時系列階差のブートストラップ解析による折れ線回帰

山中　杏奈　(指導教員：吉田 裕亮)

1　はじめに

本研究では折れ線近似可能な等間隔時系列データに対し, モデルの適切な折曲点の数及び場所を推定するための一つの手法を提案する．従来の推定手法のひとつに誤差分散を用いた手法があるが, これは計算量が莫大であり実用的ではない．本研究ではこの計算量の軽減に重点を置き, 標本平均の差の検定を用いて折曲点が含まれる範囲を決定したのちに, 最適な折曲点の場所を推定する手法を提案する．

2　標本平均の差の検定

標本平均の差の検定とは, 同じ事象についての調査結果である 2 群の標本に対して, それらの標本平均に有意な差があるか否かを統計的に検定する方法である．

まず 2 群の標本 A, B 間には差がないという帰無仮説の下で, A, B のデータ数 n_A, n_B, 標本平均 μ_A, μ_B, 標本分散 σ_A^2, σ_B^2 より T 統計値, すなわち

$$T = \frac{|\mu_A - \mu_B|}{\sqrt{\dfrac{(n_A-1)\sigma_A^2 + (n_B-1)\sigma_B^2}{n_A+n_B-2}\left(\dfrac{1}{n_A}+\dfrac{1}{n_B}\right)}}$$

を求める．この値に基づき t 分布より確率を求め, 予め定めておいた有意水準 α と比較し, 仮説の棄却か採択を判断する．なお, このとき標本 A, B は等分散と仮定されている．

本研究では与えられた時系列データに適当な数の等長を考え, 隣接する 2 区間に対して順次ずらして標本平均の差の検定を行う．差がないと判断された 2 区間は合わせて 1 つの区間とみなすことができ, 従ってその区間内に存在するデータには折曲点にあたるものはないと判断する．逆に, 差がある場合はその 2 区間内に折曲点があると推定できる．しかし, 標本平均を求める際ただ平均を求めただけではその区間内のデータの傾向を把握できない．そこで本研究では, データの傾向を把握した上でより精度の高い T 統計値を推定するために, この標本平均の差の検定を用いる 2 区間のデータに対しブートストラップ法を援用する．

3　ブートストラップ法

ブートストラップ法とは, 推定量の分布において復元抽出を用いた大量の繰り返し計算によって推定する方法である．“ 多数のデータから無作為にいくつかを取り出し計算する ”ことを繰り返し, 分布の状態を推定していく．

解析対象の 2 区間 A, B において, A のデータ数と同じ数の標本を重複を許して無作為に抽出し (フルサンプリング) その標本を A' とする．B に対しても同様の処理を行い B' とする．この標本 A', B' に対して T 統計値を求める．同じ 2 区間で 1000 回同じ操作を行い, 平均値をこの 2 区間の T 統計値とする．標本平均の差の検定を行う 2 区間の区間長を予め設定し, すべてのデータに対ずらしながら次々に T 統計値を算出する．更に本研究では T 値の移動平均を取り平滑化も行うことにする．この T 統計値の変動から, 折曲点があるとされる区間を推定する．

4　折曲点の推定

折曲点があるとされる範囲がいくつか推定されたとき, 1 つの範囲に 1 つの折曲点が存在するとみなすため, 区間内で取り得る限りの折曲点の組み合わせを考える．すべての組み合わせに対し折れ線モデル考え, 最も当てはまりの良いモデルにより最適な折曲点を推定する．その区間内における折曲点の組み合わせについて, その折曲点を通る回帰直線を繋いだ折れ線モデルを考える．それぞれの領域で元のデータとの誤差分散（$V_1, V_2, V_3, \cdots$）を求める．各折曲点から折曲点までを重みつき誤差分散式

$$S = \frac{n_1V_1 + n_2V_2 + n_2V_3 + \cdots}{N},$$

$V_i : i$ グループの誤差分散, N:総数

を評価し, S が最小となる折曲点の組み合わせを, そのデータに対する最適な折曲点と推定する．

5　有効性の検証実験

5.1　シミュレーション実験の概要

以下のように折曲点が 3 点のデータを用意し, 折曲点を含んでいるとされる範囲を推定してから折曲点を推定

することができるか, また標本平均の差の検定を行う区間の長さを 25, 50, 100 と 3 パターン設定し, 区間の長さに対して折曲点の範囲の推定にどのような差が現れるかを調べた.

$$\begin{cases} y_{n+1} = y_n + 0.3 + \varepsilon & (1 \le n < 500, \\ & 1000 \le n < 1500), \\ y_{n+1} = y_n - 0.3 + \varepsilon & (500 \le n < 1000, \\ & 1500 \le n < 2000). \end{cases} \quad (1)$$

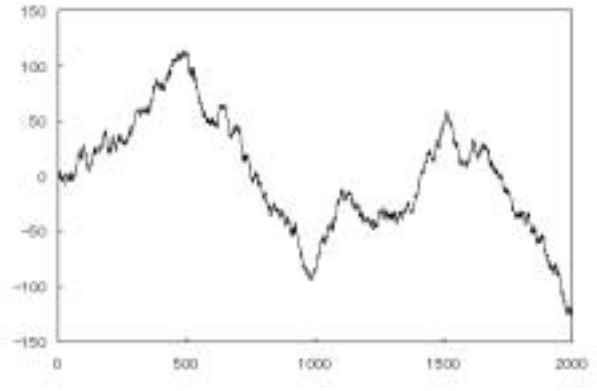

図 1: シミュレーションデータ

5.2 実験結果

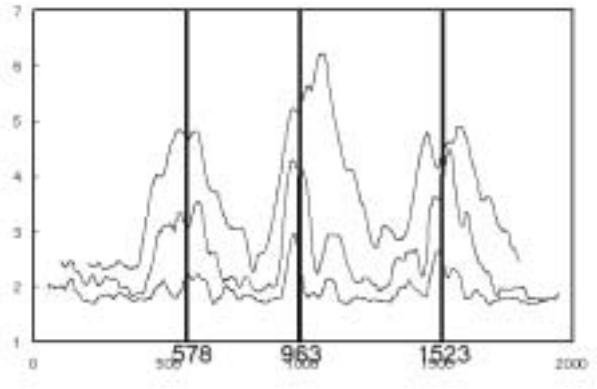

図 2: シミュレーションデータの T 値の変化

5.3 実験の考察

区間長が短いほど多くの折曲点の範囲の可能性が推定でき, 各範囲が小さく推定出来る. 区間長を長くすると推定できる折曲点の範囲の数は減少していくが設定した折曲点が強調されていく. しかし, 各範囲はおおまかにしか推定できない. その結果, 各区間長で推定した折曲点の範囲を比較し共通して推定された範囲を, そのデータの最適な折曲点の場所を含んだ範囲と決定することが適していると分かった. 範囲を決定するには区間長が少ないもの (今回の実験の場合 25) の T 統計値の変動から判断した方が範囲を狭く推定でき, 誤差分散式での計算量を軽減できる. 決定した範囲から誤差分散の式を用いた折曲点の推定では 578, 963, 1523 と得られた.

6 実データへの応用

6.1 為替データ

本研究で提案した手法の応用として, 98.10.28〜09.1.30 の JPY-USD 相場の変動データに対して本研究で提案した手法により折曲点を推定する. データの詳細は次の加茂下氏の予稿, 図 3 を参照. このデータの対数階差に本研究の手法を適用した

6.2 結果

推定された折曲点は 6 点, 折曲点を含んでいる各範囲は ±100 以内で決定でき, その範囲から折曲点の推定を行った. (図 3) 実際のデータと目視で比較しても折れ点を含む範囲を推定できたと判断できる. 重要視していた計算量も従来の方法に比べて, 十分に軽減できた.

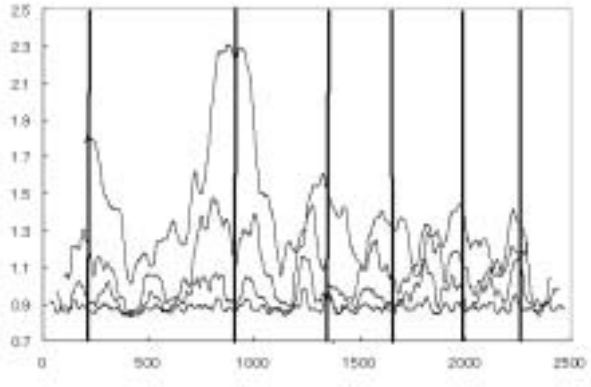

図 3: JPY-USD 相場の変動データの T 値の変化

7 まとめ

ブートストラップ法による標本平均の差の検定を用いることにより, 折れ線近似可能な等間隔時系列データに対し折曲点が存在する妥当な範囲を推定することができた. これは単に誤差分散のみを用いた方法に比べ計算量が少なく, かつ簡便な推定手法といえる. しかし推定される範囲がまだ広いため, 折曲点が多い場合には対応しきれない可能性がある. これを今後の課題としたい.