JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ALGORITHM TO OBTAIN CLOSEST NON-INCREASING DENSITY ESTIMATOR TO AN ARBITRARY DENSITY ESTIMATOR

Pitambar Y. Patil,
Associate Professor of Statistics,
Devchand College, Arjunnagar (via-Nipani), Dist:- Kolhapur-591237, India
Email:- pypatil_stats@rediffmail.com

Abstract: We came across estimation of an unknown probability density function by using random sample from it. This unknown density function itself may have some known constraints and hence it is trivial to expect the same constraints on its density estimator too. In this article we assume that density estimator is given or already estimated but it may not satisfy requirement of non-increasing property as its domain moves away from its modal point. Hence, we provide an algorithm to obtain closest non-increasing function to a given density estimator closest under some general norm.

Keywords: Density estimator, Pool Maximum Violation Region Algorithm (PMVRA), Distance function.

I. Introduction

Many times we came across estimation of unknown density function [8]. Though density function is unknown, it may have some known restrictions on it, like symmetry or unimodality or decreasing nature as its domain moves away from its modal value etc [1, 3, 5, 6]. Therefore, these will be the trivial expected constraints on its estimators too.

When there is a constraint of decreasing or non-increasingness nature on a density estimator on its positive support but if an estimator does not satisfy this constraint then we have to modify it in such a way that it becomes non-increasing as well as closet to a given density estimator. Here we assume that an arbitrary nonparametric density estimator of an unknown density is already provided for us [2, 4, 8]. In the next section we propose a general algorithm to obtain closest non-increasing function to a given density estimator closest under some general norm.

II. Pool Maximum Violation Algorithm (PMVRA)

Let $f(x) : R^+ \rightarrow R^+$ be an arbitrary function. In the following, for notational convenience we refer the *non-increasing property of a function* as P. Let V (f) = {x : f(x) is strictly increasing at x}, the set on which f(.) violates P. An interval (u, v) is said to be violation interval of f(.) if the function f(.) does not satisfy P on the interval (u, v). An interval (a, b) is said to be a *maximal violation interval* of the function f(.) if f(b) - f(a) ≥ f(b`) - f(a`), for any other violation interval (a`, b`). We modify the function f(.) to a function f_1(.) by modifying the function f(.) by a suitable constant L on maximal violation interval of f(.). However, such constants at each stage of modification can be selected optimally under a specific norm d(.) (distance function). If f(.) does not violates P then there is no need to modify f(x). If f(.) violates the property P, then there exists at least one violation interval. We assume that f(.) satisfies the following conditions:

(i) f(x) is finite.

(ii) $\int_0^\infty f(x)dx < \infty$, ($\int_0^\infty f(x)dx = 1$, (without loss of generality)).

(iii) f(x) has a finite number of turning points (points of local maxima or minima).

2.1 Development of the algorithm

Following is an algorithm to modify an arbitrary function f(.) to a non-increasing function. This modified function is piecewise constant on the set D, where D is the set on which f(.) is being modified.

Step-1: *Select the function f(.).*

Step-2: *Test the given function f(.) for its violation of P*.

If there is no violation of P (that is, V (f) = Φ) then stop. Else go to Step-3.

Step-3: *Determination of modified function*:

Let (a, b) be a maximal violation interval of the function f(.).

For L, $f(a) \le L \le f(b)$,

let $a_1(L) = \inf\{x : f(x) \le L\}$, $b_1(L) = \sup\{x : f(x) \ge L\}$,

$A(L) = \{x : a_1(L) \le x < \infty, L \le f(x) < \infty\}$, $B(L) = \{x : 0 \le x \le b_1(L), 0 \le f(x) \le L\}$

$D(L) = A(L) \cup B(L) = [a_1(L), b_1(L)]$.

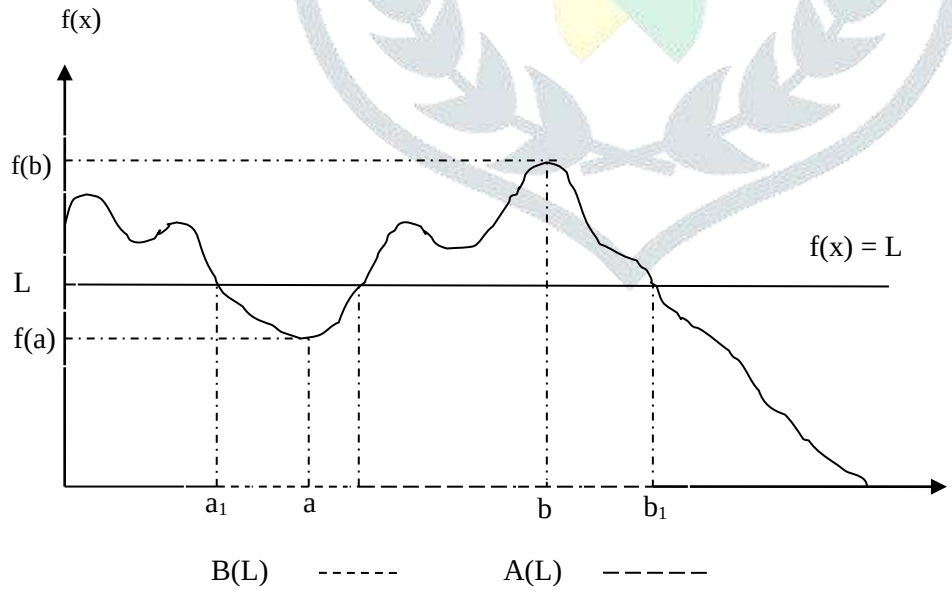
In the following, for notational simplicity we write $a_1(L)$, $b_1(L)$, A(L), B(L) and D(L) as a_1, b_1, A, B and D respectively. Note that, $a_1 \le a < b \le b_1$.

Define:

$$f_1(x, L) = \begin{cases} L, & \text{if } x \in D \\ f(x), & \text{if otherwise} \end{cases} \qquad \textbf{(2.1)}$$

In the above, $f_1(x, L)$ (= $f_1(x)$ say) is the modified function and D is the domain for modification. It is to be noted that, $f_1(x) > L$ for $x \le a_1$ and $f_1(x) < L$ for $x > b_1$. A typical function f(x) and the corresponding sets A(L) and B(L), for some L, (f(a) < L < f(b)) are described in Figure-1.

Figure 2.1: A Typical function with A(L) and B(L) for arbitrary L



Here, the interval (a, b) is *maximal violation interval* of the function f(.).

Step-4: *Identification of two functions on disjoint intervals (if exist)*:

If V (f_1) = Φ then declare that f_1(.) is non-increasing and stop, else identify the two functions on the disjoint intervals (0, a_1) and (b_1,∞) given by:

$f_{11}(x) = f_1(x)$ for $0 < x < a_1$ and $f_{12}(x) = f_1(x)$ for $x > b_1$.

Stop identifying the domain for modification if V (f_{11}) ∪ V (f_{12}) = Φ, else go to Step-1 and replace f(.) by f_{11}(.) and/ or f_{12}(.) as the case may be.

In the above, the choice of constants L's is not unique, however these constants at each stage can be selected optimally so that the resulting modified function $f_1(x)$ is closest to the original function f(.) under a given distance function d(.). The algorithm described above is referred as the *Piecewise Maximum Violation Region Algorithm for the distance function d(.)* (PMVRA-d). In the following, we describe methods to obtain L^* for a given distance measures.

We note that the sets A(L) are decreasing in L, decrease from A(f(a)) to A(f(b)) = Φ, whereas B(L) are increasing in L, increase from B(f(a)) = Φ to B(f(b)). Furthermore, the difference (μ{A(L)} − μ{B(L)}) is decreasing and it has one change of sign at L (= L^* , say), where μ(.) is an appropriate measure of a set; for example the Lebesgue measure.

Corresponding to the maximal violation interval (a, b) we find L^* , (f(a) < L^* < f(b)) such that $f_1(x, L^*)$ is closest (under a norm) to the function f(.). We note that, f_1(.) depends on L and the given function f(.); and hence for given f(.), d(f, f_1), the distance between f(.) and f_1(.) depends only on L. Let

$$\delta(L) = d(f, f_1), \text{ for } f(a) \leq L \leq f(b) \qquad \textbf{(2.2)}$$

Hence, to find the closest function f_1(.) to f(.), it is enough to find L^* such that

$$\delta(L^*) = \text{Infimum}\{\delta(L); L \in [f(a), f(b)]\} \qquad \textbf{(2.3)}$$

Depending upon choices of distance measures d(.), one can obtain respective L^* 's and develop the respective PMVRA-d.

As $f_1(x, L) = f(x)$ for $x \notin D$, we have, $d(f, f_1) = d(f^D, f_1^D)$, where f^D (f_1^D) is the confined function defined on the domain D obtained from f (f_1). Further, as D = A ∪ B and A ∩ B = Φ.

3. Comments and Remarks:

3.1. Performance of PMVRA:

Note that, removal of violation only on [u, v] may result in violation at u and v. But, in the proposed algorithm, we remove violation over [u, v] along with rectification of function over a super set of [u, v]. As such, in PMVRA an iteration removes at least one turning point, and hence the number of iterations to attain the non-increasing property will be lesser than the number of turning points.

3.2. Termination of PMVRA:

To ensure the termination of PMVRA, we assume that f(.) has k, a finite number of turning points. The PMVRA identifies the interval of maximum violation (if any) and on a certain super set of this interval the function is modified by a suitable constant, that depends on the choice of the norm. In the subsequent stage, modification if required will be on a domain excluding the interval of maximum violation. As such, after each modification the domain of the function that needs to be considered reduces very significantly. It is to be noted that there are at most (k - 1) violating intervals for f_1(.). Hence, the algorithm requires at most k iterations.

3.3. PMVRA-d's:

The algorithm described above is referred as the *Piecewise Maximum Violation Region Algorithm for the distance function d(.)* (PMVRA-d). Therefore, one can obtain different L^*'s corresponding to different distance measures d(.)'s, such as Sup-Norm, L_1-Norm and L_2-Norm.

References:

[1] Chaubey Y.P., Li J., Sen A. and Sen P.K.(2012): A New Smooth Density Estimator for Non-negative Random Variables, Journal of the Indian Statistical Association, GOLDEN JUBILEE ISSUE, Vol. 50,(1, 2), pp. 83-104.
[2] Gibbons J.D. and Chakraborti S. (2003): Nonparametric Statistical Inference, IV Edition, Marcel Dekker, Inc.
[3] Lo S. H. (1985): Estimation of a Symmetric Distribution, The Annals of Statistics, Vol. 13(3), pp. 1097-1113.
[4] Randles R. and Wolfe D. (1979) : Introduction to the Theory of Nonparametric Statistics, John Wiley and Sons, New York.
[5] Robertson T., Wright F.T. and Dykstra R.L.(1988): Order Restricted Statistical Inference, John Wiley and Sons, New York.
[6] Schuster E.G. (1975): Estimating the Distribution Function of a Symmetric Distribution, Biometrika , pp. 631 -635.
[7] Schuster E.G. (1991): Minimizing L_1-Distance Between Distribution Functions, Probability and Mathematical Statistics, Vol. 12(2) pp. 265 -270.
[8] Silverman B.W. (1986): Density Estimation for Statistics and Data Analysis, Chapman and Hall, New York.