ON CHEBYSHEV'S POLYNOMIALS OF NON-SMOOTH FUNCTIONAL AND ITS NONPARAMETRIC ESTIMATION

MOSES KOLOLI MUKHWANA, ROMANUS O. OTIENO, ORWA O. GEORGE and MUNG'ATU K. JOSEPH
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCES,
JOMO KKENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY, KENYA

N. B. OKELO*
SCHOOL OF MATHEMATICS AND ACTUARIAL SCIENCE,
JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY,
P. O. BOX 210-40601, BONDO-KENYA

Abstract

In statistical inference, one of the basic problems is that of estimating functionals. This problem is considered in the nonparametric set-up. The quality of estimation depends on smoothness properties of the functional F. However, non smooth functionals lack some degree of properties traditionally relied upon in estimation. This highlights the reason why standard techniques fail to yield sharp results. In estimating non smooth functionals, the lower and upper bounds are constructed for the MiniMax Risk. When working in the context of MiniMax estimation, the lower bounds are important. A single- value MiniMax lower bound is established by applying the general lower bound technique based on testing two composite hypotheses. A vital step is the construction of two special priors and bounding the chi-square distance between two normal mixtures. An estimator is constructed using approximation theory and Hermite polynomials and is shown to be asymptotically sharp MiniMax when the means are bounded by a given value.

Introduction

In statistical inference, a sample is obtained from the population. A population is a set of units (usually people, objects, transactions or events) of interest in a study. It is the entire group of interest. The statistic obtained from a sample is used to estimate the population parameters. This statistic is called an estimator. Populations are characterized by parameters such as the mean and the variance. The corresponding quantities for a sample are called statistics. A statistic used to estimate the values of a parameter is called an estimator. An estimator is a function of the sample, while an estimate is the realized value that is obtained when a sample is actually taken. An estimator is denoted by a capital letter while the estimate is denoted by a small letter. A desirable estimator is the one which has a high probability of being near to the unknown parameter which it estimates. This is quantified by evaluating the probability that the estimator lies within a given range of the parameter. In many practical cases, it may not be possible to have all the information available about a certain population. For example, the distribution or certain parameters may be unknown. However, some information about a sample from a given population may be available. The information from the sample is therefore used to infer information about a population. For example, in a normal distribution, the probability of a measurement occurring which is more than two standard deviations away from either side of the mean is approximately one in twenty. The probability of a measurement occurring which is more than three standard deviations away from either side of the mean, is approximately three in a thousand. The likelihood of such a measurement to occur is therefore very small. If such a measurement is found, the event is said to be unusual and it is referred as significant (S.M. Uppal et al, 2012). Nonparametric statistical procedures are applied when the distribution from which the sample is drawn is unknown (Lepski et al, 1999). If the distribution is known, parametric statistical procedures are used to obtain the estimators. These methods make more assumptions and if correct, accurate and precise estimates are produced (Bagdonavicious et al, 2011). However, if incorrect, these methods can be misleading. For this reason, they are not considered robust. Nonparametric methods use less information in their calculation and can be used on all types of data which are nominally scaled or are in rank form as well as interval or ratio scaled. Nonparametric procedures are easy to apply on a small sample size, which would demand the distributions to be known precisely in order for parametric tests to be applied. Furthermore, nonparametric tests often concern different hypotheses about population than do parametric tests (J.P Marques De sa', 2007). Also in parametric statistics, the mean and standard deviation are two most commonly used to describe a normal distribution. However, they are of little value to exploratory data analysis since they are affected by extreme values and are not easily understood by people with less knowledge in statistics. Nonparametric statistics is therefore an excellent supplement to the conventional mean and standard deviation for the communication of statistical information to a non technical audience (Corder and Foreman, 2009).Estimators are obtained through a number of approaches such as Point estimation and Interval estimation. Point estimation involves the use of sample data to calculate a single value (statistic) which is to serve as a 'best guess' or 'best estimate' of an unknown (fixed/random) population parameter. The methods used in Point estimation include: Method of moments, Maximum likelihood, Bayes' estimators and best unbiased estimators. Interval estimation provide an interval within which the parameter should lie with certain degree of certainty. It is a statistical procedure that specifies statistical methods of using sample information to calculate two values $C1$ and $C2$ that forms the end points of the interval. The two limits calculated from the sample observations $C1(x_1, x_2, ..., x_n)$ and $C_2(x_1, x_2, ..., x_n)$ and $Pr[C_1 \leq \theta \leq C_2] = 1 - \alpha$. Where α is usually small, that is 0.05, 0.01 or 0.001. A confidence coefficient is $1 - \alpha$. A confidence interval affords more information about the unknown parameter than an estimate because the confidence coefficient and interval width give an indication of how close the estimator is close to the parameter.

Methods and Techniques

The following theorem (Weierstrass Approximation Theorem) plays an important role in the approximation of functions. The theorem states that any continuous function can be approached as close as possible with polynomials, assuming that the polynomials can be of any degree. The theorem is formulated in $L\infty$ form and it also holds in the L_2 sense. Let π_n denote the space of polynomials of degree $\leq n$. The Weierstrass Approximation Theorem states that: Let f(x) be a continuous function on [a, b]. Then there exists polynomials $P_n(x)$ that converges uniformly to $f(x)$ on [a, b] that is, $\forall\epsilon > 0$, there exists an $N \in \mathrm{N}$ and polynomials $P_n(x) \in \pi_n$, such that $\forall x \in [a, b]$ $|f(x) - p_n(x)| < \epsilon \ \forall n \geq N$ At the interpolating points, the error between the function and the interpolating polynomial is zero. However, between the interpolating points, the error between the function and the interpolating polynomial gets worse for higher order polynomials. This is known as Runge's phenomenon. This is a problem of oscillation at the edges of an interval that occurs when using polynomial interpolation with polynomial points. This shows that going to higher degrees does not always improve accuracy. According to Weierstrass theorem, it is expected that using more points would lead to a more accurate reconstruction of f(x). However,

the polynomial functions are not guaranteed to have the property of uniform convergence. The theorem only shows that a set of polynomial functions exists,
but it does not give a general method of finding one.

Polynomial of Best Approximation

Assume that the function f(x) is continuous on [a,b], and assume that $P_n(x)$ is a polynomial of degree $\leq n$. $L\infty$ is the distance between f(x) and $P_n(x)$ on the interval [a, b] given by $\|f - p_n\|_\infty = \max_{a\leq x\leq b} |f(x) - p_n(x)|$ …………………………………………………….. (3.1)

Polynomials with an arbitrary large distance from f(x) (in L_∞ *sense*) can be constructed. It is important to address how close to get to f(x) with polynomials of a given degree. Define $d_n(f)$ as the infimum of (3.1) over all polynomials of degree $\leq$ *n* that is, $d_n(f) = \inf_{P_n\in\Pi_n} \|f - p_n\|\infty$ …………………………………………………………………….(3.2)

The goal is to find a polynomial $p_n^*(x)$ for which the infimum of (3.2) is actually obtained, that is, $d_n(f) = \|f - p_n\|\infty$ ………………….…………………………………………………… (3.3)

A polynomial $p_n^*(x)$ that satisfies (3.3) is referred as the polynomial of best approximation or the minimax polynomial. It's minimal distance is referred as the minimax error.

The Chebyshev Polynomial

To approximate a continuous function f on an interval [a,b] the minimax approximation need to be considered. The minimax polynomial approximations exist and they are unique when f is continuous, although they are not easy to compute. Therefore a more effective approach to consider is to use a near minimax approximation based on the Chebyshev polynomial. The Chebyshev polynomials are orthogonal. Orthogonal polynomials can be used to make the polynomial coefficients uncorrelated and minimize the sensitivity of calculations to roundoff error. Two polynomials P_i and P_j are orthogonal if $P_i(x)$ and $P_j(x)$ are uncorrelated as x varies over the same distribution. They also have the property of bounded variation. The local maxima and minima of Chebyshev polynomials on [-1,1] are exactly equal to 1 and -1 respectively regardless of the order of the polynomial. Chebyshev polynomials have the largest possible leading coefficient, but subject to the condition that their absolute value is bounded on the interval by 1. These makes them important in the approximation theory. The Chebyshev polynomials are used because their roots of the first kind , which are also called Chebyshev nodes, are used as nodes in polynomial interpolation. Polynomial interpolation provides an approximation that is close to the polynomial of best approximation to a continuous function under the maximum norm. The Chebyshev polynomial (of the first kind) of degree k is defined as $T_k(x) = \sum_{j=0}^{\frac{k}{2}} (-1)^j \frac{k}{k-j} \binom{k-j}{j} 2^{k-2j-1} x^{k-2j}$ …….………………………………………………… (4.1)

(Cai and Low, 2011) Rivlin in 1974 showed the following expansion

$$|x| = \frac{2}{\Pi} T_0(x) + \frac{4}{\Pi} \sum_{k=1}^{K} (-1)^{K+1} \frac{T_{2k}(X)}{4K^2 - 1} \quad \text{……………......……………………......…………………… (4.2)}$$

where T_{2k} (x) is the Chebyshev polynomial of degree $2k$. Truncating equation 3.5, the following expansion is obtained

$Y_k(x)=\frac{2}{\Pi}T_0(x)+\frac{4}{\Pi}\sum_{k=1}^{K}(-1)^{K+1}\frac{T_{2k}(X)}{4K^2-1}$ ……………………………………………………….(4.3)

$Y_k(x)$ can be written as $Y_k(x)=\sum_{k=1}^{k} y_{2x}x^{2k}$ ……………………..…………………………...(4.4)

Suppose (3.7) is the best polynomial approximation of degree $2k$ to $|x|$ and

$Y_k(x)$ = $\frac{2}{\Pi}T_0(x)+\frac{4}{\Pi}\sum_{k=1}^{K}(-1)^{K+1}\frac{T_{2k}(X)}{4K^2-1}$ then $\max_{x\notin[-1,1]}\left|Y_k^*(x)-|x|\right|\le\frac{\beta_*}{2k}(1+0(1))$……………….(4.5)

$\max_{x\notin[-1,1]}\left|Y_k^*(x)-|x|\right|\le\frac{2}{\Pi(2k+1)}$ The equation (3.8) shows the uniform error bounds proved by Bernstein (1913). The coefficients y_{2k}^* and y_{2k} satisfy for all $0 \le k \le K$, $\left|y_{2k}^*\right|\le 2^{3k}$ and $\left|y_{2k}\right|\le 2^{3k}$. The proof of the uniform error bounds on the coefficients y_{2k}^* and y_{2k} is given in appendices. Let the non-smooth functional to be estimated be of the form $\frac{1}{n}\sum_{i=1}^{n}|\theta_i|$ from an observation $Y \sim N(\theta, I_n)$. In estimating this functional, the lower and upper bounds are constructed for the MiniMax Risk. When working in the context of minimax estimation, the lower bounds are important. A single-value minimax lower bound is established by applying the general lower bound technique based on testing two composite hypotheses. For any two priors μ_0 and μ_1, on the parameter space, a lower bound on the expected squared bias with respect to μ_1 under a constraint on the upper bound of the expected MSE with respect to μ_0 is obtained. The lower bound depends on the difference between the expected value of T over each of the priors and over the variance of T under μ_0. It also depends on the chi-square distance between marginal distributions over μ_0 and μ_1. The optimal rates of convergence for estimating linear and quadratic functionals are often algebraic. Let

$L(\theta)$ = $\frac{1}{n}\sum_{i=1}^{n}\theta_i$ ………………………………………………………………………... (4.6)

$Q(\theta)$ = $\frac{1}{n}\sum_{i=1}^{n}\theta^2{}_i$ ……………………………………………………………………………..(4.7)

The parametric rate n^{-1} for estimating $L(\theta)$ can be obtained by $\bar{y}$ and for estimating $Q(\theta)$ can be obtained over $\Theta n(M)$ using the unbiased estimator $\hat{Q}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i^2-1\right)$ (Cai and Low, 2011). The estimator obtained after the singularity being smoothened at the origin by the best polynomial approximation is improved further using the Hermite polynomials to construct an unbiased estimator for each term in the expansion. These polynomials are orthogonal and are uncorrelated when X is standard normal on (−∞,+∞). They are used to construct an unbiased estimator for each term in the expansion.

Results

In this section we consider the bounded case and construct an estimator that relies on the best polynomial approximation and the use of Hermite polynomials. The estimator is then shown to be asymptotically sharp minimax. Optimal estimator construction is involving and this is partly due to the nonexistence of an unbiased estimator for $|\theta_i|$. Our strategy is to smooth the singularity at the origin by a polynomial approximation and construct an unbiased estimator for each term in the expansion using Hermite polynomials. A drawback of using y_k^* is that it is not convenient to construct. Therefore an explicit and nearly optimal polynomial approximation Y_k can be obtained using the Chebyshev polynomials. The Chebyshev polynomial of degree k is

$$T_k(x)=\sum_{j=0}^{\frac{k}{2}}(-1)^j\frac{k}{k-j}\binom{k-j}{j}2^{k-2j-1}x^{k-2j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots...(5.1)$$

The following expansion can also be found, see Rivlin (1974)

$$|x| = \frac{2}{\Pi}T_0(x)+\frac{4}{\Pi}\sum_{k=1}^{K}(-1)^{K+1}\frac{T_{2k}(X)}{4K^2-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.2)$$

Where $T_{2k}(x)$ is the Chebyshev polynomial of degree $2k$. The above expression can be truncated to give

$$Y_k(x) = \frac{2}{\Pi}T_0(x)+\frac{4}{\Pi}\sum_{k=1}^{K}(-1)^{K+1}\frac{T_{2k}(X)}{4K^2-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ldots\ldots\ldots (5.3)$$

We can also write $Y_k(x)$ as

$$Y_k(x)=\sum_{k=1}^{k}y_{2x}x^{2k} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5.4)$$

When we consider $M = 1$. The case of a general M involves an additional rescaling step. When $M = 1$, it follows that each $|\theta_i|$ can be well approximated

by $y_k^*(\theta_i)=\sum_{k=0}^{k}y_{2k}^*\theta_i^{2k}$ …………………..…………………………………………………… (5.5)

on the interval $[-1, 1]$ and hence the functional $T(\theta) = \frac{1}{n}\sum_{i=1}^{n}|\theta_i|$ can be approximated by

$$\hat{T}(\theta)=\frac{1}{n}\sum_{i=1}^{n}y_k^*(\theta)= \frac{1}{n}\sum_{k=1}^{k}y_{2k}^*b_{2k}(\theta) \text{ where } b_{2k}(\theta)\equiv\frac{1}{n}\sum_{i=1}^{n}\theta_i^{2k}$$

Note that $\hat{T}(\theta)$ is a smooth functional and we shall estimate $b_{2k}(\theta)$ separately for each k by using Hermite Polynomials. Let ϕ be the density function of a standard normal variable. Recall that for positive integers k, Hermite Polynomial H_k is defined by

$$\frac{d^k}{dy^k}\Phi(y)=(-1)^kH_k(y)\Phi(y) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots...\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5.6)$$

Where H_k is a Hermite Polynomial with respect to ϕ. It is well known that if $X \sim N(\mu, 1)$, $Hk(x)$ is an unbiased estimate of μk for any positive integer k, that is $E_\mu H_K(x) = \mu^k$. Also, $\int H_k^2(y)\phi(y)dy = K!$ and $\int H_k(y)H_j(y)\phi(y)dy = 0$ and define the estimator of $T(\theta)$ by when $k\neq$

j Since $H_k(yi)$ is an unbiased estimate of θ_i^k for each i, we can estimate $b_k(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n}\theta_i^k$ by

$\overline{B}_k = \frac{1}{n}\sum_{i}^{n} H_k(y_i)$ ……………………..…………………………………………………(5.7)

and define the estimator of θ by $\hat{T}_k(\theta) = \sum_{k=0}^{K} y_{2k}^{*}\,\overline{B_{2k}}$ …………………………...................(5.8)

For estimating the functional $T(\theta)$ over the bounded parameter space $\Theta(M)$ for a general $M > 0$, we shall first rescale each θ_i and then approximate $|\theta_i|$ term by term. More specifically, let $|\theta_i| = \frac{\theta_i}{M}$ then $|\Theta_i| \le 1$ for $i =$ 1, . . . , n and

$\left|\Theta_i' - Y_K^{*}(\theta_i')\right| \le \frac{\beta_*}{2k}(1 + 0(1))$ for all $|\theta_i'| \le 1$ Hence, $\left|\Theta_i' - Y_K^{*}(\theta_i')\right| \le \frac{\beta_*}{2k}(1 + 0(1))$ for all $|\theta_i'| \le M$

where $\tilde{Y}_K^{*}(x) = \sum_{k=0}^{k}\tilde{y}_{2k}^{*}x^{2k}$ with $\tilde{y}_{2k}^{*} = \tilde{y}_{2k}^{*}\, M^{-2K+1}$ Again, $H_K(y_i)$ is an unbiased estimate of θ_i^k. We estimate $\overline{B}_{2k} = \frac{1}{n}\sum_{i}^{n} H_{2k}(y_i)$ …………………………………………………………….....(5.9)

and define the estimator of $T(\theta)$ by $\hat{T}_K(\theta; M) = \sum_{K=0}^{K} \tilde{y}_{2k}^{*}\, M^{-2K+1}\,\hat{B}_{2K}$ ………………...……. (5.10)

The performance of the estimator $\hat{T}_K(\theta, M)$ clearly depends on the choice of the cut off K. We choose $K = K_* \equiv \frac{\log n}{\log\log n}$ ………………………………………………………….…(5.11)

and define the final estimator of $T(\theta)$ by

$\hat{T}_{*i}(\theta) = \hat{T}_{K*}(\theta; M) = \sum_{k=0}^{K*} y_{2k}^{*}\,\overline{B_{2k}}$ ……………………………………………………….………….....(5.12)

which gives the desired result.

REFERENCES

1. Bagdonavicious,V. Kruopis,J.,Nikulin,M.S (2011). Nonparametric tests for complete data,Iste and Wiley: London
2. Bernstein S.N. (1913).Sur la meilleure approximation de |*x*| par les polynomes degres donnes.*Acta Math*.**37**, 1-57
3. Bickel,P.J. and Ritov, Y. (1988).Estimating integrated squared density derivatives: sharp best order convergence estimates.*Sankya Ser*.**A 50**, 381-393.
4. Birge,L.,Massart,P. (1995).Estimation of integral functionals of a density. *Ann.Statist.***23**,11-29
5. Cai, T. and Low, M. (2005).Non-quadratic estimators of a quadratic functional.*Ann. Statist.***33**, 2930-2956.
6. Cai, T. and Low, M. (2011).Testing composite, Hermite polynomials, and Optimal estimation of a non smooth functional.*Ann statist*.

7. Corder,G.W and Foreman,D.I (2009). Nonparametric Statistics for Non-statisticiatians: *A step-by-step Approach*,Wiley
8. Donoho, D.L. and Liu, R.C. (1991).Geometrizing Rates of Convergence II.*Ann. Statist*.**19**, 633-667.
9. Ibragimov,I.A,Khasminski,R (1991).Asymptotic normal families of distributions and effective estimation.*Ann. Statist.***19**,1681-1724.
10. Ibragimov,I.,Nemirovski,A.,Khasminski,R. (1986).Some problems on nonparametric estimation in Gaussian white noise.*Theory Probab.Appl.*,**31**, 391-406.
11. Jean,D.G. and Subhabrata,C. (2003). Nonparametric Statistical Inference, 4*thedition*, Marcel Dekker INC, New York.
12. Kaplan,E.L and Meier, P, (1958).Nonparametric estimation from incomplete observation,*American Statistical Journal*, 458-479.
13. Korostelev,A.P.,Tsybakov,A.B. (1994).Minimax Theory of Image econstruction. Lecture Notes in Statist.Springer
14. Lehmann E.L. and Casella G. (1998). Theory of Point Estimation. Springer Texts in Statistics. Springer-Verleg, New York, Second edition.
15. Lepski,O.,Nemirovski,A. and Spokoiny, V. (1999). On estimation of the Lr norm of a regression function.*Probab. Theory Relat. Fields***113**, 221-253.
16. Le Cam, L. (1973). Convergence of estimation under dimensionality restrictions.*Ann. statist*.**1**, 38-53.
17. L.Debnath and P. Mikusinski, Introduction to Hilbert Spaces with Applications, (Boston: Academic Press,1990).
18. Spokoiny,V. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics*, **26**, 2477-2498.
19. Terry Rockafellar, Mathematical Programming: State of the Art 1994 (J.R.Birge and K.G. Murty, editors), University of Michigan Press, *AnnArbor*, 1994, 248-248
20. Thomas,S.F.,(1967). Mathematics Statistics: A decision Theoretic Approach. Probability and Mathematical Statistics,**Vol.1**. Academic press, New York.