MJS, Vol. 6, No. 2, July - December 2007, pp. 34-44
ISSN 0975-3311
https://doi.org/10.12725/mjs.11.2

DETECTION OF OUTLIERS THROUGH INFLUENCE FUNCTION ON AFFINITY

P. Rajalakshmi & P. Geetha***

ABSTRACT

Outliers are the atypical observations that lie at abnormal distances from the other observations in a random sample. Such outliers are often seen as contaminating the data. In general, the rejection of influential outliers improves the accuracy of the estimators and so the results with the identification of outliers have become the most important aspect in any data analysis. Outlier detection finds many applications in the areas such as data cleaning, fraud detection, network intrusion, pharmaceutical research and exploration in science data bases. The distance based outlier detection is the most commonly used method. In this paper, the influence function for affinity is explained and the detection of outliers in classification problems using influence function for affinity is illustrated for univariate data through a few examples.

1. Introduction

Outliers in a set of data are the observations which appear to be inconsistent with the remainder of that set of data. They lie at abnormal distance from other observations and arouse suspicions that they were generated by a different mechanism. These outliers may contaminate the data, cause difficulties in the

* *Department of Statistics, Bangalore University*

** *Department of Statistics, Christ College, Hosur Road, Bangalore 560 029. Phone: (0) 98867 56637 Email: geetha.p@christcollege.edu*

attempt to represent the population, distort the estimates of parameters and reduce the discrimination power. As a result the identification of outliers has become an important aspect in data analysis.

The area of diagnostic measures in regression analysis has been widely explored. The article by Cook (1977) had a strong influence on the study of outliers and model diagnostics. The books of Belslely, Kuh and Welsch (1980), Cook and Weisberg (1982) and Atkinson (1981) surveyed the field with applications to regression and other models. Daniel Pena (2005) has proposed a new statistic based on the effect of deletion of one observation on every other observation in the sample. This statistic follows asymptotically normal distribution and it is able to detect a group of high leverage outliers.

Detecting outliers and influential observations in discriminant analysis was first proposed by Campbell (1978) ; he developed the influence function for Mahalanobis distance D^2. This paper served as foundation to many research articles in this field which include Radhakrishnan (1983), Johnson (1987), Critchley and Vitiello (1991), Fung (1992, 1994 and 1995)

Avner Bar-Hen (1996) proposed a preliminary test in Discriminant Analysis that would test whether an unclassified entity x belongs to one of the predefined groups. Wai-Yin Poon (2004) proposed that the regression model diagnostic measures developed from the local influence perspective can be used for identifying observations in a data set that exert undue influence on linear discriminant analysis.

This paper proposes a classification diagnostic measure through affinity of influence function. In Section 2 we define the concepts of affinity and influence function. In Section 3 we explain the procedure for diagnostic measure and illustrate it in the final section.

2. Affinity and Influence Function

Matusita (1955) proposed the concept of affinity to measure the closeness of populations and investigated its statistical applications to problems in areas like decision theory, multivariate analysis and non parametric procedures.

The affinity between the two populations π_1 and π_2 with density functions $f_1(x)$ and $f_2(x)$ respectively is defined by $\rho_{12} = \int \sqrt{f_1(x) f_2(x)}\, dx$.

Affinity measure lies between 0 and 1. If affinity is equal to 1 then the populations are identical. A classification procedure is often defined through a measure of distance and affinity is inversely related to distance. Hence it may be used to measure the distance between two populations.

Influence function for a parameter is obtained by perturbing the distribution by adding a small contribution from a unit mass at the point x and finding the difference between the values of the parameter in the perturbed and unperturbed distributions.

Thus the influence function of x on the parameter θ is $I(x:\theta)=\lim_{\varepsilon\to 0}\frac{\tilde{\theta}(x)-\theta}{\varepsilon}$ where θ is the parameter of the unperturbed distribution, $\tilde{\theta}(x)$, the parameter of the perturbed distribution and ε, a small real value lying between 0 and 1.

When the parameter of interest θ involves more than one population, the theoretical influence function is determined by perturbing only one of the distribution functions and evaluating the parameter $\tilde{\theta}$.

In this paper we deal with the case of only two populations and derive the influence on affinity measure by perturbing each population.

3. Detection of Outliers through Influence Function on Affinity

Suppose that there are two populations π_1 and π_2 with parameters θ_1 and θ_2 respectively. Matusita's affinity ρ_{12} is a function of the parameters of the distribution functions F_1 and F_2 of these two populations. First we perturb the parameter of the first population and evaluate the influence on ρ_{12} which is denoted by $I_1(x:\rho_{12})$. The distribution of $I_1(x:\rho_{12})$ is identified. If the influence on affinity for an observation from this first population is less, then it can be considered as an outlier and can be discarded. If the influence on affinity is more, then the observation x will have more classification power and it can be retained. The cut off value k is determined such that it is the $100\alpha^{th}$ percentile of the distribution of $I_1(x:\rho_{12})$. i.e. $P(I_1(x:\rho_{12})\le k)=\alpha$ i.e k is determined such that

$$\int_{-\infty}^{k} f(I_1(x:\rho_{12}))dx=\alpha$$

where $f(I_1(x:\rho_{12}))$ is the probability density function of $I_1(x:\rho_{12})$. Whenever $I_1(x:\rho_{12})$ is less than or equal to k we can discard the corresponding observation x as an outlier. A graph can be plotted with the observations in the X axis and $I_1(x:\rho_{12})$ in the Y axis. The cut off point k is a straight line parallel to X axis so that the observations lying below k will be treated as outliers. Note that the value of α is decided depending upon the seriousness of the data set up. When α is large more number of observations would be diagnosed as outliers.

Similarly the second population can be perturbed and influence on affinity for an observation $I_2(x:\rho_{12})$ can be obtained. This proposed technique is explained through exponential and uniform distributions in the next section.

While dealing with data we perturb the first population, estimate the parameter, identify the outliers and eliminate them and then the same procedure is followed for the detection and elimination of outliers in the second population also.

4. Examples

4.1 Exponential Distribution

Consider the exponential populations

$$\pi_1 : f_1(x)=\frac{1}{\theta_1}e^{-\frac{x}{\theta_1}} \quad x,\theta_1>0 \text{ and } \pi_2 : f_2(x)=\frac{1}{\theta_2}e^{-\frac{x}{\theta_2}} \quad x,\theta_2>0$$

Without loss of generality, assume that $\theta_1 < \theta_2$.

Then the affinity $\rho_{12} = \int_0^\infty \left[\frac{1}{\theta_1}e^{-\frac{x}{\theta_1}}\frac{1}{\theta_2}e^{-\frac{x}{\theta_2}}\right]^{\frac{1}{2}} dx = \frac{2\sqrt{\theta_1\theta_2}}{\theta_1+\theta_2}$

Let F_1 and F_2 be the distribution functions of π_1 and π_2 respectively and let $\tilde{F}_1$ be a perturbation of F_1. Let $\tilde{\theta}_1$ be the perturbed value of θ_1 under $\tilde{F}_1$.

Then $\tilde{\theta}_1 = (1-\varepsilon)\theta_1 + \varepsilon x$ where $0 < \varepsilon < 1$. The perturbed value of ρ_{12} is obtained by replacing θ_1 by $\tilde{\theta}_1$ in ρ_{12} and it is denoted by $\tilde{\rho}_{12}$.

The influence function of x on ρ_{12} is

$$I_1(x:\rho_{12}) = \lim_{\varepsilon\to 0}\frac{\tilde{\rho}_{12}-\rho_{12}}{\varepsilon} = \frac{\rho_{12}\delta z}{2\theta_1(\theta_1+\theta_2)}$$

where $\rho_{12}=\frac{2\sqrt{\theta_1\theta_2}}{\theta_1+\theta_2}$, $\delta=\theta_2-\theta_1$, $z=x-\theta_1$.

This influence function on affinity $I_1(x:\rho_{12})$ is a linear function of the form $Y_1=aX+b$

where $a=\sqrt{\frac{\theta_2}{\theta_1}}\frac{\theta_2-\theta_1}{(\theta_1+\theta_2)^2}$ and $b=\frac{\sqrt{\theta_1\theta_2}(\theta_1-\theta_2)}{(\theta_1+\theta_2)^2}$.

The pdf of $Y_1=I_1(x:\rho_{12})$ is a two parameter exponential distribution with density function $f(y_1)=\frac{1}{a\theta_1}e^{-\frac{y_1-b}{a\theta_1}}$, $y_1>b$

The cut off value k for the detection of outliers is the $100\alpha^{th}$ percentile of the distribution of $I_1(x:\rho_{12})$ and is $k=b-a\theta_1\ln(1-\alpha)$.

Similarly the influence function on affinity when the second population is perturbed is $I_2(x:\rho_{12})=-\frac{\rho_{12}\delta z^*}{2\theta_2(\theta_1+\theta_2)}$ where $z^*=x-\theta_2$. The cut off value for the outliers is $k=a\theta_1\ln\alpha-b$.

4.2 Uniform Distribution

The following is an example of classification for a non-regular family of distribution with the range spaces of the parent populations being different.

Consider the Uniform populations

$$\pi_1 : f_1(x) = \frac{1}{\theta_1}, 0 < x < \theta_1 \text{ and } \pi_2 : f_2(x) = \frac{1}{\theta_2}, 0 < x < \theta_2$$

Let $\theta_1 < \theta_2$. Then the affinity $\rho_{12} = \int_0^{\theta_1} \frac{1}{\sqrt{\theta_1 \theta_2}} dx = \sqrt{\frac{\theta_1}{\theta_2}}$

When the first distribution is perturbed, the mean of the perturbed distribution is

$\frac{\tilde{\theta}_1}{2} = \frac{\theta_1}{2} + \varepsilon(x - \frac{\theta_1}{2})$. The influence on affinity is $I_1(x : \rho_{12}) = \frac{\rho_{12}}{\theta_1} x - \frac{\rho_{12}}{2}$ which a linear function of the form is $Y_1 = aX + b$ where $a = \frac{\rho_{12}}{\theta_1}$ and $b = -\frac{\rho_{12}}{2}$. Further note that Y_1 follows Uniform distribution with density function $f(y_1) = \frac{1}{a\theta_1}$, $b < y_1 < a\theta_1 + b$.

The cut off value k which is the $100\alpha^{th}$ percentile is obtained as $P(I_1(x : \rho_{12}) \le k) = \alpha$ and is equal to $\rho_{12}(\alpha - .5)$.

When the second distribution is perturbed, the influence on affinity is $I_2(x; \rho_{12}) = -\frac{\rho_{12}}{\theta_2} x + \frac{\rho_{12}}{2}$ and the cut off value k is $\rho_{12}(\alpha - .5)$.

In fact the cut off point for the Uniform distribution is $\rho_{12}(\alpha - .5)$ irrespective of the distribution that is perturbed.

The tables showing the cut off values for different values of α are given below for Uniform and Exponential distributions.

Table 4.1: Exponential Distribution, First Population is perturbed

θ_1	$\alpha = .20$					$\alpha = .30$				
	θ_2					θ_2				
	9	11	12	14	15	9	11	12	14	15
2	-.1907	-.1940	-.1942	-.1927	-.1914	-.1579	-.1607	-.1608	-.1596	-.1585
5	-.1064	-.1350	-.1458	-.1620	-.1682	-.0881	-.1118	-.1207	-.1342	-.1393
8	-.0228	-.0606	-.0761	-.1019	-.1126	-.0189	-.0502	-.0630	-.0844	-.0933

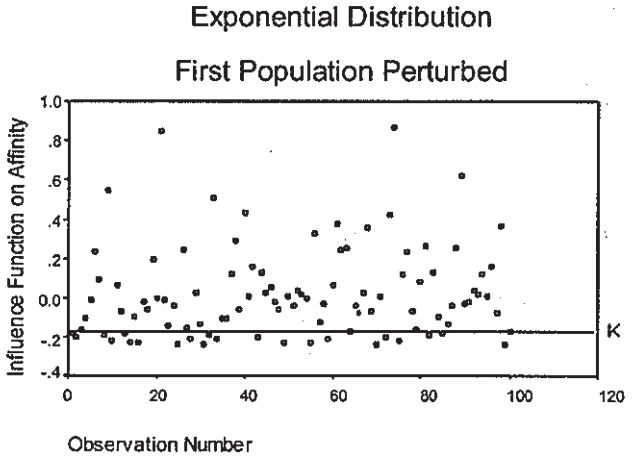
Table 4.2: Exponential Distribution, Second Population is perturbed

θ_1	$\alpha = .20$					$\alpha = .30$				
	θ_2					θ_2				
	9	11	12	14	15	9	11	12	14	15
2	-.1496	-.1522	-.1523	-.1512	-.1502	-.0501	-.0509	-.0510	-.0506	-.0503
5	-.0834	-.1059	-.1143	-.1271	-.1319	-.0279	-.0355	-.0383	-.0425	-.0442
8	-.0179	-.0475	-.0597	-.0800	-.0883	-.0060	-.0159	-.0200	-.0268	-.0296

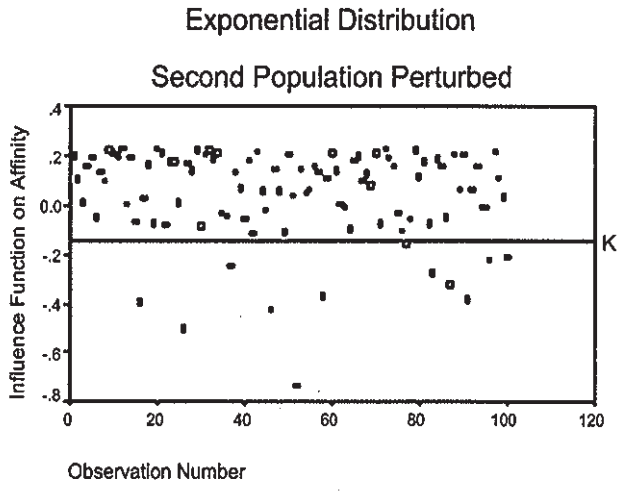
Table 4.3: Uniform Distribution, Either the First or the Second population is perturbed

θ_1	$\alpha = .20$					$\alpha = .40$				
	θ_2					θ_2				
	9	12	14	16	20	9	12	14	16	20
1	-.1000	-.0866	-.0802	-.0750	-.0671	-.0333	-.0289	-.0267	-.0250	-.0224
3	-.1732	-.1500	-.1389	-.1299	-.1162	-.0577	-.0500	-.0463	-.0433	-.0387
5	-.2236	-.1936	-.1793	-.1677	-.1500	-.0745	-.0645	-.0598	-.0559	-.0500

The graphs showing the influence function on affinity for exponential distribution and uniform distribution are given. The points falling below the cut off value k are treated as outliers.



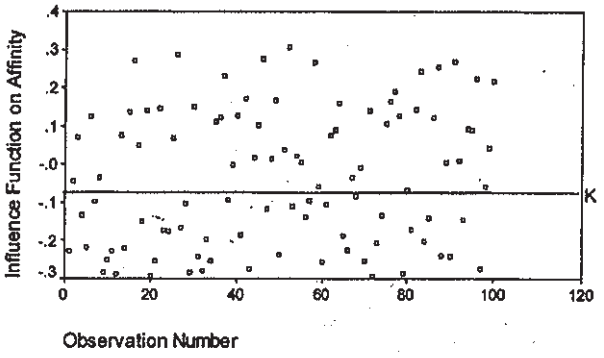
Cutoff Value K for 20th Percentile is -0.1864
$\Theta_1 = 1$ and $\Theta_2 = 9$



Cutoff Value K for 20th Percentile is -0.1463
$\Theta_1 = 1$ and $\Theta_2 = 9$

Uniform Distribution

First Population Perturbed

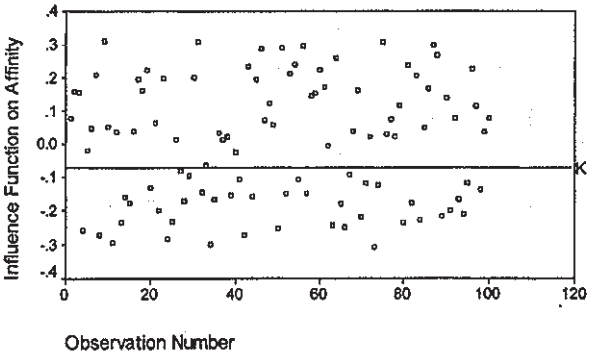


Cutoff Value K for 20th Percentile is -0.1897

$\Theta_1 = 2$ and $\Theta_2 = 5$

Uniform Distribution

Second Population Perturbed



Cutoff Value K for 20th Percentile is -0.1897

$\Theta_1 = 2$ and $\Theta_2 = 5$

Conclusion

In this article we propose the influence function on affinity as a classification diagnostic measure to detect outliers. The measure is obtained theoretically for uniform and exponential distributions and the concept is explained through simulated data from these populations. For the numerical data, the points are plotted on the graphs along with the lines indicating the cut off values.

References

1. Atkinson A.C. (1981), "Two graphical Displays for outlying and influence observations in regression", Biometrika, 68, 13-20

2. Avner Bar-Hen (1996), "A preliminary test in discriminant analysis", Journal of Multivariate Analysis, 57, 266-276

3. Belsley D. A., Kuh, E. and Welsch, R. E. (1980), "Regression Diagnostics: Identifying influential data and sources of collinearity", New York: Wiley

4. Campbell N. A. (1978), "The influence function as an aid in outlier detection in discriminant analysis", Applied Statistics, 27, 251-258

5. Cook R. D. (1977), "Detection of influential observations in linear regression", Technometrics, 19, 15-18

6. Cook R. D. and Weisberg (1982), Residuals and Influence in regression", New York: Chapman & Hall

7. Critchley and Vitiello C. (1991), "The influence of observations on misclassification probability estimates in linear discriminant analysis", Biometrika, 78, 3, 677-690

8. Daniel Pena (2005), "A new statistic for influence in linear regression", Technometrics, 47, 1, 1-12

9. Fung W. K. (1992), "Some diagnostic measures in discriminant analysis", Statistics and Probability Letters, 13, 279-285

10. Fung W. K. (1995a), "Diagnostics in linear discriminant analysis", Journal of American Statistical Association, 90, 952-956

11. Fung, W. K. (1995b), "Influence on classification and probability of misclassification", Sankhya, The Indian Journal of Statistics, Series B, 57, 337-384

12. Fung, W. K., (1996a), "Influence of observations for local log odds in linear discriminant analysis", Communications in Statistics Theory and Methods, 25, 257-268

13. Fung W. K. (1996b), "The influence of observations on misclassification probability in multiple discriminant analysis", Communications in Statistics Theory and Methods, 25, 1917-1930

14. Johnson (1987), "The detection of influential observations for allocation, separation and the determination of probabilities in a Bayesian framework", Journal of Business and Economic Statistics, 5, 369-381

15. Matusita K. (1955), "Decision rules based on the distance for problems of fit, two samples and estimation", Ann. Inst. Statist. Math., 7, 67-80

16. Radhakrishnan R (1983), "Influence functions for certain parameters in discriminant analysis", Metron, XLI-N, 1-2, 30, 183-194

17. P. Rajalakshmi (1990), 'Affinity and Applications to Classification Problems'(Phd thesis)

18. Wai Yin Poon (2004), "Identifying influential observations in discriminant analysis", Statistical Methods in Medical Research, 13, 291-308