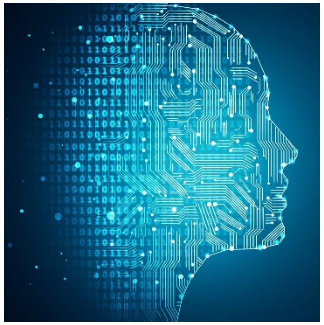
Machine Learning

Kernel Density Estimation

Kevin Moon (kevin.moon@usu.edu)

STAT/CS 5810/6655

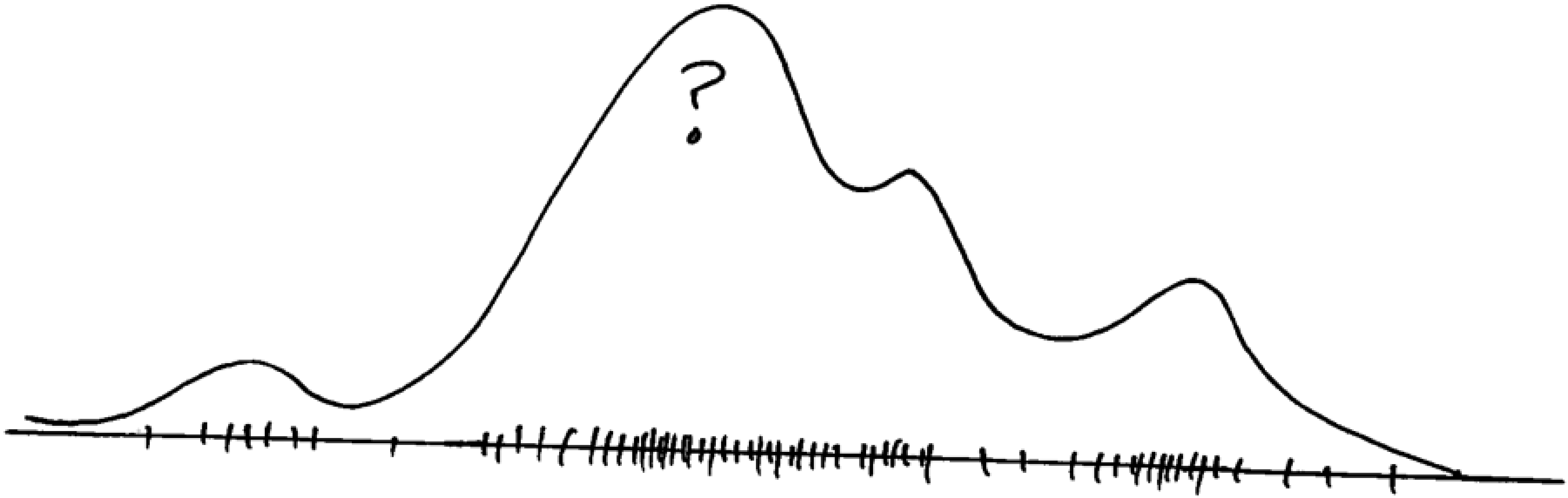


Outline

1. Density estimation
2. Kernel density estimation
3. Bias and Variance analysis
4. Model selection
5. k-nn density estimation

Density Estimation

- Random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim f$
  - f is an unknown pdf
- *Density estimation* is an unsupervised learning problem
  - No labels in the training data
- Goal is to estimate f from the random sample
  - In general, the $\boldsymbol{X}_i$s may be multidimensional



Why Density Estimation?

Classification

- Can construct a "plug-in" classifier using the Bayes classifier formula

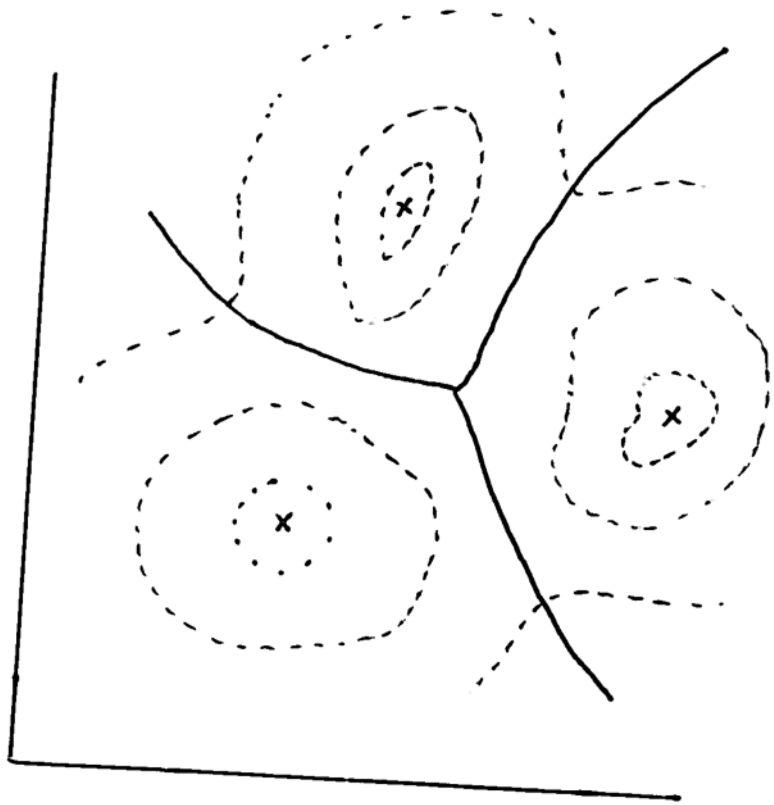
$$\arg\max_{k} \hat{\pi}_k \hat{g}_k(\boldsymbol{x})$$

  - $\hat{g}_k$ is an estimate of the class-conditional density

Why Density Estimation?

Clustering

- Clusters can be defined by the modes (i.e. peaks) of the density
- Given a point $\boldsymbol{x}$ climb the density until you reach a mode
- All $\boldsymbol{x}$ reaching the same mode form a cluster
- Referred to as *mode-based clustering*
    - Commonly implemented using the *mean-shift algorithm*



Why Density Estimation?

Novelty/anomaly Detection

- Goal: detect points that are significantly different from a training sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim f$

- Form an estimate $\hat{f}$ of f from the training data

- Check if a future observation $\boldsymbol{x}$ comes from the same distribution or not:

$$\begin{aligned} \hat{f}(\boldsymbol{x}) < \gamma \quad &\Rightarrow \quad \boldsymbol{x} \text{ is an anomaly} \\ \hat{f}(\boldsymbol{x}) > \gamma \quad &\Rightarrow \quad \boldsymbol{x} \text{ is not an anomaly} \end{aligned}$$

Kernel Density Estimation

- A *kernel density estimate* (KDE) has the form:

$$\hat{f}_h(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} k_h(\boldsymbol{X}_i - \boldsymbol{x})$$

  - $k_h(\boldsymbol{y})$ is called a *kernel*
  - $h > 0$ is a parameter called the *bandwidth*
- k_h has the form

$$k_h(\boldsymbol{y}) = h^{-d} k\left(\frac{\boldsymbol{y}}{h}\right)$$

- k is usually chosen to satisfy the following properties:
  1. $\int k(\boldsymbol{y}) d\boldsymbol{y} = 1$
  2. $k(\boldsymbol{y}) \geq 0,\ \forall \boldsymbol{y} \in \mathbb{R}^d$
  3. $k(\boldsymbol{y}) = \psi(||\boldsymbol{y}||)$ for some $\psi: [0, \infty) \to \mathbb{R}$
     - Property 3 makes k a "radial" kernel

Kernel Density Estimation

Kernel Examples

1. Gaussian kernel

$$k(\boldsymbol{y}) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}\|\boldsymbol{y}\|^2\right)$$

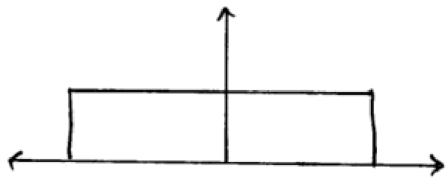
2. Uniform kernel

$$k(\boldsymbol{y}) = \frac{1}{C}\mathbf{1}_{\{\|\boldsymbol{y}\|\leq 1\}}$$

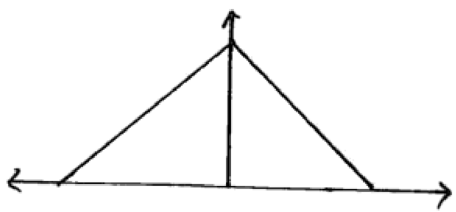
- C = volume of the unit sphere in $\mathbb{R}^d$

Kernel Density Estimation

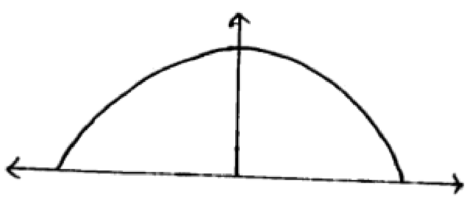
Kernel Examples for $d = 1$



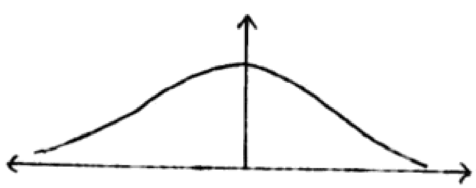
Uniform



Triangular



Epanichnikov (parabolic)



Cauchy

Kernel Density Estimation

Remarks

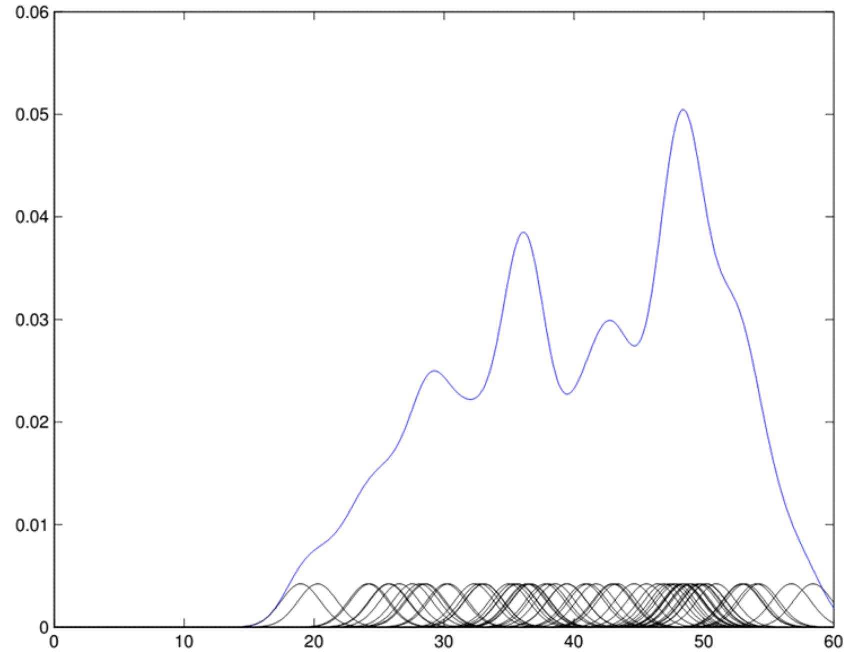
1. This notion of kernel is distinct from that of an inner product/positive definite kernel
2. The KDE is sometimes called the Parzen window or a Parzen estimate. It was originally proposed by Rosenblatt (1956) and Parzen (1962)
3. The KDE is clearly nonparametric
4. The KDE integrates to 1

Kernel Density Estimation

- Why does it work?

- KDE can be viewed as the superposition of shifted kernel functions
- The more X_i in a given region of space, the more these shifted kernels accumulate

KDE of midterm exam scores for an ML class at a different university (60 pts max)



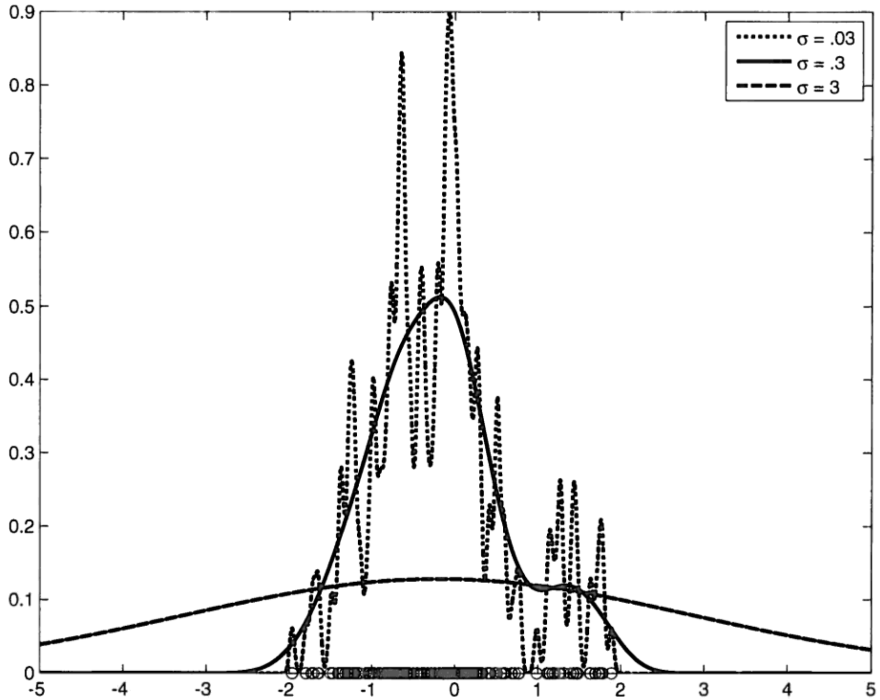
Group Exercise

1. Conceptually compare and contrast KDE with a histogram. What are the similarities and differences?
2. Discuss how you would implement a KDE on real data.

1. Both have a tuning parameter that determines the width (binwidth and the bandwidth). KDE doesn't have fixed bins. KDE can be smoother if a smooth kernel is selected.
2. First, determine the points where you want to estimate the density (i.e., the $\boldsymbol{x}$ values). For each $\boldsymbol{x}$ value and for each $\boldsymbol{X}_i$ point from the training data, calculate $k_h(\boldsymbol{X}_i - \boldsymbol{x})$. For each $\boldsymbol{x}$ value, take the sample mean of $k_h(\boldsymbol{X}_i - \boldsymbol{x})$ over the $\boldsymbol{X}_i$s.

Model Selection

- How do we know whether $\hat{f}_h$ is a "good" estimate of f?
  - Similarly, how do we select a "good" bandwidth h?
- Visually, we see that h has a major effect on the KDE (in the figure, $\sigma = h$)
- We need some measure of performance of our KDE in terms of the true density f



Mean Squared Error Analysis

- Estimation performance is typically measured in terms of *mean squared error* (MSE)
  - For a given (fixed) $\boldsymbol{x}$, the MSE of $\hat{f}_h$ is
$$\mathbb{E}\left[\left(\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right]$$
- If you expand out the MSE equation, it can be shown that the MSE is equal to the *estimation variance* plus the *squared bias*
  - Estimation Variance: $\mathbb{V}\left[\hat{f}_h(\boldsymbol{x})\right] \coloneqq \mathbb{E}\left[\left(\hat{f}_h(\boldsymbol{x})^2 - \mathbb{E}\left[\hat{f}_h(\boldsymbol{x})\right]^2\right)\right]$
  - Bias: $\mathbb{B}\left[\hat{f}_h(\boldsymbol{x})\right] \coloneqq \mathbb{E}\left[\hat{f}_h(\boldsymbol{x})\right] - f(\boldsymbol{x})$
  - $\text{MSE}\left(\hat{f}_h(\boldsymbol{x})\right) = \mathbb{V}\left[\hat{f}_h(\boldsymbol{x})\right] + \mathbb{B}\left[\hat{f}_h(\boldsymbol{x})\right]^2$
  - Let’s analyze the bias and estimation variance of the KDE

Bias Analysis

- Assume that $d = 1$ for now, the $\boldsymbol{X}_i$ are i.i.d., f is thrice differentiable, and k is symmetric
  - Last assumption $\Rightarrow \int_{-\infty}^{\infty} k(y) y^p dy = 0$ for p odd
- We need to find $\mathbb{E}\left[\hat{f}_h(x)\right]$:

$$
\begin{aligned}
\mathbb{E}\left[\hat{f}_h(x)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} k_h(\boldsymbol{X}_i - x)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[k_h(\boldsymbol{X}_i - x)] \\
&= \mathbb{E}[k_h(\boldsymbol{X}_i - x)] \\
&= \int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{z-x}{h}\right) f(z) dz
\end{aligned}
$$

Bias Analysis

- Change of variables $u = (z - x)/h$ to get

$$\mathbb{E}[\hat{f}_h(x)] = \int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{z-x}{h}\right) f(z) dz$$

$$= \int_{-\infty}^{\infty} k(u) f(x + hu) du\,.$$

- We can approximate this with a Taylor expansion of $f(x + hu)$ in the hu argument which is valid as $h \to 0$

$$f(x + hu) = f(x) + f^{(1)}(x)hu + \frac{1}{2} f^{(2)}(x) h^2 u^2 + o(h^2)$$

Bias Analysis

- Use the notation $\kappa_p(k) \coloneqq \int_{-\infty}^{\infty} k(y) y^p dy$
- Applying the Taylor expansion with the facts that $\kappa_0(k) = 1$ and $\kappa_1(k) = 0$ gives:

$$
\begin{aligned}
\mathbb{E}\left[\hat{f}_h(x)\right] &= \int_{-\infty}^{\infty} k(u) f(x + hu) du \\
&= \kappa_0(k) f(x) + \kappa_1(k) f^{(1)}(x) h + \frac{1}{2}\kappa_2(k) f^{(2)}(x) h^2 + o(h^2) \\
&= f(x) + \frac{1}{2}\kappa_2(k) f^{(2)}(x) h^2 + o(h^2)
\end{aligned}
$$

- Thus the bias is

$$\mathbb{B}\left[\hat{f}_h(x)\right] = \frac{1}{2}\kappa_2(k) f^{(2)}(x) h^2 + o(h^2)$$

- A similar expression holds when $d > 1$
  - Do a multivariate Taylor Series expansion

Estimation Variance Analysis

- *Remark*: the estimation variance is not the same as the variance of a random variable drawn from the estimated density
- It is the variance of the density estimator at a point
- Assume that $d = 1$ for now, the $\boldsymbol{X}_i$ are i.i.d., f is twice differentiable, and k is symmetric
- Since the KDE is an i.i.d. sum:

$$\begin{aligned} \mathbb{V}\left[\hat{f}_h(x)\right] &= \frac{1}{n}\mathbb{V}[k_h(\boldsymbol{X}_i - x)] \\ &= \frac{1}{n}\mathbb{E}[k_h(\boldsymbol{X}_i - x)^2] - \frac{1}{n}(\mathbb{E}[k_h(\boldsymbol{X}_i - x)])^2 \end{aligned}$$

Estimation Variance Analysis

$$\mathbb{V}\left[\hat{f}_h(x)\right] = \frac{1}{n}\mathbb{E}[k_h(\boldsymbol{X}_i - x)^2] - \frac{1}{n}(\mathbb{E}[k_h(\boldsymbol{X}_i - x)])^2$$

- From the bias analysis, we know that $\mathbb{E}[k_h(\boldsymbol{X}_i - x)] = f(x) + o(1)$
  - $\Rightarrow$ second term is $O\left(\frac{1}{n}\right)$
- For first term, make a change-of-variables and a first-order Taylor Expansion (similar to the bias analysis)

Estimation Variance Analysis

$$\begin{aligned}
\mathbb{E}[k_h(\boldsymbol{X}_i - x)^2] &= \frac{1}{h^2}\int_{-\infty}^{\infty} k\left(\frac{z-x}{h}\right)^2 f(z)dz \\
&= \frac{1}{h}\int_{-\infty}^{\infty} k(u)^2 f(x+hu)du \\
&= \frac{1}{h}\int_{-\infty}^{\infty} k(u)^2\big(f(x)+O(h)\big)du \\
&= \frac{1}{h} f(x)R(k) + O(1)
\end{aligned}$$

- $R(k) = \int_{-\infty}^{\infty} k(u)^2 du$ is the *roughness* of the kernel

- Combining the two terms gives

$$\mathbb{V}\left[\hat{f}_h(x)\right] = \frac{f(x)R(k)}{nh} + O\left(\frac{1}{n}\right)$$

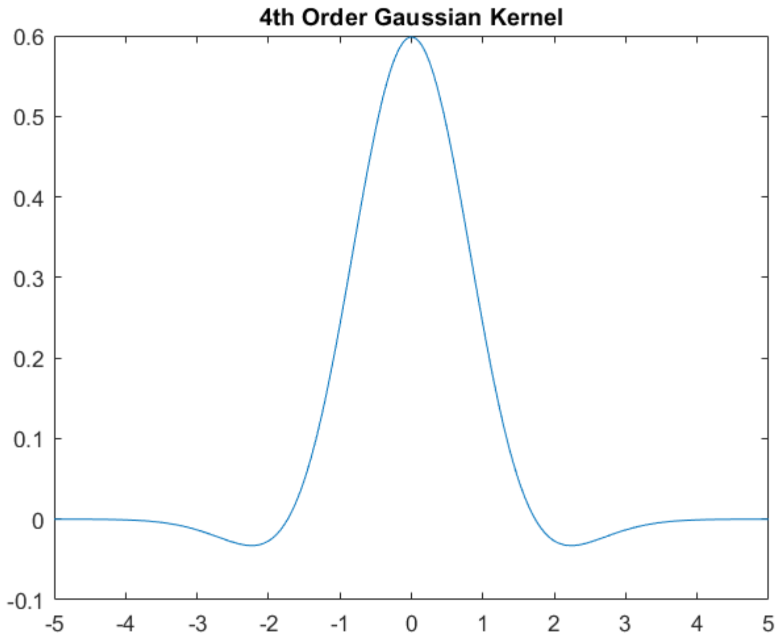
Multivariate case

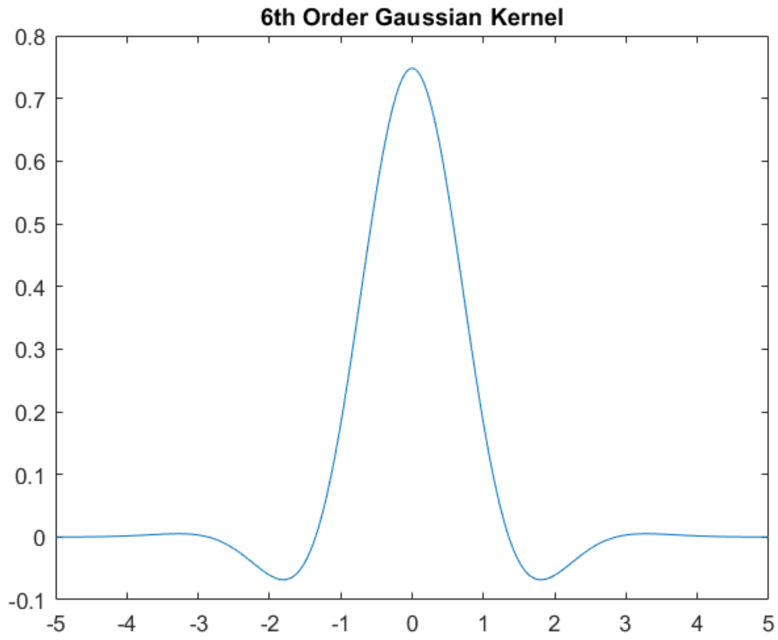
- Assume d is arbitrary, the $\boldsymbol{X}_i$ are i.i.d., the third partial derivatives of $f(\boldsymbol{x})$ exist, and K is a product of symmetric univariate kernels k
  - I.e. $K_h(\boldsymbol{u}) = k_h(u_1)k_h(u_2)\ldots k_h(u_d)$
- Bias: $\mathbb{B}\left[\hat{f}_h(x)\right] = \frac{1}{2}\kappa_2(k)C\left(f(\boldsymbol{x})\right)h^2 + o(h^2)$
  - $C\left(f(\boldsymbol{x})\right)$ is a function of the second partial derivatives of f
- Variance: $\mathbb{V}\left[\hat{f}_h(x)\right] = \frac{f(x)R(K)}{nh^d} + O\left(\frac{1}{n}\right)$

Group Exercise

1. Given the same assumptions as the previous slide (including arbitrary d), find the bandwidth h^* that minimizes the MSE of the KDE.
2. Plug in the optimal bandwidth h^* to find the optimal MSE rate. You can use Big O notation for this.
3. What additional assumptions could be imposed on the kernel k and the density f to improve the MSE rate?

Bias-canceling kernel examples

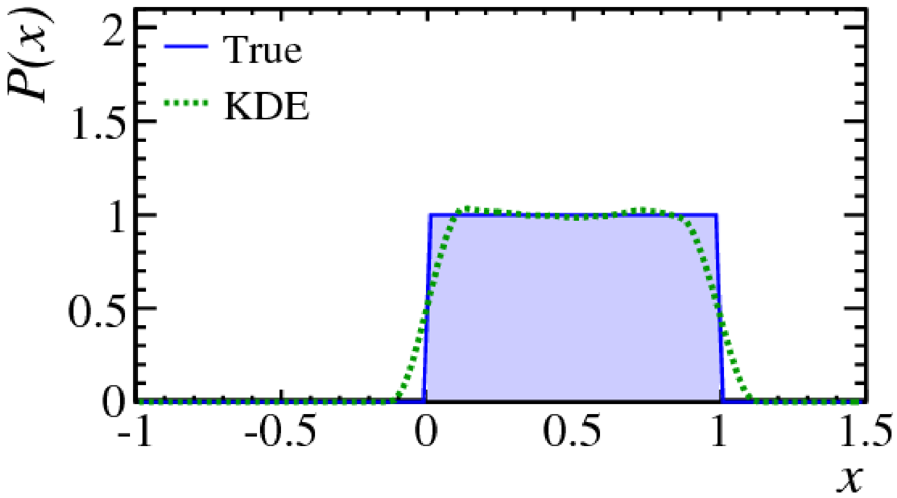


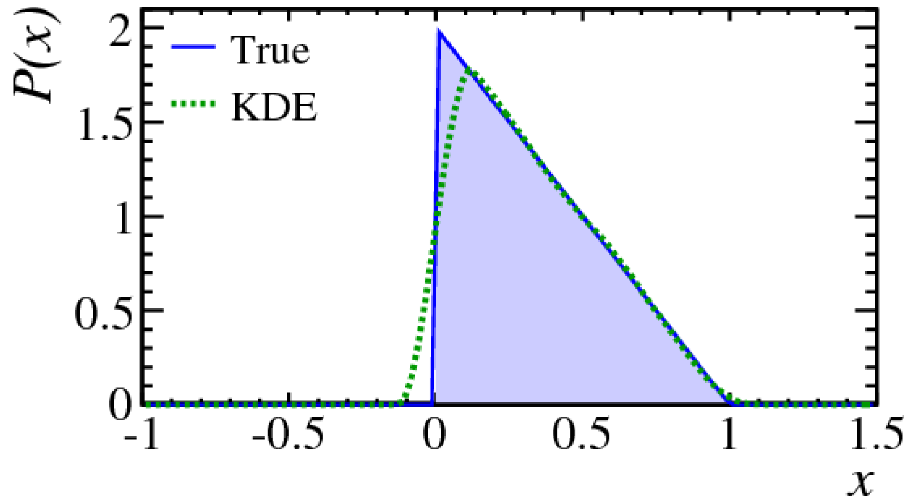


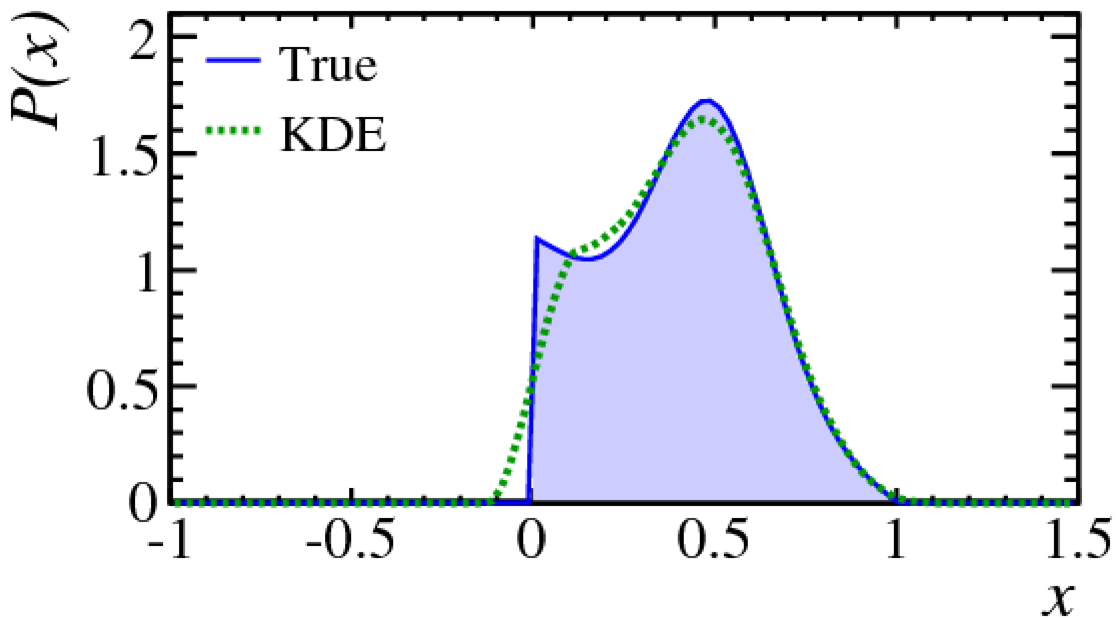
Asymptotic MSE

- **Remark**: to derive the previous results, we assumed that h is close to zero (for the Taylor series expansion to be valid)
- Thus the previously derived results give the *asymptotic MSE* and not the exact MSE
  - The results can still be useful in helping us understand the behavior of KDEs
- **Another remark**: we assumed that the support of the density is unbounded. In general, the bias of the KDE at the boundary of the support (e.g. the boundary of a uniform distribution) does not decay to zero.
  - To get the bias to decay, you typically have to do mirror kernel density estimation, which requires you to know where the boundary is

Bias at the Boundary Examples





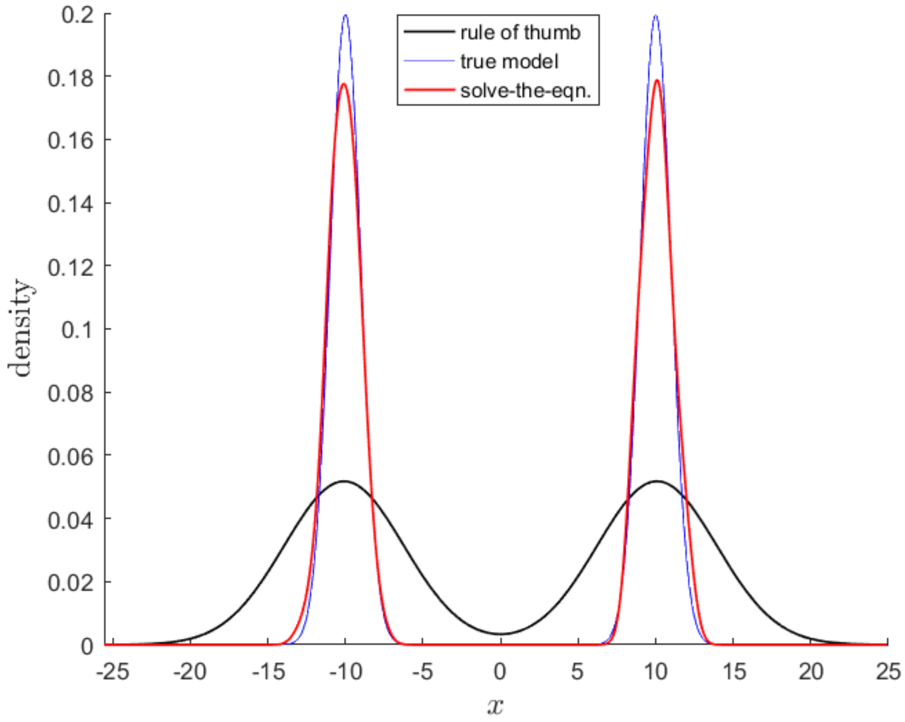


Back to Model Selection

- The optimal bandwidth that we derived depends on the density f and its derivatives… which are unknown.
- How do we choose the bandwidth?
  - A problem referred to as *model selection*
- One approach: Silverman's rule of thumb
  - Use Gaussian kernels and assume the density is Gaussian
  - Univariate case:
$$h^* = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}$$
  - $\hat{\sigma}$ is the sample standard deviation

Model Selection

- Unfortunately, Silverman's rule of thumb fails when the true density is not close to normal.
- The rule of thumb bandwidth tends to *oversmooth* the density estimate
- We need a different approach



KDE Wikipedia article

Model Selection

- MSE gave us a bandwidth value that depends on the thing (f) that we're trying to estimate
    - Thus MSE is hard to use in practice even though it's useful for theoretical analysis
- Choose a somewhat different performance measure: the *integrated squared error*
    - In real analysis, this is also referred to as the L^2 distance

$$
\begin{aligned}
ISE(h) &= \int \left(\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 d\boldsymbol{x} \\
&= \int \hat{f}_h(\boldsymbol{x})^2 d\boldsymbol{x} - 2\int \hat{f}_h(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} + \int f(\boldsymbol{x})^2 d\boldsymbol{x}
\end{aligned}
$$

Model Selection

$$ISE(h) = \int \hat{f}_h(\boldsymbol{x})^2 d\boldsymbol{x} - 2\int \hat{f}_h(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} + \int f(\boldsymbol{x})^2 d\boldsymbol{x}$$

- Can we minimize this with respect to h?
- Last term is independent of h so we can ignore it
- First term can be computed explicitly for many kernels
- **Example**: k_h is the Gaussian kernel

$$\begin{aligned}\int \hat{f}_h(\boldsymbol{x})^2 d\boldsymbol{x} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int k_h(\boldsymbol{x}-\boldsymbol{X}_i) k_h(\boldsymbol{x}-\boldsymbol{X}_j) d\boldsymbol{x} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k_{\sqrt{2}h}(\boldsymbol{X}_i-\boldsymbol{X}_j)\end{aligned}$$

  - Last step follows from the fact that convolving Gaussian densities amounts to adding Gaussian RVs

Model Selection

- Second term:

$$\int \hat{f}_h(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{X}\sim f}\left[\hat{f}_h(\boldsymbol{x})\right]$$

- **Idea**: estimate the expectation using the training data
- Could try $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_h(\boldsymbol{X}_i)$
  - This leads to overfitting ($h \to 0$)

Model Selection

- Try a leave-one-out estimator instead:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_h^{(-i)}(\boldsymbol{X}_i)$$

where

$$\hat{f}_h^{(-i)}(\boldsymbol{x}) = \frac{1}{n-1}\sum_{j\neq i} k_h(\boldsymbol{x} - \boldsymbol{X}_j)$$

- Put it all together (Gaussian kernel):

$$\hat{h} = \arg\min_h \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\sqrt{2}h}(\boldsymbol{X}_i - \boldsymbol{X}_j) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k_h(\boldsymbol{X}_i - \boldsymbol{X}_j)$$

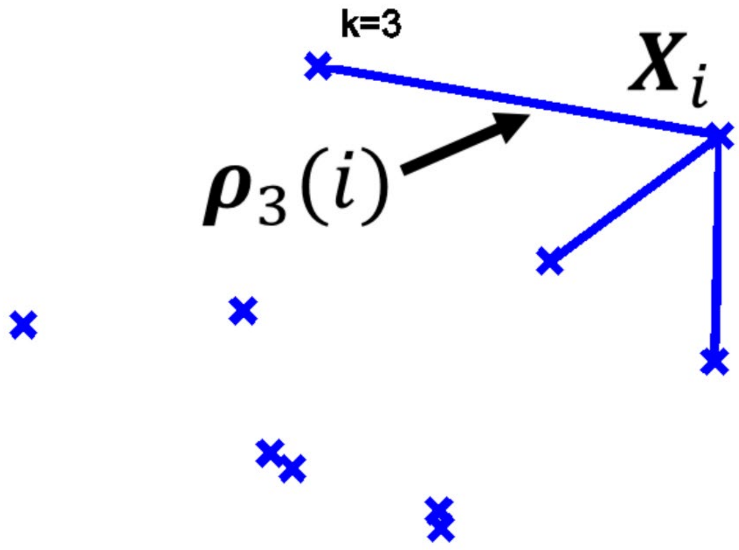
- This procedure is called least squares leave-one-out cross-validation (LS-LOOCV)
  - Could do k-fold CV instead of LOOCV
  - Other methods for bandwidth selection exist

K-nn Density Estimation

- Standard KDE assumes the bandwidth h is fixed
- This may not make sense for all densities
  - Some densities may have a mixture of wide and narrow peaks
- Using an *adaptive* bandwidth can help with this
- One way to do this is to use k-nearest neighbor distances

K-nn Density Estimation

- Random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim f$
  - f is an unknown pdf
- Let $\rho_k(\boldsymbol{x})$ be the distance from $\boldsymbol{x}$ to its kth nearest neighbor in the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$



K-nn Density Estimation

- Define the uniform kernel (sorry for change in notation but we can't use k)

$$w(\|\boldsymbol{u}\|) = c_d^{-1}\mathbf{1}_{\{\|\boldsymbol{u}\|]\leq 1\}}$$

  - $c_d =$ volume of unit ball in $\mathbb{R}^d$
- Use $\rho_k(\boldsymbol{x})$ as the bandwidth with this kernel:

$$\begin{aligned}\hat{f}_k(\boldsymbol{x}) &= \frac{1}{n\rho_k(\boldsymbol{x})^d}\sum_{i=1}^{n} c_d^{-1}\mathbf{1}_{\{\|\boldsymbol{x}-\boldsymbol{X}_i\|]\leq\rho_k(\boldsymbol{x})\}} \\ &= \frac{k}{n\rho_k(\boldsymbol{x})^d c_d}\end{aligned}$$

- Can have a smooth estimator by using a smoother kernel w
  - E.g. the Gaussian kernel

K-nn Density Estimation

- Analysis of the k-nn density estimator is harder because $\rho_k(\boldsymbol{x})$ is random
- It can be shown (see reference) that

$$\mathbb{B}\left[\hat{f}_k(\boldsymbol{x})\right] \approx \frac{\kappa_2(w)C\big(f(\boldsymbol{x})\big)}{2\big(c_d f(\boldsymbol{x})\big)^{\frac{2}{d}}}\left(\frac{k}{n}\right)^{\frac{2}{d}}$$

$$\mathbb{V}\left[\hat{f}_k(\boldsymbol{x})\right] \approx \frac{R(w)c_d f(\boldsymbol{x})^2}{k}$$

  - $R(w)$ is the roughness of w, $\kappa_2(w)$ is the 2^{nd} moment of w, and $C\big(f(x)\big)$ is a function of the second derivatives of f

K-nn Density estimation

- MSE:

$$MSE\left(\hat{f}_k(\boldsymbol{x})\right) = O\left(\left(\frac{k}{n}\right)^{\frac{4}{d}} + \frac{1}{k}\right)$$

- This is minimized by setting

$$k \propto n^{\frac{4}{d+4}}$$

- This gives an optimal MSE of

$$MSE\left(\hat{f}_k(\boldsymbol{x})\right) = \mathrm{O}\left(n^{-\frac{4}{d+4}}\right)$$

KDE vs k-nn density estimation

- k-nn approach is adaptive
  - Can give a more robust result in some applications
- However, the adaptive bandwidth from k-nn means the estimate may not integrate to 1
  - Need to rescale the estimate if integration to 1 is needed
  - However, many applications don't require this as the relative estimates are enough
    - E.g. anomaly detection
  - So it isn't always a disadvantage

KDE vs k-nn density estimation

- The two estimates also differ in the tails of the distribution
- Assume x is in the tail of f

$$\begin{aligned}
\mathbb{B}\left[\hat{f}_k(\boldsymbol{x})\right] &\approx \frac{C\left(f(\boldsymbol{x})\right)}{f(\boldsymbol{x})^{\frac{2}{d}}} \\
\mathbb{B}\left[\hat{f}_h(\boldsymbol{x})\right] &\approx C\left(f(\boldsymbol{x})\right) \\
\mathbb{V}\left[\hat{f}_k(\boldsymbol{x})\right] &\approx f(\boldsymbol{x})^2 \\
\mathbb{V}\left[\hat{f}_h(\boldsymbol{x})\right] &\approx f(\boldsymbol{x})
\end{aligned}$$

- In the tails, $f(\boldsymbol{x})$ is small
  - k-nn estimate will have larger bias but smaller variance than the KDE
  - Hard to say which is better in this case

Summary

- KDE and k-nn can be used to estimate probability densities
- Both approaches have been well-analyzed
- The density estimate can be used directly in many applications
  - Anomaly detection
  - Clustering
- The density estimate can also be used in other problems
  - Plug-in classifier
  - Estimating information measures (next time)

Further reading

- ESL Section 6.6
- Lecture Notes on Nonparametrics by Bruce E. Hansen: [Link]
- Nearest Neighbor Methods Lecture notes by Bruce Hansen: [Link]