PLISKA
STUDIA MATHEMATICA BULGARICA

ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **14** (2003), 5–13

PLISKA
STUDIA MATHEMATICA
BULGARICA

ABOUT THE CONCEPT OF WEIGHTS OF WLTE(K) ESTIMATORS

Dimitar Atanasov

The concept for trimming and weighing the terms in the Method of Maximum Likelihood gives us a very flexible and useful way to improve the robustness of MLE. Till now the studies were focused mainly on the trimming factor. The theory of d-fullness gives us a powerful method to determine this property in the cases of WLTE, MLE and LTE.

The aim of this study is to consider the weights of the WLTE estimators and to compare the results obtained by using different algorithms for calculating weights.

1. Introduction

After the introduction to $WLTE(k)$ (Vandev & Neykov, 1993) a lot of works were focused on the properties of these estimators. The aim of these works was to study the breakdown properties and the trimming factor. The obtained results are similar to the results obtained for $LTE(k)$ (Neykov and Neytchev, 1990). Here we consider the properties of the second factor included in Weighted Least Trimmed Estimators - the weights.

Let $x_1, \cdots, x_n$ be a sample of n iid observations with a probability density function $\varphi(x_i, \theta)$, where θ is an unknown vector parameter. The $WLTE(k)$ estimators of θ are defined as

$$WLTE(k) = \operatorname*{argmin}_{\theta \in \Theta^p} \sum_{i=1}^{k} w_i f_{\nu(i)}(\theta), \tag{1}$$

2002 *Mathematics Subject Classification:* 62F35, 62F15

Key words: Robust statistics, WLT Estimators, Bayes Estimator

Partially supported by Mathind IST-2000-26015 and by Pro-enbis GTC1-2001-43031.

where $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \cdots \leq f_{\nu(n)}(\theta)$ are the ordered values of $f_i(\theta) = -\log\varphi(x_i,\theta)$ at θ, and $\nu = (\nu(1),\cdots,\nu(n))$ is the corresponding permutation of the indexes, which may depend on θ. The weights $w_i \geq 0$, $i = 1,\cdots,k$ are such that there exists an index $k = \max\{i : w_i > 0\}$.

These estimators were introduced independently by Hadi and Luceño (1997), and Vandev and Neykov (1998), as an extension of the Maximum Likelihood Estimators (MLE).

Vandev and Neykov (1998) proved that the finite sample breakdown point of the $WLTE(k)$ estimators is not less than $(n-k)/n$ if the set $F = \{f_i(\theta), i = 1,\cdots,n\}$ is d-full, $n \geq 3d$ and $(n+d)/2 \leq k \leq n-d$. We remind the reader that, according to Vandev (1993), a finite set F of n functions is called d-full if for each subset of cardinality d of F, the supremum of this subset is a subcompact function. A real valued function $g(\theta)$ is called subcompact, if its Lesbegue sets $L_{g(\theta)}(C) = \{\theta : g(\theta) \leq C\}$ are compact for any constant C (see Vandev and Neykov, 1993).

We remind the reader that the finite sample BP of an estimator T, at the finite sample $X = \{x_i; i = 1,\cdots,n\}$, is defined as the largest fraction m/n for which the $\sup_{\tilde{X}} \left\|T(X) - T\left(\tilde{X}\right)\right\|$ is finite, where $\tilde{X}$ is a sample obtained from X with replacing any m of the points in X by arbitrary values (see Hampel et al. 1986, Rousseeuw and Leroy 1987).

Vandev and Neykov (1993) determined the value d for the set of log likelihoods for the multivariate normal case while Vandev and Neykov (1998) did the same for different regression models, including the grouped binary logistic regression, and showed that the corresponding sets of log likelihoods are $(p+1)$-full.

Another very powerful property of the estimators are weights. Using them one can obtain different types of estimators. If all the weights are equal to 1 we have $LTE(k)$. The weights can also be proportional to the distribution in the estimated point, or in general proportional to something which depends on the distribution as errors. We can obtain different weights according to the used algorithm. For example, here the case when the weights are proportional to the number of times in which a given observation enters into the model is considered.

There is another interesting point of view to the concept of weights. They can be used not only to find the outliers in the data, but also to make a distinction between two sets of observations. For example, if there is some kind of discrete heterogeneity in the data we can differentiate the cases according to the values of the weights.

2. Analysis of different types of weights

Now we will consider the behaviour of $WLTE(k)$. To do this we will use a numerical example. The set X consists of 200 observations in two subsets X_1 and X_2, which are generated in different ways. So we can study the conduct of the estimator changing the proportion of these sets. We will suppose that the observations have a normal distribution and the distribution of the observations in the set X_2 is $N(0,1)$. We will change the variance of the observations in set X_1 but keep the mean equal to 3. All the estimations were done using one minimization procedure with appropriate parameters for different models. The optimization algorithm is based on Golden Section search and parabolic interpolation.

First we will consider the case when all the weights are equal to 1. As it was mentioned above, in this case the $WLTE(k)$ estimator is equivalent to LTE(k).

The used algorithm can be summarized as follows:

1. Setting the initial value for the unknown parameter

2. Sorting the observations according to the log-density function at the current value of the unknown parameter

3. The weights are equal to 1

4. Finding the value which satisfies (1)

5. If the exit conditions are not satisfied than go back to 2

In Table 1 some of the obtained results are presented. They are compared to the results from the Maximum Likelihood Estimation for the set X_1.

m	k	DX_1	MLE	St. err MLE	Result	St. err	Iter
10	100	1	2.81	0.13	2.73	0.15	6
10	150	1	2.87	0.10	2.88	0.11	4
10	190	1	3.1	0.087	3.13	0.098	4
10	100	2	2.93	0.20	2.90	0.25	5
10	150	2	2.90	0.17	2.29	0.19	4
10	190	2	3.02	0.16	3.01	0.19	3
90	100	1	2.93	0.21	2.45	0.32	3
90	110	1	2.87	0.19	2.43	0.32	3
90	100	2	2.97	0.19	2.31	0.37	3
50	100	1	3.11	0.21	3.02	0.27	4
50	100	2	3.09	0.23	3.01	0.35	3
50	150	1	2.97	0.18	3.10	0.21	4
50	150	2	3.01	0.23	2.96	0.29	5

Table 1

With m we denote the number of observations in the set X_2. The first six rows of the table show the behaviour of the estimator if there are some outliers in the data (10 observations). It is seen that with the increasing of the trimming factor k the error of the result decreases as in the case of MLE (Figure 1.).

Also the error increases whit the increasing of the variance of the observations. The standard error of the Maximum Likelihood Estimator is less than the same of LTE due to the existence of the trimming factor and using the smaller number of observations.

This type of estimator can be used to distinct between two sets of observations. This is shown in the next three rows of the table. Just about 15% of the observations are misplaced in their groups.

In the last three rows is shown that when the cardinality of set X_2 is smaller, the estimations are better than in the previous case.

Next we will consider the case when the weights are proportional to the number of iterations in which given observation enters into the model. The optimization algorithm differs from the previous one only in step 3, when the weights are calculated. At each iteration step the algorithm remembers which of the observations are in the model and at the next step they have bigger weights. We keep the sum of the weights equal to 1. We thought that this will allow us to use the weights in a very flexible way, because they are not directly dependent on the probability distribution. But in fact the weights are dependent on the distribution, because they are calculated using the probability density function to order the observations and to choose which of them to enter into the model. The results, obtained using such a model, are in Table 2.

m	k	DX_1	MLE	St. err MLE	Result	St. err	Iter
50	100	2	2.93	0.27	2.34	0.96	15
50	100	1	2.96	0.24	2.57	0.88	4
50	150	1	2.93	0.21	2.96	0.73	16
50	150	2	3.01	0.26	2.51	0.93	6
10	100	2	2.94	0.19	2.56	0.82	8
10	150	2	2.90	0.16	2.65	0.76	6
10	190	2	3.02	0.14	2.84	0.63	7
10	100	1	3.01	0.17	2.57	0.79	15
10	150	1	3.05	0.15	2.97	0.78	21
10	190	1	3.12	0.13	3.12	0.71	10
90	100	2	3.13	0.31	2.65	1.02	21
90	100	1	3.07	0.25	2.71	0.93	7

Table 2

Here it is seen that the standard error of the estimator is bigger than in previous case. Also the number of iterations is bigger than in the case of LTE and the value of the unknown parameter is not as close to the true value as the values estimated with MLE and LTE.

The first four rows in Table 2 show that the standard error of the estimator increases with the increasing of the variance of the set X_1 and decreases with the increasing of the trimming factor k.

When there are just a few outliers in the set (rows 5 - 10) the obtained results are not better. The standard error is big and the estimation is not good enough.

When the number of outliers is near the 50% of the observations the estimation is very bad.

We suppose that these results are due to the smooth likelihood function in this case. Adding the weights we smooth out the likelihood, so the number of iterations and the standard error are larger.

The last case studied here is the case when the weights are proportional to the value of the probability density function for the observation at the current value of the estimated parameter. The obtained results are similar to the results in the case when the weights are proportional to the number of times in which they are in the model. The results are presented in Table 3. If there are just a few outliers in the data the results are close to the results from MLE.

m	k	DX_1	MLE	St. err MLE	Result	St. err	Iter
50	100	2	2.95	0.26	2.22	0.91	2
50	150	2	2.93	0.22	2.15	0.88	2
50	100	1	2.98	0.21	2.95	0.77	6
50	150	1	2.99	0.19	2.98	0.59	5
10	100	1	3.10	0.18	2.92	0.33	8
10	150	1	2.97	0.16	2.92	0.24	6
10	190	1	3.05	0.14	3.05	0.22	7
90	100	1	3.10	0.41	2.28	0.98	21
90	100	2	2.91	0.55	2.04	1.13	7

Table 3

We do believe that we will have similar results if the weights depended in some way on the probability density function. So we can assume that adding the weights to the log-likelihood function makes the function more smooth and it is more difficult to have a fast and good optimization procedure. As a result of that the standard error will be bigger than the standard error of the estimator obtained using MLE or LTE.

3. Bayesian Modification of the LTE

In fact, the values of the likelihood function for a given observation weigh this observation in the likelihood. So it is not necessary to add other weights, especially if the weights depend on the distribution of the observations.

To improve the performance of the LTE estimator we can use the concept of Bayesian estimation of the unknown parameter.

Let us suppose that the density of a priori distribution of the unknown parameter is $h(\theta)$ and the probability density function of the observations is $\phi(x, \theta)$. Then the a posteriori distribution of the parameter is

$$h(\theta \mid x) = \frac{\phi(x, \theta)h(\theta)}{\phi(x)},$$

where

$$\phi(x) = \int\limits_{-\infty}^{\infty} \phi(x, \theta)h(\theta)d\theta.$$

Using this notation, we can consider the estimator

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta^p} \sum_{i=1}^{k} - \log h(\theta \mid x_{\nu(i)}), \tag{2}$$

where $x_{\nu(i)}, i = 1, \cdots, n$ are the observations, ordered according to the value of $-\log h(\theta \mid x)$, $-\log h(\theta \mid x_{\nu(1)}) \leq -\log h(\theta \mid x_{\nu(2)}) \leq \cdots \leq -\log h(\theta \mid x_{\nu(n)})$.

Now let us consider the breakdown properties of the proposed estimator. Using the definition (2) we have

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta^p} \sum_{i=1}^{k} - \log \frac{\phi(x_{\nu(i)}, \theta)h(\theta)}{\phi(x_{\nu(i)})} =$$

$$= \operatorname*{argmin}_{\theta \in \Theta^p} \sum_{i=1}^{k} (- \log \phi(x_{\nu(1)}, \theta) - \log h(\theta) + \log \phi(x_{\nu(i)})).$$

Let us note the function

$$g_i(\theta) = - \log \phi(x_{\nu(1)}, \theta) - \log h(\theta) + \log \phi(x_{\nu(i)}).$$

According to Vandev (1993) we need to study the index of fullness for the set of functions $F = \{g_i(\theta)\}$. As $\log \phi(x_{\nu(i)})$ is a constant, using the unequality

$$-\log \phi(x_{\nu(1)}, \theta) - \log h(\theta) \geq -\log h(\theta)$$

we have that $g_i(\theta)$ is subcompact function if and only if $-\log h(\theta)$ is a subcompact one. We can use the following theorem:

Theorem 1. (Atanasov & Neykov, 2001) *Let D be an open subset of R^n, θ_0 belong to the boundary of D and $g(\theta)$ be a real valued continuous function defined on D. Then $g(\theta)$ is subcompact if and only if for any sequence $\theta_i \to \theta_0$ $g(\theta_i) \to \infty$ when $i \to \infty$.*

Corollary 1. If $h(\theta)$ is a continuous probability density function then according to the Theorem the function $-\log h(\theta)$ will be a subcompact one. So the index of fullness of the set F is $d = 1$.

Therefore, the breakdown point of the $WLTE(k)$ for these models is equal to $\frac{(n-k)}{n}$ when $\frac{(n-1)}{2} \leq k \leq n-1$, according to Vandev and Neykov (1998) and some additional arguments given there. This gives us a very flexible way to manage such an estimator, because we can choose the minimal possible value for k.

For example let us consider the case when the observations have a normal distribution $N(\theta, 1)$ and the distribution of the unknown parameter is also normal $N(0, \tau^2)$, for an appropriate value for τ. Under these conditions we have

$$h(\theta \mid x) = \frac{\phi(x,\theta)h(\theta)}{\phi(x)} =$$

$$= \frac{e^{-\frac{(x-\theta)^2}{2} - \frac{\theta^2}{2\tau^2}}}{\int\limits_{\infty}^{\infty} e^{-\frac{(x-\theta)^2}{2} - \frac{\theta^2}{2\tau^2}} d\theta} = \frac{e^{-\frac{(x-\theta)^2}{2n} - \frac{\theta^2}{2\tau^2}}}{\frac{\sqrt{2\pi\tau^2} e^{-\frac{x^2}{2(\tau^2+1)}}}{(\tau^2-1)}} =$$

$$= \frac{1}{\sqrt{2\pi(1+\frac{1}{\tau^2})}} e^{\frac{-\left(\theta - \frac{x}{1+\frac{1}{\tau^2}}\right)}{2\left(1+\frac{1}{\tau^2}\right)^{-1}}}.$$

So the a posteriori distribution of the unknown parameter is $N\left(\frac{x}{1+\frac{1}{\tau^2}}, \left(1 + \frac{1}{\tau^2}\right)^{-1}\right)$ and we can use the ordinary LTE estimator with a calculated a posteriori probability density function.

The obtained results are given in Table 4.

m	k	DX_1	MLE	St. err MLE	Result	St. err	Iter
90	100	2	2.88	0.36	2.71	0.38	2
90	110	2	2.83	0.32	2.65	0.33	2
90	100	1	3.05	0.28	2.95	0.29	3
90	110	1	2.99	0.26	2.98	0.26	5
50	100	2	2.91	0.19	2.92	0.21	5
50	150	2	2.97	0.15	2.94	0.19	3
50	100	1	3.02	0.14	3.05	0.18	5
50	150	1	2.95	0.12	3.01	0.16	3
10	150	1	3.04	0.21	2.99	0.26	3
10	190	1	2.96	0.18	2.98	0.23	3
10	190	2	2.87	0.22	2.91	0.24	2

Table 4

The first four rows show the case when the observations consist of two subsets with almost equal sizes. The estimation of the parameter is close to the maximum likelihood estimation. When the outliers are not so much the estimation is better and the standard error is smaller.

It is seen that using this estimator we have a very small number of iterations needed to obtain a result. Also we can say that the procedure is strongly dependent on the starting point, so we need some kind of a priori information for it.

This estimator is more flexible than LTE and gives better results than $WLTE(k)$ models, considered above.

REFERENCES

[1] D. V. Atanasov, N. M. Neykov. About the Finite Sample Breakdown Point of the WLTE(k) Estimators. In: Proceedings of the XXV Summer School Sozopol'99, (eds. Cheshankov, B.I., Todorov, M.D.) (2000), 105–106.

[2] D. V. Atanasov, N. M. Neykov. On the Finite Sample Breakdown Point of the WLTE(k) and d-fullness of a Set of Continuous Functions. In: Proceedings of the VI International Conference "Computer Data Analysis And Modeling", Minsk, Belarus. (2001), 1–2.

[3] A. Hadi, A. Luceño. Maximum Trimmed Likelihood Estimators: A Unified Approach, Examples and Algorithms. *Comput. Statist. and Data Analysis* **25** (1997), 251–272.

[4] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel. Robust Statistics: The Approach Based on Influence Functions. John Wiley and Sons, New York, 1986.

[5] N. M. Neykov, P. N. Neytchev. A Robust Alternative of the Maximum Likelihood Estimators. *COMPSTAT 1990* (1990), 99–100.

[6] P. J. Rousseeuw, A. Leroy. Robust Regression and Outlier Detection. John Wiley and Sons, New York, 1987.

[7] D. L. Vandev. A Note on Breakdown Point of the Least Median of Squares and Least Trimmed Estimators. *Statistics and Probability Letters* **16** (1993), 117–119.

[8] D. L. Vandev, N. M. Neykov. Robust Maximum Likelihood in the Gaussian Case. In: New Directions in Statistical Data Analysis and Robustness. (eds. S. Morgenthaler, E. Ronchetti, W.A. Stahel), Birkhauser Verlag, Basel, 1993.

[9] D. L. Vandev, N. M. Neykov. About Regression Estimators with High Breakdown Point. *Statistics* **32** (1998), 111–129.

Faculty of Mathematics and Informatics
Sofia University
5 J. Boucher Str.1407
1113 Sofia, Bulgaria
e-mail: datanasov@fmi.uni-sofia.bg