

# 統計的学習の基礎

## 第5章 基底展開と正則化（～5.6）

森 浩太

2016年8月23日

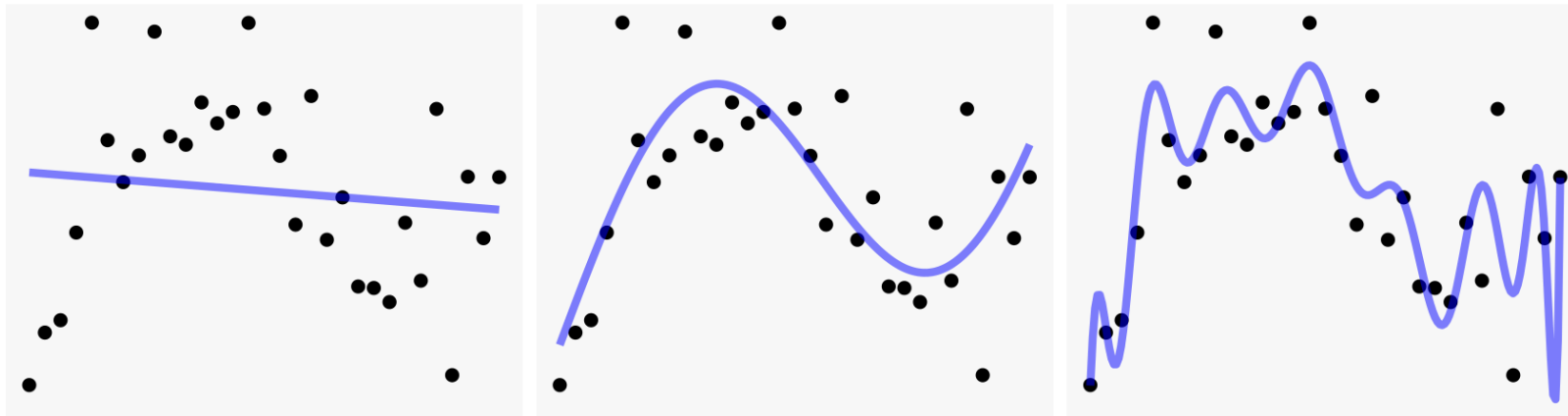
# 概要

## 基底展開 (Basis expansion)

説明変数  $X$  に事前に変数変換を施すことで、より柔軟な関係を表現する

## 正則化 (Regularization)

複雑な（自由度の高い）モデルに対して制約を課して過適合を是正する



# ロードマップ

1. 基底展開とは？ (5.1)
2. 区分的多項式、スプライン、3次自然スプラインとは？ (5.2)
3. 3次自然スプラインを用いた特徴抽出 (5.3)
4. 平滑化スプライン (5.4 ~ 5.5)
5. ロジスティック回帰への応用 (5.6)

\* 5.1の一部と5.3を除いて、 $X$ が1次元の場合を扱う

基底展開とは？

# 基底展開とは？

説明変数  $X$  に事前に変数変換を施すことで、より柔軟な関係を表現する

$$f(X) = \sum_j \beta_j X_j$$
$$\implies f(X) = \sum_k \beta_k h_k(X)$$

変換後の変数について線形モデル → これまでの手法がそのまま使える

# 基底展開の例

- 多項式回帰は基底展開の例

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots$$

問題に関する知識や想定に基づいて設定することも

- $X$ の効果は減衰していくはず・・・

$$f(X) = \alpha + \beta \log(X)$$

変数を減らす（まとめる）ために変換を施すことも

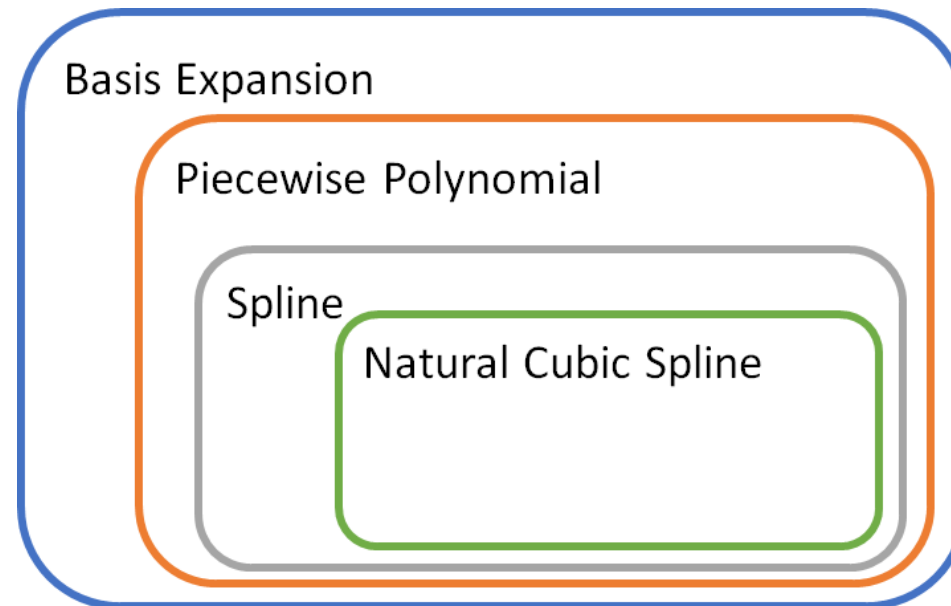
- 教育支出が世帯収入に依存するとすると・・・

$$f(Inc) = \alpha + \beta(Inc_f + Inc_m)$$

# 区分的多項式・スプライン・3次自然スプライン

# 区分的多項式・スプライン・3次自然スプライン

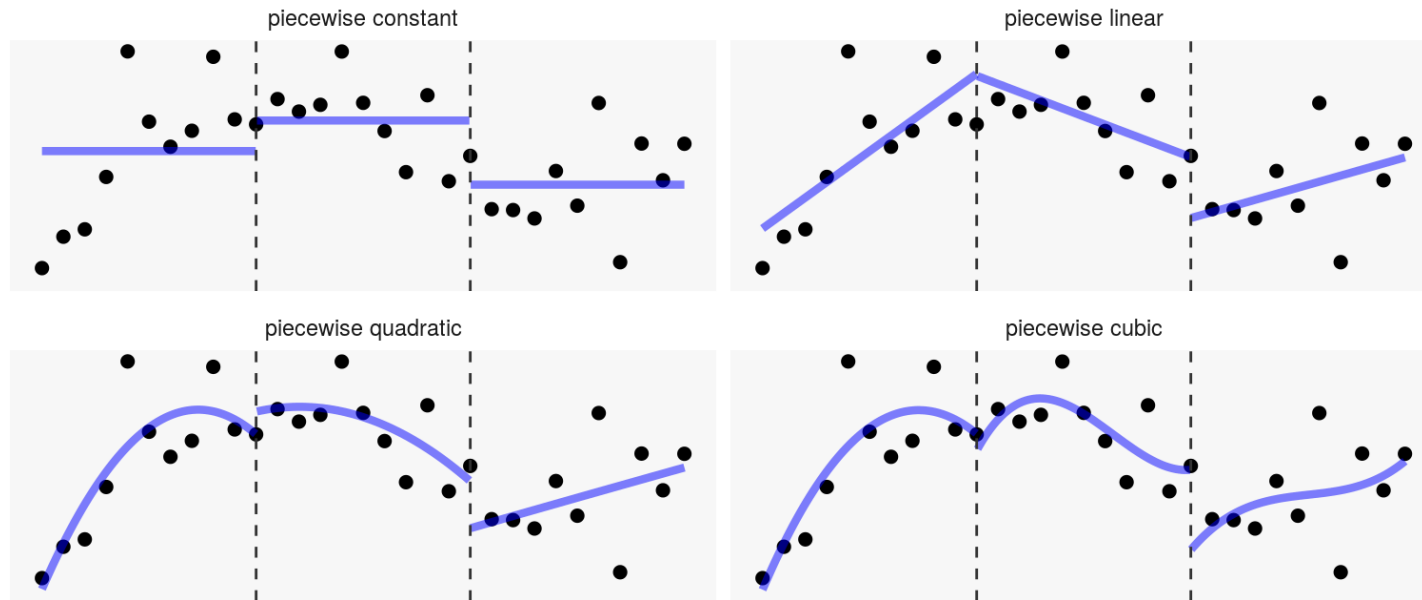
基底展開  $\supset$  区分的多項式  $\supset$  スプライン  $\supset$  3次自然スプライン





# 区分的多項式

区間ごとに異なる多項式モデルに従う



自由度(i.e. パラメータ数) = 区間数  $\times$  (次元 + 1)

E.g. 3次  $\cdot$  3区間  $\Rightarrow 3 \cdot (3 + 1) = 12$  パラメータ

自由度が高い  $\Leftrightarrow$  モデルが複雑

# Rで区分的多項式モデル

区間を表す因子変数をつくり、交差項で回帰する

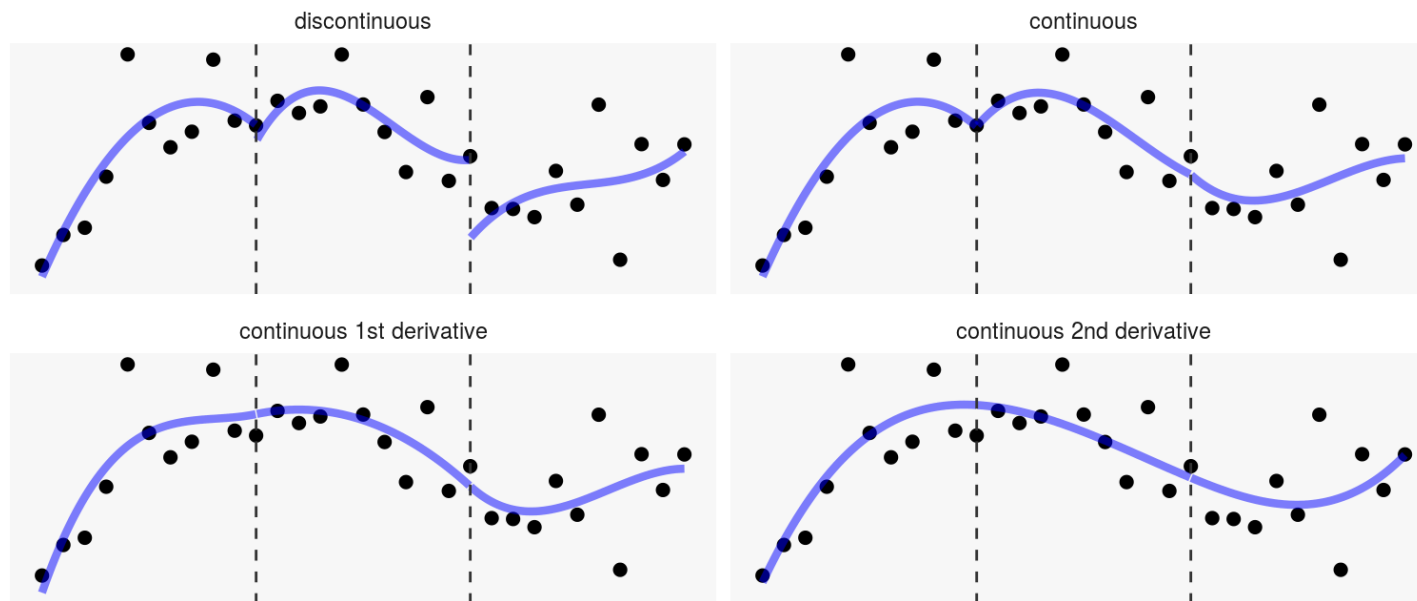
```
X <- runif(50)
Y <- runif(50)
segment <- rep(1, length(X))
segment[x > quantile(X, 1/3)] <- 2
segment[x > quantile(X, 2/3)] <- 3
segment <- factor(segment)
lm(Y ~ X*segment + I(X^2)*segment + I(X^3)*segment)
```

```
##
## Call:
## lm(formula = Y ~ X * segment + I(X^2) * segment + I(X^3) * segment)
##
## Coefficients:
##      (Intercept)              X      segment2      segment3
##          0.3431         2.4631         0.1595         0.2981
##          I(X^2)         I(X^3)      X:segment2      X:segment3
##         -8.2592         6.7109        -0.5746        -4.4755
## segment2:I(X^2) segment3:I(X^2) segment2:I(X^3) segment3:I(X^3)
##              NA          13.7797              NA          -10.6711
```

# 連続微分条件で滑らかにする

区分的多項式モデル  $f(X)$  に次のような制約をつけることで、滑らかなモデルを表現

1. 各区分点で  $f(X)$  が連続
2. 各区分点で  $f'(X)$  が連続
3. 各区分点で  $f''(X)$  が連続 .....



# 区分的多項式を表現する基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

この12の基底で3次・3区間の区間的多項式モデルを表現できる

$\xi_1, \xi_2$  は区分点を表す ( $\xi_1 < \xi_2$  とする)

$1(\cdot)$  はカッコ内が真なら 1、そうでなければ0になる関数

$(\cdot)_+$  は正の部分を表す (負の数はゼロに切り捨て)

# 区分的多項式を表現する基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

$X \leq \xi_1$  のとき、網掛け部分はゼロ

上段の4変数で3次多項式を表現

# 区分的多項式を表現する基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

$\xi_1 < X \leq \xi_2$  のとき、

新たに加わった中段の 4 変数で異なる 3 次多項式を表現

# 区分的多項式を表現する基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

$X > \xi_2$  のとき、

新たに加わった下段の4変数で異なる3次多項式を表現

結果、3区間で異なる3次多項式を表現することができる

# 連続性制約を課した場合の基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

オレンジのセルは  $X = \xi_1$  において非連続

緑のセルは  $X = \xi_2$  において非連続

⇒ これらのセルの係数はゼロにする必要がある



# 連続性制約を課した場合の基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

この10変数で、連続制約付きの3次区分的多項式を表現できる

# 1次導関数までの連続性を課す場合の基底

	$1$	$X$	$X^2$
	$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$
	$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$

各項を微分したものを考える（係数は無視してよい）

オレンジのセルは  $X = \xi_1$  において非連続

緑のセルは  $X = \xi_2$  において非連続

# 1次導関数までの連続性を課す場合の基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

この8変数で、1階連続微分制約付きの3次区分的多項式を表現できる

## 2次導関数までの連続性を課す場合の基底

		$1$	$X$
		$1(X > \xi_1)$	$(X - \xi_1)_+$
		$1(X > \xi_2)$	$(X - \xi_2)_+$

各項をもう1度微分する

オレンジのセルは  $X = \xi_1$  において非連続

緑のセルは  $X = \xi_2$  において非連続

## 2次導関数までの連続性を課す場合の基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

この6変数で、2階連続微分制約付きの3次区分的多項式を表現できる (Exercise 5.1)

## 3次導関数までの連続性を課す場合の基底

			$1$
			$1(X > \xi_1)$
			$1(X > \xi_2)$

各項をもう1度微分する

オレンジのセルは  $X = \xi_1$  において非連続

緑のセルは  $X = \xi_2$  において非連続

## 3次導関数までの連続性を課す場合の基底

$1$	$X$	$X^2$	$X^3$
$1(X > \xi_1)$	$(X - \xi_1)_+$	$(X - \xi_1)_+^2$	$(X - \xi_1)_+^3$
$1(X > \xi_2)$	$(X - \xi_2)_+$	$(X - \xi_2)_+^2$	$(X - \xi_2)_+^3$

この4変数で、3階連続微分制約付きの3次区分的多項式を表現できる

これは、通常の3次多項式モデルと同じ

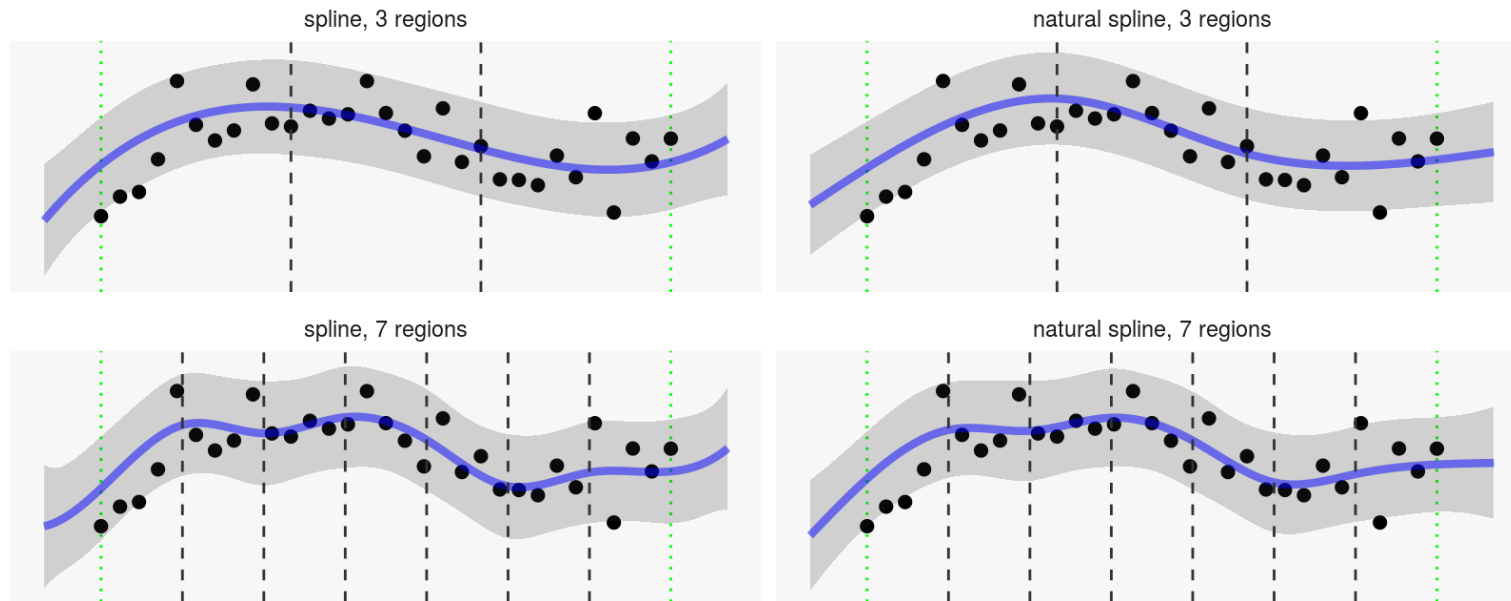
# スプラインと3次自然スプライン

- 次元数と同じ階数の連続微分を課すと、ただの多項式モデルに帰着する
- (次元数 - 1) 階の連続微分を課した区分的多項式モデルのことを「スプライン (Spline)」という
- 3次スプラインがよく使われる
- スプラインに、両端の外側には線形制約（2次以上の項なし）を課したものを「自然スプライン (Natural spline)」という
- 自然スプラインは、両端の予測誤差が大きくなりがちな傾向を緩和する
- 自然スプラインの自由度は、

$$\text{区間数} \times 4\text{変数} - (\text{区間数} - 1) \times 3\text{制約} - 4 = \text{区間数} - 1$$



# スプラインと3次自然スプライン



- 帯は90%の予測区間
- 緑の線がスプラインの両端. その外側では自然スプラインは線形
- 自然スプラインの方が両端での誤差が小さい

# Rでスプライン基底を生成

```
library(splines)
X <- runif(50)
B <- bs(X, knots = c(0.33, 0.67)); head(B, 3) # 3次スプライン
```

```
##           1           2           3           4           5
## [1,] 0.00000000 0.005987736 0.13146215 0.571423120 0.291127
## [2,] 0.60055869 0.199224763 0.01274458 0.000000000 0.000000
## [3,] 0.08561474 0.554658208 0.35185185 0.007875204 0.000000
```

```
N <- ns(X, knots = c(0.33, 0.67)); head(N, 3) # 3次自然スプライン
```

```
##           1           2           3
## [1,] 0.19938265 0.3630249 0.4316047
## [2,] -0.08647356 0.3044277 -0.2052096
## [3,] 0.18511228 0.5194767 -0.3448619
```

- knots は区分点の位置
- bs の自由度は 6 (= 5 変数 + 定数項)
- ns の自由度は 4 (= 2 ノット + 2 両端ノット)

# 3次自然スプラインを用いた特徴抽出

このセクションでは、スプラインをちょっと変わった方法で使っているので注意

- これまで：スプラインを基底として回帰変数に用いる
- ここ　　：スプラインを基底展開に利用する

# 基底展開は特徴抽出でもある

$x \in \mathbb{R}^p$ ,  $H \in \mathbb{R}^{p \times M}$  として、基底  $x^* \in \mathbb{R}^M$  を

$$x^* = H^T x$$

と定義する

$p > M$  であれば、 $x^*$  は  $p$  個の情報を  $M$  個の変数に**集約**したものになる

- 変数が減る  $\Rightarrow$  バイアスが増えて分散が減る
- 既存の知識から適切な集約方法がわかっている時には、バイアスはあまり減らないので特に有用

## 音素認識の例 (5.2.3)

- 説明変数  $x.1, \dots, x.256$ : 各周波数 (1~256) の波の強さ (i.e.  $p = 256$ )
- 被説明変数  $g$ : 音が “aa” か “ao” か (2値分類問題)
- 256変数は多すぎる + 互いに相関が強い  $\Rightarrow$  集約したい
- 3次自然スプラインが使える！

```
##      x.1      x.2      x.3      x.4      x.5      x.6
## 1 12.96705 13.69454 14.91182 18.22292 18.45390 17.25760
## 2 10.95324 11.20585 16.17634 18.59300 17.50922 10.27798
## 3  9.37324 11.29505 17.15139 18.03336 14.95980 14.97031
## 4  9.48477 11.38758 16.74884 17.36141 14.67661 15.02621
```

```
## ....
```

```
##      x.251  x.252  x.253  x.254  x.255  x.256  g
## 1 6.65202 7.69109 6.93683 7.03600 7.01278 8.52197 aa
## 2 8.79901 8.22345 7.63610 8.44448 8.28905 8.04018 aa
## 3 6.46091 4.18584 6.27844 7.73464 7.41363 0.53710 ao
## 4 7.03513 9.24298 8.77743 8.20530 9.75466 8.49344 ao
```

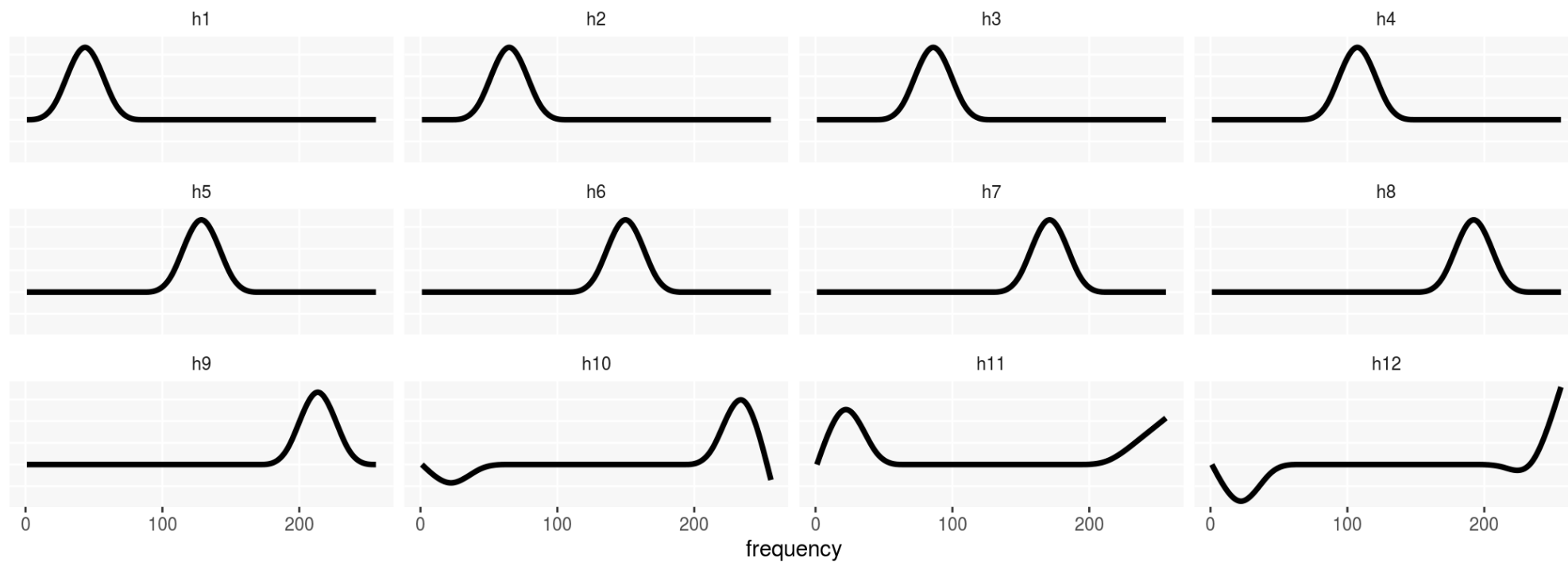
## 音素認識の例 (5.2.3)

1. 集約する変数の数  $M$  を決める (例では  $M = 12$ )
2. 行列  $H$  を 1~256 の周波数に対する、自由度  $M$  のスプライン行列とする
3. 特徴量を  $x^* = H^T x$  に集約する

```
H <- ns(1:256, df = 12)
X <- as.matrix(phenome[, paste("x.", 1:256, sep = "")])
Z <- X %*% H
head(Z, 3)
```

```
##           1           2           3           4           5           6           7
## [1,] 363.6290 310.8161 254.2466 201.9448 187.0347 163.7845 154.1996
## [2,] 315.5019 255.9451 261.6739 253.5809 212.2368 185.8311 181.2271
## [3,] 314.2025 257.6343 227.8591 203.5561 184.4343 162.5885 155.8591
##           8           9          10          11          12
## [1,] 154.6513 159.2664 38.56804 356.4331 -137.5750
## [2,] 183.8734 186.2612 65.72325 342.4778 -117.9531
## [3,] 158.3552 157.6853 54.50231 331.4313 -130.0473
```

## 音素認識の例 (5.2.3)



各スプライン基底は特定の範囲の周波数を重点的に集約している

# 平滑化スプライン



# 平滑化スプライン

$x \in [a, b]$  において  $y$  を予測するモデルについて、次の最適化問題を考える

$$\min_{f \in C^2} \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt$$

- $C^2$ : 連続2階微分可能な関数の集合
- $\lambda > 0$ : 平滑化パラメータ

## トレードオフ

- モデルを複雑にすれば、 $\sum_{i=1}^N (y_i - f(x_i))^2$  は小さくなる
- モデルを複雑にすると、 $\int_a^b (f''(t))^2$  は大きくなる
- $\lambda \rightarrow 0 \Rightarrow f$  は 各データ点を通る曲線
- $\lambda \rightarrow +\infty \Rightarrow f$  は 線形モデル

# 平滑化スプライン

$$\min_{f \in C^2} \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt$$

の解は、「 $(x_1, x_2, \dots, x_N)$  をノットとする**3次自然スプライン**」になる！

- あらゆる関数（2次連続微分可能）の中で、3次自然スプラインが最も「好ましい」
- これを**平滑化スプライン**という.

## 証明 (Exercise 5.7)

$\tilde{g}(x)$  を最適化問題の解とする.  $\tilde{g}$  が、各  $x_i$  において通る点を  $z_i = \tilde{g}(x_i)$  と置く.

一方、 $g(x)$  を  $(x_1, x_2, \dots, x_N)$  をノットとし、かつ  $N$  個の点  $(x_i, z_i)$  を通るような 3 次自然スプラインとする. 3 次自然スプラインの自由度は  $N$  なので、このような  $g$  は必ず見つかる.

この時、全ての  $x \in [a, b]$  において、 $g(x) = \tilde{g}(x)$  であることを示す.

**Step (1):**  $h(x) = \tilde{g}(x) - g(x)$  とすると、 $\int_a^b g''(x)h''(x)dx = 0$ .

部分積分によって、

$$\int_a^b g''(x)h''(x)dx = [g''(x)h'(x)]_a^b - \int_a^b g'''(x)h'(x)dx$$

ここで、 $g$  は自然スプラインなので、 $g''$  は両端ではゼロになる. したがって、第 1 項はゼロ.

## 証明 (Exercise 5.7)

また、 $g$ は3次の区分的多項式であるから、 $g'''(x)$ は各区間内で定数になる。そこで、区間  $(x_j, x_{j+1})$  における  $g'''(x)$  の値を  $g_j'''$  と置く。すると

$$\begin{aligned}\int_a^b g''(x)h''(x)dx &= - \int_a^b g'''(x)h'(x)dx \\ &= - \sum_{j=1}^{N-1} g_j''' \int_{x_j}^{x_{j+1}} h'(x)dx \\ &= - \sum_{j=1}^{N-1} g_j''' [h(x)]_{x_j}^{x_{j+1}} \\ &= - \sum_{j=1}^{N-1} g_j''' (h(x_{j+1}) - h(x_j)) = 0\end{aligned}$$

最後の等式は、 $h(x_j) = \tilde{g}(x_j) - g(x_j) = z_j - z_j = 0$  を用いている。

## 証明 (Exercise 5.7)

**Step (2):**  $\int_a^b \tilde{g}''(x)^2 dx \geq \int_a^b g''(x)^2 dx.$

$$\tilde{g}''(x) = h''(x) + g''(x)$$

$$\tilde{g}''(x)^2 = h''(x)^2 + 2h''(x)g''(x) + g''(x)^2$$

$$\int_a^b \tilde{g}''(x)^2 dx = \int_a^b h''(x)^2 dx + \int_a^b 2h''(x)g''(x) dx + \int_a^b g''(x)^2 dx$$

$$\int_a^b \tilde{g}''(x)^2 dx = \int_a^b h''(x)^2 dx + \int_a^b g''(x)^2 dx$$

$$\int_a^b \tilde{g}''(x)^2 dx \geq \int_a^b g''(x)^2 dx$$

また、両辺が等しいのは  $h(x)$  が全ての  $x \in [a, b]$  においてゼロの時のみ.

## 証明 (Exercise 5.7)

### Step (3)

一方、 $\tilde{g}$  は最適化問題の解であるので、 $\int_a^b \tilde{g}''(x)^2 dx \leq \int_a^b g''(x)^2 dx$ . よって、 $\int_a^b \tilde{g}''(x)^2 dx = \int_a^b g''(x)^2 dx$ . したがって、全ての  $x \in [a, b]$  について、 $h(x) = \tilde{g}(x) - g(x) = 0$ . (証明終)

# 平滑化スプラインの行列表現

3次自然スプラインの基底を  $(N_1, \dots, N_N)$  とおくと

$$f(x) = \sum_{j=1}^N N_j(x) \theta_j$$

行列  $\mathbf{N}$ ,  $\mathbf{\Omega}$  を、  $\mathbf{N}_{ij} = N_j(x_i)$ ,  $\mathbf{\Omega}_{jk} = \int N_j''(t) N_k''(t) dt$  と定義すると、最小化問題は

$$\min_{\theta} (y - \mathbf{N}\theta)^T (y - \mathbf{N}\theta) + \lambda \theta^T \mathbf{\Omega} \theta.$$

この解, 予測値は

$$\hat{\theta} = (\mathbf{N}^T \mathbf{N} + \lambda \mathbf{\Omega})^{-1} \mathbf{N}^T y,$$

$$\hat{y} = \mathbf{N} (\mathbf{N}^T \mathbf{N} + \lambda \mathbf{\Omega})^{-1} \mathbf{N}^T y,$$

と書ける.

# 平滑化スプラインの自由度

推定値が  $\hat{y} = \mathbf{H}y$  と書けるタイプ（線形平滑化）のモデルについて、その自由度を

$$\text{trace}(\mathbf{H})$$

で定義する.

平滑化スプラインの場合：  $\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}^T \mathbf{N} + \lambda \mathbf{\Omega})^{-1} \mathbf{N}^T$  と置いて、

$$\text{df}_\lambda = \text{trace}(\mathbf{S}_\lambda)$$

性質：

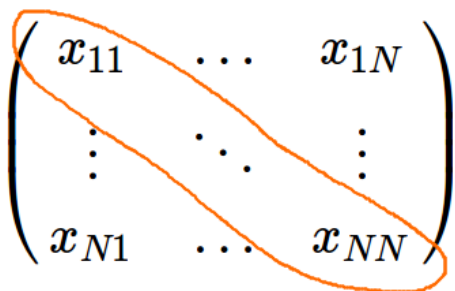
1.  $\text{df}_\lambda$  は  $\lambda$  が大きくなるにつれて減少していく
2.  $\lambda \rightarrow 0$  のとき、 $\text{df}_\lambda \rightarrow N$
3.  $\lambda \rightarrow \infty$  のとき、 $\text{df}_\lambda \rightarrow 2$



# Traceオペレータ

- 正方行列 $\mathbf{A}$ について、 $\text{trace}(\mathbf{A})$ はその対角要素の和：

$$\text{trace}(\mathbf{A}) = \sum_i A_{ii}$$


$$\begin{pmatrix} x_{11} & \dots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NN} \end{pmatrix}$$

# Traceオペレータ

- $\text{trace}(\mathbf{A})$  は行列 $\mathbf{A}$ の固有値の和に等しい (証明略)
- 行列  $\mathbf{A}(n \times m)$ ,  $\mathbf{B}(m \times n)$ について、  $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$  が成り立つ (証明略)
- $\lambda = 0$ のとき、  $\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}^T \mathbf{N})^{-1} \mathbf{N}^T$ . よって、  
$$\text{trace}(\mathbf{S}_\lambda) = \text{trace}(\mathbf{N}(\mathbf{N}^T \mathbf{N})^{-1} \mathbf{N}^T) = \text{trace}(\mathbf{N}^T \mathbf{N}(\mathbf{N}^T \mathbf{N})^{-1}) = \text{trace}(\mathbf{I}_N) = N.$$

## **Reinsch形式**

$$\mathbf{S}_\lambda = (\mathbf{I} + \lambda \mathbf{K})^{-1} \text{ と書ける}$$

ただし、 $\mathbf{K}$ は $\lambda$ に依存しない

導出 (Exercise 5.9):

$\mathbf{N}$ は $N$ 次の正方行列.  $\mathbf{N}$ が正則なら、

$$\begin{aligned}\mathbf{S}_\lambda^{-1} &= (\mathbf{N}^T)^{-1}(\mathbf{N}^T \mathbf{N} + \lambda \mathbf{\Omega})\mathbf{N}^{-1} \\ &= \mathbf{I} + \lambda (\mathbf{N}^T)^{-1} \mathbf{\Omega} \mathbf{N}^{-1} \\ &= \mathbf{I} + \lambda \mathbf{K}\end{aligned}$$

ここで  $\mathbf{K} = (\mathbf{N}^T)^{-1} \mathbf{\Omega} \mathbf{N}^{-1}$ .

# $\mathbf{S}_\lambda$ の固有値

$d_k$  を  $\mathbf{K}$ の固有値とすると

- $\rho_k(\lambda) = (1 + \lambda d_k)^{-1}$ は $\mathbf{S}_\lambda$ の固有値
- $(d_1, \dots, d_N)$ の中には少なくとも2つのゼロが含まれる
- $d_k \geq 0$

上記から次の性質が示される.

- $\mathbf{S}_\lambda$  の固有値は $\lambda$ について単調減少する
- $\text{df}_\lambda = \text{trace}(\mathbf{S}_\lambda) = \sum_k \rho_k(\lambda)$  なので、自由度も $\lambda$ について単調減少する
- $\lambda \rightarrow 0$ のとき、 $\text{df}_\lambda \rightarrow N$
- $\lambda \rightarrow \infty$ のとき、 $\text{df}_\lambda \rightarrow 2$

# 固有値について

- 正方行列 $\mathbf{A}$  について、ベクトル $v$  とスカラー $\rho$  が

$$\mathbf{A}v = \rho v$$

を満たすとき、 $\rho$ を $\mathbf{A}$ の**固有値**、 $v$ を対応する**固有ベクトル**と呼ぶ.

- $\rho \neq 0$ が $\mathbf{A}$ の固有値で、かつ $\mathbf{A}$ が正則なら、 $\rho^{-1}$ は $\mathbf{A}^{-1}$ の固有値 (証明略)
- $\mathbf{A}$ が対称行列の場合、固有値は全て実数で、かつすべての固有ベクトルが互いに直交するように選ぶことができる (証明略)
- $\mathbf{A}$ が対称行列の場合、 $\mathbf{A}$ は直交行列 $\mathbf{Q}$ と対角行列 $\mathbf{\Lambda}$ を用いて次のように分解することができる (証明略). これを**固有値分解**という.

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$$

# 固有値について

- 正方行列**A**について、任意のベクトル $x$ について

$$x^T A x \geq 0$$

が成り立つ場合、**A**は**半正定値** (positive semidefinite)という.

- 同様に  $x^T A x \leq 0$  なら**半負定値** (negative semidefinite),  $x^T A x > 0$  なら**正定値** (positive definite),  $x^T A x < 0$  なら**負定値** (negative definite)という.
- 半正定値行列の固有値は全て非負 (証明略)

## $\rho_k(\lambda) = (1 + \lambda d_k)^{-1}$ は $\mathbf{S}_\lambda$ の固有値

$d_k$  は  $\mathbf{K}$  の固有値なので、 $v$  を対応する固有ベクトルとすると  $d_k \mathbf{K} = d_k v$  と書ける。

$$\mathbf{S}_\lambda^{-1} v = (\mathbf{I} + \lambda \mathbf{K}) v = v + \lambda K v = v + \lambda d_k v = (1 + \lambda d_k) v.$$

よって、 $(1 + \lambda d_k)$  は  $\mathbf{S}_\lambda^{-1}$  の固有値 (固有ベクトルは  $v$ )。したがって、 $(1 + d_k)^{-1}$  は  $\mathbf{S}_\lambda$  の固有値。

$(d_1, \dots, d_N)$ の中には少なくとも2つのゼロが含まれる

$(N_1, \dots, N_N)$ はスプライン基底なので、定数項と1次の項を含む. 簡単化のため、 $N_0, N_1$ を定数項と1次の項とすると、 $N_1'' = 0, N_2'' = 0$ . したがって、全ての $j$ について、 $\Omega_{1,j} = \Omega_{2,j} = \Omega_{j,1} = \Omega_{j,2} = 0$

ここで、 $v_1 = (1, 0, 0, \dots)'$ ,  $v_2 = (0, 1, 0, \dots)'$  とすれば、

$$\Omega v_1 = 0 = 0 \times v_1, \Omega v_2 = 0 = 0 \times v_2$$

となるから、 $\Omega$ は固有値ゼロを2つ以上持つ.

さらに、 $\mathbf{K} = (\mathbf{N}^T)^{-1} \Omega \mathbf{N}^{-1}$ の固有値を考える. ベクトル $u_1 = \mathbf{N} v_1$ とすると、

$$\mathbf{K} u_1 = (\mathbf{N}^T)^{-1} \Omega \mathbf{N}^{-1} \mathbf{N} v_1 = (\mathbf{N}^T)^{-1} \Omega v_1 = 0 \times u_1$$

となる. 同様に、 $u_2 = \mathbf{K} v_2$  についても  $\mathbf{K} u_2 = 0 \times u_2$  なので、 $\mathbf{K}$ は固有値ゼロを2つ以上持つ.



$$d_k \geq 0$$

各要素が  $N_j''(x)$  になっているようなベクトル  $N''$  を考えると  $\mathbf{\Omega} = \int N''(t)N''(t)^T dt$  と書ける. 任意のベクトル  $v$  について、

$$v^T \mathbf{\Omega} v = \int v^T N''(t) N''(t)^T v dt = \int (N''(t)^T v)^2 dt \geq 0$$

したがって  $\mathbf{\Omega}$  は半正定値.

$$\mathbf{K} = (\mathbf{N}^T)^{-1} \mathbf{\Omega} \mathbf{N}^{-1} = (\mathbf{N}^T)^{-1} \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \mathbf{N}^{-1} = \mathbf{P}^T \mathbf{P}$$

ここで  $\mathbf{P} = \mathbf{\Lambda}^{1/2} \mathbf{Q}^T \mathbf{N}^{-1}$ . 任意のベクトル  $v$  について、

$$v^T \mathbf{K} v = v^T \mathbf{P}^T \mathbf{P} v = (\mathbf{P} v)^T (\mathbf{P} v) \geq 0$$

したがって  $\mathbf{K}$  も半正定値なので、その固有値は非負.

# 平滑化スプラインの自由度（まとめ）

1.  $\text{df}_\lambda = \text{trace}(\mathbf{S}_\lambda)$  で自由度を定義する
2.  $\text{df}_\lambda$  は  $\lambda$  が大きくなるにつれて減少していく
3.  $\lambda \rightarrow 0$  のとき、 $\text{df}_\lambda \rightarrow N$
4.  $\lambda \rightarrow \infty$  のとき、 $\text{df}_\lambda \rightarrow 2$

# λの自動決定

バイアスと分散のトレードオフ

1. 自由度が高い  $\Rightarrow$  複雑なモデルを表現  $\Rightarrow$  バイアスが小さいが、分散が大きい
2. 自由度が低い  $\Rightarrow$  シンプルなモデル  $\Rightarrow$  バイアスが大きい分散が小さい

何らかの基準で両者のバランスを取るよう自由度を調整する

**Leave-one-out cross validation**

$$\text{CV}(\hat{f}) = \frac{1}{N} \sum_{i=1}^N \left( y_i - \hat{f}_{\lambda}^{(-i)}(x_i) \right)^2$$

- $\hat{f}_{\lambda}^{(-i)}$  は*i*番目のデータを除いたモデルを表す
- 1つだけデータを除いてテスト用データとしてモデルの評価に利用する
- $\text{CV}(\hat{f})$ が小さくなるようにλを調整する

# Leave-one-out cross validation の計算

各 $i$ について、個別に $\hat{f}_{\lambda}^{(-i)}$ を計算するのは時間がかかる

次の公式を用いると、計算が効率的になる

$$\begin{aligned}\text{CV}(\hat{f}) &= \frac{1}{N} \sum_{i=1}^N \left( y_i - \hat{f}_{\lambda}^{(-i)}(x_i) \right)^2 \\ &= \frac{1}{N} \sum_{i=1}^N \left( \frac{y_i - \hat{f}_{\lambda}(x_i)}{1 - S_{\lambda}(i, i)} \right)^2\end{aligned}$$

証明の参考 : <http://robjhyndman.com/hyndsight/loocv-linear-models/>

# スプラインとロジスティック回帰

2値分類問題にもスプラインは使える.

$$\Pr(Y = 1|X = x) = p(x) = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

とモデル化する.

正則項付きの最大化問題は

$$\max_{f \in C^2} \sum_{i=1}^N \left[ y_i f(x_i) - \log(1 + e^{f(x_i)}) \right] - \frac{1}{2} \lambda \int f''(t)^2 dt$$

と書ける. 回帰分析のケースと同様の手順で、最適な  $f(x)$  は3次自然スプラインであることを示すことができる.

そこで、スプライン基底を用いて  $f(x) = \sum_{j=1}^N N_j(x) \theta_j$  と置き、最適な  $(\theta_1, \dots, \theta_N)$  を求める問題に帰着する