Multivariate analysis – method development and novel applications in spectrometry

Multivariat analyse – metodeutvikling og
nye anvendelser innen spektrometri

Philosophiae Doctor (PhD) Thesis

Kristian Hovde Liland

Dept. of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences

Ås 2009

Thesis number 2009: 37
ISSN 1503-1667
ISBN 978-82-575-0906-4

Preface

This thesis is submitted to attain the doctoral degree Philosophiae Doctor (PhD) at the Norwegian University of Life Sciences (UMB), Norway.

The work has been carried out in the period 2006-2009 at the Department of Chemistry, Biotechnology and Food Science (IKBM) under the supervision of Trygve Almøy. This thesis consists of an introduction and 6 enclosed papers.

During my time as a PhD student I have worked together with many highly competent scientists and students. First I would like to thank my main supervisor Trygve Almøy and my supervisor from my time as a master student Ulf Geir Indahl, both of whom I have written several papers together with and received excellent guidance from. In addition I have also learnt a lot from my fellow authors Tomas Isaksson, Elling-Olav Rukke, Ellen Mosleth Færgestad, Morten Skaugen and Tormod Næs.

Reaching this point in my education would of course have been impossible without my parents, family and fellow students supporting me and believing in me. There are many who can take some of the credit for my achievements, all the way from lighting my passion for the natural sciences as a child, steering me away from becoming an electrician in secondary high (dad), finding a perfect bachelor study after realising that I did not want to become a guitar teacher (Mona Liland Aabel), introducing me to statistics (Henrik Dahl), making me aware of the existence of UMB and accompanying me there (Alf Synstad) to giving me opportunities to holding lectures in statistics both during my bachelor and PhD studies. Finally I would like to thank my mother for proof reading of this thesis.

This thesis and all accompanying papers have been written in the open source document processor LyX which is based on the document markup language and preparation system LaTeX. Programming, calculations and plotting has been done both in the numerical computing environment MATLAB and in the open source statistical language and environment R.

Ås, October 2009
Kristian Hovde Liland

List of papers

I. Liland, K.H., Indahl, U.G. (2009). *Powered partial least squares discriminant analysis*, Journal of Chemometrics **23**(1), pp. 7-18

II. Indahl, U.G., Liland, K.H., Næs, T. (2009). *Canonical partial least squares — a unified PLS approach to classification and regression problems*, Journal of Chemometrics **23**(9), pp. 495-504

III. Liland, K.H., Færgestad, E.M. (2009). *Testing effects of experimental design factors using multi-way analysis*, Chemometrics and Intelligent Laboratory Systems **96**(2), pp. 172-181

IV. Liland, K.H., Mevik, B.H., Rukke, E.O., Almøy, T., Skaugen, M., Isaksson, T. (2009). *Quantitative whole spectrum analysis with MALDI-TOF MS, Part I: Measurement optimisation*, Chemometrics and Intelligent Laboratory Systems **96**(2), pp. 210-218

V. Liland, K.H., Mevik, B.H., Rukke, E.O., Almøy, T., Isaksson, T. (2009). *Quantitative whole spectrum analysis with MALDI-TOF MS, Part II: Determining the concentration of milk in mixtures*, Chemometrics and Intelligent Laboratory Systems **99**(1), pp. 39-48

VI. Liland, K.H., Almøy, T., Mevik, B.H. *Optimal choice of baseline correction for statistical analysis of spectra*, *submitted manuscript*

Summary

The papers included in this thesis cover a range of theoretical and applied aspects in the fields of multivariate and multi-way statistics, design and analysis of experiments and quantitative spectrometry. Papers I and II contribute to the development of multivariate analysis methods by extending and combining existing methodology. The dimension reducing and data compressing method partial least squares (PLS) has earlier been generalised by the inclusion of powers in the construction of vectors of loading weights used to make linear combinations of the original variables. We have adapted the use of powers to discrete responses and combinations of discrete and continuous responses through a common optimisation criterion based on canonical correlation analysis (CCA). The latter also introduces the possibility of including additional responses in the modelling, while still optimising only on a subset of the responses.

Generalised multiplicative analysis of variance (GEMANOVA) has earlier been suggested as a supplement to ordinary analysis of variance (ANOVA), enabling analysis of data without replicates or having missing data points, while opening up for more complex interactions and handling of several underlying phenomena simultaneously. It has not been clear how to validate the models produced by GEMANOVA, partially because of difficulties in estimating the degrees of freedom consumed by the models. We have proposed a model-based bootstrap procedure for estimating the variability of factor levels without the need for degrees of freedom. This is meant as an aid in explorative analysis for assessing the consistency of effects when building models.

Matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) is a well established mass spectrometry (MS) technique ordinarily used for identification of proteins, peptides and other ionisable compounds. In Papers IV and V we propose an optimisation strategy and perform a full scale analysis of milk mixtures based on whole spectra from MALDI-TOF. Through this work we demonstrate how MALDI-TOF MS can be used in a quantitative way with a minimum of effort and knowledge about proteins, avoiding peak or variable selections, while obtaining good predictions and repeatability. The work is continued in Paper VI where the focus is on optimal selection of baseline correction algorithms and their respective parameter values for spectra used in statistical analyses. In this paper the suggested procedure is tested on data from Raman spectroscopy and the same MALDI-TOF data demonstrating some of the potential advantages and pitfalls of baseline correction.

Sammendrag

Artiklene (paper) i denne avhandlingen dekker en rekke teoretiske og anvendte aspekter innen multivariat og multi-way statistikk, design og analyse av eksperimenter og kvantitativ spektrometri. Paper I og II bidrar til utviklingen av multivariate analysemetoder ved å utvide og kombinere eksisterende metodologi. Den dimensjonsreduserende og datakomprimerende metoden partial least squares (PLS) har tidligere blitt generalisert ved å innføre potenser i beregningen av ladningsvektvektorer brukt for å lage lineærkombinasjoner av de originale variablene. Vi har tilpasset bruken av potenser til kategoriske responser og kombinasjoner av kategoriske og kontinuerlige responser gjennom et felles optimeringskriterium basert på kanonisk korrelasjonsanalyse (CCA). Sistnevnte åpner også muligheten til å inkludere ekstra responser i modelleringen, mens det fortsatt kun optimeres på en undergruppe av responsene.

Generalised multiplicative analysis of variance (GEMANOVA) har tidligere vært foreslått som et supplement til ordinær variansanalyse. Dette gjør det mulig å analysere data uten replikater eller med manglende datapunkter, mens det åpner opp for mer komplekse samspill og behandling av flere underliggende fenomener samtidig. Det har ikke vært klart hvordan en skulle validere modellene produsert av GEMANOVA, delvis fordi en har vanskeligheter med å beregne antallet frihetsgrader som blir brukt av modellene. Vi har foreslått en modellbasert bootstrap-prosedyre for å estimere hvor variable faktornivåene er, for å kunne si noe om hvor konsistent effektene er, når en bygger modeller.

Matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) er en veletablert massespektrometriteknikk som vanligvis brukes til å identifisere proteiner, peptider og andre ioniserbare stoffer. I Paper IV og V foreslår vi en optimeringsstrategi og utfører en fullskala analyse av melkeblandinger basert på hele spektre fra MALDI-TOF. Gjennom dette arbeidet viser vi hvordan MALDI-TOF massespektrometri kan brukes på en kvantitativ måte med et minimum av innsats og liten kjennskap til proteiner, mens man unngår topp- og variabelseleksjon, og oppnår gode prediksjoner og repeterbarhet. Arbeidet fortsettes i Paper VI hvor fokus er på optimale valg av grunnlinjekorreksjonsalgoritmer og deres respektive parameterverdier for spektra brukt i statistisk analyse. I denne artikkelen blir den foreslåtte prosedyren testet på data fra Raman spektroskopi og de samme MALDI-TOF-dataene slik at noen av de potensielle fordelene og fallgruvene til grunnlinjekorreksjon blir belyst.

Contents

1 Aim of the study

The main objective of the PhD project has been to further develop, explore and understand aspects of multivariate statistics. Continuing the work from my master thesis on powered partial least squares discriminant analysis (PPLS-DA), extending it and applying the power methodology to new problems, has been a natural part of this objective. As this PhD is in the field of applied statistics and has been carried out in a biostatistics group, applications are of high importance both for assessing the performance of newly developed statistical methods, for increasing the understanding of biological samples and for extending the applications of existing analysis methods and instrumentation.

2 Applied statistics

In recent years samples from biology, chemistry, industrial processes and other areas have become more and more multivariate. Instead of measuring only one or a few properties of the samples or processes, hundreds and sometimes thousands of attributes are often recorded. Usually these attributes (or variables) are highly connected, resulting in complex covariance structures and a need for statistical methods able to reduce the number of variable dimensions or stabilise the covariance. Typical examples of samples containing many, correlated variables are the spectra from Raman, NIR, MALDI-TOF and other spectrometric instruments. Here neighbouring variables are often highly correlated while variables farther away from each other have little or no correlation.

Other types of data containing a vast amount of variables are gene sequences and micro arrays. Here the interactions and interdependencies are even more intricate. Analysis methods that are able to extract relevant information from the massive amount of noise, multicollinearity, interacting and interfering phenomena and multiple dependence structures is therefore of high importance. Gene sequences are sometimes called megavariate when the number of base pairs extracted are counted in hundreds of thousands. In such cases the risk of finding spurious correlations is so large that splitting or simplifying the variable space before the main analysis can become unavoidable.

With an almost exponential growth in the production of ever more complex and specialised data an increasing demand for applied statisticians and chemometricians is evident, both for method development and data analysis. Regrettably for project managers and funders there is no uniformly best prediction method or black box analysis to generate interpretable results

from any data. Furthermore the complexity of the analyses are often increasing with the complexity of the data, so that statistical expertise is required for implementation and quality control.

In the following subsections we will present some of the typical steps in the process from planning an experiment to assessing the goodness of predicted responses in multivariate statistics. As the number of available methods for each step of the process is too high to be included in this thesis, we will focus on the methods used directly in the accompanying papers or building up to them. All predictions in this paper are based on linear models. The basic linear regression model containing n samples and p explanatory variables is:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of explanatory variables, β is a $p \times 1$ vector of regression coefficients indicating the linear relationship between $\mathbf{X}$ and $\mathbf{y}$, and ϵ is the error ($\epsilon = \mathbf{y} - \mathbf{X}\beta$). Here $\mathbf{X}$ is in practise either presumed centred or augmented with a column of ones ($p + 1$ columns in total) accounting for the intercept (β is adjusted accordingly).

2.1 Factorial and mixture designs

Factorial designs

Factorial designs [1] (and various modification) are among the most widely used experimental designs in the literature. In its most basic form we find the orthogonal 2^k factorial designs where the experimental factors are varied over two levels in k factor dimensions. These can be visualised as squares, cubes or hypercubes, and are usually simple to analyse through analysis of variance (ANOVA) and testing with contrasts because of their orthogonal nature and often independent samples.

Among the benefits of the factorial designs is the ability to span relevant portions of many experimental setups in an economical manner. These are also easily expanded with axis and centre runs to form central composite designs enabling second order modelling that describes curvature in the response. In addition fractions of the 2^k factorial designs can be used to further economise the number of samples needed. A careful choice of fraction will ensure that the most interesting effects and interactions will be possible to distinguish, while others may be confounded.

In some situations it is not practical to generate the full set of samples needed to fill the whole factorial design in all factors. When one or more factors consist of varying dilutions or mixtures of dilutions these should for instance be mixed once for each combination of

all factors in the design. A simpler way of doing it is to only mix as many dilutions as are needed to fill the (hyper)cube generated by only the factors containing dilutions (or a multiple of this number) and reuse the mixtures in a structured way across the other factors. This introduces, so called, error strata which split the error of the ANOVA into two or more errors and associates all factors and interactions to these strata. In practise this gives a lower sensitivity in the ANOVA, but might also mean that the experiment becomes affordable or feasible.

Mixture Designs
Another widely used design is the mixture design [1], often in the form of a simplex lattice or simplex centroid design. As the name indicates this category of designs deals with liquids, baking ingredients and other types of compounds that will be analysed as mixtures. The basic form of a simplex design is an equilateral triangle having pure compounds in the vertices and two-compound mixtures along the edges. Depending on the resolution of the lattice, a number of grid lines can cut through the triangle between the vertices and the opposite edge (parallel to this edge). All crosses between these sets of grid lines will symbolise one mixture where all compounds sum to a concentration of 100 %. The simplex centroid design also introduces centre runs and possible axial points. All simplex designs can be extended with more compounds, giving rise to tetrahedrons (triangular pyramids) or generalisations in more dimensions. When the application means that a full simplex is impossible, expensive or impractical, some points can be removed, e.g. pure compounds, usually at the cost of lower coverage and sensitivity.

2.2 Prediction modelling

Modelling of data for prediction is a key element in this thesis. In most of the covered applications emphasis is on predicting a response accurately and robustly rather than discovering the relations and mechanisms generating the data. In this context we come across both continuous responses (regression) and discrete responses (classification). Examples of continuous responses can be concentrations or other quantities that are measurable on a continuous scale. Discrete responses can be oil types, positive/negative diagnoses or other non-overlapping qualities or states.

When working with continuous responses, e.g. concentrations, the aim of the prediction is to produce models that give predictions deviating as little from the true response as possible. A common summary for such deviations is the square root of the expected deviation between the true response and the predicted values, $\theta = \sqrt{E(\mathbf{y} - \hat{\mathbf{y}})^2}$. This can be estimated by the root mean square error (RMSE):

$$\hat{\theta}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i),k}\right)^2}. \quad (2)$$

Here k is a parameter controlling the model complexity, n is the number of samples, y_i is the i-th observation and $\hat{y}_{(i),k}$ is the prediction of the i-th observation using a model where observation i has not influenced the modelling. RMSE is often plotted as a curve with complexity increasing along the x-axis.

Having a discrete response means predictions are either correct or not, as the classes are unordered and non-overlapping. The response is often coded as a dummy matrix having as many columns as classes and as many rows as observations. Each observation has a 1 in the column corresponding to its class and zeros elsewhere. A common summary for such predictions is the mis-classification rate or conversely the proportion/percentage correctly classified observations. From some classification methods we can extract the estimated probabilities of each object originating from each of the classes. This can for instance be summarised by the root mean square deviation between the estimated probabilities and the dummy response matrix.

2.3 Feature extraction

The curse of dimensionality

In classical statistics it is usually a requirement that the analysed data have more samples (n) than variables (p) and that the variables are not too multi-collinear. Linear regression between an $n \times p$ data matrix $\mathbf{X}$ and an $n \times 1$ response vector $\mathbf{y}$ involves computing a $p \times 1$ regression vector $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ where the $p \times p$ covariance matrix $\mathbf{X}'\mathbf{X}$ (centred $\mathbf{X}$ matrices) needs to be non-singular. This condition is not met by the vast majority of spectrometric and genetic data that are produced nowadays. Often data sets contain a few tens or hundreds of samples while the variables are measured in hundreds, thousands or even more. This means that classical statistical methods like linear regression fail, and we have to rely on methods performing variable selection, dimension reduction or covariance stabilisation. In this thesis focus will be on dimension reduction, though variable selection will be applied in one of the papers.

PCA

One of the most basic forms of feature extraction is principal component analysis where the data $\mathbf{X}$ are decomposed by singular value decomposition (SVD) as $\mathbf{X} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}'$. The left singular orthogonal vectors in $\mathbf{U}$ ($n \times n$) are called scores and relate to the samples, while the right singular orthogonal vectors in $\mathbf{V}'$ ($p \times p$) are called vectors of loading weights and relate to the variables. In addition there is an $n \times p$ matrix $\mathbf{S}$ having ordered singular values

on its diagonal, decreasing toward the lower right corner. The singular values are related to the amount of variation in $\mathbf{X}$ described by each pair of singular vectors through the product $\mathbf{u}_i \cdot s_i \cdot \mathbf{v}'_i$. In practise PCA compresses the original variables down to a few representative pseudo variables that are linear combinations of the original variables. This usually means that most of the variation in the variable space is now made available by a few variables spanning a relevant subspace of $\mathbf{X}$. When PCA is used in regression it is called principal component regression (PCR), and the pairs of singular vectors are called score vectors and vectors of loading weights, respectively, together forming so called principal components.

PLS

When feature extraction is used for prediction modelling, a more effective decomposition of the variable space can often be done by partial least squares (PLS) [2]. Instead of only focusing on the data matrix, $\mathbf{X}$, the response, $\mathbf{y}$, is taken into account and the empirical covariance between $\mathbf{X}$ and $\mathbf{y}$ is maximised. Through a series of feature extractions and deflations of $\mathbf{X}$ and $\mathbf{y}$ vectors of loading weights, $\mathbf{w}_i$ $(p \times 1)$, spanning the direction of highest covariance between the residual $\mathbf{X}_{\{i\}}$ and $\mathbf{y}_{\{i\}}$, and corresponding score vectors, $\mathbf{t}_i$ $(n \times 1)$, are produced. PLS is widely used in multivariate statistics, chemometrics and other fields where data matrices are wide and the relation to a known response is of interest. When PLS is employed in regression (PLSR), the $\mathbf{X}$ matrix, used when calculating the regression vector in linear regression, is replaced by an appropriate number (k) of score vectors combined in an $n \times k$ matrix $\mathbf{T}$ and vectors of loading weights combined in a $p \times k$ matrix $\mathbf{W}$ to form $\hat{\beta} = \mathbf{W}' \left(\mathbf{T}'\mathbf{T}\right)^{-1} \mathbf{T}'\mathbf{y}$ $(p \times 1)$. The number of components (k) is usually decided by applying cross-validation as described earlier. Using PLSR the complexity parameter in RMSECV becomes the number of components included in the model (k).

When there is more than one continuous response comprising a response matrix, $\mathbf{Y}$ $(n \times q)$, ordinary PLS cannot be used directly without splitting the response and computing one PLS model for each. Instead we have to rely on the multi response version, PLS2, to obtain a common model for all responses. Whereas the vectors of loading weights in PLS are simply the products of the residual $\mathbf{X}_{\{i\}}$s and $\mathbf{y}_{\{i\}}$s, these now become the first right singular vector of the SVDs of the products of residual $\mathbf{X}_{\{i\}}$s and $\mathbf{Y}_{\{i\}}$s: $\mathbf{X}'_{\{i\}}\mathbf{Y}_{\{i\}}$ $(p \times q)$. This gives a common model for the relation between the data matrix and each of the response vectors, as opposed to making one PLS model for each of the responses.

PLS-DA

While PLS and PLS2 are used when the responses are continuous, PLS discriminant analysis (PLS-DA) [3] is the discrete counterpart (g distinct groups). If the response is discrete, maximisation of covariance is not optimal, and repeated deflation of the response does not make sense. PLS-DA instead maximises the empirical between-groups covariance of the residual $\mathbf{X}_{\{i\}}$ matrix and the dummy response matrix $\mathbf{Y}$. This is done by repeatedly

performing SVD on the $p \times p$ between-groups sum of squares and cross-products matrix $\mathbf{B} = \mathbf{X}'_{\{i\}}\mathbf{Y}\left(\mathbf{Y}'\mathbf{Y}\right)^{-1}\mathbf{Y}'\mathbf{X}_{\{i\}}$, or equivalently on the usually much smaller $g \times p$ matrix $\bar{\mathbf{X}}_g = \left(\mathbf{Y}'\mathbf{Y}\right)^{-1}\mathbf{Y}'\mathbf{X}_{\{i\}}$, and using the first right singular vector as the vector of loading weights. Indahl et al. [4] showed how arbitrary group weighting can be applied to alter the focus of the feature extraction. They also showed how a transformation of the data could be used to enable maximisation of correlation instead of covariance through the Rayleigh coefficient $\mathbf{T}^{-1}\mathbf{B}$ associated with Fisher's canonical discriminant analysis (FCDA). The score vectors and vectors of loading weights produced by PLS-DA are usually combined with a classification method like linear discriminant analysis (LDA) for the final classification of the samples and calculation of probabilities of the individual samples coming from each group.

PPLS

Indahl [5] has developed a generalised version of PLS called powered PLS (PPLS) where the vectors of loading weights are factorised into $\mathbf{y}$-$\mathbf{X}$ correlations and $\mathbf{X}$ standard deviations and parameterised with powers:

$$\mathbf{w}(\gamma) = K_\gamma \cdot \begin{bmatrix} s_1 \cdot |corr\left(\mathbf{y}, \mathbf{x}_1\right)|^{\frac{\gamma}{1-\gamma}} \cdot std\left(\mathbf{x}_1\right)^{\frac{1-\gamma}{\gamma}} \\ \vdots \\ s_p \cdot |corr\left(\mathbf{y}, \mathbf{x}_p\right)|^{\frac{\gamma}{1-\gamma}} \cdot std\left(\mathbf{x}_p\right)^{\frac{1-\gamma}{\gamma}} \end{bmatrix}. \tag{3}$$

Here K_γ is a constant absorbing the $\mathbf{y}$ standard deviation and assuring unit length of the vector, s_j are the signs of the correlations and γ is a parameter optimised over an interval $U \subseteq [0, 1]$ through $\underset{\gamma}{max}\left(corr\left(\mathbf{y}_{\{i\}}, \mathbf{X}_{\{i\}}\mathbf{w}\left(\gamma\right)\right)\right)$, for each component. This enables sharpened focus on variables having high standard deviations or high correlations to the response. Ordinary PLS can be achieved by keeping $\gamma = 0.5$ for all components. The increased flexibility means PPLS can often model data using fewer components than are needed by PLS. This and the possibility of focusing on fewer variables, even forcing more focus by limiting the interval U of available γ values, can add to the interpretability of the models produced, e.g. because of less complicated vectors of loading weights and regression vectors.

CCA

Canonical correlation analysis (CCA) [6] is an extension of ordinary correlation between two vectors to situations with several vectors organised in matrices $\mathbf{X}$ ($n \times p$) and $\mathbf{Y}$ ($n \times q$). The problem is reformulated as finding the vectors $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ maximising the correlation between $\mathbf{Xa}$ and $\mathbf{Yb}$. Assuming centred $\mathbf{X}$ and $\mathbf{Y}$ CCA maximises the equivalent expressions

$$\mathsf{corr}(\mathbf{Xu}, \mathbf{Yv}) = \frac{\mathbf{u}^t\mathbf{X}^t\mathbf{Yv}}{\sqrt{\mathbf{u}^t\mathbf{X}^t\mathbf{Xu}}\sqrt{\mathbf{v}^t\mathbf{Y}^t\mathbf{Yv}}} \quad \text{and} \quad f(\mathbf{r}, \mathbf{t}) = \mathbf{r}^t(\mathbf{X}^t\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-\frac{1}{2}}\mathbf{t}$$

over all possible $\mathbf{u}, \mathbf{r} \in \mathbb{R}^p$ and $\mathbf{v}, \mathbf{t} \in \mathbb{R}^q$. Maximisation of $f(\mathbf{r}, \mathbf{t})$ can be done by SVD on $(\mathbf{X}^t\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-\frac{1}{2}}$, where the dominant left singular vector $\mathbf{r}_0$ corresponds to the dominant eigenvector of the matrix $\mathbf{T}^{-\frac{1}{2}}\mathbf{B}\mathbf{T}^{-\frac{1}{2}}$ ($\mathbf{T} = \mathbf{X}^t\mathbf{X}$ and $\mathbf{B} = \mathbf{X}^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{X}$). By defining $\mathbf{a} = \mathbf{T}^{-\frac{1}{2}}\mathbf{r}_0$ we also achieve that $\mathbf{a}$ becomes a dominant eigenvector of the before mentioned Rayleigh coefficient $\mathbf{T}^{-1}\mathbf{B}$.

Multi-way analysis

In multi-way analysis [7] the data is arranged in cubes or hyper cubes having three or more dimensions (modes). These can for instance be samples, excitations and emissions in fluorescence spectroscopy or simply the factors of a factorial design experiment. The section of a hyper cube relating to one level of one of the modes is called a slice or slab, e.g. all excitations and emissions belonging to a single sample. Before multi-way methods were developed such data would be unfolded into two dimensional matrices before analysis by PCA or related methods. The choice of sequence of unfolding in the different modes means there is no unique way of unfolding, and internal structures between neighbouring points in the hyper cube are broken.

Some of the fundamental tri-linear methods working directly on multi-way arrays are parallel factor analysis (PARAFAC) [8] and the Tucker methods [9]. PARAFAC can be seen as a generalisation of PCA, expressing loadings (and scores) as tensors instead of vectors, still including unique solutions and adding an assortment of possible constraints to the modelling. The multi-way arrays are noted with underscore, meaning a three mode PARAFAC model with R components in tensor product notation and scalar notation looks like:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{a}_r \Delta \mathbf{b}_r \Delta \mathbf{c}_r + \underline{\mathbf{E}} \quad \text{and} \quad x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}, \tag{4}$$

respectively. PARAFAC is often used for identifying elution profiles or pure spectra and concentrations from mixed samples. Since the modelling is based on alternating least squares (ALS), it is quite robust against missing data, especially when these are missing at random or missing in areas non-central to the profiles of interest. The benefit from analysing the data in its original structure without unfolding is so substantial in many applications that its called the three-way advantage.

Generalised multiplicative analysis of variance (GEMANOVA) [10] adds the possibility of constraining PARAFAC by including components containing subsets of the factors (loadings). This is achieved by forcing selected loadings to be vectors of ones in the model building. As the name suggests, GEMANOVA has been proposed as a supplement to ordinary analysis of variance (ANOVA), enabling construction of more complex models, possibly containing missing data and/or data without replicates. The use of several components also increases

the flexibility and scope of the analysis, facilitating modelling of several phenomena simultaneously.

2.4 Validation

Applying prediction models to the data that generated them can yield some strange results. Often it seems that the more complex model one chooses, the better the fit of the data becomes. In practise most models will be used for predicting new data or similar data not included in the model building. To get a realistic impression of the complexity needed and the level of error one can expect some kind of validation is needed. The two main types of validation used in this thesis are cross-validation and test set validation, often used in conjunction. In addition bootstrapping is used where neither cross-validation nor test set validation is applicable.

Cross-validation

Cross-validation [11] is an internal validation for the main data set, i.e. the calibration data. In its basic form the data are split into K segments, where $K \leq n$ (the number of samples), and each segment is kept aside once as a validation set that is predicted by the model generated from the remaining $K-1$ segments. The K segments can consist of consecutive, interleaved or random samples, or another splitting can be done taking into account replicates or other relations between the samples. Typically RMSE for cross-validation (RMSECV) is calculated as in Equation 2. The model complexity is usually chosen by minimising the RMSECV or inspecting plots of RMSECV curves to find where they start flattening. More objective choices of complexity can be done by statistical testing of the difference of fit from the model obtaining minimum RMSECV to more conservative models (with regard to complexity) [5]. Another measure available when cross-validating is the coefficient of determination for prediction [12]:

$$R^2_{pred\,(k)} = \left(1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i),k}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}\right), \tag{5}$$

which gives an impression of the amount of variation in new data one can expect to be explained by the model.

External validation

While cross-validation is mostly used to get an idea of the level of complexity suitable for good prediction of new data, it is also used by many as an estimate of the level of error one can expect on new predictions. A more widely accepted strategy is to have separate validation data, preferably generated independently of the calibration data, for the purpose of estimating the future level of error. In many applications an independent data set is not available so the

main data set is split into calibration data and validation data, typically assigning more observations to the calibration than the validation. When doing this it is extremely important that associated observations are kept together in one of the data sets, e.g. allocating all replicates of a sample to either the calibration data or the validation data. Using calibration data for modelling and prediction of validation data, one can calculate RMSE for prediction (RMSEP) to assess the goodness of fit. When generating separate validation data one should try to make them as realistically different from the calibration data as one would expect future samples to be.

Bootstrapping
Bootstrapping [13] is a computer intensive resampling approach to statistical inference. In its basic form new data sets are sampled with replacement from the calibration data a large number of times. The statistic of interest is calculated based on each new data set resulting in a distribution from which the statistic can be described and uncertainty can be assessed. Among other types of bootstrapping we also find resampling only of residuals, sometimes called model-based bootstrapping. Here a model, e.g. regression model, is fitted once first. Samples of the residuals are added to the fitted values and the model is refitted a large number of times. This can be used to analyse any parameter estimate or statistic regarding the model at hand.

3 Spectrometry

As several of the data sets and applications in this thesis are spectrometric, we include some words about spectrometry and mass spectra in general and say a little about MALDI-TOF, Raman and NIR. Spectroscopy is a term having multiple definitions, but usually describes measurements of light (or energy) as a function of either wavelength or frequency. A spectrum (plural spectra) is a representation of the measurements as a data vector or plot. The term spectrometry is often even more widely defined, incorporating any spectroscopic or other measurement technique resulting in spectra that can be used to describe properties of a sample. The use of light, e.g. visible, x-ray, ultraviolet, infrared or near infrared (NIR), is applied when recording spectra based on absorption, fluorescence, Raman and other interaction phenomena. Other types of spectrometric techniques are nuclear magnetic resonance (NMR) and MALDI-TOF, though these are based on spin based resonance in atoms or ionisation and electrical acceleration of molecules.

MALDI-TOF
As indicated above and by its name matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) [14] is based on ionisation of molecules and calculation of masses based on

their time-of-flight in a closed system. MALDI-TOF is a mass spectrometry technique used mainly for identifying proteins, peptides, sugars and other ionisable compounds. In resent years also quantitative applications have been attempted, though mostly from comparing selected peaks and standards. Samples are either mixed with or covered by a matrix which shields the sample from direct laser, helps ionising it and eases the vapourisation of the sample. An electrical field is applied after ionisation to accelerate the ions before they fly through high vacuum, optionally through a reflector, and to a detector measuring the time-of-flight. Reproducibility can suffer from unevenness in the sample spots and variable laser hits, while mass resolution can sometimes be high enough to detect ratios of isotope variations in molecules.

Raman

Raman spectroscopy [15] is a spectrometric technique used to study vibrational, rotational, and other low-frequency modes in a system. Monochromatic light (laser) interacts with these excitations resulting in energy shifts (Raman shift), by spontaneous Raman scattering, that can be recorded. Other phenomena in the system are typically in orders of magnitudes stronger. Rayleigh scattering (radiation from particles much smaller than the laser wavelength) can usually be filtered out by a monochromator, while fluorescence (triggering of photons of lower energy than those of the laser) often has to be removed by baseline correction of the resulting spectra. Because vibrational information is specific to chemical bonds and the symmetry of molecules, Raman spectroscopy is often used in chemistry/chemometrics. For organic molecules the range of wave numbers used for fingerprinting is usually around 500-2000 cm^{-1}, though recording from 0 to more than 3000 cm^{-1} is common for detecting other excitation phenomena.

NIR

Near infrared spectroscopy is based on molecular overtones and combination vibrations resulting in broad bands and thus complex spectra. A polychromatic near infrared light source is used in combination with a dispersive element allowing for detection of a range of wavelengths, typically around 800 to 2500 nm. Though the sensitivity of NIR is not particularly high, it has the benefit of high penetration in samples, enabling applications where samples need a minimum of preparation and can be analysed non-invasively.

4 Paper summaries

4.1 Paper I – *Powered PLS discriminant analysis*

In chemometrics and multivariate statistics partial least squares (PLS) is considered one of the standard methods for compressing multi-collinear data into fewer variables holding the

key information of the relationship between a matrix of data, $\mathbf{X}$, and a response, $\mathbf{y}$. PLS can be applied to both continuous responses through PLS regression (PLSR) and categorical responses through PLS discriminant analysis (PLS-DA) [3]. In Paper I we adapt the power methodology used in powered PLS (PPLS) [5] to classification to develop powered PLS discriminant analysis (PPLS-DA). The power methodology parametrises the $\mathbf{X}$-$\mathbf{y}$ correlations and $\mathbf{X}$ standard deviations with a power parameter and uses linear optimisation on a given criterion. In PPLS this criterion is $\underset{\gamma}{max}\left(corr\left(\mathbf{X}\mathbf{w}(\gamma),\mathbf{y}\right)\right)$, while PPLS-DA in practise uses canonical correlation (CCA) by $\underset{\gamma}{max}\left(cca\left(\mathbf{X}\mathbf{W}_0(\gamma),\mathbf{Y}\right)\right)$, where $\mathbf{W}_0(\gamma)=\mathbf{S}(\gamma)\mathbf{C}(\gamma)\mathbf{P}$ is a matrix of candidate loading weight vectors associated with each of the classes/groups in the response. Here $\mathbf{W}_0$ is factorised into $\mathbf{X}$-standard deviations, $\mathbf{X}$-$\mathbf{Y}$-correlations and $\mathbf{Y}$-standard deviations and group weights. This is equivalent to maximisation of the Rayleigh quotient associated with Fisher's canonical discriminant analysis (FCDA), $\mathbf{T}^{-1}\mathbf{B}$, meaning PPLS-DA maximises correlation instead of covariance. In contrast to PPLS, where the ordinary PLS solution is achieved when forcing the algorithm to set the power parameter $\gamma=0.5$ for all components, the ordinary PLS-DA solution is not found for any value of γ in PPLS-DA. This is a consequence of changing from maximisation of between group variance ($\mathbf{B}$) to maximisation of the Rayleigh quotient associated with FCDA. The inclusion of powers gives a flexible extension to PLS-DA able to focus on single variables or groups of variables with high correlation to the response or high standard deviation. PPLS-DA also adds the possibility of imposing restrictions on the powers available for optimisation to force attention on certain characteristics. The result is often models needing fewer components than ordinary PLS-DA, having more structured vectors of loading weights and thus giving simpler interpretations of the models.

4.2 Paper II – *Canonical PLS*

In Paper II the framework from PPLS-DA is taken a few steps further – still employing CCA maximisation and the matrix of candidate loading weight vectors. Including power this gives rise to canonical powered PLS (CPPLS). The choice of maximisation criterion has several advantages. Firstly, both continuous and categorical responses can be used, even modelled simultaneously. Secondly, even though $\mathbf{Y}$ (called $\mathbf{Y}_{prim}$ here) can only consist of the responses one wants to focus on, $\mathbf{W}_0(\gamma)$ can include additional responses expanding the space of possible loading weight vectors. These additional responses can be additional measurements, design factors, some kind of summary of the data or other discrete or continuous values that contain information about the data set. Another possibility is to model multi-response data using one response at the time as $\mathbf{Y}_{prim}$ and include the rest of the responses in $\mathbf{W}_0(\gamma)$. Several applications on real data have shown that informative additional responses can stabilise predictions and simplify the corresponding models. Choices of responses and

limitations on the power parameter in CPPLS can produce PLS, PLS-DA (with correlation maximisation), PPLS, PPLS-DA and CPLS (no powers), possibly extending them with additional responses. The CPLS method used with multiple continuous responses is related to PLS2 but can be considered as finding a supervised linear combination of the columns in $\mathbf{W}_0$, instead of unsupervised. Having a more aggressive optimisation criterion (and possible observation weights) CPLS can yield more parsimonious models than PLS2, but will not be suited for $\mathbf{Y}$ matrices with too many columns.

4.3 Paper III – *Testing effects using multi-way analysis*

In Paper III the use of multi-way analysis of experimental designs is combined with bootstrapping to make visual and numerical guides to the consistency of effects. This is proposed as a supplement to ordinary analysis of variance (ANOVA). The procedure consists of applying generalised multiplicative analysis of variance (GEMANOVA) to multi-way arrays containing factorial design responses and applying a non-parametric model-based bootstrap. After fitting a first GEMANOVA, the bootstrap adds random samples of the residuals (with replacement) to the fits to construct new data sets and refits the model to this a large number of times, e.g. 1000 times. The result is estimated distributions of all levels of all effects. These can be used to estimate p-values for the difference of levels or for plotting. Using GEMANOVA instead of ANOVA means more complex interactions can be studied, single samples can be analysed, less parameters need to be estimated and randomly missing data are less harmful. Some of the benefits are demonstrated through application to a $2^4 \cdot 3$ factorial design pot experiment on nitrogen-sulfur ratios in wheat.

4.4 Paper IV – *Quantitative MALDI-TOF; Measurement optimisation*

Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS) is traditionally considered a purely qualitative technique. Its main application has been for detection and identification of proteins, peptides and other ionisable compounds in biological samples. In Paper IV instrumental settings and sampling procedures are optimised for quantitative use of the spectra. This is done through a screening experiment and an optimisation experiment on cow, goat and ewe milk where factorial designs are applied to span relevant level combinations. In this paper the success of the optimisation is assessed by repeatability of spot replicates on the target plate and signal-to-noise ratios. From the experiments we can conclude that the optimisation results in spectra that should be well suited for quantitative analysis. We also note that some peak resolution has been sacrificed to obtain stable repeatability, meaning the resulting spectra will be less suitable for qualitative applications.

4.5 Paper V – *Quantitative MALDI-TOF; Prediction of concentrations*

In Paper V the results from Paper IV are put into practical use on a simplex lattice mixture experiment with cow, goat and ewe milk. Mixture concentrations of each species are used as the response in partial least squares regression (PLSR). Sets of 45 mixtures $\times$ 4 spot replicates are produced twice (calibration and validation) from the same bulk milk. This means the validation data is a technical replicate incorporating some of the variation one can expect from new samples. A mistake with regard to the laser intensity for the validation data was made, resulting in more noise and thus a more challenging validation than intended. Because ewe milk contains twice as much proteins as cow and goat milk, a square root transformation was used to limit the non-linearity in the ewe response. Results from whole spectrum PLSR predictions are compared to three different variable selection strategies, showing how the predictions get more accurate and stable when many variables are included. The main conclusion is that MALDI-TOF MS shows great potential for being used in quantitative analysis.

4.6 Paper VI – *Optimal choice of baseline correction*

In Papers IV and V baseline correction of spectra was carried out with a well known method using default parameter values. The focus in Paper VI is on choosing the optimal baseline correction algorithm and corresponding parameter values for a given statistical analysis. Two data sets are used: Raman spectra on fish oils and the MALDI-TOF spectra on milk from Paper V. Seven baseline correction algorithms are applied with a wide range of parameter values to show some of the potential of the procedure and to stress some of the pitfalls of baseline correction. Root mean square error of prediction is used as an example of a quality measure to optimise on. GEMANOVA is applied to simplify the search for minimum RMSECV for the milk data, since it has three responses. As expected the results indicate that no single algorithm can be proclaimed as the uniformly best baseline correction, and that individual optimisation is most likely required for different combinations of spectrometric methods, statistical analysis, quality measure and sample type.

5 Discussion

5.1 Contribution

In this thesis the main purpose has been to develop multivariate analysis methods, applying them to existing data and to find new applications for existing analysis methods. Partial least squares (PLS) is the main multivariate analysis method used in Papers I, II, V and VI. Even

though PLS, and its multi-response (PLS2) and discrete (PLS-DA) counterparts, have proved over and over their value as simple and effective analysis methods for multicollinear data, there are many situations where more specialised and aggressive methods can contribute both with respect to prediction and interpretation. Through Papers I and II the focus has been on extending the power methodology from powered PLS to classification (PPLS-DA), multiresponse regression and combinations of these (CPPLS) through the use of canonical correlation analysis. Both PPLS-DA and especially CPPLS expand the areas of application for the PLS family through flexible modelling able to focus on areas of special interest, possible individual sample weights and the inclusion of stabilising additional responses.

As pointed out in Section 2.1 design of experiments is a key element when producing data for any type of statistical analysis. Factorial designs are generated and exploited in Papers III and IV, while all papers analyse data generated by factorial designs. The use of factorial designs has several purposes throughout the papers. In Paper I and II a factorial design with centre points is used in one of the data sets to span as much variation as possible (within reasonable boundaries) to see if the oil types used can still be detected by the use of near infrared spectroscopy. The pot experiment in Paper III is a factorial design analysed with multi-way methods in its original $2^4 \times 3$ hyper-cube form. In Paper IV two factorial designs are used to span relevant process and instrumental variables to optimise the generation of MALDI-TOF data for quantitative analysis. A mixture design is used in Paper V to see how predictions of concentrations by PLS regression behave across a relevant sample of mixtures of three milk types.

As Paper III deals with multi-way data, it can be considered an outlier in this thesis. It proposes the use of bootstrapping of multi-way models for assessing the consistency of effects in designed experiments. In many biological analyses too much averaging and simplification can hide interesting phenomena connected to multi-way interactions and variations between samples. Our procedure can be considered a supplement to ANOVA intended for explorative analysis using generalised multiplicative analysis of variance (GEMANOVA) to examine complex interactions and single replicates. The procedure is also used in Paper VI to simplify complex simultaneous optimisations to a series of one dimensional minimisations enabling assessment of parameter resolutions around the minima.

While MALDI-TOF MS is traditionally perceived as a purely qualitative method, we have shown through Papers IV and V that, with a minimum of effort, it has great potential for being applied in quantitative analyses. Using an experimental setup where mixtures of cow, goat and ewe milk are analysed a procedure for optimising experiments for quantitative analysis is suggested. A set of whole spectra from a mixture experiment based on the optimised experimental setup is analysed by use of PLS regression and compared to an assortment of variable selection methods. The simple whole spectrum strategy is clearly preferable to most

variable selection strategies with respect to predictions. To give an overview of the behaviour of the predictions across the different mixtures in the experiment, a novel plotting method is used where predicted mixtures are plotted as a simplex lattice mixture design. This enables a simple comparison to the original design and an easy way of assessing prediction performance in different situations.

5.2 Future perspectives

The power methodology combined with CCA is a promising development that should be explored further and applied in other contexts. Various additional responses, e.g. in the form of data or sample summaries or transformations of the main response, is one area that might hold potential advantages worth investigating. Individual weighting of observations has been incorporated both in the theory and implementation of CPPLS, though it has so far only been employed for weighting groups. Combined with outlier detection individual weights may be used to reduce the influence of observations that inflict unwanted bias to the models. Some observations also contribute more than others in the optimisation of powers. Tests using internal cross-validation and averaging over candidate powers in the CPPLS algorithm indicate that models generated by difficult data can be stabilised and robustified. Formalising and further testing is needed to confirm the results. Different multi-block PLS methods are becoming important for consolidating data from several sources, such as spectrometric instruments. These might benefit from CPPLS's ability to focus a subset of variables and utilising many responses.

During the process of working with the MALDI-TOF MS data on milk we have observed phenomena resembling resonances of peaks and clusters of peaks. These are found where one would expect double and triple charged ions from certain proteins to occur. If these are indeed proteins where some proportion consistently receives extra charges, we might be able to model the phenomenon, either utilising it to improve predictions or simply using it as an aid in describing the uncertainty of peak identifications. Looking at the algorithms used for baseline correction we also see potential for making more adaptable algorithms by combining existing strategies with new ideas conceived during the work with this thesis. Multivariate statistics is like a wasp's nest; the more you poke around in it the more problems you generate.

6 References

[1] D. C. Montgomery, Design and Analysis of Experiments, 6th Edition, John Wiley & Sons, Inc, 2005.

[2] H. Martens, T. Næs, Multivariate calibration, John Wiley and Sons, Chichester, UK, 1989.

[3] H. Nocairi, E. M. Qannari, E. Vigneau, D. Bertrand, Discrimination on latent components with respect to patterns. application to multicollinear data, Computational Statistics & Data Analysis **48**, 139–147 (2005).

[4] U. G. Indahl, H. Martens, T. Næs, From dummy regression to prior probabilities in pls-da, Journal of Chemometrics **21**, 12, 529 – 536 (2007).

[5] U. Indahl, A twist to partial least squares regression, Journal of Chemometrics **19**, 32–44 (2005).

[6] K. Mardia, J. Kent, J. Bibby, Multivariate Analysis, Academic Press, 1979.

[7] A. Smilde, R. Bro, P. Geladi, Multi-way Analysis with Applications in the Chemical Sciences, John Wiley & Sons, Ltd., 2004.

[8] R. Harshman, Foundations of the parafac procedure: Model and conditions for an 'explanatory' multi-mode factor analysis, UCLA Working Papers in phonetics **16**, 1 (1970).

[9] L. R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika **31**, 3, 279–311 (1966).

[10] R. Bro, M. Jakobsen, Exploring complex interactions in designed data using gemanova. color changes in fresh beef during storage, Journal of Chemometrics **16**, 294–304 (2002).

[11] M. Stone, Cross-validatory choice and assesment of statistical predictions, Journal of the Royal Statistical Society, Series B—Methodological **36**, 111–147 (1974).

[12] D. C. Montgomery, E. A. Peck, G. G. Vining, Introduction to Linear Regression Analysis, Wiley Interscience Publication, 2001.

[13] A. Davison, D. Hinkley, Bootstrap Methods and their Application, Cambridge University Press, 1997.

[14] M. Karas, D. Bachmann, F. Hillenkamp, Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules, Analalytical Chemistry **57**, 2935 – 2939 (1985).

[15] D. Gardiner, Practical Raman spectroscopy, Springer-Verlag, 1989.