No. 1041

B-スプライノンパラメトリック回帰モデルにおける過剰適合の回避について

by

柳原宏和・大瀧慈

May 2003

B-スプラインノンパラメトリック回帰モデルにおける過剰適合の回避について

柳原 宏和[1], 大瀧 慈[2]

要旨：複雑なトレンドを持つデータに関して平滑化をおこなう場合, B-スプラインノンパラメトリック回帰モデルを用いた手法は, 単純でかつ短い計算時間で結果を得ることができるといった利点を持つが, このモデルは, 本来柔軟なモデルであるため, 柔らか過ぎる適合, 過剰な適合を起こす場合がある. 本論分では, 基底関数の個数と平滑化パラメーターを決定する情報量基準, または推定法の改良することで, この過剰適合を回避する手法を提案する.

Keywords：M-推定, 局所線形適合, 散布図平滑化, 情報量基準, 平滑化パラメーターの決定.

1. はじめに

誤差を伴いかつ複雑なトレンドを持つデータに関して平滑化をおこなう場合, B-スプラインノンパラメトリック回帰モデルを用いた手法は, その他のノンパラメトリックな手法に比べて, その推定法が従来の線形モデルにおける係数推定問題に帰着できるために, 単純でかつ短い計算時間で結果を得ることができるといった利点を持つ. また平滑化パラメーターを伴なう場合でも, 係数の推定量自体はリッジ推定量と同じ形をし, その罰則パラメーターの決め方も情報量基準を用いた手法が提案されており, さほど複雑な手順を踏む必要が無い. さらに, 線形モデルの係数推定問題に帰着できるという特性から, 低い次数 (例えば直線や 2 次曲線) でのパラメトリックモデルと B-スプラインノンパラメトリックモデルを共通の情報量で比

[1]〒 305-8573 つくば市天王台 1-1-1 筑波大学 社会工学系, *e-mail*: *yanagi@sk.tsukuba.ac.jp.*
[2]〒 734-8553 広島市南区霞 1-2-3 広島大学 原爆放射線医科学研究所, *e-mail*: *ohtaki@hiroshima-u.ac.jp.*

較し, それらパラメトリックモデルをも候補のモデルに含んだ, 自動最適化アルゴリズム (Satoh, Yanagihara and Ohtaki (2003)) も提案されている.

しかしながら, B-スプラインノンパラメトリック回帰モデルは, 本来柔軟なモデルであるため, ときに図 1 のような柔らか過ぎる適合, つまり過剰な適合を起こすことがある. 本論文ではこのような適合を過剰適合と呼ぶ. そのような過剰適合をひきおこす原因として以下のようなことが考えられる:

1. 平滑化パラメーターが小さめに決定されている.
2. 基底関数の個数が多めに選択されている.
3. 外れ値の影響を受けている (正規性の仮定の崩れ).
4. 説明変数 x の散らばりに関して粗密がある.
5. 等分散の仮定が崩れている.

4 が原因であるような過剰適合に関しては, Yanagihara and Ohtaki (2003) で簡単に求めることができる節点の配置によって, 過剰適合を回避する最適化アルゴリズムを提案している. この節点の配置は, 複雑な計算を必要とせず, 非線型最小二乗法の問題を解いて最適な接点の配置を決める方法 (de Boor and Rice (1968)), 逐次分割法 (吉本, 市田, 清野 (1979)) また節点の配置も基準量で決める方法 (Friedman (1991)) などに較べはるかに簡単に短時間で求めることができるという利点をもつ. 5 が原因になるものに関しては, Imoto (2001) により部分的に異分散であるモデルを導入することにより過剰適合を回避する手法が提案されている.

本論文の主目的は 1, 2, 3 が原因であると考えられる過剰適合の回避にある. 基本的にはどのような原因に依存する過剰適合であっても, 節点の配置をずらしたり, 基底関数の個数の上限を変更することである程度回避できると思われる. しかしながら我々はこの B-スプラインノンパラメトリック回帰モデルを用いて最終結果を得るということよりもむしろ, ACE アルゴリズム (Breiman and Friedman (1985)) を用いた重回帰モデルへ拡張することや, 射影追跡法 (Friedman and Tukey (1974)) への適用, また, 一度に複数回モデルを適合して, その平滑化結果を実データ解析

に用いる (大瀧 他 (2000)), など B-スプラインを用いた平滑化を一つの関数としてとらえ, その他の手法に適用することを中心に考える. そのため, 一度の出力であまり間違っていない結果 (ここでは過剰適合をおこしていない結果) を得ることができる方が望ましい. 以上のことから, 簡単にかつ効果的に複数回の調整を必要とすることなく過剰適合を回避する手法が必要となる. 以下, 第 2 節では B-スプラインノンパラメトリックモデルと推定法を紹介する. 第 3 節では, 推定分散に意味のある下限を与え, それよりも小さくなる場合に推定分散をその下限と入れ替えることで 1, 2 が原因である過剰適合を回避する方法を提案する. さらに, 第 4 節では, 重み付き推定法 (Silverman (1985), Green and Silverman (1994) p 40-44) を用いることで 3 を原因とする過剰適合を回避する手法を提案する. 本論文では, その重みに Andrews (1974) で提案されている関数を適用した.

2. B-スプラインノンパラメトリック回帰モデル

2.1. モデル

説明変数と目的変数の組 (X, Y) に関して n 個のデータ $\{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$ が観測されたとする. ここでは, データは

$$y_i = \mu_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

であると仮定する. ただし ε_i はそれぞれ独立な確率変数で, 平均 $\mathrm{E}(\varepsilon_i) = 0$, 分散 $\mathrm{Var}(\varepsilon_i) = \sigma^2$ をもつ分布に従うとする. 多くの場合, この誤差項に正規性を仮定して議論を進めていくが, より広いクラスへの適用を考え, 誤差項の具体的な分布の仮定を外して議論を進める. このようなモデルに関して, 複雑な非線形構造を有するデータ分析に対して用いられるノンパラメトリック回帰モデルでは, 平均構造を十分滑らかな関数 g を用いて,

$$\mathrm{E}(Y_i \mid x_i) = \mu_i = g(x_i), \quad i = 1, 2, \ldots, n,$$

と仮定する.

B-スプラインを用いた非線形回帰モデルでは, μ を既知の基底関数の和としてとらえる. つまり, m 個の基底関数を用いた B-スプライン非線型回帰モデルでの平均構造は以下のようになる.

$$\mu_i = \sum_{j=1}^{m} a_j B_j(x_i), \quad (i = 1, 2, \ldots, n).$$

ただし, $B_j(x)$ は節点が決まると一意に決定される B-スプライン基底関数である. 例えば, 節点の幅が等しい場合の 3 次 B-スプライン基底関数は以下のように定義できる.

$$B_j(x) = B_0\{h, a + h(j-2), x\}, \qquad j = 1, \cdots, m,$$

ただし, $h = (b-a)/(m-3)$ $(m \geq 4)$, $a = \min_{i=1,\ldots,n}(x_i)$, $b = \max_{i=1,\ldots,n}(x_i)$. また, B_0 は x_0 に関する対称関数で,

$$B_0(h, x_0, x) = \begin{cases} \dfrac{1}{6h}\left\{\left(2 - \dfrac{|x-x_0|}{h}\right)^3 - 4\left(1 - \dfrac{|x-x_0|}{h}\right)^3\right\} & (|x-x_0| \leq h) \\ \dfrac{1}{6h}\left(2 - \dfrac{|x-x_0|}{h}\right)^3 & (h < |x-x_0| \leq 2h) \\ 0 & (\text{その他}) \end{cases}$$

以下, 本論文では等間隔配置の節点により決定される基底関数を取り扱う. 図 2 に 10 個の等間隔な節点で定義される基底関数の例をあげる. 他の次数での B-スプライン基底関数の構成法については 市田, 吉本 (1979, p 18-26) 等を参照していただきたい.

今,

$$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)' = \begin{pmatrix} B_1(x_1) & \cdots & B_m(x_1) \\ \vdots & \ddots & \vdots \\ B_1(x_n) & \cdots & B_m(x_n) \end{pmatrix},$$

とし, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)'$ とおくと, モデルは,

$$\boldsymbol{y} = B'\boldsymbol{a},$$

のように書きかえることができる. 以上のことにより, B-スプラインを用いた平滑化は, 係数 $\boldsymbol{a}$, σ^2 の線形モデルにおける推定問題に帰着されることがわかる.

2.2. 最適化に関するアルゴリズム

ある程度多くの基底関数を用いた場合, モデルのパラメータ $\boldsymbol{a}$ を従来の最小二乗法によって推定すると, 強度にデータに依存したモデルが推定される. そこで, この B-スプラインノンパラメトリック回帰モデルでは, 残差平方和に曲線の局所変動の程度を考慮した罰則付き残差平方和,

$$\begin{aligned} RSS_\lambda(\boldsymbol{a}) &= \sum_{i=1}^{n}(y_i - \boldsymbol{b}_i'\boldsymbol{a})^2 + \lambda \boldsymbol{a}' D_k' D_k \boldsymbol{a} \\ &= (\boldsymbol{y} - B\boldsymbol{a})'(\boldsymbol{y} - B\boldsymbol{a}) + \lambda \boldsymbol{a}' D_k' D_k \boldsymbol{a}, \end{aligned} \tag{2.1}$$

の最小化に基づいてパラメータを推定する. ここで, $\lambda\ (\geq 0)$ は平滑化パラメータと呼ばれる推定曲線の局所変動の程度を制御するパラメータであり, $\boldsymbol{a}' D_k' D_k \boldsymbol{a}$ は回帰曲線の変動に対する罰則項である. ただし D_k は k 階差分を与える $(m-k)\times m$ の行列で, ${}_kC_i$ を 2 項係数とすると以下のように定義される.

$$D_k = \begin{pmatrix} (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & 0 & \cdots & 0 \\ 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k \end{pmatrix}.$$

例えば, 2 階差分を与える $(m-2)\times m$ 行列は,

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix},$$

となる. ここで, λ が与えられたとき係数 $\boldsymbol{a}$ の推定量は,

$$\hat{\boldsymbol{a}}_\lambda = (B'B + \lambda D_k' D_k)^{-1} B'\boldsymbol{y}, \tag{2.2}$$

と書ける. また σ^2 の推定量は,

$$\begin{aligned} \hat{\sigma}_\lambda^2 &= \frac{1}{n - 2\mathrm{tr}(H_\lambda) + \mathrm{tr}(H_\lambda' H_\lambda)} \sum_{i=1}^{n}(y_i - \boldsymbol{b}_i'\hat{\boldsymbol{a}}_\lambda)^2 \\ &= \frac{1}{n - 2\mathrm{tr}(H_\lambda) + \mathrm{tr}(H_\lambda' H_\lambda)} \boldsymbol{y}'(I_n - H_\lambda)'(I_n - H_\lambda)\boldsymbol{y}, \end{aligned} \tag{2.3}$$

である. ただし H_λ はハット行列と呼ばれ, 以下のように定義される.

$$H_\lambda = B(B'B + \lambda D_k' D_k)^{-1} B'. \tag{2.4}$$

図 3 は平滑化パラメーターの変化による推定曲線の違いを示した図である. $\lambda = 0$ のときにデータに依存した結果になっていることがわかる.

最適な平滑化パラメーター λ, 基底関数の個数 m は, 与えられた λ, m のもとで $\hat{\boldsymbol{a}}_\lambda$, $\hat{\sigma}_\lambda^2$ を計算し, それをもとに情報量基準 $IC(m, \lambda)$ を求めその情報量の最小化により決定する. 以下その最適化アルゴリズムをあげる.

<u>m と λ の最適化アルゴリズム</u>

Step 1 : 基底関数の個数 (節点の個数) m を決める.

Step 2 : 固定した m のもとで $IC(m, \lambda)$ を最小にする λ を探す (図 4 参照).

Step 3 : m を変えて Step 2 を行い, 各 m ごとの最小 $IC(m, \lambda)$ を比較し, もっとも情報量を最小にする m, λ の組を最適な値とする.

正規性のもとでの罰則付き尤度にもとづく推定法の最適化アルゴリズムに関しては, 井元, 小西 (1999a, 1999b), Imoto (2001) 等を参照していただきたい. また Step 2 における最適な λ を探す方法として, Yanagihara and Ohtaki (2003) と同様に, 以下のようなスパイダーアルゴリズム (Ohtaki and Izumi (1999)) を用いて計算時間の短縮をはかった.

<u>スパイダーアルゴリズム</u>

Step 1 : τ_i を探索範囲とする. ここで基底関数の個数 m と平滑化パラメーターの初期値 λ_0 はあらかじめ決めておく.

Step 2 : $\tau_0 = \lambda_0$ とし $g_0 = \min\{\, IC(m, \lambda) \mid \lambda = \lambda_0 - \tau_0, \lambda_0, \lambda_0 + \tau_0 \}$ を計算する. このとき τ_1 と λ_1 を以下の出順に従って更新する.

(i) もし $g_0 = IC(m, \lambda_0)$ であれば $\tau_1 = \tau_0/2$, $\lambda_1 = \lambda_0$;

(*ii*) もし $g_0 = IC(m, \lambda_0 - \tau_0)$ であれば $\tau_1 = \tau_0/2$, $\lambda_1 = \lambda_0/2$;

(*iii*) もし $g_0 = IC(m, \lambda_0 + \tau_0)$ であれば $\tau_1 = 2\tau_0$, $\lambda_1 = \lambda_0 + \tau_0$.

Step 3 : τ_i と λ_i を i 回の繰り返しで得られる探索範囲と平滑化パラメーターとし, $g_i = \min\{\ IC(m, \lambda) \mid \lambda = \lambda_i - \tau_i, \lambda_i, \lambda_i + \tau_i\}$ とする. このとき τ_{i+1} と λ_{i+1} を以下の手順に従って更新する.

(*i*) もし $g_i = IC(m, \lambda_i)$ であれば $\tau_{i+1} = \tau_{i+1}/2$, $\lambda_{i+1} = \lambda_i$;

(*ii*) もし $g_i = IC(m, \lambda_i - \tau_i)$ であれば,

1. $\lambda_i - \tau_i = 0$ のとき $\tau_{i+1} = \tau_i/2$, $\lambda_{i+1} = \lambda_i/2$;
2. $\lambda_i - \tau_i \neq 0$ のとき $\tau_{i+1} = \tau_i$, $\lambda_{i+1} = \lambda_i - \tau_i$;

(*iii*) もし $g_i = IC(m, \lambda_i + \tau_i)$ であれば $\tau_{i+1} = 2\tau_i$, $\lambda_{i+1} = \lambda_i + \tau_i$.

Step 4 : 不等式 $|\hat{\sigma}^2_{\lambda_{i+1}} - \hat{\sigma}^2_{\lambda_i}|/\hat{\sigma}^2_{\lambda_i} < d$ を満たすまで Step 3 を繰り返す. このときの λ_{i+1} を最適な平滑化パラメーターとする.

2.3. m と λ を決定するための情報量基準

2.2 節で紹介したアルゴリズムで使用される情報量規準として様々なものが提案されているが, 本論文においてはMallows' C_p 基準やCross Validation 基準に代表される分布の仮定を必要としない, 予測残差平方和に基づく基準量を用いる. 本節ではそれらの情報量基準を紹介する.

(i) C_p 基準 :

Hastie and Tibshirani (1990, p 48) では B-スプラインノンパラメトリック回帰モデルにおける C_p 基準を以下のように定義している.

$$C_p(\lambda) = \sum_{i=1}^{n} \frac{(y_i - b_i'\hat{a}_\lambda)^2}{\hat{\sigma}^2_{\lambda_*}} + 2\mathrm{tr}(H_\lambda) - n. \tag{2.5}$$

ただし, ハット行列 H_λ は (2.4) によって定義され, フルモデルにおける分散の推定量に相当する $\hat{\sigma}^2_{\lambda_*}$ は, 与えられた λ_* に基づいて (2.3) によって与えられる σ^2 の推定量である.

通常 $\lambda_* = 0$ としてこの基準量を計算するが, C_p 基準に関して言えば, 基準化する分散の推定量はすべての基準量を通して共通であることが望ましいと考えられる. しかしながら B-スプラインノンパラメトリック回帰モデルでは, フルモデルの次数 (つまり最大の基底関数の個数) はいくらでも多くとることができ, 基底関数の個数の上限によって, 基準量の大きさが変化することになりあまり好ましくない. 以上のことから, この基準量には少し問題があると思われる.

(ii) Cross Validation 基準 :

$\hat{\boldsymbol{a}}_{\lambda,(-i)}$ を i 番目の標本の組 (x_i, y_i) を除いた標本に基づいて推定された推定量とする. このとき Cross Validation 基準は以下のように定義される.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{b}_i\hat{\boldsymbol{a}}_{\lambda,(-i)})^2. \tag{2.6}$$

h_{ii} をハット行列 H_λ の第 i 番目の対角成分とすると, Green and Silverman (1994) より $CV(\lambda)$ は以下の様に書きかえることができる.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{b}_i'\hat{\boldsymbol{a}}_\lambda}{1 - h_{ii}}\right)^2. \tag{2.7}$$

(iii) Generalized Cross Validation 基準 :

Craven and Wahba (1979) では Generalized Cross Validation 基準は以下のように定義されている.

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \boldsymbol{b}_i'\hat{\boldsymbol{a}}_\lambda}{1 - \mathrm{tr}(H_\lambda)/n}\right\}^2. \tag{2.8}$$

Green and Silverman (1994) によれば, $GCV(\lambda)$ は以下のように書きかえることができる.

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1 - h_{ii}}{1 - \mathrm{tr}(H_\lambda)/n}\right)^2\left(y_i - \boldsymbol{b}_i\hat{\boldsymbol{a}}_{\lambda,(-i)}\right)^2\right\}.$$

つまり $GCV(\lambda)$ は (2.6) 式における $CV(\lambda)$ に対して重み付き平均を取っているとみなすことができる.

3. 過剰適合を回避させるための情報量基準

B-スプラインノンパラメトリック回帰モデルを用いた平滑化は, 基底関数の個数の増加に伴いそのモデルが柔軟になり, 過剰適合を起こすことがある. これは推定分散が小さくなりすぎてしまい, 従来のバイアス補正項ではうまく補正しきれなくなることに原因があると思われる. このような状況を回避するために, 分散の推定値に意味のある下限を与えることで過剰適合を回避する手法を提案する. 本節では, そのような下限として局所線形適合に基づいた推定量 (Gasser, Sroka and Jennen-Steinmetz (1986)) を用いる. もちろん, その他にも局所線形適合を用いた推定量もあるが (Ohtaki (1990) 等参照), 今回は単純な形でありかつ漸近正規性も保証されているこの推定量を選んだ. 今, 説明変数 x_i は $x_1 < x_2 < \cdots < x_n$ といった大小関係が成立しているとする. このとき局所線形適合に基づいた分散の推定量は,

$$S_\varepsilon^2 = \frac{1}{n-2}\sum_{i=2}^{n-1} c_i^2 \tilde{\varepsilon}_i^2, \tag{3.1}$$

である. ただし,

$$\begin{aligned}\tilde{\varepsilon}_i &= \frac{x_{i+1}-x_i}{x_{i+1}-x_{i-1}}y_{i-1} + \frac{x_i - x_{i-1}}{x_{i+1}-x_{i-1}}y_{i+1} - y_i \\ &= p_i y_{i-1} + q_i y_{i+1} - y_i, \quad (2 \le i \le n-1),\end{aligned}$$

であり, $c_i^2 = (p_i^2 + q_i^2 + 1)^{-1}$. また $g(\cdot)$ が連続で ε_i の 4 次モーメントが存在し, かつ $|x_i - x_{i-1}| = \mathrm{O}(n^{-1})$ であれば, $\mathrm{E}(S_\varepsilon^2) = \sigma^2 + \mathrm{O}(n^{-2})$ である. 更に $g(\cdot)$ が 2 回微分可能であれば, $\mathrm{E}(S_\varepsilon^2) = \sigma^2 + \mathrm{O}(n^{-4})$ となる. 図 5 に局所線形適合に関する残差の図解を示す. もし x_i の値が重複している, $x_{i-1} = x_i = x_{i+1}$ であれば $p_i = 1/2$, $q_i = 1/2$, また $x_{i-k} < x_{i-k+1} = \cdots = x_i < x_{i+1}$ であれば, $p_i = (x_{i+1} - x_i)/(x_{i+1} - x_{i-k})$, $q_i = (x_i - x_{i-k})/(x_{i+1} - x_{i-k})$ と変形する.

この局所線形フィットはある意味で真のモデルを含んだフルモデルに対応していると考えられるので, この推定量を C_p 基準におけるフルモデルの下での分散の推定量とし, 以下のような基準量を提案する.

$$C_p(\lambda, S_\varepsilon^2) = \frac{\widetilde{RSS}(\hat{a}_\lambda)}{S_\varepsilon^2} + 2\mathrm{tr}(H_\lambda) - n. \tag{3.2}$$

ただし,

$$\widetilde{RSS}(\hat{a}) = \begin{cases} \sum_{i=1}^{n}(y_i - b_i'\hat{a})^2 & (\tilde{\sigma}^2 > S_\varepsilon^2) \\ nS_\varepsilon^2 & (\tilde{\sigma}^2 \leq S_\varepsilon^2) \end{cases}, \quad \left(\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_i'\hat{a})^2\right).$$

図 6 にこの基準量を最適化に用いることで過剰適合を回避した例を示す. 図左側が $C_p(\lambda, S_\varepsilon^2)$ を用いたときの結果であり, 右側が従来の基準量 $GCV(\lambda)$ を用いたときの結果である. また 1 段目はそれぞれの基準量によって選ばれた最適な m, λ をもちいた平滑化結果であり, 2, 3 段目はそれぞれ基底関数の個数に対する情報量, 平滑化パラメーターの動きである. この図から $C_p(\lambda, S_\varepsilon^2)$ を最適化に用いることで基底関数の増加による情報量の変動が少なくなり, 過剰適合が回避されていることがわかる.

バイアス項の計算は以下のように行う. $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ とおく. 今, U_i を互いに独立で, Y_i と独立にかつ同一な分布に従う確率変数とすると, リスクは以下のようになる.

$$R_p = \mathrm{E}_U\mathrm{E}_Y\left[\sum_{i=1}^{n}\frac{(U_i - b_i'\hat{a})^2}{\sigma^2}\right] = n(1+\theta) + \mathrm{tr}(H_\lambda' H_\lambda).$$

ただし,

$$\theta = \frac{1}{n\sigma^2}\boldsymbol{\mu}'(I_n - H_\lambda)'(I_n - H_\lambda)\boldsymbol{\mu}.$$

この θ は非負の値をとり, モデルが真のモデルを含んでいるという条件, つまり $H_\lambda\boldsymbol{\mu} = \boldsymbol{\mu}$ (この条件については Hurvich, Simonoff and Tsai (1998) を参照), の下で 0 になる. また Gasser, Sroka and Jennen-Steinmetz (1986) によれば S_ε^2 には適当な条件のもとで漸近正規性が成り立つ. つまり,

$$\sqrt{n}(S_\varepsilon^2 - \sigma^2) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \zeta^2(\sigma^2, \kappa_4)), \quad (n \to \infty).$$

ただし $\zeta^2(\sigma^2,\kappa_4)$ は ε_i の分散 σ^2 と 4 次キュムラント κ_4 に依存する係数である. この性質から, $\sqrt{n}(S_\varepsilon^2-\sigma^2)=\zeta(\sigma^2,\kappa_4)Z$ とおくと,

$$\begin{aligned}\mathrm{E}\left[\sum_{i=1}^{n}\frac{(Y_i-\boldsymbol{b}_i'\hat{\boldsymbol{a}})^2}{S_\varepsilon^2}\right]&=\mathrm{E}\left[\sum_{i=1}^{n}\frac{(Y_i-\boldsymbol{b}_i'\hat{\boldsymbol{a}})^2}{\sigma^2}\right.\\&\qquad\left.\times\left\{1-\frac{\zeta(\sigma^2,\kappa_4)Z}{\sigma^2\sqrt{n}}+\frac{\zeta(\sigma^2,\kappa_4)^2Z^2}{\sigma^4 n}+\mathrm{O}_p(n^{-3/2})\right\}\right]\\&=n(1+\theta)-2\mathrm{tr}(H_\lambda)+\mathrm{tr}(H_\lambda'H_\lambda)+R(H_\lambda,\kappa_3,\kappa_4)+\mathrm{o}(1),\end{aligned}$$

ただし, $R(H_\lambda,\kappa_3,\kappa_4)$ は ε の 3, 4 次キュムラントとハット行列 H_λ に依存する定数である. この $R(H_\lambda,\kappa_3,\kappa_4)$ は, 回帰モデルの場合小さいもので, ほんどの場合には正の値をとる (藤越, 栁原, 若木 (2002)). つまり $R(H_\lambda,\kappa_3,\kappa_4)$ をバイアスとして考慮に入れない場合, 基準量はリスクに対して過大推定をする傾向が強くなる. 本論文で取り扱う場合では, 過剰適合の回避という目的に対してはより大きなバイアス補正項が必要となるので, バイアス項を過大推定している項は補正しない方が望ましい. 以上により, $R(H_\lambda,\kappa_3,\kappa_4)$ を省略すると, バイアスの第一項は $2\mathrm{tr}(H_\lambda)$ になることがわかる. 実際には, 推定分散を入れ替えるという作業を行っているため, 上記の期待値の計算は厳密な意味では正しくない. しかしながら, θ は真のモデルと候補のモデルのずれに関する非心パラメーター (Fujikoshi and Satoh (1997)) であり, $\mathrm{E}(\tilde{\sigma}^2/S_\varepsilon^2)=1+\theta+\mathrm{o}(1)$ であるので, $\tilde{\sigma}^2/S_\varepsilon^2-1$ はその θ の推定量として考えることができる. θ は非負であるため

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\boldsymbol{b}_i'\hat{\boldsymbol{a}}_\lambda)^2}{S_\varepsilon^2}\geq 1,$$

とする制限は妥当であると思われる.

図 7 に下限を下回った場合に分散の推定量を入れ替えた場合と入れ替えなかった場合の結果を示す. 図左側が分散の推定量を入れ替えた場合の結果であり, 右側が入れ替えなかった場合の結果である. また 1 段目はそれぞれの基準量によって選ばれた最適な m, λ をもちいた平滑化結果であり, 2, 3 段目はそれぞれ基底関数の個数に対する情報量, 平滑化パラメーターの動きである. 図より分散の推定量を入れ替えることにより情報量の動きが小さくなっていることがわかる. それにより

過剰適合を回避していることが見てとれる.

図 8 は固定された基底関数の個数の下で最適化された λ に基づく推定分散とその基底関数の個数の関係を表した図である. これをみると, 基底関数の個数が増えると, 分散が過小評価されていることがわかる. また分散の推定量を入れ替えることにより, 小さい値の平滑化パラメーターを最適な値として選ばなくなり, 分散の過小評価もなくなっている. この例の場合では, 推定分散を下限値と入れ替えなかった場合, 情報量基準が最小になるのは $m = 25$ のときである. このときの分散の推定量と情報量の動きを図 9 に示す. これをみると分散を過小評価したために情報量の動きがおかしくなり, 小さい平滑化パラメーターを最適なものであると判断している. 分散の推定量を入れ替えるということで, この平滑化パラメーターの過小評価を回避していることになる.

この節を終える前に, シミュレーション実験により, $C_p(\lambda, S_\epsilon^2)$ を用いた最適化アルゴリズムが過剰適合を回避していることを確かめる. 誤差分布としては,

(1) N(0, 1) ($\kappa_3 = 0$, $\kappa_4 = 0$),

(2) 自由度 5 の t 分布 ($\kappa_3 = 0$, $\kappa_4 = 6$),

(3) U(−5, 5) ($\kappa_3 = 0$, $\kappa_4 = -1.2$),

(4) 自由度 2 の χ^2 分布 ($\kappa_3 = 2$, $\kappa_4 = 6$),

(5) LN(0, 1/4) ($\kappa_3 = (e^{1/4}+2)(e^{1/4}-1)^{1/2} \approx 1.75$, $\kappa_4 = e+2e^{3/4}+3e^{1/2}-6 \approx 5.90$),

の 5 つの分布を平均 0, 分散 2 に基準化したものを使った. 平均のトレンドとしては,

$$g(x) = 35\left\{\frac{x}{6} - \left(\frac{x}{6}\right)^2 - \left(\frac{x}{6}\right)^3\right\}\exp\left\{-\left(\frac{x}{6}\right)^2\right\},$$

を使った. 説明変数は x_i は 0 ～ 25 までの範囲の一様乱数で発生させ, すべての繰り返しにおいて同一なものを使った. そのように発生させたデータに対し平滑化

を行い, MSE を以下のように計算した.

$$MSE = \frac{1}{50}\sum_{j=1}^{50}\left\{g(z_j) - \tilde{b}_j\hat{a}_\lambda\right\}^2,$$

ただし,

$$z_j = x_1 + \frac{1}{49}(x_n - x_1)(j-1),$$

であり, $\tilde{b}_j$ は z_j に基づく基底関数ベクトルである. 図 10, 11 はそのようなデータそれぞれ $n = 20$, $n = 50$ で 100 回繰り返し発生させ, それぞれの MSE を箱ひげ図にしたものである. $n = 20$ のときの基底関数の個数を $m = 16$, $n = 50$ のときは $m = 30$ を上限値とし, 最適化を行った. 基底関数の個数と平滑化パラメーターの決定にために使用した情報量基準は $C_p(\lambda, S_\varepsilon^2)$, $C_p(\lambda)$, $CV(\lambda)$, $GCV(\lambda)$ である. また表 1 はそれぞれの繰り返しにおける MSE の算術平均である. 表 1 と図 10, 11 から, $C_p(\lambda, S_\varepsilon^2)$ を用いた平滑化法が最も過剰適合を回避していることがわかる. また標本数を増やしても, $C_p(\lambda)$ や $CV(\lambda)$ を用いた平滑化法では, 過剰適合の回数は減ってないのように思われる. これは標本数が多いからといって, 安易に基底関数の個数を増やしてはいけないということを意味する.

4. M-推定を用いた平滑化法

次に M-推定法を用いた外れ値に対して頑健な手法を考察する. このような手法を用いることで, 3 が原因であるような過剰適合を回避することができると考える. M-推定に基づく頑健な推定方法は, Sinha and Schunck (1992), Shi and Li (1995), Shi and Zhang (1995), Karczewicz and Gabbouj (1998) らによって特に画像処理の分野で熱心に研究されている. しかしながらこれらの手法は, 最適化アルゴリズムに複雑な手法を用いていたりと, まだまだ改良の余地があると思われる. 本節では簡単なアルゴリズムによって最適化可能である頑健な手法を提案する.

今回, 未知パラメーターの推定に関して, 罰則付き残差平方和 RSS_λ (2.1) ではなく, Silverman (1985) や Green and Silverman (1994), p 40-44, 等で紹介されて

いる罰則付き重み付き残差平方和 $WRSS_\lambda$,

$$WRSS_\lambda(\boldsymbol{a}) = \sum_{i=1}^{n} w_i(y_i - \boldsymbol{b}_i'\boldsymbol{a})^2 + \lambda \boldsymbol{a}' D_k' D_k \boldsymbol{a}$$
$$= (\boldsymbol{y} - B\boldsymbol{a})' W (\boldsymbol{y} - B\boldsymbol{a}) + \lambda \boldsymbol{a}' D_k' D_k \boldsymbol{a}, \qquad (4.3)$$

を最小にすることで推定量を構成する. ただし $W = \mathrm{daig}(w_1, w_2, \ldots, w_n)$. この重みは何でも構わないが, 本論文では, Andrews (1974) で提案された重み関数を適用する. つまり, $r_i = y_i - \boldsymbol{b}_i'\boldsymbol{a}$ $(1 \leq i \leq n)$ とし,

$$r_m = \underset{i=1,\ldots,n}{\mathrm{median}}(|r_i|)$$

とすると, 重みは以下のように決定される.

$$w_i = \begin{cases} \sin\left(\dfrac{r_i}{cr_m}\right)/r_i & \left(\left|\dfrac{r_i}{r_m}\right| \leq c\pi\right) \\ 0 & (\text{その他}) \end{cases} \qquad (1 \leq i \leq n), \qquad (4.4)$$

ただし c は与えられた定数であり, Andrews (1974) では $c = 1.5$ をとることを推奨している. 図 12 に重みを与える関数の形を示しておく. これらの重みを $w_i = \rho'(r_i/r_m)$ としたとき, (4.3) は, 以下のように書きかえることができ, この重み付き推定は罰則付き M-推定に基づく推定方法である事がわかる.

$$WRSS_\lambda(\boldsymbol{a}) = \sum_{i=1}^{n} \rho(y_i - \boldsymbol{b}_i'\boldsymbol{a}) + \lambda \boldsymbol{a}' D_k' D_k \boldsymbol{a}.$$

実際の最適化には以下のアルゴリズムを用いる.

<u>ロバスト推定に関するアルゴリズム</u>

Step 1 : 係数 c を決定する (奨励は $c = 1.5$).

Step 2 : 基底関数の個数 m を決める.

Step 3 : 平滑化パラメーター λ をあらかじめ与えておく.

Step 4 : 初期値 $\hat{\boldsymbol{a}}_{\lambda,(0)} = (B'B)^{-1}B'\boldsymbol{y}$ を計算し, それをもとに $\hat{\sigma}^2_{\lambda,(0)}$ を求める.

Step 5 : $\hat{r}_i^{(l-1)} = y_i - \boldsymbol{b}_i'\hat{\boldsymbol{a}}_{\lambda,(l-1)}$ $(1 \leq i \leq n,\ l = 1, 2, \cdots)$,

$$\hat{r}_m^{(l-1)} = \underset{i=1,\ldots,n}{\text{median}} \left(\left| \hat{r}_i^{(l-1)} \right| \right),$$

として, 重み

$$\hat{w}_i^{(l-1)} = \begin{cases} \sin\left(\dfrac{\hat{r}_i^{(l-1)}}{c\hat{r}_m^{(l-1)}} \right) / \hat{r}_i^{(l-1)} & \left(\left| \dfrac{\hat{r}_i^{(l-1)}}{\hat{r}_m^{(l-1)}} \right| \leq c\pi \right) \\ 0 & (\text{その他}) \end{cases},$$

を更新する.

Step 6 : $\widehat{W}_{(l-1)} = \text{daig}(\hat{w}_1^{(l-1)}, \hat{w}_2^{(l-1)}, \ldots, \hat{w}_n^{(l-1)})$ とし, 係数 $\hat{\boldsymbol{a}}_{\lambda,(l)}$ を

$$\hat{\boldsymbol{a}}_{\lambda,(l)} = (B'\widehat{W}_{(l-1)}B + \lambda D_k'D_k)^{-1}B'\widehat{W}_{(l-1)}\boldsymbol{y},$$

で更新し, $\hat{\sigma}^2_{\lambda,(l)}$ を計算する.

Step 7 : 不等式 $|\hat{\sigma}^2_{\lambda,(l)} - \hat{\sigma}^2_{\lambda,(l-1)}|/\hat{\sigma}^2_{\lambda,(l-1)} < d$ を満たすまで Step 5, 6 を繰り返す. このときの $\hat{\sigma}^2_{\lambda,(l)}$ と $\hat{\boldsymbol{a}}_{\lambda,(l)}$ を推定量とする.

Step 8 $\hat{\sigma}^2_{\lambda,(l)}$ または $\hat{\boldsymbol{a}}_{\lambda,(l)}$ をもちいて情報量基準 $IC(m,\lambda)$ を計算する.

Step 9 : m と λ を変えて Step 2 ～ 8 を反復させ $IC(m,\lambda)$ が最小になる m, λ を最適な値とする.

この重み付き推定に関して, ハット行列は

$$H_\lambda = B(B'WB + \lambda D_k'D_k)^{-1}B'W,$$

となっていることに注意する.

図 13 にこの頑健な手法を用いて得た平滑化結果を示す. 3 節と同様に, 固定した m の下での λ の最適化にはスパイダーアルゴリズム (Ohtaki and Izumu (1999)) を用いた. 図の左側は従来の B-スプラインノンパラメトリック回帰モデルに基づいた平滑化結果であり, 右側は今回提案した頑健な手法を用いて得た平滑化結果である. 上段, 下段は $C_p(\lambda, S_\varepsilon^2)$, $GCV(\lambda)$ を用いて得た最適な m, λ によって得られた結果である. これを見ると従来の手法に較べて, 今回提案した手法は外れ値の影響が小さくなっていることがわかる.

5. おわりに

本論文では, B-スプラインノンパラメトリック回帰モデルを用いた平滑化における過剰適合を, 基底関数の個数と平滑化パラメーターを決定する情報量基準を改良し, また更に M-推定法を用いることで回避する方法を提案した. B-スプラインを用いた平滑化法は, そのモデルが柔軟なためにしばしば過剰適合を起こす. 平滑化が最終目的であれば, 微妙なチューニングを何度も施し, 過剰適合していない結果を得る事もできるが, 本来このような平滑化はオートマチックに基底関数の個数と平滑化パラメーターを選び, 最適化を行うものであると我々は認識している. そのため, 基底関数の個数を増やせば過剰適合を起こすような手法だと非常に問題がある. 何故ならば, 基底関数の個数の上限値は標本数に依存させる必要があるからである. どのような標本数でも一定の上限値であれば, 少ないときには柔らかすぎ, 多いときには硬すぎる結果を与えてしまうからである. 実際には誤差の分散にも依存させる必要があるが, 誤差の分散は未知であり, 推定量はモデルに依存しているため, 誤差分散を考慮に入れた基底関数の上限値を構成することは難しい. そのため, 標本数だけ依存した個数の上限値を使わざるを得ない. 以上のことから, 推定分散に何らかの意味のある下限値を起き, 制約を与えることは, 例え理論上の特性が少なからず崩れようとも, 実用上非常に意味のあることであると考える. 特に Gasser, Sroka and Jennen-Steinmetz (1986) での誤差分散の推定量は, 一致推定量であり, その下限としては意味のあるものであると思われる.

従来の手法をそのままオートマチックに使うと過剰適合を起こす危険がある. それを回避するためのには何らかの制約を入れないといけない. 我々が提案したものは簡単であるため, 実用的な手法となりうると考える.

参考文献

Andrews D. F. (1974): A robust method for multiple linear regression. *Technometrics*, **16**, 523--531.

Breiman L. and Friedman J. H. (1985): Estimating optimal transformations for multiple regression and correlation (with comment). *J. Amer. Statist. Assoc.*, **80**, 580–599.

de Boor C. and Rice J. R. (1968): Least squares cubic spline approximation I-fixed knots and II-fixed knots. *Purdue Univ. Reports*, CSD TR 20 and 21.

Craven P. and Wahba G. (1979): Smoothing noisy data with spline functions. *Numerishe Mathematik*, **31**, 377–403.

Friedman J. H. (1991): Multivariate adaptive regression splines. With discussion and a rejoinder by the author. *Ann. Statist.*, **19**, 1–141.

Friedman J. H. and Tukey J. W. (1974): A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **23**, 881–890.

Fujikoshi Y. and Satoh K. (1997): Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707-716.

藤越康祝・柳原宏和・若木 宏文 (2002): 多変量非正規回帰モデルにおける変数選択について. 2002 年度統計関連学会連合大会講演報告集, 405–406.

Gasser T., Sroka L. and Jennen-Steinmetz C. (1986): Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

Green P. J. and Silverman B. W. (1994): *Nonparametric Regression and Generalized Linear Models.* Chapman & Hall, London.

Hastie T. J. and Tibshirani R. J. (1990): *Generalized Additive Models.* Chapman & Hall, London.

Hurvich C. M., Simonoff J. S. and Tsai C. L. (1998): Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy Statist. Soc. Ser.* B, **60** , 271–293.

市田浩三・吉本富士市 (1979): スプライン関数とその応用. 教育出版.

Imoto, S. (2001): *B-Spline Nonparametric Regression Models and Information Criteria.* Ph. D. thesis. Kyushu University.

井元清哉・小西貞則 (1999a): 情報量規準に基づく B-スプライン非線型回帰モデルの推定. 応用統計学, **28**, 137–150.

井元清哉・小西貞則 (1999b): B-スプラインによる非線型回帰モデルと情報量規準. 統計数理, **47**, 359–373.

Karczewicz M. and Gabbouj M. (1998): Robust B-spline image modeling with application to image processing. *IEEE Trans. Image Process.*, **7**, 912–917.

Ohtaki M. (1990): Some estimators of covariance matrix in multivariate nonparametric regression and their applications. *Hiroshima Math. J.*, **20**, 63–91.

Ohtaki M. and Izumi S. (1999): Globally convergent algorithm without derivatives for maximizing a multivariate function. 複雑非線形現象の統計理論の開発と応用～ランダム変動を伴う非線形データ構造の探索及び解析研究会～予稿集.

大瀧慈・川崎裕美・佐藤健一・栁原宏和・山口直人 (2000): ノンパラメトリック平滑化処理による市区町村別 SMR 疾病地図アニメーションの作製. 第 68 回日本統計学会予稿集, 261–262.

Satoh K., Yanagihara H. and Ohtaki M. (2003): Bridging the gap between B-spline and polynomial regression model. *Comm. Statist. Simulation Comput.*, **32**, 179–190.

Shi P. and Li G. (1995): Global convergence rates of B-spline M-estimators in nonparametric regression. *Statist. Sinica*, **5**, 303–318.

Shi P. and Zhang Z. (1995): Robust nonparametric regression based on L_1-norm and B-splines. *Systems Sci. Math. Sci.*, **8**, 187–192.

Silverman B. W. (1985): Some aspects of spline smoothing approach to non-parametric regression curve fitting. *J. Roy Statist. Soc. Ser.* **B**, **47**, 1–52.

Sinha S. S. and Schunck B. S. (1992): A two-stage algorithm for discontinuity-preserving surface reconstruction. *IEEE Trans. Pattern Anal. Machine Intell.*, **14**, 36–55.

Yanagihara H. and Ohtaki M. (2003): Knot-placement to avoid over fitting in B-spline scedastic smoothing. *Comm. Statist. Simulation Comput.*, **32** (in print).

吉本富士市・市田浩三・清野武 (1977): 区分的 3 次関数を用いた 2 次元データの平滑化の自動的方法. 情報処理, **18**, 128–134.

表 1: それぞれの誤差分布での MSE の算術平均

n	誤差分布	MSE $C_p(\lambda, S_\varepsilon^2)$	$C_p(\lambda)$	$CV(\lambda)$	$GCV(\lambda)$
20	Normal	3.15	15.51	5.17	8.75
	t	3.49	15.54	3.07	9.23
	Uniform	4.20	15.25	6.76	9.29
	Chi-squared	3.05	15.24	4.01	7.62
	Log-normal	4.22	14.91	5.18	9.62
50	Normal	0.79	14.96	6.15	1.95
	t	0.65	14.86	4.79	2.62
	Uniform	0.64	15.08	7.48	0.93
	Chi-squared	0.63	15.27	7.07	1.35
	Log-normal	0.60	15.40	4.56	1.49

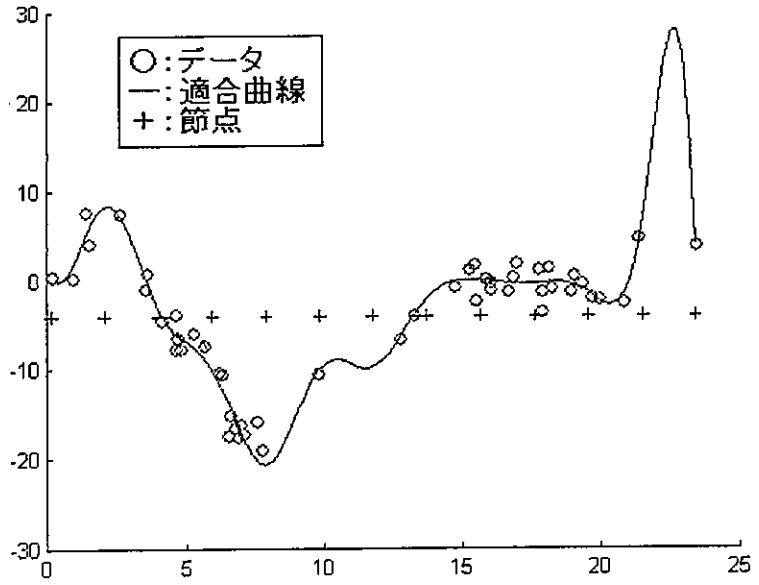

図 1: B-スプラインノンパラメトリック回帰モデルでの過剰適合の例

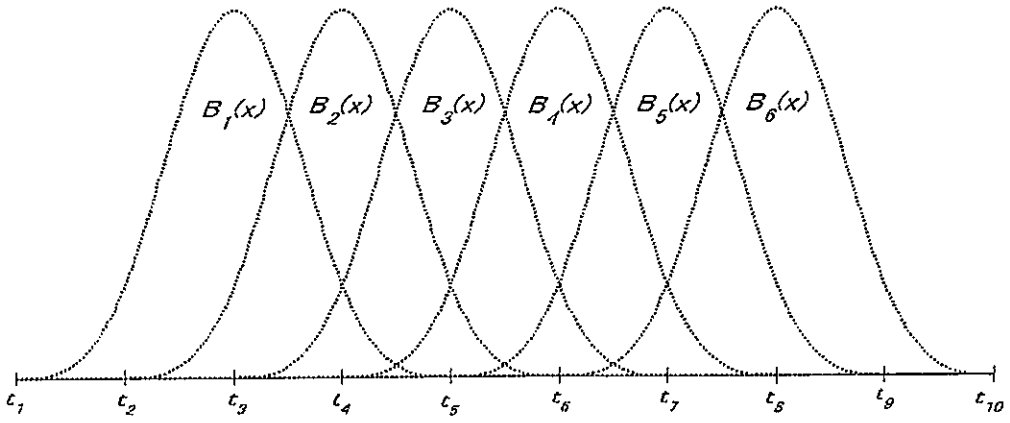

図 2: 10 個の節点 (等間隔配置) を用いた場合の基底関数

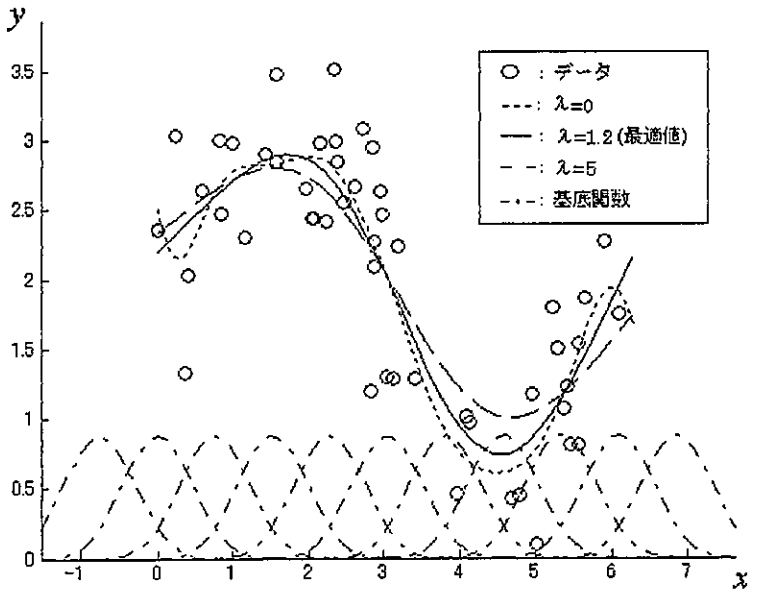

図 3: それぞれの平滑化パラメーターでの結果と基底関数

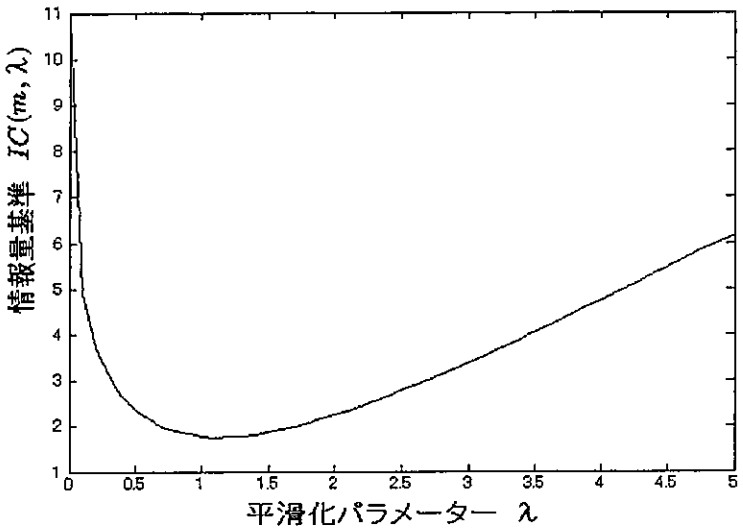

図 4: m を固定したときの平滑化パラメーターの変化に対する情報量基準の変化

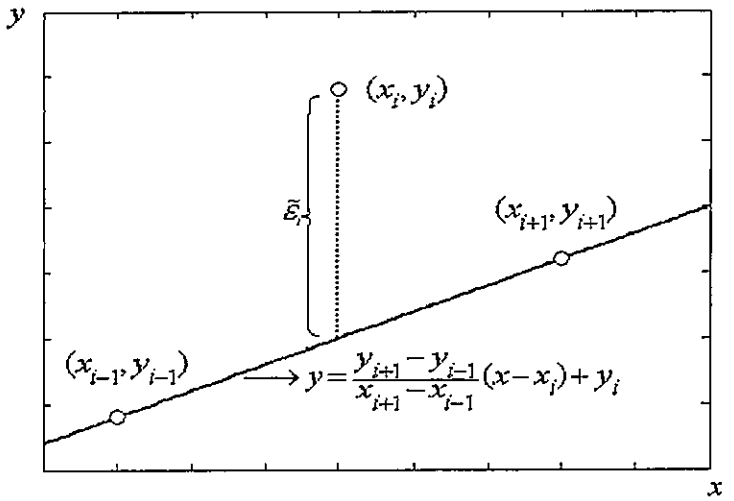

図 5: 局所線形適合における残差

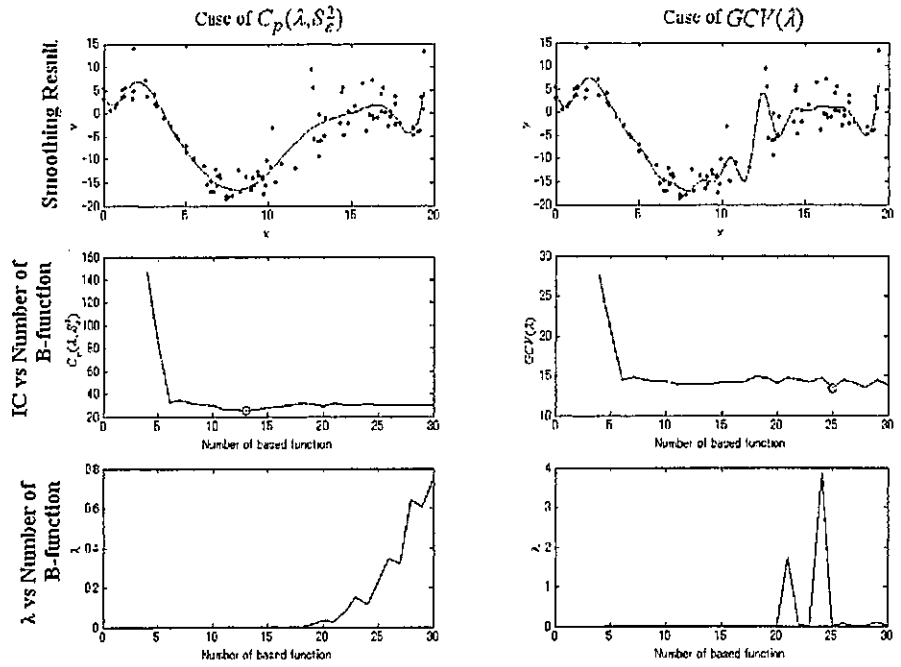

図 6: $C_p(\lambda, S_\varepsilon^2)$ を使用して過剰適合を回避した例

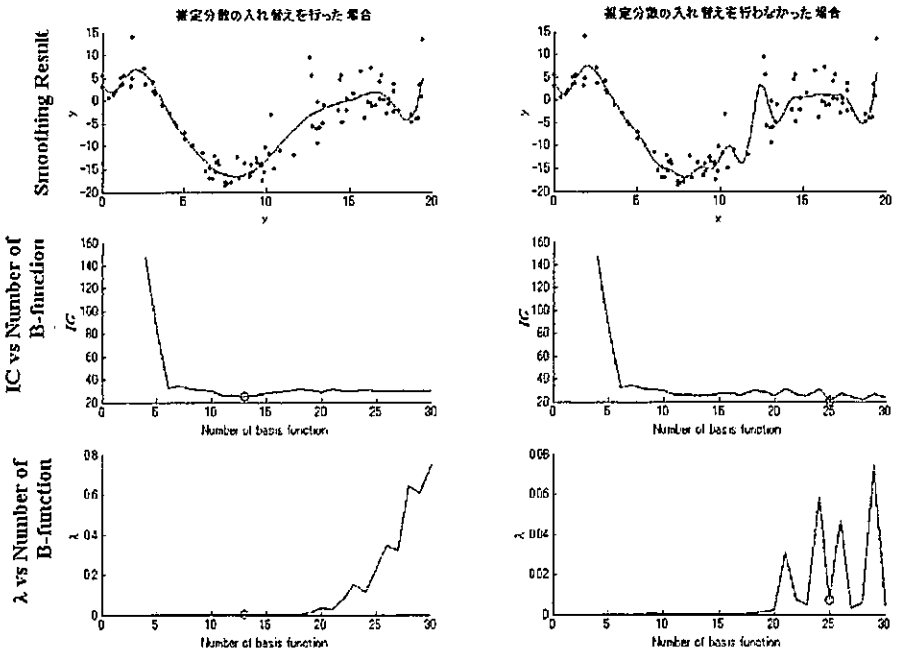

図 7: 下限値に関する入れ替えをした場合としなかった場合の比較

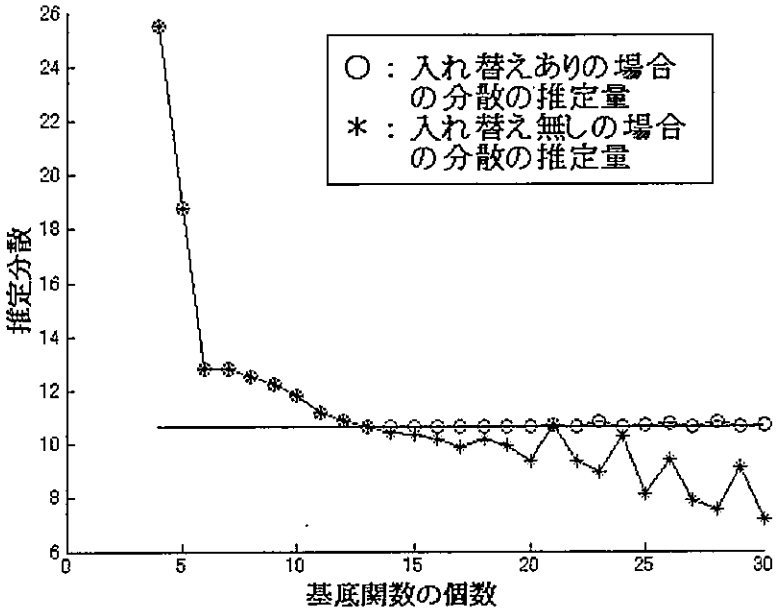

図 8: 入れ替えありの場合と無しの場合の分散の推定量の動き

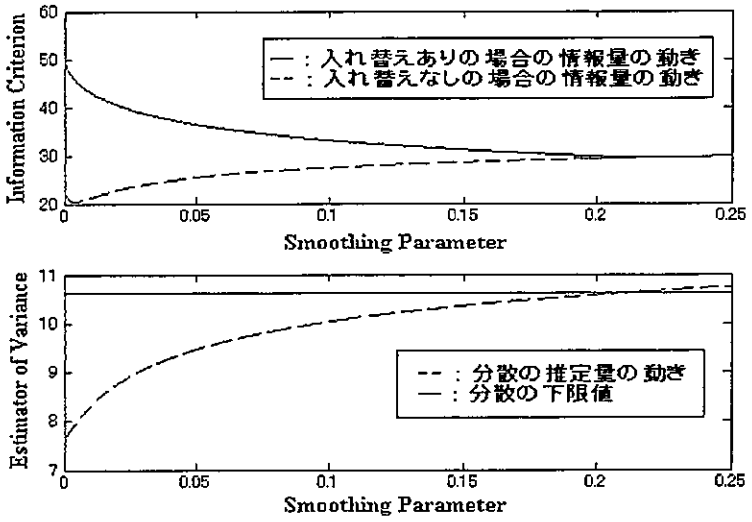

図 9: 分散の推定量の動きと情報量基準の大きさ

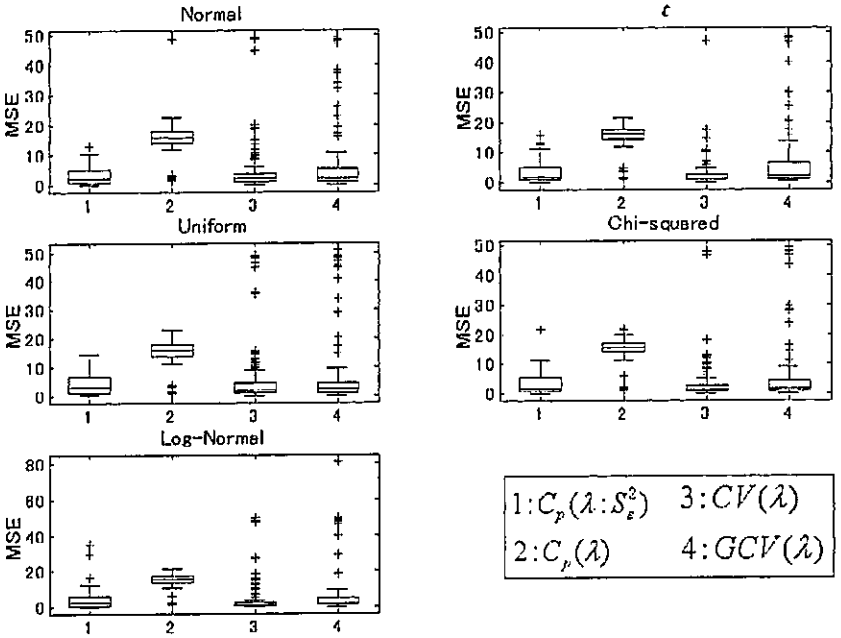

図 10: $n = 20$ の場合でのそれぞれの誤差分布における MSE

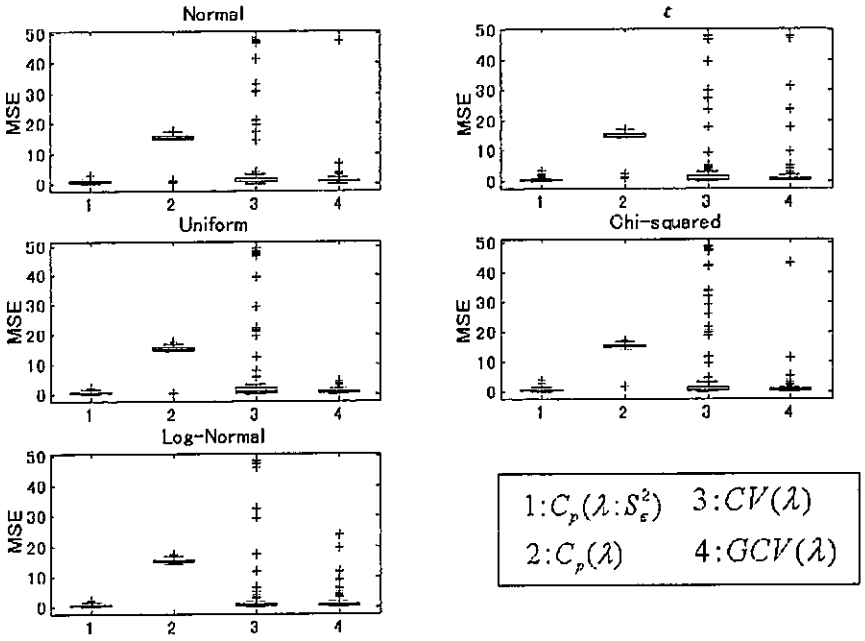

図 11: $n = 50$ の場合でのそれぞれの誤差分布における MSE

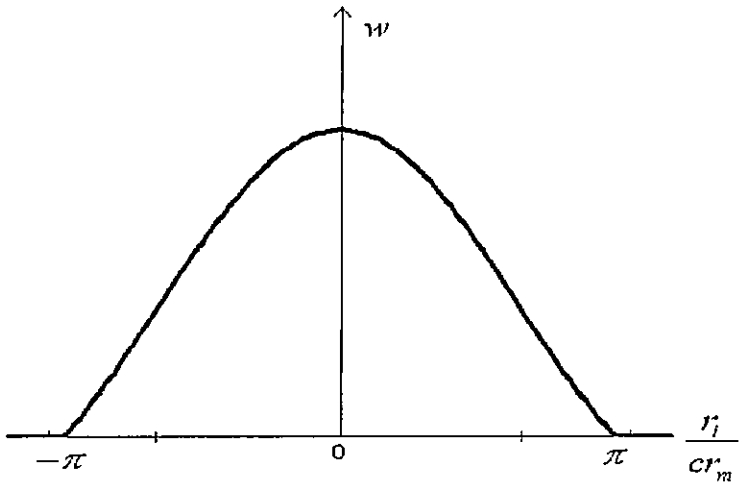

図 12: Andrews の重み関数

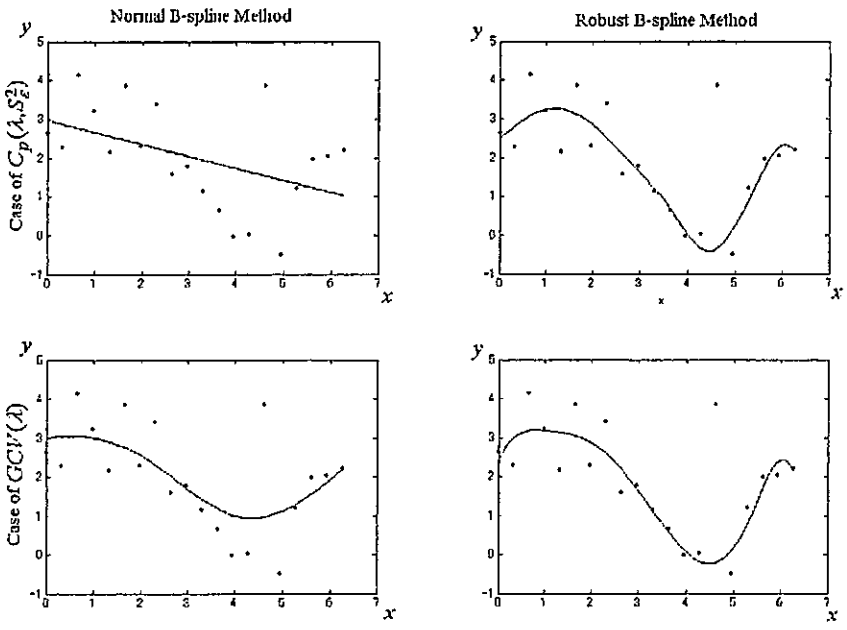

図 13: 重み付き平滑化法と重み付きでない平滑化法との比較