Multivariate analysis of NMR-based metabolomic data

Julia Debik1 | Matteo Sangermani1 | Feng Wang1,2 | Torfinn S. Madssen1 | Guro F. Giskeødegård2,3

1Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology—NTNU, Trondheim, Norway
2Clinic of Surgery, St. Olavs Hospital HF, Trondheim, Norway
3K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology—NTNU, Trondheim, Norway

Correspondence
Julia Debik, Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology—NTNU, 7491 Trondheim, Norway. Email: julia.b.debik@ntnu.no.
Guro F. Giskeødegård, K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology—NTNU, 7491 Trondheim, Norway. Email: guro.giskeodegard@ntnu.no

Funding information
The Norwegian Cancer Society, Grant/Award Numbers: 6834362, 202021; Stiftelsen Dam, Grant/Award Number: 2020/FO298770; Joint Research Committee between St. Olavs hospital and the Faculty of Medicine and Health Sciences, NTNU, Grant/Award Numbers: 28346, 28328; Swiss National Foundation, Grant/Award Number: p400PM_194482

Nuclear magnetic resonance (NMR) spectroscopy allows for simultaneous detection of a wide range of metabolites and lipids. As metabolites act together in complex metabolic networks, they are often highly correlated, and optimal biological insight is achieved when using methods that take the correlation into account. For this reason, latent-variable-based methods, such as principal component analysis and partial least-squares discriminant analysis, are widely used in metabolomic studies. However, with increasing availability of larger population cohorts, and a shift from analysis of spectral data to using quantified metabolite levels, both more traditional statistical approaches and alternative machine learning methods have become more widely used. This review aims at providing an overview of the current state-of-the-art multivariate methods for the analysis of NMR-based metabolomic data as well as alternative methods, highlighting their strengths and limitations.

KEYWORDS
ASCA, clustering, deep learning, machine learning, PCA, PLS-DA, validation

1 | INTRODUCTION

Metabolomics provides high-throughput identification and quantification of small-molecule biochemical compounds in biological samples such as biofluids, cell extracts, or tissues.1 Due to the dynamic nature of the metabolome, it can provide a more integrated view of the biological state of the sample in response to both endogenous and environmental factors, compared with genomics and proteomics.2 Metabolomics is therefore a promising tool for identifying biomarkers for diseases, and for discovering potential drug targets. Proton nuclear magnetic resonance (1H-NMR)
spectroscopy and mass spectrometry (MS) are the two main analytical approaches for metabolomic studies, and are complementary to each other. While MS-based approaches have a higher sensitivity (with a detection limit of picograms versus nanograms), NMR spectroscopy is highly reproducible and requires minimal sample preparation.\(^3\)

NMR metabolomic data can be analyzed either as spectra or as absolute or relative concentrations from metabolite quantification. Due to overlap in spectral peaks, metabolite quantification has traditionally been performed by peak fitting, which is a time-consuming procedure, or by integration of spectral areas, which does not allow reliable quantification of highly overlapping peaks. However, during recent years methods allowing batch quantification, such as Chenomx, BATMAN,\(^4\) and Mnova,\(^5\) have become available, enabling more efficient quantification. Commercial platforms for automatic metabolite and lipoprotein quantification from NMR analysis, such as Nightingale Health and Bruker BioSpin, have also recently become available, making metabolite quantification of large-scale data sets possible. For instance, NMR metabolic profiling was used to assess all-cause mortality risk in a cohort of 44,000 individuals,\(^6\) revealing serum metabolites that could predict mortality better than conventional risk factors.

For data analysis, metabolomic data are commonly stored in a matrix where each row represents one sample and each column represents a variable. When using spectra for analysis each column will correspond to data points of the NMR spectra, while when analyzing quantified metabolites each column will represent the concentrations of a specific metabolite. We may in addition have one or more vectors of variables describing outcomes or characteristics that we wish to relate to the metabolomic data. Model building can be either supervised or unsupervised. In unsupervised learning the model is built solely based on variation patterns inherent in the data, without any sample labels, whereas in supervised learning such labelling information is used to build the model. Supervised learning can be used for both regression and classification problems.

A central feature of metabolomic data, in particular spectral data, is its high degree of collinearity. Because metabolites act together in complex metabolic networks, optimal biological insight can only be achieved when using methods that take into account the correlation between metabolites. For this reason, latent-variable-based methods such as principal component analysis (PCA) and partial least-squares (PLS) regression have been widely used in metabolomics. However, the increased use of metabolite concentrations over NMR spectra for data analysis, in addition to the increasing availability of larger data sets, allows for use of other machine learning methods in addition to statistical methods that have previously been considered unsuitable for NMR metabolomic data. Machine learning refers to data analysis in which the computer may learn from the data and identify patterns in an automated way, with minimal human interaction. This review aims at covering the current status, methodological aspects, and future outlook of multivariate analysis of NMR-based metabolomic data.

## 2 | PREPROCESSING OF NMR SPECTRA

Prior to data analysis, the raw NMR spectral data must be converted through various computational preprocessing procedures.\(^7,8\) This is crucial to extract meaningful information and correct biological interpretation in metabolomic studies. Besides the initial steps of preprocessing such as phasing and baseline correction, peak alignment and binning are commonly used data preprocessing techniques. In order to adjust peak shifts in the raw spectra and thus to make data comparable in multivariate analysis, peak alignment can be performed by shifting the spectra to a reference peak and in addition by applying a peak alignment algorithm.\(^9,10\) Alternatively binning (or bucketing) can be utilized to correct for interindividual differences in peak positions on raw data or small misalignments on the aligned spectra. Binning sums up spectral intensities of predefined regions (bins), thus reducing the spectral resolution. While standard equidistant binning divides each spectrum into equally sized bins, several alternative approaches, such as non-equidistant binning, adaptive binning, and intelligent binning, allow variable bin sizes or automatically determine bin edges without user-defined arbitrary parameters or reference spectra.\(^11–13\) The reduced spectral resolution from binning may be beneficial for further computer-intensive analysis; however, this is usually not a problem with the current availability of high computational power. Both peak alignment and binning may have drawbacks, such as the lack of ability to handle overlapping peaks correctly, allocating peaks into the wrong bins or leading to peak splitting, thus it is important to visually inspect the results before further analysis.

Normalization can be applied to remove systematic variation between the samples to make them more directly comparable to each other.\(^14\) This variation could be caused by differences in sample weights, volumes, or dilution. Normalization is a row operation, and applies a scale factor to each row of the data matrix. A commonly used normalization method is the total area or integral normalization.\(^7\) With this method, each spectral point is divided by the total spectral area of each sample, so that the total spectral intensity is kept constant across different samples. Metabolites present in high concentrations, or with wider peaks (such as lipids), will contribute more to the total area affecting the normalization and may introduce a bias. Probabilistic quotient normalization (PQN) and quantile normalization are two alternative normalization techniques commonly used in metabolomic studies.\(^15,16\) It is important to keep in mind that normalization may greatly influence the results, thus the normalization method used should be taken into account when interpreting the results.

As described previously, NMR data may be analyzed in the form of spectra or as quantified metabolites. Quantification can be performed by peak integration or using commercial quantification algorithms, and metabolite concentrations may be relative or absolute. Relative concentrations can be achieved by peak integration from normalized spectra, while a concentration calibrant, such as an internal or external reference standard, or an artificial reference signal (such as ERETIC), is necessary to achieve absolute metabolite quantification.\(^17,18\) Internal reference standards
such as TSP and DSS are commonly used and have demonstrated high quantification accuracy and precision in simple solutions. They may however interact with protein and macromolecules, causing peak distortion and inaccurate quantification, thus should be used with caution in complex biological samples such as biofluids.

Preprocessed spectra or metabolite concentrations may be further pretreated by centering and scaling.\(^\text{14}\) Mean-centering the data is common practice prior to analysis with latent variable projections such as PCA. By removing the offset from the data, mean-centering allows statistical models to focus on the between-sample variation. Furthermore, highly abundant metabolites may dominate the model, masking potentially important, but low-abundance metabolites. Scaling, which is a column operation, gives low-abundance metabolites equal opportunity to influence the multivariate models. Several scaling methods, such as auto-scaling (standardization or scaling to unit variance), Pareto scaling, range scaling, and VAST scaling, are commonly applied in NMR-based metabolomic studies.\(^\text{19}\)

## 3 | TRADITIONAL STATISTICAL METHODS

### 3.1 | General linear models (GLMs)

General linear models (GLMs) are commonly used in metabolomics to assess if there are significant differences in metabolite concentrations between groups, or if a metabolite has a significant correlation to a property of interest. Simple linear regression finds the best linear relationship between a covariate \(x\) and the response variable \(y\), while multiple linear regression (MLR) finds the best linear relationship between multiple covariates \((x_1, x_2, ..., x_p)\) and the response \(y\). Given \(p\) covariates, the GLM can be described as

\[
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \epsilon_i.
\]

Here \(y_i\) is the response variable of sample \(i (i = 1, ..., N)\), \(x_q\) are the corresponding covariates \((j = 1, ..., p)\), \(\beta_0\) is the intercept, \(\beta_j\) is the coefficient for covariate \(j\), and \(\epsilon_i\) are the residuals. This equation may alternatively be written in a matrix form:

\[
y = X\beta + \epsilon
\]

where \(X\) is an \(N \times (p + 1)\) model matrix, \(\beta\) is a \((p + 1) \times 1\) vector of coefficients to be estimated from the data, and \(\epsilon\) is an \(N \times 1\) vector of residuals. Depending on the context and research question, metabolomics variables can be included in the model matrix \(X\) as covariates, or they can be included in \(y\) as a response variable. For example, t-tests and analysis of variance (ANOVA) compare mean metabolite concentrations across different groups and are equivalent to fitting a linear regression model with categorical covariates for group, and the metabolite level as the response variable \(y\).\(^\text{20}\) Alternatively, metabolite levels can be included in \(X\) as covariates in order to model or predict the response \(y\), e.g., a disease marker.

For example, linear regression has been applied to model the relationship between age and serum metabolic concentrations, adjusted for BMI.\(^\text{21}\)

The covariates in \(X\) are referred to as fixed effects, because they are treated as fixed quantities that are to be estimated from the data. Linear mixed models (LMMs) are an extension of GLMs allowing for both fixed and random effects. Random effects account for random variability across levels of pre-specified factors and are adjusted for, but not directly estimated. The general equation for the LMM can be written as

\[
y = X\beta + Z\gamma + \epsilon
\]

where \(X\) and \(Z\) are design matrices of dimensions \(N \times (p + 1)\) and \(N \times q\) for fixed and random effects, respectively. \(N\) is the number of samples, while \(p + 1\) and \(q\) indicate the numbers of fixed and random effects, respectively. LMMs are commonly used to account for dependences between samples, which for example occur in longitudinal studies where each subject is measured repeatedly.\(^\text{22,23}\)

### 3.2 | Logistic regression

In this and the following section, the metabolites are usually the covariates of the model, while the response may for example be presence or absence of a disease or a treatment type. When the response variable is binary \((y_i \in \{0, 1\})\), for instance positive or negative for disease, logistic regression can be used to model the probability of an event given one or more covariates. The probability of success (the sample is positive for the disease) is \(\pi_i(x_i) = P(y_i = 1 | x_i)\), and a logistic regression model has a linear form in the logit of this probability:

\[
\logit(\pi_i(x_i)) = \log \left( \frac{\pi_i(x_i)}{1 - \pi_i(x_i)} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}
\]
where \( \mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip}) \) is a vector containing the covariates for individual \( i \). This formula implies that \( x_i \) increases or decreases in an S-shaped function, and thus assures that the probability lies in the interval between zero and one. The parameters \( \beta_j (j = 1, ..., p) \) determine the rate of increase, and the sign indicates whether the curve ascends or descends.\(^{24}\) Logistic regression is widely used in metabolomic studies with large patient cohorts and a case cohort study design, for instance to identify associations between metabolites and the risk of developing cancer.\(^{25,26}\) The odds ratio (OR), which is a measure of association between a covariate and the response, is commonly reported. An OR greater than 1 indicates a positive association between the covariate and the response, while an OR less than 1 indicates an inverse association.

### 3.3 | Cox regression

Cox proportional hazard regression\(^{27}\) (Cox regression) is the most commonly used multivariate approach for survival analysis in medical research. Survival analysis involves analyzing the effect of one or more covariates on the expected duration of time until an event such as death, disease occurrence, or recurrence.\(^{28}\) The Cox regression model may be expressed mathematically for individual \( i (i = 1, ..., N) \) as

\[
h_i(t|\mathbf{x}_i) = h_0(t)e^{(\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip})}
\]

where \( t \) is the survival time and \( h_i(t) \) is the hazard function determined by a set of covariates \( \mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip}) \) whose impact is determined by the coefficients \( \beta_j (j = 1, ..., p) \). The baseline hazard, \( h_0(t) \), is the value of the hazard when all the covariates \( (x_i) \) are equal to zero.

A hazard ratio for the \( j \) th covariate may be calculated as \( e^{\beta_j} \) to investigate the direction of the association. A hazard ratio above unity indicates that as the \( j \) th covariate increases the event hazard and the length of the survival decreases, thus indicating that there is a negative association between the covariate and the length of survival, \( t \). Similarly, a hazard ratio below unity indicates a positive association between the variable and the length of survival. Cox regression may be used to predict the probability of survival at any given point in time for a particular risk group. For example, recently Cox regression with an included stepwise procedure was applied to identify 14 metabolic predictors in the circulation of long-term all-cause mortality based on a cohort of 44,168 individuals, of whom 5512 died during follow up.\(^6\) Cox regression analysis has also been used to identify the associations of lysophosphatidylcholine 18:0 with lower risks of breast, prostate, and colorectal cancers in the EPIC-Heidelberg cohort.\(^{29}\)

There are some common issues for the traditional statistical models described in previous sections that make them less suitable for metabolomic studies. The main problem is the need for a higher number of samples than variables for the models to converge \( (N > p) \). These models also have problems with covariates that are highly correlated with each other (ie multicollinearity), which is often the case for metabolomic data, although this issue is smaller for quantified metabolite concentrations than spectral data. Further, these models have traditionally been used for statistical inference, where \( p \)-values have been central for assessing the significance of variables included in the model. This is different from finding the best prediction models. To prevent overfitting and to create a model that is more interpretable, the most parsimonious model that still explains the data should be selected. The task of assessing the importance of variables in the model, determining which variables should be included, and verifying the model is not trivial.\(^{28}\) The selection of variables has traditionally been performed in a stepwise manner.\(^{30}\) However, the drawbacks of stepwise regression\(^{31,32}\) have led to the popularity of regularization methods,\(^{33,34}\) such as the LASSO,\(^{35}\) Ridge\(^{26}\) or elastic net,\(^{37}\) making the process of model fitting and testing very similar to that of machine learning methods. Strengths and limitations of the traditional statistical methods are summarized in Figure 1.

### 4 | LATENT-VARIABLE-BASED METHODS

Omics technologies in biology provide the ability to measure hundreds or thousands of variables all at once. Dimensionality reduction is often required to extract meaningful information from the data.\(^{38,39}\) Latent-variable-based methods are methods that describe the data in terms of latent variables, which are constructed as linear combinations of the original variables. In this way, metabolomic data sets consisting of hundreds or thousands of variables can often be visualized and explained by a few new variables. These latent variables can be thought of as variables that cannot be directly measured, but can be inferred by analyzing other observable and measurable variables. They reflect an underlying phenomenon that can be extracted from dominant correlation patterns present between measurable variables. For example, “health” can be conceptualized as a latent variable that cannot be directly measured, but can be extrapolated from the sum of the contributions of other measurable proprieties (eg blood metabolite levels). In metabolomic studies, however, the interpretation of latent variables is that they show the direction of the largest variance (PCA) that optimizes the covariance between groups of variables (PLS). PCA and PLS are the two bedrock dimensionality reduction techniques based on latent variable analysis (Figure 2).
Principal component analysis (PCA) is an unsupervised method that projects the data to a lower-dimensional space so as to maximally preserve and explain the variance in the data. If $X$ is the matrix containing the metabolite concentrations or spectral variables for the metabolomic data, then PCA decomposes it into two matrices, the principal component (PC) scores matrix ($T$) and loading matrix ($P$):

$$X = TP^T + E$$

where $E$ is the residual matrix containing the variance not explained by the model, and $P^T$ denotes the transpose of the matrix $P$. Thus the original data may be described in a lower-dimensional space, defined by the PCs, which are inherently ordered in decreasing ability to capture the total variance of the data. The score values represent the coordinates of the samples in the lower-dimensional space defined by the PCs. The loadings are linear combinations of the contributions that each measured variable has in defining the PC, and all PCs are orthogonal to one another. Plotting the sample scores in this lower-dimensional space allows us to visualize and highlight differences between groups of samples, identify naturally occurring clusters, and detect outliers. The corresponding loadings can be assessed to determine which metabolites contribute to the observed trends. Alternatively, the scores and loadings may be plotted in a biplot, to visualize the associations between samples and covariates.

Due to its unsupervised nature, and its ability to handle multicollinearity, PCA is routinely used as an exploratory method in metabolomic studies. For longitudinal studies, where samples are taken at more than one time point for each individual, PCA trajectory plots are useful for investigating and illustrating trends in the scores across the repeated measures, where points from the same subject are connected by lines in the scores plot. This allows us to investigate between- and within-subject variation, and to assess if systematic changes for the study cohorts are present. This approach has recently been used to connect samples from the same patients in a randomized clinical trial, with three sampling time points.
points, clearly demonstrating systematic changes in the first period of the intervention period, but not the second. Alternatively, the average score of a PC against time may be plotted. For example, this method was used to visualize and highlight patterns in time-of-day variation in the urinary metabolome between controls and sleep-deprived subjects.

### 4.2 Principal component regression (PCR)

Principal component regression (PCR) is a supervised method, which is in essence a PCA followed by an MLR. As the regression is performed on the PCs, which are orthogonal and thus free of collinearity, it overcomes many of the previously mentioned problems with MLR for metabolomic applications. Moreover, it is easier to ensure that the number of included covariates is lower than the number of samples, by selecting only the first few components that explain the most variance. PCR is however not much used in metabolomic studies, as the first few components will not necessarily explain the problem of interest. PLS is another related approach, which is more efficient in that it requires fewer components to achieve the same prediction capability compared with PCR, and is easier to interpret.

### 4.3 Partial least squares (PLS)

Partial least squares (PLS) regression, also commonly referred to as projection to latent structures, is an improvement and extension of PCR. PLS creates a multivariate regression model in a reduced-dimensional space, by decomposing the data using latent variables. In PLS the matrix $X$ containing the covariates (metabolite concentrations or spectral variables) and the response matrix $Y$ are decomposed into scores ($T$ and $U$) and loading matrices ($P$ and $Q$):

\[
X = TP^T + E \tag{7}
\]

\[
Y = UQ^T + F \tag{8}
\]

such that the covariance between the matrices $T$ and $U$ is maximized. $E$ and $F$ are the error terms. PLS is well suited for metabolomic data, as it utilizes the correlation structure of the data, makes biological interpretations easier than for many other approaches, and works well when the number of covariates exceeds the number of samples. Similarly as for PCA, groupings of the samples may be visualized in a lower-dimensional...
space, defined by the latent variables, which are linear combinations of the original variables. Scores and loading can thus be interpreted similarly as for PCA.

PLS discriminant analysis (PLS-DA) is widely used for analysis of metabolomic data, and is a variant of PLS used for classification problems, i.e. when $y$ is categorical.\textsuperscript{49} The importance of variables may be assessed through the loadings and variable importance scores such as variable importance in projection (VIP) scores and the selectivity ratio (SR),\textsuperscript{48,50} which summarize the contributions of the original variables to the model. However, PLS models are prone to overfitting and the number of latent variables should be carefully selected by validation, which is described in Section 7. PLS-DA was for example used to show the effect of maternal fish oil supplementation on the child’s metabolome.\textsuperscript{51} It has also been applied in multi-omic studies, for example using metagenomics of the gut microbiome and metabolomics of biofluids, whereby it was possible to explore potential relationships between the microbiome and the metabolism of the human host.\textsuperscript{52,53}

Orthogonal projection to latent structures (OPLS) is a method to reduce model complexity and improve interpretation of PLS-DA models.\textsuperscript{54} OPLS removes systematic variation from the data matrix $X$ that is not correlated to the response $y$; in other words the variability in $X$ that is orthogonal to $y$. This causes a rotation, such that the first component shows the between-class difference. OPLS can either be achieved by integrating orthogonalization into the PLS algorithm as originally described by Trygg and Wold,\textsuperscript{54} or by post-processing of the PLS model.\textsuperscript{55} The orthogonal variation can be analyzed separately, e.g. by a PCA, whereby the scores and loading plots can be interpreted without considering their relation to $y$. It is however important to note that OPLS provides no predictive advantage over PLS in terms of model performance.\textsuperscript{55,56} OPLS has recently been applied to identify a multiorgan pathological signature of SARS-CoV-2 infection, based on metabolic phenotyping by NMR and MS.\textsuperscript{57}

Multilevel PLS-DA is a method for multivariate comparison of paired data. Metabolites in a biological sample are influenced by many factors, such as age, disease state, or genetics. This causes variation between samples of different individuals often to be higher than the variations between the samples of one individual. As a result, subtle treatment-induced changes can be overshadowed by the between-subject variations. In a multilevel PLS the variation between individuals (the between-subject variation) is separated from the variation within the samples (the within-subject variation).\textsuperscript{58} Thereafter, PLS-DA is employed on the within-subject variation, focusing on the similarity in the treatment effect between different individuals. Multilevel PLS-DA may be described as similar to a multivariate paired $t$-test compared with a regular $t$-test. For example, in an NMR metabolomic-based nutritional intervention study the authors compared the ability of multilevel PLS with ordinary PLS-DA to discriminate treatment effects. In this example, the treatment effect was a small fraction of the total variation and only multilevel PLS was able to find a statistical significant treatment effect.\textsuperscript{59}

5 \hspace{1em} COMBINING GENERAL LINEAR (MIXED) MODELS AND LATENT VARIABLE ANALYSIS

Metabolomic studies often follow an experimental design where experimental conditions, or factors, are manipulated systematically in order to determine each factor’s independent effect on the metabolite profile. A factor is divided into two or more levels. For example, a factor for treatment might have two levels, “Treatment” and “Control,” while a factor for time may have one level for each timepoint. Latent-variable-based models are limited in this situation, as they are generally blind to the experimental design and other sources of known variation, and PLS-based methods are generally only used to address one factor at a time. In contrast, GLM and LMM can simultaneously test and control for the effects of several variables (both continuous and categorical) on a response variable; however, they are generally only used for one response variable at a time. In this section we will briefly describe models where general linear (mixed) models and latent variable analysis methods are combined (Figure 3), namely PC-ANOVA, ANOVA simultaneous component analysis (ASCA) and related methods, and linear mixed model (LIMM)-PCA. For detailed descriptions and tutorials, the reader is referred to papers that comprehensively explain each of these methods.\textsuperscript{60–62}

5.1 \hspace{1em} Principal component ANOVA (PC-ANOVA)

One of the more straightforward ways of combining GLMs and latent variable analysis is PC-ANOVA. Suppose we have collected $N$ NMR spectra, where each spectrum consists of $p$ spectral points. We can then collect these spectra in an $N \times p$ matrix. In this context we are treating the spectra as responses, and we therefore denote this matrix as $Y$. PC-ANOVA involves first applying PCA to $Y$, and then analyzing the scores for separate PCs using for instance ANOVA:

$$t_0 = X\beta + e.$$ \hspace{1em} (9)

This can be repeated for as many components as desired, to determine which of the experimental factors have a significant influence on the variation described by the component. The impact of the design factors (the fixed effects included in $X$) on the scores can be assessed through...
standard GLM/ANOVA metrics. Note that, although the method has “ANOVA” in its name, there is no obstacle to analyzing the scores using other model types, such as ANCOVA or mixed models.

5.2 | ANOVA simultaneous component analysis (ASCA/ASCA+)

In ASCA, the order of PCA and ANOVA is reversed compared with PC-ANOVA. First an ANOVA model is used to decompose the response matrix \( Y \) (i.e., the metabolite concentrations) into effect matrices according to the experimental design. Then PCA is applied to each effect matrix separately, thereby focusing the multivariate analysis on the variation from that specific factor. This provides a multivariate extension of the ANOVA/GLM framework to an arbitrary number of response variables, which makes it highly relevant for metabolomic data. Two other methods, ANOVA principal component analysis (APCA) and ASCA-E, are identical to ASCA except in how they handle the residuals during visualization of the effects.

Suppose that \( y \) is a single response variable measured in a design with two factors, denoted \( \alpha \) and \( \beta \), which have \( K (k = 1, \ldots, K) \) and \( H (h = 1, \ldots, H) \) levels, respectively, and there are \( N \) samples. We can specify a full ANOVA model for this design as

\[
y_{ikh} = \mu + \alpha_k + \beta_h + (\alpha\beta)_{hk} + e_{ikh}
\]  \hspace{1cm} (10)

where \( \alpha_k \) is the contribution of being in the \( k \)th level of \( \alpha \), \( \beta_h \) is the contribution of being in the \( h \)th level of \( \beta \), \( (\alpha\beta)_{hk} \) is their interaction, and \( e_{ikh} \) is the residual term.

In ASCA, this decomposition is applied separately to each response variable, and the response matrix is decomposed as

\[
Y = M_\mu + M_\alpha + M_\beta + M_{\alpha\beta} + E
\]  \hspace{1cm} (11)

where each matrix has \( N \) rows and \( p \) columns. These \( M \) matrices, called effect matrices, contain the estimated level averages associated with each factor, and therefore describe the multivariate effect from that factor. The matrix \( M_\mu \) is an offset matrix where each row contains the overall mean spectrum. The error matrix \( E \) contains the residuals.
Applying PCA separately on each matrix can allow us to visualize the multivariate differences between the factor levels:

\[
Y = M_0 + T_0 P_0 + T_p P_p + T_r P_r + E
\]

where \( T \) and \( P \) are scores and loadings from PCA, respectively. By applying PCA on the effect matrices as opposed to the full response matrix \( Y \), each multivariate sub-model will be optimal for describing the variation contributed by that specific factor, and it is possible to study their effects in isolation. This methodology has been applied in a wide variety of research fields. It is commonly applied in analysis of designed experiments, to study the independent effects of experimental conditions on multivariate data. For example, it has been widely used in the field of nutritional metabolomics to study the effect of nutritional interventions, but is also seeing increasing use in human clinical trials, as well as in observational settings. It can also be used as a pre-processing step, to remove undesired variation from the data before training classification models.

As long as the design is balanced and the proper constraints are put on the model (ie using classical ANOVA effect coding), the effects described by the different effect matrices will be fully orthogonal. In this situation it is possible to quantify and decompose the sums of squares from each factors and thereafter calculate the explained variance for a given factor. However, other parametrizations are also possible, which give a different view of the data. For example, the effects can be expressed relative to a reference level, which is done in scaled-to-maximum, aligned, and reduced trajectory (SMART) analysis and principal response curves (PRCs).

In classical ASCA, the decomposition is done using simple ANOVA calculations which are based on differences in the means. A more general estimation method is to use GLMs to estimate the effects, as proposed by Thiel et al. This method is termed ASCA+, and generalizes the ASCA methodology so that it can be used for any GLM, including continuous covariates.

### 5.3 Extensions to mixed models

One of the recent developments of this modeling approach is the extension of the ASCA framework to mixed models. These extensions build on the ASCA+ framework, but use LMMs in place of GLMs to estimate the effect matrices. One such method is called LiMM-PCA. This approach also involves an orthogonalization and dimensionality reduction of \( Y \) by applying PCA, so that the score matrix \( T_A \) is used as the response matrix instead of \( Y \). The LiMM-PCA model then becomes

\[
T_A = M_0 + \sum_{f=1}^{k} M_f + \sum_{r=1}^{l} M_r + E
\]

where \( f \) and \( r \) are fixed and random effects, respectively. The effect matrices for both fixed and random effects can now be analyzed using PCA, as shown previously. The orthogonalization of \( Y \) in LiMM-PCA simplifies several aspects related to the variance decomposition and statistical inference, which are complicated by the presence of random effects. However, the loadings for each sub-model have to be transformed back into the original variable space \( Y \) before visual interpretation of the results, in order to obtain loadings with the same interpretation as in ASCA/ASCA+.

Another method that uses LMMs is repeated measures ASCA+ (RM-ASCA+). This method is aimed at analysis of longitudinal metabolomic data, and involves the use of repeated measures LMMs to estimate temporal multivariate trajectories, as well as time-dependent treatment effects. In contrast to LiMM-PCA, RM-ASCA+ is applied directly to the response matrix \( Y \). The matrix is then decomposed into effect matrices for time (\( T \)), treatment (\( G \)), and the time \( \times \) treatment interaction (\( T \times G \)), as well as a random effect matrix for subject (\( \gamma \)).

\[
Y = M_0 + M_T + M_G + M_TG + M_\gamma + E
\]

Because the \( F \)- and \( t \)-tests used in GLMs have no multivariate counterparts when the number of predictors exceeds the number of samples, ASCA methods commonly use resampling methods to evaluate the statistical significance of the model factors. The most commonly used approach is to use permutation tests, which involve repeated random re-assignment of factor levels to evaluate statistical significance, as described in Section 7. Another approach is to use bootstrapping or jackknifing to construct confidence intervals. These methods are highly flexible, although it is not always possible to construct exact permutation tests for all types of effect, eg interaction effects. Another proposed method involves construction of confidence ellipsoids, which permits multiple testing of effect level differences. In LiMM-PCA, statistical significance is assessed using a multivariate extension of the log likelihood ratio test, which can also be used to test random effects.
6 | OTHER MACHINE LEARNING METHODS

PCA and PLS-DA are currently the most widely used methods to analyze metabolomic data sets. This is primarily because latent-variable-based methods are particularly suitable for analyzing spectral data, where the number of covariates virtually always exceeds the number of samples, but also due to their easy interpretation. During the recent years, several other machine learning methods, both supervised and unsupervised, have grown in popularity and have successfully been applied. The sizes of cohorts in metabolomic studies have in general increased, thus other machine learning methods may provide valuable alternatives for analysis. While the above-mentioned methods can only model linear relationships, many of these alternative models are suitable for modeling complex non-linear relations. In this section we provide an overview and explore the most promising machine learning methods as alternatives to latent-variable-based methods (Figure 4). Similarly as for latent-variable-based methods, metabolite concentrations or spectral variables are here treated as covariates in modeling (the X matrix).

6.1 | Clustering methods

Clustering methods divide samples into clusters (subgroups) such that the pairwise dissimilarities between samples within a cluster are smaller than those between clusters. Many cluster analysis methods exist, of which perhaps the two best-known approaches are k-means clustering and hierarchical clustering (HCA). Similar to PCA, these methods are unsupervised and useful for detecting natural groupings in the data.

In k-means clustering, the desired number of clusters must be pre-specified. The algorithm starts by randomly placing k cluster centroids in the decision space. Then a distance measure (eg the Euclidean distance) is applied to calculate the distance from each point of the decision space to the centroids, and each observation is assigned to the cluster whose centroid is closest. For each of the k clusters, a new cluster centroid is calculated from the mean values of all of the samples within the cluster. Each observation is re-assigned to the cluster whose centroid is closest, and the process is repeated until no observation is reassigned. Some of the disadvantages of k-means clustering include the need to specify the number of clusters on beforehand, its sensitivity to outliers, and the fact that different initial placements of the centroids may result in different final clusters.

HCA is an alternative approach that does not require a pre-specification of the number of clusters. The HCA treats all samples as one cluster, and starts by merging the two clusters lying closest to each other in the decision space. This procedure is repeated at each iteration, until a single cluster remains. At the same time a tree-based representation of the observation called a dendrogram, which illustrates the similarity between the samples, is created. By cutting the dendrogram at a specific height, the number of distinct clusters may be chosen. HCA may alternatively be constructed by a top-down approach. For HCA, sometimes the correct number of clusters is difficult to define (interpretation is subjective) and an erroneous split at the beginning may not be corrected at a later stage. There are several examples of successful application of these methods to metabolomic data. For example, they have been used to identify metabolic clusters of breast cancer in relation to gene and protein expression subtypes. Different clustering approaches have been applied to NMR spectra of cancer cell line extracts and urine samples of Type 2 diabetes patients and animal models, for which a variant of k-means clustering resulted in the most accurate sample classification.

6.2 | Dimensionality reduction

An alternative approach to PCA for dimensionality reduction and exploring high-dimensional data is t-distributed stochastic neighbor embedding (t-SNE). Briefly, this algorithm consists of three steps: (1) the original high-dimensional data set is converted into a matrix of pairwise similarities between the samples, (2) a similar matrix of similarities is defined for the samples in the low-dimensional space, and (3) the difference in similarities in the higher-dimensional and lower-dimensional spaces is minimized so that a refined representation of the samples in the lower-dimensional space is obtained. t-SNE involves several hyperparameters such as perplexity, learning rate, and number of steps, which may strongly influence the visualization and allow for more flexibility. However, this advantage comes at the cost of easy implementation and interpretation. t-SNE can also reveal non-linear patterns and has for example been successful for visualizing periodic trends in serum metabolites and muscle transcripts during a time-restricted feeding intervention. Inference based solely on the output of t-SNE is however not possible, as it does not provide latent variables. While PCA is deterministic, meaning that the same input will always give the same result, this is not true for t-SNE. Similarly to PCA, t-SNE may be applied to reduce the dimensions of the original data prior to another machine learning approach.

Clustering methods and t-SNE are unsupervised, with their main use in exploratory data analysis and identification of naturally occurring clusters. In this section we explore supervised machine learning methods, which are able to model the relationship between a set of covariates (such as metabolite concentrations in a metabolomics setting) and the response variable based on training data. Such supervised machine learning models may be used to make predictions for new observations (samples).
**K-means clustering**

*Unsupervised*

*Usage: Clustering*

*Strengths / limitations:*
- Simple
- Sensitive to outliers
- No interpretation
- The number of clusters must be pre-specified
- Clusters depend on the starting locations of cluster centroids
- Non-deterministic

**Hierarchical clustering (HCA)**

*Unsupervised*

*Usage: Clustering*

*Strengths / limitations:*
- The number of clusters does not need to be pre-specified
- Produces a dendogram
- Shows the hierarchical relation between the clusters
- No interpretation

**t-distributed stochastic neighbor embedding (t-SNE)**

*Unsupervised*

*Usage: Exploratory analysis, dimensionality reduction*

*Strengths / limitations:*
- Suitable for non-linear problems
- Robust to outliers
- Preserves local and global structure of the original data well
- Low interpretability
- Non-deterministic

**Support vector machine (SVM)**

*Supervised*

*Usage: Classification (Regression)*

*Strengths / limitations:*
- Robust to outliers
- Work well on non-linear problems
- Low interpretability

**Tree ensembles (RF and GBM)**

*Supervised*

*Usage: Classification, regression*

*Strengths / limitations:*
- May model non-linear relationships
- May be prone to overfitting (GBM)
- Sensitive to imbalanced data

**Artificial neural network (ANN)**

*Supervised*

*Usage: Classification, regression*

*Strengths / limitations:*
- Has good performance when applied to complex tasks (such as image recognition)
- May model non-linear relationships
- Needs in general large data sets to train
- Low interpretability

**FIGURE 4** An overview of machine learning approaches discussed in this review, highlighting their strengths and limitations.
6.3 Tree-based ensemble methods

Tree-based ensemble methods have also been successfully applied to metabolomic data. Classification and regression trees (CARTs) are constructed by recursive binary splitting.84 When building a decision tree, the split that provides the most homogeneous two subgroups out of all possible splits, is chosen in a greedy manner. “Greedy” means that at each iteration the algorithm picks the best choice among the available choices. For predictive regression modeling, the best split is chosen so as to minimize the residual sum of squares of the tree, while for classification trees the Gini index is used, which is a measure of the total variance across all of the classes of the response variable.79 The Gini index for a node (split in the tree) is given by \( \sum_{k=1}^{K} p_k (1 - p_k) \), where \( p_k \) is the probability that an observation from class \( k \) has been correctly classified, thus a value close to zero indicates high homogeneity. Splitting is repeated until a stopping criterion is met, which may be the minimum number of samples at each leaf, or a maximum tree depth. A decision tree is easily interpretable, because the prediction is obtained by starting at the root node and follows a set of logical rules, for instance metabolite concentration cut-off values, until a terminal node is reached. Single decision trees, however, may suffer from poor performance since they are prone to overfitting, because they are based on a greedy algorithm (which may not find the overall best solution), and may be biased towards dominating classes. They are also not robust, meaning that a small change in the training data may give a different model.

A random forest85 (RF) is an ensemble classifier, based on multiple CARTs, and provides a substantial improvement in the performance. In an RF each tree is built on bootstrapped (see Section 7 for an explanation of bootstrapping) training samples. In addition, when building these trees, only a random sample of covariates is chosen as split candidate from the full set of predictors. To obtain a prediction for a new observation, the mean value of all individual predictions from each tree is used, or the majority vote, for regression and classification trees, respectively. RFs are in general easy to train, and have only two tuning parameters, the number of trees in the ensemble, and the number of predictors allowed to be chosen at each split.

Another popular tree ensemble method is boosted trees. Adaptive Boosting86 (AdaBoost) was the first successful boosting algorithm, and from it many different algorithm variants have emerged over recent years.87–89 Many of these are further improvements of gradient boosting machines (GBMs).90 In contrast to RF, in which trees are independently built on bootstrapped versions of the original data, GBMs are built by adding models (decision trees) to the ensemble in a sequential manner. Each new tree in the sequence is fit to the residuals of the previous tree, thus improving the ensemble in areas where the previous tree had the weakest performance. GBMs contain two boosting hyperparameters—number of trees and learning rate, where the latter determines the contribution of each tree in the final model—in addition to tree-specific parameters.

An ensemble of trees is no longer as easily interpretable as a single decision tree. The interpretation of the contribution of the different covariates in tree ensembles may be done through variable importance measures, which in essence sum the contribution of each covariate averaged over the ensemble. This is commonly done by permuting the values of one covariate at a time, and computing the reduction in the predictive performance. Several machine learning approaches have been tested to diagnose Alzheimer-type dementia based on blood metabolite levels, of which extreme gradient boosting (XGBoost) showed the best performance.91 Similarly, XGBoost was more successful than PLS-DA and support vector machines (SVMs) in predicting autism spectrum disorder in children.92

6.4 Support vector machines (SVMs)

An SVM is a supervised machine learning algorithm that finds a hyperplane in a \( p \)-dimensional space, where \( p \) is the number of covariates, which separates samples from different classes.93 We here give only a brief introduction to SVMs as a more comprehensive explanation of SVMs may be found elsewhere.94 Given linearly separable data with a binary response \( y_i = \{-1, 1\} \) (two-class classification), the SVM finds a weight vector \( w \) such that the boundary separating the classes is given by

\[
w^*x_i + b = 0
\]

where \( b \) is the bias. It does so by determining the rotation of the hyperplane so that the margin, given by \( \frac{2}{\|w\|} \), between the classes is maximized. Every \( x_i \) for which \( w^*x_i + b > 0 \) will be classified as \( y_i = 1 \), while every \( x_i \) for which \( w^*x_i + b < 0 \) will be classified as \( y_i = -1 \). The margin is the distance between the samples lying closest to the observation of the other group, and the thickness of the margin (ie how wide it is) may be tuned by a parameter (\( C \)) to allow for more flexibility, and hence avoid over- or underfitting. With a soft margin (low \( C \)), some misclassifications are allowed, while a hard margin (high \( C \)) allows for only a minimum number of misclassifications (NMC). SVMs may gain more flexibility in order to accommodate non-linear decision boundaries between the classes by the use of kernel. By mapping the input data into a higher-dimensional space using a kernel function, we obtain an enlarged feature space, in which a linear separation might be possible. A kernel function is simply a function that takes as inputs vectors in the original space, and returns the dot product of the vectors in an enlarged space. There is a number of choices for the kernel function, of which the most commonly used is the radial basis function (RBF), also known as the Gaussian kernel.94
Gaussian kernel, on two samples \(x_1\) and \(x_2\), is of the following format: \(K(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2}\), where \(\gamma\) lies in the interval between 0 and 1. When fitting an SVM with an RBF the parameter \(\gamma\) must be tuned. A small \(\gamma\) gives less complexity and the model may underfit, while a high \(\gamma\) gives more complexity, and may lead to overfitting. The two hyperparameters \(C\) and \(\gamma\) are interdependent, and thus they need to be optimized simultaneously.

The advantages of SVMs are that they are robust, allow for classifying multiple classes, and in general deal well with high-dimensional data. SVMs can also be employed for both classification and regression purposes. Although SVMs are robust, meaning that their performance will be similar when tested on a new independent data set, they are sensitive to outliers. Another disadvantage is that finding the right kernel and optimal choices for the parameters may be computationally intensive. Moreover, SVMs are black-box methods, meaning that the models do not provide information on which covariates are important. Despite this, SVMs might be useful if the main purpose is prediction. For example, least-squares SVM has been successfully applied to create a predictive model for diagnosis of chronic obstructive pulmonary disease using serum metabolic biomarkers obtained by NMR. Additionally, applying SVM on an NMR-based urine biomarker panel, a model could distinguish patients with Alzheimer’s disease from healthy controls with high accuracy.

### 6.5 Artificial neural networks/deep learning

An artificial neural network is an algorithm designed to emulate the way the human brain works. There are many types of artificial neural network, but common to all of them is that they consist of one or more layers of interconnected nodes (artificial neurons). These nodes are computational units, which process a weighted sum of the received input values, and propagate the result to the nodes in the next layer. Deep learning is a subset of artificial neural network algorithms, which use artificial neural networks with several hidden layers. While machine learning has widely been applied in metabolomic studies for decades, the application of deep learning has only recently emerged. The majority of applications of deep learning have been towards data preprocessing, although it has also been applied to data analysis, metabolic phenotyping, and biomarker discovery, and shows potential for multi-omic integration. A recent study used 10 publicly available metabolomic data sets to compare deep learning with six other machine learning algorithms for binary classification, and found that deep learning performed similarly to PLS-DA. However, when using deep learning there was a loss of interpretability. Another study compared deep learning with six other commonly used machine learning methods to predict the estrogen status in breast cancer patients, with deep learning giving the best results. These conflicting results indicate that the best method may depend on the data and the complexity of the problem.

As previously described, data sets in metabolomics have traditionally been of limited sizes, often with a larger number of variables than samples. Deep learning models in general require a large number of samples to be properly trained. The recent trend in metabolomics towards studies on large scale population-cohorts enabled by companies that provide fully labelled metabolomic data sets opens new opportunities for using deep learning within metabolomics. Another challenge with deep learning, similarly to SVMs, is the lack of interpretability. Pomyen et al have recently evaluated applications of deep learning in metabolomics, and have concluded that several challenges should be addressed, such as metabolome-specific deep learning architectures, dimensionality problems and model evaluation regimes, before deep learning can be effectively applied to metabolomics.

### 7 VALIDATION

Validation is of utmost importance for supervised multivariate analysis, as these models tend to overfit to the data. Overfitting means that the model has been fitted too closely to the data, including the noise. This will make the model less suitable for new data. Proper validation is therefore crucial to assess the validity of the model. Validation can be performed using a separate test set or using different types of cross-validation procedure (Figure 5). In \(k\)-fold cross-validation, the data set is divided into \(k\) sets of equal size. A model is built using \(k - 1\) sets and tested on the one set kept out of model building, and the procedure is repeated \(k\) times so that all sets are used for testing once. \(k\) can range between 2 and \(N\), with \(N\) being the number of samples, and where \(k = N\) corresponds to “leave-one-out” cross-validation. An alternative approach is “leave-\(n\)-samples-out” cross-validation, also called Monte Carlo resampling. Here \(n\) samples are randomly chosen for the test set, and the remaining samples are used to build a model. The procedure is repeated several times, and in contrast to \(k\)-fold cross-validation a sample can appear in the test set several times. Héberger and Kollár-Hunek recently compared different validation procedures, concluding that no validation variant was superior in all circumstances. Leave-one-out cross-validation is however well known to cause overoptimistic results, and should only be used for very small data sets (\(N < 20\) samples). If the data set has a hierarchical structure, such as in a longitudinal setting where each subject is measured repeatedly, validation should be performed in a way that preserves the hierarchical structure of the data in both the test and training sets.

For models where parameter tuning is part of the model optimization (such as optimizing the number of latent variables in PLS, or the number of trees in RF), it is recommended to perform twofold validation in order to achieve unbiased estimates: one for optimizing parameters and one
for assessing the validity of the final model.\textsuperscript{111,112} This can be done by (1) dividing the data into training, test, and validation sets, (2) dividing the data into training and validation sets and performing cross-validation within the training data, or (3) applying double cross-validation.\textsuperscript{113}

For classification problems, the parameters giving the best separation between the classes should be selected. Commonly used measures for optimal separation within metabolomics are the NMC,\textsuperscript{114,115} the area under the receiver operating characteristic curve (ROC AUC),\textsuperscript{115} and $Q^2$ statistics.\textsuperscript{117,118} NMC is an intuitive measure, simply counting the number of samples that are wrongly classified by the model, and can be presented as a ratio to the total number of samples (the classification error), where 0 will represent perfect classification and 1 represents complete misclassification. The ROC AUC is a measure to optimize the sensitivity and specificity of the model. $Q^2$ is a standard measure of the model fit for regression, and is based on the difference between the predicted and true output variables, or more specifically one minus the ratio of the prediction error sum of squares (PRESS: the squared sum of the predicted ($\hat{y}$) minus the true $y$ value, $\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$) and the total sum of squares (TSS: the squared sum of the true minus the mean $y$ value, $\sum_{i=1}^{N}(y_i - \bar{y})^2$).\textsuperscript{119} $Q^2$ is not as intuitive to interpret, as it can take values from 1 (perfect prediction) to $\infty$, and it is difficult to determine if the resulting $Q^2$ value of a model corresponds to good separation. Consider a classification problem using class labels 1 and -1 for Classes A and B, respectively. A Class B sample predicted as 2.5 would be correctly classified using a threshold of 0; however, it would contribute to lower the $Q^2$ value in the same way as a Class B sample predicted as 0.5. The discriminant $Q^2$ (DQ$^2$) has been described as a method to overcome this issue,\textsuperscript{120} as DQ$^2$ does not penalize class predictions beyond the class label value. Szymańska et al showed that NMC was more powerful than both $Q^2$ and DQ$^2$ for detecting small group differences using PLS-DA.\textsuperscript{121} For regression problems, the $Q^2$ value or the root mean squared error (RMSE) is commonly used to assess the model performance. The RMSE is a measure of how different the predicted value $\hat{y}$ is from the true $y$ value; more specifically, it is defined as the root of the PRESS divided by $N$, and an RMSE of 0 depicts a perfect predictive model.

There are several suggestions for how to choose the best model parameters based on the measures described above. A common procedure is to choose the model parameters (for instance the number of latent variables in PLS) giving the first minimum in classification error or RMSE on the test data for discrimination problems or regression problems, respectively. Filzmoser et al have described a method using the so-called standard error method that takes into account the variability of the prediction performance when choosing the optimal number of latent variables in PLS.\textsuperscript{113} Kvalheim et al have described a related approach using Monte Carlo resampling.\textsuperscript{122}

Bootstrapping is a resampling method commonly used for model validation. In non-parametric bootstrapping, samples are randomly chosen from the original data set with replacement, meaning that the same sample might appear several times in one selection. Bootstrapping can also be
parametric, in which case a distribution is assumed for the data, thus model parameters are estimated, and random samples are drawn from the fitted model. Bootstrapping can be used to estimate model accuracies, such as confidence intervals and variations in model accuracies, by building separate models on bootstrapped sample sets.

The significance of the final, validated multivariate model can be assessed by permutation testing. Permutation testing involves randomization, or reshuffling, of the response variable ($y$), and building a model on these permuted $y$ values. This model using permuted $y$ values will correspond to chance predictions. Repeating this procedure (typically 1000-10 000 times) generates a null distribution, that is, a distribution corresponding to the null hypothesis in statistical testing, which assumes that there is no relation between the matrix $X$ and the response $y$. The significance of the true model can be assessed by comparison with the null distribution, typically in the sense of a $p$-value for the model. An alternative approach for assessing model significance is analysis of variance of cross-validated residuals (CV-ANOVA). As in ANOVA, this test compares the fitted residuals from two models using an $F$-test. In CV-ANOVA, the fitted residuals result from cross-validation to achieve more reliable results, and an $F$-test is used to determine if the PLS model has significantly smaller residuals than the variation around the global $y$ average.

Ideally, multivariate analysis models for metabolomic data should be further validated using external data sets. External validation would allow us to assess if the results are robust across different laboratories, using different NMR instruments, when using different quantification procedures, or even different analytical platforms. External validation is so far more common in other -omic fields, such as transcriptomics, where large publicly available data sets can be easily accessed and used for validation. This possibility is limited within metabolomics due to the smaller number of publicly available data sets. Data sharing is encouraged or even required by several journals and funding agents, and would allow increased use of external validation within metabolomics, an important step for clinical translation of novel biomarkers. A suitable database is MetaboLights, the main open repository for data sharing within metabolomics, accepting data from different species and analytical techniques.

### 8 CHALLENGES AND FUTURE ASPECTS

This review has provided an overview of the current state of multivariate methods for analyzing NMR-based metabolomic data. The choice of method should be based on the sample size, the complexity of the problem, and the aim of the analysis. Unsupervised methods such as PCA, clustering methods, and t-SNE are well suited for exploratory analysis. If the aim instead is hypothesis testing, to identify specific metabolites that are associated with some outcome or characteristic (e.g., treatment, the risk of developing a disease, or survival), then traditional statistical models such as general linear (mixed) models, logistic regression, or Cox regression are commonly used. For experiments where the response variable is multivariate, methods such as ASCA, which combine latent variable analysis with GLMs, can take both the experimental design and the multivariate structure of the data into account. For predictive modeling, many options are available, and the researcher must take both the sample size and the importance of interpretability into consideration when deciding which modeling approach to take. Even though SVMs and artificial neural networks may provide models with a high predictive performance, these methods generally do not allow for biological interpretation of the discrimination model. PLS-DA often has comparable performance, while also allowing easy interpretation; however, the latent-variable-based method cannot handle complex non-linear problems. Thus, researchers are encouraged to test different approaches.

When performing NMR-based metabolomics, especially as a part of medical research, the aim is often insight into disease mechanisms. This may to some extent be achieved by investigating which covariates have the highest importance in a discrimination model, or have a statistically significant association with a disease. Several methods also exist for pathway analysis of metabolomic data, such as MetaboAnalyst or MetPA, which has not been covered in this review. Pathway analysis allows for the identification of altered metabolic pathways, and provides a useful tool for interpreting metabolomics findings. A limitation for using pathway analysis, especially for NMR-based metabolomics, is the limited number of detected metabolites. A high number of metabolites often interact together in one metabolic pathway, many of which cannot be detected through NMR spectroscopy. However, pathway analysis of NMR-based metabolomics can provide hypothesis for mechanistic insight. Another possible tool for modeling the complex interdependencies between metabolites is the use of causal networks. Because it is typically not feasible to experimentally manipulate specific metabolites, interactions between metabolites, as well as between metabolites and other biological variables, must typically be inferred from observational data. Moreover, most machine learning models do not by themselves provide insights into the causal relationships between the variables in the model, and therefore cannot identify which biological variables are the most suitable targets for interventions. Causal network modeling techniques, such as structural equation modeling (SEM) or a method developed for inferring the topology of -omic data (iTOP), involve leveraging assumptions about causal relationships between variables in order to estimate causal effects based on observational data, and to determine whether they are compatible with the posited causal model. These methods have also been applied to metabolomic data. Usually the causal model must be pre-specified by the researcher based on domain knowledge. However, as the causal model is generally not known, various methods for generating causal networks algorithmically have also been developed. Although these methods have limitations, and are so far not much used within metabolomics, causal network analysis may provide a useful framework for understanding the causal role of metabolites in biological systems.
A single -omic technique, such as metabolomics, will detect alterations in biomolecules of only one -omic level. This allows identification of only a subset of the components of a pathway, and may be insufficient for understanding the disease mechanisms. Multi-omics is an approach where data sets of different -omic levels are combined, such as the genome, transcriptome, proteome, or microbiome, and is becoming a critical component of metabolomic research. The main motivation for multi-omic studies is often to obtain a comprehensive understanding of the biological system under study; however, they may also be used to associate the -omic-based molecular measurements with a clinical outcome of interest. Although there is no standardized way of combining -omics data for analysis, a range of different computational tools for integrating metabolomic data with different -omics have been developed. The -omic levels may be combined in terms of eg significant statistical associations, biomarker discovery, or correlation analysis, eg to identify common and distinct components. Multi-omic integration has eg been applied to identify a survival signature and several potential therapeutic targets for hepatocellular carcinoma. Here, the -omic levels were combined by employing biological network analyses. For the integration of metabolomic data with other -omic data, deep learning methods show potential. However, multi-omic integration still faces several challenges, including data quality, and insufficient cross-communication between different research communities in different -omic fields.

Available tools for analyzing metabolomic data ranges from basic point-and-click data analysis to advanced coding software. Due to freely available algorithms developed specifically for preprocessing of NMR metabolic data, such as peak alignment algorithms or MetaboLab, MATLAB (MathWorks) has been popular for analyzing NMR metabolic data. The main disadvantage is the need for a MATLAB license, and the need for expensive toolboxes, such as PLS Toolbox. R is a programming environment dedicated to statistical computing and graphics. Several built-in and publicly available packages cover a wide variety of statistical analysis methods, including deep learning. Python is now the fastest-growing major programming language, and is currently the most popular programming language for developing deep learning methods. Libraries developed specifically for data analysis, such as Scikit-learn and TensorFlow, allow for a wide variety of bioinformatics. There are also a number of easy-to-use proprietary tools for analyzing metabolomic data, such as MetaboAnalyst, SIMCA, or Unscrambler. These allow for multivariate analysis of NMR-based metabolomic data for researchers inexperienced with programming. It is however very important that researchers familiarize themselves with the software in order to apply it correctly. If the researcher is confident with programming, applying a programming language for data analysis provides much more flexibility. This also makes it easier to share how data analysis has been performed with other researchers by publishing custom-made scripts together with scientific papers, which is strongly recommended.

To conclude, several different methods are available for multivariate analysis of metabolomic data, all with their pros and cons. The choice of method should be based on the research question, the study design, and the sample size, and a combination of several methods may be appropriate to fully extract the meaningful information from a metabolomic data set.

ACKNOWLEDGEMENTS
This work was supported by The Norwegian Cancer Society (grants 6834362 and 202021), Stiftelsen Dam (Project 2020/FO298770), the Joint Research Committee between St. Olavs hospital and the Faculty of Medicine and Health Sciences, NTNU (28346 and 28328), and the Swiss National Foundation (grant p400PM_194482).

DATA AVAILABILITY STATEMENT
Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID
Julia Debik https://orcid.org/0000-0002-2765-8043
Guro F. Giskeødegård https://orcid.org/0000-0003-2157-8824

REFERENCES
1. Giskeødegård GF, Madssen TS, Euceda LR, Tessem M-B, Moestue SA, Bathen TF. NMR-based metabolomics of biofluids in cancer. NMR Biomed. 2018;32(10):e3927. https://doi.org/10.1002/nbm.3927
2. Zhang A, Sun H, Wang P, Han Y, Wang X. Modern analytical techniques in metabolomics analysis. Analyst. 2012;137(2):293-300. https://doi.org/10.1039/c1an15605e
3. Wurtz P, Kangas AJ, Soininen P, Lawlor DA, Davey Smith G, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: a primer on -omic technologies. Am J Epidemiol. 2017;186(9):1084-1096. https://doi.org/10.1093/aje/kww016
4. Hao J, Liebeke M, Astle W, De Iorio M, Bundy JG, Ebbels TM. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. Nat Protoc. 2014;9(6):1416-1427. https://doi.org/10.1038/nprot.2014.090
5. Mestrelab Research Manova. https://mestrelab.com/software/manova/. Accessed November 27, 2020.
6. Deelen J, Kettunen J, Fischer K, et al. A metabolomic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. Nat Commun. 2019;10(1):3346. https://doi.org/10.1038/s41467-019-11311-9
7. Vettukatti R. Preprocessing of raw metabonomic data. Methods Mol Biol. 2015;1277:123-136. https://doi.org/10.1007/978-1-4939-2377-9_10
8. Euceda LR, Giskeødegård GF, Bathen TF. Preprocessing of NMR metabolomics data. Scand J Clin Lab Invest. 2015;75(3):193-203.
Chung NC, Mirza B, Choi H, et al. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. Methods. 2019;166:66-73. https://doi.org/10.1016/j.ymeth.2019.03.004

Asakura T, Date Y, Kikuchi J. Application of ensemble deep neural network to metabolomics studies. Anal Chim Acta. 2018;1037:230-236. https://doi.org/10.1016/j.aca.2018.02.045

Date Y, Kikuchi J. Application of a deep neural network to metabolomics studies and its performance in determining important variables. Anal Chem. 2018;90(3):1805-1810. https://doi.org/10.1021/acs.analchem.7b03795

Mendez KM, Reinke SN, Broadhurst DI. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. Metabolomics. 2019;15(12):150. https://doi.org/10.1007/s11306-019-1612-4

Alakwaa FM, Chaudhary K, Garmire LX. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. J Proteome Res. 2018;17(1):337-347. https://doi.org/10.1021/acs.jproteome.7b00595

Héberger K, Kollár-Hunek K. Comparison of validation indices by sum of ranking differences and ANOVA. J Chemom. 2019;33(6):e3104. https://doi.org/10.1002/jchem.201903104

Rodriguez-Pérez R, Fernández L, Marco S. Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: a systematic study. Anal Bioanal Chem. 2018;410(23):5981-5992. https://doi.org/10.1007/s00216-018-1217-1

Westerhuis JA, Hoefsloot HJC, Smit S, et al. Assessment of PLSDA cross validation. Metabolomics. 2008;4(1):81-89. https://doi.org/10.1007/s11306-007-0099-6

Euceda LR, Haukaas TH, Bathe TF, Giskeødegård GF. Prediction of clinical endpoints in breast cancer using NMR metabolic profiles. In: Cancer Systems Biology: Methods and Protocols. New York: Springer; 2018:167-189.
How to cite this article: Debik J, Sangermani M, Wang F, Madssen TS, Giskeødegård GF. Multivariate analysis of NMR-based metabolomic data. *NMR in Biomedicine*. 2022;35(2):e4638. doi:10.1002/nbm.4638