数学特別講義：機械学習の数理 (note-01)

線形回帰モデルとカーネル法

1. カーネル関数を用いた回帰分析の方法

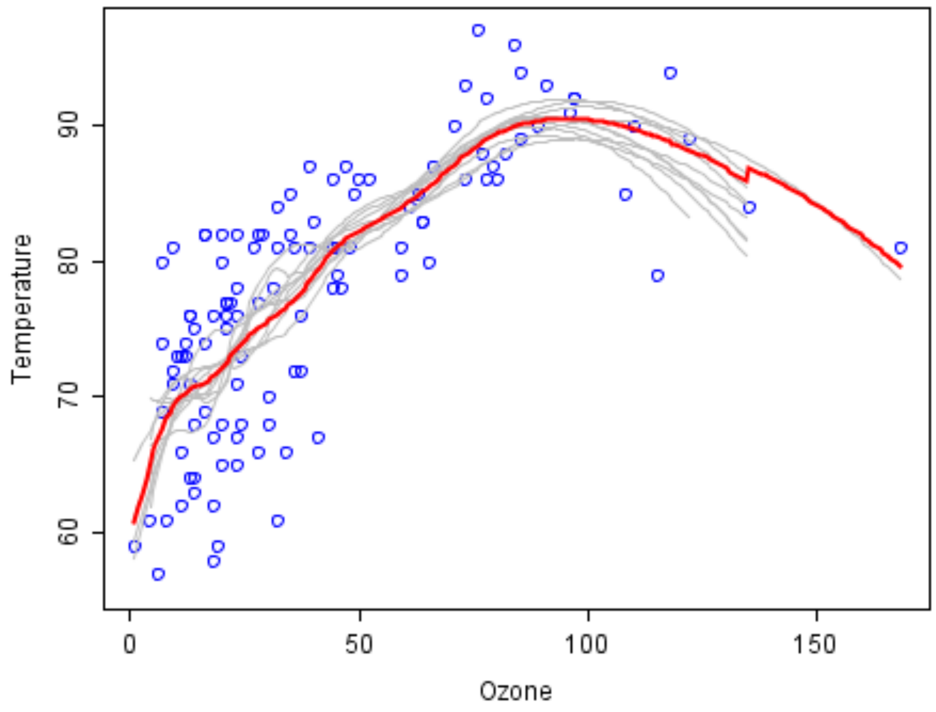
2. 正則化と交差検証法

3. カーネル関数と再生核ヒルベルト空間

4. 平滑化スプラインとカーネル法

回帰分析

- データ：$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n), \quad \boldsymbol{x}_i \in \mathbb{R}^d,\ y_i \in \mathbb{R}$.
- y を $\boldsymbol{x}$ の関数で説明，予測.
- 用語の説明
  - $\boldsymbol{x}$：入力（独立変数・説明変数）
  - y：出力（従属変数・目的変数）

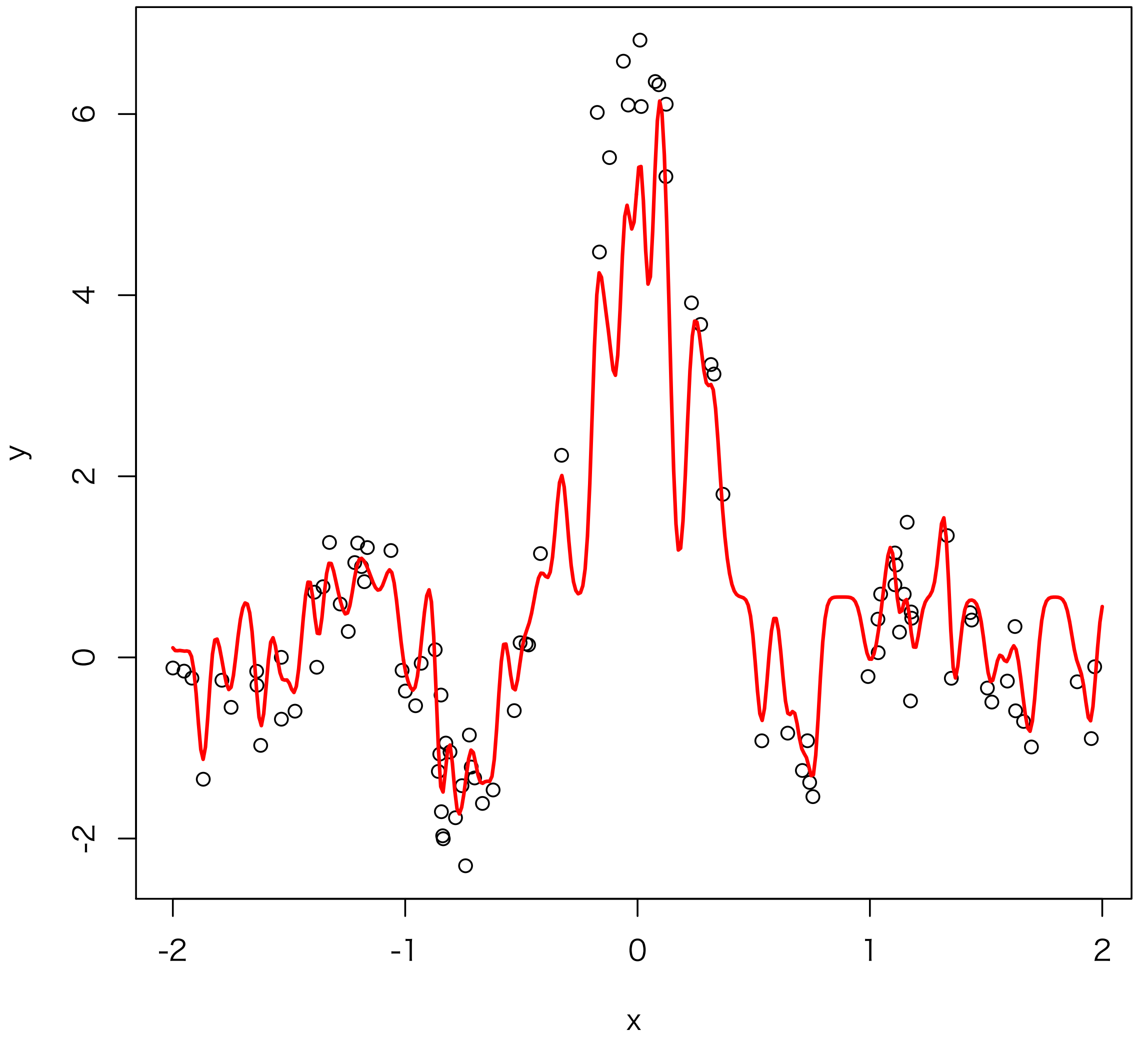


統計モデルの設定

- 単純な統計モデルでは対応できない（かもしれない）
  $\longrightarrow$ 複雑なモデルを使いたい

- 複雑なモデルを使うと
  - $*$ 計算が大変
  - $*$ データへの過剰適合（**overfitting**，過学習）が心配

過剰適合



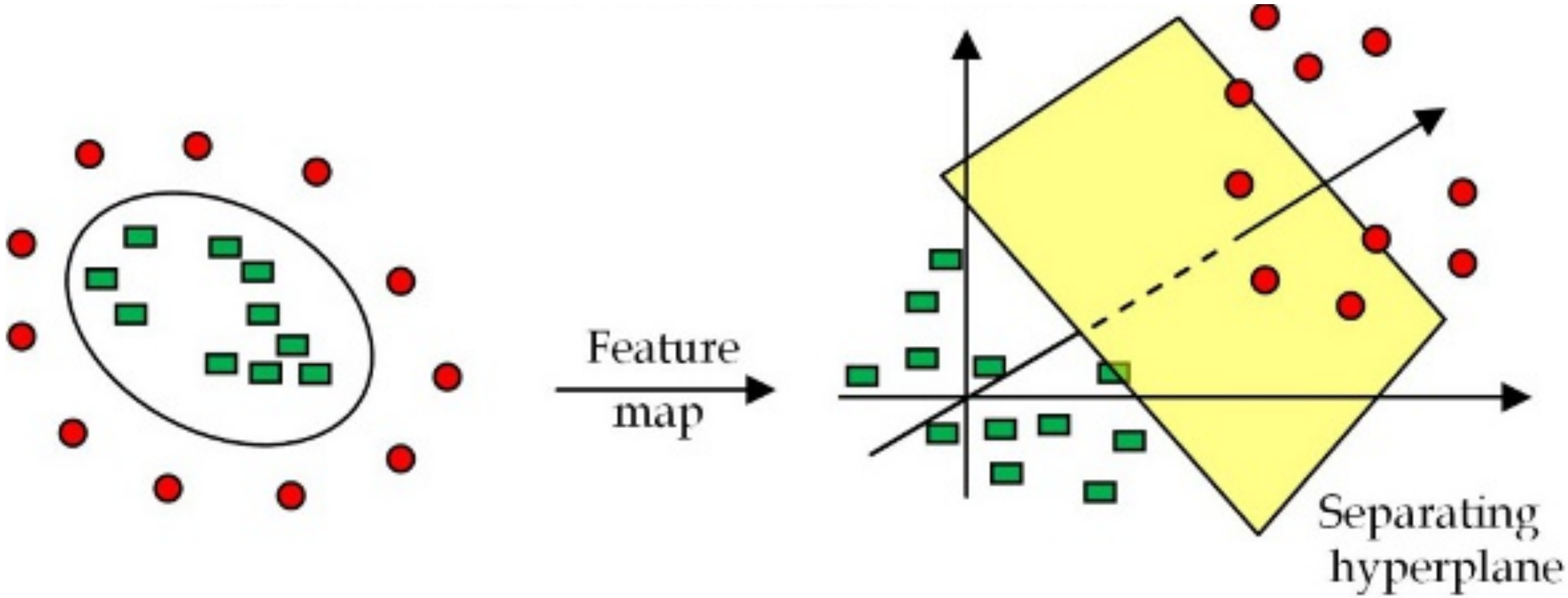
- データ点に適合し過ぎている．予測精度は低い．

表現力が高いモデルの構成

- 例：データ $x \in \mathbb{R}$ を高次元空間 $\mathbb{R}^{1000}$ に非線形変換

$$x \in \mathbb{R} \;\longmapsto\; \boldsymbol{\phi}(x) \in \mathbb{R}^{1000}$$

- 高次元空間における線形モデルでデータ解析



「複雑な統計モデル」を「非線形変換＋線形モデル」で実現

- 適切な変換 $\boldsymbol{x} \longmapsto \phi(\boldsymbol{x})$ を選ぶと，計算量はあまり増えない.
- 過学習には「正則化」**&**「交差検証法」で対処.

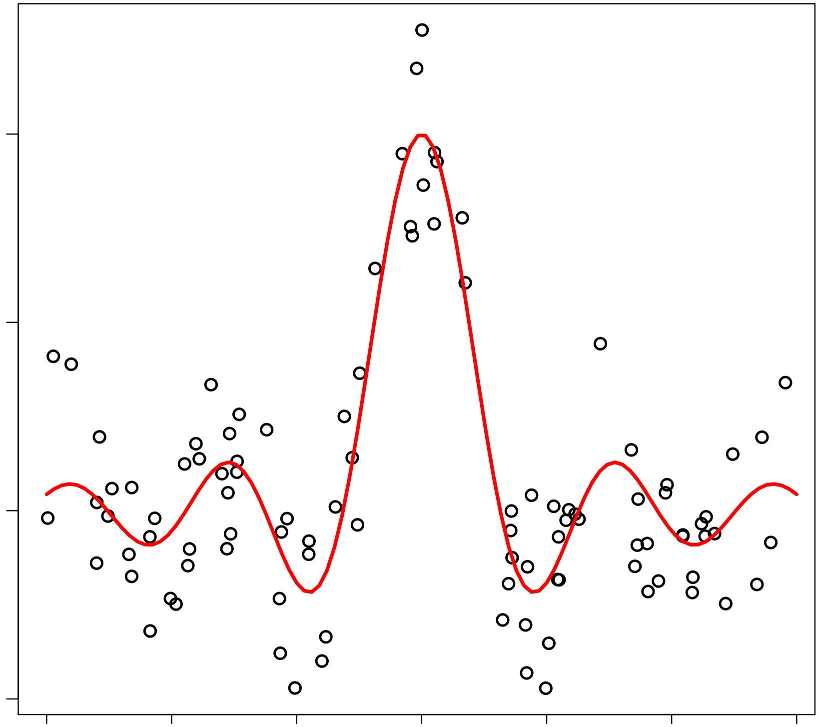
— カーネル関数を用いた回帰分析の方法 —

- 最小２乗法のカーネル表現

例：多項式回帰

データの例：

$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$.



- 多項式回帰：$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_{D-1} x^{D-1} + \varepsilon$
  で当てはめ．（ε：ノイズ項）

  パラメータ $(a_0, a_1, \ldots, a_{D-1}) \in \mathbb{R}^D$ に関して線形なので線形回帰モデル．

- 多項式回帰の解釈：
  - $*$ データ x を，以下の $\phi(x)$ で高次元にマップ.

$$x \in \mathbb{R} \longmapsto \boldsymbol{\phi}(x) = (\phi_0(x), \phi_1(x), \ldots, \phi_{D-1}(x))^T \in \mathbb{R}^D,$$

$$\phi_k(x) = x^k, \quad k = 0, 1, \ldots, D-1$$

  - $*$ データ $(\boldsymbol{\phi}(x_1), y_1), \ldots, (\boldsymbol{\phi}(x_n), y_n)$ の関数関係を高次元空間 $\mathbb{R}^D$ 上の線形モデルで推定：

$$y = a_0\phi_0 + a_1\phi_1 + a_2\phi_2 + \cdots + a_{D-1}\phi_{D-1} + \varepsilon$$

$$\longrightarrow \text{ 推定パラメータ } \widehat{\boldsymbol{a}} = (\widehat{a}_0, \ldots, \widehat{a}_{D-1})^T$$

  - $*$ $\widehat{f}(x) = \widehat{\boldsymbol{a}}^T\boldsymbol{\phi}(x)$ を使って x に対する y の値を予測.

最小２乗法：デザイン行列を用いた表現

- デザイン行列 $\boldsymbol{\Phi} = (\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_n)) \in \mathbb{R}^{D\times n}$.
- 観測データ $\boldsymbol{y} = (y_1, \ldots, y_n)^T$.

$$\sum_{i=1}^{n}(y_i - \boldsymbol{a}^T\boldsymbol{\phi}(x_i))^2 = \|\boldsymbol{y} - \boldsymbol{\Phi}^T\boldsymbol{a}\|_2^2 \;\longrightarrow\; \boldsymbol{a} \text{ について最小化}$$

$\mathrm{rank}\,\boldsymbol{\Phi} = D$ のとき：

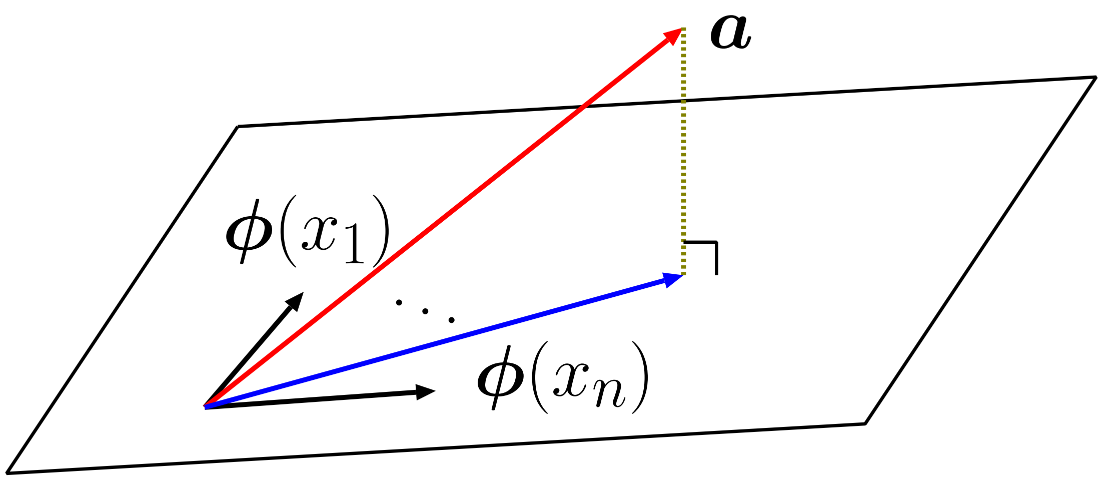
最小２乗法の解：$\widehat{\boldsymbol{a}} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{\Phi}\boldsymbol{y}$

最小２乗法：カーネル関数を用いた表現

$$\min_{\boldsymbol{a}\in\mathbb{R}^D} \quad \sum_{i=1}^{n}(y_i - \boldsymbol{a}^T\boldsymbol{\phi}(x_i))^2$$

- $\boldsymbol{a} \in \mathrm{span}\{\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_n)\}$ の範囲で考えれば十分.

直交成分は２乗損失に
影響しない.



- $\boldsymbol{a} = \sum_{j=1}^{n} \beta_j \boldsymbol{\phi}(x_j) = \Phi\boldsymbol{\beta}, \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$
  $\longrightarrow$ 2乗誤差に代入

$$\sum_{i=1}^{n} (y_i - \boldsymbol{\phi}(x_i)^T \boldsymbol{a})^2 = \|\boldsymbol{y} - \boldsymbol{\Phi}^T \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

極値条件： $\boldsymbol{\Phi}^T \boldsymbol{\Phi}\boldsymbol{\Phi}^T \underbrace{\boldsymbol{\Phi}\widehat{\boldsymbol{\beta}}}_{\widehat{\boldsymbol{a}}} = \boldsymbol{\Phi}^T \boldsymbol{\Phi}\boldsymbol{y}$

- $k(x,x') := \boldsymbol{\phi}(x)^T\boldsymbol{\phi}(x') \in \mathbb{R}$ とおく　(カーネル関数).
  行列 $K \in \mathbb{R}^{n\times n}$ を $K = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$ とおく.

$$K_{ij} = \boldsymbol{\phi}(x_i)^T\boldsymbol{\phi}(x_j) = k(x_i,x_j),$$

$$\text{極値条件：}\ \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\widehat{\boldsymbol{\beta}} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{y} \iff K^2\widehat{\boldsymbol{\beta}} = K\boldsymbol{y}$$

- 極値条件の式を解く：K, $\boldsymbol{y}$ から $\widehat{\boldsymbol{\beta}}$ を求める.

$$\widehat{f}(x) = \boldsymbol{\phi}(x)^T\underbrace{\sum_{i=1}^{n}\boldsymbol{\phi}(x_i)\widehat{\beta}_i}_{\widehat{\boldsymbol{a}}} = \sum_{i=1}^{n}k(x,x_i)\widehat{\beta}_i$$

グラム行列

データ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, 関数 $k(\boldsymbol{x}, \boldsymbol{x}')$.

$$\text{グラム行列}: \quad K = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

要素で書くと $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

グラム行列の性質：非負定値性

$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}'),\ \mathbf{\Phi} = (\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_n))$ とすると

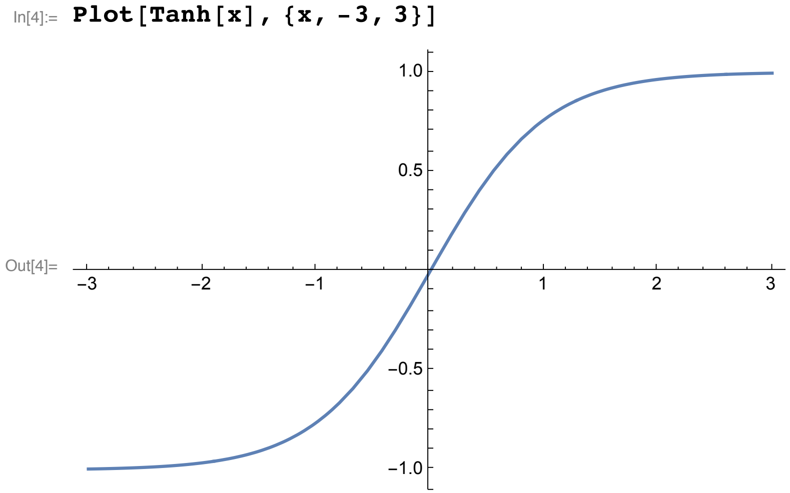
$$K = \mathbf{\Phi}^T \mathbf{\Phi}.$$

K は対称非負定値行列

$$\forall \boldsymbol{c} \in \mathbb{R}^n,\ \boldsymbol{c}^T K \boldsymbol{c} \geq 0, \quad K \succeq O \text{と書く}.$$

$$\boldsymbol{c}^T K \boldsymbol{c} = \boldsymbol{c}^T \mathbf{\Phi}^T \mathbf{\Phi} \boldsymbol{c} = \|\mathbf{\Phi} \boldsymbol{c}\|_2^2 \geq 0$$

- $k(\boldsymbol{x}, \boldsymbol{x}') = \tanh(1 + \boldsymbol{x}^T \boldsymbol{x}'), \ \tanh(z) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$
カーネル関数でない.



$*$ $\boldsymbol{x}_1 = (1, 0, \ldots, 0)^T, \boldsymbol{x}_2 = (2, 0, \ldots, 0)$ とすると

$$K = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) \end{pmatrix} \doteq \begin{pmatrix} 0.96403 & 0.99506 \\ 0.99506 & 0.99991 \end{pmatrix}.$$

固有値は $1.97718, -0.0132481$. 非負定値でない.

カーネル回帰分析の方法

- データ：$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n), \quad \boldsymbol{x}_i \in \mathbb{R}^d,\ y \in \mathbb{R}$.
- 統計モデル：$y = \boldsymbol{a}^T \boldsymbol{\phi}(\boldsymbol{x}) + \varepsilon$

カーネル回帰分析

1. カーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}')$ を定義.
2. グラム行列 K, $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, n$ を計算.
3. $K^2 \widehat{\boldsymbol{\beta}} = K \boldsymbol{y}$ を $\widehat{\boldsymbol{\beta}}$ について解く.
4. 回帰関数 $\widehat{f}(x) = \sum_{i=1}^{n} k(\boldsymbol{x}, \boldsymbol{x}_i) \widehat{\beta}_i$

以下は同じ推定：

- 線形回帰モデル $y = a_1\phi_1(\boldsymbol{x}) + \cdots + a_D\phi_D(\boldsymbol{x}) + \varepsilon$ で最小２乗推定.
- カーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}')$ でカーネル回帰.

$\boldsymbol{\phi}(x)$ ではなく $k(x, x')$ を使う理由？

カーネル回帰分析の特徴

$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}')$ だけから推定量を計算できる.

- 写像先 $\mathbb{R}^D$ が無限次元 $(D = \infty)$ でも，$k(\boldsymbol{x}, \boldsymbol{x}')$ が簡単に計算できれば **O.K.**

  例：ガウシアンカーネル **(**後述**)**

- **1** 次式モデル $y = \boldsymbol{a}^T \boldsymbol{x} + \varepsilon$ とほぼ同じ計算手順．表現力が大幅アップ！

カーネル関数の例： $\boldsymbol{x} \in \mathbb{R}^d \longmapsto \boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^D$.

- 線形カーネル：$D = d$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}', \quad (\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x})$$

統計モデル： $y = \boldsymbol{a}^T \boldsymbol{\phi}(\boldsymbol{x}) + \varepsilon = \boldsymbol{a}^T \boldsymbol{x} + \varepsilon$

- 多項式カーネル：$D = \frac{(\gamma+d)!}{\gamma!\,d!}$

$$k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^T \boldsymbol{x}')^\gamma, \quad \gamma = 1, 2, 3, \ldots$$

統計モデル：$y = \boldsymbol{a}^T \boldsymbol{\phi}(\boldsymbol{x}) + \varepsilon$.

$\boldsymbol{\phi}(\boldsymbol{x})$：$\gamma$ 次以下のすべての単項式からなるベクトル.

$d = 2,\ \gamma = 2$ のとき

$$\begin{aligned}
\boldsymbol{\phi}(\boldsymbol{x}) &= (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)^T. \\
\boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{z}) &= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\
&= (1 + x_1z_1 + x_2z_2)^2
\end{aligned}$$

- ガウシアンカーネル：$D = \infty$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\sigma \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\}, \quad \sigma > 0$$

$d = 1$, $\sigma = 1$のとき：

$$\boldsymbol{\phi}(x) = (\phi_0(x), \phi_1(x), \phi_2(x), \cdots),$$

$$\phi_j(x) = \frac{x^j e^{-x^2/2}}{\sqrt{j!}}, \quad x \in \mathbb{R}$$

- R言語によるカーネル回帰

```
> # not run. x: データ行列 (サイズ = サンプル数 * 次元)
>
> # 多項式カーネル(degree=2)
> K <- kernelMatrix(polydot(degree=2),x)
>
> # ガウシアンカーネル(sigma=3)
> K <- kernelMatrix(rbfdot(sigma=3),x)
```

- 初期設定

```
> library(kernlab)   #ライブラリ読込
> deg <- 6           #多項式カーネルの次数を設定
> n <- 100           #データ数 n=100
```

- データの生成

```
> x <- matrix(runif(n,min=-3,max=3))
> y <- sin(pi*x)/x+rnorm(n,sd=0.4)
```

- 多項式カーネル (polydot) のグラム行列の生成

```
> K <- kernelMatrix(polydot(degree=deg),x)
```

- 推定量の計算（２乗誤差の最小化）

```
> #２乗誤差とその勾配
> f <- function(z)sum((K%*%z-y)^2)
> g <- function(z)as.vector(2*K%*%(K%*%z-y))
> #２乗誤差の最小化
> beta <- optim(rep(0,n),f,g,method="BFGS")$par
```

- 予測値の計算

```
> px <- matrix(seq(-3,3,l=300)) #予測点
> # px上でのyの予測値
> py<-kernelMult(polydot(degree=deg),px,x,beta)
```

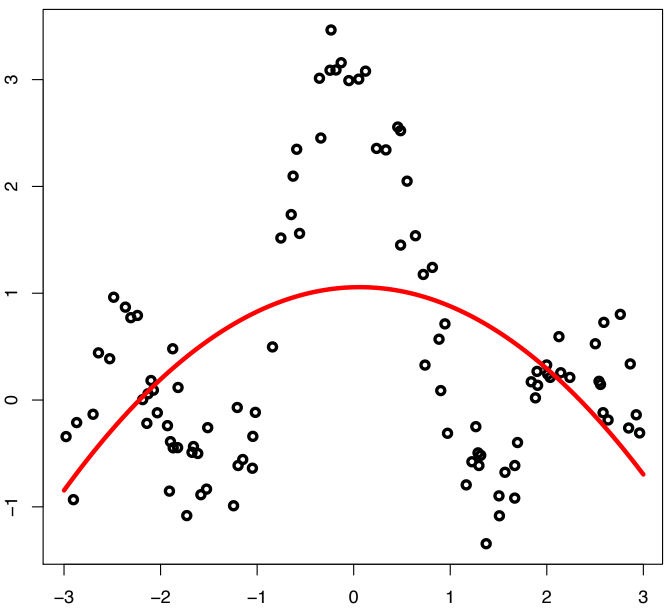
補足：`f<-kernelMult(kernel,x1,x2,beta)` は

$f_i = \sum_{j=1}^{n} k(x1_i, x2_j)\beta_j$ を計算.

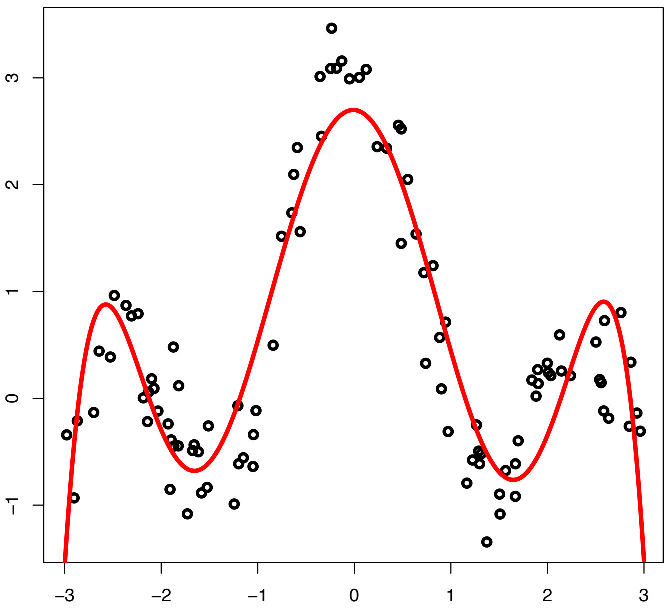
- R-code：２乗損失の最小化でパラメータ $\widehat{\beta}$ を計算.
  (`beta <- optim(···)` の行)
- 一般化逆行列（`MASS::ginv`）でも計算可だが数値的に不安定.

多項式カーネルによる回帰分析の結果：

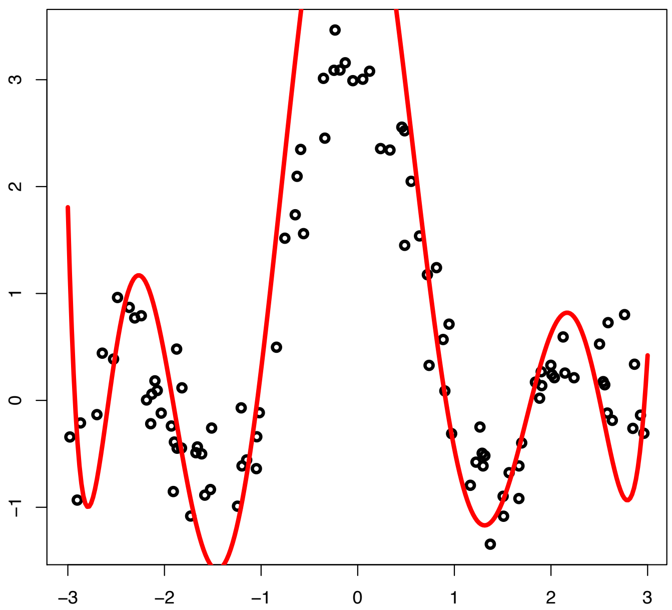
```
> plot(x,y,lwd=3);lines(px,py,lwd=4,col=2) #プロット
```



degree= 2



degree= 6



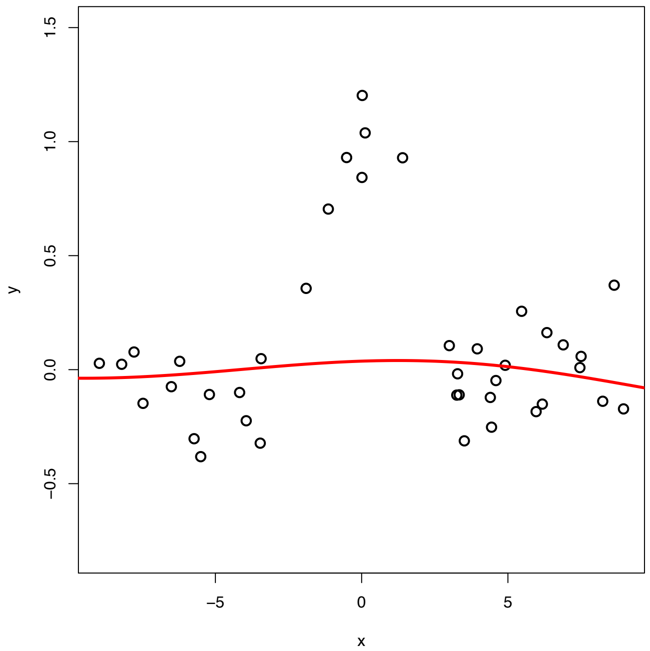
degree= 9

— データへの過剰適合と正則化 —

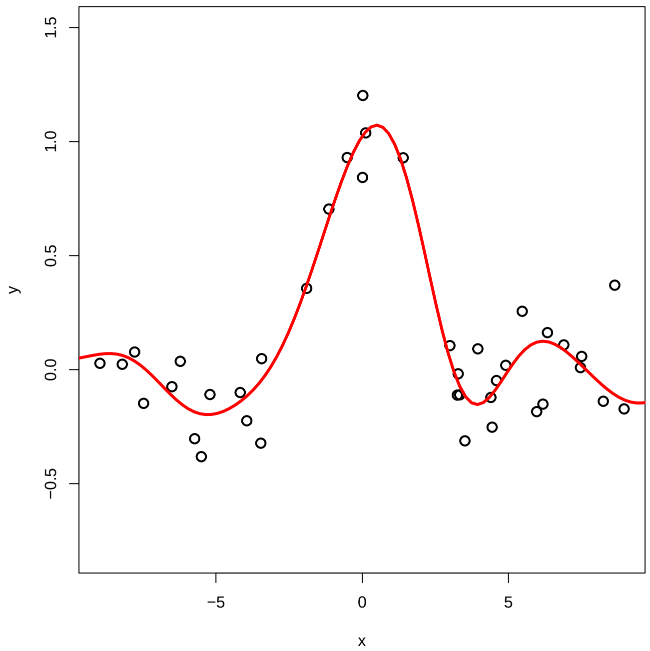
データへの過剰適合（過学習）

- 単純なモデルでは対応できなさそうなデータの解析
  $\longrightarrow$ 自由度の大きなモデルを使う.
- モデルの自由度が大きすぎてもうまくいかない.

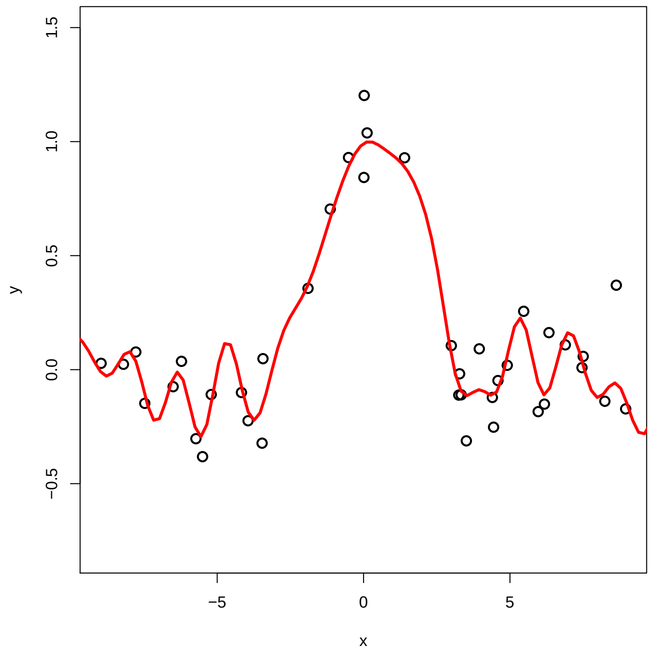
自由度：小



自由度：中



自由度：大



正則化：モデルの自由度を調整

考え方：自由度の大きいモデル ＋ パラメータを適切に制約

データ：　$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$

線形回帰モデル：　$y = \boldsymbol{a}^T \boldsymbol{\phi}(\boldsymbol{x}) + b + \varepsilon$

$$\iff \quad y = \sum_{j=1}^{n} \beta_j k(\boldsymbol{x}_j, \boldsymbol{x}) + b + \varepsilon, \quad \boldsymbol{a} = \sum_{j=1}^{n} \beta_j \boldsymbol{\phi}(\boldsymbol{x}_j)$$

正則化：$\|\boldsymbol{a}\|_2^2 = \|\sum_{j=1}^{n} \beta_j \boldsymbol{\phi}(\boldsymbol{x}_j)\|_2^2$ の大きさを制約する.

$$\|\boldsymbol{a}\|_2^2 = \sum_{i,j} \beta_i \beta_j \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_i) = \sum_{i,j} \beta_i \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

正則化パラメータ $\lambda > 0$.

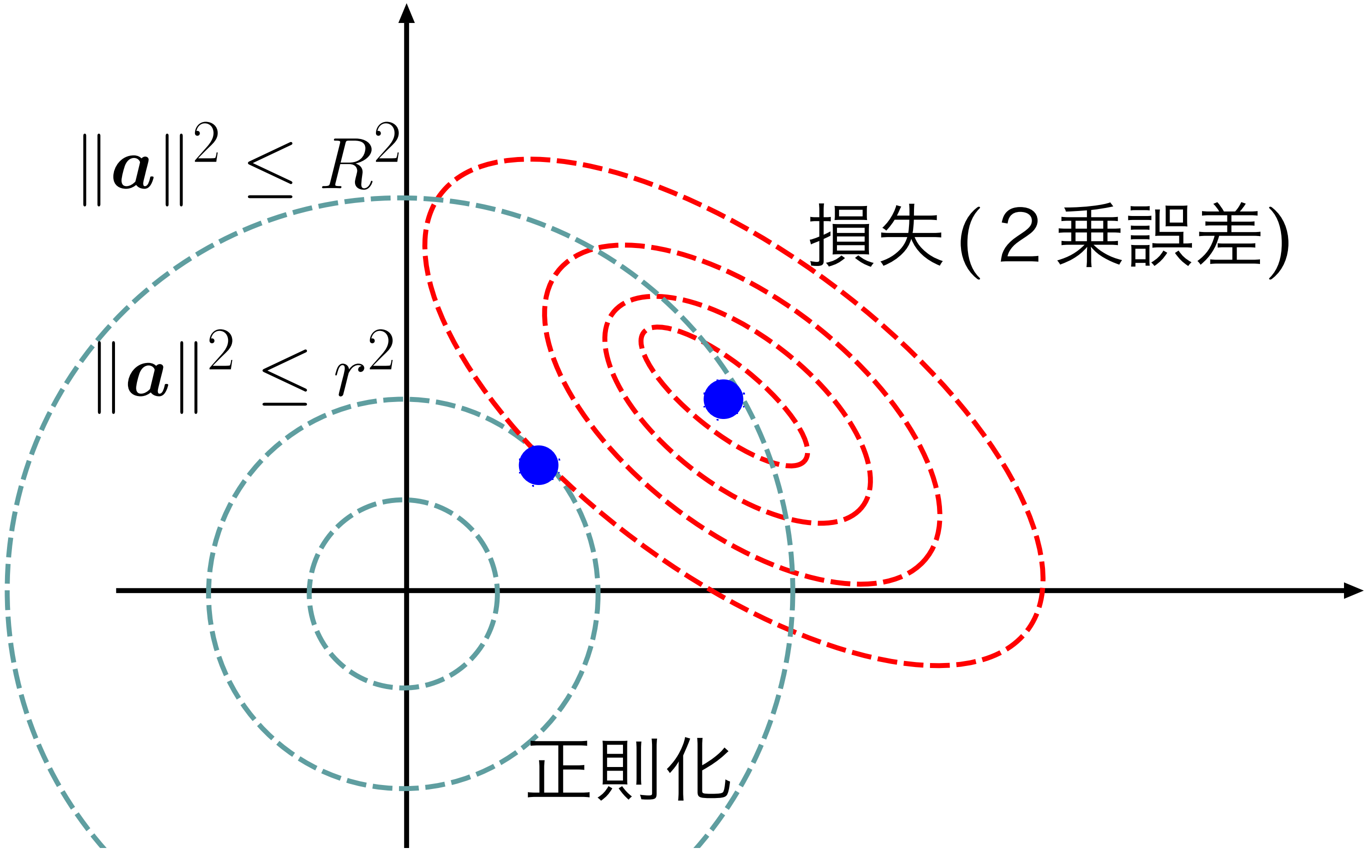
$$\min_{\boldsymbol{\beta},\, b} \sum_{i=1}^{n} \left\{ y_i - \left(\sum_{j=1}^{n} \beta_j k(\boldsymbol{x}_j, \boldsymbol{x}_i) + b \right) \right\}^2 + \lambda \sum_{i,j} \beta_i \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

正則化項

$$\Longrightarrow \ \text{最適解}\ \widehat{\beta_1}, \ldots, \widehat{\beta_n},\ \widehat{b}. \quad \widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} \widehat{\beta}_j k(\boldsymbol{x}_j, \boldsymbol{x}) + \widehat{b}$$

ベクトル・行列で表すと

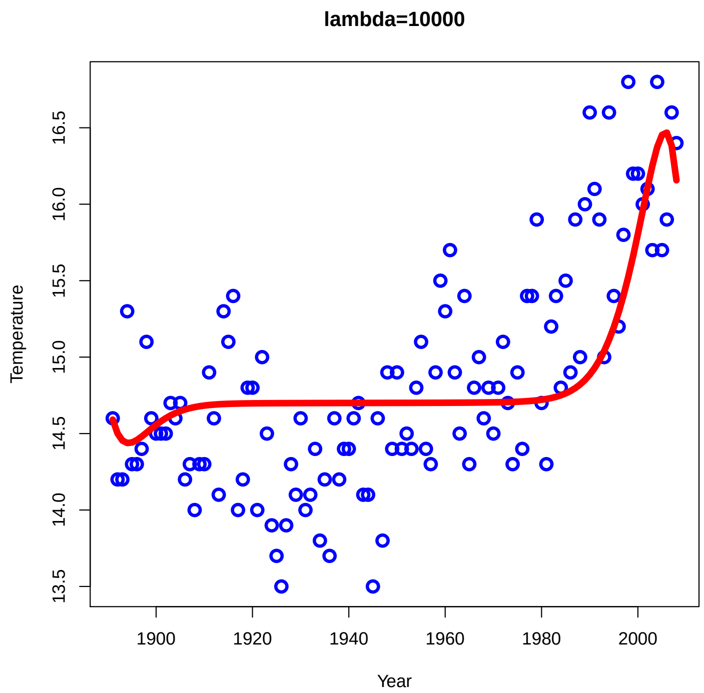
$$\min_{\boldsymbol{\beta}, b} \|\boldsymbol{y} - (K\boldsymbol{\beta} + b\mathbf{1})\|_2^2 + \lambda \boldsymbol{\beta}^T K \boldsymbol{\beta}$$



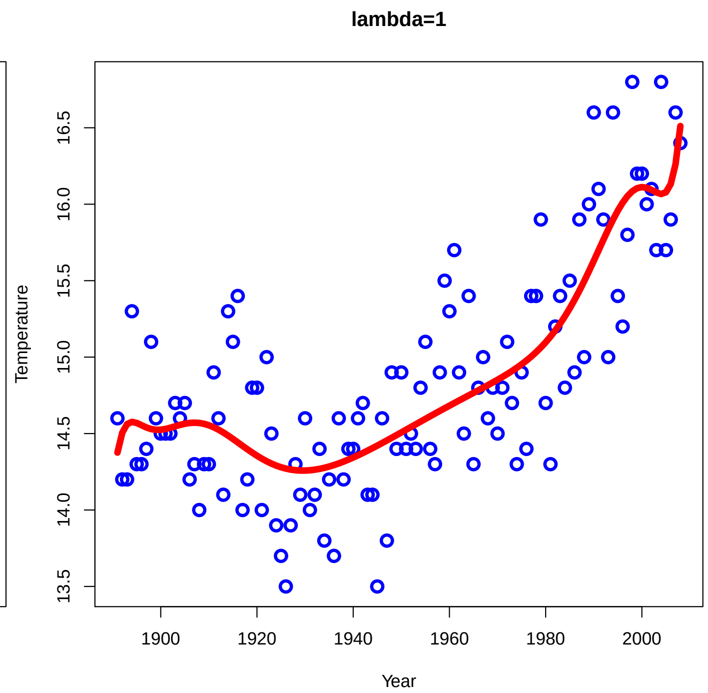
$\|\boldsymbol{a}\|^2 \leq R^2$
$\|\boldsymbol{a}\|^2 \leq r^2$
損失(２乗誤差)
正則化

正則化パラメータ $\lambda > 0$.

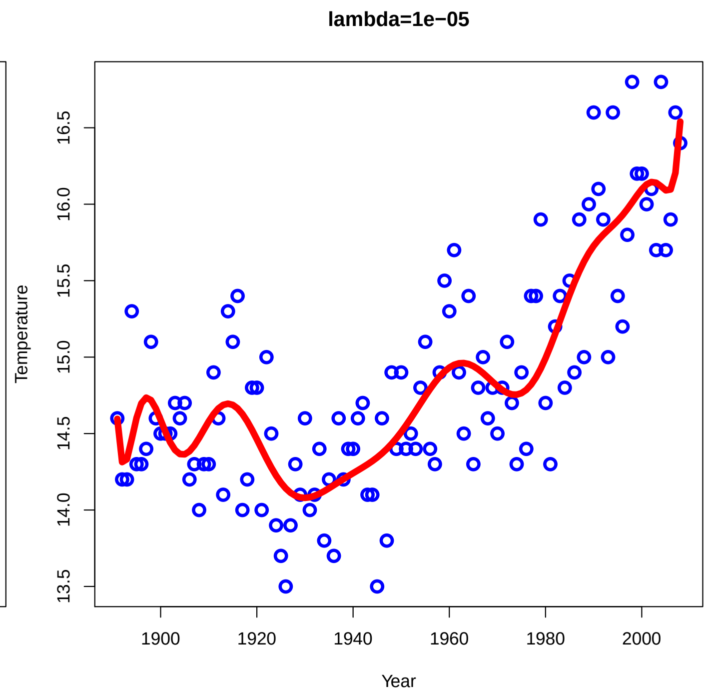
λ：大



λ：中



λ：小



小さい $\longleftarrow$ モデル自由度（表現力） $\longrightarrow$ 大きい

- 極値条件 **(**凸最適化なので大域解**)**

$$\min_{\boldsymbol{\beta},b} \|\boldsymbol{y} - K\boldsymbol{\beta} - b\mathbf{1}\|_2^2 + \lambda\boldsymbol{\beta}^T K\boldsymbol{\beta}$$

$$\Longrightarrow \quad \begin{pmatrix} \mathbf{1} & K + \lambda I \\ n & \mathbf{1}^T K \end{pmatrix} \begin{pmatrix} \widehat{b} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y} \\ \mathbf{1}^T\boldsymbol{y} \end{pmatrix}, \qquad K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^n \widehat{\beta}_i k(\boldsymbol{x}_i, \boldsymbol{x}) + \widehat{b}$

- カーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}')$ だけから計算できる.
  （$\phi(\boldsymbol{x})$ は単体では現れない）

R 言語による正則化付きカーネル回帰

- 初期設定

```
> library(kernlab)  # パッケージ読込
> sig <- 3; l <- 1; # パラメータ: sigma, λ
> n <- 100          # データ生成：データ数 n=100
> x <- matrix(runif(n,min=-4,max=4))
> y <- sin(pi*x)/x + rnorm(n,sd=0.5)
```

- ガウシアンカーネル(rbfdot)のグラム行列

```
> K <- kernelMatrix(rbfdot(sigma=sig),x)
```

- 推定量の計算：２乗誤差最小 [optim() 関数]

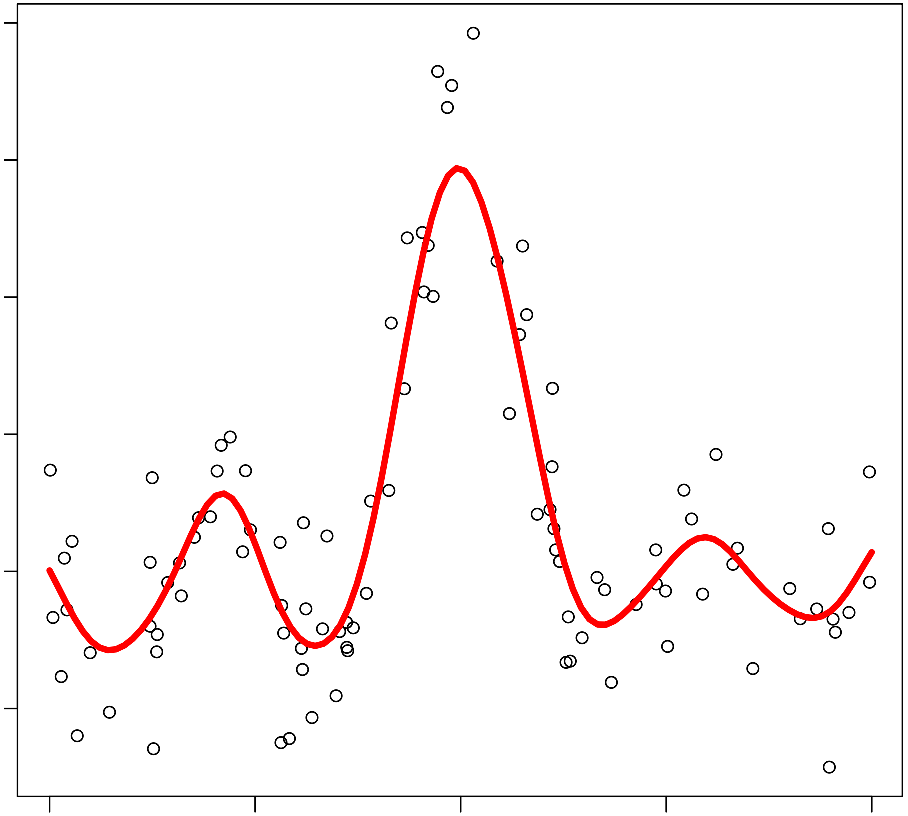
```
> f <- function(z){w <- z[-(n+1)];
+  (sum((cbind(K,1)%*%z-y)^2)+l*w%*%K%*%w)/2}
> g <- function(z){w <- z[-(n+1)];
+  rbind(K,1)%*%(cbind(K,1)%*%z-y)+c(l*K%*%w,0)}
> s <- optim(rep(0,n+1),f,g,method="BFGS")$par
> beta <- s[1:n]; b <- s[n+1]
```

- 予測

```
> px <- seq(-4,4,l=100) # 予測点
> py <- kernelMult(rbfdot(sigma=sig),px,x,beta)+b
> plot(x,y);lines(px,py,lwd=4,col=2) #プロット
```

データ点と推定結果のプロット

- ガウシアンカーネルのカーネルパラメータ：$\sigma = 3$
- 正則化パラメータ：$\lambda = 1$



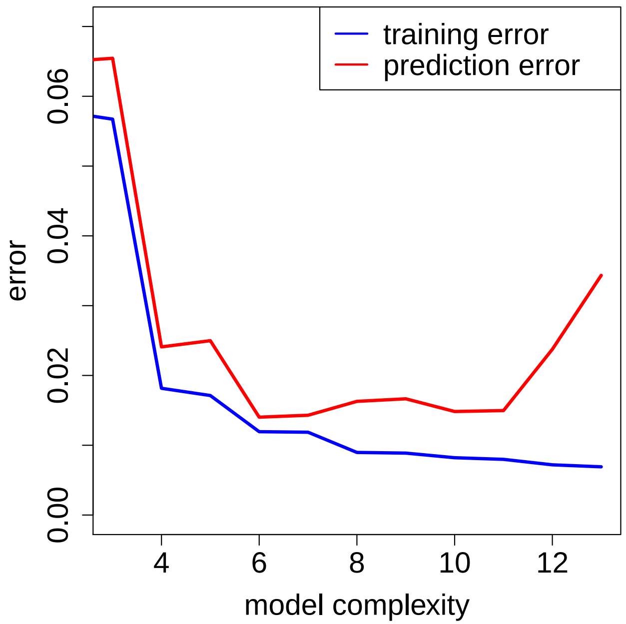
モデル選択：学習誤差と予測誤差

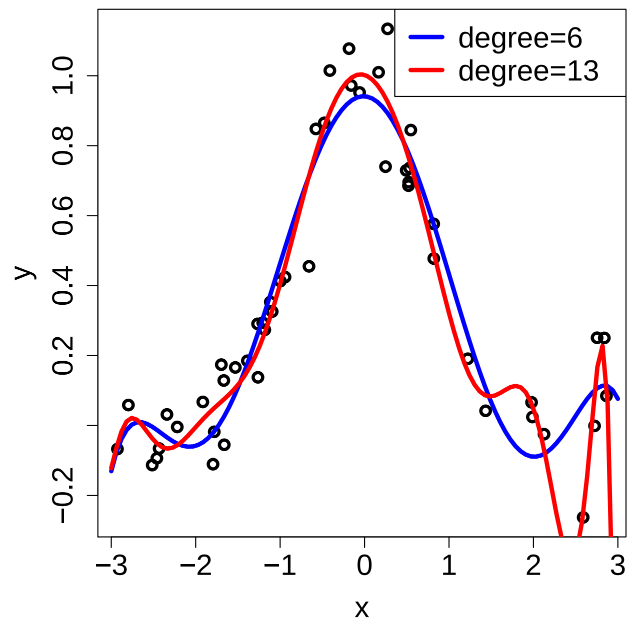
- 学習データ：$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \sim_{i.i.d.} P$
- 学習データから推定された回帰関数 $\widehat{f}(x)$

$$\text{学習誤差：} \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}(\boldsymbol{x}_i) - y_i\right)^2 \quad \text{(データから計算可能)}$$

$$\text{予測誤差：} \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\left(\widehat{f}(\boldsymbol{x}) - y\right)^2\right] \quad \text{(データだけでは分からない)}$$

多項式回帰





過学習

- 多項式の次数高いと・・・
  - ∗ 学習誤差は小さくなる
  - ∗ 予測誤差は大きくなる

$\longrightarrow$ モデル選択：モデルの自由度を適切に設定する.

カーネル法におけるモデル選択

- カーネル回帰 $\widehat{f}(\boldsymbol{x})$ の自由度を調整するパラメータ：

  * カーネルパラメータ：`kpar`
    - 多項式カーネルの $\gamma = 1, 2, 3, \ldots$
    - ガウシアンカーネルの $\sigma > 0$

  * 正則化パラメータ：$\lambda > 0,\ \mathrm{loss}(\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T K \boldsymbol{\beta}$

- これらを適切に定める.

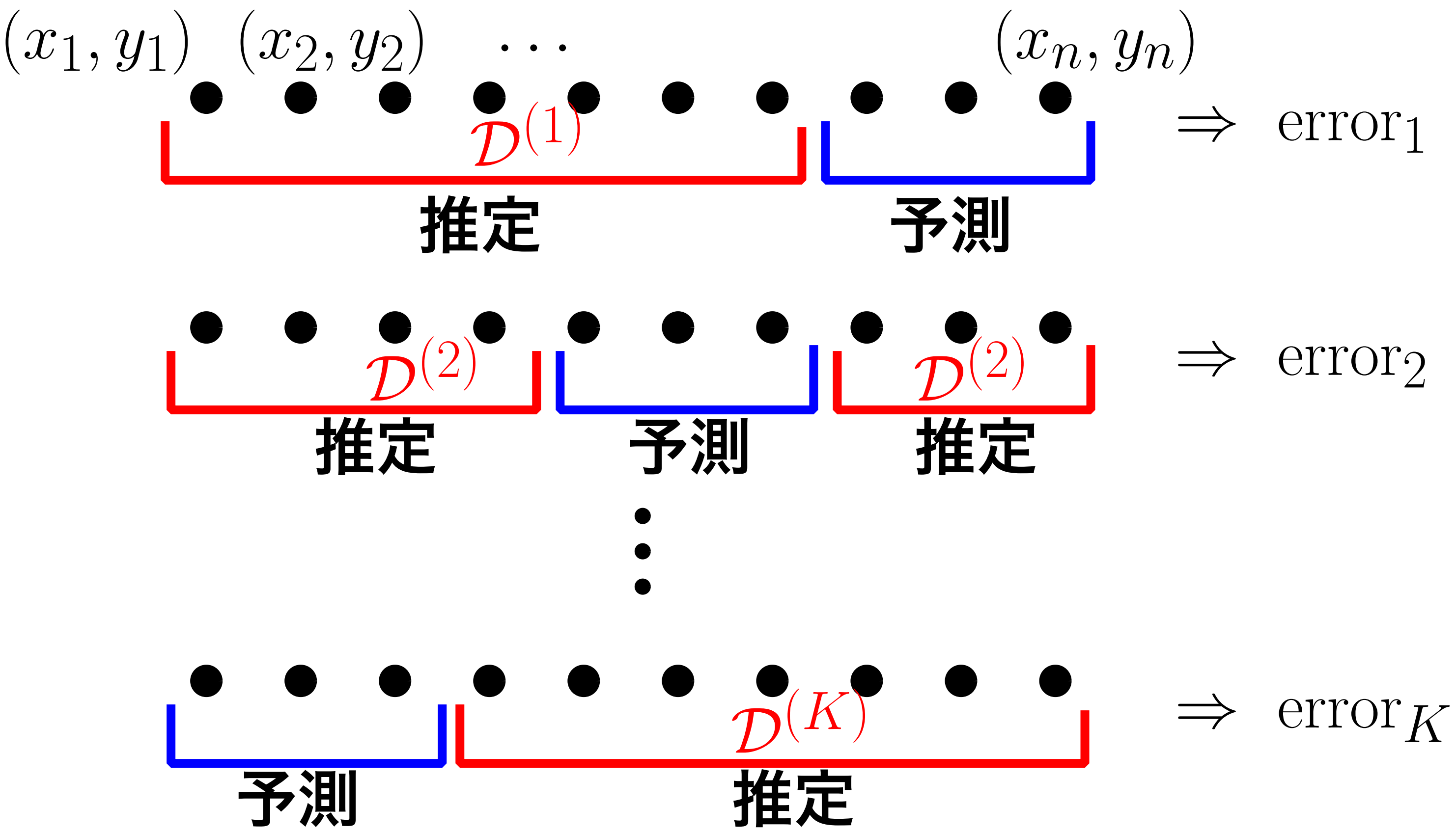
モデルパラメータ推定の方針

- いろいろな $\texttt{mpar} = (\texttt{kpar}, \lambda)$ について・・・
  - $*$ mpar に対する回帰関数：$\widehat{f}_{\mathrm{mpar}}(\boldsymbol{x})$
  - $*$ データから $\widehat{f}_{\mathrm{mpar}}(\boldsymbol{x})$ の予測誤差を推定 $\longrightarrow \widehat{e}_{\mathrm{mpar}}$
- $\widehat{e}_{\mathrm{mpar}}$ を最小にするモデルパラメータ $\texttt{mpar}$ を選択.

$\widehat{e}_{\mathrm{mpar}}$ の計算法：（K 重）交差検証法

K 重交差検証法 (K-fold cross validation)

- データを K 個のグループに分割



- 予測誤差の推定値 $\widehat{e}_{\mathrm{mpar}}$: $\mathrm{error}_1, \ldots, \mathrm{error}_K$ の平均.

モデルパラメータmparをもつ回帰関数に対するK-CV

1. データ $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ をほぼ同じサイズの K 個のグループ $\mathcal{D}_1, \ldots, \mathcal{D}_K$ に分割．$\mathcal{D}^{(k)}$ を以下のように定める．

$$\mathcal{D}_i \cap \mathcal{D}_j = \emptyset \quad (i \neq j), \quad \cup_{i=1}^{K} \mathcal{D}_i = \mathcal{D}, \quad \mathcal{D}^{(k)} := \mathcal{D} \setminus \mathcal{D}_k = \cup_{i \neq k} \mathcal{D}_i.$$

2. $\ell = 1, \ldots, K$ に対して以下を繰り返す．

(a) $\mathcal{D}^{(\ell)}$ を用いてモデルパラメータ mpar の回帰関数を学習：$\widehat{f}_\ell(\boldsymbol{x})$

(b) $\mathcal{D}_\ell$ に対する $\widehat{f}_\ell(\boldsymbol{x})$ の誤差を e_ℓ とする．

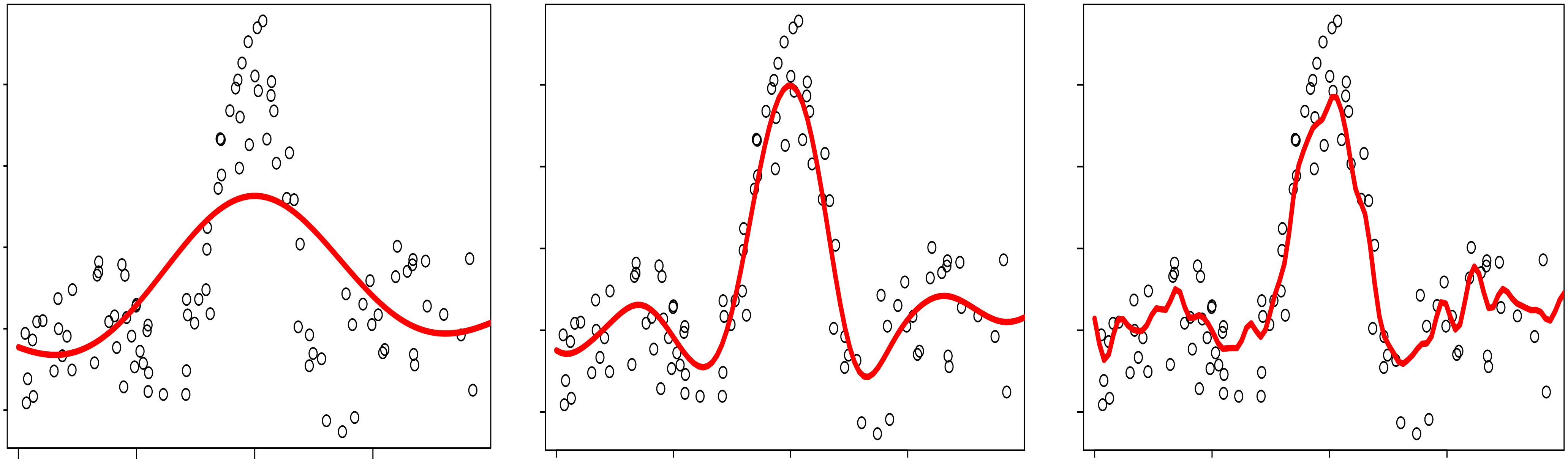
$$e_\ell := \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,y) \in \mathcal{D}_\ell} (y - \widehat{f}_\ell(\boldsymbol{x}))^2$$

（推定に使っていないデータに対する予測誤差）

3. 出力：予測誤差の推定値 $\widehat{e}_{\mathrm{mpar}} = \dfrac{1}{K} \displaystyle\sum_{\ell=1}^{K} e_\ell.$

推定結果：ガウシアンカーネル

- カーネルパラメータ σ：交差検証法で決定
- 正則化パラメータ：$\lambda = 1$ と固定

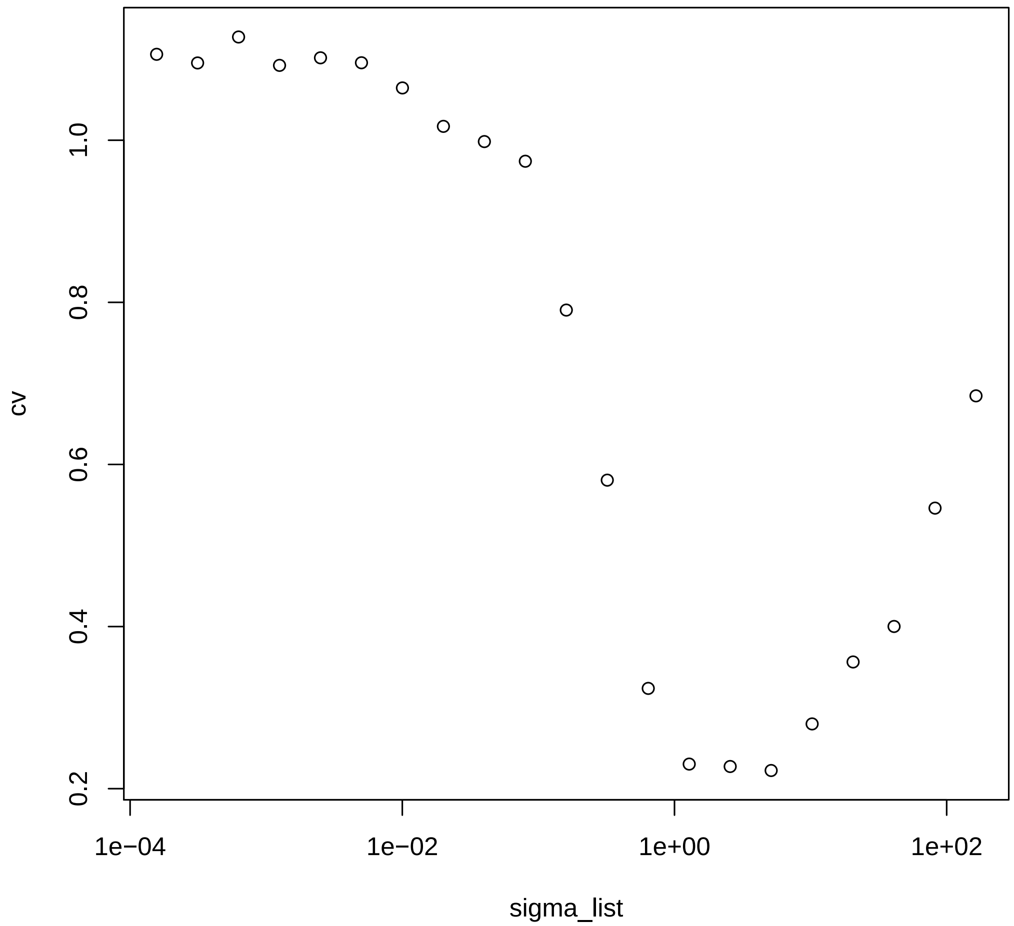


$(\sigma, \lambda) = (0.1, 1)$　　$(\sigma, \lambda) = (2.5, 1)$　　$(\sigma, \lambda) = (30, 1)$

$K = 5$の交差検証法.

$\sigma \cong 2.5$ で $\widehat{e}_{\mathrm{mpar}}$ が最小.

$\widehat{e}_{\texttt{mpar}}$ のプロット



ヒューリスティクス

- K-交差検証法：計算コストが大きい.
  - $*$ K 回学習 $\Longrightarrow$ ひとつの mpar に対する $\widehat{e}_{\mathrm{mpar}}$ が求まる.
  - $*$ モデルパラメータの選択： K 回 $\times$ mpar の候補数
- モデルパラメータが **2** 次元以上の場合
  - $*$ いくつかのパラメータ値を適当な値に固定.
    残りのパラメータを交差検証法で決める.

ガウシアンカーネルの σ の選び方

ガウシアンカーネル：

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\sigma\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\}$$

- 数値計算の安定性のため，$\sigma\,\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ が「ほどほどの値」を取るようにする.

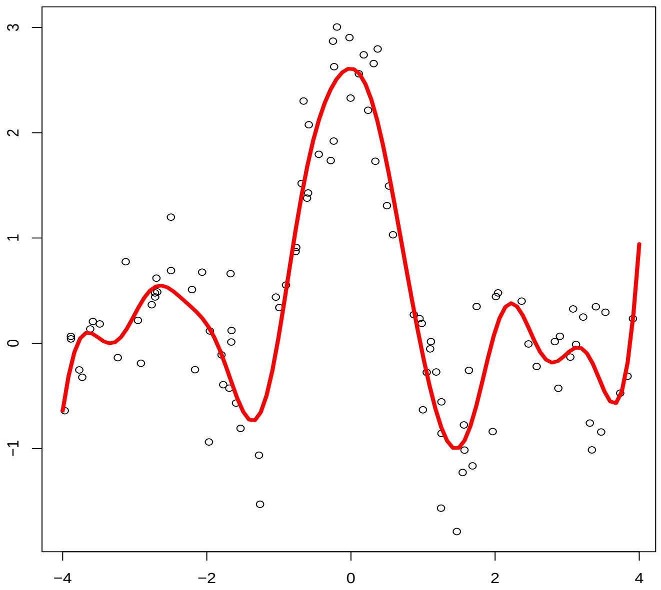
$$d^2 := \texttt{median}\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \,|\, i < j\} \quad (\text{中央値}),$$

$$\sigma = 1/d^2$$

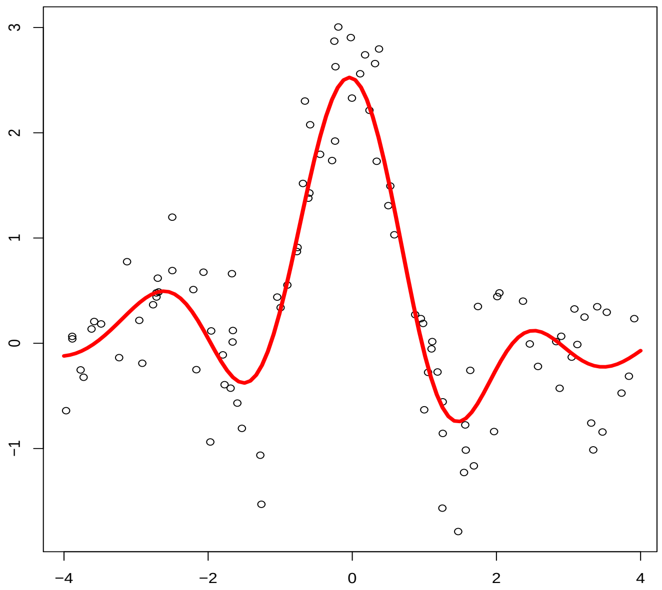
$\Longrightarrow$ $\sigma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ は 1 のまわりに分布.

推定結果

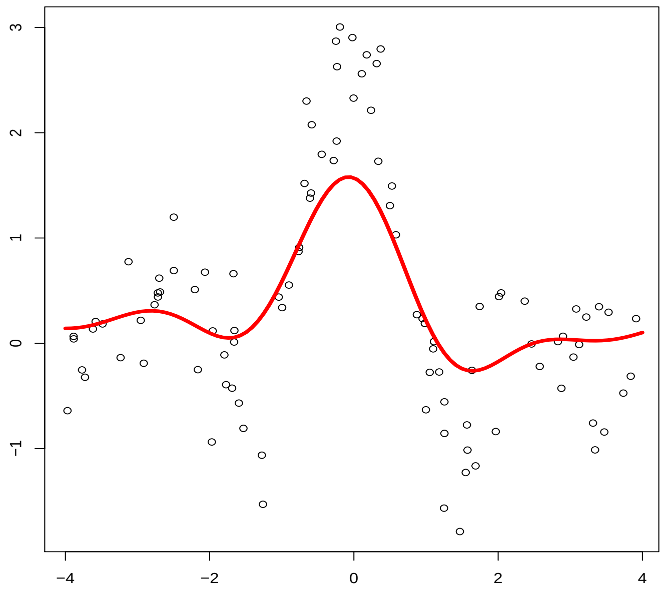
- カーネルパラメータ σ：ヒューリスティクス
- 正則化パラメータ λ：いろいろ変える.



$(\sigma, \lambda) = (0.98, 0.1^5)$

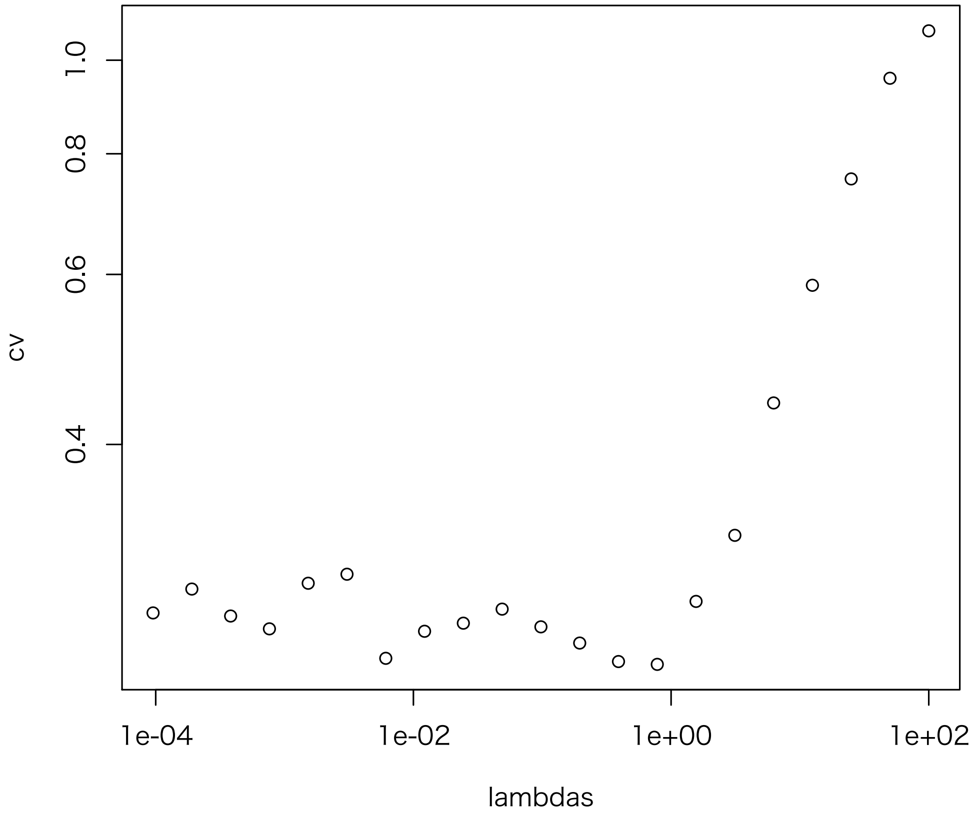


$(\sigma, \lambda) = (0.98, 0.78)$



$(\sigma, \lambda) = (0.98, 10)$

- $\widehat{e}_{\mathtt{mpar}}$ のプロット



- $\sigma = 0.984, \quad \lambda = 0.781$ が選択された.

カーネル関数と再生核ヒルベルト空間

- **(**非負定値**)** カーネル関数
- 再生核ヒルベルト空間
- 表現定理

カーネル法による学習：ある関数空間**(**再生核ヒルベルト空間**)** を統計モデルに設定することと等価.

参考文献：

- 福水，カーネル法入門，朝倉書店
- 赤穂，カーネル多変量解析，岩波書店

Definition 1 ((非負定値) カーネル関数).
集合 $\mathcal{X}$(例えば $\mathbb{R}^d$ の部分集合) に対して，以下を満たす関数 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ を <u>**(非負定値)** カーネル関数</u> という：

1. 対称性： $k(x, x') = k(x', x), \quad {}^\forall x, x' \in \mathcal{X}$

2. 非負定値性： $\forall n \in \mathbb{N}, \forall x_1, \ldots, x_n \in \mathcal{X}$

$$n \times n \text{ 対称行列} \quad K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

は非負定値行列

$$\left(\iff \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \quad {}^\forall c_1, \ldots, c_n \in \mathbb{R} \right)$$

復習：非負定値行列

非負定値**(**対称**)** 行列 $A \in \mathbb{R}^{n\times n}$. 以下の**1,2**は同値

1. $\forall \boldsymbol{c} \in \mathbb{R}^n,\ \boldsymbol{c}^T A \boldsymbol{c} \geq 0$.

2. Aの固有値は全て非負. **(**対称行列なので固有値は実数**)**.

Aの対角化：固有値 $\lambda_i \geq 0$, 固有ベクトル $\boldsymbol{u}_i \in \mathbb{R}^n$

$$A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \quad i = 1, \ldots, n, \quad \boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij},$$

スペクトル分解： $$A = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T.$$

カーネル関数の性質

$k, k_1, k_2, \ldots,$ がすべてカーネル関数のとき

1. $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^D$ に対して $k(x, x') = \boldsymbol{\phi}(x)^T \boldsymbol{\phi}(x')$ はカーネル関数

2. 非負定数 $a, b \geq 0$ に対して $a + b \cdot k(x, x')$ はカーネル関数

3. $k_1(x, x') + k_2(x, x')$ はカーネル関数

4. $k_1(x, x')k_2(x, x')$ はカーネル関数

5. $\bar{k}(x, x') = \lim_{m \to \infty} k_m(x, x')$ (各点収束) が存在するなら $\bar{k}$ はカーネル関数.

proof of 4.: n次非負定値行列 $K_1, K_2 \succeq O$ に対して

$$K_{ij} = (K_1)_{ij}(K_2)_{ij}$$

とする．$K \succeq O$を示す．スペクトル分解 $K_1 = \sum_{r=1}^n \lambda_r \boldsymbol{x}_r \boldsymbol{x}_r^T$ に対して

$$\begin{aligned}
\sum_{i,j} c_i c_j K_{ij} &= \sum_{i,j} c_i c_j (K_1)_{ij}(K_2)_{ij} \\
&= \sum_{i,j} c_i c_j \sum_{r=1}^n \lambda_r x_{ri} x_{rj} (K_2)_{ij} \\
&= \sum_{r=1}^n \lambda_r \sum_{i,j} (c_i x_{ri})(c_j x_{rj})(K_2)_{ij} \geq 0
\end{aligned}$$

Theorem 1. ガウスカーネルはカーネル関数である.

Gaussian kernel: $\boldsymbol{x} \in \mathbb{R}^d$ に対して

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\sigma \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\}, \qquad (\sigma > 0)$$

は実用上よく利用される. **Gaussian kernel** がカーネル関数の定義を満たすことを示す.

Proof. 対称性は明らか．非負定値性を示す．

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, $c_1, \ldots, c_n \in \mathbb{R}$ に対して $d_i = c_i \exp\{-\sigma \|\boldsymbol{x}_i\|_2^2\}$ とすると

$$\sum_{i,j=1}^{n} c_i c_j \exp\{-\sigma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\} = \sum_{i,j=1}^{n} d_i d_j \exp\{2\sigma \boldsymbol{x}_i^T \boldsymbol{x}_j\}.$$

ただし $d_i = c_i \exp\{-\sigma \|\boldsymbol{x}_i\|_2^2\}$．したがって，$k_1(\boldsymbol{x}, \boldsymbol{x}') = \exp\{2\sigma \boldsymbol{x}^T \boldsymbol{x}'\}$ がカーネル関数であること示せばよい．$e^z = 1 + z + z^2/2 + \cdots + z^n/n! + \cdots$ より

$$\begin{aligned}&\exp\{2\sigma \boldsymbol{x}^T \boldsymbol{x}'\} \\ &= 1 + 2\sigma \boldsymbol{x}^T \boldsymbol{x}' + (2\sigma)^2 (\boldsymbol{x}^T \boldsymbol{x}')^2/2 + \cdots + (2\sigma)^n (\boldsymbol{x}^T \boldsymbol{x}')^n/n! + \cdots\end{aligned}$$

上の各項 $(2\sigma)^n (\boldsymbol{x}^T \boldsymbol{x}')^n/n!$ はカーネル関数であり，その総和が $k_1(\boldsymbol{x}, \boldsymbol{x}')$ に各点収束するので，k_1 もカーネル関数である． ■

カーネル関数から生成される関数空間

- カーネル関数から生成される線形空間

$$\mathcal{H}_0 := \left\{ f(\cdot) = \sum_{j=1}^{m} \gamma_i k(x_i, \cdot\,) \;\middle|\; m \in \mathbb{N},\ x_i \in \mathcal{X},\ \ \gamma_i \in \mathbb{R} \right\}$$

$\mathcal{H}_0$ 上に $\langle k(x, \cdot),\ k(x', \cdot)\rangle = k(x, x')$ を満たす双線型形式を定義
$\rightarrow$ 実は内積.

補足

- $\langle f, g \rangle$, $f, g \in \mathcal{H}_0$ が f, g の表し方に依らないこと：

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x),\ \ g(x) = \sum_{j=1}^{m} \beta_j k(y_j, x) \text{ のとき}$$

$$\begin{aligned}\langle f, g \rangle &= \sum_{i,j} \alpha_i \beta_j \langle k(x_i, \cdot), k(y_j, \cdot) \rangle \\ &= \sum_{i,j} \alpha_i \beta_j k(x_i, y_j) = \sum_j \beta_j f(y_j) = \sum_i \alpha_i g(x_i)\end{aligned}$$

よって f, g の表し方に依らない.

Theorem 2. 以下の等式が成り立つ：

$$[\text{再生性}]:\quad \langle f,\, k(z,\cdot)\rangle = f(z) \qquad f \in \mathcal{H}_0, \qquad z \in \mathcal{X}$$

Proof. $f(x) = \sum_{j=1}^{m} \gamma_j k(x_j, x)$ とすると

$$\langle f, k(z,\cdot)\rangle = \sum_{j=1}^{m} \gamma_j \langle k(x_j,\cdot),\, k(z,\cdot)\rangle = \sum_{j=1}^{m} \gamma_j k(x_j, z) = f(z)$$

■

Theorem 3. $\langle f, g\rangle$ は $\mathcal{H}_0$ 上の内積である.

Proof. 双線形性，対称性は **O.K.** 以下を証明する：
$\langle f, f\rangle \geq 0$,　$\langle f, f\rangle = 0 \Longleftrightarrow f = 0$.

グラム行列の非負定値性より $f(\cdot) = \sum_{i=1}^n \gamma_i k(x_i, \cdot)$ に対して

$$\langle f, f\rangle = \sum_{i,j=1}^n \gamma_i\gamma_j \langle k(x_i, \cdot), k(x_j, \cdot)\rangle = \sum_{i,j=1}^n \gamma_i\gamma_j k(x_i, x_j) \geq 0.$$

$f = 0$ なら $\langle f, f\rangle = 0$ は明らか．逆を示す．$\langle f, f\rangle \geq 0$ を使うとコーシー・シュワルツの不等式が成り立つ（注）.

$$|f(x)|^2 = |\langle f, k(x, \cdot)\rangle|^2 \leq \langle f, f\rangle\langle k(x, \cdot), k(x, \cdot)\rangle = \langle f, f\rangle k(x, x)$$

よって $\langle f, f\rangle = 0$ なら任意の $x \in \mathcal{X}$ で $|f(x)| = 0$ よって $f = 0$ となる. ■

（注）　$\langle f, g\rangle$ が対称・双線形で $\langle f, f\rangle \geq 0$ を満たすなら
コーシー・シュワルツの不等式

$$|\langle f, g\rangle|^2 \leq \langle f, f\rangle\langle g, g\rangle \tag{1}$$

が成り立つ．条件［$\langle f, f\rangle = 0 \Leftrightarrow f = 0$］は(1) を示すのに必要ない．

ヒルベルト空間$\mathcal{H}_k$ の構成

1. $\mathcal{H}_0$ を完備化 $\to$ ヒルベルト空間**(**完備な内積空間**)** $\widetilde{\mathcal{H}}$.
完備：コーシー列は収束する. $\mathcal{H}_0$のコーシー列の集合を,

$$\text{同値関係}\quad \{f_n\} \sim \{g_n\} \Leftrightarrow \lim_{n\to\infty} |f_n - g_n| = 0$$

で割る.

i.e. $\widetilde{\mathcal{H}} = \{\{f_n\} \subset \mathcal{H}_0 : \{f_n\}\text{はコーシー列}\}/\sim$. 内積も自然に定義される.

2. $\widetilde{\mathcal{H}}$の要素 $[\{f_n\}] \Leftrightarrow$ 関数 $f(x) = \lim_{n\to\infty} f_n(x)$と対応
$\Longrightarrow$ ヒルベルト空間$(\mathcal{H}_k, \langle\cdot,\cdot\rangle)$, $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$, $\|f\| = \sqrt{\langle f, f\rangle}$.

- $\mathcal{H}_k$ で収束性の議論ができる.

- $\mathcal{H}_k$ の再生性：$\forall x \in \mathcal{X},\ \forall f \in \mathcal{H}_k,\ \langle f, k(x,\cdot)\rangle = f(x)$

- ∞-ノルムとの関連： $f \in \mathcal{H}$ に対して

$$\|f\|_\infty \le \|f\| \sup_{x\in\mathcal{X}} \sqrt{k(x,x)}$$

Proof.

$$\begin{aligned}
\|f\|_\infty &= \sup_{x\in\mathcal{X}} |f(x)| = \sup_{x\in\mathcal{X}} |\langle f, k(x,\cdot)\rangle| \le \sup_{x\in\mathcal{X}} \|f\|\|k(x,\cdot)\| \\
&= \|f\| \sup_{x\in\mathcal{X}} \sqrt{k(x,x)}
\end{aligned}$$

■

カーネル関数とヒルベルト空間との対応：

定義：$(\mathcal{H}, \langle \cdot, \cdot \rangle)$, $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ をヒルベルト空間とする．次の **(i)(ii)** を満たす $\bar{k}(x, x')$ が存在するとき，$\mathcal{H}$ を再生核ヒルベルト空間 **(RKHS)** といい，$\bar{k}$ を再生核という．

(i) $\forall x \in \mathcal{X}, \bar{k}(x, \cdot) \in \mathcal{H}$

(ii) $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, \bar{k}(x, \cdot) \rangle = f(x)$

Theorem 4. カーネル関数 $\Longleftrightarrow$ RKHS (対一対応). $k \leftrightarrow \mathcal{H}_k$.

Proof. 1. $\Rightarrow$) カーネル関数 $k \to$ Hilbert 空間 $\mathcal{H}_k$. このとき k は $\mathcal{H}_k$ の再生核. よって $\mathcal{H}_k$ は RKHS.

$\Leftarrow$) RKHS $\mathcal{H}_1 = \mathcal{H}_2$ (内積も $\langle \cdot, \cdot \rangle_1 = \langle \cdot, \cdot \rangle_2$), 再生核を k_1, k_2 とする. このとき $k_1 = k_2$. なぜなら

$$\begin{aligned} k_1(x,y) &= \langle k_1(x,), k_2(y,) \rangle = \langle k_2(y,), k_1(x,) \rangle \\ &= k_2(y,x) = k_2(x,y) \end{aligned}$$

再生核 k はカーネル関数.

対称性：$k(x,x') = \langle k(x,\cdot), k(x',\cdot) \rangle = \langle k(x',\cdot), k(x,\cdot) \rangle = k(x',x)$.
グラム行列の非負定値性：$f(\cdot) = \sum_i c_i k(x_i,\cdot)$ と $\langle f, f \rangle \geq 0$ から分かる. ■

フーリエ変換によるRKHSの表現

$$\text{フーリエ変換}: f(x) \longmapsto \widehat{f}(t) = \mathcal{F}[f](t) = \frac{1}{2\pi}\int_{\mathbb{R}} f(x) e^{-\sqrt{-1}xt} dx,$$

$$\text{フーリエ逆変換}: \widehat{f}(t) \longmapsto f(x) = \mathcal{F}^{-1}[\widehat{f}](x) = \int_{\mathbb{R}} \widehat{f}(t) e^{\sqrt{-1}tx} dt$$

$f \in L_2(\mathbb{R})$ なら $\mathcal{F}^{-1}[\widehat{f}] = f$ **(**適切に定義すれば**)**

- $\rho \in L_1(\mathbb{R}) \cap C(\mathbb{R})$, $\rho > 0$. 次の$\mathcal{H}$は**RKHS.**

$$\mathcal{H} = \left\{ f \in L_2(\mathbb{R}) \,\middle|\, \int |\widehat{f}(t)|^2/\rho(t)dt < \infty \right\},$$

$$\text{内積：}\ \langle f, g \rangle = \int \frac{\widehat{f}(t)\overline{\widehat{g}(t)}}{\rho(t)}dt.$$

- 対応するカーネル：

$$k(x,y) = k_x(y) = \int_{\mathbb{R}} e^{-\sqrt{-1}t(x-y)}\rho(t)dt = \mathcal{F}^{-1}[e^{-\sqrt{-1}tx}\rho(t)](y),$$

$$\widehat{k}_x(t) = e^{-\sqrt{-1}tx}\rho(t).$$

再生性：$\langle f, k_x \rangle = \int \frac{\widehat{f}(t)\overline{\widehat{k_x}(t)}}{\rho(t)} dt = \int \widehat{f}(t) e^{\sqrt{-1}tx} dt = f(x)$.

- ガウスカーネル：$\rho(t) = \frac{\sigma}{2\pi} e^{-\sigma^2 t^2/2} \Rightarrow k(x,y) = e^{-(x-y)^2/(2\sigma^2)}$,

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}, dx) \,\middle|\, \int |\widehat{f}(t)|^2 e^{\sigma^2 t^2/2} dt < \infty \right\}$$

- ラプラスカーネル：

$$\rho(t) = \frac{1}{2\pi} \frac{1}{t^2 + \beta^2} \Rightarrow k(x,y) = \frac{1}{2\beta} e^{-\beta|x-y|},$$

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}, dx) \,\middle|\, \int |\widehat{f}(t)|^2 (t^2 + \beta^2) dt < \infty \right\}$$

上記の $\mathcal{H}$：コンパクト集合 $\mathcal{X}$ への制限 $\{f|_{\mathcal{X}} \mid f \in \mathcal{H}\}$ は $C(\mathcal{X})$ で稠密 **(universal kernel)**.

カーネル関数を用いた推定

1. カーネル関数 $k(x, x') \mapsto$ **RKHS** $\mathcal{H}_k$.

2. データから関数 $f(x) \in \mathcal{H}_k$ を推定

- カーネル関数 k を用いて学習：統計モデルは **RKHS** $\mathcal{H}_k$
  - $\mathcal{H}_k$：統計的性質 (収束性など) を調べるのに役立つ.
  - アルゴリズムは (大抵) $\mathcal{H}_0$ で記述可能

Theorem 5 (表現定理)**.**
$\lambda > 0$, $x_1, \ldots, x_n \in \mathcal{X}$ とする．最適化問題

$$\min_{f,b} L(f(x_1), \ldots, f(x_n), b) + \lambda \|f\|^2, \quad f \in \mathcal{H}_k, \quad b \in \mathbb{R}$$

の最適解は$\mathrm{span}\{k(x_1, \cdot), \ldots, k(x_n, \cdot)\}$ に含まれる.

Remark . 表現定理より，$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$ として α_i を求めればよい.

$\mathcal{H}_k$ の次元が無限でも，有限次元最適化問題を解けばよい.

- カーネル回帰分析で同様の議論をした.

Proof. $\mathcal{H}$ のなかで関数 $k(x_1, \cdot), \ldots, k(x_n, \cdot)$ で張られる部分空間を $\mathcal{M}$ とする： $\mathcal{M} = \left\{ \sum_{i=1}^{n} \gamma_i k(x_i, \cdot) \;\middle|\; \gamma_1, \ldots, \gamma_n \in \mathbb{R} \right\}$.

関数 $f \in \mathcal{H}_k$ を $\mathcal{M}$ の成分と直交成分 $\mathcal{M}^\perp$ に分解する：

$$f = g + h, \qquad g \in \mathcal{M}, \quad h \in \mathcal{M}^\perp$$ **(Hilbert 空間の射影定理)**

このとき次式がなりたつ：

$$\begin{aligned}
\|f\|_{\mathcal{H}}^2 &= \langle g+h, g+h \rangle = \langle g, g \rangle + \langle h, h \rangle + \langle g, h \rangle + \langle h, g \rangle \\
&= \langle g, g \rangle + \langle h, h \rangle = \|g\|_{\mathcal{H}}^2 + \|h\|_{\mathcal{H}}^2 \geq \|g\|_{\mathcal{H}}^2, \\
f(x_i) &= \langle f, k(x_i, \cdot) \rangle = \langle g+h, k(x_i, \cdot) \rangle \qquad \text{(再生性)} \\
&= \langle g, k(x_i, \cdot) \rangle + \langle h, k(x_i, \cdot) \rangle = \langle g, k(x_i, \cdot) \rangle = g(x_i).
\end{aligned}$$

したがって $f \notin \mathcal{M}$ なら

$$\lambda\|f\|_{\mathcal{H}}^2 > \lambda\|g\|_{\mathcal{H}}^2,$$
$$L(f(x_1), \ldots, f(x_n), b) = L(g(x_1), \ldots, g(x_n), b)$$

となるので

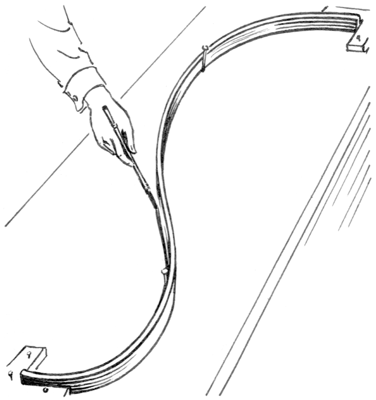
$$\begin{aligned} & L(f(x_1), \ldots, f(x_n), b) + \lambda\|f\|_{\mathcal{H}}^2 \\ > \ & L(g(x_1), \ldots, g(x_n), b) + \lambda\|g\|_{\mathcal{H}}^2. \end{aligned}$$

よって $\mathcal{M}$ のなかに最適解が存在する. ■

平滑化スプラインとカーネル法

参考：**Fasshauer and Ye, Reproducing kernels of Sobolev spaces via a green kernel approach with differential operators and boundary operators, arXiv:1109.5755, 2013.**

- スプライン：与えられた点を滑らかにつなぐカーブ.
- 回帰分析への応用：データ点を(ノイズを考慮しつつ)滑らかにつなぐ.



平滑化スプライン

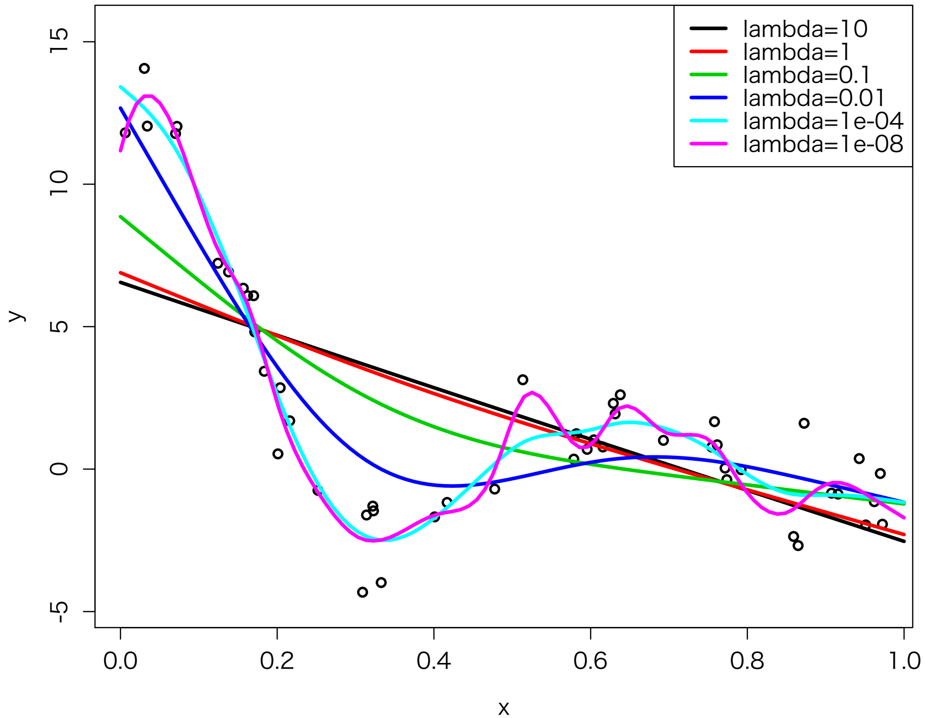
学習データ：$(x_1, y_1), \ldots, (x_n, y_n) \in (0,1) \times \mathbb{R}$.

$$\text{モデル：} \quad y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \text{ はノイズ.}$$

- データへのフィッティング：$\displaystyle\sum_{i=1}^{n} (y_i - f(x_i))^2 \longrightarrow$ 小さく
- 滑らかさ：$\displaystyle\int_0^1 (f''(x))^2 dx \longrightarrow$ 小さく

$$\min_f \quad \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx \quad \longrightarrow \quad \widehat{f}(x)$$

$\lambda \longrightarrow \infty \implies f(x) = ax + b$ **(単回帰)**



$$n = 50, \ y = \frac{\sin(4\pi x)}{x} + N(0,1), \ x \sim U(0,1)$$

正則化とカーネル関数

- 正則化項 $\int_0^1 (f''(x))^2 dx$：**RKHS** $\mathcal{H}_k$ の $\|f\|^2$ と解釈.

$$L(f) + \lambda \|f\|^2, \quad f \in \mathcal{H}_k \implies \text{表現定理を使って解く}$$

- 対応するカーネル関数を求める.

$$\mathcal{H} = \left\{ \sum_i \beta_i k(x_i, x) \;:\; \beta_i \in \mathbb{R},\ x_i \in (0,1) \right\} \ \textbf{(\& 完備化)}$$

$\subset$ 適当な**Sobolev** 空間

境界条件 : $g \in \mathcal{H} \implies$ **(極限で)** $g(0) = g'(0) = g(1) = g'(1) = 0$

推定量 : $f(x) = a + bx + g(x), \quad g \in \mathcal{H}$

- $\mathcal{H}$の内積. $\langle f, g\rangle := \langle f'', g''\rangle_{L_2} = \int_0^1 f''(x)g''(x)dx$.

  正則化項：$\|f\|^2 = \int_0^1 (f'')^2 dx$.

- 境界条件を含めると$\langle f, f\rangle = 0 \Rightarrow f = 0$ **(内積になっている)**

再生性が成り立つように k を定める. $g \in \mathcal{H}$ に対して

$$\begin{aligned}\langle k_z, g\rangle &= \int_0^1 k_z''(x)g''(x)dx = \Big[g'k_z'' - gk_z'''\Big]_0^1 + \int_0^1 g(x)k_z^{(4)}(x)dx \\ &= \int_0^1 g(x)k_z^{(4)}(x)dx\end{aligned}$$

$$\Longrightarrow \int_0^1 g(x)k_z^{(4)}(x)dx = g(z) \text{ となってほしい}$$

デルタ関数 δ (物理で登場)

$$\text{関数 } g(x) \text{ に対して} \quad \int_I g(x)\,\delta(x-z)dx = \begin{cases} g(z), & z \in I, \\ 0, & z \notin I. \end{cases}$$

次式を満たす $k(x,z) = k_z(x)$ をさがす.

$$\frac{d^4}{dx^4}k_z(x) = \delta(x-z),$$
$$k_z(0) = k_z'(0) = k_z(1) = k_z'(1) = 0.$$

$\to$ 部分積分を使ってどんどん積分すれば求まる

cf. 微分作用素 $\dfrac{d^4}{dx^4}$ のグリーン関数

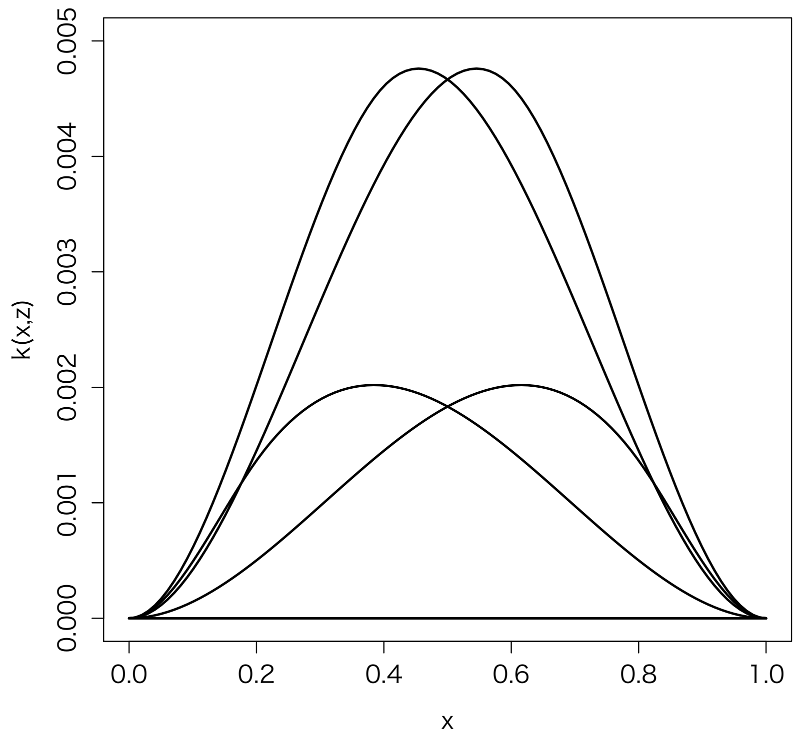
$$k(x,z)=\frac{([x-z]_+)^3}{6}+(x,z\text{の}3\text{次式})\quad(\longleftarrow\ \text{境界条件})$$

$$=\frac{([x-z]_+)^3}{6}-\frac{z^3}{6}+\frac{z^2}{2}x+x^2\left(\frac{z^3}{2}-z^2\right)+x^3\left(\frac{z^2}{2}-\frac{z^3}{3}\right),$$

$$([a]_+=\max\{a,0\}),\quad k(x,z)=k(z,x)\ \text{も}\ \textbf{O.K.}$$

プロット：

$k_z(x)$, $z=0,\,0.2,\,0.4,\,\ldots,\,1.$



モデリング：

- $f(x) = a + bx + g(x),\ g \in \mathcal{H}_k$ で推定. **note:** $(a+bx)'' = 0$.
- $a + bx$ で $f(0), f(1)$ を調整, $f'(0) = f'(1) = b$.

スプラインにおける表現定理：

$$\min_f \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx,$$

$$\text{s.t.} \quad f = a + bx + g(x),\ g \in \mathcal{H}_k$$

$$\Longrightarrow \ \text{最適解：} \quad \widehat{f}(x) = ax + b + \sum_{i=1}^n \beta_i k(x_i, x)$$

自然スプラインは少し違う条件 $(f''(0) = f''(1))$.

Proof. $\mathcal{M} = \mathrm{span}\{k(x_1,\cdot),\ldots,k(x_n,\cdot)\}$.

$g \in \mathcal{H}$ を $g = g_1 + g_2, \quad g_1 \in \mathcal{M},\ g_2 \in \mathcal{M}^\perp$ と分解.

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - f(x_i))^2 &= \sum_{i=1}^{n}(y_i - a - bx_i - \langle k_{x_i},\, g\rangle)^2 \\
&= \sum_{i=1}^{n}(y_i - a - bx_i - \langle k_{x_i},\, g_1\rangle)^2. \\
\int_0^1 (f''(x))^2 dx &= \int_0^1 (g''(x))^2 dx \\
&= \langle g, g\rangle = \langle g_1, g_1\rangle + \langle g_2, g_2\rangle \geq \langle g_1, g_1\rangle.
\end{aligned}$$

■

最適化問題：

$$\min_{a,b,\beta_i} \sum_{i=1}^{n} \Big\{ y_i - a - bx_i - \sum_{j=1}^{n} \beta_j k(x_j, x_i) \Big\}^2 + \lambda \sum_{i,j} \beta_i \beta_j k(x_i, x_j)$$

補足：多次元スプラインのカーネル **(Matérn kernel)**

例. $f : \mathbb{R}^2 \to \mathbb{R}$.

$$Af = (\sigma^2 f, \sqrt{2}\sigma f_x, \sqrt{2}\sigma f_y, f_{xx}, \sqrt{2} f_{xy}, f_{yy}), \quad \sigma > 0$$

正則化項(滑らかさの尺度)：$\displaystyle\int_{\mathbb{R}^2} \|Af\|_2^2 d\boldsymbol{x}$

$$\text{再生性：}\quad \int_{\mathbb{R}^2} Ak_{\boldsymbol{z}}(\boldsymbol{x})Ag(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^2} A^*Ak_{\boldsymbol{z}}(\boldsymbol{x})\, g(\boldsymbol{x})d\boldsymbol{x} = g(\boldsymbol{z}).$$

$$A^*Ak_{\boldsymbol{z}}(\boldsymbol{x}) = \delta(\boldsymbol{x}-\boldsymbol{z}), \quad \boldsymbol{x} \in \mathbb{R}^2$$

$$\Leftrightarrow (\sigma^2 - \Delta)^2 k_{\boldsymbol{z}}(\boldsymbol{x}) = \delta(\boldsymbol{x}-\boldsymbol{z}), \quad \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

$$\Rightarrow \textbf{Matérn kernel} \quad k(\boldsymbol{x},\boldsymbol{z}) \propto \frac{\|\boldsymbol{x}-\boldsymbol{z}\|_2}{\sigma} B_1(\sigma\|\boldsymbol{x}-\boldsymbol{z}\|_2)$$

$B_\nu(x)$：第 **II** 種 ν 次変形ベッセル関数.
(ν は次元や微分回数に依存して決まる**)**

note: 1次元スプライン：$A = \frac{d^2}{dx^2},\ A^* = \frac{d^2}{dx^2},\ A^*A = \frac{d^4}{dx^4}$