Statistics and Machine Learning 1

Lecture 2C: Density Estimation

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 2

Estimating Population Density: Histograms

A histogram provides an estimate of the underlying probability density. Given a set of $q+1$ bin-boundaries, $\mathbf{b} = (b_0, b_1, , \ldots, b_q)$ with

$$b_0 < b_1 \cdots < b_q,$$

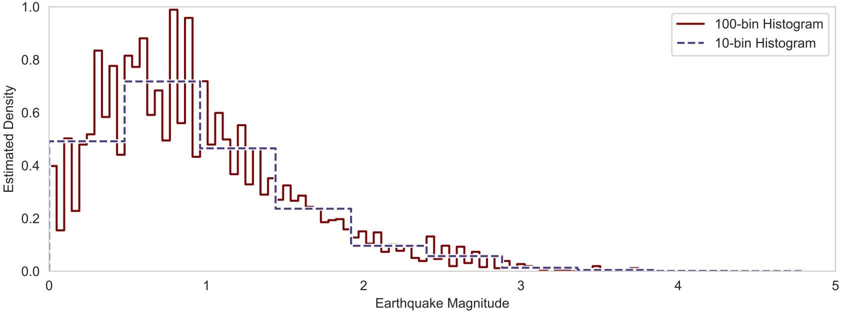
chosen so that $b_0 < \min(\mathbf{x})$ and $\max(\mathbf{x}) < b_q$, we can express the histogram-based density estimate in terms of the following sum

$$\begin{aligned} \hat{f}(s \mid \mathbf{b}) &= \frac{1}{n} \sum_{a=1}^{q} \frac{\mathbb{1}_{\{b_{a-1} \leq s < b_a\}}}{b_a - b_{a-1}} \left(\sum_{i=1}^{n} \mathbb{1}_{\{b_{a-1} \leq x_i < b_a\}} \right) \\ &= \sum_{a=1}^{q} \frac{\mathbb{1}_{\{b_{a-1} \leq s < b_a\}}}{b_a - b_{a-1}} \left\langle \mathbb{1}_{\{b_{1=1} \leq x < b_a\}} \right\rangle . \end{aligned} \tag{1}$$

which arranges that the bar above the bin into which s falls has an area equal to the fraction of the data points x_i that lie in that bin.

Histogram as a Density Estimate

The resulting estimate can be very sensitive to the length-q vector of bin-boundaries $\mathbf{b}$:



These histograms summarise data from 5392 earthquakes and so the bins in the 100-bin histogram contain, on average, only about 54 counts.

Estimating a Density with Kernels

Kernel density methods are an alternative to histograms, and yield an estimate:

$$
\begin{aligned}
\hat{f}(s \mid w) &= \frac{1}{n} \sum_{j=1}^{n} \frac{1}{w} K\left(\frac{s - x_j}{w}\right) \\
&= \left\langle \frac{1}{w} K\left(\frac{s - x}{w}\right) \right\rangle
\end{aligned}
\tag{2}
$$

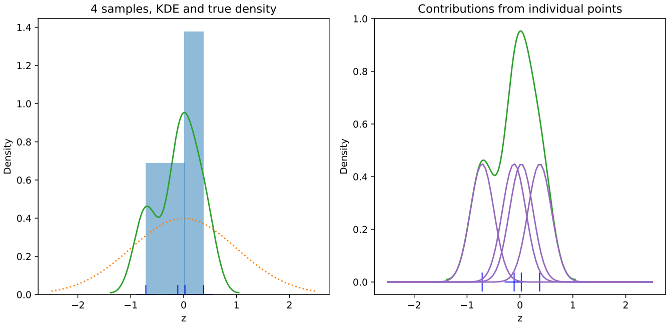
The main players in this formula are

$K(x)$: the *kernel*, typically some bump-shaped function such as a Gaussian or a parabolic bump. It should be normalised in the sense that

$$\int_{-\infty}^{\infty} K(x)\, dx = 1.$$

w : the *bandwidth*, which sets the width of the bumps.

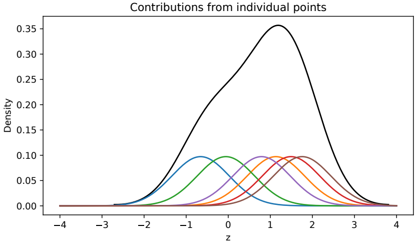
KDEs: the idea

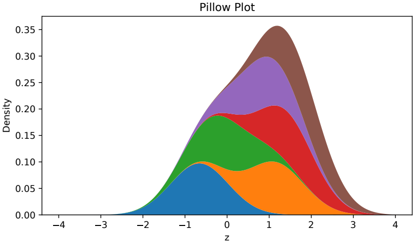


At left: 4 samples from this distribution (blue rug), the corresponding histogram estimator (blue bars) and the kernel density estimate (green curve).

At right: The *kernel density estimate* (KDE, green curve), samples (blue dashes) and their contributions to the KDE (purple bumps).

KDEs: stacking pillows

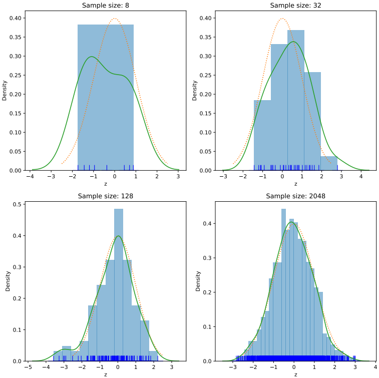




At left: The kernel density estimate (black curve) based on six samples from a standard normal distribution. Their contributions to the KDE are shown as variously coloured curves.

At right: A *pillow plot* showing how the coloured bumps shown at left add up to give the shape of the KDE.

Convergence of a KDE



The bandwidth w is a free parameter and there are various approaches to choosing its value. For parabolic and Gaussian kernels it is usually chosen so that

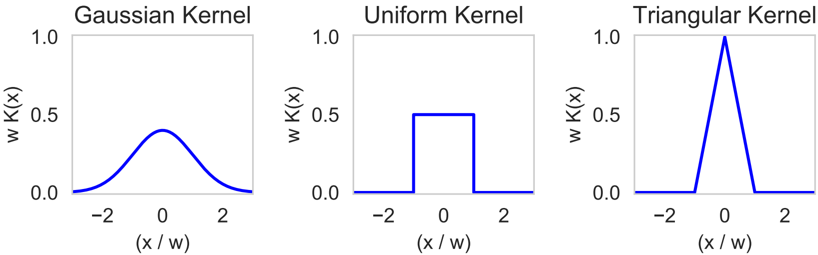
$$w \propto \frac{1}{N^{\frac{1}{5}}},$$

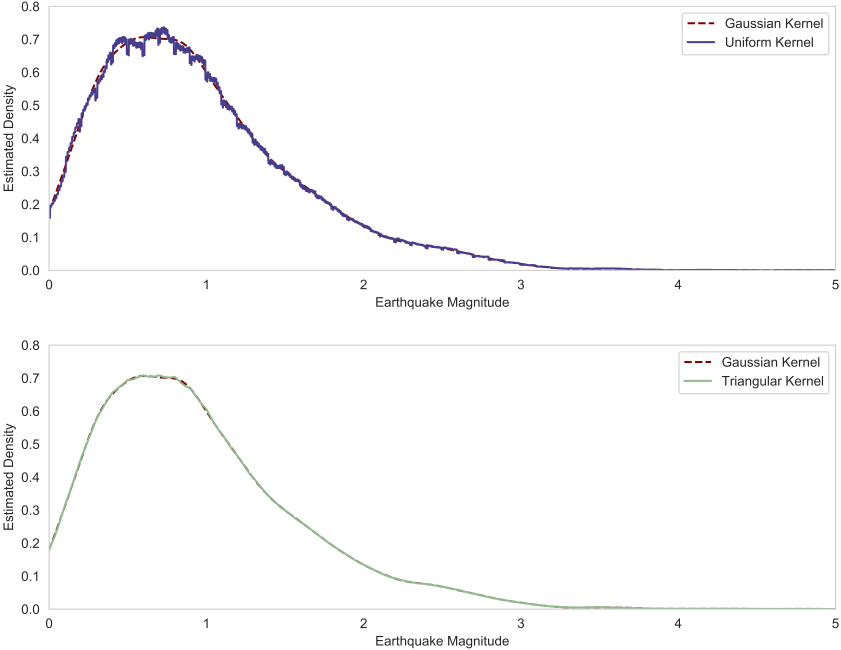
which, under mild assumptions on the smoothness of f, means that the KDE converges to the true distribution like $1/N^{4/5}$.

The plots at left above show a standard normal distribution (orange, dotted), variously-sized samples from that distribution (blue rug), a histogram-based density estimate (blue bars) and a KDE whose kernel is a Gaussian bump.

Commonly Used Kernels

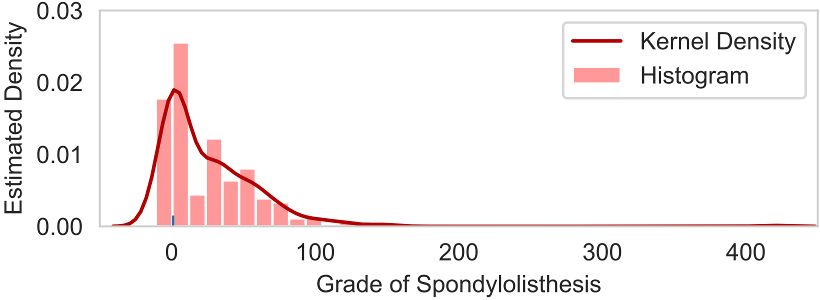
The Gaussian is the most popular choice, and default in many libraries, but well-motivated alternatives exist and often give different estimates for modes and other features:




0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Estimated Density
0
1
2
3
4
5
Earthquake Magnitude
Gaussian Kernel
Uniform Kernel
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Estimated Density
0
1
2
3
4
5
Earthquake Magnitude
Gaussian Kernel
Triangular Kernel

Spinal Dataset

- The UCI Machine Learning Repository is a great resource for datasets.
- We consider data from this repository collected by Dr. Henrique da Mota on values for six biomechanical features used to classify orthopaedic patients into 2 classes (normal or abnormal).
- The fifth column of this dataset consists of $n = 310$ observations of the Grade of Spondylolisthesis, which lies between $\min(x) = -11.06$ and $\max(x) = 414.54$, and is visualised below:

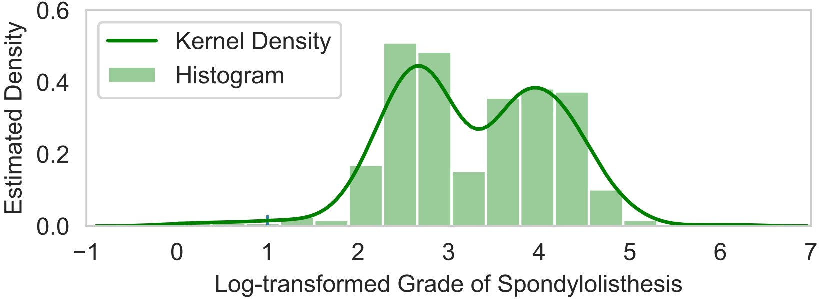


Multimodality?

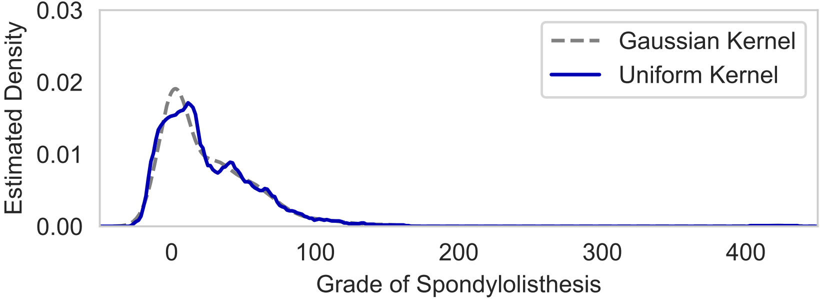
- So are there two modes in these data, as the histogram suggests?
- If we make a logarithmic transform,

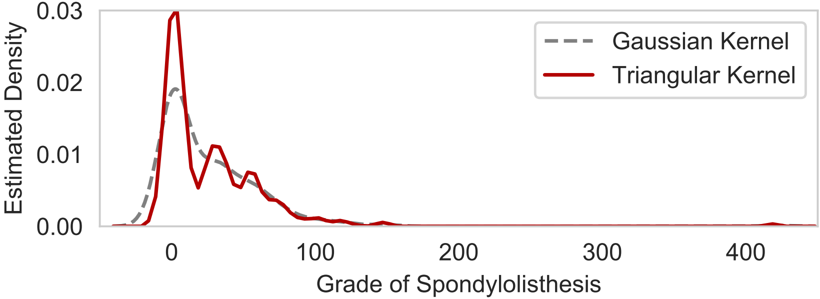
$$y_i = \ln(x_i - \min(x) + 1) \,, \tag{3}$$

then the presence of two modes looks clear:



- But different kernels ’yield” different answers—in general, there is no ‘right’ answer for this kind of *unsupervised learning* problem.


0.03
0.02
0.01
0.00
Estimated Density
0
100
200
300
400
Grade of Spondylolisthesis
Gaussian Kernel
Uniform Kernel


0.03
0.02
0.01
0.00
Estimated Density
0
100
200
300
400
Grade of Spondylolisthesis
Gaussian Kernel
Triangular Kernel

Some other considerations

- Often we should *weight* the datapoints for various reasons—*i.e.* assign a w_i to each datapoint such that

  $$\sum_{i=1}^{n} w_i = 1.$$

  Then most things go through as above but with the following definition of expectation:

  $$\langle f(x) \rangle = \sum_{i=1}^{n} w_i f(x_i). \tag{4}$$

- Typically some data will be *missing* (e.g. recorded as `NaN`, `-1`, or a blank in the data file) or *badly wrong* (for example a physically impossible value). If these occur uniformly at random, then they can just be taken out, but if not then care must be taken that the EDA should not be misleading (which is easier said than done!).

Further reading

- For a comparative discussion of histograms and KDEs, you might enjoy a `post` on the computational biologist Michael G. Lerner's blog, *Biophysics and Beer*.
- For a related introduction to KDEs in python, see this `post` on Jake VanderPlas's blog, *Pythonic Perambulations*.
- A more technical discussion appears in Section 6.6 of

  T. Hastie, R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, Springer-Verlag ISBN: 978-0-387-84858-7.
- I also found some lecture notes by the University of Wisconsin's Bruce Hansen `https://www.ssc.wisc.edu/ bhansen/718/NonParametrics1.pdf` helpful
- The standard theoretical reference about KDEs is still:

  B. W. Silverman (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC. ISBN: 978-0-4122-4620-3.