INVESTIGATION OF METABOLOMICS TECHNIQUES BY ANALYSIS OF MS PROPOLIS DATA: WHICH PRE-TREATMENT METHOD IS BETTER?

Abdulaziz Alghamdi*[1], Alison Gray[2] & David Watson[3]

Abstract

Metabolomics data usually undergoes both pre-processing of the raw data and then further pre-treatment before any statistical analysis is carried out. Different pre-treatment methods emphasise various aspects of the data, and each method has advantages and disadvantages. The choice of pre-treatment method depends on the biological question of interest, characteristics of the data and the chosen data analysis. In this short paper, we investigate the effects of different pre-treatment methods on four metabolomics data sets arising from chemical analysis of propolis samples collected from honey bee colonies in three different locations in Scotland, and also samples from Libya. Propolis has a variety of biological properties including anti-protozoal and anti-inflammatory effects. As a complex mixture, its biological activity depends on its exact composition, which can be investigated via metabolomic analysis. Two techniques of pre-treatment were applied, namely transformation and scaling. The choice of method was found to greatly affect the results of the principal components analysis (PCA) used to explain the variation in the data. The results indicated that there was no notable (if any) improvement to be made by using any transformation techniques. It was also found for all four data sets that Pareto scaling, incorporating mean centring, performed better than the other scaling approaches considered here in terms of PCA, the analysis of interest, because the results explain more of the variation in the data.

2010 Mathematics Subject Classification: 62-07, 62H25, 62P10

Keywords and phrases: metabolomics data, propolis, pre-treatment, principal components analysis (PCA), transformation, centring, standardisation, vast scaling, Pareto scaling, range scaling, level scaling.

1 Introduction

The metabolomics approach was pioneered by Nicholson et al. (1999) and Nicholson and Wilson (2003). This provides an extremely powerful analytical technique for investigation of properties of biological samples in pharmaceutical, medical and other applications. High-resolution mass spectrometry (MS) or nuclear magnetic resonance spectroscopy (NMR) is used for metabolic profiling of samples, followed by multivariate data analysis. The main purpose is identification of metabolites that may be associated with changes in physiological or environmental conditions.

The aim of this study is to discuss effects of different pre-treatment methods used on MS-based metabolomics data. The optimum pre-treatment methods may depend on the statistical analysis to be carried out on the pre-treated data. In this study the effects of pre-treatment are examined in terms of their effects on data to be examined using principal components analysis (PCA). Four metabolomics data

[1]* Corresponding author: Abdulaziz Alghamdi, Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK. E-mail: Abdulaziz.alghamdi@strath.ac.uk

[2]Alison Gray, Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK.

[3]David Watson, Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK.

sets were used to illustrate the results, namely results of MS analysis of propolis samples collected from honey bee colonies in three different locations in Scotland as well as a further set of samples collected from different locations in Libya. Propolis is a sticky resinous substance produced by honey bees, which consists of a combination of beeswax and resins gathered by the bees from exudates of various surrounding plants. It is used by the bees to seal and maintain their hives, but is also an anti-infective substance which may protect against disease (Saleh et al., 2015; Simone-Finstrom and Spivak, 2010, 2012). Propolis has various important biological properties, including anti-protozoal and anti-inflammatory effects. As a complex mixture, its biological activity depends on its exact composition, which can be examined using metabolomic analysis (Siheri et al., 2016).

Pre-treatment is an important part of any chemometric data analysis. It involves the application of certain operations to data, to remove noise or unwanted variation or to reduce this to an acceptable level. Pre-treatment follows pre-processing. Pre-processing is the general term for processes used to convert the raw instrumental data arising from the chemical analysis into clean data, to make it suitable for pre-treatment and further statistical or chemometric analysis which typically use multivariate analyses. Pre-processing methods include noise filtering, deconvolution, peak detection, alignment and baseline correction, among others (Goodacre et al., 2007). On the other hand, pre-treatment involves transformation and/or scaling of the pre-processed data to prepare it for data analysis (Brereton, 2009).

Metabolomics data are usually presented as a table or array, in which each column represents the chromatographic peak areas or heights (spectral peak intensities) for a putatively identified compound (metabolite) and each row represents a single sample or chemical analysis. Pre-treatment methods include transformation of individual data elements, row scaling operations to make comparable the areas under the spectrum for each sample, and/or column scaling operations on the data for each metabolite. Figure 1 shows sample data used in this paper.

	A	B	C	D	E	F	G
1	row ID	row m/z	row rete	Name	CABIN-	CABIN-E	CABIN-(
2	6504	121.0295	21.5	benzoic acid	1.29E+08	7.04E+07	6.22E+07
3	38	121.0295	11.7	benzoic acid	3.26E+08	7.86E+07	8.51E+07
4	4008	135.0452	8.2	phenylacetic acid	1917576	3085732	1517065
5	6946	135.0452	14.8	phenylacetic acid	3698759	7009556	1234282
6	1282	135.0452	12.3	phenylacetic acid	2.30E+07	2490245	6813808
7	6518	135.0452	16.0	phenylacetic acid	3060457	2932880	2128639
8	6581	135.0452	3.2	phenylacetic acid	440710.4	9163289	1816607
9	81	135.0452	9.3	phenylacetic acid	1.74E+08	1.87E+07	9.21E+07
10	3949	135.0452	7.4	phenylacetic acid	1.49E+07	2145642	2050572
11	1263	135.0452	2.3	phenylacetic acid	1486823	1589281	5794787
12	1266	135.0452	6.1	phenylacetic acid	2051629	3750220	1.32E+07
13	4996	135.0452	2.0	phenylacetic acid	1661488	2180287	1049138

Figure 1: An example of a metabolomics data set from propolis, where column A shows the ID from the MassBank library (Horai et al., 2010) for each spectral peak, column B shows $\frac{m}{z}$ total ion chromatogram displayed for the detected peaks, column C shows retention time, column D shows the names of components where available, and columns E, F, G relate to a label for the hive or apiary. These data were transposed before being processed.

The most common pre-treatment operations are logarithmic transformation, mean-centring, and standardisation. Pre-treatment may have positive or negative effects on the outcome of further data analysis. This paper describes the most important and commonly used pre-treatment methods used on metabolomics data, and examines these using several metabolomics data sets described above and in more detail below.

Section 2 presents materials and methods, including sample collection, transformation, and scaling, and introduces principal component analysis (PCA), of interest for the analysis of these data. Section 3 presents results, and Section 4 provides discussion and conclusions.

2 Materials and methods

2.1 Sample collection

Data from Scotland

Samples of raw propolis were collected during July and August 2014 by beekeepers from several hives of honey bee colonies in three different areas of Scotland, namely Aberdeenshire in the north-east of Scotland, Dunblane in central Scotland, and Fort William in the north-west of Scotland (see Figure 2). These were profiled using liquid chromatography-high-resolution mass spectrometry. The propolis samples

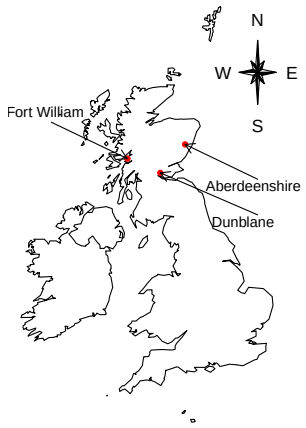


Figure 2: Map of the UK, including the locations of the apiaries supplying the analysed Scottish propolis samples.

contain several hundred compounds, many of which are still unknown structures. The Aberdeenshire data contains twenty seven samples (referred to as data set I). There are fourteen samples from Fort William (data set II) and nine samples from Dunblane (data set III).

Data from Libya

Twelve raw propolis samples were available from different geographical localities in Libya (see map in Figure 3); Tukra (Al Aquriyah), a small village located about 70km east of Benghazi city, Libya (P1); Qaminis (53km south of Benghazi) (P2); Bayda (east of Benghazi city) (P3); Quba (east of Benghazi city) (P4); Kufra A (south-east Libya) (P5); Kufra B (south-east Libya) (P6); Kufra C (south-east Libya) (P7); Ghadames (south-west Libya) (P8); Tripoli (north-west Libya) (P9); Kasser Khiar (located 80 km east of Tripoli) (P10); Khumas (located 120km east of Tripoli) (P11); and Khumas (120km east of Tripoli) (P12). Samples P1-P12 were all used in this study.
Samples P1 and P2 were collected in December 2012, samples P3-P7 were collected in July 2013 and the other samples P8-P12 are from March 2014.

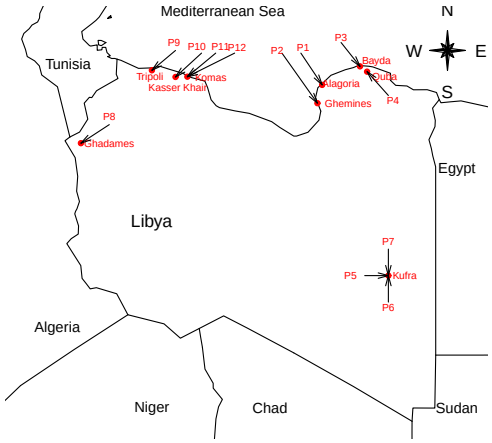


Figure 3: Map of Libya (after Siheri et al., 2016) including the locations of the apiaries supplying the analysed Libyan propolis samples: P1 (Al Aquriyah), P2 (Qaminis), P3 (Bayda), P4 (Quba), P5 (Kufra (A)), P6 (Kufra (B)), P7 (Kufra (C)), P8 (Ghadames), P9 (Tripoli), P10 (Kasser Khiar), P11 (Khumas (A)), P12 (Khumas (B)).

2.2 Transformations

In general, metabolomics data show heteroscedasticity and are often skewed. Also, interactions between the different metabolites are not necessarily additive but can be multiplicative (Boccard et al., 2010). The multivariate statistical methods used for the analysis of metabolomics data are more effective when the data are symmetric, and many statistical significance tests assume that the distribution of the data is approximately normal. Therefore, it is useful to convert the data to approximate normality as closely as possible (Brereton, 2009). Thus, transformations of the elements of metabolomics data sets are important in helping to achieve this aim. There are two common transformations used in this context, i.e. logarithmic and power transformations.

1. **Logarithmic Transformation**

   Logarithmic transformation is important as it reduces heteroscedasticity, converts multiplicative models to additive ones, and reduces the influence of large data values such as outliers and occasional high peaks. This is achieved by replacing a data element x_{ij} by its natural log, $\log(x_{ij})$. Although this has advantages, it has some limitations such as problems handling zeroes or near-zero values, especially when these values are very close to the limit of detection. If data values are below the limit of detection then they are considered as zero, and therefore their logarithms are not defined (Brereton, 2009). To overcome this, usually a small value is added to x_{ij} in the event of x_{ij} being zero, before taking the log.

2. **Power Transformation**

   Power transformation is performed by replacing x_{ij} with x_{ij}^n, e.g. for n = 1/2, this is the square root transformation, and so on. This has strengths such as (Brereton, 2009):

   (a) It reduces the influence of large values such as outliers and occasional high peaks.

(b) It can cope with zero values, removing the need to replace very small values below the limit of detection.

(c) Any uncertainties in small values do not affect data analyses as much as in the case of logarithmic transformation. The smaller a value is relative to other values, the smaller its influence on the the n^{th} root transformed data will be.

Drawbacks of this transformation can be summarised as:

(a) All values should be positive.

(b) If the distribution of the data is approximately log-normal, then power transformation cannot convert it to a symmetric one.

(c) There are many possible values of the power. Trial and error is needed to identify the most appropriate choice. Especially with multivariate data such as in metabolomics, where each metabolite may have a different distribution, it can be difficult to decide on a suitable power.

2.3 Scaling

Before any exploratory analysis of metabolite data, the data must be cleaned, normalised and scaled if there is any removable noise. Two approaches can be applied in data scaling: row scaling (scales each row) and column scaling (scales each column). Column scaling techniques are applied after any pre-processing, transformation and possible data normalisation by row scaling. Many approaches can be used for scaling (row scaling or column scaling). Mean centring and scaling to unit variance (standardisation) are two of the most popular methods. Here we focus on column scaling, assuming that the columns each represent one spectral intensity across all samples (and we did not find row scaling to be necessary). In mean centring, each column of the data table is scaled to a mean of zero by subtracting the column mean from each value in the column. Mean centring may also be applied before standardisation. In standardisation, each column of the data table is scaled to have unit variance by dividing each value in the column by the standard deviation of the column (Craig et al., 2006). Scaling affects the results of multivariate analysis, since it determines which correlations are important. This has implications for methods of analysis such as PCA, which examines the covariance or correlation structure of multivariate data. Other scaling methods include range scaling, Pareto scaling and vast scaling. Pareto scaling is very similar to standardisation, but uses the square root of the standard deviation as the scaling factor instead of the standard deviation itself. Vast scaling is an extension of standardisation. The details of all these methods are given below.

1. **Centring** Generally, centring pre-treatment allows the researcher to focus on the differences not the similarities in the data. The focus is on isolating and removing systematic variation in the data. However, care is needed when data are heteroscedastic, as the effects of centring may not be sufficient on their own. Usually, centring is applied in combination with other pre-treatment methods. It belongs to the column scaling methods (Goodacre et al., 2007). Centring converts the metabolite concentrations to fluctuate around zero instead of around the mean concentration for that metabolite (column). Therefore it is used to focus on the fluctuations in the data (Jackson (1991) and Bro and Smilde (2003)), leaving only the relevant variation (between the samples) for analysis. Centring can be applied on its own or as part of any of the methods described below.

2. **Scaling based on Data Dispersion**
   Several scaling techniques were tested that use a dispersion measure as a scaling factor. These also

belong to the column-scaling techniques, as the scaling is applied to each column of the data set (van den Berg et al., 2006). We consider standardisation (Jackson, 1991), Pareto scaling (Eriksson et al., 1999), range scaling (Smilde et al., 2005), and vast scaling (Keun et al., 2003).

In these methods, the mean and standard deviation are defined as:

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \qquad \text{and} \quad s_j = \sqrt{\frac{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2}{N-1}}, \tag{1}$$

where i and j index the data rows and columns respectively, and N is the number of rows.

Standardisation

This is a form of scaling performed by mean centring each metabolite value using the respective mean of all sample values for that metabolite, and then dividing by the standard deviation of all the sample values for that metabolite. The formula is given by

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}. \tag{2}$$

Standardisation is also called auto-scaling or unit variance scaling, as, after standardisation, all metabolites have a standard deviation of one and have comparable scales. The main advantage is that all metabolites become equally important, but this approach can increase the influence of measurement errors (van den Berg et al., 2006). After standardisation, the data becomes dimensionless.

Range Scaling

The scaling factor in the range scaling method is the range within each metabolite. In this case, the formula is

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{x_{j_{max}} - x_{j_{min}}} \tag{3}$$

referring to metabolite j. Range scaling allows comparison of metabolites with respect to their biological response range. In this approach, all metabolites are equally important, and their scaling is related to the biology. However, increased measurement errors and sensitivity to outliers may be noticed when using this scaling method (van den Berg et al., 2006). As in the case of standardisation, the data becomes dimensionless.

Pareto Scaling

Here the square root of the standard deviation is used as the scaling factor. It aims to reduce the influence of large values without losing important information concerning the structure of the data. The formula is:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j}}. \tag{4}$$

Pareto scaled data is closer to the original data than standardised data, but this depends very much on the large values in the data set.

Vast Scaling

This is an extension of standardisation. It aims to give more importance to those metabolites that appear to have small variances. To achieve that, the method uses the coefficient of variation statistic as a scaling factor. The formula is given by:

$$\tilde{x}_{ij} = \frac{(x_{ij} - \bar{x}_j)}{s_j} \cdot \frac{\bar{x}_j}{s_j}, \tag{5}$$

where $\frac{\bar{x}_j}{s_j}$ is the inverse of the coefficient of variation of the jth metabolite (column j). This method is not useful when large induced variation exists, and there is no group structure in the data.

All of these scaling methods belong to the column scaling methods, as the scaling is applied to the columns (metabolites) in the data set.

3. **Scaling based on Data Level**

   Level Scaling

   Scaling based on average values uses a size measure instead of a measure of spread as the scaling factor. Level scaling is one such method. It converts metabolite concentrations into changes relative to the size of the average concentration of the metabolite, by using the mean concentration as the scaling factor. The formula for level scaling of metabolite j is given by:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\bar{x}_j}. \tag{6}$$

   This method is suitable for the identification of biomarkers. It is however prone to increase measurement errors (van den Berg et al., 2006). Level scaling, like the scaling methods based on data dispersion, also belongs to the column scaling methods.

2.4 Principal component analysis (PCA)

Metabolomics data records information on many compounds. To account for all of these manageably in analysis, we must reduce the multidimensional information to a lower-dimensional space. Typically two or three dimensions are used for visualisation of data. One way to achieve dimension reduction is by using principal component analysis (PCA; Jackson, 1991). PCA is applied commonly in widely differing areas from neuroscience to computer graphics, because it is relatively simple and effective.

PCA is one of several multivariate statistics techniques that can be used to visualise clusters and patterns in data, by plotting it in lower-dimensional space, and this space is defined in PCA by the first few principal component (PC) vectors.

PCA is a mathematical process that uses orthogonal transformations to convert a set of observations of correlated variables to a set of values of uncorrelated variables (principal components (PCs); Vandeginste et al., 1998). It describes the main directions of variation in a data set. The PCs are mutually orthogonal vectors consisting of linear combinations of the input variables, ordered according to the amount of variation which each PC represents, so that the first few PCs can be used as derived variables to represent a high dimensional space by a lower-dimensional space accounting for most of the variation in the original data space (McVean, 2009). These first few PCs can also be used to plot the data.

The number of principal components is determined by the rank of the data matrix, which is less than or equal to the smaller of the number of rows and columns in the data. In the metabolomics context this is likely to be less than or equal to the number of samples in the original data. In this study, the number of PCs is equal to the number of samples.

PCA is sensitive to the scaling of the original variables, so is commonly used on the correlation matrix of the variables or the covariance matrix of scaled variables rather than the covariances of the original data. This overcomes the effect of possibly widely differing variances in the original variables, even when those variables may be measured on the same scale. Not doing so leads to more dispersed variables dominating the analysis.

Each PC is characterised by two sets of information, the scores and the loadings. Scores plots often give useful information about the relationships between the samples (rows in the data). These plots can

be produced as the projections of the samples to a single eigenvector (PC) versus sample number or onto the plane formed by the first two eigenvectors (first two PCs). A projection of the samples to the two eigenvectors associated with the largest eigenvalues shows the greatest amount of information about the relationship between the samples that can be shown in two (linear) dimensions. The original space here consists of metabolite expression profiles. Scatter plots of the first two or three PCs will reflect most of the information in the original data set of higher dimension, and are a useful way to see any clusters of groups of similar observations in the data.

The eigenvectors are found from an eigenanalysis of the covariance matrix or the correlation matrix of the data matrix X or a singular value decomposition of X itself. PCA decomposes the variation of data matrix X into scores T, loadings P, and a residuals matrix E, where P is a $I \times A$ matrix containing the A selected loadings and T is a $J \times A$ matrix containing the accompanying scores.

$$X = PT^T + E, \tag{7}$$

where $PT^TP = I$, the identity matrix. The scores represent the values of the original observations in the new space of variables defined by the PCs, and the loadings give the coefficients of each original variable (centred and/or scaled, as appropriate) in the PCs (van den Berg et al., 2006).

3 Results

All calculations were performed using the *R* software (R Core Team, 2019). One data matrix X was input to the *R* software for each set of propolis samples, after pre-processing of the data. Here matrix X is MS data in which row labels are sample identifiers and columns are variables. After pre-processing, we apply scaling techniques on the different data sets contained in each matrix X. Our goal is to carry out PCA on these data, therefore the optimal scaling technique is considered as the one for which the maximum variance is explained by the first few orthogonal principal components. Therefore, PCA is performed to evaluate this, as an objective numerical measure to evaluate the methods, and PC scores plots and loadings plots are also presented, to compare any patterns of clustering arising after the different scaling methods are used.

3.1 Application of pre-treatment

As described, pre-treatment takes place in the second stage of data processing, after pre-processing, to remove or reduce any uninduced variation (from sampling, sample work-up and analytical measurement errors) as much as possible, and, if it exists, heteroscedasticity of the data. The usual order of applying pre-treatment methods to a data set is to first transform the individual elements of the data set, then to apply row scaling (if used), and finally to scale the columns, as was done here. Scaling methods (row or column) have been classified as centring, scaling based on data dispersion and scaling based on average values. The two most common methods of transforming the elements of a data matrix are the log and the power transformations, and several commonly used scaling approaches have been described.

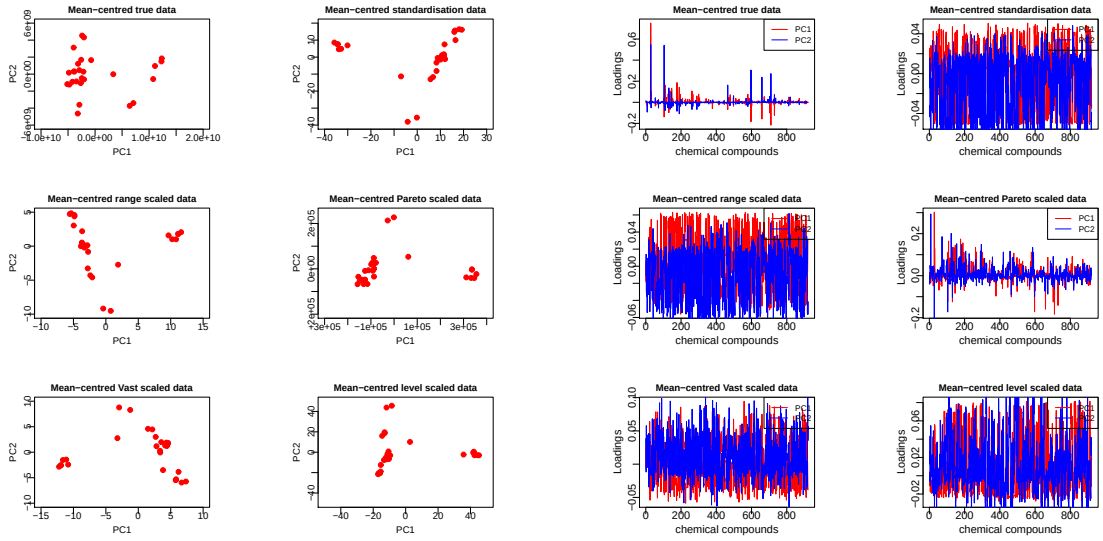
No element transformation was chosen in this case. Using a power transformation had very little effect. Although using a log transformation brought the data much closer to normality, it also led to much less interpretable results from PCA, for each of the data sets I, II, III and Libya. Row scaling was also found to be unnecessary (results not shown).

Figures 4, 5, 6 and 7, respectively, show the scores and loadings plots for the propolis samples in data sets I, II, III (Aberdeenshire, Fort William and Dunblane) and Libya. We can compare these across

methods to examine the effects of column scaling on the PCA scores and loadings from the different data sets. We are looking for patterns of clustering of samples in the scores plots, and for important metabolites to be highlighted as spikes in the loadings plots.

The scores plots in Figures 4 to 6 (left plots) indicate that there are some differences among the scores of the six scaling methods used on data sets I, II and III, for Scotland. Concerning the loadings plots in Figures 4 to 6 (right plots), the loadings on PC1 and PC2 for standardisation, range, vast and level scaling have a similar shape. For the other two scaling methods, the mean-centred true (raw) data and Pareto, the plots of the loadings on both PCs have similar shapes. In general, the shapes of loadings for the mean-centred true and Pareto approaches have the highest similarity among all the plotted loadings, and the other approaches give different results from these but similar results to each other. Similar conclusions can be drawn from Figure 7 for the Libyan data. These conclusions also apply to the scores plots, where the results for the mean-centred true (raw) data and Pareto are similar but different from the other results.

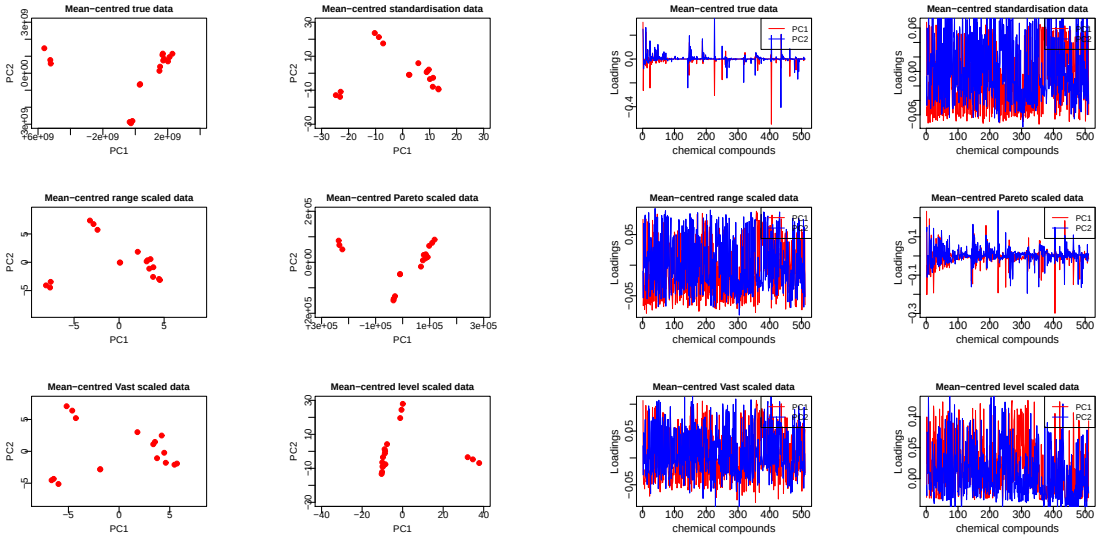
In summary, the application of different pre-treatment methods had a large effect on the resulting data used as input for data analysis, as shown by the varying effects seen in Figures 4 to 7. For instance, standardisation, range, vast and level scaling showed very many large peaks in the loadings, while after Pareto scaling (or mean centring alone) relatively fewer large peaks were present, thus giving more interpretable results indicating more important metabolites. While the results for the mean-centred true data and Pareto-scaled data are similar, Pareto scaling identifies more peaks than mean centring alone, while still being interpretable. It is clear that different results will be obtained when the differently pre-treated data sets are used as input for data analysis.



(a) scores plots (b) loadings plots

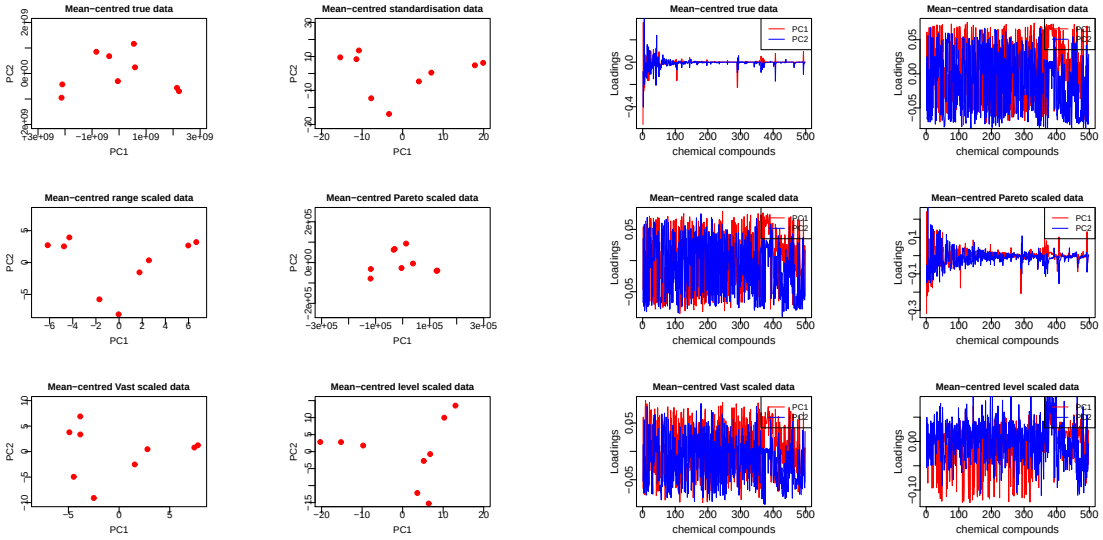
Figure 4: Plots of PC1 vs PC2 scores and loadings for the scaled Aberdeenshire data.

PCA constructs orthogonal uncorrelated linear combinations that explain as much common variation in the data as possible. Tables 1, 2, 3 and 4 show the percentage of variation explained by the first two PCs, used as an objective numerical comparison between all the scaling methods except for mean centring alone, for PCA of data sets I, II, III and Libya. Each of these scaling methods includes mean centring. From these tables, it can be observed that Pareto scaling performed much better than the other scaling methods in terms of PCA, because it explains more of the variation in the data sets in every case. The



(a) scores plots (b) loadings plots

Figure 5: Plots of PC1 vs PC2 scores and loadings for the scaled Fort William data.

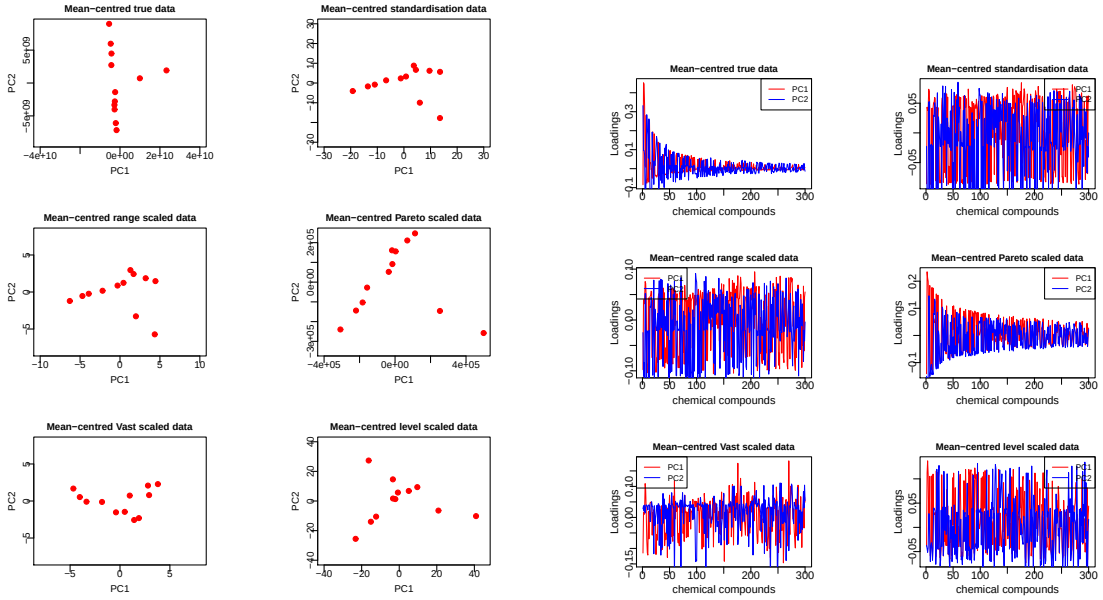


(a) scores plots (b) loadings plots

Figure 6: Plots of PC1 vs PC2 scores and loadings for the scaled Dunblane data.

first two PCs explain 79.00%, 74.78%, 76.91% and 71.79% of the total variation of data sets I, II, III and Libya respectively.

We conclude that Pareto scaling (incorporating mean centring) is the most useful type of scaling for each of these samples. We see clearly from Tables 1 to 4 that the Pareto option leads to much the most variation in data sets I, II, III and Libya being explained by the first two PCs, and therefore gives the most informative lower-dimensional analysis of the data.



(a) scores plots (b) loadings plots

Figure 7: Plots of PC1 vs PC2 scores and loadings for the scaled Libya data.

Pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
proportion variance of PC1	40.18	47.21	68.90	46.28	38.34
proportion variance of PC2	19.48	17.00	10.10	15.66	19.93
Cumulative Proportion	59.66	64.21	**79.00**	61.94	58.27

Table 1: Percentage of variance explained by the first two PCs of the Aberdeenshire data; best results are shown in bold, and similarly below.

Pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
proportion variance of PC1	35.68	38.65	52.73	36.58	41.30
proportion variance of PC2	25.64	25.41	22.05	23.57	21.99
Cumulative Proportion	61.32	64.06	**74.78**	60.15	63.29

Table 2: Percentage of variance explained by the first two PCs of the Fort William data.

Pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
proportion variance of PC1	33.47	35.45	53.38	32.81	43.01
proportion variance of PC2	30.33	30.67	23.53	30.16	26.03
Cumulative Proportion	63.80	66.12	**76.91**	62.97	69.04

Table 3: Percentage of variance explained by the first two PCs of the Dunblane data.

pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
proportion variance of PC1	37.87	40.24	54.87	47.16	33.46
proportion variance of PC2	19.55	19.58	16.91	15.80	21.80
Cumulative Proportion	57.42	59.82	**71.79**	62.68	55.25

Table 4: Percentage of variance explained by the first two PCs of the Libya data.

4 SUMMARY AND CONCLUSIONS

In this article we have given a short review of pre-treatment methods used in processing metabolomics data, including transformation and scaling methods. There are two approaches that can be applied for data scaling: row scaling, and column scaling, with different scaling methods available for either. Scaling methods can be classified as centring, scaling based on data dispersion and scaling based on average values. The advantages and disadvantages of applying these pre-treatment methods were discussed. Column scaling makes the columns more comparable to each other for subsequent analysis, and makes sense in the context of metabolomics data where the columns represent different metabolites detected in the samples analysed.

We considered mean centring alone, and standardisation, range scaling, Pareto scaling, vast scaling and level scaling, each including mean centring. We applied these six different column scaling methods to four different metabolomics data sets arising from MS analysis of samples of honey bee propolis from different locations in Scotland and Libya, to compare the effects of these on the results of principal component analysis (PCA) of the treated data. We examined PC scores plots and loadings plots, for graphical representations of the data in terms of the first two PCs, as well as the percentage of variance explained by the first two principal components.

We conclude from the results that it is best for these data sets to be mean-centred and Pareto-scaled prior to using PCA, where these operations are carried out on the columns of the data (metabolites). This approach led to much the highest percentage of variance being explained by the PCA, relative to other scaling approaches, for every data set. We did not find it necessary to first scale the rows of the data to a constant total, nor to carry out data transformation first, as this had little effect on the data used here.

These conclusions are likely to be true of other similar data sets as well, so we recommend these choices for analysis of such MS-based metabolomics data, notably mean centring and Pareto scaling of the data columns. Mortazavi-Tabatabaei et al. (2013), in a less extensive study, concluded, using NMR metabolomics data from human blood samples, that mean centring separated two different patient groups more clearly than auto-scaling, but did not examine Pareto scaling. The authors in van den Berg et al. (2006) studied the effects of several pre-treatment methods and concluded that the pre-treatment approach crucially affected the outcome of the data analysis, for functional genomics data. Therefore, what is the best approach may depend on the context, as also concluded by Craig et al. (2006). It would be interesting to carry out similar analyses on a wider variety of such data sets.

Acknowledgements

The authors are very grateful to the beekeepers providing the propolis samples, including Magnus Peterson, Willow Lohr and Sarah Kennedy in Scotland. They are also pleased to acknowledge the support and funding of a scholarship from King Abdulaziz University and the Ministry of Education in Saudi Arabia, given to Abdulaziz Alghamdi.

References

[1] J. Boccard, J. Veuthey and S. Rudaz, Knowledge discovery in metabolomics: An overview of MS data handling, Metabolomics 3(3) (2010), 231-241.

[2] R. Brereton, Chemometrics for Pattern Recognition. John Wiley and Sons, West Sussex, UK, 2009.

[3] R. Bro and E. K. Smilde, Centering and scaling in component analysis, J. Chemom. 17 (2003), 16-33.

[4] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson and J. C. Lindon, Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Anal. Chem. 78 (2006), 2262-2267.

[5] L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS), Umetrics, Umea, Sweden, 1999, pp. 213-225.

[6] R. Goodacre, D. Broadhurst, A. K. Smilde, B. S. Kristal, J. D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjostrom, J. Trygg and F. Wulfert, Proposed minimum reporting standards for data analysis in metabolomics, Metabolomics, 3(3) (2007), 231-241.

[7] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrometry 45(7)(2010), 703-714.

[8] J. E. Jackson, A User's Guide to Principal Components. John Wiley and Sons, Inc.,New York, 1991.

[9] H. C. Keun, T. M. D. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, E. Holmes, J. C. Lindon and J. K. Nicholson, Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. Anal. Chim. Acta, 490 (2003), 265-276.

[10] G. McVean, A genealogical interpretation of principal components analysis. PLoS Genetics, 5(10) (2009), e1000686.

[11] S. A. Mortazavi-Tabatabaei, F. Fathi, F. Ektefa, M. Tafazzoli, A. A. Oskouie, M. Rezaie-Tavirani, M. R. Zali, M. R. Nejad and K. Rostami, Investigation of metabonomics technique by analyze of NMR data, which method is better? Mean center or auto scale? J. Paramedical Sciences 4 (1) (2013), ISSN 2008-4978.

[12] J. Nicholson, J. Lindon and E. Holmes. Metabolomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29 (1999), 1181-1189.

[13] J. Nicholson and I. Wilson, Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat. Rev. Drug Discov. 2 (2003), 668-676.

[14] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/

[15] K. Saleh, T. Zhang, J. Fearnley and D. G. Watson, A comparison of the constituents of propolis from different regions of the United Kingdom by Liquid Chromatography-high Resolution Mass Spectrometry using a metabolomics approach. Current Metabolomics, 3 (2015), 42-53.

[16] W. Siheri, T. Zhang, G. U. Ebiloma, M. Biddau, N. Woods, M. Y. Hussain, C. J. Clements, J. Fearnley, R. E. Ebel, T. Paget, S. Muller, K. C. Carter, V. A. Ferro, H. P. De Koning and D. G. Watson, Chemical and antimicrobial profiling of propolis from different regions within Libya. PLoS ONE 11(5)(2016), e0155355. doi:10.1371/journal.pone.0155355

[17] M. Simone-Finstrom and M. Spivak, Propolis and bee health: the natural history and significance of resin use by honey bees. Apidologie 41 (3) (2010), 295-311.

[18] M. D. Simone-Finstrom and M. Spivak, Increased resin collection after parasite challenge: a case of self-medication in honey bees? PloS ONE 7 (3) (2012), e34601. doi:10.1371/journal.pone.0034601
[19] A. K. Smilde, M. J. Van der Werf, S. Bijlsma, B. J. C. van der Werff-van der Vat and R. H. Jellema, Fusion of mass-spectrometry-based metabolomics data. Anal. Chem. 77 (2005), 6729-6736.
[20] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. D. Jong, P. J. Lewi and J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics. Elsevier, Amsterdam, 1998.
[21] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde and M. J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7(1) (2006), 142.