# データ解析
# 第十回「ノンパラメトリック密度推定法」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

# 休講情報

7/7 は休講

# 今日の講義内容

- カーネル密度推定

# 構成

1. カーネル密度推定
   - カーネル密度推定の推定手法
   - バンド幅の選択
   - 理論
   - 実データでの実験

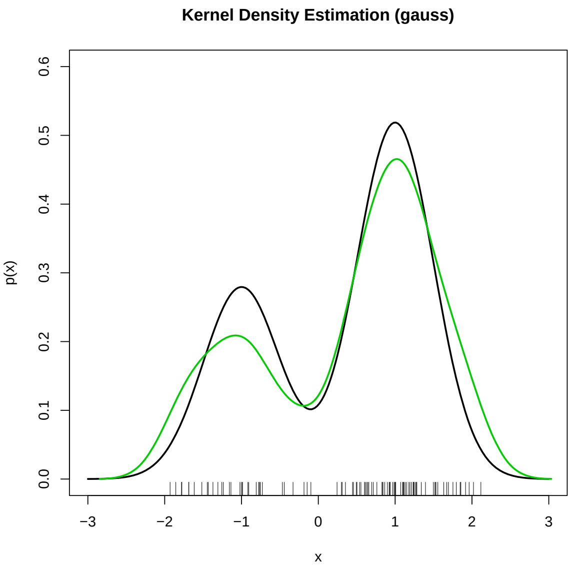
# カーネル密度推定の目的

**ある分布の密度関数を推定したい.**

いままでは分布のパラメトリックモデルを想定した手法を紹介してきた.
(正規分布，指数分布，ガンマ分布，...)

もし分布を適切なパラメトリックモデルで記述できなかったら？
↓
**ノンパラメトリック推定**
カーネル密度推定はその代表的な方法.

# カーネル密度推定で何が得られる？



# カーネル密度推定の長所・短所

- （長所）計算が簡単．単にデータ上でカーネル関数を足し合わせるだけ．
- （長所）漸近的な性質が理論的に導出可能．
- （長所）分布に対する仮定が少ない．滑らかさくらい．
- （短所）データを全て保持しておく必要があり，メモリ効率が悪い．
- （短所）新しいデータ点での密度の計算が $O(n)$ かかる．
- （短所）高次元データでは精度が悪い（次元の呪い）．

- （注意点）適切なパラメトリックモデルが得られているのなら，そちらを用いるべき．

# 構成

1. カーネル密度推定
   - カーネル密度推定の推定手法
   - バンド幅の選択
   - 理論
   - 実データでの実験

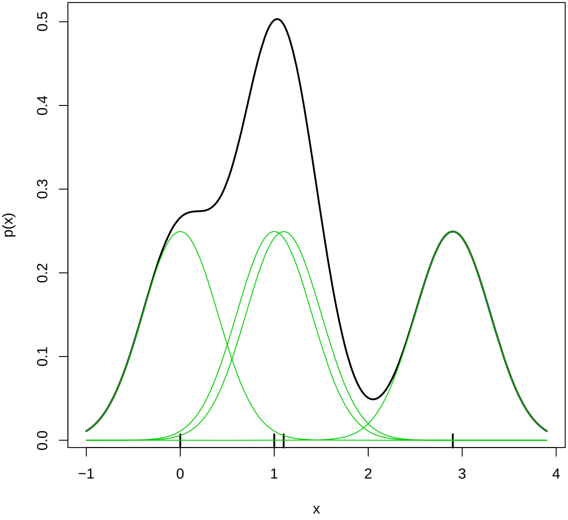
# カーネル密度推定量

$\{X_i\}_{i=1}^n$: データ (一次元とする)
カーネル密度推定量: ある**カーネル関数** $K$ を用いて,

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

1. $h > 0$ のことをバンド幅と呼ぶ. 適切に選択する必要がある.
2. $K$ は次の性質を満たすものとする:

$$\int K(x)\mathrm{d}x = 1, \quad \int xK(x)\mathrm{d}x = 0, \quad \int x^2K(x) > 0.$$

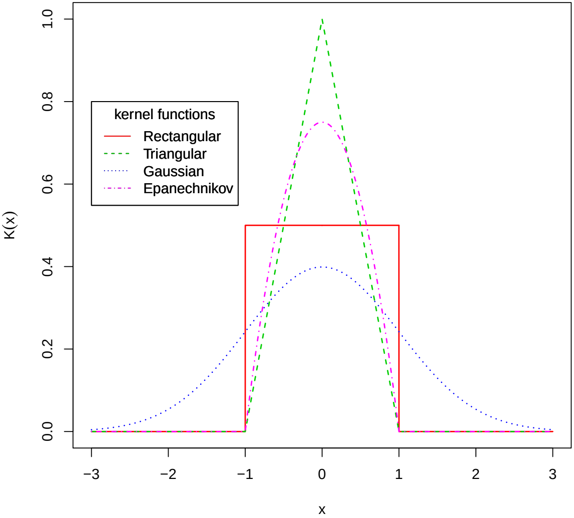

0.5
0.4
0.3
0.2
0.1
0.0
p(x)
−1
0
1
2
3
4
x

# カーネルの種類

次のようなカーネル関数がよく用いられる.

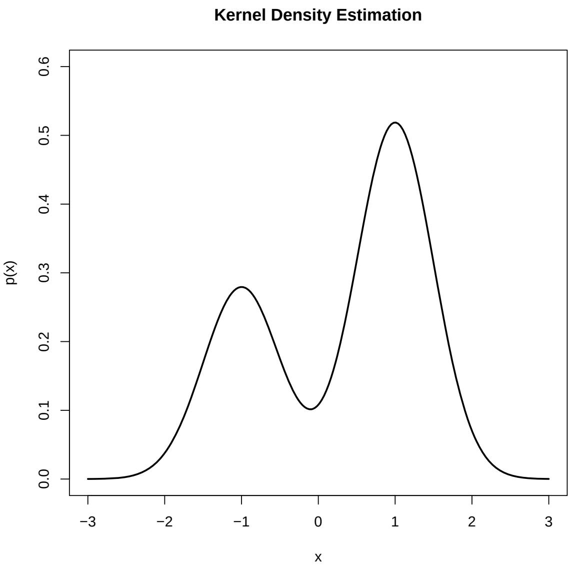
1. Gaussian:
$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$
2. Rectangular:
$$K(x) = \begin{cases} \frac{1}{2} & (|x| \leq 1), \\ 0 & (\text{otherwise}). \end{cases}$$
3. Triangular:
$$K(x) = \begin{cases} |x| & (|x| \leq 1), \\ 0 & (\text{otherwise}). \end{cases}$$
4. Epanechnikov:
$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & (|x| \leq 1), \\ 0 & (\text{otherwise}). \end{cases}$$

# カーネルの種類

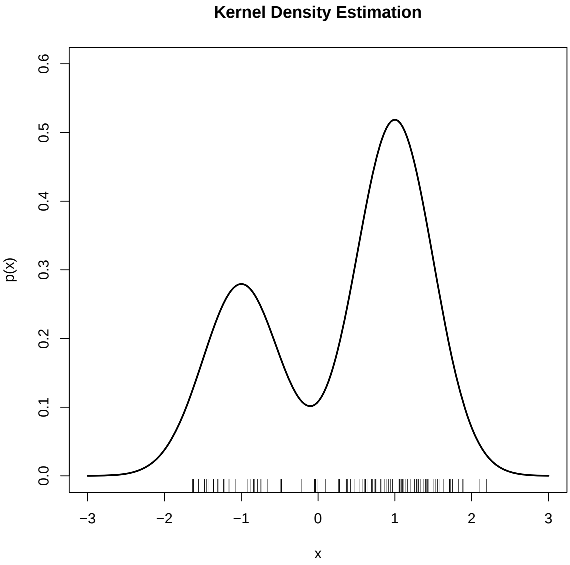


kernel functions
Rectangular
Triangular
Gaussian
Epanechnikov
K(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

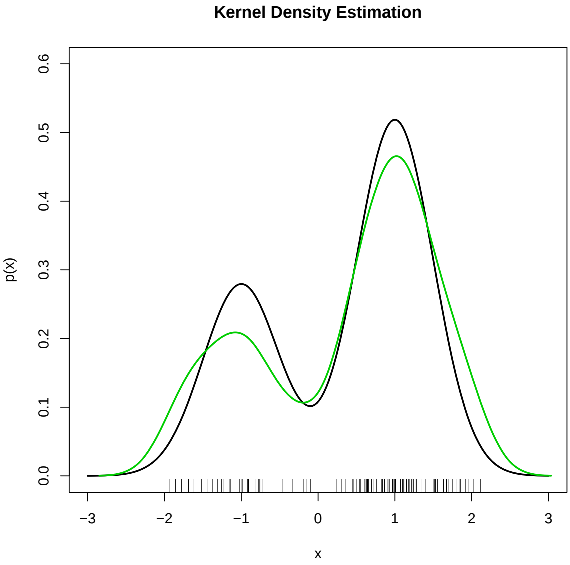
# カーネル密度推定量の様子



# カーネル密度推定量の様子

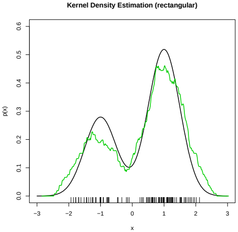


# カーネル密度推定量の様子

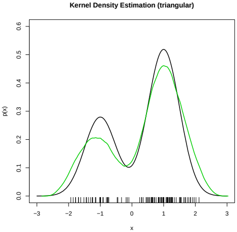


# カーネル関数の種類と推定結果

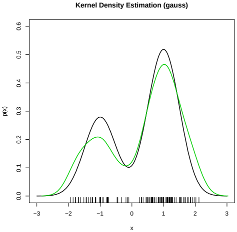
$$n = 100$$




rectangular



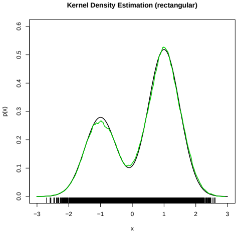

triangular



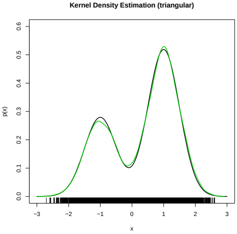

gauss

# カーネル関数の種類と推定結果

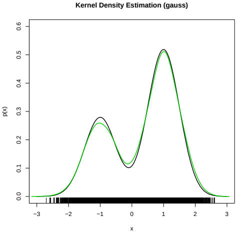
$n = 10000$




rectangular



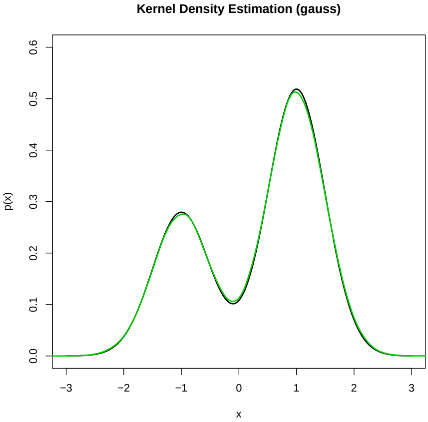

triangular




gauss

# カーネル関数の種類と推定結果

$n = 100000$

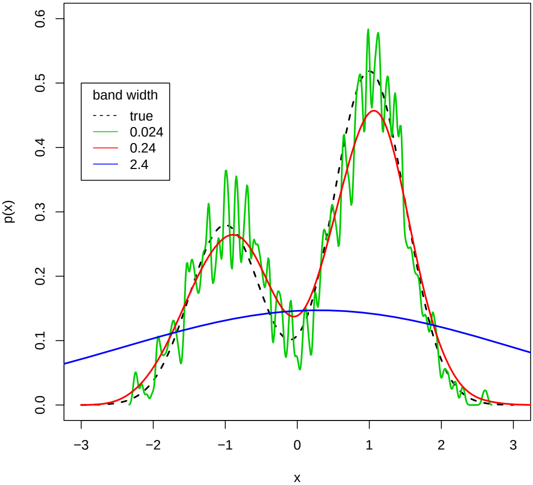



gauss

# 構成

1. カーネル密度推定
   - カーネル密度推定の推定手法
   - バンド幅の選択
   - 理論
   - 実データでの実験

# カーネル密度推定量とバンド幅




バンド幅は適切に選ぶ必要がある．

# 最小二乗クロスバリデーション

LSCV, Least Squares Cross Validation

真の密度との二乗距離 ($\hat{p}_h$ をバンド幅 $h$ のカーネル密度推定量とする)

$$\begin{aligned} &\int (\hat{p}_h(x) - p(x))^2 \mathrm{d}x \\ &= \underbrace{\int \hat{p}_h(x)^2 \mathrm{d}x - 2\int \hat{p}_h(x) p(x) \mathrm{d}x}_{=:J(h)} + \int p^2(x) \mathrm{d}x. \end{aligned}$$

$J(h)$ を最小化すれば良い．しかし $p(x)$ による積分がわからない→サンプルで代用．

ただし，手元にあるサンプルと $\hat{p}_h$ は相関がある のでクロスバリデーションする．

$\hat{p}_{h,(-i)}$: $i$ 番目のサンプル $X_i$ を抜いて推定した密度関数.

$$\widehat{J}(h) = \int \hat{p}_h(x)^2 \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \hat{p}_{h,(-i)}(X_i).$$

$\widehat{J}(h)$ を最小にする $h$ を採用すればよい.

# ヒューリスティクス

**Silverman の方法**

1. サンプル標準偏差: $\hat{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
2. サンプル分位点 $q(0.25)$: サンプルの 0.25 分位点.

$$\hat{\sigma} = \min\left\{\hat{s}, \frac{\hat{q}(0.75) - \hat{q}(0.25)}{1.34}\right\}$$

として，次のようにする:

$$\hat{h} = \frac{1.06\hat{\sigma}}{n^{1/5}}.$$

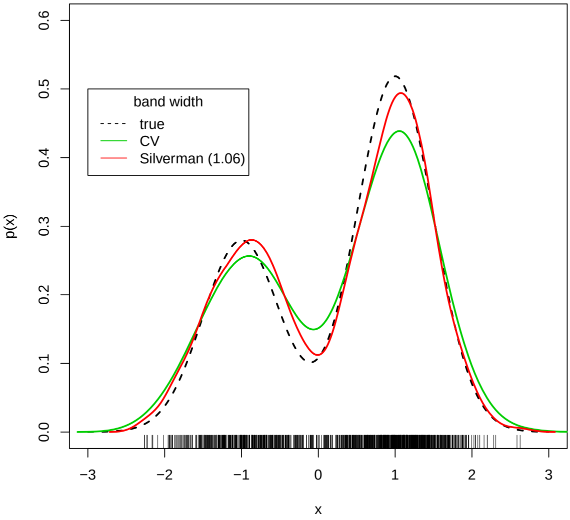
1.06 は人によって別の定数に置き換えたりする (1.06 を Scott のルール，0.9 を Silverman のルールと言う).

---

特に理論的根拠はないが，二乗誤差を漸近展開すると，

$$\hat{h} = \left[\frac{C_K}{n(\int (p''(x))^2 \mathrm{d}x)}\right]^{1/5},$$

(ただし $C_K = \frac{\int K(x)^2 \mathrm{d}x}{(\int x^2 K(x) \mathrm{d}x)^2}$) が漸近的に最小二乗誤差を与えることが示せて，上のヒューリスティクスは未知の値 $\int p''(x)^2 \mathrm{d}x$ に当たりを付ける経験則とみなせる.




ヒューリスティクスも良い推定結果を出している.

# 多変量カーネル密度推定

カーネル密度推定は多変量へ拡張できる．$\frac{1}{h}K(\frac{x-X_i}{h})$ の代わりに，

$$\prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x - X_i^{(j)}}{h_j}\right)$$

のようにする．なお，$X_i^{(j)}$ はサンプル $X_i$ の $j$ 番目の座標.

```
library(MASS)
d=kde2d(x,y,c(bandwidth.nrd(x),bandwidth.nrd(y)),n=80)
image(d,xlab="latitude",ylab="longitude")
```

# 構成

1. カーネル密度推定
   - カーネル密度推定の推定手法
   - バンド幅の選択
   - 理論
   - 実データでの実験

# 平均二乗誤差

あるバンド幅 $h > 0$ に対して，カーネル密度推定量を

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

とする．

- 推定量 $\hat{p}$ の点 $x$ における**平均二乗誤差 (Mean Squared Error)**：

$$\mathrm{MSE}(\hat{p}(x), h) := \mathrm{E}[(\hat{p}(x) - p(x))^2].$$

  ただし，期待値 $\mathrm{E}[\cdot]$ はサンプル $\{X_i\}_{i=1}^n$ の出方についてとる．
- $x$ に関して積分した**積分平均二乗誤差 (Integrated MSE)**:

$$\mathrm{IMSE}(\hat{p}(x), h) := \int \mathrm{E}[(\hat{p}(x) - p(x))^2]\mathrm{d}x.$$

**Q:**バンド幅 $h$ に対して平均二乗誤差はどのように振る舞うか？

# バイアスとバリアンスの分解

$h$ はサンプルサイズ $n$ に応じて小さくなっていくものとする $h \to 0$.

$$
\begin{aligned}
&\mathrm{MSE}(\hat{p}(x), h) \\
&= \mathrm{E}[(\hat{p}(x) - p(x))^2] \\
&= \mathrm{E}[(\hat{p}(x) - \mathrm{E}[\hat{p}(x)] + \mathrm{E}[\hat{p}(x)] - p(x))^2] \\
&= \mathrm{E}\left[(\hat{p}(x) - \mathrm{E}[\hat{p}(x)])^2 - 2(p(x) - \mathrm{E}[\hat{p}(x)])(\hat{p}(x) - \mathrm{E}[\hat{p}(x)]) + (p(x) - \mathrm{E}[\hat{p}(x)])^2\right] \\
&= \underbrace{\mathrm{E}\left[(\hat{p}(x) - \mathrm{E}[\hat{p}(x)])^2\right]}_{\text{バリアンス項}} + \underbrace{\mathrm{E}[(p(x) - \mathrm{E}[\hat{p}(x)])^2]}_{\text{バイアス項}},
\end{aligned}
$$

バイアス項とバリアンス項をそれぞれ評価しよう.

- バイアス項の評価

まず

$$
\begin{aligned}
\mathrm{E}[\hat{p}(x)] &= \mathrm{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right] \\
&= \mathrm{E}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right] = \int K(u)p(x-hu)\mathrm{d}u
\end{aligned}
$$

である.

さて，ここで $p(x-hu)$ をテイラー展開すると

$$p(x-hu) = p(x) - hu\frac{\mathrm{d}p(x)}{\mathrm{d}x} + \frac{h^2u^2}{2}\frac{\mathrm{d}^2p(x)}{\mathrm{d}^2x} + o(h^2)$$

となる．よって，

$$\begin{aligned}&\int K(u)p(x-hu)\mathrm{d}u - p(x)\\&= -h\frac{\mathrm{d}p(x)}{\mathrm{d}x}\int uK(u)\mathrm{d}u + \frac{h^2}{2}\frac{\mathrm{d}^2p(x)}{\mathrm{d}^2x}\int u^2K(u)\mathrm{d}u + o(h^2)\end{aligned}$$

である．ここで，カーネル関数の条件より

$$\int uK(u)\mathrm{d}u = 0$$

なので，

$$\text{バイアス項} = \left(\frac{h^2}{2}\frac{\mathrm{d}^2p(x)}{\mathrm{d}^2x}\int u^2K(u)\mathrm{d}u\right)^2 + o(h^2).$$

# バリアンス項の評価

- バリアンス項の評価

分散の定義より，バリアンス項は $\hat{p}(x)$ の分散 (分散を $\mathrm{Var}(\cdot)$ と書く) である．よって

$$
\begin{aligned}
&\mathrm{E}\left[(\hat{p}(x) - \mathrm{E}[\hat{p}(x)])^2\right] \\
&= \mathrm{Var}(\hat{p}(x)) \\
&= \mathrm{Var}\left(\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)\right) \\
&= \frac{1}{n} \mathrm{Var}\left(\frac{1}{h} K\left(\frac{x - X}{h}\right)\right) \\
&= \frac{1}{n}\left\{\mathrm{E}\left[\left(\frac{1}{h} K\left(\frac{x - X}{h}\right)\right)^2\right] - \mathrm{E}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\right]^2\right\} \\
&= \frac{1}{nh} \int K(u)^2 p(x - hu)\mathrm{d}u - \frac{1}{n}\mathrm{E}[\hat{p}(x)]^2 \\
&= \frac{p(x)}{nh} \int K(u)^2 \mathrm{d}u + o(1/(nh)).
\end{aligned}
$$

以上より，

$$\mathrm{MSE}(\hat{p}(x), h) = \frac{p(x)}{nh}\int K(u)^2 \mathrm{d}u + \frac{h^4}{4}\left(\frac{\mathrm{d}^2 p(x)}{\mathrm{d}^2 x}\int u^2 K(u)\mathrm{d}u\right)^2 + o(1/(nh) + h^4).$$

両辺 $x$ で積分することで

$$\mathrm{IMSE}(\hat{p}(x), h) = \frac{1}{nh}\int K(u)^2 \mathrm{d}u + \frac{h^4}{4}\int\left(\frac{\mathrm{d}^2 p(x)}{\mathrm{d}^2 x}\int u^2 K(u)\mathrm{d}u\right)^2 \mathrm{d}x + o(\frac{1}{nh} + h^4).$$

右辺の第一項と第二項を最小化する $h$ は

$$h^* = \frac{1}{n^{1/5}}\left(\frac{\int K(u)^2 \mathrm{d}u}{\int \left(p''(x)\right)^2 \mathrm{d}x \left(\int u^2 K(u)\mathrm{d}u\right)^2}\right)^{1/5}$$

で与えられる．ただし，係数には $p''(x)$ が入っているので，クロスバリデーションやヒューリスティクスを用いて係数を決める必要がある．

最適なバンド幅 $h^*$ を用いたとき，IMSE は

$$\mathrm{IMSE}(\hat{p}(x), h^*) = \frac{5}{4n^{4/5}} \frac{\left[\int (p''(x))^2 \,\mathrm{d}x \left(\int u^2 K(u)\mathrm{d}u\right)^2\right]^{1/5}}{\left[\int K(u)^2 \mathrm{d}u\right]^{4/5}} + o(1/n^{4/5})$$
$$= O(n^{-4/5})$$

となる．

# 次元の呪い

$d$ 次元の場合も同様にして最適なバンド幅に対して,

$$\mathrm{IMSE} = O\left(\frac{1}{n^{4/(d+4)}}\right)$$

が示される. $d$ が大きくなると収束レートが非常に遅くなる.
これを「次元の呪い」と言う.

# 構成

1. カーネル密度推定
   - カーネル密度推定の推定手法
   - バンド幅の選択
   - 理論
   - 実データでの実験

# カーネル密度推定を行える R 関数

使い方はスクリプトを参照

1 次元のカーネル密度推定

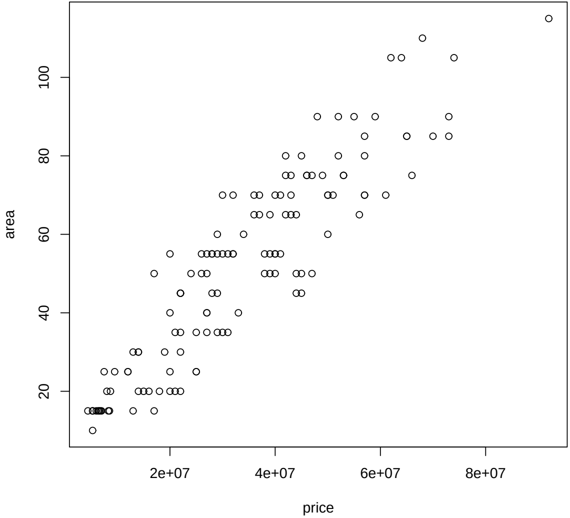
- density(x,bw,kernel):
  - バンド幅: bw = "nrd0" がデフォルト (シルバーマンの方法で 0.9 を採用), bw="'nrd' で 1.06. bw="ucv" で (バイアス修正した) クロスバリデーション, bw="bcv" でクロスバリデーション.
  - カーネル関数: kernel="gaussian" がデフォルト. 他にも"epanechnikov", "rectangular", "triangular", "biweight","cosine", "optcosine" が指定可能.
- bkde: 'KernSmooth' パッケージに入っている.

二次元以上

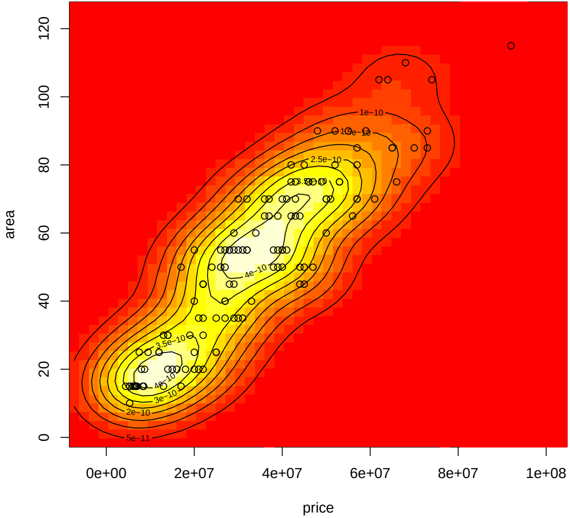
- kde2d: 'MASS' パッケージに入っている. 2 次元用.
- bkde2d: 'KernSmooth' パッケージに入っている. 2 次元用.
- kde: 'ks' パッケージに入っている. 3 次元以上の密度推定も行える.

MASS パケージの bandwidth.nwd は bw.nwd の 4 倍の値を返す.

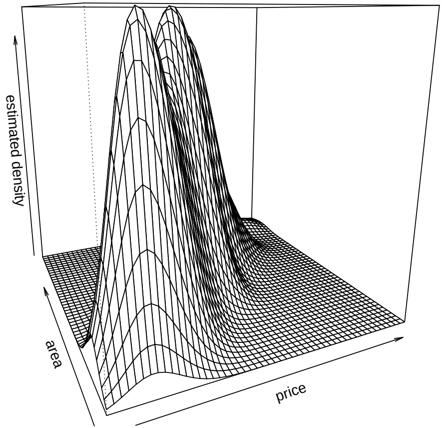
# 世田谷区中古マンション価格データ



# 世田谷区中古マンション価格データ



# 世田谷区中古マンション価格データ



# まとめ

ノンパラメトリックな密度推定手法を紹介した.

1. カーネル密度推定

バンド幅を CV などを用いてうまく調整する必要があった.

# 講義情報ページ

`http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/dataanalysis.html`