Chapter 5

Outlier Detection for Mass Spectrometric Data

HyungJun Cho and Soo-Heang Eo

Abstract

Mass spectrometry data are often generated from various biological or chemical experiments. However, due to technical reasons, outlying observations are often obtained, some of which may be extreme. Identifying the causes of outlying observations is important in the analysis of replicated MS data because elaborate pre-processing is essential in order to obtain successful analyses with reliable results, and because manual outlier detection is a time-consuming pre-processing step. It is natural to measure the variability of observations using standard deviation or interquartile range calculations, and in this work, these criteria for identifying outliers are presented. However, the low replicability and the heterogeneity of variability are often obstacles to outlier detection. Therefore, quantile regression methods for identifying outliers with low replication are also presented. The procedures are illustrated with artificial and real examples, while a software program is introduced to demonstrate how to apply these procedures in the **R** environment system.

Key words Outlier detection, Data preprocessing, Standard deviation, Interquartile range, Quantile regression

1 Introduction

Mass spectrometry (MS) data are often generated from various biological or chemical experiments. Such large amounts of data are usually analyzed automatically in a computing process that consists of pre-processing, significance testing, classification, and clustering. Elaborate pre-processing is essential to obtain successful analyses with reliable results. A key pre-processing step is the detection of outliers, which may have extreme values due to technical reasons [1]. Possible outlying observations need to be examined carefully, and then corrected for or eliminated if necessary. However, as the manual examination of all observations for outliers is time-consuming, possible outliers must be detected automatically.

An outlier is an observation that falls well above or well below the overall bulk of the data [2–4]. A natural approach to detect outliers is to investigate the distribution of the observations and evaluate the outlying degrees of potential outliers. The investigation can

Klaus Jung (ed.), *Statistical Analysis in Proteomics*, Methods in Molecular Biology, vol. 1362,
DOI 10.1007/978-1-4939-3106-4_5, © Springer Science+Business Media New York 2016

be conducted for each peptide because the distributions of observations of peptides may differ substantially. It is natural to measure the variability of observations for each peptide by calculating the standard deviation (SD) or interquartile range (IQR) of each sample [5].

The SD and IQR criteria may produce unreliable outcomes in the case of a few replicates. Furthermore, they are not applicable for duplicated samples. Another, perhaps naive, approach for detecting outliers statistically involves constructing lower and upper fences of differences between two samples for all peptides. A suspected outlier is then an observation whose value is either smaller than the lower fence or greater than the upper fence. However, this may generate a spurious result because variability is heterogeneous in high-throughput data generated even from MS experiments. Naive outlier detection methods such as these ignore the heterogeneity of variability, and may often miss true outliers at high levels and select false outliers at low levels. If a number of technical replicates for each peptide under the same biological condition can be obtained in MS experiments, a search for outliers can be conducted for each peptide. However, only a small number of replicates are usually subjected to MS experiments due to the high cost of experiments and the limited supply of biological samples. Instead, a more elaborate approach for detecting outliers with low false-positive and false-negative rates in MS data is to utilize quantile regression, which is especially useful when the number of technical replicates is small. The outlier detection procedures are illustrated in the next section, with artificial and real datasets in the **R** environment system.

2 Outlier Detection Methods

Suppose that there are n replicated samples and p peptides in an MS dataset. Then let xij be the i-th replicated observation for the j-th peptide from experiments under the same biological or experimental condition, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$ and let $y_{ij} = \log_2\left(x_{ij}\right)$. Typically, n is small and p is very large in high-throughput data, i.e., $p \gg n$. We introduce the standard deviation, interquartile range, and quantile regression approaches for identifying outliers in this section.

2.1 Standard Deviation Criteria

The standard deviation describes the distance between the data and the mean, thus providing a measure of the variability of the data. The standard deviation s is defined as the square root of the sum of squared deviations divided by the sample size minus 1, i.e., $s = \sum_i \left(y_i - \bar{y}\right)^2 / \left(n - 1\right)$. The z-score for an observation is the number of standard deviations that it falls away from the mean. A positive z-score indicates the observation is above the mean, while a negative *z*-score indicates that the observation is below the mean. For sample data, an observation from a bell-shaped distribution is a potential outlier if its *z*-score < −3 or > +3. The *z*-score criterion for identifying outliers is summarized below:

1. Compute the standard deviation, *sj* for each peptide *j*, and then *z*-score $z_{ij} = \left(y_{ij} - \overline{y}_j\right) / s_j$, where $\overline{y}_j$ and *sj* are the sample mean and standard deviation, respectively.
2. For each peptide *j*, observation *yij* is flagged as an outlier if $z_{ij} < -k$ or $z_{ij} > k$, where $k = 2$ or 3.
3. This *z*-score criterion works well when the data follows a bell-shaped, normal distribution. Thus, the thresholds *k* = 2 and 3 indicate that 95 and 99.7 % of the observations fall within 2 and 3 SDs of the mean, respectively.

Grubbs et al. [6] developed a more elaborate procedure, where the threshold is more precise, and outliers are removed recursively. This is the Grubbs' test, and its method for identifying outliers is summarized below:

1. Compute the test statistic $G_{ij} = \max_{i=1,\ldots,n} | y_{ij} - \overline{y}_j | / s_j$, where the sample mean is $\overline{y}_j$ and standard deviation is *sj* for peptide *j*.
2. For each peptide *j*, observation *yij* is flagged as an outlier if $G_{ij} > c$, where *c* is the critical value (*see* **Note 1**).
3. Remove the detected outlier, and then repeat steps 1–3 until no further outliers are detected.

If $n = 2$, the statistic is always $1 / \sqrt{n}$; thus, this test is applicable for $n > 2$. Grubbs' test is based on the assumption of normality; therefore, one should first verify that the data could be reasonably approximated by a normal distribution before applying the test. Grubbs' test detects one outlier at a time. This outlier is expunged from the dataset and the test is reiterated until no further outliers are detected. However, multiple iterations change the probabilities of detection, and the test should not be used for sample sizes of six or less since it frequently tags most of the points as outliers [7].

2.2 Interquartile Range Criteria

The *p*-th percentile is a value such that *p* percentages of the observations fall at, or below, a certain value. Three useful percentiles are the quartiles. The first quartile Q_1 is the 25th percentile, where the lowest 25 % of the data fall below it. The second quartile Q_2 is the 50th percentile, which is the median. The third quartile Q_3 is the 75th percentile, and the highest 25 % of the data exists above it. The quartiles split the data into four parts, each containing quarter (25 %) of the observations. The interquartile range (IQR) is the distance between the third and first quartiles, i.e., $\text{IQR} = Q_3 - Q_1$. An observation is declared an outlier if it is greater than 1.5 IQR below the first quartile or more than 1.5 IQR above the third quartile. Thus, the lower and upper fences for outliers are $Q_1 - 1.5\,\text{IQR}$ and $Q_3 + 1.5\ \text{IQR}$ [8]. This IQR criterion for identifying outliers is summarized as follows:

1. Compute the first and third quartiles, Q_1j and Q_3j, for each peptide j, and then its IQR: $\text{IQR}_j = Q_{3j} - Q_{1j}$.
2. For each peptide j, observation yij is flagged as an outlier if $y_{ij} < Q_{1j} - k\ \text{IQR}_j$ or $y_{ij} > Q_{3j} + k\ \text{IQR}_j$, where $k = 1.5$ or 3.

In this IQR criterion, a coefficient k determines the strictness of capturing outlying observations. Values of $k = 1.5$ or 3 are often used. A larger value of k selects outlying observations more conservatively.

The distribution of observations may not be symmetric about the median, but instead may be skewed to the left or the right, implying that the middle of the first and third quartiles is in fact not the median. Thus, the distance from the first quartile to the median is significantly different of that from the third quartile to the median. In this situation, IQR can be too large for one side and too small for the other.

As an alternative, the semi-interquartile range (SIQR) can be more effective. That is, the left and right SIQRs are used rather than IQR. This SIQR criterion for identifying outliers is summarized as follows:

1. Compute the first, second, and third quartiles, Q_1j, Q_2j, and Q_3j, for each peptide j, and then its SIQR: $\text{SIQR}j^{\text{L}} = Q_2j - Q_1j$ and $\text{SIQR}j^{\text{U}} = Q_3j - Q_2j$.
2. For each peptide j, observation yij is flagged as an outlier if $y_{ij} < Q_{1j} - 2k\ \text{SIQR}_j^{\text{L}}$ or $y_{ij} > Q_{3j} + 2k\ \text{SIQR}_j^{\text{U}}$, where k=1.5 or 3.

2.3 Quantile Regression Approaches

The above IQR and SD criteria require for the data to follow a normal distribution, and for the sample sizes to be large enough (*see* **Note 2**). However, the assumptions may not be satisfied for some MS analyses, and in particular, the sample size is often small (*see* **Note 3**).

In duplicated experiments ($n = 2$), two observed values for each peptide should be theoretically identical, but are not identical in practice due to their variability; however, they should not differ substantially. The tolerance of the difference between the two observed values from the same condition is not constant because their variability is heterogeneous. The variability of high-throughput data depends on the intensity levels.

Lower and upper fences can be constructed for detecting outliers using quantile regression in an M–A plot with M and A values in vertical and horizontal axes, respectively, where Mj is the difference between replicated samples for j and Aj is the average, i.e., $M_j = y_{1j} - y_{2j} = \log_2\left(x_{1j} / x_{2j}\right)$ and

$A_j = \left(y_{1j} + y_{2j}\right)/2 = \left(1/2\right)\log_2\left(x_{1j}x_{2j}\right)$ to detect the outliers accounting for the heterogeneity of variability [9]. By applying the regression, we compute the 0.25 and 0.75 quantile estimates, $Q_1(A)$ and $Q_3(A)$, of the differences, M, depending on the levels, A. Then we construct the lower and upper fences: $Q_1\left(A\right) - 1.5\ \mathrm{IQR}\left(A\right)$ and $Q_3\left(A\right) + 1.5\ \mathrm{IQR}\left(A\right)$, where $\mathrm{IQR}\left(A\right) = Q_3\left(A\right) - Q_1\left(A\right)$. To obtain quantile estimates that depend on the levels more flexibly, nonlinear or nonparametric quantile regression can be utilized [10]. This quantile regression approach [1], called the *OutlierD* algorithm, is summarized as follows:

1. Generate an M–A plot with M and A values in vertical and horizontal axes, respectively, where Mj is the difference between replicated samples for j and Aj is the average.
2. Apply linear, nonlinear, or nonparametric regression and then compute the 0.25 and 0.75 quantile estimates, $Q_1(A)$ and $Q_3(A)$, of the differences, M, depending on the levels, A.
3. Construct the lower and upper fences: $Q_1\left(A\right) - k\ \mathrm{IQR}\left(A\right)$ and $Q_3\left(A\right) + k\ \mathrm{IQR}\left(A\right)$, where $\mathrm{IQR}\left(A\right) = Q_3\left(A\right) - Q_1\left(A\right)$ and $k = 1.5$ or 3.
4. Peptide j is claimed as containing an outlying observation if $M_j < Q_1\left(A_j\right) - k\ \mathrm{IQR}\left(A_j\right)$ or $M_j > Q_3\left(A_j\right) + k\ \mathrm{IQR}\left(A_j\right)$, where $k = 1.5$ or 3.

A larger value of k selects outliers more conservatively. In this approach, one of the two samples is outlying, but which one is not known.

In multiple experiments $\left(n \geq 2\right)$, it is natural to search for outliers based on all observed values in a high-dimensional space. An outlier will be at a very large distance from the center of the distribution of a peptide. The cutoffs of distances for classification of outliers depend on the degree of variability from the center. The degree of variability is dependent on intensity levels, and the center can be defined as a 45° line from the origin. More flexibly, the center can be obtained by principal component analysis (PCA) [11]. The first principal component (PC) becomes the center of each intensity level, i.e., a new axis for intensity levels. The experiments are replicated under the same biological and technical condition; hence, the first PC can explain most variations. It implies that it is enough to use the first PC practically. An outlier will be at a large distance from its projection. Following the notations for applying quantile regression, we can define the distance of peptide j to the projection as Mj and the length of the projection on the new axis as Aj. Then the first and third quantiles can be obtained by applying quantile regression on an M–A plot with M and A on the vertical and horizontal axes, respectively. The quantile regression algorithm that uses this projection [7] is called the *OutlierDM* algorithm, and is summarized as follows:

1. Shift the sample means to the origin (0,…,0), i.e., $y_{ij}^{*} = y_{ij} - \overline{y}_{i}$.
2. Find the first PC vector v using principal component analysis (PCA) on the space of $y_1^*,\ldots,y_n^*$.
3. Obtain the projection of a vector $\boldsymbol{y}_j^* = \left(y_{1j}^*,,\ldots,,y_{nj}^*\right)$ of each peptide *j* on v, where $j = 1,\ldots,p$.
4. Compute the signed length, *Aj*, of the projection and the length, *Mj* of the difference between a vector of peptide *j* and the projection, where $j = 1,\ldots,p$.
5. Obtain the first and third quantile values $Q_1(A)$ and $Q_3(A)$, on an M–A plot using a quantile regression approach. Then calculate $\mathrm{IQR}(A) = Q_3(A) - Q_1(A)$.
6. Construct the lower and upper fences, $\mathrm{LB}(A) = Q_1(A) - k\ \mathrm{IQR}(A)$and $\mathrm{UB}(A) = Q_3(A) + k\ \mathrm{IQR}(A)$, where $k = 1.5$ or 3.
7. Peptide *j* is claimed as containing one or more outlying observations if it is located above the upper fence or under the lower fence.

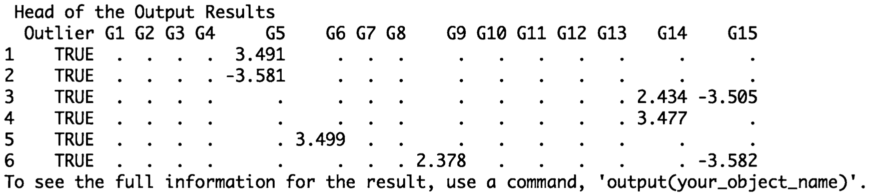
This projection quantile regression approach utilizes all of the multiple replicates simultaneously, and a high-dimensional problem reduces to two-dimensional one that can easily be solved. Note that the quantile regression approaches only determines whether each peptide contains one or more outliers, but not which observation is an outlier. A visual approach (*see* **Note 4**) is useful to identify which observation(s) of the selected peptide is (are) outlying, is illustrated in the next section.

3 Illustrations

In this section, we illustrate how to detect outliers in two cases with artificial and real examples by using an analysis written in **R** package *OutlierDM* [12] (*see* **Note 5**). The first case is illustrated with an artificial dataset to detect outlying samples for each peptide, while the second case uses a real dataset to detect the peptides containing at least one outlying observation when the number of replicates is small.

3.1 When the Number of Replicates Is Sufficiently Large

The primary purpose of outlier detection in MS data is to determine which observations for each peptide are outlying. If the number of replicates is large enough (*see* **Note 2**), one of the SD and IQR criteria can seek out the outliers within each peptide. For illustration, an artificial data set with 200 peptides and 15 samples is generated [7]. This dataset (called “toy”) contains ten peptides,



```
 Head of the Output Results
  Outlier G1 G2 G3 G4     G5    G6 G7 G8    G9 G10 G11 G12 G13   G14    G15
1    TRUE  .  .  .  .  3.491     .  .  .     .   .   .   .   .     .      .
2    TRUE  .  .  .  . -3.581     .  .  .     .   .   .   .   .     .      .
3    TRUE  .  .  .  .      .     .  .  .     .   .   .   .   . 2.434 -3.505
4    TRUE  .  .  .  .      .     .  .  .     .   .   .   .   . 3.477      .
5    TRUE  .  .  .  .      . 3.499  .  .     .   .   .   .   .     .      .
6    TRUE  .  .  .  .      .     .  .  . 2.378   .   .   .   .     . -3.582
To see the full information for the result, use a command, 'output(your_object_name)'.
```

Fig. 1 Outlier detection using the Grubbs' test for the first six peptides of the toy dataset; the test statistics are given for the detected outliers and the *dots* are given for non-outliers

and each of them has one outlying observation. This toy dataset can be called within the **R** package *OutlierDM* by using the following commands:

> \> library(OutlierDM)
>
> \> data(toy)

To detect outlying observations using the Grubbs' test with significance level 0.01, the function *odm()* of *OutlierDM* can be called as follows:

> \> fit = odm(x = toy, method = "grubbs", alpha = 0.01)
>
> \> fit

These **R** commands create an object *fit* using the three input arguments, dataset used (**x** = toy), outlier detection method (**method** = "grubbs"), and significance level (**alpha** = 0.01), and then display a table consisting of dots (test statistics for the detected outliers) for the first six peptides as an output (Fig. 1).

In the output, the first column is the row number and the second column indicates whether each peptide contains one or more outlying observations, shown as TRUE. Columns G1–G15 give the test statistics for the detected outliers, while the dots for non-outliers. To see all the peptides, the function *output(fit)* can be conducted in the **R** environment. In this example, 12 peptides were flagged as containing one or more outlying observations, two of which were flagged falsely. In the first six peptides shown, peptide 3 found two outlying observations, but one of them was flagged falsely. The other five peptides detected all the outlying samples correctly. The detected outlier for each peptide can be shown graphically by the function *oneplot()*:

> \> oneplot(fit, i = 1)

The object *fit* was generated from the function *odm()* and index "i" indicates the row number corresponding to a peptide. Figure 2 shows the dot plot of log2-transformed data points with one outlier (marked by an asterisk) detected by the Grubbs' test with significance level 0.01 (*see* **Note 4**).

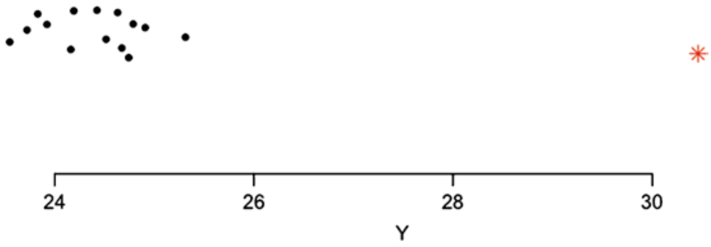


Fig. 2 Outlier detection using the Grubbs' test for the first peptide of the toy dataset; the outlier is indicated as an *asterisk*

3.2 When the Number of Replicates Is Small

We would like to know which observations for each peptide are outlying, but for cases where the number of replicates is small (*see* **Note 3**). In these events, a quantile regression approach can be utilized to detect the peptides having at least one outlying observation. For illustration, we consider a real-life dataset obtained from three replicated LC/MS/MS experiments with 922 peptides ($n = 3$ and $p = 922$). The details regarding the experiment can be found in refs. 1 and 7. This dataset can be called up by the following command:

```
> data(lcms3)
```

We first illustrate how to detect outliers under the duplicated experiment ($n = 2$). For instance, consider the first two replicates of the "lcms3" dataset and apply the *OutlierD* algorithm to the duplicated data set:

```
> fit2 = odm(x = lcms3[,1:2], method = "pair", k = 3)
> outliers(fit2)
> plot(fit2)
```

The argument method = "pair" is for the *OutlierD* algorithm and $k = 3$ is a threshold (i.e., a coefficient) used within IQR. Using the function *outliers(fit2)* generates the output shown in Fig. 3. In this output, the first column indicates the row numbers of the peptides containing an outlier observation. The next columns consist of log2-transformed values (N_1 and N_2), A and M values, the first and third quartiles (Q_1 and Q_3), and lower and upper bounds (LB and UB), respectively. Figure 4 shows the M–A plot from the object *fit2* and the superimposed lines separate outlying peptides from normally observed peptides.

Next, we use all the three replicates simultaneously to detect outliers ($n = 3$). The number of replicates is still small, so the SD and IQR criteria are not applicable. In this case, the *OutlierDM* algorithm is applied to the lcms3 dataset:

```
> fit3 = odm(lcms3, method = "proj", k = 3)
> outliers(fit3)
> plot(fit3)
```

```
    Outlier   N1   N2    A      M      Q1     Q3     LB    UB
66     TRUE 32.3 32.8 32.6 -0.600 -0.0317 0.0317 -0.222 0.222
94     TRUE 23.7 18.2 21.0  5.220 -0.6227 0.6227 -4.359 4.359
145    TRUE 23.9 19.1 21.5  4.635 -0.5960 0.5960 -4.172 4.172
236    TRUE 24.9 19.8 22.4  4.878 -0.5505 0.5505 -3.854 3.854
319    TRUE 15.0 22.2 18.6 -7.473 -0.7407 0.7407 -5.185 5.185
324    TRUE 26.9 21.7 24.3  4.946 -0.4531 0.4531 -3.172 3.172
413    TRUE 19.7 24.2 21.9 -4.742 -0.5740 0.5740 -4.018 4.018
448    TRUE 31.7 32.1 31.9 -0.542 -0.0679 0.0679 -0.475 0.475
458    TRUE 25.3 28.6 26.9 -3.522 -0.3186 0.3186 -2.230 2.230
460    TRUE 28.7 17.5 23.1 10.955 -0.5122 0.5122 -3.586 3.586
541    TRUE 25.9 20.4 23.2  5.319 -0.5111 0.5111 -3.578 3.578
661    TRUE 15.5 22.7 19.1 -7.436 -0.7188 0.7188 -5.032 5.032
751    TRUE 18.0 23.5 20.7 -5.773 -0.6343 0.6343 -4.440 4.440
782    TRUE 28.1 24.9 26.5  3.029 -0.3419 0.3419 -2.393 2.393
796    TRUE 25.3 17.0 21.1  8.116 -0.6134 0.6134 -4.294 4.294
906    TRUE 30.6 20.2 25.4 10.297 -0.3975 0.3975 -2.783 2.783
```

Fig. 3 A list of the outliers detected by the *OutlierD* algorithm for the lcms3 dataset

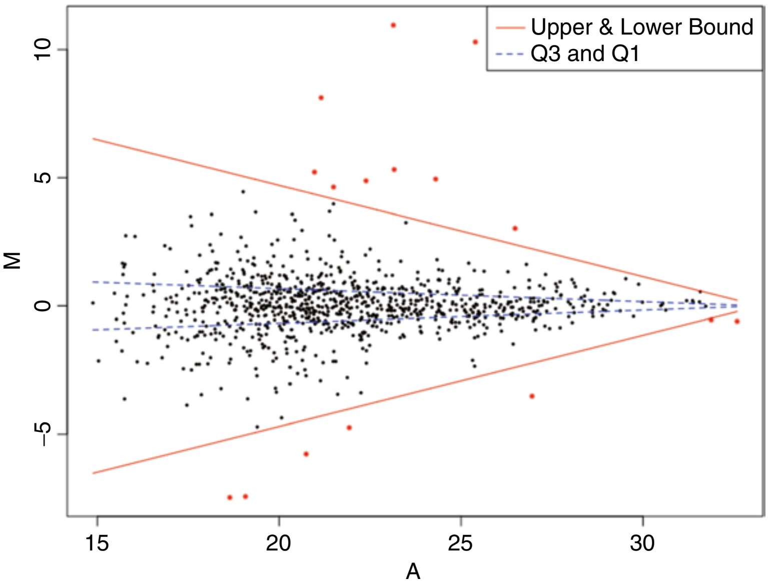
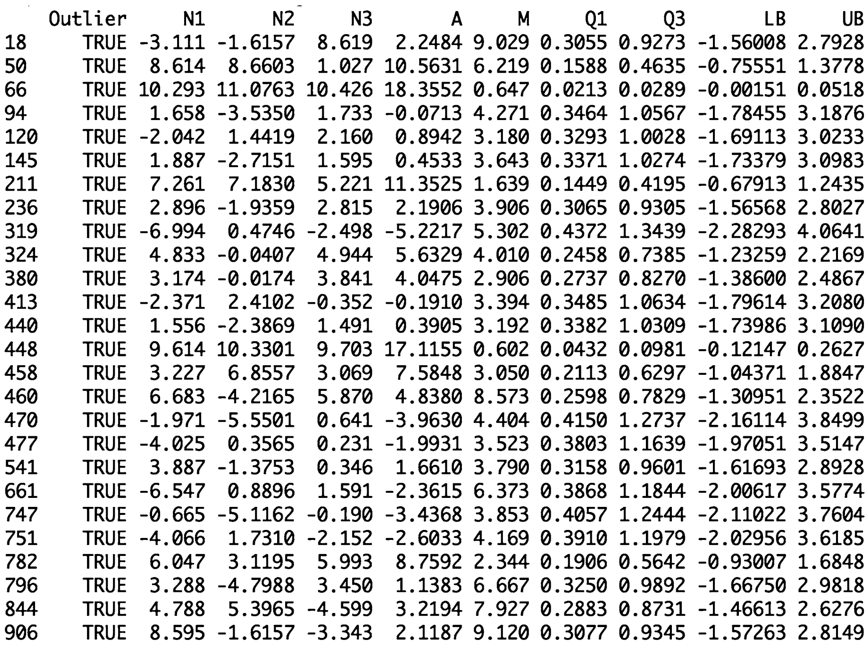


Fig. 4 Outlier detection using the *OutlierD* algorithm in a linear quantile regression analysis for the first two replicates of the lcms3 dataset; the outliers are shown as *red asterisks*

The argument method = "proj" is for *OutlierDM* and $k = 3$ is again a threshold used by IQR. Using the function *outliers(fit3)* generates the output Fig. 5. In this output, the first column indicates the row numbers of the peptides containing an outlier observation. The next columns consist of log2-transformed values (N_1, N_2, and N_3), A and M values, the first and third quartiles (Q_1 and Q_3), and lower and upper bounds (LB and UB), respectively. Figure 6 shows the M–A plot from the object *fit3* and the superimposed lines separate outlying peptides from normally observed peptides.



```
    Outlier     N1      N2     N3       A     M     Q1     Q3       LB     UB
18     TRUE -3.111 -1.6157  8.619  2.2484 9.029 0.3055 0.9273 -1.56008 2.7928
50     TRUE  8.614  8.6603  1.027 10.5631 6.219 0.1588 0.4635 -0.75551 1.3778
66     TRUE 10.293 11.0763 10.426 18.3552 0.647 0.0213 0.0289 -0.00151 0.0518
94     TRUE  1.658 -3.5350  1.733 -0.0713 4.271 0.3464 1.0567 -1.78455 3.1876
120    TRUE -2.042  1.4419  2.160  0.8942 3.180 0.3293 1.0028 -1.69113 3.0233
145    TRUE  1.887 -2.7151  1.595  0.4533 3.643 0.3371 1.0274 -1.73379 3.0983
211    TRUE  7.261  7.1830  5.221 11.3525 1.639 0.1449 0.4195 -0.67913 1.2435
236    TRUE  2.896 -1.9359  2.815  2.1906 3.906 0.3065 0.9305 -1.56568 2.8027
319    TRUE -6.994  0.4746 -2.498 -5.2217 5.302 0.4372 1.3439 -2.28293 4.0641
324    TRUE  4.833 -0.0407  4.944  5.6329 4.010 0.2458 0.7385 -1.23259 2.2169
380    TRUE  3.174 -0.0174  3.841  4.0475 2.906 0.2737 0.8270 -1.38600 2.4867
413    TRUE -2.371  2.4102 -0.352 -0.1910 3.394 0.3485 1.0634 -1.79614 3.2080
440    TRUE  1.556 -2.3869  1.491  0.3905 3.192 0.3382 1.0309 -1.73986 3.1090
448    TRUE  9.614 10.3301  9.703 17.1155 0.602 0.0432 0.0981 -0.12147 0.2627
458    TRUE  3.227  6.8557  3.069  7.5848 3.050 0.2113 0.6297 -1.04371 1.8847
460    TRUE  6.683 -4.2165  5.870  4.8380 8.573 0.2598 0.7829 -1.30951 2.3522
470    TRUE -1.971 -5.5501  0.641 -3.9630 4.404 0.4150 1.2737 -2.16114 3.8499
477    TRUE -4.025  0.3565  0.231 -1.9931 3.523 0.3803 1.1639 -1.97051 3.5147
541    TRUE  3.887 -1.3753  0.346  1.6610 3.790 0.3158 0.9601 -1.61693 2.8928
661    TRUE -6.547  0.8896  1.591 -2.3615 6.373 0.3868 1.1844 -2.00617 3.5774
747    TRUE -0.665 -5.1162 -0.190 -3.4368 3.853 0.4057 1.2444 -2.11022 3.7604
751    TRUE -4.066  1.7310 -2.152 -2.6033 4.169 0.3910 1.1979 -2.02956 3.6185
782    TRUE  6.047  3.1195  5.993  8.7592 2.344 0.1906 0.5642 -0.93007 1.6848
796    TRUE  3.288 -4.7988  3.450  1.1383 6.667 0.3250 0.9892 -1.66750 2.9818
844    TRUE  4.788  5.3965 -4.599  3.2194 7.927 0.2883 0.8731 -1.46613 2.6276
906    TRUE  8.595 -1.6157 -3.343  2.1187 9.120 0.3077 0.9345 -1.57263 2.8149
```

Fig. 5 A list of the outliers detected by the *OutlierDM* algorithm for the lcms3 dataset

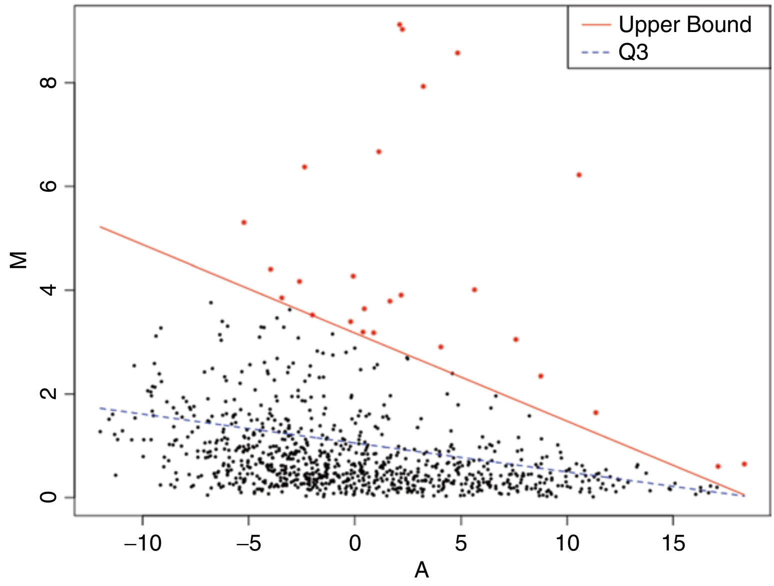


Fig. 6 Outlier detection using the *OutlierDM* algorithm in a linear quantile regression analysis on the lcms3 dataset; the outliers are shown as *red asterisks*

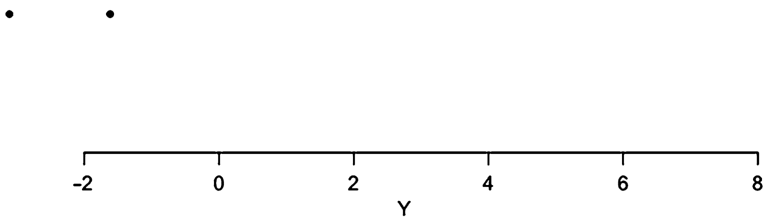


Fig. 7 A dot plot for the 18th peptide of the lcms3 dataset

After detecting the outlying peptides, their raw data points can be plotted to see which observations are furthest from the others using (*see* **Note 4**):

```
> oneplot(fit3, i=18)
```

This generates the dot plot of the log2-transformed values for the 18th peptide, as shown in Fig. 7. It is seen that one observation is far from the other two for the 18th peptide.

4 Notes

1. In Grubbs' test, the critical value c is $\frac{n-1}{\sqrt{n}}\sqrt{t^2_{\frac{\alpha}{2n},n-2}\Big/\left(n-2+t^2_{\frac{\alpha}{2n},n-2}\right)}$, where $t^2_{\frac{\alpha}{2n},n-2}$ is a t-distribution with a degree of freedom $n-2$ and significance level $\alpha/2n$.
2. The standard deviation (SD) and IQR criteria are used to detect outliers for each peptide. These require a sample size greater than six: $n>6$.
3. The quantile regression approaches are used to detect peptides containing one or more outliers when a sample size is small, usually, $n \leq 6$. They also work for a sample size of two ($n=2$).
4. After detecting peptides containing one or more outliers using a quantile regression approach, a visual analysis such as a dot plot can be used to reveal which observations are outlying for a selected peptide.
5. A software program *OutlierDM* [7] based in the **R** environment system is available at http://www.r-project.org/package=OutlierDM for conducting outlier detection.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0007936).

References

1. Cho H, Lee JW, Kim Y-J et al (2008) OutlierD: an R package for outlier detection using quantile regression on mass spectrometry data. Bioinformatics 24:882–884
2. Su X, Tsai C-L (2011) Outlier detection. WIREs Data Mining Knowl Discov 1:261–268
3. Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, New York
4. Aggarwal CC (2013) Outlier analysis. Springer, New York
5. Zimek A, Schubert E, Kriegel H-P (2012) A survey on unsupervised outlier detection in high-dimensional numerical data. Stat Anal Data Min 5:363–387
6. Grubbs FE (1950) Sample criteria for testing outlying observations. Ann Math Statist 21:27–58
7. Eo S-H, Pak D, Choi J, Cho H (2012) Outlier detection using projection quantile regression for mass spectrometry data with low replication. BMC Res Notes 5:246
8. Tukey JW (1976) Exploratory data analysis. Addison-Wesley, Boston, MA
9. Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sin 12:111–139
10. Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge
11. Jolliffe IT (2005) Principal component analysis, 2nd edn. Springer, New York
12. Min H-K, Hyung S-W, Shin J-W et al (2007) Ultrahigh-pressure dual online solid phase extraction/capillary reverse-phase liquid chromatography/tandem mass spectrometry (DO-SPE/cRPLC/MS/MS): a versatile separation platform for high-throughput and highly sensitive proteomic analyses. Electrophoresis 28:1012–1021