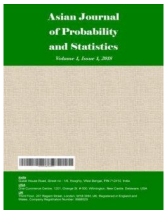

Asian Journal of Probability and Statistics

5(1): 1-15, 2019; Article no.AJPAS.50976
ISSN: 2582-0230

Estimation of Non-smooth Functionals in Hilbert Sample Space Using the Edgeworth Expansions

M. M. Kololi[1*], G. O. Orwa[2], J. K. Mung'atu[2] and R. O. Odhiambo[3]

[1]Department of Mathematics, Kibabii University, P.O.Box 1699-50200 Bungoma, Kenya.
[2]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, P.O.Box 62000-00200 Nairobi, Kenya.
[3]Meru University, Kenya.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2019/v5i130126
Editor(s):
(1) Dr. Oguntunde, Pelumi Emmanuel, Department of Mathematics, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria.
Reviewers:
(1) Fevzi Yaşar, Gaziantep University, Turkey.
(2) Olumide Adesina, Onabanjo University, Nigeria.
Complete Peer review History: http://www.sdiarticle3.com/review-history/50976

Received: 13 June 2019
Accepted: 23 August 2019
Published: 31 August 2019

Original Research Article

Abstract

An arbitrary non-smooth functional is estimated using a nonparametric set-up. Exploratory data analysis methods are relied on to come up with the functional form for the sample to allow both robustness and optimality to be achieved. An infinite number of parameters are involved and thus the Hilbert sample space is a natural choice. An important step in understanding this problem is the normal means problem, $T(\theta) = \frac{1}{n}\sum_{i=1}^{n}|\theta_i|$. The basic difficulty of estimating $T(\theta)$ as defined can be traced back to the non differentiability of the absolute value function, at the origin. Accordingly, constructing an optimal estimator is not easy partly due to the nonexistence of an unbiased estimate of the absolute value function. Therefore, best polynomial approximation was used to smooth the singularity at the origin and then an unbiased estimator for every term in the expansion constructed by use of Hermite polynomials when the averages are bounded by a given constant M > 0 say. The expansion of the Gaussian density function in terms of Hermite polynomials gives a clear and almost accurate estimate that admits cumulant generating function; the Saddle point approximation. Additional precision is obtained by using a higher order Taylor series expansion about the mean resulting in Edgeworth expansion techniques.

[*] *Corresponding author: E-mail: kololimusa@gmail.com;*

Keywords: Nonparametric; non-smooth functional; edgeworth expansion.

1 Introduction

Estimating statistical functionals play a major role in applied statistical analysis. For instance, estimating the probability density functions play a major role in many varied fields of science, engineering and the earth sciences. Functional estimation is one of the major tests used when dealing with nonparametric statistics. Having the knowledge of the functional form, further statistical issues such as estimation of quantities and construction of stochastic models for various applications can be handled.

Finding an estimator for a non-smooth functional in the NP set-up involve highly structured problems which arise differently from the ones in which mainline numerical methods have been designed [1]. They have different optimal rates of convergence [2,3]. Also, the estimation of non-smooth functional is faced with the curse of dimensionality.

Researchers have shown that the most difficult testing problem is to find the accurate critical test statistic value for small sample sizes and when sample sizes are moderately large. In such circumstances an approximation method for assessing the density of the test statistic is used to estimate the exact critical value. In the previous studies, polynomial smoothers have been suggested. However, such smoothers tend to perform poorly in the tail areas [4].

Parametric approach is based on the assumptions that the parent population has a known distribution. The distribution family is usually decided upon for a particular situation, and fitting a PDF then means determination of parameters by some estimation technique. The method of maximum likelihood and the method of moments are known examples of parametric methods [5]. The likelihood level is expressed in terms of the variance of the estimator which usually approach zero at the speed of the square root of the sample size [6].

In contrast to the foregoing is the NP smoothing technique. The variance alone does not exhaustively quantify the convergence of curve estimators. There is also a bias, which is common in the framework smoothing techniques. This is why accuracy is calculated by combining the variance and squared bias. This smoothing technique effectively reveals important structures from noisy data.

Approximations are never feasible when done on the global scales, but locally. This is possible since concentration of measure at a certain point allows approximation to be done locally [6]. Intuitively, the bias is generally due to the set of points where the function $f(\alpha)$ changes suddenly. For $f(\alpha) = -\alpha \ln \alpha$ or $p^{\alpha}, \alpha > 0$, the most non-smooth point is p = 0. Briefly, approximating locally around the nonsmooth points reduce bias. The approximation can also be obtained through series expansions, e.g. Gram-Charlier and Edgeworth series [7].

In regard to estimating the functional $F(f)$, the squared error risk, L_2 of an arbitrary estimator $\hat{F}$ is defined as $\mathbb{E}\big(F(f) - \hat{F}\big)^2$ where $\mathbb{E}_f$ is expectation with respect to the distribution f that gives the observations used by F. The squared error risk is a function of both f and $\hat{F}$, and is used to minimize the error of estimation. Thus, it is difficult to find the best guess which minimizes the risk function since it relies on unknown parameter.

In the recent years, efforts have been made to find such an estimator. This includes the use of the MiniMax criterion to give estimators whose maximal risk is minimal among all estimators, i.e. $sup_{f\in\mathcal{P}}\mathbb{E}_f\big(F(f) - \hat{F})^2$ Where $\mathcal{P}$ denotes the set of all discrete distributions. See for example [2,8,9]. MiniMax estimators have also been utilized as a standard to measure any estimator.

When the absolute value function is used to obtain an optimal estimator, it is noted that the risk function is non-smooth at the origin [5]. The function is smoothened at the origin using the MiniMax polynomial estimation. The presence of the polynomial factor in an approximate of the density can be worse in tail areas; the approximate density may even be negative and cannot be integrated to 1. In this study therefore, an estimator that gives clear and accurate results is proposed; the saddle point approximation. The estimator provides accurate approximation of densities and tail probabilities even in small samples [10,11].

The rest of this paper is organized as follows. In section 2, we review the Normal means model. The Saddle point approximation is considered in section 3. In section 4 the estimator is developed using the Edgeworth expansions. The asymptotic properties of the developed estimator are discussed in section 5. In section 6 the empirical study is considered while section 7 considers the conclusions [12].

2 The Normal Means Model

The problem of density estimation can be related to the Normal means problem since the Normal means problem unifies some NP problems. For instance, suppose $Z^n = Z_1, \dots, Z_n$where

$$Z_1 = \theta_i + \sigma_n \epsilon_i, i = 1, \dots, n, \tag{1}$$

$\epsilon_i, \dots, \epsilon_n$ are independent, $N(0,1)$ random variables, and $Z_i = n^{-1} \sum_{j=1}^{n} X_{ij}$

$$\theta^n = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$$

is a vector of unknown parameters and σ_n is assumed known. The model appears to be parametric but the number of parameters increases at the same rate as the size of the sample. The above model has all the complexities of a NP problem. In order to realize this, an infinite-dimensional version of equation (1) is considered,

$$Z_1 = \theta_i + \sigma_n \epsilon_i, i = 1,2, \dots \tag{2}$$

Where now the unknown parameter is $\theta = \theta_1, \theta_2, \dots$

Table 1. The normal means $X_{ij} = \theta_i + N(0, \sigma^2)$ and $Z_i = n^{-1} \sum_{j=1}^{n} X_{ij} = \theta_i + \sigma_n \epsilon_i$ Where $\sigma_n = \frac{\sigma}{\sqrt{n}}$ Estimating the parameters $\theta_1, \dots, \theta_n$ from the n column means $Z_1, \dots, Z_n$lead to the model $Z_i = \theta_i + \sigma_n \epsilon_i$ with $\sigma_n = \frac{\sigma}{\sqrt{n}}$

$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	…	$\boldsymbol{\theta_i}$	…	$\boldsymbol{\theta_n}$
X_{11}	X_{21}	$\cdots$	X_{i1}	$\cdots$	X_{n1}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
X_{1j}	X_{2j}	$\cdots$	X_{ij}	$\cdots$	X_{nj}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
X_{1n}	X_{2n}	$\cdots$	X_{in}	$\cdots$	X_{nn}
Z_1	Z_2		Z_i		Z_n

Given an estimator $\hat{\theta}^n = \hat{\theta}_1, \dots, \hat{\theta}_n$, the squared error loss

$$L(\hat{\theta}^n, \theta^n) = \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i) = (||\hat{\theta}^n - \theta^n||)^2 \tag{3}$$

is used with risk function;

$$R(\hat{\theta}^n, \theta^n) = \mathbb{E}_\theta (L(\hat{\theta}^n, \theta^n) = \sum_{i=1}^{n} \mathbb{E}_\theta (\hat{\theta}^n - \theta^n)^2 \tag{4}$$

One of the choices for an estimator of θ^n is $\hat{\theta}^n = Z^n$. This estimator is the MLE, minimum variance unbiased estimator and the Bayes estimator under a uniform prior. However, $\hat{\theta}^n = Z^n$ is a poor estimator as its risk is

$$R(\hat{\theta}^n, \theta^n) = \sum_{i=1}^{n} \mathbb{E}_\theta (Z_i - \theta_i)^2 = \sum_{i=1}^{n} \sigma_n^2 = n\sigma_n^2 \tag{5}$$

Estimators with significantly small risks than $R(Z^n, \theta^n)$ can be obtained. The Normal means problem in relation to NP regression and density estimation is important to improve this estimator. In order to obtain such an estimator, a review on the Hilbert function space is required [13,14].

2.1 Hilbert sample space

The Hilbert sample space is the closest analogy to the Euclidean space in infinite dimension.

The Hilbert space is a complete inner product space denoted by $L(\langle .,. \rangle)$ and abbreviated by L_2 [4]. The square integrable functions form a complete metric space under the metric induced by the inner product. The importance of forming a complete metric space is to allow the sequences converge, and find a point to which they converge within space.

An inner product always generates a norm. And where there is a norm, there is a metric. A norm defined by the inner product $\langle .,. \rangle$ will define the following metric;

$$d(x, y)=\| x - y \|= \sqrt{\langle x - y, x - y \rangle} \ \forall x, y \in X \tag{6}$$

Inner product norms satisfy a number of properties that are not satisfied by all norms.

The coordinates of an element of the Hilbert space can be specified uniquely with respect to a set of coordinate axes, which is in analogy with the Cartesian coordinates in the plane. A sequence of functions $\phi_1, \phi_2, \ldots$ is called orthonormal if $\| \phi_j \|$=1 for all j and $\int_a^b \phi_i(x)\phi_j(x)\, dx = 0$ for $i \neq j$. This sequence is complete if the only function that is orthogonal to each ϕ_j is the zero function. A complete, orthogonal set of functions forms a basis, meaning that if $f \in L_2(a, b)$ then f can be expanded in the basis. An example of an orthonormal basis is the Chebyshev and Hermite polynomials defined on $(-1,1)$ [15].

However, indicated that there is a limitation for the use of Hilbert space on the real line [3]. First, the standard Hilbert space requires that the unknown function approaches zero at infinity. This makes it unreasonable to be used on some models like the economic model. To overcome this, weighted Hilbert spaces are used, since they enforce weaker limiting requirements at infinity. Second, the Hilbert space is restricted to bounded sample spaces in nonparametric estimation [15].

3 The Saddle Point Approximation

The saddle point estimate can be gotten from any statistic that concedes a cumulant generating function. The estimate gives a relative error which gets small with the density in the tails of the distribution. In saddle point estimation, the empirical distribution is used to give a relative error of order $n^{-\frac{1}{2}}$. This results in an improvement over the absolute error in the tails of the distribution.

Suppose $f(\alpha)$ denotes the density being approximated, the asymptotic approximation can be written

$$\hat{f}_{asy}(\alpha) = f(\alpha) + O(n^{-\frac{1}{2}}) \tag{7}$$

This means $\hat{f}_{asy}(\alpha) - f(\alpha)/(n^{-\frac{1}{2}})$ is bounded as α approaches some constant. The empirical saddle point approximation can be written

$$\hat{f}_{sad}(\alpha) = f(\alpha)\left\{1 + O(n^{-\frac{1}{2}})\right\} \tag{8}$$

Also, the saddle point density replaces the Central Limit Theorem (CLT) in the conventional general strategy for moment's distribution theory [16]. While the CLT uses data about the location and convexity of the general method of moment's objective function at the global minimum, the saddle point approximation uses data about the shape of the objective function at each point in the parameter space.

The CLT is typically displayed as far as assessing the mean of a distribution. It demonstrates that the sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \sim (\mu, \frac{\sigma^2}{n})$. So as to demonstrate how quick the distribution converges to normality, the moment generating function of $\alpha, M_\alpha(t) = e^{t\alpha}$ is used. The MGF remarkably recognize a specific distribution. It is likewise possible to make adjustments to this function to build the accuracy of the approximations. The Edgeworth expansions technique utilizes information about higher order moments to build accuracy of the approximations [16].

One issue with the MGFs is that they don't always exist. This is understood through the characteristic function which is the complex extension of the MGF and defined as $\phi_\alpha(t) = E(e^{i\alpha t})$, where $i^2 = -1$. The characteristic function just like the MGF uniquely identify distributions and can be used to show limiting results.

3.1 The characteristic function

The characteristic function of a probability measure μ on the line is defined for real t by

$$\psi(t) = \int_{-\infty}^{\infty} e^{it\alpha}\mu(d\alpha)$$

$$= \int_{-\infty}^{\infty} cost\alpha\mu(d\alpha) + i\int_{-\infty}^{\infty} sint\alpha\mu(d\alpha) \tag{9}$$

A random variable α with the distribution μ has the characteristic function

$$\psi(t) = \mathbb{E}[e^{it\alpha}] = \int_{-\infty}^{\infty} e^{it\alpha}\mu(d\alpha) \tag{10}$$

The characteristic function in non probabilistic settings is the Fourier transform. The Fourier transform is used in the Central Limit Theorem to send smooth functions to bounded functions and bounded functions to smooth functions.

For instance, if $g: \mathbb{R} \to \mathbb{R}$, the Fourier transform of g is defined by

$$\hat{g}(\beta) = \int_{-\infty}^{\infty} e^{-i\alpha y} g(\alpha)(d\alpha) \tag{11}$$

Where g is a Schwartz function, it is C^∞ and all of its derivatives decay at ±∞ faster than every polynomial [17]. In the event that g is Schwartz, then $\hat{g}$ is Schwartz. This gives the inversion formula

$$\hat{g}(\beta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\beta}\hat{g}(\beta)(d\beta) \tag{12}$$

holds for Schwartz functions. A straightforward computation shows that

$$\widehat{e^{-\alpha^2/2}} = \sqrt{2\pi}\, e^{-\beta^2/2}$$

Suppose $\alpha\sim G(\alpha)$ and $\psi_\alpha(t)$ denotes the characteristic function of α. If $\int_{-\infty}^{\infty}|\psi_\alpha(t)| < \infty$, then $g(\alpha) = \acute{G}(\alpha)$ exists and

$$g(\alpha) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\alpha}\psi_\alpha(t)(dt) \tag{13}$$

This gives the inversion formula for characteristic functions. Also

$$\begin{aligned}\int_{-\infty}^{\infty} f(\alpha)\hat{g}(\alpha)d\alpha &= \int_{-\infty}^{\infty} f(\alpha)\left[\int_{-\infty}^{\infty}\hat{g}(\beta)\, e^{it\beta}d\beta\right]\\ &= \int_{-\infty}^{\infty} g(\beta)\big[f(\alpha)e^{it\beta}d\alpha\big]d\beta\\ &=\int_{-\infty}^{\infty}\hat{f}(\beta)g(\beta)d\beta\end{aligned} \tag{14}$$

This is valid given that the interchange of integrals can be justified; specifically, it holds for Schwartz f, g.

If $X\sim N(0,1)$, then by completing the square in the exponential, the characteristic function for a standard normal distribution can be obtained;

$$\phi(t) = e^{it\alpha}\frac{1}{\sqrt{2\pi}}e^{-\alpha^2/2}d\alpha = e^{-t^2/2}$$

Then for any $k > 0$,

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it\alpha}e^{-t^2/2}(it)^k dt) &= \frac{(-1)^k}{2\pi}\frac{d^k}{d\alpha^k}\int_{-\infty}^{\infty} e^{-itx}e^{-t^2/2}dt\\ &= (-1)^k\frac{d^k}{d\alpha^k}\phi(\alpha)\\ &= H_k(\alpha)\phi(\alpha)\end{aligned} \tag{15}$$

Where (3.1.7) pursues from (3.1.5) since $e^{-t^2/2}$ is the characteristic function for a standard Normal distribution [17].

As seen earlier, the CLT is based on a two-term Taylor series expansion of a higher degree expansion to get extra precision. This results in Edgeworth expansion technique that gives significantly better estimates at the mean of a distribution. However, the quality of the estimate can weaken especially for the values away from the mean.

So as to get an improved estimate at an arbitrary value in the parameter space, the original distribution is changed to a conjugate distribution. A specific conjugate distribution is selected so that its mean is changed back to the original distribution at the value of interest.

4 The Developed Estimator

Suppose F is a function to estimate $T(\theta) = \sum_{i=1}^{n}|\theta_i|$to such an extent that $F[a, b] \to \mathbb{R}$: and g(x) is a polynomial approximation of F as for the weight function $\omega(x)$. Let $X_1, X_2, \dots, X_n$ be independent Normal random variables where $X_i\sim N(0,1)$ from F(x). Assume $E(X_1) = 0$ and $Var(X_1) = 1$; for otherwise, each X_1 is replaced by $(X_j - \mathbb{E}(X_1)/\sqrt{Var(X_1)})$. Let likewise $\gamma = \mathbb{E}(X_j^3)$ and $\tau = \mathbb{E}(X_j^4)$ and assume $\tau < \infty$. In this paper, the estimate of the distribution of the standardized sum,

$$S_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j$$

is developed. By the CLT, $S_n \sim N(0,1)$ [11]. Also for each x, $P(S_n \leq x) \rightarrow \Phi(x)$ where $\Phi(x)$ shows the standard Normal distribution function. In this work, a better estimation to $P(S_n \leq x)$ than $\Phi(x)$ is developed by the characteristic function of S_n.

$$\psi S_n(t) = \mathbb{E}\left[exp\left\{it/\sqrt{n}\textstyle\sum_j X_j\right\}\right] = \left[\phi_X(t/\sqrt{n})\right]^n \qquad (16)$$

At this point the Taylor's expansion of $exp\{itx/\sqrt{n}\}$ is used: As $n \rightarrow \infty$,

$$\psi_t\left(\frac{t}{\sqrt{n}}\right) = \mathbb{E}\left\{1 + \frac{itX}{\sqrt{n}} + \frac{(it)^2X^2}{2n} + \frac{(it)^3X^3}{6n\sqrt{n}} + \frac{(it)^4X^4}{24n^2}\right\} + \mathcal{o}\left(\frac{1}{n^2}\right)$$

$$= \left(1 - \frac{t^2}{2n}\right) + \frac{(it)^3\gamma}{6n\sqrt{n}} + \frac{(it)^4\tau}{24n^2} + \mathcal{o}\left(\frac{1}{n^2}\right) \qquad (17)$$

where $\mathcal{o}\left(\frac{1}{n^2}\right)$ is the error in the Taylor's expansion and $i = -1$. Raising this tetra nominal to the nth power most terms are $\mathcal{o}\left(\frac{1}{n}\right)$:

$$\left[\psi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left(1 - \frac{t^2}{2n}\right)^n + \left(1 - \frac{t^2}{2n}\right)^{n-1} + \left(\frac{(it)^3\gamma}{6\sqrt{n}} + \frac{(it)^4\tau}{24n}\right)$$

$$+ \left(1 - \frac{t^2}{2n}\right)^{n-2}\frac{(n-1)(it)^6\gamma^2}{72n^2} + \mathcal{o}\left(\frac{1}{n}\right) \qquad (18)$$

Using (17) and the binomial theorem equation for a fixed nonnegative integer k:

$$\left(1 + \frac{a}{n}\right)^{n-k} = e^a\left(1 - \frac{a(a+k)}{2n}\right) + \mathcal{o}\left(\frac{1}{n}\right) \text{ as } n \rightarrow \infty$$

$$\psi_{S_n}(t) = e^{-\frac{t^2}{2}}\left[1 - \frac{t^4}{8n} + \frac{(it)^3\gamma}{6\sqrt{n}} + \frac{(it)^4\tau}{24n} + \frac{(it)^6\gamma}{72n}\right] + \mathcal{o}\left(\frac{1}{n}\right)$$

$$= e^{-\frac{t^2}{2}}\left[1 + \frac{(it)^3\gamma}{6\sqrt{n}} + \frac{(it)^4(\tau-3)}{72n} + \frac{(it)^6\gamma^2}{72n}\right] + \mathcal{o}\left(\frac{1}{n}\right) \qquad (19)$$

Combining equation (19) with (15) the density function below is obtained as an approximation to the distribution of S_n:

$$g(x) = \frac{1}{2\pi}\left(\int_{-\infty}^{\infty} e^{-itx}e^{\frac{-t^2}{2}} + \frac{\gamma}{6\sqrt{n}}\int_{-\infty}^{\infty} e^{-itx}e^{\frac{-t^2}{2}}(it)^3dt + \frac{\gamma-3}{24n}\int_{-\infty}^{\infty} e^{-itx}e^{\frac{-t^2}{2}}(it)^4dt \quad + \frac{\gamma^2}{72n}\int_{-\infty}^{\infty} e^{-itx}e^{\frac{-t^2}{2}}(it)^6dt\right) \qquad (20)$$

Putting (20) with (14) gives:

$$g(x) = \phi(x)\left(1 + \frac{\gamma H_3(x)}{6\sqrt{n}} + \frac{(\tau-3)H_4(x)}{24n} + \frac{\gamma^2H_6(x)}{72n}\right) \qquad (21)$$

Using (21), the antiderivative of $g(x)$ equals:

$$G(x) = \Phi(x) - \phi(x)\left(\frac{\gamma H_2(x)}{6\sqrt{n}} + \frac{(\tau-3)H_3(x)}{24n} + \frac{\gamma^2H_5(x)}{72n}\right)$$

$$= \Phi(x) - \phi(x)\left(\frac{\gamma(x^2-1)}{6\sqrt{n}} + \frac{(\tau-3)(x^3-3x)}{24n} + \frac{\gamma^2(x^5-10x^3+15x)}{72n}\right) \qquad (22)$$

The letter ϕ denotes the probability density function and the corresponding distribution function is denoted Φ.

5 Asymptotic Properties of the Developed Estimator

In this section, the focal point is describing properties of estimators when sample sizes are arbitrarily large. The properties of an estimator not withstanding when the sample size is finite are like when the sample size is arbitrarily large. The assumptions are made on the stochastic procedure that generates the sample. The two fundamental tools of asymptotic are asymptotic normality and consistency [18].

5.1 Asymptotic normality of the developed estimator

Asymptotic normality shows that as more data are obtained, averages of random variables behave such as normally distributed random variables. The probability tools for establishing asymptotic normality are the Central Limit Theorems (CLTs). The CLT demonstrates that in the event that a large enough sample is drawn from a population, at that point the distribution of the sample mean is approximately normal, regardless what population the sample was drawn from.

For instance, let $x_1, \dots, x_n$ represent sample from a Binomial distribution generated randomly with fixed p, as n gets larger the p.m.f. looks increasingly like a normal p.d.f. Suppose $Y_1,\ldots,Y_n \sim_{i.i.d} Bern(p)$ then the distribution of $\sum Y_i$ looks increasingly Normal as $n \to \infty$ Also, for fixed n, the p.m.f. looks more Normal when $p = 0.5$ than when $p = 0.05$. This is because convergence under the CLT is faster when the distribution of each Y_i is more symmetric.

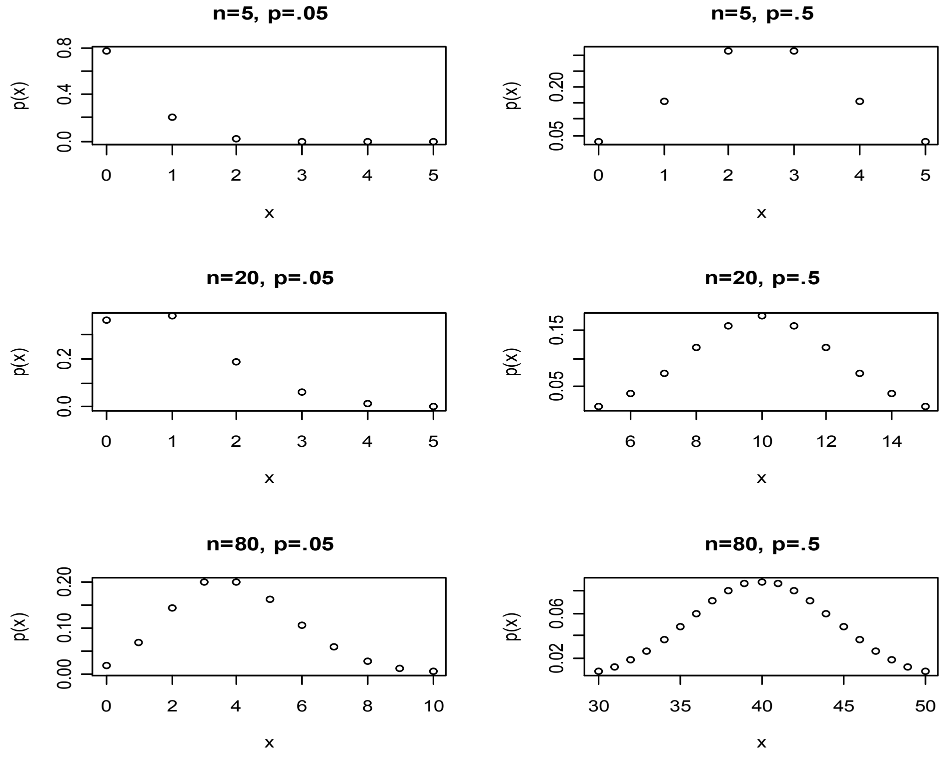

Fig. 1. The Binomial Probability mass function

5.2 Consistency of the developed estimator

Consistency demonstrates that the more data are obtained at that point one gets closer to knowing the truth. The probability theory tools for creating consistency of estimators are Laws of Large Numbers (LLNs). A LLN is a result that expresses the conditions under which a sample mean of random variables tends to a population mean. It concerns conditions under which the sequence of sample mean converges either in probability or almost surely. The results of LLN include: Chebyshev’s LLN, Kolmongorov’s/Khinchine’s LLN, Markov LLN.

The Chebyshev inequality is used to demonstrate that the developed estimator $G(x)$ is a consistent estimator for $T(\theta)$. Let X_n be i.i.d normal random variables where $x_i \sim N(\theta_i, 1)$. Let X be a random variable with mean θ and variance σ^2 then for any number $\epsilon > 0$,

$$P_x\{|\bar{X} - \theta| \geq \epsilon\} \leq \frac{var(\bar{Y})}{\sigma^2} \qquad \forall \sigma^2 < \infty$$

For a normal distribution; $\bar{X} = \frac{\sum x_i}{n}, Var(x) = var\left(\frac{\sum x}{n}\right) = \frac{1}{n^2}\sum var(x) = \frac{1}{n^2}\sum \sigma^2 = \frac{\sigma^2}{n}$

By Chebyshev’s inequality,

$$\lim_{n\to\infty} P\{|\bar{X} - \theta| > \epsilon\} \leq \frac{var(\bar{Y})}{\epsilon^2}$$

$$= \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

Since it gives zero, then $\bar{X}$ is a consistent estimator for θ.

5.3 Performance of the developed Estimator using the MiniMax criterion

The MiniMax criterion aims at finding an estimator $\hat{\theta}(x)$ for the parameter θ which minimizes the given risk function $R(\theta, \hat{\theta})$. The risk function L_2 is the expectation of some loss function $L(\theta, \hat{\theta})$ with respect to the conditional probability distribution $P(x|\theta)$. The risk of an arbitrary estimator $\bar{F}$ is defined as $\mathbb{E}_P(F(P) - \bar{F})^{\ 2}$ this risk cannot be minimized directly [19].

Calculating the exact MiniMax risk and MiniMax estimator for $F(f)$ is difficult. Moreover, even if the MiniMax Risk is computed exactly, it will depend on a parameter which is unknown. Hence, the requirement is slightly relaxed and seeks to obtain MiniMax rate estimators with maximum risk equal to the MiniMax risk up to a multiplicative constant. The MiniMax rate-optimal estimators are closely related to the problem of best polynomial approximation, which is a convex optimization problem. The connection between the two is important in approximation theory [19].

According to [8,9], a MiniMax estimator performs best in the worst possible case allowed in the problem. That is

$$sup_\theta\, R\big(\theta, \hat{\theta}\big) = \inf_{\hat{\hat{\theta}}} \sup_{\theta \in \Theta}\ R(\theta, \hat{\theta}) \tag{23}$$

Where the infimum is over all estimators θ.

Let θ be the parameter and $\hat{\theta}$ its estimator defined as

$$\hat{\theta} = G(x) = \textstyle\sum_{i=1}^{n} \propto_i\, T_i\left(2\frac{x-a}{b-a} - 1\right) \tag{24}$$

The MiniMax Risk for the developed estimator is

$$R_n \equiv R_n(\theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \tag{25}$$

The aim of this paper is to determine the MiniMax risk, R_n and find an estimator

that achieves this risk:

$$sup_{\theta \in \Theta} R\big(\theta, \hat{\theta}\big) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \ R(\theta, \hat{\theta}) \tag{26}$$

It is difficult to obtain an estimator that satisfies (26) and therefore an estimator

that achieves the MiniMax rate is used.

$$sup_{\theta \in \Theta} R\big(\theta, \hat{\theta}\big) \asymp \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \ R(\theta, \hat{\theta}) \quad n \to \infty \tag{27}$$

Where $(.) \asymp (..)$ means that both $(.)/(..)$ and $(..)/(.)$ are both bounded as $n \to \infty$.

Calculating the MiniMax Risk formalizes the idea of the existence of a best rate of convergence. The rate refers to point wise convergence uniformly over models in small neighborhood by some particular model of interest. However, the MiniMax Risk converges to zero at a slow logarithmic rate [20].

6 Simulations

A simulation experiment is performed in order to approximate a continuous function $f(x)$ by Chebyshev polynomial approximation, T_n as well compare the error of approximation as n increases. Let $f(x) = x^2 sin10x$ defined on the interval $[-1,1]$ be approximated by T_n where $n = 1,4,9,16$ and $P_n \in \mathcal{P}_n$. The following diagrams represent plots of $f(x)$ being approximated by T_n.

Fig. 2 shows that the error of approximating $f(x)$ by T_n decreases as n increases. To measure how good f(x) is approximated by the Chebyshev polynomials $P_n \in \mathcal{P}_n$, the uniform norm, $\parallel f - P_n \parallel = \max_{-1 \le x \le 1} |f(x) - P_n(x)|$ is used. This gives the error of approximation as the greatest distance between $f(x)$ and $T_n(x)$ with $x \in [-1,1]$ [21]. The coefficients of the function and the Chebyshev polynomials are used to calculate the absolute difference between $f(x)$ and $T_n(x)$. Table 2 show the coefficients of the function, $f(x)$ and the Chebyshev polynomials, P_n.

The absolute difference between $f(x)$ and $T_n(x)$ calculated from the coefficients using Matlab are shown in Table 3. From Table 3, the absolute difference between $f(x)$ and $T_n(x)$ decreases as n increases. Also, the error of approximation between $f(x)$ and $T_1(x)$ is larger than the error measure between $f(x)$ and $T_{16}(x)$. Thus, the Chebyshev polynomial T16(x) gives the smallest error to the function $f(x)$ and $T_1(x)$.

6.1 Real Data

Table 4 shows the masses of fifteen 100 gram tinned blue band weighed by an electric weighing machine. The table consists of repeated measurements of tinned blue band.

From Table 4 one notes that the masses of tinned blue band vary. This indicates that randomness is involved. The mean and variance of the observed data is 110:1 and 3:61667 respectively. However, these summary statistics do not give all the information in the observed masses. When the masses are put in an array, the elements lie between 106:8 and 112:5. The middle element (8th) is 110:5 which is closer to the maximum value 112:5 than the minimum value 106:8.

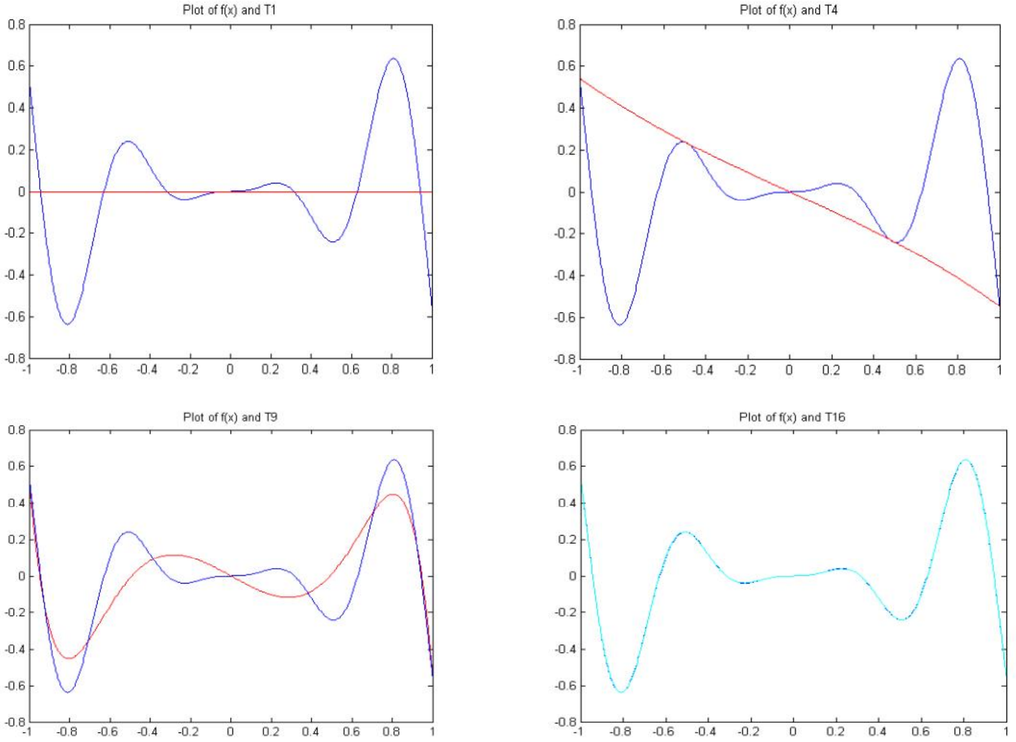

Fig. 2. Plots of $f(x)$ and its Chebyshev polynomial approximation T_n; n=1,4,9,16

Table 2. Coefficients of $f(x)$, T_n, $n = 1, 4, 9, 16$

f(x)	P_1	P_4	P_9	P_{16}
0.0000	0.0000	O.3750	0.0000	0.1964
0.0360	1.0000	0.0000	0.4922	0.0000
0.0000	-	0.5000	0.0000	0.3491
-0.1537	-	0.0000	0.3281	0.0000
0.0000	-	0.1250	0.0000	0.2444
-0.3716	-	-	0.1406	0.0000
0.0000	-	-	0.0000	0.1333
-0.1878	-	-	0.0352	0.0000
0.0000	-	-	0.0000	0.0555
0.1219	-	-	0.0390	0.0000
0.0000	-	-	-	0.0171
0.0373	-	-	-	0.0000
-	-	-	-	0.0037
-	-	-	-	0.0000
-	-	-	-	0.0005
-	-	-	-	0.0000
-	-	-	-	0.0000

Table 3. Absolute difference between f(x) and P_n (x)

n	**1**	**4**	**9**	**16**
$\|f(x) - P_n(x)\|$	0.8301	0.5301	0.4301	0.3889

Table 4. Masses of 15 tinned Blue band in Grams

n	1	2	3	4	5	6	7	8
Mass	112.5	108.5	111.6	110.4	110.8	110.6	111.7	112.1
n	9	10	11	12	13	14	15	
Mass	106.8	110.5	107.9	108.1	112.5	108.6	107.9	

Also, the observed data are univariate and consist of one particular quantity. Such data can be represented graphically by kernel density estimates and the empirical distribution function to give a visual impression of the main features of a dataset [22].

Fig. 3(a) is a histogram of the masses of 100 gram tinned blue band. From this figure one gets the impression that there are two major varieties of 100 gram tinned blue band, one with about 108 grams and the other with about 110 grams. Fig. 3(a), (b), (c) reveals the asymmetry of the observed data and the fact that elements accumulate somewhere near 108 and 110. From the shape of the histograms it seems reasonable to assume that the data are not normally distributed.

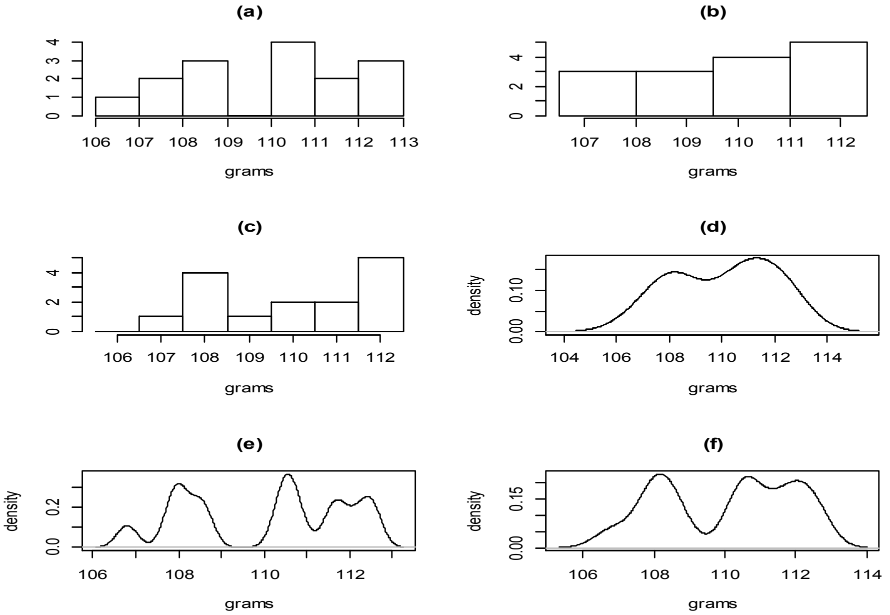

Fig. 3. (a), (b), (c): histograms of 100g tinned Blue band; (d),(e),(f): density estimates of 100g tinned Blue band

Fig. 3(b) is another histogram of the same data but with different bandwidth, h = 1:5. It gives an impression that the mass is evenly distributed with a small number of lower mass. Fig. 3(c) gives a similar impression as 4.2(b). These histograms therefore illustrate that one's impression can be influenced by both bandwidth and cut-off points. Fig. 3(d), (e) and (f) show density estimates for different bandwidths. Plot 4.2(d) was produced by the R-software default bandwidth; plots (e) and (f) were produced with $^1/_4$ and $^1/_2$ of the default bandwidth respectively. It is clear that larger bandwidth makes a smoother estimate of f; smaller bandwidth makes it rougher.

More observations can be made using Fig. 3(d), (e) and (f). First, for a given bandwidth, the number of observed modes determines the number of individual structures in the density estimate. Secondly, the mode density is an indication of the compactness of the related structure. Thirdly, the difference between the mode density and the saddle point density indicates separation of the observed structure. Lastly, when the mode density is equal to the saddle point density, the observed structures are merged into a new one.

Clustering using the NP estimation of the data density is attained by recognizing local maxima and their basins of attractions in the multivariate surface of the data density function [14]. All the data found in the basin of attraction of a mode will form a separated cluster. Contour plots are used to view the surface of data. They show lines of constant surface values similar to topographical maps. Using MATLAB a surface with peaks and depressions that can be used to illustrate contour plots for the Blue band data is shown below.

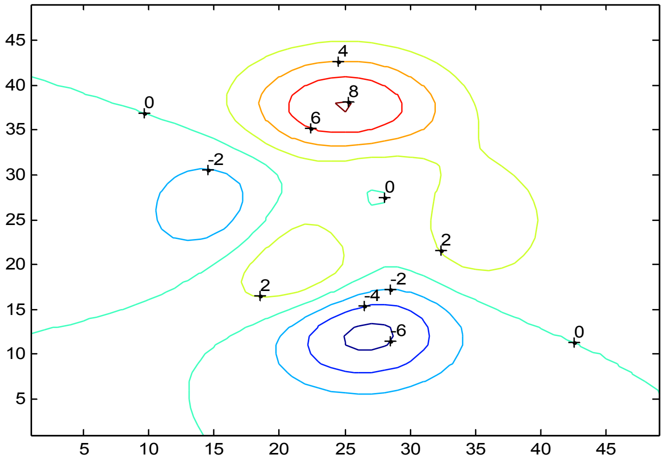

Fig. 4. Contour plot of the peaks function. The peaks make it easier to understand the hills and valleys in the surface

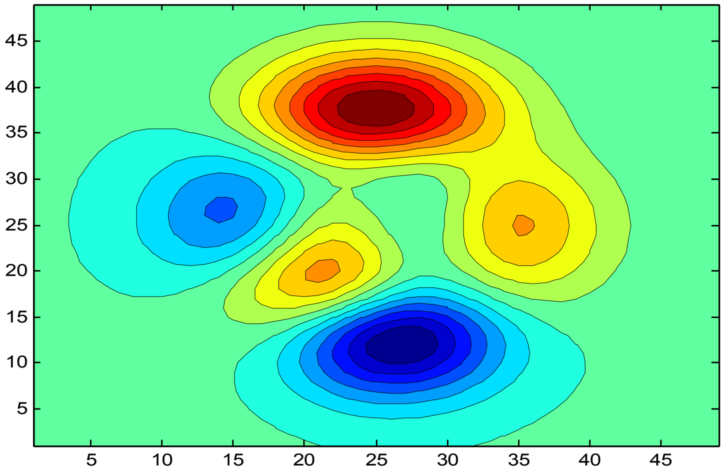

Fig. 5. Filled contour plot of the peaks surface. It is created using the contourf function

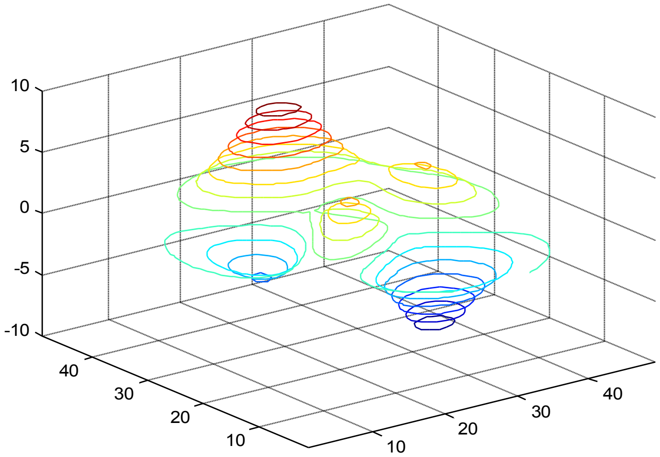

Fig. 6. Contour plots for the blue band data

7 Conclusions

Estimation of the density function in the NP set-up involves highly structured problems and also faced with the curse of dimensionality. The squared error risk, L_2 of an estimator is used to minimize the error of estimation. The L_2 is a function of both f and $\hat{F}$ hence, it is difficult to find the best estimate which minimizes the risk function since it relies on an unknown parameter. Effort has been made to use MiniMax estimators but again calculating the exact MiniMax estimator for $F(f)$ is difficult. Thus, the requirement is slightly relaxed to obtain MiniMax rate estimators. The MiniMax polynomial is therefore used to smooth the risk function at the origin. The presence of the polynomial factor in the approximate can be worse in the tail areas. The Edgeworth expansion technique is used to give clear and almost accurate results.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

References

[1] Birge JR, Murty KG. Mathematical programming: State of the art 1994 (No. CONF-9408161-). Univ. of Michigan, Ann Arbor, MI (United States); 1994.

[2] Cai T, Low M. Testing composite, Hermite polynomials, and optimal estimation of a non-smooth functional. Ann Statist. 2011;39(2):1012-1041.

[3] Bergstrom R. The estimation of nonparametric functions in a Hilbert space. Economic Theory. 1985;1:7-26.

[4] Wasserman L. All of nonparametric statistics. Department of Statistics, Carnegie Mellon University, Pittsburgh, USA; 2006.

[5] Casella G, Berger RL. Statistical Inference, 2nd Edition, Wadworth Group, New York; 2002.

[6] Polonik W. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. The Annals of Statistics. 1995;23(3):855-881.

[7] DasGupta A. Asymptotic theory of statistics and probability. Springer, New Yolk; 2008.

[8] Donoho DL, Liu RC. Geometrizing rates of convergence II. Ann. Statist. 1991;19:633-667.

[9] Ibragimov IA, Khasminski R. Asymptotic normal families of distributions and effective estimation. Ann. Statist. 1991;19:1681-1724.

[10] Daniels HE. Tail probability approximations. International Statistical Review. 1987;55:37-48.

[11] Lugannani R, Rice S. Saddle point approximation for the distribution of the sum of independent random variables. Advances in Applied Probability. 1980;12(2):475-490.

[12] Hall P. The bootstrap and Edgeworth expansion. Springer Science & Business Media; 2013.

[13] Lepski O, Nemirovski A, Spokoiny V. On estimation of the L r norm of a regression function. Probability Theory and Related Fields. 1999;113(2):221-253.

[14] Chaudhuri P, Marron JS. Scale space view of curve estimation. Annals of Statistics. 2000;408-428.

[15] Debnath L, Mikusinski P. Introduction to Hilbert spaces with applications. Academic Press; 2005.

[16] Le Cam L. Asymptotic methods in statistical decision theory. Springer Science & Business Media; 2012.

[17] Nyambega HO, Orwa GO, Mung'atu JK, Odhiambo RO. Bayesian estimation of survivor function for censored data using lognormal mixture distributions; 2017.

[18] Lehmann EL, Casella G. Unbiasedness. Theory of Point Estimation. 1998;83-146.

[19] Korostelev AP, Tsybakov AB. Minimax theory of image reconstruction. Lecture Notes in Statist. Springer; 1994.

[20] LeCam L. Convergence of estimates under dimensionality restrictions. The Annals of Statistics. 1973;1(1):38-53.

[21] Rivlin T. Chebyshev polynomials from approximation theory to algebra and number theory. Pure Appl. Math. (NY); 1990.

[22] Dekking FM, et al. A modern introduction to probability and statistics. Springer-Verlag, New York; 2010.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle3.com/review-history/50976