A Weighted Likelihood Approach Based on Statistical Data Depths

Claudio Agostinelli

1 Department of Mathematics, University of Trento, Trento, Italy
claudio.agostinelli@unitn.it

February 16, 2018

Abstract

We propose a general approach to construct weighted likelihood estimating equations with the aim of obtain robust estimates. The weight, attached to each score contribution, is evaluated by comparing the statistical data depth at the model with that of the sample in a given point. Observations are considered regular when the ratio of these two depths is close to one, whereas, when the ratio is large the corresponding score contribution may be downweighted. Details and examples are provided for the robust estimation of the parameters in the multivariate normal model. Because of the form of the weights, we expect that, there will be no downweighting under the true model leading to highly efficient estimators. Robustness is illustrated using two real data sets.

keyword: Asymptotic Efficiency; Estimating Equations; Robustness; Statistical Data Depth; Weighted Likelihood

1 Introduction

Weighted Likelihood Estimating Equations (WLEE) are often used with the aim of obtaining robust estimators. Green [1984] is perhaps one of the earliest example. Field and Smith [1994] proposes a WLEE with weights that depends on the tail behavior of the distribution function, Markatou et al. [1997, 1998]
defines WLEE with weights derived from the estimating equations of a disparity minimization problems. Kuchibhotla and Basu [2017] further improve this approach by providing a strict connection between WLEE and disparity minimization problem. Biswas et al. [2015] get ideas from both Field and Smith [1994] and Markatou et al. [1998] and provide a similar approach based on distribution functions. Their approach is very natural and easy to implement, however one of the drawback it that the resulting estimators are not affine equivariant.

Statistical data depth was first introduced for multivariate observations, see Liu et al. [2006] and the reference therein for a review. The main goal is to provide a center-outward ordering of multivariate observations that can be used for several purposes, e.g. find centers: depth maximizers; regions containing most of the observations: depth regions and quantile depth based regions; comparing distributions: Depth-Depth plot and tests based on depth, and many others. One of the main features of a depth is its invariance under affine transformations. Furthermore, results shows that for a given model both the empirical distribution and the distribution function are completely characterized by the statistical data depth.

Our main goal is to propose a simple WLEE whose weights are based on statistical data depth. Section 2 provides an overview of the weighted likelihood strategies and our new proposal. Section 3 shows how we can use our new method in the multivariate normal model and Section 4 illustrates the methodology with two real examples. Section 5 reports some comments and conclusions.

2 Weighted Likelihood based on Data Depth

Let $x = (x_1, \cdots, x_n)$ be a random sample from a $p$-random vector $X$ with unknown distribution function $F$ and corresponding density function $f$. We assume a model for $X$ by $M = \{M(x; \theta); \theta \in \Theta \subset \mathbb{R}^p, p \geq 1\}$ and we denote $m_\theta(x) = m(x; \theta)$ the probability density function. Let $\tilde{F}_n$ be the empirical distribution function. Lindsay [1994] introduced the concept of Pearson residuals $\delta$ defined by comparing the true density to the model density as

$$\delta(x) = \delta(x; M_\theta, F) = \frac{f(x)}{m(x; \theta)} - 1.$$
so that, when \( f = m_\theta \) for a given \( \theta \in \Theta \) the Pearson residuals are equal to zero for all \( x \), whereas if \( f \neq m_\theta \), in regions where the density \( f \) is higher than \( m_\theta \) the Pearson residual are large indicating the disagreement between the number of observations in that regions and the expected observations under the model. The finite sample version of the Pearson residuals is given by \( \delta_n(x) = \delta(x; M_\theta, \hat{F}_n) = \frac{\hat{f}_n(x)}{m(x; \theta)} - 1 \), which compares \( \hat{f}_n(x) \), a non parametric estimate of \( f \), to the model density \( m(x; \theta) \). \cite{Lindsay1994} studied a class of estimators based on the Pearson residuals for discrete models, while \cite{BasuLindsay1994} and \cite{Markatou1998} discussed proposals for continuous models. In particular, \cite{Markatou1997, Markatou1998} introduced weights evaluated by

\[
 w(x) = w(\delta_n(x)) = \frac{A(\delta_n(x)) + 1}{\delta_n(x) + 1}, \tag{1}
\]

where \( A(\cdot) \) is the Residual Adjustment Function \cite{RAF, Lindsay1994, Park2002} obtaining a Weighted Likelihood Estimating Equations (WLEE)

\[
 \frac{1}{n} \sum_{i=1}^{n} w(x_i; M_\theta, \hat{F}_n) u(x_i; \theta), \tag{2}
\]

where \( u(y_i; \theta) \) denotes the \( i \)-th contribution to the score function. These estimating equations are derived from a disparity measure however, there is no exact link between the two approaches. Recently, \cite{KuchibhotlaBasu2017} provide a WLEE in the same spirit which corresponds to a disparity measures, they also formally prove the asymptotic and robust properties of their approach. In the attempt to avoid the use of non parametric estimators \cite{Biswas2015} proposes to define Pearson residuals as ratio of distribution functions such as

\[
 \delta_n(x) = \begin{cases} 
 \frac{\hat{F}_n(x)}{M(x; \theta)} - 1, & \text{if } M(x; \theta) \leq 0.5 \\
 \frac{1 - \hat{F}_n(x)}{1 - M(x; \theta)} - 1, & \text{if } M(x; \theta) > 0.5 
\end{cases}
\]

Let \( H(\delta) \) be a smooth function defined on \([-1, \infty)\), which assumes its maximum value 1 at \( \delta = 0 \) and descends smoothly in either tail as \( \delta \) moves away from 0, i.e., \( H(0) = 1 \) and \( H'(0) = 0 \) and the next higher non-zero derivative at \( \delta = 0 \) has a negative sign; an example is

\[
 H(\delta, a, c) = \begin{cases} 
 \exp(-a\delta^2), & \text{if } \delta \leq c \\
 0, & \text{otherwise}
\end{cases}
\]
where $a > 0$ and $c > 0$ are constants. Biswas et al. [2015] defines a weight function as

$$w(x) = \begin{cases} 
1, & \text{if } p < M(x, \theta) < 1 - p \\
H(\delta, a, \infty), & \text{otherwise.}
\end{cases}$$

Clearly, also weights in the form of (1) are possible. Their approach is general and can be used in multivariate setting as they show a bivariate example. However, their definition makes the Pearson residuals not affine invariant and hence the final estimates are not affine equivariant. We are going to propose a general approach to construct weights in the same spirit using statistical data depth which is more natural and is affine invariant.

Let $D(x; F_X)$ be a statistical data depth [Zuo and Serfling, 2000a, Liu et al., 2006] for the point $x \in \mathbb{R}^p$ according to the distribution $F_X$ of the r.v. $X \in \mathbb{R}^p$. Let $D_n(x; \hat{F}_n)$ be the finite sample version based on the empirical distribution function $\hat{F}_n$ of the sample $x$. Denote $\mathcal{F}$ the class of distributions in $\mathbb{R}^p$, assume that $D(x; F_X)$ satisfies the following properties [Liu, 1990, Zuo and Serfling, 2000a]

\begin{itemize}
  \item [P1.] **Affine Invariance.** $D(Ax + b; F_{AX+b}) = D(x; F_X)$ for any distribution function $F_X \in \mathcal{F}$, any $p \times p$ nonsingular matrix $A$ and any $p$-vector $b$.
  
  \item [P2.] **Maximality at Center.** For any $F_X$ having “center” $\mu$ (e.g. the point of symmetry relative to some notion of symmetry), $D(\mu; F_X) = \sup_{x \in \mathbb{R}^p} D(x; F_X)$.
  
  \item [P3.] **Monotonicity Relative to Deepest Point.** For any $F_X$ having deepest point $\mu$ (i.e., point of maximal depth), $D(x; F_X) \leq D(\mu + \alpha(x-\mu); F_X), \alpha \in [0, 1]$.
  
  \item [P4.] **Vanishing at Infinity.** $D(x; F_X) \to 0$ as $||x|| \to \infty$, for each $F_X \in \mathcal{F}$.
\end{itemize}

We define the Pearson residual as

$$\delta(x) = \frac{D(x; F)}{D(x; M_\theta)} - 1$$

and the finite sample version as

$$\delta_n(x) = \frac{D_n(x; \hat{F}_n)}{D(x; M_\theta)} - 1$$
This Pearson residual have the desired behaviour being equal to 0 whenever $F = M_\theta$ for some $\theta \in \Theta$, and of attaining large values in regions where the two distributions are mismatched; furthermore, because of the invariance property $P1$. of $D$, the Pearson residual is also invariant to affine transformations. For most depths a uniform convergence of $D_n(x; \hat{F}_n)$ to $D(x; F)$ holds almost surely; hence we expect that, at the model, the proposed method is highly efficient. Using weights based on (1) or (3) leads to a WLEE that can be solved by an iterative reweighting algorithm.

3 Application to the Multivariate Normal model

Consider $M_\theta$ be a Multivariate Normal model where $\theta = (\mu, \Sigma)$, $\mu$ is the mean vector and $\Sigma$ is the variance-covariance matrix. We discuss how to evaluate the proposed Pearson residual using the halfspace depth. We first review the concept of halfspace depth [Tukey, 1975; Donoho and Gasko, 1992]. Let $HS_u(a)$ be the closed halfspace \{ \(z \in \mathbb{R}^p : u^\top z \geq a \}\}. Here $u$ is a $p$-vector satisfying $u^\top u = 1$. Note that $HS_u(a)$ is the positive side of the hyperplane $HP_u(a) = \{ z \in \mathbb{R}^p : u^\top z = a \}$. The negative side $HS_{-u}(a)$ of $HP_u(a)$ is similarly defined. Halfspace depth is the minimum probability of all halfspaces including $x$.

**Definition 1** The halfspace depth $d_{HS}(x; X)$ maps $x \in \mathbb{R}^p$ to the minimum probability, according to the random vector $X$, of all closed halfspaces including $x$, that is

$$d_{HS}(x; X) = \inf_{u : u^\top u = 1} \Pr(HS_u(u^\top x); X)$$

$$= \inf_{u : u^\top u = 1} \Pr(z \in \mathbb{R}^p : u^\top z \geq u^\top x; X).$$

This statistical data depth is particularly useful since its properties. In particular, [Struyf and Rousseeuw, 1999] Theorem 1] show that the finite sample halfspace depth characterizes the empirical distribution. [Kong and Zuo, 2010; Corollary 3.1] show that the halfspace depth characterizes the underlying distribution. Furthermore, [Zuo and Serfling, 2000b, Theorem 3.3, Corollary 4.3] shows that the halfspace depth for a multivariate normal model can be easily obtained since $D(x; M_\theta) = (1 - F_{\chi^2_p}(d(x; \theta))/2$ where $d(x, \theta) = (x - \mu)^\top \Sigma^{-1}(x - \mu)$ is the squared Mahalanobis distance and $F_{\chi^2_p}$ is the distribution function of a $\chi^2_p$ chi-squared with $p$ degrees of freedom.
random variable. Finally, calculation of the finite sample halfspace depth in a dimension $p$ can be performed efficiently using the algorithms proposed in [Liu 2017] and [Dyckerhoff and Mozharovskyi 2016] and available in the R package `ddalpha` by [Pokotylo et al. 2016]. Similar results are available for any model belonging to the elliptically symmetric family of distributions.

4 Examples

We consider the data set `pb52` available in the R package `phonTools` [Barreda 2015] which contains the Vowel Recognition Data considered in [Peterson and Barney 1952], see also [Boersma and Weenink 2012]. In a first example a bivariate data set is illustrated, where we consider the vowels “u” (close back rounded vowel) and “æ” (near-open front unrounded vowel, “Æ” as x-sampa symbol) with a sample of size 304 equally divided for each vowel and the log transformed F1 and F2 frequencies measured in Hz. Our procedure use the following settings $\alpha = 0.5$, $a = 0.05$, $c = 200$, and uses 500 subsamples of size 6 as starting values for finding the roots. We also consider the Maximum Likelihood (MLE), the Minimum Covariance Determinant (MCD), the Minimum Volume Ellipsoid (MVE) and the S-Estimates (S) as implemented in the R package `rrcov` by [Todorov and Filzmoser 2009], the last three procedures were used with the exhaustive subsampling explorations. Figure 1 in the left panel reports the three estimates found by our methods. While the first root coincides with the MLE, the other two nicely identify the two subgroups. This is not the case for all the other investigated methods, see Figure 1 right panel, where the estimates are approximately all coincident with the MLE.

In a second example a trivariate data set is used considering vowels “u” (close back rounded vowel) and “ɔ” (open-mid central unrounded vowel, “ɔ” or “ɔ” as x-sampa symbol) with a sample of size 304 equally divided for each vowel and the log transformed F1, F2 and F3 frequencies measured in Hz. Figure 2 shows the results. For the classical robust procedures only MCD is able to somehow recover the structure of the observations for the vowels “ɔ”, while the other behave like the MLE. Our procedure finds nicely all the two substructure. For this data our procedure was set using $\alpha = 0.25$, $a = 0.1$, $c = 30$, and 1000 subsamples of size 6 as starting values for finding the roots. Using the setting of the first examples only the first two roots are found.
Figure 1: Vowel Recognition Data. Bivariate example, vowels “u” and “æ”. Left: the three roots of the proposed method. Right: estimates provided by MLE and robust procedures. Ellipses are 95% regions.

5 Conclusions

We have outlined a new form of weighted likelihood estimating equations where weights are based on comparing statistical data depth of the sample with that of the model. This approach avoids the use of nonparametric density estimates which can lead to problems for multivariate data, while retains nice characteristics of the classical WLEE approach, that are high efficiency at the model, affine equivariance and robustness. In the future, we hope to formally establish the theoretical properties of the proposed estimator.

References

S. Barreda. phonTools: Functions for phonetics in R., 2015. R package version 0.2-2.1.

A. Basu and B.G. Lindsay. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Annals of the Institute of Statistical Mathematics, 46(4):683–705, 1994.

A. Biswas, T. Roy, S. Majumder, and A. Basu. A new weighted likelihood approach. Stat, 4(1):97–107, 2015.
Figure 2: Vowel Recognition Data. Trivariate example, vowels “u” and “ə”.
Left: the three roots of the proposed method. Right: estimates provided by
MLE and robust procedures. Ellipses are 95% regions.
D.L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20:1808–1827, 1992.

R. Dyckerhoff and P. Mozharovskyi. Exact computation of the halfspace depth. *Computational Statistics & Data Analysis*, 98:19–30, 2016.

C. Field and B. Smith. Robust estimation – a weighted maximum likelihood approach. *International Statistical Review*, 62:405–424, 1994.

P.J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B*, 46:149–192, 1984.

L. Kong and Y. Zuo. Smooth depth contours characterize the underlying distribution. *Journal of Multivariate Analysis*, 101:2222–2226, 2010.

A.K. Kuchibhotla and A. Basu. A minimum distance weighted likelihood method of estimation. Technical report, Interdisciplinary Statistical Research Unit (ISRU), Indian Statistical Institute, Kolkata, India, 2017.

B.G. Lindsay. Efficiency versus robustness: The case for minimum hellinger distance and related methods. *The Annals of Statistics*, 22:1018–1114, 1994.

R.Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

R.Y. Liu, R.J. Serfling, and D.L. Souvaine. *Data depth: robust multivariate analysis, computational geometry, and applications*. AMS Bookstore, 2006.

X. Liu. Fast implementation of the tukey depth. *Computational Statistics*, 32(4):1395–1410, 2017.

M. Markatou, A. Basu, and B.G. Lindsay. Weighted likelihood estimating equations: the discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*, 57:215–232, 1997.
M. Markatou, A. Basu, and B.G. Lindsay. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93(442):740–750, 1998.

C. Park, A. Basu, and B.G. Lindsay. The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. *Computational Statistics & Data Analysis*, 39(1):21–33, 2002.

G.E. Peterson and H.L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184, 1952.

O. Pokotylo, P. Mozharovskyi, and R. Dyckerhoff. Depth and depth-based classification with r-package ddalpha. *arXiv:1608.04109*, 2016.

A. Struyf and P.J. Rousseeuw. Halfspace depth and regression depth characterize the empirical distribution. *Journal of Multivariate Analysis*, 69:135–153, 1999.

V. Todorov and P. Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009.

J.W. Tukey. Mathematics and picturing of data. In *Proceedings of International Congress of Mathematics*, volume 2, pages 523–531, 1975.

Y. Zuo and R.J. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000a.

Y. Zuo and R.J. Serfling. Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics*, 28(2):483–499, 2000b.