ON MINIMAX RISK OF NON-SMOOTH FUNCTIONAL, ITS ASYMPTOTIC PROPERTIES AND POLYNOMIAL 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

Nonparametric estimation of non-smooth functionals deals with highly structured problems which arise, modeled or cast differently from the ones for which mainline numerical methods have been designed. Non-smooth functional estimation problems show some features that are different from those of estimating smooth functionals. This is in terms of the optimal rates of convergence as well as the technical tools needed for the analysis of the MiniMax lower bounds and the construction of the optimal estimators. The main difficulty of estimating the non-smooth functionals is traced back to the non differentiability of the absolute value function at the origin. This is reflected both in the derivation of the lower bounds and the construction of optimal estimators. The construction of the optimal estimators of the non-smooth functionals is more complicated than those for linear and quadratic functionals. In this study we consider asymptotic properties, polynomial estimation and MiniMax risk involving non-smooth functionals.

1 Introduction

In making statistical inference for the non-smooth functional, nonparametric procedures are preferred. This is because the distribution from which the MiniMax is drawn is unknown. These methods also 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. The MiniMax Risk is used as bench mark to evaluate the performance of estimators. An estimator that minimizes the maximum risk is the MiniMax estimator (Lehmann and Casella, 1998). When the risk functions are compared, it is seen that neither risk dominates the other. This highlights the need to compare risk function. To do so, a one-number summary of

the risk function is required. Functional estimation plays a major role in the theory of nonparametric function estimation (Cai and Low, 2011). It is a mathematical relation which maps one or more functions in one number.

2 An estimator for the non smooth functional

An estimator is a function of the observations, what gives rise that the estimator results in a random variable. The quality of an estimator is therefore given in probabilistic terms. A desirable estimator is the one which has a high probability of being near to the unknown parameter which it estimates. A single number often used to estimate a population parameter is of little value. It is therefore helpful to know a set of values normally given as a reasonable set of values for that parameter (Uppal et al, 2012). The information relating to some magnitudes of the random variables in the sample is used in making NP inferences. The actual observations can be replaced for example with their relative rankings within the sample and the probability distribution of some function of these sample ranks determined by postulating only general assumptions about the basic population sampled, this function provides a NP technique for estimation or hypothesis testing (Jean D. G. and Subhabrata C., 2003). The NP and parametric hypotheses are analogous, both relating to location, and identical in the case of continuous and symmetrical population. In statistical inference, performance is a matter of concern. However, generalizations about reliability are always difficult because of various factors like the size of the sample, significance level, cost and non existence of a definite and universely acceptable criterion for good performance. According to (Box and Anderson, 1955), the needs of the experimenter are fulfilled if the statistical criteria:

• Is powerful. That is the criteria are sensitive to the change in the specified factors tested.

• Is robust. That is the criteria are insensitive to changes of a magnitude likely to occur in practice, in extraneous factors.

Parametric tests are derived in such a way that they satisfy the first requirement. However, since such tests are not valid unless the assumptions are met, robustness is of concern in parametric statistics. The nonparametric tests are robust because their construction requires only general assumptions. It is therefore good to look at robustness as the performance criterion for parametric tests and power for nonparametric tests. Calculations of power for any test require information of the probability distribution of the test statistic under the alternative. The alternative in NP problems is general. When the assumptions are met, many of the classical parametric tests are known to be most powerful. The NP tests are almost as powerful especially for small samples and are considered desirable whenever there is any doubt about assumptions (Jean and Subhabrata C. 2003). Two limits

calculated from the sample observations $L_1(x_1, x_2 ..., x_n)$ and $L_2(x_1, x_2 ..., x_n)$ and $P[L_1 \leq \theta \leq L_2] = 1 - \alpha$ where α is usually small 0.05, 0.01 or 0.001. This $1 - \alpha$ is the confidence coefficient. The confidence coefficient has more information about the unknown parameter than an estimate. This is because the confidence coefficient and interval width give an indication of how close the estimator is to the parameter. The bias and variance are important functions associated with any estimator, θ. These functions are useful on how well the estimator is doing. They have a useful relationship between them with the mean square error. The mean square error decomposes into a bias and a variance terms. $E_\theta\left(\left(\theta - \hat{\theta}\right)\right) = \beta_{\hat{\theta}}(\theta)^2 + V_{\hat{\theta}}(\theta)$. The MSE of an estimator can therefore be described as a sum of a term measuring how far off the estimator is on average and a term measuring the variability of the estimator. It is shown that in estimating functionals based on iid, bounds on Mini-Max estimation based on testing two composite hypotheses, (Cai and Low, 2011). Composite hypotheses are common in problems where constraints, presumed convexities, stochastic dominance etc may lead to one or more inequalities. These hypotheses do not point to a unique probability measure to be used in hypotheses testing, this makes it more challenging to test composite hypotheses than simple hypotheses. If two priors say μ_0 and μ_1 are used to obtain a lower bound on the expected mean squared error (MSE) with respect to μ_0. The lower bound depends on the difference between the expected value of T over each of the priors and also on the variance of T under μ_0 (Cai and low, 2011).The bound also depends on the Chi-square distance between the two marginal distributions of the observations, one over μ_0 and the other over μ_1. The difficulty of composite testing problem was shown in (Le Cam, 1973 and 1986) to depend between convex hulls of the two composite hypotheses. Using the prior's μ_0 and μ_1 it is seen as picking points in the convex hulls of the two subsets of the parameter space and the bounds on their risk can be shown. When these priors are chosen carefully, sharp MiniMax lower bounds for estimating l_1 norm of the means of normal random variables are obtained. Techniques used in (Lepski et al, 1999) focus on estimating the l_1 norm of a regression function with a bound given for the Kullback-Leibler while those found in Cai and Low, 2011, it is seen that a chi-square distance bounded. However, it is not easy to provide good bounds directly for the Kullback-Leibler distance. This can be shown using cases which correspond to parameter spaces with increasing bounds. The lower bounds provided by Kullback-Leibler can only be used in cases where the parameter space has fixed bounds.(Cai and Low, 2011).

3 Nonparametric Estimation

Most general methods of estimation, such as minimum chi-square or maximum likelihood, may be interpreted as procedures for selecting from known class of distributions one which, in a particular case, best fits the observations (Kaplain and Meier, 1958) and (Wolfowitz, 1942). These methods make more assumptions and if correct, accurate and precise estimates are produced (Bagdonavicious et al, 2011). The most frequently used methods of parametric estimation for distributions of lifetimes are perhaps fitting of a normal distribution to the observations or their logarithms by calculating the mean or variance (Kaplain and Meier, 1958). Such Assumptions are advantageous if correct; the estimates are simple and relatively efficient. However, if incorrect, these methods can be misleading. For this reason, they are not considered robust. If the distribution is unknown, nonparametric statistical procedures are used (Lepski et al, 1999). These 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. They are easy to apply on a small sample size (Bagdonavicious et al, 2011). The assumption most frequently required is that the population is continuous. More restrictive assumptions are some sometimes made, for example, that the population is symmetrical. The information used in NP inferences generally relates to some functions of the actual magnitudes of the random variables in the sample, this function provides a distribution-free technique for estimation or hypothesis testing. Box and Anderson (1955) state that to fulfill the needs of the experimenter, statistical criteria should:

1. Be sensitive to change in the specific factors tested (power);
2. Be insensitive to the changes of a magnitude likely to occur in practice, in extraneous factors (robust)

Parametric tests are derived in such a way that the powerful criteria are satisfied for an assumed specific probability distribution. On the other hand, NP tests are robust because their construction requires only general assumptions. The mean and standard deviation are two parametric statistics that are 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. NP 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).

4 Polynomial Approximation

A polynomial is a function that can be written in the form $P(x) = c_0+c_1x+...+c_nx^n$ with some coefficients $c_0, ..., c_n$. If $c_n \neq 0$, then the polynomial is said to be of order n. A first order (linear) polynomial is the equation of a straight line, while the

second order (quadratic) polynomial describes a parabola. Polynomials are mathematical functions that exist, requiring multiplication and additions for their evaluation. They also have the flexibility to represent general nonlinear relationships (G.K. Smyth, 1998). The purpose of polynomial approximation in statistics is to approximate a difficult to evaluate function, such as a density or a distribution function, with the aim of fast evaluation on a computer. Here the interest is not on the curve but on how closely the polynomial can follow the special function, and how small the maximum error can be made. The function is first transformed so as to make it more amenable to polynomial approximation. The orthogonal polynomials can be used to make the polynomial coefficients uncorrelated, to minimize the sensitivity of calculations to round off error (G.K. Smyth, 1998). Two polynomials P_i and P_j are said to be orthogonal if $P_i(x)$ and $P_j(x)$ are uncorrelated as X varies over some distribution. For instance, Hermite polynomials are uncorrelated when X is standard normal on $(-\infty,+\infty)$. The orthogonal polynomials changes sign (and has a zero) n times in the interval of interest

5 Asymptotic Properties

These are properties of estimators which hold as n increases. For example, unbiased, sufficiency, consistency, efficiency and minimum variance unbiased. In statistical inference, the standard error of the mean is given by $\frac{\delta^2}{n}$ .This is a standard deviation of the sampling distribution and it measures the precision of any estimator. The smaller the standard error of the sampling distribution the greater the precision. A statistic with the propertity $E(\bar{x}) = \mu$ is said to be unbiased. An unbiased estimate is not necessary a good one. The distribution of the means of samples of say size 100 has a smaller variance $\frac{\delta^2}{100}$ than that of a distribution of size 10, $\frac{\delta^2}{100}$. In many cases the Minimum Variance Unbiasness cannot be obtained. The estimator which is asymptotically MVU is usually obtained. That is the estimator is MVU for large n. Such estimators are useful when n is large and will often be good when n is small.

An unbiased estimator is measured relative to the square of standard error of the best unbiased estimator. If the squared standard estimator error of one unbiased estimator is given by $\frac{\delta^2}{n}$ and the squared standard error of the best unbiased estimator $\frac{\delta^2}{2n}$. Then the efficiency of the first estimator is defined as $E_f = \frac{\delta^2}{2n} \Big/ \frac{\delta^2}{n} =$

$\frac{1}{2}$ It is not really necessary for T to be unbiased provided that most of its probability distribution is near θ. A measure of the tendency of T to be displaced from θ is given by $E(T-\theta)^2$ which is called the mean square error (MSE). Thus an alternative to the unbiased estimator is to find an estimator which has the minimum MSE among the set of possible estimates. $MSE = E(T-\theta)^2 = E[[\{T - E(T)\} + \{E(T) - \theta\}]^2] = V(T) + \{E(T) - \theta\}^2$ This equation shows that MSE have a minimum MSE. An estimate T of θ is the probability of $Pr(|T-\theta|) < \epsilon \rightarrow 1$ as $n \rightarrow \infty$ for $\epsilon > 0$. If T has zero bias and variance $\rightarrow 0$ as n increases then T is consistent. One possible criterion for optimality of estimators is their maximum risk in certain classes of functions, leading to MiniMax estimators. Numerous asymptotic MiniMax results concerning the rate of convergence can be found in the literature. It turns out that most commonly used estimators (kernel, spline, orthonormal and wavelet) can be tuned to achieve such an optimal rate of convergence. This suggests that emphasis on the rate of convergence is often too weak to find a method which is 'best'. Recent research focus on both the optimal constant in the asymptotic MiniMax Risk. This has been inspired by the work of (Pinker, 1980) density estimation, nonparametric regression with Gaussian errors, (Nussbaum,1985), which was extended to non-normal error distribution (Golubev and Nussbaum, 1990). Recently, the results of this type were derived in the context of testing Composite Hypotheses, Hermite Polynomials and Optimal Estimation of a Non-smooth Functional (Cai and Low, 2011). It has been shown that the asymptotic MiniMax Risk can be obtained by optimally tuned polynomial approximation.

6 MiniMax Risk

An estimator is called MiniMax if its maximal risk is minimal among all estimators. This is an estimator which performs best in the worst possible case allowed in the problem (Donoho and Liu,1991) and (Ibragimov and Khasminski,1991). Let $\hat{}\theta = \hat{}\theta(Xn)$ be an estimator for the parameter $\theta \in \Theta$ and let its loss function be $L(\theta, \theta)$. The loss function of an estimator, measures how good is the estimator. Examples of loss function include:

$L(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$ -Square error loss

$L(\theta,\hat{\theta}) = |\theta - \hat{\theta}|^2$ absolute error loss,

$L(\theta,\hat{\theta}) = |\theta - \hat{\theta}|^2$ p-Lp- loss

$L(\theta,\hat{\theta}) = \int \log\left(\frac{p(x;\theta)}{p(x;\theta)}\right) p(x;\theta)dx$ -Kullback-Leibler loss

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector then some common loss function are

$$L(\theta,\hat{\theta}) = \left|\theta-\hat{\theta}\right|^2 = \sum_{j=1}^{k}\left(\hat{\theta}_j-\theta_j\right)^2$$

$$\left|\theta-\hat{\theta}\right|^2 p = \left(\sum_{j=1}^{k}\left|\hat{\theta}_j-\theta_j\right|^2 p\right)^{\frac{1}{p}}$$

The risk of an estimator θ is $R(\theta,\ \hat{\theta}) = E_\theta\left(\theta-\hat{\theta}\right) = \int L\left(\theta,\hat{\theta}(x_1,...,x_n)\right) = p(x_1,\ ...,\ x_n,\ \theta)dx$

When the loss function is a squared error, the risk is just the MSE (mean squared error) $R\left(\theta,\hat{\theta}\right) = E_\theta\left(\theta-\hat{\theta}\right)^2 = \mathrm{var}_\theta\left(\hat{\theta}\right) + bias^2$. The minimax risk is $R_n = \inf\sup R\left(\theta,\hat{\theta}\right)$ where the infimum is over all estimators. An estimator θ is a minimax estimator if $sup\ R\left(\theta,\hat{\theta}\right) = infsup\ R\left(\theta,\hat{\theta}\right)$(Lehmann and Casella, 1998).

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.