ベイズ型スプライン回帰の応用

高橋　啓　　統計思考院　特任助教

【スプライン回帰】

スプライン回帰とは，区分的な多項式（スプライン）関数により，変量 y_iと x_i $(i = 1, 2, \ldots, N)$ の間に潜む関係を見出す手法である．ここで，用いるスプライン関数を B-スプライン関数とし，ノットを幅 d区間数 Mの一様ノットとすると，

$$\Omega(y_i|\alpha) = \sum_{j=1}^{M} \alpha_j s_j(x), \quad d \equiv \frac{c-b}{M}$$

なるスプラインとなる．ここで，c, b は，分割する上限，下限である．ここで，この滑らかな曲線 $\Omega(y_i|\alpha)$を同定する問題は，各区間 j の重み α_j を推定する問題となる．具体的にスプライン関数 $s_j(x)$ を特定する前に，次の記号を定義する．i 番目のノットt_iは d により，次のように表される：

$$t_i = b + di \quad i = 0, 1, \ldots, M$$

さらに $j(x)$を区間 $(b, c]$ に対して次のように定義する：

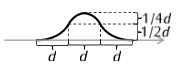
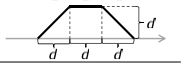
$$j(x) = j \quad \text{if } q \in (t_{j-1}, t_j]$$

ここで，$(t_{j-1}, t_j]$を j–区間と呼び，この区間の中点$d(j)$を次のように定義する：

$$d(j) = \frac{t_{j-1} + tj}{2} = b + d\left(j - \frac{1}{2}\right)$$

具体的なスプライン関数は，2次，1次（折れ線）の場合，次のとおりとなる：

$$s_j(x) = \begin{cases} \Psi(x - d(j-1)) & \text{if } j(x) = j-1 \\ \Phi(x - d(j)) & \text{if } j(x) = j \\ \Psi(d(j+1) - x) & \text{if } j(x) = j+1 \\ 0 & \text{otherwise} \end{cases}$$

	$\Phi(x)$	$\Psi(x)$	概形
2次の場合	$\Phi(x) = -\frac{1}{d^3}x^2 + \frac{3}{4d}$	$\Psi(x) = \frac{1}{2d^3}\left(x + \frac{d}{2}\right)^2$	
1次の場合	$\Phi(x) = d$	$\Psi(x) = x + \frac{d}{2}$	

【スプライン回帰のベイズ推定】

一般的に，スプライン回帰のパラメータ推定は，尤度関数：

$$L(\alpha, \sigma^2, M) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \sum_{j=1}^{M} \alpha_j s_j(x_i)\right)^2\right)$$

を最大化することにより行われる．しかし，場合によっては任意のノットにデータが存在しなくなり，極端な場合パラメータが不定となる．これを解決するために，$\alpha_1, \alpha_2, \ldots, \alpha_M$の2階の階差：

$$\Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2} \quad j = 1, 2, \ldots, M \qquad (1)$$

が0近傍に分散u^2で分布するとする．条件 (1) のみの場合には，事前分布：

$$f(\alpha|u^2, M) = \left(\frac{u}{\sqrt{2\pi\sigma}}\right)^M \exp\left(-\frac{u^2}{2\sigma^2}\sum_{j=1}^{M}|\Delta^2\alpha_j|^2\right)$$

を仮定していることとなり，ABIC最小化により，解析的にα, u, M を求めることができる．また，これ以外に，スプライン曲線の滑らかさを増すための条件：

$$\Delta_2 \alpha_j = \alpha_j - \alpha_{j-2} \quad j = 1, 2, \ldots, M \qquad (2)$$

が，0近傍に分散s^2で分布するとする．この場合は解析的に求めるのが困難であるため，MCMC（R + Rstan）により，解を求める．

【天文学における応用：CCDカメラのノイズ除去】

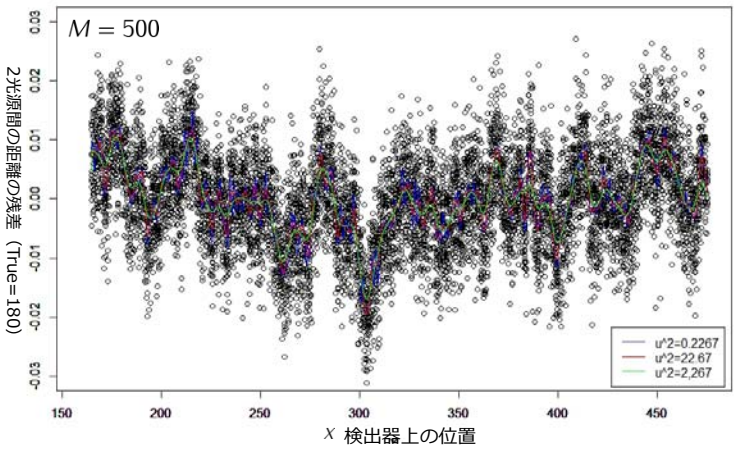
これは，ベイズ型スプライン回帰により，人工衛星に搭載する望遠鏡のCCDカメラのノイズを除去したものである．CCDカメラのノイズは，検出器の各ピクセルごとの光の干渉，感度特性のムラ，暗電流などにより，ある点を中心として，同心円状に存在し，一定距離以上離れるとほとんど影響がなくなるという特性がある．あるピクセル従来，天体観測におけるこの種のノイズ除去には非線形最小二乗法が用いられてきたが，この手法では次のような欠点があった：

・何次関数を用いるべきか，

・それが例え決まったとしても，全体最適解はなかなか求まらない

・計算に時間がかかる

本研究では，**2次スプライン**を用い，(1)の制約のもとで，スプライン曲線を描き，これらのノイズを除去（ダーク・フレームの作成）している．用いるデータは，水沢実験と呼ばれる地上における一次元のものであり，データ数は9,000観測である．



M	10	100	500	1,000	2,000
ABIC	−150, 104	−155, 438	−156, 853	−156, 856	−156, 857
σ	7.592×10^{-3}	6.510×10^{-3}	6.159×10^{-3}	6.159×10^{-3}	6.157×10^{-3}
u	9.382×10^{-2}	9.398×10^{-1}	4.762×10^{0}	2.807×10^{1}	1.551×10^{2}

【水文学における応用：流量ー水位曲線】

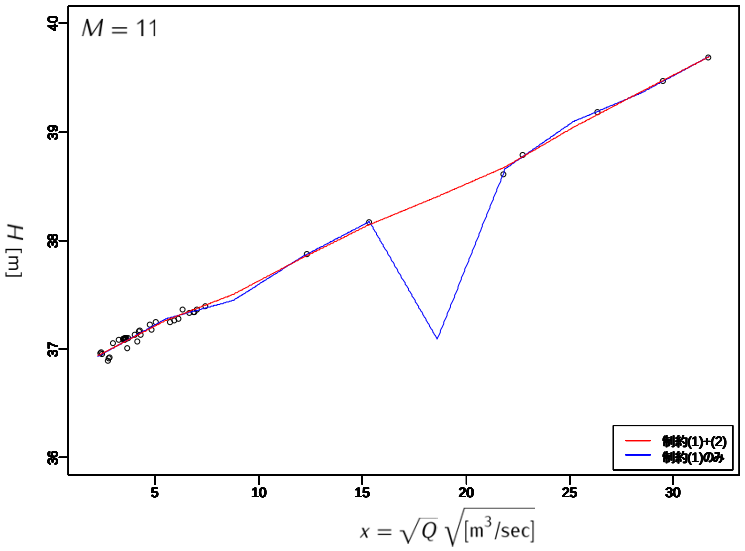
日本の河川管理において基本的資料となる流量-水位曲線（Q-H Curve）をベイズ型スプライン回帰により描いたものである．Q-H Curve の推定は，Qについて1/2 乗したうえで，任意の区間に分割し，それぞれの区間で線形回帰することで行われている．そして各直線の採用の可否は，相関係数が0.8以上というあいまいな基準で行われている．しかし，この一連の手法は，次のような欠点が指摘されている：

・区間分割が任意に行われるため，作成者により解が異なる

・各区間の直線が交わらない場合がある

・データが少ない直線では過適合のおそれがある

本研究では，**1次スプライン(折れ線)**を用い，これらの問題点を解決する手法を提案している．制約としては，(1)だけではなく(2)も用いている．(2)も用いない場合，流量が増えても水位が下がるという折れ線が描かれてしまう．用いるデータは，豊平川（中流域）の日データ（43観測）である．なお，より対象領域における正確性を増すために，上下両方に全くデータ点の存在しないノットをとり，表示する際にこの部分の直線は消去している．



$x = \sqrt{Q}\ \sqrt{[\mathrm{m}^3/\mathrm{sec}]}$

【今後の課題】

・CCDカメラ　…二次元スプラインへの拡張，経年変化によるノイズの除去

・流量ー水位曲線…季節性の考慮

・その他　…交通流における Flow-Density Curve の推定

両方とも共同研究スタートアップからスタートした研究です