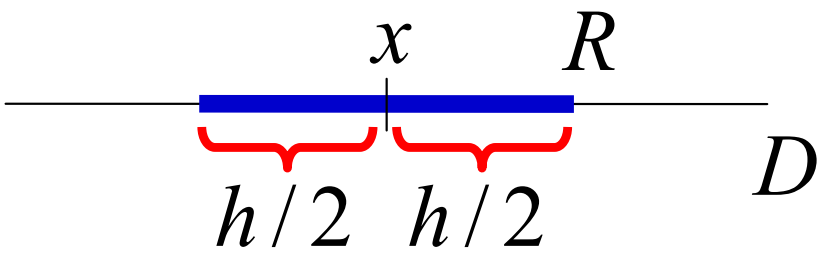
# 核密度推定法

# パーゼン窓関数

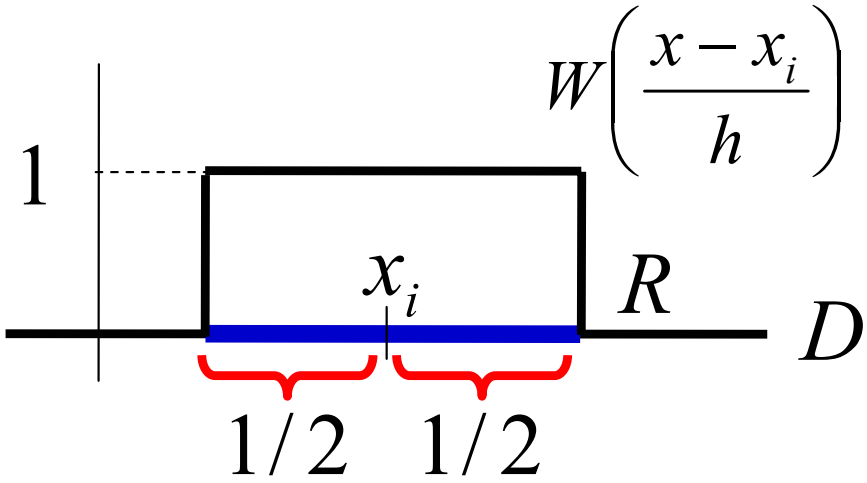
- 領域 $R$ として，ある点 $x$ を中心とする一辺の長さが $h$ の超立方体(hypercube)を用いる.
  - 体積 $V = h^d$
  - $R$ に入る標本数

$$k = \sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right)$$




$$W(x) = \begin{cases} 1 & \max|x^{(i)}| < \frac{1}{2} \\ 0 & \text{それ以外} \end{cases}$$

パーゼン窓関数
(Parzen window function)

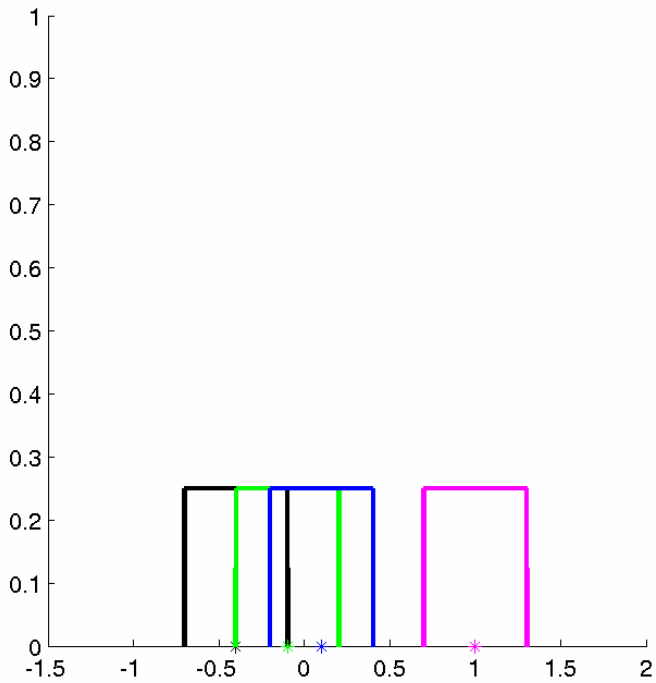


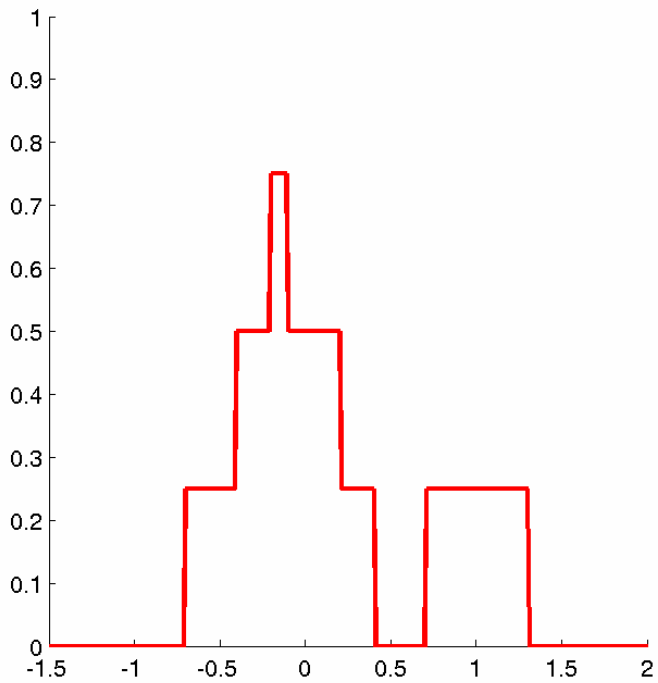
$$p(x) \approx \frac{k}{nV}$$

# パーゼン窓法

■ パーゼン窓法(Parzen window method):

$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} W\left( \frac{x - x_i}{h} \right)$$



# 核密度推定法

- パーゼン窓法で得られる $\hat{p}(x)$ は不連続.
- 核関数(kernel function) $K(x)$ : 滑らかな関数

$$\int_D K(x)dx = 1 \qquad K(x) \geq 0 \text{ for all } x \in D$$

- 核密度推定法(kernel density estimation method):

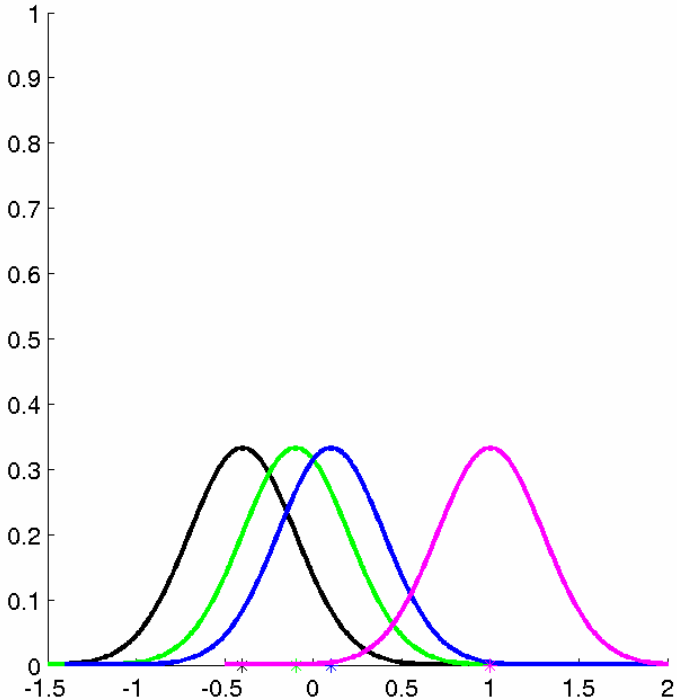
$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

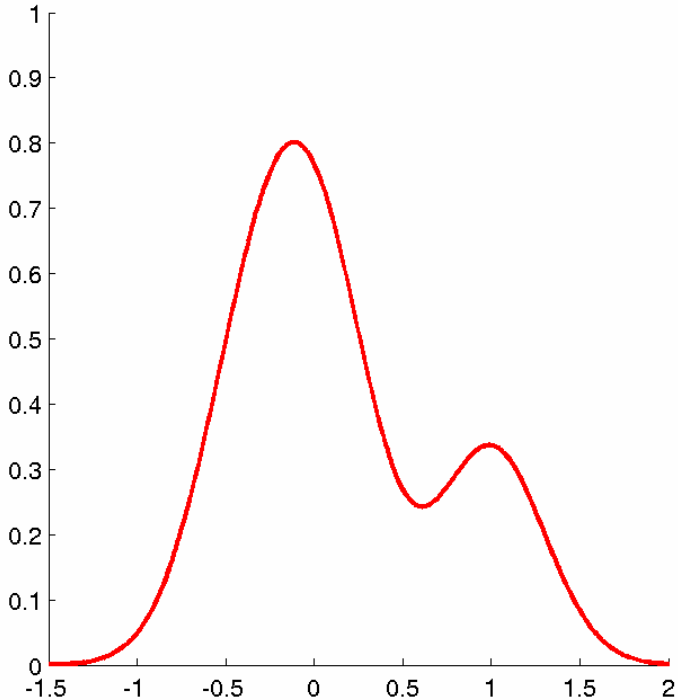
$h$ :幅(bandwidth)

# 核密度推定法の例

■ガウス核関数

$$K(x) = \frac{1}{(2\pi)^{d/2}} \exp(-\tfrac{1}{2} x^T x)$$



# 一般化

$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left( \frac{x - x_i}{h} \right)$$

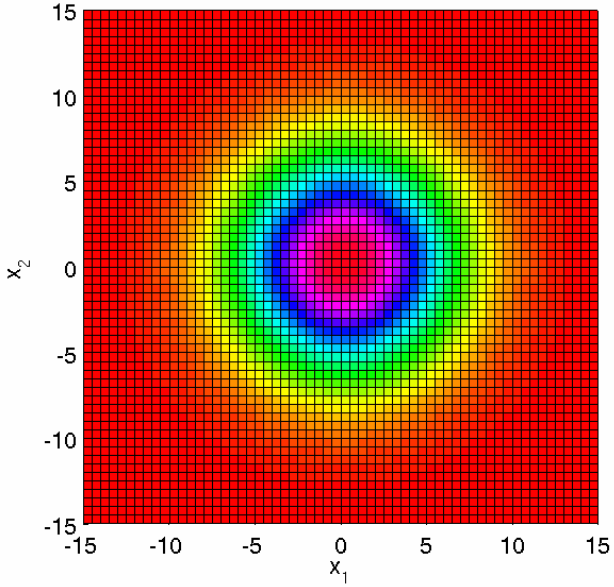
- 幅を一つのパラメータ $h$ で制御すると自由度が小さい.

- 幅行列(bandwidth matrix) $H$ :正値対称行列

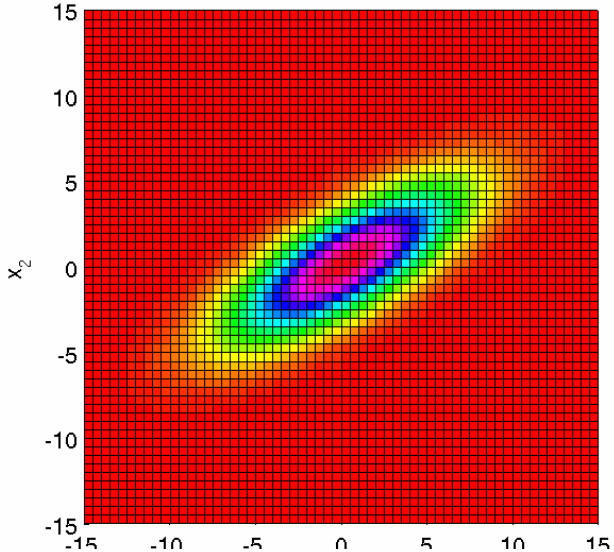
$$\hat{p}(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K\left( H^{-1}(x - x_i) \right)$$

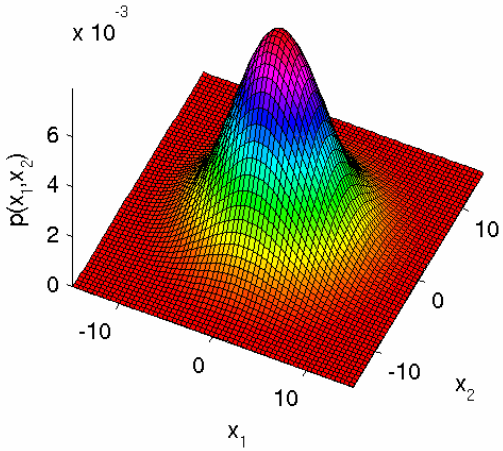
# 例

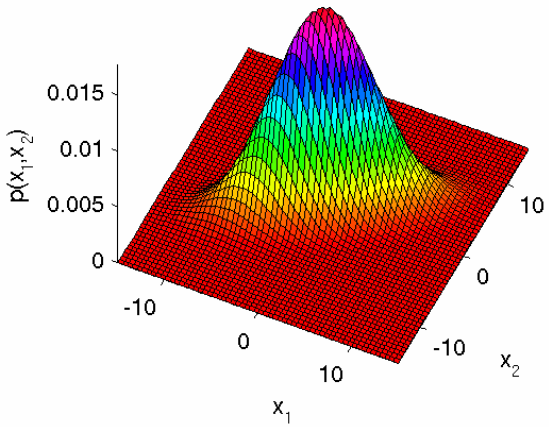
$h$


15
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
15
$x_1$
$x_2$

$H$

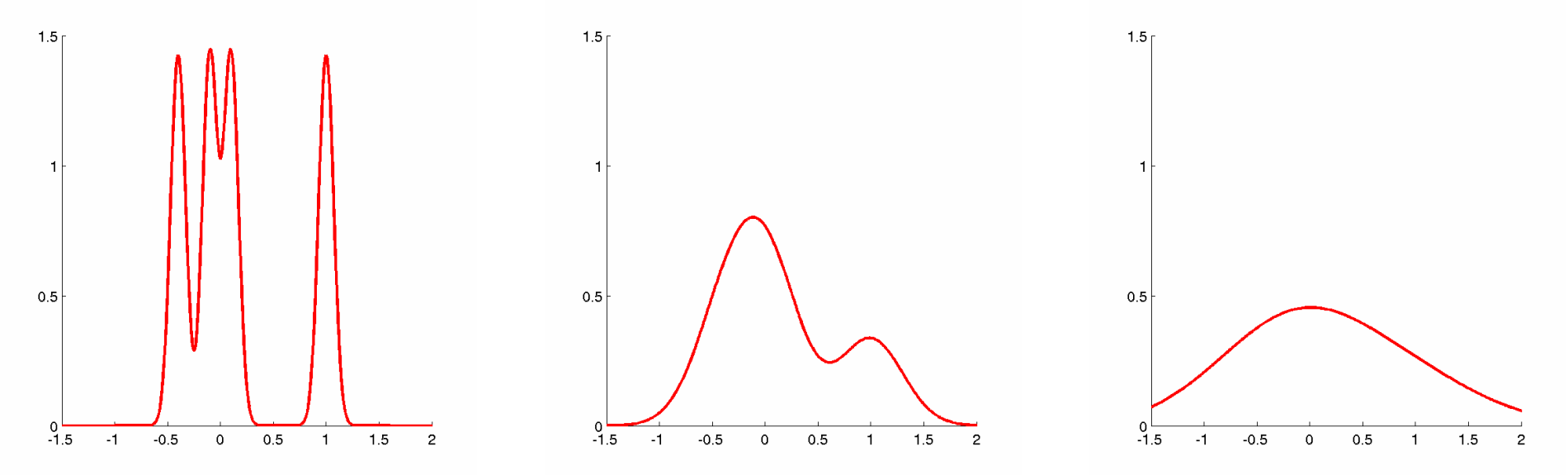

15
10
5
0
-5
-10
-15
$x_2$


x 10^-3
6
4
2
0
$p(x_1,x_2)$
-10
0
10
$x_1$
10
0
-10
$x_2$


0.015
0.01
0.005
0
$p(x_1,x_2)$
-10
0
10
$x_1$
10
0
-10
$x_2$

# 幅の決定法

- 核密度推定法によって得られる推定結果は，核関数の幅の決め方に依存する.




- 訓練標本を用いて幅を決定する.
  - 最小二乗交差確認法
  - シルバーマンの方法

# 最小二乗交差確認法(1)

- $\hat{p}(x)$ のよさを測る規準:積分二乗誤差 (integrated squared error)

$$ISE(h) = \int_D \left(\hat{p}(x) - p(x)\right)^2 dx$$
$$= \underbrace{\int_D \left(\hat{p}(x)\right)^2 dx}_{\text{計算可}} - 2\int_D \hat{p}(x)p(x)dx + \underbrace{\int_D \left(p(x)\right)^2 dx}_{\text{定数なので無視}}$$

- 第二項目を推定する.

# 最小二乗交差確認法(2)

- $\hat{p}_j(x)$ : $j$ 番目の訓練標本を推定に使わずに核密度推定法で推定した確率密度関数

$$\hat{p}_j(x) = \frac{1}{(n-1)|H|}\sum_{i \neq j} K\left(H^{-1}(x - x_i)\right)$$

- $\hat{p}_j(x_j)$ は $\int_D \hat{p}(x)p(x)dx$ の推定になっている.
- これを全ての $j$ について平均するとよりよい推定量が得られる.

$$\frac{1}{n}\sum_{j=1}^{n} \hat{p}_j(x_j)$$

- この方法を最小二乗交差確認法(least-squares cross-validation method)という.

# 最小二乗交差確認法(3)

- 実際の幅の最適化:
  - 幅行列の候補をいくつか用意
  - それぞれの候補に対してISEを推定し，最小のものを選ぶ
- 特徴
  - 真の確率密度関数がどんなものでもOK
  - どのような核関数でもOK
- 欠点
  - 計算に時間がかかる

# シルバーマンの方法

$$H = diag(h_1, h_2, \ldots, h_d)$$

- 真の確率密度関数が共分散のない正規分布の場合，ガウス核関数の最適な幅 $h_j$ は，

$$h_j \approx \left( \frac{4}{(d+2)n} \right)^{1/(d+4)} \sigma_j$$

$\sigma_j$ :正規分布の分散

- $\sigma_j$ を訓練標本から推定

$$\hat{\sigma}_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left( x_i^{(j)} - \bar{x}^{(j)} \right)^2} \qquad \bar{x}^{(j)} = \frac{1}{n} \sum_{i=1}^{n} x_i^{(j)}$$

- これをシルバーマンの方法(Silverman’s method)という.