IIT KHARAGPUR

Bachelor's Thesis

Mechanical Engineering

Kernel Density Estimation

Author:
Anshit Singh

Supervisor:
Prof S.R.Khare

April 6, 2018

Department of Mechanical Engineering

Indian Institute of Technology Kharagpur

Certificate

This is to certify that the project entitled **" Kernel Density Estimation"** submitted by **Anshit Singh** [Rollno 14ME10009] , to Indian Institute of Technology, Kharagpur, for the award of the Bachelor of Technology, is a record of research work carried out by him under my supervision and guidance. The thesis, in my opinion, is worthy of consideration for the award of the degree of Bachelor of Technology in Engineering in accordance with the rules and regulations of the Institute.

Prof S.R.Khare
(Project Guide)

Prof. Anandroop Bhattacharya and Prof. Atul Jain
(Course Coordinator)

Date:6th April 2018

Abstract

When we try to estimate the density or the pattern behind an underlying data there are two ways by which we approach this problem.The first is try to model the data using standard distributions. Some standard distributions such as normal and a mixture of gaussians are used. In such a situation we usually have a fixed number of parameters such as the mean the variance and other parameters which if found can lead to the whole distribution. This involves a big assumption though which is that we assume that the data comes from that distribution.This assumption if proved incorrect can lead to very incorrect predictions.The other method is not assuming that the data comes from any fixed distribution and varying the number of parameters as per the data we have.This involves a lot of visual techniques such as plotting and counting the number of observable peaks.This on the contrary gives a more general representation. However as we move on to higher dimensions that is when we have more number of variables to incorporate plotting and counting peaks is not possible.We use other techniques such as histograms and kernels to estimate the underlying density .

1 Introduction

Parametric Density estimation is the use of an underlying assumption that the data is derived from a density which can be modelled by estimating a fixed number of parameters.Examples of this are assuming the data comes from a normal distribution or multiple of them as in the case of a mixture gaussian model.These models are applied in practice to many kinds of data. Parametric studies provide neat summaries and an easy representation of the data.When no simple parametric model solves our problem, we opt for fully"unparametric" methods that may be loosely collected under the heading of exploratory data analysis. Such analyses are highly graphical, but in a complex non-Normal setting, a graph may provide a more concise representation than a parametric model, because a parametric model of adequate complexity may involve hundreds of parameters. For example Kernel density estimation is applied often to the steel surface data (Bowyer, 1980; Silverman, 1986) .The data are measurements from an arbitrary origin of the actual height of a machined fiat surface at a grid of 15,000 points. The bandwidths were selected so that the values at the mode matched.So non parametric density estimation give a pretty good estimate of the irregularities of the steel surface post machining

DIFFFERENCES BETWEEN PARAMETRIC AND NON PARAMETRIC ESTIMATION

- There are some significant differences between parametric and nonparametric modeling. The focus on optimality in parametric modeling is not emphasised as much in the nonparametric world.
- There is no fixed superiority among algorithms in the nonparametric case but is dependent not only on the unknown density but also on the sample size.
- For example, the histogram might be proved to be an inadmissible estimator, but that theoretical fact should not be taken to suggest histograms should not be used.
- No nonparametric estimate is considered wrong; only different components of the solution are emphasized.
- The "curse of optimality" says that if the notion that optimality is all important is adopted, then the focus becomes matching the theoretical properties of an estimator to the assumed properties of the density function.
- Unlike optimal parametric estimates that are useful for many purposes, nonparametric estimates must be optimized for each application. The extra work is justified by the extra flexibility.
- This attitude reflects not sloppy thinking, but rather the imperfect relationship between the practical and theoretical aspects of our methods.

2 Literature Review

The book on Multivariate Density Estimation by David W Scott provides a good theoretical and practical insight into density estimation.It provides a good detailed analysis on what exactly defines a non parametric approach and how it can be used for the estimation of the function which represents the data.It introduces how and why histograms came across as the first non parametric estimation approach and how the modern methods such as Kernel Density Estimation developed. It provides a thorough treatment of the maths behind the estimation as well as touching upon the intuition.An important insight derived from the book is how the origin of the bin plays very little role in the estimation whereas the main governing factor is the size of the bins.

The paper Density Estimation for Statistics and Data Analysis by B.W Silverman also uses Kernel Density Estimation for a wide variety of data sets such as suicide data, transformed mast cell data etc and finds non parametric methods to be a very good estimate.

The book by Kuo-Lung Wu and Miin-ShenYang Mean shift-based clustering provides a very thorough and intuitive explanation of how kernel density estimation is used in clustering problems.

The paper Comaniciu, D. and Meer, P. (1997)Robust Analysis of FeatureSpaces: Color Image Segmentation is very insightful showing how mean shift algorithm can be used in the segmentation of images into different regions which can be further used to understand the features of an image.

3 Problem Description

Given a dataset to try and find the density estimate which fits the underlying data best.Often parametric methods are not able to represent all kinds of data.So our focus is to try to understand the nonparametric methods of data estimation to find the best estimate.

4 Mathematical Formulation

4.1 Histogram Representation

Histogram form of representation even though a nonparametric form has 2 main parameters

- Location of Origin
- Bin Width

Often the origin is chosen as t0=0.

4.2 Stuges rule for Bin Width Approximation

He took the binomial distribution and used it as a model of an optimally constructed histogram. I_{th} bin had C_i^{k-i} as the bin count and it was seen that with increase in k the distribution assumed a normal shape.

The Stuges rule for number of bins was

$$k = 1 + log_2 n \tag{1}$$

However if the data is skewed or kurtotic then additional bins may be required.

4.3 Frequency and Density Histograms

Frequency histograms is built using blocks of size 1 stacked and width h.The integral is clearly equal to nh. The density histogram uses building blocks of height 1/nh so that the integral under the histogram is 1.

4.4 Histogram representation

Let v_k be the count of the number of observations of the k_{th} bin. Then histogram is defined as

$$\hat{f}(x) = \frac{v_k}{nh} = \frac{1}{nh} \sum_{i=1}^{n} I_{[t_k, t_{k+1}]}(x_i) \tag{2}$$

It is also recognized that the bin counts are random variables

$$v_k \sim \ B(n, p_k) \, where \, p_k = \int_{B_k} f(t)dt \tag{3}$$

Finding the MSE .ie mean and variance

$$Var \ \hat{f}(x) = \frac{Var \ v_k}{nh^2} = \frac{(p_k)(1-p_k)}{nh^2} \tag{4}$$

$$Bias \ \hat{f}(x) = E\hat{f}(x) - f(x) = \frac{Ev_k}{nh} - f(x) = \frac{p_k}{h} = f(x) \tag{5}$$

By mean value theorem and further simplification we have

$$MSE \ \hat{f}(x) \leq \frac{f(\epsilon_k)}{nh} + \gamma^2 h^2 \tag{6}$$

So h is the smoothing parameter This is very intuitive. If we have the bin size as small then the variance will be high as the probability of a new sample falling in the same bin will be low and hence the variation around the mean estimate will be high. Also the bias will be low as if the bin size is small the estimate will almost be a constant so the difference between the mean of the estimate and the actual distribution will be less. And vice versa for a big h.

4.5 Rate of convergence of an estimate

This is the rate at which the MSE reaches 0 as n tends to infinity i.e as our sample size becomes larger and larger and includes all possible values i.e becomes very general.

PARAMETRIC-The optimal rate of parametric estimate convergence is O(n-1) and is not attained by any nonparametic estimate.

4.6 Global L2 Histogram error

MISE is calculated by summing the MSE over each bin and then summing over all bins. Calculating

$$ISB = \frac{h^2}{12} \int_{-\infty}^{\infty} {f'_x}^2 dx + o(h^2) \tag{7}$$

The main term in this is the Asymptotic integrated square bias. So $ISB = AISB + O(h_2)$

$$AISB = \frac{h^2}{12} \int_{-\infty}^{\infty} {f'_x}^2 dx = \frac{h^2}{12} R(f') \tag{8}$$

Where R() is a measure of roughness

Likewise we have the AIV the main term of IV and the total sum i.e MISE and its main term AMISE

$$\begin{aligned} AMISE(h) &= \frac{1}{nh} + \frac{h^2}{12} R(f') \\ h &= [6/R(f')]^{1/3}\ n^{-1/3} \\ AMISE &= (3/4)^{2/3} R(f')^{1/3} n^{-2/3} \end{aligned} \tag{9}$$

Where h is the point where AMISE is minimum and AMISE is the minimum AMISE
So the optimal error AMISE decreases at a rate of $O(n^{-2/3})$ far from the desirable rate

NORMAL DENSITY REFERENCE RULE

Scott proposed using the normal density as a reference and then changing accordingly if the data is skewed or kurtotic.

$$Normal\ bin\ width\ reference\ rule: \quad \hat{h} = 3.5\hat{\sigma}n^{-1/3} \tag{10}$$

If the data was skewed or kurtotic it needed to be multiplied by a skewness or a kurtosis factor.

INFLUENCE OF BIN EDGE LOCATION

Bin edge doesn't affect the MISE significantly.

4.7 Optimally Adaptive Bin Sizes

A fixed size bin width appears rough in the tails so we seek varying bin sizes But these need to be properly constructed. Adaptive histograms constructed in an ad-hoc fashion often give poorer results than fixed bin width size histograms.

4.8 Oversmoothed Bin Widths

There exists an upper bound on the bin width or a lower bound on the number of bins

$$Number\ of\ bins = \frac{b-a}{h \geq \frac{b-a}{h_{OS}} = \sqrt[3]{2n}} (11)$$

Where a,b are the limits of an interval where f1 is linearly mapped and hos is the oversmoothed bin width After further simplification and solving an optimization problem we get

$$h^{\leq}(\frac{6}{nRf'(2)})^{1/3} = (\frac{686\sigma^3}{5\sqrt{7}n})^{1/3} = 3.729\sigma n^{-1/3} = h_{OS} \tag{12}$$

BIASED AND UNBIASED CROSS VALIDATION
The UCV function is noisier on an average but its minimizer is correct.THE BCV is biased towards larger bin widths

L1 ERROR
The primary difficulty with L1 error is that it cannot be broken down into a bias variance tradeoff

4.9 Kernel Density Estimators

The basic kernel estimator may be written as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i) \tag{13}$$

Virtually all non parameteric estimators are asymptotically kernel density estimators.The motivation for this comes from various ideas such as the limiting case of the averaged shifted histograms,numerical analysis and finite differences etc.

THEORETICAL PROPERTIES

MISE ANALYSIS

Calculating the MISE from bias and variance we get
Bias

$$\begin{aligned} E\ K_h(x,X) &= \int \frac{1}{h}K(\frac{x-t}{h})f(t)dt \\ &= \int\ K(w)f(x-hw)dw \\ &= f(x)\ \int K(w)dw - hf^{'}(x)\int wK(w)dw + \frac{1}{h^2}\int w^2K(w) \end{aligned} \tag{14}$$

To make the expectation of $\hat{f(x)}$ equal to f(x) of the order of h^2 the kernel K should satisfy

$$\int K(w)dw = 1 \quad \int wK(w) = 0 \quad and \int w^2K(w) = \sigma^2 > 0 \tag{15}$$

Then the bias becomes

$$Bias(x) = \frac{1}{2}{\sigma_k}^2h^2f"(x) + O(h^4) \Rightarrow ISB = \frac{1}{4}{\sigma_k}^4h^4f"(x) + O(h^6) \tag{16}$$

Variance is

$$Var\ K_h(x,X) = E\ [\frac{1}{h}K(\frac{x-X}{h})]^2 - [E\ \frac{1}{h}K(\frac{x-X}{h})]^{\ 2} \tag{17}$$

So for a nonnegative kernel estimator we have AMISE and the minimum AMISE (AMISE) and the optimal bandwidth h as

$$\begin{aligned} AMISE(h) &= \frac{R(k)}{nh} + \frac{\sigma_k^4h^4}{4}R(f") \\ h &= [R(k)/\sigma_k^4R(f")]^{1/5}\ n^{-1/5} \\ AMISE &= \ 5/4[\sigma_kR(K)]^{4/5}R(f")^{1/5}n^{-4/5} \end{aligned} \tag{18}$$

From the requirements for kernel we can think of taking the kernel as a probability density if only non negative kernels are to be taken.

ESTIMATION OF DERIVATIVES

The derivatives of the density function are often required when looking for modes and bumps.Derivatives of the kernel estimate behave well if the kernel is differentiable and the the bandwidths are large enough. Taking the estimator of the r_{th} derivative of the function as the r_{th} derivative of

the kernel estimate we have

$$\hat{f}^r(x) = \frac{d^r}{dx^r}\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) = \frac{1}{nh^{r+1}}\sum_{i=1}^{n} K^r(\frac{x-x_i}{h}) \tag{19}$$

Calculating the AMISE for this estimate and calculating the optimal AMISE and the optimal h value we have

$$\begin{aligned} AMISE(h) &= \frac{R(K^(r))}{nh^{2r+1}} + \frac{\sigma_k^4 h^4}{4} R(f^{(r+2)}) \\ h &= [(2r+1)\ R(K^(r))/\sigma_k^4\ R(f^{(r+2)})]^{1/(2r+5)}\ n^{-1/(2r+5)} \\ AMISE &= \frac{2r+5}{4} R(K^{(r)})^{\frac{4}{2r+5}} [\sigma_K^4 R(f^(r+2))/(2r+1)]^{\frac{2r+1}{2r+5}} n^{\frac{-4}{2r+5}} \end{aligned} \tag{20}$$

So while the order of the bias term remains $O(h^4)$, each additional derivative introduces 2 extra powers in the variance.

CHOICE OF KERNEL

Even though the quality of the estimate is widely recognised to be primarily governed by the bandwidth of the kernel estimator ,the choice of kernel analysis is still done to see what all inferences we get and what is the best choice of kernel.

HIGHER ORDER KERNELS

The normal kernel density estimates use weighted averages so they tend to understimate the peaks and overestimate the valleys.To further reduce the contribution of bias to the MISE the idea of higher order kernels was considered

OPTIMAL KERNELS

We find the best kernel among various kernels available by minimising its contribution to the AMISE.
Since

$$AMISE = [\sigma_k R(k)]^{4/5} \tag{21}$$

We do the minimisation as

$$min_k \ \ R(k) \ \ s/t \ \ {\sigma_k}^2 = \sigma^2 \tag{22}$$

The solution is a scaled version of the so called Epanechnikov kernel

$$K_2(t) = \frac{3}{4}(1-t^2)I_{[-1,1]}t \tag{23}$$

It is interesting to note that the optimal kernel has a finite support and is non differentiable at the boundaries. Since AMISE is also proportional to $n^{-4/5}$// We have

$$\frac{\sigma_k R(k)}{\sigma_{K_2} R(K_2)} = \frac{\sigma_k R(k)}{3/5\sqrt{5}} \tag{24}$$

Other kernels require the above mentioned number times as much data to achieve the same AMISE as the Epanechnikov kernel.
However the optimal kernel shows only a modest improvement in the AMISE so other kernels can be chosen over this giving more importance to other factors such as as computational efficiency, differentiabililty etc
Often the normal kernel is chosen over others for its computational efficiency.
Here are some common kernels and their relative efficiencies

5 Applications of Kernel Density Estimation

5.1 Estimation of the Gradient

Nonparametric density gradient estimation is a very important use of kernel density estimation which has lead to many algorithms involving clustering.This is especially useful in higher dimensions wherein the clusters are not visible to the naked eye via plots. Conditions on the kernel functions are derived to guarantee asymptotic unbiasedness, consistency, and uniform consistenby of the estimates. The results are generalized to obtain a simple mean-shift estimate that can be extended in a k-nearestneighbor approach. Applications of gradient estimation to pattern recognition are presented using clustering and intrinsic dimensionality problems, with the ultimate goal of providing further understanding of these problems in terms of density gradients.

5.2 Intuition

Nonparametric estimation of probability density functions is based on the concept that the value of a density function at a continuity point can be estimated using the sample observations that fall within a small region around that point.

In most pattern recognition problems, very little if any information is available as to the true probability density function or even as to its form.

Due to this lack of knowledge about the density, we have to rely on nonparametric techniques to obtain density gradient estimates. A straightforward approach for estimating a density gradient would be to first obtain a differentiable nonparametric estimate of the probability density function and then take its gradient Similarly, the gradient of a probability density function can be estimated using the sample observations within a small region.

5.3 Proposed Gradient Density Estimate

Let X_1,X_2,...X_n be a set of N independent and identically distributed n-dimensional random vectors

$$\hat{f_n}(x) = \frac{1}{Nh^n} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{25}$$

where k(X) is a scalar function satisfying

$$\begin{aligned} \int K(w)dw &= 1 \\ \int \|K(w)\| &< \infty \\ sup \ \|K(w)\| &< \infty \end{aligned} \tag{26}$$

and where $\|.\|$ is the ordinary Euclidean norm

Motivated by this general class of nonparametric density estimates, we use the differentiable kernel function and then estimate the density gradient as the gradient of the density estimate . This gives the density gradient estimate

$$\hat{\nabla}_x f_n(x) = \frac{1}{Nh^n} \sum_{i=1}^{n} \nabla_x K(\frac{x - x_i}{h}) \tag{27}$$

where

$$\nabla K(w) = (\frac{\partial K(w)}{\partial w_1}, \frac{\partial K(w)}{\partial w_2} ... \frac{\partial K(w)}{\partial w_n})^T \tag{28}$$

and ∇ is the usual gradient operator with respect to X_1,X_2,...X_n

Equation is the general form of our density gradient estimates. As in density estimation, this is a kernel class of estimates, and various kernel functions k(Y) may be used. Conditions on the kernel functions and h(N) will now be derived to guaranteed asymptotic unbiasedness,consistency, and uniform consistency of the estimate

5.4 Asymptotic Unbiasedness

The basic result needed for the proof of asymptotic unbiasedness.

$$\int \|K(w)\| < \infty \tag{29}$$

then the sequence of functions $f_n(X)$ defined by

$$\hat{f_n}(x) = \frac{1}{Nh^n} \int K(\frac{x - x_i}{h}) f(x - y) dy \tag{30}$$

converges to every point of continuity of f(X) to

$$\lim_{x \to \infty} f_n(x) = f(x) \int f(y) dy \tag{31}$$

5.5 Uniform Consistency

The gradient estimate is uniformly consistent if The conditions to be satisfied are listed as follows

$$\begin{aligned} \lim_{x \to \infty} h(N) &= 0 \\ \lim_{x \to \infty} Nh^{2n+2}(N) &= \infty \end{aligned} \tag{32}$$

Also the fact that the gradient of f must be continuous.

5.6 Mean Shift Gradient Estimates

The Gaussian kernel function is perhaps the best known differentiable multivariate kernel function satisfying the conditions for asymptotic unbiasedness, consistency, and uniform consistency of the density gradient estimate. The Gaussian probability density kernel function with zero mean and identity covariance matrix is

$$(x) = \frac{1}{2\pi^{n/2}} exp(\frac{-1}{2} X^T X) \tag{33}$$

Taking the gradient of and substituting it to get the gradient estimate

$$\hat{\nabla}_x f_n(x) = \frac{k}{N * P(x)} \frac{n+2}{h^2} \sum (\frac{x_i - x}{k}) \tag{34}$$

where

$$p(x) = \frac{h^n \pi^{n/2}}{\Gamma(n + 2/2)} \tag{35}$$

The above provides us with an excellent interpretation of the gradient estimation process.The last term in it is the sample mean shift

$$M(x) = \sum (\frac{x_i - x}{k}) \tag{36}$$

summed over the volume of the region

$$S_h(X) = Y : (Y - X)^T (Y - X) < h^2 \tag{37}$$

and k is the number of observations falling within $S_h(X)$ and, therefore, the number in the sum.

Interpretations

Clearly, if the gradient or slope is zero, corresponding to a uniform density over the region $S_h(X)$, the average mean shift would be zero due to the symmetry of the observations near X. However, with a nonzero density gradient pointing in the direction of most rapid increase of the probability density function, on the average more observations should fall along its direction than elsewhere in $S_h(X)$.

Correspondingly, the average mean shift should point in that direction and have a length proportional to the magnitude of the gradient

5.7 Mean Shift Normalized Gradient Estimate

Examining the proportionality constant in (34), we see that it contains a term identical to the probability density estimate using a uniform kernel function over the region

$$\hat{f}_n(x) = \frac{k}{N * P(x)} \tag{38}$$

By taking this to the left side of (34) and using the properties of the function In y, 'we see that the mean shift can be used as an estimate of the normalized gradient

$$\frac{\nabla_x f_n(x)}{f(x)} = \nabla_x \ln P(x) \tag{39}$$

This mean-shift estimate of the normalized gradient has a pleasingly simple and easily calculated expression . It can also be given the same intuitive interpretation that was just given in the previous section for the gradient estimate

5.8 Gradient Clustering Algorithm

One method of clustering a set of observations into different classes would be to assign each observation to the nearest mode along the direction of the gradient at the observation points. To accomplish this, one could move each observation a small step in the direction of the gradient and iteratively repeat the process on the transformed observations until tight clusters result near the modes.
Another approach would be to shift each observation by some amount proportional to the gradient at the observation point. This transformation approach is the one we will investigate in this section since it is intuitively appealing and will be shown to have good physical motivation behind its application.
Letting

$$X_j^0 = X_j \quad j = 1, 2, ...N \tag{40}$$

we will transform each observation recursively according to the clustering algorithm

$$X_j^{i+1} = X_j^i + a\nabla_x \ln p(X_j^i) \tag{41}$$

where a is an appropriately chosen positive constant to guarantee convergence.This is the n-dimensional analog of the linear iteration technique for stepping into the roots of the equation

$$\nabla_x p(X) = 0 \tag{42}$$

and equivalently the modes of the mixture density p(X).

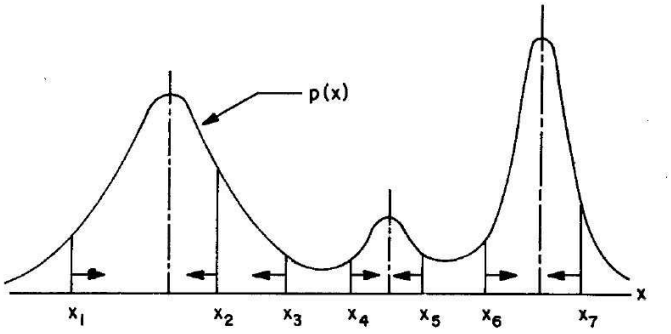

Figure 1: Gradient mode clustering

5.9 K-Means vs Mean Shift

Meanshift looks very similar to K-Means, they both move the point closer to the cluster centroids. We may wonder how is this different from K-Means?

- K-Means is faster in terms of runtime complexity
- The key difference is that Meanshift does not require the user to specify the number of clusters. In some cases, it is not straightforward to guess the right number of clusters to use. In K-Means, the output may end up having too few clusters or too many clusters to be useful. At the cost of larger time complexity, Meanshift determines the number of clusters suitable to the dataset provided.
- Another commonly cited difference is that K-Means can only learn circle or ellipsoidal clusters. However, this is not true. The reason that Meanshift can learn arbitrary shapes is because the features are mapped to another higher dimensional feature space through the kernel.
- K-means is very sensitive to initializations, while Mean shift is sensitive to the selection of bandwidth h

6 Methodology

6.1 Data Visualization

Visualization is an important component of nonparametric data analysis. Data visualization is the focus on methods ranging from simple.scatterplots to interactive displays. Function visualization is a significant component of nonparametric function estimation,and is used extensively in the nonparametric approach to understand what the data looks like.

6.2 General Flow/Approach

There is a natural flow among the parametric, exploratory, and nonparametric procedures that represents a rational approach to statistical data analysis.

- We begin with the aim to obtain an overview of the data
- If a probabilistic structure is present, estimate that structure nonparametrically and explore it graphically and using plots.
- If a linear model appears adequate, we adopt a fully parametric approach.
- We try to extract important information, finally reducing the dimension of the solution to a handful of interesting parameters
- Some people work in the reverse order, progressing to exploratory methodology as a diagnostic tool for evaluating the adequacy of a parametric model fit

6.3 Nonparametric Approach

Eventually after having explored the data visually and graphically and taking into account whatever insights we have so far achieved we try to fit kernel density estimates of various kinds and see which fits best.

7 Results and Discussion

SIMULATION 1

Self generated /simulated data which consists of a mixture of 2 gaussians.The first gaussian has 1000 points with mean 0 and variance 1. The second gaussian has 1500 points with mean 3 and variance 1.Although the gaussian kernel could be thought of somewhat appropriate but here we try different kernels.

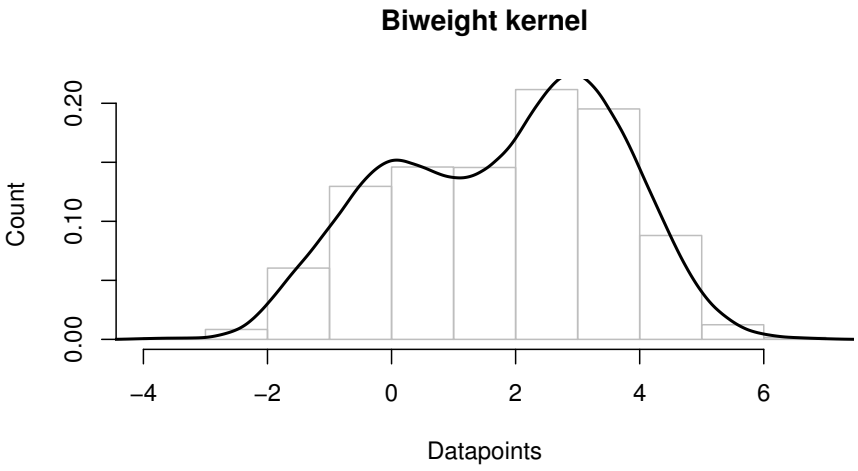

Figure 2: The biweight kernel

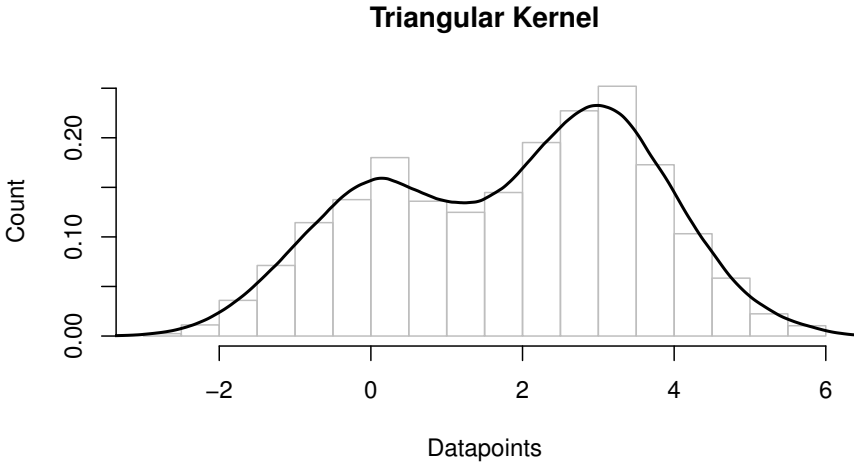

Figure 3: Triangular kernel also gives a good approximation to the original distribution

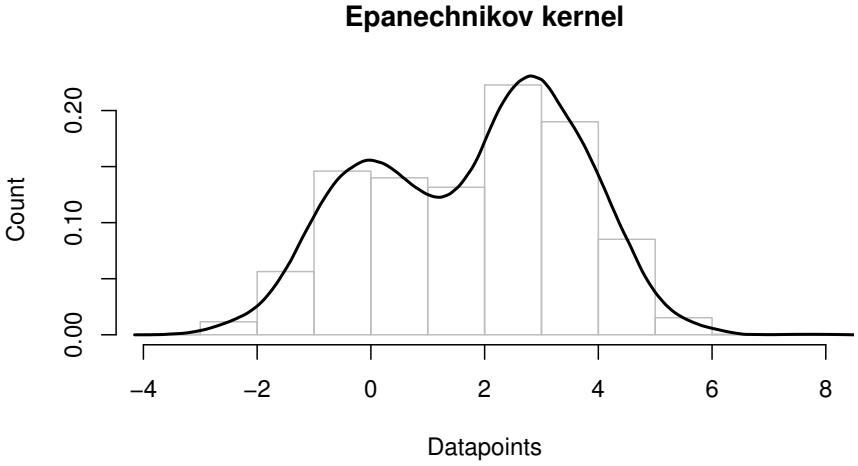

Figure 4: The epanechnikov fits the best in most cases it has been found out

SIMULATION 2

MEAN SHIFT WITH EXAMPLE

Given a set of datapoints, the algorithm iteratively assign each datapoint

towards the closest cluster centroid. The direction to the closest cluster centroid is determined by where most of the points nearby are at. So each iteration each data point will move closer to where the most points are at, which is or will lead to the cluster center. When the algorithm stops, each point is assigned to a cluster.

Unlike the popular K-Means algorithm, meanshift does not require specifying the number of clusters in advance. The number of clusters is determined by the algorithm with respect to the data.

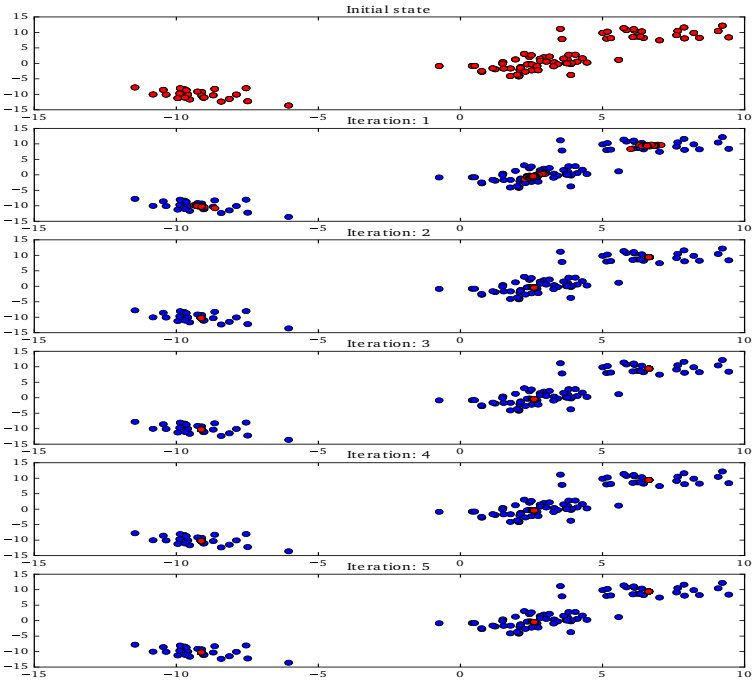

Figure 5: Diagram that shows what happens step-by-step in Meanshift

Iteration:

Initial state.

The red and blue datapoints overlap completely in the first iteration before the Meanshift algorithm starts.

- All the red datapoints move closer to clusters. Looks like there will be 4 clusters.
- End of iteration 1. All the red datapoints move closer to clusters. Looks like there will be 4 clusters.
- End of iteration 2. The clusters of upper right and lower left seems to have reached convergence just using two iterations. The center and lower right clusters looks like they are merging, since the two centroids are very close.
- End of iteration 3. No change in the upper right and lower left centroids. The other two centroids' have pulled each other together as the datapoints affect each clusters. This is a signature of Meanshift, the number of clusters are not pre-determined.
- End of iteration 4. All the clusters should have converged.
- End of iteration 5. All the clusters indeed have no movement. The algorithm stops here since no change is detected for all red datapoints.

Meanshift found 3 clusters here,as is evident. The original data is actually generated from 4 clusters of data, but Meanshift thinks 3 can represent the set of data better, and it's not too bad.
The code for the same is in the bibliography.

SIMULATION 3

IMAGE SEGMENTATION

Image Processing

Meanshift is used as an image segmentation algorithm.The idea is that similar colors are grouped to use the same color. Here is what Meanshift can do for us:

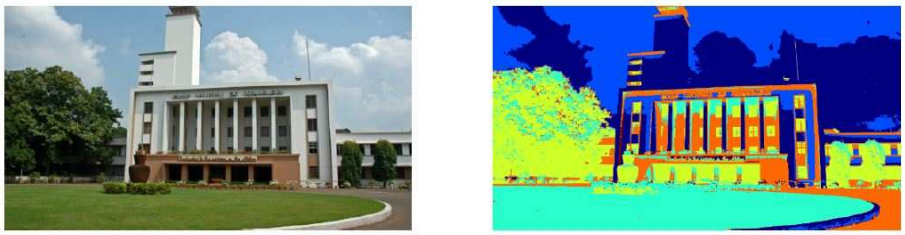

Figure 6: Institute main building image segmentation

As we can clearly see there are four clusters or groups and the colors representing those are orange,green ,light blue and dark blue.The building , the golc,the trees and the back of the building plus roads are the clustered segments of the image.

Figure 7: Nehru Museum photo segmented

As we can clearly see there are three clusters or segments and the colors representing those are green ,light blue and dark blue.The sky,the building and the back part of the building are the different clustered segments.

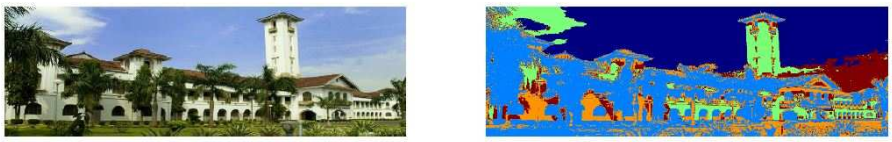

Figure 8: Nehru Museum side photo segmented

As we can clearly see there are three clusters or segments and the colors representing those are green ,light blue and dark blue.The sky, the building and the trees are the different clustered segments in the image.

The code for the same is in the bibliography.

This segmentation method has widespread applications such as in object tracking algorithm.

- Object tracking is an important task in computer vision
- Mean shift is recently used in tracking object (visual tracking)
- It requires a data of video frame to use the mean shift algorithm in visual (video) tracking
- The idea is that you need to identify the target to track, and build a color histogram of this target, and then keep on sliding the tracking window to the closest match (the cluster center) to keep up with the target to track.

8 Conclusion

As is clear from the plots and the various simulations the non parametric estimates give a good estimate of the underlying data. And as we have seen that these give a more general representation as compared to the parametric estimates. This is very helpful as there are many datasets where the parametric methods fail to give a good estimate. It has exceptional clustering applications using the estimation of the gradient of the kernel density estimator and using mean shift algorithm for the same. The results it gives is comparable and at times better than k-means clustering.We saw how this has immense applications in field such as computer vision wherein the first step of any big feature such as object detection/recognition is segmentation of the image.The mean shift algorithm implemented on various images and visualized as well using the iterative approach performs exceedingly well and has led to many computer vision applications to be based on this. These are just a few of the fields of the application of kernel density estimation.This has applications in almost any field where clustering [and even regression] is involved and it performs comparable to the other parametric approaches.And more importantly this is a more general approach as compared to the parametric approaches and is hence more widely applicable.

References

[1] Kopka Helmut, W. Daly Patrick, *Guide to LaTeX*, 4^{th} Edition. Available at `http://www.amazon.com`.

[2] Graetzer George, *Math Into LaTeX*, Birkhauser Boston; 3 edition (June 22, 2000).

[3] David W.Scott,Multivariate Density Estimation-Visualization,Theory and Practice,2nd Edition (March 2015) Available at `http://www.amazon.com`

[4] B. W. Silverman, Density Estimation for Statistics and Data Analysis,Monographs on Statistics and Applied Probability, Chapman and Hall (1986);

[5] Periklis Andritsos. Data Clustering Techniques `University of Toronto. Department of Computer Science`

[6] Comaniciu, D. and Meer, P. (1997)Robust Analysis of Feature Spaces: Color Image Segmentation. Available at `IEEE Conference on Computer Vision and Pattern Recognition. CVPR 1997, pages 750{755.`

[7] Kuo-Lung Wu and Miin-ShenYang. Mean shift-based clustering Available at `http://www.amazon.com`

[8] My codes Available at `http://https://github.com/Anshitsingh/BTP`