# Chapter 6
# Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis

**Ibrahim Karaman**

**Abstract** From data acquisition to statistical analysis, metabolomics data need to undergo several processing steps, which are crucial for the data quality and interpretation of the results. In this chapter, methods for preprocessing, normalization, and pretreatment of metabolomics data generated from nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) are presented and discussed. Preprocessing is reported for both NMR and MS analysis. The challenges in preprocessing such complex data are highlighted. Subsequently, normalization methods such as total area normalization, probabilistic quotient normalization, and quantile normalization are explained. Finally, several scaling and data transformation methods are discussed for metabolomics data pretreatment, which is an important step prior to statistical analysis.

**Keywords** Preprocessing • Alignment • Normalization • Pretreatment • Scaling • Transformation

I. Karaman
Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, St. Mary's Campus, Norfolk Place, W2 1PG London, UK
e-mail: i.karaman@imperial.ac.uk

© Springer International Publishing AG 2017

A. Sussulini (ed.), *Metabolomics: From Fundamentals to Clinical Applications*, Advances in Experimental Medicine and Biology,
DOI 10.1007/978-3-319-47656-8_6

## Abbreviations

| | |
|---|---|
| ANOVA | Analysis of variance |
| CPMG | Carr-Purcell-Meiboom-Gill |
| GC | Gas chromatography |
| glog | Generalized log |
| LC | Liquid chromatography |
| LOESS | Locally estimated smoothing |
| *m/z* | Mass-to-charge ratio |
| MS | Mass spectrometry |
| NMR | Nuclear magnetic resonance spectroscopy |
| PCA | Principal component analysis |
| PLSR | Partial least squares regression |
| $R^2$ | Linear regression coefficient |
| RSD | Relative standard deviation |
| RT | Retention time |
| QCs | Quality control samples |
| TSP | 3-trimethylsilylpropionic acid |

## 6.1 Introduction

Metabolomics analysis in clinical applications is often performed by either NMR or LC/GC-MS [1–4]. These platforms generate highly complex high-throughput data when biofluids are analyzed. NMR is a non-destructive and reproducible technique because the sample and the instrument do not physically interact [5]. In contrast, LC/GC-MS are destructive and less reproducible techniques [6]. However, LC/GC-MS have higher sensitivity compared to NMR [7]. Advanced technologies in the instrumentation offer fast and inexpensive solutions for metabolomics analysis and provide the opportunity of analyzing more than a thousand samples in an experimental run [8, 9]. The complexity of the data escalates, and consequently the data must go through various preprocessing and quality control steps prior to statistical analysis. There are many available methods and softwares for preprocessing, and new methods are developed or novel softwares are released constantly. The recent methodologies in omics data preprocessing can be followed via online web-based tools [10]. In Fig. 6.1, a generalized workflow for metabolomics data analysis is shown, considering the steps from data acquisition to statistical analysis. After the samples are analyzed by the instrument, the raw data of each sample need to be converted and processed in order to be summarized in a data table. The rows and the columns must be as comparable as possible after all the processing steps. This chapter focuses on the three blocks at the middle of this workflow and aims to give better understanding to the reader about the various steps of preprocessing and pretreatment. The following sections are fashioned according to Fig. 6.1 as Sects. 6.2, 6.3, 6.4, and 6.5.

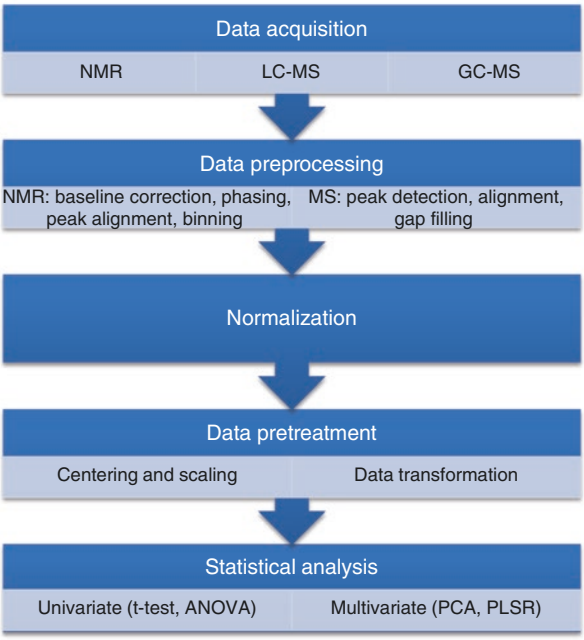


**Fig. 6.1** General processing steps of metabolomics data analysis, from data acquisition to statistical analysis

## 6.2 Preprocessing of LC/GC-MS Data

In MS-based analyses, the measured variables are mass-to-charge ratios (*m/z*). When MS is combined with LC or GC, an additional dimension is added to the variable space, which is the chromatographic retention time. Therefore, raw LC/GC-MS data consist of a 3D structure of *m/z*, retention time (RT), and intensity count. In Fig. 6.2, an LC-MS profile of a blood serum sample is demonstrated in a specific retention time interval. Raw LC-MS data have many data points in one sample, as the data are often acquired in high-resolution instruments. Most of the data are generally either spectral noise or not biologically relevant (column material, contaminants, etc.). Therefore, it is necessary to convert each 2D sample profile into a 1D vector of peak areas/intensities.

The aim of preprocessing is to generate a 2D data table of features where the rows correspond to the study samples and the columns to *m/z*-RT pairs. There are several preprocessing steps in order to achieve this, and various softwares are available to perform the preprocessing, such as MarkerLynx (Waters), MassHunter (Agilent), MarkerView (AB Sciex), XCMS [11], MZmine 2 [12], and Progenesis QI (Waters). The LC/GC-MS data preprocessing steps are:

(a) *Peak picking/detection and deconvolution*: Peak picking is a crucial step of the preprocessing pipeline. It aims to detect each measured ion in a sample and to

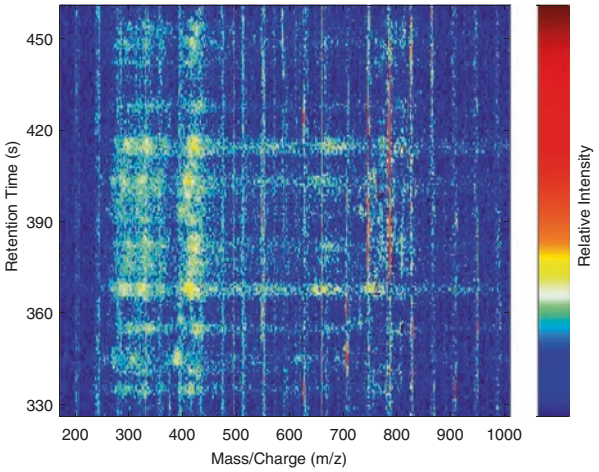


**Fig. 6.2** Representative LC-MS profile of blood serum in a specific retention time interval for better visualization

assign to a feature (*m/z*-RT pair). In this step, the peak picking algorithm captures and deconvolutes peaks from the extracted ion chromatograms taking possible baseline and noise structures into account. If necessary, smoothing such as moving average or Savitzky-Golay filters can be applied during this step.

(b) *Alignment*: Improvements in the technology of mass spectrometers provide good reproducibility in the *m/z* dimension; however, reproducibility may be a problem in the RT dimension especially for LC-MS experiments. During chromatographic separation, RT shifts can occur due to changes in the mobile phase and the column stationary phase, variations in temperature and pressure, column aging, or effects related to sample matrix. Therefore, a metabolite can be eluted in slightly different retention times across the samples. This problem is crucial when hundreds or thousands of samples are analyzed in a long experimental run. Alignment algorithm aims to group detected peaks across the samples with respect to a *m/z* and a RT window. The grouped peaks are subsequently integrated as peak height or peak area and assigned to a feature in the data table.

(c) *Gap filling*: The data table after peak picking and alignment will contain missing values (gaps) in some of the samples. The reason of the presence of missing values is generally the existence of badly shaped peaks, which can be missed during the peak picking process, and peaks with low intensity, which cannot be detected during the peak picking process. Some of the preprocessing algorithms have gap-filling algorithms where peak structures are searched in the raw data on the defined *m/z* and RT window. This approach is useful when large peaks are missed during peak picking. There are also missing value estimation methods in literature [13], such as k-nearest neighbor imputation method. Care must be taken when using such methods because the imputations are based on the complete part of the whole data, which may not be the best representation for imputing the missing values.

After the initial preprocessing steps, the data table is complete without missing values. The next step is to assess the quality of the features in the data table. For metabolomics studies, it is recommended to analyze quality control samples (QCs) after every couple of (between 5 and 10) study samples in the entire sample run in order to monitor the experiment [14, 15]. The QCs are prepared by pooling the study samples; therefore, they represent the whole sample set. By looking at the QCs, it is possible to assess each feature in the data table for:

(a) *Presence in the QCs*: Some preprocessing softwares provide the number of samples, which are present in a predefined sample group (the QCs in this case). Features that are not present in a certain number of QCs can be filtered out from the data table. This filtering step assumes the sample set is well represented by the QCs.
(b) *Intensity drifts*: As data acquisition takes a significant amount of time, it is common to observe intensity drifts, which cause intra- and inter-batch variation throughout the analysis. These drifts are specific to each feature and cannot be handled by sample normalization. Therefore, each feature has to be examined separately. There are methods available to remove intensity drifts [14, 16, 17], and a common method is to fit a nonlinear locally estimated smoothing (LOESS) curve to the intensities of the QCs along the experimental run order. Thereafter, a correction factor for each study sample is estimated by interpolating the LOESS curve to the experimental run of the study samples. These correction factors are used to remove intensity drifts in each feature by dividing the intensity by the correction factor. In Fig. 6.3, the effect of the drift correction on the data is demonstrated. This drift correction step is important, and care must be taken when applying, because there should not be outliers among the QCs, and they may require normalization beforehand.
(c) *Repeatability*: Each feature in the QCs should have low relative standard deviation (RSD) across the QCs throughout the experimental run in order to have a good repeatability. RSD for each feature is calculated by dividing the sample standard deviation by the sample mean. The features with high percent RSD values should subsequently be removed from the data table. The suggested

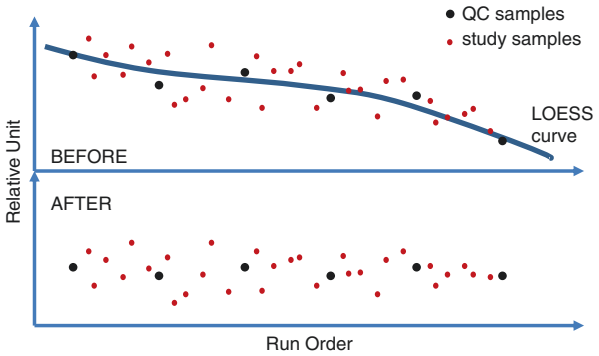


**Fig. 6.3** Graphical representation of intensity drift correction using LOESS curve

threshold is 20% for LC-MS and 30% for GC-MS [14], but it may be more flexible depending on the size of the sample set.

(d) *Linearity*: A series of QC samples with varying dilutions can be prepared and analyzed within the experimental run [18]. The dilution factors can thus be regressed against the corresponding intensities of each feature in the data table. The features with low $R^2$ and with negative beta coefficients are thus removed from the data table. An $R^2$ threshold between 0.5 and 0.7, which is not very stringent, is suggested; nonetheless, it depends on the study sample size. Inspecting the distribution of the $R^2$ values may provide help in deciding the threshold.

At the end of the preprocessing steps, the data table is generated by features of *m/z*-RT pairs after filtering based on the QCs and correcting for instrumental drifts. The columns of the data table are to the best extent made comparable for further analysis.

## 6.3 Preprocessing of $^1$H NMR Data

Preprocessing of metabolomics data acquired using $^1$H NMR is crucial and challenging in clinical studies when blood (serum/plasma) and urine samples are analyzed. In NMR metabolomics, sample spectra can be acquired by different NMR experiments, such as standard 1D $^1$H NMR experiment, 1D $^1$H Carr-Purcell-Meiboom-Gill (CPMG) spin-echo NMR experiment, and 2D $^1$H-$^1$H J-resolved NMR experiment. Each of these experiments contains water presaturation. CPMG experiment is specifically used for blood samples because it removes the broad baselines with respect to the macromolecules, such as the phospholipids and lipoproteins, in the blood.

In general, initial steps of preprocessing in $^1$H NMR experiments involve apodization, Fourier transform, phasing, baseline correction, and chemical shift calibration. These steps are currently automated by the instrument vendor software and can be applied either manually or automatically according to the scientific problem.

Figure 6.4 demonstrates representative 1D $^1$H NMR spectra for blood serum and urine after proper initial preprocessing steps. The spectral data acquisition range for these samples is δ -0.50–10.00 ppm by the instrumental setting because no bona fide metabolite signals are expected outside this region. By looking at the CPMG (Fig. 6.4a) and standard 1D (Fig. 6.4b) $^1$H NMR spectra of blood serum, the broad baselines under the sharp peaks on the standard 1D $^1$H NMR spectrum draw attention. The latter spectrum contains several broad resonances from macromolecules, and these broad resonances are highly overlapped with the low molecular mass metabolites with sharp peaks. Nevertheless, the information captured from both experiments is complementary. On the other hand, standard 1D $^1$H NMR spectrum of urine (Fig. 6.4c) exhibits numerous sharp peaks throughout the spectral range. The broad baselines are observed only locally.

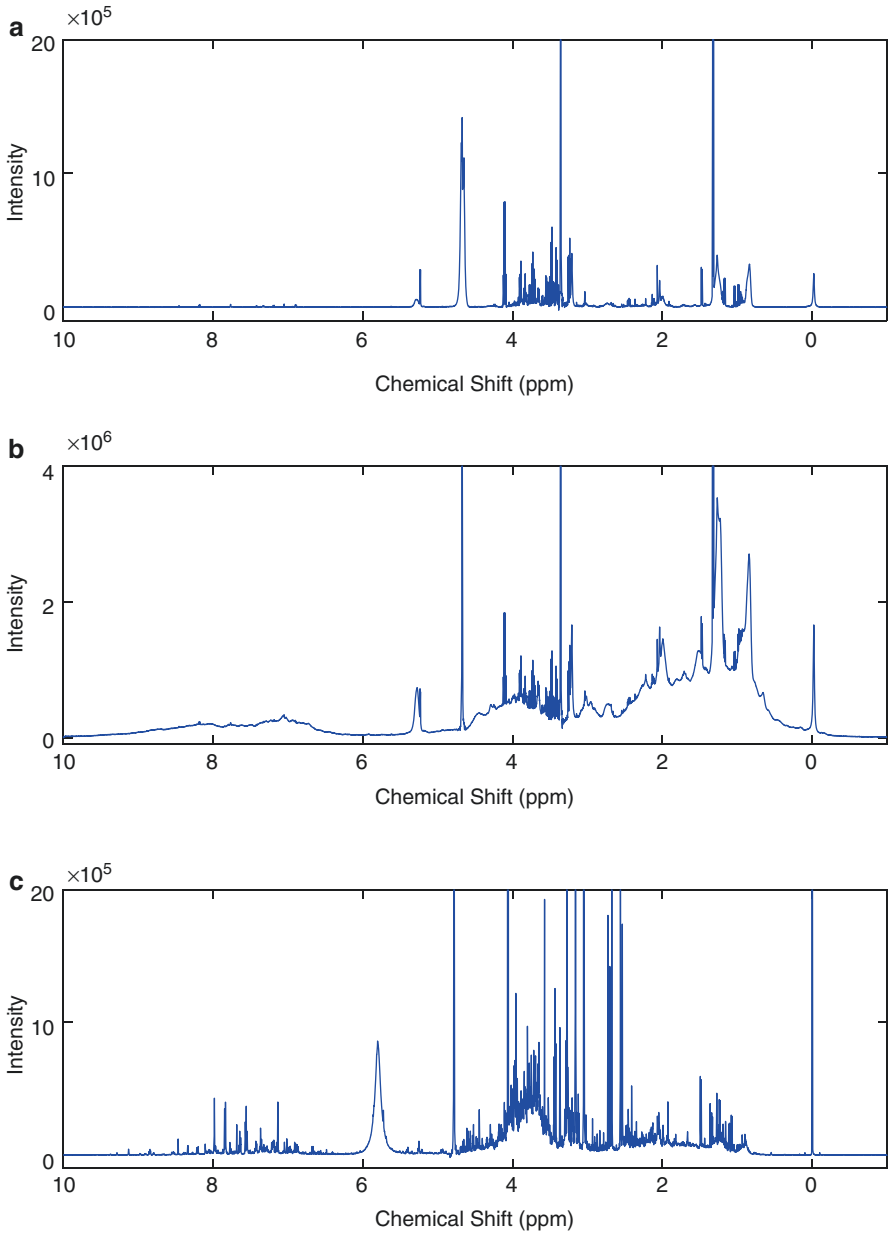


**Fig. 6.4** Representative (**a**) $^{1}$H NMR CPMG spectrum of blood serum, (**b**) $^{1}$H NMR standard 1D spectrum of blood serum, and (**c**) $^{1}$H NMR standard 1D spectrum of urine

Once all of the sample spectra are available, one can generate a data table where the rows correspond to the study samples and the columns to spectral data points. The columns of the data table should correspond to the same information after the preprocessing. During spectral data acquisition, peak shifts can be observed due to changes in pH, temperature, or fluctuations in the magnetic field. This will cause the columns of the data incomparable along the samples. Therefore, metabolite signals should be aligned and made comparable prior to statistical analysis. Below, some of the most commonly used approaches for NMR preprocessing are highlighted:

(a) *Using high-resolution $^{1}H$ NMR spectra*: The data can be analyzed as raw spectra or after applying a peak alignment algorithm. When peak shifts are systematic, they can be corrected by calibrating the spectra toward a reference peak, such as the singlet at δ 0.00 ppm due to the internal standard TSP added to every sample (see the sharp peak at δ 0.00 on the spectra in Fig. 6.4). For blood serum/plasma, the glucose doublet at δ 5.23 ppm is a suitable alternative to TSP since TSP may bind to proteins in serum/plasma samples, which will cause changes in peak shape and position. However, there may still be small but significant shifts in the peak positions between the samples. Applying peak alignment algorithms can correct shifts in the peak positions to some extent. Figure 6.5 depicts a hypothetical example of spectral data alignment. There are several algorithms available in the literature [19–21]. Most of these methods locate the position of the peaks in a sample spectrum and fit the corresponding chemical shift in a reference spectrum. Some of the common ones are icoshift and recursive segment-wise peak alignment algorithms. Both algorithms require a reference spectrum. A reference spectrum can be randomly selected, or a sample spectrum, which is the closest to the rest of the sample spectra, can be used. Alternative to using a sample spectrum as reference is creating a reference spectrum by calculating the mean or median spectrum from the entire sample set or the QCs. The major drawback of peak alignment methods is that they may not handle overlapping peaks correctly, especially when two adjacent peaks overlap

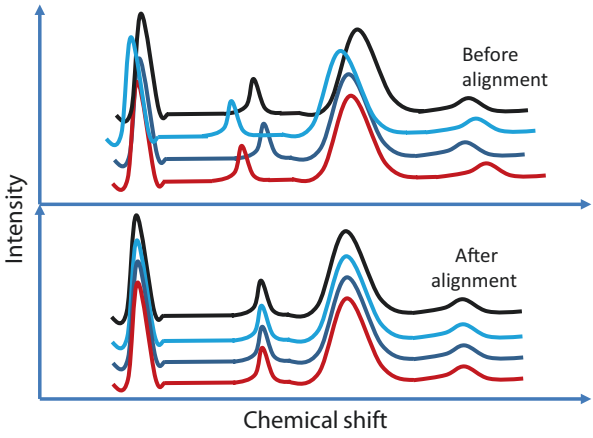


**Fig. 6.5** Schematic representation of spectral data alignment

or even swap in position between samples due to different amount of peak shifts. Nonetheless, statistical analysis of spectra after carefully applying peak alignment algorithms can be a good compromise to analyzing raw spectra.

(b) *Using binned $^1$H NMR spectra*: Binning or bucketing can be applied to raw spectra or to aligned spectra in order to correct for shifts in the peak positions on the raw spectra or small misalignments on the aligned spectra. By binning, spectral resolution is lowered by converting segments of the spectrum into a bin where the spectral data inside each segment are summed as area under the curve and represented by one single value. The data also become more compact and easy to handle computationally. There are several methods available in literature for binning [22, 23]. Although binning is attractive for lowering the resolution and handling the misalignments, care must be taken when applying. The binning method and the parameters should be selected with caution in order to obtain good data. Otherwise, peaks may fall into the wrong bins, peaks may be split, and obviously binning does not handle overlapping peaks.

(c) *Quantifying known metabolites*: When targeted metabolomics data analysis is the case, converting the spectral data table into a table of annotated and quantified metabolites is convenient even though new metabolites are not generally available for analysis. Automated quantification of metabolite levels from spectra is not an easy process due to peak overlapping, variations in peak positions, and spectral noise. In addition, building a calibration model is time consuming and requires standard sample spectra for calibration. There are methods in literature such as Bayesian automated metabolite analyzer [24] where peaks from 1D $^1$H NMR spectra are deconvoluted and assigned to specific metabolites from a known metabolite list.

As one can see above, all of the approaches have benefits and weaknesses. They aim to make the columns of the data table comparable between the samples. It is up to the analyst to decide how to proceed with the preprocessing, keeping in mind the various consequences. Assuming the columns of the data table are comparable after the preprocessing steps, the analyst can move to the next step, which is the normalization of the samples. If high-resolution or binned $^1$H NMR spectra are used for further analysis, it is important to remove interfering spectral regions related to water suppression residual (δ 4.40–5.00 ppm) and possible contaminants in the samples. Peak of urea at around δ 5.80 ppm should be also removed from urine spectra. The reason is that they are not changing proportionally with the changes in concentration, and they may adversely influence the normalization of the samples.

## 6.4 Normalization

The term normalization here is used for the division of each row of data table by a normalization factor. Normalization procedure removes unwanted variation between the samples and allows quantitative comparison of the samples. In metabolomics,

study samples are biofluids in most cases, and they exhibit differences in the concentration of metabolites due to varying dilution factors for different samples. For example, metabolite concentrations in different urine samples may differ with respect to the amount of water as the solvent. Therefore, the measured metabolite concentrations will reflect to dilution instead of the changes in metabolic responses. In order to remove such variations between the samples, a normalization factor should be computed for each row of the data table. There are several ways for performing normalization [25–29]:

(a) *Addition of internal/external standard(s)*: A standard with known concentration can be added to every sample, and the samples can be normalized using the peak area/intensity of this standard. However, this method is not convenient for untargeted metabolomics because the source of unwanted variation is not only related to sample introduction to the instrument but also the variations in the dilution factors.
(b) *Total area normalization*: The normalization factor for each sample is computed by summing all of the features in the corresponding row. The disadvantage of this normalization is that changes in metabolite concentrations across the samples will affect the normalization factor because the technique assumes the total metabolite concentration in a sample does not change across the samples. High-concentration metabolites contribute to the total area, i.e., the normalization factor, more than the small-concentration metabolites. In the presence of a significant change in the peak intensity/area of a high-concentration metabolite, the normalization factor will be affected.
(c) *Probabilistic Quotient Normalization*: This method assumes that metabolite peaks affected by dilution will have the same fold changes between two samples. Fold changes of a sample are computed for every feature against a target spectrum/profile, which can be the median spectrum/profile. The normalization factor for that sample is the median value of the fold changes. This method is not affected by large changes in a few metabolites because it uses the median of many fold change values instead of an estimated single sum as for total area normalization.
(d) *Quantile normalization*: This method forces all samples in a sample set to have identical peak intensity/area distribution. The difference of quantile normalization from the previous ones is that there is no estimated normalization factor for each sample. First, each row of the data table is sorted from lowest to highest. Thereafter, mean/median of each column is calculated from the sorted data table. These mean/median values form the target spectrum/profile. All rows of the data table are replaced with the target spectrum/profile. Finally, the data table is restored into its original order before sorting. The rows of the new data table are composed by the normalized samples. This method can be problematic with high-value features in the data table because they can dramatically differ from sample to sample.

## 6.5 Data Pretreatment

When clean and normalized metabolomics data are ready for statistical analysis, it is important to use the appropriate data pretreatment method before starting [30, 31]. The data are converted into different forms by data pretreatment. The effects of technical and measurement errors are aimed to be reduced, whereas the relevant biological variations are aimed to be enhanced.

The choice of data pretreatment method depends on the scientific question and the data analysis method to be used. If univariate analysis is used, generally there is no need for a pretreatment. However, when multivariate analysis methods are considered, data pretreatment plays an important role in obtaining and interpreting the results. In Sects. 6.5.1 and 6.5.2, ways for data pretreatment are explained with a few example methods. A publically available exemplar LC-MS data set [32] was used for demonstrating the outcome of each data pretreatment method described. The preprocessed data table consists of 28 rows/samples and 168 columns/features. Although metabolomics data from clinical applications contain thousands of samples and features, the exemplar data set used here is sufficient for the readers to understand how data pretreatment works.

### *6.5.1 Centering and Scaling*

In untargeted metabolomics studies with a purpose of biomarker discovery, multivariate analysis techniques based on latent variable projections such as PCA or PLSR are used. Such methods extract information from the data by projecting onto the direction of the maximum variance. Analyzing the data from NMR and MS platforms directly by latent variable projection techniques will focus on the average spectrum/profile, and any type of biological variation in the data will be masked. Therefore, mean-centering the data table, where the mean of a feature is subtracted from each element of the feature vector, is a common practice before PCA and PLSR, and generally it is applied by default. By mean centering, it is aimed to remove the offset from the data and focus on the biological variation, as well as similarities/dissimilarities among the samples in the data.

Metabolites that are more abundant will exhibit high values in the data table and subsequently show large differences among samples compared to the low-abundant metabolites. NMR and MS platforms are effective in quantifying low-abundant metabolites, as well as the highly abundant metabolites. As PCA and PLSR are focusing on the maximum variance, centering the data alone may not be enough to find biomarkers because the highly abundant metabolites will dominantly contribute to the model. The biologically important but low-abundant metabolites thus can be masked, and the results of the statistical analysis may become biased. Consequently, scaling each feature in the data table, which potentially corresponds

to a metabolite, needs to be carefully considered. In the following, the scaling operations are explained for one feature, i.e., column, in the whole set of features in the data set.

(a) *Auto-scaling (unit variance scaling)*: The mean and the standard deviation of the feature are calculated. The feature is first mean-centered. Thereafter each element in the mean-centered feature is divided by the standard deviation. The aim of auto-scaling is to give equal weights to all of the features. Therefore, metabolites with both low and high abundance will equally contribute to the multivariate model. The drawback of auto-scaling is that noisy and uninformative features will also be as important as the interesting features. Moreover, the measurement errors on the metabolites with low abundance will inflate as they are more affected. One needs to make sure that the features in the data table have good quality, i.e., noisy features or features with low repeatability/linearity are filtered in case of analyzing MS data. When NMR data analysis is considered, auto-scaling may be better used after removing noisy and outlying/contaminant regions from the spectra. Auto-scaling can be also useful when multivariate analysis is combined with variable selection.
(b) *Pareto scaling*: This is similar to auto-scaling but in this case, each element in the mean-centered feature is divided by the square root of the standard deviation. Pareto scaling is a compromise between mean-centering and auto-scaling because Pareto-scaled metabolites with high abundance are less dominant compared to the corresponding mean-centered ones. Nonetheless, the Pareto-scaled data are kept closer to the mean-centered data, and the drawbacks of using only mean-centering count also for Pareto scaling. Therefore, multivariate analysis may be still prone to focus on the metabolites with high abundance.
(c) *Range scaling*: The mean and the range of the feature are calculated. The range is defined as the difference between the minimum and the maximum values in the feature. Each element in the mean-centered feature is divided by the range in range scaling. Using the range as the scaling factor is risky as it is sensitive to only a few outlying samples in a large sample set. It can still be an alternative to auto-scaling when range is estimated robustly.
(d) *Vast (variable stability) scaling*: Each element in the auto-scaled feature is divided by the coefficient of variation, which is the ratio of the standard deviation and the mean. In contrast to auto-scaling where each feature equally contributes to the statistical model, the focus falls onto the more stable features after vast scaling. The assumption here is that important metabolic features should have small coefficient of variation, i.e., relative standard deviation, so that they will be more stable.

In Fig. 6.6, the effect of centering and scaling to features in sample 17 from the exemplar LC-MS data set is depicted. In panel (a), most of the features seem to have low abundance. There are a few very highly abundant features. In panel (b), mean-centering moved the features to distribute around zero, but the same highly abundant features are still present and dominating the data. This is also visible in panel (d) after the features were Pareto-scaled even though low-abundant features were

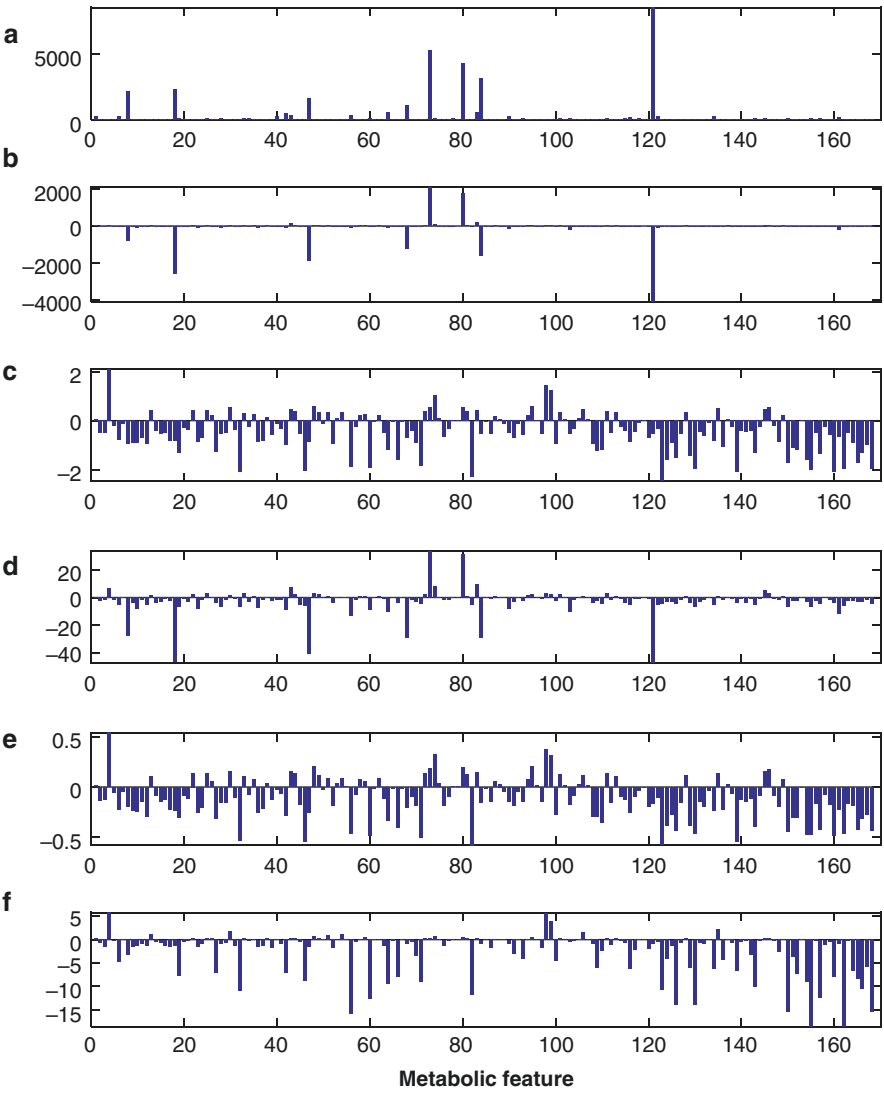


**Fig. 6.6** Graphical representation of (**a**) untreated, (**b**) mean-centered, (**c**) auto-scaled, (**d**) Pareto-scaled, (**e**) range-scaled, and (**f**) vast-scaled features of sample 17 from the exemplar data set. Y-axes of each plot were left unlabeled because they are varying with respect to the pretreatment method

inflated to some extent. On the other hand, the features seem to be more comparable with each other after auto-scaling and range scaling in panel (c) and (e). In panel (f), the vast-scaled features seem to be more comparable compared to mean-centered and Pareto-scaled data; however, care must be taken because the features with high coefficient of variation were penalized.

### 6.5.2 Data Transformation

The data from NMR and MS platforms are generally subject to heteroscedastic noise from various sources where the amount of noise increases as a function of increased signal intensity. Statistical analysis tools assume the noise is homoscedastic where the noise is consistent across all features. Therefore, the data table may need to be transformed into a form in which the noise structure is no more heteroscedastic. Furthermore, the distributions of the features can be skewed and may need to be made close to normal prior to any type of statistical analysis. Transformations aim to correct for heteroscedasticity and skewness. They also have pseudo scaling effect on the features because the differences of the features with high and low abundances are substantially diminished. Notwithstanding, it may still be necessary to apply centering and scaling after transformation. In the following, a few common transformation operations are explained for each element of the entire data table.

(a) *Log transformation*: Logarithm of each element in the feature is calculated and replaced with the original data. In case of the presence of values between 0 and 1, 1 can be added to each element in the feature before the logarithm operation. Log transformation aims to convert multiplicative noise into additive noise.
(b) *Glog transformation*: This is similar to log transformation but logarithm operation is applied to $x+\sqrt{x^2+\lambda}$ instead of $x$ directly where $x$ is the untransformed element in the data, and $\lambda$ is the transform parameter. Glog transformation can be used as a scaling method after optimizing the transform parameter using a series of technical replicate samples [33]. Therefore, only biological variation will predominantly remain in the data table after glog transformation.
(c) *Power transformation*: Square root of each element in the feature is calculated and replaced with the original data. Although it does not convert the multiplicative noise into additive noise, it has similar effects as log transformation.

In Fig. 6.7, the effect of data transformation to features from the exemplar LC-MS data set is depicted. Homoscedastic data are supposed to have a flat distribution on such plots. In panel (a), some features with high average seem to have high standard deviation as well. This means the data set is heteroscedastic and needs to be made homoscedastic by data transformation. In panel (b) and (c), log and glog transformations seem to work well on this data set because the transformed data have a flat distribution. On the other hand, the power-transformed data do not have flat distribution, as can be seen on panel (d). The reason might be the presence of multiplicative error.

## 6.6 Concluding Remarks

In this chapter, the main preprocessing steps involved in metabolomics data analysis for NMR and LC/GC-MS platforms were summarized. Descriptions for commonly used methods for each step were briefly provided with

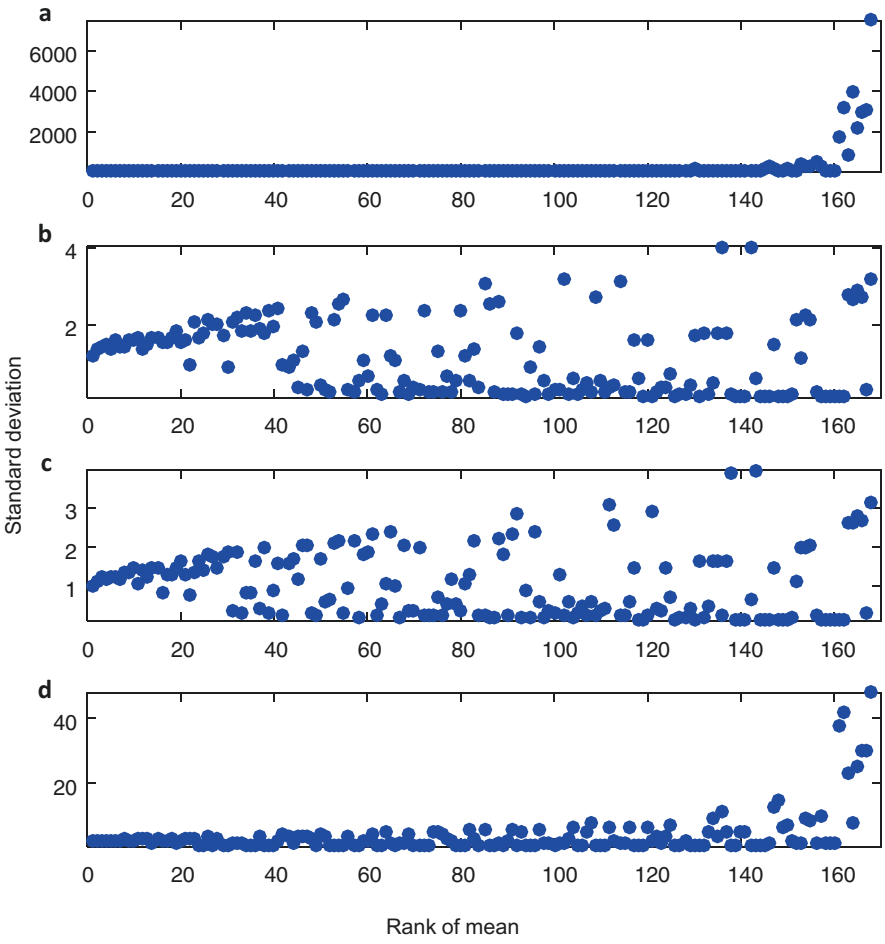


**Fig. 6.7** Graphical representation of (**a**) untreated, (**b**) log-transformed, (**c**) glog-transformed, and (**d**) power-transformed features from the exemplar data set. Transform parameter was set to 10 when glog transformation was applied

discussions on their advantages and disadvantages. The purpose of preprocessing and normalization procedures is to extract clean and comparable data across the samples from the raw data. Pretreatment aims to focus on the biologically relevant information in the data. It is important to use methods that are convenient to the data set at hand in order to remove artifacts and variation without biological importance. The choice should not be biased according to the biological question; therefore, the methods must be chosen with respect to the assumptions and the limitations of the methods.

**Acknowledgements** The author thanks Rui Pinto for helpful discussions in the preparation of this book chapter.

## References


1. Emwas A-HM, Salek RM, Griffin JL, Merzaban J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. Metabolomics. 2013;9(5):1048–72. doi:10.1007/s11306-013-0524-y.
2. Robertson DG, Watkins PB, Reily MD. Metabolomics in toxicology: preclinical and clinical applications. Toxicol Sci. 2011;120(Suppl1):S146–S70. doi:10.1093/toxsci/kfq358.
3. Vermeersch KA, Styczynski MP. Applications of metabolomics in cancer research. J Carcinog. 2013;12:9. doi:10.4103/1477-3163.113622.
4. Yin P, Xu G. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography–mass spectrometry with special emphasis in clinical applications. J Chromatogr A. 2014;1374:1–13. doi:http://dx.doi.org/10.1016/j.chroma.2014.11.050.
5. Lacy P, McKay RT, Finkel M, Karnovsky A, Woehler S, Lewis MJ, et al. Signal intensities derived from different NMR probes and parameters contribute to variations in quantification of metabolites. PLoS One. 2014;9(1):e85732. doi:10.1371/journal.pone.0085732.
6. Gika HG, Theodoridis GA, Wingate JE, Wilson ID. Within-day reproducibility of an HPLC – MS-based method for metabonomic analysis: application to human urine. J Proteome Res. 2007;6(8):3291–303. doi:10.1021/pr070183p.
7. Pan Z, Raftery D. Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. Anal Bioanal Chem. 2007;387(2):525–7. doi:10.1007/s00216-006-0687-8.
8. Lewis MR, Pearce JTM, Spagou K, Green M, Dona AC, Yuen AHY, et al. Development and application of ultra-performance liquid chromatography-TOF MS for precision large scale urinary metabolic phenotyping. Anal Chem. 2016. doi:10.1021/acs.analchem.6b01481.
9. Dona AC, Jiménez B, Schäfer H, Humpfer E, Spraul M, Lewis MR, et al. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. Anal Chem. 2014;86(19):9887–94. doi:10.1021/ac5025039.
10. Henry VJ, Bandrowski AE, Pepin A-S, Gonzalez BJ, Desfeux A. OMICtools: an informative directory for multi-omic data analysis. Database. 2014. doi:10.1093/database/bau069.
11. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78(3):779–87.
12. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf. 2010;11:395.
13. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. Metabolomics. 2012;8(1):161–74. doi:10.1007/s11306-011-0366-4.
14. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011;6(7):1060–83. doi:http://www.nature.com/nprot/journal/v6/n7/abs/nprot.2011.335.html#supplementary-information.
15. Kamleh MA, Ebbels TMD, Spagou K, Masson P, Want EJ. Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. Anal Chem. 2012;84(6):2670–7. doi:10.1021/ac202733q.
16. Fernández-Albert F, Llorach R, Garcia-Aloy M, Ziyatdinov A, Andres-Lacueva C, Perera A. Intensity drift removal in LC/MS metabolomics by common variance compensation. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu423.
17. Kirwan JA, Broadhurst DI, Davidson RL, Viant MR. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. Anal Bioanal Chem. 2013;405(15):5147–57. doi:10.1007/s00216-013-6856-7.
18. Eliasson M, Rännar S, Madsen R, Donten MA, Marsden-Edwards E, Moritz T, et al. Strategy for optimizing LC-MS data processing in metabolomics: a design of experiments approach. Anal Chem. 2012;84(15):6869–76. doi:10.1021/ac301482k.

19. Veselkov KA, Lindon JC, Ebbels TMD, Crockford D, Volynkin VV, Holmes E, et al. Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. Anal Chem. 2009;81(1):56–66. doi:10.1021/ac8011544.
20. Savorani F, Tomasi G, Engelsen SB. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. J Magn Reson. 2010;202(2):190–202. doi:http://dx.doi.org/10.1016/j.jmr.2009.11.012.
21. Wong JWH, Durante C, Cartwright HM. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. Anal Chem. 2005;77(17):5655–61. doi:10.1021/ac050619p.
22. Blaise BJ, Shintu L, Elena B, Emsley L, Dumas M-E, Toulhoat P. Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics. Anal Chem. 2009;81(15):6242–51. doi:10.1021/ac9007754.
23. Sousa SAA, Magalhães A, Ferreira MMC. Optimized bucketing for NMR spectra: Three case studies. Chemom Intell Lab Syst. 2013;122:93–102. doi:http://dx.doi.org/10.1016/j.chemolab.2013.01.006.
24. Hao J, Liebeke M, Astle W, De Iorio M, Bundy JG, Ebbels TMD. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. Nat Protoc. 2014;9(6):1416–27.
25. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem. 2006;78(13):4281–90. doi:10.1021/ac051632c.
26. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93. doi:10.1093/bioinformatics/19.2.185.
27. Veselkov KA, Vingara LK, Masson P, Robinette SL, Want E, Li JV, et al. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. Anal Chem. 2011;83(15):5864–72. doi:10.1021/ac201065j.
28. Sysi-Aho M, Katajamaa M, Yetukuri L, Orešič M. Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinf. 2007;8(1):1–17. doi:10.1186/1471-2105-8-93.
29. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem. 2006;78(2):567–74.
30. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142.
31. Bro R, Smilde AK. Centering and scaling in component analysis. J Chemom. 2003;17(1):16–33.
32. Acar E, Papalexakis EE, Gürdeniz G, Rasmussen MA, Lawaetz AJ, Nilsson M, et al. Structure-revealing data fusion. BMC Bioinf. 2014;15(1):1–17. doi:10.1186/1471-2105-15-239.
33. Parsons HM, Ludwig C, Günther UL, Viant MR. Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. BMC Bioinf. 2007;8(1):1–16. doi:10.1186/1471-2105-8-234.