社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－MPS－73 (17)
2009／3／5

平滑化帯域幅の決定を含む移動平均の一手法

田中 佑生子
お茶の水女子大学大学院 人間文化創成科学研究科

本研究では, 新たな移動平均法を提案すると共に, 等間隔時系列データにおける平滑化パラメータの最適値を推定するひとつの方法を示す. 平滑化の良さを測る指標として, 平滑化済データに求められる相反する二つの性質 "元データに対する忠実性" と "滑らかさ" を按配するある規準式を導入する. 二つをどの程度ずつ考慮するかという兼ね合いは, 元データに含まれるノイズ量およびそれらが最も理想的に平滑化されたデータにおける先の二つの特徴量を推定した値によって決定している. この規準式を用いて, 先の二つの性質を最も理想的な配分で備えている平滑化済データを見つけることにより, 平滑化パラメータの最適値を決定する.

A smoothing method with a certain MA model by determing the values of the parameters

Yukiko Tanaka
Graduate school of Humanities and Sciences, Ochanomizu University

We shall show a new method of smoothing with MA model and of determing values of the parameters in the model. The goodness of the smoothed data is expressed by the fidelities to the original data and its smoothness, but unfortunately these two factors are in the relation of trade-off. Here we introduce the formula for mesuring the quantity of the goodness of the smoothed data, which is based on the value which strikes a balance between the above two factors. Using our criterion, each the best value of the pameters on our model used for the smoothed data will be selected as the best.

1 研究背景と目的

科学計測機器からの出力や種々の現象の観測から得られたデータには通常, 目的の信号以外にも様々なノイズ成分が含まれている. これらの除去手法の代表なものとして, 以下に示す移動平均法がある.

信号を N 個の離散値 $\{g(n)\}$ $(n = 1, 2, \cdots, N)$ で表したとき, これらの平滑値を次のように求める.

$$s(m) = \frac{1}{W}\sum_{i=-b}^{b} w(i)y(m+i)$$

$$\left(\begin{array}{l} m = b+1, b+2, \cdots, N-b \\ W = \displaystyle\sum_{i=-b}^{b} w(i) \qquad (w(i) : \text{任意関数}) \end{array} \right)$$

$2b+1$ 個の離散点上の任意関数 $w(i)$ は重み関数とよばれ, また $2b+1$ は平滑化の帯域幅とよばれている. これらには過去の研究で推奨されている既定値もしくはデータの発生機構や経験に基づき ad hoc に決定された値が用いられることが多い.

本研究では, これら平滑化パラメータの最適化を含むある移動平均法を提案する.

2 提案する平滑化法

2.1 モデル

系列 $\{g(n)\}$ $(n = 1, 2, \cdots, N)$ の平滑化成分を次のように定める.

$$s(m) = \frac{1}{b}\sum_{i=1}^{b} R_{m-b+i,m-1+i}(b-i+1)$$

ただし

$$m = b, b+1, \cdots, N-(b-1), \quad b \geq 4$$

であり, k と l が与えられたとき, $R_{k,l}(j)$ $(j = 1, 2, \cdots, l-k+1)$ は $\{g(n)\}_{n=k}^{l}$ の回帰曲線系列の j 番目とする.

この移動平均モデルは有する平滑化パラメータが帯域幅ひとつであることが特徴であり, 従って以後はこの帯域幅の最適化について考える. 一つの平滑値が決まるのに必要とされるデータ数を平滑化帯域幅とすると, このモデルにおける帯域幅は $2b-1$ と表される. ただし本稿では, モデル式の特徴を踏まえ, 多項式回帰を施すデータ数つまり window の意味に近い b に着目した議論を行う.

2.2 帯域幅決定の規準式

2.2.1 用いる記号について

以降, 与えられたデータと平滑化されたデータの誤差分散を E, データの第 1 階差二乗和の平均を D, 分散を V とおき, これらが算出される対象となるデータを次のように表す.

E_*, D_*, V_* について

$$\left(\begin{array}{l} *=all \quad (D,V \text{ のみ}): \\ \quad \text{与えられたデータ } (N \text{ 個}), \\ *=all\_S: \\ \quad all \text{ の理想的な平滑化済データ } (N \text{ 個}), \\ *=p(b) \quad (D,V \text{ のみ}): \\ \quad \text{平滑化されるデータ } (N-2(b-1) \text{ 個}), \\ *=p(b)\_S: \\ \quad p(b) \text{ の平滑化済データ } (N-2(b-1) \text{ 個}). \end{array}\right.$$

ただし, $*=all\_S$ に関しては実際に計算することはできない. また算出に用いたデータの違いを明確に示すため, 本稿では $*=all\_S$ および $p(b)\_S$ の場合には $\widetilde{E}_*, \widetilde{D}_*, \widetilde{V}_*$ と表すこととする.

2.2.2 規準式

平滑化済データには, "元データに対する忠実性" と "滑らかさ" の相反する 2 つの性質が求められる. 本稿では前者は $\widetilde{E}_{p(b)\_S}$, 後者は $\widetilde{D}_{p(b)\_S}$ で表せると考え, 平滑化の良さを測る指標として次の系列 $\{SG(\alpha,b)\}_b$ すなわち

$$SG(\alpha,b)=\alpha\cdot\log\left(C_1\cdot\widetilde{E}_{p(b)\_S}\right)+\log\left(\frac{\widetilde{D}_{p(b)\_S}}{D_{all}}\right) \quad ,$$

ただし,

$$C_1=\left[\frac{\widetilde{D}_{all\_S}}{\widetilde{E}_{all\_S}}\text{の推定値}\right] . \qquad \text{(次節参照)} \qquad (1)$$

ここで, (α,b) は以下の条件を満たす.

(i) $\alpha\in\mathbf{Z}$, $\alpha_{\mathbf{q}^{\star_1}}\le\alpha\le\alpha_{\mathbf{p}^{\star_1}}$

$\left(\alpha_\natural:\ SG(b=\natural) \text{ に最小値を与える } \alpha\right)$

(ii) $b=p^{\star_1},p^{\star_1}+1,\cdots,q^{\star_1}-1,q^{\star_1}$

$\left(p^{\star_1}\sim q^{\star_1}:\widetilde{D}_{p(p^{\star_1})\_S}\fallingdotseq\cdots\fallingdotseq\widetilde{D}_{p(q^{\star_1})\_S} \text{ を満たす範囲}\right)$

このとき, 最も多く $\{SG(\alpha,b)\}_b$ に最小値を与える b を $b^\sharp_\alpha$, つまり

$$b^\sharp_\alpha=\arg\max^\sharp\left\{\arg\min_b\{SG(\alpha,b)\}\right\}$$

とすると, 与えられたデータの平滑化に最適な帯域幅は $2\,b^\sharp_\alpha-1$ によって与えられるものとする.

3 $\dfrac{\widetilde{D}_{all\_S}}{\widetilde{E}_{all\_S}}$ の推定

(1) に示した C_1 は以下のように求める.

$$\begin{aligned} C_1&=\left[\frac{\widetilde{V}_{all\_S}}{\widetilde{E}_{all\_S}}\text{の推定値}\right]\times\left[\frac{\widetilde{D}_{all\_S}}{\widetilde{V}_{all\_S}}\text{の推定値}\right] \\ &=\left[\frac{1}{q^{\star_2}-p^{\star_2}+1}\sum_{b=p^{\star_2}}^{q^{\star_2}}\frac{\widetilde{V}_{p(b)\_S}}{\widetilde{E}_{p(b)\_S}}\right]\times \\ &\quad\left[\frac{1}{K_2-K_1+1}\sum_{k=K_1}^{K_2}\left(\beta\left(x^\star{}_b,k\right)\cdot\frac{D_{all}}{V_{all}}\right)\right], \end{aligned}$$

ただし,

$$\begin{aligned} \beta(x,k)&=\left(\frac{\widetilde{D}_{all\_S}}{\widetilde{V}_{all\_S}}\right)\Big/\left(\frac{D_{all}}{V_{all}}\right) \\ &=\left(\frac{D_{all}}{V_{all}}\cdot\frac{x}{k}\right)\Big/\left(\frac{D_{all}}{V_{all}}\right)=\frac{x}{k} . \end{aligned} \qquad (2)$$

$$x^\sharp{}_b=\arg\max^\sharp\left\{\arg\min\left\{\left(\frac{D_{p\_S(b)}}{V_{p\_S(b)}}-\beta(x,k)\right)^2\right\}\right\}$$

$$\left(\begin{array}{l} x=1,2,\cdots,k-1 \\ k: \text{十分に大きな自然数} \end{array}\right) .$$

このとき, 与えられたデータは

$$0 < \frac{\widetilde{D}_{all\_S}}{\widetilde{V}_{all\_S}} < \frac{D_{all}}{V_{all}}$$

を満たす ((2) で使用) .

ここで, 以下の条件が満たされるものとする.

(i) $(b =) p^{\star 2}, p^{\star 2}+1, \cdots, q^{\star 2}-1, q^{\star 2}$:

次の 2 式を対数で満たす範囲

$$\frac{\widetilde{V}_{p(p^{\star 2})\_S}}{\widetilde{E}_{p(p^{\star 2})\_S}} \fallingdotseq \cdots \fallingdotseq \frac{\widetilde{V}_{p(q^{\star 2})\_S}}{\widetilde{E}_{p(q^{\star 2})\_S}}$$

$$\frac{\widetilde{D}_{p(p^{\star 2})\_S}}{\widetilde{V}_{p(p^{\star 2})\_S}} \fallingdotseq \cdots \fallingdotseq \frac{\widetilde{D}_{p(q^{\star 2})\_S}}{\widetilde{V}_{p(q^{\star 2})\_S}}$$

(ii) K_1, K_2 :

$K_1 < K_2$ を満たす十分に大きな自然数

4 数値実験

4.1 実験概要

適当な箇所にピークを 2 つ入れた正弦曲線系列 $\{ORG_n\}$ を用意し, これに分散の異なる正規ノイズを加えた 3 つのデータ系列 A_n, B_n, C_n を作った. 1 つの系列を構成するデータ点数は 600 である. これらの平滑化を行い, 元データに含まれるノイズの分散によって, 推定される $\widetilde{D}_{all\_S}/\widetilde{E}_{all\_S}$ や最適な平滑化帯域幅の値にどのような違いがみられるかをみる. また, これらの推定値を用いて得られた平滑化済データの形状を, 特にピーク部分に注目しながら元データと比較する.

4.2 実験結果

推定された $\widetilde{D}_{all\_S}/\widetilde{E}_{all\_S}$ と平滑化帯域幅の値, および元のデータと平滑化済データを以下に示す.

	$\widetilde{D}_{all\_S}/\widetilde{E}_{all\_S}$		帯域幅
データ	設定値	推定値	
A_n	0.325101	0.353806	33
B_n	0.040086	0.050802	51
C_n	0.009908	0.015954	109

Tab. 1: 数値実験の結果

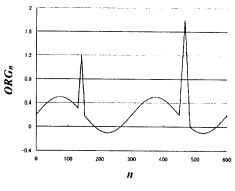

Fig. 1: ORG_n

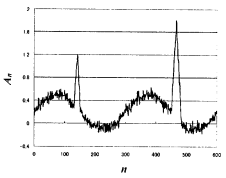

Fig. 2: A_n

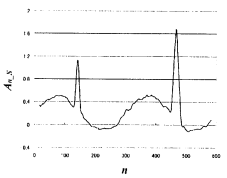

Fig. 3: 平滑化済 A_n

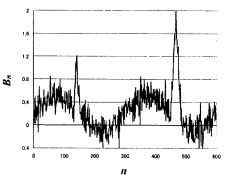

Fig. 4: B_n

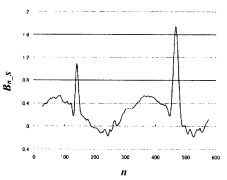

Fig. 5: 平滑化済 B_n

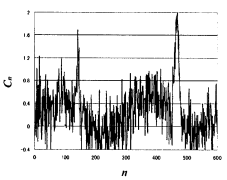

Fig. 6: C_n

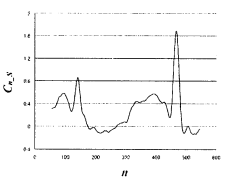

Fig. 7: 平滑化済 C_n

4.3 考察

$\widetilde{D}_{all\_S}/\widetilde{E}_{all\_S}$ と帯域幅の推定値を用いることにより, どのデータの平滑化においても ORG_n に近い形状をもつ平滑化済データを得ることができた. ピーク部分もが大幅に均されることなく平滑化が行われている. また, $\widetilde{D}_{all\_S}/\widetilde{E}_{all\_S}$ の設定値を用いて計算した場合も, 得られる帯域幅の推定値は Tab. 1 に示した値と変わらなかった. 従って $\widetilde{D}_{all\_S}/\widetilde{E}_{all\_S}$ は, 本稿で提案している手法において, 設定値の代用が可能な範囲で推定できていると考えられる. 帯域幅は分散の大きなノイズをもつデータほど大きな値が推定されたことも理論的な結果といえる.

5　実データへの応用

発光しているプラズマの可視光の発光強度, および月単位で観測された 50 年分の太陽黒点相対数のデータの平滑化を行い, 以下のような結果を得た. 観測データの点数は前者が 800 , 後者が 600 である.

データ	$\widetilde{D}_{all\_S}/\widetilde{E}_{all\_S}$	帯域幅
プラズマの発光強度	0.729483	9
太陽黒点相対数	0.037214	65

Tab. 2: 実データ実験の結果

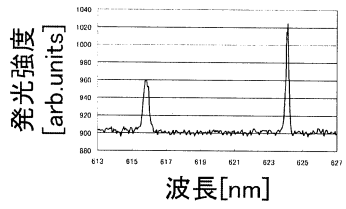

Fig. 8: プラズマの発光強度の観測データ

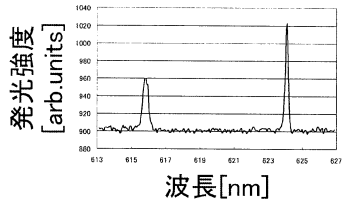

Fig. 9: プラズマの発光強度の平滑化済データ

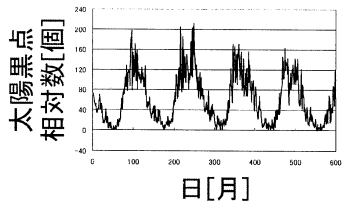

Fig. 10: 太陽黒点相対数のデータ

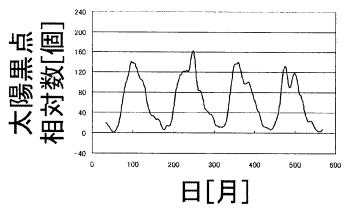

Fig. 11: 太陽黒点相対数の平滑化済データ

6　まとめと今後の課題

モデル化に必要なパラメータを定式し, かつ単純な移動平均法では均らされやすいピークを大幅に消すことなくデータを平滑化することができた. しかし, このような部分的な特徴と全体的な形状の両方を十分に捉えるにはさらなる改良が必要である. 特に, 5 節で扱ったプラズマの発光強度のデータのように, ノイズに比してピーク値が極度に大きい場合に関しては課題が多く残る.

謝辞

本研究を進めるにあたり, 実データの提供をいただきました株式会社デンソーの棚橋裕基氏に感謝申し上げます.

参考文献

[1] 南茂夫, 科学計測のための波形データ処理, CQ 出版社 (1986).

[2] Tanaka Y., Yoshida H., “An estimation of the most suitable domain on the linear regression for partially linearizable data”, preprint(2008).

[3] Tanabe K., Hiraishi J., “A Comparison Between Kawata-Minami and Savizky-Golay Smoothing Methods with Raman Spectral Data”, Computer Enhanced Spectroscopy. vol.2, no.1, pp.17-20, Nov.1984.

[4] Whittaker E. T., “On a New Method of Graduation”, Proc. Edinburgh Math. Soc., vol.41, pp.63-75, 1923.

[5] Henderson R., “A New Method of Graduation”, Transactions of the Actuarial Society of America, vol.25, pp.29-40, 1924.