NONPARAMETRIC BOOTSTRAP TECHNIQUES FOR IMPLICITLY WEIGHTED ROBUST ESTIMATORS

Jan Kalina

Abstract

The paper is devoted to highly robust statistical estimators based on implicit weighting, which have a potential to find econometric applications. Two particular methods include a robust correlation coefficient based on the least weighted squares regression and the minimum weighted covariance determinant estimator, where the latter allows to estimate the mean and covariance matrix of multivariate data. New tools are proposed allowing to test hypotheses about these robust estimators or to estimate their variance. The techniques considered in the paper include resampling approaches with or without replacement, i.e. permutation tests, bootstrap variance estimation, and bootstrap confidence intervals. The performance of the newly described tools is illustrated on numerical examples. They reveal the suitability of the robust procedures also for non-contaminated data, as their confidence intervals are not much wider compared to those for standard maximum likelihood estimators. While resampling without replacement turns out to be more suitable for hypothesis testing, bootstrapping with replacement yields reliable confidence intervals but not corresponding hypothesis tests.

Key words: robust statistics, econometrics, correlation coefficient, multivariate data

JEL Code: C14, C12, C63

Introduction

Numerous standard statistical methods are well known to be too sensitive to the presence of outliers. This is true for various estimators in various models, including linear regression, correlation coefficient, estimates of the mean and scatter matrix in multivariate data, nonlinear regression, dimensionality reduction, classification analysis etc.

Robust statistical estimators can be described as tools resistant to the presence of outliers, which commonly have the form of a modification (robustification) of standard statistical estimators. Their overview was presented e.g. by Filzmoser & Todorov (2011). Our attention is however focused entirely on implicitly weighted estimators, which seem to yield promising results in a variety of applications. They can be described as analogues of the least weighted

squares (LWS) estimator in the linear regression (Víšek, 2002). Roelant et al. (2009) used the term minimum weighted covariance determinant estimator (MWCD) for its analogue, which has been tailor made for mean and covariance matrix estimation in multivariate data analysis. A robust correlation coefficient based on the implicit weighting was investigated by Kalina & Schlenker (2015). All these implicitly weighted estimators, for which the weights are assigned based on ranks of residuals, can be computed by an analogue of the approximate FAST-MCD algorithm of Rousseeuw & van Driessen (1999). Also other approaches to rank methods are known to yield robust results, mainly in regression setup (Saleh et al., 2012).

Nevertheless, robust estimators require to be accompanied by a variety of diagnostic tools, including hypothesis tests and confidence intervals. In robust statistics, they have been investigated mainly for linear regression estimators (Víšek, 2011), while they have not penetrated to other important econometric models (Salini et al., 2016). Because they are often too complicated to be investigated theoretically, we take resort to resampling techniques.

Bootstrap estimation (bootstrapping) has become popular in a variety of statistical tasks, mainly in estimating the variance of estimators. Practical approaches to bootstrapping, i.e. resampling with replacement, were investigated by Efron & Tibshirani (1994). Incorporating the basic principles of bootstrapping, one may develop a great variety of resampling techniques that provide us with new possibilities of analyzing data by means of residual bootstrap, semiparametric bootstrap, Bayesian bootstrap etc.

Permutation tests can be also interpreted as an important class of resampling methodology (without replacement). They can be interpreted as a flexible nonparametric technique suitable if the asymptotic behavior is not known but if exchangeability of individual observations is ensured (Pesarin & Salmaso, 2010). Sometimes, permutation tests are also called invariance tests or conditional tests, where the latter concept stresses conditioning of the procedure by the observed data. If only a random sample of permutations is used, the approach is often denoted as permutation bootstrap.

In this paper, Section 1 proposes a new two-stage procedure for assigning weights for robust estimators based on implicit weights. Section 2 is devoted to a robust correlation coefficient based on the least weighted squares, which is accompanied by a permutation test and by a bootstrap-based confidence interval. Their performance is illustrated on an economic data set in Section 3. Section 4 illustrates a bootstrap estimator of variance of the MWCD estimator of parameters of multivariate data on real data.

1 Choice of the weights for implicitly weighted estimators

The choice of suitable weights is an important parameter in the process of applying implicitly weighted estimators in various models. This section recalls several available possibilities for the choice of weights and also proposes novel kernel-based weights, which will be used in the examples throughout the paper.

Some examples of weights, which are suitable for various implicitly weighted estimators (e.g. the LWS-based correlation coefficient or the MWCD estimator) include:

- Zero-one weights, used in e.g. least trimmed squares (LTS) or minimum covariance determinant (MCD) estimators,
- Linearly decreasing weights (Kalina, 2012),
- Weights generated by a (given) non-increasing function,
- Data-dependent adaptive weights of Čížek (2011).

The first three choices of fixed weights are not sufficiently flexible, while the only adaptive proposal is computationally rather complicated. Therefore, we will now propose novel data-dependent weights assigned by a two-stage rule as an alternative. The weights are denoted as kernel-based weights.

First, an initial highly robust estimator must be chosen and computed for the given data. Residuals of individual observations will be denoted as $u_1^0, \dots, u_n^0$. In the second stage, weights are obtained by means of a given kernel K as $w_i = K(u_i^0)$ for $i = 1, \dots, n$.

The construction of kernel-based weights resembles kernel-based nonparametric regression estimators (see e.g. Matioli et al. (2017)), particularly the popular Nadaraya-Watson estimator. The proposal of kernel-based weights is simple, its computation is straightforward and the weights allow a clear interpretation. As an important useful example, let us mention the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \quad x \in \mathbb{R}, \tag{1}$$

which will be used in examples throughout the paper.

2 Robust correlation coefficient

This section is devoted to a robust correlation coefficient based on the LWS. We accompany this implicitly weighted correlation coefficient by a permutation test as well as by a more detailed description of two types of a bootstrap confidence interval. These methods will be illustrated in Section 3 on a numerical example.

Independent identically distributed (i.i.d.) measurements coming from two random variables $X = (X_1, \dots, X_n)^T$ and $Y = (Y_1, \dots, Y_n)^T$ from a continuous distribution are considered. Their true (i.e. population) but unknown value of the correlation coefficient will be denoted by ρ. The robust correlation coefficient r_{LWS} (Kalina & Schlenker, 2015) is obtained as the weighted Pearson correlation coefficient with such weights, which are found as optimal by the LWS estimator in the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, \dots, n. \tag{2}$$

Known properties of r_{LWS} include its high breakdown point or asymptotic normality; an asymptotic test of $H_0: \rho = 0$ based on r_{LWS} was proposed by Kalina & Schlenker (2015).

2.1 A permutation test

A permutation test of $H_0: \rho = 0$ against $H_1: \rho \neq 0$ is a standard tool considering all possible permutations of the pairs of observations. Let us consider a permutation π of indices $1,2, \dots, n$, i.e. a bijection from $\{1, \dots, n\}$ to $\{1, \dots, n\}$ and the i-th coordinate of X is $X_{\pi(i)}$. The test is based on a repeated evaluating of r_{LWS} between $(X_{\pi(1)}, \dots, X_{\pi(n)})^T$ and $(Y_1, \dots, Y_n)^T$. Thus, the standard permutation test can be exploited for any estimator of ρ and we do not to describe here the particular version based on the highly robust coefficient r_{LWS}.

2.2 A bootstrap confidence interval

Further, we propose two versions of a bootstrap confidence interval for the population correlation coefficient ρ, which will be based again on r_{LWS}. First, a naïve confidence interval can be constructed as

$$\left[r_{LWS} - 1.96\sqrt{var\ r_{LWS}}\ , r_{LWS} + 1.96\sqrt{var\ r_{LWS}}\right], \tag{3}$$

where the variance of the robust correlation coefficient is estimated by bootstrap in a standard way as the variance of the bootstrap distribution, i.e. empirical distribution of the bootstrap samples.

A more standard confidence interval exploiting the bootstrap distribution, i.e. empirical distribution obtained from bootstrap samples, is proposed in the following Algorithm.

Algorithm 1. Bootstrap confidence interval for ρ based on r_{LWS}.

Input: Data $(X_1, \dots, X_n)^T$ and $(Y_1, \dots, Y_n)^T$, number of repetitions K

Output: Bootstrap confidence interval for ρ

1. For $k = 1$ to K do
2. Generate n bootstrap samples

$$\left(X_j^{(k)}, Y_j^{(k)}\right), \quad j = 1, \dots, n, \tag{4}$$

by sampling with replacement from the original set of data rows (X_i, Y_i) with $i = 1, \dots, n$.

3. Compute r_{LWS} between

$$\left(X_1^{(k)}, \dots, X_n^{(k)}\right) \text{ and } \left(Y_1^{(k)}, \dots, Y_n^{(k)}\right), \tag{5}$$

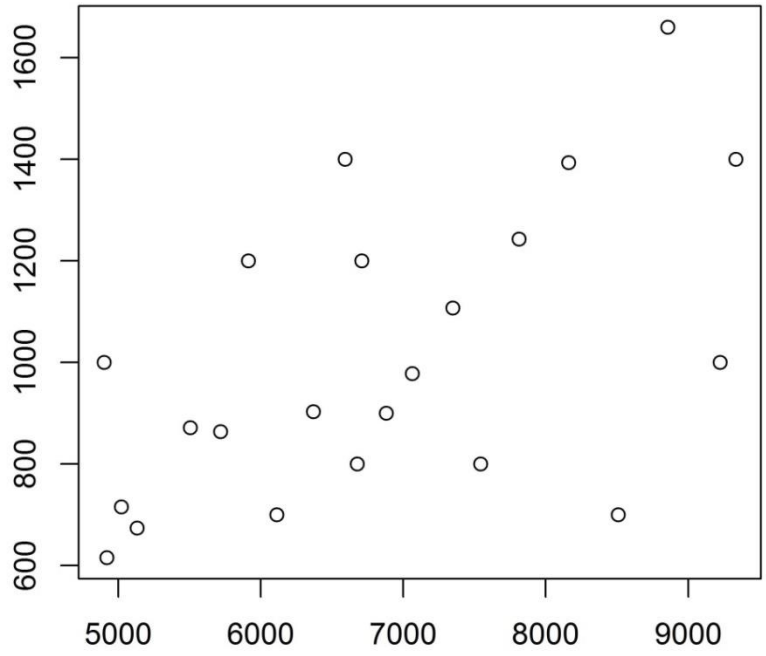
store the value and denote it as r_k.

4. End for
5. Arrange the values in ascending order as $r_{(1)} < \cdots < r_{(K)}$.
6. Construct the 95 % confidence interval as

$$\left[r_{(\lfloor h \rfloor)}, r_{(\lfloor n-h \rfloor)}\right], \tag{6}$$

where $h = \lfloor 0.025n \rfloor$ and $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.

Fig. 1: The data set investigated in Section 3.



Source: an artificial data set created by the author

In general, bootstrap estimates require to derive their properties for each particular task, although this is typically ignored in practical applications, mainly because it is rather complicated to derive consistency or expression for the bias for bootstrap estimates or

confidence intervals. Such formal investigation of properties of the bootstrap confidence intervals would be however necessary also in our case, in spite of the fact that r_{LWS} is a consistent estimator of ρ under the assumption of bivariate normality of $(X, Y)^T$.

3 Example: robust correlation coefficient

An artificial data set of Figure 1 is considered with 22 measurements of two continuous variables. The data were manually constructed as a mixture of a linear trend with severe contamination. We computed the Pearson's correlation coefficient as well as r_{LWS} with the weights according to the Gaussian kernel (1), where the correlation coefficient based on the least trimmed squares with $h = 11$ is used as the initial robust estimate.

Further, we applied methods of Section 2. The permutation test of $H_0: \rho = 0$ against $H_1: \rho \neq 0$ as well as the bootstrap confidence intervals for ρ based on r or r_{LWS} were computed. The asymptotic test based on r is the standard t-test, while the asymptotic test based on r_{LWS} exploits the formula of Kalina & Schlenker (2015). The classical confidence interval based on r based on the asymptotic normality, was computed using the function cor.test() in R software. We implemented all other confidence intervals in R software.

Tab. 1: Results of the example of Section 3.

Method	Result of tools based on r	Result of tools based on r_{LWS}
Correlation coefficient	0.55	0.60
Permutation test	$p = 0.0042$	$p = 0.0033$
Asymptotic test	$p = 0.0076$	$p = 0.0059$
Classical (asymptotic) confidence interval	[0.17, 0.79]	[0.18, 0.86]
Bootstrap confidence interval (Alg. 1)	[0.23, 0.79]	[0.21, 0.87]
Naïve bootstrap confidence interval (3)	[0.21, 0.89]	[0.19, 0.83]

Source: own computation

The results of all computations, which were performed in R software, are shown in Table 1. Both the permutation and asymptotic test are highly significant. The number of observations is not high enough to have the confidence intervals narrower, although it suffices for a significant corresponding test. The naïve interval (3) is the only one centered around r_{LWS}, which is however not a desirable property due to asymmetry of the distribution of r_{LWS}; the symmetry is redeemed by an assumption of normality which is not fulfilled. In fact, we

understand the standard bootstrap confidence interval to be the best among the three presented intervals, because it is the narrowest, non-asymptotic and also uses the complete information about the whole bootstrap distribution. It is at the same time computationally well feasible although demanding more bootstrap samples compared to (3).

4 Comparison of multivariate estimators

The aim of this section is to compare robust multivariate estimators on a real data set. While the variance of estimators is their important characteristic, the comparison will be based on bootstrap estimates of the variance of the MCD and the minimum weighted covariance determinant (MWCD) estimators.

I.i.d. multivariate data will be assumed coming from an elliptical distribution like in Roelant et al. (2009), who proposed the MWCD estimator of the population mean denoted as $\bar{X}_{MWCD}$. The MWCD, which can be understood as a compromise between classical estimates and the MCD, minimizes the determinant of the weighted covariance matrix over all possible permutations of weights. We are interested in bootstrap estimation of $var\ \bar{X}_{MWCD}$, while the MWCD is known to lower the bias of the MCD (Roelant et al., 2009).

The asymptotic variability of the MWCD-mean is too complicated to be formally proven. Therefore, a bootstrap estimator was applied already by Willems and van Aelst (2004) for estimating $var\bar{X}_{MWCD}$ corresponding to the minimum covariance determinant (MCD) estimator. The MCD corresponds to the MWCD with zero-one weights. It is obtained as the classical mean computed only for such h observations, which minimize the determinant of the covariance matrix over all possible h-subsets of observations.

Nevertheless, neither consistency nor asymptotic bias has been formally investigated for their bootstrap estimator. The bootstrap estimator for the variance of the MWCD-mean can be computed in a standard way as the variance of the bootstrap distribution.

The following example has the aim to compare multivariate estimators on a real data set as well as to compare to illustrate the bootstrap estimates of their variance. The 3-dimensional phosphorus data set, which is publicly available (Rousseeuw & Leroy, 1987), contains $n = 18$ measurements of 3 variables, namely inorganic, organic and plant phosphorus content in soil. The data set does not contain severe outliers. We compute three estimators of the mean and covariance matrix of the data:

- The maximum likelihood estimators. For the mean $\bar{X}$, we used $var\ \bar{X} = \Sigma/n$, which was estimated by S/n.

- The MCD estimator with default parameters (i.e. its reweighted version).
- The MWCD estimator with kernel-based weights (1) using the MCD with $h = 10$ as the initial estimator.

Table 2 presents the results of the computations performed in R software. The main result of the computations is a clear loss of the efficiency of the MCD estimator much compared to maximum likelihood estimates. On the other hand, the MWCD is able to yield reliable results with variability estimates much closer to the maximum likelihood estimates. Thus, the results reveal that the MWCD estimator is able to combine the robustness with efficiency, which is a very desirable property.

Tab. 2: Results of the example of Section 4. Estimates of the population mean of the three variables measuring phosphorus of different origin in soil. Three estimators are presented together with bootstrap estimates of their standard deviations.

	X_1 (inorganic)	X_2 (organic)	X_3 (plant)
Maximum likelihood	11.9	42.1	81.3
	(0.56)	(0.76)	(1.50)
MCD	11.7	39.7	76.1
	(2.84)	(3.60)	(5.01)
MWCD	11.7	40.3	77.9
	(0.89)	(1.22)	(2.46)

Source: own computation

Conclusion

This paper is devoted to two implicitly weighted robust methods applicable to econometrics and fills the gap of additional tools for these methods, namely an LWS-based robust correlation coefficient r_{LWS} and the MWCD estimator.

While the data of Section 3 do not contain severe outliers, the inference based on r_{LWS} yields results resembling those obtained with the Pearson's correlation coefficient r. We can say that the approach based on r_{LWS} does not seem to lose much information and seems suitable for non-contaminated data as well. Robust properties of r_{LWS} are granted (Kalina & Schlenker, 2015). For r_{LWS}, we elaborated possible resampling approaches in a more detailed way and compared two different approaches to bootstrap confidence intervals.

The example of Section 4 with multivariate data brings arguments in favor of the MWCD estimator. It is very desirable that robust methods are efficient for non-contaminated data,

which seems to be true in the example. The MWCD estimator is guaranteed to improve bias (Roelant et al., 2009) and the example shows that it can improve efficiency as well.

Two different resampling techniques, namely a permutation test and bootstrap estimation, are conceptually rather different, although both are based on computationally demaning resampling, While the permutation principle (i.e. resampling without replacement) allows to construct reliable hypotheses tests, bootstrapping yields reliable confidence intervals but not *p*-values of corresponding hypothesis tests. Also different techniques must be used to derive their properties, while bootstrap estimates require to prove their properties for each particular situation, while permutation tests are generally valid.

Sometimes, permutation tests are claimed to require to perform all permutations. If however the user performs only a (sufficiently large) random sample of permutations, then the test would be sometimes called a bootstrap test without replacement rather than a permutation test. In fact, bootstrapping has not much penetrated to hypothesis testing although systematic comparisons of permutation tests and bootstrap tests can be also found in the literature. However, such comparisons have been however performed in rather particular problems, see e.g. the comparison by Hušková and Kirch (2012) for the change-point problem.

As a future work, it is intended to perform a simulation study showing how fast the permutation test based on r_{LWS} converges to the asymptotic test. In addition, we would like to propose an improved approximate algorithm for computing the MWCD estimator and to investigate its computational aspects.

Acknowledgment

The work was supported by the project 17-07384S of the Czech Science Foundation.

References

1. Čížek, P. (2011): Semiparametrically weighted robust estimation of regression models. *Computational Statistics & Data Analysis*, 55 (1), 774-788.
2. Efron, B. & Tibshirani, R.J. (1994): *An introduction to the bootstrap*. Chapman & Hall/ CRC, Boca Raton.
3. Filzmoser, P. & Todorov V. (2011): Review of robust multivariate statistical methods in high dimension. *Analytica Chimica Acta* 705 (1-2), 2-14.
4. Hušková, M. & Kirch, C. (2012): Bootstrapping sequential change-point tests for linear regression. *Metrika*, 75 (5), 673-708.

5. Kalina, J. & Schlenker, A. (2015): A robust supervised variable selection for noisy high-dimensional data. *BioMed Research International*, 2015, Article No. 320385.
6. Kalina, J. (2012): Highly robust statistical methods in medical image analysis. *Biocybernetics and Biomedical Engineering*, 32 (2), 3-16.
7. Matioli, L.C., Santos, S.R., Kleina, M., & Leite, E.A. (2018): A new algorithm for clustering based on kernel density estimation. *Journal of Applied Statistics*, 45, 347-366.
8. Pesarin, F. & Salmaso, L. (2010): *Permutation tests for complex data: Theory, applications and software*. Wiley, New York.
9. Roelant, E., Van Aelst, S., & Willems, G. (2009): The minimum weighted covariance determinant estimator. *Metrika*, 70 (2), 177-204.
10. Rousseeuw, P.J., & Leroy, A.M. (1987): *Robust regression and outlier detection*. Wiley, New York.
11. Rousseeuw, P.J., & Van Driessen, K. (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41 (3), 212-223.
12. Saleh, A.K.M.E., Picek, J., & Kalina, J. (2012): R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika*, 75, 311-328.
13. Salini, S., Cerioli, A., Laurini, F., & Riani, M. (2016): Reliable robust regression diagnostics. *International Statistical Review*, 84 (1), 99-127.
14. Víšek, J.Á. (2011): Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47, 179-206.
15. Willems, G. & Van Aelst, S. (2004): A fast bootstrap method for the MCD estimator. In Antoch J. (Ed.): *Proceedings in Computational Statistics COMPSTAT 2004*. Springer, Heidelberg, 1979-1986.

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic
& The Czech Academy of Sciences, Institute of Information Theory and Automation
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
kalina@cs.cas.cz