UMEÅ UNIVERSITY

Multivariate integration and visualization of multiblock data in chemical and biological applications

Tomas Skotare

Department of Chemistry, Umeå University
Umeå, Sweden, 2019

Dissertation for PhD
ISBN: 978-91-7855-069-2
Electronic version available at: http://umu.diva-portal.org/
Printed by: VMC-KBC, Umeå
Umeå, Sverige, 2019

Till πa, $\sim$ och μ.

Contents

Abstract

Thanks to improvements in technology more data than ever before is generated in almost all fields of science and industry. The data is analyzed to hopefully provide valuable information and knowledge about a product or process, such as how to improve the quality of a manufactured product.

Analysis of collected data is often performed on a single dataset or data source at a time. In this thesis, I have focused on multiblock analysis, a concept that includes multiple sources or data blocks. Analogous to how the human senses combine to let us experience the world around us, multiblock analysis integrates multiple data sources, providing a fuller examination of the product or process under study.

My thesis introduces Joint and Unique Multiblock Analysis, JUMBA, a complete analysis workflow for data integration. I describe each step of JUMBA, including data pre-treatment, model building and validation as well as model interpretation. Special focus is put on several newly developed visualizations for model validation and interpretation to make it as easy as possible to draw conclusions from the analysis.

By reading my thesis, the reader will gain a working understanding of the process of performing multiblock analysis, including solutions to common problems that are often encountered.

Sammanfattning på svenska

Tack vare tekniska framsprång genereras det idag stora mängder data inom forskning och industri. Genom att analysera sådan data kan det i slutändan leda till att värdefull kunskap om en produkt eller process erhålls och kvaliteten på de studerade produkterna därmed kan ökas.

Analysen av data sker ofta på en enda datakälla, som då representeras av en matris, även kallat ett *datablock*. I denna avhandling har jag istället fokuserat på koncept som kan analysera flera datakällor samtidigt och integrera dessa. I likhet med hur människans sinnen låter oss uppleva världen runt omkring medför integrerandet av flera datakällor att undersökningen av en produkt eller process blir mer omfattande.

I min avhandling introduceras arbetsflödet JUMBA (*Joint and Unique Multiblock Analysis*, eng), som är ämnat för att utföra en fullständig integration av data. Jag beskriver varje enskilt steg av JUMBA, allt från förbehandling av data till byggande och validering av modeller samt deras tolkning. Jag har lagt särskild vikt vid att beskriva flera nyskapade typer av visualiseringar som underlättar att korrekta slutsatser kan dras från analysen.

Jag hoppas att läsaren av min avhandling kommer få förståelse för hur man utför analys av flera datablock och denne hittar även lösningar på problem man normalt sett kan ställas inför vid genomförandet.

Papers included in this thesis

This thesis is based upon the following appended papers, which are referred to using the corresponding Roman numerals. For convenience, Papers I-IV sometimes refer to the studies reported in them as well as the papers per se.

Paper I *Visualization of descriptive multiblock analysis*,
Tomas Skotare, Rickard Sjögren, Izabella Surowiec, David Nilsson, Johan Trygg
In press, Journal of Chemometrics, 2018
DOI: 10.1002/cem.3071

Paper II *Joint and Unique Multiblock Analysis for Integration and Calibration Transfer of NIR Instruments*,
Tomas Skotare, David Nilsson, Shaojun Xiong, Paul Geladi, Johan Trygg
Analytical Chemistry, 2019, 91 (5), pp 3516–3524
DOI: 10.1021/acs.analchem.8b05188

Paper III *Multi-Tissue Metabolomics Integration Utilising Data Integration Methods and Hierarchical Modelling*,
Frida Torell[†], **Tomas Skotare**[†], Johan Trygg
Submitted to journal, PLOS ONE

Paper IV *Joint and Unique Multiblock Analysis of biological data – multiomics malaria study*,
Izabella Surowiec, **Tomas Skotare**, Rickard Sjögren, Sandra Gouveia-Figueira, Judy Orikiiriza, Johan Normark, Sven Bergström, Johan Trygg
Accepted, Faraday Discussions, 2019
DOI: 10.1039/C8FD00243F

[†]These authors made equal contributions

Papers not included in the thesis

The author of this this thesis also contributed to the following papers during the course of his doctoral studies, but they are not appended.

Paper V *Multivariate patent analysis—Using chemometrics to analyze collections of chemical and pharmaceutical patents*,
Rickard Sjögren, Kjell Stridh, **Tomas Skotare**, Johan Trygg
Journal of Chemometrics, 2018,
DOI: 10.1002/cem.3041

Paper VI *A multi-omics approach reveals function of Secretory Carrier-Associated Membrane Proteins in wood formation of Populus trees*,
Ogonna Obudulu, Niklas Mähler, **Tomas Skotare**, Joakim Bygdell, Ilka N. Abreu, Maria Ahnlund, Madhavi Latha Gandla, Anna Petterle, Thomas Moritz, Torgeir R. Hvidsten, Leif J. Jönsson, Gunnar Wingsle, Johan Trygg, Hannele Tuominen
BMC Genomics, 2018, 19:11
DOI: 10.1186/s12864-017-4411-1

Paper VII *OnPLS-Based Multi-Block Data Integration: A Multivariate Approach to Interrogating Biological Interactions in Asthma*,
Stacey N. Reinke, Beatriz Galindo-Prieto, **Tomas Skotare**, David I. Broadhurst, Akul Singhania, Daniel Horowitz, Ratko Djukanović, Timothy S.C. Hinks, Paul Geladi, Johan Trygg, Craig E. Wheelock
Analytical Chemistry, 2018, 90 (22), pp 13400–13408
DOI: 10.1021/acs.analchem.8b03205

Abbreviations

CV	Cross Validation
DA	Discriminant Analysis
DoE	Design of Experiments
IR	Infrared Spectroscopy
JUMBA	Joint and Unique Multiblock Analysis
LV	Latent Variables
MIA	Multivariate Image Analysis
MLR	Multiple Linear Regression
MSC	Multiplicative Signal (or Scatter) Correction
MVA	Multi-Variate Analysis
NIPALS	Non-linear Iterative Partial Least Squares
NIR	Near Infrared Spectroscopy
OSC	Orthogonal Signal Correction
O-PLS	Orthogonal Projections to Latent Structures
O2PLS	Bidirectional O-PLS
OnPLS	Multiblock O-PLS
PCA	Principal Component Analysis
PLS	Projections to Latent Structures
Q^2	Goodness of Prediction
R^2	Goodness of Fit
RMSEP	Root Mean Square Error of Prediction
SNV	Standard Normal Variate
SS	Sum of Squares
SVD	Singular Value Decomposition
TOP	Transfer using Orthogonal Projections
UV	Unit Variance

Chapter 1

Fundamental concepts in data analysis

> If I have seen further it is by standing on the shoulders of Giants.
>
> *Sir Isaac Newton*

The amount of data generated in almost all fields of life has increased manyfold in the past decades, often thanks to computerization in almost all areas of industry and science. It is getting increasingly cheaper to measure and store data related to almost any type of process. Hence, there is currently more data than ever, but it needs to be analyzed and understood before it adds value, i.e. converted into information that can provide answers, for example to social, biological or chemical questions.

There are a multitude of problems facing anyone aiming to extract the information, however. Data can contain large amounts of noise, i.e. unstructured data that is not related to the focal problem. Such noise can have several origins, such as invalid sample treatment before measurement or instrumental inaccuracies.

Another type of interference can be in the form of signal that is not related to studied process but rather to some other process or question. Data can also be skewed or biased, for example with regards to gender or age distributions in clinical studies. Missing data can also be a problem, especially if such data is not missing at random but rather missing due to some other factor, such as one specific group avoiding answering certain questions in a questionnaire.

Furthermore, data from a single source (e.g. an instrument) is often unlikely to completely explain the variation in all the potentially relevant features of samples. An obvious example is that humans and other animals acquire, process and integrate complementary information from multiple senses (taste, sight, touch, smell, and sound) in their interactions with the world. Merging or *integrating* multiple sources of data has more recently become increasingly relevant as the total amount of data grows and our ability to measure multiple facets samples increases. [1].

Such integration of data from multiple sources is the main focus of this thesis, and the information in this chapter provides foundations for understanding the following chapters.

1.1 Chemometrics

Chemometrics [2, 3] can generally be considered a field within the broader scope of *machine learning*. Machine learning is a very broad field which general consists of different algorithms and statistical models which aim to extract information from almost any type of data, without being explicitly told exactly how.

These aspects of machine learning are also included in chemometrics, and they share many ideas and overall properties. As implied by the name, however, chemometrics is a field in which various tools and ideas are applied to extract information from chemical systems. As summarized by Svante Wold [4]:

> Chemometrics is the branch of chemistry concerned with the analysis of chemical data (extracting information from data) and ensuring that experimental data contain maximum information (the design of experiments).

Both parts (design of experiments and extraction of data) are briefly covered in this chapter.

1.2 Design of experiments

Design of experiments (DoE) [5, 6], sometimes called *experimental design*, is a concept that permeates chemometrics. The core of DoE is to gain as reliable and representative data as possible with respect to a specific question while minimizing the number of experiments required to do so.

Acquisition of representative data is absolutely crucial for successful analysis. Also, reducing the number of experiments has obvious benefits when there are cost, time or availability of material constraints. The aim of DoE is to optimize *factors* (experimental settings such as concentration, temperature or sample treatment) with regards to some desired effect, such as maximizing efficiency or minimizing cost [5, 6].

Using design of experiments provides the best possible conditions for further analysis as the resulting data will be reliable and contain variation relevant to the intended goal.

1.3 The concept of models

To address questions concerning a complex system, a *model* of the system can be used. Models are simplified reflections of reality and are not expected to perfectly represent the real world, but are still useful and can be used to both interpret and predict real-world phenomena [7, 8, 9]. There are various types of models, and their foundations may range from limited information about the phenomena they are intended to represent to 'deep knowledge' or first principles of the field, such as fundamental physical laws governing phenomena such as energy balances and heat transfer [10]. This thesis focuses on the former type: empirical models created from data generated by measuring samples. Models based on first principles generally have very broad scope, i.e., they enable accurate extrapolation, while empirical models at best have verifiable validity within the experimental space covered by the samples.

1.4 Data structures, matrices and data blocks

Within most fields of science, conducted experiments produce some way to characterize observations (samples, individuals, etc.) using some sort of variable (property or characteristic) such as temperature, density and so on. Measured observations and variables can be combined into a data matrix or *block*. A block is an $m \times n$ matrix, where each row (m) represents a single observation (e.g. sample, time point, etc.) and each column (n) represents a measured property (e.g. wavelength, weight, concentration). Consequently, measurements of height, weight and age of 10 patients would result in a matrix of size 10×3. Each entry in the block is represented by a numerical value, and in this thesis all analyzed data are assumed to be in this format. If the measured property is categorical or logical (true or false), the values must first be converted into numbers, e.g. 1 and 0 for true and false, respectively. If there are multiple options, the separate values are assigned a new column with 1 where applicable and 0 for the other columns, so-called *dummy variables* [11].

By convention, when a single block is being analyzed it is called X, and consists of the measured observations and variables. If another block is involved it is called Y, which often consists of response variables and generally the goal of the analysis is to explain their responses, i.e., identify the underlying causes and factors responsible for observed variation in the Y variables, and if possible predict their variation in other cases. For example, if X consists of measurements from a

near-infrared instrument, Y may be the concentrations of some constituent(s) in the analyzed samples.

1.5 Supervised and unsupervised data analysis methods

For the methods used in this thesis, the concept of unsupervised and supervised methods is crucial. Unsupervised methods only consider the input data and without further guidance seek signals, trends and/or clusters in the data. They include Principal Component Analysis (PCA) [12], described in *1.8.1*, *Principal Component Analysis*, clustering methods (e.g. hierarchical cluster analysis [13, 14] and K-means clustering [15]) and autoencoders [16].

Supervised models are instead guided towards either regression or classification. Regression models predict quantitative numerical values such as quantities or size, while classification models predict qualitative discrete or categorical output, for example gender or type. For supervised methods to work, each sample must have a corresponding label or *response* which should be predictable given the input data. A dataset or block pertaining to a set of samples could consist of near-infrared spectra of apples, and the responses could be their sugar contents (modelled by regression) [17] or information on whether they are bruised or not (classification) [18]. Examples of supervised methods include use of support vector machines (SVM) [19], various artificial neural networks (e.g. convolutional neural networks (CNN) [20]) and Projection to Latent Structures (PLS) [21].

The type of method that should ideally be applied always depends on the problem that needs to be solved, but it is often preferable to initially analyze data using an unsupervised method, such as PCA (see *1.8.1*, *Principal Component Analysis*), which has no inherent bias. Some suitable cases for using the methods are described below.

1.6 The concept of latent variables and their models

A fundamental principle in chemometrics is the idea of latent variables. These are 'hidden' variables which are not directly observed but can be inferred from a combination of other variables that can be measured [7, 11, 22]. An example of a latent variable is the physical health of a patient, which can be estimated by measuring related metrics such as cholesterol level, weight and blood pressure.

The concept of latent variables enables us to reduce a complex system characterized by a large number of measured variables in terms of a lower number of latent variables. During analysis, latent variables replace the original

variables, reducing complexity and simplifying analysis while still retaining most of the original content in the data.

The methods described here are linear methods, i.e. an implicit assumption is that the latent variables have linear relationships with corresponding original variables.

In contrast, nonlinear methods, as the name implies, are used to study nonlinear relationships. Examples of such methods include spline-PLS (SPL-PLS) [23] and various techniques involving use of neural networks [24], but they are not addressed in this thesis. However, even if the analyzed block(s) contain(s) data with nonlinear relationships, it is usually possible to change or transform them into relationships that are sufficiently close to linear to allow use of linear methods. For more details see *1.7, Data pre-treatment prior to modeling*.

The linear methods in this thesis involve use of so-called *components* to represent linear latent variables. Essentially, variables measuring the same thing in a focal set of samples are merged into a single variable, or score vector (denoted t), which has one value per observation in the original block. Contributions of variables to the score are known as loading vectors (denoted p), or simply loadings, with one value per original variable. Note that there is normally no guarantee that a single component will correspond to a single latent variable, although some methods try to find such correspondence (e.g. single-Y O-PLS, see *1.8.3, Orthogonal Projections to Latent Structures*).

The merging of variables relates to the problem of *multicollinearity*. Multicollinearity happens when variables have a strong linear relationship with another, i.e. *correlate* strongly and are not independent. It will inevitably occur when the number of variables is larger than the number of observations (i.e. $n > m$) [25]. However, even if $n < m$, the variables measured on real-world samples often intrinsically correlate [26] as they (indirectly) measure the same latent variable. The presence of such correlating variables prohibits use of some regression methods, and introduces problems in assessing the relative importance of variables with other methods [25, 27]. Solutions in these cases include removal of affected variables until no multicollinearity remains, or summation of variables that correlate. However, multicollinearity has much less impact on results obtained using other methods, including the latent variable methods used in the studies underlying this thesis [28, 26]. In fact, principal component analysis, a latent variable method described below, can be used as a pre-treatment to solve multicollinearity when other methods are used [29].

1.7 Data pre-treatment prior to modeling

Sometimes data must be modified to improve its suitability for handling by the methods described here as they are based on minimizing errors in a least-squares sense. In most cases, the objective of the modification is to allow each variable to have equal influence on the final model, and this will not be the case if some

variables have much higher values or much higher spread than the others. For example, if weight is measured in grams and height is measured in meters, reasonable values for a person in a sampled population may be 80000 g and 1.8 m, respectively. If the data are not modified, the weight will completely dominate a height and weight model of the population simply because its values are larger.

To solve these problems, data are generally pre-treated to give each variable equal influence. Two common methods are column-centering or *centering* and scaling to unit variance (UV), sometimes called standard deviation scaling. UV scaling means dividing values of each variable by its standard deviation, causing each variable to have the same spread, regardless of the original scale. To column-center variables, the mean of each variable is subtracted from all of its values. Applying both methods results in the mean of each variable being 0, and its standard deviation 1, a combination often called autoscaling, as visualized in **Figure 1.1**.

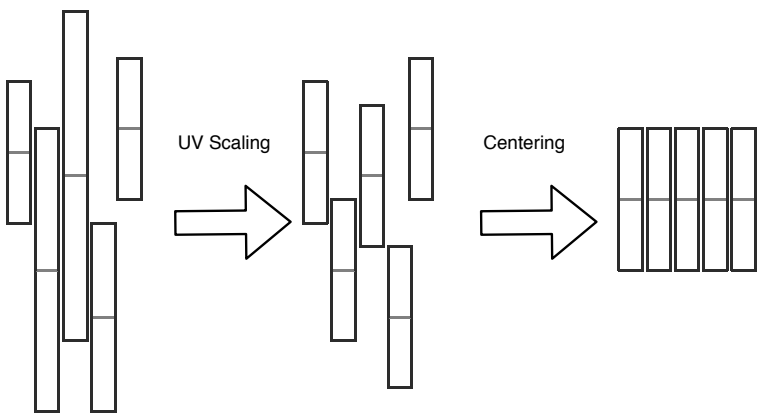


Figure 1.1: Representation of how variables are scaled using a combination of UV scaling and column-centering. Each box represents a variable, with the central line representing the mean and the length between the middle and each end the standard deviation. Initially (left) the variables differ in terms of both mean and standard deviation. After applying unit variance (UV) scaling, the means are unchanged but the standard deviations of all the variables are the same (middle), and after column-centering both their means and standard deviations are equal (right).

Data are almost always column-centered because otherwise the first component will largely reflect the average point of each variable relative to the origin, essentially centering the data, rather than the greatest variation. UV scaling can however be omitted in cases where all variables are known to be on the same scale and the loadings of the model should be on the original scale, e.g. when analyzing spectral data.

Other pre-treatments that may be applied include various transformations of non-linear data that increase their approximation to linearity, e.g. log

transformation [30], used in *Paper IV*, and square root transformation [12]. A combination of components may be able to explain non-linear data, but not necessarily [7], therefore these pre-treatments are often preferable as they help reduce model complexity.

More specialized pre-treatments include spectra-specific transformations, such as several intended to remove scatter effects in NIR spectra, for example Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) correction [31]. The latter was used in *Papers I* and *II*. Both approaches are useful for correting slope and baselines of spectra, often caused by physical or instrumental variations. Another example is cross-contribution compensating multiple standard normalization (CCMN) [32] used in *Paper IV* to remove run order and batch effects.

1.7.1 Separating interfering variation

Real-world datasets often contain variation that is related to the focal phenomenon/process and other variation that is not related to it. The unrelated, or *orthogonal* variation, can interfere with the interpretation of models. Orthogonal variation may, for example, be caused by differences in the gender or age of patients, when the aim of the study is to find variation related to a disease. Even if such variation has structure, if it is unrelated to the response it should be analyzed separately. For this reason, solutions that separate orthogonal variation have been developed with the intention to facilitate interpretation of models.

An example is orthogonal signal correction, OSC [33]. OSC is applied to the data before analysis to remove orthogonal variation. After the orthogonal variation has been removed, the remaining variation is then analyzed using conventional methods such as PLS (described below).

1.8 Common modeling methods

This section describes several of the most commonly used models within the field of chemometrics.

1.8.1 Principal Component Analysis

Principal Component Analysis (PCA) [12, 34] is an unsupervised method, i.e. it focuses solely single block denoted X. PCA is widely used in chemometrics and is often considered the "workhorse" [35, 36] that is used as the first step in data analysis.

The aim of PCA is commonly to identify outliers in the data, groupings and trends between both observations and variables.

PCA does this by decomposing X into several components that explain successively decreasing amounts of variation, essentially explaining the many variables in X using a smaller set of new variables. Because of this feature, PCA can be used as a dimensionality reduction technique [37, 38, 39]. The feature also allows it to be resistant to problems with multicollinearity (see *1.6, The concept of latent variables and their models*) as similar variables are combined, as long as a reasonable number of components are extracted (see below).

The decomposition of the original block of data X can be written as:

$$X = TP^T + E \tag{1.1}$$

where T represents a matrix of score vectors, P represents a matrix of loading vectors and E is the residual. Note that the scores values are constrained to be orthogonal, i.e. unrelated to each other. The residual represents the variation in the data that could not be extracted into components, and is usually referred to as unstructured variation or noise. The PCA decomposition is visualized in **Figure 1.2**.

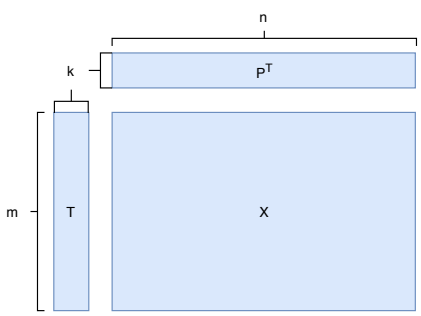


Figure 1.2: Decomposition of X of size $m \times n$ by PCA using k components. The scores (T) represent the observations in the original block, while the loadings (P) represent the original variables. The scores are of size $m \times k$ while the loadings are of size $n \times k$. Note that the loadings have been transposed for this visualization.

Extracted components can be visualized and analyzed separately, providing insight into both the relationships between the observations in the original data, and between the original variables.

In PCA, the maximum number of components that can be extracted from X corresponds to the number of linearly independent columns or rows there are in X, i.e. the *rank* of X. However, extracting the maximum number of components is not desirable, except in special cases, as it leads to overfitting [40]. As components are extracted in order of the amount of variation they explain, extraction generally stops when further components explain too little variation. The threshold for this is case-specific, nontrivial to determine, and often done using some sort of model validation technique (see *1.9, Model metrics and validation*). Recent literature has described additional methods to determine the ideal number of components to extract [41].

1.8.2 Projections to Latent Structures

Projections to Latent Structures (PLS), originally known as Partial Least Squares [21, 42] is a supervised linear regression method, capable of finding the linear relationship between two blocks: X (sometimes called the *predictors*) and Y (the *responses*). PLS is widely used to predict values of Y given X in diverse industrial, medical, chemical and many other fields. This often involve cases where X can be easy to measure but Y is expensive, slow or in some other way cumbersome to determine [42].

In PLS, similarly to PCA, each block is decomposed as shown in Equation (1.2).

$$\begin{aligned} X &= TP^T + E \\ Y &= TC^T + F \end{aligned} \tag{1.2}$$

Here, T is a matrix of score vectors for X, while P and C are the matrices of loading vectors for X and Y, respectively.

The PLS algorithm works by finding the underlying structure in both X and Y, and seeking the largest *covariance* between them. Finding the underlying structure in each block that maximizes the covariance leads to the ability to predict unknown values of the response Y for a measured sample in X. As in PCA, the loadings can be analyzed to determine which original variables in X are most strongly related to Y.

A special case of PLS, PLS - Discriminant Analysis (PLS-DA), is intended for *classification*, i.e. determine the class of each observation [43]. This is useful in cases where the response values are not numbers but categories, for example gender or country of origin.

1.8.3 Orthogonal Projections to Latent Structures

Orthogonal Projections to Latent Structures, or O-PLS [44], is an extension of PLS, where an orthogonal filter (similar to OSC) is added to the main PLS algorithm. This has the benefit of separately extracting the orthogonal information, i.e. the information that is not related to the response of interest, while retaining the predictive ability of regular PLS [45]. The decomposition is thus:

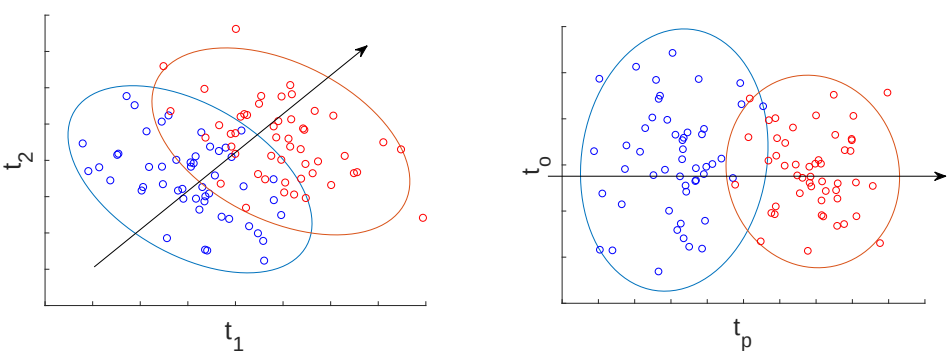
$$\begin{aligned} X &= T_o P_o^\mathsf{T} + T_p P_p^\mathsf{T} + E \\ Y &= T_p Q^\mathsf{T} + F \end{aligned} \tag{1.3}$$

While the decomposition of Y is equivalent to PLS, X is instead split into the predictive scores and loadings (T_p and P_p) and the orthogonal scores and loadings, T_o and P_o. The different parts can be visualized separately to obtain insight into the different types of variation in the data. The predictive part contains variation

related to Y, while the orthogonal part contains variation unrelated (orthogonal) to Y. The difference is visualized in **Figure 1.3**, which shows that PLS uses a mix of two components to determine class, but a single predictive component is sufficient for O-PLS. In the example, we can see that the orthogonal component (the Y axis, marked t_o) contains only variation that is unrelated to the problem at hand. Having separated the variation, conclusions can be drawn regarding the relevance of each variable. For example, as the predictive component only contains variation relevant to the study, the most important variables can be determined by interpreting the predictive loadings P_p, using for example an S plot or shared and unique structures (SUS) plot [46, 47]. Similarly, the orthogonal loadings P_o can be interpreted to determine which variables are not relevant to the study but rather to some other process or phenomena. By combining the information on variable relevancy from both sources, variables that are irrelevant to the study can be determined, i.e. variables that are not used in any direction (e.g. noisy) or variables that provide no predictive benefit. Such variables could then be omitted from future studies.

Extracting the orthogonal variation separately ensures that a maximum of one predictive component can be extracted in the case that there is only one response vector y, and thus all response-relevant information is found in a single predictive component. This property of O-PLS further simplifies data analysis.

Like PLS, O-PLS can also be used for discriminant analysis and is then called OPLS-DA [48].



(a) PLS (b) O-PLS

Figure 1.3: Visualized difference between PLS (a) and O-PLS (b) using identical simulated data. Two classes are marked using different colors (red and blue), and the direction of prediction is shown using an arrow. We can see that PLS uses both components (t_1 and t_2) to determine the class, but O-PLS can remove the orthogonal variation, so thus a single predictive component (t_p) is sufficient to determine the class.

1.8.4 O2-PLS

A further development of PLS and O-PLS, bidirectional orthogonal PLS (O2-PLS) [49], is a symmetric method. This means that instead of modeling the one-way relationship of Y to X, as in (O-)PLS, the relationship is modeled in both directions and neither block is the 'primary' block.

The variation in each block is separated into unique variation, which only resides in one block, and the shared variation, which is shared between the blocks and the residual, as seen in Equation (1.4).

$$\begin{aligned} X &= X_J + X_U + E \\ Y &= \underbrace{Y_J}_{\text{Joint}} + \underbrace{Y_U}_{\text{Unique}} + \underbrace{F}_{\text{Residual}} \end{aligned} \tag{1.4}$$

Dividing the variation in each block into different parts simplifies analysis, providing greater understanding of the types of variation in the data. O2-PLS has been applied successfully in several studies, e.g. to integrate omics datasets such as NMR [50, 51, 52].

1.9 Model metrics and validation

Model validation is an essential step in data analysis to ensure that a model is reliable. To facilitate validation, numerous metrics have been developed that indicate the quality of models' performance. Generally, these metrics are intended to describe how well a model describes the data it is generated from, how well the model describes future data, or a combination of both.

If too little of the relevant variation in the data is explained by the model it might be underfitted, meaning it is too simple to properly represent the data. In contrast, an overfitted model is overly complex, fitting the data excessively well and generally including noise. In most cases a model should be complex enough (i.e. have enough components) to properly represent the data while avoiding excessive modeling noise. Modeling the noise impairs predictions of observations outside the training set as they will not contain the same noise, so the predictions will be inaccurate.

1.9.1 Sum of squares

The models described here are based on minimizing errors in a least-squares sense, first the sum of squares (SS) metric must be introduced. To calculate SS for a matrix X, the matrix X is first column-centered (See *1.7, Data pre-treatment prior to modeling*). Each element in the column-centered matrix X is then separately squared and the resulting products are summed, resulting in the SS value. So, $SS = \sum (X - \bar{X})^2$, where X is the matrix and $\bar{X}$ is the column mean of X. This value is often specifically labelled the *total sum of squares*, but in this thesis if

not otherwise specified, SS refers to the total sum of squares. Dividing SS by the degrees of freedom in X results in the variance σ^2, and SS by itself is a measure of the amount of variation in the whole matrix X.

1.9.2 Coefficient of determination (R2)

When assessing the performance of models described here, a common metric is the coefficient of determination, R^2 ("R-squared"), or simply R2, which is the fraction of variation that a model explains (in a least-squares sense), with values ranging from 0 to 1. The calculation of R2 is shown in Equation (1.5). In this equation, the *predicted block* refers to the reconstructed block being measured e.g. $X_p = TP'$ for PCA.

$$R^2 = 1 - \frac{SS_p}{SS_o} \tag{1.5}$$

where: SS_p = Sum of squares of the predicted block
SS_o = Sum of squares of the original block

This equation can be applied to any block included in the analysis, such as R2X for X or R2Y for Y. It can also be calculated for a component or for the whole model. If the variables measured are relevant to the study, components explaining the largest percentage of the sum of squares are more likely to be relevant to the analysis. If some variables are not relevant to the study, the largest components may explain some unrelated process or phenomena.

There is also no set cutoff point at which a component is no longer relevant, as that depends largely upon the signal to noise ratio in the data. However, it should to be stressed that R2 is not a measurement of how well a model can predict future values, but strictly a metric of how well the model fits the data used to construct it.

1.9.3 Cross-validated coefficient of determination (Q2)

Another model metric is Q^2 ("Q-squared") or simply Q2, which indicates a model's predictive ability. Q2 often used in combination with R2 for determining the optimal number of components, and the two are expected to have similar values, with Q2 being slightly lower than R2 but still above zero.

Q2 is calculated using cross-validation, i.e. dividing the observations in X into several parts, often 5-7 [11]. Each part is excluded once, and a model is built using the remaining parts. The model based on the remaining observations is then used to predict the excluded observations. When all observations have been excluded and predicted once, the predicted observations can be combined into a predicted block, labelled X_p, which is of the same size as the original block. The Predicted

Residual Error Sum of Squares (PRESS) value can then be calculated, as shown in Equation (1.6).

$$PRESS = \sum (X - X_p)^2 \tag{1.6}$$

where: X = The original block
X_p = The predicted block

PRESS describes the squared sum of the predictive errors. As this metric is described in relation to the size of the original variation, it can be difficult to compare. Thus it is usually represented instead in the form of Q2 as shown in Equation (1.7).

$$Q^2 = 1 - \frac{PRESS}{SS} \tag{1.7}$$

where: $PRESS$ = The predicted residual error sum of squares
SS = The sum of squares of the original data

Calculating Q2 normalized PRESS values, with 1 representing a perfect model. Zero or negative values are possible, and indicate that the model have no predictive ability. Possible reasons for negative Q2 values include lack of relation between the measured data and response, or numbers of extracted components being either too low or too high. Q2 values may also be low even for a model of a well-designed dataset containing relevant information, if each iteration of cross-validation removed observations containing unique information that was impossible to predict using the remaining observations. In such cases, cross-validation will give artificially low results.

On the other hand, if the dataset contains redundant observations, such as technical replicates, Q2 can be artificially inflated if replicates are not excluded in the same iteration. In essence, the variation covered by the replicates will still be available during model building and overfitting will generate a high Q2, although the high Q2 will not reflect a real improvement in predictive ability. There is no specific (positive) cutoff point beyond which Q2 is considered good or bad. In some types of studies, a Q2 value of 0.5 is considered very good (for example metabolomics) while in others the same value would be unacceptable.

1.9.4 Root mean squared error metrics

In some cases, when predicting responses, metrics that estimate the model error in the same unit as the response are also useful. These includes the root mean square error of estimation (RMSEE), calculated using the whole fitted model, root mean square error from cross validation (RMSECV) based on the cross-validated response and root mean square error of prediction (RMSEP) [11]. RMSEP is applied in *Paper II* and uses an external set of data, a *prediction set*, to estimate the predictive ability of the model. The prediction set should contain observations

that are not included in model building. However, if the prediction set contains significantly different observations from the the set used to create the model (the *training set*) performance will suffer, as the model will be forced to extrapolate. Additionally, the number of components in the model should not be based on the model's performance with the prediction set, which could essentially be regarded as one cross-validation iteration (as described above). If these problems are avoided, the RMSEP should reflect the real-world predictive performance of the model.

In conclusion, different metrics have different uses for determining models' quality. However, some circumstances can lead to the metrics misrepresenting their quality [53], including the presence of outliers and inappropriate treatment of data (see *1.9.7*, *Handling outlying observations*, and *1.7*, *Data pre-treatment prior to modeling*).

1.9.5 Correlation coefficients

Correlation coefficients (e.g. Pearson's [54], Spearman's [55] and, Kendall's [56]) are metrics that describes the relationship (or association) between two variables, i.e. how one variable changes when the other changes. Pearson's correlation coefficient (r), which was used in the studies this thesis is based upon, has a value between -1 and 1, depending on the covariance of the variables divided by the product of their standard deviations, and the correlation between vectors x and y is given in Equation (1.8).

$$r = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{1.8}$$

where: x, y = vectors of the same length
cov = covariance
σ = standard deviation

In essence, values of r close to 0 imply that there is little or no linear relationship between the vectors x and y, while values of r close to -1 and 1 indicate a nearly perfect negative and positive relationship, respectively.

However, although a correlation coefficient is a very helpful tool, it can easily be misinterpreted for several reasons. First, correlation does not imply causation: changes in many strongly correlated variables are driven by the same underlying cause and neither directly causes the other to change. Causation cannot easily be resolved by statistics, at least without a time series of data (which enables tests of Granger causality in some circumstances [57]), and specific analysis is generally required. Additionally, if the number of observations is low, one or a few observations can have a high impact and cause high correlation, even (in extreme cases) when there are many uncorrelated observations in the dataset (see *1.9.7*, *Handling outlying observations*).

It is also possible that some clusters or trends in the variables may cause the correlation coefficient to be misrepresentative [58]. If there is any doubt, the

variables could be plotted against each other (e.g. using a scatter plot) and the results visually verified.

1.9.6 Variable selection

The variables measured in a dataset are not always relevant for the studied problem and can instead interfere during both model building and its interpretation [59]. It can therefore be useful to identify such variables and perhaps exclude them.

In some cases, the procedure is simple as the variables may be noisy because of known instrumental problems. For example, in *Paper II*, wavelengths near the detection limits of the sensors used to acquire NIR data were excluded.

In other cases, where the reason for noisy variables is less clear, statistical metrics that indicate the importance of variables for the resulting models are available, such as variable importance on projection (VIP), selectivity ratio (SR) and significance multivariate correlation (sMC) [60, 59]. Variables with less than a certain threshold significance, according to one or more of these metrics, can be discarded.

Applying dimensionality reduction techniques (e.g. PCA [61]) is also an option. These methods address the problem that if the number of observations is lower (in many cases, much lower) than the number of variables, the variables cannot be linearly independent as the rank of the matrix is limited by the number of observations (see *1.6, The concept of latent variables and their models*). The variables can therefore be replaced by a smaller set that represent the variation in the original variables [62]. In PCA, if a suitable number of components are extracted, the structured variation will remain, but the unstructured data (noise) will be removed. A downside of this approach is that the loadings of the new model now relate to these new variables (in PCA, the scores), losing the relationship with the original variables and hindering interpretation.

An alternative to simply removing the variables is to use regularized methods (e.g. regularized PLS [63]) which penalize variables deemed irrelevant without removing them from the dataset. This causes components to be sparser (not involve all variables) and thus easier to interpret.

In all cases of variable selection, it is essential to remember that removing variables because they do not agree with the hypothesis being tested (without any further justification) will result in a biased model that might not reflect reality, and such a model would require extensive validation.

1.9.7 Handling outlying observations

Outliers in data are generally described as specific observations that deviate to a significant degree from the norm. This may be due to measurement error or actual deviation in the sample properties. Dealing with outliers is nontrivial and widely discussed [64, 65, 66]. Outliers be classified as moderate or strong [11]. The

difference between the them is largely dependent on subjective interpretation, but generally they are considered moderate if they are unlikely to strongly influence the model, while strong outliers can change the direction of a component. Moderate outliers are usually removed before the modeling step, but strong outliers are often detrimental to any analysis and therefore removed. In both cases, the underlying reason for the outlier must be investigated.

If the outliers carry relevant information that represents intrinsic properties of the samples, rather than measurement errors, they must be accounted for in some way, or there are risks of misrepresenting the data.

Two metrics are often used in chemometrics to find outliers: Hotelling's t-squared distribution (or simply Hotelling's T2) [67] and observation residuals, here represented by the distance to model in X (DmodX) [11]. Hotelling's T2 is a generalization of Student's t-statistic for multivariate purposes and can be used to generate a confidence interval for observations. The confidence interval is often plotted as an ellipse in score scatter plots, and an observation lying outside this ellipse can be considered an outlier.

DmodX is a metric that measures the residuals E, i.e., the unexplained variation in each observation. A large residual implies that the corresponding observation is not well explained by the model. The standard deviation of the model residuals can be calculated and a significance threshold can be set (e.g. 95% or 99%). If the amount of residual variation in an observation is above the threshold, the observation is not well explained by the model and is therefore flagged as a possible outlier.

1.9.8 Handling missing values

Real-world datasets often contain missing values for various reasons (e.g. levels of variables may be below the detection range of an instrument, or not measured, or measurements may be lost) [68, 69, 70]. Some analytical procedures, such as PCA implementations using the NIPALS algorithm [71], can inherently handle missing data, usually at a cost of computational performance. When other algorithms, e.g., regular singular value decomposition (SVD) algorithms [72] are used, missing data must be properly handled, but the optimal approach depend on the reason that data are missing.

If the data are not missing at random, the reason for the absence (for example, values being below a detection limit or questions about sensitive issues, such as income, in a personal questionnaire not being answered at all) must be identified, and accounted for if necessary. Determining whether data are missing at random or not is challenging especially if the reason for their absence is not known in advance.

If data are found to be missing at random, entries for the missing values can be obtained by interpolation from existing data. A commonly used approach is to replace missing values with the mean or median value of the variable. Other approaches include creation of regression models to approximate the values [73],

or use of context-based rules that are application-dependent (e.g. replacing values below a biomarker detection limit [74]). Another option is to simply exclude the variable completely from the analysis. The main advantage of removal is that no invalid conclusions will be drawn based on potentially badly replaced values. If too many values of a variable are missing (a rule of thumb sometimes applied is more than 20% [71]), and the underlying cause is unknown this may be the safest approach. However, the downside is that potentially relevant information is lost.

Ultimately, whatever approach is taken, missing data will always impair the analysis as the missing values could have had great significance. Replacing the values before multivariate analysis ensures that the most statistically sound replacement method can be used, and (if properly reported) allows for greater transparency.

1.10 Ways to improve model quality

For several reasons, a model's performance may not meet expectations.

The metrics and approaches mentioned in this chapter have been mainly concerned with measuring or addressing problems in the modeling, e.g., choosing a model of appropriate complexity, transforming the data appropriately or dealing with problematic values. However, poor model performance may not be due to deficiencies in model building, other possible reasons include lack of representative data and too much noise in the modeled dataset [7]. It may become apparent that samples (observations) lack representativeness of the whole population when a model is externally validated, e.g. using a validation set. If the measured data are simply not capturing information relevant to the response, due for instance to problems with measuring instruments or items in a questionnaire, the models will always perform inadequately, possibly with high R2 but low Q2. This is also true if the data or responses contain too much noise and no clear signal can be found, both R2 and Q2 will suffer, and the residual E will be large.

If samples are non-representative, the solution is often to expand the measured space (see *1.2, Design of experiments*) and measure additional samples. On the other hand, if the samples are representative but simply too diverse, it may be better to use two or more *local* models, based on subsets of the samples. The smaller models will be more likely to accurately represent the variation within the subsets. If measured data are not representative, i.e. relevant information is simply not available, the study should be re-started using more relevant measurements.

Chapter 2

Data integration by multiblock analysis

> It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
>
> *Arthur Conan Doyle*

Integrated analysis of more than two blocks simultaneously is referred to in this thesis as *multiblock analysis*. The concept is known by other names, such as data integration [75], multiview analysis [76], data fusion [77] and various similar names [1]. In some contexts, there are differences in objectives between these procedures, e.g., in data fusion new data are added to data previously used for modeling in order to obtain a more accurate portrayal of the focal process or phenomenon, while integration may involve combination of data from different sources to gain knowledge about it that would otherwise be unavailable. The second type of procedure, data integration, is the focus of this thesis. The goal is to combine complementary information from multiple sources to improve understanding of the studied system.

Use cases for integration can be when measuring the same sample on several instruments, each able to detect different phenomena. It could also be measuring the same sample at different time points or after different treatments. By combining the information from multiple sources the studied process can be

better understood.

2.1 OnPLS

OnPLS, sometimes called Multiblock OPLS [78, 79, 80], is an extension of O2-PLS that allows analysis of two or more blocks simultaneously. The advantage of the OnPLS approach is that structures in the data (the components) of each block are split into different parts, which can be separately analyzed. The parts include the globally joint variation found in all blocks, the locally joint variation found in at least two but not all blocks, and the unique variation found in only a single block, as shown in Equation (2.1).

$$X_i = \underbrace{X_G}_{\text{Globally joint}} + \underbrace{X_L}_{\text{Locally joint}} + \underbrace{X_U}_{\text{Unique}} + \underbrace{E}_{\text{Residual}} \tag{2.1}$$

OnPLS has been successfully applied, for example, to analyze omics data in characterization of *Populus* plants stress responses [81] and lung cancer [82]. Mathematical details of the rationale behind the inner workings in OnPLS are beyond the scope of this thesis and the reader is referred to previous literature [78, 79, 80], which provides a thorough mathematical description of OnPLS.

Other multiblock algorithms

Several comparable techniques to OnPLS have been designed to integrate multiple blocks of data. Early attempts include consensus PCA [83] and hierarchical PCA and PLS [84, 35]. The two hierarchical methods operate on different *levels*, essentially creating several separate PCA or PLS models to separate the content and improve interpretation. Other methods include DISCO-GCA and PCA-GCA [85], SO-PLS and PO-PLS [86, 87], MB-OPLS [88] and JIVE [89]. Several of the methods and their working mechanisms have been previously compared and explained [85, 90]. In addition, results of three approaches (JIVE, OnPLS and hierarchical modeling) are compared in *Paper III*.

Generally, all of these multiblock methods are likely to be better than the others for handling some real or simulated datasets. However, when properly applied they are all also likely to indicate similar patterns and trends to OnPLS, at least when there are significant patterns in the analyzed data. Hence, the others are mentioned here for the sake of completeness, but otherwise are beyond the scope of this thesis.

2.1.1 New implementation of OnPLS

To evaluate different approaches and solutions to problems, it is often necessary to change parameters such as data pre-treatment and algorithm thresholds. Clearly, faster feedback (i.e. faster execution time) is always preferable when dealing with

data analysis, as obvious problems can quickly be remedied. The main concern is simply that the problem space, i.e., possible component placement combinations, grows quickly with increases in the complexity of the data (number of components and especially number of involved blocks.

To combat this problem, two variants were previously introduced for component selection, called the *Full model* and the *Partial model* [80], with the latter intended to mitigate the problem. In both cases, the components containing all included blocks, the *globally joint* components, were extracted first. Thus, the combinations considered were only those involving the *locally joint* components, joint components not including all blocks. The full model approach evaluates all possible combinations of locally joint components, a task which quickly becomes unfeasible with increasing problem size. The partial model approach instead 'prunes' the possible options by quickly discarding options that are unlikely to prove useful, and is therefore less computationally intensive. The decision whether or not a joint component should be pruned is based on several metrics, including explained variance and correlation between the joint components.

These steps enabled creation of OnPLS models without resorting to use of computer clusters or similar solutions. However, the execution was still undesirably slow when analyzing large datasets pertaining to five or more samples, even using the faster method. This led to a new implementation of OnPLS, which was used in all the studies underlying this thesis, which is designed to retain the accuracy of the previous algorithm while improving execution speed.

The new implementation only considers the next selection step when placing joint components, assuming that selections with poor performance are immediately pruned. This has given good results during tests, and was responsible for much of the reduced execution speed of the algorithm.

Another large improvement in speed provided by prohibiting missing values in the data. This simple restriction has led to the ability to nearly always use faster algorithms such as Singular Value Decomposition (SVD) [91] rather than Non-linear iterative partial least squares (NIPALS) [22], which is otherwise commonly used for calculating PCA models with missing values. For more information on missing values, see *1.9.8*, *Handling missing values*.

The end result of these changes is a significant increase in speed, at the cost of some accuracy as some options are not evaluated. To compensate for this, my colleagues and I (hereafter we) provided a set of visualizations showing the structure of the model along with the tools required to create an acceptable model, as described in the following sections.

2.2 Joint and Unique Multiblock Analysis (JUMBA)

Joint and unique multiblock analysis (JUMBA) is a structured analtical workflow based on OnPLS. It was used in all of the Papers I-IV , but only specifically described as such in *Papers II* and *IV*. The steps were initially described briefly in *Paper I*, but formalized in *Paper II*.

For each considered block, X_i, the variation decomposed by JUMBA is summarized as shown in Equation (2.2).

$$X_i = \underbrace{X_G}_{\text{Joint}} + \underbrace{X_U}_{\text{Unique}} + \underbrace{E}_{\text{Residual}} \tag{2.2}$$

This is a simplified schema of the decomposition applied by OnPLS (Equation (2.1)), as the components labelled *globally joint* and *locally joint* have been combined into a single set of *joint* components. The combination reflects the reduced emphasis in JUMBA of globally joint components, as there is no longer a constraint that forces joint components containing all blocks to be extracted first. An additional benefit is that the naming schema is simplified as all joint components are numbered in the order they appear (see *2.4*, *JUMBA naming convention* for more details).

As with the other linear models discussed in this thesis, blocks are decomposed into scores summarizing observations and loadings describing variables' influence on the observations. In JUMBA, joint components have highly correlated but separate sets of scores for each block.

2.3 The JUMBA workflow

The JUMBA workflow involves the following steps:

1. Pre-treatment of data
    - Align datasets (i.e. ensure observations match between blocks)
    - Run PCA analysis for overview and quality control of each dataset
2. Pairwise joint component determination
    - Run regression (e.g. PLS or O-PLS) between the pairs of blocks and use appropriate metrics to determine a suitable number of joint components
3. JUMBA model construction
    - Perform the OnPLS modeling steps described previously (*2.1.1*, *New implementation of OnPLS*) resulting in a JUMBA model

4. JUMBA model evaluation
    - Create and inspect the correlation matrix plot (see *3.3.2*, *Correlation Matrix Plots*) for unwanted correlations and check for outliers in the model components
    - Potentially revise or repeat the workflow problems in the created model are identified (e.g. outliers)
5. JUMBA model interpretation
    - Visualize and interpret the multiblock model using appropriate tools as described in *3.3*, *Multiblock visualizations*
    - Use standard chemometric tools such as loading plots to interpret specific contents in each block

This description differs slightly from the one presented in *Paper II* as the original steps 5 and 6 were merged. Each of the steps is described below.

2.3.1 Step I - Data pre-treatment

As described in *1.7*, *Data pre-treatment prior to modeling*, data pre-treatment is essential to ensure that the dataset to be analyzed is suitable for the planned analysis. JUMBA is a linear method and the analytical steps share many properties with regular PCA. Generally therefore, the procedures used to prepare data for the other methods mentioned in this thesis could also be applied in JUMBA. If the data are skewed, transformations such as log can be used in attempts to improve the structure. UV scaling is also often used, except in such cases when it is unnecessary (as in *Papers I* and *II*, where the spectral information is in the same ranges).

Compression of datasets, for example by replacing the blocks by PCA scores, may be a valid approach if the number of variables is much larger than the number of observations, as described in *1.9.6*, *Variable selection*. This can improve performance and given that the PCA model explains most of the source data, the model will be very similar, but interpretation will suffer as the loadings now reflect the new variables (i.e. PCA scores).

For the type of multiblock analysis presented in this thesis, observations must be aligned. This means that rows in the input matrix of one block must match the same rows in the other blocks being analyzed. The nature this match may vary enormously. For example, it may be by time (values for the same samples at different measuring times) or treatment (values for the same samples after a treatment or processing step) or any other identical feature that allows the analyst to consider them equivalent.

It should be noted that the current implementation of JUMBA does not allow any missing values (discussed in *1.9.8*, *Handling missing values*), in order to improve execution speed. Multiple ways to replace missing values have been considered, but the two most common methods are simple removal (if many

samples are missing the value of any single variable, or values of most variables of an observation are missing) or PCA replacement. PCA replacement simply means fitting a model by replacing missing values with corresponding means, then predicting the missing values using the resulting PCA model. Replacing values this way is known to cause the affected variables to have artificially lowered standard deviations, as the model will only be able to predict values that are within the range of the existing values [92]. Hence, if large amounts of data are missing, it may be prudent to employ methods with better properties of the replaced data. However, none of the datasets used in *Papers I* to *IV* had amounts of missing values that were considered significant enough to warrant the use of such methods.

2.3.2 Step II - Pairwise joint component determination

When looking where to place joint variation, it is first necessary to determine how much of the variation is joint. If we extract too little as joint variation, some joint variation will instead incorrectly be considered unique variation. If we extract too much, later joint components are unlikely to correlate well and actually contain unique variation.

As previously described [80] joint variation is determined by extracting PCA components from the covariation matrix between each possible pair of blocks. The number of *pairwise joint components* extracted is therefore critical.

To determine the number of components to extract it is useful to analyze the pair separately, both by regular PCA analysis to determine the number of usable components in each block (see *1.9*, *Model metrics and validation*) and PLS or O-PLS analysis of the pair of blocks, in both directions. The number of components possible to extract using PCA sets an upper limit to the number of possible pairwise joint components. By performing PLS/O-PLS in both directions a cross-validated limit to the number of components can be found.

This information can be enough to decide the number of pairwise joint components to use, but additional domain-specific knowledge of the relationship between the blocks can also be used. Such knowledge may indicate that components should be either added or removed, depending (for example) on noise levels of the blocks or conceptional similarity (if blocks contain data pertaining to a system in very similar, or very different, states or times).

2.3.3 Step III - JUMBA model construction

This section briefly explains how the models used in JUMBA are constructed. The procedure is based on the OnPLS algorithm [80] with differences described in *2.1.1*, *New implementation of OnPLS*.

The procedure for determining joint variation (here called the *joint spaces*) in each block is visualized for the three-block case in **Figure 2.1**. Essentially, SVD is

applied to the correlation matrix between each possible pair of block combinations. This results in the equivalent of a set of PLS weights vectors (depending on the number of components extracted, see the previous step) labelled W, one for each block included in any combination. Each block will have several sets of these *local* W, depending on the number of possible pairwise combinations.

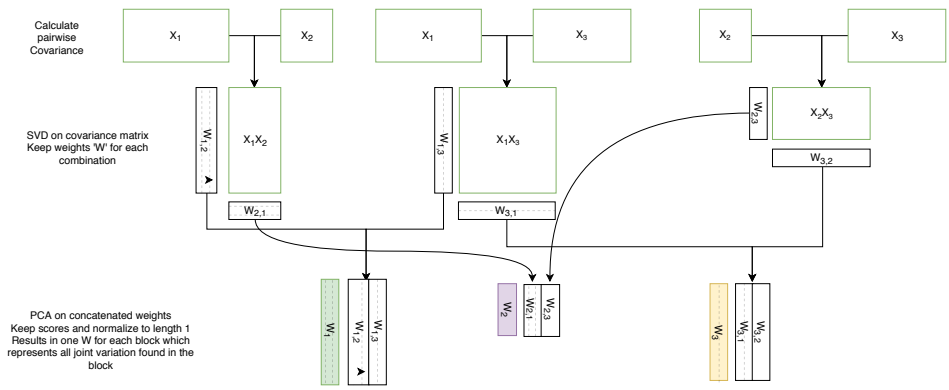


Figure 2.1: Determination of pairwise joint variation: Calculate the covariance matrix, extract the previously determined number of pairwise joint components, concatinate and compress the weights W, resulting in the *joint spaces* for each block.

A new *global* W for each block is then calculated by first concatenating all the *local* Ws belonging to the block. The concatinated local Ws are then analyzed using PCA, and a suitable number of components are extracted (see *1.9*, *Model metrics and validation*). Doing so removes redundant information from the pairwise joint components (as X_1 can share the same variation with X_2 and with X_3). The score vectors from these components are then normalized to equal length (i.e. $SS = 1$), creating the new global W for the block. This step is repeated until all blocks have a global W.

The global W for each block (from here only labelled W i.e. W_1, W_2 and W_3 in **Figure 2.1**) represents the variation found to be joint in the corresponding block. W is then used as a filter to determine and remove unique components from the blocks. The remaining variation, which is the joint (and residual) is then placed as described in *2.1.1*, *New implementation of OnPLS*.

Optionally, remaining variation after placement (residual, or possible joint variation which was not placed, see *Step IV* for further discussion on this) could be inspected for structured variation, which if found would be considered unique. The remaining variation in each block after the optional step is considered residual (E).

2.3.4 Step IV - JUMBA model evaluation

Model evaluation in JUMBA is largely based on the plots described previously. While metrics such as R2 (see *1.9*, *Model metrics and validation*) are generally included in plots (e.g. correlation matrix plots) validating multiblock models can be complicated, and the current implementation does not include a Q2 metric.

In essence, the structure of models created using parts of a dataset during cross-validation do not necessarily coincide with the structure of a model of the same type based on all observations, as the amounts and placements of joint and unique components may differ. However, this also arguably applies to procedures such as O-PLS, which partitions two types of variation, and possible approaches for calculating a metric equivalent to Q2 for JUMBA models are discussed in *2.6*, *Using JUMBA models on new data*. In the meantime, the visualizations described in *3.3*, *Multiblock visualizations* are used to visually assess their validity.

For these steps, the visualizations are generally applied to catch generic problems that could be present in, or arise when handling, any dataset. Essentially, little knowledge of the data itself (i.e., type of data) is required in this stage (further contextual analysis is applied in the next step, if the model passes visual inspection). This step involves evaluation of the correlation matrix plot, which can provide immediate feedback on some model problems, such as lack of correlation of joint components and incorrect extraction of joint variation as unique variation (see *3.3.2*, *Correlation Matrix Plots* for more details).

The model is then checked for outliers, in the same manner as when checking for outliers in PCA modeling, but using multiblock scatter plots rather than regular scatter plots (see *3.3.4*, *Multiblock score scatter plots*), which also indicate if separate observations display deviating behavior. In addition, multiblock loading scatter plots (see *3.3.6*, *Multiblock loading scatter plot*) or separate line plots (e.g., of spectroscopic data) can be inspected to ensure that no deviating variables are responsible for the direction of each component, however such variables should generally be handled before multiblock analysis begins.

The residuals can also be further inspected, per observation, to determine outliers using conventional tools, for example using DmodX (see *1.9*, *Model metrics and validation*).

As a final note, it is often not possible to neatly split all variation into either joint or unique. When nearing the noise level of the dataset, it becomes increasingly likely that noise is included when extracting the pairwise joint components. This problem can be minimized using the procedures described above, but often a choice must be made whether to keep the variation as joint or leave it as unique.

Depending on the choice of number of pairwise joint components in *Step II* or joint component placement in *Step III*, correlation in the last components will suffer if all variation is extracted, or unique components will correlate to some degree with other joint or unique components if the variation is left as unique. The correct choice will depend entirely on the circumstances, e.g. data quality (noise level) or goal of the study and whether it is possible to determine the cause

of the variation or not. One example where a component was kept is discussed in *3.3.7*, *Score scatter plot matrices* (**Figure 3.6**b).

2.3.5 Step V - JUMBA model interpretation

This step is a combination of the plots generated specifically for multiblock models (covered in *3.3*, *Multiblock visualizations*) and conventional multivariate tools or visualizations (e.g. separate loading line plots or S-plots, further investigation of residuals and so on).

The exact nature of interpretation will largely be context specific, i.e. depend on the goal of the current study. As long as some type of metadata is available a good starting point for interpretation is the metadata correlation plot (described in *3.3.3*, *Metadata correlation plots*). *Metadata* in this context literally means data about the data, and is information related to the samples that is not explicitly included in any block. Strong correlation between score(s) and metadata indicate some relationship(s). JUMBA models can be considered unsupervised with regards to metadata as, by the definition above, metadata should not be included in any analyzed block. Also discussed previously (see *1.6*, *The concept of latent variables and their models*) there is no guarantee that a single component explain a single underlying latent variable. Instead, a latent variable can be explained by a combination of several components and thus correlate to varying degrees to all of them.

If the number of components correlating meaningfully to a metadata variable is at most two, a single multiblock score plots can be used (i.e. multiblock score scatter plot or multiblock bar score plot). If more components correlate to the metadata there are some options, which include simply looking at pairs at a time (e.g. several multiblock score plots) or perhaps try to extract only the variation related to the metadata.

The latter option could for example accomplished by extracting, from one block only, the joint scores relating to the metadata and use the scores X and the metadata as y in a single-response O-PLS model. The resulting O-PLS model will produce a single predictive component which in turn contains the variation relevant for the specified response (i.e. the metadata variable). The single predictive component can then be separately analyzed. A downside of this approach is that, instead of one single model to validate and interpret, there are now two or more, which complicates the overall analysis.

In addition to the described visualizations, the JUMBA model contains scores and loadings, which can be separately interpreted using conventional methods and plots. Multiblock variants of loading scatter plots is also possible (see *3.3.6*, *Multiblock loading scatter plot*), and line plots showing loadings from the same joint components (used in *Paper II*).

2.4 JUMBA naming convention

In order to precisely specify features of a model (component, loading, score and so on) a naming convention is required. As any number of blocks can potentially be involved in JUMBA, each block is named X_i, where i is an integer from 1 to the number of included blocks. The overall structure is illustrated in **Figure 2.2**.

In JUMBA, labels of components (which have both a loading and a score) start with a capital X followed by the type, e.g. X_J for joint components and X_U for unique components. To refer to a specific component, it is specified after the type (joint or unique), e.g. X_{J2} for the second joint component.

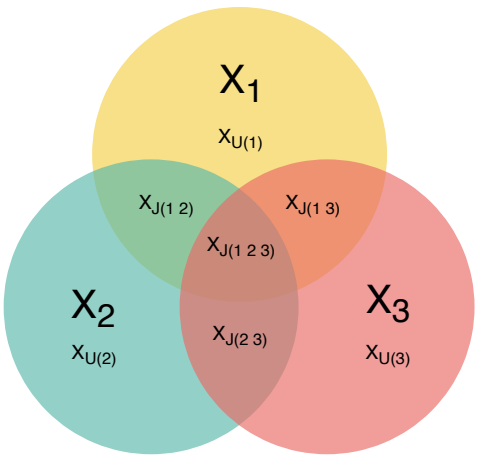


Figure 2.2: Venn diagram of parts of the joint variation explained by a three-block JUMBA model. The center part containing all blocks is sometimes called the *globally joint* variation, and joint parts covering fewer than all blocks are called the *locally joint* variation. Note that residual variation (if present) is not shown as it is not modeled.

It is also often necessary to specify the block(s) a component belongs to (i.e., has been extracted from). This is done by adding a parenthesis after the component number e.g. $X_{J2\,(1\,2\,3)}$ for the second joint component extracted from blocks 1, 2 and 3. The numbers are always in ascending order. This notation is intended to describe a whole component, i.e., all scores and loadings for all the included blocks as a whole unit, *a model component*.

To refer to a specific block in a specific joint component, the notation within the parenthesis changes. Instead of ascending order, focal block starts the numbering, followed by a comma and then the remaining blocks, e.g. $X_{J2\,(2,\,1\,3)}$ for the loading and score vectors for the second block and second joint component.

For a unique component it is always necessary to specify which block it belongs to, and this is done in parenthesis at the end, e.g. $X_{U2(1)}$ for the second unique component of block 1.

Additionally, to refer solely to the score vectors or loading vectors rather than whole components (for example, in a multiblock scatter plot), X can be replaced by T (for example: $T_{J2\,(1\,2\,3)}$) or P, respectively. Note that upper case letters are used when referring to a matrix or several vectors, and lower case letters when referring to a single vector, such as when specifying a single score vector (e.g. the score for the second block of the second joint component $t_{J2\,(2,\,1\,3)}$).

2.5 Block scores vs superscores

JUMBA uses highly correlated, but not identical, scores for each joint component, which we call block scores [75]. Other multiblock methods, such as JIVE [89] and DISCO [85] instead use the same scores for each block, *superscores*. When using superscores, the correlation between the scores in the same joint component is 1. Hence, the explained variance is placed in the loadings for each component as the scores are always normalized to equal length. The variation is placed in the scores when using block scores, and when comparing scores from different blocks this variation is usually normalized to ensure that the scales are comparable.

While the plots specific to multiblock methods presented within this thesis can be used for both types of scores, additional information can often be found in just how block scores relate to each other. For example, as the scores are not identical, the nature in which they do not correlate to (e.g. a specific observation not following the trend) can provide additional information which can then be interpreted.

2.6 Using JUMBA models on new data

Multivariate modeling techniques, for example PLS, are often used to predict responses. In multiblock analysis, none of the analyzed blocks are specifically designated as responses. However, a multiblock model can also be used in combination with new measurements that can be added to one of the constituent blocks, in order to predict what the contents of the other (not measured) blocks would have been if they had been measured simultaneously.

JUMBA prediction can use a previously created JUMBA model on data measured on one of the blocks to estimate the contents of the unavailable blocks. The solution, used in *Paper II* for calibration transfer, is presented in Equation (2.3).

$$X_M = (Tn_E \times Mf_{M,E}) \times P_M \tag{2.3}$$

where: X_M = Predicted missing block
Tn_E = Predicted existing block scores normalized to length 1
$Mf_{M,E}$ = Multiplication factor
P_M = Missing loadings (only joint with existing block)

We argue that this approach, which is based on integration of data from multiple sources, is inherently advantageous for prediction, and the results of *Paper II* support this hypothesis. The rationale for this is the expectation that a signal or trend detected in multiple sources is more likely to reflect an actual underlying structure in the data than a signal detected in a single source, as manifested in the nearly identical loading spectra seen in *Paper II*.

In essence, the *new data* is pre-treated identically to the original data (e.g. same means are used when column mean-centering and so on). The joint scores are predicted and unique variation stripped from the new data, as explained in *2.3.3*, *Step III - JUMBA model construction*, using the corresponding set of joint weights W from the model. By exploiting the assumption that the joint scores correlate strongly, the matching joint scores can be predicted. The already predicted scores are normalized to length 1 (i.e. $SS = 1$) and then multiplied so that a length equivalent of the original score (i.e. the variation is reintroduced). These shared joint scores can be used with the corresponding loadings from the model to estimate the other blocks, as seen in Equation (2.3).

By definition, unique variation cannot be used to predict another block, nor can unique variation in another block be predicted. Joint components where the new data is not included cannot be estimated, similar to how unique variation cannot be estimated. For example, when trying to predict X_1 using X_3 only the parts overlapping between X_1 and X_3 can be predicted (see in **Figure 2.2**), i.e. $X_{J(1\,3)}$ and $X_{J(1\,2\,3)}$, but not $X_{J(1\,2)}$ or any unique component.

Chapter 3

Multiblock model visualization and interpretation

> Important stories live in our data and data visualization is a powerful means to discover and understand these stories, and then to present them to others.
>
> *Stephen Few*

If findings, including important findings that could change lives, cannot be understood by anyone other than the author they will have no impact, so clear presentation of results is crucial, especially results that are inherently complex, such as outputs of multiblock multivariate data analysis. Thus, visualization of multiblock models was the focus of *Paper I*, with the intention to modify previously used plots and visualizations to enable clear portrayal of multiblock data.

3.1 Color selection

The colors used in visualizations are often important, and not only to enable colorblind people to see highlighted features. It is important to select a valid color scale for a given application, such as sequential (from one value to another, e.g. 0-100), divergent (from a mid-point to two boundaries, such as positive or negative correlations) or qualitative, where the values are distinct (classification). A plot

may be unnecessarily difficult to interpret if an inappropriate color scale is used. A very useful resource for selecting good colors, based on previous research [93], can be accessed at `http://colorbrewer2.org`.

3.2 Previous attempts of multiblock analysis visualizations

'A straightforward approach for visualizing outputs of linear multiblock models involving extraction of components (e.g., OnPLS models) is to extract the joint scores in all blocks, concatenate them, then analyze the resulting block using PCA [94] and present the results in score and loading scatter plots.

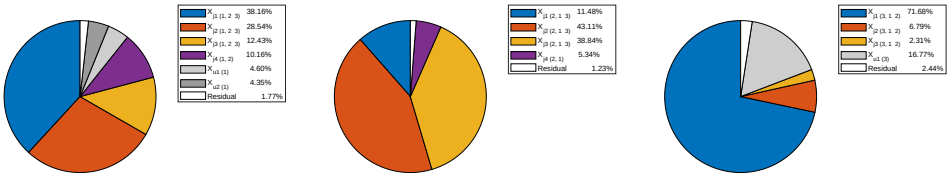
This has the advantage of providing an overview of the relationship between the observations (scores) and multiblock components (loadings). The downside is that the output is predetermined, depending on how the components are scaled, as the first multiblock component is expected to be the first PCA component, and so on. Basically, if the scores are not scaled before PCA analysis, the largest joint components (in sum of squares sense) will be extracted first. Otherwise, if each variable (i.e. joint score) is UV-scaled, joint components with the most included blocks will be extracted first. In both cases, the outcome can be difficult to interpret as the relation to the original variables has been lost. This visualization approach has therefore been largely superseded by a combination of the approaches presented below.

3.3 Multiblock visualizations

The following kinds of visualizations are plots that have been customized for presenting multiblock models, while remaining similar to conventional multivariate plots. In this section, the visualization approaches are illustrated with multiblock models generated using block scores, but many of them have also been shown to work with models generated using superscores (see *Paper I* for more on this).

3.3.1 Pie chart

The pie chart (**Figure 3.1**) is a useful tool for quickly visualizing the overall structure of the model, specifically how much variation is shared between blocks and how much is unique, and has been used in previous multiblock studies [80]. Pie charts provide the same information as a table summarizing the explained variation per component, but are visually more appealing. In general, each block should be first considered separately to get an understanding of how the model has divided the variation within the block.



(a) IR (b) NIR (c) Raman

Figure 3.1: Three pie charts (a-c) from the same multiblock model, the carageenan model from *Paper I*. Slices of the pie charts sharing the same color belong to the same joint component. Unique components are presented in grayscale and should not be compared between blocks. Unmodelled (residual) variation is presented as white.

Comparing the relative amount of explained variation is often not relevant when the blocks originate from completely different sources. If the amount of total variation in different blocks is not similar, comparing the size of each slice of the pie chart is not meaningful.

If a visualization is not deemed necessary, an alternative is to present the same information in table format, as seen in **Table 3.1**.

Table 3.1: Summation of the variation explained by the carrageenan model in *Paper I*, as a part of the whole. Unique components are summed and not presented separately. A dash (-) indicate that the block is not included in the component.

	IR	NIR	Raman
$T_{J1\,(1\,2\,3)}$	0.38	0.11	0.72
$T_{J2\,(1\,2\,3)}$	0.29	0.43	0.07
$T_{J3\,(1\,2\,3)}$	0.12	0.39	0.02
$T_{J4\,(1\,2)}$	0.10	0.05	-
Unique	0.09	0	0.17

3.3.2 Correlation Matrix Plots

The correlation matrix plot (**Figure 3.2**) provides an overview of a linear multiblock model that enables inspection of the model's overall structure and detection of several common errors (described below). As the name suggests, such plots show coefficients of correlation (often, but not necessarily Pearson's) between all the scores (joint and unique). The depth of color and size of the circles indicate the strength of the correlations.

We expect the *model scores* (i.e., scores from each of the blocks in the same joint component) to be highly correlated with each other, but not to any other component, either joint or unique.

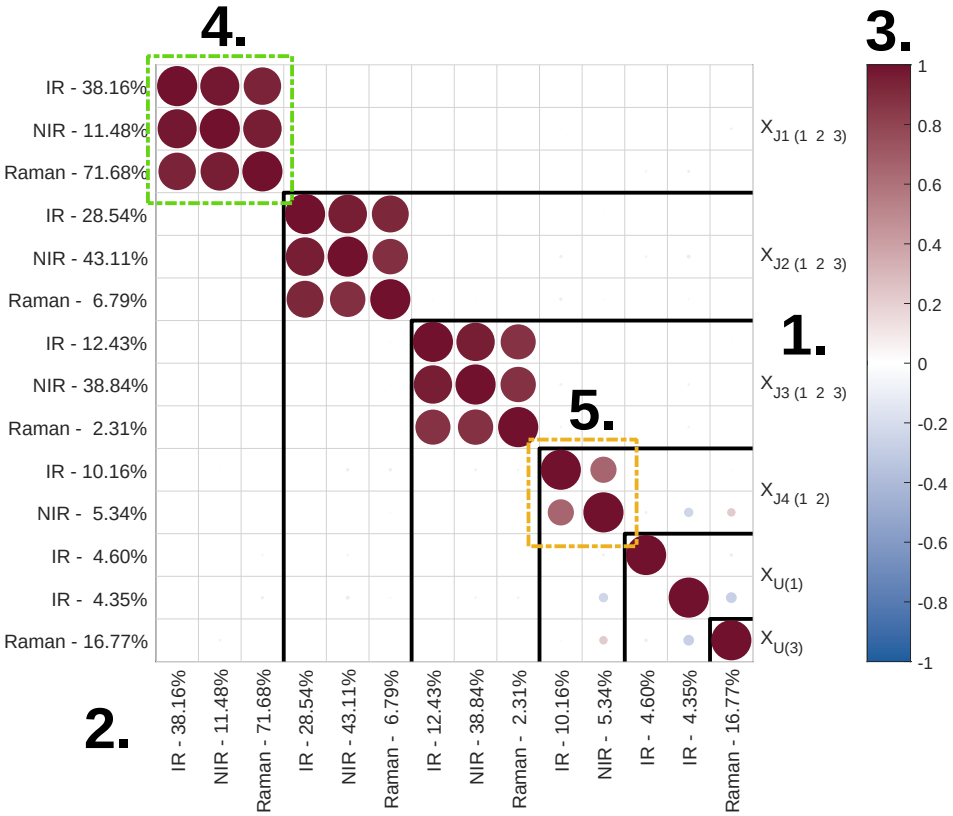


Figure 3.2: Correlation matrix plot of the carrageenan dataset described and modeled in *Paper I*. Each component is labeled on the right hand side of the plot (1), starting with the first extracted joint component at the top. A thick black line separates first the scores in the same joint components from other joint components, and then the unique components belonging to each block. The block names and explained variance are shown on the X and Y axes (2), with the same order top to bottom and left to right. Hence, the diagonal shows correlations of components with themselves, and the lower left part is a mirror of the upper right part. The Pearson correlation scale is also shown (3). The first joint component (4) is clearly strong, with strongly correlated scores, while the fourth and last (5) is substantially weaker but still considered acceptable.

The unique components are expected to be uncorrelated with any other component. Any deviation from this pattern is a cause for concern, but provides feedback regarding the underlying cause. Model scores that are not highly correlated indicate that at least one included block should be excluded from the joint component. If scores in a joint component correlate to the scores in another joint component, the two joint components should likely be merged, forming a new joint component. Similarly, correlation to a unique component suggests that the block containing the unique information should be included in the joint

component. Finally, strong correlation of unique components implies that the number of pairwise joint components between the corresponding blocks is too low and some joint variation was incorrectly classified as unique.

However, there is the additional possible risk that the number of pairwise joint components is too high, i.e. the model is overfitted. Overfitting the number of pairwise joint components and extracting all possible joint variation will lead to non-joint signals and noise incorrectly classified as joint variation. This can be observed in the correlation matrix plot as perfect correlation within (but not between) all joint components, and is essentially a result too good to be true. Regardless, this is unlikely to happen by mistake except in the cases where the number of observations or variables (essentially the rank) in each block is very low, and therefore extra care must be taken when extracting pairwise joint variation in these cases (*Paper III*).

Due to the combination of these attributes, a correlation matrix plot is a powerful tool for inspecting a new multiblock model, and obtaining information regarding both model errors and possible ways to solve them. To ensure that none of the correlations are due to errors (see *1.9.5*, *Correlation coefficients*) the components can be validated using, for example, multiblock score scatter plots or score scatter plot matrices (see *3.3.4*, *Multiblock score scatter plots* and *3.3.7*, *Score scatter plot matrices*, respectively)

3.3.3 Metadata correlation plots

A useful tool for identifying components that warrant further analysis is a metadata correlation plot (**Figure 3.3**). In this context, *metadata* literally means data about the data, particularly information related to the samples that is not explicitly included in any block. Examples of such information may include the run order of experiments, age or gender of subjects, who took measurements and so on. As per the definition above, these variables do not directly influence the direction of the model, and the multiblock methods used here can be considered to be unsupervised with regards to the metadata.

The intent of the metadata correlate plot is to show whether these values correlate with any found joint or unique score in the model, or any combination of such scores. Strong correlation indicate a relationship between a score and a metadata variable, and such relationships should be further investigated using for example a multiblock score scatter plot (see *3.3.4*, *Multiblock score scatter plots*) and coloring observations according to the relevant metadata variable. It needs to be emphasized that there is no guarantee that any single component is related to only a single metadata variable, or vice versa.

A metadata correlation plot of a model has the same number of rows as the corresponding correlation matrix plot, described above, and can be put next to it. This provides an even more complete overview of the data structure, showing the model's construction as well as its relationship to the metadata. However, even by itself, it is possible to determine the correlations of the model scores using a metadata correlation plot. For example, if all scores in a joint component correlate

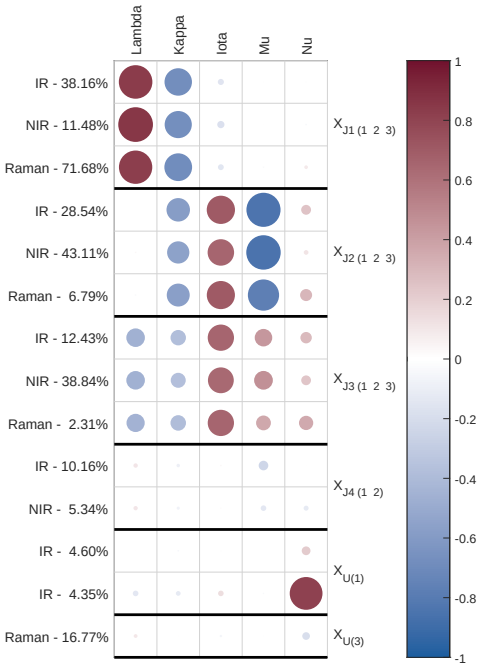


Figure 3.3: A metadata correlation plot (of the carrageenan model described in *Paper I*), showing correlations between model components and external metadata. Large circles with strong colors indicate strong correlation and therefore a strong relationship. However, care must be taken to ensure that the correlation is not caused by outliers.

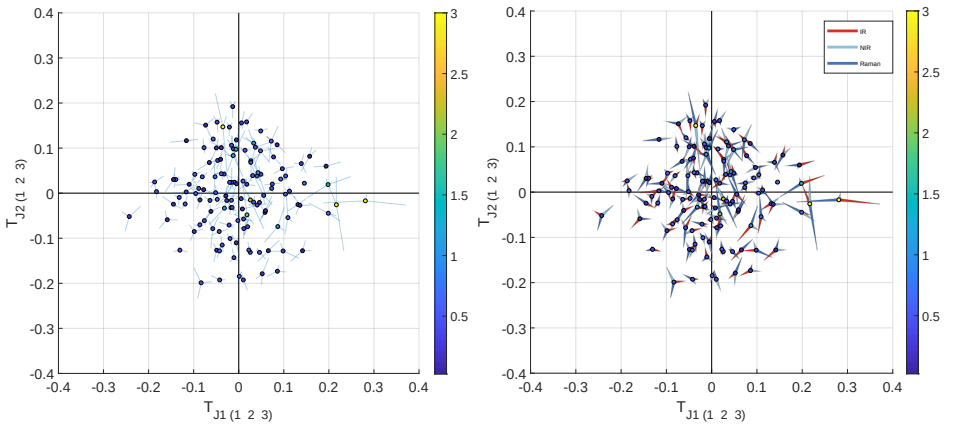
with a given metadata variable (i.e. the respective circles have very similar size and coloration in the plot) they are likely to correlate internally as well. However, if a strong correlation only appears for one of the joint scores, it may mean that the internal correlation is weak for that joint component.

As previously mentioned, each joint component is also likely to relate to more than one metadata variable (as mentioned in *1.6, The concept of latent variables and their models*, variables often correlate with each other, because they are influenced by the same factors), and vice versa. The same is true for metadata, and in some cases interpretation can be made easier by removing redundant metadata variables, for example by using PCA to select representative metadata variables which was done in *Paper II*.

3.3.4 Multiblock score scatter plots

A multiblock score scatter plot (illustrated in **Figure 3.4**) is used in the same way as a traditional score scatter plot, for example to look for trends, outliers and clusters in the data. It can also visualize scores from all blocks included in the joint components at the same time, thereby also providing insight into the inner

relations within each joint component, i.e. how well the scores correlate [11, 95]. If scores in the joint component correlate perfectly, it will be identical to a regular score scatter plot. It is necessary to normalize scores to the same length, otherwise scores explaining the most variation will dominate the plot.



(a) Whisker style (b) Arrow style

Figure 3.4: Two variants of multiblock scatter plots, one using whiskers (a) and one using arrows (b). Both variants show the same data: the first two joint components extracted from the carrageenan dataset described in *Paper I*.

Two variants of this kind of plot are used in *Paper I*: one where separate scores are represented using whiskers and one using arrows. In both cases, a center point is drawn to represent each mean score value. The first version (**Figure 3.4a**), using whiskers, bears more resemblance to traditional score scatter plots. In the other version (**Figure 3.4b**) colored arrows are used instead of whiskers, with the color matching the block responsible for the score. Using different colors enables easy visualization of whether one included block differs systematically from the others, e.g. is always located in a certain direction away from the other blocks. The underlying reason for such a difference should be determined.

Regardless of the variant used, the scatter points can be colored according to the similarity of the constituent scores, as shown in **Figure 3.4**. Scores that deviate significantly from the norm (and thus have longer whiskers or arrows than the others) may need further investigation. Reasons for such deviation may include (for instance) incorrect measurements of one or more samples in one of the blocks.

By default, the plot requires blocks included in the in the two visualized components to be the same. It is possible, however, to override this behavior and only plot joint scores which are included in both components. Doing so comes with the understanding that not all joint scores are actually represented, which should be taken into consideration during evaluation. Instead of going for this option, which risks misrepresenting the data, it may be preferable to instead plot the joint components separately, for example using the multiblock bar plot described below.

3.3.5 Multiblock score bar plot

The multiblock score bar plot (**Figure 3.5**) is used when inspecting the scores of a single joint component. It operates largely as a regular score bar plot would, but similar to the multiblock score scatter plot described above the scores from each block has been normalized to the same length, and each bar represents the normalized mean of each observation.

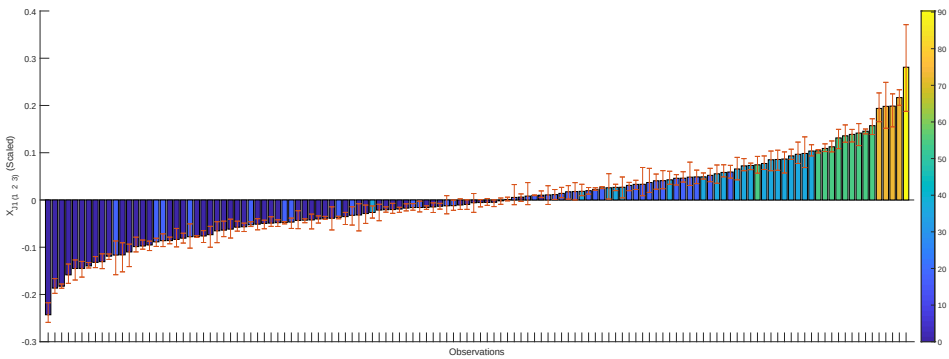


Figure 3.5: A multiblock score bar plot of the first joint component of the carageenan dataset (used in *Paper I*), colored by the Lambda type carageenan and sorted by mean score value. Each value represents the normalized mean score of the constituent joint components, with error bars showing the highest and lowest value for each observation.

The bar plot can be colored according to some metadata, and sorted (as seen in the example), and error bars show the minimum and maximum score values for each observation.

3.3.6 Multiblock loading scatter plot

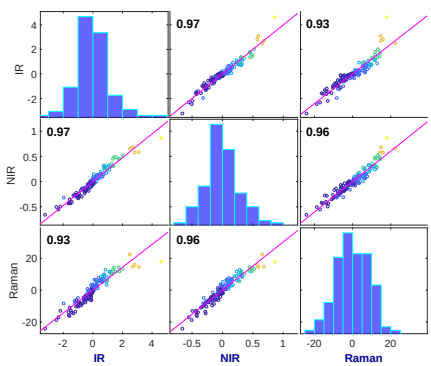
Loadings of a multiblock model can be inspected to improve understanding of the structures in the original data that are largely responsible for each component. A multiblock loading scatter plot is a regular loading plot, but shows the loadings from multiple blocks simultaneously. However, as the loadings from each block have equal variation (i.e., $SS = 1$), blocks with more numerous variables will have smaller values. To ensure that they are comparable, it can therefore be prudent to normalize their scale, or use the original variable / score correlation [46] (correlation loadings, or *p(corr)*), which was done in *Paper IV*.

The scores are often colored according to the block they belong to. This enables easy verification of expected relationships between variables, both within and between blocks.

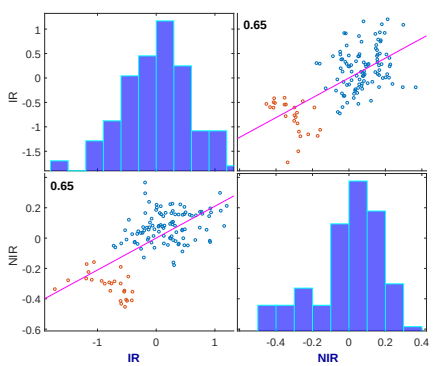
3.3.7 Score scatter plot matrices

Plotting the scores in a joint component in a scatter plot matrix allows more detailed inspection of the scores' relations to each other within a joint component than, for example, a multiblock score scatter plot. It is essentially a set of scatter plots showing each possible pairwise combination of the joint component scores. It is especially useful for model validation, to ensure that a joint component does not solely explain a single outlying observation. The scatter plots can additionally be colored according to some known metadata variable, as exemplified in **Figure 3.6**.

In the example (**Figure 3.6**a), colored by Lambda, a clear trend from low (blue) to high (orange-yellow) can be seen in the scores, with no easily distinguishable outliers. Most values are close to the line showing that the overall correlation in the component is high. In the fourth joint component (**Figure 3.6**b) the correlation is weaker, also indicated by the correlation matrix plot (**Figure 3.2**) for the same model. There is a clear separation between data obtained on sampling day (marked in red) and the other days (marked in blue). Although the correlation is not perfect the component contains relevant structure and is therefore kept in the model (see *2.3.4, Step IV - JUMBA model evaluation* for more discussion on when to keep joint components).



(a) X_{J1}



(b) X_{J4}

Figure 3.6: Score scatter plot matrix plot of the carrageenan model described in *Paper I*. The diagonal shows the histograms for each block in order. The scatter plots show the actual score values plotted against each other, and if the correlation was perfect all the scores would be on a straight diagonal line. The correlation coefficient is shown in the upper left corner of each plot. The first joint component of the aarrageenan model (a) is colored according to the lambda fraction, selected by inspecting the metadata correlation plot (**Figure 3.3**) for the strongest correlation for that joint component.

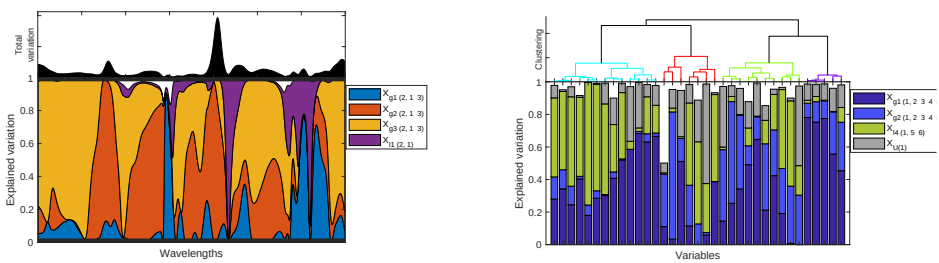
3.3.8 Modelled variation plot

A modelled variation plot provides an overview of the variables that were most influential, i.e., which variables contributed most strongly to the components.

Two variants of this kind of plot, with spectral (a) and separate variable (b) styles, are shown in **Figure 3.7**. The spectral style is most useful in situations where variables are inherently spatially related, wavelengths for example, as each wavelength is closely related to the ones before and after it. The variation of each variable is shown in the upper half of the plot. In the lower part, the explained variation is shown, with different colors corresponding to different variables.

The other variant, shown in (b), is used when the variables are independent and can be reordered. In the example shown, they have been reordered according to the result of hierarchical cluster analysis [13], but this is optional. The dendrogram corresponding to detected clusters is shown in the upper part, while the lower part shows the explained variation per component.

Knowing where a model focuses, in combination with previous knowledge about the components' relations with metadata variables (obtained, for example, using a metadata correlation plot) is highly valuable for determining where relevant information is located, and this kind of plot provides an overview of this information. Additionally, the distribution in multiple blocks can be simultaneously shown side by side, with matching colors per component. This shows the relationships between variables in different blocks.



(a) Spectral style

(b) Independent variable style

Figure 3.7: Two variants of a modelled variation plot: in spectral style (a, from *Paper I*) and independent variable style (b). In (a), the upper part visualizes the magnitude of the variation per variable. In (b), the upper part presents a dendrogram showing how the variables cluster together.

Chapter 4

Results

> What makes a scientist great is the care that he takes in telling you what is wrong with his results, so that you will not misuse them.
>
> *W. Edwards Deming*

In this chapter we provide a brief description of the results from each paper, as well as comments on specific choices we made during the creation of them.

4.1 Paper I: Multiblock visualizations and their applications

In the first paper, we presented several novel visualizations, which are covered in *3.3*, *Multiblock visualizations* and will not be repeated here. OnPLS was selected to represent multiblock models using block scores, and JIVE was selected to represent superscores (see *4.1.1*, *Choice of alternative multiblock method* below). We showed how the visualizations can be applied on both types of multiblock model, including how to interpret them, using two separate datasets.

The first dataset was a vibrational spectroscopy dataset (comprising measurements of carrageenan samples with NIR, IR and Raman instruments) measured on carageenan samples, intended to determine the composition of each sample [96]. In this set, we were able to show the how the model could be interpreted to find relationships between extracted components and known study

metadata variables, in this case day of measurement, which was deviating. As the alternative method, JIVE, is not designed to extract joint components not including all blocks (sometimes called *locally joint* components), it was not able to extract structure describing this relationship. It needs to be emphasized that this is not an inherent limitation in superscore methods as a whole, and is specific to the current implementation of JIVE. The deviating day of measurement found by us confirmed conclusions drawn by earlier analysis of the same dataset [97]. Generally, when disregarding the inability to separately extract joint components, we found that the JIVE extracted similar (but not identical) structures as those extracted by OnPLS.

The second dataset is a multiomics set (comprising transcript, metabolite and protein measurements of hybrid aspen plants) designed to identify factors the plants' growth rate [98]. Using OnPLS and relevant visualizations, we were able to separate and visualize the different genotypes and internodes in the samples using a single model.

4.1.1 Choice of alternative multiblock method

To prove that it was possible to use the same visualizations on models using superscores, we needed an example model. Several options were possible, but there were two main reasons for selecting JIVE. These were accessibility, as it had publicly available source code, and the fact that JIVE had been used in a number of previous studies. We also believed that the contrast between OnPLS and JIVE (i.e. that JIVE only separates globally joint variation) would be of interest.

4.2 Paper II: Introduction of JUMBA and use of JUMBA for calibration transfer

Measuring the same sample on for example multiple similar spectroscopic instruments may result in spectra that are slightly different. Small instrumental differences are often enough to cause significant errors when one instrument attempts to predict a response using a model created on another instrument [99, 100, 101]. Applying multiblock models to solve this problem is the focus of *Paper II*, using a methodology called *calibration transfer* [102].

Calibration transfer refers to a set of approaches applied in attempts to allow a model created on a specific instrument to be used on another instrument while retaining similar accuracy (e.g. in the form of RMSEP) [99]. This is highly desirable as recreating the calibration model can be difficult for example due to costs or limited supply of the source material. The original instrument used for creating the calibration model is often referred to as the *master* instrument, and the new instrument is the *slave* instrument. To solve the calibration problem, several approaches have been taken, including direct standardization and piecewise direct standardization [103, 104, 105], spectral space transformation [106] and transfer using orthogonal projections (TOP) [107] used in *Paper II*. They all involve creation of a model linking outputs of the slave and master instruments. The exact workings of each method are beyond the scope of this thesis, but it is worth mentioning that only TOP exploits measurements of the same samples by multiple instruments. The other methods instead use only two instruments, the master and the slave.

In *Paper II*, we introduced JUMBA calibration transfer, which uses JUMBA prediction (explained in *2.6*, *Using JUMBA models on new data*) and proved that it can be used to transfer calibrations between a set of instruments, using two different datasets.

We conclude that the performance of JUMBA calibration transfer matches the best performance of the other tested methods, and does so while being more stable than the others (measured as described in *4.2.2*, *Calibration transfer results*). We believe this stability is thanks to a stabilizing effect granted by using joint structures (which are always represented in several different instruments) as a basis for the transfer.

4.2.1 Selecting representative sets

The selection of the different sets, i.e. calibration, transfer and validation sets, were crucial for a representative study in *Paper II*.

A calibration set consists of data pertaining to the samples used to create the original model for the master system, and the performance of this model is generally considered a benchmark for future measurements. The transfer set consists of observations available for model building by the transfer methods.

Normally, this consists of measurements of the same samples taken using

both master and slave instruments, and the validation set is used for testing the resulting transfer performance. In the spent mushroom dataset (SMD) we found obvious clusters in the data (see *Paper II*). The calibration set is the most important as without a good representation of samples the resulting master calibration model would never be accurate. Therefore, we did not randomize selection for the calibration set and instead used the Kennard-Stone selection algorithm (KS) [108] to iteratively select observations with the greatest distance to previously selected observations. This ensured that the calibration samples were representative of the overall variation in the whole dataset.

For the SMD, an additional constraint was that only samples with corresponding responses could be considered for the calibration set. This was not applicable for the corn dataset, as all measured samples had corresponding measured responses.

The transfer set was then selected at random, and remaining samples were used for the validation set. There are other selection algorithms which could have been used, including modifications of the Kennard-Stone algorithm [109], which would likely have proven perfectly viable as well. However, in the end all transfer methods used the same sets, thus there was no inherent bias in that regard.

4.2.2 Calibration transfer results

As previously mentioned (e.g. in *1.9.4*, *Root mean squared error metrics*), optimizing with regards to RMSEP essentially causes the validation set to no longer be useful for validation, but will instead act as one iteration of cross-validation, and will therefore not be representative of real-world future use. However, in this case we deemed optimizing this way acceptable, as the aim of the study was to compare transfer methods and not to actually produce the best models for predicting future samples, and all methods had the same advantage.

We elected to use two different measurements, both the upper confidence interval and the mean of the RMSEP. The upper confidence interval was used to measure the stability of each method when different combinations of master and slave is used for the same response. We believe that in many real-world cases, stability can be more important even than mean value, and in this aspect JUMBA performed very well, which was reflected in low values for the upper confidence interval.

If we consider both different instruments (i.e. master/slave combinations) and different responses, only TOP was able to match JUMBA transfer in terms of stability, and even then only on the corn dataset. In the other dataset (i.e. SMD), TOP was the least stable and JUMBA transfer the most stable. This might be due to the clusters found in the SMD, which we believe TOP was unable to accurately model.

4.2.3 Multi-instrument calibration

Multi-instrument calibration (MIC) is one possible future application for JUMBA transfer. MIC describes a procedure where samples characterized by different instruments can be used to create a new calibration model, as long as each sample has corresponding measured responses.

As long as a representative set (i.e. the transfer set) of samples are measured on all involved instruments and a JUMBA transfer model has been fitted, measurements from one instrument can always be transferred to all other instrument. Thus, it is possible to combine samples measured on any combination of instruments to fit a calibration model on any involved instrument, as long as the samples had corresponding measured responses. This would allow models to be more easily updated with new samples, which can be especially useful in cases where the same sample is difficult to measure on more than one instrument, perhaps due to the instruments' physical location being far apart. Evaluating the performance of this approach should be the target of future research.

Paper III: Multi-tissue metabolomic integration

In *Paper III* we use three different approaches to integrate multi-tissue metabolomic data from mice. The three approaches were hierarchical modeling, JIVE and OnPLS. However, the OnPLS model was constructed according to the JUMBA workflow approach described in *2.3*, *The JUMBA workflow*, and will therefore be considered a JUMBA model.

In the study, we used six different blocks, each block representing the metabolic profile extracted from one tissue type (gut, kidney, liver, muscle, pancreas and plasma). Different methods were used to integrate and interpret the resulting model or models to determine the overall metabolic patterns. The different methods were selected to represent established methodology (in this case hierarchical modeling) and two more recent approaches to multiblock analysis in the form of JIVE and OnPLS. JIVE and OnPLS were used for the same reasons they were used in *Paper I* (see *4.1.1*, *Choice of alternative multiblock method*).

We need to emphasize that the intent with the study is not a comprehensive comparison of all multiblock methods for the general case, but rather a comparison of the information gained by integrating metabolomic data with limited samples using a few representative methods.

4.3 Handling of analysis involving few samples

The main concern from an analytical standpoint was that the number of samples was limited to eight.

While we believe that the data is of sufficient quality due to how the study was performed (and has been the basis of previous research [110, 111]), it would be very easy to by accident overfit the models, and results from cross-validation may not be representative. As a result, carefully selecting a suitable number of components in all models was critical.

For JUMBA, we had to be especially conservative with the pairwise joint components. Allowing each pair of blocks more than one pairwise joint component would, when taken in aggregate, mean that nearly all variation would be considered joint. As a result, even though no pair was allowed more than one component, we were still able to extract several globally joint components (after the steps described in *2.3.3*, *Step III - JUMBA model construction*). Even after this limitation, the final model contained 7 joint components, although the maximum any single block was involved in was 5. As a consideration to the noise level in the data and the expected number of components (as determined by analysis using PCA), we omitted the last 3 joint component from further analysis, resulting in a total of 4 joint components.

A modified approach to outliers is also required when dealing with so few samples. The first joint component of the JUMBA model was to a large degree dedicated to explaining the deviating behavior of one of the mice. We were, however, able to determine the underlying cause of this behavior, and it was therefore not considered an outlier.

Ultimately, even though data was limited, we were still able to interpret the multiblock dataset, with all methods able to find structure relating to mouse which expressed a deviating metabolic profile as well as structures relating to mouse size. Only the JUMBA model was able to extract additional locally joint variation relating to the mouse genotype. We believe this provides evidence that, as long as special care is taken, the multiblock modelling approach we describe in this thesis is applicable even in cases where the number of observations is limited.

Paper IV: Multiblock analysis of malaria samples

In this paper, JUMBA was used for multiblock analysis of lipidomic, metabolomic and oxylipin datasets obtained from profiling of plasma samples from children infected with P. falciparum malaria.

For this paper, we created the model following the JUMBA procedure described in *2.3*, *The JUMBA workflow*. The metabolomics data was pre-treated using CCMN [32] and the oxylipin data was log-transformed. Three samples in total was removed after initial analysis as one was completely missing from one block and two others displayed deviating behavior we deemed to be due to invalid measurement. However, this still left 57 profiled samples in the dataset and model fitting was otherwise straightforward.

Interpretation of the model allowed us to confirm previously known trends. Furthermore, previously unknown trends were also found, most probably related to food intake and personal differences in immunological response and general metabolism. These new trends which would either not have been possible or much harder to verify if only single blocks had been analyzed separately. Therefore, the analysis concludes that JUMBA was able to successfully integrate, visualize and interpret data from three different analytical platforms.

Chapter 5

Conclusions and future perspective

All's well as ends better.

J.R.R. Tolkien
The Lord of the Rings

Having the ability to extract information and useful knowledge from data has never been more important than it is now. The wide availability of all types of data has increased greatly over the last decades, and the growth shows no signs of stopping. New methods that are able to improve on some aspect of such information extraction or *data analysis* are constantly being developed.

Multiblock methods and models intend to provide additional information and useful knowledge by integrating several data sources or *blocks*. The work presented in this thesis has attempted to provide the fundamentals required to construct, validate and interpret multiblock models using the JUMBA workflow (*Papers I* and *II*), and also how to use previously created models to predict new data (*Paper II*). A central theme throughout this thesis has been the use of visualizations throughout the different steps in the analysis, and making these visualizations both as simple and informative as possible. These visualizations and the use of the JUMBA workflow has been demonstrated on several real-world cases with different circumstances (*Papers I* to *IV*), proving that the solution is widely applicable.

I believe that the future will bring new methods that take advantage of the stabilizing effect extracting the joint variation has (as seen in *Paper II*) to

simplify further analysis (by providing better validation metrics) and to better predict future events. I also hope that better tools will be developed, such as new visualizations and interactive software solutions, which can further help guide this type of multiblock analysis and make it even more accessible to a broader audience.

Acknowledgements

Så kommer man till slut tills den svåraste, sista och mest lästa delen av hela avhandlingen! Jag tycker att ni som direkt hopppade hit kan i alla fall kanske läsa sammanfattningen på svenska? Den är inte så lång!

Först, ett stort tack till min huvudhandledare **Johan**, för att du lät mig vara en del av din grupp och trott på mig, för alla tips och idéer om hur man ska lösa problem av alla slag, förklarat hur det här med forskning och publicering fungerar och för att du gett mig stor frihet att jobba självständigt.

Tack även till min vice handledare och rumskamrat **David**. Tack för alla diskussioner, inspiration och idéer genom åren, både inom jobbet och kanske ett fåtal saker utanför jobbet såsom datorer, telefoner, spel, musik, filmer, serier, teknik i allmänhet, mat och kanske viktigast, *Pokémon Go*. Utan dig hade jag aldrig börjat med det här doktorerandet, och utan din hjälp hade det heller aldrig gått vägen.

Tack till mina mentorer, **Jun** och **Paul**, som visat mig rätt spår framåt.

Tack även till mina andra rumskamrater som stått ut med den musik jag spelat i mina öppna hörlurar! **Rickard**, som alltid varit så bra att bolla idéer mot och inte varit blyg att säga när något kan förbättras, det har varit otroligt hjälpsamt! **Nabil**, for showing me the ropes at the start of my studies and always encouraging me, without you all this would have been even more confusing than it was!

Till gamla och nuvarande medlemmar i Johans grupp: **Frida**, för din optimism och att du visat var skåpet ska stå när det gäller produktivitet. **Izabella**, som sett till att saker blivit organiserade och faktiskt hänt! Speciellt tack för all konstruktiv kritik du har gett, det har varit otroligt lärorikt! **Daniel**, som jag visat nya sätt att bränna både pengar och tid. Jag hoppas på att kunna visa dig ännu fler såna livsnödvändiga saker i framtiden. **Matilda**, för att du alltid visat den ljusa sidan av allt! **Sergiu**, for all our discussions and for showing me many interesting types of food which I would never have discovered on my own.

Hans och **Pär**, som båda varit otroligt hjälpsamma och haft många insikter i alla möjliga kemometriska problem. Jag ska genast använda OPLS-EP på alla

äpplen, päron, apelsiner och andra frukter jag ser.

Stort tack till alla jag ofta ätit lunch tillsammans med! **Benny**, **Kristina**, **Elin**, **Joakim**, tack för alla viktiga diskussioner om allt möjligt vi haft genom åren. Jag är säker att vi snart har löst det här med fred i världen och sånt.

Resten av gänget i korridoren och i närheten, nuvarande och tidigare: **Henrik**, **Anna**, **Calle**, **Rui** och **Olena**. **Mackan** som försökt lära mig allt jag nu kan om kemi, och hur man spelar *Pokémon Go*. Det senare var nog mer framgångsrikt än det första!

Jag vill även tacka folket i administrationskorridoren, som fått allting att rulla genom åren: **Lars**, **Maria**, **Sara**, **Rosita**, **L-G** och alla andra. Tack också till **Lars Åberg** som hjälpt mig att skriva ut denna avhandling.

Ett speciellt tack till mamma **Maria** och pappa **Tommy**, min bror **Erik**, min släkt och ingifta släkt, vänner och andra som hjälpt mig på olika sätt på vägen!

Min älskade fru **Pia**, för att du funnits där. **Tilde** och **My**, som båda dök upp mitt i hela den här processen, för att ni finns och delar med er av den glädje ni hittar var dag.

Jag hoppas jag inte glömt någon, och om jag gjort det lovar jag på att bjuda på en pepsi! Tack till er allihop!

-Tomas

Bibliography

[1] D. Lahat, T. Adali, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects," *Proceedings of the IEEE*, vol. 103, pp. 1449–1477, Sept. 2015.

[2] M. A. Sharaf, D. L. Illman, B. R. Kowalski, *et al.*, *Chemometrics*, vol. 82. John Wiley & Sons, 1986.

[3] R. G. Brereton, *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons, 2003.

[4] S. Wold, "Chemometrics; what do we mean with it, and what do we want from it?," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, pp. 109–115, Nov. 1995.

[5] R. A. Fisher, *The Design of Experiments*. Oliver And Boyd; Edinburgh; London, 1937.

[6] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold, "Design of experiments," *Principles and Applications, Learn ways AB, Stockholm*, 2000.

[7] H. Martens, T. Naes, and T. Naes, *Multivariate Calibration*. John Wiley & Sons, 1992.

[8] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and Analysis of Computer Experiments," *Statistical Science*, vol. 4, no. 4, pp. 409–423, 1989.

[9] G. E. P. Box and N. R. Draper, *Empirical Model-Building and Response Surfaces*. Empirical Model-Building and Response Surfaces., Oxford, England: John Wiley & Sons, 1987.

[10] B. Chandrasekaran, "Models versus rules, deep versus compiled content versus form: Some distinctions in knowledge systems research," *IEEE Expert*, vol. 6, no. 2, pp. 75–79, 1991.

[11] L. Eriksson and U. AB, eds., *Multi- and Megavariate Data Analysis*. Umetrics Academy - Training in Multivariate Technology, Umeå: Umetrics, 2nd rev. and enl. ed ed., 2006.

[12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37–52, Aug. 1987.

[13] W. Revelle, "Hierarchical Cluster Analysis And The Internal Structure Of Tests," *Multivariate Behavioral Research*, vol. 14, pp. 57–74, Jan. 1979.

[14] F. B. Baker and L. J. Hubert, "Measuring the Power of Hierarchical Cluster Analysis," *Journal of the American Statistical Association*, vol. 70, pp. 31–38, Mar. 1975.

[15] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] B. M. Nicolaï, K. I. Theron, and J. Lammertyn, "Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, pp. 243–252, Feb. 2007.

[18] J. Xing, C. Bravo, D. Moshou, H. Ramon, and J. De Baerdemaeker, "Bruise detection on 'Golden Delicious' apples by vis/NIR spectroscopy," *Computers and Electronics in Agriculture*, vol. 52, pp. 11–20, June 2006.

[19] L. Wang, *Support Vector Machines: Theory and Applications*, vol. 177. Springer Science & Business Media, 2005.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[21] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.

[22] H. Wold, "Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach," *Journal of Applied Probability*, vol. 12, pp. 117–142, Sept. 1975.

[23] S. Wold, "Nonlinear partial least squares modelling II. Spline inner relation," *Chemometrics and Intelligent Laboratory Systems*, vol. 14, pp. 71–84, Apr. 1992.

[24] S. Grossberg, "Nonlinear neural networks: Principles, mechanisms, and architectures," *Neural Networks*, vol. 1, pp. 17–61, Jan. 1988.

[25] T. Naes and B.-H. Mevik, "Understanding the collinearity problem in regression and discriminant analysis," *Journal of Chemometrics*, vol. 15, pp. 413–426, May 2001.

[26] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, pp. 27–46, Jan. 2013.

[27] R. Goldstein, "Conditioning Diagnostics: Collinearity and Weak Data in Regression," *Technometrics*, vol. 35, pp. 85–86, Feb. 1993.

[28] O. M. Kvalheim, "History, philosophy and mathematical basis of the latent variable approach - from a peculiarity in psychology to a general method for analysis of multivariate data: History, philosophy and mathematical basis of the latent variable," *Journal of Chemometrics*, vol. 26, pp. 210–217, June 2012.

[29] S. Lafi and J. Kaneene, "An explanation of the use of principal-components analysis to detect and correct for multicollinearity," *Preventive Veterinary Medicine*, vol. 13, pp. 261–275, Sept. 1992.

[30] C. Feng, H. Wang, N. Lu, T. Chen, H. He, Y. Lu, and X. M. Tu, "Log-transformation and its implications for data analysis," *Shanghai Archives of Psychiatry*, vol. 26, pp. 105–109, Apr. 2014.

[31] M. Dhanoa, S. Lister, R. Sanderson, and R. Barnes, "The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra," *Journal of Near Infrared Spectroscopy*, vol. 2, no. 1, pp. 43–47, 1994.

[32] H. Redestig, A. Fukushima, H. Stenlund, T. Moritz, M. Arita, K. Saito, and M. Kusano, "Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data," *Analytical Chemistry*, vol. 81, pp. 7974–7980, Oct. 2009.

[33] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg, and S. Wold, "An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 44, pp. 229–244, Dec. 1998.

[34] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[35] L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, and S. Wold, "Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics

(gpm)," *Analytical and Bioanalytical Chemistry*, vol. 380, pp. 419–429, Oct. 2004.

[36] J. Trygg, E. Holmes, and T. Lundstedt, "Chemometrics in Metabonomics," *Journal of Proteome Research*, vol. 6, pp. 469–479, Feb. 2007.

[37] M. D. Farrell and R. M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 2, pp. 192–195, 2005.

[38] L. Cao, K. S. Chua, W. Chong, H. Lee, and Q. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1-2, pp. 321–336, 2003.

[39] A. Globerson and N. Tishby, "Sufficient Dimensionality Reduction," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1307–1331, 2003.

[40] M. Defernez and E. Kemsley, "The use and misuse of chemometrics for treating classification problems," *TrAC Trends in Analytical Chemistry*, vol. 16, pp. 216–221, Apr. 1997.

[41] M. Gavish and D. L. Donoho, "The optimal hard threshold for singular values is 4/sqrt(3)," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 5040–5053, 2014.

[42] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 109–130, Oct. 2001.

[43] L. Ståhle and S. Wold, "Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study," *Journal of Chemometrics*, vol. 1, no. 3, pp. 185–196, 1987.

[44] J. Trygg and S. Wold, "Orthogonal projections to latent structures (O-PLS)," *Journal of Chemometrics*, vol. 16, pp. 119–128, Mar. 2002.

[45] T. Verron, R. Sabatier, and R. Joffre, "Some theoretical properties of the O-PLS method," *Journal of Chemometrics*, vol. 18, pp. 62–68, Feb. 2004.

[46] S. Wiklund, E. Johansson, L. Sjöström, E. J. Mellerowicz, U. Edlund, J. P. Shockcor, J. Gottfries, T. Moritz, and J. Trygg, "Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models," *Analytical Chemistry*, vol. 80, pp. 115–122, Jan. 2008.

[47] N. E. Madala, L. A. Piater, P. A. Steenkamp, and I. A. Dubery, "Multivariate statistical models of metabolomic data reveals different metabolite distribution patterns in isonitrosoacetophenone-elicited Nicotiana tabacum and Sorghum bicolor cells," *SpringerPlus*, vol. 3, no. 1, p. 254, 2014.

[48] M. Bylesjö, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg, "OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification," *Journal of Chemometrics*, vol. 20, pp. 341–351, Aug. 2006.

[49] M. Bylesjö, D. Eriksson, M. Kusano, T. Moritz, and J. Trygg, "Data integration in plant biology: The O2PLS method for combined modeling of transcript and metabolite data," *The Plant Journal*, vol. 52, pp. 1181–1191, Dec. 2007.

[50] S. el Bouhaddani, J. Houwing-Duistermaat, P. Salo, M. Perola, G. Jongbloed, and H.-W. Uh, "Evaluation of O2PLS in Omics data integration," *BMC Bioinformatics*, vol. 17, Dec. 2016.

[51] M. Rantalainen, O. Cloarec, O. Beckonert, I. D. Wilson, D. Jackson, R. Tonge, R. Rowlinson, S. Rayner, J. Nickson, R. W. Wilkinson, J. D. Mills, J. Trygg, J. K. Nicholson, and E. Holmes, "Statistically Integrated Metabonomic-Proteomic Studies on a Human Prostate Cancer Xenograft Model in Mice," *Journal of Proteome Research*, vol. 5, pp. 2642–2655, Oct. 2006.

[52] R. Consonni, L. R. Cagliani, M. Stocchero, and S. Porretta, "Evaluation of the Production Year in Italian and Chinese Tomato Paste for Geographical Determination Using O2PLS Models," *Journal of Agricultural and Food Chemistry*, vol. 58, pp. 7520–7525, July 2010.

[53] K. Kjeldahl and R. Bro, "Some common misunderstandings in chemometrics," *Journal of Chemometrics*, vol. 24, pp. 558–564, July 2010.

[54] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, pp. 1–4, Springer, 2009.

[55] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 15, p. 72, Jan. 1904.

[56] M. G. Kendall and J. D. Gibbons, *Rank Correlation Methods*. London : New York, NY: E. Arnold ; Oxford University Press, 5th ed ed., 1990.

[57] C. Granger, "Testing for causality," *Journal of Economic Dynamics and Control*, vol. 2, pp. 329–352, Jan. 1980.

[58] J. Hauke and T. Kossowski, "Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data," *Quaestiones Geographicae*, vol. 30, pp. 87–93, June 2011.

[59] T. N. Tran, N. L. Afanador, L. M. Buydens, and L. Blanchet, "Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC)," *Chemometrics and Intelligent Laboratory Systems*, vol. 138, pp. 153–160, Nov. 2014.

[60] M. Farrés, S. Platikanov, S. Tsakovski, and R. Tauler, "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation: Comparison of variable selection methods," *Journal of Chemometrics*, vol. 29, pp. 528–536, Oct. 2015.

[61] M. Partridge and R. Calvo, "Fast dimensionality reduction and simple PCA," *Intelligent Data Analysis*, vol. 2, no. 1-4, pp. 203–214, 1998.

[62] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative review," tech. rep., Tilburg centre for Creative Computing, 2009.

[63] G. I. Allen, C. Peterson, M. Vannucci, and M. Maletić-Savatić, "Regularized partial least squares with an application to NMR spectroscopy," *Statistical Analysis and Data Mining*, vol. 6, pp. 302–314, Aug. 2013.

[64] I. Guttman and D. E. Smith, "Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution II: Estimation of the Variance," *Technometrics*, vol. 13, pp. 101–111, Feb. 1971.

[65] R. Ratcliff, "Methods for dealing with reaction time outliers.," *Psychological Bulletin*, vol. 114, no. 3, pp. 510–532, 1993.

[66] I. Stanimirova, M. Daszykowski, and B. Walczak, "Dealing with missing values and outliers in principal component analysis," *Talanta*, vol. 72, pp. 172–178, Apr. 2007.

[67] H. Hotelling, "The Generalization of Student's Ratio," *The Annals of Mathematical Statistics*, vol. 2, pp. 360–378, Aug. 1931.

[68] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[69] T. Aittokallio, "Dealing with missing values in large-scale studies: Microarray data imputation and beyond," *Briefings in Bioinformatics*, vol. 11, pp. 253–264, Mar. 2010.

[70] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, vol. 793. Wiley, 2019.

[71] P. R. Nelson, P. A. Taylor, and J. F. MacGregor, "Missing data methods in PCA and PLS: Score calculations with incomplete observations," *Chemometrics and Intelligent Laboratory Systems*, vol. 35, pp. 45–65, Nov. 1996.

[72] J. H. WILKINSON, F. L. Bauer, and C. Reinsch, *Linear Algebra.* Berlin, Heidelberg: Springer Berlin / Heidelberg, 2013. OCLC: 1066197562.

[73] R. G. Downey and C. V. King, "Missing Data in Likert Ratings: A Comparison of Replacement Methods," *The Journal of General Psychology*, vol. 125, pp. 175–191, Apr. 1998.

[74] M. D. Ruopp, N. J. Perkins, B. W. Whitcomb, and E. F. Schisterman, "Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection," *Biometrical Journal*, vol. 50, pp. 419–430, June 2008.

[75] T. Skotare, R. Sjögren, I. Surowiec, D. Nilsson, and J. Trygg, "Visualization of descriptive multiblock analysis: Visualization of descriptive multiblock analysis," *Journal of Chemometrics*, p. e3071, July 2018.

[76] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view Discriminant Analysis," in *Computer Vision – ECCV 2012* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), vol. 7572, pp. 808–821, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[77] D. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, Jan./1997.

[78] T. Löfstedt and J. Trygg, "OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation," *Journal of Chemometrics*, 2011.

[79] T. Löfstedt, *OnPLS Orthogonal Projections to Latent Structures in Multiblock and Path Model Data Analysis*. Umeå: Department of Chemistry, Umeå University, 2012. OCLC: 939796521.

[80] T. Löfstedt, D. Hoffman, and J. Trygg, "Global, local and unique decompositions in OnPLS for multiblock data analysis," *Analytica Chimica Acta*, vol. 791, pp. 13–24, Aug. 2013.

[81] V. Srivastava, O. Obudulu, J. Bygdell, T. Löfstedt, P. Rydén, R. Nilsson, M. Ahnlund, A. Johansson, P. Jonsson, E. Freyhult, J. Qvarnström, J. Karlsson, M. Melzer, T. Moritz, J. Trygg, T. R. Hvidsten, and G. Wingsle, "OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants," *BMC Genomics*, vol. 14, no. 1, p. 893, 2013.

[82] F. Pettersson, P. A. Stewart, R. J. Slebos, E. A. Welsh, L. Cen, Y. Zhang, Z. Chen, C.-H. Cheng, G. Zhang, B. Fang, V. Izumi, S. Yoder, K. Fellows, Y. A. Chen, J. K. Teer, S. Eschrich, J. M. Koomen, A. Berglund, and E. B. Haura, "Abstract 1565: OnPLS-based integrative proteogenomics analysis of lung squamous cell cancer," *Cancer Research*, vol. 77, pp. 1565–1565, July 2017.

[83] J. A. Westerhuis, T. Kourti, and J. F. MacGregor, "Analysis of multiblock and hierarchical PCA and PLS models," *Journal of Chemometrics*, vol. 12, pp. 301–321, Sept. 1998.

[84] S. Wold, N. Kettaneh, and K. Tjessem, "Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection," *Journal of Chemometrics*, vol. 10, pp. 463–482, Sept. 1996.

[85] A. K. Smilde, I. Måge, T. Naes, T. Hankemeier, M. A. Lips, H. A. L. Kiers, E. Acar, and R. Bro, "Common and distinct components in data fusion: Common and distinct components in data fusion," *Journal of Chemometrics*, vol. 31, p. e2900, July 2017.

[86] T. Næs, O. Tomic, N. K. Afseth, V. Segtnan, and I. Måge, "Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 32–42, May 2013.

[87] I. Måge, E. Menichelli, and T. Næs, "Preference mapping by PO-PLS: Separating common and unique information in several data blocks," *Food Quality and Preference*, vol. 24, pp. 8–16, Apr. 2012.

[88] B. Worley and R. Powers, "A sequential algorithm for multiblock orthogonal projections to latent structures," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 33–39, Dec. 2015.

[89] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types," *The Annals of Applied Statistics*, vol. 7, pp. 523–542, Mar. 2013.

[90] F. M. van der Kloet, P. Sebastián-León, A. Conesa, A. K. Smilde, and J. A. Westerhuis, "Separating common from distinctive variation," *BMC Bioinformatics*, vol. 17, Dec. 2016.

[91] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*, pp. 134–151, Springer, 1971.

[92] J. A. Saunders, N. Morrow-Howell, E. Spitznagel, P. Dore, E. K. Proctor, and R. Pescarino, "Imputing Missing Data: A Comparison of Methods for Social Work Researchers," *Social Work Research*, vol. 30, pp. 19–31, Mar. 2006.

[93] M. Harrower and C. A. Brewer, "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps," *The Cartographic Journal*, vol. 40, pp. 27–37, June 2003.

[94] O. Obudulu, N. Mähler, T. Skotare, J. Bygdell, I. N. Abreu, M. Ahnlund, M. Latha Gandla, A. Petterle, T. Moritz, T. R. Hvidsten, L. J. Jönsson, G. Wingsle, J. Trygg, and H. Tuominen, "A multi-omics approach reveals function of Secretory Carrier-Associated Membrane Proteins in wood formation of Populus trees," *BMC Genomics*, vol. 19, Dec. 2018.

[95] P. Geladi, M. Manley, and T. Lestander, "Scatter plotting in multivariate data analysis," *Journal of Chemometrics*, vol. 17, pp. 503–511, Aug. 2003.

[96] M. Dyrby, "Towards on-line monitoring of the composition of commercial carrageenan powders," *Carbohydrate Polymers*, vol. 57, pp. 337–348, Sept. 2004.

[97] L. Eriksson, J. Trygg, and S. Wold, "A chemometrics toolbox based on projections and latent variables: A chemometrics toolbox based on projections and latent variables," *Journal of Chemometrics*, vol. 28, pp. 332–346, May 2014.

[98] M. Bylesjö, R. Nilsson, V. Srivastava, A. Grönlund, A. I. Johansson, S. Jansson, J. Karlsson, T. Moritz, G. Wingsle, and J. Trygg, "Integrated Analysis of Transcript, Protein and Metabolite Data To Study Lignin Biosynthesis in Hybrid Aspen," *Journal of Proteome Research*, vol. 8, pp. 199–210, Jan. 2009.

[99] J. J. Workman, "A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy," *Applied Spectroscopy*, vol. 72, pp. 340–365, Mar. 2018.

[100] O. E. de Noord, "Multivariate calibration standardization," *Chemometrics and Intelligent Laboratory Systems*, vol. 25, pp. 85–97, Nov. 1994.

[101] V. H. da Silva, J. J. da Silva, and C. F. Pereira, "Portable near-infrared instruments: Application for quality control of polymorphs in pharmaceutical raw materials and calibration transfer," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 134, pp. 287–294, Feb. 2017.

[102] R. N. Feudale, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, and J. Ferré, "Transfer of multivariate calibration models: A review," *Chemometrics and Intelligent Laboratory Systems*, vol. 64, pp. 181–192, Nov. 2002.

[103] Y. Wang, D. J. Veltkamp, and B. R. Kowalski, "Multivariate instrument standardization," *Analytical Chemistry*, vol. 63, pp. 2750–2756, Dec. 1991.

[104] M. L. Griffiths, D. Svozil, P. Worsfold, and E. Hywel Evans, "The application of piecewise direct standardisation with variable selection to the correction of drift in inductively coupled atomic emission spectrometry," *Journal of Analytical Atomic Spectrometry*, vol. 21, no. 10, p. 1045, 2006.

[105] T. M. Alam, M. K. Alam, S. K. McIntyre, D. E. Volk, M. Neerathilingam, and B. A. Luxon, "Investigation of Chemometric Instrumental Transfer Methods for High-Resolution NMR," *Analytical Chemistry*, vol. 81, pp. 4433–4443, June 2009.

[106] W. Du, Z.-P. Chen, L.-J. Zhong, S.-X. Wang, R.-Q. Yu, A. Nordon, D. Littlejohn, and M. Holden, "Maintaining the predictive abilities of multivariate calibration models by spectral space transformation," *Analytica Chimica Acta*, vol. 690, pp. 64–70, Mar. 2011.

[107] A. Andrew and T. Fearn, "Transfer by orthogonal projection: Making near-infrared calibrations robust to between-instrument variation," *Chemometrics and Intelligent Laboratory Systems*, vol. 72, pp. 51–56, June 2004.

[108] R. W. Kennard and L. A. Stone, "Computer Aided Design of Experiments," *Technometrics*, vol. 11, p. 137, Feb. 1969.

[109] A. Saptoro, M. O. Tadé, and H. Vuthaluru, "A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models," *Chemical Product and Process Modeling*, vol. 7, Jan. 2012.

[110] F. Torell, K. Bennett, S. Cereghini, S. Rännar, K. Lundstedt-Enkel, T. Moritz, C. Haumaitre, J. Trygg, and T. Lundstedt, "Multi-Organ Contribution to the Metabolic Plasma Profile Using Hierarchical Modelling," *PLOS ONE*, vol. 10, p. e0129260, June 2015.

[111] F. Torell, K. Bennett, S. Cereghini, M. Fabre, S. Rännar, K. Lundstedt-Enkel, T. Moritz, C. Haumaitre, J. Trygg, and T. Lundstedt, "Metabolic Profiling of Multiorgan Samples: Evaluation of MODY5/RCAD Mutant Mice," *Journal of Proteome Research*, vol. 17, pp. 2293–2306, July 2018.