Received: 19 February 2021 | Accepted: 6 June 2022

DOI: 10.1111/rssc.12580

ORIGINAL ARTICLE

Robust correspondence analysis

Marco Riani[1] | **Anthony C. Atkinson**[2] | **Francesca Torti**[3] | **Aldo Corbellini**[1]

[1]Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma, Parma 43100, Italy

[2]The London School of Economics, London, WC2A 2AE, UK

[3]European Commission, Joint Research Centre (JRC), Ispra 21027, Italy

Correspondence
Anthony C. Atkinson, The London School of Economics, London WC2A 2AE, UK.
Email: a.c.atkinson@lse.ac.uk

Abstract

Correspondence analysis is a method for the visual display of information from two-way contingency tables. We introduce a robust form of correspondence analysis based on minimum covariance determinant estimation. This leads to the systematic deletion of outlying rows of the table and to plots of greatly increased informativeness. Our examples are trade flows of clothes and consumer evaluations of the perceived properties of cars. The robust method requires that a specified proportion of the data be used in fitting. To accommodate this requirement we provide an algorithm that uses a subset of complete rows and one row partially, both sets of rows being chosen robustly. We prove the convergence of this algorithm.

KEYWORDS

automobile comparisons, contingency table analysis, informative plotting, minimum covariance determinant estimation, outlier detection, robustness

1 | INTRODUCTION

Correspondence analysis is a method for displaying information from two-way tables of count data. Typically, the rows are subjects (in our first example the 28 countries of the European Union) and the columns are response categories (in that case the cost range of clothes). The main result is a two-dimensional plot showing the structure of the data. The theory and practice of correspondence analysis are presented in several books by Greenacre, most recently Greenacre (2017).

Little attention seems to have been given to the effect of outliers on correspondence analysis nor to the desirability and practice of robust estimation.

Chapter 12 of Greenacre (2017) discusses some aspects of outliers in data tables, including the suggestion of 'supplementary points', which are included in the plotted summary of the data analysis, but are excluded from parameter estimation. Bendixen (1996), in an expository analysis of data on breakfast foods, discusses the effects of outliers and identifies one in his data. However, neither author suggests a systematic procedure for determining which observations should have zero weight. The problems in outlier detection procedures that start from an analysis of all the data and then work backwards by the deletion of apparent outliers have been widely discussed, for example in Atkinson and Riani (2000) Chaps. 3 and 4. One problem is that of 'masking' when the presence of several outliers so affects the parameter estimates that the outliers are not evident. A related problem is 'swamping' in which the outliers cause an uncontaminated observation to appear outlying.

In this paper we adapt the minimum covariance determinant estimator (MCD), a hard-trimming method for multivariate normal data (Rousseeuw & Van Driessen, 1999), to the analysis of contingency tables using correspondence analysis. A complication is that MCD for multivariate data specifies the proportion of observations to be included in data fitting; the highest breakdown point coming when this proportion is just over 50%. For the contingency table this corresponds to some rows being completely included in the analysis with one row of the table usually only partially included. To allow an arbitrary, but pre-specified, proportion of the total observations to be given zero weight, we therefore work in terms of individual counts rather than whole rows of the table. In our examples of data analysis we mainly use the most robust MCD, that is with a breakdown point of 50%, in order to exhibit most strongly the effect on correspondence analysis of outliers and their detection and deletion.

We start in Section 2 with a description of correspondence analysis. Section 3 introduces the example we use to exemplify our robust procedure and provides the non-robust correspondence analysis. We use plots of (squared) Mahalanobis distances to search for outlying rows. Use of an empirical simultaneous confidence interval indicates seven outlying countries. Calculation of the interval requires simulation of 10,000 contingency tables with the same marginal distributions as the data. We discuss approximations to this confidence band.

Robust methods are introduced in Section 4. Section 4.1 describes the MCD for multivariate data and Section 4.2 provides the extension to the analysis of contingency tables. The proof of the convergence of the concentration step of the algorithm for robust fitting is in Section 4.3. The robust analysis of the clothes data is given in Section 5.1. The MCD analysis of the data in Section 5.1 reveals seven outlying countries, five of which are the same as those suggested by the non-robust analysis in Section 3. However, the order and amount of outlyingness is different. The correspondence analysis of the data with the seven outlying countries deleted is given and discussed in Section 5.2. A brief analysis is given in Section 6 of data on seven perceived characteristics of 38 makes of vehicle. The robust analysis leads to the deletion of, again, seven rows and to a correspondence analysis plot that greatly clarifies the interpretation of the data. The paper proper ends with a brief Section 7 in which alternative robust approaches to correspondence analysis are considered.

An on-line supplement (Riani et al., 2022) contains further analyses of the two examples. In the clothes data the countries are in approximately the same positions relative to the axes in the robust and non-robust plots. We include a further, fictitious country, a distinct outlier, to show the different effect that another form of outlier can have. The supplement also includes enhanced correspondence analysis plots of the car data. The discussion focuses on the structure of the characteristics of the cars, rather than, as in Section 6, on the makes of vehicle.

2 | BACKGROUND

2.1 | Contingency tables

Our notation is based on that of Greenacre (2017), particularly Appendix A. The $I \times J$ contingency table N contains count data. Element $(i,j) = n_{ij}, \geq 0,\ \ i = 1,\ 2,\ \ldots\ ,\ I$ and $j = 1, 2, \ldots, J$. The row sums of N are $n_{i.}$ with column sums $n_{.j}$. The grand total $n = \sum_i \sum_j n_{ij} = 1_I' N 1_J$, where 1_I is a column vector of ones of length I and 1_J is of length J. It is customary to rescale N to give P, the correspondence matrix of relative frequencies with element $(i,j) = f_{ij} = n_{ij}/n$, so that $1_I' P 1_J = 1$.

Under the independence hypothesis the expected frequencies are estimated by the multiplication of the row and column totals to give the matrix $\hat{N}$ with $\hat{n}_{ij} = n_{i.}n_{.j}/n$. The test statistic for this hypothesis is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}. \tag{1}$$

The asymptotic null distribution of (1) is χ^2 with degrees of freedom $(I - 1) \times (J - 1)$. The statistic tests whether the row profiles are the same and, given the symmetry of rows and columns, whether the profile columns are similar to each other. The value of X^2/n is often called the (total) inertia of the table. Exhibit 4.2 of Greenacre (2017) illustrates this nomenclature in terms of decreasing similarity of row (and column) profiles giving increasing inertia. It is helpful for our robust procedure to re-express X^2/n as a weighted sum of Mahalanobis distances.

Let the vector r of length I contain the row masses so that $r = P1_J = (f_{1.}, f_{2.}, \ldots, f_{I.})'$. Then $r'1_I = 1$. Also r is the centroid of column profiles. From r we form the $I \times I$ matrix $D_r = \text{diag}(r)$. Similarly for the column masses $c = P'1_I = (f_{.1}, f_{.2}, \ldots, f_{.J})'$, which is also the centroid of row profiles. We now form the $J \times J$ matrix $D_c = \text{diag}(c)$. For use in Section 2.2, we require the $I \times J$ matrix R containing row profiles, with element (i,j) given by $f_{ij}/f_{i.} = n_{ij}/n_{i.}$. Then

$$R = D_r^{-1} \times P = \begin{pmatrix} \tilde{r}_1' \\ \ldots \\ \tilde{r}_I' \end{pmatrix}.$$

Let S be the $I \times J$ matrix containing the weighted standardized row profiles,

$$D_r^{1/2}(R - 1_I c')D_c^{-1/2}.$$

These are the signed square roots of the elements of X^2/n. Then, for example (Greenacre, 2017, A.4), given that $R = D_r^{-1}P$ and $1_I = D_r^{-1}r$

$$S = D_r^{1/2}(R - 1_I c')D_c^{-1/2} = D_r^{1/2}(D_r^{-1}P - D_r^{-1}rc')D_c^{-1/2} = D_r^{-1/2}(P - rc')D_c^{-1/2}. \tag{2}$$

The sum of the squares of the elements of S can be found as

$$\text{trace}(SS') = \text{trace}[D_r^{-1/2}(P - rc')D_c^{-1/2}\{D_r^{-1/2}(P - rc')D_c^{-1/2}\}']$$

$$= \sum_{i=1}^{I} f_{i.}(\tilde{r}_i - c)' D_c^{-1}(\tilde{r}_i - c) \tag{3}$$

$$= \sum_{i=1}^{I} f_{i.} d_i^2(c). \tag{4}$$

Thus the total is a weighted sum of squared distances of the profile points to their respective centroids. Because D_c is diagonal these Mahalanobis distances $d_i^2(c)$ are the squared weighted Euclidean distances of the i-the profile row from its centroid c, with weights defined by the $f_{i.}$. This representation is instrumental in identifying the observations to be downweighted by the robust procedure of Section 4.2.

2.2 | Correspondence analysis

The I row profiles $\tilde{r}_1, \dots, \tilde{r}_I$ define I points in the $J - 1$ dimensional column space. The weight of $\tilde{r}_i$ is $f_{i.}$. The matrix of centred row profiles is $R - 1_I c'$ and can be seen as the matrix of row profiles after removing the zero dimensional subspace (point c) which is closest to points in the $J-1$ dimensional space weighted by D_c^{-1} (Greenacre (2017)). The purpose of correspondence analysis is to perform the singular value decomposition (SVD) of matrix $R - 1_I c'$ in such a way that the left and right singular vectors are orthonormalized with respect to D_r and D_c^{-1}. Thus we require the SVD of

$$U\Gamma V' = R - 1_I c'$$

in such a way that $U' D_r U = I_J = V' D_c^{-1} V$. Here Γ is the diagonal matrix containing the singular values in non-increasing order. In order to achieve this purpose we perform the usual SVD of matrix S in (2)

$$S = D_r^{-1/2}(P - rc')D_c^{-1/2} = D_r^{1/2}(R - 1_I c')D_c^{-1/2} = U^*\Gamma V^{*\prime}.$$

Then

$$(R - 1_I c') = D_r^{-1/2} U^* \Gamma (D_c^{1/2} V^*)' = U\Gamma V', \tag{5}$$

where $U = D_r^{-1/2} U^*$ and $V = D_c^{1/2} V^*$. The matrices U and V are orthonormalized with respect to D_r and D_c^{-1}. In the plot of correspondence analysis the row coordinates which are shown (principal coordinates or row points) are the first two columns of the matrix

$$D_r^{-1/2} U^* \Gamma.$$

The first two columns of the matrix

$$D_c^{-1/2} V^* \Gamma$$

are likewise the coordinates of the column points (principal coordinates of column points). Alternative representations in the low-dimensional space are discussed for example in Greenacre (2013).

3 | A MOTIVATING EXAMPLE: CLOTHES DATA

This section presents the data we shall use to introduce our robust principal component analysis and to illustrate some properties of the method. The data are in the context of European Union (EU) international trade, which is regulated in compliance with the World Trade Organization (1994), and are available publicly in the COMEXT database[1] of the European statistical office (Eurostat).

Table 1 shows occurrences from the 28 members of the EU of the trade flow of clothes not coming from the European Union, according to five price segments. The categories are identified using a methodology discussed in Cerasa and Cerioli (2017) which is based on the robust regression of value against quantity. The resulting index depends not only on price, but also adjusts for quantity; trade flows with quantity below a threshold are removed. The lowest price segment is x_1, with x_5 the highest. The data considered cover a period of several years, ending in 2017.

The main interest in the analysis is in the row profiles: does the proportion of flows in each price segment vary in any meaningful way between countries? We start by looking at the breakdown of X^2 afforded by the Mahalanobis distances $d_i^2(c)$ which are plotted in the upper panel of Figure 1. For ease of interpretation we have included an empirical simultaneous confidence band. This was generated under the independence hypothesis of frequencies that depend only on row and column totals, by simulating 10,000 tables using the algorithm of Boyett (1979) as modified by Patefield (1981) to produce tables with the same row and column totals as the data. A Bonferroni approximation was used to find the 99% simultaneous confidence band for the rows of the table. This required the $100(1{-}0.99/28) = 99.964$ percentage point of the empirical distribution for each country. To be conservative we took the 9997th ordered value out of 10,000. The jagged nature of the envelope is caused by the varying row masses—in particular the largest spike is for unit 25, LU, for which $f_{i.}$ is only 0.0025.

The lower panel plots the distance weighted by the row masses $f_{i.}$, with the appropriately weighted envelope. The seven largest distances correspond to SK, GB, RO, IE, BG, LV and ES, (Table 2) all lying well outside the envelope. AT, which appears as relatively extreme in the unweighted plot appears much less extreme in the weighted plot. Again it is a country with low mass; $f_{i.} = 0.0137$.

We now consider approximations to the distribution of the Mahalanobis distances. Since the statistic X^2 has an asymptotic χ^2 distribution on $(I-1)(J-1)$ degrees of freedom, it is natural to consider χ^2_{J-1} as a first approximation. Simulations based on the marginal properties of Table 1 show that this distribution is slightly too long tailed, even at the 99% level. In order that the row X^2 values have the correct sum of expectations, we next used simulation to explore gamma distributions with parameter $\{(J-1)\times(I-1)/I, 2\}$, tending to χ^2_{J-1} as I becomes large. This approximation gave increased agreement at the 99% level. However, at the 99.99% level the approximation was still too long tailed. Since this is the range of levels that we need for simultaneous intervals, we continue the process and found, by further simulation that, for our table $\Gamma\{(J-1)\times(I-2)/I, 2\}$ gave a good approximation at this extreme level, although it was slightly short tailed at lower levels like 95%. These effects increase for small I and decrease for larger I. Of course, there are also the usual cautions about there being an adequate number of observations

[1] The COMEXT data on the EU external trade can be downloaded from http://epp.eurostat.ec.europa.eu/newxtweb/ or, in comprehensive csv files, from https://ec.europa.eu/eurostat/data/bulkdownload.

TABLE 1 Clothes data. Contingency table between the 28 member states of the European Union (data collected well before Brexit) and five price segments. These are occurrences of country trade flows for a wide set of clothes: x_1 denotes the lowest price segment and x_5 the highest price segment and c' is the row vector of column masses. In all there are 4373 counts

	Price segment					Row mass
Country	x_1	x_2	x_3	x_4	x_5	r
GB	134	76	43	50	49	0.0805
SK	173	62	20	23	16	0.0672
BG	67	76	48	36	23	0.0572
IE	11	21	31	36	52	0.0345
BE	25	32	57	60	58	0.0531
ES	32	42	40	67	67	0.0567
PL	20	35	31	41	41	0.0384
FI	10	16	23	23	24	0.0220
GR	54	28	29	30	23	0.0375
HU	12	19	14	15	20	0.0183
SI	9	10	14	20	23	0.0174
NL	52	43	38	47	54	0.0535
IT	21	36	33	30	36	0.0357
RO	85	74	55	31	22	0.0611
AT	3	8	12	12	25	0.0137
FR	28	33	40	31	45	0.0405
HR	9	17	23	19	34	0.0233
SE	18	36	44	35	40	0.0396
CZ	12	24	22	25	37	0.0274
DK	16	32	35	39	38	0.0366
DE	28	39	36	41	54	0.0453
LT	3	15	22	25	24	0.0204
PT	30	40	28	20	26	0.0329
EE	8	10	12	13	17	0.0137
LU	2	1	2	3	3	0.0025
MT	29	10	16	8	9	0.0165
LV	47	51	29	19	12	0.0361
CY	7	19	20	26	9	0.0185
c'	0.2161	0.2070	0.1868	0.1887	0.2015	1

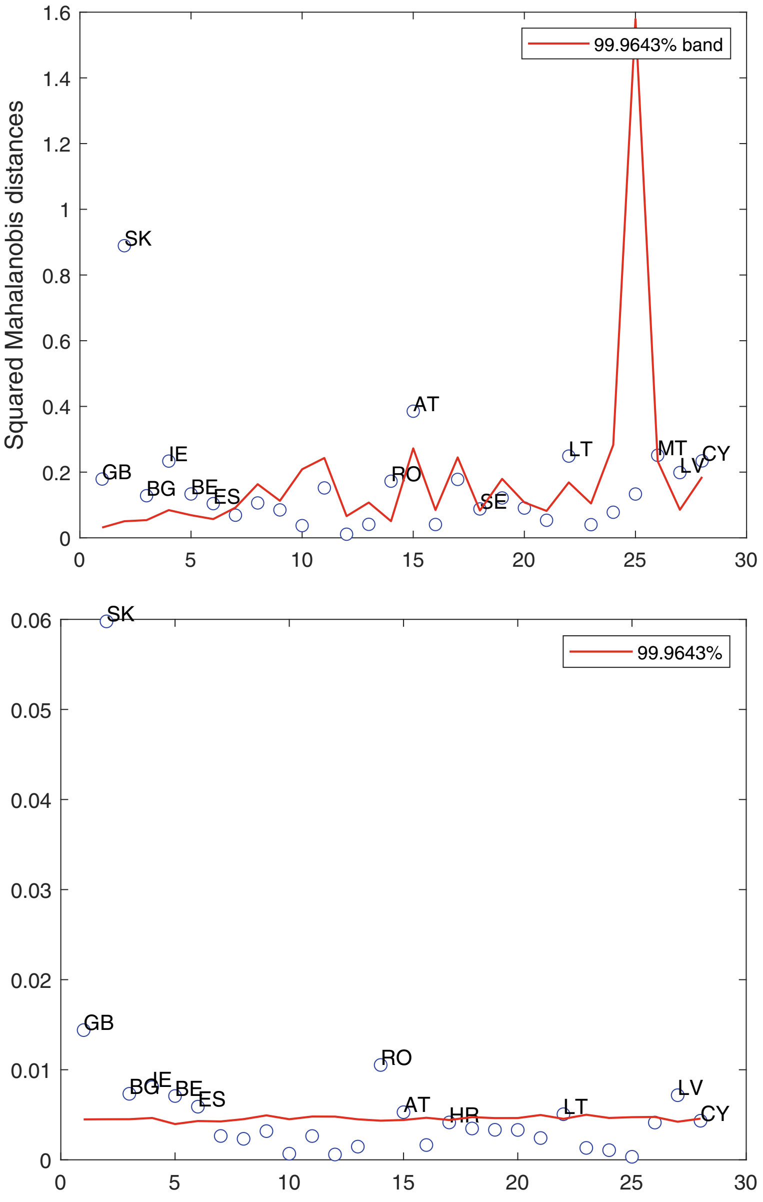

FIGURE 1 Clothes data. Squared Mahalanobis distances of row profiles from c with simulation envelopes. Upper panel: unweighted distances; lower panel: distances weighted by row masses $f_{i.}$ In both panels the numbers on the x axis correspond to the rows of Table 1.

in each cell, for example Agresti (2013). However, 10,000 simulations of our table took about 40 seconds, so such distributional approximations may only be necessary for much larger tables.

We work with the terminology of robustness and outlier detection. It is important to be clear what an outlier means in this context. In regression and the analysis of multivariate normal data, an outlier is an observation which is suspected of being ‘wrong’, perhaps coming from a different population or corrupted by a numerical error or by a systematic measurement error. Here the aim

TABLE 2 Clothes data. Seven outlying countries, ordered from the most outlying, according to four analyses: MHD—squared Mahalanobis distances, CA—correspondence analysis; distances measured by projection onto principal axis

Analysis	Ordered countries						
Traditional MHD (Figure 1)	SK	GB	RO	IE	BG	LV	ES
Traditional CA (Figure 2)	SK	MT	GB	RO	LV	BG	GR
Robust MHD (Figure 3)	SK	GB	RO	BG	LV	GR	MT
Robust CA (Figure 4)	SK	GB	MT	LV	RO	BG	GR

is to find the structure of the rows. In this case an outlier is a row which does not agree with the multiplicative model assuming independence fitted to the data. In the case of Figure 1 this model has been fitted to all I rows, so that each of the potentially outlying rows listed above has had an effect on the estimation of the multiplicative model.

The two panels of Figure 1 respectively indicate results from 11 and 9 countries lying above the 95% Bonferroni bound. There is a strong indication of the presence of outlying observations. However, in the presence of multiple outliers some good observations may appear outlying and some outlying observations may not be evident. These configurations can be discovered by the deletion of observations in a structured way using robust methods. Analyses of numerous outlier-free data sets for other market segments indicate, perhaps surprisingly, the unstructured independence of country profiles. It is therefore meaningful to search for outlying behaviour from this null hypothesis of independence.

We now turn to the correspondence analysis of these data. This should provide information not only on the relationship between the row profiles, but also their relationship to the five ordered price categories. The plot of the correspondence analysis is in Figure 2, with the countries represented by circles. This is not of the standard structure seen in such plots as Exhibits 9.2 and 10.2 of Greenacre (2017) in which the two axes intersect near the centre of a point cloud of the projections of row profiles. Here, on the other hand, the intersection of the axes leaves a cloud of points to its right, with a straggle of points to the left of the intersection. In order, the most remote on the first axis, which accounts for 83.3% of the total variability, are SK, MT, GB, RO, LV, BG and GR. Five of these are indicated as outlying by the plots of Mahalanobis distances in Figure 1. In addition, the order is changed and AT now appears as a member of the central group. The comparison is clarified in Table 2 (Section 5.1).

Also included in Figure 2 are confidence regions for the positions of the column points, derived using a general bootstrapping method suggested by Greenacre (2017, p. 226), which gives elliptical confidence intervals. The website https://www.rdocumentation.org/packages/FactoMineR/versions/2.4/topics/ellipseCA provides algorithms for three methods of constructing the ellipses, which vary in the way the bootstrap samples are generated. In Figure 2 we have plotted the 99.9% region for all x_j, taking the original contingency table as a reference. In the method we used 10,000 new data tables drawn from a multinomial distribution with theoretical frequencies equal to n_{ij}/n. These new data tables are projected as supplementary rows or supplementary columns in the reference table. Confidence ellipses are then drawn using the centroid and the covariance matrix of the 10,000 projected points. In this example the alternative methods based on generating tables bootstrapping row by row or column by column gave virtually indistinguishable ellipses. The figure shows that the main feature of these regions is that all price levels would be considered as significant, since none of the regions includes the origin of the axes.

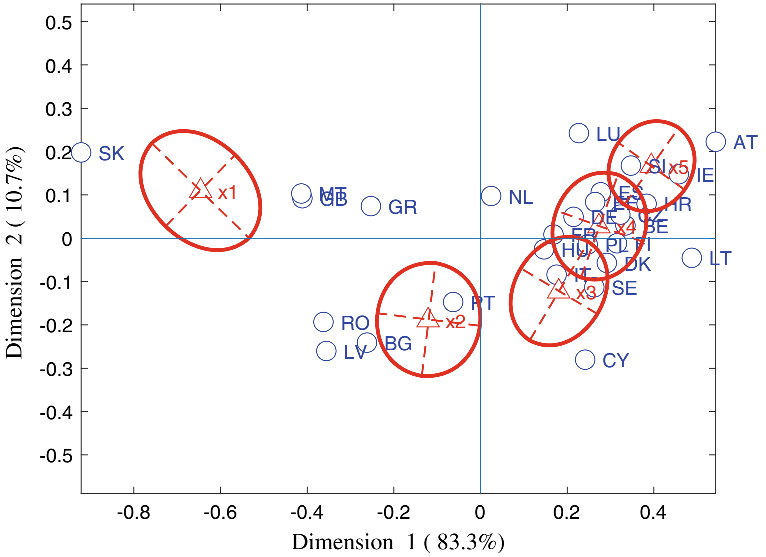

FIGURE 2 Clothes data. Correspondence analysis plot of the data in Table 1 with the points displayed in principal coordinates. The rows are represented by circles and the columns by triangles. Confidence level of the column ellipses is 99.9%.

These two analyses indicate the presence of several outliers among the data from the 28 countries. It is, however, not clear, in the low-dimensional representation, which countries are outlying, nor what effect the outliers are having on the positions of the seemingly non-outlying rows and on the inertia explained.

4 | ROBUST CORRESPONDENCE ANALYSIS

We first discuss the MCD estimator for the parameters μ and Σ of multivariate normal data and then describe the extensions that are needed to apply the MCD to the analysis of contingency tables, leading to the detection of outlying rows, which may be clustered, or ordered, by the column variable.

4.1 | The MCD estimator

The squared Mahalanobis distance of a v-dimensional multivariate normal random variable y_i is

$$d_i^2(\mu, \Sigma) = \{y_i - \mu\}'\Sigma^{-1}\{y_i - \mu\}, \quad i = 1, \dots, n. \tag{6}$$

The contours of constant squared Mahalanobis distances form ellipsoids in v-dimensional space. This simple geometric interpretation suggests the MCD estimator of μ and Σ (Rousseeuw & Van Driessen, 1999), found by hard trimming, which yields a subset of h observations, intended to be outlier free, that provides parameter estimates for the analysis of the data. The number of untrimmed observations $h, \lfloor (n + v + 1)/2 \rfloor \leq h \leq n$ is decided before the data are analysed. If the minimum value of h is chosen, the procedure has a breakdown point of 50%, although the efficiency is very low (Hubert & Debruyne, 2010).

The $n-h$ observations to be trimmed are found numerically. Let $d^2_{(1)}, \leq \ldots, \leq d^2_{(n)}$ be the ordered values of the Mahalanobis distances $d^2_i(\mu, \Sigma)$. Then the MCD estimator of μ and Σ minimizes the trimmed sum

$$S_{\text{MD}}(h) = \sum_{i=1}^{h} d^2_{(i)}. \tag{7}$$

For robust correspondence analysis we adapt a slightly simplified version of the algorithm of Rousseeuw and Van Driessen (1999). This starts by taking a large number (generally 2000) of random subsets of $v+1$ observations. The parameters are estimated from each subset and the Mahalanobis distances calculated for all n observations. These parameter estimates can be improved by a 'concentration' step by ordering the Mahalanobis distances and using the subset containing the smallest h distances to provide new parameter estimates. The procedure is repeated a few times for each random subset. Although the parameter estimates from all subsets could be concentrated to convergence, it is customary to bring to full convergence just the five smallest values of $S_{\text{MD}}(h)$ out of 2000. The subset h^* that provides the smallest overall value of $S_{\text{MD}}(h)$ is used for the robust analysis of the data. Details and proof of convergence of the concentration steps for correspondence analysis are in Section 4.3.

4.2 | MCD correspondence analysis

The subset h of observations in the algorithm described in Section 4.1 can have any integer value between $(n+v+1)/2$ and $n-1$. In correspondence analysis the entries of the table are rows, the ith of which contains $n_{i.}$ observations. In order to calculate the MCD in this situation we need to formulate correspondence analysis when only an arbitrary number h of the n observations is included. We start by rewriting (4) by replacing c by the notation c_n. Then c_n is the vector which satisfies the following minimization

$$\min_{c_n} \sum_{i=1}^{I} f_{i.}(\tilde{r}_i - c_n)' D^{-1}_{c_n}(\tilde{r}_i - c_n) = \min_{c_n} \sum_{i=1}^{I} f_{i.} d^2_i(c_n). \tag{8}$$

In robust estimation for normal data, the trimming of extreme observations leads to biased estimation of variances and covariances. With such data it is desirable to apply a consistency correction to the estimation of the covariance matrix in the Mahalanobis distance to correct for this bias; (see Rousseeuw & Van Driessen, 1999, p. 218). The trimming, however, does not introduce bias into the estimation of the mean. In (8) the covariance matrix is diagonal, but separate estimates of the variances are not required. Because, for the Poisson distribution, the mean and variance are equal, we use sample means in the computation of the robust Mahalanobis distances and so do not need to apply a consistency factor.

Equation (8) can be rewritten in terms of all n observations as

$$\min_{c_n} \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{n_{i.}} (\tilde{r}_{i,n} - c_n)' D^{-1}_{c_n}(\tilde{r}_{i,n} - c_n). \tag{9}$$

The notation $\tilde{r}'_{i,n}$, stresses that this is the ith row of the $I \times J$ matrix R of row profiles based on n observations. Note that the summand is not a function of k.

When only h observations are of interest, the vector c_h satisfies

$$\min_{c_h} \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{n_{i.}} a_n\{O_{ik}(c_h)\}(\tilde{r}_{i,h} - c_h)' D_{c_h}^{-1} (\tilde{r}_{i,h} - c_h) = \min_{c_h} \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{n_{i.}} a_n\{O_{ik}(c_h)\} d_i^2(c_h). \qquad (10)$$

Here $O_{ik}(c_h)$ is the rank order of $d_{ik}^2(c_h)$ among all n distances based on c_h

$$\underbrace{d_1^2(c_h), \ \dots \ , d_1^2(c_h),}_{n_{1.}} \ \dots \ , \underbrace{d_I^2(c_h), \ \dots \ , d_I^2(c_h)}_{n_{I.}} \qquad (11)$$

and $a_n(s) = I(s \leq h), s = 1, 2, \dots , n$.

The computation of c_h is generally based on a subset of l rows which are fully represented with overall mass $h_\ell / n < h/n$ and a partially represented row p with original mass $f_{p.} = h_p / n$. This is represented in the subset of size h with $h - h_\ell$ units. The original frequencies of row p of the contingency table $n_{p1}, n_{p2}, \ \dots \ , \ n_{pJ}$, with row total $n_{p.}$ are modified as follows

$$(h - h_\ell)\frac{n_{p1}}{n_{p.}} \quad (h - h_\ell)\frac{n_{p2}}{n_{p.}} \quad \dots \quad (h - h_\ell)\frac{n_{pJ}}{n_{p.}}. \qquad (12)$$

Calculations for the main part of the MCD algorithm of Section 4.1 then depend on ordering all the n distances based on c_h

$$\underbrace{d_1^2(c_h), \ \dots \ , d_1^2(c_h),}_{n_{1.}} \ \dots \ , \underbrace{d_I^2(c_h), \ \dots \ , d_I^2(c_h)}_{n_{I.}} \qquad (13)$$

and obtaining the value of $S_{\mathrm{MD}}(h)$.

The algorithm starts, as in Section 4.1, from a randomly selected basic subset, in this case of J rows of the table, each of which is completely represented. Unlike traditional MCD analysis here the sizes h_0 of the elemental subsets may vary. An alternative, which we have not followed here, is to use a basic subset of size J containing one observation per row. Let these J rows be the subset J_0, with $h_0 = J$. Then, in the notation of (12), $(h - h_\ell) = 1$ for all selected rows $j \in J_0$ and the cell frequencies are $n_{p1}/n_{p.}, p \in J_0$. In either case, calculation of c_{h0} leads to ordering the n Mahalanobis distances as in (13) with c_{h0} replacing c_h. Note that here the basic subset is of dimension J, rather than $v + 1$, as we do not need an extra degree of freedom to estimate the variance.

4.3 | Concentration steps for contingency table analysis with MCD

We now prove the convergence of the concentration steps for the MCD analysis of contingency tables.

Let $H^{(k)}$ be the current h subset at iteration k associated with a contingency table based on h observations $N_h^{(k)}$ (with $\sum_{i=1}^{I_{(k)}} \sum_{j=1}^{J} n_{ij}^{(k)} = h$). The vectors $c_h^{(k)}$ and $f_{(t.)}$ are respectively the centroid of row profiles and the associated row masses in this contingency table. These h observations are associated with a set of rows of the original contingency table and to another row of the original contingency table which is partially represented. The objective function is

$$S_{\mathrm{MD}}(H^{(k)}) = \sum_{t \in H^{(k)}} f_{(t.)} d^2_{(t)}(c_h^{(k)}).$$

Let $H^{(k+1)}$ be the h subset which contains the smallest weighted distances with respect to $c_h^{(k)}$. By construction:

$$\sum_{t \in H^{(k+1)}} f_{(t.)} d^2_{(t)}(c_h^{(k)}) \leq \sum_{t \in H^{(k)}} f_{(t.)} d^2_{(t)}(c_h^{(k)}) = S_{\mathrm{MD}}(H^{(k)}). \tag{14}$$

Let the new associated contingency table be $N_h^{(k+1)}$ (with $\sum_{i=1}^{I_{(k+1)}} \sum_{j=1}^{J} n_{ij}^{(k+1)} = h$). Since the centroid of the row profiles (vector of column masses) of the contingency table $N^{(k+1)}$, which we denote with $c_h^{(k+1)}$, minimizes (8) it follows that

$$S_{\mathrm{MD}}(H^{(k+1)}) = \sum_{t \in H^{(k+1)}} f_{(t.)} d^2_{(t)}(c_h^{(k+1)}) \leq \sum_{t \in H^{(k+1)}} f_{(t.)} d^2_{(t)}(c_h^{(k)}). \tag{15}$$

Combining (14) and (15), the new h subset $H^{(k+1)}$ has an objective function that is less than or equal to that of $H^{(k)}$. Note that the only way to obtain equality is if no element inside vector c_h has changed, in which case the iteration stops.

5 | MCD ANALYSIS OF THE CLOTHES DATA

We now report the results of the MCD analysis of the clothes data with maximal breakdown point of 50%, that is $h = 2189$. We first establish the outliers and then describe the results of the correspondence analysis with the outlying rows deleted.

5.1 | Outlier detection

The upper panel of Figure 3 shows the plot of the robust Mahalanobis distances by country. This is the robust version of the lower panel of Figure 1. Not only is the simultaneous envelope of course the same in the two figures, but there are again seven outliers, listed in Table 2. AT is no longer marginally outlying, but PT is very close to the envelope.

The MCD estimate with 50% breakdown is highly inefficient. If there are fewer than 50% outliers, a more efficient estimator may be found based on a larger subset than $\lfloor (n+5)/2 \rfloor$. Following a suggestion of Rousseeuw and Van Driessen (1999) for improving the MCD estimate, we use the information from Figure 3 to calculate a reweighted estimate. The data are now reanalysed without the eight rows (including PT) detected as outlying by the MCD reweighting. These contain 2672 observations, so the fit is now with a subset of size 2672 rather than 2189 for MCD. The resulting plot of Mahalanobis distances, together with the simulation envelope, is in the lower panel of Figure 3. Despite the marked difference in subset sizes used for fitting, any difference between the two plots is very hard to detect. However, there is a change in the outlying rows that are detected because, after reweighting, Portugal falls just inside the envelope rather than outside.

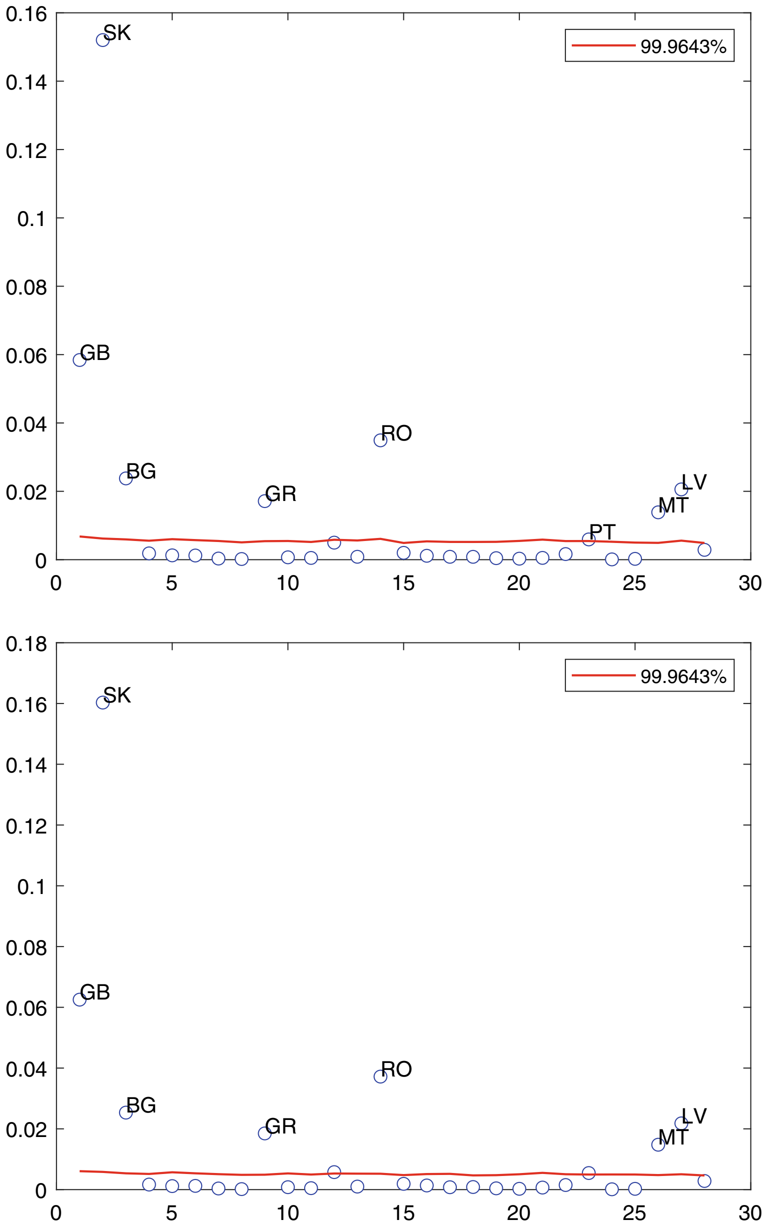

FIGURE 3 Clothes data. Top panel: MCD squared Mahalanobis distances of row profiles from c_h with envelopes. Bottom panel: reweighted MCD Mahalanobis distances. In both panels the distances are weighted by the row masses.

5.2 | Outlier-removed correspondence analysis

Outlier detection as in Section 5.1 leads to a reduced contingency table with $I(h^*)$ rows. We apply the correspondence analysis described in Section 2.2 to the reduced table with $I(h^*) < I$ rows, finding the SVD not of the matrix S in (2) of size $I \times J$, but of the smaller $I(h^*) \times J$ matrix S_* from the reduced table of size $\sum_{i \in I(h^*)} \sum_{j=1}^{J} n_{ij} = n^* < n$. Once the correspondence analysis

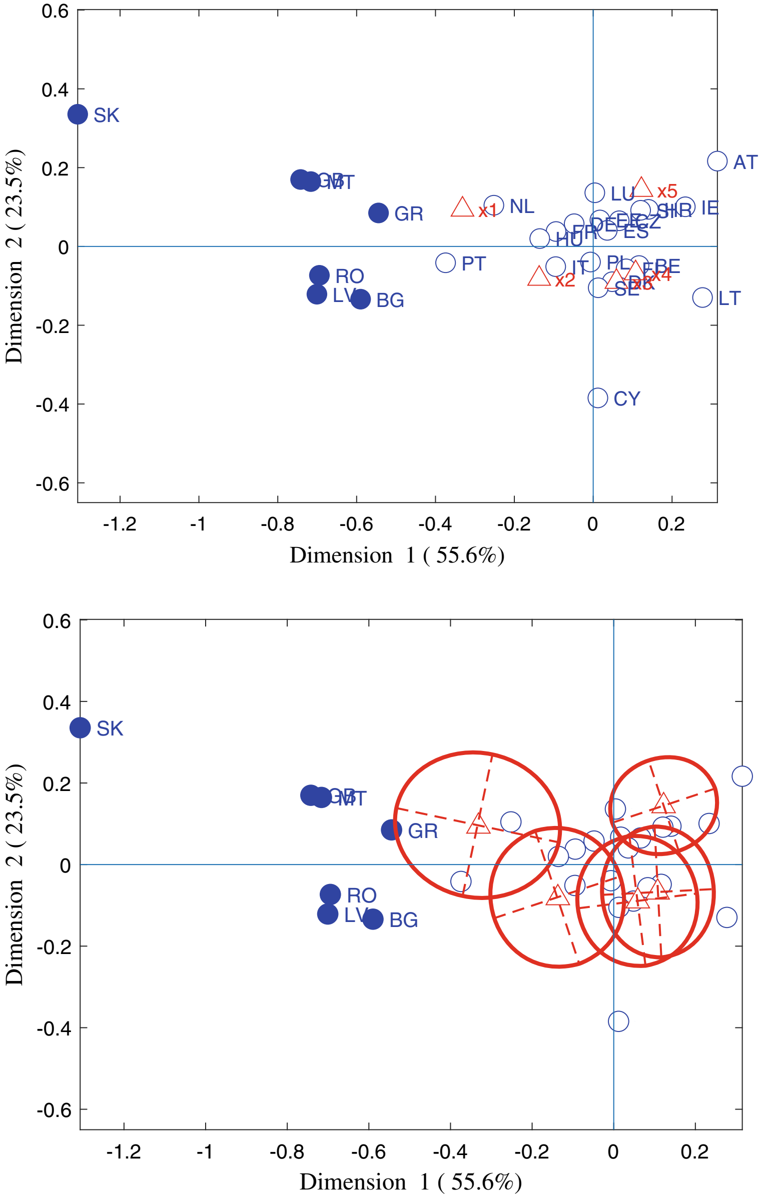

FIGURE 4 Clothes data. Correspondence analysis plot from MCD analysis. Filled circles are the seven deleted observations ('supplementary points'). The ellipses (lower panel) give a 99.9% confidence interval for the positions of $x_1, \dots, x_5$.

transformation has been calculated, we transform the outlying rows in the same way and include them in the plot under the name 'supplementary points' (Greenacre, 2017, Chap. 12).

To continue our analysis of the data we give, in the upper panel of Figure 4, the plot from the correspondence analysis of the data when the seven outlying rows are excluded from estimation of the principal coordinates. They are, however, plotted in the figure, and so have become 'supplementary points'. The points of the remaining 21 observations now form a scatter around the intersection of the axes: there are 16 countries forming a relatively tight cluster with five countries scattered slightly more remotely around them. There is then, at lower values of the first principal axis, a seeming cluster of six countries. The most remote country on this axis is Slovakia. In both this plot and Figure 2 the ordering along this axis seems to reflect an increasing proportion of purchases of the cheapest clothes. The second lowest point on this axis is that for GB. The other five outliers are for countries with much smaller economies; the other major economies of Europe at that time (FR, IT, DE, PL and ES) all have values for the first axis close to zero; that is, they are typical European countries.

The lower panel of Figure 4 repeats the upper panel, but the focus is now on the columns. As in Figure 2 we give the confidence regions for the five column points. Removing the seven outliers from estimation causes a sharp change in the assessment of the significance of columns points. Now the regions for the three central price categories, $x_2 - x_4$, overlap with the origin. After removal of the outliers the information on the effect of price comes only from the two most extreme categories. While in the non-robust representation the confidence ellipses for x_1 and x_2 are very far from those of the other three points, in the robust representation there is considerable overlap in the ellipses for each consecutive pair of price levels.

The first latent dimension can be interpreted as price level; when the outliers are removed, the projection on this dimension of the seven outliers gives values that are much lower than the projection of the column point x_1. The second latent dimension now more sharply contrasts the extreme prices (x_1 and x_5) with the intermediate prices. In the robust correspondence analysis plot, only the column points x_1 and x_5 show positive values for this latent component. It is also interesting to note that, while in the original representation the countries RO, LV and BG seem to be very close to x_2, in the robust representation they are all well to its left.

The robust analysis leads also to the revelation of further information in the data. For example, Figure 5 is a scatterplot matrix of the row profiles with the outliers highlighted. On the diagonals we give the box plots for the five values of x separated into normal and outlying rows. The structure is abundantly clear. The first row shows the high values of x_1 for all seven outlying countries, whereas the values of x_4 and x_5 in the last two rows are almost all lower for the outliers than for

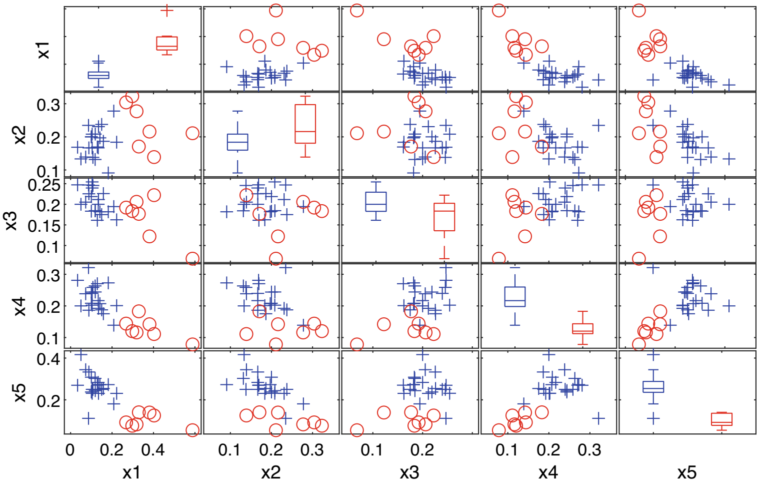

FIGURE 5 Clothes data. Scatter plot matrix of row profiles with outliers shown as circles (red in the online pdf version).

the normal countries. A panel such as that for x_1 and x_5 shows complete separation of the two groups.

The ordered listing in Table 2 of the seven outliers found by the four analyses shows that the most outlying country is SK, followed by GB. Columns 2 and 3 show that RO and MT also occur as the second or third most outlying in some analyses. However, these rankings give no information as to the magnitude, rather than ordering, of the distances between the central group and these countries. What is most important about the robust analysis is the contrast of the robust CA plot of Figure 4 with the traditional CA plot of Figure 2. The outliers are clearly displayed in the robust plot, lying remote from the central cluster, which is now at the intersection of the two principal axes.

In this example, the countries are in approximately the same positions relative to the axes in the robust and non-robust plots. The main effects of the robust analysis are to tighten the cluster of the main 21 observations and to emphasize the existence and structure of the outliers. In the on-line supplement (Riani et al., 2022) we add an extra fictitious country, also outlying, to show the distinct effect that another form of outlier can have.

6 | 2014 CAR DATA

As a second example of the application of robust correspondence analysis we again analyse an example with rows that have a clear resonance, in this case, brands of cars. The data, given in Table 3 are presented by Bora Bera at the web address https://boraberan.wordpress.com/2016/09/22. They are taken from the 2014 Auto Brand Perception survey by Consumer Reports (USA) where 1578 randomly selected adults were asked what they considered exemplary attributes for 39 different vehicle brands. Respondents picked all that they felt applied from among a list that consisted of: Style, Performance, Quality, Safety, Innovation, Value and Fuel Economy.

As in Section 5 we look at several CA plots, using plots of squared Mahalanobis distances from robust and non-robust analyses to guide us in the choice of supplementary points. The focus in the description of this analysis is in the evolving interpretability of CA plots as the analysis progresses.

We start with the non-robust CA plot in Figure 6. The most clear property is that of Safety, on which Volvo scores highly and Subaru well. Otherwise, the cars and their properties are concentrated in an uninformative group with, for example, Performance and Innovation virtually indistinguishable. Smart is the only other outlying make, which is certainly to be expected. Finally, there is one quadrant, the second (NW) one, in which the row points are all close to the origin.

We then performed a robust analysis using MCD with a breakdown point of 0.5. The index plot of Mahalanobis distances had a less regular structure than the two panels of Figure 3, with only eight makes falling below the simultaneous confidence interval. This suggests that there were few standard vehicles on the properties of which consumers were agreed. This is distinct from the first example where the citizens of many EU countries were shown to have similar profiles for the purchases of clothes. Despite this lack of structure in the car data, there were many clear outliers, the largest seven being, in order: Volvo, Toyota, Honda, Kia, Hyundai, Volkswagen and Smart. The CA analysis with these seven vehicles deleted to become supplementary points is in Figure 7. Now that Volvo and Toyota have been deleted, safety is a much less important axis and the importance of other properties is clearer. In particular, Performance and Innovation are now well separated.

TABLE 3 2014 car data. Number of adults out of 1578 stating that a particular make of vehicle had a specific quality. In all there are 11,713 counts

	Exemplary attributes							Row total
Make	**Fuel economy**	**Innovation**	**Performance**	**Quality**	**Safety**	**Style**	**Value**	$n_{i.}$
Acura	24	38	28	20	28	33	25	196
Audi	9	54	54	30	19	67	8	241
Bentley	0	16	18	25	9	27	17	112
BMW	14	83	94	55	38	93	35	412
Buick	25	48	39	58	52	52	43	317
Cadillac	14	73	50	76	40	83	36	372
Chevrolet	114	103	202	174	140	160	145	1038
Chrysler	38	65	96	54	54	103	72	482
Dodge	60	61	141	61	63	133	69	588
Ferrari	0	20	45	10	8	46	5	134
Fiat	19	21	17	20	15	7	16	115
Ford	167	180	169	179	161	157	188	1201
GMC trucks	40	40	64	57	80	50	58	389
Honda	163	68	73	118	104	50	135	711
Hyundai	97	25	31	27	35	42	82	339
Infiniti	5	39	31	15	10	17	16	133
Jaguar	0	3	18	19	3	47	12	102
Jeep	18	33	14	51	19	41	52	228
Kia	68	30	17	13	24	42	109	303
Lamborghini	5	19	37	8	6	23	24	122
Land Rover	0	43	0	5	0	47	2	97
Lexus	10	62	29	50	27	64	26	268
Lincoln	6	37	23	31	24	40	19	180
Maserati	0	6	9	0	0	41	25	81
Mazda	46	23	34	10	12	26	38	189
Mercedes-Benz	8	83	44	87	58	82	42	404
Mini	23	12	4	4	13	12	4	72
Mitsubishi	20	13	33	23	7	32	13	141
Nissan	80	68	51	53	52	55	70	429
Porsche	0	17	66	14	6	42	5	150
Ram trucks	9	22	21	10	18	1	16	97
Rolls-Royce	0	4	4	35	11	25	17	96
Scion	20	24	11	6	11	4	4	80

(Continues)

TABLE 3 (Continued)

	Exemplary attributes							Row total
Make	**Fuel economy**	**Innovation**	**Performance**	**Quality**	**Safety**	**Style**	**Value**	$n_{i.}$
Smart	38	9	3	7	0	5	10	72
Subaru	19	14	32	33	75	20	40	233
Tesla	23	35	10	12	9	15	12	116
Toyota	238	116	95	134	113	74	150	920
Volkswagen	90	30	25	37	27	22	46	277
Volvo	9	15	16	31	180	14	11	276
Total ($n_{.j}$)	1519	1652	1748	1652	1551	1894	1697	11, 713

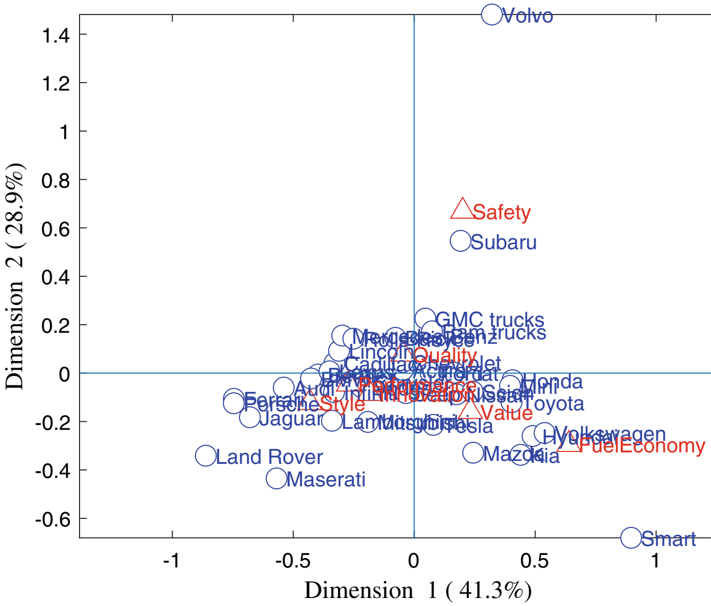

FIGURE 6 2014 car data. Non-robust correspondence analysis plot.

To explore the presence of the large number of indicated outliers in our analysis of the 2014 car data, we repeated the MCD analysis with a breakdown point of 0.25. The results from the unweighted and re-weighted analyses were similar to those for 0.5 breakdown, but did suggest that Land Rover might also be treated as a supplementary point. The resulting correspondence analysis plot is in Figure 8. This final plot contains easily understood information on consumers' perceptions of cars. This is particularly the case in the left-hand half of the plot, where the variables are Quality, Style and Performance. There are several surprising conclusions, such as the orthogonality of Performance and Fuel Economy and the presence of Innovation virtually on the intersection of the axes, suggesting the absence of consistent perceptions about innovation. In the transparent structure of this plot there is a wide range of values for the row points in all quadrants.

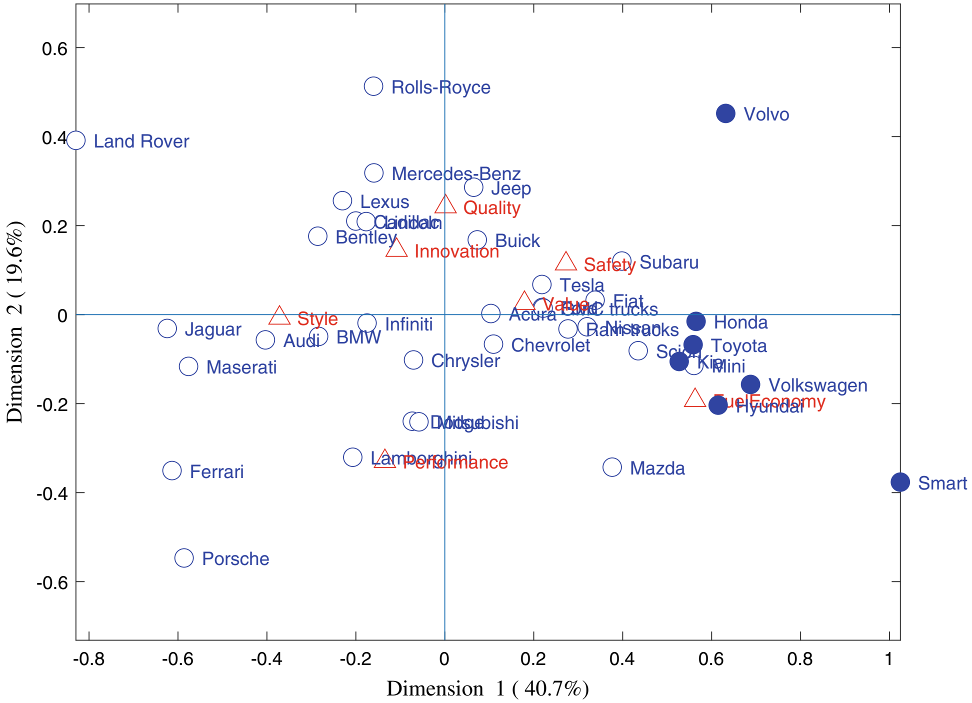

FIGURE 7 2014 car data. Correspondence analysis plot from MCD analysis, breakdown point = 0.5. Filled circles are the seven deleted observations ('supplementary points').

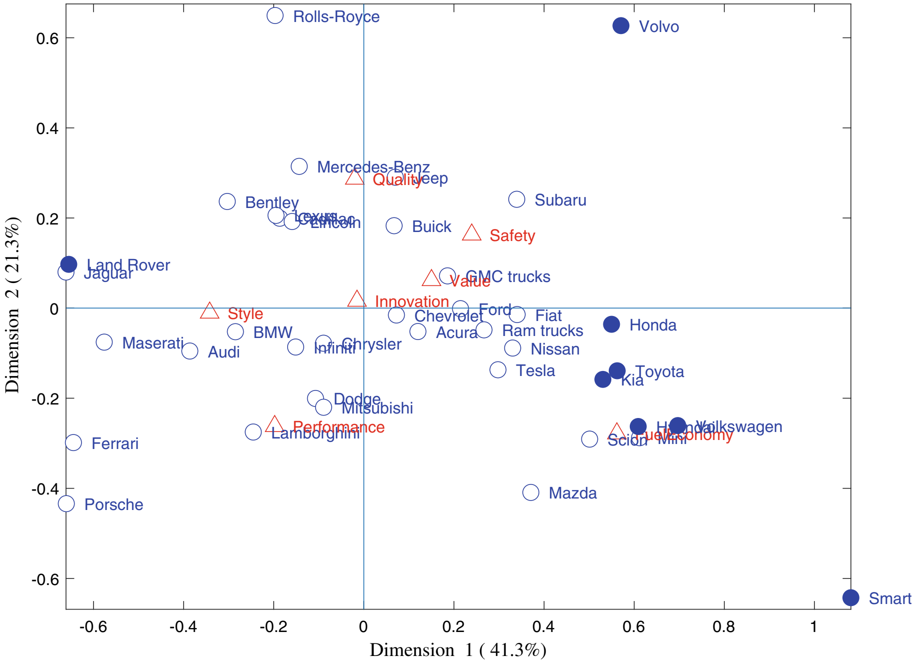

FIGURE 8 2014 car data. Correspondence analysis plot from MCD analysis, breakdown point = 0.25. Filled circles are the eight deleted observations ('supplementary points'), that is with Land Rover also treated as a supplementary point.

7 | DISCUSSION

We have applied robust methods using hard trimming to the analysis of two-way contingency tables, leading to highly informative robust displays of correspondence analyses with deleted rows. Choi and Huh (1999) described a method for correspondence analysis that uses M-estimation to downweight the effect of outlying rows. Because the rows are only downweighted, Choi and Huh cannot use deleted rows as supplementary points, a device that we have found highly informative in exhibiting the structure of the data. In particular, the analysis of the car data showed the clarity that can be found through the deletion of rows guided by the outliers detected by hard trimming.

In this paper we have exhibited the use of the MCD for hard trimming. In both examples we started with a breakdown point of 50%. Because such severe trimming can lead to unnecessarily inefficient estimation of parameters, we used a downweighting procedure to increase the size of the fitted subset of observations. We also, in the analysis of the car data, repeated the analysis with a breakdown point of 25% for the MCD. An extension of our work would be to move from these two values of h (four with reweighting) to monitoring how the properties of the fitted model and Mahalanobis distances change as the breakdown point decreases from 50% to zero, that is the non-robust fit. Cerioli et al. (2018) illustrate the procedure for multivariate normal data both for the MCD and for the Forward Search (Atkinson et al., 2010) which proceeds by automatically increasing the value of h. The resulting informative plots and test statistics lead to data-dependent estimation of the breakdown point that can then be used to give the most efficient parameter estimates for the specific data set. The algorithm of Section 4.2 allows monitoring through the fitting of subsets of increasing size, including a partial row. However, in the final analysis rows are either fully present or deleted.

Finally, we stress that we have here only treated 'simple' correspondence analysis, that is the analysis of two-way contingency tables. The extension to 'multiple' correspondence analysis, that is to the analysis of higher-way tables, is in chapter 13 of Greenacre (2017). The robust method used here extends straightforwardly to this form of correspondence analysis.

ACKNOWLEDGEMENTS

This research initiated in the context of a longstanding activity of the European Commission's Joint Research Centre in support of the anti-fraud policy of the Union, promoted and financed by the HERCULE III and the 2014–2020 JRC Work Programs in the Automated Monitoring Tool project series. The work also benefits from the High Performance Computing (HPC) facility of the University of Parma. We further acknowledge financial support from the 'Statistics for fraud detection, with applications to trade data and financial statements' project of the University of Parma.

We are grateful to the referees, whose positive comments led to improvements in the presentation of our results.

CODE

All the calculations in this paper have used the Flexible Statistics and Data Analysis (FSDA) MATLAB toolbox, which is freely downloadable from the file exchange of Mathworks at the web address https://www.mathworks.com/matlabcentral/fileexchange/72999-fsda or from github at the web address https://uniprjrc.github.io/FSDA/. More specifically, the routine to compute the MCD in correspondence analysis is called *mcdCorAna*. The associated HTML documentation can be found after installing the toolbox or directly from the web address http://rosa.unipr.it/FSDA/

mcdCorAna.html. The CA routine which plots the confidence ellipses is called *CorAnaplot* and accepts as input the structure created by routine *CorAna*.

ORCID

Marco Riani https://orcid.org/0000-0001-7886-2207
Anthony C. Atkinson https://orcid.org/0000-0001-7937-9225
Francesca Torti https://orcid.org/0000-0003-0801-1735
Aldo Corbellini https://orcid.org/0000-0002-6936-7295

REFERENCES

Agresti, A. (2013) *Categorical data analysis*, 3rd edition. New York: Wiley.

Atkinson, A.C. & Riani, M. (2000) *Robust diagnostic regression analysis*. New York: Springer-Verlag.

Atkinson, A.C., Riani, M. & Cerioli, A. (2010) The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, 39, 117–134. Available from: https://doi.org/10.1016/j.jkss.2010.02.007

Bendixen, M. (1996) A practical guide to the use of correspondence analysis in marketing research. *Research On-Line*, 1, 16–38.

Boyett, J.M. (1979) Algorithm AS 144: random R × C tables with given row and column totals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28, 329–332.

Cerasa, A. & Cerioli, A. (2017) Outlier-free merging of homogeneous groups of pre-classified observations under contamination. *Journal of Statistical Computation and Simulation*, 87(15), 2997–3020.

Cerioli, A., Riani, M., Atkinson, A.C. & Corbellini, A. (2018) The power of monitoring: how to make the most of a contaminated multivariate sample (with discussion). *Statistical Methods and Applications*, 27, 559–666. Available from: https://doi.org/10.1007/s10260-017-0409-8

Choi, Y.-S. & Huh, M.-H. (1999) Robust simple correspondence analysis. *Journal of the Korean Statistical Society*, 28, 337–346.

Greenacre, M. (2013) Contribution biplots. *Journal of Computational and Graphical Statistics*, 22(1), 107–122.

Greenacre, M. (2017) *Correspondence analysis in practice*, 3rd edition. Boca Raton, FL: Chapman and Hall/ CRC Press.

Hubert, M. & Debruyne, M. (2010) Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 36–43. Available from: https://doi.org/10.1002/wics.61

Patefield, W.M. (1981) Algorithm AS 159: an efficient method of generating random R × C tables with given row and column totals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 30, 91–97.

Riani, M., Atkinson, A.C., Torti, F. & Corbellini, A. (2022) Supplementary material for "Robust Correspondence Analysis". *Applied Statistics*, 71. (In press)

Rousseeuw, P.J. & Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

World Trade Organization. (1994) Agreement on implementation of Article VII of the General Agreement on Tariffs and Trade 1994 (Customs Valuation). *Official Journal of the European Communities*, L336, 23/12/1994, p. 119.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Riani, M., Atkinson, A.C., Torti, F. & Corbellini, A. (2022) Robust correspondence analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1–21. Available from: https://doi.org/10.1111/rssc.12580