ANLY-601
Advanced Pattern Recognition

Spring 2018

L15 --- Nonparametric Density Models – Kernel Density Estimates

GEORGETOWN UNIVERSITY

Nonparametric Density Estimates

Parametric forms involve a chosen functional form of the density, and fitting parameters by estimation from a sample --- e.g. the Gaussian

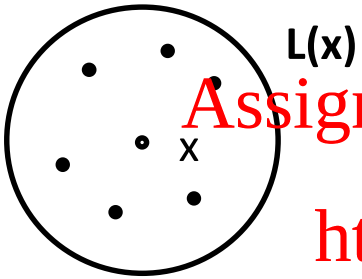
$$p(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp -\tfrac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)$$

Non-parametric methods do not use a chosen parametric functional form, but rather are unstructured. The histogram is an example. We'll look at

– Parzen windows or kernel estimate
– k-nearest neighbor estimate

Parzen

Consider a region L(x) about the point x (not necessarily a data point)



The region L(x) contains volume V, and probability mass

$$M_{L(x)} = \int_{L(x)} d^n x' \; p(x') \approx p(x) \; V$$

The approximation is more accurate as the region shrinks (to below the length scale over which p(x) varies appreciably)

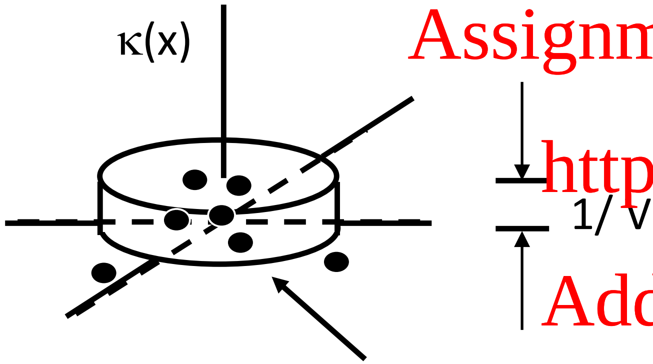
We can approximate the density at x by

$$\boxed{\hat{p}(x) \; V \approx \frac{k(x)}{N}}$$

where N is total number of data points, k(x) the number of points in L(x), and V the volume enclosed by L. This is the Parzen window estimate.

Kernel Estimates

The Parzen window can be constructed in a slightly different light. Consider data in 2-D. Let the function κ(x) have value 1/V throughout the region L(x) (of 2-D area *V*), and value zero outside



base area *V*
(data points in plane)

The Parzen estimate can be re-written

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \kappa(x - x_i)$$

where x_i, $i=1 \ldots N$ are the data points. The kernel $\kappa(x- x_i)$ takes value *1/V* for all data points x_i that fall within the base area *V* (centered on *x*), but zero outside. The summation is thus *k(x)/V*.

Kernel Estimates

Since the function $\kappa(y)$ is symmetric in y, we can put another interpretation on the kernel density estimate

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \kappa(x - x_i)$$

Place the center of a kernel at **each data point** x_i, sum up the result, and using that as a picture of the density.

This is how kernel estimates are usually interpreted. There's lots of possible kernels. They satisfy

$$\int \kappa(y)\ d^n y = 1$$

$$\kappa(-y) = \kappa(y)$$

Kernel Estimates

One possible kernel is the Dirac delta function -- an infinitely narrow, infinitely tall spike that

is symmetric $\delta(x) = \delta(-x)$

integrates to 1 $\int_R \delta(x-y)\, d^N x = \begin{cases} 1, & y \text{ in } R \\ 0, & \text{otherwise} \end{cases}$

has the sifting property $\int_R \delta(x-y)\, f(x)\, d^N x = \begin{cases} f(y), & y \text{ in } R \\ 0, & \text{otherwise} \end{cases}$

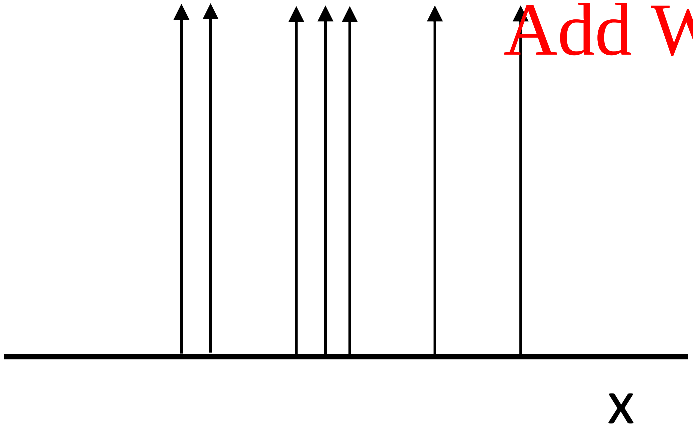
The corresponding density estimate is a set of spikes, one at each data point.

Kernel Estimates

The delta kernel density estimate

$$\hat{p}_s(x) = 1/N \sum_{i=1}^{N} \delta(x - x_i)$$

is a set of spikes - one at each data point.

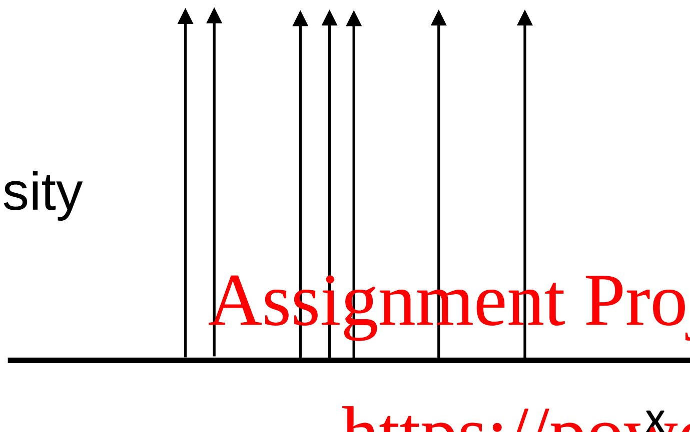


This needs to be smoothed to give a reasonable picture of the density p(x).

To smooth it, we convolve this spike density with a smoothing kernel function k(x).

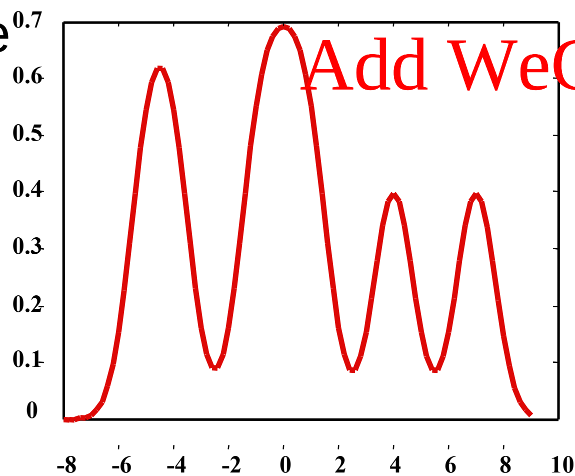
Kernel Estimates

Take the spike density



$$\hat{p}_S(x) \;=\; 1/N \sum_{i=1}^{N} \delta(x - x_i)$$

and convolve (smear) it with the kernel κ(x) to get the smoothed version



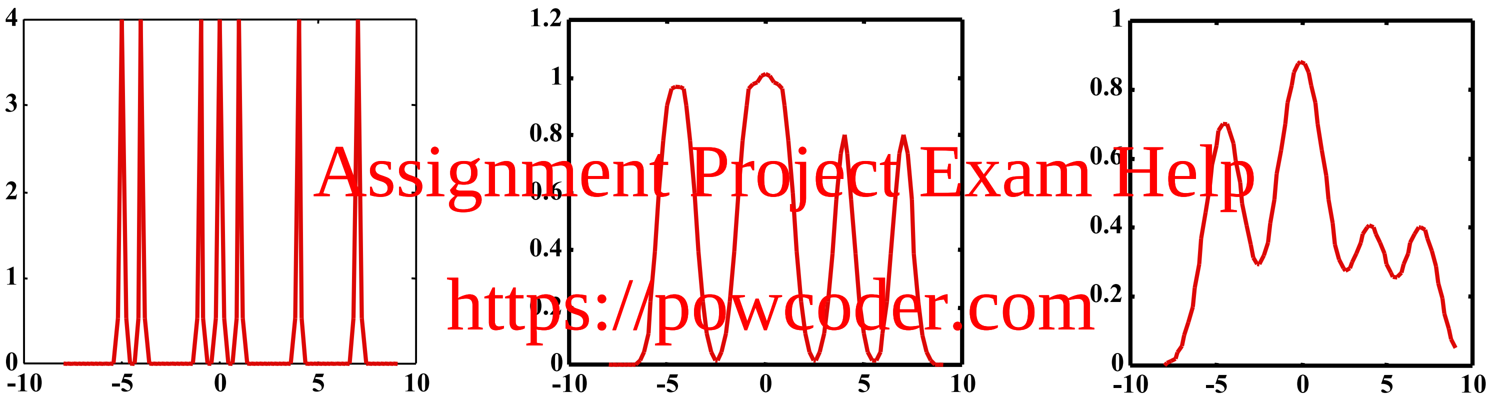
$$\hat{p}(x) \;=\; \kappa(x) * \hat{p}_S(x)$$

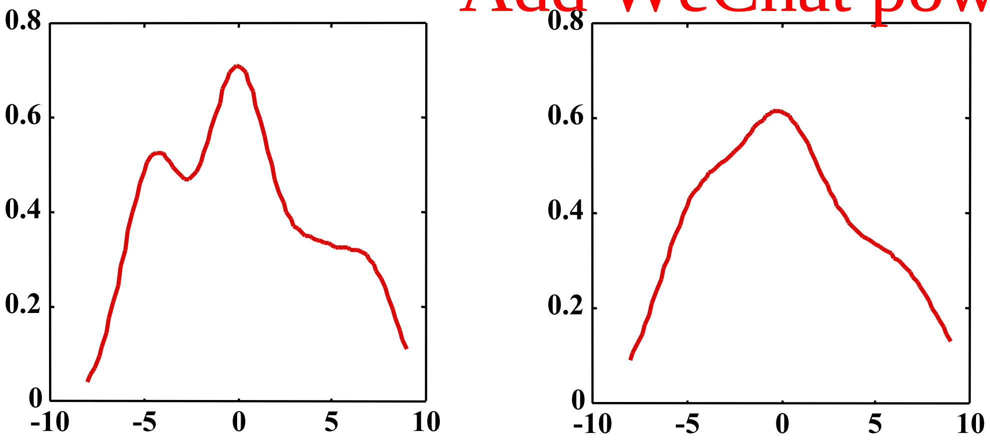
$$\equiv \int \kappa(x - y)\;\; 1/N \sum \;\delta(y - x_i)\;\; dy$$

$$= \;\; 1/N \sum_{i=1}^{N} \kappa(x - x_i)$$

Kernel Estimates

The width of the kernel function determines how much smoothing results - here's a progression from narrow to wide kernels:





If the kernel is narrower than the distance between the data points, the density estimate will be too bumpy. If the kernel is too wide, we wipe out detail in the density.

Family of Kernels

Here's a convenient family of kernels

$$\kappa(x) = \frac{m\,\Gamma(n/2)\,\Gamma^{n/2}\left(\frac{n+2}{2m}\right)}{(n\pi)^{n/2}\,\Gamma^{n/2+1}\left(\frac{n}{2m}\right)} \;\frac{1}{r^n\,|A|^{1/2}}\;\exp -\left[\frac{\Gamma\left(\frac{n+2}{2m}\right)}{n\,\Gamma\left(\frac{n}{2m}\right)}\; x^T \left(r^2 A\right)^{-1} x\right]^m$$

where

m -- determines shape (m=1 --> normal curve)

$r^2 A = \Sigma_\kappa$ is the covariance

r -- determines the spatial extent of the kernel

A -- matrix determines the directional variation of the kernel

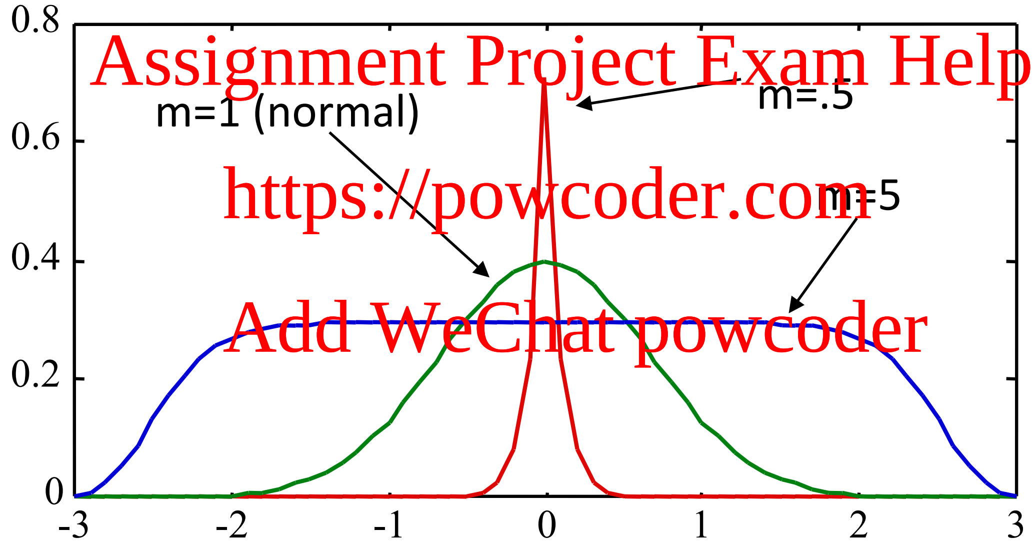
This family satisfies

$$\int \kappa(x)\; d^n x = 1$$

$$\Sigma_\kappa = \int x\,x^T\;\kappa(x)\; d^n x = r^2 A$$

Kernels

Here's what they look like (1-d) for different values of m:

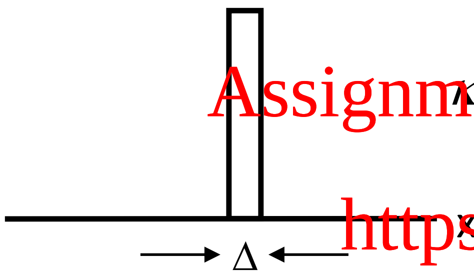


and as $m \to \infty$ the kernel becomes uniform

Histograms and Kernel Density Estimates

Histograms are a crude type of kernel density estimate obtained by

– Take a rectangular kernel

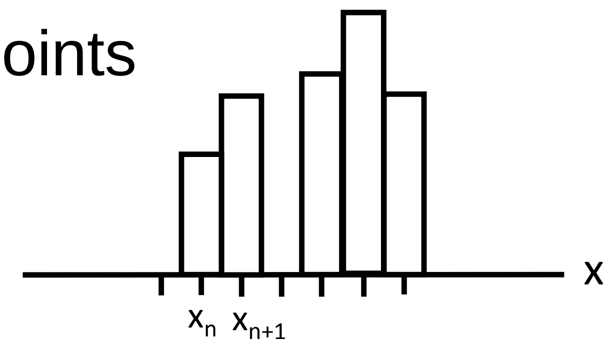


$$\kappa(x - x') = \begin{cases} 1/\Delta & if \quad |x - x'| < \Delta/2 \\ 0 & otherwise \end{cases}$$

– Form the kernel density estimate $\hat{p}(x) = 1/N \sum_{i=1}^{N} \kappa(x - x_i)$

– Sample the estimate at the discrete points

$$x_n = n\,\Delta, \quad n = ,\ldots, -2, -1, 0, 1, 2, 3, \ldots$$



Bias and Variance of Kernel Estimates

The kernel density estimate $$\hat{p}(x) = 1/N \sum_{i=1}^{N} \kappa(x - x_i)$$

is an estimator dependent on the sample data points x_i . Like all such statistical estimators we can ask about its bias and variance.

Start with the delta-function kernel estimate

$$\hat{p}_S(x) = 1/N \sum_{i=1}^{N} \delta(x - x_i)$$

and take its expectation over all sets of x_i

GEORGETOWN UNIVERSITY

Bias and Variance of Kernel Estimates

Start with the delta-function kernel estimate

$$\hat{p}_s(x) = 1/N \sum_{i=1}^{N} \delta(x - x_i)$$

and take its expectation over all sets of x_i

$$E_D[\ \hat{p}_s(x)\ ] = \int \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i) \quad p(x_1, x_2, ..., x_N)\ d^n x_1\ ...\ d^n x_N$$

$$= \int \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i) \quad p(x_1)\ p(x_2) ... p(x_N)\ d^n x_1\ ...\ d^n x_N$$

$$= \frac{1}{N}\sum_{i=1}^{N} p(x) = p(x)$$

so it’s <u>unbiased</u>!

Bias of Kernel Density Estimate

For a symmetric but otherwise arbitrary kernel, the expectation is

$$E_D\ [\ \hat{p}_\kappa(x)\ ] = \int \ \tfrac{1}{N}\sum_{i=1}^{N} \kappa(x-x_i) \quad p(x_1)\ \ p(x_2)\ \ldots\ p(x_N)\ \ d^n x_1 \ \ldots\ d^n x_N$$

$$= \int \int \ \tfrac{1}{N}\sum_{i=1}^{N} \kappa(x-y)\ \ \delta(y-x_i)\ \ p(x_1)\ \ p(x_2)\ \ldots\ p(x_N)\ \ d^n x_1 \ \ldots\ d^n x_N \quad d^n y$$

$$= \int \ \kappa(x-y) \quad p(y) \quad d^n y \quad \equiv \quad p * \kappa$$

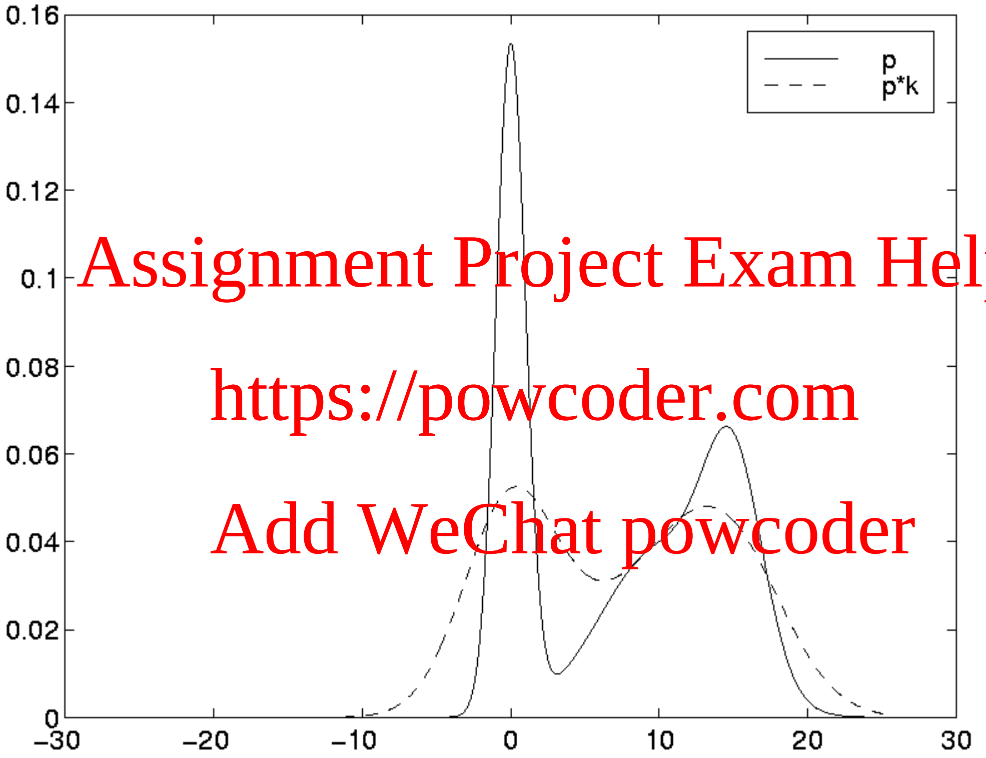
$$E_D[\hat{p}_\kappa(x)] = p * \kappa$$

and this is, in general, biased.

What does the bias look like?

Bias of Kernel Density Estimate

Convolution with the kernel smoothes the parent distribution p(x). Here's an example of convolving a density with a Gaussian kernel



Smoothing reduces the "contrast" in the curve; the smoothed density underestimates the true density where the latter is high, and overestimates the true density where the latter is low.

Bias of Kernel Density Estimates

- The <u>expected</u> kernel density estimate is smoother than the real density for all finite-width kernels.

- The extent of the smoothing, and hence of the bias, <u>increases</u> as the kernel width increases.

- Only delta function kernels give an unbiased estimate of p(x).

Variance of Kernel Estimate

The variance involves the second moment

$$E_D\left[\ \hat{p}^2(x)\ \right] = \int \ \frac{1}{N^2}\sum_{i=1}^{N} \kappa^2(x-y_i) \quad p(y_i)\ d^N y_i$$

$$+ \quad \frac{1}{N^2}\sum_{i\neq j} \int\int \kappa(x-y_i)\ \kappa(x-z_j)\ p(y_i)\ p(z_j)\ d^N y_i \quad d^N z_j$$

$$= \frac{1}{N}\,\kappa^2 * p \ + \left(\frac{N-1}{N}\right)\left(\kappa * p\right)^2$$

$$\mathrm{var}[\ \hat{p}(x)\ ] = \tfrac{1}{N}\left(\ \kappa^2 * p \ - \ (N-1)\left(\kappa * p\right)^2\right)$$

<u>Note that this decreases as 1/N</u>.

Bias and Variance

Intuitively, the bias decreases as the kernel gets narrower.

The variance is difficult to understand, apart from its decrease with increasing dataset size. Fukunaga (1) gives approximate forms for the bias and variance for the family of kernels on pp10-11 of these notes.

So we are able to control the bias and variance by adjusting the kernel width *r.*

One prescription is to maximize the likelihood of a validation set $\{y_i\}$ under a kernel model with kernel centers at the training dataset points $\{x_i\}$.

$$\hat{p}(\{y\}) = \prod_{i=1}^{Q} \hat{p}(y_i) = \prod_{i=1}^{Q} \frac{1}{N} \sum_{j=1}^{N} \kappa(y_i - x_j \;\; ; r)$$

Why not maximize the likelihood of $\{x_i\}$?

GEORGETOWN UNIVERSITY

GEORGETOWN UNIVERSITY