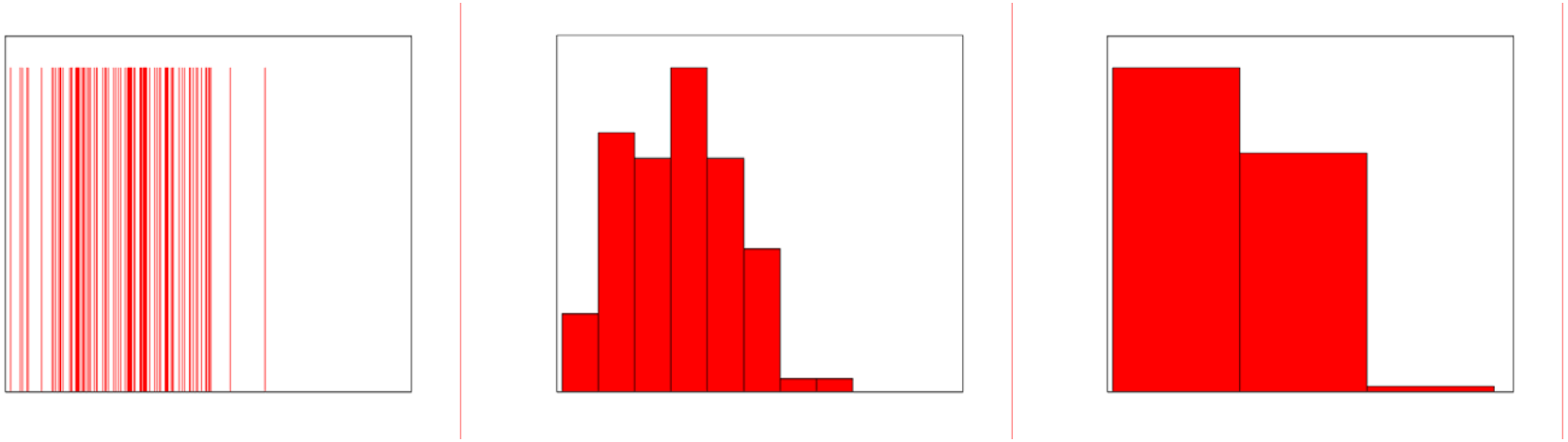
情報認識
「核密度推定法」

- 担当教員： 杉山　将（計算工学専攻）
- 居室： W8E-505
- 電子メール: sugi@cs.titech.ac.jp

ヒストグラム法の問題点



- 領域間で不連続
- 領域の分割の仕方を決めるのが難しい

もう少し工夫した方法が必要

ノンパラメトリック法の基礎

- 推定したい確率密度関数 $p(x)$ が R 内でほぼ一定値を取るとき，確率 P は領域 R 内のある点 x を用いて

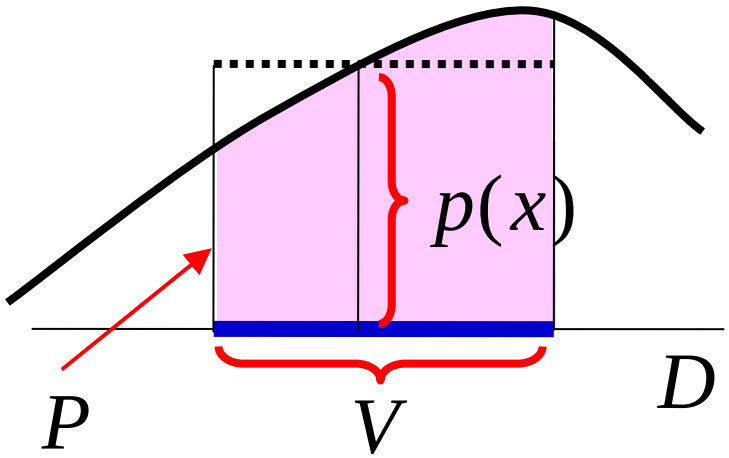
$$P \approx Vp(x)$$

- k ：n個の訓練標本のうち R に入っている個数

$$P \approx k/n$$

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

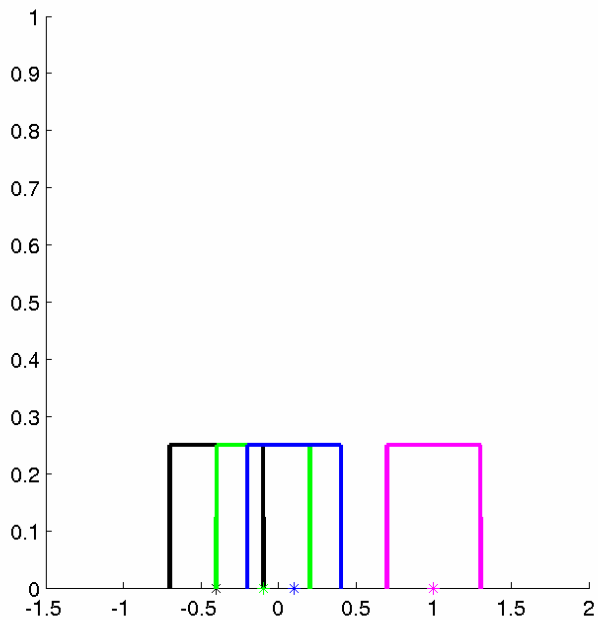
積分の長方形近似

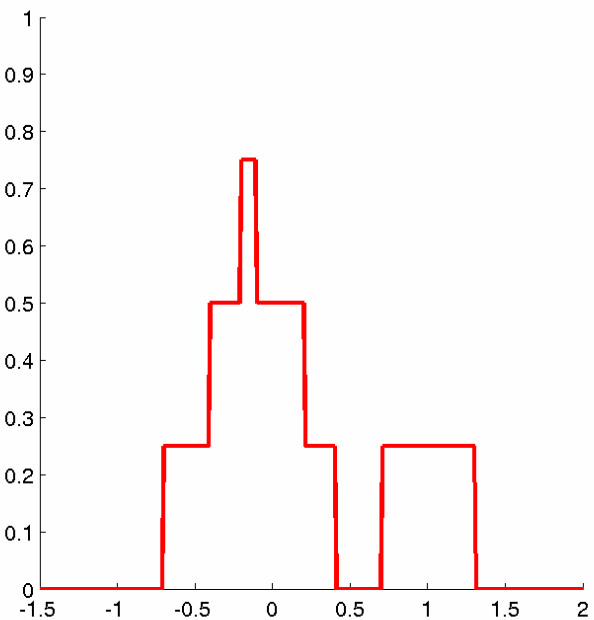


パーゼン窓法

- V を固定したもとで k を標本から決定：ある点 x を中心とする一辺の長さが h の超立方体を用いる

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





パーゼン窓法の特徴

- 与えられた標本からパターン領域の分割を適応的に決定できるため，ヒストグラム法のようにあらかじめ領域の分割の仕方を決定しておく必要が無い．
- 領域間での不連続性は未解決．
- 領域の分割の仕方を決める必要は無いが，パーゼン窓関数の幅 h を適切に決める必要がある．

核密度推定法

- パーゼン窓法の不連続性を解決したい
- 核関数(kernel function) $K(x)$：

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

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

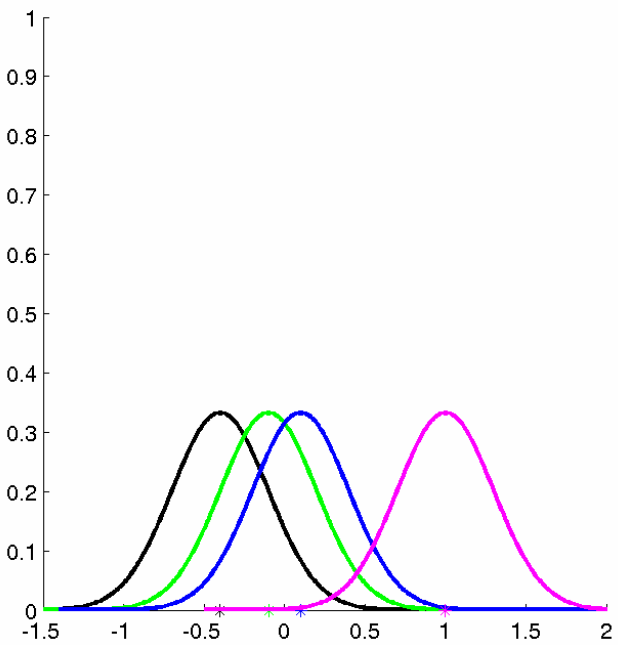
$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

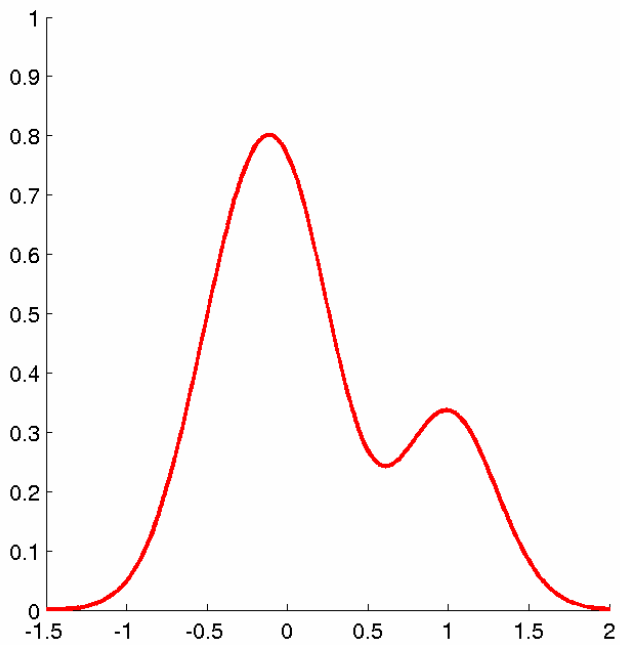
h :幅(bandwidth)

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

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

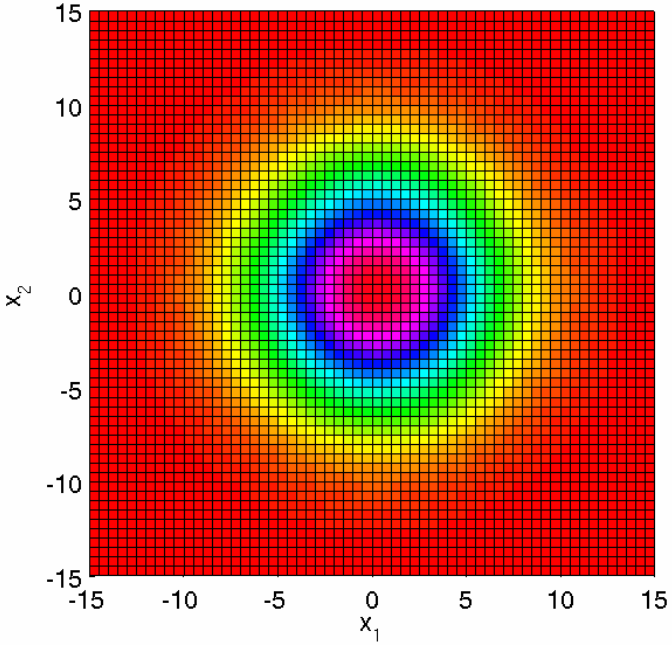
- 幅を一つのパラメータ h で制御すると自由度が小さい.

- 幅行列(bandwidth matrix) H：正値対称行列

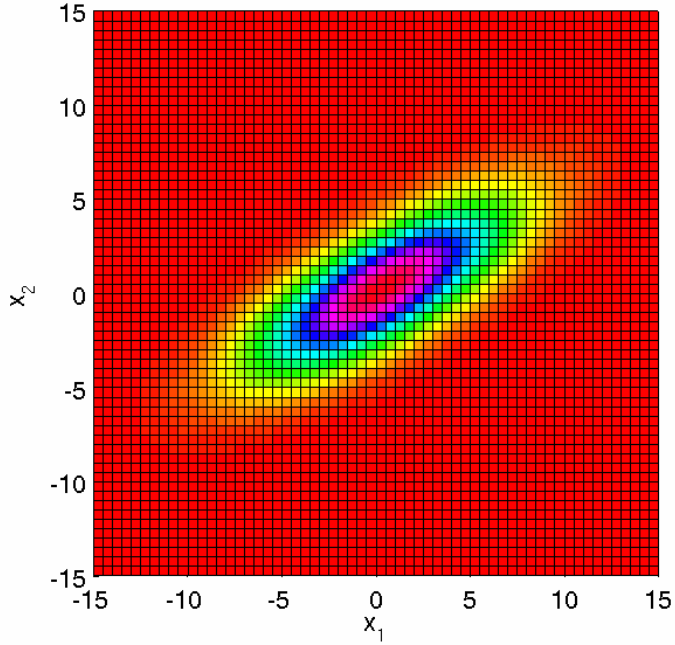
$$\hat{p}(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K\left(H^{-1}(x - x_i) \right)$$

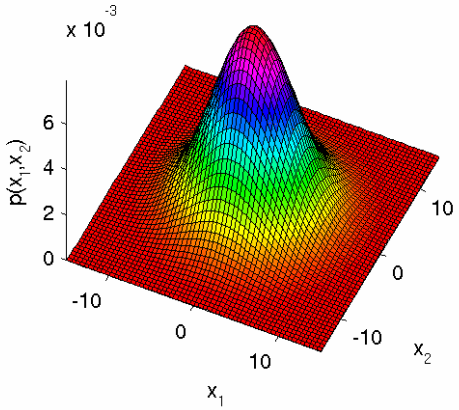
例

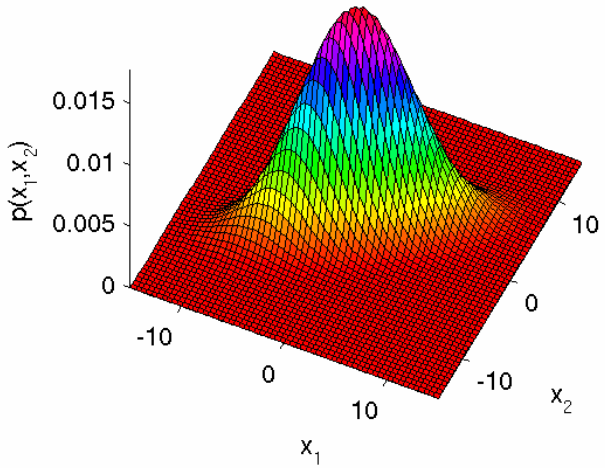
h


15
10
5
0
-5
-10
-15
-15 -10 -5 0 5 10 15
x1
x2

H

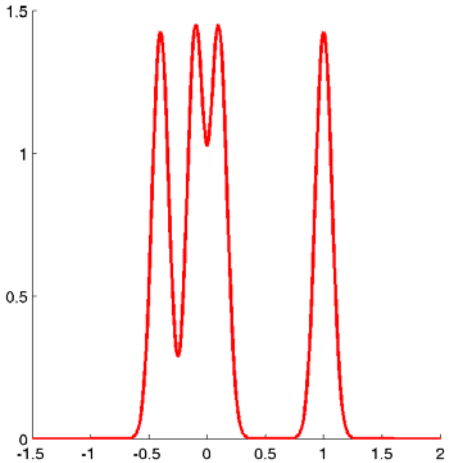

15
10
5
0
-5
-10
-15
-15 -10 -5 0 5 10 15
x1
x2

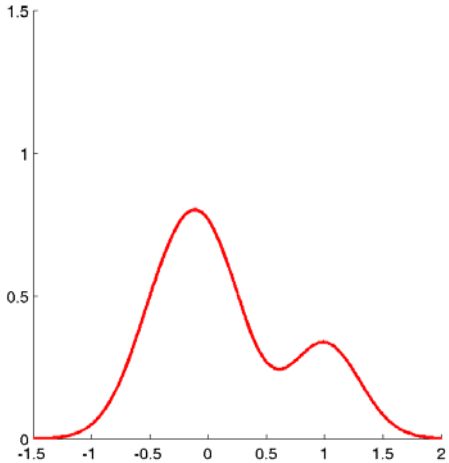

x 10-3
6
4
2
0
p(x1,x2)
-10 0 10
x1
10 0 -10
x2

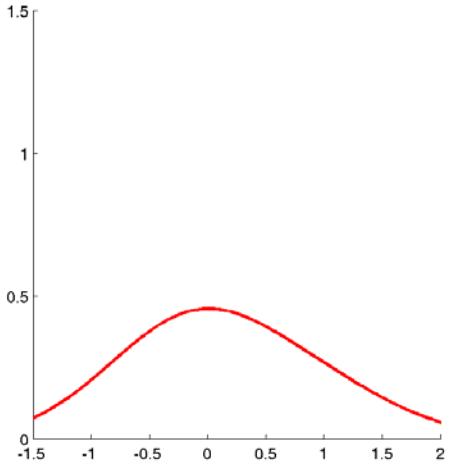

0.015
0.01
0.005
0
p(x1,x2)
-10 0 10
x1
10 0 -10
x2

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，

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

σ_j :正規分布の分散

- σ_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．

$$x = (x^{(1)}, x^{(2)}, \ldots x^{(d)})^T$$

小レポート：計算機実験

- 核密度推定法を実装し，適当なデータを用いて確率密度関数を推定せよ．
- 最小二乗交差確認法，及び，シルバーマンの方法を実装し，各関数の幅を決定せよ．
- データ標本数，真の確率分布，核関数，幅などの条件を変化させたとき，どのように推定結果が変わるかを考察せよ．