DOI: 10.18267/j.pep.411

ON MULTIVARIATE METHODS IN ROBUST ECONOMETRICS

Jan Kalina*

Abstract:

This work studies implicitly weighted robust statistical methods suitable for econometric problems. We study robust estimation mainly for the context of heteroscedasticity or high dimension, which are up-to-date topics of current econometrics. We describe a modification of linear regression resistant to heteroscedasticity and study its computational aspects. For a robust version of the instrumental variables estimator we propose an asymptotic test of heteroscedasticity. Further we describe robust statistical methods for dimension reduction and classification analysis. We propose the robust quadratic classification analysis based on a new minimum weighted covariance determinant (MWCD) estimator. In general the robust methods based on down-weighting less reliable observations are resistant to outlying values (outliers) and insensitive to the assumption of Gaussian normal distribution of the data. The methods are illustrated on econometric data examples.

Keywords: least weighted squares, heteroscedasticity, multivariate statistics, model selection, diagnostics, computational aspects

JEL Classification: C14, C13, C51.

1. Introduction

The least weighted squares (LWS) regression is a robust regression method based on the idea of down-weighting less reliable observations proposed by Víšek (2001). In this paper we use the idea of the LWS estimator to modify popular econometric methods, which are sensitive to outliers.

Recent trends in robust statistics and econometrics aim at a systematic treatment of heteroscedasticity and dimension reduction for high-dimensional data. Hekimoglu *et al.* (2009) performed a simulation study to examine robustness properties of regression estimators to different types of outliers. Alqallaf *et al.* (2009) considered a contamination model for high-dimensional data allowing to modell different types of outliers and showed that standard estimators are very vulnerable to contamination for high-dimensional data. Gelper *et al.* (2009) carried out robust online estimation of variance in a univariate time series in a moving window using three consecutive observations.

* Jan Kalina, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, CZ – 182 07, Praha 8 (kalina@cs.cas.cz). The research is supported by Center of Biomedical Informatics, Project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic.

DOI: 10.18267/j.pep.411

Only recently there has been studied highly robust multivariate estimation. García-Escudero and Gordaliza (2005) proposed to detect outliers in multivariate elliptical data rigorously by means of a robustified Mahalanobis distance from the robust midpoint of the data. Salibián-Barrera and Yohai (2006) proposed new algorithms for robust estimation of location and dispersion for high-dimensional multivariate datasets. Riani *et al.* (2009) defined outlier identification rules for the multivariate model. Still the theory of diagnostic tools for robust statistical methods is not developed and robust multivariate analysis is a hot topic with promising applications to econometrics. Therefore our work aims to implement the idea of the least weighted squares regression to substantial topics of modern robust statistics and econometrics, mainly to diagnostic tools and multivariate methods.

This paper has the following structure. Section 2 recalls the least weighted squares estimator and its properties. Section 3 is devoted to robust regression resistant under heteroscedasticity, which is a model with different variances of individual disturbances. Section 4 proposes a diagnostic test for the robust instrumental weighted variables. The methods of both Sections 3 and 4 are special cases of a robust generalized method of moments estimator, which is popular in econometrics. Section 5 considers robust ways of reducing the dimension for high-dimensional data. Finally Section 6 studies robust quadratic classification analysis. Implicit weighting turns out to be a promising concept to obtain robust methods suitable for econometric applications.

Some of the methods of this paper belong to a general context called robust generalized method of moments (GMM). This is true for the robust regression efficient under heteroscedasticity and instrumental weighted variables estimator. The generalized method of moments estimator is a general tool for statistical estimation given by orthogonality conditions is defined for a general parametric situation involving instrumental variables by Hansen (1982) and we refer to Greene (2002) or other econometric textbooks for an overview. The connection to the classical method of moments is shown by Wooldridge (2001). While there is an extensive number of references on robust inference for the linear regression, only a few econometric papers extend the results to the more general GMM estimator: Ronchetti and Trojanni (2001) studied robustness properties of GMM estimation with applications to the ARCH model; Wagenvoort and Waldmann (2002) proposed an instrumental variables estimator with a bounded influence function; and Víšek (2005) used the idea of down-weighting less reliable observations to robustify the GMM estimator.

The methods of this paper also form an integral part of the task of model selection, which is a crucial topic in current multivariate econometrics and statistics. The task of selecting the suitable model requires to include suitable independent variables or instruments to the resulting model, which will be a submodel of the original full model. The following properties are required from reliable methods of model selection: accurate predictions; clear interpretation; robustness and stability; low bias in parameter estimation or hypotheses testing in the resulting model.

DOI: 10.18267/j.pep.411

2. Least weighted squares regression

This section recalls the least weighted squares (LWS) regression estimator of Víšek (2001) and summarizes its properties and advantages. Let us consider the linear regression model in the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

which can be rewritten in the usual matrix notation as $Y = X\beta + e$. The least weighted squares estimator is one of robust estimation methods with a high breakdown point, which is a statistical measure of sensitivity against noise or outliers in the data (see Rousseeuw and Leroy, 1987). The LWS estimator down-weights less reliable observations based on the values of squared residuals. The magnitudes of nonnegative weights $w_1, w_2, \ldots, w_\mathrm{n}$ must be specified before the computation of the estimator. These are, however, assigned to particular data points after a permutation, which is determined automatically only during the computation based on the residuals. It is reasonable to choose $w_1, w_2, \ldots, w_\mathrm{n}$ as a non-increasing sequence so that the most reliable observations obtain the largest weights, while outliers with large values of the residuals get small (or zero) weights. One possibility is to choose linearly decreasing weights. The data-adaptive weights of Čížek (2008) are another choice.

Let us denote the *i*-th order value among the squared residuals for a particular value of the estimate *b* of the parameter β by $u^2_{(i)}(b)$. The least weighted squares estimator b_{LWS} for model (1) is defined as

$$b_{LWS} = argmin \sum_{i=1}^{n} w_i u^2_{(i)}(b), \tag{2}$$

where the minimum is computed over all possible values of *b*.

The least trimmed squares (LTS) regression proposed by Rousseeuw and Leroy (1987) represents a special case of least weighted squares with weights equal to zero or one only. The computation of the LWS estimator is intensive and computational aspects are studied by Kalina (2008).

The least weighted squares estimator has interesting properties and applications. In particular it is robust for contaminated data sets; here it is usually assumed that the disturbances represent a mixture of normally distributed random variables with outliers. At the same time the estimator is reliable for data with normally distributed disturbances without contamination; this reliability is measured in terms of a high efficiency, which compares the asymptotic variability of the estimator relatively to the variability of the least squares estimator. Theoretical properties including the robustness of the estimator are studied by Čížek (2008), who conjectures that the LWS estimator is a reasonable compromise between the least squares and least trimmed squares. Diagnostic tools for the disturbances *e* (random regression errors) are available. For example tests of heteroscedasticity and autocorrelation of the disturbances can be computed employing weighted residuals (see Kalina, 2007); such diagnostic tests are asymptotically equivalent with classical tests for least squares regression. Another advantage of the

DOI: 10.18267/j.pep.411

estimator is that no detection of outliers is actually needed to compute it, because outlying data are down-weighted automatically. The LWS estimator has a small local sensitivity compared to the LTS.

3. Robust Regression Efficient under Heteroscedasticity

Cragg (1983) proposed a modification of the least squares regression, which has become popular in econometrics for its efficiency also under heteroscedasticity. This section proposes a robust analogy of Cragg's approach to linear regression resistant to heteroscedasticity and studies computational aspects of the proposed estimator. This may be a suitable method for high-dimensional data.

Let us consider the linear regression model in the form (1) with heteroscedastic disturbances *e*. This means the violation of the classical assumption $var\ e_i = \sigma^2$. The least squares estimator b_{LS} of the regression parameters β is not efficient under heteroscedasticity of the disturbances *e* and further the estimator of $var\ b_{LS}$ is biased (see Greene, 2002). It follows that the confidence intervals and hypothesis tests concerning β are not valid.

Cragg (1983) proposed a useful transformation allowing to obtain a more reliable estimator of β and mainly of its variance even without testing if the heteroscedasticity is present in the model (1). The idea is to use some auxiliary variables which could possibly contribute to explaining the variability of the disturbances *e*. The usual choice contains squares of all independent variables from (1) and also products of always two different independent variables. This corresponds to using the matrix (let us say *Q*) consisting of all columns of the original design matrix *X* and the auxiliary variables as additional columns. The model (1) is transformed to

$$Q^T\ Y = Q^T X\beta + Q^T e, \tag{3}$$

where ^{T} denotes transposition of a matrix. The parameters β in (3) can be estimated by the generalized least squares (GLS) estimator of Aitken (1935), which will be denoted by b_{GLS}. In general Aitken estimator is based on a known variance matrix of the disturbances. Because in our case the variance matrix of $Q^T e$ is unknown, we estimate it and plug in this estimator into the formula for Aitken estimator, which yields the so-called admissible (or estimated) Aitken estimator.

Let us denote the diagonal matrix with squared residuals in (3) by *S*. This is an estimator of the variance matrix of the disturbances *e* based on a naïve estimator of β. Now it is possible to obtain an estimator for $var\ b_{GLS}$ in the form

$$X^T Q (Q^T S Q)^{-1} Q^T X. \tag{4}$$

The number of rows in the model (3) is equal to the number of columns in the matrix *Q*, which is the number of variables in the original model (1) together with the number of auxiliary variables contained in columns of *Q*. This is the reason for the success of the method for data with a high dimension, which is reduced for the task of estimating

DOI: 10.18267/j.pep.411

the regression parameters. At the same time it is not needed to include the additional columns to the original model to replace (1) by the model $Y = Q\beta + e$.

Now we describe a robustification of Cragg's approach, which is alternative to Víšek (2005). The idea is to down-weight less reliable observations (possible outliers) while the most credible and typical data points obtain the largest weights. The weights are assigned to particular data points automatically during the computation of the estimator. In general the weights represent an important diagnostic tool explaining the outlyingness of individual data points. Particularly the linearly decreasing weights

$$w = 1 - \frac{i-1}{n}, \quad i = 1, ..., n, \ \text{standardized to} \sum_{i=1}^{n} w_i = 1, \tag{5}$$

are a good choice well-established also for the LWS in the linear regression context (see Kalina, 2008). However, the optimal choice of the weights for implicitly weighted estimators remains to be an open problem.

We introduce weights directly to the model (3) in the form

$$Q^T WY = Q^T WX\beta + Q^T We, \tag{6}$$

where W is a weight matrix containing weights determined by the least weighted squares in the original model (1). This down-weights outliers both in the response and the independent variables. The whole procedure starts by the LWS and then uses the (classical) weighted regression to estimate β in the transformed model (4). The method can be described as a two-stage estimator:

1. The least weighted squares regression is used in the model (1). The matrix W is obtained. The matrix S is obtained as the diagonal matrix containing squares of residuals.
2. Using the transformation (6), the estimator of β is obtained as the admissible Aitken estimator, where *var* $(Q^T We)$ is approximated by $Q^T WSWQ$.

Example 1. The procedure is illustrated on a data set from Maddala (1988). Consumption expenditures of the total number of $n = 20$ families are modelled as linear response of the income of each family. The least squares estimate of β is $b = (0.847, 0.899)^T$ with standard errors $(0.703, 0.025)^T$. While the linear trend is correctly estimated, standard errors of b are overestimated due to heteroscedasticity.

We apply Cragg‘s approach to the least squares regression in our example. Let the matrix Q contain the square of the income as an auxiliary variable. The formula (6) is a linear system with three rows and two parameters, equivalent to solving a linear regression with two regressors without intercept. The scatter plots of $Q^T Y$ against both columns of $Q^T X$ show that the three points are very close to being collinear. Using the Cragg‘s approach the regression parameters are estimated by $(0.628, 0.910)^T$ with standard errors $(0.298, 0.020)^T$. The estimate of β is very similar to the classical least squares, while there is reduction in the variability. The new estimate of β is therefore more accurate than the classical estimate, which is deceived by heteroscedasticity.

The least weighted squares estimate of β with linear weights is equal to $(0.399, 0.939)^T$ with asymptotic standard errors $(0.812, 0.029)^T$. The two-stage LWS regression with data-adaptive weights estimates β by $(0.691, 0.904)^T$. This regression line is very close to the least squares regression line. The largest weights correspond to data very close to the estimated line. The asymptotic standard errors of b_{LWS} are $(0.704, 0.904)^T$; we point out that the asymptotic variance of LWS is derived for independent identically distributed normal disturbances and is deceived by heteroscedasticity.

The modification of the two-stage least weighted squares using (6) with the square of the income as auxiliary variable and with data-adaptive weights gives the estimate of β equal to $(0.645, 0.906)^T$ with standard errors equal to $(0.047, 0.0003)^T$, where the improvement is remarkable compared to asymptotic variance for the least weighted squares.

4. Diagnostics for the Instrumental Weighted Variables Estimator

The instrumental variables (IV) estimator has become a standard tool for estimation in econometrics (see Greene, 2002). This section presents an asymptotic version of Szroeter's heteroscedasticity test for a robust instrumental variables estimator.

The instrumental variables estimator does not assume that the disturbances e are uncorrelated with the independent variables, while there is the total number L of instrumental variables available satisfying $L \geq p$. Let the vector $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{iL})^T$ of values of the instruments correspond to the i-th observation. Let the matrix Z contain the values of the instruments so that data vectors Z_i are contained in rows of Z. The estimation in the model (1) with instrumental variables starts with explaining the regressor X by the instruments Z, in other words there is assumed a linear model $X = Z\gamma + v$, where γ is a vector of (arbitrary) regression parameters and v is a vector of disturbances. The instrumental variables estimator is a popular method in econometrics and there is paid an intensive attention to the task of selecting suitable instruments. At the same time the method is suitable also for high-dimensional problems, because it is not required to include the instruments to the original model (1), while the number of instruments can be large.

Víšek (2006) proposed a robust version of the instrumental variables estimator called instrumental weighted variables (IWV) estimator b_{IWV}. The IWV estimator is based on the idea of down-weighting less reliable observations, similarly with the least weighted squares regression. Víšek (2006) studied the consistency and asymptotic normality of the estimator for the special case $L = p$ and proved the asymptotic representation of b_{IWV}. The FAST-LTS algorithm proposed by Rousseeuw and van Driessen (1999) can be modified to compute an approximation to the estimator.

In this work we present the asymptotic Szroeter‘s test of heteroscedasticity for the instrumental weighted variables estimator, which can be carried out using the classical instrumental variables. Similarly to the original work of Szroeter (1978) we propose a class of tests of heteroscedasticity assuming $var\ e_i \geq var\ e_{i-1}$ for $i = 2, ..., n$. It is required to specify constants $h_1, ..., h_n$ satisfying $h_i \leq h_j$ for $i < j$. Szroeter (1978)

DOI: 10.18267/j.pep.411

described several possibilities for the choice of $h_1, ..., h_n$. One possibility is to select them as indicators assigning data to groups similarly with the Goldfeld-Quandt test (see Kmenta, 1986). Another choice

$$h_i = 2\left[1 - \cos\left(\frac{i\pi}{n+1}\right)\right], \quad i = 1,...n, \tag{7}$$

leads to such form of the test, which has the same critical values as the Durbin-Watson test of independence of the disturbances e.

Let us introduce the notation B for the diagonal matrix $B = diag\{h_1,... , h_n\}$, I_n for the unit matrix with dimension (n,n) and $M = I_n - X(Z^TX)^{-1}Z^T$. Let us express the residuals u of the IWV estimator as $u = Y - Xb_{IWV}$ and let us denote their mean by $\bar{u}$. The asymptotic Szroeter‘s test for the IWV residuals is based on the result of Víšek (2006), who proves the asymptotic representation for the IWV estimator under technical assumptions; here we assume them to be fulfilled.

Theorem 1. Let us assume the assumptions of Víšek (2006) to be fulfilled. Then the test statistic

$$(u-\bar{u})^T B(u-\bar{u}) / (u-\bar{u})^T \ (u-\bar{u}) \tag{8}$$

is asymptotically equivalent in probability with

$$e^TM^TBMe/e^TM^TMe. \tag{9}$$

The proof follows from Kalina (2007), who proves the asymptotic equivalence of the Durbin-Watson test statistic computed with residuals of the least weighted squares regression with the Durbin-Watson test statistic computed with residuals of the least squares regression. Here, however, the asymptotic representation for the IWV estimator is used, which is derived by Víšek (2006) under general assumptions; apart from technical conditions the weights are assumed to be generated by a nonincreasing function and there is assumed a unique solution of the normal equations defining the IWV estimator. Based on Víšek (2006), (8) is approximated by

$$(\kappa-\bar{\kappa})^T B(\kappa-\bar{\kappa})^T / (\kappa-\bar{\kappa})^T \ (\kappa-\bar{\kappa}), \tag{10}$$

where the asymptotic approximation gives $u = \kappa + o_P(n^{-1/2})$ with the term $o_P(n^{-1/2})$ negligible in probability.

The Szroeter‘s test statistic is scale-invariant under the null hypothesis. The asymptotic test for the residuals of the instrumental weighted variables estimator can be computed using simulations. Random disturbances following normal distribution with zero expectation and any variance can be repeatedly generated to obtain the exact p-value of the test (8). This is an approximation to the p-value of the asymptotic test.

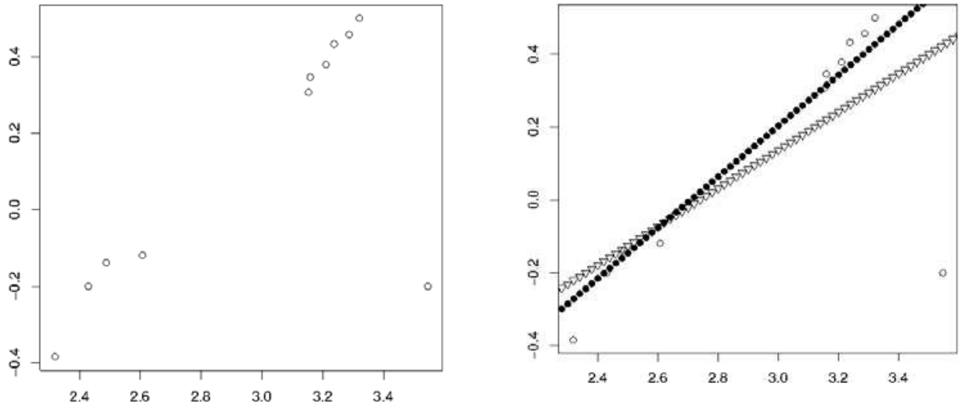
We illustrate the IWV estimator on two examples which reveal its advantages. The robustness of the instrumental weighted variables estimator will be demonstrated by comparing it with a non-robust estimator (Example 2) and by cross-validation (Example 3).

DOI: 10.18267/j.pep.411

Example 2. We work with furniture data from Kmenta (1986). The response and the independent variable are both measured in furniture manufacturing industry in 11 different countries of the world. However, to compare the classical and robust approaches, we introduce an outlier, namely we replace the response of the first observation (the value *0.768*) by the value *-0.2*. The data after this modification are shown in Figure 1 (left), where the outlier is now in the right bottom corner of the plot. Firstly the linear regression of the response against the regressor is considered. The least squares estimate of b is $(-1.48,\ 0.540)^T$, while the least weighted squares with data-adaptive weights is different with $(-1.91,\ 0.706)^T$. Kmenta (1986) warned that additional variables may contribute to the variability of the response, so the disturbances may not be uncorrelated with the regressor. Therefore he recommended to use an instrumental variable, namely the regressor measured in knitting mill industry. The relationship between the regressor and the instrument is very nicely linear.

Figure 1

Furniture Data of Example 2



Left: original data with an outlier. Right: estimated values fitted by the classical instrumental variables estimator (grey triangles) and by the robust instrumental variables estimator (dark circles).

The estimator of b using instrumental variables estimation is similar to the result of the least squares estimate computed without instruments, namely $(-1.44,\ 0.526)^T$. The robust instrumental variables estimator is computed as the two-stage estimator with the least weighted squares in each stage with data-adaptive choice of weights. The estimate of b equals $(-1.89,\ 0.699)^T$, similarly with the LWS estimate without instruments. Figure 1 (right) shows the classical and robust estimates using instrumental variables. We can conclude that the robust approach truly brings an improvement over the classical one, which is deceived by the outlier.

Example 3. We simulate $n = 60$ observations $(X_i,\ Y_i,\ Z_i)^T$, $i = 1, \ldots, n,$ in the following way. $X = (X_i, \ldots, X_n)^T$ are equidistant values between *0.167* and *10.0* and $Z = (Z_1, \ldots, Z_n)^T$ are generated as $Z_i = X_i + v_i$, where $v_1, \ldots, v_n$ are independent identically distributed random disturbances following the normal distribution *N(0, 0.25).*

The response values $Y = (Y_1, ..., Y_{55})$ are created as $Y_i = X_i + e_i$, where the independent identically distributed random disturbances e_i follow the normal distribution *N(0, 0.25);* to introduce outliers to the data set we generate Y_{56}, ..., Y_{60} as independent identically distributed random variables with the normal distribution *N(2,1).* The plot of *Y* against *X* and the plot of *X* against *Z* are shown in Figure 2.

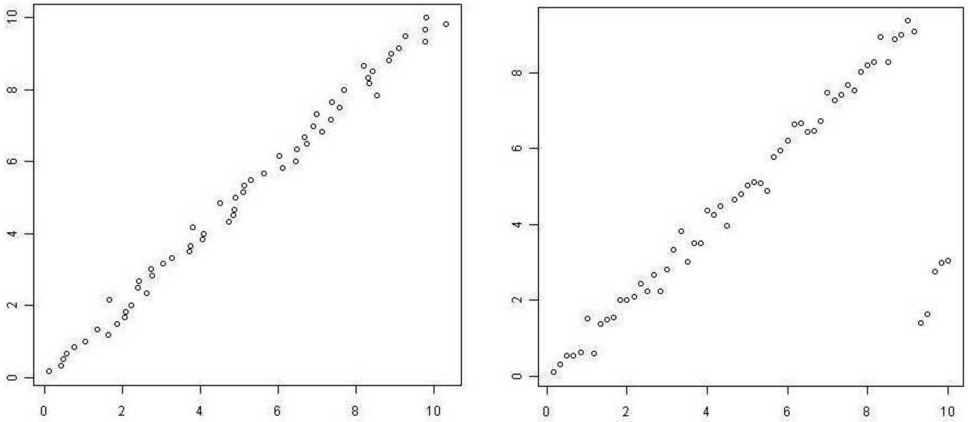
Our aim is to estimate the parameters β in the model

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad i = 1, 2, ..., n, \tag{11}$$

using *Z* as an instrument for *X*. The IWV estimator based on the LWS with linearly decreasing weights (5) yields the estimate $(0.112, 0.960)^T$ much closer to the true value $b = (0,1)^T$ compared to the classical IV estimator yielding $(1.089, 0.679)^T$.

Figure 2

Simulated Data of Example 3



Left: plot of the regressor *X* against the instrument *Z*. Right: plot of the response *Y* against the regressor *X*.

To compare the prediction accuracy of the classical and robust approaches to the IV estimation we apply cross-validation. We use the leave-1-out cross-validation method, which involves fitting the model over the observations after trimming away 1 randomly selected observation; similarly the leave-3-out method is trained over all data after removing 3 randomly selected data points. The results are presented in Table 1 for the leave-1-out (*k = 1*) and leave-3-out methods (*k = 3*) for 1000 experiments corresponding to random selections of removed observations.

DOI: 10.18267/j.pep.411

Table 1
Leave-k-out Cross-Validation for Simulated Data of Example 3

	K=1	k=1	k=3	k=3
	IV	IWV	IV	IWV
MSE	5.06	5.99	5.08	7.68
SEP	2.27	2.40	2.26	2.67
Trimmed MSE	1.70	0.24	1.70	0.29
Trimmed SEP	1.19	0.48	1.18	0.52

Note: The classical instrumental variables (IV) and robust instrumental weighted variables (IWV) estimation procedures are compared for *k=1* and *k=3.* Results are evaluated by means of the mean square error (MSE) and standard error of prediction (SEP) for *n=60* (top) and their trimmed analogs trimming away *5/60* of the data (bottom).

Top half of Table 1 summarizes the mean values of the mean square error (MSE) and the standard error of prediction (SEP), which are characteristics of the predictive quality of the particular model; see Varmuza and Filzmoser (2009) for details. Large values of MSE and SEP of the robust IV estimator are caused by an extremely poor fit for the outliers, while the classical IV estimation is intrigued by them. Only the trimmed analogs of the MSE and SEP show realistic outcomes yielding the robust IV estimator as a clear winner in the quality of the fit and prediction. Further we use trimmed analogs of MSE and SEP, which are presented in the bottom half of Table 1. These are computed as the classical MSE and SEP after a complete ignoring of 5 mostly outlying results; therefore only $55/60 * 100\% \approx 91.7\%$ of the original observations are used to compute these robust characteristics. Particularly we find observations yielding the largest absolute prediction error and compute the MSE and SEP only for the remaining cases. The difference between $k = 1$ and $k = 3$ is not very dramatic. To summarize, the big improvement of the robust IV estimation compared to the classical IV estimator is a significant argument in favour of the instrumental weighted variables estimator.

5. Dimension Reduction for High-Dimensional Data Sets

Robust methods for dimension reduction include for example robust versions of the principal components analysis or instrumental variables estimation. Here we describe useful methods applicable to econometrics. The purpose of this section is to review usual steps, which are often carried out as a preprocessing before applying (robust) classification analysis, which will be our aim in Section 6.

Croux (2000) described a robust modification of the principal component analysis (PCA) method based on robust estimation of the covariance or correlation matrix and replacing eigenvalues and eigenvectors by their robust counterparts. M-estimators and S-estimators are examined together with their influence functions.

A more robust approach with respect to outlier values in terms of the breakdown point is proposed by Hubert (2005), applying the concept of the projection pursuit technique. This is a general method of Rousseeuw and Leroy (1987) for finding the

DOI: 10.18267/j.pep.411

most informative directions or components for multivariate (high-dimensional) data. Such classification is based on a robust measure of spread of the data, taking into account the outlyingness of each data point. Candidate directions for the principal components are selected by a grid algorithm optimizing such objective function only in a plane, while subsequent components are added in later steps. The fast algorithm of Rousseeuw and van Driessen (1999) for the method is implemented in library *pcaPP* of the *R* software package (Ihaka and Gentleman, 1996), which makes the method appealing for high-dimensional data. Other methods for dimension reduction include partial least squares, lasso, least angle regression or logic regression, which are not very popular in econometrics yet.

6. Robust Classification and Discrimination

We propose the minimum weighted covariance determinant (MWCD) estimator, which is a robust multivariate estimator based on the idea of down-weighting less reliable observations. Then a robust quadratic classification method is proposed based on the MWCD estimator.

The minimum covariance determinant (MCD) estimator is a high-breakdown estimator of multivariate location and scatter (Rousseeuw and van Driessen, 1999). It requires to choose the trimming constant *h* $(n/2<h<n)$; while *n-h* observations are ignored completely, only the *h* remaining data points are used to compute the estimator. Particularly for the estimation of the multivariate location, the MCD estimator is defined as the trimmed mean with such *n-h* observations trimmed away, yielding the smallest possible determinant of the trimmed variance matrix. Such trimming involves the complete rejection of *n-h* data points, while the trimmed mean and trimmed variance matrix are computed as classical mean and variance matrix using only the *h* remaining observations.

Let us consider *p*-dimensional data $X_1, X_2, ..., X_n$. We propose to define the minimum weighted covariance determinant (MWCD) estimator as a weighted analogy of the MCD estimator, down-weighting less reliable data points. It is required to specify the sizes of the weights, $w_1, w_2 ..., w_n$ and again we recommend to use linearly decreasing weights (5). For fixed weights $w_1, w_2 ..., w_n$ the weighted mean $\bar{X}_W$ and the weighted variance matrix

$$S_{MWCD} = \Sigma_i w_i (X_i - \bar{X}_W)(X_i - \bar{X}_W)^T \qquad (12)$$

can be computed. In our case we consider all possible permutations of the weights. We find such permutation of the weights, which yields the minimal determinant of the weighted variance matrix. We define the minimum weighted covariance determinant (MWCD) estimator of location as the weighed mean $\bar{X}_W$ of the data with these optimal weights and the corresponding estimator of the variance matrix is the weighted variance matrix (10) with the optimal weights.

Rousseeuw and van Driessen (2006) proposed an approximate algorithm for the computation of the MCD estimator. Now we propose a modification for the computation of the MWCD estimator. We introduce weights to the algorithm and the complete rejection of data points is replaced by assigning these weights. The complete algorithm for computing the MWCD estimator will be now described.

Algorithm 1.

(i) Initialize the value of the loss function as + ∞.

(ii) Randomly select an initial set of p observations. Compute the mean T and variance matrix C based on these p observations.

(iii) Compute robust Mahalanobis distances for each of the observations $X_1, X_2, ..., X_n$ in the form $d(i; T, C) = [(X_i - T)^T C^{-1}(X_i - T)^{1/2}]$ for each observation X_i. Sort these distances in ascending order. This determines a permutation $\pi(1), \pi(2), ..., \pi(n)$ of the indexes $1,2, ..., n,$ which fulfils $d(\pi(1); T, C) \leq d(\pi(2); T, C) \leq ... \leq d(\pi(n); T, C)$. Assign the weights to individual observations according to the ranks of the Mahalanobis distances. In other words, for example the observation $X_{\pi(1)}$ obtains the weight w_1.

(iv) The loss function is evaluated as the determinant of the matrix C. If the loss is smaller than the previously obtained value, continue with step (v). Otherwise continue with step (vi).

(v) Store the values of the weights. Compute the weighted mean and weighted variance matrix using these weights. Continue with steps (ii), (iii) and (iv). This is repeated as long as the value of the loss decreases.

(vi) Repeat the steps (i) to (v) 10 000 times. The optimal weights are those which yield the minimal value of the loss function over all repetitions of steps (i) to (v).

In the algorithm, the permutation of the data arranges the data according to the ascending order of the Mahalanobis distances. Therefore observations with a small Mahalanobis distance obtain larger weights.

Our aim is a robust classification method based on the MWCD estimator. This method is inspired by Croux (2000) or Hubert (2005). A comparison of standard robust approaches to classification analysis is given by Todorov and Pires (2007). The MCD estimator is a modification of the LTS regression to the multivariate context, while the MWCD estimator corresponds to the LWS regression. Similarly with the LTS regression, the MCD estimator suffers from a high local sensitivity. On the other hand the LWS regression has a small local sensitivity (see Víšek, 2001), which motivates the usage of implicitly weighted estimators such as MWCD.

Let us consider multivariate data in a total number of J groups. We use the notation μ_j for the mean and Σ_j for the variance matrix of the data in the j-th group ($j = 1,..., J$). The quadratic classification analysis (QDA) is based on the quadratic classification function (see Johnson and Wichern, 1982)

DOI: 10.18267/j.pep.411

$$Qj = \bar{X}_j^T S_j^{-1} X - \frac{1}{2}(\log|S_j| + \bar{X}_j^T S_j^{-1} \bar{X}_j + X^T S_j^{-1} X) \qquad (13)$$

which is derived from the likelihood function of the multivariate normal distribution and estimates μ_j by $\bar{X}_j$ and Σ_j by S_j. Now we estimate the population characteristics μ_j and Σ_j by robust estimators, namely we replace them by the MWCD estimator of multivariate location $\bar{X}_{j,\ MVCD}$ and variance matrix $S_{j,\ MVCD}$. This estimation is done separately in each group of data points. The robust estimation of the population characteristics allows us to define the robust quadratic classification method based on the MWCD estimator.

Definition 2. The quadratic MWCD-classification assigns a new observation *X* to the *j*-th group, if the quadratic classification function

$$Q_j^* = \bar{X}^T_{j,MWCD} S^{-1}_{j,MWCD}\, X - \frac{1}{2}\left(\log|S_{j,MWCD}| + \bar{X}^T_{j,MWCD}\, S^{-1}_{j,MWCD}\, \bar{X}_{j,MWCD} + X^T S^{-1}_{j,MWCD} X\right) \qquad (14)$$

is equal to $max\{Q_1^*, ..., Q_j^*\}$.

The classification analysis performed on robust principal components by Croux (2000) or Hubert (2005) allows to preserve as much information relevant for the classification as possible while reducing the computational complexity.

Here the new approach can profit from the properties of the MWCD estimator, inherited from the idea of down-weighting less reliable data points. The estimator namely turns out to be very robust for highly contaminated data sets and efficient for normal data without contamination, and at the same time robust also with respect to the local sensitivity. Therefore it overcomes an important drawback of locally sensitive LTS and MCD estimators and the MWCD estimator turns out to be one of implicitly weighted estimators with desirable properties.

References

Aitken, A. C. (1935), “On Least Squares and Linear Combination of Observations”. *Proceedings of the Royal Statistical Society* 55, pp. 42-48.

Alqallaf, F.; van Aelst, S.; Yohai, V. J.; Zamar, R. H. (2009), “Propagation of Outliers in Multivariate Data.” *Annals of Statistics* 37, No. 1, pp. 311-331.

Čížek, P. (2008), “Efficient Robust Estimation of Time-Series Regression Models.” *Applications of Mathematics* 53, No. 3, pp. 267-279.

Cragg, J. G. (1983), “More Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form.” *Econometrica* 51, No. 3, pp. 751-763.

Croux, C.; Haesbroeck, G. (2000), “Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies.” *Biometrika* 87, pp. 603-618.

García-Escudero, L. A.; Gordaliza, A. (2005), “Generalized Radius Processes for Elliptically Contoured Distributions.” *Journal of the American Statistical Association* 100, pp. 1036-1045.

Gelper, S.; Schettlinger, K.; Croux, C.; Gather, U. (2009), “Robust Online Scale Estimation in Time Series: a Model-Free Approach.” *Journal of Statistical Planning and Inference* 139, pp. 335-339.

Greene, W. H. (2002), *Econometric Analysis.* Fifth edition. New York: Macmillan.

DOI: 10.18267/j.pep.411

Hansen, L. P. (1982), "Large Samples Properties of Generalized Method of Moments Estimators." *Econometrica* 50, No. 4, pp. 1029-1054.

Hekimoglu, S.; Erenoglu, R. C.; Kalina, J. (2009), "Outlier Detection by Means of Robust Regression Estimators for Use in Engineering Science." *Journal of Zhejiang University – Science A (JZUS-A)* 10, No. 6, pp. 909-921.

Hubert, M.; Rousseeuw, P. J.; Vanden Branden, K. (2005), "ROBPCA: A New Approach to Robust Principal Component Analysis." *Technometrics* 47, pp. 64-79.

Ihaka, R.; Gentleman, R. R. (1996), "A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5, pp. 299-314.

Johnson R. A.; Wichern, D. W. (1982), *Applied Multivariate Statistical Analysis.* Prentice-Hall, Englewood Cliffs.

Kalina, J. (2007), "Asymptotic Durbin-Watson Test for Robust Regression." *Bulletin of the International Statistical Institute* 62, pp. 3406-3409.

Kalina, J. (2008), "Robustní regrese a diagnostické nástroje." In Kupka, K. (Ed.), *Data analysis 2007/II, Progressive methods of statistical data analysis and modelling for research and technical practice.* Trilobyte Statistical Software, Pardubice, pp. 31-41. (In Czech.)

Kmenta, J. (1986), *Elements of Econometrics*. New York: Macmillan.

Maddala, G. S. (1988), *Introduction to Econometrics*. New York: Macmillan.

Riani, M.; Atkinson, A. C.; Cerioli, A. (2009), "Finding an Unknown Number of Multivariate Outliers." *Journal of the Royal Statistical Society, Series B*, 71, pp. 447-466.

Ronchetti, E.; Trojani, F. (2001), "Robust Inference with GMM Estimators." *Journal of econometrics* 101, pp. 37-69.

Rousseeuw, P. J.; Leroy, A. M. (1987), *Robust Regression and Outlier Detection.* New York: Wiley.

Rousseeuw, P. J.; van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics* 41, No. 3, pp. 212-223.

Rousseeuw, P. J.; van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets." *Data Mining and Knowledge Discovery* 12, pp. 29-45.

Salibián-Barrera, M.; Yohai, V. J. (2006), "A Fast Algorithm for S-Regression Estimates." *Journal of Computational and Graphical Statistics* 15, pp. 414-427.

Szroeter, J. (1978), "A Class of Parametric Tests of Heteroscedasticity in Linear Econometric Models." *Econometrica* 46, pp. 1311-1328.

Todorov, V.; Pires, A. M. (2007), "Comparative Performance of Several Robust Linear Discriminant Analysis Methods." *REVSTAT Statistical Journal* 5, pp. 63-83.

Varmuza, K.; Filzmoser, P. (2009), *Introduction to Multivariate Statistical Analysis in Chemometrics.* Boca Raton: Taylor & Francis - CRC Press.

Víšek, J. Á. (2001), "Regression with High Breakdown Point." In Antoch, J.; Dohnal, G. (Eds.), *Proceedings of ROBUST 2000, Summer School of JČMF*, JČMF and Czech Statistical Society, pp. 324-356.

Víšek, J. Á. (2005), "Robustifying Generalized Method of Moments." In Kupka K. (Ed.): *Data analysis 2004/II, Progressive Methods of Statistical Data Analysis and Modelling for Research and Technical Practice*, Trilobyte, Pardubice, pp. 171-193.

Víšek, J. Á. (2006), "Instrumental Weighted Variables." *Austrian Journal of Statistics* 35, No. 2&3, pp. 379-387.

Wagenvoort, R.; Waldmann, R. (2002), "On B-Robust Instrumental Variable Estimation of the Linear Model with Panel Data." *Journal of Econometrics* 106, pp. 297-324.

Wooldridge, J. M. (2001), "Applications of Generalized Method of Moments Estimation." *Journal of Economic Perspectives* 15, No. 4, pp. 87-100.