Kernel density estimation

Victor Kitov

v.v.kitov@yandex.ru

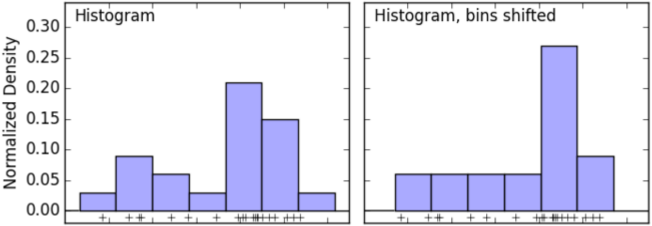
Nonparametric density estimation

- Need a non parametric density estimation.

[1] example by Jake Vanderplas

Nonparametric density estimation

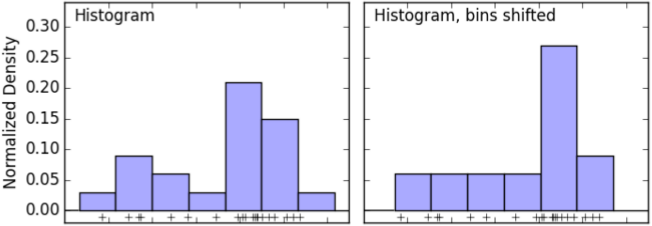
- Need a non parametric density estimation.
- Histogram [1]



[1] example by Jake Vanderplas

Nonparametric density estimation

- Need a non parametric density estimation.
- Histogram [1]



- Bins selection problem!

[1] example by Jake Vanderplas

Kernel density estimation

Idea

Center each block on the point it represents.

Kernel density estimation

Idea

Center each block on the point it represents.

Implementation:

Define individual block

$$K(u) = \frac{1}{2}\mathbb{I}[|u| \le 1]$$

It has the properties

$$K(u) \ge 0, \quad \int_{-\infty}^{+\infty} K(u)du = 1$$

Kernel density estimation

Idea

Center each block on the point it represents.

Implementation:

Define individual block

$$K(u)=\frac{1}{2}\mathbb{I}[|u|\le 1]$$

It has the properties

$$K(u)\ge 0,\quad \int_{-\infty}^{+\infty}K(u)du=1$$

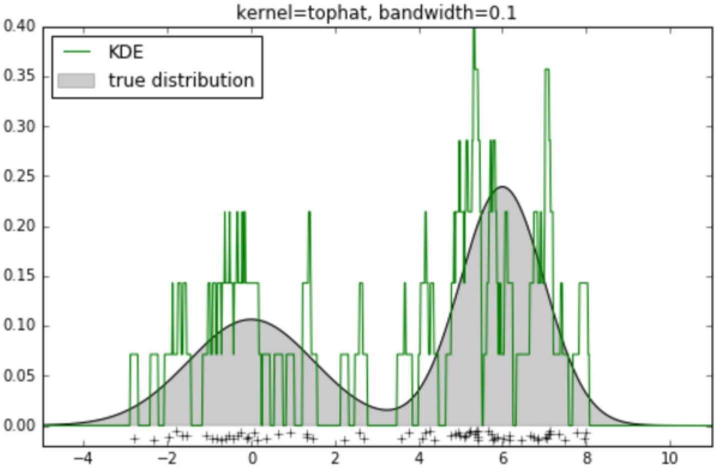
Kernel density estimation (KDE)

$$\widehat{p}(x)=\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x-x_i}{h}\right)$$

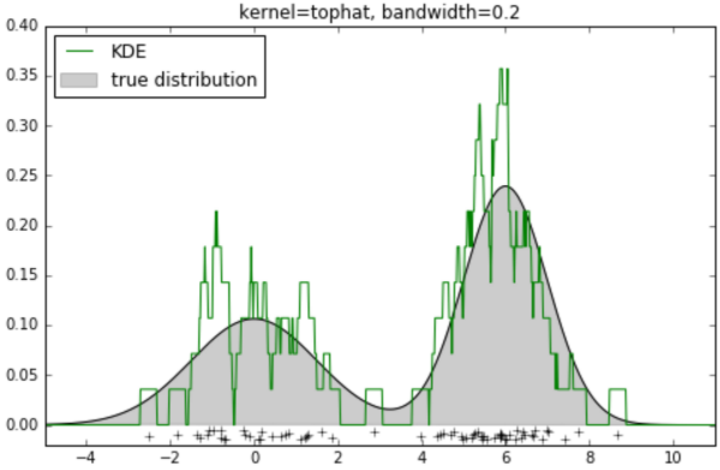
Comments

- $K(u)$ is called *kernel*
- $K(u) = \frac{1}{2}\mathbb{I}[|u| \le 1]$ is called *tophat kernel*
- h is called *bandwidth*
- h controls the smoothness of KDE - how?

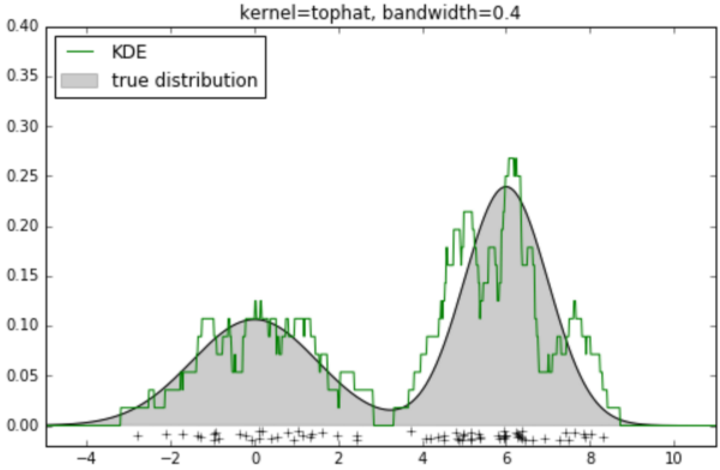
Example: tophat KDE



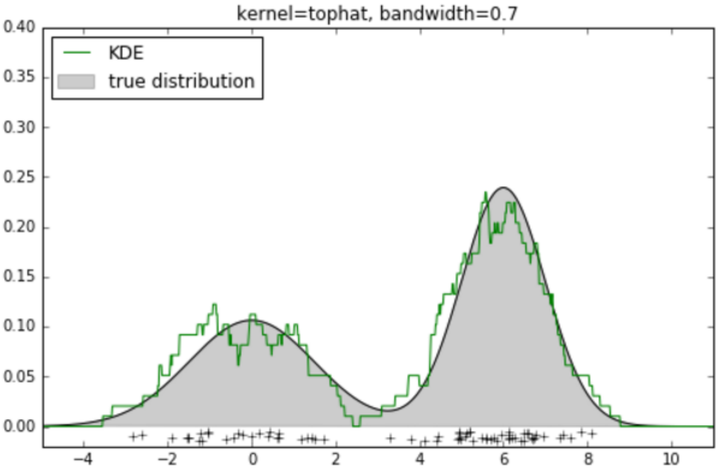
Example: tophat KDE



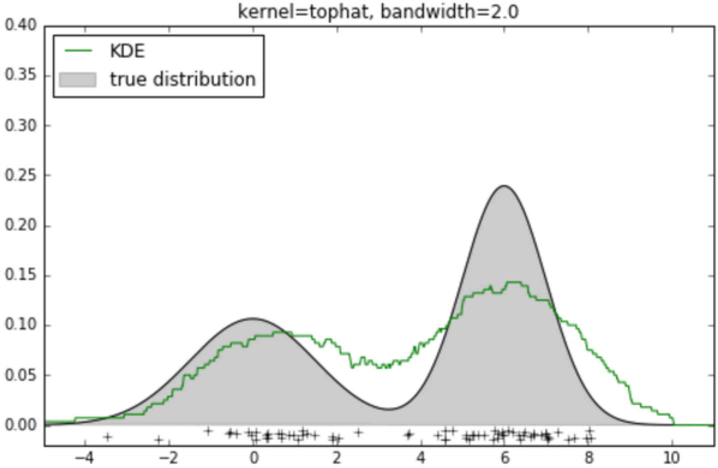
Example: tophat KDE



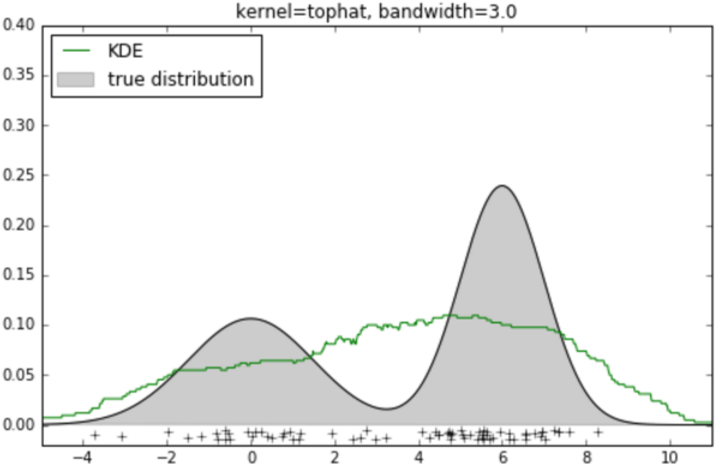
Example: tophat KDE



Example: tophat KDE



Example: tophat KDE



Extension of tophat kernel

Problems of tophat:

- Resulting KDE $\widehat{p}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)$ is **discontinious.**
- Impact of close points x_i does not change with $\rho(x, x_i)$

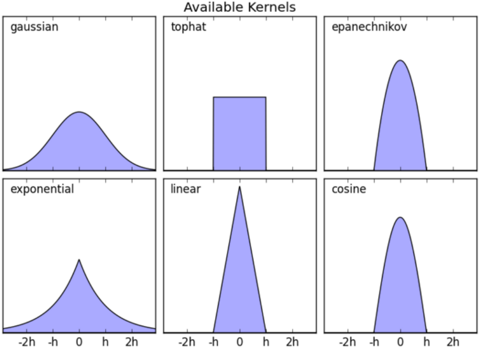
[2] example by Jake Vanderplas

Extension of tophat kernel

Problems of tophat:

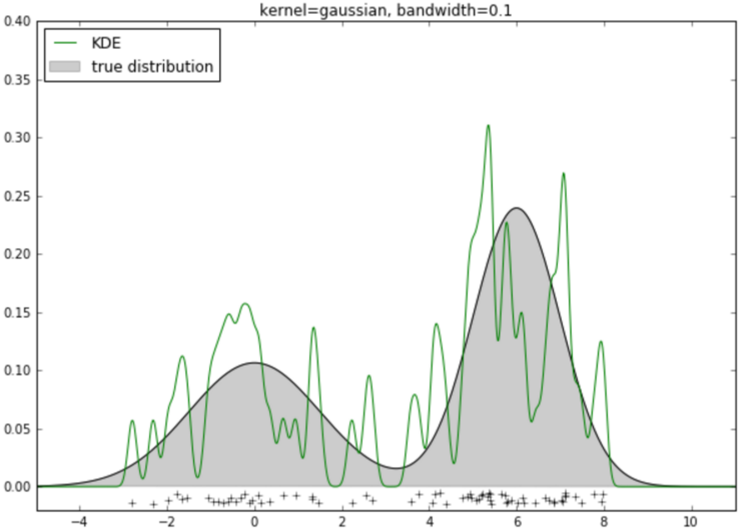
- Resulting KDE $\widehat{p}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)$ is **discontinious.**
- Impact of close points x_i does not change with $\rho(x, x_i)$

We can use smooth unimodal kernels[2]:

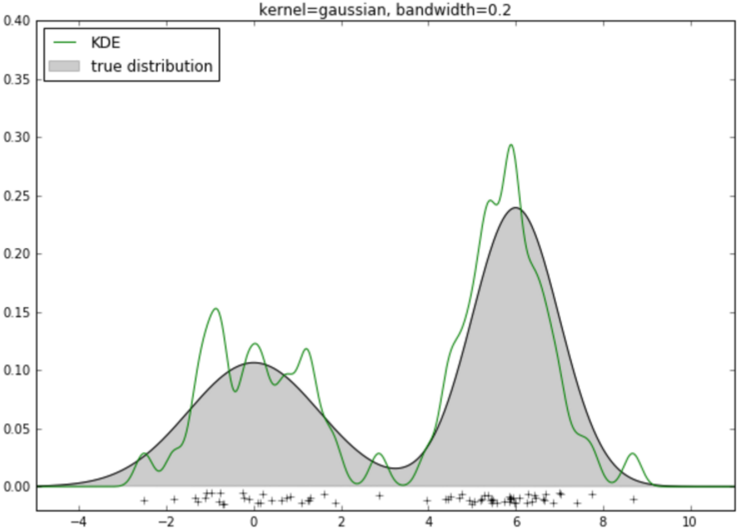


[2] example by Jake Vanderplas

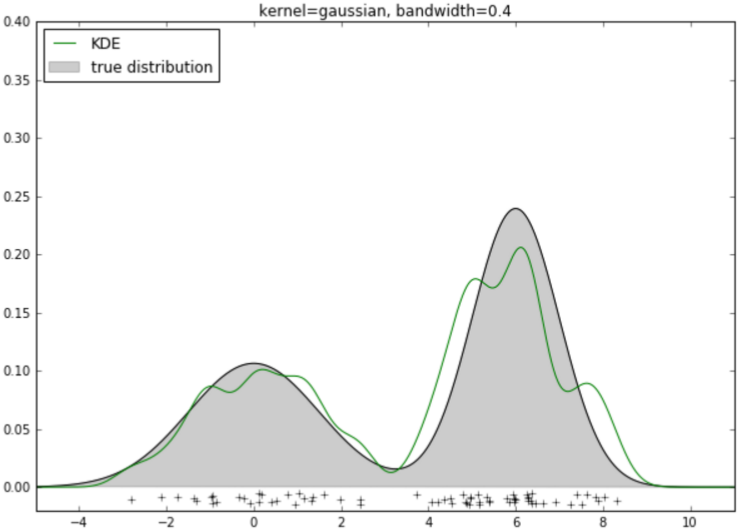
Example: Gaussian KDE



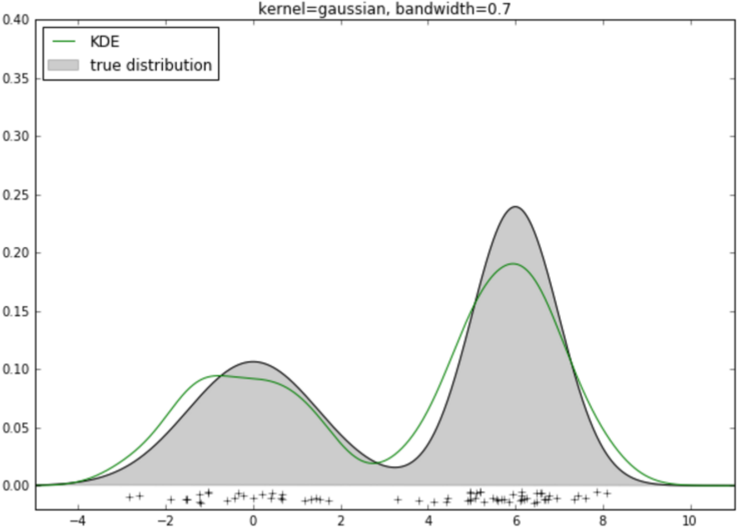
Example: Gaussian KDE



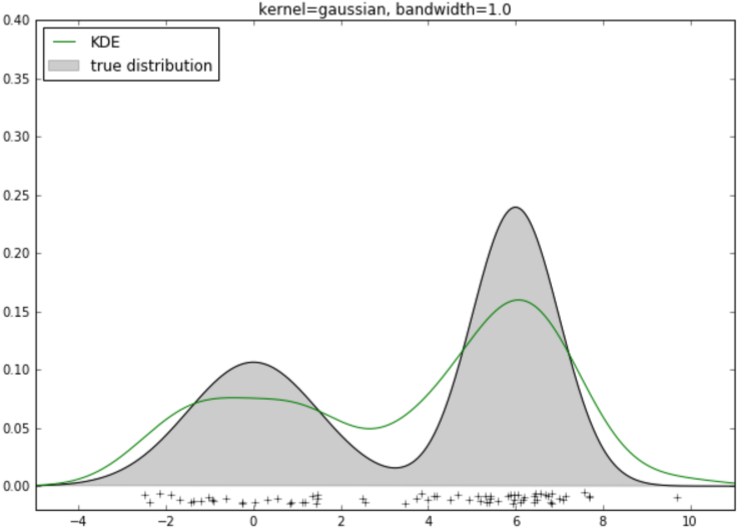
Example: Gaussian KDE



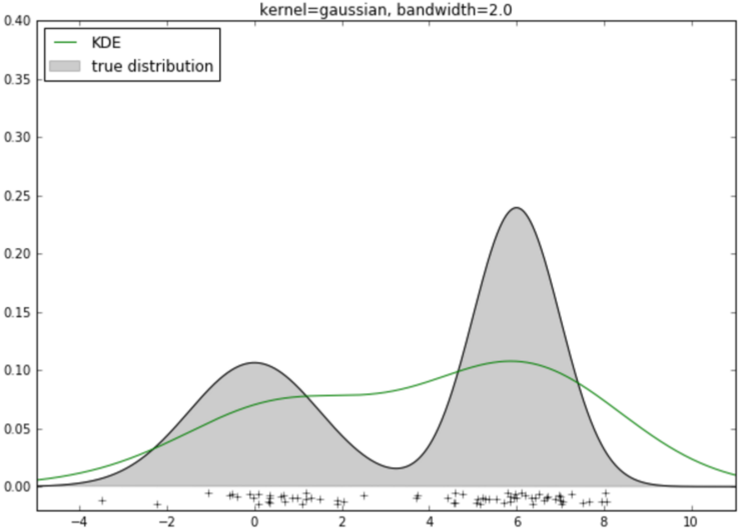
Example: Gaussian KDE



Example: Gaussian KDE



Example: Gaussian KDE



Mathematical definitions of kernels

name	definition of $K(u)$
tophat (rectangular)	$\frac{1}{2}\mathbb{I}[\lvert u\rvert \le 1]$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$
biweight	$\propto \left[(1-u^2)^2\right]_+$
triangular	$[1-\lvert u\rvert]_+$
Epanechnikov	$\propto \left[1-u^2\right]_+$

Mathematical definitions of kernels

name	definition of $K(u)$
tophat (rectangular)	$\frac{1}{2}\mathbb{I}[\|u\| \le 1]$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$
biweight	$\propto \left[(1-u^2)^2\right]_+$
triangular	$[1-\|u\|]_+$
Epanechnikov	$\propto \left[1-u^2\right]_+$

Comments:

- type of kernel affect smoothness but not accuracy of approximation.
- selection of bandwidth is more important for accuracy

Techincal conditions for consistency

Kernel consistency

Kernel density estimation $\widehat{p}(x)$ is consistent if
$\mathbb{E}[(\widehat{p}(x) - p(x))^2] \xrightarrow{N\to\infty} 0$ for $\forall x$

Techincal conditions for consistency

Kernel consistency

Kernel density estimation $\widehat{p}(x)$ is consistent if
$\mathbb{E}[(\widehat{p}(x) - p(x))^2] \xrightarrow{N\to\infty} 0$ for $\forall x$

Sufficient conditions for consistency:

- Bandwidth convergence:
- $\lim_{N\to\infty} h(N) = 0$
    - $\lim_{N\to\infty} Nh(N) = \infty$

Techincal conditions for consistency

Kernel consistency

Kernel density estimation $\widehat{p}(x)$ is consistent if

$\mathbb{E}[(\widehat{p}(x) - p(x))^2] \xrightarrow{N\to\infty} 0$ for $\forall x$

Sufficient conditions for consistency:

- Bandwidth convergence:
- $\lim_{N\to\infty} h(N) = 0$
    - $\lim_{N\to\infty} Nh(N) = \infty$
- Kernel regularity:
    - $\int |K(u)|du < \infty$
    - $\int K(u)du = 1$
    - $\sup_u K(u) < \infty$
    - $\lim_{u\to\infty} |uK(u)| = 0$

Multivariate extension

Multivariate kernels:

$$\widehat{p}(\mathbf{x}) = \frac{1}{Nh^D}\sum_{i=1}^{N} K\left(\frac{1}{h}(\mathbf{x}-\mathbf{x_i})\right)$$

name	definition of $K(\mathbf{u})$
Gaussian	$\frac{1}{(2\pi)^{D/2}}e^{-\frac{\mathbf{u}^T\mathbf{u}}{2}}$
Epanechnikov	$\propto \left[1-\mathbf{u}^T\mathbf{u}\right]_+$
Product of univariate kernels	$\prod_{d=1}^{D} K_d\left(\frac{\mathbf{x}^d-\mathbf{x}_n^d}{h}\right)$

Multivariate extension

Distance based kernels:

$$\widehat{p}(\mathbf{x}) = \frac{1}{Nh^D} \sum_{i=1}^{N} K\left(\rho(\mathbf{x}, \mathbf{x_i})\right)$$

name	definition of $K(\rho(x, x_i))$
Gaussian	$\frac{1}{(2\pi)^{D/2}} e^{-\frac{\rho^2(x,x_i)}{2}}$
Epanechnikov	$\propto \left[1 - \rho^2(x, x_i)\right]_+$

Selection of h

Bandwidth selection guidline

The more dense is sample points distribution, the smaller should be h.

Selection of h

Bandwidth selection guidline

The more dense is sample points distribution, the smaller should be h.

Constant bandwidth:

- $h = \frac{1}{N}\sum_{i=1}^{N} d_{iK}$, d_{iK}-distance from x_i to its K-th nearest neighbour
- out-of-sample maximum likelihood

Variable bandwidth (for significantly variable densities):

- $h(x)$ - distance to $K-th$ nearest neighbour from x

Parzen window method

Estimate $p(x|y)$ with KDE:

$$p(x|y) = \frac{1}{N_y h^D} \sum_{i:y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right)$$

Bayes decision rule gives:

$$\widehat{y} = \arg\max_y p(y|x) \propto p(y)p(x|y)$$
$$= \arg\max_y \frac{N_y}{N} \frac{1}{N_y h^D} \sum_{i:y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right)$$
$$\widehat{y}(x) = \arg\max_y \sum_{i:y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right)$$

k-NN

- For fixed x and k take $K(u) = \mathbb{I}[|u| \le 1]$ and $h(x) = \arg\min_{R:\ \#\{x_n : \rho(x,x_n) \le R\} = K} R$ (distance to k-th nearest neighbour).
- Then Parzen window method reduces to k-nearest neighbour:

$$\widehat{y}(x) = \arg\max_y \sum_{i:y_i=y} K\left(\frac{\rho(x,x_i)}{h(x)}\right) = \arg\max_y \sum_{i:\rho(x,x_i)\le h(x)} \mathbb{I}[y_i = y]$$

- Density estimation comparison:
  - With fixed h we fix area and see how many points fall inside
  - In k-NN we fix k and average by area spreaded over k nearest neighbours$\{x :\ \rho(x, x_i) \le h(x)\}$