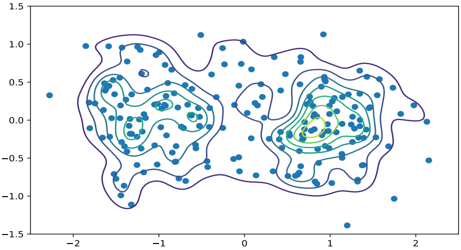
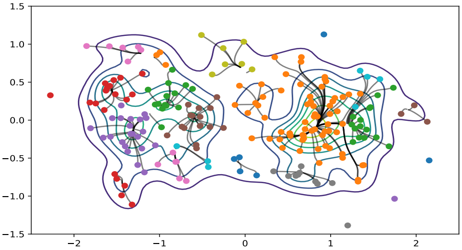
Karlsruher Institut für Technologie

Kernel Density Estimates and Mean Shift Clustering

Jonas Spinner | February 4, 2019

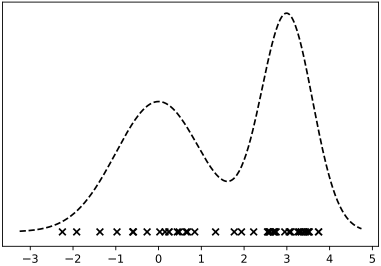
ANALYTICS AND STATISTICS AT THE INSTITUTE OF OPERATIONS RESEARCH





Outline

Density estimation



- **Task**: Given samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, estimate the underlying density.

The kernel density estimate

The kernel density estimate is defined as:

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

- K is a *kernel function*.
- h is a *bandwidth parameter*.

- When $\int_{\mathbb{R}^d} K(\boldsymbol{x})dx = 1$ and $K(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$ then $\hat{f}(\boldsymbol{x})$ is a valid probability density function.

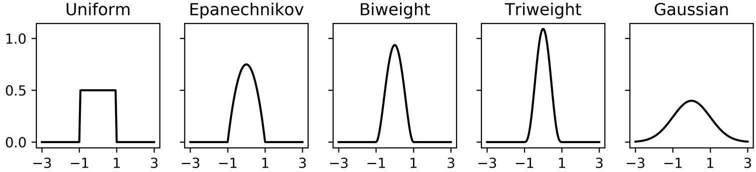
Popular kernel functions

Radially symmetric kernel functions are kernel functions which can be represented as

$$K(\boldsymbol{x}) = c_{k,d} \; k\left(\|\boldsymbol{x}\|^2\right)$$

- $k(u) : [0, \infty) \to [0, \infty)$ is called the profile of K.
- For example the gaussian kernel has the profile $k(u) = \exp(-\frac{1}{2}u)$.
- Nearly all popular kernel belong to this class of kernels.

Popular kernel functions



Name	Profile support	$k(u)$	$-k'(u)$	$K(\boldsymbol{x})$
Uniform	$u \in [0, 1]$	1	0	$\mathrm{vol}\,(S_d)^{-1}$
Epanechnikov	$u \in [0, 1]$	$1 - u$	1	$\frac{1}{2}\mathrm{vol}(S_d)^{-1}(d+2)\left(1 - \lVert\boldsymbol{x}\rVert^2\right)$
Biweight	$u \in [0, 1]$	$(1-u)^2$	$2(1-u)$	$\propto \left(1 - \lVert\boldsymbol{x}\rVert^2\right)^2$
Triweight	$u \in [0, 1]$	$(1-u)^3$	$3(1-u)^2$	$\propto \left(1 - \lVert\boldsymbol{x}\rVert^2\right)^3$
Gaussian	$u \in [0, \infty)$	$\exp\left(-\frac{1}{2}u\right)$	$\frac{1}{2}\exp\left(-\frac{1}{2}u\right)$	$(2\pi)^{-d/2}\exp\left(-\frac{1}{2}\lVert\boldsymbol{x}\rVert^2\right)$

Bandwidth

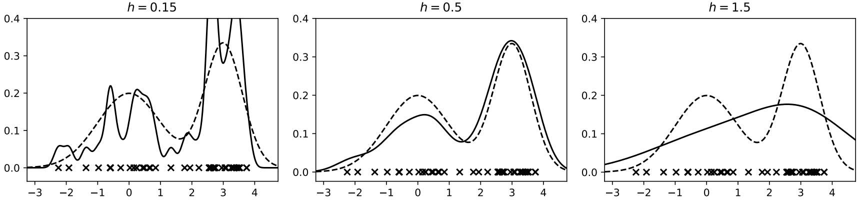


Abbildung: Gaussian kernel, $n = 50$.

- The choice of bandwidth is a bias-variance tradeoff for the estimate $\hat{f}(\boldsymbol{x})$.
- A small bandwidth results in high variance, a large bandwidth introduces a bias.

Bandwidth

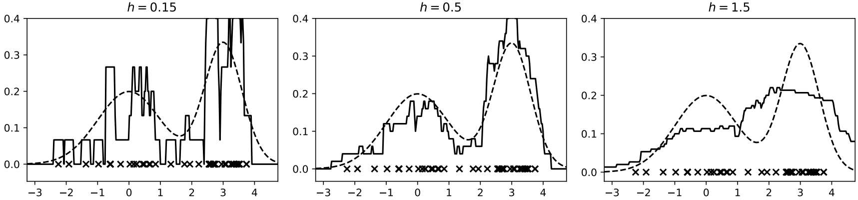


Abbildung: Uniform kernel, $n = 50$.

- The choice of bandwidth is a bias-variance tradeoff for the estimate $\hat{f}(\boldsymbol{x})$.
- A small bandwidth results in high variance, a large bandwidth introduces a bias.

Bandwidth

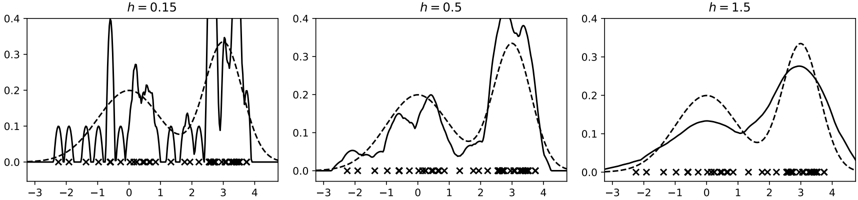


Abbildung: Epanechnikov kernel, $n = 50$.

- The choice of bandwidth is a bias-variance tradeoff for the estimate $\hat{f}(\boldsymbol{x})$.
- A small bandwidth results in high variance, a large bandwidth introduces a bias.

History

- Introduced by Fukunaga & Hostetler (1975)
- Popularized for computer vision tasks Comaniciu & Meer (2002) and Comaniciu et al. (2003). Image segmentation, image filtering and object tracking

Locally weighted mean

The weighted mean

$$\boldsymbol{\mu}^* = \frac{\sum_{i=1}^n \boldsymbol{x}_i w_i}{\sum_{i=1}^n w_i}$$

The locally weighted mean has weights depending on the distance to a data point $\boldsymbol{x}$

$$\boldsymbol{\mu}^*(\boldsymbol{x}) = \frac{\sum_{i=1}^n \boldsymbol{x}_i \; g\left(\left\| \frac{\boldsymbol{x}-\boldsymbol{x}_i}{h} \right\|^2\right)}{\sum_{i=1}^n g\left(\left\| \frac{\boldsymbol{x}-\boldsymbol{x}_i}{h} \right\|^2\right)}$$

$$\boldsymbol{x}^{(t+1)} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i \; g\left(\left\|(\boldsymbol{x}^{(t)} - \boldsymbol{x}_i)/h\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|(\boldsymbol{x}^{(t)} - \boldsymbol{x}_i)/h\right\|^2\right)} \qquad \text{for } i = 1, 2, \ldots$$

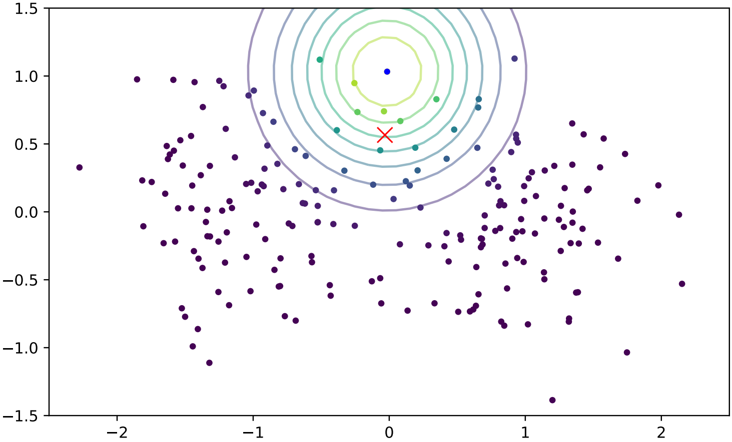
$$= \boldsymbol{x}^{(t)} + \boldsymbol{m}\left(\boldsymbol{x}^{(t)}\right)$$

Connecting the mean shift vector and kernel density estimation

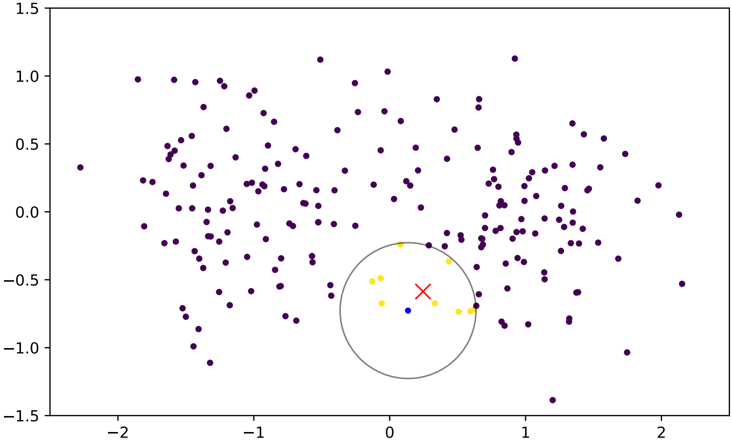
- **Result**: the mean shift vector $\boldsymbol{m}(\boldsymbol{x})$ points into the gradient direction of a kernel density estimate.

$$\boldsymbol{m}(\boldsymbol{x}) = \mu^*(\boldsymbol{x}) - \boldsymbol{x} = \frac{h^2 c_{g,d}}{2c_{k,d}} \frac{\nabla \hat{f}_{h,K}(\boldsymbol{x})}{\hat{f}_{h,G}(\boldsymbol{x})}$$

Convergence – Gaussian kernel



Convergence – Uniform kernel

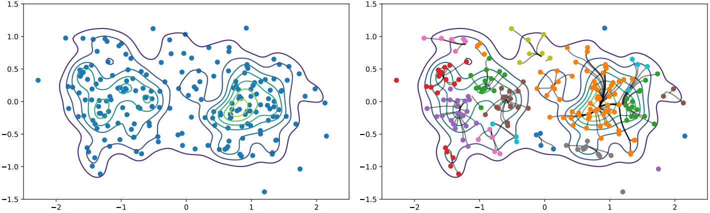


Bandwidth effects

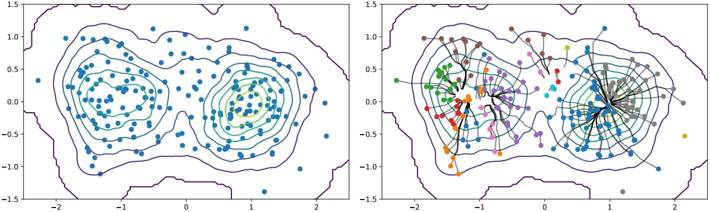
- The choice of bandwidth influences the density estimation and therefore the clustering outcome

- Small bandwidth $\Rightarrow$ many density peaks / clusters
- Large bandwidth $\Rightarrow$ few peaks / clusters

Small bandwidth

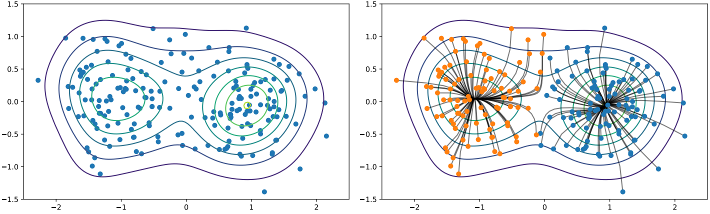


$h = 0.15$, gaussian kernel

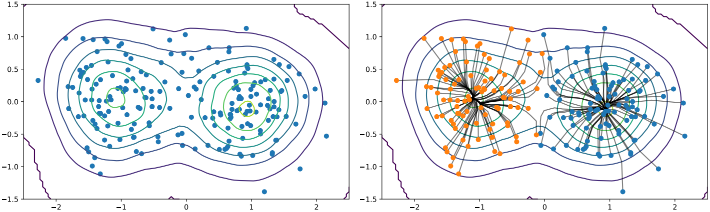


$h = 0.5$, uniform kernel

Middle bandwidth

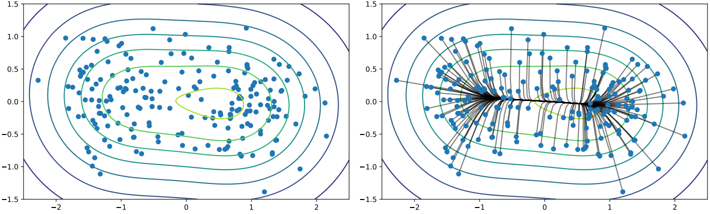


$h = 0.35$, gaussian kernel

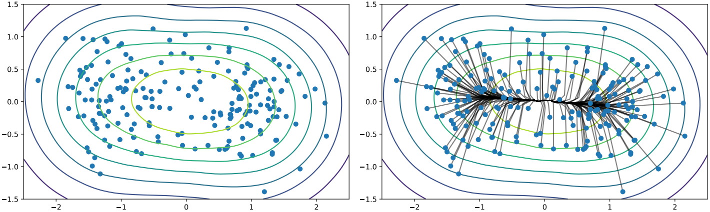


$h = 0.8$, uniform kernel

Large bandwidth



$h = 0.8$, gaussian kernel



$h = 1.7$, uniform kernel

Discussion

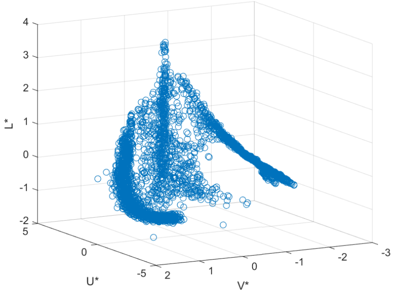
Advantages

- No prior assumption on cluster shapes. Complex and non-convex shapes are possible
- Only has one tuning parameter, the bandwidth
- No restriction on number of clusters
- Outliers do not affect the clustering

Disadvantages

- Density estimation fails for high dimensions ($\approx d > 5$)
- Bad computational complexity $\mathcal{O}(Tn^2)$
- Finding a good bandwidth is hard

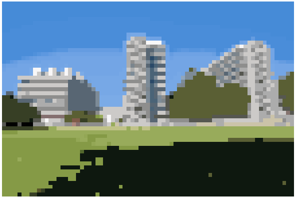
Application – Image segmentation



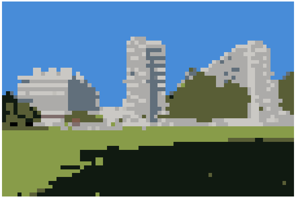
- Image data can be represented as data points. The pixels can be clustered and the clustering results in a segmentation of the original image.

Image segmentation – Gaussian kernel

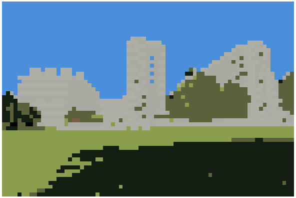
(a) $h = 0.1$



(b) $h = 0.2$

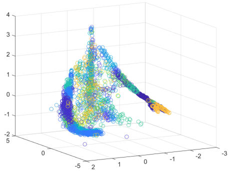


(c) $h = 0.3$

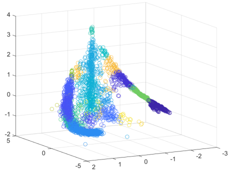


(d) $h = 0.4$

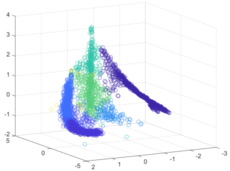
Image segmentation – Gaussian kernel



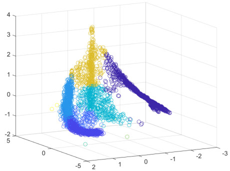
(a) $h = 0.1$



(b) $h = 0.2$



(c) $h = 0.3$



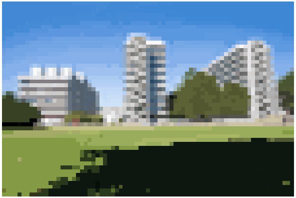
(d) $h = 0.4$

Abbildung: caption

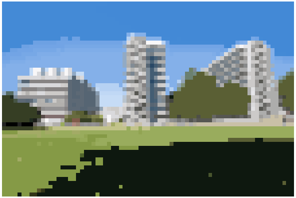
Image segmentation – Uniform kernel

(a) $h = 0.1$

(b) $h = 0.2$

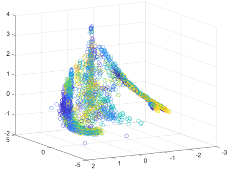


(c) $h = 0.3$

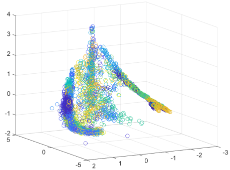


(d) $h = 0.4$

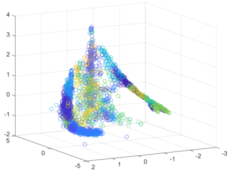
Image segmentation – Uniform kernel



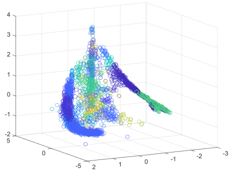
(a) $h = 0.1$



(b) $h = 0.2$



(c) $h = 0.3$



(d) $h = 0.4$

References I

Comaniciu, D. & Meer, P. (2002), 'Mean shift: a robust approach toward feature space analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(5), 603–619.

Comaniciu, D., Ramesh, V. & Meer, P. (2003), 'Kernel-based object tracking', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(5), 564–577.

Fukunaga, K. & Hostetler, L. (1975), 'The estimation of the gradient of a density function, with applications in pattern recognition', *IEEE Transactions on Information Theory* **21**(1), 32–40.

Literatur

-