

# Kernel Density Estimation for the Eigenvalues of Variance Covariance Matrix of FFT Scaling of DNA Sequences: An Empirical Study of Some Organisms

Salah H. Abid<sup>1,\*</sup>, Jinan H. Farhood<sup>2</sup>

<sup>1</sup>Mathematics Department, Education College, University of Al-Mustansiriyah, Baghdad, Iraq

<sup>2</sup>Mathematics Department, Education College, University of Babylon, Babylon, Iraq

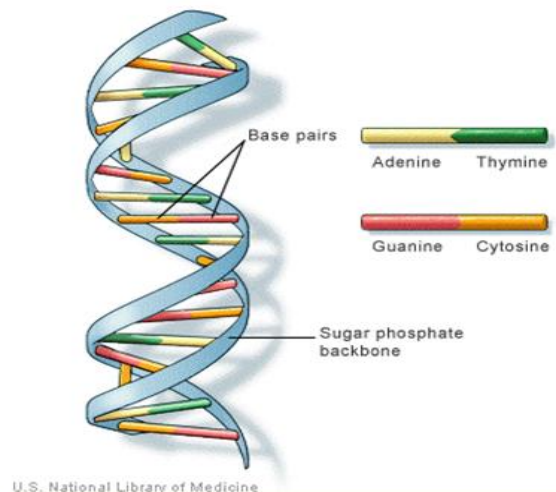
**Abstract** Many studies discussed different numerical representations of DNA sequences. In this paper, we discussed the kernel density estimation for the first, second, third and fourth eigenvalues of variance covariance matrix of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper. We computed an empirical values for the kernel density estimation for data series according to the following Kernels, Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine. To determine the valuable of our work, it should be noted that it is the first time that the variance covariance matrix eigenvalues of (FFT) for numerical values representation of DNA sequences, is used in an analysis like this and related analyzes.

**Keywords** FFT scaling, DNA, Kernel Density Estimation, Bandwidth, Eigenvalues

## 1. Introduction

In the process of developing the technology, many possible interesting adaptations became apparent: One of the most interesting directions was the use of the technology in the analysis of long DNA sequences. A benefit of the techniques was that it combined rigorous statistical analysis with modern computer power to quickly search for diagnostic patterns within long DNA sequences. Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar, and a phosphate group. There are four different bases that can be grouped by size, the pyrimidines, thymine (T) and cytosine (C), and the purines, adenine (A) and guanine (G). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the 5' carbon of one sugar linked to the 3' carbon of the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inward. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand; refer to Fig. 1. Thus, a strand of DNA can

be represented as a sequence  $\{X_t; t = 1, 2, \dots, n\}$  of letters, termed base pairs (bp), from the finite alphabet  $\{A, C, G, T\}$ . The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA (Polovinkina et al. (2016)).



**Figure 1.** The general structure of DNA and its bases

A common problem in analyzing long DNA sequence data is in identifying CDS that are dispersed throughout the sequence and separated by regions of noncoding (which

\* Corresponding author:

abidsalah@uomustansiriyah.edu.iq (Salah H. Abid)

Published online at <http://journal.sapub.org/ajms>

Copyright © 2019 The Author(s). Published by Scientific & Academic Publishing

This work is licensed under the Creative Commons Attribution International

License (CC BY). <http://creativecommons.org/licenses/by/4.0/>

makes up most of the DNA). Another problem of interest that we will address here is that of matching two DNA sequences, say  $X_{1t}$  and  $X_{2t}$ . The background behind the problem is discussed in detail in the study by Waterman and Vingron (1994). For example, every new DNA or protein sequence is compared with one or more sequence databases to find similar or homologous sequences that have already been studied, and there are numerous examples of important discoveries resulting from these database searches.

One naive approach for exploring the nature of a DNA sequence is to assign numerical values (or scales) to the nucleotides and then proceed with standard time series methods. It is clear, however, that the analysis will depend on the particular assignment of numerical values. Consider the artificial sequence ACGTACGTACGT. . . Then, setting  $A = G = 0$  and  $C = T = 1$ , yields the numerical sequence 0101010101. . ., or one cycle every two base pairs (i.e., a frequency of oscillation of  $\omega = 1/2$  Cycle/bp, or a period of oscillation of length  $1/\omega = 2$  bp=cycle). Another interesting scaling is  $A = 1$ ,  $C = 2$ ,  $G = 3$ , and  $T = 4$ , which results in the sequence 123412341234. . ., or one cycle every four bp ( $\omega = 1/4$ ). In this example, both scalings of the nucleotides are interesting and bring out different properties of the sequence. It is clear, then, that one does not want to focus on only one scaling. Instead, the focus should be on finding all possible scalings that bring out interesting features of the data. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a DNA sequence of virtually any length in a quick and automated fashion. In addition, the technique can determine whether a sequence is merely a random assignment of letters (Polovinkina et al. (2016)).

Fourier analysis has been applied successfully in DNA analysis; McLachlan and Stewart (1976) and Eisenberg et al. (1994) studied the periodicity in proteins using Fourier analysis.

Stoffer et al. (1993) proposed the spectral envelope as a general technique for analyzing categorical-valued time series in the frequency domain. The basic technique is similar to the methods established by Tavar'e and Giddings (1989) and Viari et al. (1990), however, there are some differences. The main difference is that the spectral envelope methodology is developed in a statistical setting to allow the investigator to distinguish between significant results and those results that can be attributed to chance.

The article authored by Marhon and Kremer 2011, partitions the identification of protein-coding regions into four discrete steps. Based on this partitioning, digital signal processing DSP techniques can be easily described and compared based on their unique implementations of the processing steps. They compared the approaches, and discussed strengths and weaknesses of each in the context of different applications. Their work provides an accessible introduction and comparative review of DSP methods for the identification of protein-coding regions. Additionally, by breaking down the approaches into four steps, they

suggested new combinations that may be worthy of future studies. A new methodology for the analysis of DNA/RNA and protein sequences is presented by Bajic in 2000. It is based on a combined application of spectral analysis and artificial neural networks for extraction of common spectral characterization of a group of sequences that have the same or similar biological functions. The method does not rely on homology comparison and provides a novel insight into the inherent structural features of a functional group of biological sequences. The nature of the method allows possible applications to a number of relevant problems such as recognition of membership of a particular sequence to a specific functional group or localization of an unknown sequence of a specific functional group within a longer sequence. The results are of general nature and represent an attempt to introduce a new methodology to the field of biocomputing. Fourier transform infrared (FTIR) spectroscopy has been considered by Han et al. in 2018 as a powerful tool for analysing the characteristics of DNA sequence. This work investigated the key factors in FTIR spectroscopic analysis of DNA and explored the influence of FTIR acquisition parameters, including FTIR sampling techniques, pretreatment temperature, and sample concentration, on calf thymus DNA. The results showed that the FTIR sampling techniques had a significant influence on the spectral characteristics, spectral quality, and sampling efficiency. Ruiz et al. 2018 proposed a novel approach for performing cluster analysis of DNA sequences that is based on the use of Genomic signal processing GSP methods and the K-means algorithm. We also propose a visualization method that facilitates the easy inspection and analysis of the results and possible hidden behaviors. Our results support the feasibility of employing the proposed method to find and easily visualize interesting features of sets of DNA data. A novel clustering method is proposed by Hoang et al. in 2015 to classify genes and genomes. For a given DNA sequence, a binary indicator sequence of each nucleotide is constructed, and Discrete Fourier Transform is applied on these four sequences to attain respective power spectra. Mathematical moments are built from these spectra, and multidimensional vectors of real numbers are constructed from these moments. Cluster analysis is then performed in order to determine the evolutionary relationship between DNA sequences. The novelty of this method is that sequences with different lengths can be compared easily via the use of power spectra and moments. Experimental results on various datasets show that the proposed method provides an efficient tool to classify genes and genomes. It not only gives comparable results but also is remarkably faster than other multiple sequence alignment and alignment-free methods. One challenge of GSP is how to minimize the error of detection of the protein coding region in a specified DNA sequence with a minimum processing time. Since the type of numerical representation of a DNA sequence extremely affects the prediction accuracy and precision, by this study Mabrouk in 2017 aimed to compare different DNA numerical

representations by measuring the sensitivity, specificity, correlation coefficient (CC) and the processing time for the protein coding region detection. The proposed technique based on digital filters was used to read-out the period 3 components and to eliminate the unwanted noise from DNA sequence. This method applied to 20 human genes demonstrated that the maximum accuracy and minimum processing time are for the 2-bit binary representation method comparing to the other used representation methods. Results suggest that using 2-bit binary representation method significantly enhanced the accuracy of detection and efficiency of the prediction of coding regions using digital filters. Identification and analysis of hidden features of coding and non-coding regions of DNA sequence is a challenging problem in the area of genomics. The objective of the paper authored by Roy and Barman in 2011 is to estimate and compare spectral content of coding and non-coding segments of DNA sequence both by Parametric and Nonparametric methods. Consequently an attempt has been made so that some hidden internal properties of the DNA sequence can be brought into light in order to identify coding regions from non-coding ones. In this approach the DNA sequence from various Homo Sapien genes have been identified for sample test and assigned numerical values based on weak-strong hydrogen bonding (WSHB) before application of digital signal analysis techniques. The statistical methodology applied for computation of Spectral content are simple and the Spectrum plots obtained show satisfactory results. Spectral analysis can be applied to study base-base correlation in DNA sequences. A key role is played by the mapping between nucleotides and real/complex numbers. In 2006, Galleani and Garelo presented a new approach where the mapping is not kept fixed: it is allowed to vary aiming to minimize the spectrum entropy, thus detecting the main hidden periodicities. The new technique is first introduced and discussed through a number of case studies, then extended to encompass time-frequency analysis.

For analyzing periodicities in categorical valued time series, the concept of the spectral envelope was introduced by Stoffer et al., 1993 as a computationally simple and general statistical methodology for the harmonic analysis and scaling of non-numeric sequences. However, The spectral envelope methodology is computationally fast and simple because it is based on the fast Fourier transform and is nonparametric (i.e., it is model independent). This makes the methodology ideal for the analysis of long DNA sequences. Fourier analysis has been used in the analysis of correlated data (time series) since the turn of the century. Of fundamental interest in the use of Fourier techniques is the discovery of hidden periodicities or regularities in the data. Although Fourier analysis and related signal processing are well established in the physical sciences and engineering, they have only recently been applied in molecular biology. Since a DNA sequence can be regarded as a categorical-valued time series it is of interest to discover ways in which time series methodologies based on Fourier

(or spectral) analysis can be applied to discover patterns in a long DNA sequence or similar patterns in two long sequences. Actually, the spectral envelope is an extension of spectral analysis when the data are categorical valued such as DNA sequences.

An algorithm for estimating the spectral envelope and the optimal scalings given a particular DNA sequence with alphabe  $\xi = \{b_1, b_2, \dots, b_{r+1}\}$ , is as follows (Stoffer 2012).

1. Given a DNA sequence of length  $n$ , from the  $r \times 1$  vectors  $Y_t, t = 1, 2, \dots, n$ ; namely, for  $j = 1, 2, \dots, r, Y_t = e_j$  if  $X_t = b_j$  where  $e_j$  is a  $r \times 1$  vector with a 1 in the  $j$ th position as zeros elsewhere, and  $Y_t = 0$  if  $X_t = b_{j+1}$ .
2. Calculate the Fast Fourier Transform FFT of the data,  $d(j/n) = \sum_{t=1}^n Y_t \exp(-2\pi i t j / n) / \sqrt{n}$ . Note that  $d(j/n)$  is a  $r \times 1$  complex-valued vector. Calculate the periodogram,  $\tilde{f}(j/n) = d(j/n) d^*(j/n)$ , for  $j = 1, 2, \dots, [n/2]$ , and retain only the real part, say  $f^{\sim re}(j/n)$ .
3. Smooth the real part of the periodogram as preferred to obtain  $f^{\sim re}(j/n)$ , a consistent estimator of the real part of the spectral matrix.
4. Calculate the  $r \times r$  variance-covariance matrix of the data,  $S = \sum_{t=1}^n (Y_t - \bar{Y})(Y_t - \bar{Y})' / n$ , where  $\bar{Y}$  is the sample mean of the data.
5. For each  $\omega = j/n, j = 1, 2, \dots, [n/2]$ , determine the largest eigenvalue and the corresponding eigenvector of the matrix  $2S^{-1/2} f^{\sim re}(\omega_j) S^{-1/2} / n$ .
6. The sample spectral envelope  $\hat{\lambda}(\omega_j)$  is the eigenvalue obtained in the previous step.
7. The optimal sample scaling is  $\hat{\beta}(\omega_j) = S^{-1/2} v(\omega_j)$ , where  $v(\omega_j)$  the eigenvector obtained in the previous step.

In this paper, we discussed the kernel density estimation for the first, second, third and fourth eigenvalues of variance covariance matrix of Fast Fourier Transform for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper. We computed an empirical values for the kernel density estimation for data series according to the following Kernels: Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine. To determine the valuable of our work, it should be noted that it is the first time that the variance covariance matrix eigenvalues of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences, is used in an analysis like this and related analyzes.

## 2. Kernel Density Estimation

The problem of estimation of a probability density function  $f(x)$  is interesting for many reasons, among which are the possible applications in the field of discriminant analysis or the estimation of functions of the density. The parametric approach to density estimation assumes a functional form for the density and then estimates the unknown parameters using techniques such as the maximum likelihood estimation or Pearson system based on the estimation of the skewness and the Kurtosis (Oja (1981)). Unless the form of density is known a priori, assuming a functional form for a density very often leads to erroneous inference. It is clear that, nonparametric methods do not make any assumptions as to the form of the underlying density. Today, a rich basket of nonparametric density estimators (Kernel, orthogonal series, histogram, etc.) exists (Bowman and Azzalini (1997); Hall (1982); Silverman (1986)). On other words, the parametric methods assumes that the data is drawn from a known parametric family of distributions, for example a normal distribution but the nonparametric methods does not make this assumption. The main restriction of parametric methods is that they impose restrictions on the shapes that  $f$  can have (Lopez-Novoa et al. (2015)). This work focuses on kernel density estimators (KDE).

The kernel density estimator represents one of the important scientific subjects in spectrum theory. It is evident that, the kernel density estimator is a very useful nonparametric method and is very robust to the shape of distribution. An attractive feature of the kernel density estimator is that the resulting estimator can always be bimodal for an appropriate choice of bandwidth, which controls the smoothness of the estimator (Bae and Kim (2008)). Several techniques have been proposed for optimal bandwidth selection. The best known of these contain rules of thumb, oversmoothing, least squares cross-validation, direct plug-in methods, solve-the-equation plug-in method, and the smoothed bootstrap Jones, Marron and Sheather, (1996).

KDE is a common tool in many research areas, used for a variety of purposes, some of the relevant scientific literatures are as follows.

André's Ferreyra et al. in (2001) used density estimates to forecast weather and other factors as part of a model for optimizing maize production. In the same field, it has been applied to evaluate the signature of climate change in the frequency of weather regimes (Corti et al., 1999). Furthermore, in the field of evolutionary computation, density estimation has been used to estimate a distribution of the problem variables in estimation of distribution algorithms (Bosman and Thierens, 2000).

Li et al. in (2007) suggested a nonparametric estimator of the correlation function for data, using kernel methods. They developed a pointwise asymptotic normal distribution for the suggested estimator, when the number of subjects is fixed and the number of vectors or functions within each subject

goes to infinity. Based on the asymptotic theory, they suggested a weighted block bootstrapping method for making inferences about the correlation function, where the weights account for the inhomogeneity of the distribution of the times or locations. The method is applied to a data set from a colon carcinogenesis study, in which colonic crypts were sampled from a piece of colon segment from each of the rats in the experiment and the expression level of p27, an important cell cycle protein, was then measured for each cell within the sampled crypts.

Troudi et al. in (2008) suggested a faster procedure than that of the common plug-in method. The mean integrated square error (MISE) depends directly upon  $J(f)$  which is linked the second-order derivative of the pdf. They presented an analytical approximation of  $J(f)$ , such that the pdf is estimated only once, at the end of iterations. Thus, these two kinds of algorithm are tested on different random variables having distributions known for their difficult estimation.

Bae and Kim in 2008 presented a kernel density estimation approach for the segmentation of the microarray spot. They estimated the density of  $n$  pixel intensities for a given target area by the kernel density estimation, and the resulting kernel density estimate gives bimodal density by appropriate choice of the smoothing parameter. They proposed two modes of the kernel density estimate for  $n$  pixel intensities as estimates of the foreground (mode with larger value) and the background (mode with smaller value) intensity, respectively.

Fan et al. in (2010) studied nonparametric estimation of genewise variance for microarray data. Microarray experiments are one of widely used technologies nowadays, allowing scientists to monitor thousands of gene expressions simultaneously. They presented a two-way nonparametric model, which is an extension of the famous Neyman-Scott model and is applicable beyond microarray data. The problem itself posed interesting challenges because the number of nuisance parameters is proportional to the sample size and it is not obvious how the variance function can be estimated when measurements are correlated. In such a high-dimensional nonparametric problem, they suggested two novel nonparametric estimators for genewise variance function and semiparametric estimators for measurement correlation, via solving a system of nonlinear equations. Their asymptotic normality is established. The finite sample property is demonstrated by simulation studies. The estimators also improve the power of the tests for detecting statistically differentially expressed genes.

Weyenberg et al. in (2014) suggested and implemented KDETREES, a non-parametric method for estimating distributions of phylogenetic trees, with the goal of identifying trees that are significantly different from the rest of the trees in the sample. This method compared favorably with a similar recently published method, featuring an improvement of one polynomial order of computational complexity (to quadratic in the number of trees analyzed), with simulation studies proposing only a small penalty to classification accuracy. Application of KDE TREES to a set

of apicomplexa genes identified several unreliable sequence alignments that had escaped previous detection, as well as a gene independently reported as a possible case of horizontal gene transfer. They also analyzed a set of *Epichloë* genes, fungi symbiotic with grasses, successfully identifying a contrived instance of paralogy. An extensive list of application fields of KDE can be found in Sheather (2004).

Colbrook et al. in (2018) considered a new type of boundary constraint, in which the values of the estimator at the two boundary points are linked. They provided a kernel density estimator that successfully incorporates this linked boundary condition. This is studied via a nonsymmetric heat kernel which generates a series expansion in nonseparable generalised eigenfunctions of the spatial derivative operator. Despite the increased technical challenges, the model is proven to outperform the more familiar Gaussian kernel estimator, yet it inherits many desirable analytical properties of the latter KDE model.

The probability distribution of a random variable  $X$  is described through its probability density function (PDF)  $f$ . This function  $f$  gives a natural description of the distribution of  $X$ , and allows us to determine the probabilities associated with  $X$  using the relationship

$$P(a < X < b) = \int_a^b f(X) dX$$

Given several observed data points (samples) from a random variable  $X$ , with unknown density function  $f$ , density estimation is used to create an estimated density function  $\hat{f}$  from the observed data. Here, we use the non-parametric approach (KDE).

Furthermore, one of the most common techniques for density estimation of a continuous variable is the histogram,

which is a representation of the frequencies of the data over discrete intervals (bins). It is widely used due to its simplicity, but it has several shortcomings, such as the lack of continuity. The KDE technique relies on assigning a kernel function  $K$  to each sample (observation in the dataset), and then summing all the kernels to obtain the estimate. In contrast to the histogram, KDE constructs a smooth probability density function, which may reflect more accurately the real distribution of the data. We now describe the KDE technique in more detail as follows (Lopez-Novoa et al. (2015)).

Let  $x_1, x_2, \dots, x_n$  represent a random sample of size  $n$  from a random variable with density  $f(\cdot)$ . Silverman in 1986 defined the following kernel density estimate of  $f$  at the point  $x$  by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (1)$$

where the smoothing parameter  $h$  is called the bandwidth and  $K$  is generally chosen to be a unimodal probability density symmetric about zero. That is,  $K$  is assumed to be an even regular function with unit variance and zero mean. The Kernel  $K$  is called regular if it is a square integrated density.

For a practical implementation of KDE, the choice of the bandwidth  $h$  is very important. Small  $h$  leads to an estimator with a small bias and large variance, whereas large  $h$  leads to a small variance at the expense of increase: the bandwidth has to be optimally chosen (Troudi et al. (2008)).

In this case,  $K$  satisfies the following conditions, as may be seen below (Sheather and Jones (1991) and Sheather (2004)),

$$\int K(y) dy = 1, \quad \int yK(y) dy = 0, \quad \int y^2 K(y) dy > 0.$$

**Table (1).** Kernels under considerations for the kernel density estimation of  $f$  (Wolodzko, 2017)

Kernel	Equation	Notes
Gaussian	$K(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$	$-\infty < y < \infty$
Epanechnikov	$K(y) = \frac{3}{4}(1 - y^2)1_{\{ y  < 1\}}$	$1_{\{ y  < 1\}} = \begin{cases} 1 & \text{if }  y  < 1 \\ 0 & \text{otherwise} \end{cases}$
Rectangular	$K(y) = \frac{1}{2}$	$ y  \leq 1,$
Triangular	$K(y) = (1 -  y )$	$ y  \leq 1.$
Biweight (Quartic)	$K(y) = \frac{15}{16}(1 - y^2)^2$	$ y  \leq 1$
Cosine	$K(y) = \frac{1}{2}(1 + \cos(\pi y))$	$ y  \leq 1$
Optcosine	$K(y) = \frac{\pi}{4} \cos\left(\frac{\pi}{2} y\right)$	$ y  \leq 1$

In practice, there are a lot of popular choices of kernel function  $K$  such as: Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, Optcosine and others, the focus of attention was on this kernels (Silverman (1986)).

The most commonly kernels used have been considered as in Table (1).

In addition, the bandwidth controls the smoothness of the density estimate and highly influences its appearance. Selecting a suitable  $h$  is a pivotal step in estimating  $f(x)$ .

Supposing that the underlying density is sufficiently smooth and that the kernel has finite fourth moment, by using Taylor series it can be clearly demonstrated that (Sheather (2004)),

$$\text{Bias}\{\hat{f}_h(x)\} = \frac{h^2}{2} \mu_2(K) f''(x) + o(h^2)$$

$$\text{Var}\{\hat{f}_h(x)\} = \frac{1}{nh} f(x) R(K) + o\left(\frac{1}{nh}\right),$$

where  $R(K) = \int K^2(y) dy$

Adding the leading variance and squared bias terms produces the asymptotic mean squared error (AMSE) as,

$$\text{AMSE}\{\hat{f}_h(x)\} = \frac{1}{nh} f(x) R(K) + \frac{h^4}{4} \mu_2^2(K) (f''(x))^2,$$

An extensively used choice of an overall measure of the discrepancy between  $\hat{f}$  and  $f$  is the mean integrated squared error (MISE), which is defined by,

$$\begin{aligned} \text{MISE}\{\hat{f}_h\} &= E\left\{\int (\hat{f}_h(y) - f(y))^2 dy\right\} \\ &= \int \text{Bias}(\hat{f}_h(y))^2 dy + \int \text{Var}(\hat{f}_h(y)) dy \end{aligned}$$

Under an integrability assumption on  $f$ , integrating the expression for AMSE gives the expression for the asymptotic mean integrated squared error (AMISE), that is,

$$\text{AMISE}\{\hat{f}_h\} = \frac{1}{nh} R(K) + \frac{h^4}{4} \mu_2^2(K) R(f''), \quad (2)$$

where  $R(f'') = \int [f''(y)]^2 dy$

In this way, the value of the bandwidth that minimizes the AMISE is obtained by the following

$$h_{\text{AMISE}} = \left[ \frac{R(K)}{R(f'') \mu_2^2(K)} \right]^{1/5} n^{-1/5}$$

supposing that  $f$  is sufficiently smooth, we can use integration by parts to demonstrate that,

$$R(f'') = \int [f''(y)]^2 dy = - \int f^{(4)}(y) f(y) dy$$

Thus, the functional  $R(f'')$  is a measure of the underlying roughness or curvature. In particular, It is clear that the larger the value of  $R(f'')$  is, the larger is the value

of AMISE (i.e., the more difficult it is to estimate  $f$ ) and the smaller is the value of  $h_{\text{AMISE}}$  (i.e., the smaller the bandwidth needed to capture the curvature in  $f$ ).

*Following, we will briefly review some famous methods for choosing a global value of the bandwidth  $h$ ,*

## 2.1. Rules of Thumb

The computationally simplest method for choosing a global bandwidth  $h$  is based on replacing  $R(f'')$ , the unknown part of  $h_{\text{AMISE}}$ , by its value for a parametric family expressed as a multiple of a scale parameter, which is then estimated from the data. The method seems to date back to Deheuvels, (1977) and Scott, (1979), who each introduced it for histograms. The method was popularized for kernel density estimates by Silverman, 1986, who used the normal distribution as the parametric family.

Let  $\sigma$  and  $Q$  denote the standard deviation and interquartile range of  $X$ , respectively. Take the kernel  $K$  to be the usual Gaussian kernel. Assuming that the underlying distribution is normal, Silverman, 1986 showed that (2) reduces to,

$$h_{\text{SROT}} = 0.9 A n^{-1/5}$$

where  $A = \text{Min}\{s, q/1.34\}$ , and  $s, q$  are sample standard deviation and sample interquartile range respectively. This rule is commonly used in practice and it is often referred to as Silverman's rule of thumb.

## 2.2. Cross-Validation Methods

A measure of the closeness of  $\hat{f}$  and  $f$  for a given sample is the integrated squared error (ISE), which is defined by Bowman, 1984

$$\begin{aligned} \text{ISE}\{\hat{f}_h\} &= \int (\hat{f}_h(y) - f(y))^2 dy \\ &= \int (\hat{f}_h(y))^2 dy - 2 \int \hat{f}_h(y) \cdot f(y) dy + \int f^2(y) dy \end{aligned}$$

It is clear that, the last term on the right-hand side of the previous expression does not involve  $h$ . Bowman, 1984 presented choosing the bandwidth as the value of  $h$  that minimizes the estimate of the two other terms in the last expression, namely,

$$\frac{1}{n} \sum_{i=1}^n \int (\hat{f}_{-i}(y))^2 dy - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i)$$

where  $\hat{f}_{-i}(y)$  denotes the kernel estimator constructed from the data without the observation  $x_i$ . However, the method is commonly referred to as least squares cross-validation, since it is based on the so-called leave-one-out density estimator  $\hat{f}_{-i}(y)$ .

Scott and Terrell, 1987 introduced a method called biased cross-validation (BCV), which is based on choosing the bandwidth that minimizes an estimate of AMISE rather than an estimate of ISE. The BCV objective function is just the estimate of AMISE obtained by replacing  $R(f'')$  in (2) by



$$R(\hat{f}_h'') - \frac{1}{nh^5} R(K'')$$

where  $\hat{f}_h''$  is the second derivative of the kernel density estimate (1) and the subscript  $h$  denotes the fact that the bandwidth used for this estimate is the same one used to estimate the density  $f$  itself. The BCV objective function is thus defined by Scott and Terrell, 1987

$$BCV(h) = \frac{1}{nh} R(K) + \frac{\mu_2^2(K)}{2hn^2} \sum_{i < j} \sum \varphi\left(\frac{X_i - X_j}{h}\right)$$

Where  $\varphi(c) = \int K''(w) K''(w+c) dw$ . We denote the bandwidth that minimizes  $BCV(h)$  by  $h_{BCV}$ . It can be seen that, the above methods are well-known first generation methods for bandwidth selection. These methods were mostly introduced before 1990.

Following some of second generation methods, which are introduced after 1990.

### 2.3. Smoothed Bootstrap

One approach to this method is to consider the bandwidth that is a minimizer of a smoothed bootstrap approximation to the MISE. Early versions of this were presented by Faraway and Jhun, 1990. An interesting feature of this approach is that unlike most bootstrap applications, the MISE in the "bootstrap world can be calculated exactly, instead of requiring the usual simulation step, which makes it as computationally fast as other methods discussed here. Furthermore, the basic idea behind bootstrap smoothing is very simple, we estimate  $MISE(h)$  by a bootstrap version of the form,

$$MISE_*(h) = E_* \int (\hat{f}_h^*(y) - \hat{f}_g(y))^2 dy$$

where  $E_*$  denotes the expectation with respect to the bootstrap sample  $x_1^*, x_2^*, \dots, x_n^*$ ,  $g$  is some pilot bandwidth,  $\hat{f}_g(y)$  is a density estimate which depends on the original sample  $x_1, x_2, \dots, x_n$  and  $\hat{f}_h^*(y)$  is an estimate, based on  $x_1^*, x_2^*, \dots, x_n^*$ . Then, we choose the value  $h$  minimizing  $MISE_*(h)$ . The basic differences among the various versions of this bootstrap methodology lie on the choice of the auxiliary window  $g$  and on the procedure (smoothed or not) for generating the resampled data  $x_1^*, x_2^*, \dots, x_n^*$ . Here we introduced a smoothed bootstrap procedure (*Smoothed bootstrap with pilot bandwidth*) which is considered by Faraway and Jhun, 1990 (the bootstrap sample is taken from  $\hat{f}_g$ ) where  $g$  is chosen by least-squares cross-validation from  $x_1, x_2, \dots, x_n$ . These authors do not use an exact expression for  $MISE_*(h)$ . They approximate  $MISE_*(h)$  by resampling; that is,  $B$  bootstrap samples are drawn from

$\hat{f}_g$ , and  $MISE_*(h)$  is approximated by,

$$BMISE(h) = B^{-1} \sum_{j=1}^B \int (\hat{f}_{h(j)}^*(y) - \hat{f}_g(y))^2 dy$$

Where  $\hat{f}_{h(j)}^*(y)$  denotes the value of the estimator for the  $j$ -th bootstrap sample. The resulting bandwidth  $h$ , is defined as the value of  $h$  which minimizes  $BMISE(h)$ .

### 2.4. Plug-in Methods

The slow rate of convergence of LSCV and BCV encouraged much research on faster converging methods. A popular approach, commonly called plug-in methods, is to replace the unknown quantity  $R(f'')$  in the expression for  $h_{AMISE}$  given by (2) with an estimate. We next describe the "solve the-equation" plug-in approach developed by Sheather and Jones, 1991, since this method is widely recommended (e.g., Bowman and Azzalini, 1997). The Sheather and Jones, 1991 approach is based on writing  $g$ , the pilot bandwidth for the estimate  $R(\hat{f}'')$  as a function of  $h$ , namely,

$$g(h) = C(K) \left[ \frac{R(f'')}{R(f''')} \right]^{1/7} h^{5/7}$$

The Sheather–Jones plug-in bandwidth  $h_{SJ}$  is the solution to the above equation.

## 3. Bivariate Kernel Density Estimation

In the bivariate case the data points are considered by two vectors  $x_1 = [x_{11}, x_{12}, x_{13}, \dots, x_{1n}]$  and  $x_2 = [x_{21}, x_{22}, x_{23}, \dots, x_{2n}]$  where  $x_i = (x_{1i}, x_{2i})$  is a sample from a bivariate distribution  $f$ . In analogy with the univariate case, the bivariate kernel density estimate is given by the following equation (Billock et al. (2016))

$$\hat{f}_H(x) = \frac{1}{nH} \sum_{i=1}^n K\left(\frac{x - x_i}{H}\right) \quad (3)$$

Here the bandwidth is the positive definite matrix,

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix},$$

and the kernel function  $K$  is a symmetric and non negative function fulfilling such that  $\int_{R^2} K(y) dy = 1$ .

As in the univariate case the bivariate kernels used in this work have been the Gaussian kernel,

$$K(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^T y}, \text{ and the Epanechnikov kernel}$$

$$K(y) = \frac{2}{\pi} (1 - y^T y) 1_{\{|y^T y| < 1\}},$$

and other bivariate kernels. To evaluate the closeness of a kernel density estimator to the target density an error criteria must be used. A common error estimate for kernel density estimation is the Mean Integrated Square Error (MISE) which is defined by (Billock et al. (2016)):

$$MISE\{\hat{f}\} = E\left\{\int (\hat{f}_H(x) - f(x))^2 dx\right\}$$

In this case, since the Mean Integrated Square Error depends on the true density  $f$  it can only be computed for data sets drawn from known distributions  $f$ . The Mean Integrated Square Error can be approximated with the Integrated Mean Square Error IMSE. The expression for the Integrated Mean Square Error IMSE is extracted by moving the expectation value in above equation inside the integral. The IMSE can be computed numerically using Monte Carlo integration, for example. In the bivariate case the plug in method aims to minimize the bivariate AMISE, that is:

$$AMISE\{\hat{f}(H)\} = n^{-1} |H|^{-\frac{1}{2}} R(K) + \frac{1}{4} \mu_2(K)^2 \left( \text{vech}^T H \psi_4 \text{vech} H \right),$$

where  $\text{vech}$  represents the following operation

$$\text{vech}H = \text{vech} \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix} = [h_{11} \ h_{12} \ h_{22}]^T.$$

The  $3 \times 3$  matrix  $\psi_4$  is given as

$$\psi_4 = \begin{pmatrix} 1\psi_{40} & 1\psi_{31} & 1\psi_{22} \\ 2\psi_{31} & 4\psi_{22} & 2\psi_{13} \\ 1\psi_{22} & 2\psi_{13} & 1\psi_{04} \end{pmatrix},$$

where  $\psi_{\eta_1 \eta_2} = \int_{R^2} f^{(\eta_1, \eta_2)}(x) f(x) dx$

and  $f^{(\eta_1, \eta_2)}(x) = \frac{\partial^4}{\partial \eta_1 x_1 \partial \eta_2 x_2} f(x)$

is the partial derivatives of  $x$  with respect to  $x_1$  and  $x_2$ .

Thus, As in the univariate case  $\psi_{\eta_1, \eta_2}$  has to be estimated. A commonly used estimate is

$$\hat{\psi}_{(\eta_1, \eta_2)}(G) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n K_G^{(\eta_1, \eta_2)}(X_i - X_j),$$

where  $K_G(y) = \frac{1}{G} K\left(\frac{y}{G}\right)$

and  $G$  is the pilot bandwidth matrix. In Doung and Hazelton (2003), it is proposed that this matrix should be on the form  $G = g^2 I$ . Choosing  $g$  can be done in a similar way as in

the univariate case. For each entry  $\psi_{(\eta_1, \eta_2)}$  in  $\hat{\psi}_4$ ,  $g = g_{AMSE}$  is chosen such that it minimises the Asymptotic Mean Square Error approximation

$$AMSE\hat{\psi}_{(\eta_1, \eta_2)}(g) = 2n^{-2} g^{-2(\eta_1 + \eta_2) - 2} \psi_0 R(K^{(\eta_1 + \eta_2)}) + \left( n^{-1} g^{-(\eta_1 + \eta_2) - 2} K^{(\eta_1 + \eta_2)}(0) + \frac{1}{2} g^2 \mu_2(K) (\psi_{\eta_1 + 2, \eta_2} + \psi_{\eta_1, \eta_2 + 2}) \right)^2$$

This method may produce  $\psi_4$  matrices that are not positive definite. In that case a minimum to the objective function does not exist. To solve this issue, many researchers such as Doung and Hazelton propose a new approach as opposed to finding one optimal  $g$  for each entry in  $\hat{\psi}_4$ . Instead,  $g = g_{4, SAMSE}$  that minimizes the sum

$$SAMSE - \sum_{\eta_1 + \eta_2 = 4} AMSE\hat{\psi}_{(\eta_1, \eta_2)}(g)$$

should be calculated and used as a common  $g$  for all entries in  $\hat{\psi}_4$ . A closed form expression for  $g_{4, SAMSE}$  is given in Doung and Hazelton (2003). In analogy with the univariate case, the estimates of  $g$  depends on  $\psi_{\eta_1, \eta_2}$  and therefore an easy estimate of  $\hat{\psi}_{\eta_1, \eta_2}$  has to be made at some stage (Billock et al. (2016)).

## 4. The Proposed Algorithm

The following algorithm steps is performed to achieve our aims

1. Generate the DNA sequence for five organisms, Human, E. coli, Rat, Wheat and Grasshopper with corresponding information in table (2).

**Table (2).** Relative proportions (%) of Bases in DNA

Organisms	A	T	G	C
Human	30.9	29.4	19.9	19.8
E. coli	26.0	23.9	24.9	25.2
Rat	28.6	28.4	21.4	21.5
Wheat	27.3	27.1	22.7	22.8
Grasshopper	29.3	29.3	20.5	20.7

2. The sequence size is  $n=500$  and run size is  $k=205$ .
3. Transform DNA sequence to numerical values by setting one to the base that appears and zero to the other bases.
4. Transform the sequence of numerical values to the corresponding FFT values.
5. Calculate the eigenvalues of variance covariance matrix for each run results, and then we get 205 fourth order vectors of eigenvalues for each organism. Each vector contains the four eigenvalues, rank from the largest one to the smallest.



## 5. An Empirical Study and Results Discussion

### 5.1. Univariate Density Estimation

The nonparametric method (kernel density estimation) has been applied of the first, second, third and fourth eigenvalues of variance - covariance matrix of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper. It should be noted that it is the first time that the variance covariance matrix eigenvalues of Fast Fourier Transform (FFT) for numerical values representation of DNA sequences, is used in an analysis like this and related analyzes.

We compute an empirical values for the kernel density estimation for data series according to the following Kernels: Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine. The results of simulaion experiment are recorded in variety of tables as may be seen below that is, the results is displayed in a series of images. This section summarizes the results of this work by table (8) in appendix (1) and represented in Figures (12 – 31) in appendix (2).

**Table (3).** Min Max density (at corresponding x-value) and bandwidth for a kernel among kernels under consideration of the first eigenvalue for each organism

Organism	Bandwidth	Min max	x-value	Kernel
Human	1.85671	0.06740	154.88260	Optcosine
Grasshopper	1.97662	0.06025	151.28073	Biweight
E-Coli	1.93667	0.06154	138.54199	Biweight
Rat	2.55799	0.04759	147.27848	Triangular
Wheat	2.11323	0.05259	146.07245	Cosine

**Table (4).** Min Max density (at corresponding x-value) and bandwidth for a kernel among kernels under consideration of the second eigenvalue for each organism

Organism	Bandwidth	Min max	x-value	Kernel
Human	2.16573	0.05499	141.80183	Epanechnikov
Grasshopper	2.35564	0.04921	142.70795	Epanechnikov
E-Coli	1.28408	0.09281	130.25672	Triangular
Rat	2.40422	0.04378	134.74596	Biweight
Wheat	1.72088	0.06842	128.31104	Gaussian

**Table (7).** Empirical values for the Bandwidth and Correlation used in KDE of each of five organisms

Organism	First and Second Eigenvalues			First and Fourth Eigenvalues		
	x-axis Bandwidth	y-axis Bandwidth	Correlation	x-axis Bandwidth	y-axis Bandwidth	Correlation
Human	1.85392	2.66480	-0.06517	1.85392	1.97014	-0.57761
Grasshopper	2.41621	2.47512	0.05680	2.41621	2.46776	-0.49434
E-Coli	2.36051	1.48344	0.05752	2.36051	2.32454	-0.58715
Rat	3.13377	2.90817	-0.12735	3.13377	2.64811	-0.66749
Wheat	2.57170	2.08196	-0.19545	2.57170	2.28586	-0.55060

**Table (5).** Min Max density (at corresponding x-value) and bandwidth for a kernel among kernels under consideration of the third eigenvalue for each organism

Organism	Bandwidth	Min max	x-value	Kernel
Human	1.85671	0.06343	106.89360	Epanechnikov
Grasshopper	2.29603	0.05181	108.56873	Cosine
E-Coli	1.39526	0.08385	122.81705	Optcosine
Rat	2.01501	0.06203	115.36226	Epanechnikov
Wheat	1.39526	0.08040	118.20472	Gaussian

**Table (6).** Min Max density (at corresponding x-value) and bandwidth for a kernel among kernels under consideration of the fourth eigenvalue for each organism

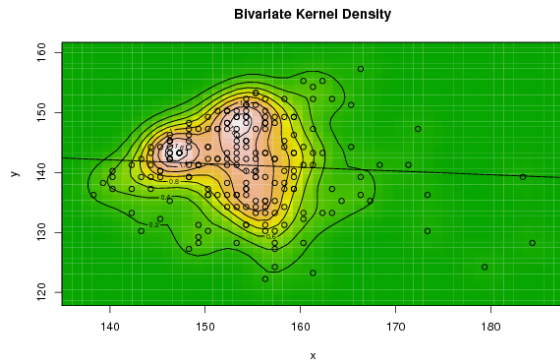
Organism	Bandwidth	Min max	x-value	Kernel
Human	1.85671	0.06863	97.64718	Optcosine
Grasshopper	1.86035	0.06063	97.56647	Epanechnikov
E-Coli	2.15815	0.05620	109.65297	Epanechnikov
Rat	2.32544	0.05368	104.59309	Gaussian
Wheat	1.95486	0.06345	109.84217	Cosine

Tables (3 - 6) shows an optimal kernel estimation among kernels under considerations, which are commonly used for these purposes worldwide.

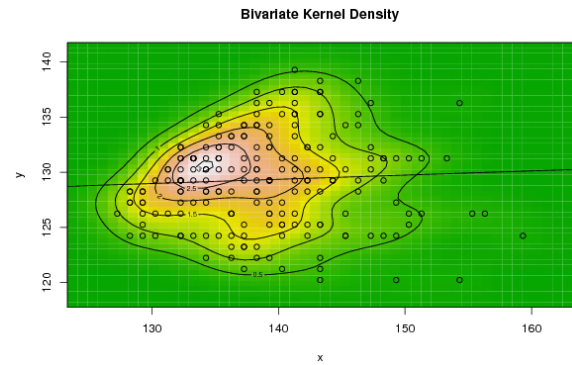
### 5.2. Bivariate Density Estimation

The bivariate kernel density estimation has been applied of the first and second eigenvalues from a side and the first and fourth eigenvalues from another side. These eigenvalues of variance covariance matrix of Fast Fourier Transform for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper.

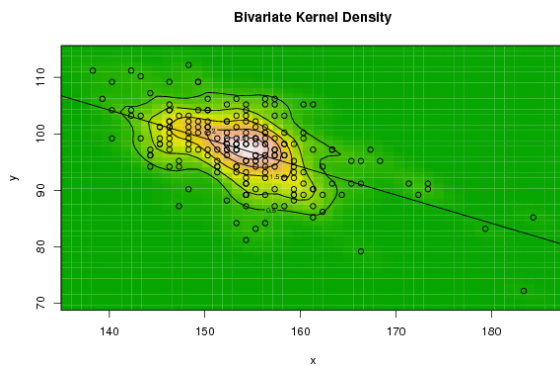
Table (7) contain the results referred to below.



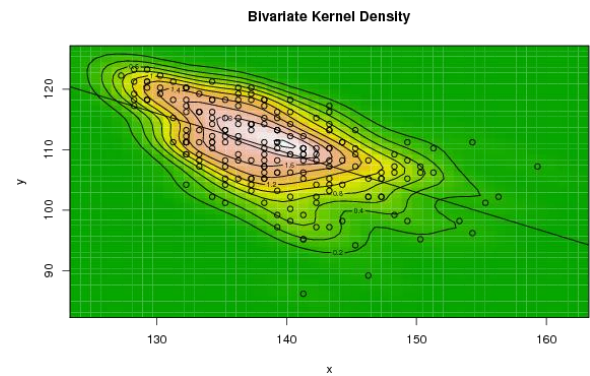
**Figure 2.** Plot of bivariate kernel density for Human according to first and second eigenvalues



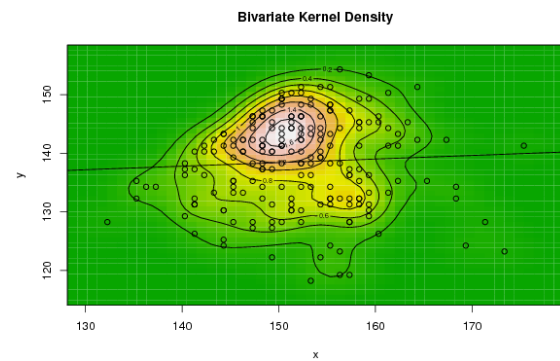
**Figure 6.** Plot of bivariate kernel density for E-Coli according to first and second eigenvalues



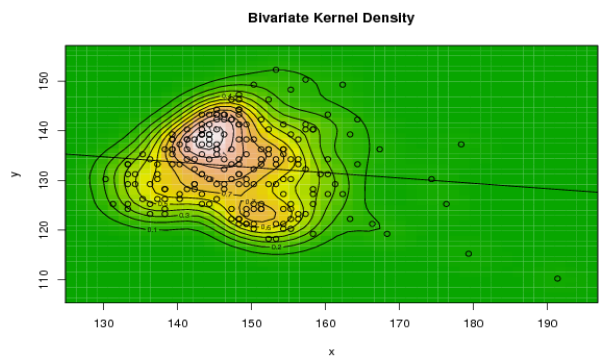
**Figure 3.** Plot of bivariate kernel density for Human according to first and fourth eigenvalues



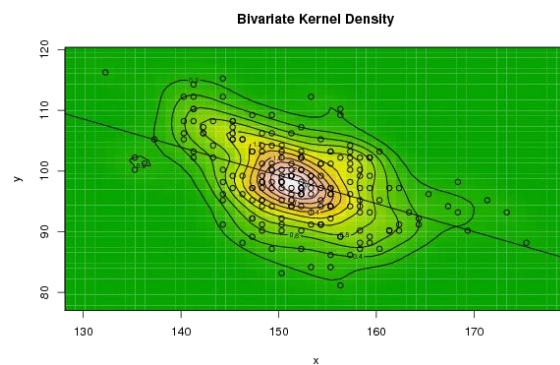
**Figure 7.** Plot of bivariate kernel density for E-Coli according to first and fourth eigenvalues



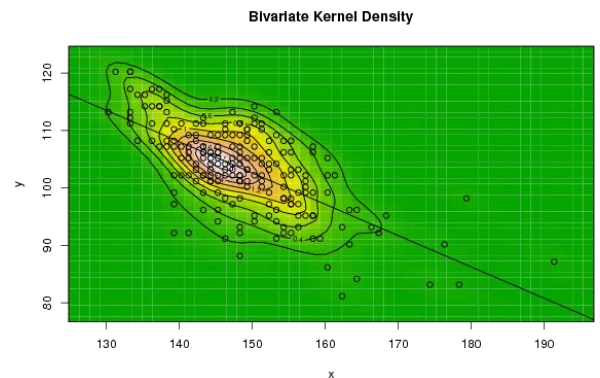
**Figure 4.** Plot of bivariate kernel density for Grasshopper according to first and second eigenvalues



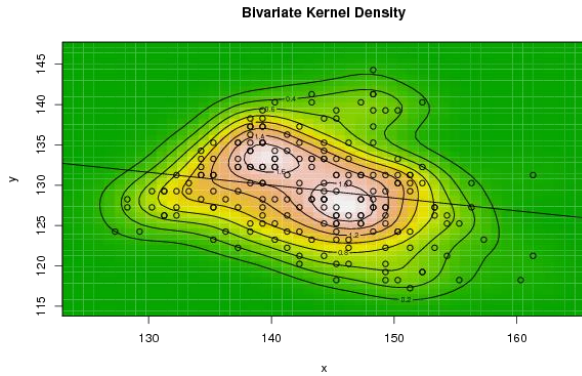
**Figure 8.** Plot of bivariate kernel density for Rat according to first and second eigenvalues



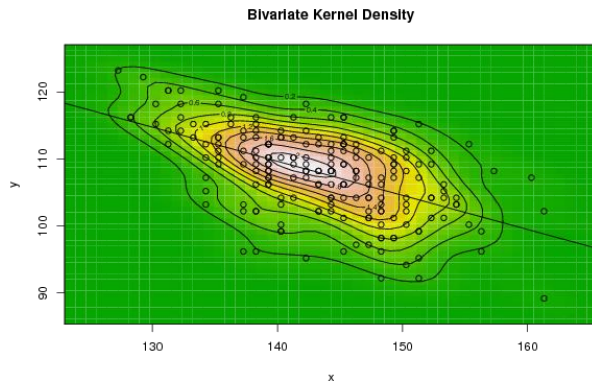
**Figure 5.** Plot of bivariate kernel density for Grasshopper according to first and fourth eigenvalues



**Figure 9.** Plot of bivariate kernel density for Rat according to first and fourth eigenvalues



**Figure 10.** Plot of bivariate kernel density for Wheat according to first and second eigenvalues



**Figure 11.** Plot of bivariate kernel density for Wheat according to first and fourth eigenvalues

The results in table (7) and plots will be crucial to determine DNA affiliation.

## 6. Summary

## Appendix (1)

**Table (A.1).** Empirical values of maximum density and bandwidth for the (first, second, third and fourth) eigenvalues of each of five organisms

Kernel	x-value	Max. Density	Bandwidth(h)	Organisms
Gaussian	154.77060	0.06857	1.85671	Human (first eigenvalue)
Epanechnikov	154.88260	0.06748		
Rectangular	155.33061	0.07119		
Triangular	154.32260	0.06812		
Biweight	154.99460	0.06768		
Cosine	154.88260	0.06784		
Optcosine	154.88260	0.06740		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	142.55431	0.05680	2.16573	Human (second eigenvalue)
Epanechnikov	141.80183	0.05499		
Rectangular	140.67311	0.05721		
Triangular	142.27213	0.05610		
Biweight	142.36619	0.05550		
Cosine	142.36619	0.05570		
Optcosine	142.74243	0.05511		

One of important properties of FFT, which is return the variations to their original sources help us tremendously to understand, discriminant and deal with DNA sequences.

Probability behavior of DNA data, represented by the density estimation, is very important. The references stated here, highlights some aspects of that importance.

Univariate density estimation is fitted for first, second, third and fourth eigenvalues of variance covariance matrix of Fast Fourier Transform for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper.

Bivariate density estimation is fitted for first and second eigenvalues from a side and first and fourth eigenvalues from another side. We treat here with first and second eigenvalues from a side and first and fourth eigenvalues from another side because we think that the relations between them are more interactive than others. These eigenvalues of variance covariance matrix of Fast Fourier Transform for numerical values representation of DNA sequences of five organisms, Human, E. coli, Rat, Wheat and Grasshopper.

The methods used here are aimed to discriminant among different organisms using another point of view. This point of view is based on eigenvalues of variance covariance matrix of FFT for numerical values representation of DNA sequences. It should be noted that, it is the first time this point of view is used to achieve aims like ours.

Empirical studies are conducted to show the value of our point of view and the applications based on. So we recommended that,

- (1) Other empirical studies should be done for other organisms and statistical methods by using the point of view adopted here.
- (2) Aspects stated here must be used in an applied manner for DNA sequences discrimination.

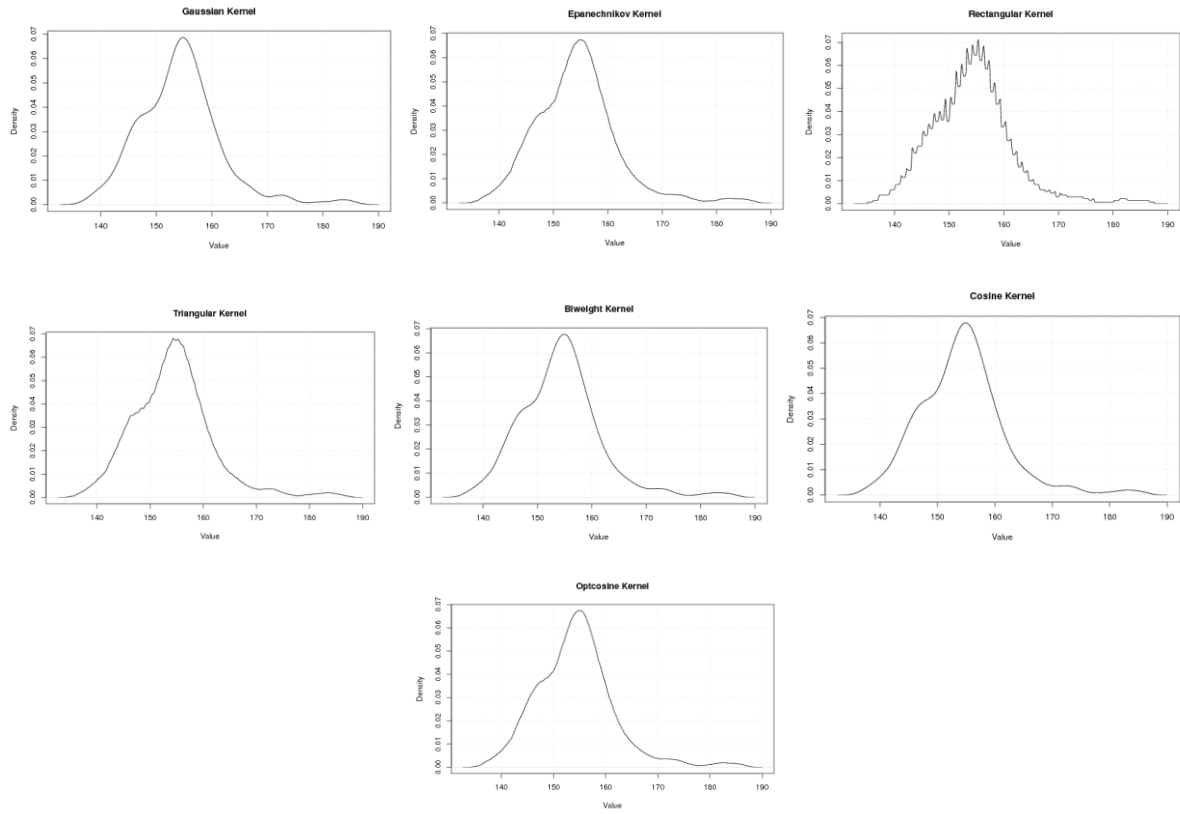
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	106.63408	0.06423	1.85671	Human (third eigenvalue)
Epanechnikov	106.89360	0.06343		
Rectangular	107.23964	0.06825		
Triangular	106.20153	0.06488		
Biweight	106.98011	0.06344		
Cosine	106.72059	0.06361		
Optcosine	107.15313	0.06348		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	97.84765	0.06978	1.85671	Human (fourth eigenvalue)
Epanechnikov	97.64718	0.06865		
Rectangular	97.24624	0.07204		
Triangular	98.14836	0.06965		
Biweight	97.74742	0.06880		
Cosine	97.74742	0.06897		
Optcosine	97.64718	0.06863		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	151.28073	0.06032	1.97662	Grasshopper (first eigenvalue)
Epanechnikov	151.38825	0.06048		
Rectangular	151.17320	0.06115		
Triangular	151.28073	0.06053		
Biweight	151.28073	0.06025		
Cosine	151.28073	0.06026		
Optcosine	151.38825	0.06048		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	142.41320	0.04975	2.35564	Grasshopper (second eigenvalue)
Epanechnikov	142.70795	0.04921		
Rectangular	142.31495	0.05294		
Triangular	142.31495	0.04973		
Biweight	142.60970	0.04933		
Cosine	142.51145	0.04937		
Optcosine	142.41320	0.04930		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	108.77560	0.05200	2.29603	Grasshopper (third eigenvalue)
Epanechnikov	108.36187	0.05199		
Rectangular	107.53440	0.05450		
Triangular	108.25843	0.05212		
Biweight	108.46530	0.05182		
Cosine	108.56873	0.05181		
Optcosine	108.25843	0.05201		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	97.47599	0.06161	1.86035	Grasshopper (fourth eigenvalue)
Epanechnikov	97.56647	0.06063		
Rectangular	97.20457	0.06494		
Triangular	97.20457	0.06210		
Biweight	97.38552	0.06075		
Cosine	97.56647	0.06089		
Optcosine	97.29504	0.06070		

Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	138.45650	0.06199	1.93667	E-Coli (first eigenvalue)
Epanechnikov	138.20004	0.06173		
Rectangular	138.20004	0.06461		
Triangular	138.28552	0.06231		
Biweight	138.54199	0.06154		
Cosine	138.54199	0.06163		
Optcosine	138.20004	0.06175		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	129.99505	0.09295	1.28408	E-Coli (second eigenvalue)
Epanechnikov	129.89038	0.09363		
Rectangular	130.20438	0.10188		
Triangular	130.25672	0.09281		
Biweight	129.99505	0.09310		
Cosine	130.04738	0.09301		
Optcosine	129.89038	0.09353		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	123.42334	0.08471	1.39526	E-Coli (third eigenvalue)
Epanechnikov	122.81705	0.08420		
Rectangular	121.87394	0.08663		
Triangular	123.28861	0.08651		
Biweight	123.22124	0.08399		
Cosine	123.28861	0.08409		
Optcosine	122.81705	0.08385		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	110.04454	0.05686	2.15815	E-Coli (fourth eigenvalue)
Epanechnikov	109.65297	0.05620		
Rectangular	107.59723	0.05864		
Triangular	110.14243	0.05663		
Biweight	109.84876	0.05632		
Cosine	109.84876	0.05642		
Optcosine	109.65297	0.05634		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	147.12883	0.04778	2.55799	Rat (first eigenvalue)
Epanechnikov	146.97918	0.04774		
Rectangular	147.12883	0.04892		
Triangular	147.27848	0.04759		
Biweight	146.97918	0.04772		
Cosine	146.97918	0.04770		
Optcosine	146.82954	0.04779		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	135.18831	0.04419	2.40422	Rat (second eigenvalue)
Epanechnikov	133.30834	0.04387		
Rectangular	132.31306	0.04850		
Triangular	134.30362	0.04434		
Biweight	134.74596	0.04378		
Cosine	134.96713	0.04386		
Optcosine	134.19303	0.04381		

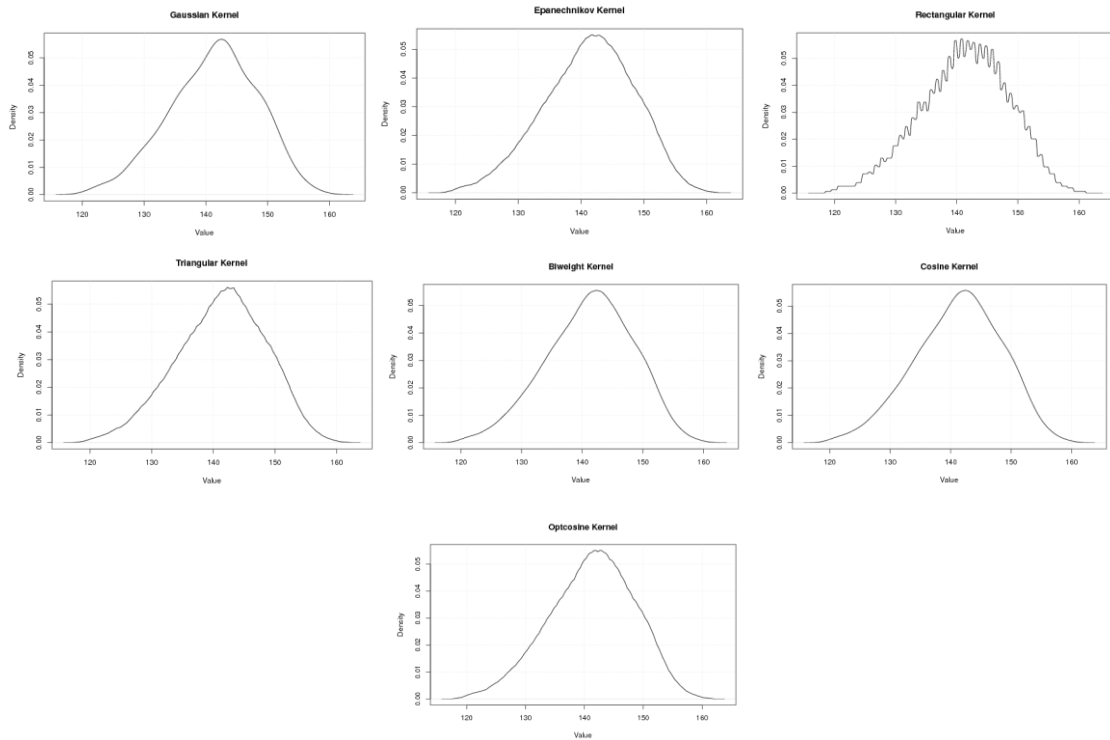
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	116.33431	0.06344	2.01501	Rat (third eigenvalue)
Epanechnikov	115.36226	0.06203		
Rectangular	114.74368	0.06576		
Triangular	116.15757	0.06259		
Biweight	115.89247	0.06229		
Cosine	116.15757	0.06249		
Optcosine	116.06921	0.06204		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	104.59309	0.05368	2.32544	Rat (fourth eigenvalue)
Epanechnikov	104.80065	0.05447		
Rectangular	104.48931	0.05564		
Triangular	105.21576	0.05373		
Biweight	104.69687	0.05410		
Cosine	104.69687	0.05398		
Optcosine	105.00821	0.05425		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	145.52356	0.05280	2.11323	Wheat (first eigenvalue)
Epanechnikov	146.25542	0.05280		
Rectangular	141.77278	0.05921		
Triangular	145.34059	0.05298		
Biweight	146.16393	0.05262		
Cosine	146.07245	0.05259		
Optcosine	145.88949	0.052707		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	128.31104	0.06842	1.72088	Wheat (second eigenvalue)
Epanechnikov	128.74993	0.06928		
Rectangular	128.67678	0.07106		
Triangular	128.23789	0.06893		
Biweight	128.89623	0.06864		
Cosine	128.53048	0.06856		
Optcosine	128.74993	0.06918		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	118.20472	0.08040	1.39526	Wheat (third eigenvalue)
Epanechnikov	118.71227	0.08186		
Rectangular	118.90260	0.08459		
Triangular	119.21982	0.08108		
Biweight	118.45850	0.08089		
Cosine	118.39505	0.08072		
Optcosine	118.64883	0.08143		
Kernel	x-value	max. density	Bandwidth(h)	Organisms
Gaussian	109.57330	0.06352	1.95486	Wheat (fourth eigenvalue)
Epanechnikov	110.02142	0.06382		
Rectangular	108.94594	0.06619		
Triangular	110.20066	0.06380		
Biweight	109.84217	0.06356		
Cosine	109.84217	0.06345		
Optcosine	110.02142	0.06380		



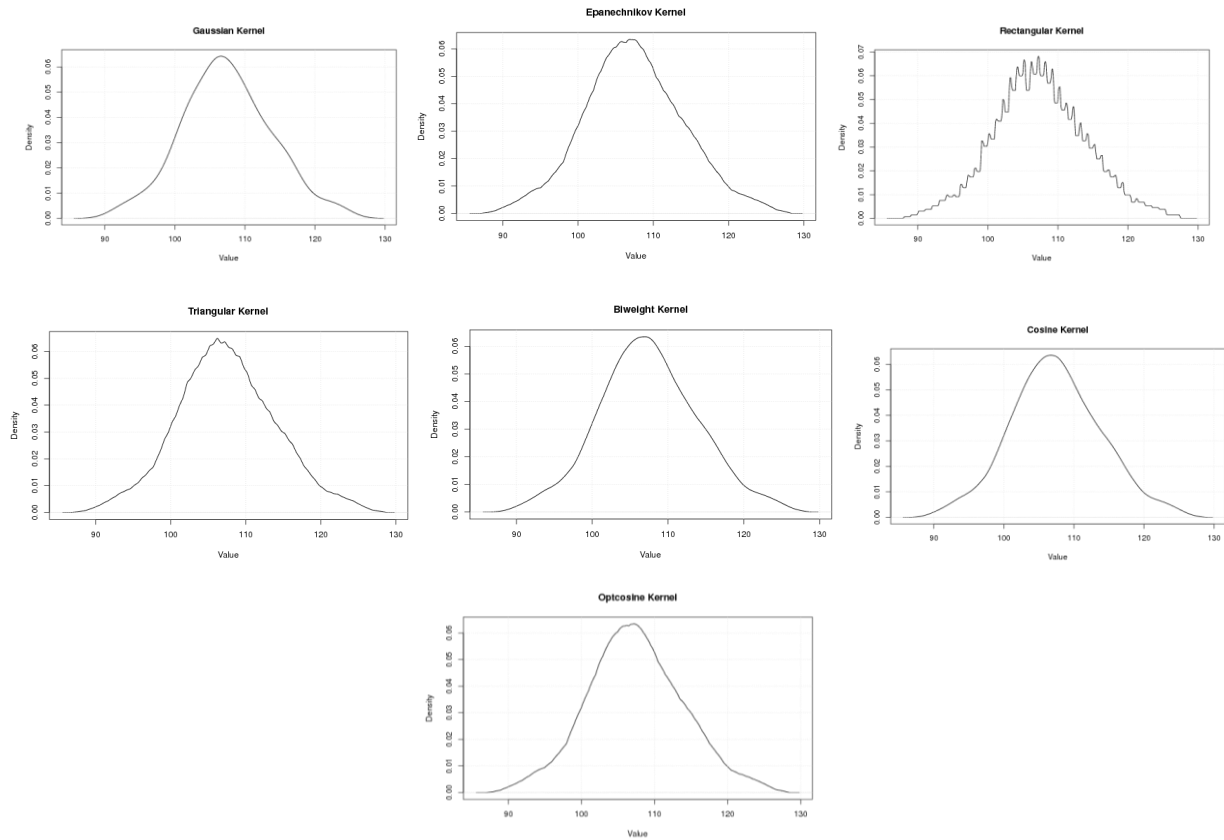
## Appendix (2)



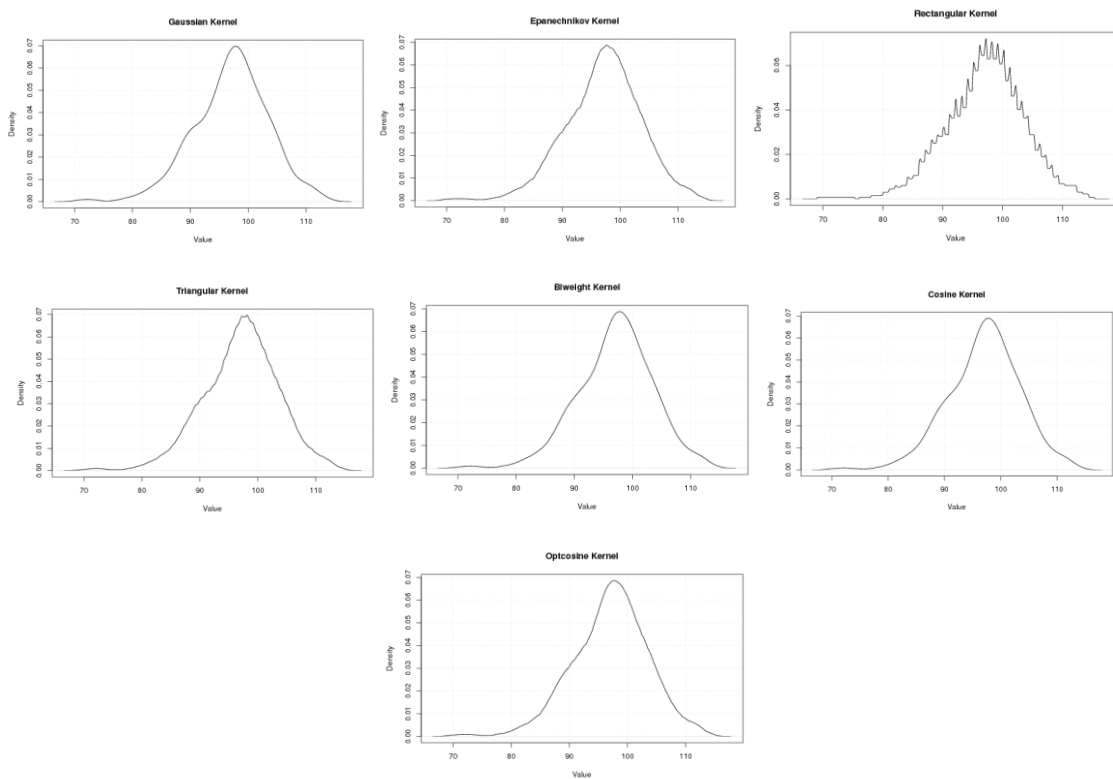
**Figure A.1.** Kernel density estimation plots for the first eigenvalue of Human according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



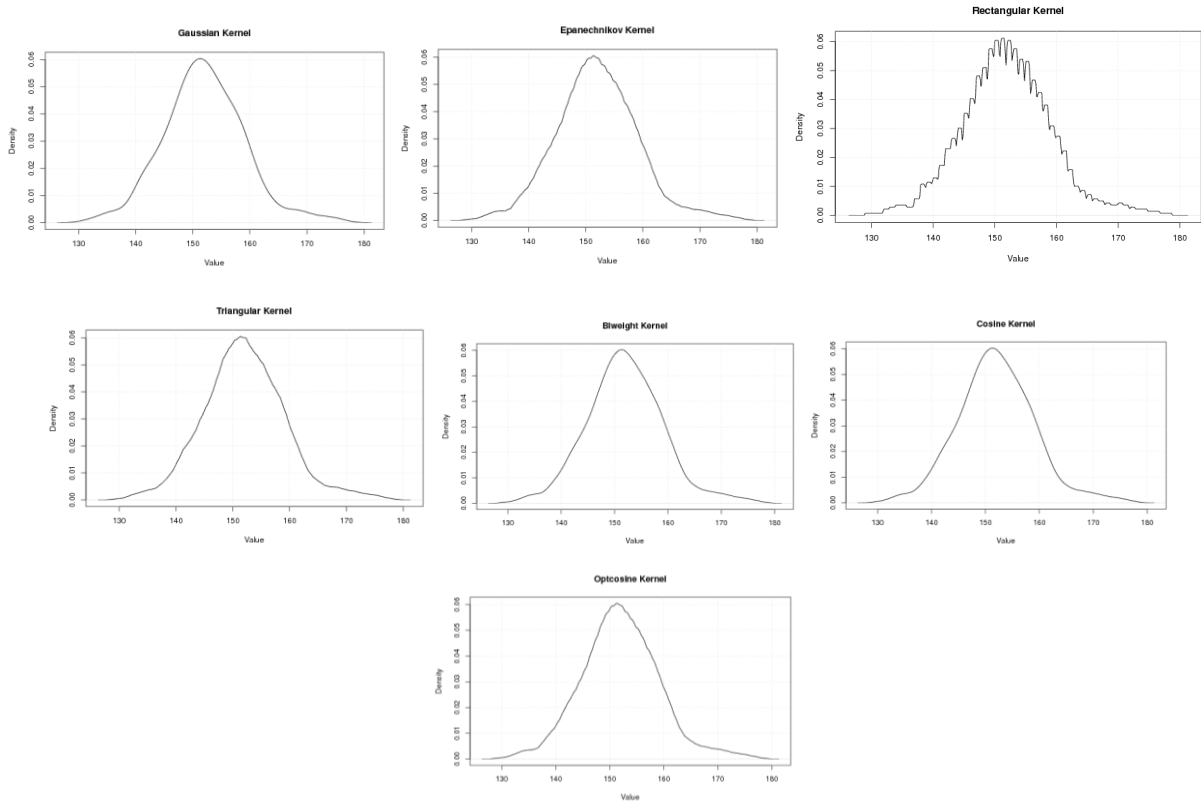
**Figure A.2.** Kernel density estimation plots for the second eigenvalue of Human according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



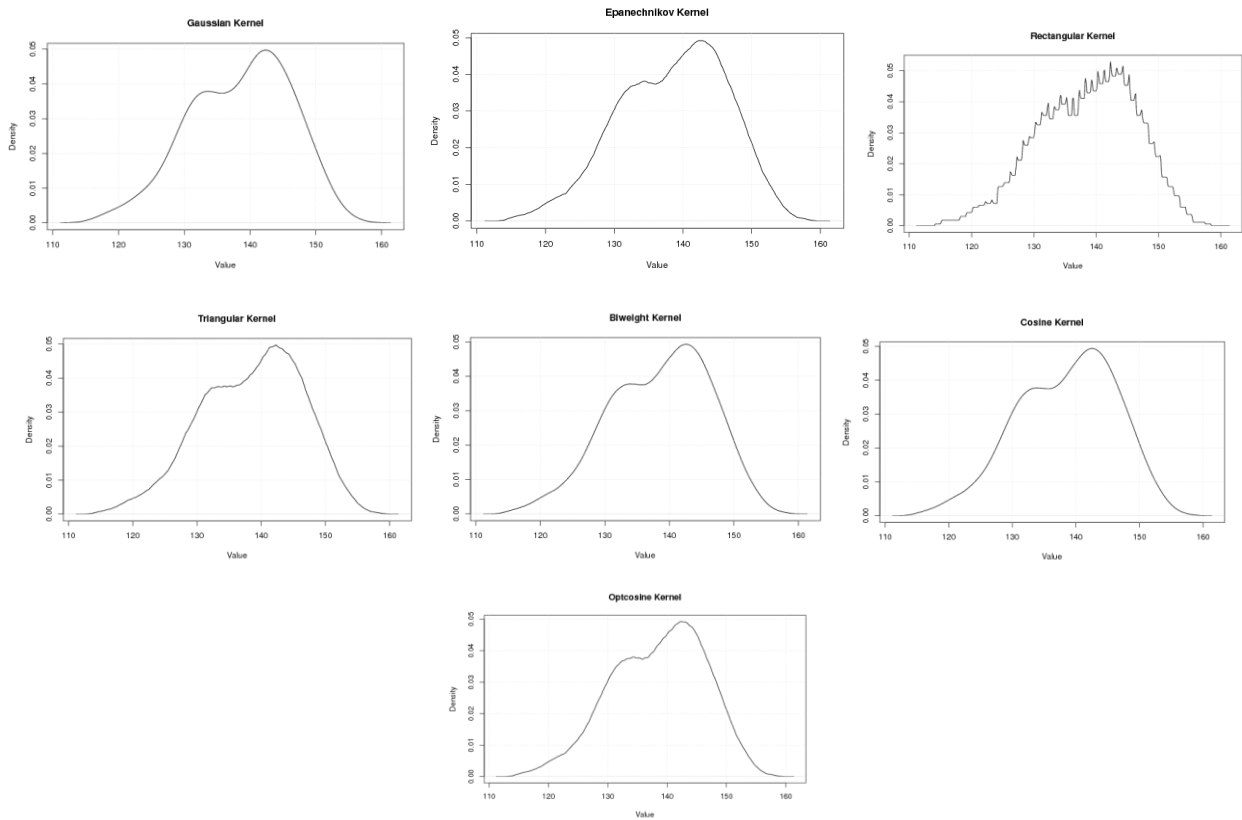
**Figure A.3.** Kernel density estimation plots for the third eigenvalue of Human according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



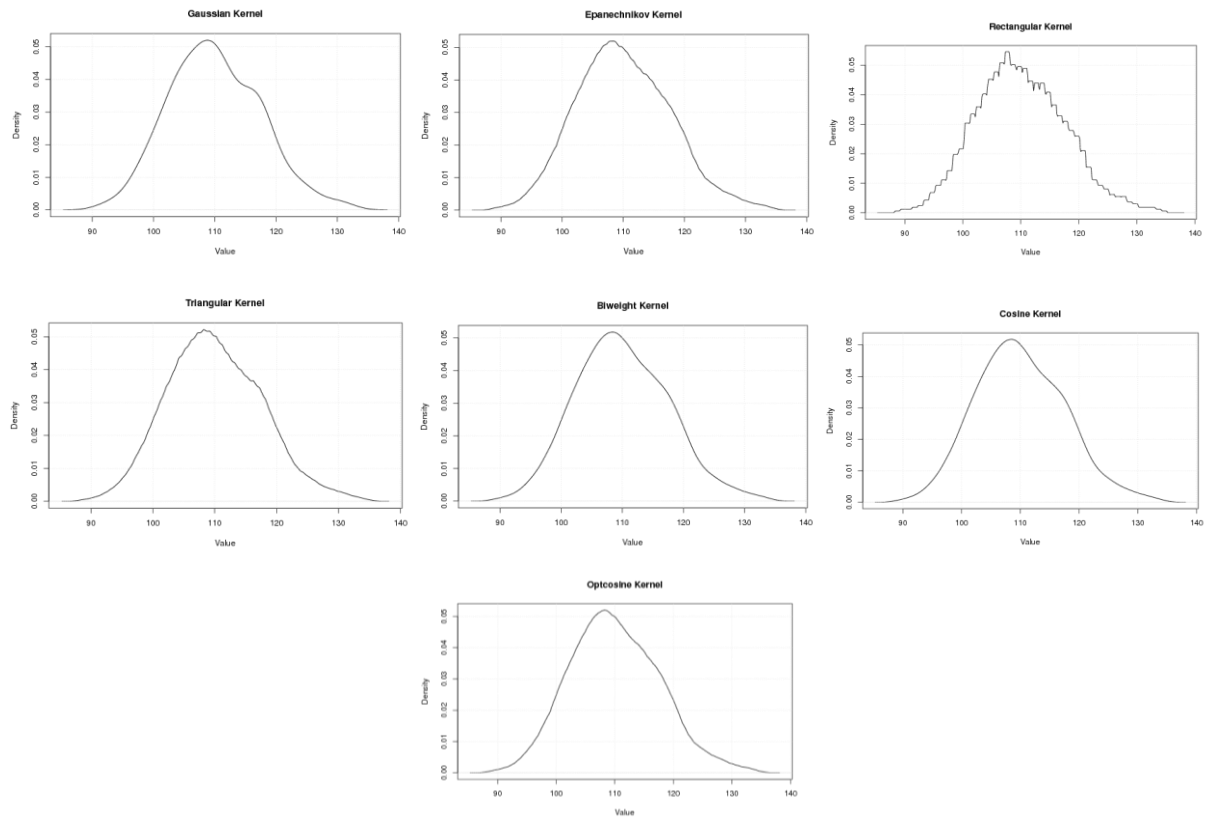
**Figure A.4.** Kernel density estimation plots for the fourth eigenvalue of Human according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



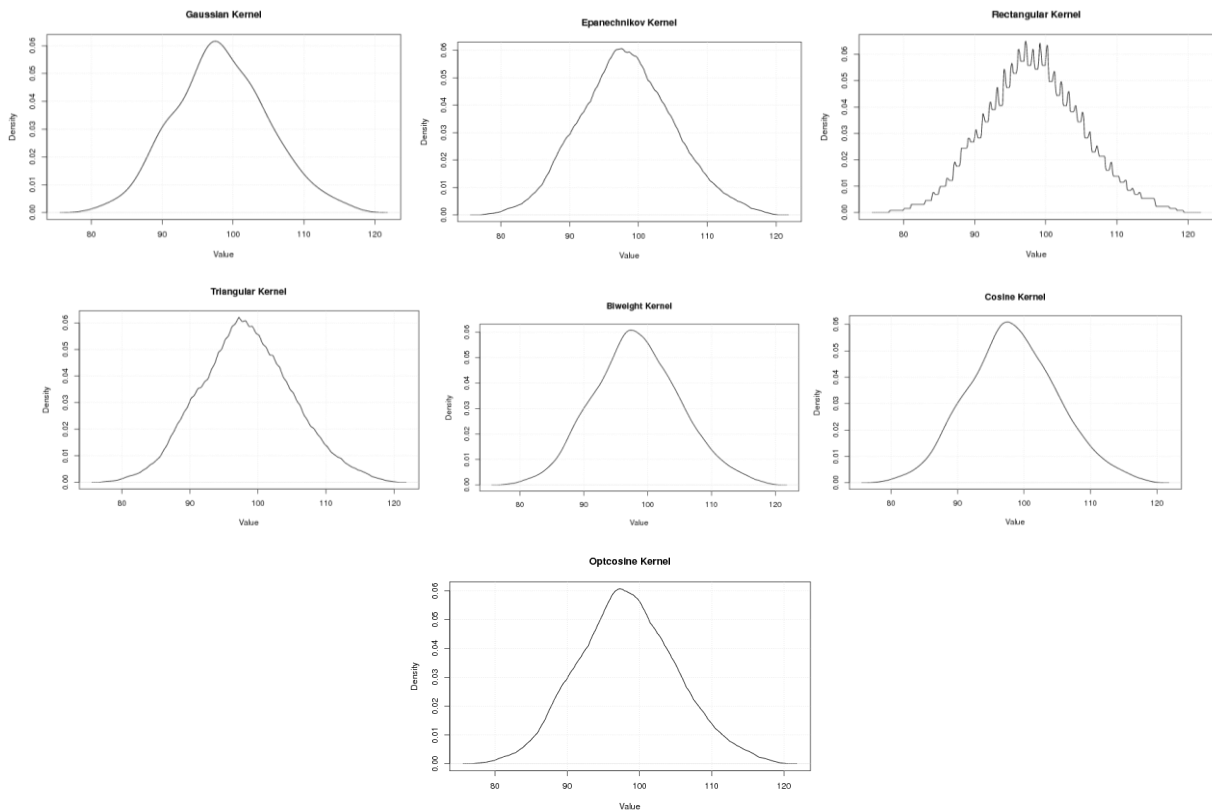
**Figure A.5.** Kernel density estimation plots for the first eigenvalue of Grasshopper according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



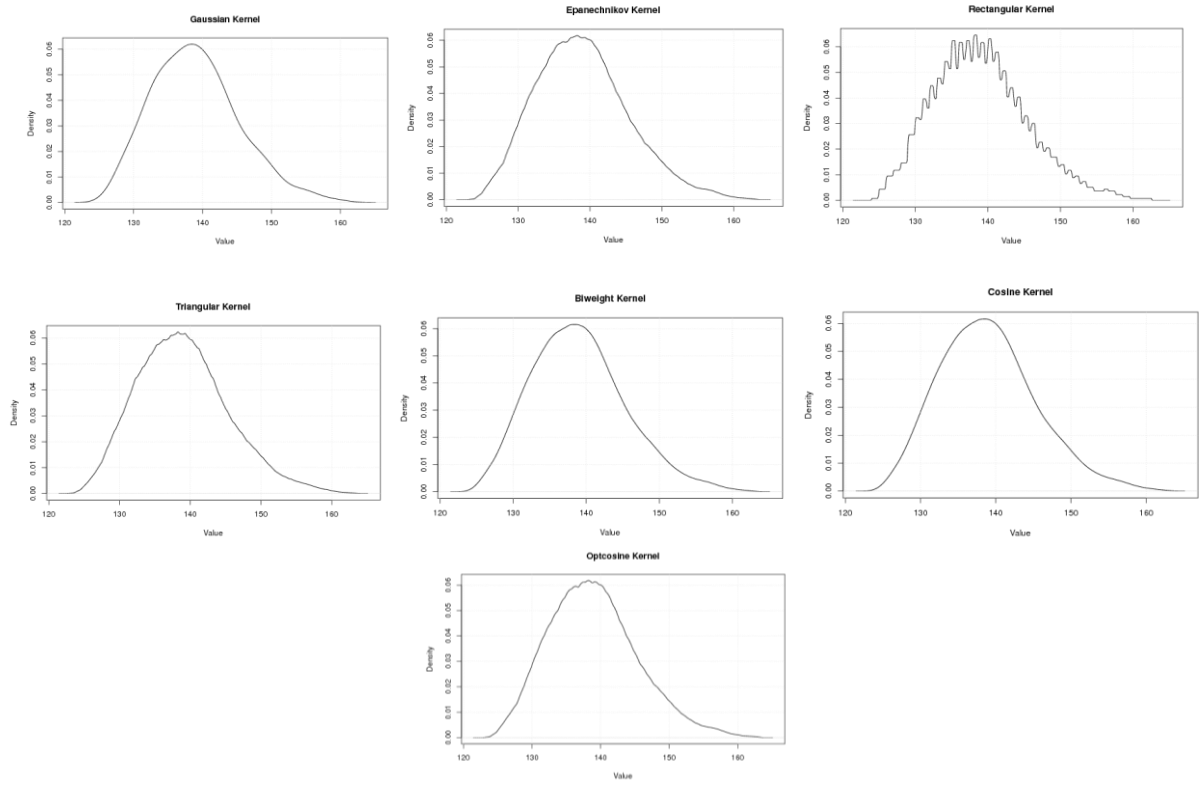
**Figure A.6.** Kernel density estimation plots for the second eigenvalue of Grasshopper according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



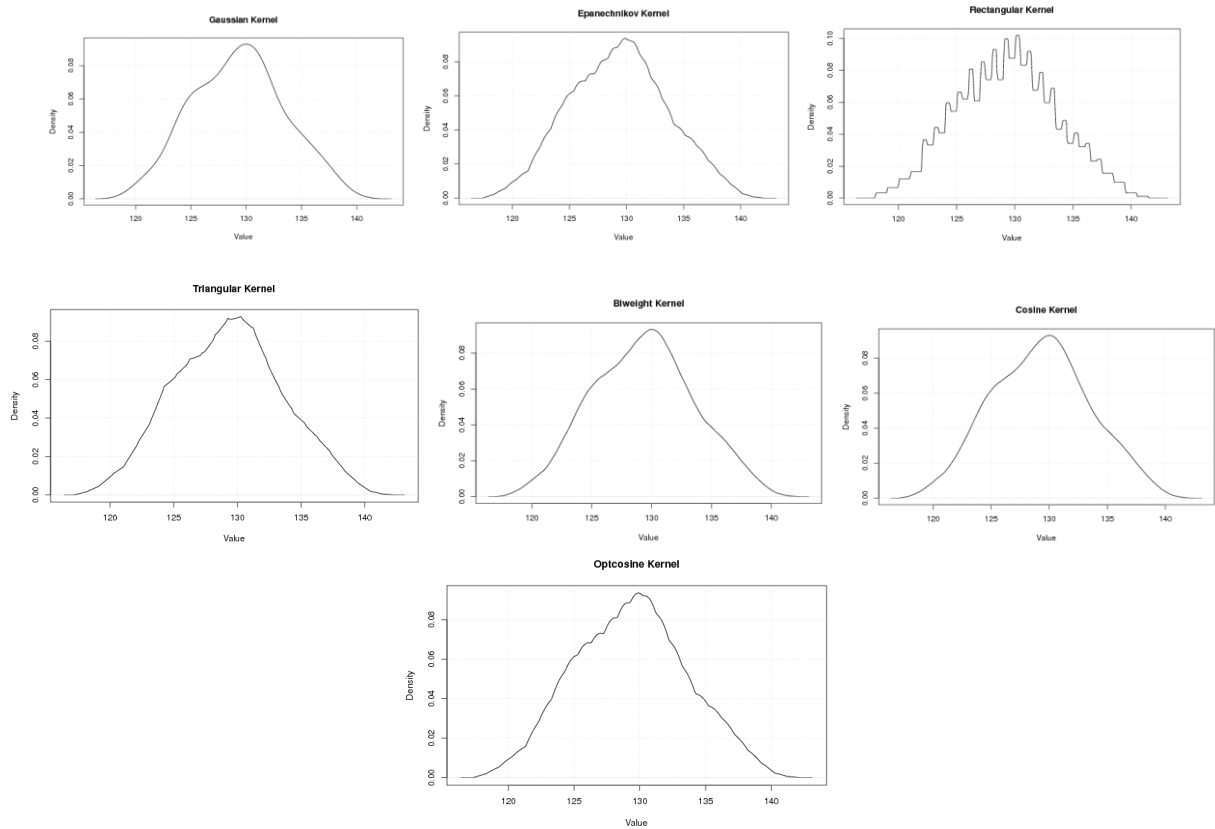
**Figure A.7.** Kernel density estimation plots for the third eigenvalue of Grasshopper according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



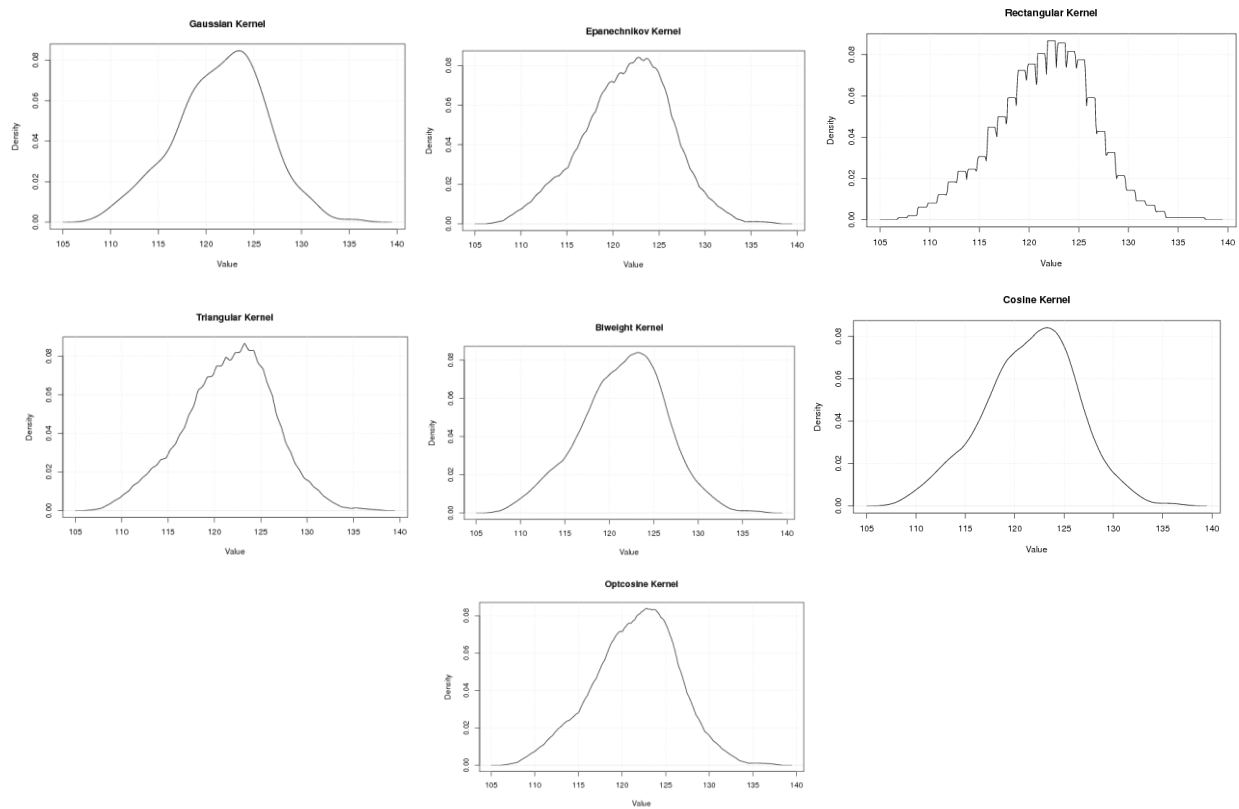
**Figure A.8.** Kernel density estimation plots for the fourth eigenvalue of Grasshopper according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



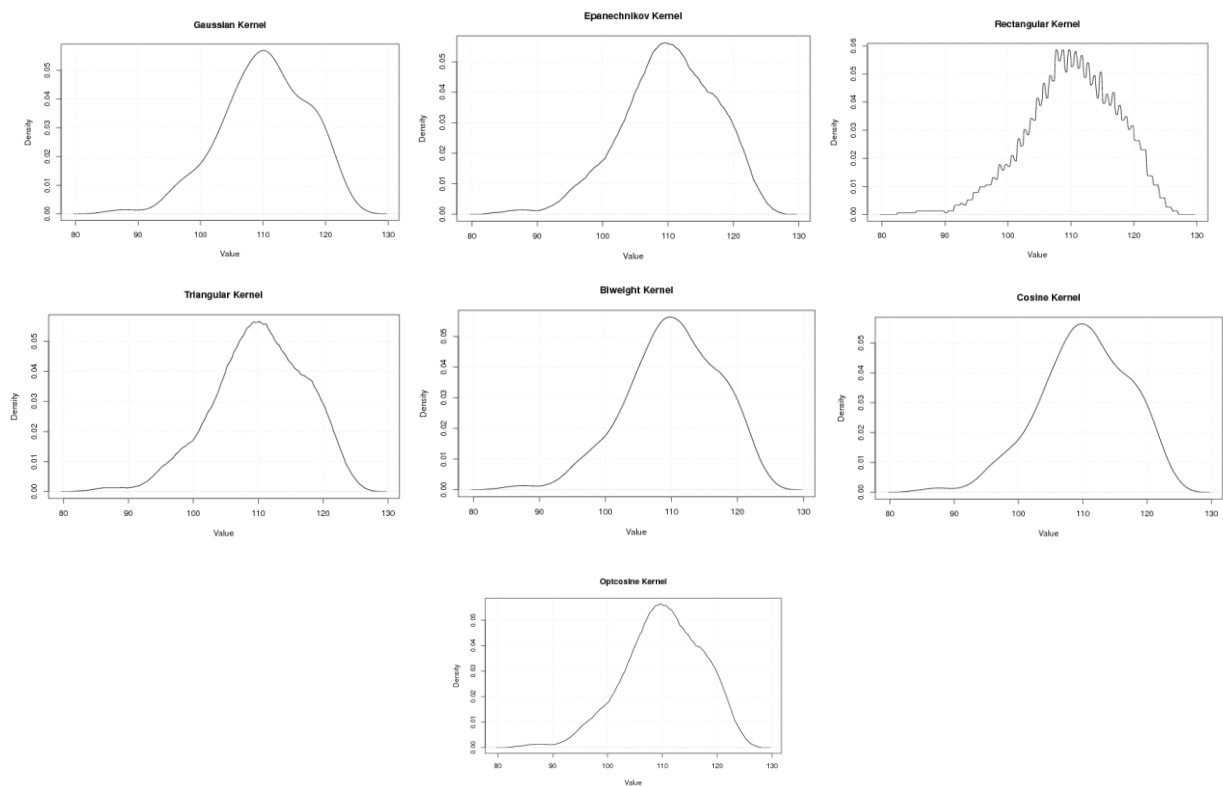
**Figure A.9.** Kernel density estimation plots for the first eigenvalue of E-Coli according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



**Figure A.10.** Kernel density estimation plots for the second eigenvalue of E-Coli according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels

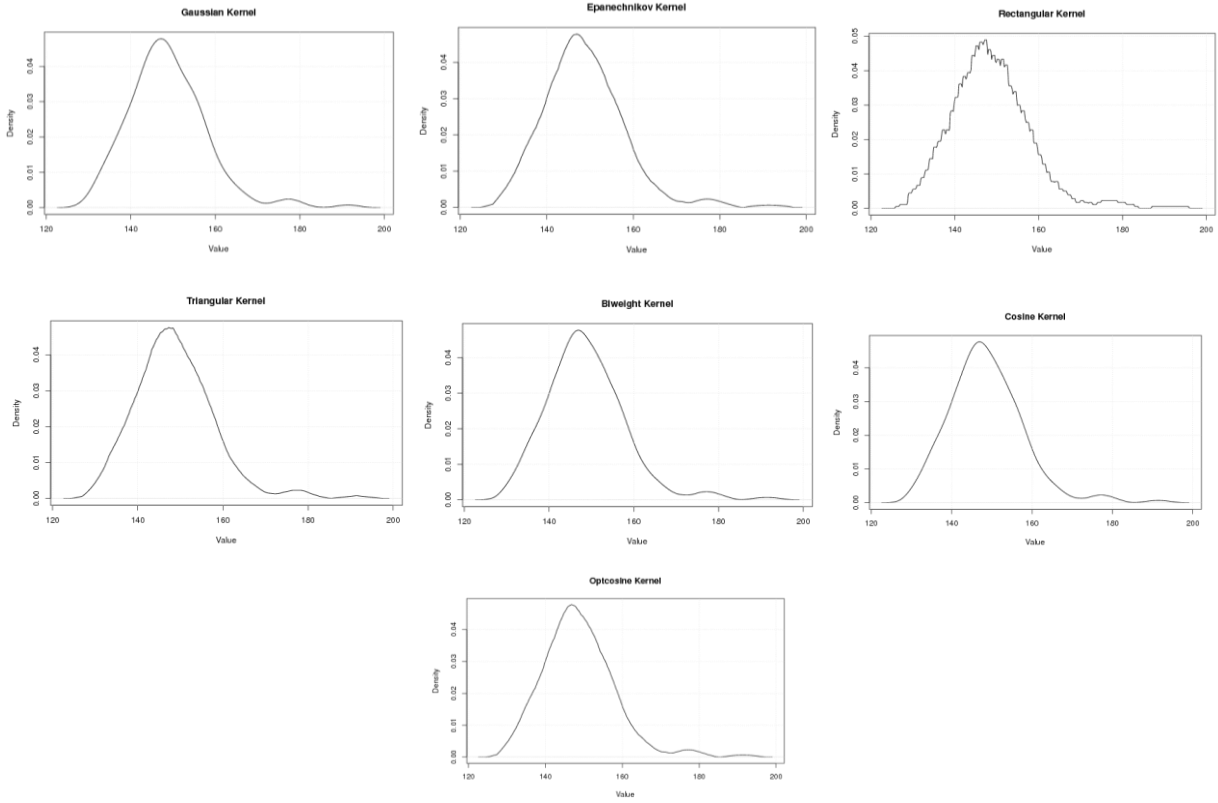


**Figure A.11.** Kernel density estimation plots for the third eigenvalue of E-Coli according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels

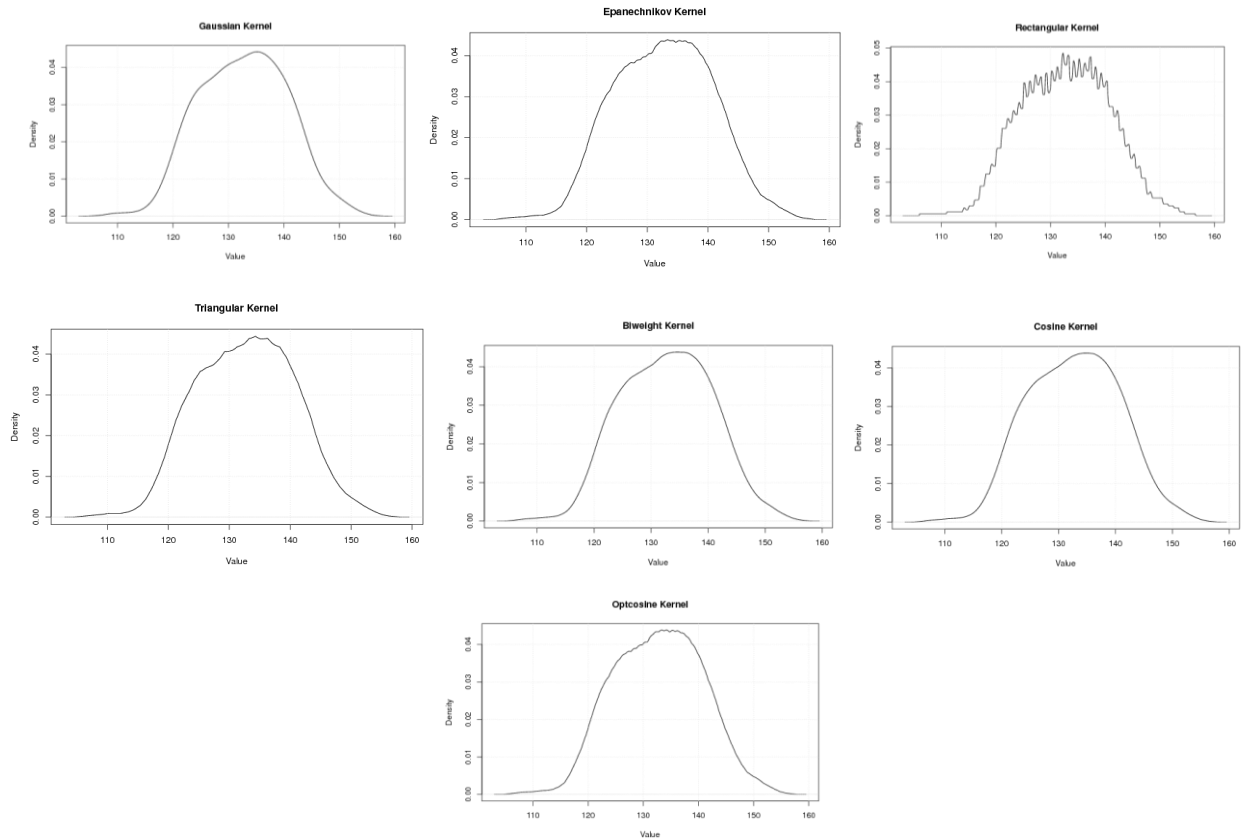


**Figure A.12.** Kernel density estimation plots for the fourth eigenvalue of E-Coli according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels

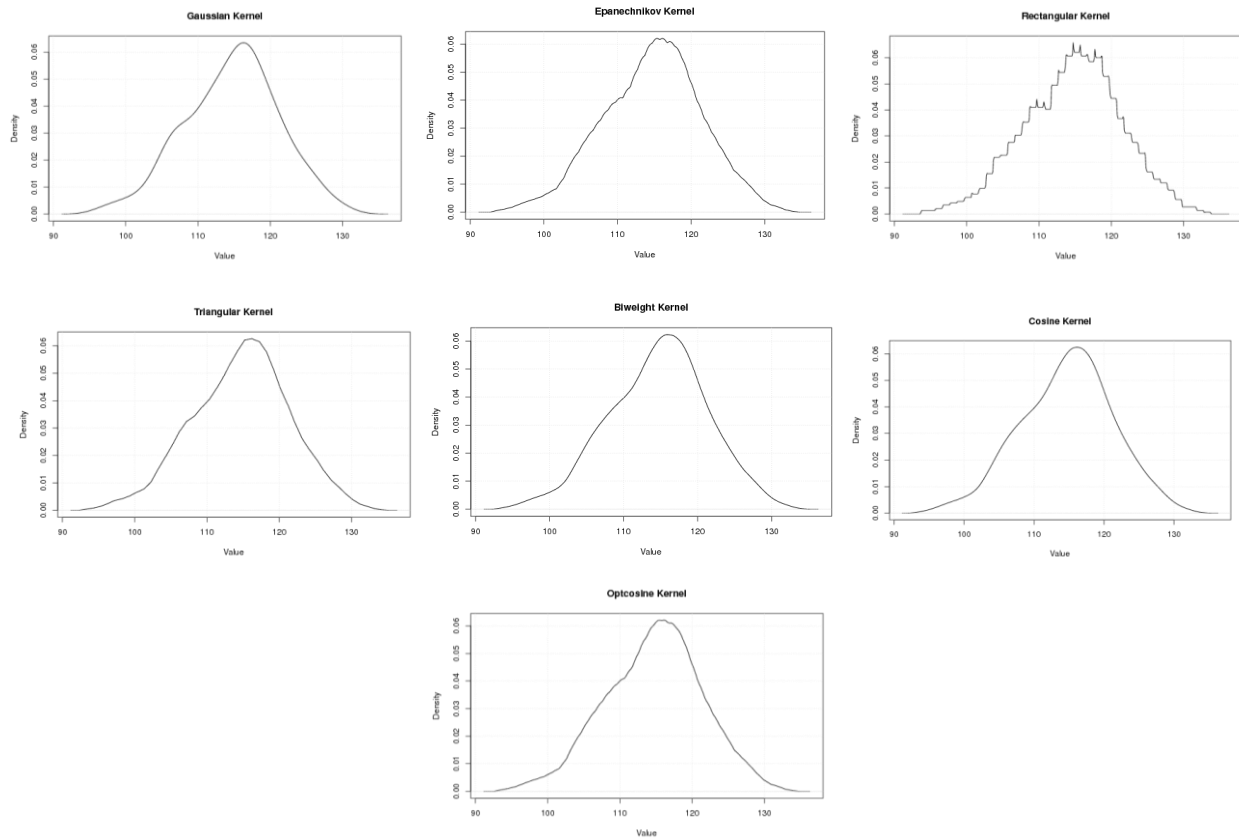




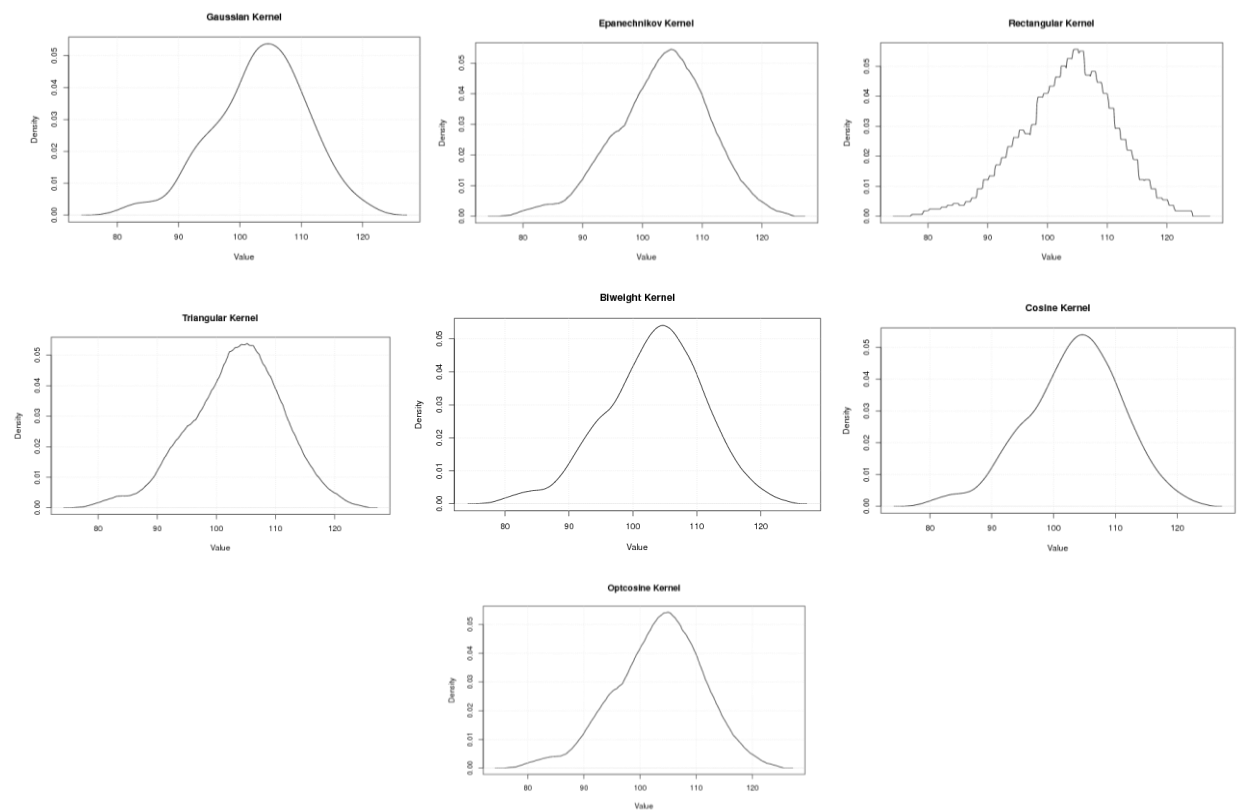
**Figure A.13.** Kernel density estimation plots for the first eigenvalue of Rat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



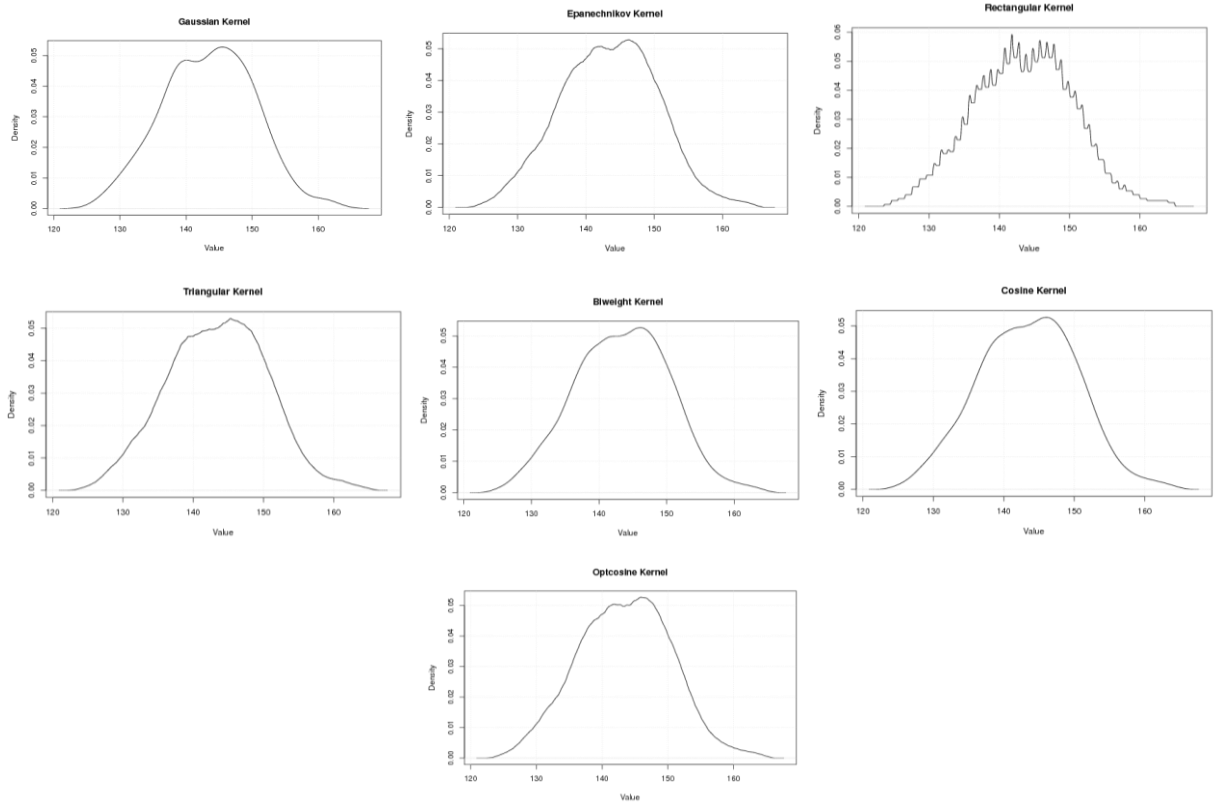
**Figure A.14.** Kernel density estimation plots for the second eigenvalue of Rat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



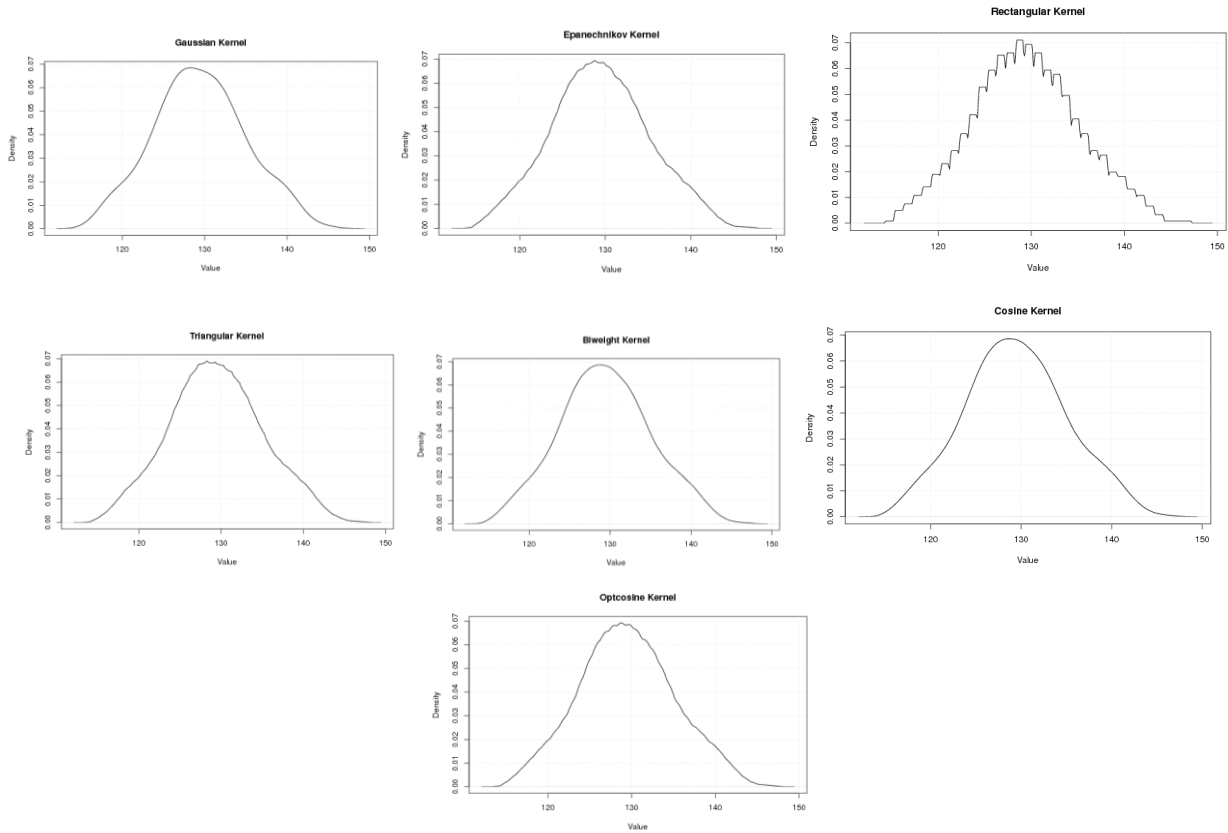
**Figure A.15.** Kernel density estimation plots for the third eigenvalue of Rat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



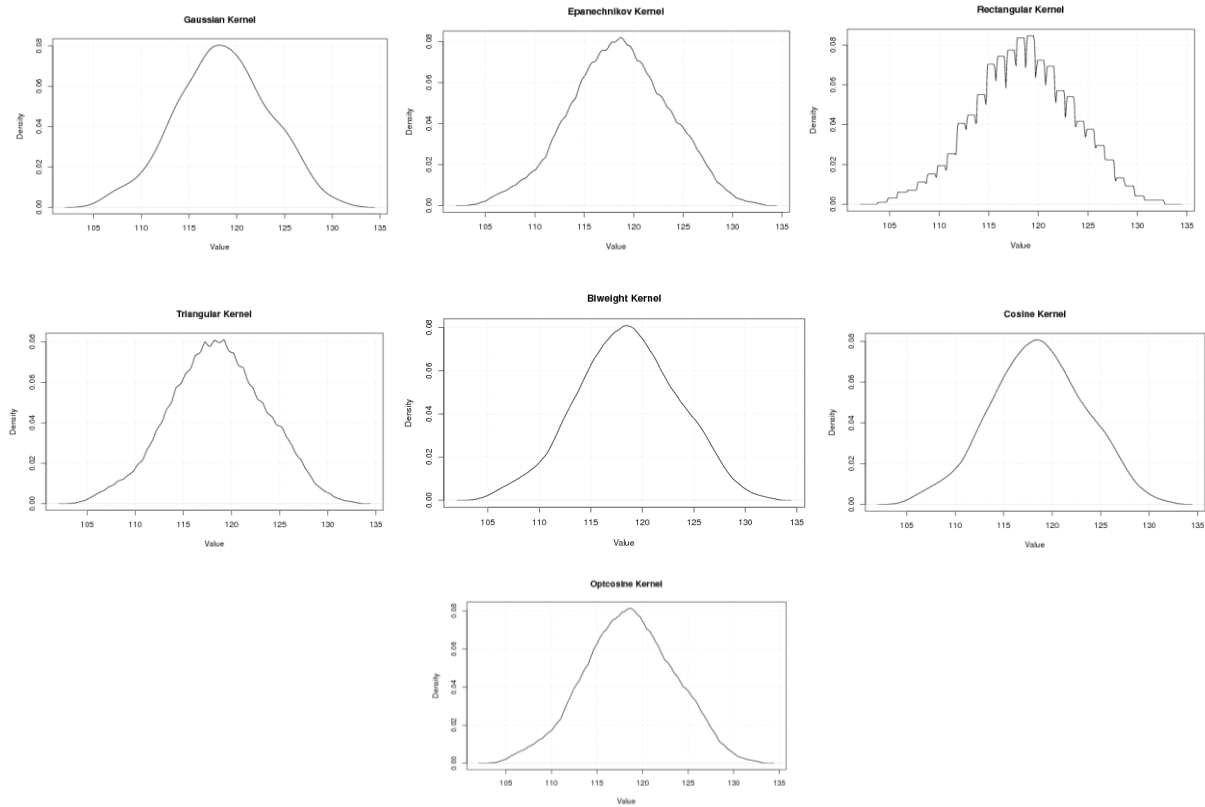
**Figure A.16.** Kernel density estimation plots for the fourth eigenvalue of Rat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



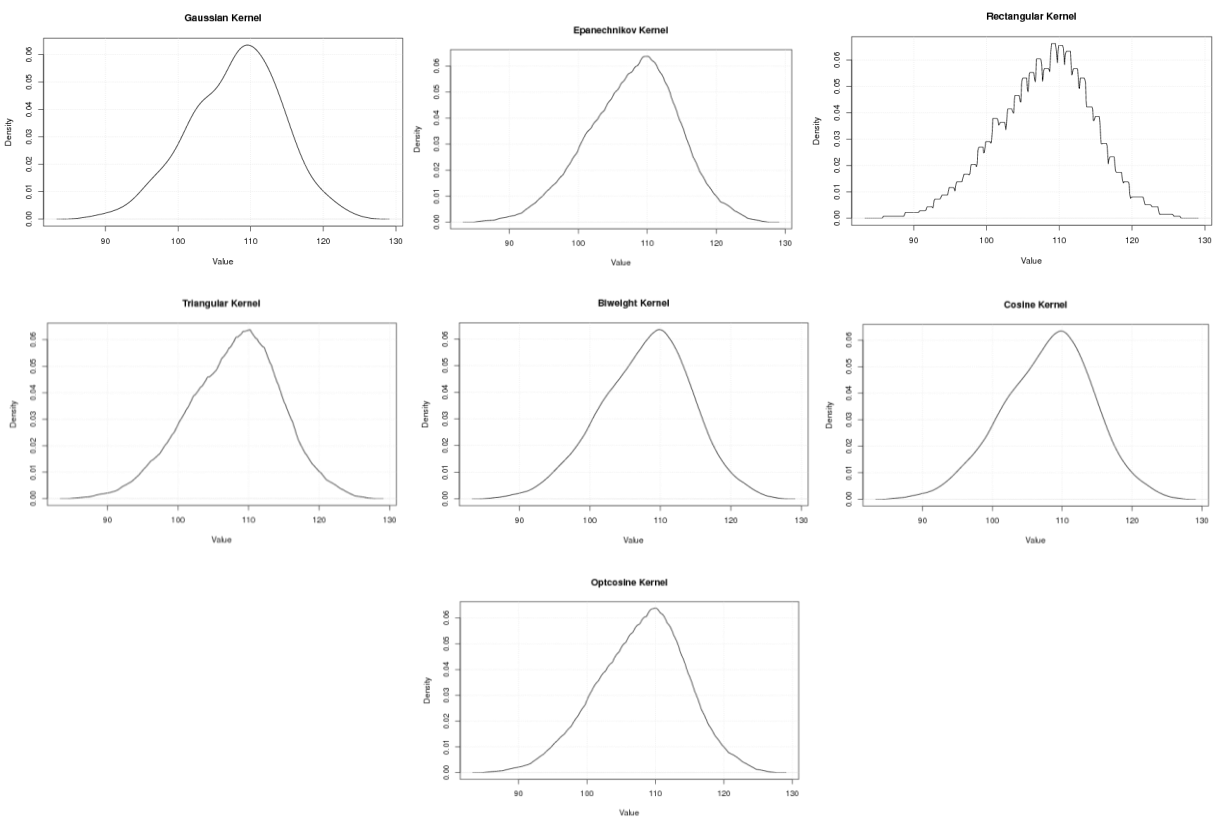
**Figure A.17.** Kernel density estimation plots for the first eigenvalue of Wheat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



**Figure A.18.** Kernel density estimation plots for the second eigenvalue of Wheat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



**Figure A.19.** Kernel density estimation plots for the third eigenvalue of Wheat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels



**Figure A.20.** Kernel density estimation plots for the fourth eigenvalue of Wheat according to Gaussian, Epanechnikov, Rectangular, Triangular, Biweight, Cosine, and Optcosine kernels

## REFERENCES

- [1] Andre´s Ferreyra, R., Podesta´, G.P., Messina, C.D., Letson, D., Dardanelli, J., Guevara, E., et al. (2001) "A linked-Modeling Framework to Estimate Maize Production Risk Associated with Enso-related Climate Variability in Argentina", *Agricultural and Forest Meteorology*, Vol. 107, No. 3, PP. 177–192.
- [2] Bae, W. and Kim, C. (2008) "A simple Segmentation Method for DNA Microarray Spots by Kernel Density Estimation", 30, PP. 223–234, DOI 10.1007/s00291-007-0091-6.
- [3] Bajic, V., Bajic, I. and Hide, W. (2000) "A new Method of Spectral Analysis of DNA/RNA and Protein sequences" Centre for Engineering Research.
- [4] Bilock, A., Jidling, C. and Rydin, Y. (2016) "Modelling Bivariate Distributions Using Kernel Density Estimation", Project in Computational Science, Department of information technology.
- [5] Bosman, P.A. and Thierens, D. (2000) "Ideas Based on the Normal Kernels Probability Density Function", Technical Report 11, Department of Computer Science, Utrecht University, The Netherlands.
- [6] Bowman, A. (1984) "An alternative method of cross-validation for the smoothing of density estimates", *Biometrika*, Vol.71, PP. 353–360.
- [7] Bowman, A.W. and Azzalini, A. (1997) "Applied Smoothing Techniques for Data Analysis", Oxford University Press, Oxford, UK.
- [8] Colbrook, M., Botev, Z.I., Kuritz, K. and MacNamara, S. (2018) "Kernel Density Estimation with Linked Boundary Conditions", arXiv:1809.07735v1 [math.ST] 20, PP. 1-50.
- [9] Corti, S., Molteni, F. and Palmer, T. (1999) "Signature of Recent Climate Change in Frequencies of Natural Atmospheric Circulation Regimes", *Nature*, Vol. 398, No. 6730, PP. 799–802.
- [10] Deheuvels, P. (1977) "Estimation Nonparamétrique de la Densité Par Histogrammes Generalizes", *Rev. Statist. Appl.*, Vol. 25, PP. 5–42.
- [11] Doung, T. and Hazelton, M. (2003) "Plug-In Bandwidth Matrices for Bivariate Kernel Density Estimation, Nonparametric Statistics".
- [12] Eisenberg, D., Weiss, R.M., Terwillger, T.C., (1994) "The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity". *Proc. Natl. Acad. Sci.*, Vol. 81, PP. 140–144.
- [13] Fan, J., Feng, Y. and Niu, Y.S.(2010) "Nonparametric Estimation of Genewise Variance for Microarray Data", *NIH Public Access*, Vol. 38, No. 5, PP. 2723–2750, doi:10.1214/10-AOS802.
- [14] Faraway, J. and Jhun, M. (1990) "Bootstrap choice of bandwidth for density estimation", *J. Amer. Statist. Assoc.*, Vol. 85, PP.1119-1122.
- [15] Galleani, L. and Garello, R. (2006) "Spectral Analysis of DNA Sequences by Entropy Minimization", 14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September, PP. 4-8.
- [16] Hall, P. (1982) "Comparison of two orthogonal series methods of estimating a density and its derivatives on interval", *Journal of Multivariate Analysis*, Vol. 12, No. 3, PP. 432–449.
- [17] Han, Y., Han, L., Yao, Y., Li, Y. and Liu, X. (2018) "Key Factors in FTIR Spectroscopic Analysis of DNA: The Sampling Technique, Pretreatment Temperature and Sample Concentration", *Analytical Methods*, Issue Vol. 21, No. 10, PP. 2436-2443.
- [18] Hoang, T., Yin, C., Zheng, H. Yu, C., Lucy He, R. and Yau, S. (2015) "A new Method to Cluster DNA Sequences Using Fourier Power Spectrum", *J Theor Biol.* 7; 372:135-45.
- [19] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996) "A brief Survey of Bandwidth Selection for Density Estimation", *Journal of the American Statistical Association*, Vol. 91, No. 433, PP. 401–407.
- [20] Li, Y., Wang, N., Hong, M., Turner, N.D., Lupton, J.R. and Carroll, R.J. (2007) "Nonparametric Estimation of Correlation Functions in Longitudinal and Spatial Data, with Application to Colon Carcinogenesis Experiments", *University of Georgia and Texas A&M University, The Annals of Statistics*, Vol. 35, No. 4, PP. 1608–1643.
- [21] Lopez-Novoa, U., Sa´enz, J., Mendiburu, A. and Miguel-Alonso, J. (2015) "An efficient implementation of kernel density estimation for multi-core and many-core architectures", *The International Journal of High Performance Computing Applications*, PP 1-17.
- [22] Mabrouk, M. (2017) "Advanced Genomic Signal Processing Methods in DNA Mapping Schemes for Gene Prediction Using Digital Filters", *American Journal of Signal Processing*, Vol. 7, No. 1, PP. 12-24.
- [23] Marhon, S. and Kremer, S. (2011) "Gene Prediction Based on DNA Spectral Analysis: A literature Review", *J Comput Biol.*, Apr, Vol. 18, No. 4, 639-76.
- [24] McLachlan, A. and Stewart, M. (1976) "The 14-fold Periodicity in Alpha-Tropomyosin and the Interaction with Actin", *J. Mol. Biol.*, Vol. 103, PP. 271–298.
- [25] Oja, H. (1981) "On Location, Scale, Skewness and Kurtosis of Univariate Distribution", *Scandinavian Journal of Statistics*, Vol. 8, No.2, PP. 164–168.
- [26] Polovinkina, A., Krylova, I., Druzhkova, P., Ivanchenko, M., Meyerova, I., Zaikina, A., and Zolotykh, N. (2016) "Solving Problems of Clustering and Classification of Cancer Diseases Based on DNA Methylation", *Data Pattern Recognition and Image Analysis*, Vol. 26, No. 1, PP. 176–180.
- [27] Roy, M. and Barman, S. (2011) "Spectral Analysis of Coding and Non-coding Regions of a DNA Sequence by Parametric and Nonparametric Methods: A comparative Approach", *Annals of Faculty Engineering Hunedoara– International Journal of Engineering*; Tome IX; Fascicule 3; PP. 57-62.
- [28] Ruiz, G., Israel, Godínez, I., Ramos, S., Ruiz, S., Pérez, H. and Morales, J. (2018) "Genomic Signal Processing for DNA Sequence Clustering" *PeerJ* v.6; DOI 10.7717/peerj.4264.
- [29] Scott, D. (1979) "On optimal and data-based histograms", *Biometrika*, Vol. 66, PP. 605–610.

- [30] Scott, D. and Terrell, G. (1987) "Biased and unbiased cross-validation in density estimation", *J. Amer. Statist. Assoc.*, Vol. 82, PP.1131–1146.
- [31] Sheather, S. and Jones, M. (1991) "A reliable Data-based Bandwidth Selection Method for Kernel Density Estimation", *J. Roy. Statist. Soc. Ser.B*, Vol. 53, PP. 683-690.
- [32] Sheather, S. (2004) "Density Estimation", *Statistical Science*, Vol. 19, No 4, PP. 588-597.
- [33] Silverman, B. (1986) "Density Estimation for Statistics and Data Analysis", Chapman and Hall, London.
- [34] Stoffer, D., Tyler, D. and McDougall, A. (1993) "Spectral Analysis for Categorical Time Series: Scaling and the Spectral Envelope", *Biometrika*, Vol. 80, PP. 611–622.
- [35] Stoffer, D. (2012) "Frequency Domain Techniques in the Analysis of DNA Sequences", *Handbook of Statistics Volume 30*, PP. 261-295.
- [36] Tavaré, S., Giddings, B. (1989) "Some Statistical Aspects of the Primary Structure of Nucleotide Sequences", In Waterman M.S. (Ed), *Mathematical Methods for DNA Sequences*. CRC Press, Boca Raton, Florida, PP. 117–131.
- [37] Troudi, M., Alimi, A.M. and Saoudi, S. (2008) "Analytical Plug-InMethod for Kernel Density Estimator Applied to Genetic Neutrality Study", Hindawi Publishing Corporation, *EURASIP Journal on Advances in Signal Processing*, Article ID 739082, 8 pages, doi :10.1155/2008/739082.
- [38] Viari, A., Soldano, H. and Ollivier, E. (1990) "A Scale-independent Signal Processing Method for Sequence Analysis. *Comput. Appl. Biosci.*, Vol. 6, PP. 71–80.
- [39] Waterman, M. and Vingron, M. (1994) "Sequence Comparison Significance and Poisson Approximation", *Stat. Sci.*, Vol. 9, PP. 367–381.
- [40] Weyenberg, G., Huggins, P.M., Schardl, C.L., Howe, D.K. and Yoshida, R. (2014) "Non parametric Estimation of Phylogenetic Tree Distributions", *Original Paper* , Vol. 30, No. 16, PP. 2280–2287, doi:10.1093/bioinformatics/btu258.
- [41] Wolodzko, T. (2017) "Smoothed Bootstrap and Random Generation from Kernel Densities", Package 'kernelboot', PP. 1-14.