

B-スプラインによる不連続曲線推定と変化点探索

九州大・数理 井元 清哉

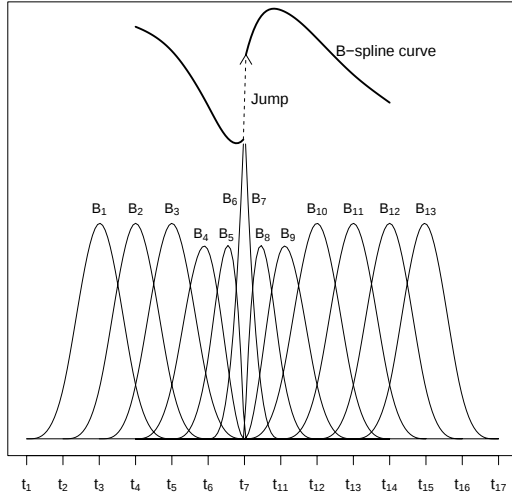
九州大・数理 小西 貞則

1. はじめに

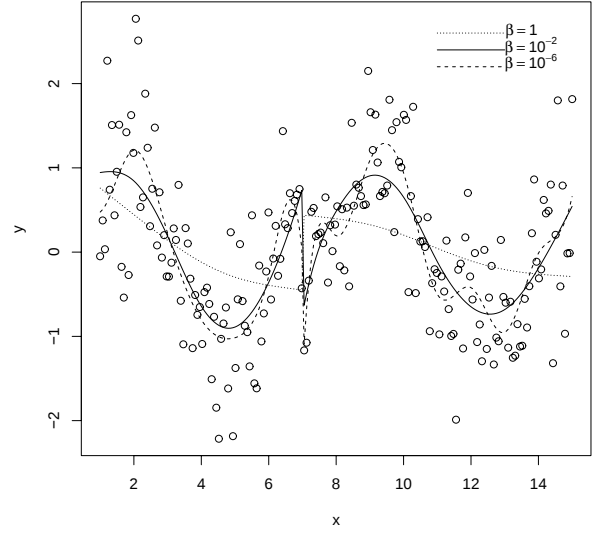
説明変数 X と目的変数 Y に関して n 組のデータ $\{(x_\alpha, y_\alpha); \alpha = 1, \dots, n\}$ が得られたとする．このとき，目的変数の平均構造 $\mu_\alpha = E[Y_\alpha | x_\alpha]$ の推定に当たっては，スプラインや核関数に基づく非線形回帰モデルが主に用いられる．しかし，ある点を境に平均構造や誤差分散が突然変化するような場合，非線形回帰モデルの当てはめは失敗し，データから有効な情報が得られないことが多い．従って，変化点を適切に推定することは，非線形回帰モデルからの情報抽出において重要な役割を担い，また，変化点自体が現象解明において貴重な情報となりうる．

非線形回帰に基づく変化点推定のこれまでの研究としては，Koo (1997)，Müller (1992)，Loader (1996) が挙げられる．Koo (1997) は，1 次スプラインに基づく非線形回帰モデルを用い，ベイズ型情報量規準 BIC に基づいて不連続点を推定する方法を提案している．しかし，1 次スプライン (折れ線) を用いているため，複雑な非線形構造の探索に対して十分な結果を得ることは難しい．Müller (1992)，Loader (1996) は，核関数に基づく非線形回帰モデルを用い，ジャンプの幅が最大となるように不連続点を推定する方法を提案している．しかし，データの非線形構造を有効に捉えるためには，核関数のバンド幅を適切に選択する必要があり，この点に問題が残る．また，これらの方法はかなり複雑な実行手順を必要とする．

本報告では， B -スプライン (de Boor, 1978) に基づく非線形回帰モデル (井元・小西, 1999) を拡張することで，変化点を推定しデータに内在する複雑な非線形構造を推定する方法を提案する．平均構造の突然の変化は， B -スプラインの節点を重複させることにより捉えることが可能となる．このとき，重複節点の位置の選択が変化点推定の本質的問題となり，情報量規準に基づく選択方法を提案する．また，誤差分散の変化を捉えるための非線形回帰モデルを構成し，そのモデリングについても考察する．提案する方法は，実データ・人工データの解析を通してその有効性を検証する．



(a)



(b)

図 1: 不連続曲線の構成例: (a) 重複節点 $t_7 = t_8 = t_9 = t_{10}$ に基づく基底関数. (b) 様々な平滑化パラメータ β の値に対する B -スプライン曲線.

2. 重複節点 B -スプライン

B -スプライン基底関数は、節点の位置を決めると一意に決定される関数である。連続で滑らかな曲線を推定する方法として、Eilers and Marx (1996), 井元・小西 (1999) は、図 2 (a) に示される等間隔節点に基づく B -スプライン基底関数を用いる方法を提案している。本節では、 B -スプライン基底関数の節点を重ね不連続な曲線を構成する方法について説明する。一般に r 次 B -スプラインに関しては、 $r + 1$ 個の節点を重ねることによって不連続な関数を構成できる。図 1 (a) には、3 次 B -スプラインにおいて 4 つの節点 ($t_7 = t_8 = t_9 = t_{10}$) を重ねた基底関数 $B_j(x)$ ($j = 1, \dots, 13$) の例を挙げる。この基底関数を用いることにより t_7 で不連続な回帰関数が構成される。また、重複させる節点の数を 1 つ減らし、 r 次 B -スプラインに対して r 個の節点を重ねると、その点で連続ではあるが 1 階導関数が不連続であるような曲線を構成することができる。

3. B -スプライン非線形回帰モデル

B -スプライン非線形回帰モデルは、一般化線形モデルの枠組みでの構成が可能であるが、ここでは簡単のため次のモデルを仮定する。

$$y_\alpha = \mu_\alpha + \varepsilon_\alpha, \quad \varepsilon_\alpha \sim N(0, \sigma_\alpha^2), \quad \alpha = 1, \dots, n. \quad (1)$$

ただし, $\mu_\alpha = E[Y_\alpha|x_\alpha]$ である. 観測点 x_α が与えられたときの目的変数 Y_α の期待値 μ_α ($\alpha = 1, \dots, n$) に対して B -スプラインによる非線形関数

$$\mu_\alpha = \sum_{j=1}^m \gamma_j B_j(x_\alpha) = \mathbf{b}_\alpha^T \boldsymbol{\gamma}, \quad \alpha = 1, \dots, n \quad (2)$$

を仮定する. ただし, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m)^T$ は m 次元パラメータベクトル, $\mathbf{b}_\alpha = (B_1(x_\alpha), \dots, B_m(x_\alpha))^T$ ($m < n$) は B -スプライン基底関数とデータから求められる既知の基底関数ベクトルとする. また, 区分的に異なる分散を持つモデルを当てはめるために, 誤差分散 σ_α^2 に対して次の構造を仮定する.

$$\sigma_\alpha^2 = \sum_{l=1}^L \tau_l^2 d(x_\alpha; R_l) = \mathbf{d}_\alpha^T \boldsymbol{\tau}, \quad \alpha = 1, \dots, n. \quad (3)$$

ただし, $d(\cdot; \cdot)$ は, $[x_{\min}, x_{\max}] = R_1 \cup \dots \cup R_L$, $R_i \cap R_j = \emptyset$ ($i \neq j$) としたとき, $d(x; R_l) = 1$ ($x \in R_l$), $= 0$ (otherwise) と定義し, $\mathbf{d}_\alpha = (d(x_\alpha; R_1), \dots, d(x_\alpha; R_L))^T$ は既知の L 次元ベクトル, $\boldsymbol{\tau} = (\tau_1^2, \dots, \tau_L^2)^T$ は L 次元パラメータベクトルとする. 誤差分散の変化に関しては, Chen and Gupta (1997) の研究が挙げられるが, データにトレンドがある場合, 適切な変換を施し平均を 0 にする必要がある.

従って, (1), (2), (3) より B -スプライン非線形回帰モデルは次式で与えられる.

$$f(y_\alpha|x_\alpha; \boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{\sqrt{2\pi \mathbf{d}_\alpha^T \boldsymbol{\tau}}} \exp \left\{ -\frac{(y_\alpha - \mathbf{b}_\alpha^T \boldsymbol{\gamma})^2}{2 \mathbf{d}_\alpha^T \boldsymbol{\tau}} \right\}. \quad (4)$$

一般化線形モデルの枠組みでの非線形回帰モデルの構成に関しては, Green and Silverman (1994), Imoto and Konishi (1999) を参照されたい.

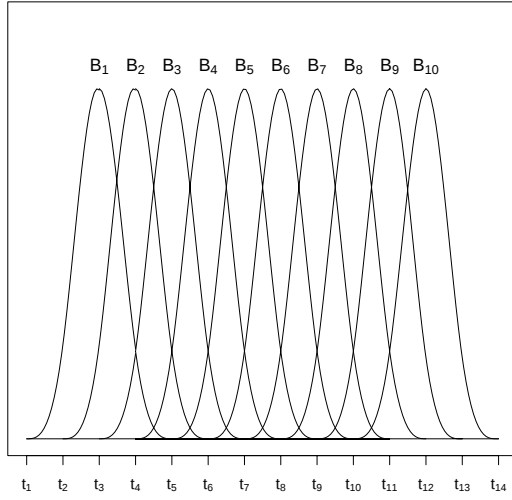
モデルのパラメータ $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$ は, 次の罰則付き対数尤度関数の最大化に基づいて推定する.

$$l_\lambda(\boldsymbol{\gamma}, \boldsymbol{\tau}) = \sum_{\alpha=1}^n \log f(y_\alpha|x_\alpha; \boldsymbol{\gamma}, \boldsymbol{\tau}) - \frac{n\lambda}{2} \boldsymbol{\gamma}^T K \boldsymbol{\gamma}. \quad (5)$$

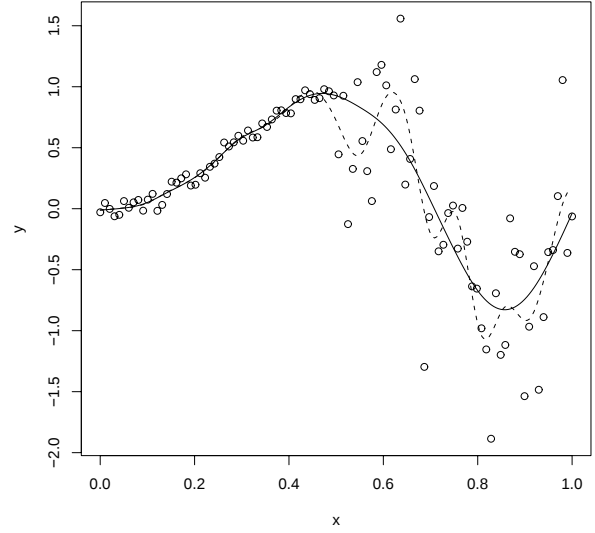
ただし, $\lambda (> 0)$ は平滑化パラメータ, K は負荷行列と呼ばれ, $m \times m$ 次非負定値対称行列である. Eilers and Marx (1996), 井元・小西 (1999) は, 連続な 3 次 B -スプライン曲線の罰則項として 2 階差分行列 D_2 を用い負荷行列を $K = D_2^T D_2$ と表した. しかし, 負荷行列をこのように設定すると, 例えば図 1 (a) の B_6 と B_7 の係数の差にも制限を加えていることになる. このような場合は, ジャンプ幅の過小推定がいくつかの実験で観測された. そこで本稿では, 負荷行列 K を次式で定義される $m \times (m-4)$ 行列 C_2 を用いて $K = C_2^T C_2$ とする.

$$C_2 = \begin{pmatrix} D_{2,m_1} & O \\ O & D_{2,m_2} \end{pmatrix}.$$

ただし, D_{2,m_i} は $m_i \times (m_i - 2)$ 行列 D_2 を表し, O はゼロ行列, $m_1 + m_2 = m$ とする. 例えば, 図 1 (a) の基底関数に対しては, $m = 13$, $m_1 = 6$, $m_2 = 7$ とする.



(a)



(b)

図 2: 区分的に異なる分散を持つモデルを当てはめた例 .

パラメータ γ, τ の推定値は $\theta = (\gamma^T, \tau^T)^T$ としたとき , 罰則付き尤度方程式 $\partial l_\lambda / \partial \theta = 0$ の解として得られ , Fisher スコア法などの繰り返し計算によって求めることができる . 特に , 分散パラメータを共通の σ^2 と置いたとき , モデルのパラメータ γ, σ^2 の推定値は解析的に

$$\hat{\gamma} = (B^T B + n\beta K)^{-1} B^T \mathbf{y}, \quad \hat{\sigma}^2 = \|\mathbf{y} - B\hat{\gamma}\|^2 / n \quad (6)$$

で与えられる . ただし $\beta = \sigma^2 \lambda$ を平滑化パラメータとして与えることで , これらの推定値を求めることができる . 図 1 (b) には , $x = 7$ で 4 つの節点を重ねたときの , 様々な平滑化パラメータ β に対する推定曲線を挙げる . 平滑化パラメータを適切に選択することで ($\beta = 10^{-2}$) , 曲線の突然の変化と非線形構造を捉えることができることが分かる .

図 2 (b) は区分的に異なる分散を持つモデルを当てはめた結果の例である . データは $y_\alpha = \sin(2\pi x_\alpha^2) + \varepsilon_\alpha$ から発生させ , 誤差分散に対して $\varepsilon_\alpha \sim N(0, 0.05^2)$ ($x_\alpha < 0.5$) , $\sim N(0, 0.6^2)$ ($x_\alpha \geq 0.5$) とした . 実線が $R_1 = \{x | x < a\}$, $R_2 = \{x | x \geq a\}$ とし区分的に異なる分散パラメータを持つモデルに基づく推定曲線であり , 破線が分散パラメータを共通の σ^2 として推定した曲線である . 各 B-スプライン基底関数は等間隔節点に基づいて構成し (図 2 (a)) , 平滑化パラメータ , 分散の変化点 $a = 0.5$ は次節で導出する情報量規準 SPIC に基づいて選択した . 区分的に異なる分散パラメータを持つモデルは , データを発生した真の構造を的確に捉えている . このとき分散パラメータの推定値は $\hat{\tau}_1^2 = 0.045^2$, $\hat{\tau}_2^2 = 0.53^2$ であった . しかし , 分散パラメータを共通としたモデルに基づく推定曲線には , 誤差分散の変化による見せかけの変動が見られる .

非線形回帰モデル (4) 式を (5) 式の罰則付き対数尤度関数の最大化に基づいて推定するとき，平滑化パラメータ，基底関数の個数，重複節点の位置，定義関数 $d(x; R_l)$ に含まれる領域 R_l の設定がモデル構築において本質的となる．従って，モデル評価規準に基づいてこれらの値を選択することが次節の目的である．

4. モデル評価規準

(5) 式の罰則付き最尤法に基づいて推定された B -スプライン非線形回帰モデル $f(y_\alpha|x_\alpha; \hat{\gamma}, \hat{\tau})$ を評価する情報量規準は

$$\text{SPIC} = -2 \sum_{\alpha=1}^n \log f(y_\alpha|x_\alpha; \hat{\gamma}, \hat{\tau}) + 2\text{tr}(\hat{I}_\lambda \hat{J}_\lambda^{-1}) \quad (7)$$

で与えられる．ただし， $\hat{I}_\lambda$ ， $\hat{J}_\lambda$ は， $(m+L) \times (m+L)$ 行列で

$$\begin{aligned} \Lambda \ (n \times n) &= \text{diag} \left[\frac{y_1 - \mathbf{b}_1^T \hat{\gamma}}{\mathbf{d}_1^T \hat{\tau}}, \dots, \frac{y_n - \mathbf{b}_n^T \hat{\gamma}}{\mathbf{d}_n^T \hat{\tau}} \right], \\ \Gamma \ (n \times n) &= \text{diag} \left[\frac{1}{\mathbf{d}_1^T \hat{\tau}}, \dots, \frac{1}{\mathbf{d}_n^T \hat{\tau}} \right], \\ B \ (n \times m) &= (\mathbf{b}_1, \dots, \mathbf{b}_n)^T; \quad \mathbf{b}_\alpha = (B_1(x_\alpha), \dots, B_m(x_\alpha))^T, \\ D \ (n \times L) &= (\mathbf{d}_1, \dots, \mathbf{d}_n)^T; \quad \mathbf{d}_\alpha = (d(x_\alpha; R_1), \dots, d(x_\alpha; R_L))^T, \\ \mathbf{1}_n \ (n \times 1) &= (1, \dots, 1)^T \end{aligned}$$

とおいたとき

$$\begin{aligned} \hat{I}_\lambda &= \begin{pmatrix} B^T \Lambda - \lambda K \hat{\gamma} \mathbf{1}_n^T, \\ D^T (\Lambda^2 - \Gamma) / 2 \end{pmatrix} \begin{pmatrix} \Lambda B, & (\Lambda^2 - \Gamma) D / 2 \end{pmatrix}, \\ \hat{J}_\lambda &= \begin{pmatrix} B^T \Gamma B + n \lambda K, & B^T \Lambda \Gamma D \\ D^T \Gamma \Lambda B, & D^T (2 \Lambda^2 - \Gamma) \Gamma D / 2 \end{pmatrix} \end{aligned}$$

で与えられる．情報量規準 SPIC が最小となる平滑化パラメータ，節点の位置，および分割 $R_1, \dots, R_L$ を選択し，対応するモデルを最適なモデルとする．また，対応する重複節点の位置を変化点の推定値とする．このスプライン情報量規準 SPIC は，想定したモデルとデータを発生した真のモデルの分離のもとで Kullback-Leibler 情報量あるいは期待平均対数尤度の推定量として一般化情報量規準 (Konishi and Kitagawa, 1996) の理論に基づき導出した．また，分散パラメータを共通の σ^2 としたとき，(7) 式の情報量規準 SPIC は，井元・小西 (1999) で得られた情報量規準に一致する．情報量規準構成の一般理論に関しては，Konishi and Kitagawa (1996)，Konishi (1999) を参照されたい．

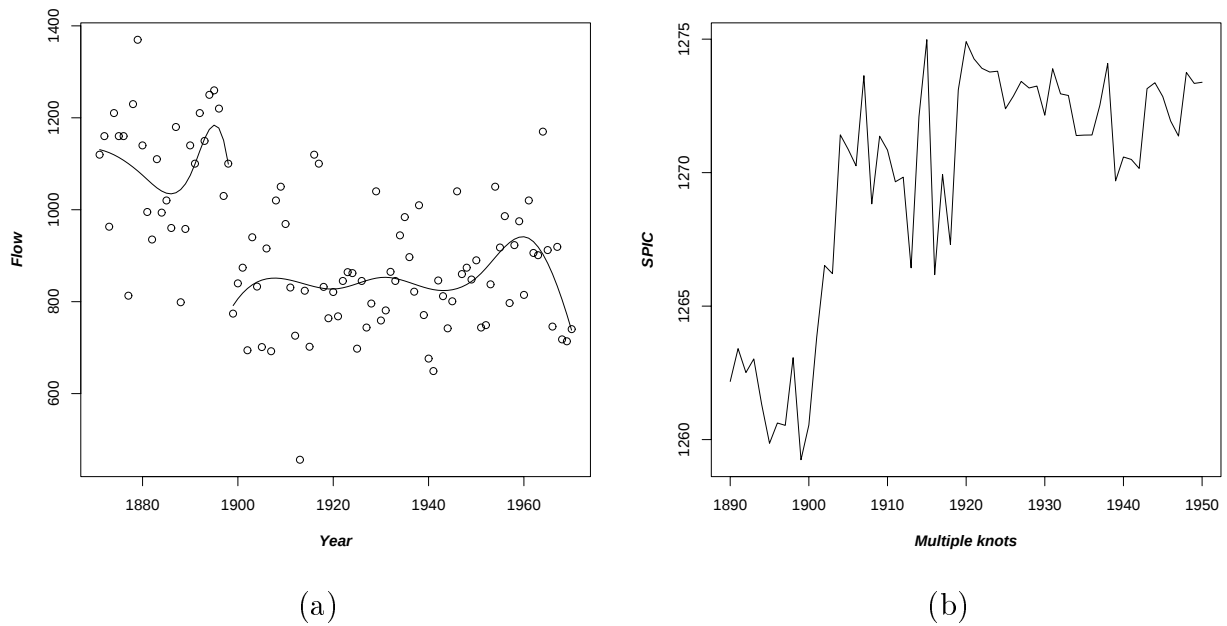


図 3: ナイルデータ

5. 適用例

• ナイルデータ

変化点問題の例としてしばしば分析されるのが、図 3 (a) に示されるナイルデータ (Cobb, 1978) である。このデータは 1871 年から 1970 年までのナイル河の年間河川流量 ($10^8 m^3$) を計測したものである。Cobb (1978) は、1898 年を境に年間河川流量が突然変化していることを示し、前世紀の変わり目における熱帯地方の降水型の変化と関連させ議論した。また、Müller (1992) も Cobb と同じく 1898 年と 1899 年の間で何らかの変化が起こったと推定している。

このデータに対して、(4) 式の B -スプライン非線形回帰モデルを 16 個の B -スプライン基底関数、および共通の分散パラメータ $\tau = \sigma^2$ で構成したモデルに基づいて分析した。基底関数を構成する節点は、まず $[x_{min}, x_{max}]$ を 10 等分し、変化点の候補点に重複節点を配置した。重複節点の位置はここではデータ点に一致させた。また、 B -スプライン基底関数の構成方法から、ここでの基底関数は右連続となる。すなわち、図 1 (a) において重複節点 t_7 に一致するデータ x に対して $B_6(x) = 0$, $B_7(x) > 0$ となる。図 3 (b) は、重複節点の位置を変えたときの情報量規準 SPIC の挙動を示している。ただし、各重複節点の位置に対して平滑化パラメータは最適化している。この図から重複節点を 1899 年に置いたとき SPIC の値が最小となり、1898 年と 1899 年の間で年間流量が突然変化したことが示唆される。図 3 (a) の曲線が対応する B -スプライン曲線である。ナイル河の年間河川流量が 1898 年と 1899 年の間で急激に変化している様子が現れている。

- 数値実験

本節では，提案する方法の変化点検出能力について検討する．データは次のモデルに従って発生させた．

$$y_\alpha = \sin(2\pi x_\alpha) + h \times I(x_\alpha \geq 0.5) + \varepsilon_\alpha, \quad x_\alpha = (2\alpha - 1)/(2n), \quad \alpha = 1, \dots, n.$$

ただし， $I(x_\alpha \geq a) = 0$ ($x_\alpha < a$), $= 1$ ($x_\alpha \geq a$)， $\varepsilon_\alpha \sim N(0, \sigma^2)$ とした．従って h が真の曲線のジャンプ幅，0.5 が正しい変化点である．実験は表 1 に挙げる設定で行った．図 4 に $n = 200$, $\sigma = 0.3$, $h = 0.75$ のデータ例を挙げる．実験の目的は，正しい変化点 $x = 0.5$ を推定し，データを発生した真の構造 $y = \sin(2\pi x) + h \times I(x \geq 0.5)$ を捉えることである．

データ数 n :	200, 400
誤差分散 σ :	0.1, 0.2, 0.3
ジャンプ幅 h :	0.5, 0.75, 1.0

表 1: 実験設定

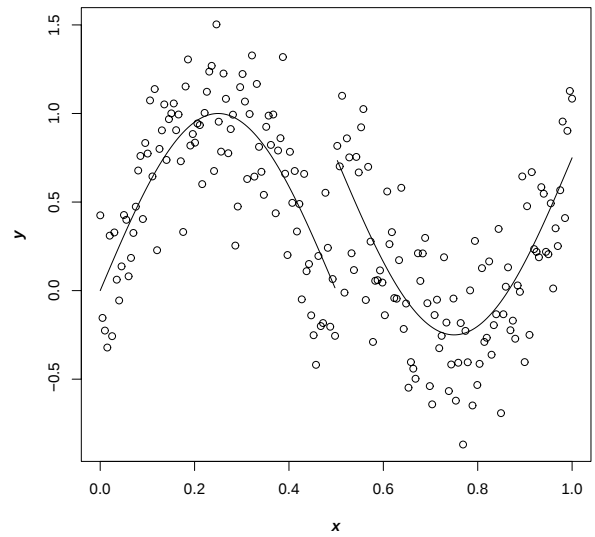


図 4: データ例： $n = 200$, $\sigma = 0.3$, $h = 0.75$

実験はそれぞれの設定において 100 回の繰り返しを行った．図 5, 6 に $n = 200$, $\sigma = 0.3$, $h = 1.0$ と $n = 400$, $\sigma = 0.3$, $h = 0.75$ に対応する結果を示した．それぞれ (a) が発生したデータの例，(b) は選択された変化点のヒストグラム，(c) が推定された曲線である．ただし，(c) に破線で示した曲線は，等間隔節点に基づく連続な B -スプライン曲線であり SPIC に基づいて選択した．図 5, 6 (b) から分かるように，提案する変化点推定法は高い検出能力を有する．示した結果は共に誤差分散が $\sigma = 0.3$ のものであるが，分散が小さくなるにつれ検出力は高くなる．特に誤差分散が $\sigma = 0.1$ の時は，実験の範囲では全てのデータの組に対して正しい変化点を推定できた．ジャンプ幅が $h = 0.5$ の場合は，視覚的には変化がほとんど捉えることができないような場合に相当する．このような場合でもデータ数が $n = 400$ であれば，誤差分散が $\sigma = 0.2$ の時，80%以上の高い検出力を得ることができた．図 6 (b) の結果に見られるように，正しい変化点の 1 つ隣のデータを変化点として推定することが数回あった．このような場合にも多くの場合，推定される曲線はデータを発生した真の構造を的確に

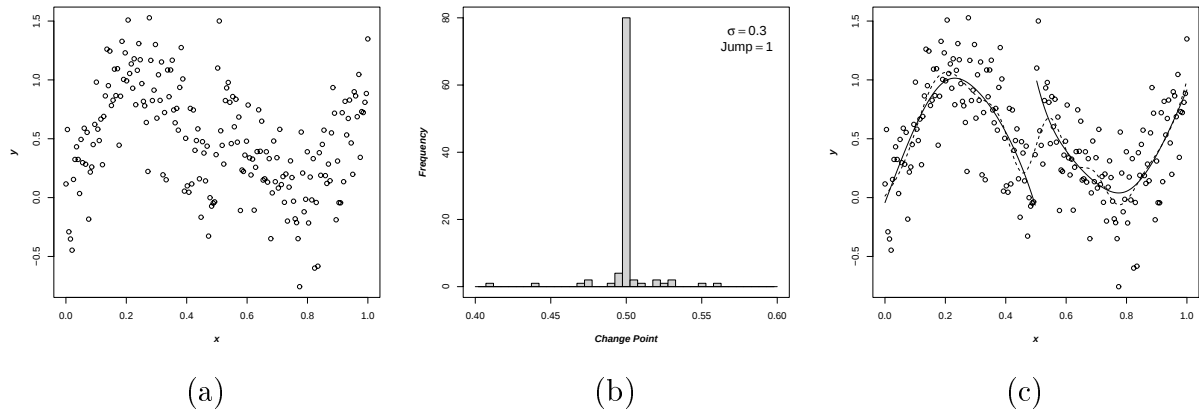


図 5: 結果 1: $n = 200$, $\sigma = 0.3$, $h = 1.0$

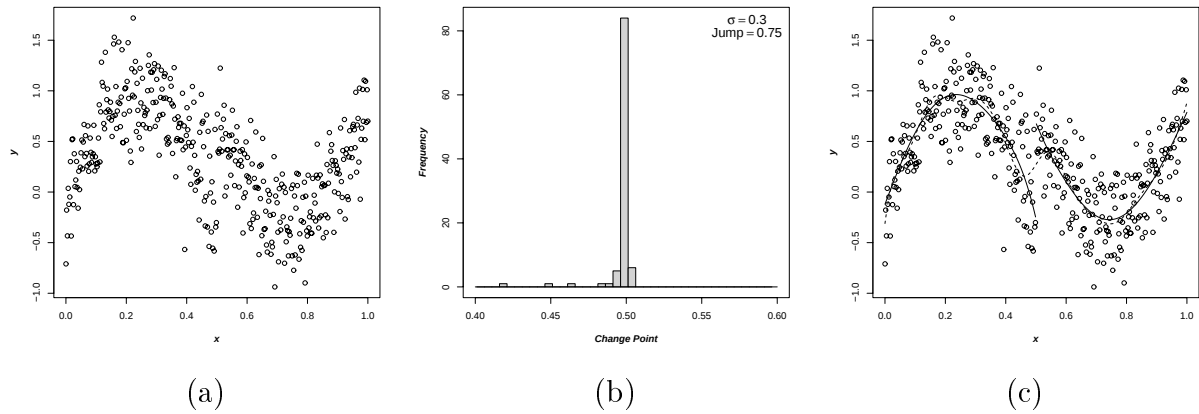


図 6: 結果 2: $n = 400$, $\sigma = 0.3$, $h = 0.75$

捉えていることが分かった．図 5, 6 (c) には滑らかな B -スプライン曲線を当てはめた結果も破線で示している．目的変数 Y_α の平均構造が不連続であるような場合，スプラインや核関数に基づく非線形回帰モデルを当てはめると，データの突然の変化を追うために平滑化パラメータを小さく選択することが多い．実際，破線で示された曲線は，平滑化パラメータの過小評価による不適切な変動が見られる．しかし，変化点を探索し不連続な B -スプライン曲線を当てはめた結果，データを発生した真の曲線に極めて近い曲線が推定されていることが分かった．

参考文献

- Chen, J. and Gupta, A. K. (1997). Testing and Locating Variance Change-points with Application to Stock Prices. *Journal of American Statistical Association*, **92** 739-747.
- Cobb, G. W. (1978). The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika*, **65** 243-251.

- de Boor, C. (1978). *A Practical Guide to Splines*. Springer, Berlin.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.
- 井元清哉, 小西貞則 (1999). 情報量規準に基づく B -スプライン非線形回帰モデルの推定. 応用統計学, 第 28 巻, 第 3 号, 137-150.
- Imoto, S. and Konishi, S. (1999). Estimation of B -spline nonparametric regression models using information criteria. *Proceedings of the 52nd. International Statistical Institute*, **2** 157-158.
- Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83** 875-890.
- Konishi, S. (1999). Statistical model evaluation and information criteria, *Multivariate Analysis, Design of Experiments and Survey Sampling* (ed. S. Ghosh), 369-399, Marcel Dekker, New York.
- Koo, J.-Y. (1997). Spline Estimation of Discontinuous Regression Functions. *Journal of Computational and Graphical Statistics*. **6** 266-284.
- Loader, C. R. (1996). Change Point Estimation using Nonparametric Regression. *Ann. Statist.* **24** 1667-1678.
- Müller, H. G. (1992). Change-points in Nonparametric Regression Analysis. *Ann. Statist.* **20** 737-761.