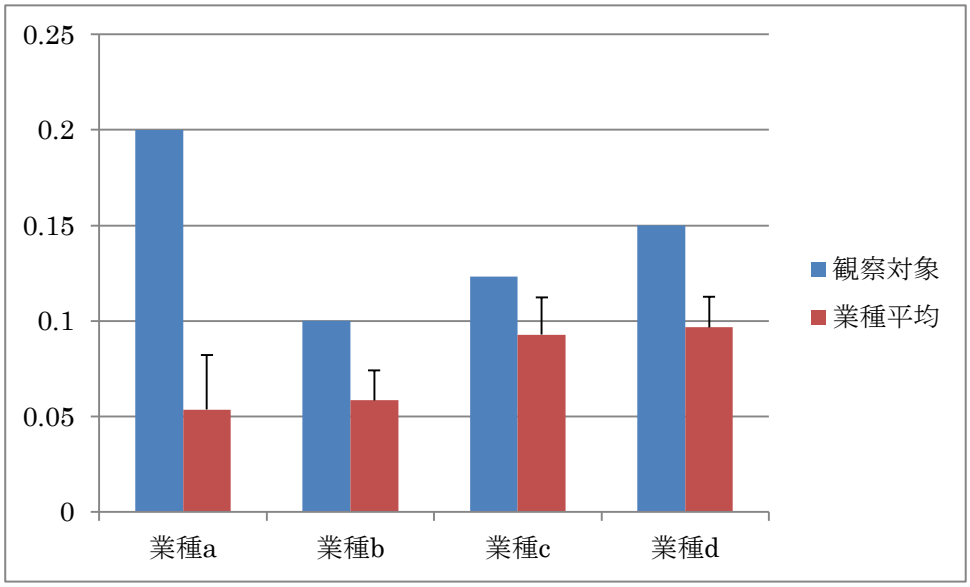

図 4.5　マハラノビス距離により閾値を定めた場合の結果（$\theta_\sigma = 4$）

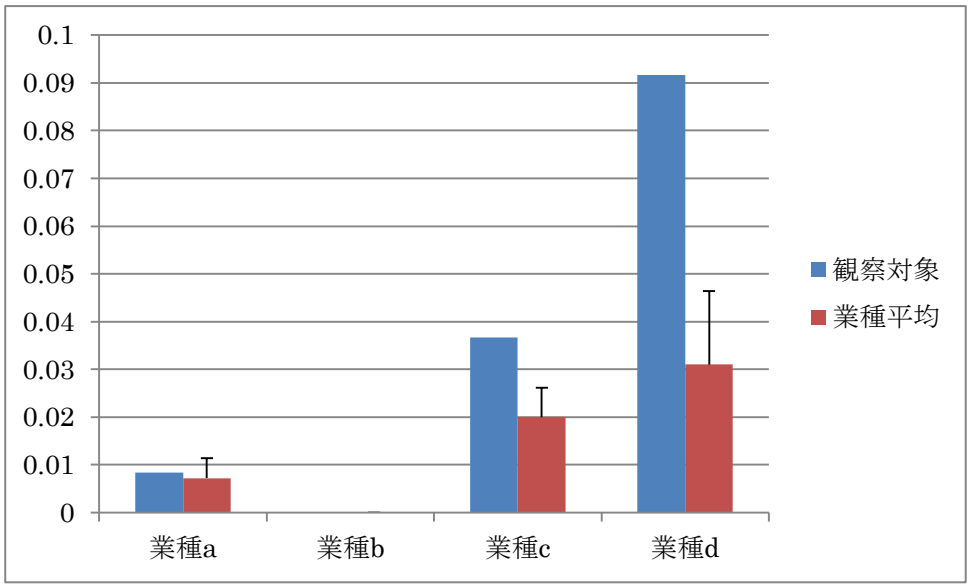

図 4.6　マハラノビス距離により閾値を定めた場合の結果（$\theta_\sigma = 5$）

４．１．３　考察

4.1.2 項に示した実験 1 の結果から、尤度によって閾値を設定したときよりもマハラノビス距離によって閾値を設定した方が全体的に観察対象銘柄の不適合率がほかの銘柄と比べ高く、良い結果を得られた。実際、業種 b や業種 d では尤度による不適合度で比べた場合はほかの銘柄と比べさほど高い不適合度は示していないが、マハラノビス距離による不適合度ではほかの銘柄と比べ高い不適合度を示している。

ある特徴ベクトルを不適合と判定する閾値を尤度により定めるのか、マハラノビス距離により定めるかの本質的な違いは、それぞれのクラスタの混合率 π を考慮に入れるか否かである。尤度により定める場合は考慮に入れ、マハラノビス距離で定める場合はすべてのクラスタに対して同等に扱っているので考慮に入れていない。

また、マハラノビス距離による閾値に関しては $\theta_\sigma = 4$ の時が、観察対象銘柄の不適合率が同業種他銘柄より高くなり、その差の不適合率に対する割合も大きくなった。ただし業種 c では $\theta_\sigma = 3$ の場合に不適合度が一番高くなったが、$\theta_\sigma = 4$ の場合、その不適合度は全体の 2 番目となった。$\theta_\sigma = 5$ まで閾値を厳しくすると、業種 a、業種 b、では不適合率が小さくなりすぎて検出には困難であった。ただし業種 d に関しては他の同業種銘柄と比べ高い不適合率を得ることができた。以上の事より今回の実験では $\theta_\sigma = 4$ がどの業種にも良いパフォーマンスを得る最良の閾値であると判断したが、この閾値を定量的にどう決定するかは今後の課題である。

クラスタ数に関する問題として、同じパラメータセットでも実験ごとにまれにクラスタ数が変化することがあった。これは k-means++アルゴリズムにおける初期値を選ぶ際に式(3.2.3)で表される確率を用いていることと、EM アルゴリズムにおいて局所最適解に陥ってしまっていることが原因であると考えられる。対数尤度

の多峰性や EM アルゴリズムがその中で必ずしも最大のものに収束しないことは[6]でも述べられている。表 4.2 にクラスタ数が異なった場合の不適合度の例を示す。なお不適合度は有効数字 2 桁になるよう四捨五入している。

表 4.2　クラスタ数と不適合度の関係

	$\theta_L = 0.0001$	$\theta_L = 0.00001$	$\theta_L = 0.000001$	$\theta_\sigma = 3$	$\theta_\sigma = 4$	$\theta_\sigma = 5$
クラスタ数：3	0.21	0.050	0.0083	0.20	0.0083	0
クラスタ数：4	0.22	0.042	0.0083	0.18	0.025	0

この表からも分かるように、クラスタ数が多く/少なくなると不適合度はどうなるかと言う関係は一般には言えなかった。なお 10 回の繰り返し中に 1 回でも他と違うクラスタ数の混合ガウスモデルができた銘柄は 24 銘柄中 6 銘柄であった。本研究では繰り返し中にクラスタ数が異なるものが出てきてしまうという問題に、不適合度は 10 回の平均を取ることで対処するが、k-means++アルゴリズムにおける初期値依存性と EM アルゴリズムにおける対数尤度の多峰性の問題は今後の課題である。なおクラスタ数が同じ場合の不適合度はほとんどが全く同じで、まれにごくわずかに異なることがあったが、これも 10 回の平均を取ることで無視できるものとなる。

４．２　実験２：最適な異常判定閾値算出のための実験

４．２．１　実験の概要

本節では、4.1 節で示した結果について、不適合度がどれだけ高ければその銘柄の入力期間を異常と判定するとすればよいかという閾値を、定量的に求めるための実験を行った。以下不適合度と言った場合は、マハラノビス距離による閾値で、$\theta_\sigma = 4$ と設定した場合の不適合度を指すこととする。最適化の目的変数 α は次式で表されるものとする。

$$\theta_{inc} = \bar{\theta} + \alpha\sigma \tag{4.2.1}$$

ただし $\bar{\theta}$ は、同一業種内の銘柄における不適合度の平均、σ はその標準偏差、θ_{inc} はある銘柄において、次式を満たした時に市場状態が変化しているというサインを出す不適合度の閾値である。

$$\text{不適合度} > \theta_{inc} \tag{4.2.2}$$

この実験に用いる指標は精度と再現率、そしてその二つから計算される F 値である。精度と再現率、F 値はそれぞれ以下の式で表される。

$$\text{精度} = \frac{\text{異常と判定され実際に異常であったものの数}}{\text{異常と判定されたものの数}} \tag{4.2.3}$$

$$\text{再現率} = \frac{\text{異常と判定され実際に異常であったものの数}}{\text{異常であったものの数}} \tag{4.2.4}$$

$$\text{F 値} = \frac{2 \cdot \text{精度} \cdot \text{再現率}}{\text{精度} + \text{再現率}} \tag{4.2.5}$$

ある判定システムの良さを表す指標として誤判定率の少なさを表す精度と、取りこぼしの少なさを表す再現率が考えられるが、F 値はその両者を考慮した指標と言え、その値が高い方が良いシステムと判断される。本実験では、4.1 節の実験 1 で得られた結果のうち、$\theta_\sigma = 4$ のものに対して、異常と判断する不適合度の閾値を式(4.2.1)における α を変化させることによって複数の値を試す。その結果に対して、式(4.2.3~4.2.5)で示した精度、再現率、F 値を求め、最適な α を求める。

４．２．２　結果

以下の図 4.7, 4.8 にその結果を示す。

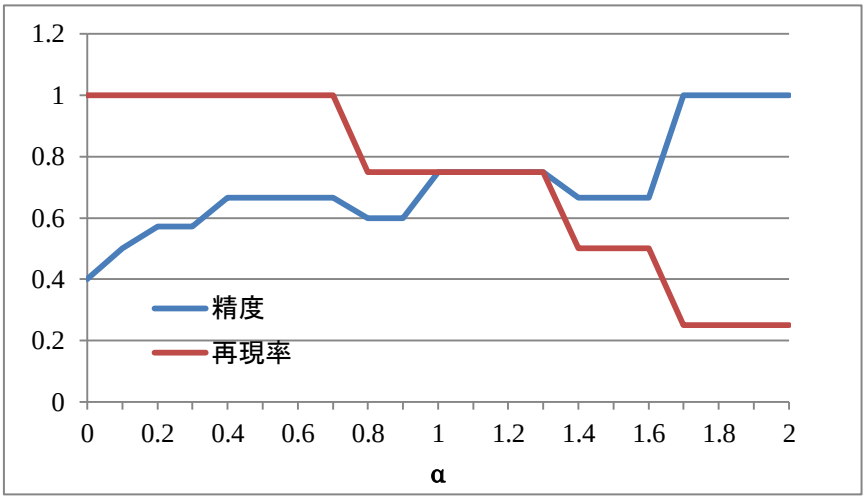

図 4.7　精度と再現率

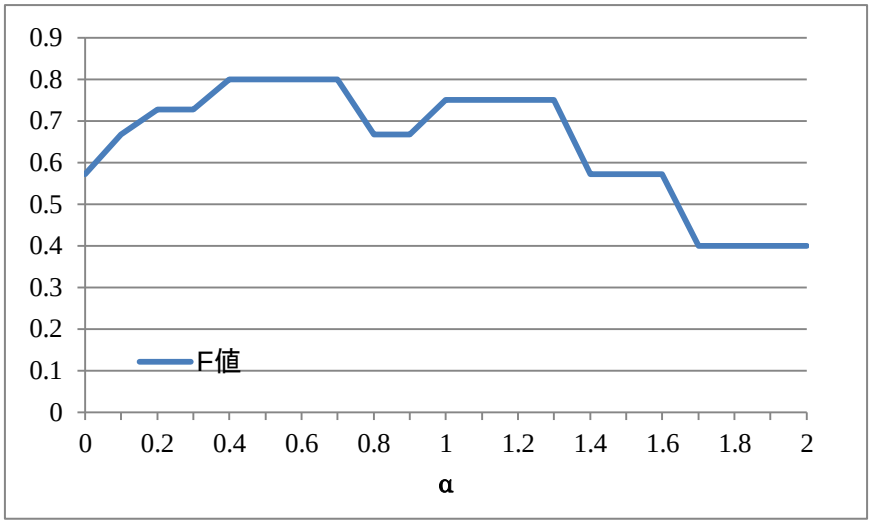

図 4.8　F 値

４．２．３　考察

以上の結果より、$0.4 \leq \alpha \leq 0.7$ で F 値は最大値を取り、その時システムは最も良いと言える。なおその際には精度＝0.67, 再現率=1 となっていた。たとえば、$\alpha = 0.5$ とすると、内部者取引が報告されている 4 銘柄の入力期間の注文状況は、すべて異常と判定される。　ただし今回の実験では対象にした観察対象の母数が 4 銘柄しかないことから、本実験で求めた精度、再現率、F 値も粗いものでしかないことは問題点である。また、業種 b などは特に、同業種銘柄が 3 銘柄しかなく、その中で標準偏差を取っている。これは標準偏差を取るには少ない値で、これも本実験における問題点である。

なお、増資をしていない比較対象の銘柄の中では、業種 a に 1 銘柄、業種 c に 1 銘柄が$\alpha = 0.5$ としたときに特異と判断された。業種 c の銘柄については、入力期間の前日である 2010 年 8 月 10 日の大引け後に、最終赤字が前年より拡大をしたと発表され、それにより市場の注文状況が変化し不適合度が高まったと考えられる。本手法ではそのようなニュースによる注文状況の変化も検出できたと言える。業種 a の銘柄については、入力期間において市場の注文状況を変化させるような大きなニュースは特に見つからなかった。

４．３　実験３：通常の増資銘柄との比較

次に、内部者取引が報告されていない通常の増資銘柄（表 4.3）に関しても同様の分析を行い、4.1 節の実験 1 の結果と比較を行った。これにより、本手法が増資のみによる注文状況の変化を抽出しているのではないことを検証した。

４．３．１　分析銘柄と期間

以下の表 4.3 に実験 3 の分析銘柄を示す。

表 4.3　通常の増資銘柄

銘柄名	業種	増資公表日
銘柄 E	e	2010/9/21
銘柄 F	a	2010/11/5
銘柄 G	g	2011/2/23
銘柄 H	c	2011/8/30

それぞれの同業種の比較対象銘柄は、銘柄 E に関しては日経 500 種平均株価採用銘柄、その他 3 銘柄は日経株価指数 300 採用銘柄を用いた。実験期間は実験 1 と同様に、入力期間を発表当日から 10 営業日前まで、学習期間をさらにそこからさかのぼり 100 営業日前までとした。

４．３．２　結果

ある特徴ベクトルを不適合と判定するための閾値は、実験 1 で良い結果を得たマハラノビス距離による基準を用い、その値は $\theta_\sigma = 4$ とした。その他のパラメータは実験 1 と同様である。以下にその実験結果を示す。

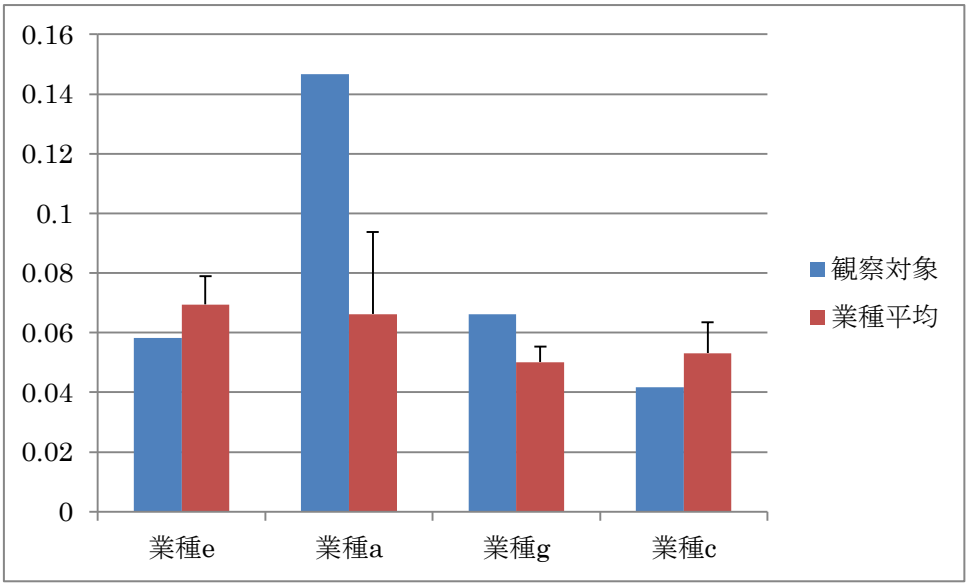

図 4.9　通常増資銘柄の結果

４．３．３　考察

以上の結果に対して、実験 2 の結果を踏まえて$\theta = \bar{\theta} + 0.5\sigma$ という閾値を用いて判定すると、通常の増資銘柄 4 つのうち銘柄 F と銘柄 G が特異とみなされた。以上二つの実験結果をまとめると以下の表のようになる。これにより、分析対象銘柄の方がそうでない通常の増資銘柄と比べ、増資公表 10 営業日前までの注文状況の特異性が高かったということが本手法による分析結果から示唆される。

表 4.4 増資公表前の注文状況が特異/非特異と判定された銘柄の数

	特異	非特異
実験 1 の結果	4	0
通常の増資銘柄	2	2

４．４　実験４：期間別比較

実験 1 では観察対象 4 銘柄について、マハラノビス距離を用いて $\theta_\sigma = 4$ という閾値を設定した場合に、同業他銘柄と比較して高い不適合度を示すことを確認した。実験 4 では同一銘柄について、内部者取引が報告された 2010 年とそれ例外の 2011 年、2012 年で提案手法を試し比較する。このことによって、特異な取引のあった 2010 年のみで高い不適合度を示すことを確かめ、提案手法の有効性を確認する。

４．４．１　分析銘柄と期間

分析銘柄とそれぞれの期間を以下表 4.5 に示す。いずれも実験 1 と同様に学習期間は 100 営業日、入力期間は 10 営業日である。

表 4.5　期間別比較　分析銘柄と期間

		銘柄 A	銘柄 B	銘柄 C	銘柄 D
2010 年	学習	2010/1/14 ~ 2010/6/11	2010/1/27 ~ 2010/6/24	2010/3/15 ~ 2010/8/10	2010/4/19 ~ 2010/9/13
	入力	2010/6/14 ~ 2010/6/25	2010/6/25 ~ 2010/7/8	2010/8/11 ~ 2010/8/24	2010/9/14 ~ 2010/9/29
2011 年	学習	2011/1/13 ~ 2011/6/10	2011/1/27 ~ 2011/6/24	2011/3/15 ~ 2011/8/10	2011/4/20 ~ 2011/9/13
	入力	2011/6/13 ~ 2011/6/24	2011/6/27 ~ 2011/7/8	2011/8/11 ~ 2011/8/24	2011/9/14 ~ 2011/9/29
2012 年	学習	2012/1/17 ~ 2012/6/11	2012/1/23 ~ 2012/6/15	2012/1/23 ~ 2012/6/15	2012/1/20 ~ 2012/6/15
	入力	2012/6/12 ~ 2012/6/25	2012/6/18 ~ 2012/6/29	2012/6/18 ~ 2012/6/29	2012/6/18 ~ 2012/6/29

※銘柄 B、銘柄 C、銘柄 D の 2012 年のデータについては、データ数の関係からほかの年と時期がずれている。また、以下の銘柄の日付のデータは欠損や破壊のため除いている。

・銘柄 A : 2010 年 5 月 6 日

・銘柄 B : 2010 年 5 月 6 日

・銘柄 C : 2010 年 5 月 6 日

・銘柄 D : 2010 年 5 月 6 日, 2012 年 2 月 2 日

４．４．２　結果

表 4.5 に示した銘柄･期間について、実験 1 と同様のパラメータを用いて、マハラノビス距離よって $\theta_\sigma = 4$ と閾値を定めた場合の不適合度を求めた。以下図 4.10 ~ 4.13 にその実験結果を示す。なお、不適合度は、10 回の実験を行いその平均を用いている。各銘柄とも、増資に際し内部者取引が行われた 2010 年を四角枠で囲っている。

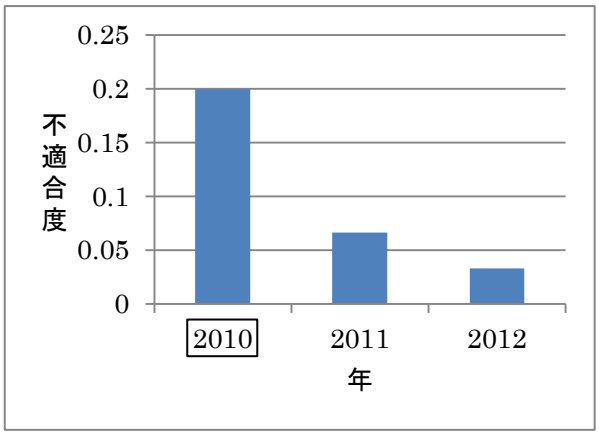

図 4.10　期間別比較　銘柄 A

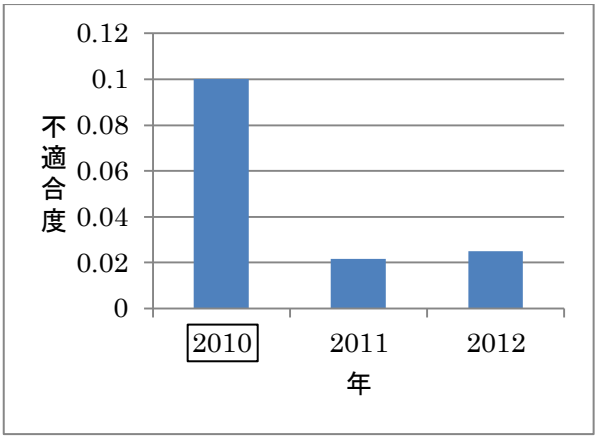

図 4.11　期間別比較　銘柄 B

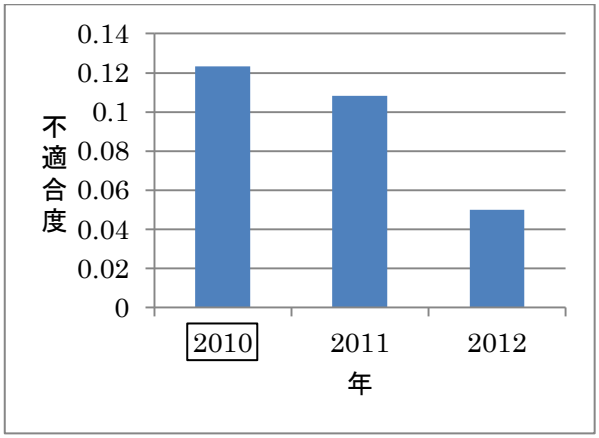

図 4.12　期間別比較　銘柄 C

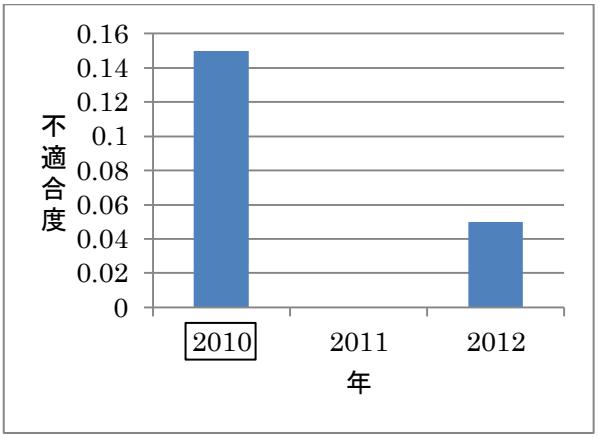

図 4.13　期間別比較　銘柄 D

４．４．３　考察

実験 4 の結果から、いずれの銘柄においても増資に係る内部者取引が報告された 2010 年に、ほかの年と比べ最も高い不適合度を示すことが分かった。このことから、実験 1 では対象銘柄の不適合度が同業種他銘柄の不適合度より高くなったが、それはその銘柄自体の特性ではなく、2010 年が特異な状態にあったといえる。

ところで、銘柄 D の 2011 年の不適合度は 0 となり、3 年のうち最小となった。これは 2011 年 3 月に起きた東日本大震災の影響で銘柄 D における注文状況が大きく変化し、学習期間のそのような特異性が大きなガウス分布を作ったためだと考えられる。このように学習期間において市場状態の変化があった際には、同じ入力ベクトルでも結果が変わってしまいうることは本手法の欠点と言える。

５．結論

本研究の目的は、株式の板情報を分析し、注文状況が通常とは異なる状態にあるかどうかの分析を支援する手法を開発することであった。本研究では、そのために梅岡らの提唱した手法を基として、板情報から抽出される特徴ベクトルに対し混合ガウスモデルを適応し、ある入力期間における不適合度という指標を求めた。

実験 1 では、分析対象銘柄と同業種他銘柄を同一期間において比較し、分析対象銘柄の不適合度がその他の銘柄よりも高くなったことを確認した。不適合度を求めるための閾値としては、尤度によるものとマハラノビス距離によるものを実験したところ、尤度よりもマハラノビス距離により閾値を定めた方が適合性が高くなった。本研究においては、これら閾値の値を経験的に求めたが、将来的には定量的に定める方法を検討する必要があり、今後の課題である。

実験 2 では、実験 1 で得られた結果に対し、不適合度がある値を超えた際にその対象を異常な状態と判定するといった不適合度の閾値を、F 値という指標を用いて定量的に求めた。その結果「不適合度が比較対象を含めた平均より、0.5 標準偏差分高ければ異常と判定する」という閾値を提案した。このように閾値を経験的ではなく定量的に示すことは、特異な取引を検出するシステムを構築する際には重要である。

実験 3 では、通常の増資銘柄に関しても実験 1 と同様に本手法を試し、不適合度の高まりが分析対象銘柄に顕著な特徴であるかを確認した。その結果、通常の増資銘柄と比べても、増資公表前 10 営業日の注文状況の不適合度が高かったことを確認した。

実験 4 では、2010 年に確認された高い不適合度が、その銘柄固有の特性によるものではないことを示すため、同一銘柄において期間別の比較（2010 年、2011 年、2012 年）を実施した。その結果、内部者取引が報告された 2010 年のみに高い不適合性が確認され、銘柄固有の特性によるものでないことを示した。

本研究の応用としては、特異な取引の自動検出システムの構築がある。そのようなシステムの構築のためには、より膨大なデータを検証することが必要であり、また上に述べた以外にも今回は固定した学習/入力期間や特徴ベクトルを抽出する時間間隔などといったパラメータを最適化する必要がある。また、今回はある一定期間内になされた注文株数の、ラベルごとの総和と言った非常に単純な量を基に特徴ベクトルを作ったが、板情報からは直近取引価格や注文間隔、注文のキャンセルなど多様な情報を得ることができる。これらを用いることでさらに高度な特徴ベクトルを作ることが可能となり、より精度の高い分析を行うことができるだろう。これらについては、今後の課題としたい。

参考文献

[1] 西岡寛兼, 鳥海不二夫, 石井健一郎, 「板情報を用いた市場変化の分析」, 『人工知能学会研究会資料 SIG-FIN-003』, pp.58-63 (2009)

[2] 西岡寛兼, 鳥海不二夫, 石井健一郎, 「板情報を用いた金融市場の相転移検出」, 『人工知能学会全国大会論文集』(2010)

[3] 梅岡利光, 鳥海不二夫, 平山高嗣, 榎堀優, 石井健一郎, 間瀬健二, 「板情報を用いた株式市場の状態変化の分析」, 第 37 回 JAFEE 大会(2012)

[4] 村井泰裕, 藤吉弘亘, 数井誠人, 「時空間特徴に基づくエスカレータシーンにおける人の異常行動検知」, 『情報処理学会研究報告 2008-CVIM-164』, pp.251-258 (2008)

[5] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, vol.10, pp.19-41 (2000)

[6] C・M・ビショップ著（元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇訳）, 『パターン認識と機械学習 - ベイズ理論による統計的予測　下』, シュプリンガー・ジャパン, 第 9 章(2008)

[7] David Arthur, Sergei Vassilvitskii, "k-means++: The Advantages of Careful Seeding", Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithm, 1027-1035 (2007)

[8] I. Katsavounidis, C. C. J. Kuo, Z. Zhang, "A new initialization technique for generalized Llyoyd iteration", IEEE Signal Processing Letters, 1(10), 144-146, (1994)

[9] 小野田崇, 坂井美帆, 山田誠二, 「k-means 法の様々な初期値設定によるクラスタリング結果の実験的比較」, 『第 25 回人工知能学会全国大会論文集　1J1-OS9-1』 (2011)

[10] 金田尚久, 新居玄武, 「混合分布問題　――その基礎からカーネル降下法まで―― Part 2」, 『学習院大学経済論集』, 第 46 巻 第 2 号, pp.127-170(2009)

[11] 水戸藍, 「情報量基準とその応用」,
http://www.seto.nanzan-u.ac.jp/msie/ma-thesis/2007/KIMURA/m06mm019.pdf

[12] 東証：FLEX Historical サービス（2013 年 1 月 21 日アクセス）
http://www.tse.or.jp/market/service/flex_historical/index.html

[13] 日経株価指数 300 採用銘柄の株価一覧（2013 年 1 月 21 日アクセス）
http://www.nikkei.com/markets/kabu/nidxprice.aspx?index=N300

[14] 日経 500 種平均株価採用銘柄の株価一覧（2013 年 1 月 21 日アクセス）
http://www.nikkei.com/markets/kabu/nidxprice.aspx?index=N500