Soft Methods in Robust Statistics

Peter Filzmoser[1]

The focus is on robust regression methods for problems where the predictor matrix has full rank and where it is rank deficient. For the first situation, various robust regression methods have been introduced, and here an overview of the most important proposals is given. For the latter case, robust partial least squares regression is discussed. The way of downweighting outlying observations is important. Using continuous weights (leading to "soft" robust methods) has advantages over 0/1 weights in terms of statistical efficiency of the estimators. This will be illustrated for both types of regression problems. Soft methods are particularly useful in high-dimensional settings.

Keywords: robust regression, partial least squares, high-dimensional data

1 Introduction

The term "soft computing" was coined by Lotfi Zadeh in 1991, and it refers to the design of intelligent systems to process uncertain, imprecise and incomplete information. Since that time, many methods for soft computing have been developed, and their application offers more robust and tractable solutions than conventional techniques. The term "robust" can be seen under various aspects. In this contribution it will be treated in the light of "robust statistics" which includes statistical approaches that are less influenced by outlying observations and deviations from strict statistical model assumptions [3]. Soft computing, and hence soft methods, are also common practice in this field, and they refer to the way how data information is prepared for the statistical methodology. While classical methods give equal weight to each data point, robust methods downweight atypical observations. The

Vienna University of Technology, Department of Statistics and Probability Theory, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria P.Filzmoser@tuwien.ac.at

weights could either be chosen as 0 or 1, corresponding to rejecting the observation or not, or continuously in the interval [0,1]. The latter case can be associated with *soft methods in robust statistics*. Such methods should ideally only discard data points if they are extremely distinct from the bulk of the data. In all other cases, the information contained in the data should to some extent be taken into account. The advantage of such a procedure is usually an increase in statistical efficiency of the resulting estimator.

In this contribution we will focus on robust regression. Section 2 provides an overview of the most important proposals and explains the choice of the weight functions. Section 3 contains methods that can be used for high-dimensional problems. Here the choice of the weights is even more important. In section 4 we compare the efficiencies of the robust regression methods by a simulation study.

2 Robust Regression

In a multiple linear regression model we consider the observations $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ of a response variable and an $n \times p$ matrix $\boldsymbol{X}$ of non-random predictor variables with elements x_{ij}. For a regression model with intercept the first column of $\boldsymbol{X}$ is a column of ones. The i-th observation of the predictor variables is denoted by the column vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t$. The linear regression model is then given by

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + e_i \quad \text{for} \quad i = 1, \ldots, n, \tag{1}$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ the unknown regression coefficients, and e_i the error terms which are assumed to be i.i.d. random variables. The goal is to estimate the regression coefficients. For a given estimator $\hat{\boldsymbol{\beta}}$ the resulting i-th residual is $r_i = r_i(\hat{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}$. The classical least squares (LS) estimator is defined as

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i \left(\boldsymbol{\beta}\right)^2. \tag{2}$$

Under the assumptions of normally distributed errors with the same variance, and if $\boldsymbol{X}$ has full rank, this estimator is known to have excellent statistical properties. However, if the assumptions are violated, and in particular if outliers are contained either in the response, in the predictors, or in both, the performance of the LS estimator can be very poor [3].

2.1 Regression M Estimates

For this reason, the M estimator for regression was introduced as

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i\left(\boldsymbol{\beta}\right)}{\hat{\sigma}}\right), \tag{3}$$

where $\hat{\sigma}$ is a robust scale estimator of the residuals, which makes the regression estimator scale equivariant [2]. The function ρ controls the weighting of the scaled residuals, and it needs to be chosen carefully. It should be a bounded function such that very large residuals will have a limited influence on the estimator. A popular choice is the *bisquare* (also called *biweight*) family, with

$$\rho\left(r\right) = \begin{cases} \left(\frac{r}{k}\right)^2 \left(3 - 3\left(\frac{r}{k}\right)^2 + \left(\frac{r}{k}\right)^4\right) & \text{for} \quad |r| \leq k \\ 1 & \text{otherwise} \end{cases}. \tag{4}$$

The value k is a tuning parameter, balancing efficiency and robustness. For $k \to \infty$, the corresponding estimate tends to LS and hence it becomes more efficient but at the same time less robust. Differentiation of (3) with respect to $\boldsymbol{\beta}$ gives a robustified version of the normal equations,

$$\sum_{i=1}^{n} w_i(\boldsymbol{\beta})(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})\boldsymbol{x}_i = \mathbf{0} \tag{5}$$

with the weights $w_i(\boldsymbol{\beta}) = \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) / \left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right)$ where $\psi = \rho'$. The solution can be found by the IRWLS (iteratively reweighted least squares) algorithm. However, the resulting estimator is only robust with respect to outliers in the residuals, but it is still not robust against outliers in the predictor variables. This can be seen in the definition of the weights w_i, where only outliers in the residual space are considered. The crucial point is the way how the residual scale ($\hat{\sigma}$ in Equation (3)) is estimated.

2.2 Regression S Estimates

A possibility to estimate the residual scale is to use an *M estimator of scale*, which is defined as the solution σ of the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\sigma}\right) = \delta, \tag{6}$$

where ρ is a bounded ρ-function (e.g. the bisquare function) and δ is a fixed constant with $\delta \in (0, \rho\left(\infty\right))$. Dividing Equation (6) by $(r_i/\sigma)^2$ yields

$$\sigma^2 = \frac{1}{n\delta}\sum_{i=1}^{n} \frac{\rho\left(\frac{r_i}{\sigma}\right)}{\left(\frac{r_i}{\sigma}\right)^2} r_i^2 = \frac{1}{n\delta}\sum_{i=1}^{n} w_i r_i^2 \tag{7}$$

with weights $w_i = \rho\left(\frac{r_i}{\sigma}\right) / \left(\frac{r_i}{\sigma}\right)^2$. Given some starting value σ_0, an iterative procedure can be implemented to find the M estimator of scale $\hat{\sigma}$. Using this robust scale estimator, a robust regression estimator can be defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{S}} = \arg\min_{\boldsymbol{\beta}} \hat{\sigma}\big(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta})\big) \tag{8}$$

resulting in the *regression S estimator* [1]. In can be shown that regression S estimators satisfy Equation (5), which implies that they can be computed by an IRWLS algorithm. Although regression S estimators achieve highest possible robustness, the efficiency of this estimator with ρ taken as the bisquare function is only 29%, and in general it cannot exceed 33%.

2.3 Regression MM Estimates

A way to obtain the highest possible robustness with controllable efficiency is given by *regression MM estimators* [8]. The procedure for the computation is as follows [5]:

- Compute an initial estimator $\hat{\boldsymbol{\beta}}_0$; this is done by a regression S estimator (8) which is robust but inefficient.
- Compute a robust scale $\hat{\sigma}$ of the residuals $r_i(\hat{\boldsymbol{\beta}}_0)$; this is done by an M estimator of scale (6).
- Compute $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ as a local solution of (3) using the IRWLS algorithm starting from $\hat{\boldsymbol{\beta}}_0$. The resulting MM estimator inherits its robustness from $\hat{\boldsymbol{\beta}}_0$, and the efficiency can be controlled by the parameter k from the bisquare function (4). Using $k = 3.44$ in this step yields an asymptotic efficiency of 0.85. A higher value is not recommended because this would lead to an increase of the bias [3].

2.4 Hard Rejection of Outliers for Regression

Regression MM estimators use weights for the observations from the interval $[0, 1]$. The further the weights are away from 1, the less information is used from these observations. A popular regression estimator using weights of 0 and 1 for hard rejection of outliers is the LTS (least trimmed sum of squares) estimator [4]. Similar to Equation (7), this estimator minimizes a measure of scale, namely the *trimmed squares scale*

$$\sigma = \left(\frac{1}{n}\sum_{i=1}^{h} |r|^2_{(i)}\right)^{1/2}, \tag{9}$$

where $|r|_{(1)} \leq ... \leq |r|_{(n)}$ are the ordered absolute values of the residuals. Here, h determines the trimming proportion, and for obtaining the highest possible robustness one has to take h equal to (the integer part of) $(n + p + 1)/2$. Similar to Equation (8), the LTS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{LTS}}$ is given by $\hat{\sigma}$ that results from minimizing (9). The asymptotic efficiency of the LTS estimator is only about 7%. Thus, although hard rejection of outliers results in a robust estimator, the efficiency is much lower than that of the MM estimator which uses "soft" weights corresponding to the "useful" data information.

3 Partial Robust Regression

There exist many problems where the number of the explanatory variables is much higher than the number of observations. This situation frequently occurs in chemometrics, biostatistics, in applications of marketing and econometrics, and in various other fields. Because of singularity, neither the LS estimator could be used here, nor any of the discussed robust regression methods. Partial least squares (PLS) regression, a method originally coming from chemometrics, can deal with this situation, see, e.g. [7]. The idea is to use only partial information for regression. Hence, rather than considering the regression model (1), a so-called latent variable model

$$y_i = \boldsymbol{u}_i^t \boldsymbol{\gamma} + e_i \quad \text{for} \quad i = 1, \ldots, n, \tag{10}$$

is used, where $\boldsymbol{u}_i$ are *score vectors* of length $h < p$, $\boldsymbol{\gamma}$ are the regression coefficients, and e_i the error terms. The scores $\boldsymbol{u}_i$ include only partial information contained in the original $\boldsymbol{x}_i$'s because they are of lower dimension. They are computed by $\boldsymbol{u}_i^t = \boldsymbol{x}_i^t \boldsymbol{A}$, with the so-called *loading matrix* $\boldsymbol{A}$ of dimension $p \times h$. The columns $\boldsymbol{a}_k$, $k = 1, \ldots, h$, of $\boldsymbol{A}$ are obtained sequentially by

$$\boldsymbol{a}_k = \arg\max_{\boldsymbol{a}} \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{a}) \tag{11}$$

under the constraints $\|\boldsymbol{a}\| = 1$ and $\mathrm{Cov}(\boldsymbol{X}\boldsymbol{a}, \boldsymbol{X}\boldsymbol{a_j}) = 0$ for $1 \leq j < k$. Once $\hat{\boldsymbol{\gamma}}$ is obtained, the final estimate for $\boldsymbol{\beta}$ for the original model (1) is directly obtained as $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\hat{\boldsymbol{\gamma}}$.

The crucial point is the estimation of 'Cov' in Equation (11). For classical PLS regression, the sample covariance is used. For the robust case, several proposals were made, including robust covariance estimation, see [7]. Here we refer to a highly robust and efficient method called partial robust M regression [6]. The idea is to use for 'Cov' the sample covariance for weighted observations $w_i \boldsymbol{x}_i$ and $w_i y_i$ with weights $w_i = \sqrt{w_i^u w_i^r}$, for $i = 1, \ldots, n$. In terms of the latent variable model (10), the weights originate from

$$\hat{\boldsymbol{\gamma}}_{\mathrm{RM}} = \arg\min_{\boldsymbol{\gamma}} \sum_{i=1}^{n} w_i^u w_i^r \left(y_i - \boldsymbol{u}_i^t \boldsymbol{\gamma}\right)^2 . \tag{12}$$

'RM' stands for *robust M regression*, because Equation (12) corresponds to an M estimator (3) with weights w_i^r for outliers in the residuals, but has additional weights w_i^u for outliers in the scores. The latter weights make the estimator fully robust against all types of contamination. The weights can be chosen according to the so-called Fair function $f(z,c) = 1/\left(1 + |\frac{z}{c}|\right)^2$, where

$$w_i^r = f\left(\frac{r_i}{\hat{\sigma}}, c\right) \quad \text{and} \quad w_i^u = f\left(\frac{\|\boldsymbol{u}_i - \tilde{\boldsymbol{u}}\|}{\mathrm{median}_i \|\boldsymbol{u}_i - \tilde{\boldsymbol{u}}\|}, c\right) \tag{13}$$

with $c = 4$, see [6]. Here, $r_i = r_i(\boldsymbol{\gamma}) = y_i - \boldsymbol{u}_i^t \boldsymbol{\gamma}$ are the residuals from (12), $\hat{\sigma}$ is a robust scale estimate of the residuals, and $\tilde{\boldsymbol{u}}$ denotes the robust center of the scores. Using initial robust weights, an iterative procedure can be formulated to obtain the solution $\hat{\boldsymbol{\beta}}_{\mathrm{RM}} = \boldsymbol{A}\hat{\boldsymbol{\gamma}}_{\mathrm{RM}}$, see [6]. The need for an iterative procedure is also the reason why this rather simple weighting scheme is recommended. An MM estimator would achieve higher efficiency, but–depending on the dimensionality of the problem–it would cause a substantial increase in computation time.

The weights in (13) are chosen from the interval $[0, 1]$, and thus this is another example of "soft weighting". It is easy to modify the weights in order to get hard rejection of the outliers by replacing $f(z, c)$ in (13) by

$$\tilde{f}(z,c) = \begin{cases} 1 & \text{if } |z| \le c \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

with $c = 2.5$. The resulting estimator has the advantage that large values of $|z|$ have no effect, but the disadvantage that intermediate outliers are either completely rejected of fully included.

4 Soft Versus Hard Rejection: A Simulation Study

The use of a continuous weight function, or of weights 0 and 1, will affect the efficiency of the regression estimator. It seems obvious that soft rejection, i.e. the use of "soft" weights, is able to include information that is potentially relevant to improve the statistical precision of the estimator, while hard rejection may fail to use this information. Note that with both types of weighting schemes it is possible to achieve highest possible robustness.

In the following simulation study the effects of different choices of the weights on the efficiency of the estimators will be illustrated. For the regression model (1) we generate standard normally distributed values, forming the elements of the $n \times p$ matrix $\boldsymbol{X}$. For the latent variable model (10) an $n \times h$

score matrix $\boldsymbol{U}$ and a $p \times h$ loading matrix $\boldsymbol{A}$ are generated, both filled with random standard normal numbers, and the predictor matrix is obtained by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{A}^t$. Thus, for $h < p$ a situation with perfect collinearity is simulated. For each considered n and p, the predictor part is fixed. For both cases, the true regression parameters are denoted by $\boldsymbol{\beta}_0$, with components randomly drawn from a standard normal distribution, leading to a model

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta}_0 + e_i \quad \text{for} \quad i = 1, \ldots, n. \tag{15}$$

The error terms e_i are simulated from various distributions: standard normal, Laplace, Student t with 5 and 2 degrees of freedom, Cauchy, and Slash. The latter two are heavy-tailed distributions. From every generated sample with specific values of n, p, and h (for the latent variable model), the estimate $\hat{\boldsymbol{\beta}}^j$ is computed for $j = 1, \ldots, m$, using $m = 1000$ replications. The precision of the estimator is measured by the mean squared error (MSE), given by

$$\text{MSE} = \frac{1}{m} \sum_{j=1}^{m} \|\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}_0\|^2. \tag{16}$$

The results are shown in Figure 1 (for the regression model) and in Figure 2 (for the latent variable model). For the regression model we compare LS, LTS, S, and MM estimation. The LS estimator performs very poor under heavier-tailed distributions, while the robust regression methods are not much affected by the different error distributions. Overall, the MM estimator shows the best efficiency among the robust estimators, and it is able to compete with LS regression under normal errors.

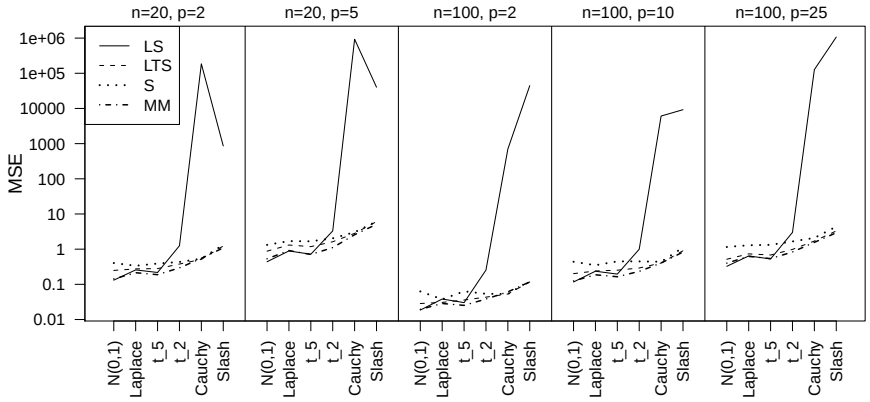

Fig. 1 Simulated MSEs for LS, LTS, S, and MM regression, using different error distributions (legend on the bottom), and different dimensions of the predictor matrix (legend on top).

For the latent variable model we compare in Figure 2 the results of classical estimation (PLS) and robust estimation using the weight function (13) for soft rejection (PRM) and the weights (14) for hard rejection (PRM01). Again, classical estimation dramatically fails for heavy-tailed error distributions. The efficiencies based on hard and soft weighting differ more and more with increasing dimensionality of the predictor matrix: while they differ by a factor of 1.1 to 1.8 for dimensions up to $p = 20$, the ratio increases to a value of 2.5 to 2.8 for $p = 1000$. "Intelligent" robustness–in contrast to robustness based on outlier rejection–thus becomes particularly important for high-dimensional problems, which occur frequently nowadays in practice.

References

1. Davies P (1987) The Annals of Statistics 15:1269–1292
2. Huber PJ (1981) Robust statistics. John Wiley & Sons, New York
3. Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. John Wiley & Sons Canada Ltd., Toronto, ON
4. Rousseeuw PJ (1984) Journal of the American Statistical Association 79:871–880.
5. Salibian-Barrera M, Yohai VJ (2006) Journal of Computational and Graphical Statistics 15:414–427
6. Serneels S, Croux C, Filzmoser P, Van Espen PJ (2005) Chemometrics and Intelligent Laboratory Systems 79(1-2):55–64
7. Varmuza K, Filzmoser P (2009) Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton, FL
8. Yohai VJ (1987) The Annals of Statistics 15:642–65

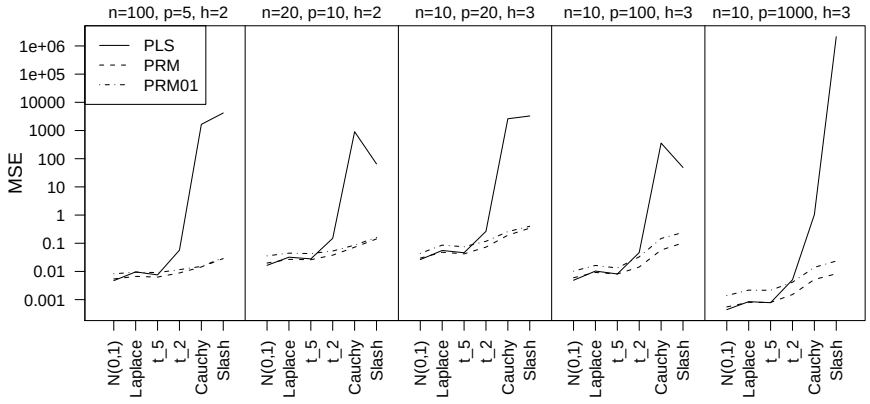

Fig. 2 Simulated MSEs for classical (PLS) and robust partial least squares regression based on soft (PRM) and hard rejection (PRM01), using different error distributions (legend on the bottom), and different dimensions and ranks of the predictor matrix (legend on top).