Tutorials in Quantitative Methods for Psychology
2012, Vol. 8(2), p. 108-121.
DOI: 10.20982/tqmp.08.2.p108

Detecting outliers in multivariate data while controlling false alarm rate

André Achim
Université du Québec à Montréal

Outlier identification often implies inspecting each z-transformed variable and adding a Mahalanobis D^2. Multiple outliers may mask each other by increasing variance estimates. Caroni & Prescott (1992) proposed a multivariate extension of Rosner's (1983) technique to circumvent masking, taking sample size into account to keep the false alarm risk below, say, $\alpha = .05$. Simulations studies here compare the single multivariate approach to "multiple-univariate plus multivariate" tests, each at a Bonferroni corrected α level, in terms of power at detecting outliers. Results suggest the former is better only up to about 12 variables. Macros in an Excel spreadsheet implement these techniques.

The impetus of the present work was to identify, in the context of a graduate course in multivariate statistics, sound statistical procedures to recommend for the examination of multivariate data for the detection of outliers, *assuming normal distributions*. The basic consideration is that the statistical criterion beyond which a piece of data would be considered an outlier must take into account both the number of cases (subjects) inspected as well as the number of variables examined if the variables are inspected one by one. This is required to adequately control the risk of falsely rejecting at least one case that actually belongs to the population. In particular, a fixed critical z-score, irrespective of number of variables or of sample size, can hardly be recommended. Beyond controlling for false alarm (FA) rate, an adequate outlier detection procedure should accommodate, for adequate sensitivity, the fact that a multiplicity of outliers makes their detection more difficult than detecting a single outlier, due to a masking effect. Furthermore, for practical considerations, an adequate procedure must be available even to students with no computer programming experience and should accommodate cases belonging to groups that could differ in means (assuming homogeneity of their covariance matrices).

Based on work by Wilks (1963) and by Rosner (1983), Caroni and Prescott (1992) documented a multivariate outlier detection procedure meant to control the FA rate even when some real outliers are present in the sample, i.e. controlling the risk of declaring outliers outside the subset actually present in the sample. Although this appears close to the optimal procedure sought to recommend, except for easy availability, no discussion was found of whether this is uniformly better than applying Rosner's (1983) procedure with a Bonferroni correction on each of the p variables (i.e., setting "variable-wise α" to "global α"/p in testing each variable, where p is the number of variables) when the outliers to be detected are actually outliers on a single variable. Initial exploratory simulations with various combinations of number of cases and number of *independent* variables indicated some advantage for multiple univariate tests over a single multivariate test, which would correspond to the usual recommendation to inspect the z scores on each variable besides inspecting the global Mahalanobis D^2.

Obviously, the multiple univariate approach alone would not detect pattern-only outliers, i.e. outlier cases in which all variables show individual scores within an acceptable range but their pattern does not fit the rest of the distribution. If a multiplicity of univariate test, with adequate control of FA rate, was to be generally superior to the single multivariate test for detecting univariate outliers within the sample, then a general procedure should apply both approaches, so as not to miss pattern-only outliers,

correcting appropriately for the extra multivariate test added to the *p* univariate tests. Exploratory work on this question indicated that counting the multivariate test as only half an extra test, for the purpose of applying a Bonferroni correction for the total number of tests, is generally appropriate.

Both the Rosner univariate outlier detection procedure and the Caroni and Prescott (CP) multivariate outlier detection procedure include a parameter *k* that specifies an upper limit on the actual number of outliers that could be present in the data and both were documented with *k* = 10 in the presence of up to five outliers in the data sets. These procedures do not require the exact number of outliers to be known, *k* is the maximum expected. But if more than *k* outlier cases are actually present in the sample, masking effects might prevent even some of the *k* most extreme cases from being detected, although they might be if a larger value of *k* was selected.

In these procedures, the successively most extreme values (or most extreme cases, for the multivariate test), from none to *k*-1, are iteratively excluded from the sample and the most extreme remaining value is tested against a suitable criterion that depends on the current sample size. All extreme values down to the latest one to exceed its own criterion (based on current sample size) are declared outliers, even if some earlier extreme values did not qualify by themselves as significant outliers, presumably because of masking, i.e. because the currently remaining outliers in the sample inflated the variance estimate and displaced the mean.

Empirical formula improvement.

Both these procedures were documented to reliably maintain the FA rate close to the nominal level for samples larger than about 25. Empirical exploration of each procedure indicated that, for smaller sample sizes, they do not produce inflated FA rates when a single outlier is to be detected (i.e. with *k* set to 1). This indicates that the criterion set for the largest deviation in a sample is correctly estimated, even for relatively small samples. It follows that the problem of inflated FA rate for small sample sizes but with *k* > 1 is associated with the correction for more than one extreme value removed. The risk of the first extreme value being significant, in the absence of real outliers, could be made less than the nominal rate so as to allow for a few instances where it is a later extreme that first exceeds the nominal value. Alternately, the progression of critical value could be such that it is really exceptional that a later extreme from a normal distribution without outliers is significant when the previous extremes were not. Based on this latter option, the respective formulas described by Rosner (1983) and by Caroni and Prescott (1992) were revisited through an educated trial-and-error procedure that introduced the original sample size, n_0 (i.e. sample size with zero observation removed), in the equation for the current critical value. In discussing this, we may by extension denote n_i the reduced sample size after the i most extreme cases sequentially identified have been excluded.

For ease of computation, Rosner's formula for a critical Student *t* value may be implemented as its square, yielding a critical *F* value (*crit*) for the maximum of n_i scores, which is itself based on $F = F_{\alpha/n_i,1,n_i-2}$, the critical value of the F distribution with 1 and n_i-2 degrees of freedom and a probability α/n_i, embedding the Bonferroni correction, where α is the selected global FA rate, typically .05 (when a single variable is to be examined). We then calculate the appropriate critical value as:

$$crit = F\frac{(n_i-1)^2}{n_i(n_i-2+F)}$$

Implementation is further simplified if the index calculated for the maximum deviation in the sample involves its division by the sum of squared deviations from the mean, instead of by the variance. This resulting index will be smaller by a factor of (n_i-1), and so should its critical value. For practical reasons, Rosner's (1983) original procedure may be implemented by squaring the maximum deviation from the mean, dividing by the sum of squared deviations and comparing the result to the following critical value:

$$crit = F\frac{n_i-1}{n_i(n_i-2+F)}$$

which represesents a variant of Rosner's formula applicable to squared deviation divided by sum of squared deviations.

Similarly, the CP procedure, which reduces to Rosner's approach for the specific case of a single variable (i.e., *p* = 1) may be implemented by calculating

$$C_j = (x_j - \overline{x})' A^{-1} (x_j - \overline{x})$$

where x_j is the vector of observations for subject *j* and A^{-1} is the inverse of the sum of cross products matrix. The maximum of this score is then compared to its critical value, C_{crit}, which is based on the critical F value with *p* and *n*-*p*-1 degrees of freedom and which is calculated as follows:

first $$G = \frac{p}{n_i - p - 1} F_{\alpha/n_i,p,n_i-p-1}$$

and then $$C_{crit} = \frac{G}{G+1} \times \frac{n_i - 1}{n_i}$$

Our empirical exploration of this formula to remedy the

inflated FA rate in relatively small samples and with k = 10, led to changing (n_i-1) in the numerator above into (n_0-1), which corrects the problem for small samples while affecting larger samples only minimally. Thus, the general multivariate formula computes G as above but follows with

$$C_{crit} = \frac{G}{G+1} \times \frac{n_0 - 1}{n_i}$$

which, for the univariate case, reduces to

$$crit = F\frac{n_0 - 1}{n_i(n_i - 2 + F)}.$$

Before proceeding with the main purpose of the present work, it was appropriate to document, through Monte Carlo simulations, the behavior of the modified formulas compared to the original ones as well as the appropriateness of a Bonferroni correction for the number of variables if the univariate outlier detection procedure is to be applied sequentially to each variable in a multivariate set and a case excluded if any of its p measurements exceeds the criterion for outlier declaration.

A first simulation study bearing on the FA rate when no outlier is actually present will be followed by the comparison of two candidate methods in terms of power at detecting true outliers and in terms of their FA rates for the remaining non outlier cases in the presence of true outliers. This latter section will include various levels of correlations among the variables, which will also, aside from the main purpose, document the effect of correlations among the variables on the FA rates.

Study 1: Confirmation that the modified formula keeps the FA rate within the nominal 5% value.

Methods

All simulations were carried in MATLAB 7.10 (R2010a) or 7.12 (R2011a) using the default pseudo random number generation algorithm, the Mersenne Twister (Matsumoto & Nishimura, 1998). All simulation studies looked for a maximum of k = 10 outlier cases in the sample, with global α set to .05. Varied numbers of variables (10 levels of p: 2:1:6, 8:2:12, 15, 20, 30) and varied cases per variable ratios (6 levels: 2, 3, 5, 9, 15 and 25) were used, to span a wide range of experimental situations. Only combinations yielding at least 15 cases and with at least 10 cases more that the number of variables were used (otherwise, removing 9 potential outliers results in a singular sum of cross products matrix). For each of the 54 valid combinations of these parameters, 10 000 simulated data sets were generated, where each variable was drawn from an N(0,1) distribution (i.e. no real outlier added). For each distribution, five outlier detection methods were applied, (1) the standard and (2) the modified Rosner procedures, both with a Bonferroni correction of the nominal α of each univariate test (i.e. dividing .05 by the number p of variables), (3) the standard and (4) the modified CP procedures (single test at α = .05) and (5) a combo procedure, applying the modified univariate test on each individual variable in addition to the modified multivariate test, with each of these tests performed at α = .05/(p+½). The latter correction followed our preliminary explorations indicating that the multivariate test, in parallel to the p univariate tests may be counted as only half an extra test for the purpose of correcting for the total number of tests performed on each subject. For each simulation condition, the number of simulated studies yielding at least one FA was tallied for each method separately.

In addition to the above, the original and modified formulas were applied to 100 000 simulations with a single variable and n = 15:5:40. The added number of simulated studies, here, aimed at a narrower estimate of the actual FA rate for eventual univariate applications of the modified procedure.

Assuming that a method actually yields its nominal FA rate, the 99% confidence interval for FA rate out of 10 000 simulated studies includes from 4.44% to 5.56% FAs. With 100 000 simulated studies, the 99% confidence interval goes from 4.82% to 5.18%. Conditions that yielded more FAs than the upper limit are of particular interest here, but there is also interest in noting whether the corrections described above make the tests conservative on relatively large samples.

Results

The *original Rosner procedure* with a Bonferroni correction for the total number of variables exceeded 5.56% FAs in all conditions with n ≤ 20. In decreasing order, these were 15.71% (n = 15, p = 3), 9.03% (n = 18, p = 2), 7.97% (n = 18, p = 6), and 6.44% (n = 20, p = 4). The limit was also slightly exceeded for n = 25, p = 5 (5.60%) and for n = 30, p = 2 (5.62%). In the simulations with only one dependent variable and 10 times as many simulated studies per condition, the observed FA rates were above the upper limit of 5.18% even for the larger sample size tested. The observed rates were 17.1% (n = 15), 7.99% (n = 20), 6.26% (n = 25), 5.69% (n = 30), 5.46% (n = 35) and 5.42% (n = 40).

With more than one variable, the *modified Rosner procedure* yielded all FA rates actually between 4.54% and 5.44%, i.e., all well within the 99% confidence interval. With a single variable, it fared better than the original version but nevertheless exceeded the upper limit of the 99% confidence interval, with observed rates of 5.67% (n = 15), 5.60% (n = 20),

5.48% (n = 25), 5.38% (n = 30), 5.27% (n = 35) and 5.30% (n = 40).

For the *original CP procedure*, 13 of the 54 conditions exceeded the confidence interval upper limit of 5.56%. Seven of these conditions had sample sizes at least 25. The maximum of the latter was 9.22% obtained for n = 40, p = 20.

The *modified CP procedure* produced FA rates between 4.58% and 5.47% except for 4.36% with n = 200, p = 8, and for 5.59% with n = 15, p = 3, which both are just outside the 99% confidence interval. Running new sets of simulations in these two conditions gave respective FA rates of 4.86% and 5.71% (but 15.99% for the original CP procedure), suggesting that the initial result for n = 15, p = 3 reflects a real FA excess, although a slight one, while the initial result for n = 200, p = 8 was a statistical accident.

Finally, the *combo procedure* produced all FA rates between 4.64% and 5.45%. Although this appears completely acceptable, the distribution of counts below and above the expected count of 500, respectively at 15 and 38, is clearly asymmetrical (χ^2 (1) = 9.98, *prob.* = .0016).

Discussion

The first conclusion form these simulations is that the modified version of both the Rosner and CP procedures improves over the original version and is highly satisfactory. The correction does not even make the tests conservative with large sample sizes. It actually appear totally satisfactory for all tested multivariate cases and, although the modified version still has a slight tendency to exceed its nominal FA rate when applied to a single dependent variable, its observed FA rate was always observed below 5.72% (for a nominal rate of 5%) when estimated with 100 000 simulated studies.

Since the modified Rosner procedure performed at nominal level for the multivariate cases with independent variables, it may be inferred that the principle of a Bonferroni correction for number of variables tested is supported by these data. Had this been an excessive correction (for independent variables), a tendency to produce significantly less that nominal FA rate would have been observed. Not observing this may not be attributed to a mere compensation effect associated with a (slightly) inflated FA rate that would apply, with a single variable, across all levels of α. Indeed, an extra univariate run with n = 20 but α = .01 indicated that the modified procedure signals outliers within the expected interval, with an observed FA rate of 0.983%. Thus, the modified Rosner procedure appears very adequate when used with α smaller than .05, which is the case with a Bonferroni correction for the number of variables tested and which the present simulations demonstrated to work as expected.

When an outlier is declared on a variable, the question arises whether the case should be removed from the sample or not in inspecting the remaining variables. In the present simulation study this did not matter as we were only concerned with the per study FA rate and it was found that very close to the expected 95% of the simulated studies included no apparent outlier at all. In actual applications in which true outliers may be present, excluding outliers detected on earlier tested variables would reduce sample size for the remaining variables and would thus provide slightly more power at detecting new outlier cases on the remaining variables (because of the embedded Bonferroni correction for sample size). The slight gain in power would, however, come at the cost of not detecting, say, a pair of outlier scores in the same subject. If the combo procedure is adopted and it is decided a priori that any subject failing any outlier detection test would necessarily be excluded from the sample, exclusion of already identified outlier cases should be applied as the sequence of tests progresses.

It should be noted here that the independent variables used in the simulations should constitute a worse case condition for multiple tests per subject. With correlated variables, the risk for a subject of being falsely declared an outlier on variable j+1, given that he/she was within limits on the first j variables should actually be lower when this variable is correlated with the ones previously tested than when it is independent from them. This should be confirmed in study 2 that uses correlated variables.

Finally, the present simulations confirm the rule of thumb derived from preliminary explorations that adding the multivariate outlier detection test to the univariate outlier detection test on each variable may be counted as only half an extra test. The asymmetry of FA rates above and below the expected value, however, hints that this may only be a rough approximation. Examination of the distribution of high and low FA counts across the conditions with different numbers of variables provided no suggestion of a tendency of either type of counts to be associated with a low of high number of variables. In particular, the mean number of variables in the simulations for which FA number was observed below the expected 500 count was 11.67 while that for FAs above 500 was 11.82. Considering that correlated variables should lead to conservative tests when a Bonferroni correction is applied, the correction with p+½ when the multivariate test is also applied should be completely adequate.

Note that from here on, the Rosner and CP procedures should be taken to mean their modified versions. The Rosner procedure (equivalently, the CP formula used for multiple univariate tests, where p = 1) will only be used within the combo procedure, since it cannot detect pattern-

only outliers.

Study 2: Comparison of approaches to outlier detection.

Given that both methods currently considered for outlier detection in multivariate data provide good control over FAs, the question remains whether one is uniformly more powerful than the other at detecting true outliers. The CP method applies a single multivariate test to all subjects and operates at α = .05 sample-wise. The combo method, on the other hand applies p+1 tests per subject, but each at the more extreme criterion of $\alpha = .05/(p+½)$ sample-wise.

The primary purpose of study 2 was to compare the CP and the combo methods when some outliers are present, including cases of pattern-only outliers, which is a meaningful concept only with correlated variables. As an extreme case for multiple tests, however, conditions of independent variables should also be included. Aside from the level of correlation between variables, the number of variables was varied since the Bonferroni correction embedded in the combo procedure (number of variables plus one half) might affect its relative power compared to the CP procedure for detecting true outliers.

Three patterns of outliers are relevant to the present investigation. First, a case may be an outlier on a single variable. Secondly, a comparable distance of a case from the means may be widely spread over many variables, which should leave the case comparatively detectable for a multivariate procedure. Pattern-only outliers are not easily matched in size with the previous two types but may be produced by sign changes on about half of the variables. The combo procedure may be expected less powerful at detecting these because its embedded multivariate CP test is applied with a much reduced α level.

Procedure adjustment

Preliminary simulations with up to five true outliers present in the data indicated a FA problem with the CP procedure under some conditions. With four or five same polarity outliers either on a single variable or each spread among several variables, but not with pattern-only outliers, the CP procedure produced excess amounts of FA among the remaining cases, a phenomenon known as swamping (Bradu & Hawkins, 1982). For instance, with a nominal α of .05, at least one FA was observed in 16.26% of 10 000 simulated studies when the sample contained five outliers on the same variable out of 12 variables reflecting three correlated factors. A reasonable speculation about these FAs is that they would come from values in the tail of the distribution opposite to the direction of slippage. With enough outliers of the same polarity present, the shift in the estimated population mean could make one of these come out as the currently most extreme case, although not currently significant. When true outliers are later detected beyond their criterion, all previous extremes are also counted as outliers by virtue of the prescribed rule. This suggests revising the outlier exclusion rule.

The original decision rule consists in comparing the statistics calculated for each successive extreme value with its own criterion (that depends on the current sample size) and to exclude all successive extreme values up to the latest significant one. A rule that solves the excess FA problem simply adds a final test on each extreme value identified before the last significant one. Starting from the subsample in which the last significant extreme value was obtained, this extreme value is replaced in turn by each preceding extreme value and the most extreme value of this subsample is then identified. The case just reintroduced is declared an outlier only if it is the current extreme and its statistics exceeds the current critical value. Cases not so rejected as outliers are not reintroduced in the sample in this final retesting phase, such that all potential outliers are retested with the same critical value.

To formally document that the original rule produces an excess of FAs when the sample contains four or five true outliers and to confirm the appropriateness of the modified rejection rule, a set of 2 000 simulated studies was run, each with 10 variables and sample size 100. The variables depended on three independent factors expressed respectively in four, three and three variables with randomly selected weight between .6 and 1.0 and with noise adjusted to give each variable unit variance in the population. For each of 2 000 simulated studies, zero to five outliers of three types were produced in each data set. Outliers were created by adding 5 to one of the first four variables or 7.4 to the factor score that is expressed in these first four variables. Pattern-only outliers were created by inverting the sign of the weights for half the variables depending on factor 1.

Main simulations.

After documenting the modified rejection rule, nine sets of simulation were run in a 3 x 3 design with 6, 15 or 30 variables that were either independent, relatively weakly correlated or relatively strongly correlated in the population. Only the modified decision rule was applied for these conditions.

Sample distributions. Stimulations for independent variables simply involved generating 100 random numbers from a N(0,1) distribution for each variable. In the remaining six sets of simulations, random correlations were produced by modifying the pair-wise orthogonality of initially independent variables, with a probability of 0.7 of reducing

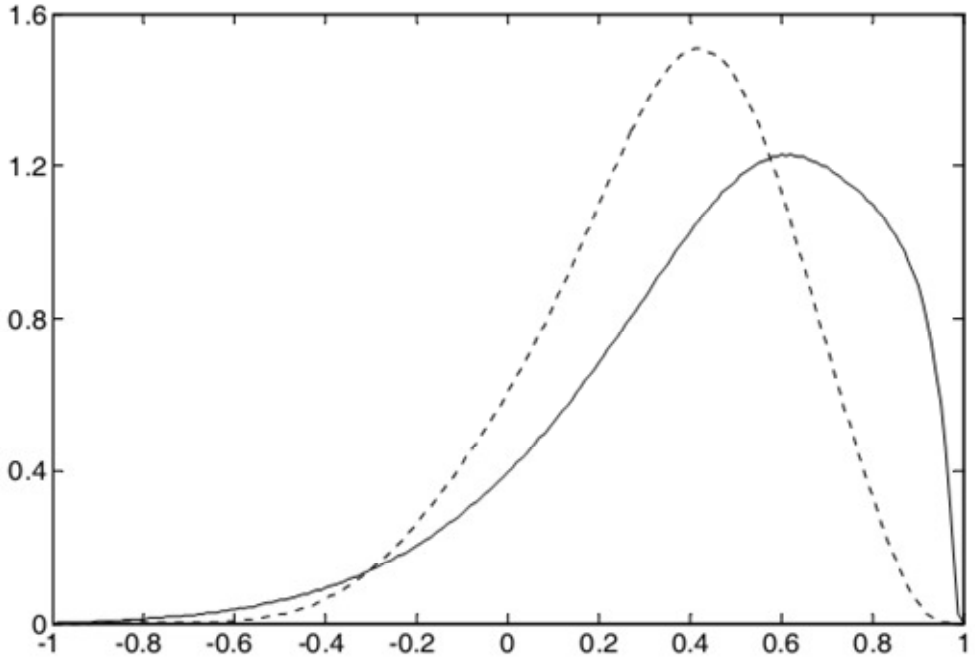

Figure 1. Distributions of correlation coefficients obtained from a 30 x 30 matrix in which the initially orthogonal angle between each pair of axes was modified by moving each axis toward (p = 0.7) or away from (p = .3) the other by a common random angle between 4 and 12 degrees (continuous line) or between 3 and 9 degrees (dotted line).

the 90° angle between the variables to make them positively correlated (and a complementary probability of increasing the angle for a negative correlation). Angular change between variables was uniformly distributed from 8° to 24° (for relatively strong correlations) or from 6° to 18° (for more moderate correlations), where each axis of a pair effected half the change. The two empirically derived distributions of expected pair-wise correlations are depicted in Figure 1. To insure the same expected distribution of correlations irrespective of the number of variables used, all axes changes were effected on a 30 x 30 matrix. For fewer than 30 variables, random subsets of the 30 randomly correlated variables were chosen to represent the population correlation matrix for a given simulated study. This population matrix was then subjected to singular value decomposition to produce a transformation matrix to be applied to independent N(0,1) normally distributed variables in order to produce correlated variables with expected unit variance.

For each simulated data set, observed values for a sample of 100 cases were first generated without any outlier and the two procedures, CP and combo in their modified rejection versions, were applied. Then one to five outliers of a given type were sequentially produced by suitably modifying the scores of the first one to five cases, this being repeated for each type of outliers starting from the same original data set. In a given simulated data set, outlier slippage on a single variable consisted in adding 5.0 to one of the variables. The same variable was used for all the single variable outliers in a given data set. The outliers whose slippage was distributed on many variables were actually outliers with the same total slippage evenly spread on the first five underlying independent variables (i.e. before multiplication by the transformation matrix), thus producing an equivalent effect from a multivariate point of view. Finally for pattern-only outliers, the sign of each odd numbered variable was inverted. There were no pattern-only outliers with independent variables.

In the combo procedure involving a sequence of outlier detection tests, i.e., p univariate tests followed by the multivariate test, cases flagged as outliers on any test were excluded from the later tests to optimize power.

With 2 000 simulated studies, the 99% confidence interval for an expected FA rate of 5% ranges from 3.75% to 6.25%. Simulated samples with at least one FA, before the introduction of outliers, were tallied to estimate the respective FA rates of the procedure with correlated data. Besides, their pair-wise divergent outcomes were tallied according to which method of the pair produced at least one FA. When true outliers were added, samples with at least

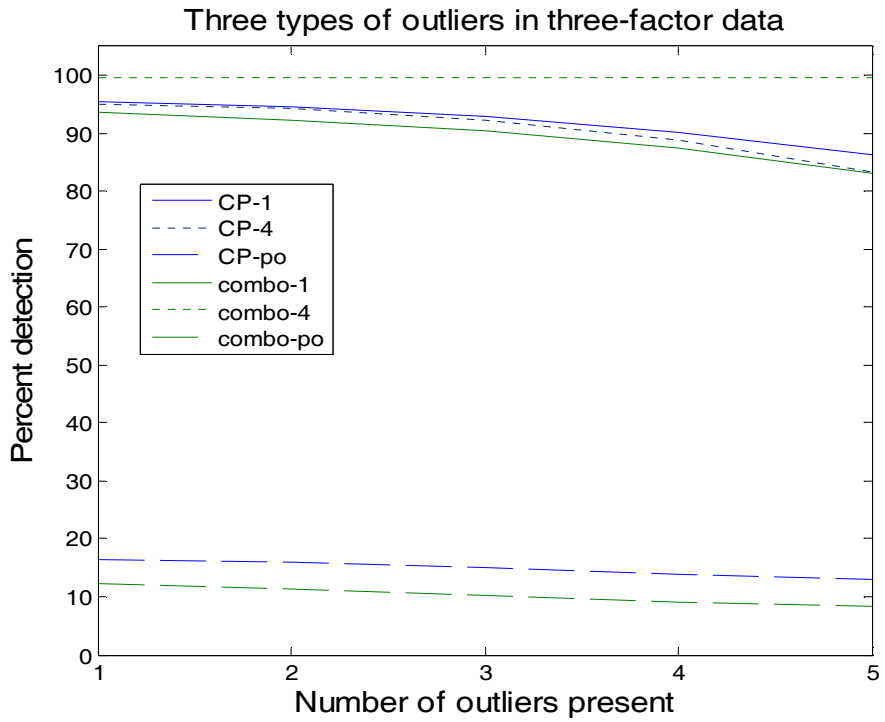

Figure 2. Detection rates in data sets with N=100 and ten variables depending on three factors, for the CP (blue) and combo (green) methods. Suffixes -1 and -4 indicate outliers on a single variables (one of those governed by the first factor) and on all four variables of the first factor, being outliers on the underlying factor score. Suffix -po indicates pattern-only outliers.

one FA in their non-outlier portion were also tallied, in order to verify the behavior of each procedure when true outliers are present (although some could be missed). All FA tallies were thus done experiment-wise (i.e. simulated samples with at least one FA were counted).

For true outlier detection, absolute counts and pair-wise divergent outcomes were tallied separately for each outlier in the sample, rather than experiment-wise, since the percentage of true outliers detected is here of interest. The divergent outcome tallies are used to test differences in sensitivity between the methods through a χ^2 test of difference of proportions for paired data. Note that the outcomes of these tests will only be reported as p values, where the assigned fractional values will prevent any confusion with number of variables p. In these various tallies, the same simulated sample could give rise both to detection of some true outliers and to FAs in their non-outlier portion.

Results

The preliminary simulation set with ten variables from three factors and which used both the original and the modified rejection rules confirmed the need for a revised rejection rule. With five outliers on the same variable present among the 100 cases, the CP procedure with the original rule gave 9.95% of the simulated studies with at least one FA, compared to 2.75% for the revised rule. For four or five outliers on the factor expressed in the first four variables, the FA rate was 7.4% and 13.6% respectively for the original CP rejection rule, but 3.1% and 2.6% with the modified rule. Without any outlier, the FA rate of the CP procedure was 5.35% (not affected by exclusion rule). The combo procedure expressed a similar tendency only with five outliers on a single variable, with a FA rate of 5.9%, which was reduced to 4.45% with the revised rejection rule. When no true outlier was present, the FA rate by the combo procedure was 5.05%.

The detection rates of this preliminary set of simulations are depicted in Figure 2. The outliers on the factor score (dotted lines) were detected almost perfectly by the combo procedure (sic) and well detected by the CP procedure. For outliers on a single variable, the order is reversed although both procedure detected a large proportion (83% or more) of the outliers present. This order also prevails for pattern-only outliers, although the detection rates are relatively low, between 16.5 and 8.45%. All the difference are statistically significant with $p < .001$.

Main simulation sets.

False alarm rates. The anticipation that the actual FA rate in the multiple test combo procedure would be lower than the nominal rate when the variables are correlated was not

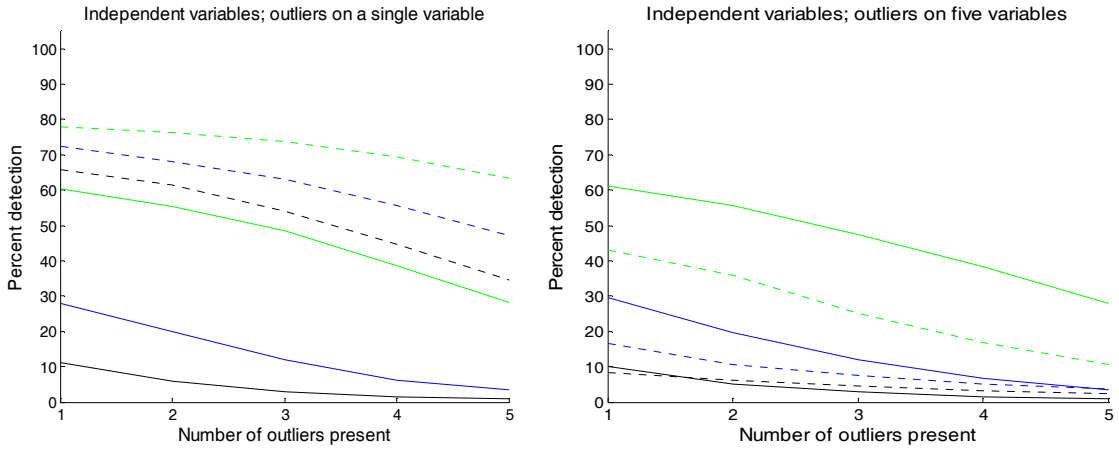

Figure 3. Outlier detection rates for 6 (green), 15 (blue) or 30 (black) independent variables by the CP (solid line) and combo (dotted line) procedures. Left graph: outliers on a single variable. Right graph: outliers on five variables.

supported in the preliminary simulation, with its observed 5.05% FA rate. This effect was observed in the main set of simulations, but only in the 30 strongly correlated variable condition. When no true outlier was present, this was the only condition with a FA rate outside the 99% confidence interval and it was not way below the lower limit of 3.75%. The observed combo procedure FA rate of 3.4% is also significantly less that that of 5.3% for the CP procedure on the same data (p = .0028). The neighbouring conditions of 15 strongly correlated variables gave 4.25% FAs and that of 30 moderately correlated variables gave 4.2% FAs, only expressing the anticipated effect as a mild trend.

With at least one outlier present, of whatever type, the FA rate among non outlier cases was generally below the 5% nominal rate, often below the 99% confidence interval. This was especially so for the CP procedure with outliers on a single variable and for the combo procedure for outliers on a subset of five underlying variables. FA rates below 2% were observed only seven times, all in the strongly correlated variable condition. Only one such case was observed with 15 variables, with 1.75% FA for CP with three outliers present. With 30 variables, 1.35% was observed for CP with four and five outliers on the same variable, 1.4% and 1.05% for CP with four and five pattern-only outliers and 1.95% and 1.7% for combo with four and five outliers on five underlying variables. If anything, thus, outliers make the tests conservative for the remaining non outlier cases.

Outlier detection

Only the revised exclusion rule is considered for comparing the CP and Combo procedures in the main set of simulations and true outlier detection is reported as proportion of detected outliers among true outliers present rather than as proportion of studies with some or all outliers detected. These detection rates are presented in Figures 3, 4 and 5 for respectively independent variables, moderately correlated variables and strongly correlated variables, each for the three types of outliers (only two for independent variables). Each sub-figure depicts the single-test CP procedure as a single (continuous) line and the multiple-test combo procedure as a dotted line. Simulations with 6, 15 and 30 variables are painted in increasing color darkness, namely green, blue and black.

For completely independent variables, an unlikely situation in multivariate analyses, the results are as could be anticipated, namely that for outliers on a unique variable the single multivariate test of the CP procedure is much less efficient than the independent tests of the combo procedure. Furthermore, both tests loose power as the affected variable is diluted among more variables. For outliers on five variables, the single test CP procedure has more power, but its advantage decreases as the five variables become a smaller portion of the total set of variables, such that, with 30 variables, the combo procedure takes the advantage when at least two outliers are present. All differences are highly significant (p < .0001), except for five outliers on five underlying variables out of 15 (p = .39) and for one or two outliers on five out of 30 variables, where CP has a slight advantage for a single outlier present (p = .0328) and the reverse holds for two such outliers (p = .016).

For moderately correlated variables and *outliers on a single variable*, the CP procedure generally outperforms the combo procedure (p < .0001, but only p = .0079 and .0013 for one and two outliers respectively in the six variable condition), with the exception of the 30 variable case where the combo procedure outperforms CP in the presence of five outliers only (p < .0001). In this condition but with fewer outliers, the difference in favor of CP is significant only at

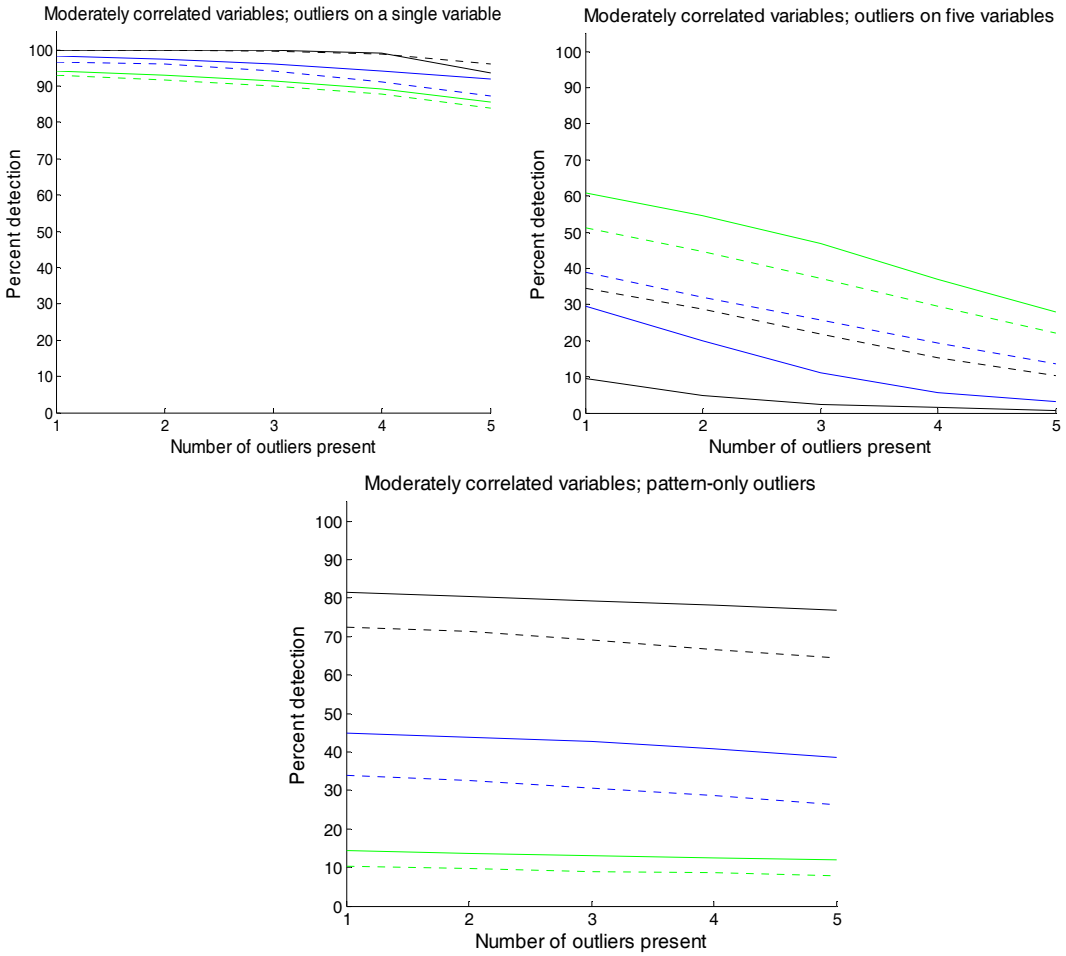

Figure 4. Outlier detection rates for 6 (green), 15 (blue) or 30 (black) moderately correlated variables by the CP (solid line) and combo (dotted line) procedures. Upper left graph: outliers on a single variable. Upper right graph: outliers on five variables. Lower graph: pattern-only outliers.

two (p = .0082) and three outliers (p < .0001). For *outliers on five (underlying) variables*, the CP procedure is best in the six-variable condition but the combo dominates with 15 and 30 variables (p < .0001). The more diluted are the five involved variables among all variables, the lower the detection rates. Finally, for pattern-only outliers, CP dominates (p < .0001) and detection increases as number of variables increases, as half the variables are inversed in sign to create these outliers.

For strongly correlated variables, the differences are in favor of CP with all three types of outliers in the six-variable condition, but for outliers on a single variable, the difference is significant only with four and five outliers present (each p = .0001). With 15 variables, CP dominates for outliers on one variable (p = .0023 for one outlier present, p < .0001 thereafter) and for pattern-only outliers (p < .0001), but combo dominates for outliers on one third of the variables (p < .0001). For 30 variables, detection was perfect up to three outliers present on the same variable and favored combo thereafter (p < .0001). Combo outperformed CP (p < .0001) for outliers on one sixth of the underlying variables. The reverse holds for pattern-only outliers (p < .0001, except p = .0082 for a single outlier present).

Discussion

Although this second study aimed at documenting which approach is more sensitive to detect outliers under various conditions, an excess of FAs in the presence of true outliers (swamping) had to be controlled first. The solution adopted, namely a revised rejection rule for extreme cases before the last significant one, proved quite satisfactory. It must be said, however, that the conditions under which the corrected rule matters are elusive. Actually, the swamping problem was not seen in any main simulation condition. Documenting that the situation can arise therefore required a different example, similar to the more complex one that manifested the phenomenon in earlier explorations.

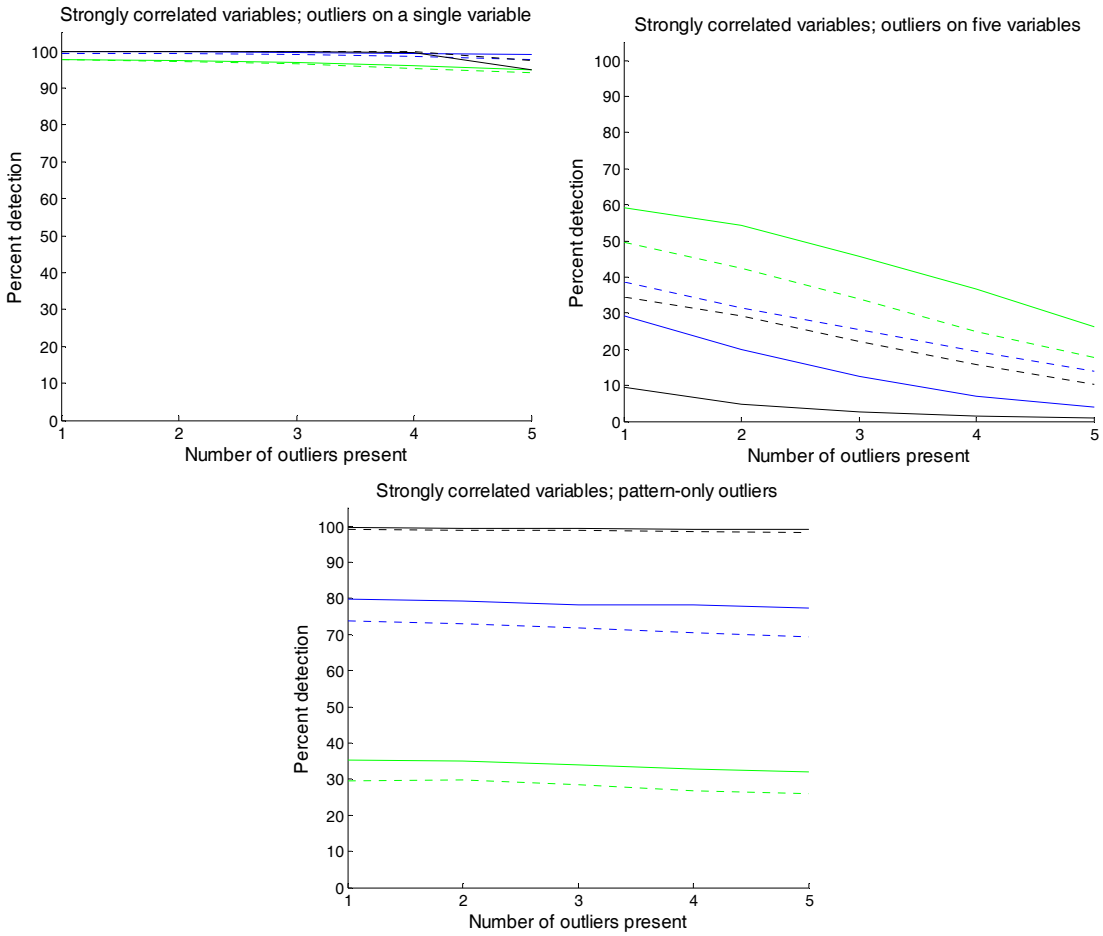

Figure 5. Outlier detection rates for 6 (green), 15 (blue) or 30 (black) strongly correlated variables by the CP (solid line) and combo (dotted line) procedures. Upper left graph: outliers on a single variable. Upper right graph: outliers on five underlying variables. Lower graph: pattern-only outliers.

The simulations with the current preliminary condition indicated that CP, based on a multivariate test embracing all variables at once, was better for outliers on a single variable while the combo procedure, with its multiple univariate tests, was better for outliers expressed on four variables. To understand this apparent mismatch of test with outlier type, we must remember that the latter type of outliers were actually outliers on the underlying factor score. They thus conformed to the general pattern of correlations among the variables but with more extreme scores. Outliers on a single variable, on their part, did not conform as well to the pattern of correlations between the four variables expressing the factor, which presumably helped the CP procedure to detect them. Although generalizing from this particular data structure would be hazardous, the results at least indicate that no one technique is universally better than the other.

For the main sets of simulations, the winner between the CP and combo procedures also depends on conditions. Even without claiming that the present conditions of simulation could be considered representative of most real data situations, it appears that the CP procedure could be preferred up to 10, perhaps 12, variables, more or less irrespective of the type of outliers to be detected. If however the data would only admit pattern-only outliers, as for data from Likert scales with reasonable spread on each item, obviously the CP method would also be preferred irrespective of the number of items. Otherwise, above twelve variables, the combo procedure could be preferable.

Practical considerations

As mentioned in the introduction, a convenient outlier detection method should preferably also be applicable to group data where the group means may differ. Simply applying the CP or combo procedure to each group separately does not need the assumption of homogeneity of covariance matrices but provides much less power, because of the fewer degrees of freedom available within a single group. Besides, for separate inspection of each group, each group size must exceed the number of variables plus k and some adjustment of the nominal alpha level for each group,

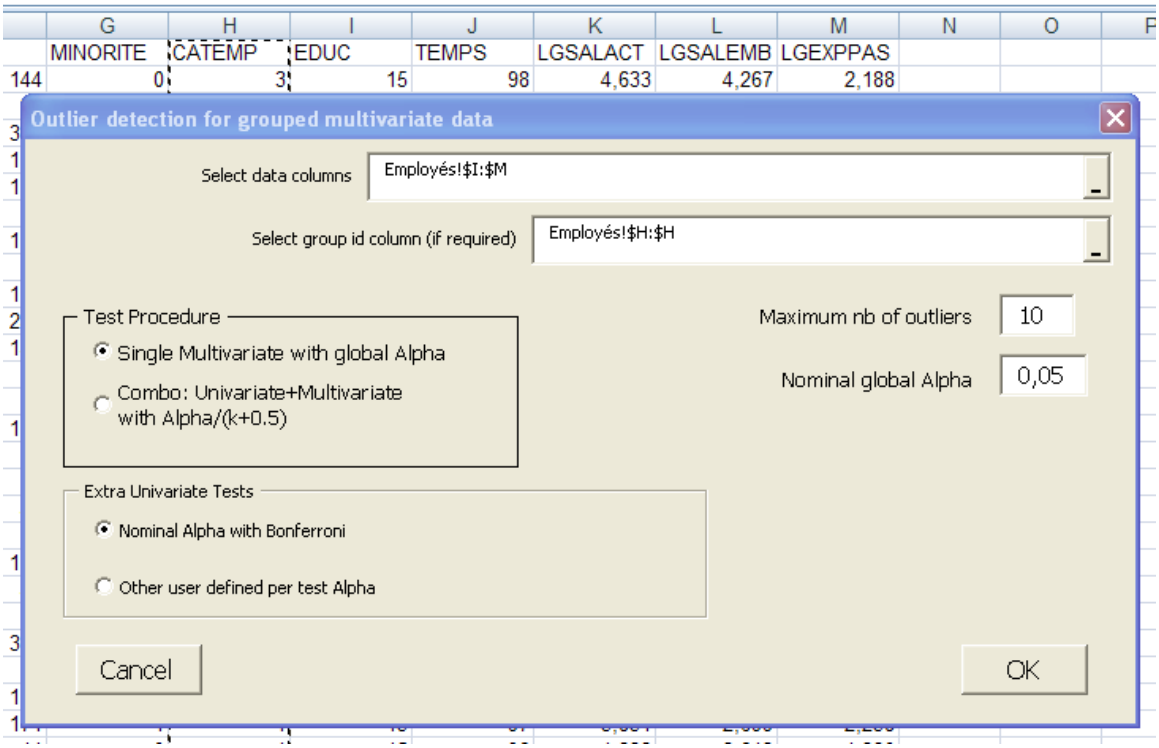

Figure 6. Illustration of the opening window of OutlierDetection.xls. Data in the background are from the SPSS example "Employee Data", after some variables have been suitable log transformed and the dependent variables to be assessed have been regrouped into consecutive column. CATEMP is employee category; its dotted contour indicates that it has just been selected as the group ID column.

preferably based on group size, is required in order to maintain to 5% the overall risk of falsely rejecting a case that is not an outlier.

Caroni (1998) investigated the effect of various levels of heterogeneity of covariance due to several variables (diffuse) or a single variable (concentrated) and concluded that "the size of Wilks' [test for a single outlier in multivariate normal samples from different subpopulations] is acceptably robust to moderate heterogeneity in covariances (25-50% difference in total variation), especially if sample sizes are small (below 20 per group)". She concluded, with reference to the CP procedure, that "an exactly similar procedure should be applicable in the multiple-group case, with potential outliers being ordered by Mahalanobis distance from their group mean". This suggestion is implemented, with the above correction to prevent inflated FA rates with relatively small samples, by using the original group size of each subject in place of n_0, the original sample size, in

$$C_{crit} = \frac{G}{G+1} \times \frac{n_0 - 1}{n_i}$$

defined earlier, while n_i remains the total number of subjects remaining in the whole sample.

When this was tested with independent variables and no outlier added, the FA rate generally lied within the 99% confidence interval of 4.44% to 5.56% for 10 000 simulated studies. The lower group size investigated was 10. The case of two groups of 10 subjects each on four variables gave, on three separate runs, 5.37%, 5.65% and 5.17% FAs. Two groups of 10 cases produced 5.18% FAs with two variables and successively 5.52% and 5.09% with eight variables. Groups of respectively 10 and 20 cases on four variables produced 5.4% FAs and, on replication, 5.32%. Two groups of 20 cases on four variables gave 5.41% and 5.46% FAs. Two groups of 30, again on four variables, gave 4.88% FAs. Four groups of 10 cases with either 2 or 8 variables gave FA counts within expected range. The respective advantages of the CP and combo methods were not investigated for more than one group, but there is no reason to doubt that similar results would be obtained.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	SEXE	DATENAIS	SALACT	SALDEB	EXP	MINORITE	CATEMP	EDUC	TEMPS	LGSALACT	LGSALEMB	LGEXPPAS
2	1	m	1952-02-03	$57 000	$27 000	144	0	3	15	98	4,633	4,267	2,188
3	2	m	1958-05-23	$40 200	$18 750	36	0	1	16	98	4,418	4,011	1,663
4	3	f	1929-07-26	$21 450	$12 000	381	0	1	12	98	3,872	3,544	2,592
5	4	f	1947-04-15	$21 900	$13 200	190	0	1	8	98	3,898	3,672	2,301
6	5	m	1955-02-09	$45 000	$21 000	138	0	1	15	98	4,491	4,097	2,170
7	6	m	1958-08-22	$32 100	$13 500	67	0	1	15	98	4,258	3,699	1,886
8	7	m	1956-04-26	$36 000	$18 750	114	0	1	15	98	4,342	4,011	2,093
9	8	f	1966-05-06	$21 900	$9 750	0	0	1	12	98	3,898	3,097	1,000
10	9	f	1946-01-23	$27 900	$12 750	115	0	1	15	98	4,143	3,628	2,097
11	10	f	1946-02-13	$24 000	$13 500	244	0	1	12	98	4,000	3,699	2,405
12	11	f	1950-02-07	$30 300	$16 500	143	0	1	16	98	4,212	3,903	2,185
13	12	m	1966-01-11	$28 350	$12 000	26	1	1	8	98	4,157	3,544	1,556
14	13	m	1960-07-17	$27 750	$14 250	34	1	1	15	98	4,138	3,760	1,643
15	14	f	1949-02-26	$35 100	$16 800	137	1	1	15	98	4,324	3,919	2,167
16	15	m	1962-08-29	$27 300	$13 500	66	0	1	12	97	4,124	3,699	1,881
17	16	m	1964-11-17	$40 800	$15 000	24	0	1	12	97	4,428	3,813	1,531
18	17	m	1962-07-18	$46 000	$14 250	48	0	1	15	97	4,505	3,760	1,763
19	18	m	1956-03-20	$103 750	$27 510	70	0	3	16	97	4,953	4,279	1,903
20	19	m	1962-08-19	$42 300	$14 250	103	0	1	12	97	4,452	3,760	2,053
21	20	f	1940-01-23	$26 250	$11 550	48	0	1	12	97	4,088	3,484	1,763
22	21	f	1963-02-19	$38 850	$15 000	17	0	1	16	97	4,395	3,813	1,431
23	22	m	1940-09-24	$21 750	$12 750	315	1	1	12	97	3,889	3,628	2,512
24	23	f	1965-03-15	$24 000	$11 100	75	1	1	15	97	4,000	3,415	1,929
25	24	f	1933-03-27	$16 950	$9 000	124	1	1	12	97	3,470	2,699	2,127
26	25	f	1942-07-01	$21 150	$9 000	171	1	1	15	97	3,854	2,699	2,258
27	26	m	1966-11-08	$31 050	$12 600	14	0	1	15	96	4,232	3,613	1,380
28	27	m	1954-03-19	$60 375	$27 480	96	0	3	19	96	4,666	4,278	2,025
29	28	m	1963-04-11	$32 550	$14 250	43	0	1	15	96	4,268	3,760	1,724
30	29	m	1944-01-28	$135 000	$79 980	199	0	3	19	96	5,083	4,854	2,320
31	30	m	1961-09-17	$31 200	$14 250	54	0	1	15	96	4,236	3,760	1,806
32	31	m	1964-02-24	$36 150	$14 250	83	0	1	12	96	4,345	3,760	1,968
33	32	m	1954-01-28	$110 625	$45 000	120	0	3	19	96	4,985	4,562	2,114
34	33	m	1961-03-18	$42 000	$15 000	68	0	1	15	96	4,447	3,813	1,892
35	34	m	1949-02-02	$92 000	$39 990	175	0	3	19	96	4,892	4,498	2,267
36	35	m	1961-08-22	$81 250	$30 000	18	0	3	17	96	4,828	4,332	1,447
37	36	f	1963-08-07	$31 350	$11 250	52	0	1	8	96	4,239	3,439	1,792
38	37	m	1954-10-09	$29 100	$13 500	113	1	1	12	96	4,179	3,699	2,090
39	38	m	1962-04-27	$31 350	$15 000	49	1	1	15	96	4,239	3,813	1,771
40	39	m	1960-06-22	$36 000	$15 000	46	1	1	16	96	4,342	3,813	1,748
41	40	f	1933-08-28	$19 200	$9 000	23	1	1	15	96	3,716	2,699	1,519
42	41	f	1961-03-18	$23 550	$11 550	52	1	1	12	96	3,980	3,484	1,792

Figure 7. Screen print showing part of the data after the CP procedure highlighted in green the cases identified as outliers and the following descriptive Rosner's procedure turned to red the values identified as outliers on their variable. Underlined values are cases with a different outcome when the corresponding procedure is applied separately to each group.

Two other practical considerations are relevant. One is the availability of a computer program to apply the procedure. This is here solved by Excel macros embedded in *OutlierDetection.xls* available on the journal's web site. The first time this file is opened, a message is displayed indicating that the security level must be lowered (from high to medium) in order for the macros to be activated. The data must be gathered in a separate Excel file with the dependent variables (in their transformed version is required) in consecutive columns. Optionally, the first row may contain text (variable heading). If the data are in groups, group membership is restricted to a single data column, but it does not matter that groups are specified by text or by numbers and, in the latter case, group numbers do not have to be consecutive. The Excel data file must be opened when OutlierDetection.xls is already opened. This provides access to its macros to the data file. Depending on Excel version, a menu item may be labeled "Complements" and clicking on this will provide access to a function labeled "Multivariate Outliers" (or "Étrangers multi-variables" if the operating system is in French) or, in older versions of Excel, a menu item will directly wear this label. Upon activating "Multivariate Outliers", the window illustrated in Figure 6 appears, which requires one to select the data columns and, if required, the group ID column. Default values are presented and may be modified if needed. The function applies the CP procedure by default (but combo may be selected instead). Cases with missing values are highlighted by a yellow background. Outliers on this procedure are flagged by changing to green the background color of all used data of the cases concerned. But it is important to identify why a case is labeled as outlier. For this purpose when CP is the selected procedure, the (modified) Rosner procedure is also applied descriptively to each variable, with per test α = global α/number of dependent variables, unless "Other user defined per test Alpha" was selected, which asks for the desired value. This CP approach differs from the combo procedure because the latter divides the global α by

(p + ½) instead of by p, including for the multivariate test, while CP does the multivariate test at the nominal α level. Cases identified as outliers on any single variable are flagged by turning their data value to red. There is no restriction that the case involved was previously identified as a multivariate outlier. If no group is specified, a group variable with constant value is temporarily created and deleted at the end of the procedure. Should an error occur, this column might be seen remaining in the data file.

When executed, the macro will either display a message to the effect that no outlier was detected or it will highlight the suspected outliers. Figure 7 illustrate a segment of the outcome. Cases with ID 24, 25, 40 and (not seen) 111 were flagged as outliers. All four cases flagged as outliers had 9 000$ as initial salary. None of the seven cases with the next lowest initial salary level (9 750$) was identified as outlier even when the maximum number of outliers to be detected was raised to 15.

The last of our practical considerations is that relatively severe lack of homogeneity of the group covariance matrices may bias the tests. In particular, subjects belonging to groups with larger dispersion run an inflated risk of being declared outliers when they are tested with the pooled covariance matrix, which underestimate their actual dispersion. When severe heterogeneity of covariance is suspected, the solution is to test each group separately. When data are in groups, OutlierDetection.xls assumes homogeneity of covariance but also checks the groups separately with an alpha level that maintains the overall experiment-wise FA rate at the nominal level (5% by default). The nominal alpha for group g is $\alpha_g = 1 - 0.95^{n_g/n}$, such that the products of all (1- α_g) is 1-α. Cases with a different outcome in this group-wise and in the original sample-wise procedures are flagged by underlying their group ID value for a difference in outcome in the CP procedure or the individual value for a difference in outcome on a variable by variable test. The colors of the underlined values or of their background reflect the global test, not the tests performed on each group separately.

In Figure 7, the underlined values of cases 24 and 25 for the group ID variable (CATEMP) and for the variable labeled LGSALEMB (log initial salary) indicate that these two cases would not have been detected if the testing had been done separately for each group. It remains the responsibility of the user to estimate if this could rather be a consequence of lack of power or of larger variance in the group labeled 1 than in the other groups.

General discussion

The original formulas provided by Rosner (1983) and by Caroni and Prescott (1992) tended to produce more than their nominal FA rate with sample size less than 25. With a single variable, a slight but significant excess of FAs was observed in the present simulations even for as many as 40 cases. This bias could be satisfactorily corrected by a slight modification of the formula. Similarly, in the presence of four or five outliers in certain data configurations, the FA rate among the remaining cases truly belonging to the population could be inflated and this could be alleviated by a modification of the rejection rule. With these modifications, good control over the experiment-wise FA rate is achieved.

Of the two approached investigated, namely only applying the multivariate based CP procedure or applying a test on each variable plus a multivariate test with a suitable correction for the number of tests (combo procedure), neither appears uniformly more powerful that the other at detecting true outliers. For sample size 100, the simulations suggest that CP is better up to perhaps 12 variables after which combo would provide more power. Since the combo procedure applies each test with a nominal α divided by number of variables plus one half, its superiority over CP must come from a more serious deterioration of the multivariate test when the outlying values are on a small portion of the variables. This obviously must depend on the data structure. It could also depend on sample size. Further studies would be required to establish whether the same relationship holds (CP better only up to 12 variables) in smaller or in larger samples.

The procedures were tested with a maximum of five outliers when *k*, the maximum to be detected, was set to 10 (as in Rosner, 1983, and Caroni & Prescott, 1992). The effect of specifying too small a value for *k* (i.e. having more than *k* outliers in the sample) might actually cause detecting much fewer that *k* outliers, because the remaining outliers would produce masking. If *k* outliers are actually reported, there is a clear indication that the iterative procedure might have stopped too early and the procedure could then be repeated with a larger *k* limit. But since k does not appear in the procedure formulas, it could have been set higher than 10, mostly at the cost of longer computing time (which matters almost only in the simulations of thousands of studies). Obviously, *n-k* must remain more that *p*, to ensure that the underlying matrix inversion can be done. Therefore *k* cannot exceed *n-p*. OutlierDetection.xls internally reduces *k* to *n-p*-1, if required, to prevent function failure.

It seems unlikely that setting *k* to an arbitrary larger value would inflate the FA rate at all. During the verification of the FA rate for data in groups in the absence of true outliers, the cumulative number of studies with at least one FA was obtained as a function of iteration cycle (i.e. testing for 1, 2, … up to 10 outliers). The maximum of FAs in each

condition was always reached by the fourth cycle. In other words, the last six most extreme cases in the sample never met the current criterion to be falsely flagged as outliers. This is likely related to the increasing density of the tails of the distributions as extreme cases are removed. Therefore, the FA rates would very likely have been identical had *k* been set to a higher value.

Finally, although this may sound off topic, it is important to insist that outlier detection must always be preceded by inspection of the distributions and that suitable transformations must be applied, particularly for skewed data distributions. If a variable is to be transformed (e.g., because its skewness is outside ±2 standard errors), then one should aim that the transformed variable skewness be within one standard error, to be confident that this new scale is close to symmetrical in the population. When a constant must be included before a logarithmic or a square root transformation, the choice of that constant is often crucial. For example, in the illustrative data, Current Salary (SalAct in Figure 7) was transformed into LgSalAct = LG10(SalAct-14000), with a skewness of 0.058, the skewness standard error being 0.112. Using constant 10000 produced a skewness of 0.565 while a constant of 15000 inverted the skewness to -0.265. The often seen recommendation of adding a fixed 0.5 or 1 before taking the logarithm is much too restrictive and was clearly inappropriate here.

References

Bradu, D., and Hawkins, D.M. (1982) Location of Outliers in Two-Way Tables Using Tetrads. *Technometrics*, 24: 103-108.

Caroni, C. (1998) Wilks' outlier test in more than one multivariate sample. *Communications in Statistics - Simulation and Computation*, 27: 79-94.

Caroni, C., and Prescott, P. (1992) Sequential application of Wilks' multivariate outlier test. *Applied Statistics*, 41: 355-364.

Matsumoto, M., and Nishimura, T. (1998) Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudorandom Number Generator. *ACM Transactions on Modeling and Computer Simulation*, 8: 3-30.

Rosner, B. (1983) Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25: 165-172.

Wilks, S.S. (1963) Multivariate statistical outliers. *Sankhya A*, 25: 407-426.

Manuscript received 16 November 2011.
Manuscript accepted 7 May 2012.