F-012

メディアンに基づく時系列データの変化点検出法
Change points detection method of the time series data based on medians

小林 えり † 伏見 卓恭 † 斉藤 和巳 † 池田 哲夫 †
Eri Kobayashi Takayasu Fushimi Kazumi Saito Tetsuo Ikeda

1. はじめに

株価や Twitter，レビューデータなど，多様な時系列データの変化点検出は，社会状況の変化の察知や，ユーザの異常行動の把握などに多く利用されている．変化点検出手法として，ユーザの活動間隔のモデルを導入し，尤度比検定を土台として再帰的に変化点を求める手法が提案されており [1]，株価の動きから経済活動の動きや影響力のある社会現象の抽出，twitter のバースト期間抽出によるネット上のイベント，炎上問題の発見などの研究に応用されている．系列データの値が大きく変化した時刻を変化点とし，何らかのイベントや変化要因の発見に利用，また，ある既知のイベント時刻が変化点と見なされていない場合，そのイベントの影響度は低いと考えられ，イベントの影響力の研究にも活用できる．この手法を基に，ガウス分布でのモデリングの下で最小記述長原理を土台に逐次的に変化点を求める手法が提案されている [2]．文献 [2] の手法はある区間を時系列データの平均値で近似したときの L2 誤差（自乗誤差）が最小となるような時刻を変化点とし，再帰的に求める．しかしながら，既存手法は L2 誤差の最小化問題であるために，外れ値による影響を受けやすいことが予測される．

これに対し，本研究では変化点間の時系列データの平均値ではなく，メディアン（中央値）を用いて L1 誤差（絶対値誤差）が最小となるような変化点を検出する手法を提案する．本稿では株価の出来高データを用いて，既存手法と提案法のとりうる変化点時刻を比較し，両者の違いを明確化し評価する．

2. 既存手法

ある時系列の時刻 t の値を x_t とし，時刻 1 から T までの時系列データは $\mathbf{x} = (x_1, \ldots, x_T)^T$ と表す．ここで，変化点の個数を K とし，それぞれの変化点を古い順に F_1 から F_K に格納される変化点時刻ベクトルを $\mathbf{F}$ とし，また便宜上 $F_0 = 0$ かつ $F_{K+1} = T$ と設定すると， 時系列ベクトル $\mathbf{F}$ は $(K+2)$-次元ベクトルで表わされる．区間 $F_{(k-1)} < t \leq F_k$ 内の平均値を μ_k 表すと，L2 誤差 $E(\mathbf{F})$ は次式で計算することができる．

$$E(\mathbf{F}) \quad = \quad \sum_{k=1}^{K+1} \sum_{t=F_{(k-1)}+1}^{F_k} (x_t - \mu_k)^2 \qquad (1)$$

よって，変化点検出問題は $E(\mathbf{F})$ を最小化する $\mathbf{F}$ を求める問題として定式化できる．詳しくは文献 [2] を参照されたい．

3. 提案手法

文献 [2] に対し，我々は平均値ではなくメディアンを用いた L1 誤差を最小化させる変化点ベクトルを検出する手法を提案する．ここでメディアンによる L1 誤差を最小にする変化点ベクトルを $\mathbf{D}$ とし，$\mathbf{F}$ 同様，$D_0 = 0$ かつ $D_{K+1} = T$ と設定し，それぞれの変化点を古い順に D_1 から D_K に格納する $(K+2)$-次元ベクトルで表す．ここで，区間 $D_{(k-1)} < t \leq D_k$ の中央値を m_k とすると L1 誤差は以下の式で定義される．

$$G(\mathbf{D}) = \sum_{k=1}^{K+1} \sum_{t=D_{(k-1)}+1}^{D_k} |x_t - m_k| \qquad (2)$$

ここで、時刻 i から j までのデータと，その区間内のメディアン $\bar{m}(i,j)$ との L1 誤差を算出する関数 $g(i,j)$ を以下のように定義する．

$$g(i,j) = \sum_{t=i+1}^{j} |x_t - \bar{m}(i,j)| \quad (0 \leq i < j \leq T) \qquad (3)$$

いま求める変化点数を $K = 1$ として考える．式 3 を用いて L1 誤差 (式 2) を書き換えると以下の式となる．

$$G(\mathbf{D}) \quad = \quad g(0, D_1) + g(D_1, T) \qquad (4)$$

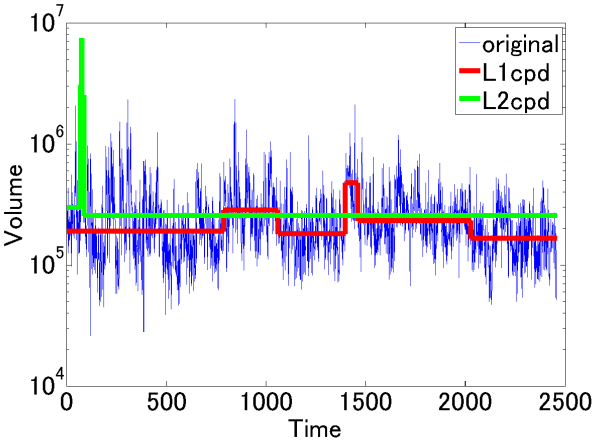
ここで関数 $g(i,j)$ と区間の終点 (または始点) の時系列データ値 x_j(または x_i) を一つ取り除いた関数値 $g(i, j-1)$(または $g(i+1, j)$) との差分値を求める．差分値は，$j - i$ の値によって以下の値を取ることが分かる．

$$g(i,j) - g(i,j-1) = \begin{cases} |x_j - \bar{m}(i,j)| & \text{if } (j-i) \text{ is odd} \\ |\bar{m}(i,j) - x_j| + |\bar{m}(i,j) - \bar{m}(i,j-1)| & \text{otherwise} \end{cases}$$

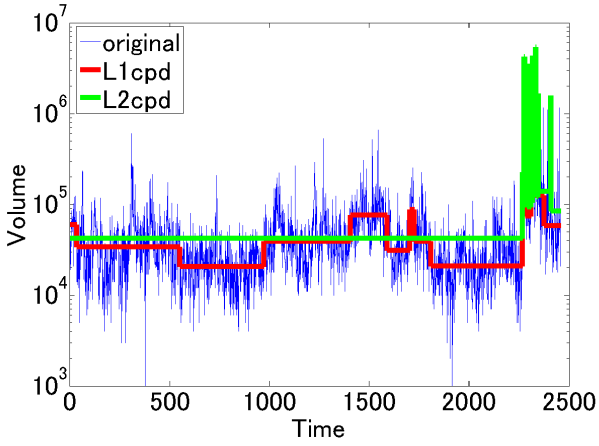
$$g(i,j) - g(i+1,j) = \begin{cases} |x_i - \bar{m}(i,j)| & \text{if } (j-i) \text{ is odd} \\ |\bar{m}(i,j) - x_i| + |\bar{m}(i,j) - \bar{m}(i+1,j)| & \text{otherwise} \end{cases}$$

この性質を利用し，$g(0,T-1), g(0,T-2) \ldots g(0,2)$ と $g(2,T), g(3,T) \ldots g(T-1,T)$ を逐次求め，対となる g のペアを用いれば，各時刻を変化点とみなした際の L1 誤差値が求まる．例えば $g(0,T-1)$ と $g(T-1,T)$ のペアを用いた場合，$g(0,T-1) + g(T-1,T)$ は時刻 $T-1$ を変化点とみなした際の L1 誤差を表す．ここで式 3 の計算に必要なのは各区間での中央値であるが，前処理として，時系列データ値 x_t をソートすることにより求めることが可能である．ここでは $K = 1$ のときのみ記述したが，$k \leq 2$ へ一般化することも可能である．よって式 2 を以下のように書き換えることができ，

†静岡県立大学大学院経営情報イノベーション研究科



ノリタケカンパニーリミテド （ガラス・土石製品）



ダントーホールディングス （ガラス・土石製品）

図 1: 評価値の高い上位 2 社

変化点検出問題は $G(\mathbf{D})$ を最小化する $\mathbf{D}$ を求める問題として定式化できる．

$$G(\mathbf{D})=\sum_{k=1}^{K+1} g(D_{(k-1)}, D_k) \tag{5}$$

4. 実験評価

実験では，東証 1 部 830 社の 2000 年 1 月 1 日から 2009 年 12 月 31 日まで (うち営業日は 2457 日) の株の取引数である出来高データを用いた．

平均値との L2 誤差を用いる従来法を L2cpd 法，メディアンとの L1 誤差を用いる提案法を L1cpd 法とし，両手法より検出した変化点の相違を確認し，それぞれの特徴を明確化することを目的とする．L1cpd 法と L2cpd 法での変化点の違いを表す評価指標として以下の式を定義する．

$$H(K) = \frac{1}{2}\sum_{k=1}^{K}\left(\min_{1\geq l\geq K}(F_k - D_l) + \min_{1\geq l\geq K}(D_k - F_l)\right) \tag{6}$$

評価値 $H(K)$ が大きいほど，L1cpd 法と L2cpd 法は互いに異なる時刻を変化点として検出したことを示す．既存研究 [2] では L2cpd 法において，最適な変化点数を求めるアルゴリズムが提案されているが，L1cpd 法と L2cpd 法では最適変化点数が異なり，公平な評価が行えないため，今回は変化点数 $K = 5$，$K = 20$ と設定した．

4.1. 実験結果

図 1 は $H(5)$ が最も高かった”ノリタケカンパニーリミテド”と，$H(20)$ が最も高かった，”ダントーホールディングス”の結果を示す．縦軸は株の出来高を対数軸で，横軸は時刻である日時を示し，青線が出来高を，赤線が L1cpd の結果を，緑線が L2cpd の結果を表している．

結果を見てみると $K = 5$，$K = 20$ の両結果とも L2cpd 法の変化点は局所的な期間に反応しており，出来高が急激に変化した期間を検出しているのが分かる．一方，L1cpd 法は外れ値の影響を受けずに全体の流れに沿うように変化点を検出しており，出来高の好調期間，不調期間を検出することが出来る．今回設定していない変化点数においても，L2cpd 法は外れ値に反応し，局所的な期間しか抽出出来ていないのに対し，L1 は全体の流れを酌んだ変化点を取る傾向にあった．

4.2. 考察

L2cpd 法は平均値との L2 誤差を誤差関数として用いているため，どうしても外れ値の影響を受けやすく，そのため変化点は局所的部分に反応する傾向にあり，バースト期間抽出に向いていることが分かる．対し，L1cpd 法はメディアンとの L1 誤差を誤差関数として用いているため，外れ値の影響を受けにくく，全体の変動の流れに沿った変化点を検出し，時系列データの長期的な変動を追うことが出来る．よって時系列データの長期的な変動期間抽出，大まかな好調不調期間の分類が可能だと期待できる．

5. おわりに

本研究では，メディアンに基づく時系列データの変化点検出法を提案し，検出される変化点時刻の観点から既存研究と比較しその性能を確認した．L2 誤差に基づく既存研究では，外れ値に強く影響を受けてしまう対し，提案法は影響を受けにくく，そのため全体の流れを把握することが出来た．今後は多様なデータセットでの検証を行っていく．

謝辞 本研究は科学研究費補助金 (No.26330138) の補助を受けた．

参考文献

[1] K. Saito, K. Ohara, M. Kimura and H. Motoda, ”Burst Detection in a Sequence of Tweets based on Information Diffusion Model, ” Proc. of DS2012, pp.239–253, 2012.

[2] 杉澤 優馬，伏見 卓恭，斉藤 和巳，” 時系列変化点の異種時系列への影響度分析，”第 12 回情報科学技術フォーラム（FIT2013），Sep.2013.