スプライン関数を利用した
ノンパラメトリック回帰

1 はじめに

- スプライン関数（以下では単に「スプライン」と記す）は、数多くの回帰係数によって表されるにもかかわらず、得られる曲線は滑らかなものになる。
- 回帰分析においては回帰係数の数を少なくすることを意図しない回帰のことをノンパラメトリック回帰と総称する。スプラインを用いた回帰の多くは、ノンパラメトリック回帰に分類できる。
- Ｂ－スプラインを利用すると、数値計算の分量は多くなるけれどもアルゴリズムは容易になる。S-Plusにおいても B-スプラインが利用されている。

2　Ｂ－スプラインによるスプラインの表現

有限区間のスプライン関数を定義するための式の中で、特によく使われるのが以下のものである。

$$s_{finite}(x) \;=\; \tilde{a}_0 + \tilde{a}_1 x + \tilde{a}_2 x^2 + \tilde{a}_3 x^3 + \sum_{j=2}^{m-1} \tilde{b}_j \cdot \left((x - \xi_{j+3})_+\right)^3$$
$$(\xi_4 \le x \le \xi_{m+3}) \tag{2.1}$$

$\{\tilde{a}_0, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{b}_2, \tilde{b}_3, \cdots, \tilde{b}_{m-1}\}$が回帰係数。

$\{\xi_4, \xi_5, \cdots, \xi_{m+3}\}(\xi_4 \;\le\; \xi_5 \;\le\; \cdots \;\le\; \xi_{m+3})$が節点$(\mathrm{knot},\ \mathrm{breakpoint})$の位置。

$((\cdot)_+)^3$は切断冪関数$(\mathrm{truncated\ \ power\ \ function})$と呼ばれるものの一種で、$(x)_+$は以下のように定義される。

$$(x)_+ = \begin{cases} x & \text{if} \quad x \ge 0 \\ 0 & \text{if} \quad x < 0 \end{cases} \tag{2.2}$$

式(2.1)を以下のように書き換えることができる。

$$s_{finite}(x) = \sum_{j=1}^{m+2} \beta_j B_j(x) \quad (\xi_4 \leq x \leq \xi_{m+3}) \tag{2.3}$$

$\{B_j(x)\}$が基底(basis)である。$B_j(x)$は$\xi_j \leq x \leq \xi_{j+4}$の範囲以外では0になる（つまり、局所的な台(support)を持つ基底）。また、$B_j(x)$が負の値をとることはない。

- 式(2.1)も式(2.3)も$(m+2)$個の回帰係数を持つ。
- 式(2.1)の形で与えられた関数（定義域が、$\xi_4 \leq x \leq \xi_{m+3}$ ）は、式(2.3)の形で厳密に表現でき、逆も成り立つ。すなわち、式(2.1)が張る関数空間と式(2.3)が張る関数空間が同じものになる
- 式(2.3)を用いると、この式の形をそのまま利用して数値計算を行っても高い精度の結果が得られることや、より複雑な型式の回帰式へ発展させることが容易であることが利点である。
- 式(2.1)を用いる場合には、$\{\xi_4, \xi_5, \cdots, \xi_{m+3}\}$という$(n-2)$個の節点を設定する（$\xi_4$と$\xi_{m+3}$はどこに設定してもスプラインの定義式が変わるだけ）のに対して、式(2.2)を用いると、$\{\xi_1, \xi_2, \cdots, \xi_{m+6}\}$の$(n+6)$個の節点を設定する必要がある。
- $\{\xi_1, \xi_2, \xi_3\}$と$\{\xi_{m+4}, \xi_{m+5}, \xi_{m+6}\}$の値は、$\xi_1 \leq \xi_2 \leq \xi_3 \leq \xi_4$と$\xi_{m+3} \leq \xi_{m+4} \leq \xi_{m+5} \leq \xi_{m+6}$を満たしさえすれば、$\xi_4 \leq x \leq \xi_{m+3}$の範囲で式(2.1)と式(2.3)が等価になる。

従って、この６つの節点の位置を以下のように設定する場合だけを考えれば、$\xi_4 \leq x \leq \xi_{m+3}$の範囲において式(2.3)を使って式(2.1)と同じ関数を表現するという目的を達成することができる。そのため、以下のようにする場合だけを考えれば充分である。

$$\xi_1 = \xi_2 = \xi_3 = \xi_4, \qquad \xi_{m+3} = \xi_{m+4} = \xi_{m+5} = \xi_{m+6} \tag{2.4}$$

S-Plusの`bs()`の標準的な使い方においては、式(2.4)が設定される。$\{\xi_5, \xi_6, \cdots, \xi_{m+}$を内部節点(internal breakpoints)と称して、その位置を`knots=`の形で指定するようになっているのはこのためである。

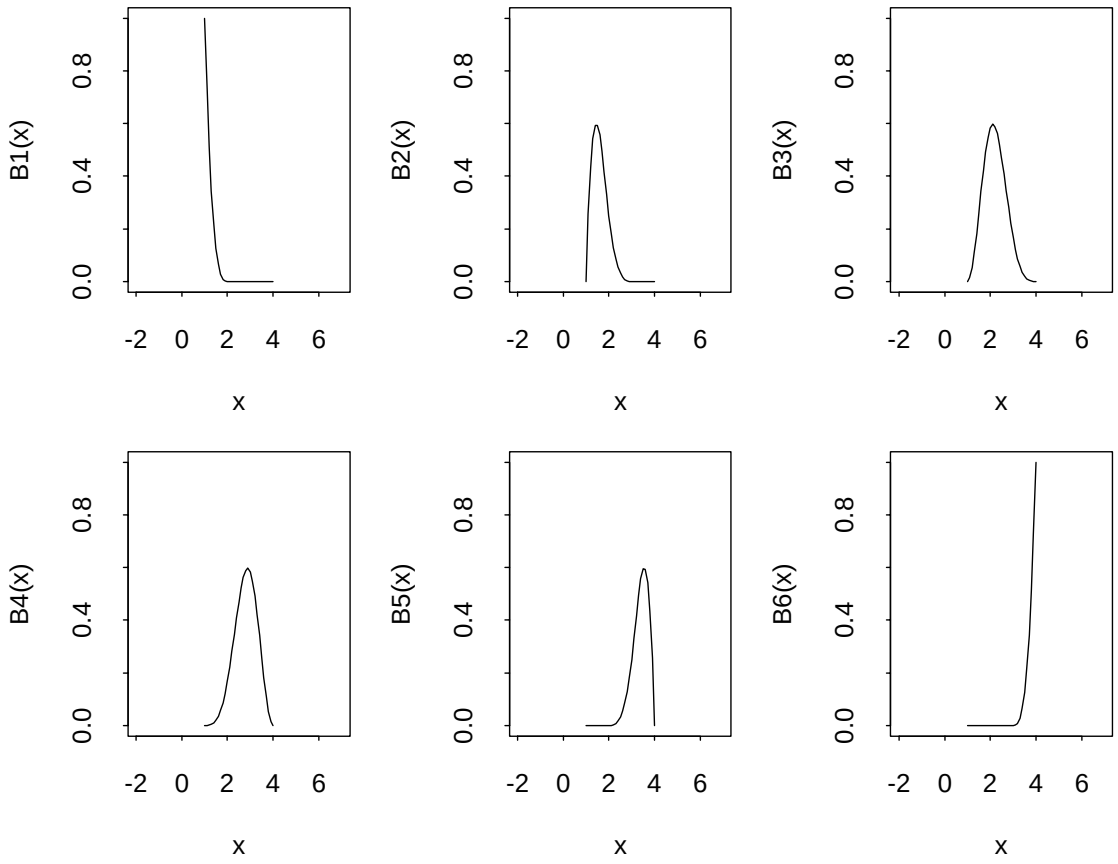

図 1: 節点を $\{1, 1, 1, 1, 2, 3, 4, 4, 4, 4\}$ に設定したときの３次のＢ－スプラインの基底の全て。bs() を用いて描いた。

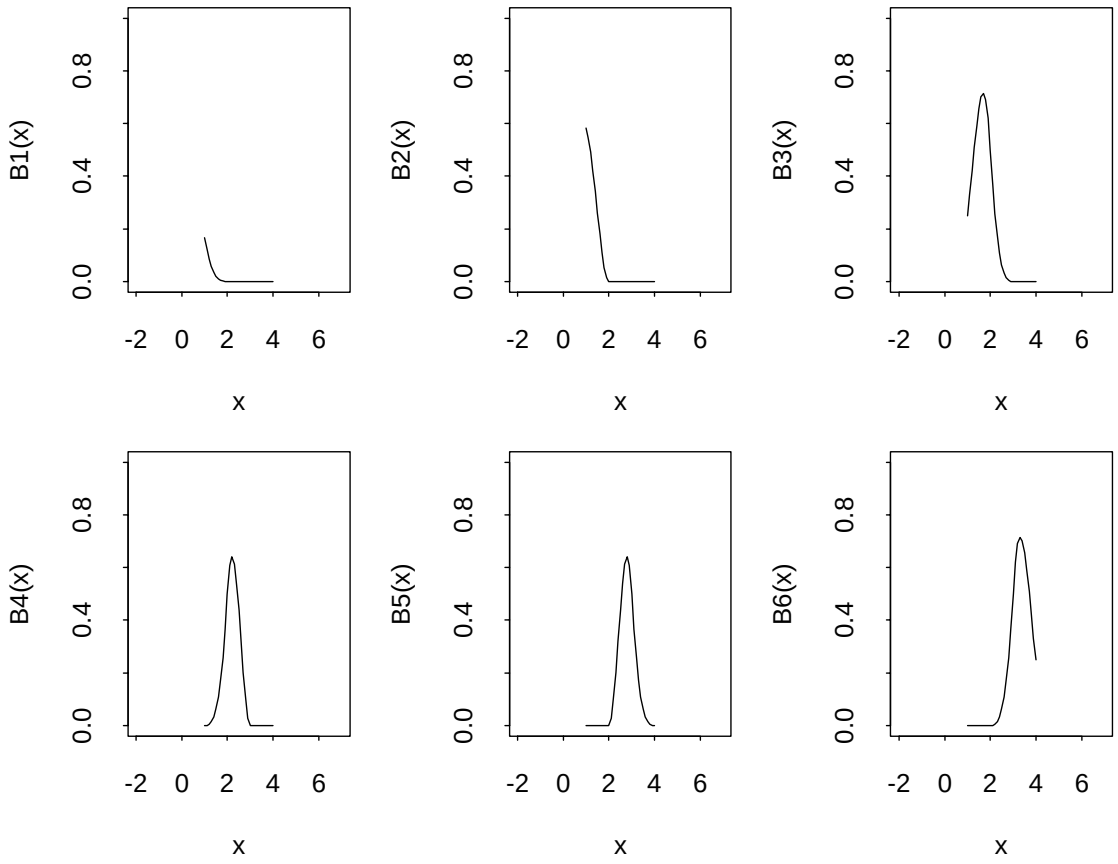

図 2: 節点を $\{-2, -1, 0, 1, 2, 3, 4, 5, 6, 7\}$ に設定したときの３次のＢ－スプラインの基底の全て。spline.des() を用いて描いた。

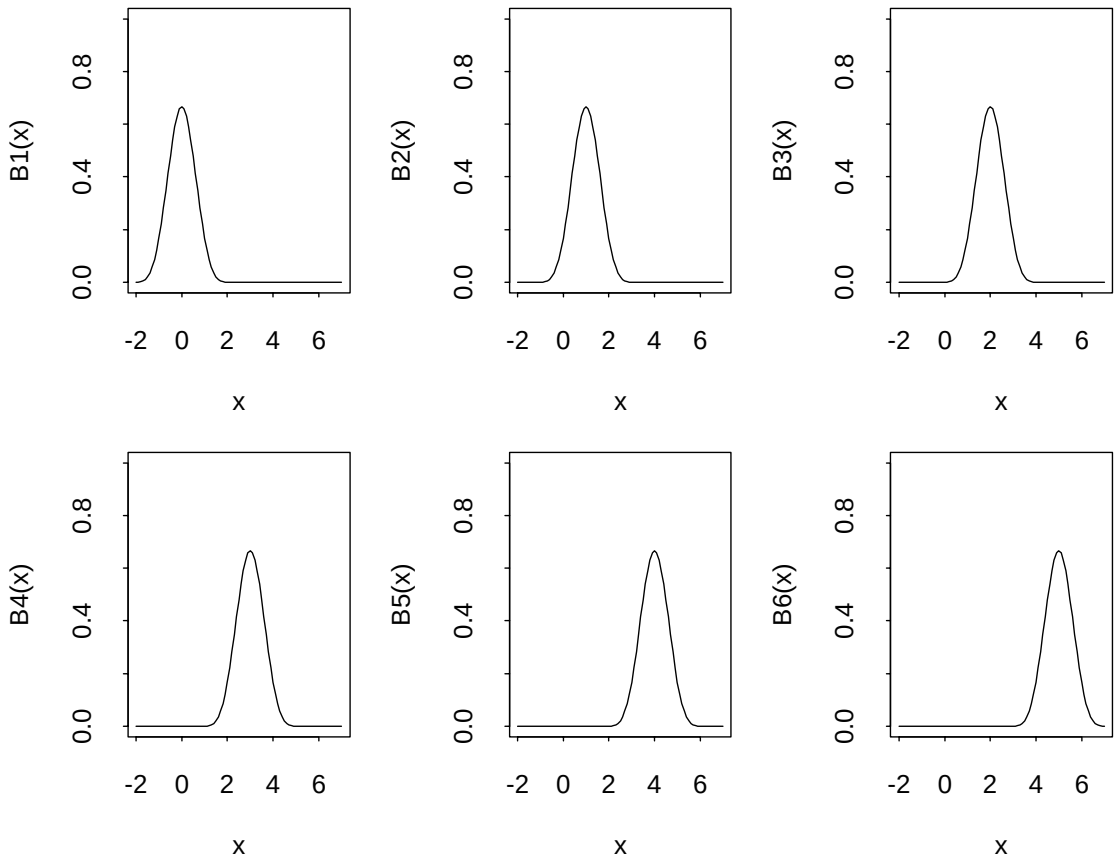

図 3: 節点を $\{-2,-1,0,1,2,3,4,5,6,7\}$ に設定したときの３次のＢ－スプラインの基底の全て。定義にしたがってＢ－スプラインの基底を描くオブジェクトを作製した。

3 Ｂ－スプラインによる平滑化スプライン

平滑化スプラインとは、以下の値を最小にする回帰関数。

$$E_{ss} = \sum_{i=1}^{n} (Y_i - m(X_i))^2 + \lambda \int_{-\infty}^{\infty} \left(\frac{d^2 m(x)}{dx^2} \right)^2 dx \qquad (3.1)$$

Ｂ－スプラインを利用して平滑化スプラインの計算を行うことの特質。

- 計算量は幾らか多くなる。
- $m(x)$ として式 (2.3) をそのまま利用すれば精度の高い結果が得られる。
- データの予測変数の部分に同じ値をとるものが含まれていた場合にも特別な措置を必要としない。
- 部分スプライン等に発展させることが容易である。

S-Plus の `smooth.spline()` も Ｂ－スプラインを利用した計算方法を用いている。

式(2.3)を利用すると、式(3.1)が以下のものになる。

$$E_{ss} = (\mathbf{y} - \mathbf{G}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{G}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^t\mathbf{K}\boldsymbol{\beta} \tag{3.2}$$

$\{\beta_i\}$のそれぞれについて偏微分して0とおくと、この値を最小にする$\boldsymbol{\beta}$を$\hat{\boldsymbol{\beta}}$とすると、$\hat{\boldsymbol{\beta}}$は以下のように書ける。

$$\hat{\boldsymbol{\beta}} = (\mathbf{G}^t\mathbf{G} + \lambda\mathbf{K})^{-1}\mathbf{G}^t\mathbf{y} \tag{3.3}$$

ここで、$\mathbf{y}$、$\mathbf{G}$、$\boldsymbol{\beta}$、$\mathbf{K}$は以下のものである。

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} \tag{3.4}$$

$$\mathbf{G} = \begin{pmatrix} K_1(X_1) & K_2(X_1) & \dots & K_{n+2}(X_1) \\ K_1(X_2) & K_2(X_2) & \dots & K_{n+2}(X_2) \\ K_1(X_3) & K_2(X_3) & \dots & K_{n+2}(X_3) \\ \vdots & \vdots & \ddots & \vdots \\ K_1(X_n) & K_2(X_n) & \dots & K_{n+2}(X_n) \end{pmatrix} \tag{3.5}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{n+2} \end{pmatrix} \tag{3.6}$$

$$\mathbf{K} = \begin{pmatrix} k_{11} & k_{12} & \dots & k_{1(n+2)} \\ k_{21} & k_{22} & \dots & k_{2(n+2)} \\ \vdots & \vdots & \ddots & \vdots \\ k_{(n+2)1} & k_{(n+2)2} & \dots & k_{(n+2)(n+2)} \end{pmatrix} \tag{3.7}$$

$\{X_i\}$がデータの予測変数の部分、$\{Y_i\}$がデータの目的変数の部分。

また、式(3.7)のk_{ij}は以下のものである。

$$k_{ij} = \int_{-\infty}^{\infty} \left(\frac{d^2 K_i(x)}{dx^2} \right) \left(\frac{d^2 K_j(x)}{dx^2} \right) dx \tag{3.8}$$

よって、式(3.2)は以下のように書ける。

$$E_{ss} = (\mathbf{y} - \mathbf{G}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{G}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^t \mathbf{K} \boldsymbol{\beta} \tag{3.9}$$

$\{\beta_i\}$のそれぞれについて偏微分して0とおくと、この値を最小にする$\boldsymbol{\beta}$を$\hat{\boldsymbol{\beta}}$とすると、$\hat{\boldsymbol{\beta}}$は以下のように書ける。

$$\hat{\boldsymbol{\beta}} = (\mathbf{G}^t \mathbf{G} + \lambda \mathbf{K})^{-1} \mathbf{G}^t \mathbf{y} \tag{3.10}$$

よって、ハット行列は以下のようになる。

$$\mathbf{H}^{ss} = \mathbf{G}(\mathbf{G}^t \mathbf{G} + \lambda \mathbf{K})^{-1} \mathbf{G}^t \tag{3.11}$$

ハット行列とは以下の性質を持つ行列である。

$$\hat{\mathbf{y}} = \mathbf{H}^{ss} \mathbf{y} \tag{3.12}$$

ここで、$\hat{\mathbf{y}}$は以下のものである。

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \hat{Y}_3 \\ \vdots \\ \hat{Y}_n \end{pmatrix} \tag{3.13}$$

$\{\hat{Y}_i\}$は、$\{Y_i\}$に対応する推定値である。

ドゥブア・コックスのアルゴリズムを用いて$\{K_i(x)\}$を求めた後、その２次導関数（１次関数）の二つの係数の値を求めるためには、`solve()`を使って連立方程式を解く方法がある。しかし、悪条件に見舞われているときにも高い精度の計算を行うためには、`lm()`と`poly()`を用いて最小２乗法による３次式のあてはめを行う方が安全である。`poly()`は直交多項式を用いた計算を行うので悪条件の場合にも安定した計算になる可能性が高いためである。

＜注意＞

S-Plus の `smooth.spline()` において、平滑化パラメータの値を設定しないと自動的に最適な平滑化パラメータが選択する。しかし、データの予測変数の部分に同じ値をとるものが含まれているときには、最適でない平滑化パラメータが選択されてしまう。

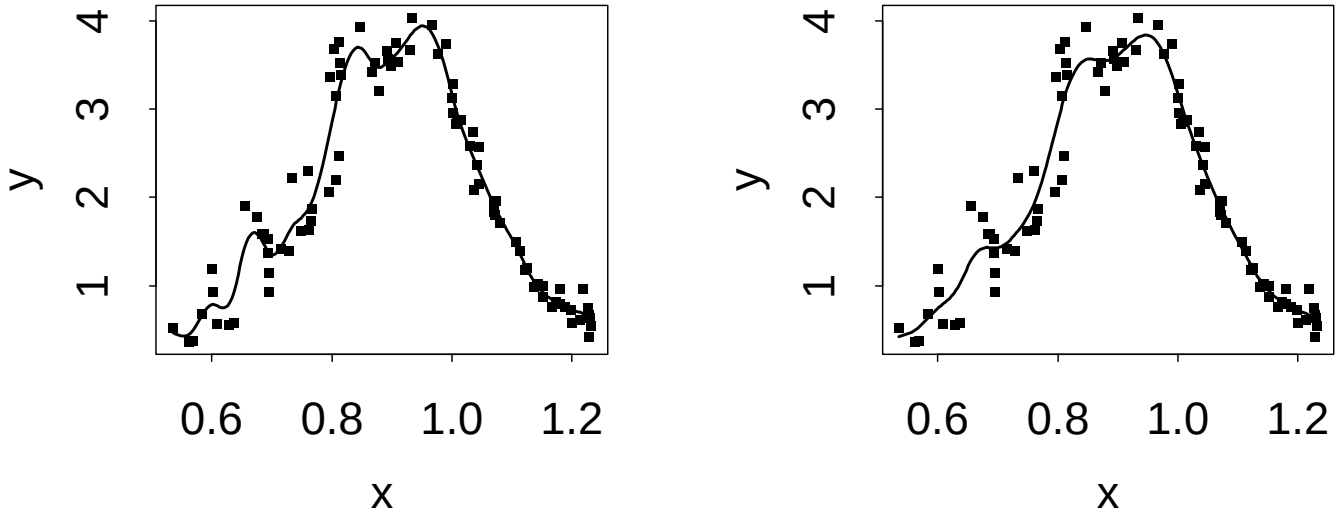

図 4: smooth.spline() の GCV による平滑化パラメータの自動最適化機能を用いた場合（「平滑化とノンパラメトリック回帰への招待」の 208 ページの図 5.28 の上部のグラフ）（左）。GCV を用いて平滑化パラメータを正しく最適化した場合（右）。

平滑化パラメータを等しくしたときの、`smooth.spline()`による推定値を$\{\hat{Y}_i^{s-plus}\}$、GCVPACKによる推定値を$\{\hat{Y}_i^{gcvpack}\}$、Bースプラインを用いた自作のオブジェクトによる推定値を$\{\hat{Y}_i^{b-sp}\}$とし、$\{d_i^1\}$と$\{d_i^2\}$を以下のように定義する。

$$d_i^1 = \hat{Y}_i^{s-plus} - \hat{Y}_i^{gcvpack} \tag{3.14}$$

$$d_i^2 = \hat{Y}_i^{b-sp} - \hat{Y}_i^{gcvpack} \tag{3.15}$$

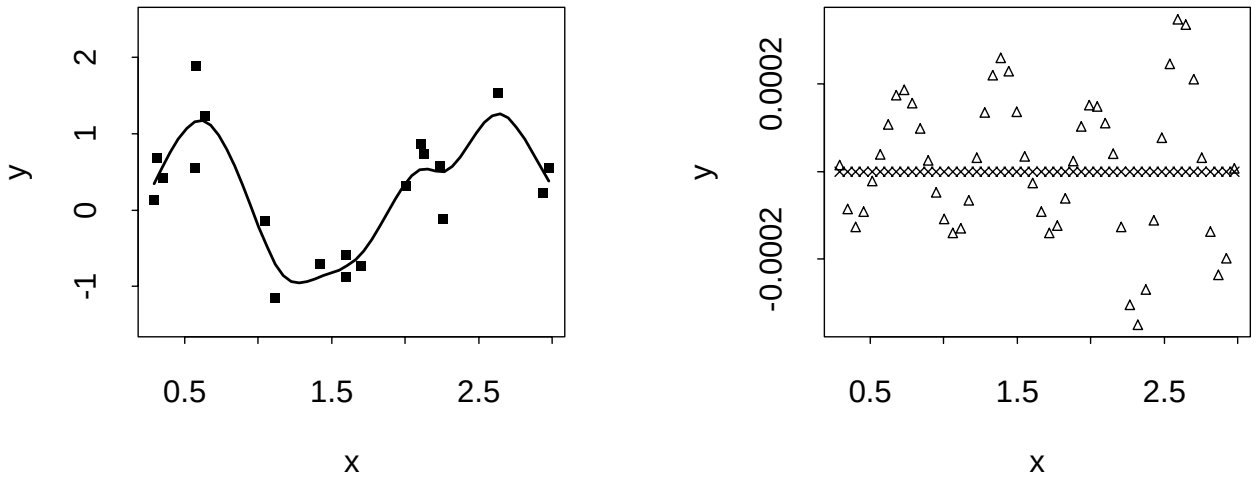

図 5: smooth.spline()による平滑化。△が$\{d_i^1\}$、×が$\{d_i^2\}$（右）。

4　部分スプラインの作製

お米の作況指数の主要な部分には、部分スプラインが用いられている。Ｂ－スプラインを用いて部分スプラインと呼ばれる回帰式を表現すると以下のものになる。

$$y(t,\{x_j\}) = \sum_{j=1}^{n+2} \beta_j \cdot g_j(t) + \sum_{j=1}^{l} a_j x_j \tag{4.1}$$

ここで、tは年次（実際の年次は、$\{T_j\}$）、$\{x_j\}$は主成分分析を使って気象要素を線形変換したものである。部分スプラインを用いる方法は1997年（平成９年）から利用されている。そこでは、Ｂ－スプラインではなく、GCVPACKを利用する方法が用いられてきた。

しかし、より信頼性の高い結果を得るための方法を模索するためには、S-PlusにおいてＢ－スプラインを利用した様々計算方法を試みるべきであるという観点からの検討が進められた。その詳細は、インターネット上で公表されている。

そこで、式(4.1)の$\{\beta_j\}$と$\{a_j\}$を求めるために、以下の値を最小にする。

$$E_{semi} = (\mathbf{y} - \mathbf{G}\boldsymbol{\beta} - \mathbf{X}\mathbf{a})^t(\mathbf{y} - \mathbf{G}\boldsymbol{\beta} - \mathbf{X}\mathbf{a}) + \lambda\boldsymbol{\beta}^t\mathbf{K}\boldsymbol{\beta} \tag{4.2}$$

ここで、$\mathbf{X}$と$\mathbf{a}$は以下のものである。

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1l} \\ X_{21} & X_{22} & \dots & X_{2l} \\ X_{31} & X_{32} & \dots & X_{3l} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nl} \end{pmatrix} \tag{4.3}$$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_l \end{pmatrix} \tag{4.4}$$

X_{ij}はi番目のデータにおいて、気象要素を主成分分析を使って線形変換したもののうち、j番目のものである。

また、ここでの$\mathbf{G}$は以下のものである。

$$\mathbf{G} = \begin{pmatrix} K_1(T_1) & K_2(T_1) & \dots & K_{n+2}(T_1) \\ K_1(T_2) & K_2(T_2) & \dots & K_{n+2}(T_2) \\ K_1(T_3) & K_2(T_3) & \dots & K_{n+2}(T_3) \\ \vdots & \vdots & \ddots & \vdots \\ K_1(T_n) & K_2(T_n) & \dots & K_{n+2}(T_n) \end{pmatrix} \tag{4.5}$$

式(4.2)を、$\boldsymbol{\beta}$と$\mathbf{a}$のそれぞれの要素で微分して0とおくと以下の式が得られる。

$$\begin{pmatrix} \mathbf{G}^t\mathbf{G} + \lambda\mathbf{K} & \mathbf{G}^t\mathbf{X} \\ \mathbf{X}^t\mathbf{G} & \mathbf{X}^t\mathbf{X} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{G}^t \\ \mathbf{X}^t \end{pmatrix} \mathbf{y} \tag{4.6}$$

このときのハット行列は以下のものになる。

$$\mathbf{H}^{semi} = \begin{pmatrix} \mathbf{G} & \mathbf{X} \end{pmatrix} \begin{pmatrix} \mathbf{G}^t\mathbf{G} + \lambda\mathbf{K} & \mathbf{G}^t\mathbf{X} \\ \mathbf{X}^t\mathbf{G} & \mathbf{X}^t\mathbf{X} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{G}^t \\ \mathbf{X}^t \end{pmatrix} \tag{4.7}$$

後進あてはめ法(backfitting)による部分スプラインの導出

$\{Y_i^{muti}\}$は推定値の重回帰の部分、$\{Y_i^{ss}\}$は推定値のスプラインの部分をそれぞれ示す。

Step 1: $\{Y_i^{muti}\}(1 \leq i \leq n)$の要素をすべて0にする。

Step 2: $\{Y_i - Y_i^{muti}\}$に年次を予測変数とする平滑化スプラインをあてはめる。$\{Y_i\}(1 \leq i \leq n)$は収量データである。得られた平滑化スプラインによる推定値を、$\{Y_i^{ss}\}$とする。

Step 3: $\{Y_i - Y_i^{ss}\}$に気象要素を用いた重回帰式をあてはめる。得られた重回帰式による推定値を$\{Y_i^{multi}\}$とする。

Step 4: Step 2とStep 3を$\{Y_i^{ss}\}$と$\{Y_i^{multi}\}$が収束するまで繰り返し、得られたスプラインと重回帰式の和を最適な回帰式とする。

S-Plusの`gam()`を使って、加法モデルや部分スプラインなどを作製すると、この方法が用いられる。

GCVPACK、式(4.2)を用いたもの（S-Plusを用いて作製）、S-Plusに標準で所収されているオブジェクト（`lm()`、後退あてはめ法を利用している）、自作の後退あてはめ法のオブジェクトを比較した。

以下の式を用いてシミュレーションデータを作製した。

$$\begin{aligned} T_i &= i \quad (1 \leq i \leq 22) \\ Y_i &= sin(\pi/22 \cdot T_i) + X_{i1} + 2X_{i2} + 3X_{i3} + \epsilon_i \end{aligned} \tag{4.8}$$

ここで、$\{T_i\}$が年次に相当し、$\{X_{i1}\}$、$\{X_{i2}\}$、$\{X_{i3}\}$が気象要素に相当する。$\{X_{i1}\}$、$\{X_{i2}\}$、$\{X_{i3}\}$は0と10の間の値をとる一様乱数の実現値である。$\{\epsilon_i\}$は平均が0、標準偏差（分散の平方根）が0.1の正規乱数の実現値である。

平滑化スプラインの部分（$c_0 + c_1 t + \frac{1}{12}\Sigma_{j=1}^{n} b_j \cdot \mid t - T_j \mid^3$、$\Sigma_{j=1}^{P} \beta_j \cdot g_j(t)$）の推定値は平均が0になるようにしている。平滑化パラメータ（λ）の値は1とした。

４つの方法による推定値を比較するために、以下の、$\{e_i^1\}$、$\{e_i^2\}$、$\{e_i^3\}$を求めた。

$$e_i^1 = \hat{Y}_i^{b-sp} - \hat{Y}_i^{gcvpack} \tag{4.9}$$

$$e_i^2 = \hat{Y}_i^{s-plus} - \hat{Y}_i^{gcvpack} \tag{4.10}$$

$$e_i^3 = \hat{Y}_i^{backfitting} - \hat{Y}_i^{gcvpack} \tag{4.11}$$

ここで、$\{\hat{Y}_i^{gcvpack}\}$はGCVPACKによる推定値、$\{\hat{Y}_i^{b-sp}\}$は式(4.6)による推定値、$\{\hat{Y}_i^{s-plus}\}$は`lm()`による推定値、$\{\hat{Y}_i^{backfitting}\}$は後退あてはめ法のオブジェクトを自作して実行した結果である。

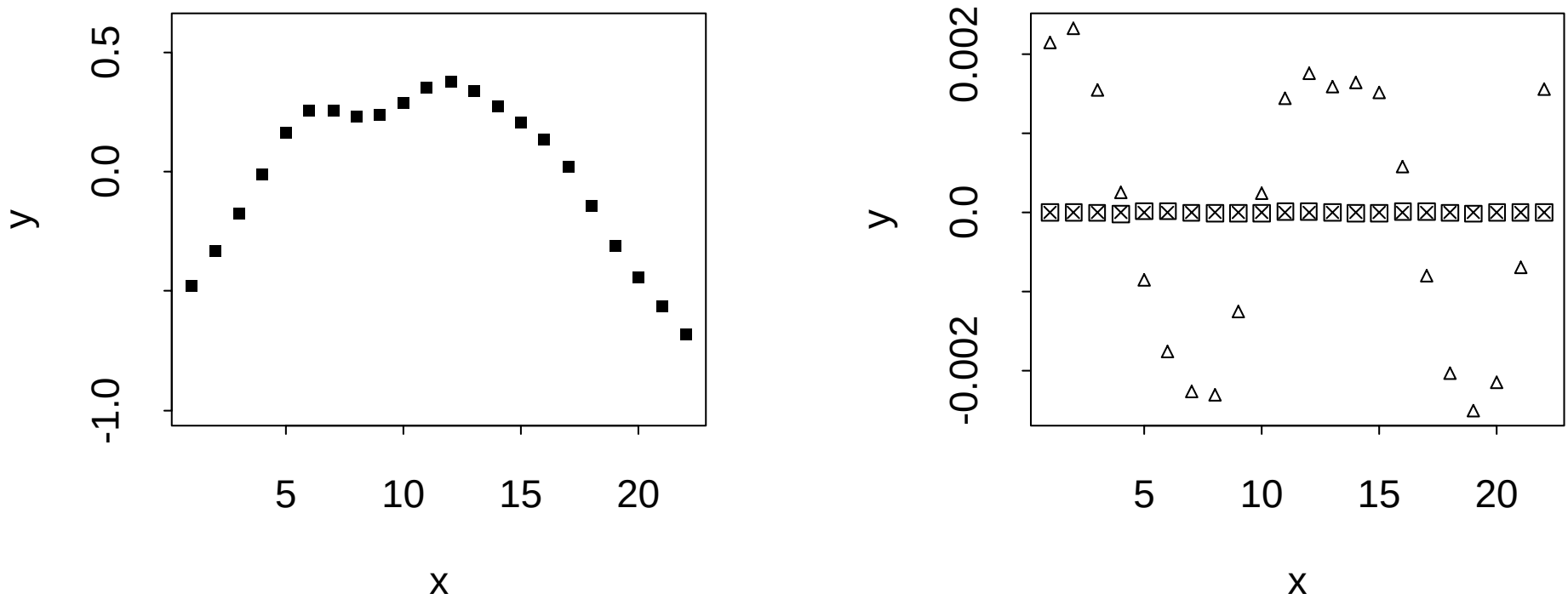

図 6: lm() を使って得られた部分スプラインのうち、スプラインの部分。×が $\{e_i^1\}$（B－スプライン）、△が $\{e_i^2\}$（S-Plus）、□が $\{e_i^3\}$（右）（自作の収束法）。

$\{\hat{Y}_i^{s-plus}\}$（S-Plus）が他の値と幾分ずれているのは、繰り返し計算の回数が少ないためか、平滑化スプラインの推定値の精度が低いためかは分からない。

5　部分スプラインにリッジ回帰を加味する方法

式(4.2)を用いると、得られる回帰係数（$\{a_j\}$）の値が作物の性質とは矛盾することがある。例えば、日照時間に対する回帰係数の値が負になることである。これは、同一年次に、日照時間以外の気象要素の値が等しい実験を行えば、日照時間が長いほど収量が小さくなることを意味している。これでは、作物の性質を表した回帰係数とは言えない。

こうした点を改善するための方法として、リッジ回帰が知られている。そこで、式(4.2)にリッジ回帰を導入することを試みる。

リッジ回帰を取り入れると、式(4.2)は以下のものになる。

$$E_{semi-r} = (\mathbf{y} - \mathbf{G}\boldsymbol{\beta} - \mathbf{X}\mathbf{a})^t(\mathbf{y} - \mathbf{G}\boldsymbol{\beta} - \mathbf{X}\mathbf{a}) + \lambda\boldsymbol{\beta}^t\mathbf{K}\boldsymbol{\beta} + k \parallel \mathbf{a} \parallel^2 \quad (5.1)$$

ここで、$\parallel\ \parallel$はベクトルの長さを表す。kは正または0の定数で、$k = 0$のとき、式(4.2)になる。k の値が大きくなるにしたがって、回帰係数の２乗和を小さくしようという指向が強くなる。

式(5.1)を、$\boldsymbol{\beta}$と$\mathbf{a}$のそれぞれの要素で微分して0とおくと以下の式が得られる。

$$\begin{pmatrix} \mathbf{G}^t\mathbf{G} + \lambda\mathbf{K} & \mathbf{G}^t\mathbf{X} \\ \mathbf{X}^t\mathbf{G} & \mathbf{X}^t\mathbf{X} + k\mathbf{I}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{G}^t \\ \mathbf{X}^t \end{pmatrix} \mathbf{y} \tag{5.2}$$

ここで、$\mathbf{I}_m$は$m \times m$の大きさの単位行列である。

このときのハット行列は以下のものになる。

$$\mathbf{H}^{(r)} = \begin{pmatrix} \mathbf{G} & \mathbf{X} \end{pmatrix} \begin{pmatrix} \mathbf{G}^t\mathbf{G} + \lambda\mathbf{K} & \mathbf{G}^t\mathbf{X} \\ \mathbf{X}^t\mathbf{G} & \mathbf{X}^t\mathbf{X} + k\mathbf{I}_m \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{G}^t \\ \mathbf{X}^t \end{pmatrix} \tag{5.3}$$

GCVPACKでは、回帰式の妥当性を調べるための統計量としてGCVが用いられている。GCVは、回帰式の妥当性を調べるための統計量としてその価値が理論的にも実証的にも充分に明らかになっていて、しかも、GCVの値の平方根をほぼ予測誤差の大きさと見なすことができるからである。しかし、他の統計量を使うこともできる。よく利用されるものは以下のものである。

(1) クロスバリデーション (Cross-Validation)

$$CV = \sum_{i=1}^{n} \frac{(\hat{m}(X_i) - Y_i)^2}{n \cdot (1 - [\mathbf{H}]_{ii})^2} \tag{5.4}$$

ここで、SSEが残差２乗和で、$[\mathbf{H}]_{ii}$がハット行列の対角要素である。

(2) 一般化クロスバリデーション (Generalized Cross-Validation)

$$GCV = \frac{SSE}{n \cdot (1 - \frac{\Sigma_{i=1}^{n} [\mathbf{H}]_{ii}}{n})^2} \tag{5.5}$$

(3) 赤池の情報量基準 (Akaike's Information Criterion)

$$AIC = n \cdot \log\left(\frac{SSE}{n}\right) + 2\sum_{i=1}^{n}[\mathbf{H}]_{ii} \tag{5.6}$$

(4) 赤池の情報量基準を修正したもの (Corrected AIC)

$$AICc = \log\left(\frac{SSE}{n}\right) + \frac{n + \Sigma_{i=1}^{n}[\mathbf{H}]_{ii}}{n - \Sigma_{i=1}^{n}[\mathbf{H}]_{ii} - 2} \tag{5.7}$$

式(5.1)を用いた回帰において上記の４つの統計量がどのような結果を導くかを調べるためには、理論的な検討だけでは充分ではない。そこで、以下のような式を用いてシミュレーションデータを作製した。

$$T_i = i \quad (1 \leq i \leq 22) \tag{5.8}$$

$$X_{1i} = \epsilon_i^1 \tag{5.9}$$

$$X_{2i} = \epsilon_i^2 \tag{5.10}$$

$$X_{3i} = 0.8X_{2i} + 0.2\epsilon_i^3 \tag{5.11}$$

$$\begin{aligned} Y_i = & \ sin(\pi/22 \cdot T_i) \\ & + X_{i1} + 2X_{i2} + 3X_{i3} + \epsilon_i^4 \end{aligned} \tag{5.12}$$

ここで、$\{T_i\}$が年次に相当し、$\{X_{i1}\}$、$\{X_{i2}\}$、$\{X_{i3}\}$が気象要素に相当する。ϵ_i^1、ϵ_i^2、ϵ_i^3は0と1の間の値をとる一様乱数の実現値である。$\{\epsilon_i\}$は平均が0、標準偏差（分散の平方根）が0.2の正規乱数の実現値である。

まとめ

- Ｂ－スプラインの基本的な性質を理解しさえすれば、ノンパラメトリック回帰の計算を気軽に行うことができる。
- スプラインや平滑化スプラインの基本的な計算には、S-Plus に標準で所収されているオブジェクトが利用できる。
- S-Plus を有効に利用すれば、部分スプラインなどの、より発展的な回帰を行うことも容易である。
- GCVPACK のような評価が確立したソフトを S-Plus の一部分として利用することも、計算結果の検証と、計算の高速化のために有益である。
- 加法モデル (Additive model) や ACE(Alternating Conditional Expectations) なども、S-Plus に所収されているオブジェクトをそのまま使うだけではなく、Ｂ－スプラインを利用したものを自作することもできる。

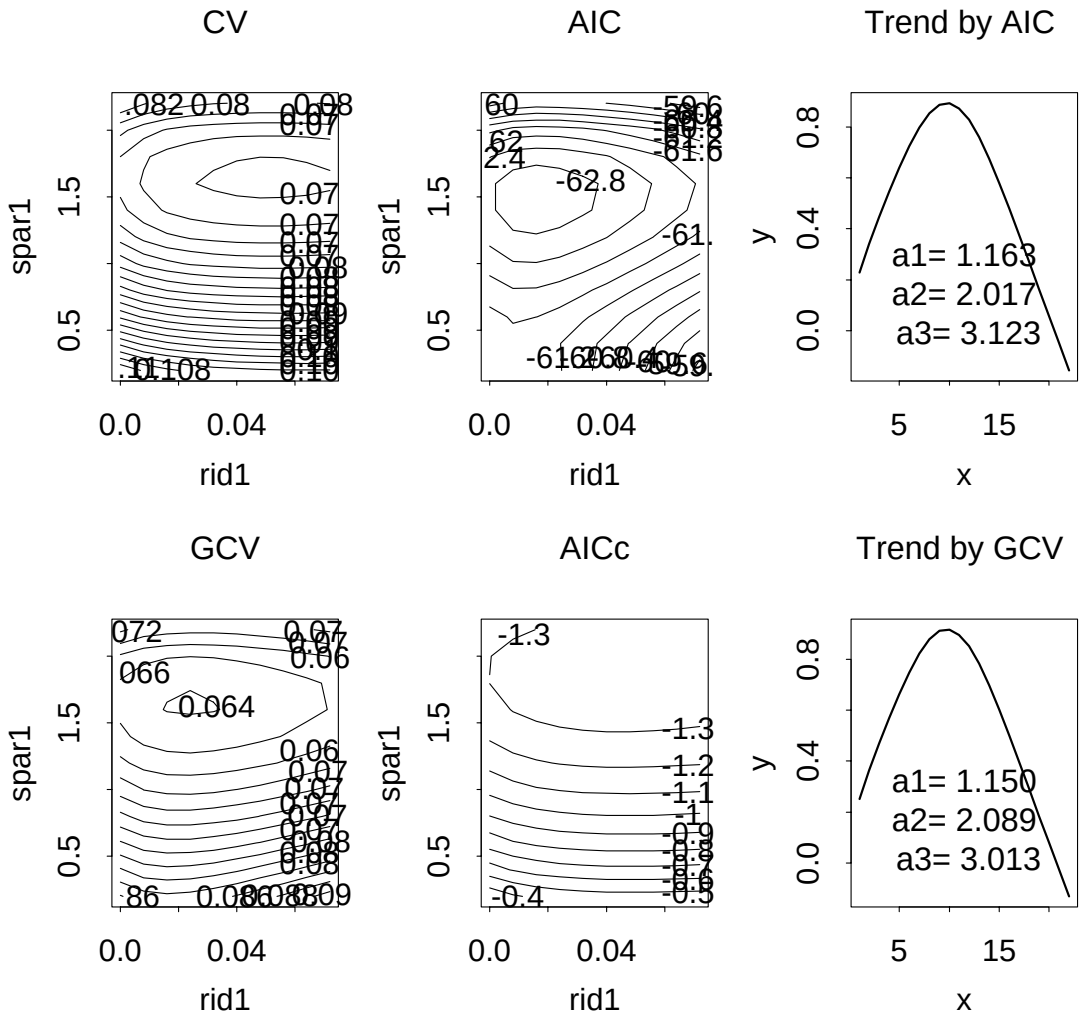

図 7: 式(5.8)を用いたシミュレーションデータによる結果の一例。グラフの中の「spar1」が λを示している。何れの統計量を用いた場合も結果はあまり違わない。

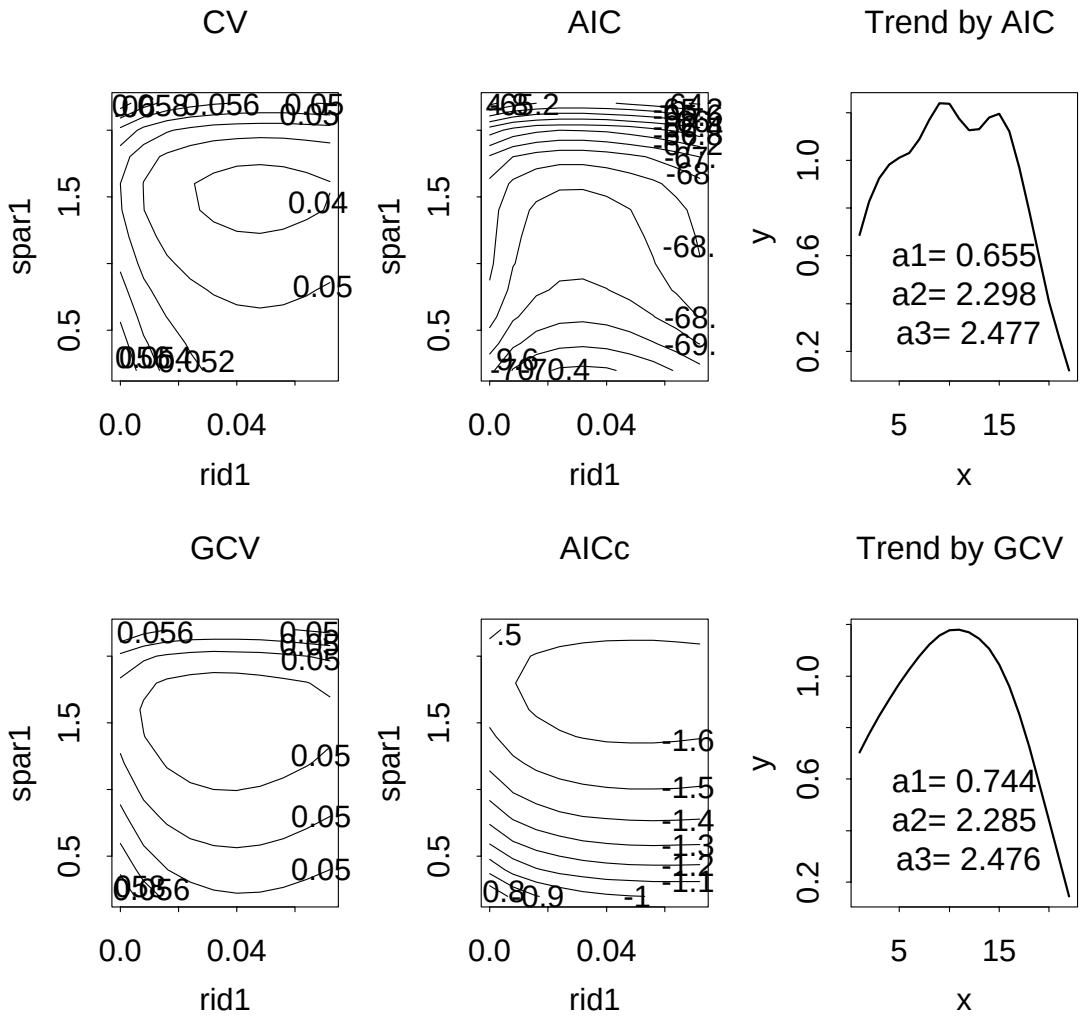

図 8: 式(5.8)を用いたシミュレーションデータによる結果の一例。グラフの中の「spar1」が λを示している。AICを用いた場合は、年次の影響を表す曲線の凸凹が大きすぎる。