https://doi.org/10.33472/AFJBS.6.1.2024.72-84

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Drichlet, Féjer and Modify Daniell Kernel to Estimate Human DNA Spectral Envelope

Ali Ghanim Abood[1], Prof. Tahir R. Dikheel[2]

[1,2]Administration & Economics, University of AL-Qadisiyah, Iraq
Email: alialmoaly88@gmail.com, tahir.dikheel@qu.edu.iq

Article Info

Volume 6, Issue 1, January 2024

Received: 17 November 2023

Accepted: 08 December 2023

Published: 05 January 2024

doi: 10.33472/AFJBS.6.1.2024.84-72

ABSTRACT

In this article, we modeled a non-stationary time series by dividing it into several segments, then analyzed it by estimating the spectral envelope density function for each segments separately to reduce the effect of the trend. Three kernel functions are used, which are: Dirichlet, Modify Daniell, Féjer. We generated data to test our algorithm and compare the kernel functions to choose the best one by calculating Signal-to-noise ratio (SNR), we conclude that Dirichlet is the most efficient. a human DNA real data is used to apply our algorithm.

Keywords: spectral envelope, Dirichlet kernel, Modify Daniell kernel, Fejér kernel, Kullback-Leibler divergence, Tree-Based Adaptive Segmentation

1. Introduction

Data processing is one of the most important scientific topics, because it is directly related to all types of science and experiments. We need a science that specializes in the problems of data processing and analysis, and this science is the science of statistics. Statistics is a broad science and has many branches, but what concerns us is time series.

The time series is a collections of observations generated sequentially through time or any parameter. There are several types of time series, including continuous ones, in which measurements or readings can be taken at every moment of the series, but most series are discrete time series, which discrete time series consists of data points separated by certain time intervals. Time series can be divided according to the type of data into several types, such as: Binary time series, Count time series, Categorical time series. Categorical data is data that can be anything but numbers, such as letters, colors, symbols, etc.

The time series can be analyzed in two ways: the first method is within the time domain, and the second method is within the frequency domain, which is done in several ways, such as spectral analysis or wavelet analysis, or spectral envelope analysis method.

In spectral envelope analysis, we need weight functions for the purpose of obtaining accurate results, which is called the kernel function, and there are many weight functions that can be used, such as: The Beta kenel, Parzen kernel, Dirichlet kenel, etc.

In this research, we will analyze a sample of human DNA, which is considered categorical data. The analysis will be within the frequency domain of the spectral envelope of the time series, and in our analysis we will use the Dirichlet Kernel estimator, modify danielle Kernel estimator and The Fejér Kernel.

The goal of this research is modling the categorical time series by estimating the spectral envelope function in several methods and functions which are: Dirichlet Kernel estimator, modify danielle Kernel estimator and The Fejér Kernel, in order to compare these methods and find out the best method.

Spectral Envelope

The spectral envelope is a frequency-based, principal components technique applied to a multivariate time series. Now if we suppose that x_t, t=0,±1,±2,… is a stationary categorical time series, with finite state-space which is $c = \{c_1, c_2, \dots, c_k\}$ values, then we suppose the numeric value α_j for any c_j. So **α** is a vector of the real values, $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_k)'$, and $p_j = Pr(x_t = c_j) > 0$, and h(x_t) be stationary time series with the real value. Now we defined Y_t as the flowing:

$$Y_t \begin{cases} z_j \; if \; x_t = c_j \; for \; j = 1,2,3\dots,k-1 \\ \qquad\qquad o \; if \; x_t = c_j \end{cases}$$

When z_j is a vector of k items, all its items are zero except jth row are one. And o are k×1 vector, all its items are zero. Then we come collusion $h(x_t) = \boldsymbol{\alpha}' Y_t$, all so $h(x_t) = \alpha_j$

The probability space of the DNA is {A, C, G, T}, so we can design Y_t as following:

$$Y_t = (1,0,0)', when \, X_t = A$$
$$Y_t = (0,1,0)', when \, X_t = C$$
$$Y_t = (0,0,1)', when \, X_t = G$$
$$Y_t = (0,0,0)', when \, X_t = T$$

So the goal is chose the best value to α so that maximize the power at each frequency ω as follow:

$$\lambda(\omega) = \max_{\alpha} \frac{f_{xx}(\omega:\alpha)}{\sigma^2(\alpha)}$$

Where

$\lambda(\omega)$ is the power of the frequency, $f_{xx}(\omega:\alpha)$ is the spectral density and $\sigma^2(\alpha) = var\{h(x_t)\}$.

suppose the vector process y_t has a continuous spectral density denoted by $f_{yy}(\omega)$ and for each ω there are $k \times k$ complex-valued Hermitian matrix. as we have $h(x_t) = \boldsymbol{\alpha}' Y_t$ suggest $f_{xx}(\omega:\alpha) = \boldsymbol{\alpha}' f_{yy}(\omega)\boldsymbol{\alpha}$. Now, if $f_{yy}^{re}(\omega)$ are the real part of $f_{yy}(\omega)$ and $f_{yy}^{im}(\omega)$ the imaginary part of $f_{yy}(\omega)$, and as $f_{yy}^{im}(\omega)$ is skew-symmetric, so $f_{yy}^{im}(\omega)' = -f_{yy}^{im}(\omega)$ and $x = x^{re} + ix^{im}$, so $\boldsymbol{\alpha}' f_{yy}(\omega)\boldsymbol{\alpha} = \boldsymbol{\alpha}' f_{yy}^{re}(\omega)\boldsymbol{\alpha}$

$$\lambda(\omega) = \frac{\boldsymbol{\alpha}' f_{yy}^{re}(\omega)\boldsymbol{\alpha}}{\boldsymbol{\alpha}' \boldsymbol{V}\boldsymbol{\alpha}}$$

where $\boldsymbol{V}$ is the variance–covariance matrix of y_t. Where $p = (p_1, p_2, \dots, p_k)'$, $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{pp}'$, and $\boldsymbol{D}$ IS K × K diagonal matrix $\boldsymbol{D} = diag\{p_1, p_2, \dots, p_k\}$. By assumption, $p_j > 0,\ j = 1, \dots, k$; so the rank$(V) = k - 1$ with the null space of being spanned by I_k, for any $k \times (k-1)$ full rank matrix $\boldsymbol{Q}$ whose columns are linearly independent of I_K. And $\acute{\boldsymbol{Q}}\boldsymbol{V}\boldsymbol{Q}$ is a $(k-1) \times (k-1)$ positive symmetric matrix.

If $\mathrm{f}_{yy}(\omega)$ is a consistent each $\mathrm{j} = 1, \dots, \mathrm{J}$, the largest root of $\mathrm{f}_{yy}^{re}(\omega)$ is distinct, then

$$\left\{\frac{\eta_n[\hat{\lambda}(\omega_j) - \lambda(\omega_j)]}{\lambda(\omega_j)},\ \ \eta_n[\hat{\propto}(\omega_j) - \propto(\omega_j)]; j = 1, \dots, J\right\} \quad (1)$$

converges jointly in distribution to independent zero-mean, normal distribution as $n \to \infty$

the value of η_n in the equation (1) depends on the type of the estimator. In our study, the smoothed periodogram matrix

$$I_n(\omega_j) = \hat{f}_{xx} = \sum_{l=-m}^{m} h_l I_n(\omega_j + l/n)$$

If we use the smoothed periodogram matrix with weight h_l then $\eta_n{}^{-2} = \sum_{l=-m}^{m} h_l{}^2$

Peak searching for the smoothed spectral envelope estimate can be aided using the following approximations. Using a first-order Taylor expansion, we have

$$\log \hat{\lambda}(\omega) \approx \log \lambda(\omega) + \frac{\hat{\lambda}(\omega) + \lambda(\omega)}{\lambda(\omega)}$$

Hence $E\left[\log \hat{\lambda}(\omega)\right] \approx \log \lambda(\omega)$ and $\mathrm{var}\left[\log \hat{\lambda}(\omega)\right] \approx \eta_n{}^{-2}$ because $\eta_n\left(\log \hat{\lambda}(\omega) - \log \lambda(\omega)\right)$ is standard normal [1].

Kernel Estimator Methods

In nonparametric statistics, a kernel is a weighting function used to estimate the spectral density and spectral envelope density where they are known as window functions [2].

The Dirichlet kernel estimator is a general stat of the beta kernel estimator. it is an asymmetric kernel Which is basically derived from dirichlet distribution [4].

The dirichlet kernel function defined as follow:

$$K_{D,n}(\theta) = \frac{1}{2\pi}\left(1 + 2\sum_{k=1}^{n} cos(k\theta)\right)$$

$$= \frac{1}{2\pi}\frac{sin(n + 0.5)\theta}{sin(\omega/2)}$$

As for Fejér kernel, its function is given as:

$$K_T(\omega) = \begin{cases} \frac{T+1}{2\pi} & \omega = 0 \\ \frac{1}{2\pi(T+1)}\left(\frac{\sin\left(\frac{T+1}{2}\omega\right)}{\sin\frac{\omega}{2}}\right)^2 & otherwise \end{cases}$$

And the properties of this function is:

1- $K_T(\omega) \geq 0\ \ and\ K_T(\omega) = K_T(-\omega)$ for all ω.

2- $K_T(\omega)$ is periodic with period 2π.

3- $\int_{-\infty}^{\infty} K_T(\omega)d\omega = 1$

4- for any $\theta \in (0, \pi)$ and for all $\theta \leq |\omega| \leq \pi$, we have [5]

$$|K_T(\omega)| \leq \frac{1}{2\pi(N + 1)sin^2\frac{\theta}{2}}$$

Modify Daniell is considered an development of Daniell kernel, Modify Daniell kernel smooths our model and reduces some of the variability that we saw in the sample spectrum. If we make the smoothing window wider, we will reduce the variability even further and it will make the forecasting difficult, but otherwise we will make the bias bigger. The abrupt change at the end points of the Daniell window could be softened by making the weights decrease at the extreme [6].

This function can be given as following:

$$\mathrm{w}_T(K) = \frac{1}{2T+1}\quad for - T \leq k \leq T$$

Stationarity in Time Series

One of the most important things that we must take into consideration is the stationarity of the time series, meaning that the series should not contain a trend., the trend can be defined "any systematical change in the level of a time series", because calculating the mean, variance, and autocorrelation in non-stationary time series will be complicated, and will cause inaccurate analysis of the time series. One of the ways that we can reduce the effect of the trend is to divide the series into a group of segments, and analyze each segment separately, because this will reduce the effect of the trend. Where we will divide the time series by: Tree-Based Adaptive Segmentation as the following:

1- The series is divided by two segmentations which is the level one, then each segmentation (Which are called blocks) are divided by two in level two till k-times in level k which is the deeper level. If T is the length of the entire series, then length of each block are $T/2^k$

2- We will denote the block $B(k, l)$, l is the l-th block in k level, and N_k is the length of blocks in k level.

3- Estimate the distance between two blocks. Let $D(k,l)$ be between two adjacent blocks, $B(K+1{,}2l)$ and $B(K+1{,}2l-1)$ Compute the estimates of the distances $D(k,l)$

4- for $k-1,\dots,0$ and $l=1,\dots,2^k$.

If $D(k,l) \leq D(K+1{,}2l-1)+D\ (K+1{,}2l)$ then mark the block $B(k,l)$ Otherwise, leave the block $B(k,l)$

5- The Final Segmentation determines the Segmentation depth. The final segmentation will be set of the highest marked blocks which is marked and its ancestor blocks are not marked [7].

LOCAL SPECTRAL ENVELOPE

Let's assume $k \times 1$ a vector value of pricewise stationary process, $\{Y_{s,T}\}_{s=0}^{T-1}$ for $T \geq 1$ where T is the length of series, is given as:

$$Y_{s,t} = \sum_{b=1}^{B} Y_{s,b} I(\mathrm{s/T}, \mathrm{U}_b)$$

Where $Y_{s,b}$ are stationary processes with continuous $k \times k$ spectral matrices $f_{s,b}(\omega)$ of b block. And $\mathrm{U}_b = [\mathrm{u}_{b-1}\mathrm{u}_{b},) \subset [0{,}1)$ is the interval. And $I(\mathrm{s/T}, \mathrm{U}_b)$ is an indicator which be $equal\ to\ 1\ if\ s/b \in \mathrm{U}_b$, and 0 otherwise.

now let rescale time in each block, so:

$\{Y_{s,b}: s/t \in \mathrm{U}_b\} \rightarrow \{Y_{t,b}: t = 0,\dots,M_b - 1\}$

and the number of observations in segment b is M_b, and $\sum_{b=1}^{B} M_b = T$. This rescaling of time represents a simple time shift to the origin where $Y_{s,b} \rightarrow Y_{t,b}$ for $s/t \in \mathrm{U}_b$ with $t = s - \sum_{i=1}^{b-1} M_i$

We shall say that a categorical time series, $\{x_{s,T}\}$, on a finite state-space and with nonzero marginal, is pieceunise stationary if the corresponding $k \times 1$point process, $\{Y_{s,T}\}$, is piecewise stationary. To assure that more observations fall within each stationary segment (or block) upon sampling the process $x_{s,T}$, we assume that the lower bound, M, for the number of observations in each block, b, satisfies $M \rightarrow \infty$ as $T \rightarrow \infty$.

If $x_{s,T}$ is a piecewise stationary categorical time series, we define the local spectral envelope as follows. The local analogue of the optimality criterion

$$\lambda_b(\omega) \underset{\boldsymbol{\alpha} \propto I_k}{sub} = \frac{\boldsymbol{\alpha}' f_{yy}^{re}(\omega)\boldsymbol{\alpha}}{\boldsymbol{\alpha}' \boldsymbol{v}_b \boldsymbol{\alpha}} \quad /$$

for $b = 1,2,\dots,B$ where $\boldsymbol{v}_b$ is the variance-covariance matrix of $Y_{t,b}$ and $\lambda_b(\omega)$ the local spectral envelope and the corresponding eigenvector $\boldsymbol{\alpha}_b(\omega)$ to be the local optimal scaling of block b and frequency ω.

now we present some asymptotic $T \rightarrow \infty$ results for estimators of the local spectral envelope and the corresponding local scaling vectors.

Next Let $\boldsymbol{Q}$ whose columns are linearly independent of I_K., namely, $\boldsymbol{Q} = [I_{K-1}|0\acute{]}$, and let $\widehat{\boldsymbol{V}}_b$ be the sample variance-covariance matrix obtained from the data in segment $b, \{Y_{s,T}: s/T \in U_b\}$, or equivalently,$\{Y_{s,b}: t = 0,1,\dots,M_b - 1\}$. So

$$Y_{s,b} \overset{def}{=} \acute{\boldsymbol{Q}} Y_{s,b}$$

this operation has the effect of removing the k-th element from$Y_{t,b}$, so that it is now a $(k-1) \times 1$ vector. In this case, we denote:

$$\widehat{\boldsymbol{V}}_b \overset{def}{=} \acute{\boldsymbol{Q}}\widehat{\boldsymbol{V}}_b\boldsymbol{Q} \quad and \quad \hat{f}_{Y,b}(\omega) \overset{def}{=} \acute{\boldsymbol{Q}}\hat{f}_{Y,b}(\omega)\boldsymbol{Q}$$

note that $\widehat{\boldsymbol{V}}_b$, and $\hat{f}_{Y,b}(\omega)$ are now the upper $(k-1) \times (k-1)$ blocks of the previously defined $\widehat{\boldsymbol{V}}_b$ and $\hat{f}_{Y,b}(\omega)$ matrices, respectively. In addition, we will use the same convention for the population values $\widehat{\boldsymbol{V}}_b$, and $\hat{f}_{Y,b}(\omega)$.

For simplicity and without loss of generality, we define the local sample spectral envelope, $\hat{\lambda}_b(\omega)$, to be the largest eigenvalue of $\hat{g}_b^{re}(\omega)$ where:

$$\hat{g}_b = \widehat{\boldsymbol{V}}_b^{-1\backslash 2}\hat{f}_{Y,b}\widehat{\boldsymbol{V}}_b^{-1\backslash 2}$$

The local sample optimal scaling, $\widehat{\boldsymbol{\alpha}}_b(\omega)$, is then defined by $\widehat{\boldsymbol{\alpha}}_b(\omega) = \widehat{\boldsymbol{V}}_b^{-1\backslash 2}\hat{u}_b(\omega)$, where $\hat{u}_b(\omega)$ is the eigenvector of $\hat{g}_b^{re}(\omega)$ associated with the root hat $\hat{\lambda}_b(\omega)$ The scale corresponding to the k-th category is held fixed at zero. Furthermore, let $\hat{u}_b(\omega)$ be normalized so $\acute{\hat{u}}_b\hat{u}_b(\omega) = 1$, and with the first nonzero entry of $\hat{u}_b(\omega)$taken to be positive.

To allow for the application of a general theory in obtaining asymptotic distributions for the estimates of the local spectral density $f_{Y,b}(\omega)$, we assume throughout this section that $Y_{t,b}$is strictly stationary for each block b, and that all local cumulant spectra, of all orders, exist for each series $Y_{t,b}$ The assumption of the existence of all local cumulant spectra is not restrictive in the categorical case because the elements of $Y_{t,b}$ take on only two values, zero or one. Rather than introduce excessive notation [8].

The local periodogram of the data $\{Y_{s,T} : s/T \in U_b\}$ in black b is given by:

$$I_b(\omega) = d_b(\omega)d_b^*(\omega)$$

Where

$$d_b(\omega) = M_b^{-1\backslash 2} \sum_{t=0}^{M_b - 1} Y_{t,b} e^{-2\pi i \omega t}$$

Is the finite fourier transform of the data $\{Y_{s,T} : s/T \in U_b\}$.

Where

$$\hat{f}_{Y,b} = (2m+1)^{-1} \sum_{i=-m}^{m} I_b(\omega + i/M_b)$$

Now the algorithm and calculations can be simplified by merging any two adjacent blocks if they have similar behavior in the spectral envelope, and the similarity in behavior can be measured using the algorithm Kullback-Leibler divergence as the follow:

$$I\big(p(X), q(x)\big) = \sum \left(\log \frac{p(x)}{q(x)}\right) p(x) \geq 0$$

Where p(x) and q(x) denote the probability density functions of random variable x.

To apply the algorithm Kullback-Leibler divergence on local spectral envelope we suppose $\hat{\lambda}_{K+1,2l}(\omega_j)$, $\hat{\lambda}_{K+1,2l-1}(\omega_j)$ are local sample spectral envelope at the frequency ω_j for the blocks $B(K+1,2l)$ and $B(K+1,2l-1)$. The we can find the Kullback-Leibler divergence between the two block as:

$$D(K,l)=\frac{1}{M_K/2+1}\sum_{j=0}^{M_K/2+1}\hat{\lambda}_{K+1,2l}(\omega_j)\log\frac{\hat{\lambda}_{K+1,2l}(\omega_j)}{\hat{\lambda}_{K+1,2l-1}(\omega_j)}$$

As Kullback-Leibler divergence are not symmetric, then we should us a symmetrised divergence, which defined as:

$$I(p,q)=\frac{1}{n}\sum_{i=1}^{n}\left(p_i\log\frac{p_i}{q_i}+q_i\log\frac{q_i}{p_i}\right)=\frac{1}{n}\sum_{i=1}^{n}\left([p_i-q_i]\log\frac{p_i}{q_i}\right)$$

So by the tow equations we get [7]:

$$D(K,l)=\frac{1}{M_K/2+1}\sum_{j=0}^{M_K/2+1}\left\{\left(\hat{\lambda}_{K+1,2l}(\omega_j)-\hat{\lambda}_{K+1,2l-1}\right)\log\frac{\hat{\lambda}_{K+1,2l}(\omega_j)}{\hat{\lambda}_{K+1,2l-1}(\omega_j)}\right\}$$

we will calculate the amount of its efficiency by calculating the SNR, which is Signal-to-noise ratio, where the higher the ratio, the lower the efficiency of the function, and vice versa. And the SNR is given as:

$$SNR=10\log\frac{P_s^2}{p_n^2}$$

Where p_s average power of signal, Where p_n average power of noise.

SIMULATION

In order to ensure the effectiveness of our algorithm, we will simulate it on data generated through equations (2) and (3):

$$X_1(t)=2\cos\left(\frac{2\pi t}{10}\right)+\cos\left(\frac{2\pi t}{3}\right)+0.3\epsilon_1(t)\qquad(2)$$

$$X_2(t)=\cos\left(\frac{2\pi t}{3}\right)+0.01\epsilon_2(t)\qquad(3)$$

and $\epsilon_1(t)$ and $\epsilon_2(t)$ are Gaussian white noise and with unit variance [8]. We repeat the experiment 500 times to reach the stability in our results. We set the deepest level at K = 4 to get best segmentation of the data set simulated.

Table (1)

SNR.DIR	SNR.FEJ	SNR.MDA	ϵ2	ϵ1
0.3716449	0.400329	0.4298197	0.1	0.1
0.3843005	0.40553	0.4310931	0.2	
0.3844411	0.4069628	0.4300496	0.3	
0.3856573	0.4010184	0.4279381	0.4	
0.377933	0.4067566	0.4311372	0.5	
0.7628395	0.8025926	0.8580769	0.1	0.2
0.7661904	0.8255708	0.8769	0.2	
0.7700758	0.8127702	0.8724363	0.3	
0.7713775	0.8341745	0.8790027	0.4	

0.7532902	0.7964639	0.8641199	0.5	
1.149613	1.208427	1.269268	0.1	0.3
1.147468	1.211061	1.295924	0.2	
1.163584	1.225195	1.286477	0.3	
1.174593	1.287403	1.344623	0.4	
1.165727	1.205203	1.283201	**0.5**	
1.552434	1.625218	1.706787	**0.1**	0.4
1.524964	1.634251	1.72714	**0.2**	
1.52105	1.618786	1.697334	**0.3**	
1.51946	1.627191	1.706705	**0.4**	
1.526252	1.587807	1.688534	**0.5**	
1.890793	2.009025	2.147682	**0.1**	0.5
1.931691	2.077368	2.198115	**0.2**	
1.925021	2.055334	2.172029	**0.3**	
1.887476	2.022909	2.14441	**0.4**	
1.904484	2.065318	2.189037	**0.5**	

Case (1) when the series length $T = 64$: This table shows the values of SNR of the three kernels (dirichlet, modify daniell and Féjer) also we notice that the Dirichlet function is the least function that has SNR, followed by the Féjer function, then the modified Daniell function, and this indicates that the the Dirichlet is the best in the first place, and Féjer is the second, and the modify daniell is the least efficient. with different values of $\epsilon_1(t)$ and $\epsilon_2(t)$, and $\epsilon_1(t) = 0.1, 0.2, 0.3, 0.4, 0.5$ and $\epsilon_2(t) = 0.1, 0.2, 0.3, 0.4, 0.5$, We will notice an increase in SNR with an increase in $\epsilon_1(t)$, while an increase in $\epsilon_2(t)$ did not affect the value of SNR.

Table (2)

SNR.DIR	**SNR.FEJ**	**SNR.MDA**	**ϵ2**	**ϵ1**
0.4916296	0.5185052	0.5388679	0.1	0.1
0.4875641	0.5159508	0.5412174	0.2	
0.4903628	0.5069003	0.5398473	0.3	
0.4934298	0.512536	0.5402371	0.4	
0.4915778	0.5032324	0.5295313	0.5	
0.9920309	1.025035	1.078214	0.1	0.2
0.9787744	1.024076	1.064123	0.2	
0.9760111	1.034798	1.067418	0.3	
0.9666547	1.019706	1.052843	0.4	
0.9622879	1.01305	1.055073	0.5	
1.44927	1.531902	1.597485	0.1	0.3
1.476582	1.5332	1.600587	0.2	
1.460798	1.536524	1.592947	0.3	
1.464646	1.535674	1.606712	0.4	
1.475798	1.516616	1.590978	0.5	
1.955448	2.038051	2.113182	0.1	0.4
1.966233	2.024986	2.124478	0.2	
1.968392	2.06911	2.16706	0.3	
1.919457	2.045421	2.158698	0.4	

1.927068	2.054976	2.147391	0.5	
2.448022	2.551684	2.652292	0.1	0.5
2.444091	2.552906	2.689711	0.2	
2.44182	2.550657	2.694096	0.3	
2.434137	2.519359	2.636933	0.4	
2.450165	2.537289	2.691139	0.5	

Case (2) when the series length $T = 128$: in the table above we have the values of SNR of the three functions of the kernels (dirichlet, modify daniell and Féjer) Through it, we can make sure that the best method is Dirichlet and the least efficient is Modify Daniell, and this depends on the small amount of SNR, the least method is more efficient, with different values of $\epsilon_1(t)$ and $\epsilon_2(t)$, and $\epsilon_1(t) = 0.1, 0.2, 0.3, 0.4, 0.5$ and $\epsilon_2(t) = 0.1, 0.2, 0.3, 0.4, 0.5$, We will notice an increase in SNR with an increase in $\epsilon_1(t)$, while an increase in $\epsilon_2(t)$ did not affect the value of SNR.

Table (3)

SNR.DIR	**SNR.FEJ**	**SNR.MDA**	**ϵ2**	**ϵ1**
0.6388364	0.6580156	0.6884488	0.1	0.1
0.6353982	0.6460467	0.6673244	0.2	
0.6299271	0.6468314	0.6688786	0.3	
0.6518823	0.6681748	0.6861094	0.4	
0.6514229	0.6657707	0.6899997	0.5	
1.304429	1.332485	1.359883	0.1	0.2
1.270372	1.281991	1.347287	0.2	
1.297633	1.313348	1.364689	0.3	
1.297604	1.315829	1.359966	0.4	
1.28379	1.317899	1.360493	0.5	
1.919325	1.982515	2.06656	0.1	0.3
1.938994	1.978266	2.066572	0.2	
1.972616	2.031568	2.093187	0.3	
1.9409	1.983267	2.039123	0.4	
1.926553	1.961336	2.022034	0.5	
2.607022	2.644573	2.752029	0.1	0.4
2.545621	2.595872	2.701873	0.2	
2.599238	2.636847	2.750788	0.3	
2.578736	2.607909	2.71972	0.4	
2.580245	2.656141	2.718683	0.5	
3.244801	3.293663	3.40992	0.1	0.5
3.217919	3.292232	3.435937	0.2	
3.194905	3.252596	3.352632	0.3	
3.230507	3.310323	3.404721	0.4	
3.260982	3.302506	3.425728	0.5	

Case (3) when the series length $T = 256$: it is the values of SNR of kernels functions affected by a change in values $\epsilon_1(t) = 0.1, 0.2, 0.3, 0.4, 0.5$ and

$\epsilon_2(t) = 0.1, 0.2, 0.3, 0.4, 0.5$, Since Dirichlet has a lower SNR, then it is the best method, followed by Féjer and then Modify Daniell.

Table (4)

SNR.DIR	**SNR.FEJ**	**SNR.MDA**	**ϵ2**	**ϵ1**
0.8528351	0.8509396	0.8609303	0.1	0.1
0.8610108	0.8632214	0.8769899	0.2	
0.8643097	0.8611453	0.8727881	0.3	
0.8574658	0.8531526	0.85884	0.4	
0.859195	0.8636694	0.8687784	0.5	
1.718979	1.734656	1.753989	0.1	0.2
1.686205	1.692459	1.723663	0.2	
1.737542	1.724064	1.770104	0.3	
1.711009	1.729026	1.767159	0.4	
1.705533	1.710395	1.740164	0.5	
2.622079	2.590571	2.644265	0.1	0.3
2.574888	2.604067	2.66691	0.2	
2.587633	2.595195	2.633649	0.3	
2.603265	2.602385	2.652459	0.4	
2.574234	2.587266	2.643174	0.5	
3.446554	3.456571	3.515744	0.1	0.4
3.449007	3.401074	3.471733	0.2	
3.438864	3.430769	3.511962	0.3	
3.463009	3.489161	3.527932	0.4	
3.424694	3.450596	3.519508	0.5	
4.295082	4.304542	4.362178	0.1	0.5
4.280181	4.260857	4.372519	0.2	
4.299988	4.251576	4.349968	0.3	
4.324139	4.321861	4.362505	0.4	
4.303323	4.326193	4.462218	0.5	

Case (4) when the series length $T = 512$: The table above shows the values of SNR of the three kernels (dirichlet, modify daniell and Féjer) also we notice that the Dirichlet function is the least function that has SNR, followed by the Féjer function, then the modified Daniell function, and this indicates that the The dirichlet is the best in the first place, and in the second place, the Féjer is the best, and the modify daniell is the least efficient., the values of with different values of $\epsilon_1(t)$ and $\epsilon_2(t)$, and $\epsilon_1(t) = 0.1, 0.2, 0.3, 0.4, 0.5$ and $\epsilon_2(t) = 0.1, 0.2, 0.3, 0.4, 0.5$. We will notice an increase in SNR with an increase in $\epsilon_1(t)$, while an increase in $\epsilon_2(t)$ did not effect on the value of SNR. And through the previous six tables, we notice that the value of SNR increases with the increase in the mole of the sreies.

Now we see the Dirichlet function is the least function that has SNR, followed by the Féjer function, then the modified Daniell function, and this indicates that the Dirichlet function is the best because the percentage of confusion in it is less, and the modified Daniell function is the least efficient among the three functions because it has the higher distortion rate, so, it is the least efficient function.

Dna Data

First, we will take a quick look at what DNA is, The DNA is a sequence of letters which represents information of the DNA strand. The DNA strand is made up of a long string of chemical building blocks called "nucleotides". Each nucleotide is made up of nitrogenous base, a five carbon sugar, and a phosphate group. There are four different nitrogenous bases, which are labeled A(Adenine), T(Thymine), G(Guanine) and C(Cytosine). Nucleotides are arranged in two long strands that form a spiral called a double helix. The strands are complementary; Adenine with Thymine and Cytosine with Guanine. So, it is sufficient to represent a DNA molecule by the sequence of nitrogen bases on one single strand.

The data collected is the first 8192 sequences of nitrogenous bases of Homo sapiens mitochondrion.

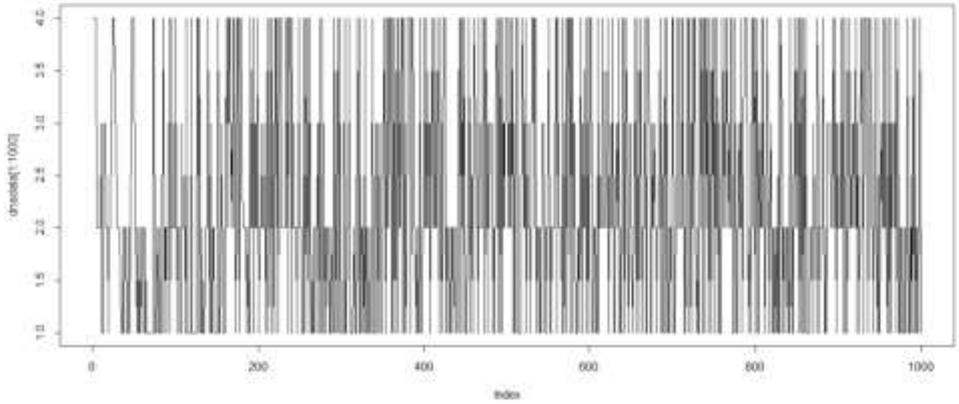

Figure (1) Representing the data in the form of a chart, where the number 1 represents the nuclide A, the number 2 represents the nuclide C, the number 3 represents the nuccludite G, and the number 4 represents the nuccludite T

After modeling and analyzing it, we obtained the following value of SNR of kernles:

The method	SNR
Dirichlet	12.42816
Féjer	12.44902
Modify Daniell	12.73957

Through the results, we notice that preference was given to the Dirichlet method, then Féjer, and finally the modify Daniell, and this is completely consistent with the simulation results.

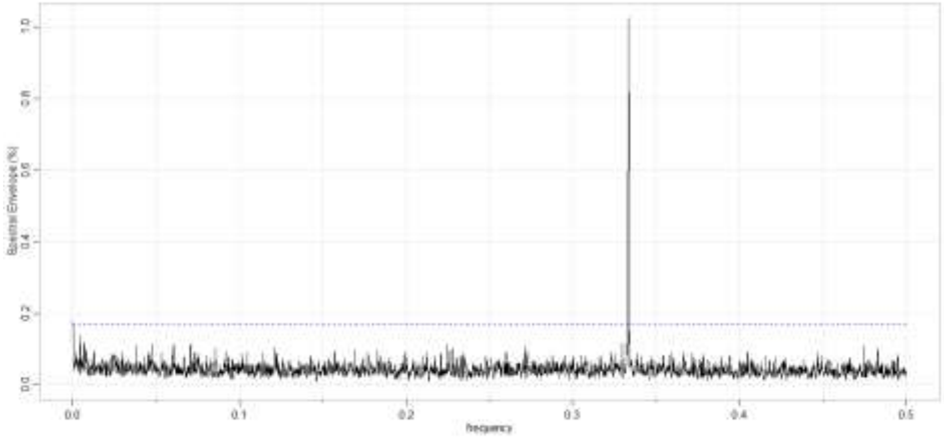

Figure (2) The SNR amount is shown under the Dirichlet method

Here we notice a peak at frequencies $\omega_{2732} = 0.33349$ which coincide to the period of $P = 18.8407$ This reference that the data shows an approximate Nineteen sequence cycle. This peak indicates the possibility of a genetic mutation. Where, during a DNA test, it is stimulated to replicate and reproduce itself.When the DNA is doubled, an incorrect association of the nitrogenous bases occurs in the DNA, and this may lead to the emergence of an unwanted genetic trait

2. Conclusions

It was found that the length of the series has an effect on increasing the SNR of the spectral envelope, and this matches the effect ofϵ_1, which is the amount of random error in function (6), while ϵ_2 had no effect, which is the amount of random error in function (7), as Equations (6) and (7) are the two functions that generate the data in the simulation. And also It was found in simulation that Dirichlet is less disruptive than Féjer and modify Daniell so it is the best method, this is what we found also in the application in real data when we applied the scenario to the DNA.

3. Bibliography

1. Albert, I., & Denis, J. B. (2012). Dirichlet and multinomial distributions: properties and uses in Jags.
2. Chan, K. S., & Cryer, J. D. (2008). *Time series analysis with applications in R*. springer publication.
3. Chatfield, C., & Xing, H. (2019). *The analysis of time series: an introduction with R*. CRC press.
4. Fr´ed´eric Ouimet and others, Asymptotic properties of Dirichlet kernel density estimators, California Institute of Technology, Pasadena, CA 91125, USA, 22 Sep 2021.
5. Jeong, H. (2012). *The spectral analysis of nonstationary categorical time series using local spectral envelope* (Doctoral dissertation, University of Pittsburgh)..
6. Li, Q., & Racine, J. S. Nonparametric Econometrics: Theory and Practice Princeton University Press (2007).
7. Manikandan. (2022). Cyber Security Issues and Solution in Vehicular Networks. Journal of Artificial Intelligence,Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172, 2(06), 43–54. https://doi.org/10.55529/jaimlnn.26.43.54
8. Nura Bawa, Hafsat Yusuf Imam, & Aishatu Jibril Bello. (2022). Undergraduate Students’ Perceptions of the Use of Moodle Learning Management System in Usmanu Danfodoyo University, Sokoto. Journal of Artificial Intelligence,Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172, 2(03), 1–8. https://doi.org/10.55529/jaimlnn.23.1.8
9. Omkar Dabade, Aditya Admane, Deepak Shitole, & Vitthal Kamble. (2022). Developing an Intelligent Credit Card Fraud Detection System with Machine Learning. Journal of Artificial Intelligence,Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172, 2(01), 45–53. https://doi.org/10.55529/jaimlnn.21.45.53
10. Ouimet, F., & Tolosana-Delgado, R. (2022). Asymptotic properties of Dirichlet kernel density estimators. *Journal of Multivariate Analysis*, *187*, 104832.
11. Panneerselvam. (2022). Framework and Challenges of Cyber Security in India: An Analytical Study. International Journal of Information Technology &Amp; Computer

Engineering (IJITC) ISSN : 2455-5290, 2(04), 27–34. https://doi.org/10.55529/ijitc.24.27.34

12. Paul Matudi Bako, & Udisifan Michael Tanko. (2022). The Place of Artificial Intelligence in Accounting Field and the Future of Accounting Profession. Journal of Artificial Intelligence,Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172, 2(05), 15–21. https://doi.org/10.55529/jaimlnn.25.15.21
13. Pérez-Cruz, F. (2008, July). Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory* (pp. 1666-1670). IEEE.
14. Rust, B. (2013). Convergence of Fourier Series.
15. Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). *Time series analysis and its applications* (Vol. 3). New York: springer..
16. Stoffer, D. S., Ombao, H. C., & Tyler, D. E. (2002). Local spectral envelope: an approach using dyadic tree-based adaptive segmentation. *Annals of the Institute of Statistical Mathematics*, *54*, 201-223.
17. Stoffer, D. S., Tyler, D. E., & McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, *80*(3), 611-622.
18. Showkat Ahmad Dar, & Dr. Naseer Ahmad Lone. (2022). Lockdowns in Jammu and Kashmir: The Human Rights Consequences. Journal of Legal Subjects(JLS) ISSN 2815-097X, 2(04), 1–11. https://doi.org/10.55529/jls.24.1.11
19. S Ramesh. (2022). A Study of Law Regarding Life Insurance Business in India. Journal of Legal Subjects(JLS) ISSN 2815-097X, 2(05), 10–14. https://doi.org/10.55529/jls.25.10.14
20. Shabir Ahmad Lone. (2022). Reflections of Dr. B.R Ambedkar's Idea of Social Justice. Journal of Legal Subjects(JLS) ISSN 2815-097X, 2(03), 6–11. https://doi.org/10.55529/jls.23.6.11
21. Talia Sopiyani, Kanti Rahayu, Erwin Aditya Pratama, Toni Haryadi, & Achmad Irwan Hamzani. (2021). Comparison of the Law of Geographical Indications between Indonesia and India. Journal of Legal Subjects(JLS) ISSN 2815-097X, 1(02), 1–7. https://doi.org/10.55529/jls12.1.7
22. Vandaele, V. V. (1983). Applied Time Series and Box-Jenkins Models Academic press. *Nevv York.*
23. Vanshika Singh. (2022). Role of Juvenile Justice System in India. Journal of Legal Subjects(JLS) ISSN 2815-097X, 2(05), 1–4. https://doi.org/10.55529/jls.25.1.4
24. Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, *15*(2), 70-73.

Cite this article as: **Ali Ghanim Abood (2024).**

Drichlet, Féjer and Modify Daniell Kernel to Estimate Human DNA Spectral Envelope

African Journal of Biological Sciences. 6(1), 72-84. *doi: 10.33472/AFJBS.6.1.2024.72-84*