A PRACTICAL APPLICATION OF A ROBUST MULTIVARIATE OUTLIER DETECTION METHOD

Sarah Franklin, Marie Brodeur, Statistics Canada
Sarah Franklin, Statistics Canada, BSMD, R.H.Coats Bldg, 11th floor, Ottawa, Ontario, Canada, K1A 0T6

Key Words: Covariance matrix; Mahalanobis distance; Principal component analysis; Editing; Annual Wholesale and Retail Trade Survey.

1. Introduction

Outliers in survey data are generally considered to be observations which are a long way from, or inconsistent with, the remainder of the data (Joliffe, 1986). They are often the result of response or capture errors during collection. Outlier detection in surveys is commonly used to macro edit respondent data. This relieves the burden of excessive micro editing by detecting errors in data through the analysis of aggregate data (Chinnappa and Outrata, 1989).

In the multivariate case, a classical way of identifying outliers is to calculate Mahalanobis' distance, using robust estimators of the covariance matrix and the mean vector. A popular class of robust estimators is M-estimators, first introduced by Huber (1964). To calculate M-estimators, the Reweighted Least Squares (RLS) algorithm has been widely used (Beaton and Tukey, 1974; Holland and Welsch, 1977; Hampel *et al.*, 1986).

For the past three years, the Annual Wholesale and Retail Trade Survey (AWRTS) at Statistics Canada has successfully employed a robust multivariate outlier detection method. Other than the simple two dimensional case, AWRTS is the only survey at Statistics Canada to use a formal multivariate method.

AWRTS identifies outliers using Mahalanobis' distance. The covariance matrix and mean vector are robustly estimated using an RLS estimator proposed by Patak (1990), where the RLS weights are scaled residuals from principal component analysis. The resultant estimators are orthogonally equivariant with a breakdown point of one half for large samples.

The benefits of this method are threefold: first, it is multivariate and therefore incorporates the correlation structure of the variables. Secondly, as configured for AWRTS, it is easily run by a subject matter expert since it requires only two input files, a parameter file and a data file. Thirdly, since the output indicates which variable is primarily causing the observation to be an outlier, the output is easily interpreted by subject matter experts.

This paper will begin with a general description of outlier detection, briefly describing the univariate methods most popularly used at Statistics Canada, and the multivariate method used by AWRTS (section 2). The principle behind robust M-estimators and the multivariate M-estimators of scale and location used by AWRTS are presented in sections 3 and 4. Section 5 is devoted to the application to AWRTS' outlier detection method, including practical considerations and findings.

2. Outlier detection

While most surveys collect multivariate data, few perform multivariate outlier detection: univariate methods are favoured for their simplicity. However, univariate methods cannot detect observations which violate the correlational structure of the dataset. This is the main reason for performing multivariate detection. Also, in AWRTS' case, we wish to automate a manual edit in which erroneous data are identified by cross-checking the reported values of other variables.

Most outlier detection methods use some measure of distance to evaluate how far away an observation is from the centre of the data. To measure this distance, the sample mean and variance may be used but since they are not robust to outliers, they can mask the very observations we seek to detect. To avoid this masking effect, robust scale and location estimators, which are inherently resistant to outliers, may be used. This is why many outlier detection methods use order statistics, such as the median or quartile.

2.1 Univariate outlier detection methods

Perhaps the most popular univariate outlier detection technique for survey data is the quartile method. This method creates an allowable range for the data using lower and upper quartiles: data falling outside of the range are outliers. The method is not only robust, but simple and non-parametric. Hidiroglou and Berthelot (1986) proposed an adaptation of the quartile method for trend data where the trends are first transformed to dampen a size masking effect.

These two quartile techniques are the ones most commonly used at Statistics Canada. Many surveys also use less formal ones, for example graphical methods. Lee *et al.* (1992) provide a comprehensive review of outlier detection methods employed at Statistics Canada.

2.2 Multivariate outlier detection methods

Consider a multivariate p-dimensional data set with n observations, where the ith observation is $x_i^T=(x_{i1}, x_{i2},...,x_{ip})$. If $x_1,...,x_n$ is a random sample from a multivariate normal distribution with mean vector u and covariance matrix V, a classical way of detecting outliers is to calculate Mahalanobis' distance for each observation as follows:

$$D(x_i)=(x_i-u)^T V^{-1}(x_i-u)$$

Mahalanobis' distance identifies observations which lie far away from the centre of the data cloud, giving less weight to variables with large variances or to groups of highly correlated variables (Joliffe, 1986). This distance is often preferred to the Euclidean distance which ignores the covariance structure and thus treats all variables equally.

A test statistic for $D(x_i)$ can be created as follows

$$F_i=\frac{(n-p)n}{(n^2-1)p}D(x_i)$$

which has an F distribution with p and n-p degrees of freedom (Afifi and Azin, 1972).

Other currently popular multivariate outlier detection methods fall under projection pursuit techniques, originally proposed by Kruskal (1969). Projection pursuit searches for 'interesting' linear projections of multivariate data sets, where a projection is deemed interesting if it minimizes or maximizes a projection index (typically a variance).

Huber (1985) cites two main reasons why principal components are interesting projections: first, in the case of clustered data, the leading principal axes pick projections with good separations; secondly, the leading principal components collect the systematic structure of the data. Thus, the first principal component reflects the first major linear trend, the second principal component, the second major linear trend, etc. So, if an observation is located far away from any of the major linear trends it might be considered an outlier.

3. Robust estimation: M-estimators

The purpose of robust estimation is to produce an efficient estimator in the presence of outliers, while minimizing bias. This is done by reducing the influence of the outliers on the estimator.

To evaluate robust estimators, the usual properties such as bias and precision are of interest, as well as others: how contamination influences the estimator (the influence curve or function), how much contamination the estimator can tolerate before it breaks down (the breakdown point) and if the estimator is affected by location or scale transformations (equivariance concepts). If an estimator is unaffected by translations it is called translation or location equivariant. An estimator which is scale and location equivariant for orthogonal transformations is called orthogonally equivariant. Rousseeuw and Leroy (1987), amongst others, provide formal definitions of these concepts.

Some of the most popular robust estimators are M-estimators (the M stands for maximum likelihood) first introduced by Huber (1964).

In the univariate case, a robust M-estimator could be created as follows: for the observation, x_i, location estimate, T, and scale estimate, S, define the residual, $r_i = (x_i-T)/S$. Next, define a function, $\rho(x,T,S)=\rho[(x-T)/S]$. Typically, the role of this function is to decrease the influence of observations with large residuals. Then perform minimization.

For example, given the location estimate, T, a univariate M-estimate of scale, S, could be obtained by solving the equation:

$$Minimize_S \sum_{i=1}^{n} \rho\left(\frac{x_i-T}{S}\right)$$

Often, $\rho(x,T,S)$ trims large residuals, resulting in a Winsorized estimator. Different $\rho(x,T,S)$ yield different M-estimators, including the usual maximum likelihood estimators.

3.1 Reweighted least squares

In traditional least squares estimators, the squared residuals are not trimmed before minimization. Thus, for residuals as defined above, $\rho(x,T,S)=(x_i-T)^2/S^2$. Clearly,

this function it is not robust: one large residual can have unbounded influence on the estimator. One way to robustify least squares is to bound the influence of a single observation through an appropriate weight, w_i, and then perform reweighted least squares (RLS).

In the multivariate case, for data from a multivariate normal distribution, the RLS estimators of the location vector, $\boldsymbol{u}$, and the scale covariance matrix, $\boldsymbol{V}$, given weight, w_i, can be expressed explicitly as follows:

$$\hat{V}=\sum_{i=1}^{n}(x_i-\hat{u})(x_i-\hat{u})^T w_i^2 \;/\; \sum_{i=1}^{n} w_i^2 \quad (1)$$
$$\hat{u}=\sum_{i}^{n} w_i x_i \;/\; \sum_{i}^{n} w_i$$

This is the estimator employed by AWRTS. The quality of the estimator hinges on the definition of the weights, as detailed in the following section.

4. Calculating the RLS weights

The RLS weight used by AWRTS is a function of residuals from principal component analysis: observations with large residuals receive weights less than one. The idea here is to down weight observations which lie far away from any major trend in the dataset.

4.1 Centring the data

Before calculating any weights, the data are first robustly centred using an L_1-estimator (the multivariate analogue of the median). For a p-dimensional data set with observations x_i, the L_1 estimate of the location, $\boldsymbol{T}$, is defined as the solution to the minimization problem:

$$\min_T \sum_{i=1}^{n} \|x_i - T\|$$

Centring ensures that the final estimate of the covariance matrix is location invariant. Denote the centred data by z_i ($z_i=x_i - T$).

4.2 Calculating the principal components

Robust principal components can be generated from an initial robust estimate of the covariance matrix. Each eigenvector $\alpha_1,\alpha_2,...,\alpha_p$ is a column vector of dimension p. The projection of $z_i^T=(z_{i1}, z_{i2}, ..., z_{ip})$ onto the jth principal component coordinate is $\alpha_j^T z_i$. The first principal component is the linear function, $\alpha_1^T z$, of $z_1,...,z_p$ which has maximum variance. The second principal component is a linear function, $\alpha_2^T z$, uncorrelated to the first, with maximum variance, etc.

4.3 Calculating the weights

Patak (1990) proposed the following RLS weights, w_i for each observation:

$$w_i=\prod_{j=1}^{p} \tilde{r}_{ij} \;/\; r_{ij} \quad (2)$$

where the denominator, r_{ij}, is the residual for the ith observation and the jth principal component and the numerator is a trimmed residual. All weights range from zero to one.

For the jth eigenvector, α_j, the residuals are calculated univariately by comparing the projection onto each principal component, $\alpha_j^T z_i$, with the median value, *med*, and scaling by the median absolute deviation, *mad* (divided by 0.674 to make it consistent with the normal distribution):

$$r_{ij} = \frac{|\alpha_j^T z_i - med(\alpha_j^T z)|}{mad(\alpha_j^T z)/0.674} \quad (3)$$

The residuals are trimmed as follows:

$$\tilde{r}_{ij} = \begin{cases} r_{ij} & if\ r_{ij} \le 1.75 \\ 1.75 & if\ 1.75 < r_{ij} \le 3.5 \\ 0 & otherwise \end{cases} \quad (4)$$

The cut-off points of 1.75 and 3.5, were chosen based on the analysis of symmetrical data sets. These cut-offs trim extreme residuals, but leave most untouched.

Patak (1990) shows that if the principal components are derived from the initial robust covariance matrix presented in section 4.4, this weight function results in an RLS estimator which is orthogonally equivariant with a breakdown point of $(n/2 - p)/n$. Thus, in large samples, up to half of the data can be outliers before the estimator breaks down.

4.4 Calculating an initial robust covariance matrix

An initial robust estimate of the covariance matrix is used to derive the robust principal components. This initial covariance matrix is calculated using the RLS equation in

(1) and weights as in (2), but this time the residuals are calculated by projecting the data onto the 'best basis' vectors, as defined by Stahel (1981) and Donoho (1982).

The best basis is the one which detects the most outliers when the data are projected onto each basis vector. It is found by randomly selecting basis vectors, and choosing the basis with the smallest weights.

Specifically, using the centred data, find the best basis by repeatedly projecting the p-dimensional data set onto a randomly generated p-dimensional orthogonal subspace.

The projection of z_i onto the jth basis vector, v_j, is $v_j^T z_i$. For each observation, z_i, look for the one-dimensional projection which yields the largest residuals ς_{ij}, defined by:

$$\varsigma_{ij} = \underset{\|v_j\|=1}{supremum} \frac{|v_j^T z_i - med(v_j^T z)|}{mad(v_j^T z)}$$

Note that these residuals are the same as those defined in (3), except here the eigenvector has been replaced with a basis vector.

Once the best basis has been found, the residuals are trimmed using equation (4), and the initial weights are calculated as in (2). The initial covariance matrix is then calculated using the RLS equation (1).

In theory, to identify the best orthogonal subspace, we would have to examine all possible subspaces, which is impractical. So instead, for AWRTS, we randomly sample only a subset. Patak found the minimum number of subspaces to be sampled to be ten per dimension.

4.5 Calculating the final covariance matrix

Use the robust principal components determined from the initial covariance matrix to calculate the final RLS weights (equations (2) to (4)). Calculate the final covariance matrix and mean vector using equation (1). Insert these location and scale estimates into Mahalanobis' distance and label as outliers observations with 'large' distances. This list of outliers is sent to subject matter experts for their review.

5. Application to the AWRTS

The purpose of the AWRTS is to collect principal statistics on Canadian Wholesalers and Retailers. The sample design is a stratified simple random sample of retail and wholesale companies. Preliminary validity edits are performed during collection, but due to the large sample size (26,943 companies), micro editing of all questionnaires is not possible, nor is it desirable.

This multivariate outlier detection method was first implemented in 1993. Performed directly after collection, it serves two purposes: to provide subject matter analysts with a list of outliers for editing and to flag outliers to prevent them from being used by imputation.

5.1 ODR input and output files

The outlier detection routine (ODR) is programmed in C. The user must provide two input files: a parameter file and a data file.

5.1.1 Input files

The parameter file contains the following information: the number of survey variables used to detect outliers, cut-offs for the weight function used by the M-estimator, two parameters limiting the number of outliers to be output per domain, and a variable indicating whether or not the data are to be transformed. The user is provided with default parameters.

The data file contains the following variables: the survey variables used to detect outliers, a unique record identifier and a domain variable.

5.1.2 Output file

The ODR output is a simple listing of outliers with the following information: unique identifier, domain, Mahalanobis' distance and which variable contributes the most to Mahalanobis' distance.

5.2 Practical considerations

Since parametric assumptions are required to create a test statistic for Mahalanobis' distance and because the cut-offs for the weights are based on symmetrical data, the original AWRTS data are first transformed to ensure symmetry by domain. The actual variables used are ratios and, where these are not symmetric, logarithms of ratios.

Referring to them by their numerator (the denominator is total operating revenues), the five ratios are: opening inventories, closing inventories, cost of goods, employee benefits and wages, and total expenditures. For the first two ratios, we use logarithmic transformations. These

variables were chosen since they are the ones used by subject matter analysts to manually micro edit the data.

Outliers are calculated separately by domain. These domains are imputation domains, defined by four variables: survey type (wholesale or retail), chain flag (indicating if the company is a retail chain, or not), company size (large, medium or small) and trade group (an aggregation of Standard Industrial Classifications, SIC). It should be noted here that these domains are very similar to the sampling strata, and consequently companies within the same domain have similar, if not identical, design weights.

Along with symmetry, another practical consideration is high item non-response: if several of the ratios are zero or missing, the results will be spurious. Consequently, the ODR is run three times. The first run includes all companies with reported data for all five ratios. The second run drops the ratio with the highest non-response rate (opening inventories) and uses only those with non-zero reported values for the remaining four ratios. For the third and last run, the two ratios with the highest non-response rates are dropped.

5.3 Results

An observation is flagged as an outlier if F_i (defined in section 2.2) is greater than the corresponding F value for the $99.9th$ percentile. Since there are hundreds of domains, we further restrict that there be no more than ten outliers per domain. To identify influential observations, the list of outliers is prioritized with respect to Mahalanobis' distance and the sampling weight.

For 1995 AWRTS data, 15,193 companies (56% of the sample) reported non-zero data for all five ratios. Of these, 650 (4%) were flagged as outliers.

For most outliers, the aberrant data were the inventory ratios. This is not surprising since respondents are known to have difficulty providing the data.

Subject matter experts corrected 80% of the outliers which were confirmed to be incorrect data, primarily due to reporting errors, followed by interviewer errors. The remaining 20% of outliers were confirmed to be unusual observations, but not in error and therefore not modified.

Table 1 provides the distribution of the outliers, indicating which ratio predominantly caused a company to be flagged as an outlier.

Table 1: Outliers and primary cause for outlier

Variable contributing the most to Mahalanobis' distance	Percentage of companies (%)
Opening Inventories	43%
Closing Inventories	35%
Total Expenditures	13%
Cost of Goods	8%
Employee Benefits and Wages	1%
Total Outliers	100%

The outstanding 11 750 sampled companies (44%) which had zero or missing values and therefore could not be tested for outliers with respect to five ratios, were tested in subsequent runs of the ODR. These subsequent runs identified outliers with respect to only four or three ratios. For these runs, the ranking of the ratio primarily causing a company to be an outlier remained unchanged.

Acknowledgements

The authors would like to thank Zdenek Patak, Hyunshik Lee and Jean-Louis Tambay for their very useful comments and insight, and Shelley Colmer for her assistance.

References

1.) Afifi, A.A., and Azen, S.P. (1972), *Statistical analysis: a computer oriented approach*, Academic Press, New York.

2.) Beaton, A.E., and Tukey, J.W. (1974), The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics*, **16**, 147-185.

3.) Chinnappa, N., and Outrata, E. (1989), General survey functions design at Statistics Canada, *I.S.I. Proceedings*, 47th Session, 219-238.

4.) Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. Qualifying Paper, Dept of Statist., Harvard University.

5.) Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

6.) Hidiroglou, M. A., and Berthelot, J.-M. (1986). Statistical edit and imputation for periodic surveys. *Survey Methodology*, **12**, 73-83.

7.) Holland, P.W., and Welsch, R.E. (1977), Robust regression using iteratively reweighted least squares,

Commun. Stat (Theory and Methods), **6**, 813-828.
8.) Huber, P.J. (1964), Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73-101.
9.) Huber, P.J. (1985), Projection pursuit, *The Annals of Statistics*, **13**(2), 435-475.
10.) Joliffe, I.T. (1986) *Principal Component Analysis*, Springer-Verlag, New York.
11.) Kruskal, J.B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In *Statistical Computation*, edited by R.C. Milton and J.A. Nelder, Academic Press, New York.
12.) Lee, H., Ghangurde, P.D., Mach, L. and Yung, W. (1992) Outliers in Sample Surveys, Working Paper No. BSMD-92-008E, Methodology Branch, Business Survey Methods Division, Statistics Canada.
13.) Patak, Z. (1990), Robust principal component analysis via projection pursuit, M. Sc. thesis, University of British Columbia, Canada.
14.) Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, U.S.A.
15.) Stahel, W. A. (1981). Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen. Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich, Switzerland.