

# Fast and Robust Classifiers Adjusted for Skewness

Mia Hubert<sup>1</sup> and Stephan Van der Veeken<sup>2</sup>

<sup>1</sup> Department of Mathematics - LStat, Katholieke Universiteit Leuven  
Celestijnenlaan 200B, Leuven, Belgium, *Mia.Hubert@wis.kuleuven.be*

<sup>2</sup> Department of Mathematics - LStat, Katholieke Universiteit Leuven  
Celestijnenlaan 200B, Leuven, Belgium, *Stephan.Vanderveeken@wis.kuleuven.be*

**Abstract.** In this paper we propose two new classification rules for skewed distributions. They are based on the adjusted outlyingness (AO), as introduced in Brys et al. (2005) and applied to outlier detection in Hubert and Van der Veeken (2008). The new rules combine ideas of AO with the classification method proposed in Billor et al. (2008). We investigate their performance on simulated data, as well as on a real data example. Throughout we compare the classifiers with the recent approach of Hubert and Van der Veeken (2010) which assigns a new observation to the group to which it attains the minimal adjusted outlyingness. The results show that the new classification rules perform better when the group sizes are unequal.

**Keywords:** robustness, classification, outlyingness

## 1 Introduction

In a classification context, a random sample from a group of  $k$  populations is given and the aim is to construct a rule to classify a new observation into one of the  $k$  populations. Many of the classification methods proposed in the literature rely on quiet strict distributional assumptions such as multivariate normality, or at least elliptical symmetry. Moreover many are sensitive to outliers in the data. Recently we proposed a classification rule based on the adjusted outlyingness (Hubert and Van der Veeken (2010)). This classifier does not rely on any distributional assumption and is robust to outliers. Observations are classified in the group to which they attain minimal adjusted outlyingness (AO). This AO can be seen as a type of projection depth, and hence this approach corresponds to assigning an observation to the group for which it attains *maximal* depth. Consequently this method generalizes the maximum depth classifiers of Ghosh and Chaudhuri (2005). In Billor et al. (2008) a slight modification to the work of Ghosh and Chaudhuri (2005) has been proposed. Observations are classified according to the group for which the *rank* of their depth is maximal. In this paper we propose two modifications of our original rule based on the AO, in the line of Billor et al. (2008). Simulation results and an application to a real data set show that we obtain lower misclassification errors when the group sizes are unequal. In

Section 2 we define the different classification rules. Section 3 contains the results of a simulation study, whereas in Section 4 we apply our rules to a real data set.

## 2 Construction of the classification rules

We assume we have sampled observations from  $k$  different classes  $X^j$ ,  $j = 1, \dots, k$ . The data belonging to group  $X^j$  are denoted by  $\mathbf{x}_i^j$  for  $i = 1, \dots, n_j$ . The dimension of the data space is  $p$  and is assumed to be much smaller than the sample sizes. In Hubert and Van der Veeken (2010) the following classification rule was proposed: for each new observation  $\mathbf{y}$  to be classified, its *adjusted outlyingness* with respect to each group  $X^j$  is calculated. Then  $\mathbf{y}$  is assigned to the group for which its adjusted outlyingness is minimal. The adjusted outlyingness is introduced in Brys et al. (2005) and studied in detail in Hubert and Van der Veeken (2008). It generalizes the Stahel-Donoho outlyingness towards skewed data. The skewness is estimated by means of the medcouple (MC), a robust measure of skewness (Brys et al. (2004)). For univariate data, the adjusted outlyingness of an observation  $x_i^j$  w.r.t. its group  $X^j$  is defined as:

$$\text{AO}^{(1)}(x_i^j, X^j) = \begin{cases} \frac{x_i^j - \text{med}(X^j)}{c_2 - \text{med}(X^j)} & \text{if } x_i^j > \text{med}(X^j) \\ \frac{\text{med}(X^j) - x_i^j}{\text{med}(X^j) - c_1} & \text{if } x_i^j < \text{med}(X^j) \end{cases} \quad (1)$$

where  $c_1$  corresponds to the smallest observation greater than  $Q_1 - 1.5e^{-4}\text{MC IQR}$ , and  $c_2$  to the largest observation smaller than  $Q_3 + 1.5e^3\text{MC IQR}$ . The notations  $Q_1$  and  $Q_3$  stand for the first and third quartile of the data, and  $\text{IQR} = Q_3 - Q_1$  is the interquartile range. This definition assumes that the data are right skewed, which is concluded when  $\text{MC} > 0$ . If the medcouple is negative, the  $\text{AO}^{(1)}$  is computed on the inverted data  $-X^j$ . In order to define the adjusted outlyingness for a multivariate data point  $\mathbf{x}_i^j$ , the data are projected on all possible directions  $\mathbf{a}$  and the  $\text{AO}^{(1)}$  is computed. The overall  $\text{AO}_i^j$  is then defined as the supremum over all univariate  $\text{AO}$ 's:

$$\text{AO}_i^j = \text{AO}(\mathbf{x}_i^j, X^j) = \sup_{\mathbf{a} \in \mathbb{R}^p} \text{AO}^{(1)}(\mathbf{a}^t \mathbf{x}_i^j, X^j \mathbf{a}). \quad (2)$$

Since in practice it is impossible to consider all possible directions, we use  $m = 250p$  directions. Random directions are generated as the direction perpendicular to the subspace spanned by  $p$  observations, randomly drawn from  $X^j$ . As such, the  $\text{AO}$  is invariant to affine transformations of the data. Note that this procedure can only be applied in our classification setting when  $p < n_j$ , and when the dimension  $p$  is not too large (say  $p < 10$ ). Otherwise taking  $250p$  directions is insufficient and more directions are required

to achieve good estimates. We do not consider this as an important drawback of our rules as it is well known that skewness is only an issue in small dimensions (when the dimensionality increases, the data are more and more concentrated in an outer shell of the distribution). Of course, the algorithm can be easily adapted to search over more than  $m$  directions, but this will come at the cost of more computation time.

To apply our new classification rules, we have to define the outlyingness of a *new observation*  $\mathbf{y}$  w.r.t. each group  $X^j$ . One approach would be to compute this outlyingness  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  according to (2), with  $\tilde{X}^j$  the augmented data set  $\tilde{X}^j = \{\mathbf{x}_1^j, \dots, \mathbf{x}_{n_j}^j, \mathbf{y}\}$ . This would of course become computationally very demanding when many new observations need to be classified, as then the augmented data set is modified for each new observation and the median, the IQR and the medcouple have to be recomputed each time on every projection. Hence, we compute the outlyingness of  $\mathbf{y}$  w.r.t. a fixed data set which does not include  $\mathbf{y}$ . A natural candidate is of course  $X^j$  itself. However, we obtain better results when we first remove the outliers from  $X^j$ . As explained in Hubert and Van der Veeken (2008) this can be easily performed by first computing the adjusted outlyingness of all observations  $\text{AO}_i^j$  in group  $X^j$ . Then the univariate outlyingness of every  $\text{AO}_i^j$  is computed. Formally we can define the *outlier score*  $\text{OS}_i = \text{AO}^{(1)}(\text{AO}_i^j, \{\text{AO}_i^j\})$ . Observations with a 'too large' outlyingness can be defined as those  $\mathbf{x}_i^j$  for which  $\text{AO}_i^j > \text{med}(\text{AO}_i^j)$  and  $\text{OS}_i > 1$  (or equivalently with  $\text{AO}_i^j > c_2$ ). Those observations are removed from  $X^j$ , yielding  $\tilde{X}^j$ . To compute the outlyingness of a new case  $\mathbf{y}$ , we then consider  $\text{AO}(\mathbf{y}, \tilde{X}^j)$ , so we fixed the median, IQR and medcouple of the projected outlier-free data from group  $j$ . Further we denote  $\{\tilde{\text{AO}}^j\}$  as the set of AO values of the outlier-free group  $\tilde{X}^j$  of size  $\tilde{n}_j$ .

We now consider the following classification rules:

- **Rule 1:** The observation  $\mathbf{y}$  is assigned to the group  $j$  for which  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  is minimal. This is the classification method proposed in Hubert and Van der Veeken (2010).
- **Rule 2:** Let  $r_y^j$  be the 'rank' of  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  with respect to the  $\{\tilde{\text{AO}}^j\}$ , formally defined as

$$r_y^j = \frac{1}{\tilde{n}_j} \sum_{i=1}^{\tilde{n}_j} I(\tilde{\text{AO}}_i^j \leq \text{AO}(\mathbf{y}, \tilde{X}^j)).$$

The observation  $\mathbf{y}$  is then assigned to the group  $j$  for which  $r_y^j$  is minimal. In case of ties, rule 1 is applied. This is the approach which follows closely Billor et al. (2008).

- **Rule 3:** To measure the position of  $\text{AO}(\mathbf{y}, \tilde{X}^j)$  within the  $\{\tilde{\text{AO}}^j\}$ , we do not use the rank, but a distance which is related to the definition of the univariate AO given in (1). Let in general

$$\text{SAO}^{(1)}(x, X) = \text{AO}^{(1)}(x, X) \text{sign}(x - \text{med}(X^j))$$

be the *signed* adjusted outlyingness of an observation  $x$  with respect to a univariate data set  $X$ . The observation  $\mathbf{y}$  is then assigned to the group  $j$  for which  $\text{SAO}^{(1)}(\text{AO}(\mathbf{y}, \tilde{X}^j), \{\tilde{\text{AO}}^j\})$  is minimal.

### 3 Simulation results

In this section, we compare the different classifiers on several simulated data sets. We consider the two-class problem ( $k = 2$ ). In all simulation settings, one uncontaminated group is generated from a normal distribution, while the other uncontaminated cases come from a skew-normal distribution (Azzalini and Dalla Valle (1996)). Using the notation  $\mathbf{0}_p = (0, 0, \dots, 0)^t \in \mathbb{R}^p$ , a  $p$ -dimensional random variable  $Z$  is said to be standard skew-normal distributed  $\text{SN}_p(\mathbf{0}_p, \tilde{\Omega}, \boldsymbol{\alpha})$  if its density function is of the form

$$f(\mathbf{x}) = 2\phi_p(\mathbf{x}; \tilde{\Omega})\Phi(\boldsymbol{\alpha}^t \mathbf{x})$$

where  $\phi_p(\mathbf{x}; \tilde{\Omega})$  is the  $p$ -dimensional normal density with zero mean and correlation matrix  $\tilde{\Omega}$ ,  $\Phi$  is the c.d.f. of the standard normal distribution and  $\boldsymbol{\alpha}$  is a  $p$ -dimensional vector that regulates the skewness. By adding location and scale parameters, we obtain  $X = \boldsymbol{\mu} + \boldsymbol{\omega}^t Z \sim \text{SN}_p(\boldsymbol{\mu}, \Omega, \boldsymbol{\alpha})$  with  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_p)^t$  and  $\Omega = \boldsymbol{\omega}^t \tilde{\Omega} \boldsymbol{\omega}$ .

We also consider contaminated training data, in which 5% of the observations come from a normal distribution. With  $\mathbf{1}_p = (1, 1, \dots, 1)^t \in \mathbb{R}^p$ , we can describe the different simulation settings as follows:

1. (a)  $p = 2, X_1 \sim N_2(\mathbf{0}_2, I_2), X_2 \sim \text{SN}_2(-2.1_2, I_2, 5.1_2)$   
 (b) 5% observations from the second group replaced with observations from  $N_2(-3.1_2, 0.2I_2)$
2. (a)  $p = 3, X_1 \sim N_3(\mathbf{0}_3, I_3), X_2 \sim \text{SN}_3(-2.1_3, I_3, 5.1_3)$   
 (b) 5% observations from the first group replaced with observations from  $N_3(-3.1_3, 0.2I_3)$ .
3. (a)  $p = 5, X_1 \sim N_5(\mathbf{0}_5, I_5), X_2 \sim \text{SN}_5(-1.5.1_5, I_5, 5.1_5)$   
 (b) 5% observations from the first group replaced with observations from  $N_5(-3.1_5, 0.2I_5)$

In the 'equal group size' setting, we use  $n_1 = n_2 = 500$ . We also perform the simulation with unequal group sizes, by taking  $n_1 = 100$  and  $n_2 = 500$ .

From each population we randomly generate  $n_j$  training observations and 100 test data. On the training data we apply the three different classifiers as defined in Section 2. The results of the simulations are summarized in terms of average misclassification errors. The misclassification error is defined as the overall proportion of wrongly classified observations in the test sets. The results listed in Table 1 and Table 2 are average misclassification errors with their respective standard errors over 100 simulations.

	Rule 1	Rule 2	Rule 3
2D, No Cont.	0.0737 (0.0018)	0.0751 (0.0019)	0.0758 (0.0019)
2D, 5% Cont.	0.0744 (0.0021)	0.0751 (0.0021)	0.0756 (0.0021)
3D, No Cont.	0.0440 (0.0015)	0.0449 (0.0016)	0.0451 (0.0016)
3D, 5% Cont.	0.0425 (0.0015)	0.0437 (0.0015)	0.0425 (0.0015)
5D, No Cont.	0.0737 (0.0015)	0.0749 (0.0017)	0.0758 (0.0018)
5D, 5% Cont.	0.0736 (0.0016)	0.0735 (0.0016)	0.0767 (0.0019)

**Table 1.** Simulation results for equal group sizes.

	Rule 1	Rule 2	Rule 3
2D, No Cont.	0.1047 (0.0033)	0.0882 (0.0026)	0.0876 (0.0026)
2D, 5% Cont.	0.0991 (0.0032)	0.0797 (0.0024)	0.0818 (0.0023)
3D, No Cont.	0.0986 (0.0032)	0.0527 (0.0015)	0.0534 (0.0015)
3D, 5% Cont.	0.0965 (0.0032)	0.0533 (0.0018)	0.0499 (0.0017)
5D, No Cont.	0.2298 (0.0042)	0.0930 (0.0026)	0.0909 (0.0028)
5D, 5% Cont.	0.2284 (0.0041)	0.0956 (0.0023)	0.0916 (0.0028)

**Table 2.** Simulation results for unequal group sizes.

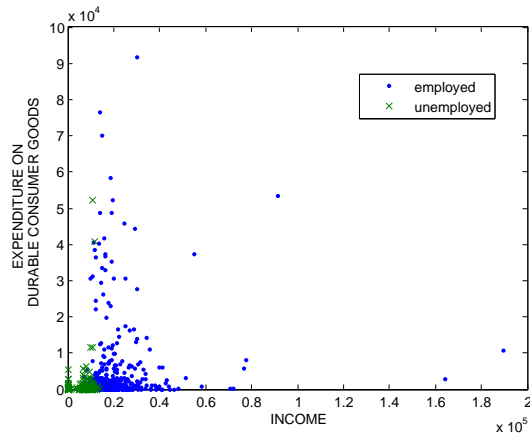
In case that the two groups are of equal size, we see that the three methods have a very comparable performance. Adding 5% outliers does not influence the results significantly. However, when the group sizes are unequal, the new rules 2 and 3 clearly outperform the first rule. This is due to the fact that the distribution of the outlyingnesses is different in both groups. The second and third rule adjust for this difference. The differences between the new rules are not significant (following  $t$ -test).

## 4 Example

The data used in this example come from the Belgian Household Survey of 2005. The Household Survey is a multi-purpose continuous survey carried out by the Social Survey Division of the Institute for National Statistics which collects information on people living in private households in Belgium.

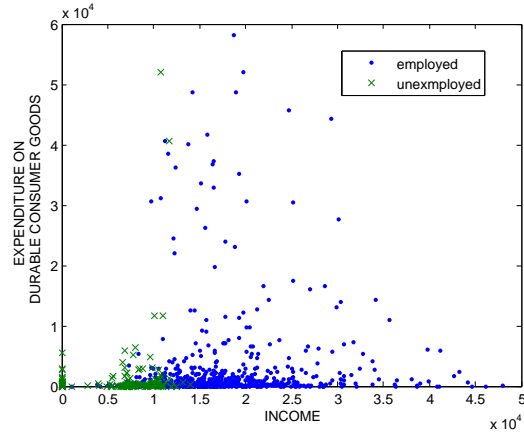
The main aim of the survey is to collect data on a range of topics such as education, welfare, family structure and health. We selected a subset of two variables from the data set: 'Income' and 'Expenditure on durable consumer goods'. In order to avoid correcting factors for family size, only single persons are considered. This group of single persons consists of 174 unemployed and 706 (at least partially) employed persons. Figure 1 is a scatterplot of the data for both groups. As the group of employed people is highly spread out, we have also plotted the lower left part of the data in Figure 2, in which both groups can be better distinguished. We notice the skewness in both classes, as well as some overlap between the groups.

Both groups are split into a training and a test set which contains 10 data points. This is done 100 times in a random way. Rule 1 results in an average misclassification error of 0.2580 with a standard error of 0.0099. Due to the fact that the group sizes are quite different, rules 2 and 3 clearly outperform this result. Rule 2 gives an average misclassification of 0.1655 (s.e. 0.0082) and rule 3 an average classification error of 0.1855 (s.e. 0.0086). This is in line with the simulation results.



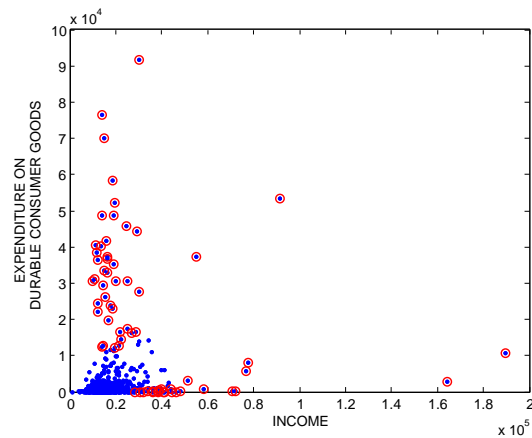
**Fig. 1.** Scatterplot of expenditure on durable consumer goods versus income (complete data set).

For illustrative purposes, we also show in Figures 3 and 4 the outliers in each group, as those observations flagged by their large AO. For the employed group (Figure 3) we find 67 outliers, whereas in the unemployed group (Figure 4) only the two most extreme observations are marked as outliers. For comparison, we also computed the Stahel-Donoho outlyingness of all observations in both groups. Then 179 of the employed and 40 of unemployed persons are flagged as outliers. This is because the skewness is not taken into account and the method searches for the most central elliptical part of the



**Fig. 2.** Scatterplot of expenditure on durable consumer goods versus income (reduced data set).

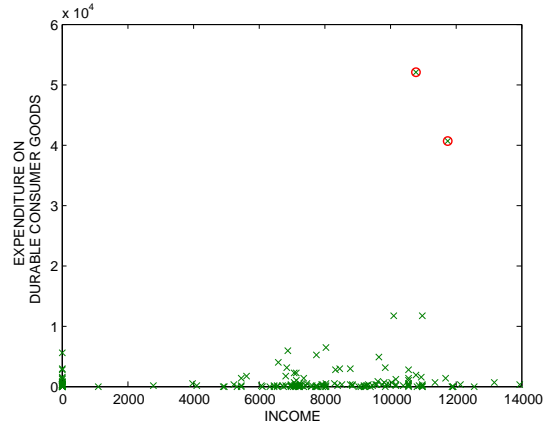
data. For skewed data, it is clearly more appropriate to use skewness-adjusted methods.



**Fig. 3.** Employed persons with outliers marked.

## References

- AZZALINI, A. and DALLA VALLE, A. (1996): The multivariate skew-normal distribution. *Biometrika* 83, 715–726.
- BILLOR, N., ABEBE, A., TURKMEN, A. and NUDURUPATI, S.V. (2008): Classification based on depth transvariations. *Journal of Classification* 25, 249–260.



**Fig. 4.** Unemployed persons with outliers marked.

- BRYNS, G., HUBERT, M. and STRUYF, A. (2004): A robust measure of skewness. *Journal of Computational and Graphical Statistics* 13, 996–1017.
- BRYNS, G., HUBERT, M., and ROUSSEEUW, P.J. (2005): A robustification of Independent Component Analysis. *Journal of Chemometrics* 19, 364–375.
- GHOSH, A.K., and CHAUDHURI, P. (2005): On maximum depth and related classifiers. *Scandinavian Journal of Statistics* 32, 327–350.
- HUBERT, M., and VAN DER VEEKEN, S. (2008): Outlier detection for skewed data. *Journal of Chemometrics* 22, 235–246.
- HUBERT, M., and VAN DER VEEKEN, S. (2010): Robust classification for skewed data. *Advances in Data Analysis and Classification, in press.*