Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry
Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry
Following the genomics revolution, proteomics, metabolomics, and lipidomics studies have emerged, among others, as a natural follow-up either in the investigation of human biology or, similarly, in animal or plant studies. The combined use of these omics fields may lead to a more comprehensive understanding of system biology. In medical research, it may generate novel biomarkers which may be used in early detection of disease and for the development of new screening programs. Examples are in practical clinical patient monitoring, and in the development of new patient disease management rules for complex diseases such as cancer and cardiovascular or inflammatory diseases, among others. Another key objective of such comprehensive system-level study is in the elucidation of the molecular biochemical process associated with these biomarkers. In plant biology, relevant applications may be in the prediction of desirable properties of novel foods or crops or in assisting genetic manipulation-based breeding programs for new varieties, for example.

Within the omics revolution, proteins play a key role in the study of living organisms as they provide the essential link with the genome on the one hand, while they are also key components of the physiological metabolic pathways of cells. Proteomics is therefore a fundamental research field which investigates the structure or function of protein complexes consisting of multiple proteins simultaneously. Metabolites, on the other hand, are small molecules formed from the breakdown products of larger molecules, such as proteins, after undergoing a metabolic process within an organism. They are involved in (cell) signaling processes through stimulatory and inhibitory effects on enzymes, among others. Comprehensive metabolic profiling or metabolomics can give an instantaneous snapshot of the physiology of the cell. Lipidomics is closely related to metabolomics but studies a specific set of non-water-soluble molecules consisting of glycerol and fatty acids. The collection of all lipids in an organism is referred to as lipidome, in analogy to genome, proteome, metabolome, and so on. Although all these molecules have differences in both structure and functions, they can all be studied experimentally using modern spectrometric technology—specifically mass spectrometry—to assess the required omics expression of interest.
Unlike measurement procedures and methodology in genomics research, which is reasonably standardized at the time of writing, mass spectrometry is itself a vast field with many forms and variants. Typical vehicles are time-of-flight (MALDI) mass spectrometry, liquid chromatography-mass spectrometry (LC-MS), and Fourier transform ion cyclotron resonance mass spectrometry (FTICRMS or FTMS), among others. The field is therefore still very much in flux, with many distinct mass spectrometric measurement technologies and hence also different study designs in use. The data types generated in these studies are also very different. They tend to have complex structures, while no consensus data analytical approaches have yet been agreed upon. For these reasons, expert knowledge gained with one specific measurement platform does not easily carry over to other mass spectrometric systems. Writing a comprehensive overview text on statistical data analytic methodology in the new mass spectrometry-based omics field would therefore not be realistic. Instead, we have chosen to bring together a group of established researchers to present their expert knowledge in their specific application area within this emerging field. With this book, we want to provide an overview of the current status of such mass spectrometry-based omics data analysis and give impetus to the emergence of a common view on the design and analysis of such data and experiments. In this way, the book could support the development of more standardized templates, research practices, and references for any data analyst such as statisticians, computer scientists, computational biologists, analytical chemists, and data scientists, both in the omics application fields we discuss and in related omics fields such as glycomics.

Materials presented in this book cover a broad range of topics. First, we discuss the preprocessing of mass spectrometry data such as data normalization, alignment, denoising, and peak detection including monoisotopic peaks. Second, it provides methods for identification of proteins from tandem mass spectrometry data. There is also a chapter on a software package for the analysis of such omics data. Additionally, it has chapters on downstream data analysis using Bayesian and frequentist statistical predictive modeling and classification techniques. Last but not the least, there are chapters on specific examples of biomarker detection using proteomics, metabolomics, and lipidomics data. We hope that this book will be suitable for the scientists, practitioners, software developers, as well as advanced students of statistics, computer science, computational biology, and biomedical sciences.

Gainesville, FL, USA
Leiden, The Netherlands

Susmita Datta
Bart J.A. Mertens
Contents

Transformation, Normalization, and Batch Effect in the Analysis of Mass Spectrometry Data for Omics Studies ......................... 1
Bart J.A. Mertens

Automated Alignment of Mass Spectrometry Data Using Functional Geometry ............................................................. 23
Anuj Srivastava

The Analysis of Peptide-Centric Mass-Spectrometry Data Utilizing Information About the Expected Isotope Distribution ........... 45
Tomasz Burzykowski, Jürgen Claesen, and Dirk Valkenburg

Probabilistic and Likelihood-Based Methods for Protein Identification from MS/MS Data ............................................. 65
Ryan Gill and Susmita Datta

An MCMC-MRF Algorithm for Incorporating Spatial Information in IMS Proteomic Data Processing ................................. 81
Lu Xiong and Don Hong

Mass Spectrometry Analysis Using MALDIquant ........................................ 101
Sebastian Gibb and Korbinian Strimmer

Model-Based Analysis of Quantitative Proteomics Data with Data Independent Acquisition Mass Spectrometry .............. 125
Gengbo Chen, Guo Shou Teo, Guo Ci Teo, and Hyungwon Choi

The Analysis of Human Serum Albumin Proteoforms Using Compositional Framework .................................................. 141
Shripad Sinari, Dobrin Nedelkov, Peter Reaven, and Dean Billheimer
Variability Assessment of Label-Free LC-MS Experiments for Difference Detection ............................................................. 157
Yi Zhao, Tsung-Heng Tsai, Cristina Di Poto, Lewis K. Pannell, Mahlet G. Tadesse, and Habtom W. Ressom

Statistical Approach for Biomarker Discovery Using Label-Free LC-MS Data: An Overview ............................................. 177
Caroline Truntzer and Patrick Ducoroy

Bayesian Posterior Integration for Classification of Mass Spectrometry Data ............................................................... 203
Bobbie-Jo M. Webb-Robertson, Thomas O. Metz, Katrina M. Waters, Qibin Zhang, and Marian Rewers

Logistic Regression Modeling on Mass Spectrometry Data in Proteomics Case-Control Discriminant Studies ........................... 213
Bart J.A. Mertens

Robust and Confident Predictor Selection in Metabolomics ................. 239

On the Combination of Omics Data for Prediction of Binary Outcomes . 259
Mar Rodríguez-Girondo, Alexia Kakourou, Perttu Salo, Markus Perola, Wilma E. Mesker, Rob A.E.M. Tollenaar, Jeanine Houwing-Duistermaat, and Bart J.A. Mertens

Statistical Analysis of Lipidomics Data in a Case-Control Study ........... 277
Bart J.A. Mertens, Susmita Datta, Thomas Hankemeier, Marian Beekman, and Hae-Won Uh
1 Spectrometry and Data Transformation

There is a long-standing literature on the application of transformation in spectrometry data, predating the advent of mass spectrometry. A good example may be found in the literature on infrared (IR) and near-infrared (NIR) spectrometry. An excellent recent introduction to statistical methods in this field was written by Naes et al. [13]. The log-transform has a special significance in traditional spectrometry. A crucial component of the appeal of the log-transform in (near) infrared spectrometry is due to Beer’s law (1852) [2], which states that absorbance of light in a medium is logarithmically related to the concentration of the material through which the light must pass before reaching the detector. In other words,

\[
\text{Absorbance} = \log\left(\frac{I_0}{I}\right) = kLC, 
\]

where \(I_0\) is the incident intensity of the light and \(I\) the measured intensity after passing through the medium, with \(C\) the concentration, \(L\) the path length the light travels through, and \(k\) the absorption constant. Another way to put this is that relative intensity is linearly related to concentration through the log-transform as

\[
\log\left(\frac{1}{I}\right) = \alpha + \beta C, 
\]
where the path length, initial intensity, and absorption constants are subsumed in the parameters $\alpha, \beta$. Similar formulae exist for light reflection spectrometry. The formula has been used to provide justification for the application of classical linear regression procedures with log-transformed univariate spectrometric intensity readings.

The above constitutes an argument in favor of the log-transform for spectrometry data based on non-linearity of spectral response. It is partly responsible for causing the early literature on statistical and chemometric approaches in the analysis of spectrometry data to be based on the log-transformed measures. Beer’s law applies only to (univariate) IR or NIR spectrometric readings at a single wavelength. There is no multivariate extension of the law to cover full spectra consisting of the spectrometric readings across an entire wavelength range. Nevertheless, the log-transform would also be routinely applied once truly multivariate spectrometry became available, jointly recording the spectrometric response at several wavelengths, as shown in Fig. 1 which plots 50 NIR reflectance spectra across the range 1100–2500 nm.

Log-transforming was heavily embedded in the statistical spectrometry literature, when modern laser-based mass spectrometry measurement became routinely available by the end of the twentieth century. Although Beer’s law does not apply to mass spectrometry, the log-transform continues to be key to mass spectrometry-based

---

**Fig. 1** Log infrared reflectance measurements on 50 samples of mustard seeds within the 1100–2500 nm wavelength range.
omics data analysis. This is because the statistical reasons and rational behind the log-transform are more enduring and powerful than the appeal to Beer’s law might reveal. We discuss four distinct arguments in favor of the log-transform.

1.1 Scale and Order of Magnitude Effects

The first argument is associated with the objective of mass spectrometry to estimate protein composition in complex mixtures consisting of large numbers of proteins which may differ substantially both in the masses of the constituent proteins in the mixture and in the abundances in which these are present. It is precisely because of this objective that mass spectrometry has become a tool of choice in modern omics research. Indeed modern mass spectrometric instruments are specifically engineered to have the capability to measure protein concentrations across large ranges of abundance, typically spanning several orders of magnitude—known as the so-called dynamic range—and to simultaneously achieve this across a wide mass range.

Unfortunately, the spectrometry engineer’s delight then becomes the statistical analyst’s nightmare as this property renders data which spreads across several orders of magnitude in spectral response. A good example may be found in Fourier-transform spectrometry, which can easily display variation across 5–6 orders of magnitude. In extreme cases this may cause numeric overflow problems in analyses, although use of modern professional statistical software may reduce this problem. Taking logarithms removes the order of magnitude effects.

1.2 Skew and Influential Observations

A second issue related to the above is that spectral measures will tend to be extremely skewed, not only within an individual (across the within-sample spectrum responses), but also across samples or patients at a single m/z point. This may render the data unsuitable for standard analyses such as linear discriminant or similar when used on the original scale.

A related issue is that the skew may cause or be associated with a limited set of highly influential observations. This is particularly troubling in omics applications, as the spectra are typically very high-dimensional observations either when storing the response on a large grid of m/z values which can easily range in the thousands, or after reduction to an integrated peak list. Influential observations may affect the robustness of conclusions reported, as results may differ substantially after removal of a single—or isolated group of—spectra from the analysis. A good example may be found in the calculation of a principal component decomposition as a dimension reduction prior to application of some subsequent data-analytic procedure. Principal component analysis is known to be sensitive to extreme observations [7], particularly in high-dimensional applications with small sample size. The same
phenomenon will, however, also tend to apply for other analysis approaches, such as regression methods and discriminant procedures. Transformation to log-scale may mitigate this problem.

1.3 Statistical Properties of Particle Counting

At some risk of oversimplification, mass spectrometers are in some sense nothing else but sophisticated particle counters, repeating a particle counting operation at each m/z position along the mass/charge range which is being investigated. As a consequence, spectrometry measures tend to have statistical properties reminiscent of those observed in Poisson (counting) processes. The variation of the spectral response tends to be related to the magnitude of the signal itself. This implies that multiplicative noise models are often a more faithfully description of the data. Multiplicative error data are, however, more difficult to analyze using standard software. Log-transforming can be used to bring the data closer to the additive error scenario.

1.4 Intrinsic Standardization

An interesting property of log-transforming is that it may lead to intrinsic standardization of the spectral response. Imagine, for example, that we are interested in calibrating the expected value \( E(Y) \) of some outcome \( Y \) based on two spectral responses \( X_1 \) and \( X_2 \) and that we have a study available recording both the observed outcome and the two spectrometry measures across a collection of samples (such as patients). Let us also assume that we can use some generalized linear model to link the expected outcome to the spectrometry data via some link function \( f \) and that the true model may be written in terms of a linear combination of the log-transformed spectral measures

\[
 f(E(Y)) = \alpha + \log(X_1) - \log(X_2)
\]

which reduces to

\[
 f(E(Y)) = \alpha + \log\left(\frac{X_1}{X_2}\right).
\]

The result is that the linear dependence of the expected outcome via the link function on the log-transformed data actually implies regression on the log-ratio of the spectral responses \( X_1 \) and \( X_2 \), such that any multiplicative effect would
Transformation, Normalization, and Batch Effect in the Analysis of Mass... 5
cancel. This can be regarded as an implicit form of standardization through the log-
transform. It is a general property which can be used in many statistical approaches
such as (generalized) linear regression and discriminant analysis.

For spectrometry data generally and mass spectrometry particularly, generally
good advice would be to replace the raw measurements with log-transformed values
at an early stage of the analysis, by application of the transformation
\[
\log(Y + a)
\]
with \(a\) a suitably chosen constant. See also [14] for comments about transformation
for mass spectrometry data and the log-transform in particular. Many statistical texts
will also mention the Box-Cox transform when discussing the log-transform. For
(mass) spectrometry data this approach is of limited value however, because the
optimal transform may lack the multiple justifications given above—which might
as well be used as a priori grounds for choosing the logarithm—but also because
the approach is by definition univariate, while a modern mass spectrometry reading
will consist of thousands of measures across m/z values and samples.

2 Normalization and Scaling

Normalization is an issue which is often encountered in omics data analyses, but
is somewhat resistant to a precise definition. We can identify what are usually
perceived as the main objectives of it and warn about the dangers associated with
the topic, so that we may avoid the most common pitfalls.

The objective of normalization can be loosely described as removing any
unwanted variation in the spectrometric signal which cannot be controlled for or
removed in any other way, such as by modifying the experiment, for example. This
sets it apart from batch effect, which we will discuss later. The latter can sometimes
be adjusted for or accommodated by changing the experiment so that its effects can
be either explicitly removed or adjusted for in subsequent analysis, by exploiting
the structure of the experimental design. Not all effects can be accounted for in this
manner, however.

Examples of such effects which may induce a need for normalization are
variations in the amount of material analyzed, such as ionization changes, e.g., small
changes in “spotting” sample material to plates, subtle fluctuations in temperature,
small changes in sample preparation prior to measurement, such as bead-based
processing to extract protein, differential sample degradation, sensor degradation,
and so forth. An important feature of such variation is that, while we may speculate
such variability sources are there and affect our experiment, they are difficult to
either control or predict, which typically means all we can do is try to post hoc
adjust for it, but prior to any subsequent analysis steps.

Important in devising an appropriate normalization strategy is that we should try
to remove or reduce these effects on the measured spectral data, while retaining the
relevant (biological) signal of interest. Unfortunately, this is typically problematic
for spectrometry. This is because, as explained in the above paragraphs on transformation, the spectrometric signal and its variability are typically linked, often even after log-transform, while the unwanted sources of variation affect all measures derived from the spectrum.

Imagine we have a study recording a mass spectrum on a dense grid of finely spaced points along the mass range or alternatively storing the data as a sequence of integrated peaks representing protein or peptide abundances and this for a collection of samples (be it patients, animal, or other). We write the ordered sequence of spectrometry measures for each $i$th sample unit as $x_i = (x_{i1}, \ldots, x_{ip})$, with $p$ the number of grid points at which the spectrum is stored or the number of summary peaks. A transformation choice favored by some analysts is to apply early on in the analysis (possibly after first application of the log-transform) standardization to unit standard deviation of all spectral measures at each grid point separately. In other words, we replace the original data at each $j$th gridpoint with the measures $x_{ij} = \frac{std(x_j)}{\sqrt{n}}/\sqrt{\sum(x_j)^2}$ where $std(x_j)$ is the standard deviation of the measures $x_j = (x_{1j}, \ldots, x_{nj})^T$ across all $n$ samples at that gridpoint. This procedure, sometimes also referred to as reduction to z-scores, is a form of scaling. It is identical to standardization to unit standard deviation of predictor variables in regression analysis ([16, pp. 124–125], [15, pp. 349, 357–358]) when predictor variables are measured at different measurement scales (different units, such as kg, cm, and mg/l). Indeed, in the early days of regression analysis, reporting standardized regression coefficients was an early attempt at assessing relative importance of effects.

For spectrometry data generally and mass spectrometry in particular, standardization is more complicated. Measurement units are by definition identical within a spectrum across the mass range. This would counsel against transforming to unit standard deviation as calibrated effects then remain directly comparable across the mass range on the original untransformed scale. There are, however, stronger arguments against this form of standardization in spectrometry. Figure 2 illustrates the issues. The plot shows mean (MALDI-TOF) spectra from a clinical case–control study, after suitable transformation. To ease comparison, we plot the negative control spectrum versus the mean case spectrum. The rectangular region highlights and enlarges a region between 1200 and 1900 Da where most of the discriminant effects are found between the cases and control groups, based on a discriminant analysis. Indicated are four key peaks at 1352.4, 1467.7, 1780, and 1867.2 Da which together summarize most of the between-group contrast between cases and controls. Figure 3 shows different statistics calculated on the same data within the same mass range. The top plot again shows the mean spectra for cases and controls within the 1200–1900 Da region as before, while the middle graph plots a graph of weighted discriminant coefficients obtained from a linear discriminant model calibrated from the data. It is obvious how the discriminant analysis identifies the peaks at 1354.2 and 1467.7 Da and contrasts these with the peaks at 1780 and 1867.2 Da. The below graph in Fig. 3 shows the first two principal components calculated on the same data and based on the pooled variance–covariance matrix.

There are several things to note in this picture. The first is how much the principal component and mean spectra curves resemble one another. The first component
Fig. 2  Mean spectra for cases and controls separately in a case–control study. The negative of the mean control group spectrum is plotted to improve readability.

Fig. 3  The top plot shows mean cases and controls spectra separately. The bottom curves are the loadings of the first two principal components across the same mass/charge range. The middle curve shows the discriminant weights from a logistic regression model calibrated to distinguish cases from controls with the same data.
closely approximates the mean control spectrum, while the second component does the same for the mean cases spectrum. At first sight, this might seem all the more remarkable, since the principal component decomposition is based on the pooled variance–covariance matrix, and hence on the “residual spectra” \( x_i - \overline{x}_{g(i)} \), where \( \overline{x}_{g(i)} \) denotes the mean spectrum of group \( g(i) \) to which the \( i \)th observation belongs, with \( g = 1, 2 \) for the cases and control groups, respectively. So the figure shows two different aspects of the data. One is the systematic (mean) spectral response (top graphs), the other are the deviations relative to the mean spectral outcome (bottom graphs). From the figure, we can see that the component decomposition tells us that the peaks at 1352.4 and 1467.7 Da are highly correlated and account for much of the variation in the spectral data, as they weigh heavily in the first principal component. Similarly, the second component summarizes much of the expression in both peaks at 1780 and 1867.2, which are again highly correlated. Because of this, the classification might as well be summarized as a contrasting between the first and second principal component, since this would contrast peaks 1352.4 and 1467.7 with the expression at peaks 1780 and 1867.2 (see Mertens et al. [11] for the full analysis).

This feature of the data where the mean expression and deviations from the mean are closely linked as shown in the above example is typical of spectrometric variation. It is the consequence of the connection between mean expression and variance we mentioned above when discussing the log-transform and can be observed in almost all spectrometry data, often even after log-transforming. To put this differently, in spectrometry data, we will find the signal where the variance is (even if we correct the variance calculation for systematic differences in expression, as shown in our above example). It is for this reason that transforming to unit standardization should be avoided with spectrometry data, unless scale-invariant methods are explicitly used to counter this problem.

In addition to the above considerations, there are also other arguments for avoiding reduction to \( z \)-scores or transformation to unit standard deviation. An important argument here is that summary measures such as means and standard deviations are prone to outliers or influential observations, which can be a particular problem in high-dimensional statistics and with spectrometry in particular. A specific problem with such form of standardization is that it may cause problems when comparing results between studies. This is because systematic differences may be introduced between studies (or similarly, when executing separate analyses between batches—see further), due to distinct outliers which affect the estimates of the standard deviations for standardizing between repetitions of the experiment.

Our final comment on the above standardization approach is that medians and inter-quartile ranges (IQR) are sometimes used instead of means and standard deviations in an attempt to alleviate some of the robustness concerns. Other authors advocate use of some function of the standard deviation, such as the square root of the standard deviation instead of the standard deviation itself. This is sometimes referred to as Pareto scaling [17]. The rational for this amendment is that it upweights the median expressed features without excessively inflating the (spectral)
baselines. An advantage of the approach may be that it does not completely remove the scale information from the data. Nevertheless, the choice of the square root would still appear to be an ad hoc decision in any practical data-analytic setting.

Some authors make a formal distinction between scaling and normalization methods and consider the first as operations on each feature across the available samples in the study [5]. Normalization is then specifically defined as manipulation of the observed spectral data measurements on the same sample unit (or collection of samples taken from the same individual) (within-spectrum or within-unit normalization). A potential issue with the above-described approaches to normalization via statistical transformation is that they are based on a borrowing of information across samples within an experiment. Another extreme form of such borrowing is a normalization approach which replaces the original set of spectral expression measures for a specific sample with the sample spectrum measures divided by the sample sum, such that the transformed set of measures adds to 1. An argument sometimes used in favor of such transformation is that it would account for systematic differences in abundance—possibly caused by varying degrees of ionization or similar effects from sample to sample—such that only the relative abundances within a sample are interpretable. Although this approach is unfortunately common, it has in fact no biological foundation [5]. Even if arguments based on either the physics or chemical properties of the measurement methodology could be found, these could not be used in favor of such data-analytic approach as described above, which we shall refer to as “closure normalization.” The problem with the approach is that it actually induces spurious—and large—biases in the correlations between the spectral measures which mask the true population associations between the compounds we wish to investigate. Figure 4 shows the effect of closure normalization on uncorrelated normal data in three dimensions. The left plot shows scatterplots between each of the three normally distributed measures. The right shows scatterplots of the resulting transformed variables after closure normalization. The absolute correlations between each variable pair has increased from 0 (for the original uncorrelated data) to 0.5 (after transformation). This becomes particularly problematic should the subsequent objective be to perform some form of network or association analysis, in which case the prior closure normalization renders results meaningless. Similarly problems would, however, also apply to regression and discriminant analysis.

A variant of the above closure approach to normalization which is sometimes also used is to adjust to the maximum peak observed in a spectrum, where the maximum is either the spectrum-specific maximum, or the maximum at that spectral location which corresponds to the maximum mean spectrum across several samples. Adjustment is often carried out by dividing each spectrum by its maximum at the maximum location, such that the spectral response gets a constant expression at the maximum location in the transformed data. Just as for closure normalization, this procedure appears appealing on intuitive grounds at first sight, but suffers from similar problems, as the variation at the maximum induces severe correlations across the spectral range in the transformed data, which cannot have biological interpretation. Both approaches should be avoided.
Data normalized by the sum of the combined expression (closure normalization) can be viewed as an instance of compositional data [1]. Hence, instead of applying such normalization, one could therefore think of using special-purpose methods from the compositional data analysis literature, or to develop or adapt such methods for application in omics applications. This has not been attempted to our knowledge at the time of writing. As an alternative, it should be recommended to take a conservative approach and refrain from excessive transformation when the consequences are not well understood or accounted for in subsequent analysis. In such cases restricting to log-transformation as discussed earlier is safer. In any case, the original untransformed data should always be at hand and stored to allow verification of results through possible sensitivity analysis.

The above is only an introduction to some of the main forms of normalization in use at this time. Many other forms exist and will undoubtedly continue to emerge. An interesting one worth mentioning is the idea of “lagging” the spectra by taking differences between subsequent values within the spectral range. With log-transformed spectrometric data, this is another approach which induces ratios between subsequent spectral intensities which eliminate multiplicative change effects. An example is found in an interesting paper by Krzanowski et al. [10]. It has the drawback that results from subsequent statistical analysis can be more difficult to interpret, but it might be of use in pure prediction problems. Other forms of standardization and normalization are also found in the literature, particularly methods which seem to inherit more from common approaches in microarray analysis, such as quantile normalization [6]. Ranking of spectral response including the extreme form of reducing to binary has also been investigated. The latter can be particularly useful as a simple approach when data are subject to a lower detection

**Fig. 4** Effect of closure normalization on uncorrelated data
limit [8]. Other forms of normalization and standardization worth mentioning at time of writing are scatter correction and orthogonal signal correction. We refer to Naes et al. [13] for a good introduction to these methods.

Which transformations should be applied first? What is a good order of applying distinct normalization or transformation steps? There is some difference of opinion between researchers on the precise sequence in which various normalization procedures are applied to the data. As a general rule it seems wise to apply logarithms early and calculate means and standard deviations only after log-transforming.

The issue of normalization is closely linked to the problem of standardization of mass spectra. Several definitions of standardization may be possible here. One option is to define the problem as “external” standardization, which would form part of the experiment itself (as opposed to the post-experimental data processing we describe before) where we somehow try to change the experiment so that part of the systematic experimental variation is either prevented from occurring or could be accounted for through post-processing of the data. Examples would be in the use of spike-in controls, on a sample plate, or even within the sample material itself, so that the spectral response can be adjusted for the expression of the known spike-in material which is added. Another example would be in the use of technical controls on a sample plate with known concentrations. Yet another example would be systematic equipment re-calibration to re-produce a (set of) known standards, so that sample-to-sample variation due to experimental drift is suppressed as much as possible. All these approaches to standardization are different from the above described methods in that they try to circumvent known sources of variation by changing the experiment itself, rather than post hoc attempting to adjust for it.

3 Batch Effect in Omics Studies

A batch effect is a source of unwanted variation within an experiment which is typically characterized by the property that the effect is due to a known structure within the experiment. Usually the structure, and thus also the existence of the associated effect, is known in advance of the experiment but cannot be avoided or eliminated from the measurement process. Furthermore, the effect of the experimental structure itself may not necessarily be predictable either, even when knowledge of the structure exists. A typical example would be distinct target plates which are used to collect sample materials prior to analysis in some mass analyzer. However, we can sometimes manipulate or modify the experiment to account for the known structure so that it is rendered innocuous and cannot unduly affect conclusions. Even better, we may be able to remove (part of) the effects due to batch structure, by taking suitable precautions at the experimental design stage.

In contrast to genetics, batch effect is much more problematic within proteomics and metabolomics, which is due to the measurement procedures involved. It should also be noted that batch effect and the accommodation of it are different from standardization and normalization issues in that we need to identify its presence
and consider its potential effects both before and after the experiment. Before the experiment, because we may want to tweak or change the experimental structure to take the presence of the batch effect into account (this may involve discussion between both statistician and spectrometrist). The objective is to change the experimental design so as to avoid confounding of the batch structure with the effect or group structure of interest. After the experiment, because we may wish to apply some data-analytic approach to remove the effect (which may depend on the experiment having been properly designed in the first place). Accounting for batch effect in proteomic studies will hence involve two key steps.

1. To control the experimental design as much as possible in advance of the experiment to maximize our options to either remove or account for batch effect at the analysis stage.
2. To exploit the chosen structure afterwards to either remove or adjust for the batch variation in the analysis.

The objectives of these steps are at least threefold.

1. To ensure experimental validity (the experiment can deliver the required results)
2. To improve robustness of the experiment (conclusions will still be valid, even when experimental execution differs from experimental planning)
3. To improve (statistical) efficiency of effect estimates based on the data (statistical summaries will have lower variation).

In the following discussion on batch effect we will make a distinction between the following two types of batch effect which may occur in practical experimentation.
The first are time-fixed batch effects. Examples of these are

- plates in mass spectrometry
- freezer used for sample storage
- change of cleaning or work-up fluid for sample processing
- batches of beads or cartridges for protein fractionation
- instrument re-calibration.

In contrast, time-dependent batch effects are due to experimental structures associated with time. Examples of these are found in the following situations:

- longitudinal experimentation, long observation tracks with repeated measurement per individual
- sample collection across extended time periods, patient accrual spread across an extended period of time
- time-indexed samples or instrument calibrations
- distinct days of measurement or sample processing.

Both types of batch effects may occur in proteomic experiments, but for different reasons and with distinct consequences. The treatment of both types of batch effect will also be different between the two.