折れ線トレンドの推定についての研究

A study on estimation methods for piecewise linear trends

経営システム工学専攻 徳永 備久
Tomohisa Tokunaga

1 はじめに

時系列解析は, 株価や気温など, 時間とともに変動する時系列データの分析に関する統計学の一分野である. 時系列データのトレンドとは長期的傾向を意味し, 周期成分などが存在しない場合, 平均値関数と同等である. なめらかなトレンドを前提とした古典的な推定としては移動平均法や多項式回帰などがある. 一方, 本研究ではなめらかさを仮定せず, トレンドに折れ線型を想定する. 折れ線トレンドは時系列の上昇・下降の変化を明確に表現することができる. 従って, 時系列の折れ線トレンドを推定することができれば, トレンドの極値, すなわち山や谷を明示的に求めることができる.

折れ線トレンドを推定する方法としてダミー変数を用いる方法が Draper and Smith[1] で紹介されている. しかし, 折れ点となる時点を与える必要があるうえに, 折れ点の数が増えるにつれてダミー変数を増やす必要がある. また, 大塚・吉原 [3] は最大 2 つまでの折れ点を持つトレンドの推定方法が提案されている. 青木 [5] は折れ点の数が一つの場合について, 折れ点の推定法を紹介している. Gallant and Fuller[2] は区分別に多項式モデルを仮定し, 境界の時点も含めて推定する手法を提案している. しかし, これらの手法では折れ点の数を多くするのは困難である. 樋口 [4] は折れ点の時点をある程度絞り込み, 最適な折れ点を探索して推定する, 複数の折れ点の場合にも対応した手法を提案している. また一方で, Kim et al.[6] は最適化問題の観点から l_1 トレンドフィルタリングという方法を提案している. この手法では折れ点の箇所を指定する必要はなく, また折れ点は多数あってもよい. しかしながら, l_1 トレンドフィルタリングでは正規化パラメータを与える必要があるがどのような値を用いればよいかについては触れられていない. また, 統計的な観点から問題があることが明らかになった. そこで本研究では, 新たな推定法の提案を目的とする. 提案手法の妥当性についてはシミュレーションによって検証する. また, 実際のデータへ適用することによって実用性を示す.

2 時系列と折れ線トレンド

2.1 折れ線トレンド

観測される時系列を $\{y_t|t = 1, 2, \ldots, T\}$ とする. y_t は次のような構造を持つものと仮定する.

$$y_t = \mu_t + z_t$$

ここで, μ_t は折れ線型の平均値関数, z_t は平均 0 の定常過程である. 目的は μ_t を推定することである. なお, μ_t の折れ点の時点および個数は未知とする. また, z_t の確率構造も未知であるものとする.

2.2 折れ線を生成するモデル

Kim et al.[6] はシミュレーションのために折れ線型トレンドを生成するモデルを用いているが, 一般的なモデルとみなすことはできない. そこで, 折れ線トレンド μ_t を生成するための確率モデルを新たに導入する.

各時点における折れ線の傾きを $\{v_t\}$ とおくと

$$\mu_{t+1} = \mu_t + v_t,\ t = 1, \ldots, T-1$$

と表すことができる. ここで $\{v_t\}$ に次のような構造を仮定する.

$$v_{t+1} = \begin{cases} v_t & \text{確率 } p \\ \rho v_t + u_{t+1} & \text{確率 } 1-p \end{cases}.$$

ただし, $\{u_t\}$ は平均 0 のホワイトノイズに従うものとする. また, p については v_t が変わって b 時点以内では 1, そうでないときは定数 p_0 とする.

2.3 評価指標

シミュレーションにおいてトレンドの推定値 $\hat{\mu}_t$ の妥当性を評価するために 2 つの指標 γ_1, γ_2 を導入する.

$$\gamma_1 = \sqrt{\frac{\sum_{t=1}^{T}(\mu_t - \hat{\mu}_t)^2}{T}},$$

$$\gamma_2 = \frac{\sqrt{\sum_{t=1}^{T}(\mu_t - \bar{\mu}_t)^2/T} - \sqrt{\sum_{t=1}^{T}(\hat{\mu}_t - \bar{\hat{\mu}}_t)^2/T}}{\mathrm{sd}(\boldsymbol{z})}$$

と定義する. ここで $\mathrm{sd}(\boldsymbol{z})$ は $\{z_t\}$ の標準偏差である.

γ_1 は真のトレンドと推定値の平均二乗誤差の平方根であり, 値が小さいほど真のトレンドに近い推定値であることを意味する. また γ_2 が 0 に近いほど, ノイズの大きさに比較してトレンド部分の推定がうまくいっていることを意味する.

3 l_1 トレンドフィルタリング

時系列データ $\{y_t\}$ に対し,

$$\frac{1}{2}\sum_{t=1}^{T}(y_t - \mu_t)^2 + \lambda \sum_{t=2}^{T-1} |\mu_{t-1} - 2\mu_t + \mu_{t+1}| \quad (1)$$

という目的関数を最小にするように μ_t を定める手法を l_1 トレンドフィルタリングと呼ぶ. (1) における μ は解を持ち, 以後これを μ^{lt} と表すことにする.

正規化パラメータ λ を与えたときの l_1 トレンドフィルタリングによって得られた μ^{lt} は折れ線トレンドとなる. しかし, l_1 トレンドフィルタリングを使用するには適当な正規化パラメータを与える必要があるが, Kim et al.[6] ではどのような値を用いれば良いかについては触れられていない. そこで本研究では AIC を用いた正規化パラメータの推定法を考案した. それによって推定した例を図 1 に示す. 図 1 において点線が原データ, 太線が推定結果, 実線が真のトレンドを表している. 罰則項の影響を受け推定結果と原データにズレが生じる場合があることが分かる. そこでこの手法とは別に新たな折れ線トレンドの推定法を提案する.

4 折れ線トレンドの同定法

4.1 折れ線型トレンドの推定

準備として, 折れ点の時点が既知であるときの推定法を紹介する. Draper and Smith[1] ではダミー変数を用いているが, 以下のように最小二乗法によって直接推定することも可能である.

折れ点の時点を $k_0, k_1, \cdots, k_L, k_{L+1}$ とする. ただし, $k_0 = 1, k_{L+1} = T$. $T \times (L+2)$ 行列の要素が

$$x_{ij} = \begin{cases} \frac{i-k_{j-2}}{k_{j-1}-k_{j-2}} & (k_{j-2} \le i < k_{j-1}) \\ \frac{k_j - i}{k_j - k_{j-1}} & (k_{j-1} \le i < k_j) \\ 0 & (\text{その他}) \end{cases}$$

$$(j = 2, \cdots, L, L+1)$$

$$x_{i1} = \begin{cases} \frac{k_1 - i}{k_1 - k_0} & (i < k_1) \\ 0 & (\text{その他}) \end{cases}$$

$$x_{i,L+2} = \begin{cases} \frac{i-k_L}{k_{L+1}-k_L} & (k_L \le i) \\ 0 & (\text{その他}) \end{cases}$$

となるよう $\boldsymbol{X}$ を定めると,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

というモデルで記述できる. ここで, $\boldsymbol{\beta}$ は折れ点の高さからなるベクトルであり, $\boldsymbol{\epsilon}$ は誤差項である. 従って $\boldsymbol{\beta}$ は

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

によって推定することができる. 折れ線トレンド $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_T)'$ の推定値は

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

となる.

モデルの評価には $\{z_t\}$ にガウス型ホワイトノイズを想定した擬似的な AIC を用いることにする. 未知パラメータは, 各折れ点の時点 L 個と, 端点を含む各折れ点の高さ $(L+2)$ 個と $\{z_t\}$ の分散 σ^2 である. 従って, L 個の折れ点を持つモデルの AIC は

$$\mathrm{AIC} = T\log(\hat{\sigma}^2) + 4L + 6$$

となる. ここで, $\hat{\sigma}^2$ は σ^2 の推定値であり, 共通項は省略されている.

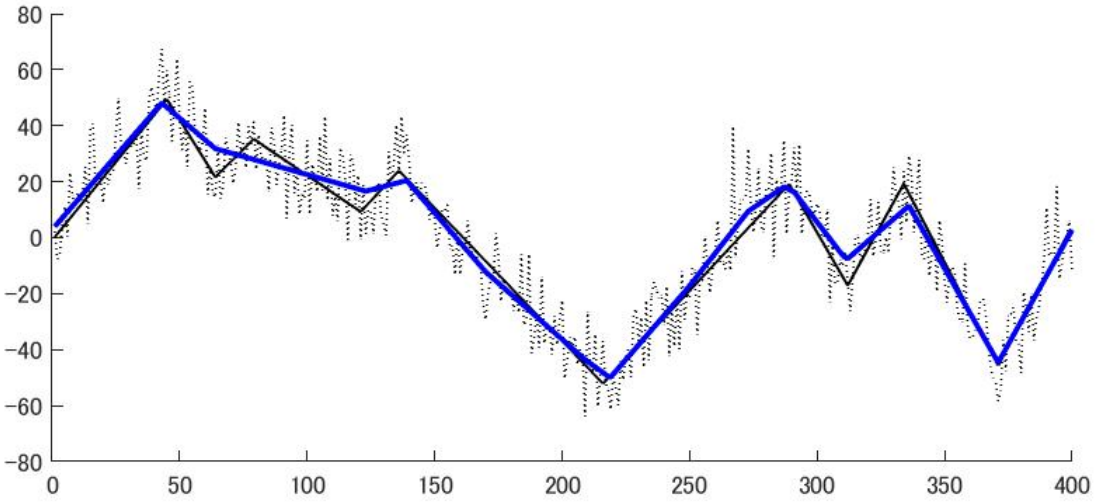

図 1: l_1 トレンドフィルタリングによる推定

4.2 同定アルゴリズム

本研究では, 折れ点の初期集合を与え,4.1 を用いた繰り返し法を提案する. 折れ点の時点の集合を $K = \{k_0, k_1, \cdots, k_L, k_{L+1}\}$ と表す. 初期集合として折れ点の間隔を $2d$ で与える. さらに折れ点の時点の間隔には下限があるものと仮定し, その下限を $2d'$ とする.

1. K の初期値を次のように定める.

$$k_\ell = 2\ell d + 1,\ \ell = 1, 2, \cdots, L.$$

ただし, $L = \left[\frac{T-1}{2d}\right] - 1$.

2. $\ell = 1, 2, \cdots, L$ について, 区間 $[k_{\ell-1} + d', k_{\ell+1} - d']$ 内で AIC が最小となる時点 k'_ℓ とその時の AIC である AIC_ℓ を求める. さらに, k_ℓ を削除したときの AIC を求め, AIC_ℓ より AIC が小さいときはそれを AIC_ℓ とし, $k'_\ell = \mathrm{null}$ とする.

3. AIC_ℓ $(\ell = 1, 2, \cdots, L)$ が最小となる ℓ について, k_ℓ を k'_ℓ で置き換える. ただし, $k'_\ell = \mathrm{null}$ のときは k_ℓ を削除し, $L = L - 1$ とする.

4. 2. と 3. を K が変わらなくなるまで繰り返す.

5 シミュレーション

提案手法の妥当性を検証するために 2.2 のモデルにより生成した人工データを用いたシミュレーションを行う.

5.1 方法

シミュレーションにおいては, 標本平均 $\sum_{t=1}^{T} y_t/T = 0$, 標本標準偏差 $s(\boldsymbol{\mu}) = \sqrt{\sum_{t=1}^{T}(\mu_t - \overline{\mu})^2/T}$ が一定となるように y_t を生成する. また $v_1 \sim N(0, \sigma_r^2)$, $u_t \sim N(0, 10)$ とし, $\rho = -1$ とした. $\{z_t\}$ はガウス型の AR(1):

$$z_t = \phi z_{t-1} + w_t$$

とし, σ_w^2 は標準偏差 $\mathrm{sd}(\boldsymbol{z}) = \sqrt{\sigma_w^2/(1-\phi^2)} = 10$ となるように定めた. $s(\boldsymbol{\mu})$, ϕ の値を変えて系列長 $T = 400$ とし, $b = 10$, $p_0 = 0.95$ それぞれについて 100 本の系列を生成してシミュレーションを行った. l_1 トレンドフィルタリングでは $0.01 \le \lambda \le 1000$ の範囲を 0.01 刻みで, 提案手法では $5 \le d \le 100, 5 \le d' \le d$ の範囲で適用し, 得られた結果の中から最適なものを選択した.

5.2 結果

生成した各系列に対し l_1 トレンドフィルタリングと提案手法を適用し, それぞれ折れ線トレンドを推定した. γ_1 の値の平均値 (下の数値は標準偏差) を表 1 と表 2 に, γ_2 の値の平均値 (下の数値は標準偏差) を表 3 と表 4 に示す.

表 1: γ_1 (提案手法)

平均 (標準偏差)		sd($\boldsymbol{\mu}$)		
		25	50	100
ϕ	0.00	3.0788 (0.5360)	3.1314 (0.5531)	3.1901 (0.5677)
	0.05	3.2388 (0.5464)	3.3395 (0.5308)	3.4144 (0.5884)
	0.10	3.4937 (0.5658)	3.6014 (0.5394)	3.6477 (0.5971)

表 2: γ_1 (l_1 トレンドフィルタリング)

平均 (標準偏差)		sd($\boldsymbol{\mu}$)		
		25	50	100
ϕ	0.00	3.0851 (0.6953)	3.2772 (0.6692)	3.4228 (0.6602)
	0.05	3.1694 (0.7592)	3.3501 (0.6715)	3.4678 (0.6140)
	0.10	3.1947 (0.7131)	3.4115 (0.6841)	3.5886 (0.6306)

6 考察

ホワイトノイズの場合, 提案手法は l_1 トレンドフィルタリングと比較して, 指標 γ_1 はより小さく, 指標 γ_2 の絶対値もより小さくなっており, 推定法としてより適当であることが分かる. ノイズに自己相関がある場合, ノイズがトレンドの推定に影響を与え, その結果 γ_1 の値が大きくなる様子が見られた. 他方で γ_2 は提案手法では平均で負の値を取っており, わずかにオーバーフィッティングの傾向にあるとみなされる. これは AIC を採用したことによるものと考えられる.

表 3: γ_2 (提案手法)

平均 (標準偏差)		sd($\boldsymbol{\mu}$)		
		25	50	100
ϕ	0.00	-0.0181 (0.0487)	-0.0108 (0.0487)	-0.0065 (0.0484)
	0.05	-0.0203 (0.0514)	-0.0123 (0.0511)	-0.0073 (0.0510)
	0.10	-0.0239 (0.0541)	-0.0140 (0.0539)	-0.0082 (0.0535)

表 4: γ_2 (l_1 トレンドフィルタリング)

平均 (標準偏差)		sd($\boldsymbol{\mu}$)		
		25	50	100
ϕ	0.00	0.0440 (0.0518)	0.0402 (0.0555)	0.0371 (0.0525)
	0.05	0.0425 (0.0566)	0.0393 (0.0544)	0.0341 (0.0552)
	0.10	0.0406 (0.0564)	0.0360 (0.0584)	0.0344 (0.0577)

本研究では時系列データにおいてなめらかさを仮定せず, トレンドに折れ線型を想定した同定法を提案した. 移動平均法や多項式回帰などといった従来の手法と異なり, 極値を明確に推定できることがこの手法の特徴である. 極値の推定や予測が求められるような状況において提案手法は有用であろう. AR ノイズを仮定した AIC の導入や他の情報量規準を用いた場合との比較は今後の課題である.

参考文献

[1] Draper, N. and Smith, H. (1966) *Applied Regression Analysis*, Wiley.

[2] Gallant, A. R. and Fuller, W. A. (1973) Fitting Segmented Polynomial Regression Models Whose Join Points Have to Be Estimated, *Journal of the American Statistical Association*, 68:341, 144-147

[3] 大塚 雍雄・吉原 雅彦 (1975) 1 ないし 2 の折曲点をもつ折れ線モデルのあてはめ, 応用統計学 Vol.5, No.1, 29-39.

[4] 樋口 知之 (1999) 大規模データの発見的全探索:大規模地球電流系構造の自動同定，統計数理 Vol. 47, No.2, 291-306.

[5] 青木 繁伸. “R – 折れ線回帰”. R による統計処理. 2009-02-04. http://aoki2.si.gunma-u.ac.jp/R/oresen.html, (参照 2020-02-16)

[6] Kim, S.-J. and et al.(2009) $l_1$trend filtering, *SIAM Review*, Vol.51, No.2, 339-360.