ISBN 978-82-575-1057-2
ISSN 1503-1667

Nofima
Osloveien 1, NO-1430 Ås, Norway
Phone: +47 64 97 01 00
www.nofima.no, e-mail: post@nofima.no

Norwegian University of Life Sciences
NO–1432 Ås, Norway
Phone +47 64 96 50 00
www.umb.no, e-mail: postmottak@umb.no

Sahar Hassani

Philosophiae Doctor (PhD) Thesis 2012:20

Norwegian University of Life Sciences • Universitetet for miljø- og biovitenskap
Department of Mathematical Sciences and Technology
Philosophiae Doctor (PhD) Thesis 2012:20

Integration of multivariate data in systems biology

Integrering av multivariate data i systembiologi

Sahar Hassani

Integration of Multivariate Data in Systems Biology

Integrering av Multivariate Data i Systembiologi

Philosophiae Doctor (PhD) Thesis

Sahar Hassani

Department of Mathematical Sciences and Technology
Norwegian University of Life Sciences

Ås 2012

Thesis number 2012:20
ISSN 1503-1667
ISBN 978-82-575-1057-2

To my husband
Shahriar

Table of contents

Acknowledgements

The work described in this thesis was carried out from June 2008 to June 2012 mainly at Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research. Financial support was provided by the Nordic Centre of Excellence on Food, Nutrition and Health “Systems biology in controlled dietary interventions and cohort studies” (SYSDIET) funded by NordForsk, and by Foundation for Research Levy on Agricultural Products in Norway.

My Ph.D. project was extremely exciting and I have learned a great deal from many kind people around me during the past four years. It would not have been possible to write this doctoral thesis without their help and support, to only some of whom it is possible to give particular mention here.

First and foremost I want to thank my supervisor **Achim Kohler** for all your contributions of time and ideas as well as your tremendous support and skilled supervision. Your patience, insightful comments, constructive feedback, enthusiastic encouragement and academic guidance were invaluable. You were always there to provide necessary assistance and advice. It was an honor to be your Ph.D. student.

I would like to thank my co-supervisor **Grethe Iren Borge** for all your support and your help with the biological concepts. I always felt comfortable asking you for help and advice.

I am especially grateful to my co-supervisor **Harald Martens** for your professional expertise, encouraging discussions and for always being friendly and inspiring. It was a privilege to work with you and learn from you.

My sincere thanks go to **El Mostafa Qannari** and **Mohamed Hanafi** for your helpful, successful and friendly collaborations. It has been a real pleasure working with you!

I would like to thank all the colleagues at the departments of ‘Raw materials and Process Optimisation’ and ‘Food and Health’. Special thanks to the department directors **Ragni Ofstad** and **Kristine Naterstad**. You all made this place a great place to work!

To my Ph.D. colleagues, thank you all for the friendly and inspiring atmosphere you created. Special thanks to **Olga Shapaval** for being there always, in good as well as in difficult times. You are the best friend and office mate ever! Thanks to **Ibrahim Karaman** as well for all your support especially your help with my Matlab algorithms. I would not be able to pass the variable selection course without your collaboration! I would also like to thank **Aida Eslami** and **Karen Wahlstrøm Sanden** for your great helps with the final steps of my thesis submission.

To my ever loving parents **Marzieh** and **Ramezan**, thank you for believing in me and for giving me love and confidence. I would not be where I am today without your help and support. Thank you so much for everything you have done for me over the past 30 years. To **Ali**, my dearest brother, thank you for your care, support, understanding and being my favorite brother.

Last, but most important, I want to thank my husband **Shahriar**. No words can express how much you mean to me and how important you have been for my Ph.D. study. Thank you for your amazing and encouraging support and for sharing your life with me.

Ås, June 2012

Sahar Hassani

Abstract

Owing to the rapid rate of development in the field of systems biology researchers have faced many new challenges with regard to handling the large amount of generated data sets originating from different –omics techniques, integrating and analyzing them and finally interpreting the results in a meaningful way. Different statistical methods have been implemented in the field of systems biology. The use of chemometrics approaches for the integration and analysis of systems biology data has recently increased. Different chemometrics methods are potentially available for integrating –omics data and detecting variable and sample patterns. An important challenge is to decide which method to use for the analysis of –omics data sets and how to pre-process the data sets for this purpose. Special attention needs to be given to the validity of the detected patterns.

In this study we have been working on developing multi-block methods for integrating different types of systems biology data and investigating the co-variation patterns among the measured variables. A special focus was given to the validation of the results of the multi-block methods CPCA and MBPLSR. Different types of graphical tools were introduced for the purpose of validation. We have also developed pre-processing techniques that could explicitly be used for lipidomics data sets. A framework was built for pre-processing, integrating, analyzing and interpreting the lipidomics data sets. The framework was then used for the analysis of a lipidomics data set from a human intervention study.

Working on the development of the validation tools required an understanding of the concept of DFs consumption during the multi-block modeling. Therefore, we ran simulation studies where we investigated the number of DFs that were consumed during the modeling processes of PCA and CPCA. Another important issue for applying multi-block methods is the choice of the deflation method. Hence, we studied different deflation strategies available for Multi-block PCA and investigated their interpretational aspects.

Norsk sammendrag

På grunn av rask utvikling innen systembiologi har forskere møtt mange nye utfordringer med hensyn til håndtering av store datamengder, som genereres med forskjellige -omics teknikker. Det er en stor utfordring både å integrere, analysere og til slutt tolke resultatene på en meningsfull måte. Ulike statistiske metoder har blitt implementert for analyse av systembiologi data. Bruk av kjemometri for integrering og analyse av biologiske data har økt mye den siste tiden. I utgangspunktet finnes det flere metoder fra kjemometri som kan brukes for å integrere data fra forskjellige –omics teknikker og for å oppdage grupperinger av objekter og variabler. En stor utfordring er å bestemme hvilken metode som skal brukes til analyse av -omics datasett og hvordan pre-prosessere datasettene. Det er også viktig å validere de grupperingene som har blitt oppdaget.

I denne studien har vi jobbet med å utvikle multiblokk metoder for å integrere ulike typer data fra systembiologi og å undersøke samvariasjon blant de målte variablene. Det har spesielt vært fokus på validering av resultatene av multiblokkmetoder som CPCA og MBPLSR. Ulike typer verktøy ble innført for å sikre valideringen. Vi har utviklet pre-prosessering teknikker som kan brukes spesielt til lipidomics datasett. Vi har bygget et rammeverk for pre-prosessering, integrering, analysering og tolkning av lipidomics datasett. Metoden er blitt brukt til å analysere et lipidomics datasett fra et human intervensjonsstudie.

Utvikling av validerings metoder krever en forståelse av bruk av antall frihetsgrader under modelleringen. Det har derfor blitt gjennomført simuleringsstudier hvor vi undersøkte antallet frihetsgrader som ble brukt under modellering med PCA og CPCA. Et annet viktig tema når man bruker multiblokk metoder er valget av deflasjonsmetoden. Det er blitt studert ulike deflasjonsstrategier som er tilgjengelige for multiblokk PCA og undersøkt deres tolkningsaspekter.

List of papers

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals. The papers are appended at the end of the thesis.

I. Hassani, S., Martens, H., Qannari, E.M., Hanafi, M., Borge, G.I., and Kohler, A. (2010). Analysis of –omics data: Graphical interpretation- and validation tools in multi-block methods. Chemometrics and Intelligent Laboratory Systems 104, 140-153.

II. Hassani, S., Martens, H., Qannari, E.M., Hanafi, M., and Kohler, A. (2011). Model validation and error estimation in multi-block partial least squares regression. Chemometrics and Intelligent Laboratory Systems. (*In press*)

III. Hassani, S., Martens, H., Qannari, E.M., and Kohler, A. Degrees of freedom estimation in Principal Component Analysis and Consensus Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems. (*Under revision*)

IV. Hassani, S., Hanafi, M., Qannari, E.M., and Kohler, A. Deflation strategies for multi-block principal component analysis revisited. Chemometrics and Intelligent Laboratory Systems. (*Under revision*)

V. Hassani, S., Martens, H., Ottestad, I., Borge, G.I., Myhrstad, M.C. and Kohler, A. Simultaneous analysis of inter- and intra-class lipid changes in lipidomics studies. (*Manuscript*)

VI. Ottestad, I., Hassani, S., Borge, G.I., Kohler, A., Gjermund, V., Hyötyläinen, T., Oresic, M., Brønner, K.W., Holven, K.B., Ulven, S.M., and Myhrstad, M.C. Fish Oil Supplementation Alters the Plasma Lipidomic Profile and Increases Long-Chain PUFAs of Phospholipids and Triglycerides in Healthy Subjects. PLoSOne. (*Under revision*)

1. Background

Integrating Systems Biology data

Systems biology is a multidisciplinary emergent field that employs several high-throughput techniques to study interactions between different components of a biological system [1]. Obtaining data along the casual chain from genotype to phenotype enables studying the samples at different levels from DNA to phenotype. A variety of –omics techniques are nowadays becoming available in the field of systems biology e.g. genomics, proteomics, metabolomics and lipidomics which is a branch of metabolomics (Fig. 1). Understanding a biological system as a whole requires integration and simultaneous analysis of such –omics data sets [2]. As it can be seen in Fig.1, different types of techniques are generally used for generating –omics data sets. Collecting data from each technique in a separate data matrix, results in multi-block multivariate data set containing different types of measurements belonging to the same samples. Samples are ordered in the same way in each data set leading to a row to row correspondence between the blocks of the multi-block data set. An example of a multi-block data set from Systems Biology is shown in Fig. 2a where different –omics techniques are applied on the same samples. As it can be seen in Fig. 2a, different blocks of a multi-block data set always contain the same sample set while they contain different variable sets. The measurement of the same samples by different –omics techniques raises the challenge of building a multi-block framework for integrating and analyzing such generated multi-block data sets.

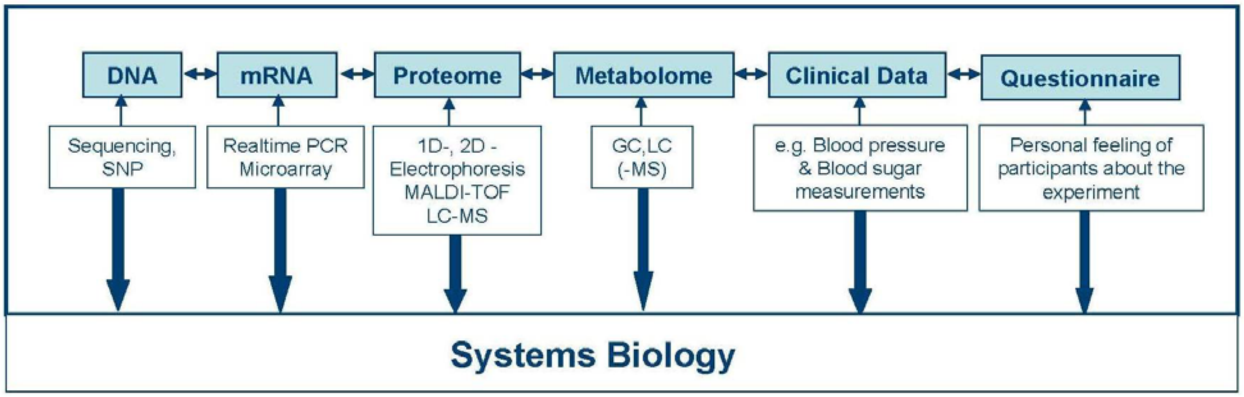

DNA
mRNA
Proteome
Metabolome
Clinical Data
Questionnaire
Sequencing, SNP
Realtime PCR Microarray
1D-, 2D - Electrophoresis MALDI-TOF LC-MS
GC,LC (-MS)
e.g. Blood pressure & Blood sugar measurements
Personal feeling of participants about the experiment
Systems Biology

Figure 1: Integration of data in Systems Biology along the casual chain from genotype to phenotype. The figure is adapted from [3].

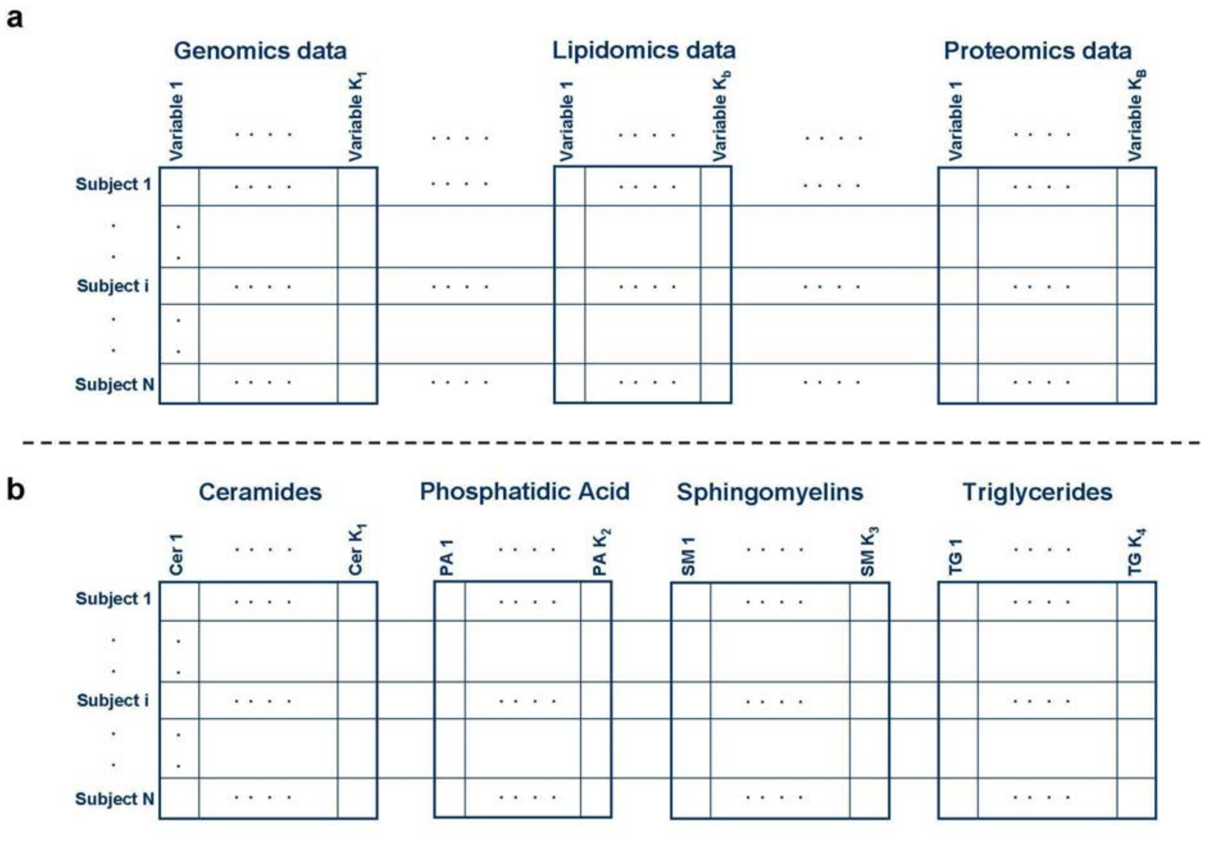


Figure 2: (a) Structure of an example multi-block data set from Systems Biology. (b) Structure of an example multi-block lipidomics data set containing four lipid classes.

Lipidomics, a branch of metabolomics, is the study of the cellular lipidome, involving detection, characterization and quantitative analysis of hundreds up to thousands of lipids (i.e. fatty molecules) using mass spectrometry instruments with high sensitivity and high specificity (mass resolution) [4]. Lipids are classified into several lipid classes and sub-classes. Such a classification of lipids results in a multi-block situation for lipidmics data sets when the data from different lipid classes are gathered in different data blocks. An example of a multi-block lipidomics data set is shown in Fig. 2b for a four-block lipidomics data set containing the following lipid classes: *Ceramides* (Cer), *Phosphatidic Acid* (PA), *Sphingomyelins* (SM) and *Triglycerides* (TG). An important challenge for analyzing such lipidomics data sets is integrating lipids from different lipid classes and analyzing them simultaneously in order to explore the lipid-lipid relationships as well as the dynamics between different lipid classes. There is also a need for integrating lipidomics data together with other types of data and studying the co-variation patterns among the lipids and co-variation patterns among lipids and other variables.

Exploratory chemometrics approaches, such as Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR), are nowadays being employed for the analysis of –omics data sets. PCA is an unsupervised chemometrics approach that is used for the purpose of modeling one-block data sets. The application of PCA for the analysis of different types of data from Systems Biology has recently increased e.g. analysis of metabolomics data [5-8], proteomics data [9-11], genomics data [12-14] and lipidomics data [15, 16]. PCA reveals the co-variation patterns among the samples and variables of a one-block data set. PLSR is a different supervised exploratory chemometrics approach that is used for modeling two-block data sets. PLSR is commonly applied for the analysis of Systems Biology data e.g. in the analysis of metabolomics studies [17-19], in proteomics studies [20, 21], in genomics studies [22, 23] and in lipidomics studies [24, 25]. PLSR is a subspace regression method that reveals the co-variation pattern between the samples and variables of a two-block data set by maximizing the covariance between the variables of two data blocks.

Due to the fact that different types of –omics data sets are generated by the same experiment in Systems Biology, there is a growing need for data analysis methods that can be used for integrating and analyzing such multi-block data sets. Consensus PCA (CPCA) [26] and Multi-block PLSR (MBPLSR) [27] are two exploratory chemometrics approaches that are capable of modeling multi-block data sets. These methods, which are based on latent variables, aim at detecting a common underlying pattern between different data matrices and revealing the contribution of every individual block to the detected pattern. CPCA and MBPLSR can therefore be adapted for the integration of multi-block –omics data sets such as lipidomics data sets. However, the application of these multi-block techniques within the Systems Biology field is at its early immature stage and only few systems biology studies have reported the use of these multi-block methods [28, 29][Paper VI].

CPCA and MBPLSR have excellent graphical visualization possibilities and therefore overview about sample and variable variation patterns can be easily gained. Global score plots of CPCA and MBPLSR illustrate the global sample patterns shared between the different blocks of a multi-block data set while block score plots show the sample patterns within every block of the multi-block data. Correlation loading plots illustrate the variable variation patterns among the variables within and between different data blocks.

Before data from different –omics data sets can be integrated, the scientists are faced with the challenge of pre-processing of –omics data sets. The pre-processing of instrumental and experimental effects contained in raw data as obtained from the –omics techniques as for example shift alignments of chromatography data is a wide field and beyond the scope of the thesis. Still, after the pre-processing of instrumental and experimental effects other pre-processing steps are necessary which have a direct effect on the integration of the data blocks in the multi-block model. An important issue to be solved here is the grouping of the variables into different data blocks in a logical way, which is related to the biological problem to be solved. Another issue for pre-processing procedure of such multi-block data sets prior to integration is the weighting of different data blocks. This is an important aspect of multi-block data analysis since it provides the researcher with the possibility of a simultaneous analysis of data blocks that may contain very different number of variables as well as very different data units.

After the pre-processing of the data blocks, data blocks may be integrated by CPCA and MBPLSR. In recent years different variants of CPCA and MBPLSR have been discussed in literature. All of them are based on the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm for CPCA and MBPLSR, but they differ in the deflation procedures employed [30, 31]. The different deflation strategies lead in general to different sample and variable variation patterns. Mathematical aspects of these deflation procedures have been discussed in literature, but it is not clear how the deflation procedure relates to the interpretation of sample and variable variation patterns. Therefore, understanding the different results obtained for the different deflation procedures needs further investigation.

As it was mentioned before, global and block score plots visualize the sample patterns between and within the blocks of a multi-block data set. However, the question of "how strong a detected pattern is" remains still unanswered. It is also hard to guess the importance of the different data blocks for the detected global pattern just by studying the score plots visually. Therefore, there is a need for statistical methods that can detect the blocks that are significantly contributing to the detected patterns and can give us a measure for the amount of such contribution.

It was explained before that lipidomics data sets are multi-block data sets due to the possibility of dividing them into lipid classes. In the analysis of lipidomics data sets scientists are mostly interested in knowing if lipid profiles are significantly changed by a design parameter or not. If lipid profiles are significantly changing, then they investigate lipid classes in order to see if so-called remodeling effects appear, i.e. if lipid metabolism is going on within specific lipid classes, or if lipid metabolisms are going on between the different lipid classes. Techniques that could visualize such remodeling effects are therefore needed.

Similar to sample variation patterns, variable variation patterns also need to be validated in order to provide the user with the possibility of detecting the significant variables and knowing the amount of their contribution to the detected patterns. The large number of variables makes the use of univariate analysis methods complicated for –omics data and leads to the multiple testing problem: when t-tests are performed on hundreds or thousands of variables the chance for false discovery is high. Penalizing p-values for multiple testing leads in average to high p-values and therefore to many false-negatives. However, such statistical tests are still the most common variable selection method employed for the analysis of –omics data sets.

Cross-validation has been frequently used for validating the results of PCA and PLSR [32-34]. Cross-validation investigates the reproducibility of the results by dividing the samples into calibration sets and test sets. The calibration models are built based on the calibration samples and then are implemented on the test samples in order to check their predictability on new samples. Validation tools for the extension of these methods (i.e. PCA and PLSR) to the multi-block situations (i.e. CPCA and MBPLSR) could be developed by extending these ideas to a multi-block situation. Special attention needs to be given to the concept of Degree of Freedom (DF) when dealing with the development of such validation tools for calculating the Mean Squared Errors of cross-validated models.

2. Aim of this study

The main purpose of this study was to develop methods that can be used for integrating and analyzing different types of multivariate data sets that are generated in the field of systems biology with a special focus on data from lipidomics. The study had the following sub-goals:

1. To adapt CPCA and MBPLSR for integrating data from lipidomics.
2. To unify the different CPCA and MBPLSR methods with respect to the different deflation strategies existed. For this purpose the interpretational aspects of the different deflation strategies needed to be investigated.
3. To develop validation tools for CPCA and MBPLSR. For this purpose the degrees of freedom consumed during validation needed to be investigated.

3. Methods

3.1 Pre-processing

Data pre-processing involves different types of treatments applied on the raw data in order to make it ready for being analyzed [35]. Raw data should be prepared prior to the data analysis procedures since it is usually harder and less efficient to analyze the raw data directly. A simultaneous analysis of data from different sources requires appropriate pre-processing methods for integrating them into the same data framework. Data pre-processing depends both on the data type and the analysis method that is being used. Three different methods for pre-processing of data are described in the following sections: 1) Mean-centering the variables, 2) scaling the blocks (two pre-processing procedures that are commonly used prior to PCA/CPCA and PLSR/MBPLSR) and 3) a special pre-processing technique for lipidomics data.

3.1.1 Mean-centering

The data is usually mean-centered prior to PCA/CPCA and PLSR/MBPLSR by subtracting the mean of the variables according to:

$$\begin{aligned}\mathbf{X}_{\text{Mean-centred}} &= \mathbf{X}_{\text{Raw}} - \mathbf{1}\cdot\bar{\mathbf{x}}'_{\text{Raw}}\\ \mathbf{Y}_{\text{Mean-centred}} &= \mathbf{Y}_{\text{Raw}} - \mathbf{1}\cdot\bar{\mathbf{y}}'_{\text{Raw}}\end{aligned} \qquad (1)$$

where, $\mathbf{X}_{\text{Mean-centred}}$ and $\mathbf{Y}_{\text{Mean-centred}}$ are mean-centered data, $\mathbf{X}_{\text{Raw}}$ and $\mathbf{Y}_{\text{Raw}}$ are the original data sets, $\mathbf{1}$ is an $N \times 1$ vector of 1s, $\bar{\mathbf{x}}_{\text{Raw}}$ and $\bar{\mathbf{y}}_{\text{Raw}}$ are vectors of sizes $K \times 1$ and $J \times 1$ respectively which contain the mean values of the variables of $\mathbf{X}_{\text{Raw}}$ and $\mathbf{Y}_{\text{Raw}}$.

The effect of mean-centering on the results of PCA analysis is illustrated in Fig 3 using a data set from spectroscopy which contains 88 samples and 498 variables. Fig. 3a shows score plot when the data is mean-centered prior to the analysis. The grouping pattern in the data is clearly detected by the first Principal Component (PC) which is describing 62.5% of the variation in the data set. Fig. 3b illustrates score plot for the same data when the data set is not mean-centered prior to the analysis. The grouping pattern in the data set is not detected by the first PC anymore. The first PC describes 96% of the variation while less than 3% of the variation (due to the grouping pattern) is described by the second PC. Therefore, one can see that it is crucial to mean-center data prior to running PCA especially since the important issue for finding PCs is the relative variation among the samples and not the absolute values.

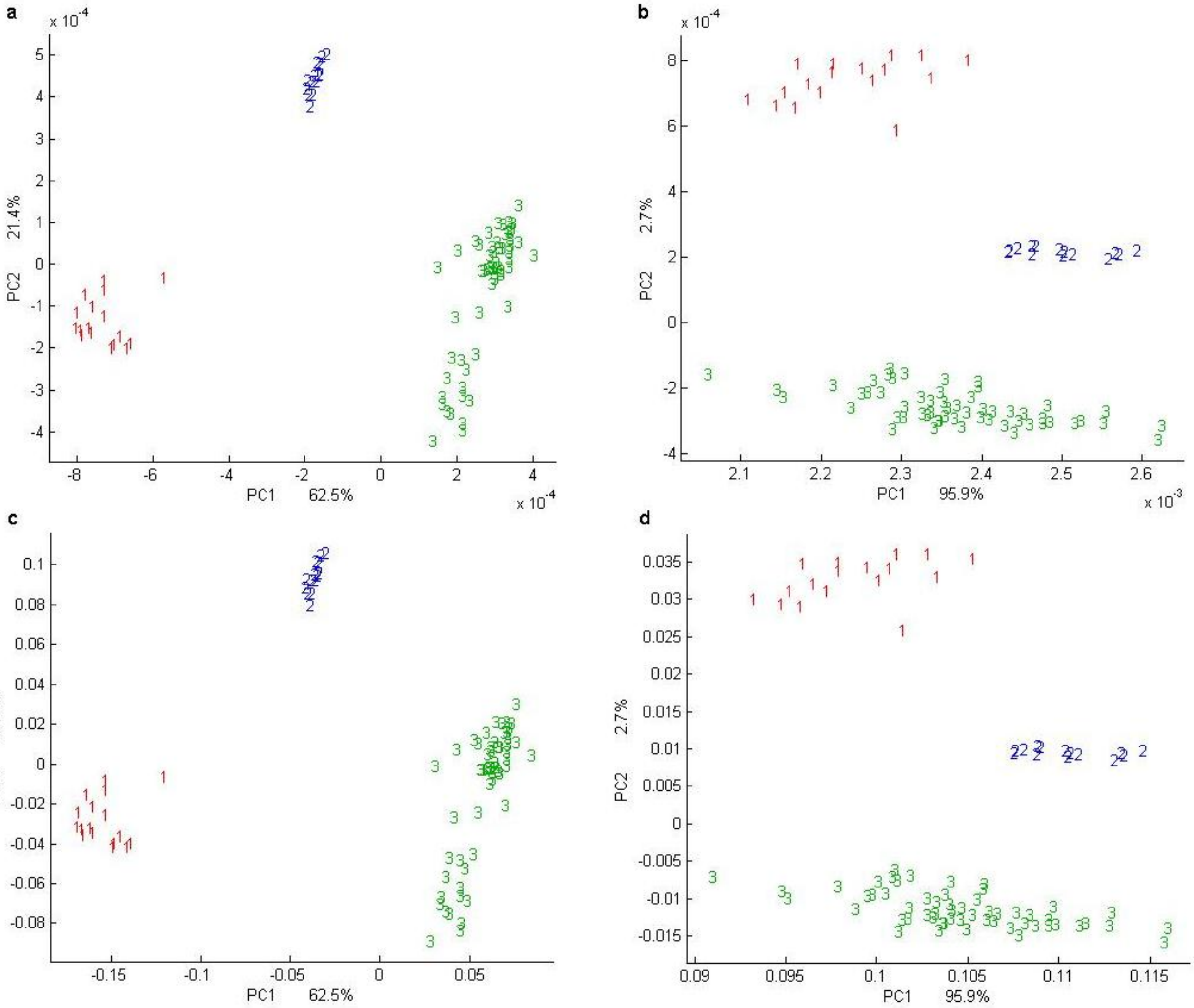


Figure 3: Score plots for the PCA analysis of data. (a) The data is mean-centered. (b) The data is not mean-centered. (c) The data is scaled and mean-centered. (d) The data is scaled but not mean-centered.

3.1.2 Scaling

Since the different data blocks in a multi-block data set are generally coming from different sources, they may have very different number of variables or their magnitudes may vary significantly from each other. In order to put all blocks of data on the same footing prior to CPCA and MBPLSR, the data blocks can be scaled by dividing the mean-centered data blocks by their Frobenius norm as in Eq. 2:

$$\mathbf{X}^b = \frac{\mathbf{X}^b_{\text{Mean-centred}}}{\sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K_b}(\mathbf{X}^b_{\text{Mean-centred}}(i,k))^2}} \quad (a)$$

$$\mathbf{Y} = \frac{\mathbf{Y}_{\text{Mean-centred}}}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{J}(\mathbf{Y}_{\text{Mean-centred}}(i,j))^2}} \quad (b) \qquad (2)$$

where $\mathbf{X}^b$ and $\mathbf{Y}$ are mean-centered and scaled data, $\mathbf{X}^b_{\text{Mean-centred}}$ and $\mathbf{Y}_{\text{Mean-centred}}$ are mean-centered data calculated by Eq.1, $\mathbf{X}^b_{\text{Mean-centred}}(i,k)$ and $\mathbf{Y}_{\text{Mean-centred}}(i,j)$ are the $(i,k)th$ entry and $(i,j)th$ entry of $\mathbf{X}^b_{\text{Mean-centred}}$ and $\mathbf{Y}_{\text{Mean-centred}}$ respectively. We denote the samples by $i = 1,...,N$, the variables in $\mathbf{X}^b$ by $k = 1,...,K_b$, the variables in $\mathbf{Y}$ by $j = 1,...,J$ and the data blocks of a multi-block data set by $b = 1,2,...,B$.

If the data set contains only one data block, then scaling the variables, using the same scale factor for all of them, does not affect the PCA results. The corresponding score plots for Fig. 3a-b are shown in Fig. 3c-d where the data is scaled according to Eq. 2a for $B = 1$. One can see that the patterns in the score plots are exactly the same as before while the scales of the axes are different. Since the score plots are tools that are used for visual identification of patterns, the scales of the axes do not have any interpretational influence on the outcome.

Scaling plays a critical role in multi-block situations (such as CPCA and MBPLSR) since different blocks of data are often scaled by different scale factors. Global score plot and block score plots for running CPCA on a two-block data set are shown in Fig. 4. The same data block that was used in Fig. 3 is the first block here. The second block is a different spectroscopic data block for the same 88 samples. Both blocks have 498 variables. The global score plot and block score plots when the data are mean-centered but not scaled are shown in Fig. 4c and Fig. 4a-b, respectively. The global score plot (Fig. 4c) is showing precisely the same pattern as the one in the second block (Fig. 4b). As it can be seen on the axes for the block score plots, the magnitudes of the axes in two blocks are different. Since the second block has extremely large numbers compared to the first block, this block is strongly affecting the global underlying pattern. In fact, the pattern in the first block is completely removed due to having small influence on the global model. Scaling the blocks according to Eq. 2a let all blocks contribute to the global model equally regardless of their magnitudes. The corresponding score plots for the scaled data are illustrated in Fig. 4d-f. Scales of the axes in Fig. 4d-f indicate that the variables from the different blocks are on the same footing. The global pattern in Fig. 4f does not anymore belong only to one of the blocks. A mixture of patterns from both blocks is seen in Fig. 4f. This is what one wishes to detect in most of the situations when running a multi-block analysis. However, there are certain instances when one would like to force a data block dictate its pattern to the global pattern or prevent a data block from influencing the global pattern e.g. when dealing with design data blocks. Such situations can be handled by over-weighting or under-weighting the corresponding data block by scaling with a very large or a very small number instead of its norm.

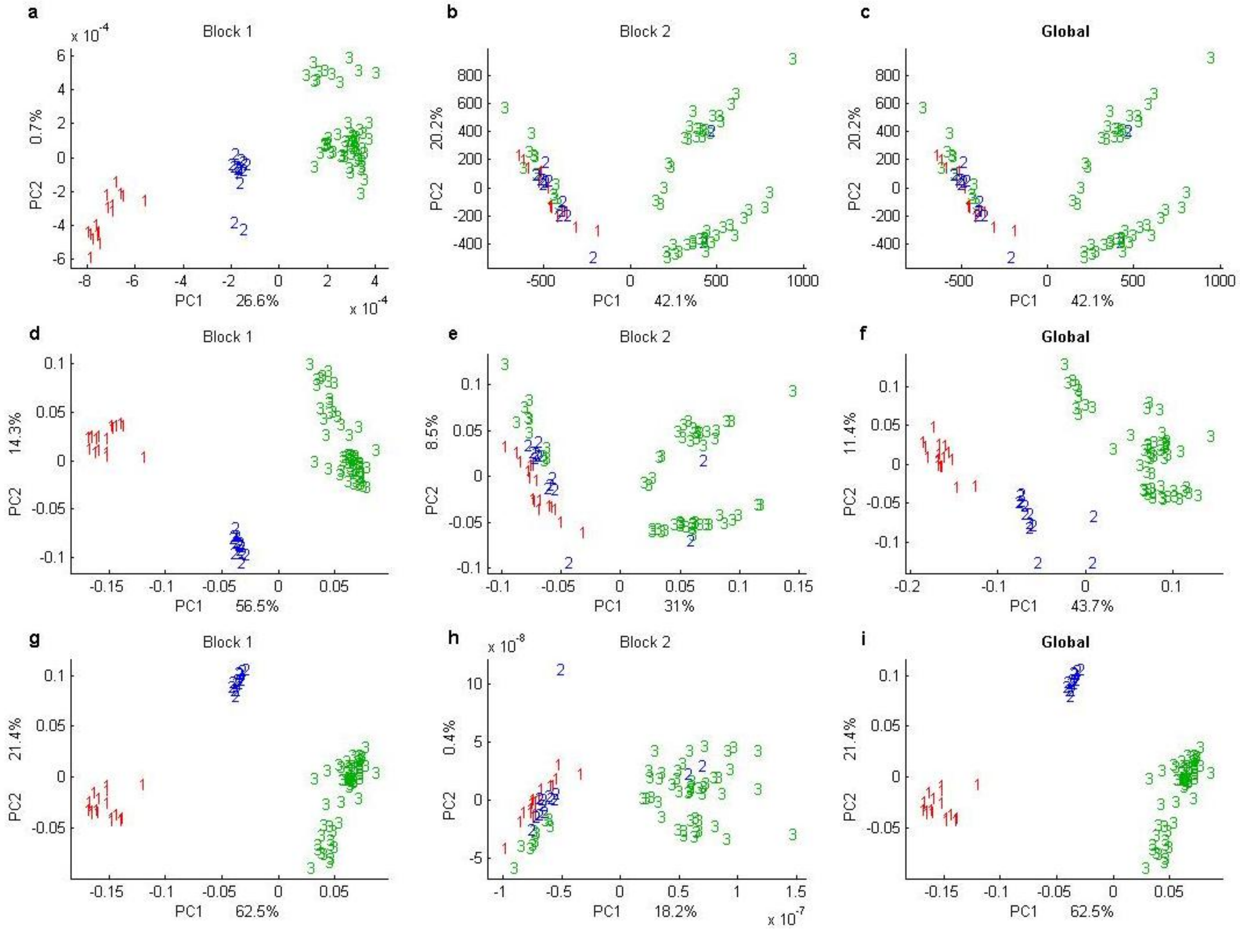


Figure 4: Score plots for the CPCA analysis of a multi-block data set. (a-c) The data blocks are not scaled. (d-f) The data blocks are scaled and are therefore on the same footing. (g-i) The data blocks are scaled first and the second data block is down-weighted afterwards.

An example for under-weighting a data block is seen in Fig. 4g-i where the second block is scaled by multiplying by 0.000001. One can see that the global pattern is dominated by the first block (Fig. 4h).

3.1.3 Pre-processing of lipidomics data

A wide variety of lipids exist from simple fatty acids to complex glycolipids (i.e. lipids with a carbohydrate attached). Lipids are categorized into eight major classes: *Fatty acids*, *Glycerolipids*, *Glycerophospholipids*, *Sphingolipids*, *Sterols*, *Prenol lipids*, *Saccharolipids* and *Polyketides* [36, 37]. This classification enables us to split the data into several data blocks according to different lipid classes. The blocking procedure is subjective toward the aim of the study and the detected lipids, therefore it is certainly possible to use any other classification of lipidomics data for the purpose of blocking (e.g. the following classes/sub-classes of lipids: *Ceramides*, *Lysophosphatidylcholines*, *Lysophosphatidylethanolamines*, *Phosphatidic Acid*, *Phosphatidylcholines*, *Phosphatidylethanolamines*, *Phosphatidylglycerols*, *Phosphatidylserines*, *Sphingomyelins* and *Triglycerides* [Paper VI]). The original lipidomics data table ($\mathbf{X}$ of size $N \times K$) is consequently split into B blocks of data for B different lipid classes ($\mathbf{X} = \left[\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^B\right]$ where $\mathbf{X}^b$ is of size $N \times K_b$ for $b = 1, 2, ..., B$). In order to get insights into different lipid species belonging to a lipid class, the data should be pre-processed within each lipid class. For this purpose the original amounts of the lipids are replaced by their relative variations within their corresponding lipid classes. This task is performed by dividing the raw data for every lipid class by the total amount of lipids in that class (i.e. sum of the data in each block). Restoring the sum values for the lipid classes in a separate data block enables a simultaneous analysis of the lipid species and lipid classes. The same pre-processing approach can be applied when dealing with a single data block (i.e. when the data set contains all lipids without classification). In that case the total amount can be added as an extra variable to the data table instead of an extra data block.

Fig. 5 shows the correlation loading plot for PCA of *Phosphatidylethanolamines* (PE) data. Fig. 5a illustrates the results when data is not pre-processed, and Fig. 5b shows results when data is pre-processed as described above (the data is mean-centered and scaled in both cases prior to PCA). The variables in Fig. 5a are all located on the lower part of the plot which gives the impression that all of the variables are influencing the second PC in the same way. One can also see that many of the variables are explaining more than 50% variance in the data set and are highly positively correlated indicating that they all increase and decrease together. This is expected in most of the situations when dealing with one specific class of lipids, since the lipid species within the same class may often increase and decrease simultaneously. However, this is not what the analysis is mainly seeking. The relative variation of the lipids within the class is an important issue for the purpose of the data analysis. Fig. 5b shows correlation loading plot for the same data set when the data is pre-processed according to the procedure described earlier in this section. One can see that the lipids are now distributed in the whole plot and are not gathered in only one spot which enables detecting the lipid species whose changes are significant relative to the other lipids in the same class.

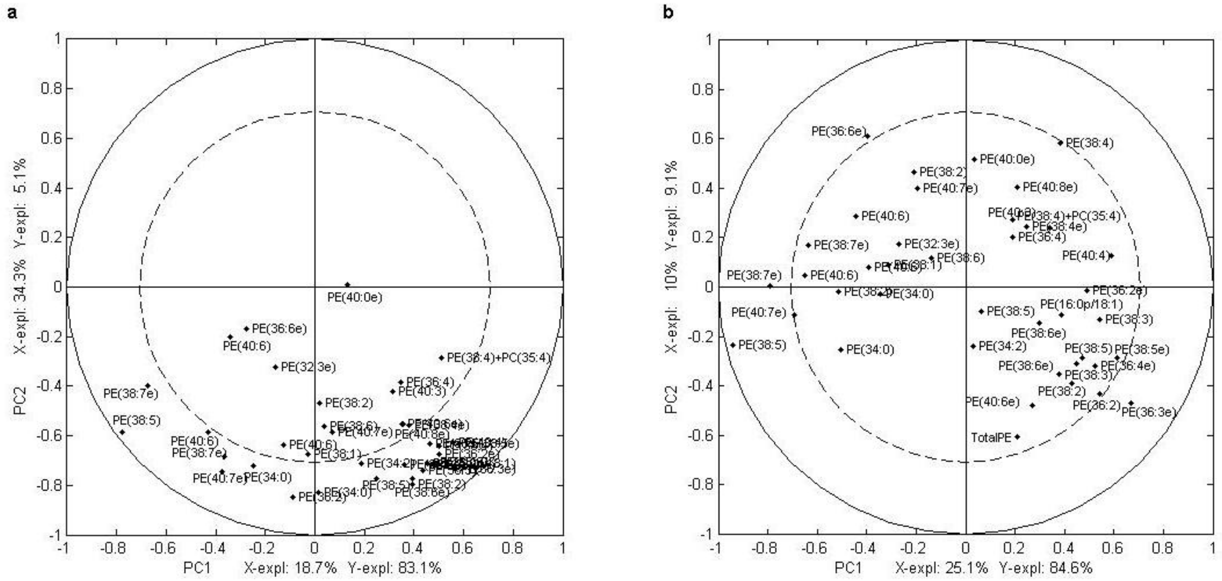


Figure 5: Correlation loading plot for the PCA analysis of a lipidomic data set. (a) The data is not pre-processed properly. (b) The data is pre-processed.

An important application of pre-processing the lipidomics data according to the proposed procedure is the ability for detecting any remodelling of the lipids within the lipid classes. Remodelling of the lipids occurs when a lipid is transformed into another lipid within the same lipid class. Replacing the original amounts of the lipids by their relative amounts enables investigating the variation of lipid species within the lipid classes and studying any increase or decrease of the lipid species with respect to the other lipids in the same class as well as a simultaneous analysis of the total amount of the lipids in each class.

3.2 Consensus PCA (CPCA) and Multi-block PLSR (MBPLSR)

Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR) have been employed for analyzing different types of systems biology data for over a decade now [13, 20, 38-40]. The integration of data from different techniques in systems biology has been in focus of several studies recently [41-44]. CPCA and MBPLSR which are extensions of PCA and PLSR to multi-block data sets can be employed for such integration of systems biology data [3, 28, 45][Paper I, Paper VI]. CPCA aims at finding a common underlying pattern among the data blocks ($\mathbf{X} = \left[\mathbf{X}^1,...,\mathbf{X}^b,...,\mathbf{X}^B\right]$) and studying individual block's contribution to the global pattern while MBPLSR tries to find the common underlying pattern among the descriptor data blocks ($\mathbf{X} = \left[\mathbf{X}^1,...,\mathbf{X}^b,...,\mathbf{X}^B\right]$) that can explain most of the variations in the response data ($\mathbf{Y}$).

A powerful visualization tool (so-called global score plot) is available for CPCA and MBPLSR that provides the users with an overview over the sample variation pattern that is shared by all the data blocks. The global score plot reveals the grouping pattern of samples with respect to the data from all of the data blocks. To what extent every block is contributing to the detected global underlying pattern may be studied by so-called block score plots. The block score plots provide an insight into different data blocks by visualizing the variation patterns that are detected by the respective data blocks. The contributions of the variables to the detected patterns can then be studied by correlation loading plots. The correlation loading plot reveals the relationships among the variables under investigation. Score plots, loading plots and correlation loading plots are described in more details in Sections 2.3.3–2.3.5. CPCA and MBPLSR algorithms are described in the following sections.

3.2.1 CPCA parameter calculation

Applying CPCA on the mean-centered and scaled multi-block data set $\mathbf{X} = \left[\mathbf{X}^1,...,\mathbf{X}^b,...,\mathbf{X}^B\right]$, models the data as sum of *A* relevant Principal Components (PCs) plus a residual matrix. The global CPCA model is given in Eq. 3:

$$\mathbf{X} = \mathbf{T}_A \mathbf{P}'_A + \mathbf{E}_A \quad (3)$$

where $\mathbf{X} = \left[\mathbf{X}^1 \mathbf{X}^2 ... \mathbf{X}^b ... \mathbf{X}^B\right]$ is the concatenated multi-block data set, $\mathbf{T}_A = \left[\mathbf{t}_1 \mathbf{t}_2 ... \mathbf{t}_a ... \mathbf{t}_A\right]$ is the global score matrix containing *A* global score vectors, the corresponding global loading vectors are collected in the global loading matrix $\mathbf{P}_A = \left[\mathbf{p}_1 \mathbf{p}_2 .. \mathbf{p}_a .. \mathbf{p}_A\right]$ and $\mathbf{E}_A = \left[\mathbf{E}_A^1 \mathbf{E}_A^2 ... \mathbf{E}_A^b ... \mathbf{E}_A^B\right]$ is the matrix of residuals for the model. The corresponding block parameters for the model in Eq. 3 are given in Eq. 4:

$$\mathbf{X}_A^b = \mathbf{T}_A^b \mathbf{P}_A^{b\prime} + \mathbf{E}_A^b \qquad (4)$$

where $\mathbf{T}_A^b = \left[\mathbf{t}_1^b \mathbf{t}_1^b ... \mathbf{t}_a^b ... \mathbf{t}_A^b\right]$ is the block score matrix for block *b* ($b = 1, ..., B$), $\mathbf{P}_A^b$ is the corresponding segment of the global loading matrix $\mathbf{P}_A$ for block *b* and $\mathbf{E}_A^b$ is the residual matrix for the block model *b*.

The algorithm for CPCA contains two main steps:

1) Parameter estimation (*a*th component, for $a = 1, 2, ..., A$): the global and block parameters (i.e. *a*th global score, *a*th global loading, *a*th block scores and *a*th block loadings) are initially calculated. Nonlinear Iterative Partial Least Squares (NIPALS) is the method which is commonly used for CPCA parameter estimation. The iterative procedure of NIPALS for the calculation of parameters for *a*th component is shown in Fig. 6. First, an arbitrary vector ($\mathbf{t}_a$) is chosen as the initial global score vector for component *a*. (i) Loading vector for every block ($\mathbf{p}_a^b$) is then obtained by column-wise projection of each data block on $\mathbf{t}_a$. (ii) Block score vector for every block ($\mathbf{t}_a^b$) is then calculated by row-wise projection of every data block to its loading vector. These block score vectors are then put together and form a matrix of block score vectors (**T**). (iii) This matrix is then projected on $\mathbf{t}_a$ in order to obtain global loading weights ($\mathbf{w}_a$). (iv) A new estimate for the global score vector is obtained by projecting **T** on $\mathbf{w}_a$. The whole process is iterated until convergence of the global score vector ($\mathbf{t}_a$).

2) Deflation (*a*th component, for $a = 1, 2, ..., A$): the data set is deflated by subtracting the variation that corresponds to the *a*th calculated parameters. Different deflation strategies are suggested to be applied when dealing with multi-block data sets. CPCA implements the deflation on global scores where the variation due to the *a*th global score is removed from every data block. Two alternative deflation methods are also available: deflation on block scores [30] and deflation on block loadings (applied by Multiple Co-inertia Analysis (MCoA) [46]).

The deflated data set is then used for calculation of the $(a+1)th$ parameters.

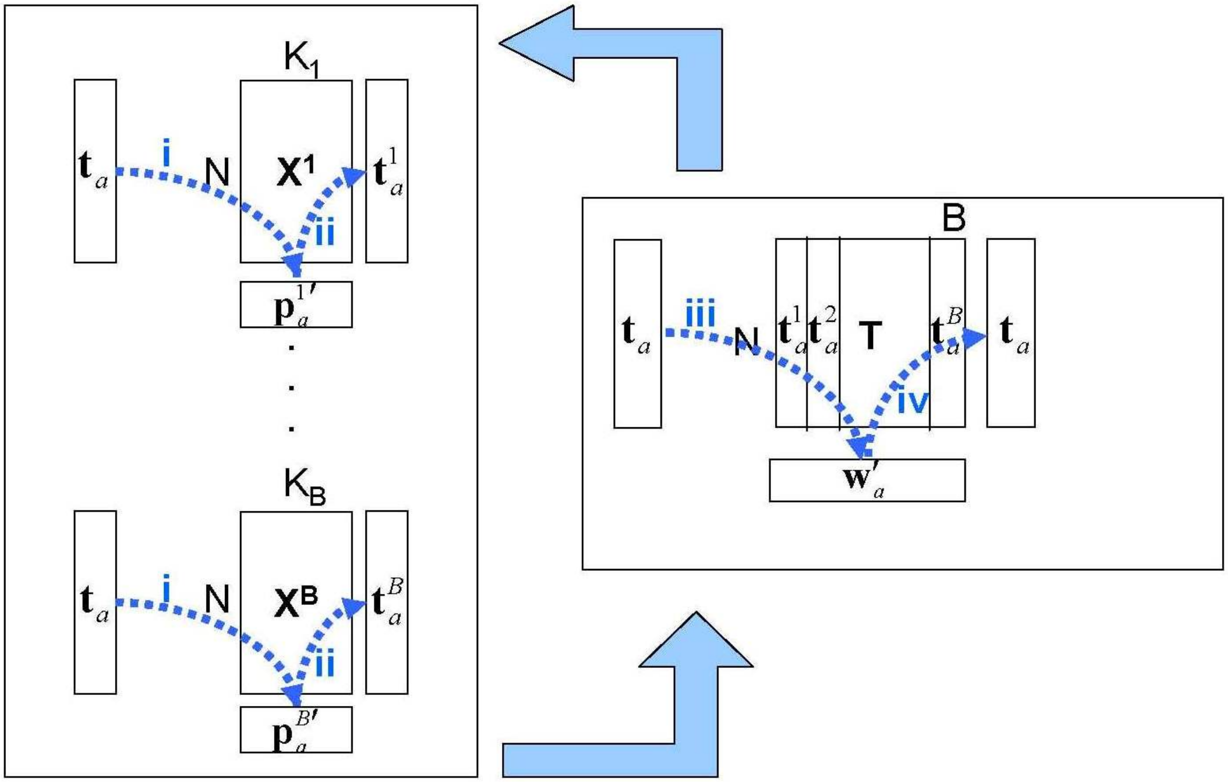


Figure 6: The iterative procedure of NIPALS for calculating CPCA parameters. The figure is adopted from [3].

3.2.2 MBPLSR parameter calculation

Analyzing the mean-centered and scaled multi-block descriptor data set ($\mathbf{X} = \left[\mathbf{X}^1,...,\mathbf{X}^b,...,\mathbf{X}^B\right]$) and response data ($\mathbf{Y}$) by means of MBPLSR models the data as sum of A relevant latent variables plus residual matrices. The global MBPLSR model is as the following:

$$\begin{aligned}
&\mathbf{T}_A = \mathbf{X}\mathbf{V}_A \\
&\mathbf{Y} = \mathbf{X}\mathbf{B}_A + \mathbf{F}_A \\
&\mathbf{X} = \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A \qquad (5) \\
&\mathbf{Y} = \mathbf{T}_A\mathbf{Q}'_A + \mathbf{F}_A \\
&\mathbf{B}_A = \mathbf{V}_A\mathbf{Q}'_A
\end{aligned}$$

where $\mathbf{X} = \left[\mathbf{X}^1\mathbf{X}^2...\mathbf{X}^b...\mathbf{X}^B\right]$ is the concatenated descriptor multi-block data set and $\mathbf{Y}$ is the response data set. $\mathbf{T}_A = \left[\mathbf{t}_1\mathbf{t}_2...\mathbf{t}_a...\mathbf{t}_A\right]$ is the matrix of A scores from $\mathbf{X}$ (so called global score vectors) defined by weight vectors $\mathbf{V}_A = \left[\mathbf{v}_1\mathbf{v}_2...\mathbf{v}_a...\mathbf{v}_A\right]$ so as to maximize the total covariance between each consecutive score vector $\mathbf{t}_a$ and $\mathbf{Y}$. $\mathbf{P}_A = \left[\mathbf{p}_1\mathbf{p}_2...\mathbf{p}_a...\mathbf{p}_A\right]$ and $\mathbf{Q}_A = \left[\mathbf{q}_1\mathbf{q}_2...\mathbf{q}_a...\mathbf{q}_A\right]$ are the loadings for $\mathbf{X}$ and $\mathbf{Y}$ respectively, $\mathbf{E}_A = \left[\mathbf{E}_A^1\mathbf{E}_A^2...\mathbf{E}_A^b...\mathbf{E}_A^B\right]$ and $\mathbf{F}_A$ are the residuals for modeling $\mathbf{X}$ and $\mathbf{Y}$ respectively and $\mathbf{B}_A$ is the regression coefficients (of size $K \times J$). The corresponding block parameters for the model in Eq. 5 are given by the following model:

$$\mathbf{X}^b = \mathbf{T}_A^b\mathbf{P}_A^{b\prime} + \mathbf{E}_A^b \qquad (6)$$

where $\mathbf{T}_A^b$, $\mathbf{P}_A^{b\prime}$ and $\mathbf{E}_A^b$ are block scores, block loadings and block residuals respectively belonging to data block $\mathbf{X}^b$.

Similar to CPCA, the algorithm for MBPLSR consists of two main steps:

1) Parameter estimation (*a*th component, for $a = 1,2,...,A$): *a*th parameters (i.e. *a*th global scores, *a*th X-block scores, *a*th X-block loading weights, *a*th Y-scores and *a*th Y-loading weights) are calculated first. Several variations of PLSR algorithm to be used when having more than one descriptor data block are available [26, 47-50]. The procedure for MBPLSR algorithm proposed

by Wangen and Kowalski [27][Paper II], that handle most types of different relationships between the data blocks, is shown in Fig. 7. An arbitrary vector is chosen as the initial *a*th Y-score ($\mathbf{u}_a$). (i) Every data block ($\mathbf{X}^b$) is then projected column-wise on $\mathbf{u}_a$ in order to obtain block loading weights ($\mathbf{w}_a^b$). (ii) Row-wise projection of each data block on its loading weights results in the block scores ($\mathbf{t}_a^b$). The block scores from all data blocks are then put together in the matrix of block scores (**T**). (iii) Projecting **T** on $\mathbf{u}_a$ gives the super loading vector ($\mathbf{w}_a^s$). (iv) **T** is then projected row-wise on $\mathbf{w}_a^s$ in order to obtain global scores ($\mathbf{t}_a$). (v) Projecting **Y** on the global scores gives Y-loading ($\mathbf{q}_a$). (vi) A new estimation for Y-score ($\mathbf{u}_a$) is then calculated from projecting **Y** on its loading. The procedure is iterated until the convergence of global scores ($\mathbf{t}_a$).

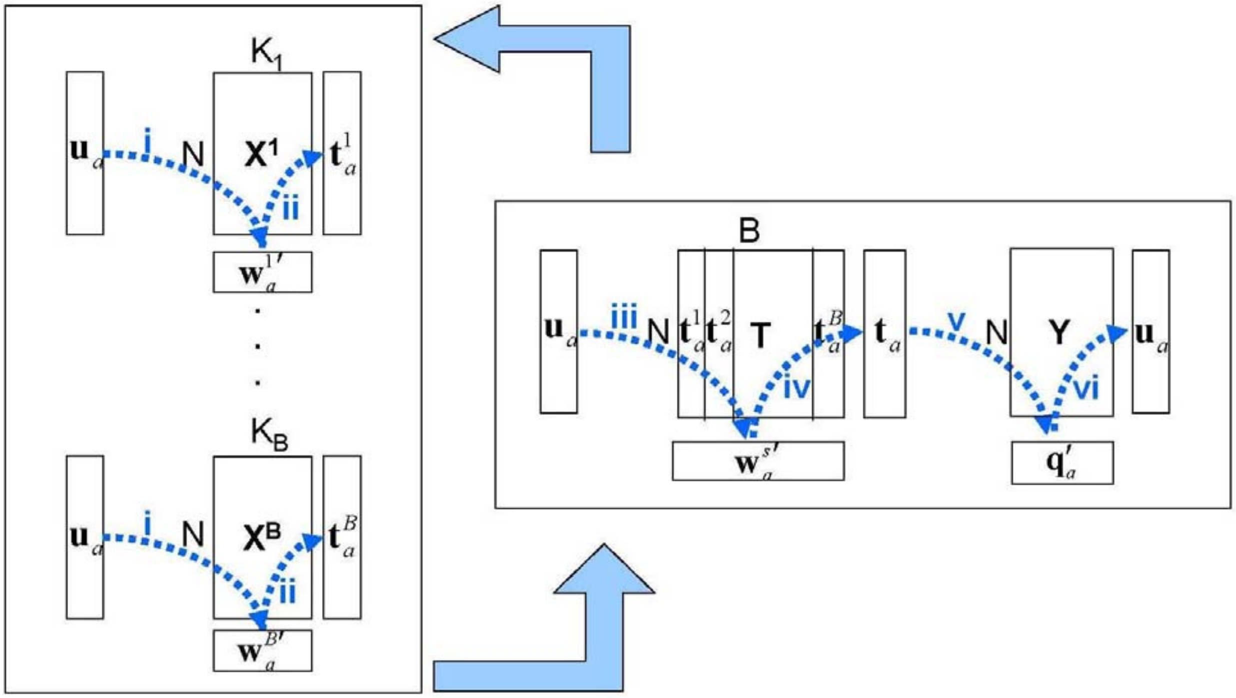


Figure 7: The iterative procedure for calculating MBPLSR parameters.

2) Deflation (*a*th component, for $a = 1, 2, ..., A$): similar to CPCA, when the *a*th parameters are calculated the data is deflated by removing the variations that corresponds to the *a*th parameters. Different deflation strategies are available for MBPLSR [30]. The method that is commonly used deflates both **X** and **Y** on global scores [51]. An alternative deflation strategy is to deflate data

blocks ($\mathbf{X}^b$) by the block scores ($\mathbf{t}_a^b$) and **Y** by the global score ($\mathbf{t}_a$) [27, 31]. The other possibility is to deflate only **X** on the global scores and not deflate **Y** at all [31].

The deflated **X** and **Y** are then used for the calculation of (*a+1*)th parameters.

3.3 Visualization tools

3.3.1 Score plot

PCs are the directions of the largest variances in a data set in a descending order (i.e. the first PC is responsible for most of the variation). PCs build a new coordinate system. In fact, the axes of the original variable space are rotated in a way that the axes in the new coordinate system are expanding the variances of the data. The coordinates of the original samples in this new rotated system are given by scores. Each score vector ($\mathbf{t}_a$) is in fact a latent variable which is a linear combination of the original variables ($\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_K]$). A score plot of two given PCs illustrates the location of the samples in the new coordinate system.

Two types of scores are calculated by CPCA and MBPLSR: global scores ($\mathbf{T}_A = [\mathbf{t}_1\mathbf{t}_2...\mathbf{t}_a...\mathbf{t}_A]$) and block scores ($\mathbf{T}_A^b = [\mathbf{t}_1^b\mathbf{t}_2^b...\mathbf{t}_a^b...\mathbf{t}_A^b]$ for $b = 1, 2, ..., B$). Consequently, two types of score plots are becoming available by these analyses: global score plot and block score plots. The global score plot provides an overview over the underlying pattern that is in common between the data from all of the blocks whereas the block score plot illustrates how much of the global underlying pattern is present in every individual block. These plots indicate the contribution of every block to the detected global underlying pattern.

It is worth noting that deflating the data by global score (i.e. the most commonly used method which is also employed by CPCA) results in having orthogonal global scores while the block scores are not orthogonal. This means that the axes for the global score plot are always orthogonal (i.e. independent), while the axes for the block score plots are not necessarily orthogonal. This can in some instances lead to strange block patterns if the block follows a very different pattern from the common underlying one. An example is given in Fig. 8 where the CPCA block and global score plots are illustrated for a five-block data set. One can notice that block five is showing a strange pattern (Fig. 8e) indicating that the first and second block scores for block five are linearly correlated to each other. Such situation can not happen in a global score plot since the global scores for different components are always orthogonal (given that the data is deflated on the global scores or on the block loadings).

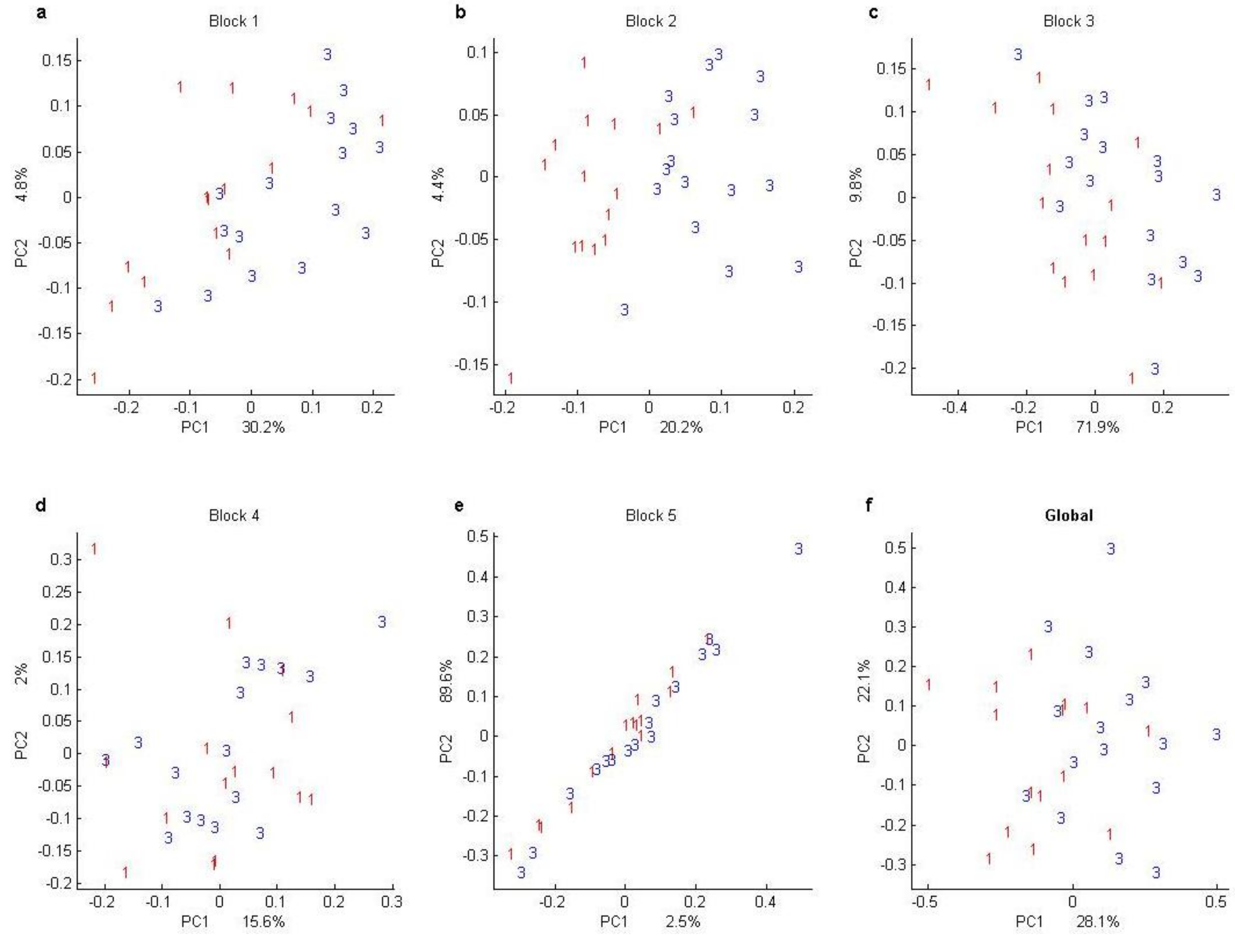


Figure 8: Score plots for the CPCA analysis of a multi-block data set. (a-e) The block score plots for PC1 and PC2. (f) The global score plot for PC1 and PC2.

3.3.2 Loading plot

PCs which are the axes of the new coordinate system are in fact latent variables that are linear combinations of the original variables. Loadings ($\mathbf{P}_A = [\mathbf{p}_1 .. \mathbf{p}_a .. \mathbf{p}_A]$) represent the weights of the original variables in this new coordinate system defining the contribution of each original variable to the new latent variables. Loadings can be visualized in different ways e.g. plotting the loadings for the first PC ($\mathbf{p}_1$) against that for the second PC ($\mathbf{p}_2$) or plotting PCs against the original variables. Loading plot corresponds to the score plots in Fig. 3a and Fig. 3c where $\mathbf{p}_1$

and $\mathbf{p}_2$ are plotted as a function of the original variables is shown in Fig. 9. It is worth mentioning that loadings are unit-free parameters and therefore the same loadings corresponds to both unscaled and scaled data in Fig. 3a and Fig. 3c. The contribution of the original variables to the new latent variables can be studied by loading plots. The location of the variable 186 is shown on the figure by a green dashed line. Since both PC loadings reach a relatively large negative peak for this variable, it can be concluded that the variable is significantly contributing to both PC1 and PC2 in the same way. Variable 225 is also marked on the figure. As it can be seen, PC1 and PC2 reach a negative and a positive peak respectively for this variable leading to the conclusion that variable 225 is significantly contributing to PC1 negatively and to PC2 positively. The location of variable 368 is also shown on the figure. It can be seen that both loadings are zero for this variable leading to the conclusion that variable 368 is not contributing either to the first or to the second PC.

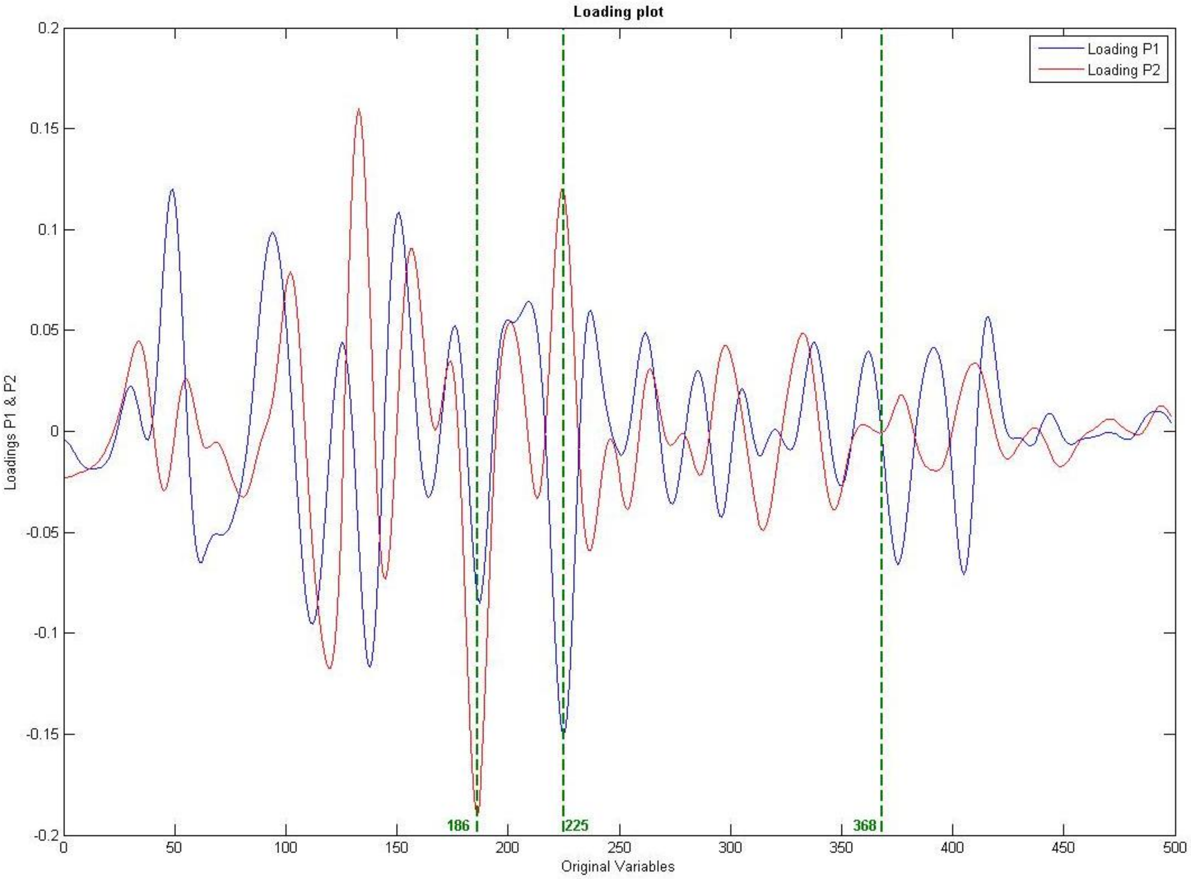


Figure 9: Loading plot for the PCA analysis of a data set: loadings for the first and second components are plotted in blue and red respectively.

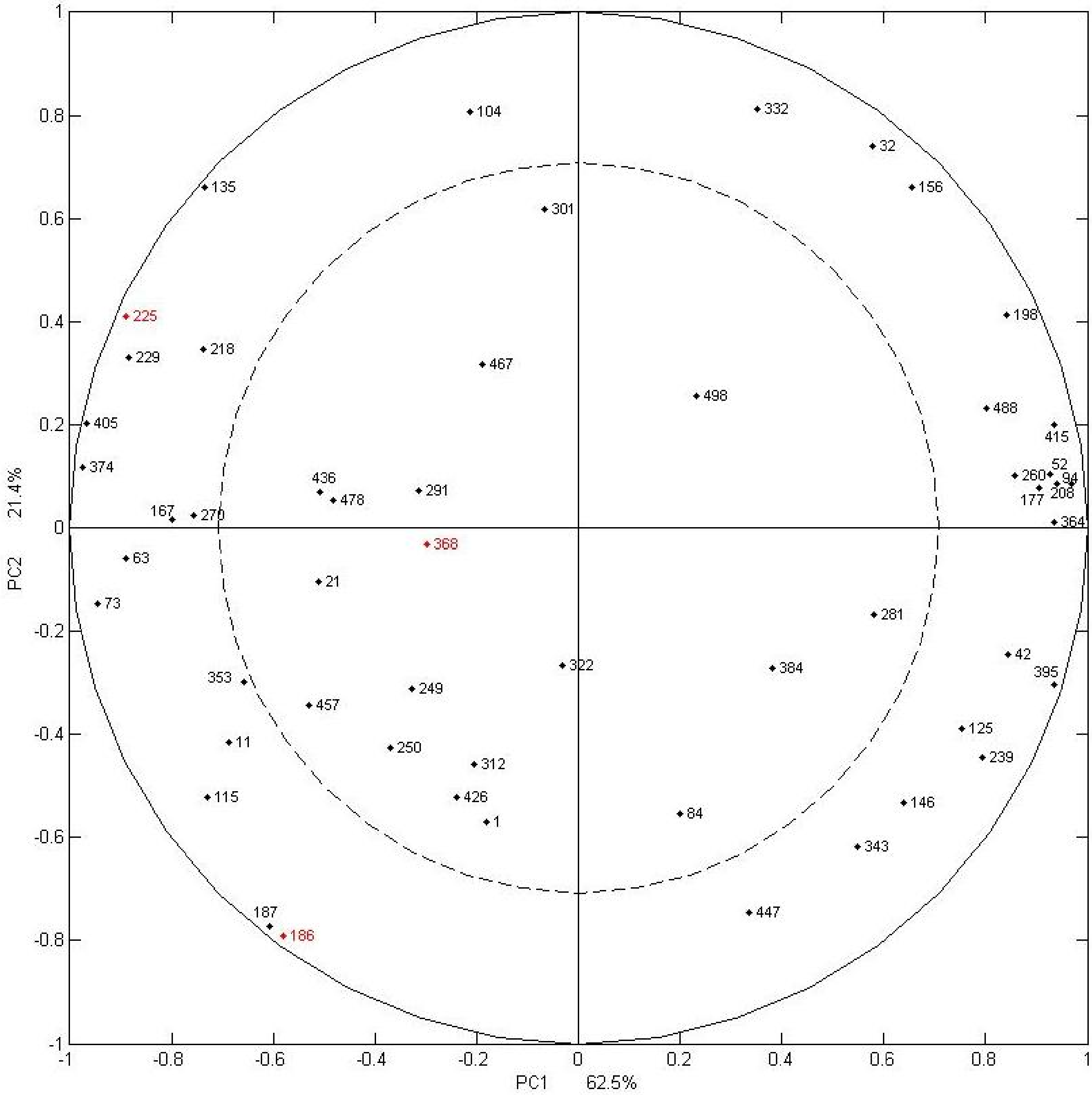


Figure 10: Correlation loading plot correspond to the loadings in Fig. 9.

3.3.3 Correlation loading plot

Correlation coefficients (r) between the latent variables and the original variables are plotted in the correlation loading plot. Correlation coefficient is a unit-free parameter and can be considered as a measure of dependencies between the latent variables and the original variables. Fig. 10 illustrates the correlation loading plot corresponding to the loading plot in Fig9. The abscissa in the correlation loading plot is the correlation coefficient (r_1) between the variable and the corresponding PC (e.g. PC1) and the ordinate is the correlation coefficient (r_2) between the variable and PC2. Since the correlation loading plot would become too crowded having all 498 variables in the data set, only few were chosen to be shown in Fig. 10. The relative amount of explained variance is represented by sum of the squared correlation coefficients (i.e. $r_1^2 + r_2^2$). Therefore, the outer and inner circles with radii of 1 and $\sqrt{0.5}$ represent 100% and 50% explained variances respectively. Three variables that were marked in the loading plot in Fig. 9 are plotted in red in Fig. 10. Similar conclusions as in Fig. 9 are also derived here. E.g. variables 186 and 225 are contributing significantly to both PCs, while variable 368 is not contributing to these PCs.

3.4 Validation

Model validation is one of the main challenges in data analysis. The goal of validating is to ensure the reliability of the model and assess the final outcomes. Different resampling techniques [52] in statistics are available for the purpose of validation e.g. Permutation test [53], Bootstrapping [54], Cross-validation [55] and Jack-knifing [56, 57]. The importance of validating the results should not be neglected when dealing with methods (e.g. CPCA and MBPLSR) that provide the user with practical visualization tools (e.g. score plots). This is because the visualization tools can be misleading for the scientist's mind due to their fascinating graphical capabilities. Cross-validation and Jack-knifing are two methods that can be applied for validating the visually identified patterns of the score plots [Paper I, Paper II]. These methods are described in the following sections.

3.4.1 Cross-validation

Cross-validation aims at verifying the reproducibility of the results by predicting how well a model will perform on future data sets. For this purpose the data is split into *M* subsets. Each subset is considered as a test set (left-out data segment) when the rest of the data is used as training set (leave-in data segment). Models are first built on the leave-in samples and are then applied on the left-out data in order to validate how well the models will work for a data not included in the modeling process. The procedure is performed *M* times for all different data segments. The error is calculated for all of these sub-models and is then used as a measure for the model goodness. This is called an *M*-fold cross-validation. If *M* is chosen equal to the total number of samples, it is called a leave-one-out (or full) cross-validation (i.e. every single sample is left out in turn to be used as test set).

Studying the Root Mean Squared Errors (RMSE) calculated from cross-validating CPCA or MBPLSR models provides an opportunity for evaluating the models and assessing the validity of visually identified patterns [Paper I, Paper II]. Moreover, comparing the RMSE calculated for the global model with those calculated for every block gives an indication for the contribution of the blocks to the global pattern.

3.4.2 Jack-knifing

Jack-knifing is employed for estimating the bias and variance of a statistic when using a random set of samples. Similar to cross-validation samples are left out in turn and the statistic is estimated based on the leave-in samples. A set of estimations for the statistic is calculated in this way. It is in fact the variation of the statistic from sub-model to sub-model that gives an estimate for the true variance of the statistic. Jack-knifing may be employed in the calculation of uncertainty t-test statistic by giving an estimate for the variance of the statistic [Paper I].

3.4.3 Permutation test

Permutation test is a resampling technique that is employed for running statistical significant tests. The test statistic under study is firstly estimated for the subjects in the experiment in their original orders (this may be called true test statistic). The subjects of the experiment are then rearranged in all possible ways and the test statistic is calculated in every permutation round. This procedure results in generating a distribution for the test statistic under study. The ranking of the true test statistic among the generated distribution gives a p-value for the significance level of the true test statistic. In the cases that there are too many possibilities for reordering of the subjects, Monte Carlo sampling technique [58, 59] can be used where a sub-set of the possible orderings is chosen randomly in order to be used for generating the distribution. The choice for the number of the elements in the sub-set depends on the accuracy of the test.

3.5 Degree of Freedom (DF)

The term DF is widely used in different fields of science (e.g. mechanics, physics, chemistry, statistics and chemometrics) referring to different yet related concepts. The concept of DF in mechanics refers to "independent displacements and/or rotations that specify the orientation of the body or system" [60] while in statistics the DF is defined as "the number of values in the final calculation of a statistic that are free to vary" [61, 62]. DF plays an important role when statistical hypothesis tests such as F-test and student's t-test are run. It is also an important issue when assessing statistical models and estimating parameters since neglecting the DF may lead to misinterpreting an "over-fitted model" as a "good model". Estimating parameters such as a variable's mean requires having knowledge of the remaining DFs in the data as well i.e. the total number of independent samples minus the number of independent estimated parameters. Martens and Næs used the term DF in the field of chemometrics in 1989 where they discussed the "degrees of freedom used in the fitting of the regression equations" [63]. However, at that time no specific definition was given for the term DF in Chemometrics. The importance of having knowledge about the correct number of consumed DF draws special attention when facing issues such as calculation of prediction uncertainty for a PCA- or PLSR-model. By the number of DFs that are being consumed by implementing a modeling technique (such as PCA or PLSR), we generally refer to the number of pieces of independent and useful information from the data that are consumed during the process. Estimating the DFs that are consumed when a data set is modeled using multivariate data modeling techniques that are based on latent variables (e.g. PCA or PLSR) is very complicated. Few studies with the focus on estimating the consumption of DF by these models can be found in literature [64-68].

4. Paper summaries

Paper I. Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods

Rapid development of systems biology leads to generating large different types of –omics data sets. The data are in general huge multi-block sets generated by applying different high–throughput techniques on the same samples e.g. proteomics–, genomics– and metabolomics–data. The ongoing challenge is to integrate these different –omics measurements, analyze them in light of the background knowledge and interpret the outcomes. A data analysis framework for analyzing such massive data sets was presented in this article. Visualization tools were presented together with their interpretational aspects. These tools enable investigating the common underlying patterns in complex multi-block data sets. They make it possible to investigate the pattern shared by all data blocks as well as the presentation of the global pattern in each block. Validation tools for evaluating the detected patterns on a block level were introduced in this article. Tools that can be used for detecting outliers at global and block levels were also introduced. The methods in this paper were introduced for Consensus Principal Component Analysis (CPCA) while the general concepts are still possible to be transferred to other multi-block analysis methods e.g. MBPLSR. The presented methods were applied and illustrated by a multi-block microbiological data set.

Paper II. Model validation and error estimation in multi-block partial least squares regression

Design of the multi-response experiments has been given special attention by many systems biology studies. Multi-block Partial Least Squares Regression (MBPLSR) can be implemented for analyzing such data sets. Consequently, investigating the effect of the design factors on the measured variables becomes an important issue for these studies. MBPLSR-Discriminant Analysis (MBPLSR-DA) can be applied in such situations where the study aims at separating different groups of observations. The significant role of MBPLSR family of methods for analyzing systems biology data is therefore clear. Extensive applications of these methods for the analysis of data require validation strategies. Tools for validating the prediction ability of the

MBPLSR models were introduced in this article. These tools can be used for validation on the block level as well as on the global level enabling the user to investigate the contribution of every block to the grouping pattern as well as studying the common grouping pattern shared by all the blocks. Moreover, tools for validating the model stability were also introduced which are available on both global and block levels. In addition we investigated the problem of choosing the number of latent variables to be included in a PLSR model. The proposed methods were illustrated with the same data set that was used in Paper I.

Paper III. Degrees of freedom estimation in Principal Component Analysis and Consensus Principal Component Analysis

In this paper, we ran simulation studies in order to investigate the true number of DFs consumed when cross-validating PCA and CPCA models. The simulation studies confirmed the formula for estimating the consumed DFs which we proposed in Paper I. In the cases that cross-validation is not implemented and the errors are therefore estimated by fitting the model from the same samples, the number of consumed DFs increases. The reason is the loss of DFs due to the search process that leads to CPCA parameter estimation. By simulating data sets with different eigenvalue structures, we showed that the DF consumption depends on the eigenvalue structure of the data to be modeled. We also proposed a method for estimating the DFs that are lost during the search processes of PCA and CPCA. The method was afterwards implemented on real data sets from spectroscopy. We estimated the consumed DFs for a real data set considering its eigenvalue structure. We showed that the estimated number of DFs can be used for a different real data set that has a similar eigenvalue structure.

In cross-validation a part of the data set – often a small part – is set aside for validation. This process is repeated until all samples are once used as a test set, without taking the same samples twice or more times as test set. Especially when a data set is small and one can afford to set aside only a small number of samples, cross-validation is attractive. Cross-validation results are questioned by some scientists since the same samples are used both for modeling and testing the models. Simulation studies in this paper indicated that the findings of cross-validation agree with those from independent test sets. The effect of the number of cross-validation segments on the results was also studied. We even showed that using a higher number of cross-validation segments does not necessarily lead to better results.

Paper IV. Deflation strategies for multi-block principal component analysis revisited

Different deflation strategies can be implemented when analyzing data sets by methods that are based on latent variables. The choice of the deflation strategy affects the estimated parameters (i.e. scores and loadings) which therefore leads to different interpretation of the results. Three different strategies are available for running multi-block Principal Component Analysis: i) deflation on global scores that is employed by Consensus PCA (CPCA), ii) deflation on block scores and iii) deflation on block loadings that is employed by Multiple Co-inertia Analysis (MCoA). In this paper we described these methods in details and compared them with each other. We studied the theoretical properties of these methods as well as their interpretational aspects. Orthogonality properties for block and global scores and for block and global loadings were also discussed. Data block's reconstruction formulas for different deflation strategies were established.

The effect of implementing different deflation strategies on the results were illustrated by an example. The interpretational aspects of different deflation strategies were also studied by the example. We showed that deflation by global scores and by block loadings have some advantages over the deflation by block scores. In order to gain insight into the multi-block data set we proposed using the deflation by global scores (i.e. the global variation pattern is subtracted from every block) and compared the results with those using the deflation by block loadings where the block variable variation pattern is subtracted. We also showed that it is difficult to interpret the block patterns in connection to the global pattern when deflating by block scores. This is because new underlying block loadings are defined for the purpose of deflation which leads to block patterns that are more similar to results of PCA of every block instead of the multi-block PCA results.

Paper V. Simultaneous analysis of inter- and intra-class lipid changes in lipidomics studies

Lipidomics is an emerging field of systems biology. Due to its rapid development there is a growing need for the methods that can integrate and analyze data from different lipid classes.

Investigating the remodeling effects (i.e. when lipids dynamics happen within a lipid class) and lipid metabolisms (i.e. when lipids from one lipid class are transferred into lipids from other lipid classes) are some of the main challenges of the lipidomics studies. In this paper we proposed a multi-block structure for the lipidomics data sets with respect to the lipid classes and lipid sub-classes. We then suggested pre-processing the multi-block lipidomics data by two different normalization strategies: 1) by normalizing the amount of lipids with respect to the total amount of lipids, 2) by normalizing the amount of lipids with respect to the total amount of lipids in each lipid class and keeping the sum values in a separate data block. Using a simulated data set we showed that the second pre-processing strategy improves the detection of remodeling effects. We proposed employing multi-block methods (i.e. CPCA and MBPLSR) for integrating and analyzing the data from different lipid classes simultaneously. In order to investigate the importance of each lipid class for the global pattern, we suggested running Monte Carlo permutation tests which led to p-values for the significance of every lipid class. The suggested methods were implemented on a real lipidomics data set and the results were interpreted from a data analysis point of view. The readers were provided with the in-house-written and standard MATLAB routines for implementing the proposed framework for the analysis of lipidomics data sets.

Paper VI. Fish Oil Supplementation Alters the Plasma Lipidomic Profile and Increases Long-Chain PUFAs of Phospholipids and Triglycerides in Healthy Subjects

Our contribution to this paper was the analysis of a lipidomics data set from a human intervention study by means of the framework that was proposed by us in Paper V. A seven-week double-blinded randomized controlled parallel-group intervention study was run to investigate the effect of fish oil supplementation on plasma lipidomic profile in healthy subjects. The subjects completed a fully controlled diet period during the first three weeks of the intervention study where they received capsules containing either fish oil or high oleic sunflower oil.

We structured the lipidomics data from the intervention study into 10 lipid classes. In order to be able to investigate the remodeling effects of lipids, we then pre-processed the lipid classes by normalizing each lipid class by its total amount. The total amounts of lipids for each lipid class were restored in a separate data block. In order to study the differences between the intervention groups the pre-processed multi-block data of lipid classes (containing 11 blocks) were then analyzed by MBPLSR-DA. As expected, a good separation of the groups after three weeks of intervention was detected. The influence of each lipid class on the global MBPLSR model was estimated by the methods from Paper II. Several data blocks (7 out of 11) showed a significant contribution to the detected pattern while the block that contained the sum values of the lipid classes did not show a significant contribution. This led us to the conclusion that remodeling of lipids was happening within each lipid class. In order to detect the significant lipids we ran uncertainty t-tests on the regression coefficients of the MBPLSR model and 75 lipids were found to be significantly altered during the intervention. The results were compared to the results from univariate t-tests and 49 lipids were found to be significant by both methods.

5. Results and discussions

5.1 Validation

Integrating and analyzing multi-block data sets from the –omics field, interpreting the findings and finally validating the results, is an increasing challenge for the data analysts. Multi-block methods that are based on latent variables, such as Consensus Principal Component Analysis (CPCA) and Multi-block Partial Least Squares Regression (MBPLSR), are tools that can be employed for the purpose of integrating and analyzing such data sets. These multivariate techniques provide the user with powerful visualization tools e.g. score plots. However, the detected patterns are subjective and need to be validated. This necessitates the existence of validation techniques for evaluating the findings of these methods. Papers I and II proposed validation tools for CPCA and MBPLSR, respectively.

5.1.1 Cross-validation

The patterns that are detected by CPCA can be validated by studying the cross-validated Root Mean Squared Errors (RMSE) of the CPCA model. In Paper I we proposed a method for calculating cross-validated RMSE for the global model as well as for each individual block. Studying the cross-validated global RMSE plot allows evaluating the patterns detected in the global score plot. The cross-validated RMSE plot also gives an indication for the number of Principal Components (PCs) to be included in a CPCA model. The contribution of each block to the global detected pattern can then be studied by investigating cross-validated block RMSE plots. In order to study the significance of variables in the multi-block model, we suggested using an uncertainty t-test (based on cross-validation and jack-knifing) for assessing the contribution of the variables to the CPCA model. This results in a p-value for each variable and helps detecting important variables. The stabilities of the samples within a PCA model can be studied by investigating the stability score plots [69]. In Paper I we extended these plots to multi-block situations and proposed methods for the calculation of block stability plots. These plots are used for assessing the extent to which the calculated scores are influenced by every individual sample. The reliability of the CPCA model can also be assessed by studying the global stability plot. An important aspect of the block stability plots is that it allows identifying outliers on a block level.

These outliers are either interesting objects with a special property that is only detectable by means of the technique related to that specific block or they have been subjected to an error in the respective block.

In Paper II we undertook a study for developing validation tools for MBPLSR. Similar to CPCA, the detected patterns here can be validated by investigating the cross-validated global RMSE which is calculated for the descriptor data set ($\mathbf{X}$). In this paper we proposed a method to calculate cross-validated RMSE for each block separately in order to study the contribution of each individual descriptor block ($\mathbf{X}^b$) to the global pattern. Investigating the cross-validated RMSE plots calculated for $\mathbf{X}$ enables evaluating the grouping patterns that are detected in the score plots. The block cross-validated RMSE plots give an indication for the contribution of every block to the global MBPLSR model. Since MBPLSR constructs predictive models, the models can be also validated for their predictive ability. For this purpose, we proposed calculating the cross-validated RMSE for the response data set ($\mathbf{Y}$) with respect to both the global model parameters as well as the block parameters. Studying the cross-validated global RMSE calculated for $\mathbf{Y}$ enables evaluating the predictive ability of the global MBPLSR model which is an indication for the predictive ability of the multi-block descriptor data set ($\mathbf{X} = \left[\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^B\right]$). By investigating the block cross-validated RMSE for $\mathbf{Y}$ the ability of every descriptor data block to predict the response data ($\mathbf{Y}$) can be studied.

5.1.2 Cross-validation vs. independent test set

Our proposed methods for assessing the reliability of the CPCA and MBPLSR models in Paper I and Paper II were based on cross-validation. However, the cross-validation itself can be criticized for the fact that the same samples are used both for modeling and for testing the models. Alternative approach for such critics is the use of an independent test set. Therefore we undertook simulation studies (Paper III) where we compared the results from running cross-validation with the results calculated by means of independent test sets. Our simulation studies show that the results from cross-validation agree with those from independent test sets. When running cross-validation, one faces the important question of "How many cross-validation segments to be used?". It is believed that, as a rule of thumb, increasing the number of cross-validation segments leads to more robust results. Especially, leave-one-out cross-validation is supposed to get the most stable results. Since the models are built based on all the samples except only one sample, they are not expected to vary significantly from each other. However, our simulation study indicates the fact that using a higher number of cross-validation segments does not always lead to better results. A 10-fold cross-validation gave similar results as an independent test set for all of our simulated data sets.

5.2 Degree of Freedom (DF)

Calculating the cross-validated MSE i.e. the average of the squared errors, as described in the previous section requires having knowledge over the DFs that are consumed while errors are estimated. In Paper I we proposed a tentative formula for the calculation of cross-validated MSE. In Paper III we wanted to investigate the validity of that tentative formula. Since DF is a challenging issue in multivariate modeling, we extended our simulation studies to the estimation of DFs that are consumed during the PCA and CPCA modeling processes. Our simulation studies show that DFs are consumed at two different stages: i) when searching for the direction of the largest variation in the data set. ii) When estimating parameters. Calculating the cross-validated errors does not involve any search process in the left-out data and the parameters are estimated based on the directions found in the leave-in data segment. However, our simulations showed that some DFs are still being consumed for the parameters estimation when using an independent test data set. The DFs consumed are equal to the number of estimated parameters (a) in PCA models [63] and equal to the leverage of the block loadings (h_A^b) for every block in CPCA models. When errors are calculated without cross-validation, more DFs are consumed due to the search process. In Paper III we proposed a method for estimating the DFs consumption of the search process. Only one formula (proposed by Faber in [68]) is available in literature for calculating the number of DFs consumed by a PCA model. We estimated the overall DFs consumption for PCA models for simulated data sets and compared our results with the results from the formula that was previously proposed by Faber. Our results indicated that the previously proposed formula does not take into account the DFs consumed by the search process.

We have shown that estimating the cross-validated errors (for PCA and CPCA) does not necessarily require running cross-validation. The eigenvalue structure of a data set defines the DFs. Data sets originating from the same type of measurements have similar eigenvalue structures, given that the size of the data set is the same. Therefore, if one is dealing with similar data sets in terms of the eigenvalue structure, it is sufficient to estimate the DFs that are consumed by the search process, once and to use the DFs for correcting the fitted RMSE calculated for comparable data sets.

5.3 Deflation strategies

Deflation plays an important role in estimating CPCA and MBPLSR parameters. Different deflation strategies for CPCA and MBPLSR can be found in literature. Deflation strategies that can be used for MBPLSR were studied in [31]. In Paper IV, we investigated three deflation strategies (i.e. deflation by global scores, by block scores and by block loadings) that are commonly used for running multi-block PCA. The choice of the deflation strategy affects the orthogonality properties of the estimated parameters (i.e. scores and loadings). It also influences the reconstruction procedure for the data blocks which leads to different explained variances for each PC. The differences of the calculated parameters can in some instances lead to very different visual patterns in the score plots. We discussed the interpretational aspects of different deflation strategies and gave an overview over their properties. When deflation is performed on the global scores, the common variation pattern among all blocks is subtracted in every deflation stage. However, deflating on block loadings subtracts the variables variation pattern belonging to every data block from itself. These two deflation methods can be used in parallel considering their different interpretational aspects. Deflation on block scores requires calculating new block loadings. Therefore the new block loadings lose their relationship to the global parameters and therefore the block patterns are not directly interpretable along with the global pattern. The calculation of the new block loadings is performed by going through an extra NIPALS step which moves the block results toward the results of running PCA on the given block. Therefore the patterns detected in the block score plots become more similar to the patterns that are seen in the PCA score plot of each block. This can be considered a drawback of this deflation strategy since the aim of running multi-block data analysis is detecting a common pattern among the different blocks and not running single block analysis.

5.4 Integrating lipidomics data

Lipidomics is an emerging –omics field which aims at investigating the role of the lipids in the biological systems. Detecting the remodeling of lipids is an important challenge for the data analysts in the lipidomics studies. In Paper V we proposed tools for pre-processing the lipidomics data sets in a way that promotes the detection of lipid remodeling. The proposed pre-processing strategy re-arranges the lipidomics data set into a multi-block data of different lipid classes. This method enables a simultaneous analysis of lipid species and lipid classes by means of multi-block data analysis techniques. Therefore, in Paper V, we further developed the multi-block methods from Paper I and Paper II into a framework for integrating and analyzing the multi-block set of lipid classes. New validation tools (based on Monte Carlo permutation tests) were added to the previous methods providing more insight into the importance of different lipid classes. We also developed new tools for analyzing the lipidomics data sets with respect to the underlying design of the experiments.

5.5 Application of the proposed methods

In Paper VI, we applied the proposed methods from Paper V on a lipidomics data set. The data was from a seven-week double-blinded randomized controlled parallel-group intervention study which aimed at investigating the effect of fish oil supplementation on plasma lipidomic profile in healthy subjects. The lipidomics data consisted of three data matrices of 568 variables (i.e. detected lipids) for 33 samples (i.e. healthy subjects) measured at three different time points. We grouped the data into 11 blocks where each lipid class was defined as an individual data block. Afterwards, the data blocks were pre-processed by our method as described above. MBPLSR was then performed on the pre-processed multi-block lipidomics data set where the intervention group indicator was used as y-variable. The lipidomic profiles of the intervention groups were well separated in the global sample variation patterns detected by global score plots. The sample variation patterns for every lipid class were investigated by studying the block score plots. Several lipid classes were identified for showing a clear separation of the intervention groups. We validated the contribution of the lipid blocks to the grouping by means of the cross-validation based methods from Paper II. We also ran significance testing (from Paper V) in order to identify the lipids that were significantly contributing to the separation of the intervention groups. We studied the correlation loading plots in order to further identify the lipids that contributed to the grouping patterns. The results of analyzing this lipidomics data set by means of the methods from Paper V were biologically explained in Paper VI. The fact that the findings were biologically relevant is an indication of the reliability of our proposed methods for investigating lipidomics data sets.

6. Conclusions and Future perspectives

The focus of the present study was to establish a framework for integrating systems biology data and analyzing them in light of background knowledge and the design of the experiment. Multi-block methods based on latent variables (i.e. CPCA and MBPLSR) were employed for this purpose. These methods provide the user with powerful visualization. In this thesis tools were developed for the validation of the perception of the identified patterns. In general techniques for investigating and validating the multi-block models from different points of view were developed. All of the methods were put together and introduced as a framework for the analysis of lipidomcis data. Afterwards, the proposed framework was used for pre-processing and analyzing a multi-block lipidomics data set from a human intervention study. The results of the analysis were biologically relevant and explainable. This was an indication for the fact that the proposed tools are appropriate to be employed when dealing with data sets from systems biology. At the moment the proposed tools are available through in-house-written and standard MATLAB routines. This requires the user to have a primary knowledge about MATLAB programming which sounds fearful to many researchers in the field of biology. Therefore, developing a user friendly interface for the tools could motivate many more biologists to try these methods on their data sets.

Developing the validation tools for multi-block data analysis methods made us undertake a simulation study in order to gain a deeper understanding of the concept of DF in multi-block models. We were actually surprised when we figured out that DF's consumption affects many different areas. In Paper III we proposed a method for estimating the consumption of DF in real data sets and we implemented our method on three data sets from spectroscopy. It will be interesting to study the consumption of DF in data sets from other fields as well e.g. different types of systems biology data. We studied the consumption of DFs for CPCA and PCA models while it is also interesting to study the same concept within the PLSR framework in order to investigate the consumption of DFs both in descriptor and in response data sets.

References

1. Snoep J, Westerhoff H, Alberghina L, Westerhoff HV: **From isolation to integration, a systems biology approach for building the Silicon Cell**. In: *Systems Biology*. vol. 13: Springer Berlin / Heidelberg; 2005: 13-30.
2. Joyce AR, Palsson BO: **The model organism as a system: integrating 'omics' data sets**. *Nat Rev Mol Cell Biol* 2006, **7**(3):198-210.
3. Kohler A, Hanafi M, Bertrand D, Oust Janbu A, Møretrø T, Naderstad K, Qannari M, Martens H: **Interpreting several types of measurements in bioscience**. In: *Modern concepts in biomedical vibrational spectroscopy*. Edited by Lasch P, Kneipp J. USA: John Wiley & Sons; 2008.
4. Wenk MR: **Lipidomics: New Tools and Applications**. *Cell* 2010, **143**(6):888-895.
5. Verouden MPH, Westerhuis JA, van der Werf MtJ, Smilde AK: **Exploring the analysis of structured metabolomics data**. *Chemometrics and Intelligent Laboratory Systems* 2009, **98**(1):88-96.
6. Quinones MP, Kaddurah-Daouk R: **Metabolomics tools for identifying biomarkers for neuropsychiatric diseases**. *Neurobiology of Disease* 2009, **35**(2):165-176.
7. Verhoeckx KCM, Bijlsma S, Jespersen S, Ramaker R, Verheij ER, Witkamp RF, van der Greef J, Rodenburg RJT: **Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis**. *International Immunopharmacology* 2004, **4**(12):1499-1514.
8. Thissen U, Coulier L, Overkamp KM, Jetten J, van der Werff BJC, van de Ven T, van der Werf MtJ: **A proper metabolomics strategy supports efficient food quality improvement: A case study on tomato sensory properties**. *Food Quality and Preference* 2011, **22**(6):499-506.
9. Mercier C, Truntzer C, Pecqueur D, Gimeno J-P, Belz G, Roy P: **Mixed-model of ANOVA for measurement reproducibility in proteomics**. *Journal of Proteomics* 2009, **72**(6):974-981.
10. Álvarez-Chaver P, Rodríguez-Piñeiro AM, Rodríguez-Berrocal FJ, García--Lorenzo A, Páez de la Cadena M, Martínez-Zorzano VS: **Selection of putative colorectal cancer markers by applying PCA on the soluble proteome of tumors: NDK A as a promising candidate**. *Journal of Proteomics* 2011, **74**(6):874-886.
11. Pieragostino D, Petrucci F, Del Boccio P, Mantini D, Lugaresi A, Tiberio S, Onofrj M, Gambi D, Sacchetta P, Di Ilio C *et al*: **Pre-analytical factors in clinical proteomics investigations: Impact of ex vivo protein modifications for multiple sclerosis biomarker discovery**. *Journal of Proteomics* 2010, **73**(3):579-592.
12. Park S, Ku YK, Seo MJ, Kim DY, Yeon JE, Lee KM, Jeong S-C, Yoon WK, Harn CH, Kim HM: **Principal component analysis and discriminant analysis (PCA-DA) for discriminating profiles of terminal restriction fragment length polymorphism (T-RFLP) in soil bacterial communities**. *Soil Biology and Biochemistry* 2006, **38**(8):2344-2349.
13. Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Osborne CK, Fuqua SAW: **Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance**. *Journal of the National Cancer Institute* 1999, **91**(5):453-459.

14. Færgestad EM, Langsrud Ø, Høy M, Hollung K, Sæbø¸ S, Liland KH, Kohler A, Gidskehaug L, Almergren J, Anderssen E *et al*: **4.08 - Analysis of Megavariate Data in Functional Genomics**. In: *Comprehensive Chemometrics.* Edited by Brown SD, Tauler R, Walczak B. Oxford: Elsevier; 2009: 221-278.

15. Yang S, Qiao B, Lu S-H, Yuan Y-J: **Comparative lipidomics analysis of cellular development and apoptosis in two Taxus cell lines**. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2007, **1771**(5):600-612.

16. Balogh G, Péter M, Liebisch G, HorvÃ¡th I, Török Z, Nagy E, Maslyanko A, Benkő S, Schmitz G, Harwood JL *et al*: **Lipidomics reveals membrane lipid remodelling and release of potential lipid mediators during early stress responses in a murine melanoma cell line**. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2010, **1801**(9):1036-1047.

17. Xu J, Cai S, Li X, Dong J, Ding J, Chen Z: **Statistical two-dimensional correlation spectroscopy of urine and serum from metabolomics data**. *Chemometrics and Intelligent Laboratory Systems* 2012, **112**(0):33-40.

18. Roux A, Lison D, Junot C, Heilier J-F: **Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review**. *Clinical Biochemistry* 2010, **44**(1):119-135.

19. Chuang H-L, Huang Y-T, Chiu C-C, Liao C-D, Hsu F-L, Huang C-C, Hou C-C: **Metabolomics characterization of energy metabolism reveals glycogen accumulation in gut-microbiota-lacking mice**. *The Journal of Nutritional Biochemistry* 2011(0).

20. Gottfries J, Sjögren M, Holmberg B, Rosengren L, Davidsson P, Blennow K: **Proteomics for drug target discovery**. *Chemometrics and Intelligent Laboratory Systems* 2004, **73**(1):47-53.

21. Smit S, Hoefsloot HCJ, Smilde AK: **Statistical data processing in clinical proteomics**. *Journal of Chromatography B* 2008, **866**(1-2):77-88.

22. Gavaghan CL, Wilson ID, Nicholson JK: **Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA**. *FEBS Letters* 2002, **530**(1-3):191-196.

23. Simoh S, Linthorst HJM, Lefeber AWM, Erkelens C, Kim HK, Choi YH, Verpoorte R: **Metabolic changes of Brassica rapa transformed with a bacterial isochorismate synthase gene**. *Journal of Plant Physiology* 2010, **167**(18):1525-1532.

24. Yetukuri L, Katajamaa M, Medina-Gomez G, Seppanen-Laakso T, Vidal-Puig A, Oresic M: **Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis**. *BMC Systems Biology* 2007, **1**(1):12.

25. Del Boccio P, Pieragostino D, Di Ioia M, Petrucci F, Lugaresi A, De Luca G, Gambi D, Onofrj M, Di Ilio C, Sacchetta P *et al*: **Lipidomic investigations for the characterization of circulating serum lipids in multiple sclerosis**. *Journal of Proteomics* 2011, **74**(12):2826-2836.

26. Wold S, Hellberg S, Lundstedt Y, Sjostrom M, Wold H: In: *Proc Symp on PLS Model Building: Theory and Application: 1987; Frankfurt am Main.*; 1987.

27. Wangen LE, Kowalski BR: **A multiblock partial least squares algorithm for investigating complex chemical systems**. *Journal of Chemometrics* 1989, **3**(1):3-20.

28. Richards SE, Dumas M-E, Fonville JM, Ebbels TMD, Holmes E, Nicholson JK: **Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework**. *Chemometrics and Intelligent Laboratory Systems* 2010, **104**(1):121-131.

29. Brás LP, Lopes JA, Santos CR, Cardoso JP, Menezes JC, Barbosa-Póvoa A, Matos H: **Modelling and identification of individual stage contributions in an industrial pharmaceutical process by multiblock PLS**. In: *Computer Aided Chemical Engineering.* vol. Volume 18: Elsevier; 2004: 601-606.
30. Qin SJ, Valle S, Piovoso MJ: **On unifying multiblock analysis with application to decentralized process monitoring**. *J Chemometrics* 2001, **15**:715-742.
31. Westerhuis JA, Smilde AK: **Deflation in multiblock PLS**. *J Chemometrics* 2001, **15**:485-493.
32. Westad F, Hersleth M, Lea P, Martens H: **Variable selection in PCA in sensory descriptive and consumer data**. *Food Quality and Preference* 2003, **14**(5-6):463-472.
33. Gidskehaug L, Anderssen E, Alsberg BK: **Cross model validation and optimisation of bilinear regression models**. *Chemometrics and Intelligent Laboratory Systems* 2008, **93**(1):1-10.
34. Westerhuis J, Hoefsloot H, Smit S, Vis D, Smilde A, van Velzen E, van Duijnhoven J, van Dorsten F: **Assessment of PLSDA cross validation**. *Metabolomics* 2008, **4**(1):81-89.
35. Kotsiantis SB, Kanellopoulos D, Pintelas PE: **Data Preprocessing for Supervised Leaning**. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE* 2006, **1**(1):111-117.
36. Yetukuri LR: **Bioinformatics approaches for the analysis of lipidomics data**. *PhD dissertation.* The Aalto University School of Science and Technology; 2010.
37. Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CRH, Shimizu T, Spener F, van Meer G, Wakelam MJO, Dennis EA: **Update of the LIPID MAPS comprehensive classification system for lipids**. *Journal of Lipid Research* 2009, **50**(Supplement):S9-S14.
38. Faergestad EM, Langsrud Ø, Høy M, Hollung K, Sæbø¸ S, Liland KH, Kohler A, Gidskehaug L, Almergren J, Anderssen E *et al*: **4.08 - Analysis of Megavariate Data in Functional Genomics**. In: *Comprehensive Chemometrics.* Edited by Stephen DB, RomÃ T, Beata W. Oxford: Elsevier; 2009: 221-278.
39. Goodpaster AM, Kennedy MA: **Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies**. *Chemometrics and Intelligent Laboratory Systems* 2011, **109**(2):162-170.
40. Kristian Hovde L: **Multivariate methods in metabolomics â€“ from pre-processing to dimension reduction and statistical analysis**. *TrAC Trends in Analytical Chemistry* 2011, **30**(6):827-841.
41. Bare JC, Koide T, Reiss D, Tenenbaum D, Baliga N: **Integration and visualization of systems biology data in context of the genome**. *BMC Bioinformatics* 2010, **11**(1):382.
42. Sullivan DE, Gabbard JL, Shukla M, Sobral B: **Data Integration for Dynamic and Sustainable Systems Biology Resources: Challenges and Lessons Learned**. *Chemistry & Biodiversity* 2010, **7**(5):1124-1141.
43. Conesa A, Prats-MontalbÃ¡n JM, Tarazona S, Nueda MJ, Ferrer A: **A multiway approach to data integration in systems biology based on Tucker3 and N-PLS**. *Chemometrics and Intelligent Laboratory Systems* 2010, **104**(1):101-111.
44. Goesmann A, Linke B, Rupp O, Krause L, Bartels D, Dondrup M, McHardy AC, Wilke A, Pühler A, Meyer F: **Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology**. *Journal of Biotechnology* 2003, **106**(2-3):157-167.

45. van den Berg RA, Rubingh CM, Westerhuis JA, van der Werf MtJ, Smilde AK: **Metabolomics data exploration guided by prior knowledge**. *Analytica Chimica Acta* 2009, **651**(2):173-181.

46. Hanafi M, Kohler A, Qannari E-M: **Connections between multiple co-inertia analysis and consensus principal component analysis**. *Chemometrics and Intelligent Laboratory Systems* 2010, **106**(1):37-40.

47. Gerlach RW, Kowalski BR, Wold HOA: **Partial least-squares path modelling with latent variables**. *Analytica Chimica Acta* 1979, **112**(4):417-421.

48. Frank IE, Kowalski BR: **prediction of wine quality and geographic origin from chemical measurements by parital least-squares regression modeling**. *Analytica Chimica Acta* 1984, **162**:241-251.

49. Frank I, Feikema J, Constantine N, Kowalski B: **Prediction of Product Quality from Spectral Data Using the Partial Least-Squares Method**. *Journal of Chemical Information and Computer Sciences* 1984, **24**(1):20-24.

50. Frank IE, Kowalski BR: **A multivariate method for relating groups of measurements connected by a causal pathway**. *Analytica Chimica Acta* 1985, **167**:51-63.

51. Westerhuis JA, Coenegracht PMJ: **Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares**. *Journal of Chemometrics* 1997, **11**(5):379-392.

52. **Efron B**: **The jackknife, the bootstrap, and other resampling plans**. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs* 1982.

53. Good P: **Permutation Tests**. New york: Springer; 1994.

54. Efron B, Tibshirani R: **An Introduction to the Bootstrap**: Chapman and Hall/CRC; 1994.

55. Stone M: **Cross-Validatory Choice and Assessment of Statistical Predictions**. *Journal of the Royal Statistical Society Series B (Methodological)* 1974, **36**(2):111-147.

56. Quenouille MH: **Notes on bias in estimation.** *Biometrika* 1956, **43**:353-360.

57. Tukey JW: **Bias and confidence in not-quite large samples.** *Annals of Mathematical Statistics and Computing* 1958, **29**.

58. Dwass M: **Modified Randomization Tests for Nonparametric Hypotheses**. *The Annals of Mathematical Statistics* 1957, **28**:181-187.

59. Nichols TE, Holmes AP: **Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples**. *Human Brain Mapping* 2001, **15**(1):1-25.

60. Pennestri E, Cavacece M, Vita L: **On the Computation of Degrees-of-Freedom: A Didactic Perspective**. *ASME Conference Proceedings* 2005, **2005**(47438):1733-1741.

61. Jaccard J, Becker MA: **Statistics for the behavioral sciences.**, 2nd edn. Belmont, CA: Wadsworth.; 1990.

62. Yu CH, Lovric M: **Degrees of Freedom, International Encyclopedia of Statistical Science**. In.: Springer Berlin Heidelberg; 2011: 363-365.

63. Martens H, Næs T: **Multivariate calibration**. Chichester: Wiley & Sons; 1989.

64. van der Voet H: **Pseudo-degrees of freedom for complex predictive models: the example of partial least squares**. In., vol. 13: John Wiley & Sons, Ltd.; 1999: 195-208.

65. Krämer N, Sugiyama M: **The Degrees of Freedom of Partial Least Squares Regression**. *Journal of the American Statistical Association* 2011, **106**(494):697-705.

66. Kato K: **On the degrees of freedom in shrinkage estimation**. *Journal of Multivariate Analysis* 2009, **100**(7):1338-1352.

67. Krämer N, Braun ML: **Kernelizing PLS, degrees of freedom, and efficient model selection**. In: *Proceedings of the 24th international conference on Machine learning.* Corvalis, Oregon: ACM; 2007.

68. Faber NM: **Degrees of freedom for the residuals of a principal component analysis -- A clarification**. *Chemometrics and Intelligent Laboratory Systems* 2008, **93**(1):80-86.

69. Martens H, Martens M: **Multivariate Analysis of Quality: An Introduction**. Chichester, UK: Wiley; 2001.

Paper I

Chemometrics and Intelligent Laboratory Systems 104 (2010) 140–153

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods

Sahar Hassani [a,c,*], Harald Martens [a,b,c], El Mostafa Qannari [d], Mohamed Hanafi [d], Grethe Iren Borge [a], Achim Kohler [a]

[a] Nofima Mat AS, Centre for Biospectroscopy and Data Modelling, Osloveien 1, N-1430 Ås, Norway
[b] CIGENE – Center for Integrative Genetics, University of Life Sciences, 1432 Aas, Norway
[c] Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway
[d] ONIRIS, Unité de Sensométrie et de Chimiométrie, site de la Géraudière, BP 82225, 44322 Nantes Cedex 3, France

ARTICLE INFO

Article history:
Received 12 October 2009
Received in revised form 9 August 2010
Accepted 12 August 2010
Available online 26 August 2010

Keywords:
Omics data
Multi-block methods
Validation tools

ABSTRACT

As systems biology develops, various types of high-throughput -omics data become rapidly available. An increasing challenge is to analyze such massive data, interpret the results and validate the findings. Data analysis for most of the omics-techniques is in a fledgling immature stage. Alone the dimensionality of the data tables calls for new ways to reveal structure in the data, without cognitive overflow and excessive false discovery rate. Multi-block methods have been developed and adapted in order to find common variation patterns in data and depict these findings on graphical displays while providing tools to enhance the interpretation of the outcomes. In particular, multi-block methods based on latent variables are powerful tools to study block and global variation patterns, e.g. by inspecting block and global score plots. These methods can be used to achieve a graphical overview over sample and variable variation patterns in an efficient way. However, a visual detection of patterns may be subjective and, therefore, there is a need for validation tools. In this paper tools for validation of visually identified patterns in multi-block results are presented. Cross-validated estimates of Root Mean Square Error (RMSE) for block results are introduced for estimating the number of relevant PCs of the Consensus Principal Component Analysis (CPCA) models. Furthermore, important variables are identified by approximate *t*-tests based on Procrustes-corrected jackknifing. For the assessment of the stability of score patterns, block stability plots are introduced. Outliers can be revealed graphically on block and global level by stability plots.

1. Introduction

Systems biology is a new biological research field where the interaction between different biological levels is studied by different-omics techniques in different scientific disciplines such as nutrigenomics and nutrigenetics, genomics, proteomics, metabonomics and metabolomics. Genomics is increasingly being used in a variety of health applications including pharmaceutical companies, healthcare industry, animal research studies and the production of livestock and crops [1–4]. Proteomics is slowly altering the biomarker discovery methods in the field of medicine while having significant applications in basic and applied biology [5–8]. An extension of genomics and proteomics leads to metabonomics which studies the metabolic responses to diets, drugs and diseases [9]. Metabolomics is a newborn science which offers a unique opportunity to study genotype-phenotype as well as genotype–environtype relationships. There are many and diverse applications of metabolomics in drug trials, toxicology, transplant monitoring and pathway discovery [10–12]. Systems biology has gained in importance in food science and food industry due to an increasing focus on food for better health. Food industry and food science in collaboration with nutrition experts are applying human dietary intervention and cohort studies in order to test the effects of foods on human health [1,3,13–16].

As systems biology develops, high-throughput techniques generating huge amounts of -omics data become rapidly available: e.g., Amplified Fragment Length Polymorphism (AFLP) is a powerful DNA fingerprinting technique for DNAs of any origin which is also highly sensitive and reproducible. It has been used for identifying the genetic variation in strains or closely related species of plants, fungi, animals, and bacteria [17]. Measurement principles for proteomics based on 2-dimensional electrophoresis are well established, but analysis of the data from these experiments still remains a challenge [18,19]. Mass spectrometry (MS), Nuclear magnetic resonance (NMR) and Fourier transform infrared (FTIR) spectroscopy are analytical techniques used in many metabolomics studies resulting in comprehensive and

[*] Corresponding author. Nofima Mat AS, Centre for Biospectroscopy and Data Modelling, Osloveien 1, N-1430 Ås, Norway.
E-mail address: sahar.hassani@nofima.no (S. Hassani).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.08.008

quantitative analysis of wide arrays of metabolites in biological samples [20–24].

The design of -omics experiments poses new challenges, since the obtained data is multi-response data: From a chosen set of combinations of experimental design parameters, genomics-, proteomics-, metabonomics- and metabolomics-data are generated, together with target parameters such as clinical data in intervention studies, phenotypical data and food quality parameters [25]. Thus, the design of multi-response experiments is an important aspect of further multi-matrix research not only with respect to the design factors but also with respect to the chosen variables and measurement techniques [26]. The effects of different designs factors on response variables are of central interest in many research studies [27–29]. Selecting an appropriate design optimizes the number of samples, variables and measurement techniques and minimizes the noise and experimental error. Different measurement principles have different error structures and so they require a different number of samples, biological and technical replicates. The incorporation of replicates is an approach that ensures greater experimental success. Replication increases the statistical power and subsequently the confidence of the conclusions drawn from the study.

Structuring of -omics data in appropriate databases and integration of the large amounts of data is gaining in importance [30]. An increasing challenge is to analyze such massive data, interpret the results and make the findings reproducible [31]. Data modelling techniques for the analysis of, e.g. genomics and transcriptomics data have been emerging during the recent years [32,33]. But still, the cross-disciplinary integration of different -omics measurements within one common data modelling approach is at its earliest beginning [34–37].

Omics experiments lead in general to multiple data matrices or data blocks, where each block refers to data from one measurement principle. When the measurements are performed on the same set of samples, the data blocks can be ordered to form a *multi-block data set* where the same rows in every block refer to the same sample, i.e. we obtain a row to row correspondence. Mathematical aspects of data analysis methods for multi-block data have been in focus of intensive research during the last three decades [38–47]. Multi-block methods have been applied in several disciplines [44,46], their use in the field of systems biology is rather new: Multi-block methods have been developed and adapted in order to find common variation patterns in data in functional genomics [45,47]. These methods can be used to achieve a graphical overview over sample and variable variation patterns within and between blocks of variables and/or sets of samples in an efficient way [47]. In general there is a lack of statistical validation methodology for multi-block methods. Moreover, concepts for variable selection need to be transferred to multi-block methods. Due to the complexity of the systems biology data, strategies for graphical visualisation and validation need to be developed and made available for the user.

In this paper, a data analysis strategy for a multi-response experiment is presented and applied to a data set obtained in a study that aimed to characterize natural variability in microbiology. Graphical representation and interpretation are presented to help the user to discover interactions and common structures in complex datasets. In order to enable the user to validate visually detected patterns when studying global and block results in multi-block analysis, new methods for the validation of block results are introduced. These validation tools are introduced for Consensus Principal Component Analysis (CPCA), but since they represent general concepts, they could be easily transferred to other multi-block methods. The paper is organised in the following way: After an introduction into the multi-block analysis of -omics data in Section 2.1, the NIPALS algorithm for CPCA is introduced in Section 2.2. This is done in order to introduce block and global parameters of CPCA that are used for visualisation. In Section 2.3 the calculation of block Root Mean Squared Errors (RMSE) is introduced for the validation of block patterns. In Section 2.4 the calculation of uncertainties by cross-validation is explained. The cross-validated loadings obtained in Section 2.4 are used in Section 2.5 for the calculation of cross-validated block loadings and block stability plots. In Section 3, a multi-block example is presented, different techniques for visualisation are explained and the new validation tools are illustrated by an example. In Section 4 we finish by a conclusion.

2. Theory

2.1. Notation

We follow the notation commonly used in chemometrics, e.g. Martens & Martens in [40]: Matrices and vectors are written as boldface, matrices as upper-case letters and vectors as lower-case letters. By the indices $b=1,...,B$ we denote blocks of variables, by $m=1,...,M$ cross-validation segments of samples and by $a=1,...,A$ the number of principal components. The total number of samples in each data set is represented by N, the total number of variables by K and the total number of variables in a given block b by K_b. By $\mathbf{X}=\left[\mathbf{X}^1,\mathbf{X}^2,...,\mathbf{X}^B\right]$ we denote the multi-block data set consisting of B blocks. Measurements belonging to the same measurement technique, e.g. AFLP, NMR, GC–MS, proteomics, FTIR and phenotypes are typically collected in the same block $\mathbf{X}^b$. In omics experiments different measurement techniques are applied to the same samples and for the multi-block analysis data need to be ordered in a way that a sample-to-sample (row-to-row) correspondence between the blocks is achieved. Only measurements originating from the same biological replicate can be related to each other by a row-to-row correspondence. If, for example, for different methods several and different biological replicates are used, only means of biological replicates can be related to each other.

2.2. CPCA

In order to find the common underlying patterns between data blocks $\mathbf{X}^b$ in a multi-block data set, Consensus Principal Component Analysis (CPCA) has been used [48]. However, the graphical and validation tools discussed herein are generic and can be easily adapted to other multi-block methods. In CPCA, principal components or latent variables describe variation patterns within and between the data blocks. Variation patterns shared by several blocks can thus be detected. The CPCA algorithm consists of two steps: (1) The first global score and loading vector, block score and loading vectors are calculated for the multi-block data set $\mathbf{X}$ where each block has been preprocessed to remove irrelevant variation types (if possible) and scaled so as to balance the various blocks. (2) In a deflation step $\mathbf{X}$ is updated by subtracting the variation that corresponds to the first global score vector. For the calculation of the second component, the procedure is repeated on the deflated matrix and so on. For the calculation of the global score and loading vectors, block score and loading vectors an iterative procedure is used, the NIPALS algorithm for CPCA [49,50]. The NIPALS algorithm is given below, where the notation according to Westerhuis et al. (1998) [45] is used.

2.2.1. Preprocessing

2.2.1.1. Mean-centering. All variables are usually mean-centered prior to CPCA. Mean-centering is done by subtracting the mean according to

$$\mathbf{X}_{\text{Unscaled}}=\mathbf{X}_{\text{input}}-\mathbf{ONES}\cdot\overline{\mathbf{x}}'_{\text{input}} \tag{1}$$

where $\mathbf{X}_{\text{Unscaled}}$ is the centered variables, $\mathbf{X}_{\text{input}}$ is the original data, $\mathbf{ONES}$ is the $N\times1$ vector whose components are equal to one and $\overline{\mathbf{x}}_{\text{input}}$, a $K\times1$ vector, is the mean along the samples of the data matrix $\mathbf{X}_{\text{input}}$.

2.2.1.2. Scaling. In CPCA the data blocks are scaled by dividing each block by its norm. Scaling is performed for mean-centered variables according to

$$\mathbf{X}^b = \frac{\mathbf{X}^b_{\mathrm{Unscaled}}}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{K_b}\left(\mathbf{X}^b_{\mathrm{Unscaled}}(i,j)\right)^2}} \tag{2}$$

where $\mathbf{X}^b$ is the data block after centering and scaling, $\mathbf{X}^b_{\mathrm{Unscaled}}$ is the mean-centered, un-scaled data block, $\mathbf{X}^b_{\mathrm{Unscaled}}(i,j)$ is the ith, jth entry of data block $\mathbf{X}^b_{\mathrm{Unscaled}}$. This scaling sets all the blocks on the same footing (i.e. same total variance) and is usually recommended because un-scaled data would give the blocks different influence, since they differ from each other with respect to their numbers of variables and the measurement units. In some instances the user may want to give a certain block a different influence than the other blocks, either because the user wants to let one or several blocks dictating the variation pattern or because the user wants to avoid that certain blocks influence the variation pattern for example it is an appropriate approach to “passify” the design block by down-weighting it with a very small number e.g. 0.000001, in order to minimize its influence on the CPCA model. The advantage of this particular scaling is that although the design is not significantly contributing to the model, the relation of design factors and measured variables can still be investigated. Individual variables within the blocks may be similarly “passified” by drastic down-scaling (Martens & Martens in [40]), but that is not employed herein.

2.2.2. Overall modelling

CPCA is used to explore the systematic variation pattern in $\mathbf{X}$. Data table $\mathbf{X}$ was modeled as sum of A relevant principal components (PCs) plus a residual matrix $\mathbf{E}$. The CPCA model for $\mathbf{X}$ is given in Eq. (3):

$$\begin{aligned}\mathbf{X} &= \mathbf{TP'} + \mathbf{E}\\ \mathbf{X}^b &= \mathbf{T}^b\mathbf{P}^{b\,\prime} + \mathbf{E}^b\end{aligned} \tag{3}$$

where $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_a, \ldots, \mathbf{t}_A]$ contains A global score vectors $\mathbf{t}_a$, and $\mathbf{P}$ is the corresponding matrix of global loading vectors $\mathbf{p}_a$, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_a, \ldots, \mathbf{p}_A]$. The global loading matrix $\mathbf{P}$ can also be written as the matrix of concatenated block loading matrices $\mathbf{P}^b$: $\mathbf{P}' = \left[\mathbf{P}^{1\prime}, \mathbf{P}^{2\prime}, \ldots, \mathbf{P}^{b\prime}, \ldots, \mathbf{P}^{B\prime}\right]$.

2.2.3. Component estimation

Several equivalent CPCA estimation algorithms are proposed in the literature. The original algorithm called NIPALS is the most popular and presents the advantage of explicitly showing how to compute in addition to the global scores, the block loadings and the block scores. NIPALS algorithm runs as follows. For each component $a = 1,2,\ldots$:

A. Initialization
   1.1 Choose an arbitrary starting global score vector, $\mathbf{t}$

B. Computation of block scores and block loadings
   1.2 $\tilde{\mathbf{p}}^b = \frac{\mathbf{X}^{b\prime}\mathbf{t}}{\mathbf{t}'\mathbf{t}}$ Preliminary block loadings
   1.3 $\mathbf{p}^b = \frac{\tilde{\mathbf{p}}^b}{\|\tilde{\mathbf{p}}^b\|}$ Block loadings, scaled to length 1 in each block
   1.4 $\mathbf{t}^b = \mathbf{X}^b\mathbf{p}^b$ Block scores

C. Computation of global scores and global loadings
   1.5 $\mathbf{T} = \left[\mathbf{t}^1\ \mathbf{t}^2 \ldots \mathbf{t}^B\right]$
   1.6 $\mathbf{w} = \mathbf{T}'\mathbf{t}$ Block weights
   1.7 Normalize $\mathbf{w}$ to $\|\mathbf{w}\| = 1$
   1.8 $\mathbf{t} = \mathbf{Tw}$ Global scores

D. Replace the starting score vector $\mathbf{t}$ by the updated vector of global scores in 1.8 and iterate until convergence of the algorithm (i.e. no significant change in $\mathbf{t}$).

Alternatively, the same solution for global scores and loadings may be obtained by performing PCA on $\mathbf{X}$. Thereafter, block scores and block loadings may be computed according to their definition in the NIPALS algorithm above.

2.3. Error estimation, cross-validation and jack-knifing

In order to estimate the stability of the CPCA model and the number of relevant PCs, RMSE is calculated by cross-validation [51] for $a = 1,\ldots,A$ components, where A is chosen sufficiently large. The set of $m = 1,\ldots,M$ segments of data consists of one or several samples which are left out, in turn, resulting in a left-out segment of data ($\mathbf{X}_m$) and a leave-in segment of data ($\mathbf{X}_{-m}$). CPCA models are determined for the leave-in segment matrices $\mathbf{X}_{-m}$ in a cross-validation procedure. Thereafter, the model is fitted to the left-out samples leading to the predicted matrix $\hat{\mathbf{X}}_m$. The residual matrix $\mathbf{E}_m$ is defined as the difference between the predicted values for the left-out samples, $\hat{\mathbf{X}}_m$, and the data in $\mathbf{X}_m$. This procedure is repeated over all segments in the data set resulting in A residual matrices $\mathbf{E}_a$. The above mentioned steps are visualized in Fig. 1.

In order to predict $\hat{\mathbf{X}}_m$, $\mathbf{P}_{-m}$ is calculated from the CPCA model of $\mathbf{X}_{\mathbf{C}_{-m}}$ which represents the mean-centered “leave-in” samples. Mean-centering is performed on $\mathbf{X}_{-m}$ according to Eq. (4):

$$\mathbf{X}_{\mathbf{C}_{-m}} = \mathbf{X}_{-m} - \mathbf{ONES} \cdot \overline{\mathbf{x}}'_{-m} \tag{4}$$

where $\mathbf{X}_{-m} = \left[\mathbf{X}^1_{-m}, \mathbf{X}^2_{-m}, \ldots, \mathbf{X}^b_{-m}, \ldots, \mathbf{X}^B_{-m}\right]$ contains leave-in samples and $\overline{\mathbf{x}}_{-m}$, a $K \times 1$ vector, is the mean along samples of $\mathbf{X}_{-m}$.

The CPCA model of $\mathbf{X}_{\mathbf{C}_{-m}}$ is given in Eq. (5):

$$\mathbf{X}_{\mathbf{C}_{-m}} = \mathbf{T}_{-m}\mathbf{P}'_{-m} \tag{5}$$

where $\mathbf{T}_{-m} = \left[\mathbf{t}_{-m,1}, \mathbf{t}_{-m,2}, \ldots, \mathbf{t}_{-m,a}, \ldots, \mathbf{t}_{-m,A}\right]$ contains A global scores calculated for leave-in samples, $\mathbf{P}_{-m}$ is the matrix of concatenated block loadings for leave-in samples $\mathbf{P}'_{-m} = \left[\mathbf{P}^{1\prime}_{-m,A}, \mathbf{P}^{2\prime}_{-m,A}, \ldots, \mathbf{P}^{b\prime}_{-m,A}, \ldots, \mathbf{P}^{B\prime}_{-m,A}\right]$ and $\mathbf{X}_{\mathbf{C}_{-m}} = \left[\mathbf{X}^1_{\mathbf{C}_{-m}}, \mathbf{X}^2_{\mathbf{C}_{-m}}, \ldots, \mathbf{X}^b_{\mathbf{C}_{-m}}, \ldots, \mathbf{X}^{\mathrm{B}}_{\mathbf{C}_{-m}}\right]$.

The set of A score values for each of the left-out samples is unknown and is determined as follows. Firstly, the leave-out samples are centered using the means estimated from the leave-in samples:

$$\mathbf{X}_{\mathbf{C}_m} = \mathbf{X}_m - \mathbf{ONES} \cdot \overline{\mathbf{x}}'_{-m} \tag{6}$$

Secondly, the matrix of loadings $\mathbf{P}_{-m}$ estimated from the leave-in samples are applied to $\mathbf{X}_{\mathbf{C}_m}$ leading to the score matrix for the leave-out samples:

$$\hat{\mathbf{T}}_m = \mathbf{X}_{\mathbf{C}_m}\mathbf{P}_{-m} \tag{7}$$

Multiplying the predicted scores for the left-out segments from Eq. (7) with the loadings estimated according to Eq. (5) results in a prediction for the left-out segment data. In order to choose the appropriate number of PCs, left-out samples were estimated for various values of $A = 0, 1, \ldots, A_{\max}$ ($A_{\max}$ is considered sufficiently large e.g. the total number of variables) and resulted in different estimations and residual matrices. Residual matrices are calculated as

$$\begin{aligned}\hat{\mathbf{X}}_{m,A} &= \hat{\mathbf{T}}_{m,A}\mathbf{P'}_{-m,A}\\ \mathbf{E}_{m,A} &= \mathbf{X}_{\mathbf{C}_m} - \hat{\mathbf{X}}_{m,A}\\ \mathbf{E}_{m,0} &= \mathbf{X}_{\mathbf{C}_m}\end{aligned} \tag{8}$$

where $\hat{\mathbf{T}}_{m,A} = \left[\hat{\mathbf{t}}_{m,1}, \hat{\mathbf{t}}_{m,2}, \ldots, \hat{\mathbf{t}}_{m,a}, \ldots, \hat{\mathbf{t}}_{m,A}\right]$ contains A global score vectors calculated for left-out samples according to Eq. (7), $\mathbf{P'}_{-m,A} = \left[\mathbf{P}^{1\prime}_{-m,A}, \mathbf{P}^{2\prime}_{-m,A}, \ldots, \mathbf{P}^{b\prime}_{-m,A}, \ldots, \mathbf{P}^{B\prime}_{-m,A}\right]$ is the matrix of concatenated block loadings for leave-in samples, $\hat{\mathbf{X}}_{m,A}$ is the prediction of the

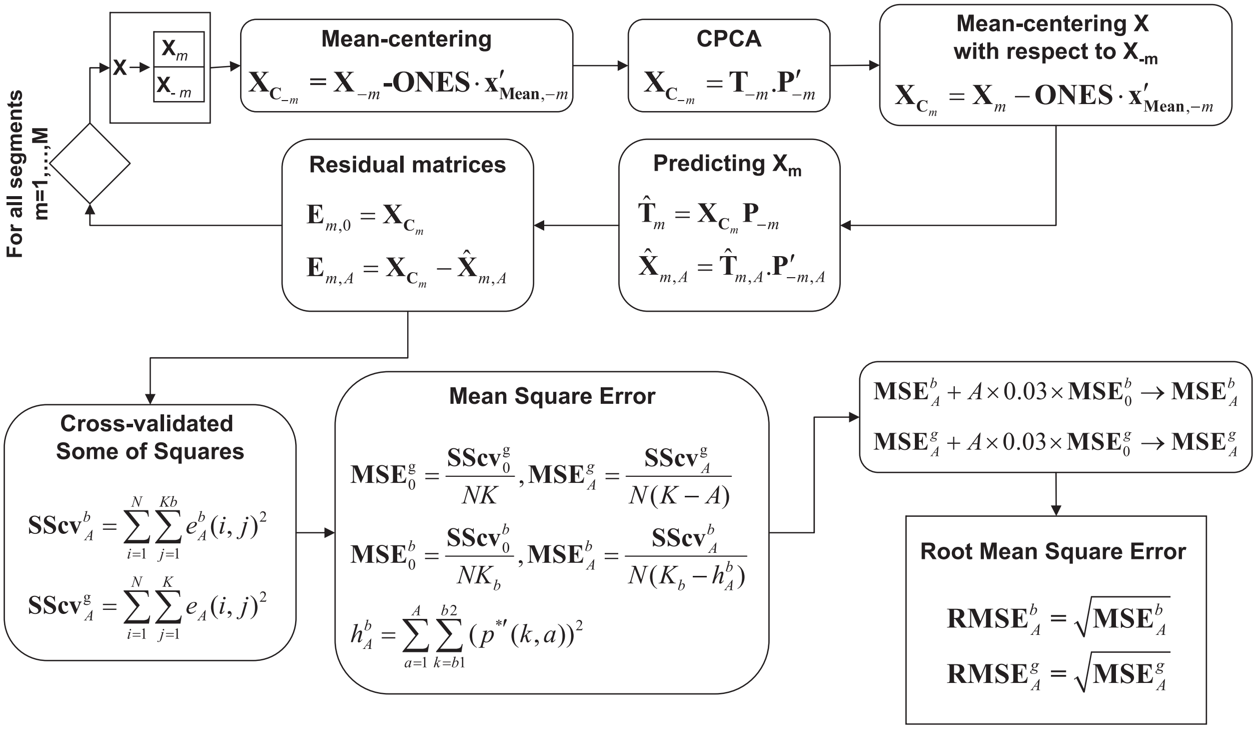

For all segments m=1,...,M
X
$\mathbf{X}_m$
$\mathbf{X}_{-m}$
Mean-centering
$\mathbf{X}_{\mathbf{C}_{-m}} = \mathbf{X}_{-m}\text{-}\mathbf{ONES}\cdot\mathbf{x}'_{\mathbf{Mean},-m}$
CPCA
$\mathbf{X}_{\mathbf{C}_{-m}} = \mathbf{T}_{-m}.\mathbf{P}'_{-m}$
Mean-centering X with respect to X-m
$\mathbf{X}_{\mathbf{C}_m} = \mathbf{X}_m - \mathbf{ONES}\cdot\mathbf{x}'_{\mathbf{Mean},-m}$
Residual matrices
$\mathbf{E}_{m,0} = \mathbf{X}_{\mathbf{C}_m}$
$\mathbf{E}_{m,A} = \mathbf{X}_{\mathbf{C}_m} - \hat{\mathbf{X}}_{m,A}$
Predicting $\mathbf{X}_m$
$\hat{\mathbf{T}}_m = \mathbf{X}_{\mathbf{C}_m}\mathbf{P}_{-m}$
$\hat{\mathbf{X}}_{m,A} = \hat{\mathbf{T}}_{m,A}.\mathbf{P}'_{-m,A}$
Cross-validated Some of Squares
$\mathbf{SScv}^b_A = \sum_{i=1}^{N}\sum_{j=1}^{Kb} e^b_A(i,j)^2$
$\mathbf{SScv}^g_A = \sum_{i=1}^{N}\sum_{j=1}^{K} e_A(i,j)^2$
Mean Square Error
$\mathbf{MSE}^g_0 = \frac{\mathbf{SScv}^g_0}{NK}, \mathbf{MSE}^g_A = \frac{\mathbf{SScv}^g_A}{N(K-A)}$
$\mathbf{MSE}^b_0 = \frac{\mathbf{SScv}^b_0}{NK_b}, \mathbf{MSE}^b_A = \frac{\mathbf{SScv}^b_A}{N(K_b-h^b_A)}$
$h^b_A = \sum_{a=1}^{A}\sum_{k=b1}^{b2}(p^{*\prime}(k,a))^2$
$\mathbf{MSE}^b_A + A\times 0.03\times\mathbf{MSE}^b_0 \rightarrow \mathbf{MSE}^b_A$
$\mathbf{MSE}^g_A + A\times 0.03\times\mathbf{MSE}^g_0 \rightarrow \mathbf{MSE}^g_A$
Root Mean Square Error
$\mathbf{RMSE}^b_A = \sqrt{\mathbf{MSE}^b_A}$
$\mathbf{RMSE}^g_A = \sqrt{\mathbf{MSE}^g_A}$

Fig. 1. Flow chart of the RMSE calculation. Square boxes are input and output, diamond is for loop and rounded boxes are instructions. The following indices are used: (*i,j*) for the *i*th, *j*th entry of the respective matrix, K_b for the total number of variables in block *b*, *K* for the total number of variables in the data set, *N* for the total number of samples in the data set, $m = 1,...,M$ for the cross-validation segments of samples, *A* for the number of principal components, *b*1 for the column number of first variable in block *b* and *b*2 for the column number of the last variable in block *b*. '*g*' stands for the global.

left-out samples based on *A* PCs, $\mathbf{X}_{\mathbf{C}_m}$ is segment *m* of the data table which was calculated by Eq. (6) and $\mathbf{E}_{m,A} = \left[\mathbf{E}^1_{m,A}, \mathbf{E}^2_{m,A}, ..., \mathbf{E}^b_{m,A}, ..., \mathbf{E}^B_{m,A}\right]$ is the corresponding residual matrix of concatenated block residual matrices $\mathbf{E}^b_{m,A}$ for segment *m* using *A* PCs.

Thereafter, the residual matrices $\mathbf{E}_{m,A}$ were concatenated vertically for all segments, which resulted in one residual matrix $\mathbf{E}_A = \left[\mathbf{E}^1_A, \mathbf{E}^2_A, ..., \mathbf{E}^b_A, ..., \mathbf{E}^B_A\right]$ for every different value of *A*. In the same way, the residual matrices $\mathbf{E}_{m,0}$ were concatenated vertically for all segments, which resulted in one residual matrix $\mathbf{E}_0 = \left[\mathbf{E}^1_0, \mathbf{E}^2_0, ..., \mathbf{E}^b_0, ..., \mathbf{E}^B_0\right]$.

The error estimation and cross-validation described above were introduced within the framework of PCA. The aim herein is to extend these procedures to the multi-block setting and propose graphical tools to enhance their interpretation.

2.3.1. Block errors

For the estimation of block errors, we suggest to calculate for each PC in each block *b*, the sum of squares of residual matrix according to Eq. (9)

$$SScv^b_A = \sum_{i=1}^{N}\sum_{j=1}^{Kb} e^b_A(i,j)^2 \tag{9}$$

where $e^b_A(i,j)$ is the *i*th, *j*th entry of residual block $\mathbf{E}^b_A$, (i.e. block *b* of the residual matrix for a model with *A* PCs included).

In order to calculate cross-validated estimates of the mean squared errors for each block we propose to calculate MSE^b according to:

$$MSE^b_A = \frac{SScv^b_A}{N(K_b - h^b_A)}$$
$$MSE^b_0 = \frac{SScv^b_0}{NK_b} \tag{10}$$

The quantity h^b_A is the partial block leverage, intended to represent block *b* contribution in the *A* degrees of freedom consumed in predicting *A* global scores $\mathbf{t}_a$. The partial block leverages h^b_A for *A* components and every block is calculated according to Eq. (11)

$$\tilde{\mathbf{p}}'_a = \left[\tilde{\mathbf{p}}^{1\prime}_a, \tilde{\mathbf{p}}^{2\prime}_a, ..., \tilde{\mathbf{p}}^{b\prime}_a, ..., \tilde{\mathbf{p}}^{B\prime}_a\right]$$
$$\mathbf{p}^{*\prime}_a = \tilde{\mathbf{p}}'_a\left(\tilde{\mathbf{p}}'_a\tilde{\mathbf{p}}_a\right)^{-1/2} \tag{11}$$
$$h^b_A = \sum_{a=1}^{A}\sum_{k=b1}^{b2} p^{*\prime}(k,a)^2$$

where $\tilde{\mathbf{p}}^{b\prime}$ is calculated in the CPCA algorithm in (Section 2.2.3), $p^{*\prime}(k, a)$ is the *(k,a)-th* entry of the loading vector matrix $\mathbf{P}^{*\prime}$ containing the loading vectors $\mathbf{p}^*_a$, defined in Eq. (11), as columns. The b_1 and b_2 are the column numbers of the first and the last variables in block *b*, respectively, considering all blocks. It is worth noting that the sum of the partial leverages equals *A*.

In order to have better RMSE plots where it is easier to choose the correct number of PCs, the mean squared errors associated with each block were augmented by 3% of the initial variance, MSE^b_0, for each new PC introduced in the model:

$$MSE^b_A + A \times 0.03 \times MSE^b_0 \rightarrow MSE^b_A \tag{12}$$

The 3% rule has been successfully used in PCA and PLSR for a number of years [52], e.g. in The Unscrambler software. We have tried several alternatives, more advanced methods to avoid incidental over optimism, e.g. trying to assess "significance" of each component etc., but we have concluded that the 3% rule seems the easiest and most reliable method. The cross-validated RMSE for each block was determined by calculating the square root of MSE^b_A for all various values of A (the number of components to be introduced in the CPCA model):

$$RMSE_A^b = \sqrt{MSE_A^b} \tag{13}$$

The percent cross-validated explained variance for each block, for all various values of A, is then calculated according to Eq. (14):

$$\text{Percent Explained Variance} = \frac{MSE_0^b - MSE_A^b}{MSE_0^b} \times 100 \tag{14}$$

The *unvalidated explained block variance* for each block and component a is given as fraction of the total variance of the respective block. The unvalidated explained variance is shown for each PC on the axes of the score plots in Figs. 3a–e, 4a–e.

2.3.2. Global errors

Global cross-validated RMSE and degree of freedom correction for RMSE is calculated as in PCA and briefly recalled in the following.

For all model ranks $A=0,1,...,A_{\max}$, the sum of squares of the cross-validated residual matrix was calculated according to Eq. (15)

$$\begin{aligned} SScv_0^g &= \sum_{i=1}^{N}\sum_{j=1}^{K} e_0(i,j)^2 \\ SScv_A^g &= \sum_{i=1}^{N}\sum_{j=1}^{K} e_A(i,j)^2 \end{aligned} \tag{15}$$

where $e_A(i,j)$ is the (i,j)–*th* entry of $\mathbf{E}_A$ and $\mathbf{E}_A = \left[\mathbf{E}_A^1, \mathbf{E}_A^2,..., \mathbf{E}_A^b,..., \mathbf{E}_A^B\right]$ is the residual matrix associated with a CPCA model which includes A PCs, $e_0(i,j)$ is the (i,j)-*th* entry of residual matrix $\mathbf{E}_0$ which is calculated in Section 2.3.

$SScv_A^g$ was then corrected for the approximate number of degrees of freedom being consumed, using Eq. (16):

$$\begin{aligned} MSE_0^g &= \frac{SScv_0^g}{NK} \\ MSE_A^g &= \frac{SScv_A^g}{N(K-A)} \end{aligned} \tag{16}$$

Eq. (16) assumes that all N objects have been sampled and measured independently, and that all K variables have been measured with independent measurement error. This is not always satisfied, and therefore, the MSE values are only considered as approximate.

Again, in order to have better RMSE plots where it is easier to choose the correct number of PCs, 3% of the initial variance is added for each new PC:

$$MSE_A^g + A \times 0.03 \times MSE_0^g \rightarrow MSE_A^g \tag{17}$$

The cross-validated global RMSE was determined by calculating the square root of MSE_A^g:

$$RMSE_A^g = \sqrt{MSE_A^g} \tag{18}$$

The global percent cross-validated explained variance is calculated according to Eq. (19):

$$\text{Percent Explained Variance} = \frac{MSE_0^g - MSE_A^g}{MSE_0^g} \times 100 \tag{19}$$

Unvalidated global explained variance is the fraction of the variance in the data set which is explained by component A. It is shown for each PC on the respective axes in Figs. 3f, 4f and 5. By examining the plot where the approximate global and block RMSEs were plotted against PC numbers, a decision about the number of relevant PCs to be retained in the CPCA model A_{opt} was made, in the whole data set and also in each block individually. The flow chart of the approximate RMSE calculation is given in Fig. 1.

2.4. Uncertainty *t*-test for the variables

In order to assess whether the variables at hand are contributing to the CPCA model, an approximate t-test was run on the loading coefficients. For the t-test an uncertainty standard deviation of the loading matrix $\mathbf{P}_{K,A_{opt}}$ was calculated by cross-validation. An important aspect involved here is that the orientation of the subspace in the variable space which is defined by the loading matrix $\mathbf{P}$ is in general not identical to the one which is defined by the loading matrix $\mathbf{P}_{-m}$ obtained by leaving out the segment m. In order to correct for rotations and flipping of directions Procrustes rotation was used. In each cross validation step, the full, global $\mathbf{P}_{-m}$, estimated from leave-in samples was rotated towards $\mathbf{P}$ by using Procrustes rotation. This resulted in a new loading matrix which is called $\hat{\mathbf{P}}_{-m}$. The calculation of the rotation matrix was done by taking the singular value decomposition of $\left(\mathbf{P}'_{-\mathbf{m}}\mathbf{P}\right)$. Related equations are detailed below:

$$\begin{aligned} &\text{SVD}\left(\mathbf{P}'_{-m,A_{opt}}\mathbf{P}_{A_{opt}}\right) = \mathbf{USV}' \\ &\mathbf{R}_{m,A_{opt}} = \mathbf{UV}' \\ &\hat{\mathbf{P}}_{-m} = \mathbf{P}_{-m,A_{opt}}\mathbf{R}_{m,A_{opt}} \end{aligned} \tag{20}$$

where A_{opt} is the number of relevant PCs, $\mathbf{P}'$ is the loading matrix for the whole data set and $\mathbf{P}'_{-m}$ is the loadings matrix of the leave-in samples. When the $A_{opt} \times A_{opt}$ rotation matrix R is estimated, this entails the estimation of $A_{opt}-1$ unknown parameters and hence consumption of $A_{opt}-1$ degrees of freedom from the original K input variables. Therefore, the variance/covariance of the rotated loadings $\hat{\mathbf{P}}_{-m}$, estimated in Eq. (21), should be corrected by the factor $\dfrac{K}{K-A_{opt}-1}$. By taking account of this degree-of-freedom correction we make sure that the Procrustes rotation does not create overfitting.

The jack-knifed estimate of the uncertainty standard deviation $\mathbf{S}_{K,A_{opt}}$ is finally calculated by comparing the loadings for the whole data set with Procrustes-rotated loadings of the leave-in samples $\hat{\mathbf{P}}_{-m}$. The $s_{k,\,a}$, which is the (k,a)-th entry of the matrix $\mathbf{S}_{K,A_{opt}}$, is calculated according to

$$s_{k,a}^2 = \sum_{m=1}^{M} \left(p_{k,a} - \hat{p}_{k,a}\right)^2 C_M D_A \tag{21}$$

Where $p_{k,\,a}$ is the (k,a)-th entry of the loading matrix $\mathbf{P}_{K,A_{opt}}$, $C_M = \frac{M}{M-1}$ is the jack-knife correction factor and $D_A = \frac{K}{K-A_{opt}-1}$ is the degrees-of-freedom correction from the Procrustes rotation.

The t-statistics are then calculated according to

$$t_{k,a} = \frac{p_{k,a}}{s_{k,a}} \tag{22}$$

where $k=1,\dots,K$ denotes the variables, $a=1,\dots,A_{\text{opt}}$ the components.

2.5. Stability plots

2.5.1. Global stability score plots

Stability plots for PCA were introduced and discussed in detail by Martens & Martens in [40]. They aim at assessing the extent to which the scores are influenced by the various samples in the datasets. We discuss herein how to extend these plots to the multi-block setting.

The stability of the scores can be visualized by so-called stability plots. The stability of scores is estimated by cross-validation. Since the stability plots compare score values for different cross-validation models with the full model, the models need to be corrected using Procrustes rotations as in the previous section. Therefore, for the calculation of the stability scores the rotated loadings $\hat{\mathbf{P}}_{-m}$ of Eq. (20) are used:

$$\hat{\mathbf{T}}_A = \mathbf{X}_\mathbf{C}\hat{\mathbf{P}}_{-m} \tag{23}$$

where $\hat{\mathbf{P}}_{-m}$ is the rotated loading matrix of the leave-in samples after having taken out segment m calculated in Eq. (20) and $\mathbf{X}_\mathbf{C}$ is obtained by mean-centering the input data $\mathbf{X}_{input}$ using the mean of the leave-in samples $\mathbf{X}_{-m}$. Finally we obtain according to Eq. (23), M cross-validated score matrices $\hat{\mathbf{T}}_A$ (For the sake of simplicity, we omit a possible index m for marking $\hat{\mathbf{T}}_A$ according to the cross-validation model m). Scores $\mathbf{T}_A$, obtained by using the full model calculated in Eq. (3), are plotted together with the M cross-validated sets of scores $\hat{\mathbf{T}}_A$ in the following way: In score plots we draw n lines, for each sample, from each pair of score values (t_i^n, t_j^n), where t_i^n refers to column (component) i and row n (sample n) in $\mathbf{T}_A$, to all M pairs of cross-validated score values $\left(\hat{t}_i^n, \hat{t}_j^n\right)$, where $\hat{t}_i^n$ refers to column (component) i and row n (sample n) in $\hat{\mathbf{T}}_A$. The length of the instability line $d_1^n = \sqrt{\left(t_i^n - \hat{t}_i^n\right)^2 + \left(t_j^n - \hat{t}_j^n\right)^2}$ and its direction shows how much the parameters for sample n for components i and j vary when segment m is left out. This provides a nice overview of possible outliers etc.: If the instability line is long, it means that the sample has unique information not represented by samples in segment m, resulting in an unstable model. The instability line expression may be modified in various ways. For instance, since the uncertainty variance of the scores t_a^n is estimated by the jack-knifing expression $s_{n,a}^2 = \sum_{m=1}^{M} \left(t_a^n - \hat{t}_{-m,a}^n\right)^2 C_M D_A$, the individual perturbation distances can be corrected by $d_2^n = d_1^n \sqrt{C_M D_A}$ in order to avoid underestimation in the case of high-dimensional Procrustes rotation. In our example this is ignored, since both factors are close to 1.

In the results section we will mark the instability line that represents the cross-validation round where the sample itself has been left out by a dot. When more than one sample are left out in one cross-validation loop, all samples that are taken out in a given cross-validation loop could be marked.

2.5.2. Block stability score plots

In order to estimate stability of samples on block level, we will propose how stability plots can be calculated for block score plots. For setting the global stability plots we predicted cross-validated scores according to Eq. (23). This was possible because the rotated global loadings $\hat{\mathbf{P}}_{-m}$ are orthogonal. Since the rotated block loadings $\hat{\mathbf{P}}_{-m}^b$, that are obtained by splitting the rotated global loadings $\hat{\mathbf{P}}_{-m}$ into blocks $\left(\text{i.e } \hat{\mathbf{P}}_{-m} = \left[\hat{\mathbf{P}}_{-m}^1, \hat{\mathbf{P}}_{-m}^2, \ldots, \hat{\mathbf{P}}_{-m}^b, \ldots, \hat{\mathbf{P}}_{-m}^B\right]\right)$, are in general not orthogonal, the block stability scores $\hat{\mathbf{T}}^b$ cannot be calculated in the same way. They need to be calculated at every deflation step as shown in the following.

We start by calculating the cross-validated and rotated loadings $\hat{\mathbf{P}}_{-m}$ according to Eq. (20). For every segment m, we perform the following calculations and deflations on $\mathbf{X}_\mathrm{C}$:

1. Start by choosing the first loading $\hat{\mathbf{p}}_{-m,a}$ from $\hat{\mathbf{P}}_{-m}$ $(a = 1)$.
2. Split $\hat{\mathbf{p}}_{-m,a}$ into rotated and cross-validated block loadings $\hat{\mathbf{p}}_{-m,a} = \left[\hat{\mathbf{p}}_{-m,a}^1, \hat{\mathbf{p}}_{-m,a}^2, \ldots, \hat{\mathbf{p}}_{-m,a}^b, \ldots, \hat{\mathbf{p}}_{-m,a}^B\right]$
3. Normalize the $\hat{\mathbf{p}}_{-m,a}$ block-wise: $\frac{\hat{\mathbf{p}}_{-m,a}^b}{\|\hat{\mathbf{p}}_{-m,a}^b\|} \rightarrow \hat{\mathbf{p}}_{-m,a}^b$
4. Calculate the rotated and cross-validated block scores $\hat{\mathbf{t}}_a^b = \mathbf{X}_\mathrm{C}^b \hat{\mathbf{p}}_{-m,a}^b$
5. Replace $\mathbf{X}_\mathrm{C}$ by $\mathbf{X}_\mathrm{C} - (\mathbf{X}_\mathrm{C}\hat{\mathbf{p}}_{-m,a})\hat{\mathbf{p}}'_{-m,a}$ and increment a by one
6. Iterate steps 2.–5. A times to obtain A components.

Following this procedure we obtain block stability scores $\hat{\mathbf{T}}^b = \left[\hat{\mathbf{t}}_1^b, \hat{\mathbf{t}}_2^b, \ldots, \hat{\mathbf{t}}_a^b, \ldots, \hat{\mathbf{t}}_A^b\right]$. Similarly to the global stability scores, the block stability scores are obtained for every cross-validation loop for all samples, also for the left-out samples (here again we omit labeling the block stability scores $\hat{\mathbf{T}}^b$ according to the cross-validation loop m for the sake of simplicity). The block stability score plots are visualized in the same way as in the global stability score plots.

Since the Procrustes rotation of the loadings with A PCs estimates $A-1$ independent parameters from the K input variables, it is

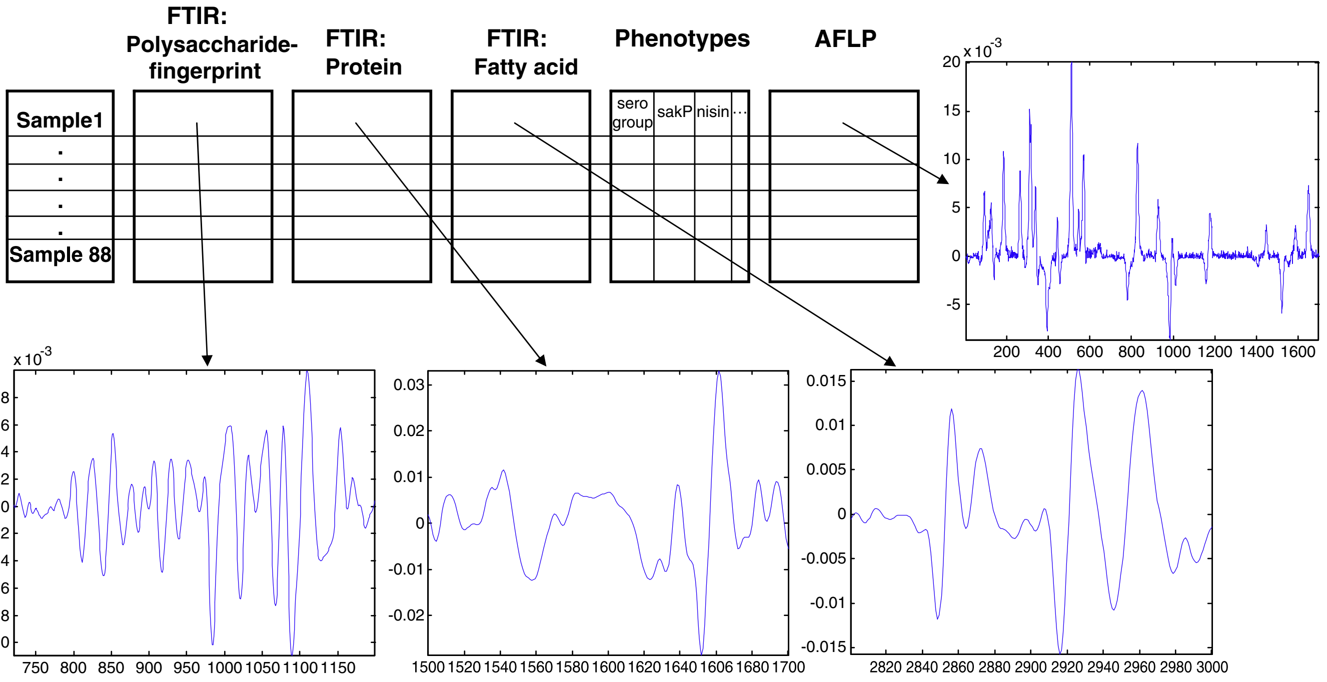


Fig. 2. Structure of "*Listeria monocytogenes* strains" data set. Three different ranges of the FT-IR spectra (1200–720 cm^{-1}, fingerprint region and polysaccharide region; 1700-1500 cm^{-1}, protein region; 3000-2800 cm^{-1}, fatty acid region) are used in order to produce three different blocks. In addition, phenotype data and AFLP data are used as two further data blocks [47,53].

considered to consume $A-1$ degrees of freedom. Hence, to avoid too small deviations in the stability plots, the length of the line segments may be corrected by the square root of the factor $\frac{K}{K-A_{opt}-1}$. But since the number of components here is low relative to the number of variables K, this correction is presently ignored.

3. Multi-block data set

The multi-block data set used in this study is described in detail in the references [47,53] and illustrated in Fig. 2. The data set consists of five data blocks with different numbers of variables in each block; all these variables being measured on the same set of 88 microbiological samples. This multi-block data set contains amplified fragment length polymorphism (AFLP) data (genetic fingerprinting), Fourier Transform Infrared (FTIR) spectra, and a collection of other, univariate phenotypes (serotype grouping, susceptibility to sakacin P, nisin and the antibacterial agent benzalkonium chloride) of 88 *L. monocytogenes* strains. FTIR spectroscopy is a rapid technique for metabolic fingerprinting of microorganisms [54]. The FTIR data block is subdivided into the following spectral region blocks: polysaccharide region and fingerprint region (720–1200 cm^{-1}) defining block $\mathbf{X}^1$ (498 variables), the protein region (1500–1700 cm^{-1}) defining block $\mathbf{X}^2$ (209 variables) and the fatty acid region (2800–3000 cm^{-1}) defining block $\mathbf{X}^3$ (208 variables). The phenotypes are collected in data block $\mathbf{X}^4$ (10 variables). The AFLP data block defines block $\mathbf{X}^5$ (1701 variables). Prior to CPCA the spectral data was pre-processed by EMSC [55,56]. The structure of the data set is shown in Fig. 2. Integrating and exploiting these omics data sets in combination was done in the reference [47] and further information about the background behind the *L. monocytogenes* strain to strain variation in general and especially the variation in susceptibility to bacteriocins was obtained. Different ways to group the 88 strains phenotypically have been studied: (1) According to sakacin sensitivity, since the strains form two distinct sensitivity groups. Half of the strains is below and half above a sensitivity threshold. (2) According to serotype. (3) According to the polysaccharide-fingerprint region in FTIR: The polysaccharide-fingerprint region of FTIR shows three distinct groups which we named *FTIR groups* [47,53]. The three polysaccharide-fingerprint groups can be obtained by running a Principal Component Analysis (PCA) on the polysaccharide-fingerprint region (720–1200 cm^{-1}) and considering the first two components. The FTIR groups will be used for graphical illustration in the following.

4. Results and discussion

4.1. CPCA of a multi-block data set

In order to find the common variation pattern between blocks in the multi-block data set, CPCA was performed on the five different

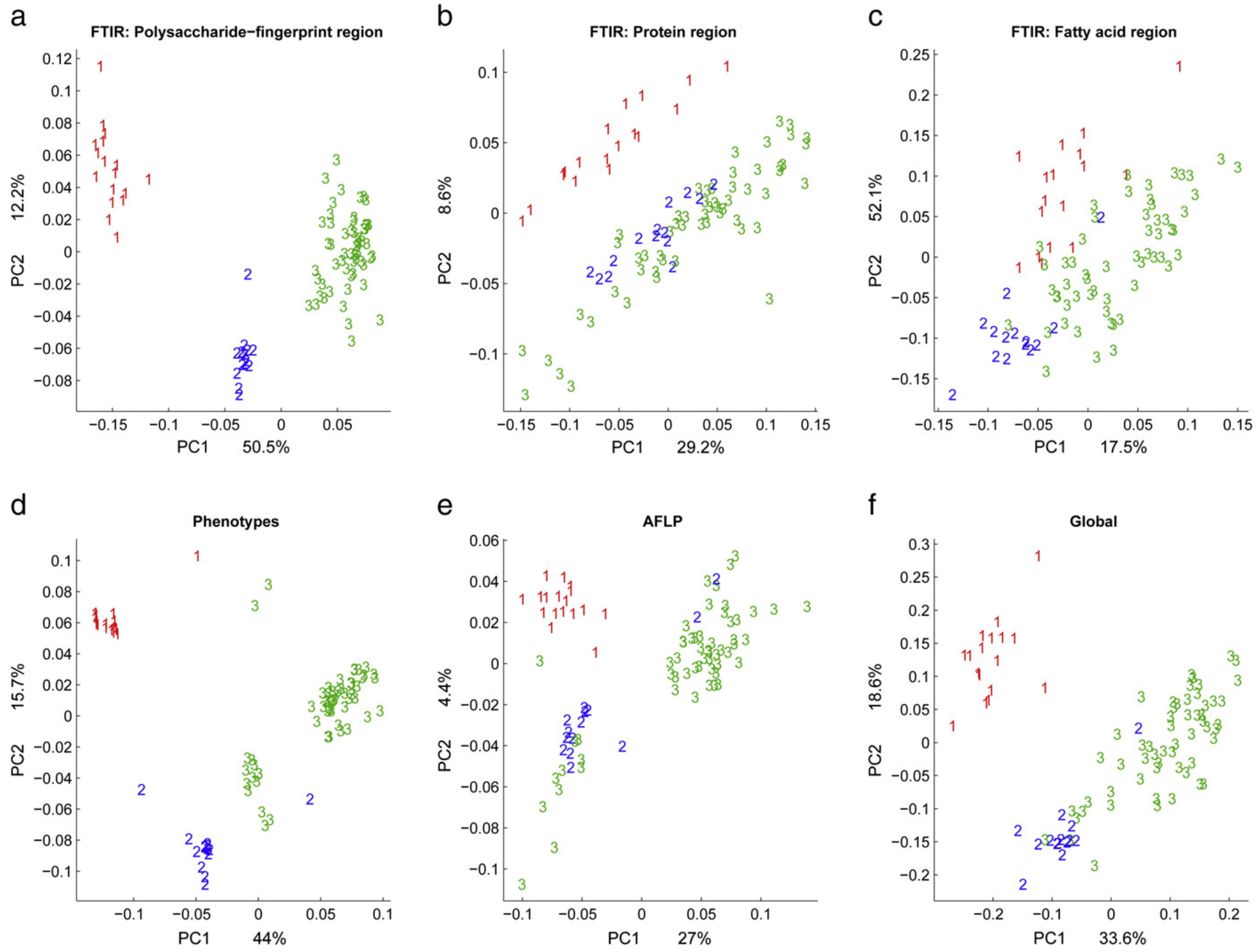


Fig. 3. Consensus Principal Component Analysis (CPCA) of the data. The samples are labeled “1” (red), “2” (blue) or “3” (green) according to FTIR polysaccharide-fingerprint groups. (a–e) First and second components of block scores. (f) First and second components of Global scores. The (un-validated) explained variance is shown on the axes.

data blocks. The structure of the data set is described above and illustrated in Fig. 2. In Fig. 3 the score plots (first and second components) for the five different blocks and the global scores are shown. The samples are labeled “1”(red), “2”(blue) or “3” (green) according to FTIR polysaccharide-fingerprint groups as defined in [47]. We see that the three polysaccharide-fingerprint groups are visible in the global pattern (Fig. 3f): The first principal component covers 34% of the total variation (unvalidated explained variance). The first component separates group 3 from groups 1 and 2. The second PC accounts for 19% of the variation in the data set and separates group 1 from groups 2 and 3. The sample variation pattern in block 1 (polysaccharide-fingerprint region) is very close to the global pattern. The phenotypes block (block 4) and AFLP block (block 5) show tendencies toward the same grouping pattern. However, blocks 2 and 3 (the protein region and fatty acid region of FTIR) show slightly different patterns. We can see that although there are some differences in the separation between groups in different blocks in Fig. 3, similar patterns can be identified in all blocks. This rises the question whether there are indeed similar co-variation patterns in all the blocks or whether one or few blocks (e.g. block 1) are dominant and impose their pattern to the other blocks. The (unvalidated) explained variance of block 1, the polysaccharide-fingerprint region, already gives an indication that block 1 is dominant. Nevertheless, all the other blocks show a relatively large explained variance. In Fig. 4 the score plots (third and fourth components) for the five different blocks and the global scores are shown. The third and the fourth global components do not explain a large variation in the data as they both recover only 16% of the variation. It can also be seen that the second block leads the global pattern for the third and fourth principal components. Indeed, 28% of the variation in the second block is explained by the third and fourth components.

The correlation loading plot (first and second components) is shown in Fig. 5, revealing correlations between the global scores **T** and the variables in the different data blocks used for CPCA (black: FTIR (polysaccharide-fingerprint region); blue: FTIR (protein region); green: FTIR (fatty acid region); red: phenotypes data; and olive: AFLP data). A circle centered at the origin and unit radius is also drawn and represents the maximum correlation. More precisely, whenever a variable is close to the circle, this highlights that it can be well predicted by the two PCs under consideration. Contrariwise, whenever a variable is close to the origin, this means that it is not connected to the two considered PCs. This graphical display also makes it possible to depict the extent to which the variables in **X** are correlated to each other and the two PCs being considered. Roughly speaking, two variables close to the circle which point to the same direction are positively correlated and vice versa.

If $\mathbf{x}_k$ is the kth variable vector (column) in the data matrix **X** and $(\mathbf{t}_i, \mathbf{t}_j)$ are score vectors of two PCs, for the correlation loading plot, the correlation coefficients (r_{ki}, r_{kj}) are calculated for $\mathbf{x}_k$ towards $\mathbf{t}_i$ and $\mathbf{t}_j$. The variable corresponding to $\mathbf{x}_k$ is then represented as a point with coordinates (r_{ki}, r_{kj}). This is done for all of the variables. Due to the

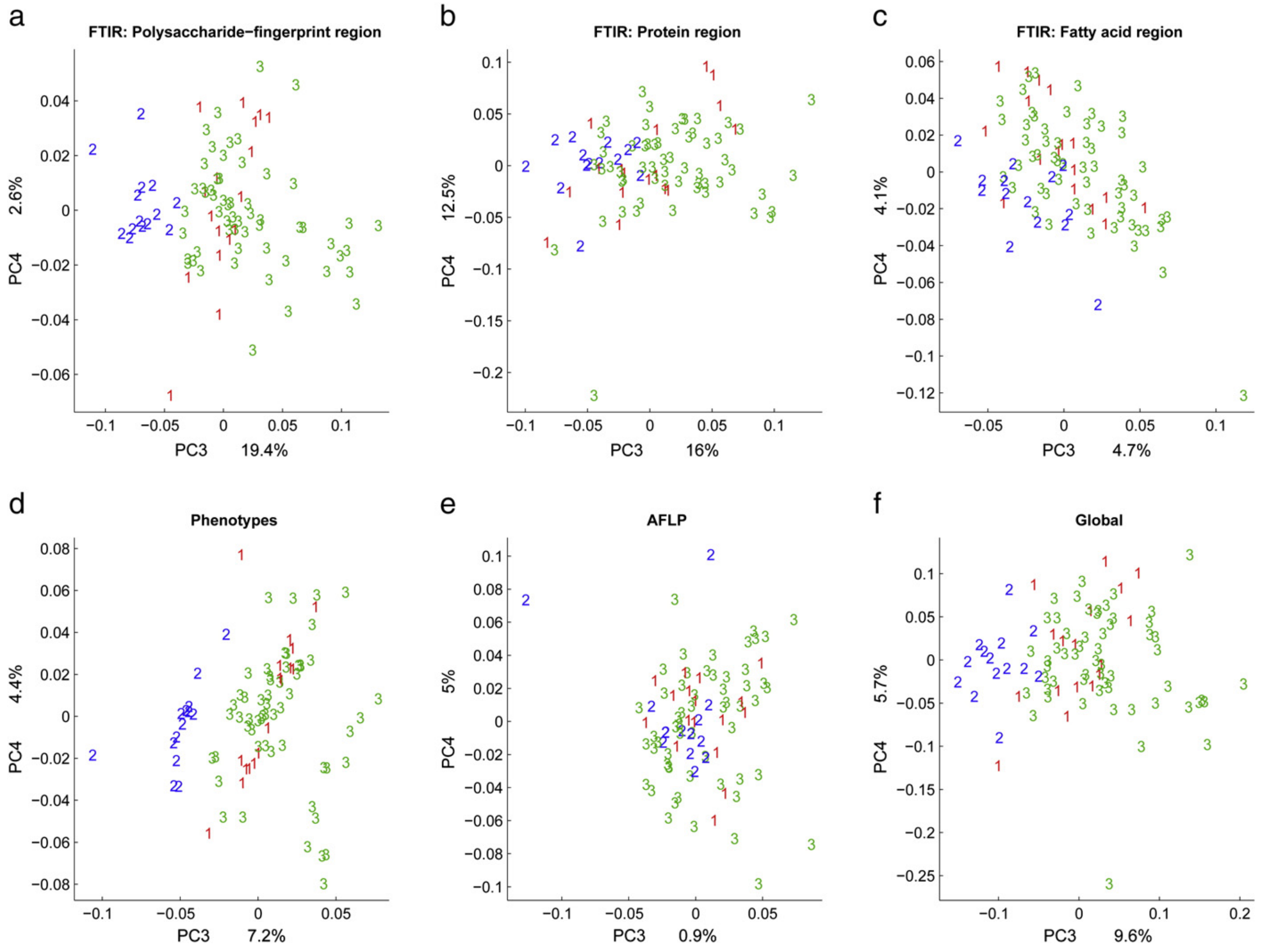


Fig. 4. Consensus Principal Component Analysis (CPCA) of the data. The samples are labeled “1” (red), “2” (blue) or “3” (green) according to FTIR polysaccharide-fingerprint groups. (a–e) Third and fourth components of block scores. (f) Third and fourth components of Global scores. The (un-validated) explained variance is shown on the axes.

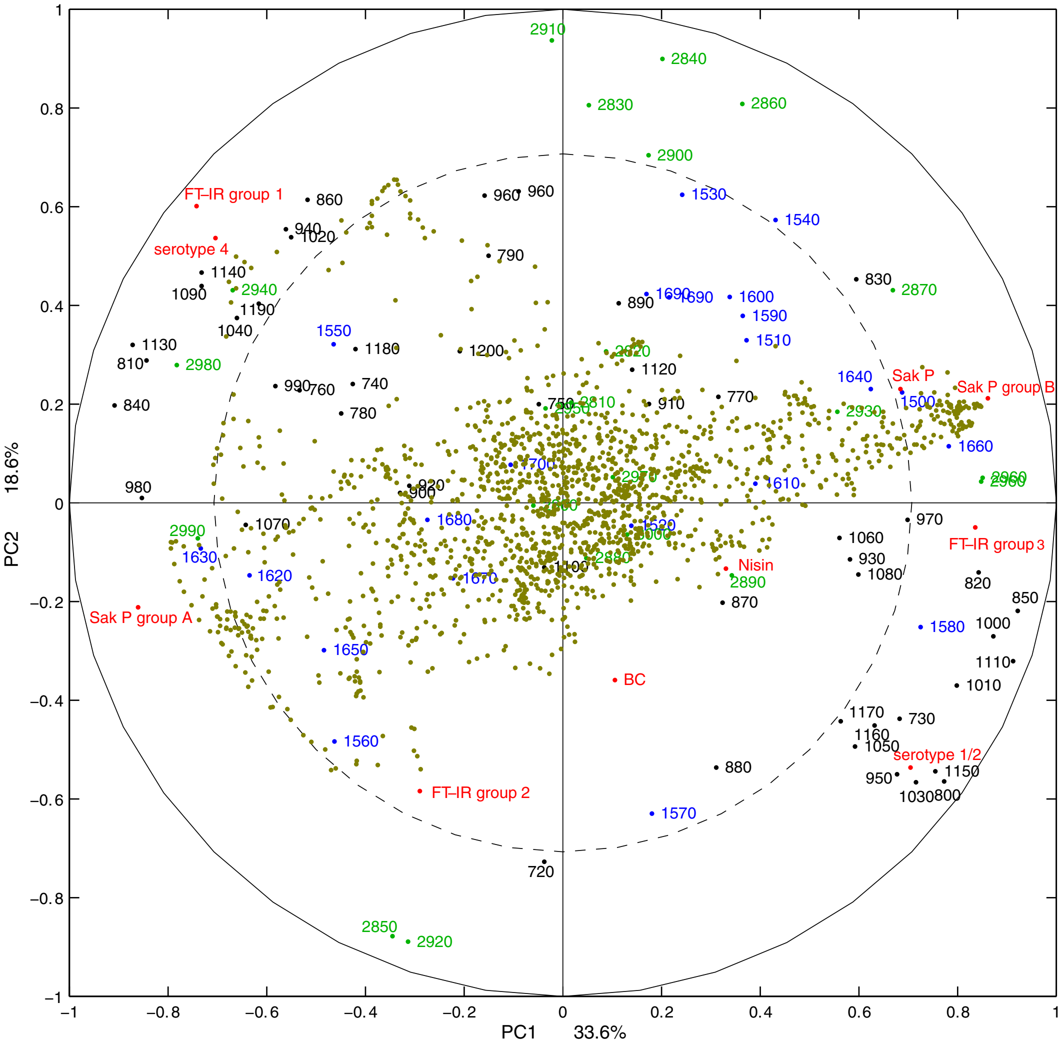


Fig. 5. Consensus Principal Component Analysis (CPCA) of the data. Correlation loading plot. The (un-validated) explained variance is shown on the axes.

orthogonality between scores, the norm of the vector (r_{ki},r_{kj}) is less than or equal to 1. Thus a circle of radius equal to 1 is plotted, on the correlation plot, with the center at the origin of the graph. If $\mathbf{x}_k$ is positioned on the circle it indicates that it is possible to exactly predict the variable k from the scores $\mathbf{t}_i$, $\mathbf{t}_j$. A variable positioned close to the origin of the graph is then not predictable from the studied pair of scores and it plays no role in the “construction” of these scores. Moreover, if two variables are close together on the correlation plot and also close to the perimeter of the circle, they are strongly positively correlated, while they are strongly negatively correlated if they are found at the opposite perimeter.

For the sake of clarity, only FTIR variables related to actual chemical bands are plotted in Fig. 5. The figure clearly shows a high correlation between polysaccharide-fingerprint region of FTIR data and serotype grouping. It can also be seen that AFLP variables have high correlations with sakacin P susceptibility and FT-IR group 3, while they do not have significant correlation with any of the serotype groups. The protein region of FTIR data is mainly distributed in the inner circle, which corresponds to 50% explained variance, revealing the fact that the protein region of FTIR data is not well explained by the first two principal components.

4.2. Root Mean Square Error (RMSE): a cross-validation approach

In order to validate visually detected variation patterns of CPCA we propose in this paper to study global RMSE and block RMSEs, calculated by Eqs. (18) and (13) respectively. Visual perception when inspecting the score plots can be misleading since scientist's mind is always looking for patterns of grouping. Colors in the score plots improve visualizing the underlying patterns and grouping whilst they can lead to false discovery. Normally, if the different blocks of variables represent different types of information about the samples, one would expect the blocks to contain different numbers of important latent structures. It is important to discover these differences in latent block structure. But on the other hand we want to compare the different blocks in a common cognitive structure; one way to do this is the CPCA with block score displays, but with deflation on the global, common scores only. The block RMSE plots then give a validated image of block score plots revealing the important contributions to the global patterns. RMSE plots for all of the five blocks together with the global RMSE are shown in Fig. 6a. We observe that block 3 has the most important contribution to the second component, which is already indicated by the unvalidated explained variance. Blocks 1 and 4 are also contributing to the component number 2 but obviously block 3 affects the second component more than any other block. The most remarkable result is that Block 5 (AFLP) has no influence on the second component (see Fig. 3) even if the block score plot (Fig. 3e) shows a nice pattern. Although a pattern was seen in block 5 with the same tendency as the global pattern for the first and second components (Section 4.1), the validation (RMSE) plot shows that block 5 does not contribute to the global variation pattern in the second PC and in fact it is the only block which does not

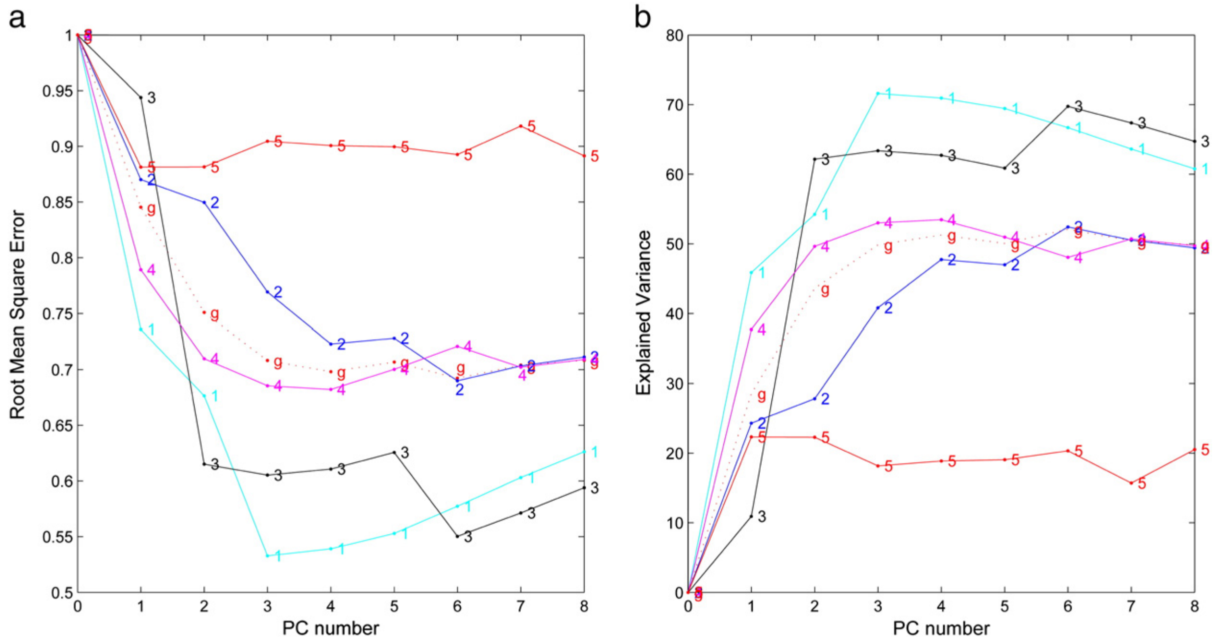


Fig. 6. (a) Root Mean Square Error (RMSE) plot for the data set. (1–5) RMSE plots for blocks 1–5. (red dotted) Global RMSE plot. (b) Explained variance plot for the data set. (1–5) for blocks 1–5. (red dotted) Global explained variance.

fit to the rest of the data. Block 4 and 1 are mostly contributing to the global pattern for the first component whilst block 3 is not important for component one at all.

RMSE plots can also be used for determining the number of relevant PCs. As it is seen in Fig. 6a the global RMSE reaches a local minimum at four PC. Although RMSE decreases for blocks 2 and 3 up to the sixth component, it does not seem to have a significant effect on the global pattern. Studying the RMSE plot, it is concluded that the first block consists of three components and that the common pattern in the data set for the first two PCs is much like the pattern in this block. Looking at the fifth block reveals that it is very much different from the other blocks, as it shares only the first PC with the other blocks and it does not fit the common pattern in the data set.

By looking at the global RMSE plot one can state that in order to extract most of the information from the multi-block data only the first four components should be retained. It is also clear that the sixth PC is just affected by the second and third blocks and it does not contain any reliable information for the other blocks. In addition to allowing us to find the necessary rank of the CPCA model, the rank estimates in the individual blocks give us new insight about these sets of variables. Of course, further rank details about the individual blocks may be obtained by PCA of individual blocks or PLSR between pairs of blocks.

4.3. Cross-validated explained variance

The cross-validated explained variance is calculated according to Eqs. (14) and (19). The corresponding cross-validated explained variance plot is shown in Fig. 6b. This plot displays the RMSE information in Fig. 6a in a different way. It depicts the percentage of the total variance in the data explained by a CPCA model as a function of the number of PCs. From this figure, one can also assess how much of the variation is left unexplained after a specific number of components have been introduced in the model. It can be seen that with three components more than 60% of the variation in the first and third blocks is explained by the model whereas the fifth block never reaches even 50% of the explained variance.

Fig. 7 shows bar plots for the cross-validated explained variance (a) together with the unvalidated explained variance (b), until rank $A=7$. It is interesting to compare them as it can be informative in some cases e.g. if one block is pulling the global pattern toward itself while the block is led by few extreme samples then the original explained variance will be high for the mentioned block while after cross-validation it would be hard for the model to estimate the extreme samples. As a consequence, this will force the cross-validated explained variance bar to go much lower than the original one. In the present study the plots do not contain such extreme cases and it can be concluded that the original explained variance is trustful enough. The explained variance is much lower and even negative for the fifth PC after cross-validation which makes the fifth PC less trustworthy than what one may think by looking at the original results.

4.4. Detecting important variables

The patterns which were detected by CPCA analysis have thus been validated by a statistical assessment, but the important variables which contributed to these patterns have not been identified yet. Looking at the correlation loading plots that show the correlation between variables and PCs, gives an idea about which variables contribute to the determination of each PC. A good way for assessing the variable importance is jack-knife-based *p*-value. An approximate *t*-test is run on the variables and results in *p*-values for each variable. In order to visualize the significance of variables we plot negative logarithmic *p*-values ($-\log_{10}(p\text{-value})$, i.e. $p=0.1$, 0.01 or 0.001 yield 1, 2 or 3, respectively). This makes it possible to highlight the most significant values. The respective plot for all variables for the second component is shown in Fig. 8. The red line defines the 5% significance level (variables above the threshold are considered to be significant). It can be seen that a lot of variables are significant in the first four blocks compare to not many significant variables in the fifth block although this latter block contains more than 1700 variables. This result confirms the conclusions from the RMSE plots regarding the weak contribution of block five to the second component. It is also clear that in the first block there are many important variables (confirming the conclusions from RMSE plots). Variables can be examined for different numbers of components and different blocks separately. E.g. in FTIR variables corresponding to specific spectral bands in the FTIR spectra were evaluated and it turned out that the

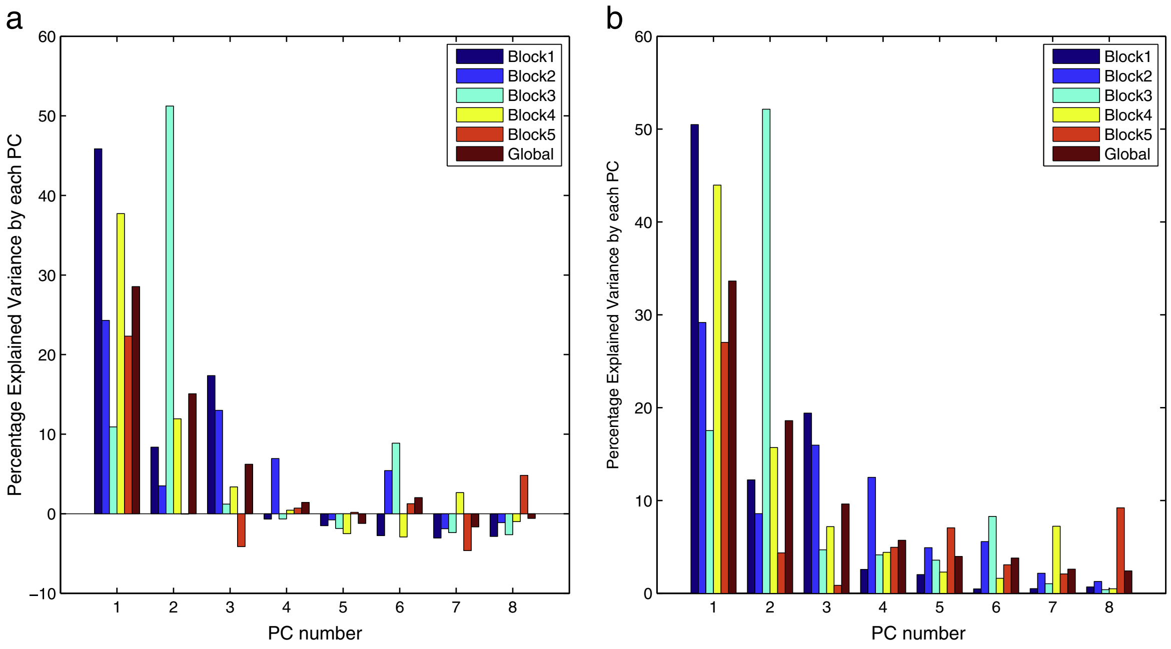


Fig. 7. (a) Cross-validated explained variance. Bar plot for the percentage cross-validated explained variance in each PC. (b) Explained variance. Bar plot for the percentage (unvalidated) explained variance in each PC.

bands at 835 cm^{-1} and 980 cm^{-1} were significant variables for the first component (results are not shown). Fig. 5 shows that the first component accounts for the variation in susceptibility toward sakacin P. The bands at 835 cm^{-1} and 980 cm^{-1} can be attributed to pyranose rings, and Lafleur [57] suggested that the variation in susceptibility toward sakacin P is connected to variations in the cell wall, probably to variations in pyranose. Since these two variables are found to be significant for explaining the variation in susceptibility towards sakacin P and, moreover, this can be explained biologically, we can be confident that high significances found for these two variables are not spurious.

4.5. Stability of the model and detecting outliers

As an additional approach to validate visually identified variation patterns in CPCA, we defined in this paper block stability scores. Fig. 9a–e shows such block stability plots for the first and second components for the five blocks. In Fig. 9f the global stability plot for the first and second components is shown. In Fig. 9 we can see how samples change positions during cross-validation loops where the model predicts left-out samples from a model determined on the basis of leave-in ones. The iteration when the sample itself was left out is marked by a dot. It can be seen that the stability plot gets less stable in the third and fourth components (Fig. 10). In Fig. 9e (plot referring to the AFLP data) a sample, belonging to the second group (blue), shows a high instability when the sample itself is left out in the construction of the model (the dot at the end of the instability line indicates the cross-validation step when the sample itself has been left out). The same sample appears to be very stable in the other stability plots. The reason for such a behavior can be either a measurement error in the AFLP or that the sample represents a 'real' (biological) outlier. In order to clarify this, additional analyses of this sample need to be done. In Fig. 10 one sample belonging to the third group (green) largely changes position during an iteration of the cross-validation loop. Since this was again the iteration, where the sample itself was left out, it can be concluded that the sample has a large effect on the model

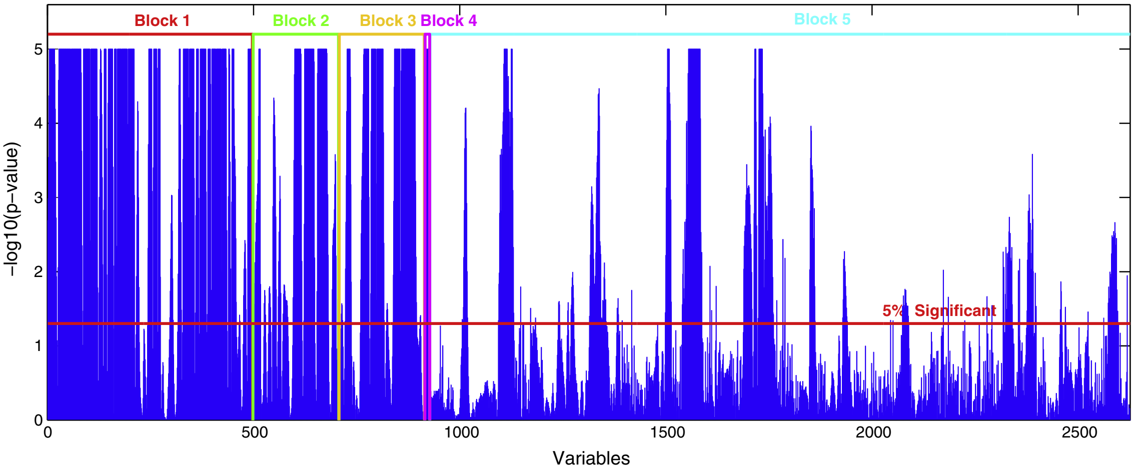


Fig. 8. Negative logarithmic plot of p-values ($-\log_{10}(p\text{-value})$) for all blocks for the first component, truncated at 5 (i.e. $p \leq 0.00001$). Blocks are specified by different colors.

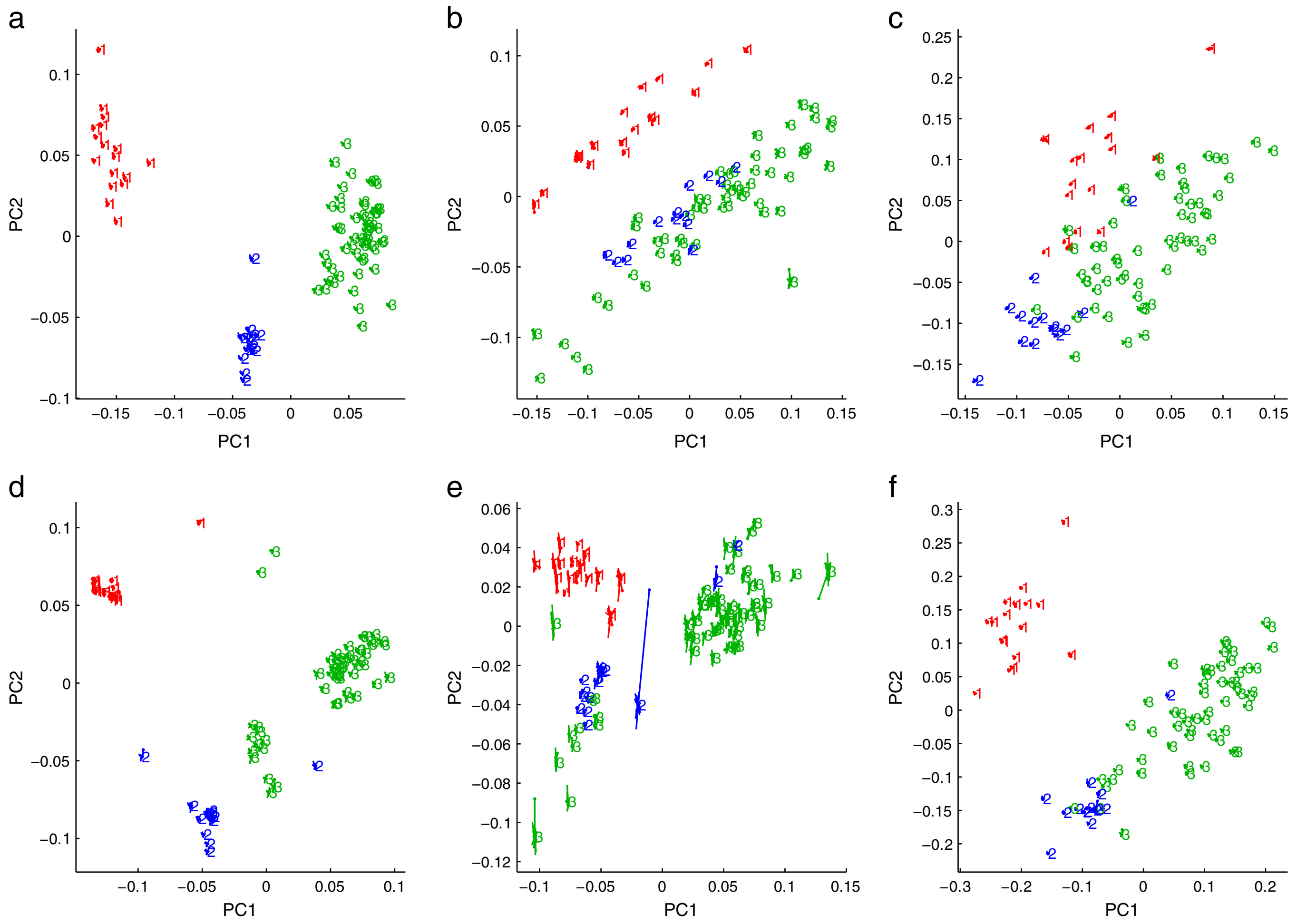


Fig. 9. Stability plot of the scores for the first and second components. The samples are labeled "1" (red), "2" (blue) or "3" (green) according to FTIR polysaccharide-fingerprint groups. The dots at the end of the perturbation lines indicate the cross-validation step when the sample itself has been left out. (a–e) First and second components of block scores stability plots. (f) First and second components of Global scores stability plot.

since its position is well estimated when it is included among leave-in samples whereas the position is inaccurately estimated when the sample is left aside. It is worth noting that the same kind of variation in position for that sample is seen in blocks two, three and four. Therefore that sample can be considered as a real biological outlier – all the data blocks consider it as such – and its behavior should be studied more carefully.

5. Conclusion

Multi-block methods provide an overview over a multi-block data set, in terms of sample- and variable- variation patterns within and between blocks of variables. The visual perception of patterns by inspection of global and block score plots is very subjective and needs validation: In explorative data modelling the scientist who designs the experiment and performs the data analysis is an essential part of the process. For better or worse our prior knowledge and hypotheses influence the perception of patterns and thereby the reported results. For instance, the recognition of structures and patterns in the score plots is a creative process that is strongly influenced by the user's expectations about the biological system. In order to reduce the subjectivity and prevent the danger of being misled by random noise effects, we present various validation techniques to help the user to critically assess the visual perception when interpreting the results from multi-block analysis. More precisely, new tools for validating block variation patterns together with the global variation patterns as obtained in CPCA are presented. They can help the users validate their discovery of visual variation interactions and common structures in complex datasets. We have also introduced the use of block RMSE plots, which are powerful graphical tools for evaluating the contribution of visually identified block patterns to the global variation patterns. We have shown that certain block components after validation by the block RMSE plots appear to contain no relevant information for the pattern detected in the global score plot. This shows that the interpretation of block patterns can be misleading and that patterns that appear in some of the block scores are only present because they are introduced by other blocks that have a very strong connection to the associated components.

By the use of global stability plots, the reliability of multi-block models as a function of the model rank can be assessed. Another important use of the stability plots is the possibility to identify outliers. In the present paper we have introduced the block stability plots, which can be used to assess outliers on block level. This has turned out to be very useful, since it made it possible to assess whether a sample is an outlier with respect to one block only or with respect to the global pattern. A sample that is identified as an outlier on block level, might either have a very special property seen only by this block (e.g. -omics technique) or be subjected to a measurement error in the block under consideration (e.g. -omics technique).

One possible way for assessing the importance of variables for a given pattern is to present significance level estimates (*p*-values),

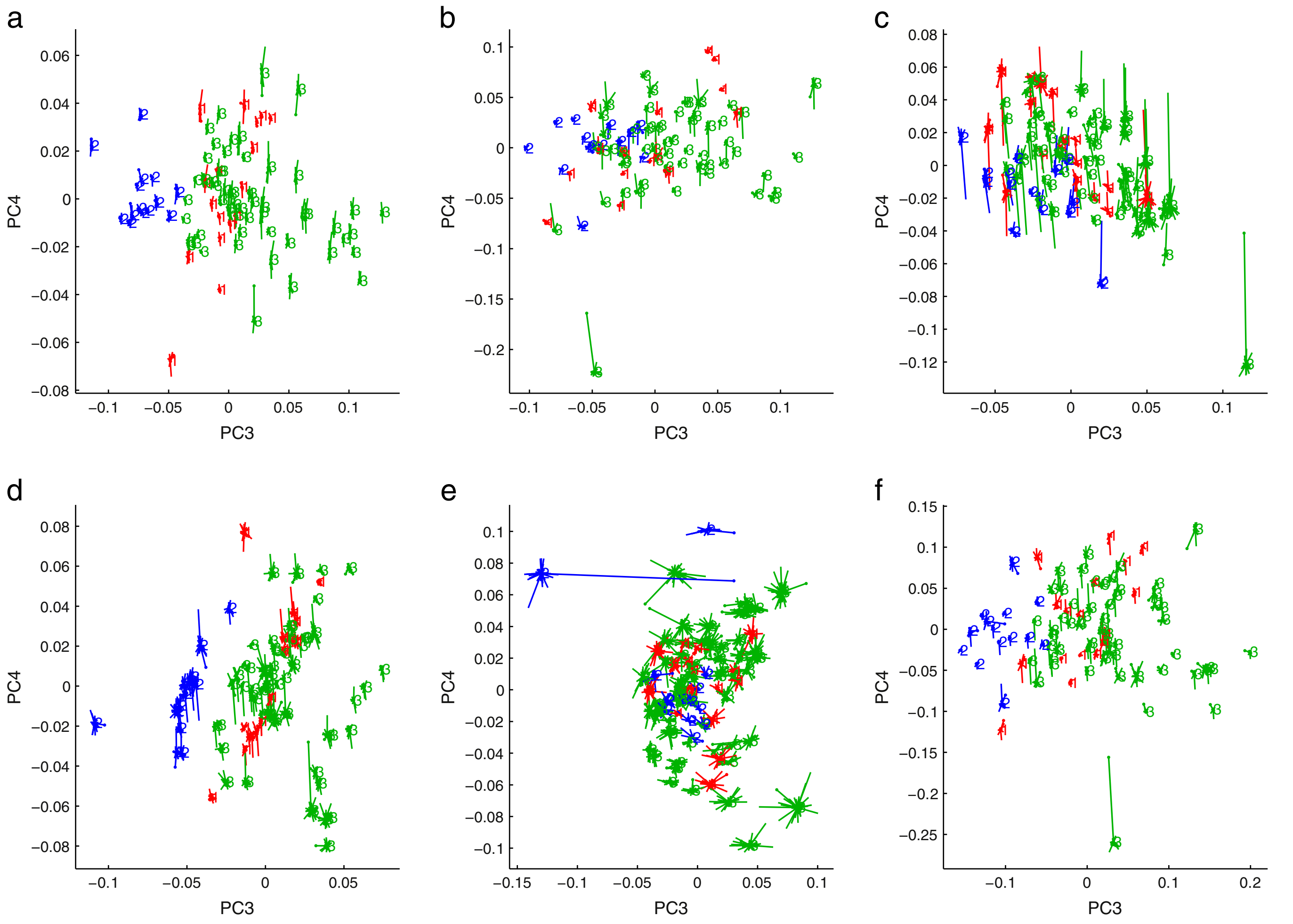


Fig. 10. Stability plot of the scores for the third and fourth components. The samples are labeled "1" (red), "2" (blue) or "3" (green) according to FTIR polysaccharide-fingerprint groups. The dots at the end of the perturbation lines indicate the cross-validation step when the sample itself has been left out. (a–e) Third and fourth components of block scores stability plots. (f) Third and fourth components of Global scores stability plot.

which reflect the probability that the observed effect could have been caused by random noise in the data. Such *p*-values are often used to assess whether computed values reflect significant departures from pre-specified hypotheses. In our context, we used them in a more informative way by plotting them as measures that highlight which variables are significantly contributing to the determination of the various principal components. The way of plotting the approximate *p*-values can either be used to get an overview on how single blocks contribute to the global model, or they can be used to assess the importance of single variables, as specific bands in the FTIR spectra.

An interesting issue in PCA and CPCA is to define the number of relevant components. Here it was achieved by plots of approximate estimates of the modelling error RMSE from cross-validation. There are other alternative methods for determining the number of relevant components which we plan to study in the future. On the other hand the present work is not limited to CPCA. It can be easily extended to other multi-block methods such as HPCA.

Acknowledgements

The authors are grateful for financial support by the Nordic Centre of Excellence on Food, Nutrition and Health "Systems biology in controlled dietary interventions and cohort studies" (SYSDIET, nr 070014) funded by NordForsk and by Foundation for Research Levy on Agricultural Products in Norway.

References

[1] G. Emilien, M. Ponchon, C. Caldas, O. Isacson, J.M. Maloteaux, Impact of genomics on drug discovery and clinical medicine, QJM 93 (7) (2000) 391–423.

[2] M.F. Rothschild, G.S. Plastow, Impact of genomics on animal agriculture and opportunities for animal health, Trends in Biotechnology 26 (1) (2008) 21–25.

[3] M. Georges, Recent progress in livestock genomics and potential impact on breeding programs, Theriogenology 55 (1) (2001) 15–21.

[4] J. Salmeron, L.R. Herrera-Estrella, Plant biotechnology: fast-forward genomics for improved crop production, Current Opinion in Plant Biology 9 (2) (2006) 177–179.

[5] M.J. Cunningham, Genomics and proteomics: The new millennium of drug discovery and development, Journal of Pharmacological and Toxicological Methods 44 (1) (2000) 291–300.

[6] A. Sinha, C. Singh, D. Parmar, M.P. Singh, Proteomics in clinical interventions: achievements and limitations in biomarker development, Life Sciences 80 (15) (2007) 1345–1354.

[7] D.A. Colantonio, D.W. Chan, The clinical application of proteomics, Clinica Chimica Acta 357 (2) (2005) 151–158.

[8] N.L. Anderson, A.D. Matheson, S. Steiner, Proteomics: applications in basic and applied biology, Current Opinion in Biotechnology 11 (4) (2000) 408–412.

[9] J.K. Nicholson, J.C. Lindon, Systems biology: metabonomics, Nature 455 (7216) (2008) 1054–1056.

[10] D.B. Kell, Systems biology, metabolic modelling and metabolomics in drug discovery and development, Drug Discovery Today 11 (23–24) (2006) 1085–1092.

[11] H. Stenlund, R. Madsen, A. Vivi, M. Calderisi, T. Lundstedt, M. Tassini, M. Carmellini, J. Trygg, Monitoring kidney-transplant patients using metabolomics and dynamic modeling, Chemometrics and Intelligent Laboratory Systems 98 (1) (2009) 45–50.

[12] G.D. Lewis, A. Asnani, R.E. Gerszten, Application of metabolomics to cardiovascular biomarker and pathway discovery, Journal of the American College of Cardiology 52 (2) (2008) 117–123.

[13] S. Naylor, A.W. Culbertson, S.J. Valentine, Towards a systems level analysis of health and nutrition, Current Opinion in Biotechnology 19 (2) (2008) 100–109.

[14] H. Kim, G.P. Page, S. Barnes, Proteomics and mass spectrometry in nutrition research, Nutrition 20 (1) (2004) 155–165.

[15] S. Rezzi, Z. Ramadan, L.B. Fay, S. Kochhar, Nutritional metabonomics: applications and perspectives, Journal of Proteome Research 6 (2) (2007) 513–525.

[16] D.S. Wishart, Metabolomics: applications to food science and nutrition research, Trends in Food Science and Technology 19 (9) (2008) 482–493.

[17] P. Vos, R. Hogers, M. Bleeker, M. Reijans, T.V.D. Lee, M. Hornes, A. Friters, J. Pot, J. Paleman, M. Kuiper, M. Zabeau, AFLP: a new technique for DNA fingerprinting, Nucleic Acids Research 23 (21) (1995) 4407–4414.

[18] J. Owens, Ian Humphery–Smith on current challenges in proteomics, Targets 2 (1) (2003) 10–13.

[19] S. Jamesdaniel, R. Salvi, D. Coling, Auditory proteomics: methods, accomplishments and challenges, Brain Research 1277 (2009) 24–36.

[20] H. Wu, A.D. Southam, A. Hines, M.R. Viant, High-throughput tissue extraction protocol for NMR- and MS-based metabolomics, Analytical Biochemistry 372 (2) (2008) 204–212.

[21] E. Werner, J.-F. Heilier, C. Ducruix, E. Ezan, C. Junot, J.-C. Tabet, Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends, Journal of Chromatography B 871 (2) (2008) 143–163.

[22] D.S. Wishart, Quantitative metabolomics using NMR, TrAC, Trends in Analytical Chemistry 27 (3) (2008) 228–237.

[23] C. Kumar, M. Mann, Bioinformatics analysis of mass spectrometry-based proteomics data sets, FEBS Letters 583 (11) (2009) 1703–1712.

[24] G.G. Harrigan, R.H. LaPlante, G.N. Cosma, G. Cockerell, R. Goodacre, J.F. Maddox, J.P. Luyendyk, P.E. Ganey, R.A. Roth, Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity, Toxicology Letters 146 (2004) 197–205.

[25] J.J.O. Lay, R. Liyanage, S. Borgmann, C.L. Wilkins, Problems with the "omics", TrAC, Trends in Analytical Chemistry 25 (11) (2006) 1046–1056.

[26] D.M. Rocke, Design and analysis of experiments with high throughput biological assay data, Seminars in Cell & Developmental Biology 15 (6) (2004) 703–713.

[27] H.C.J. Hoefsloot, D.J. Vis, J.A. Westerhuis, A.K. Smilde, J.J. Jansen, Multiset data analysis: ANOVA simultaneous component analysis and related methods, in: S. Brown, R. Tauler, R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 453–472.

[28] J. Sarembaud, R. Pinto, D.N. Rutledge, M. Feinberg, Application of the ANOVA-PCA method to stability studies of reference materials, Analytica Chimica Acta 603 (2) (2007) 147–154.

[29] R. Climaco Pinto, V. Bosc, H. Noçairi, A.S. Barros, D.N. Rutledge, Using ANOVA-PCA for discriminant analysis: application to the study of mid-infrared spectra of carraghenan gels as a function of concentration and temperature, Analytica Chimica Acta 629 (1–2) (2008) 47–55.

[30] J.M. Fostel, Towards standards for data exchange and integration and their impact on a public database such as CEBS (Chemical Effects in Biological Systems), Toxicology and Applied Pharmacology 233 (1) (2008) 54–62.

[31] A. Aderem, Systems biology: its practice and challenges, Cell 121 (4) (2005) 511–513.

[32] E.M. Færgestad, Ø. Langsrud, M. Høy, K. Hollung, S. Sæbø, K.H. Liland, A. Kohler, L. Gidskehaug, J. Almergren, E. Anderssen, H. Martens, Analysis of megavariate data in functional genomics, in: S. Brown, R. Tauler, R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 221–278.

[33] C.D. Millar, L. Huynen, S. Subramanian, E. Mohandesan, D.M. Lambert, New developments in ancient genomics, Trends in Ecology & Evolution 23 (7) (2008) 386–393.

[34] A.K. Smilde, M.T.J. van der Werf, S. Bijlsma, B.J.C. van der Werff-van der Vat, R.H. Jellema, Fusion of mass spectrometry-based metabolomics data, Analytical Chemistry 77 (20) (2005) 6729–6736.

[35] J.C. Lindon, E. Holmes, J.K. Nicholson, Global systems biology through integration of "omics" results, in: J.C. Lindon, J.K. Nicholson, E. Holmes (Eds.), The Handbook of Metabonomics and Metabolomics, Elsevier Science B.V., Amsterdam, 2007, pp. 533–555.

[36] K. Meyer, L. Suter-Dick, A. Amberg, J.-C. Gautier, M. Wendt, B. Riefke, A. Sutter, M. Raschke, H. Gmuender, S08: The challenge of integrating different "omics" technologies, Experimental and Toxicologic Pathology 61 (3) (2009) 260.

[37] M. Orei, C.B. Clish, E.J. Davidov, E. Verheij, J. Vogels, L.M. Havekes, E. Neumann, A. Adourian, S. Naylor, G. Jan van der, T. Plasterer, Phenotype characterisation using integrated gene transcript, protein and metabolite profiling, Applied Bioinformatics 3 (2004) 205–217.

[38] H. Martens, T. Næs, Multivariate calibration. I. Concepts and distinctions, TrAC, Trends in Analytical Chemistry 3 (8) (1984) 204–210.

[39] H. Martens, T. Næs, Multivariate calibration, Wiley & Sons, Chichester, 1989.

[40] H. Martens, M. Martens, Multivariate analysis of quality: an introduction, Wiley, Chichester, UK, 2001.

[41] V. Tyssø, K. Esbensen, H. Martens, UNSCRAMBLER, an interactive program for multivariate calibration and prediction, Chemometrics and Intelligent Laboratory Systems 2 (1–3) (1987) 239–243.

[42] A. Höskuldsson, K. Svinning, Modelling of multi-block data, Journal of Chemometrics 20 (8–10) (2006) 376–385.

[43] P. Geladi, Analysis of multi-way (multi-mode) data, Chemometrics and Intelligent Laboratory Systems 7 (1–2) (1989) 11–30.

[44] E.J.J. van Velzen, J.A. Westerhuis, J.P.M. van Duynhoven, F.A. van Dorsten, H.C.J. Hoefsloot, D.M. Jacobs, S. Smit, R. Draijer, C.I. Kroner, A.K. Smilde, Multilevel data analysis of a crossover designed human nutritional intervention study, Journal of Proteome Research 7 (10) (2008) 4483–4491.

[45] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, Journal of Chemometrics 12 (5) (1998) 301–321.

[46] P. Bougeard, M. Hanafi, E.M. Qannari, Multiblock latent root regression. Application to epidemiological data, Computational Statistics 22 (2) (2007) 209–222.

[47] A. Kohler, M. Hanafi, D. Bertrand, A. Oust Janbu, T. Møretrø, K. Naderstad, M. Qannari, H. Martens, Interpreting several types of measurements in bioscience, in: P. Lasch, J. Kneipp (Eds.), Modern concepts in biomedical vibrational spectroscopy, John Wiley & Sons, USA, 2008.

[48] S. Wold, S. Hellberg, Y. Lundstedt, M. Sjostrom, H. Wold, Proc. Symp. on PLS Model Building: Theory and Application, Frankfurt am Main, 1987.

[49] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, 1966, pp. 391–420.

[50] Y. Miyashita, T. Itozawa, H. Katsumi, S.-I. Sasaki, Comments on the NIPALS algorithm, Journal of Chemometrics 4 (1) (1990).

[51] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society: Series B: Methodological 36 (2) (1974) 111–147.

[52] H.M.F. Westad, Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression, Journal of Near Infrared Spectroscopy 8 (2) (2000) 117–124.

[53] A. Oust, T. Moretro, K. Naterstad, G.D. Sockalingum, I. Adt, M. Manfait, A. Kohler, Fourier transform infrared and Raman spectroscopy for characterization of *Listeria monocytogenes* strains, Applied and Environmental Microbiology 72 (1) (2006) 228–232.

[54] C.L. Winder, S.V. Gordon, J. Dale, R.G. Hewinson, R. Goodacre, Metabolic fingerprints of *Mycobacterium bovis* cluster with molecular type: implications for genotype-phenotype links, Microbiology 152 (9) (2006) 2757–2765.

[55] A. Kohler, C. Kirschner, A. Oust, H. Martens, Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin, Applied Spectroscopy 59 (6) (2005) 707–716.

[56] A. Kohler, M. Zimonja, V. Segtnan, H. Martens, Data preprocessing: SNV, MSC and EMSC pre-processing in biospectroscopy, in: S. Brown, R. Tauler, R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 139–162.

[57] M. Lafleur, Phase behaviour of model stratum corneum lipid mixtures: an infrared spectroscopy investigation, Canadian Journal of Chemistry 76 (1998) 1501–1511.

Paper II

Chemometrics and Intelligent Laboratory Systems xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Model validation and error estimation in multi-block partial least squares regression

Sahar Hassani [a,b,*], Harald Martens [a,c], El Mostafa Qannari [d], Mohamed Hanafi [d], Achim Kohler [c,a]

[a] *Nofima Mat AS, Osloveien 1, N-1430 Ås, Norway*
[b] *Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway*
[c] *Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway*
[d] *Unité de Sensométrie et de Chimiométrie, ONIRIS/INRA, BP 82225, 44322 Nantes Cedex 3, France*

ARTICLE INFO

Article history:
Received 31 March 2011
Received in revised form 6 June 2011
Accepted 8 June 2011
Available online xxxx

Keywords:
Multi-block methods
Multi-block Partial Least Squares Regression (MBPLSR)
Validation tools
Cross validation
Number of components

ABSTRACT

While validation of Partial Least Squares Regression (PLSR) models has been discussed extensively, validation tools that are tailored to Multi-block Partial Least Squares Regression (MBPLSR) have not been discussed in literature yet. This paper introduces validation tools for estimating predictive ability and model stability in MBPLSR models on block level and on global level. Predictive ability on the block level and global level are estimated by calculating the predictive power of block and global parameters. Model stability is estimated by checking the stability of block model parameters and global parameters. By comparing error plots for model stability and predictive ability the user can decide on the number of component to be used. The number of components to be chosen depends on the data set and the purpose of the investigation.

1. Introduction

Analyzing a multi-block data set can be accomplished by means of different multi-block methods e.g. Consensus Principal Component Analysis (CPCA) and Multi-block Partial Least Squares Regression (MBPLSR). Both methods provide the user with an efficient graphical overview over sample and variable variation patterns between and within the data blocks [1]. Powerful visualization tools provided by these multi-block methods make it easy to interpret the results. However, one should bear in mind that the interpretations made by practitioners on the basis of visual detection of patterns may be misleading. This remark raises the need of formal validation of the outcomes of a multi-block analysis. So far, this topic has not attracted sufficient attention. Recently, we have reported a study concerning the interpretation and validation of visually detected patterns both in the global and block results of CPCA [2]. We intend to undertake a similar study within the framework of MBPLSR. This method of analysis is a prevalent approach in the analysis of multi-block data sets. It is employed in different fields of science e.g. analysis of environmental data sets [3] and spectral data sets [4] [5], modeling of pharmaceutical processes [6] and monitoring complex chemical processes [7].

Validating the MBPLSR model can be studied from two different points of view: 1) Since MBPLSR is a data analytical technique which enables the user to set up predictive models, one possible way of validation is to validate the predictability of the model. 2) As MBPLSR is also a multi-block data analytical technique, it is of paramount interest to validate the contribution of different blocks in the overall model and to assess how the global MBPLSR model is related to each block. In this paper, two validation strategies for MBPLSR are presented and illustrated on the basis of a data set pertaining to a study which aimed at characterizing natural variability in microbiology. New methods for validating the visually identified patterns, both at global and block levels, from the results of MBPLSR are introduced thus allowing the user to formally validate these patterns. The paper is organized as following: In order to introduce block and global parameters of MBPLSR that are used for visualization, MBPLSR is described and its algorithm is given in Section 2.2. Root Mean Square Error of **X** and **Y** ($\mathrm{RMSE_X}$ and $\mathrm{RMSE_Y}$) are calculated for the validation purposes in Section 2.3. Section 3 presents a multi-block data set that has been used as an example in this paper. Global and block score plots which are important visualization tools in MBPLSR are illustrated together with our proposed validation tools in Section 4. We end the paper with a conclusion in Section 5.

2. Theory

2.1. Notation

We follow the notation commonly used in chemometrics, e.g. Martens & Martens in [8]: Matrices and vectors are written as boldface, matrices as upper-case letters and vectors as lower-case letters. By the indices $b = 1,\ldots,B$ we denote blocks of variables, by $m = 1,\ldots,M$

* Corresponding author at: Nofima Mat AS, Norwegian Food Research Institute, Osloveien 1, N-1430 Ås, Norway. Tel.: +47 97594676.
E-mail address: sahar.hassani@nofima.no (S. Hassani).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.06.001

cross-validation segments of samples and by $a = 1,\ldots,A$ the number of PLSR components. By $\mathbf{X} = [\mathbf{X}^1\,\mathbf{X}^2\ldots\mathbf{X}^b\ldots\mathbf{X}^B]$ we denote the multi-block descriptor data set consisting of B blocks. Measurements pertaining to the same measurement technique, e.g. spectroscopy, chromatography and gene fragment analysis profiling are typically collected in the same descriptor block $\mathbf{X}^b$. In omics experiments different measurement techniques are applied to the same samples and, for the multi-block analysis, data need to be ordered in a way that a sample-to-sample (row-to-row) correspondence between the blocks is achieved. Only measurements originating from the same biological replicate can be related to each other by a row-to-row correspondence. If, for example, for different methods several and different biological replicates are used, only means of biological replicates can be related to each others. By $\mathbf{Y}$ we denote the response data set which contains the target variable e.g. phenotype data in the current study. The total number of samples in each data set is represented by N, the total number of variables in a given descriptor block b by K_b, the total number of variables in the descriptor data set ($\mathbf{X}$) by K ($K = \sum_{b=1}^{B} K_b$), and, finally, the total number of variables in the response data set ($\mathbf{Y}$) by J.

2.2. MBPLSR

In order to predict a set of response variables $\mathbf{Y}$ from a multi-block set of descriptor variables $\mathbf{X}$, Multi-block Partial Least Squares Regression (MBPLSR) has been used herein. MBPLSR seeks latent variables, within and between the multi-block set of descriptor variables, which account for most of the variation in $\mathbf{X}$ while, at the same time, predicting the response variables $\mathbf{Y}$ in the best way. The multi-block descriptor data set $\mathbf{X}$ and the response data set $\mathbf{Y}$ are preprocessed prior to MBPLSR calculations and the irrelevant variation types are removed. $\mathbf{X}$ and $\mathbf{Y}$ are also scaled in order to make the data blocks balanced as described in the subsequent section.

Two sets of parameters are produced during the MBPLSR algorithm: super (i.e. global) parameters and block parameters. The super parameters are related to the global model for predicting $\mathbf{Y}$ from a multi-block set ($\mathbf{X}$) which are equivalent to the parameters calculated by running an ordinary PLSR on $\mathbf{Y}$ and concatenated $\mathbf{X}$ blocks. The extra feature of MBPLSR is the calculation of block parameters which gives an insight into the contribution of each block to the model. The MBPLSR parameters are calculated in two main steps: (1) Calculation of super score, super weight, block scores and block loading weights. (2) Deflation: $\mathbf{X}$ and $\mathbf{Y}$ (or only $\mathbf{Y}$) are updated by subtracting the variation already explained by the super score. Steps 1 and 2 are repeated on the deflated matrices for the calculation of subsequent MBPLSR components.

For the calculation of the super score and super weight vector, block score and block loading weight vectors a general MBPLSR algorithm was introduced by Wangen and Kowalski (1988) [9] which was based on the algorithm presented by Wold and Martens (1983) [10]. The MBPLSR algorithm is given below, where the notation according to Westerhuis et al. (1998) [11] is used.

2.2.1. Preprocessing

2.2.1.1. Mean-centering. All of the variables (belonging to both $\mathbf{X}$ and $\mathbf{Y}$) are mean-centered prior to MBPLSR calculations. Mean-centering is performed by subtracting the mean of each variable over all of the samples according to

$$\begin{aligned}\mathbf{X}_{\text{Unscaled}} &= \mathbf{X}_{\text{Input}} - \mathbf{1}\cdot\bar{\mathbf{x}}'_{\text{Input}}\\ \mathbf{Y}_{\text{Unscaled}} &= \mathbf{Y}_{\text{Input}} - \mathbf{1}\cdot\bar{\mathbf{y}}'_{\text{Input}}\end{aligned} \tag{1}$$

where $\mathbf{X}_{\text{Unscaled}}$ and $\mathbf{Y}_{\text{Unscaled}}$ are the mean-centered descriptor data table and response data table respectively, $\mathbf{X}_{\text{Input}}$ and $\mathbf{Y}_{\text{Input}}$ are the original non-centered descriptor data table and response data table respectively, $\mathbf{1}$ is a $N \times 1$ vector of 1s, $\bar{\mathbf{x}}_{\text{Input}}(K \times 1)$ and $\bar{\mathbf{y}}_{\text{Input}}(J \times 1)$ are the vectors of mean values of the variables along the samples in $\mathbf{X}$ and $\mathbf{Y}$ respectively.

2.2.1.2. Scaling. As part of the preprocessing, variables in $\mathbf{X}$ and $\mathbf{Y}$ are scaled block-wise, to balance the sum of square contributions for different blocks. Scaling is attained by dividing the mean-centered data tables by their norms according to:

$$\begin{aligned}\mathbf{X}^b &= \frac{\mathbf{X}^b_{\text{Unscaled}}}{\sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K_b}\left(\mathbf{X}^b_{\text{Unscaled}}(i,k)\right)^2}}\\ \mathbf{Y} &= \frac{\mathbf{Y}_{\text{Unscaled}}}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{J}\left(\mathbf{Y}_{\text{Unscaled}}(i,j)\right)^2}}\end{aligned} \tag{2}$$

where $\mathbf{X}^b$ and $\mathbf{Y}$ are the mean-centered and scaled descriptor data blocks and response data block respectively, $\mathbf{X}^b_{\text{Unscaled}}$ and $\mathbf{Y}_{\text{Unscaled}}$ are the mean-centered and non-scaled descriptor data blocks and response data block, calculated by Eq. (1), respectively. $\mathbf{X}^b_{\text{Unscaled}}(i,k)$ is the $(i,k)th$ entry of $\mathbf{X}^b_{\text{Unscaled}}$ and $\mathbf{Y}_{\text{Unscaled}}(i,j)$ is the $(i,j)th$ entry of $\mathbf{Y}_{\text{Unscaled}}$. By $i = 1,\ldots,N$ we denote samples, by $k = 1,\ldots,K_b$ we denote the variables of $\mathbf{X}^b$ and by $j = 1,\ldots,J$ we denote the variables of $\mathbf{Y}$.

The purpose of scaling is to set all of the blocks on the same footing (i.e. the same total variance in every block) so that the number of variables or the measurement unit of a certain block will not have any influence on the MBPLSR model. However, it should be noted that this scaling is flexible: in the sense that different scaling factors could be introduced according to the user's aim. For instance, one may wish to scale the blocks in such a way that a particular block dictates its variation pattern to the overall MBPLSR model or, contrariwise, the scaling can prevent a given block to influence the overall model. For instance, a common situation which motivates to drastically down-weight a block of variables occurs when dealing with the design block. In order to avoid that the design of the study influences the results of MBPLSR model, one can down-weight the design block by scaling it with a very small number (e.g. 0.000001). The advantage of down-weighting the design block with such a small number and still keeping it in the model calculations, is to investigate its relationships with the other variables. Similarly, individual variables in any block can be "down-weighted" by down-scaling them (Martens & Martens in [8]), but this is not used herein.

2.2.2. Overall modeling

In order to explore the systematic variation patterns in $\mathbf{X}$ which are likely to predict the systematic variation patterns in $\mathbf{Y}$, MBPLSR is applied. Descriptor data tables $\mathbf{X}$ and response data table $\mathbf{Y}$ are modeled as sum of A latent variables plus residual matrices $\mathbf{E}$ and $\mathbf{F}$ respectively. The MBPLSR model for mean-centered and scaled data is as below

$$\begin{aligned}\mathbf{X} &= \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A\\ \mathbf{X}^b &= \mathbf{T}_A\mathbf{P}^{b\prime}_A + \mathbf{E}^b_A\\ \mathbf{Y} &= \mathbf{T}_A\mathbf{Q}'_A + \mathbf{F}_A\\ \mathbf{Y} &= \mathbf{X}\mathbf{B}_A + \mathbf{F}_A\end{aligned} \tag{3}$$

where $\mathbf{X} = \left[\mathbf{X}^1\,\mathbf{X}^2\ldots\mathbf{X}^b\ldots\mathbf{X}^B\right]$ is the matrix of concatenated mean-centered and scaled descriptor data blocks ($\mathbf{X}^b$) in Eq. (2), $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_a, \ldots, \mathbf{t}_A]$ is the matrix of A super score vectors $\mathbf{t}_a$ and $\mathbf{P}_A = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_a, \ldots, \mathbf{p}_A]$ is the corresponding matrix of A loading vectors associated with $\mathbf{X}$. The $\mathbf{X}$ loading matrix $\mathbf{P}$ can also be written

Please cite this article as: S. Hassani, et al., Model validation and error estimation in multi-block partial least squares regression, Chemometr. Intell. Lab. Syst. (2011), doi:10.1016/j.chemolab.2011.06.001

as the matrix of concatenated block loading matrices $\mathbf{P}^b$: $\mathbf{P}' = \left[\mathbf{P}^{1'}\,\mathbf{P}^{2'}\ldots\mathbf{P}^{b'}\ldots\mathbf{P}^{B'}\right]$. $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_a, \ldots, \mathbf{q}_A]$ contains A loading vectors of $\mathbf{Y}$ ($\mathbf{q}_a$), and $\mathbf{B}_A$ is the $K \times J$ matrix of regression coefficients derived from regressing $\mathbf{Y}$ upon the super scores $\mathbf{T}_A$.

2.2.3. Component estimation

Multi-block PLSR is an extension of PLSR with the possibility for the researcher to add additional knowledge to the data by dividing them into meaningful data blocks. This division of data into blocks is likely to give more insight into the data, since variables can be grouped according to the scientist's *a priori* knowledge about the variables. This may enhance the interpretation of variation patterns that several data blocks have in common. Several extensions and variations of PLSR to more than one descriptor block have been proposed so far [12] [13] [14] [15] [16]. For this study the MBPLSR algorithm of Wangen and Kowalski (1988) [9] which can handle most types of relationships between the blocks is used. The algorithm is given below.

The following procedure is performed for each PLSR component $a = 1, 2, \ldots$:

A. Initialization
- 1.1 Choose an arbitrary starting $\mathbf{Y}$ score vector, $\mathbf{u}$

B. Computation of $\mathbf{X}$ block scores and block loading weights
- 1.2 $\tilde{\mathbf{w}}^b = \frac{\mathbf{X}^{b'}\mathbf{u}}{\mathbf{u}'\mathbf{u}}$ $\mathbf{X}$ block loading weights
- 1.3 $\mathbf{t}^b = \mathbf{X}^b\tilde{\mathbf{w}}^b\left(\tilde{\mathbf{w}}^{b'}\tilde{\mathbf{w}}^b\right)^{-1}$ $\mathbf{X}$ Block scores

C. Computation of super scores and super weights, $\mathbf{Y}$ scores and $\mathbf{Y}$ weights
- 1.4 $\mathbf{T} = \left[\mathbf{t}^1\,\mathbf{t}^2\ldots\mathbf{t}^b\ldots\mathbf{t}^B\right]$
- 1.5 $\tilde{\mathbf{w}}_s = \frac{\mathbf{T}'\mathbf{u}}{\mathbf{u}'\mathbf{u}}$ $\mathbf{X}$ super weight (relative to the contribution of each block)
- 1.6 $\mathbf{t} = \mathbf{T}\tilde{\mathbf{w}}_s\left(\tilde{\mathbf{w}}_s'\tilde{\mathbf{w}}_s\right)^{-1}$ $\mathbf{X}$ super scores
- 1.7 $\mathbf{q} = \frac{\mathbf{Y}'\mathbf{t}}{\mathbf{t}'\mathbf{t}}$ $\mathbf{Y}$ loading weights
- 1.8 $\mathbf{u} = \frac{\mathbf{Y}\mathbf{q}}{\mathbf{q}'\mathbf{q}}$ $\mathbf{Y}$ scores

D. Replacing the $\mathbf{Y}$ score vector $\mathbf{u}$ by the updated vector of $\mathbf{Y}$ scores in 1.8 and iterating until convergence of the algorithm (i.e. no significant change in super scores $\mathbf{t}$).

E. Computation of Regression Coefficients
- 1.9 $\mathbf{p}_a^b = \frac{\mathbf{X}^{b'}\mathbf{t}}{\mathbf{t}'\mathbf{t}}$ $\mathbf{X}$ Block loadings
- 1.10 $\mathbf{p}_a' = \left[\mathbf{p}_a^{1'}\,\mathbf{p}_a^{2'}\ldots\mathbf{p}_a^{b'}\ldots\mathbf{p}_a^{B'}\right]$ $\mathbf{X}$ *a*th Super Loadings
- 1.11 $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_a]$
- 1.12 $\tilde{\mathbf{w}}_a = \left[\tilde{\mathbf{w}}_a^1\,\tilde{\mathbf{w}}_a^2\ldots\tilde{\mathbf{w}}_a^b\ldots\tilde{\mathbf{w}}_a^B\right]$
- 1.13 $\mathbf{w}_a = \frac{\tilde{\mathbf{w}}_a}{\|\tilde{\mathbf{w}}_a\|}$
- 1.14 $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_a]$
- 1.15 $\mathbf{V} = \mathbf{W}(\mathbf{P}'W)^{-1}$
- 1.16 $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_a]$
- 1.17 $\mathbf{B}_a = \mathbf{V}\mathbf{Q}'$ Regression coefficients

F. Deflation on super scores
- 1.18 $\mathbf{X}_a = \mathbf{X}$ and $\mathbf{Y}_a = \mathbf{Y}$
- 1.19 $\mathbf{X} = \mathbf{X}_a - \mathbf{t}\mathbf{p}_a'$ $\mathbf{X}$ deflation
- 1.20 $\mathbf{Y} = \mathbf{Y}_a - \mathbf{t}\mathbf{q}'$ $\mathbf{Y}$ deflation

Alternatively, the same solution for super scores and weights may be obtained by performing PLSR on $\mathbf{X} = \left[\mathbf{X}^1\,\mathbf{X}^2\ldots\mathbf{X}^B\right]$ and $\mathbf{Y}$. Thereafter, block loading weights, block scores and block loadings may be computed according to their definition in the multi-block algorithm given above.

Deflating both $\mathbf{X}$ and $\mathbf{Y}$ on super score is one way of deflation in MBPLSR. Other ways of deflation are available and can be applied depending on the goal of the study [17]. Westerhuis and Smilde proposed that instead of deflating $\mathbf{X}$ and $\mathbf{Y}$, only $\mathbf{Y}$ should be deflated using the super score [18]. The validation procedures proposed in this paper will be illustrated using the algorithm proposed by Wangen and Kowalski (1988). All validation procedures can be easily transferred to other deflation types.

2.3. Error estimation and cross-validation

We propose two different strategies for validating MBPLSR models and selecting the appropriate number of latent variables to be included in the model: (1) Model stability and co-variation patterns in the X-matrix are validated by computing the Root Mean Square Error (RMSE) for the explanatory data set ($\mathbf{X}$). (2) The prediction ability of the main common variation pattern in $\mathbf{X}$ and main variation patterns in the different blocks are validated by computing the Root Mean Square Error (RMSE) for the response data set ($\mathbf{Y}$). When the RMSE is calculated for explanatory data ($\mathrm{RMSE_X}$) it highlights the contribution of every data block in the MBPLSR model for each PLSR component. The $\mathrm{RMSE_X}$ calculation is described in Section 2.3.1. When RMSE is calculated for response data ($\mathrm{RMSE_Y}$), it highlights the prediction ability of each explanatory data block for predicting response data table $\mathbf{Y}$. The $\mathrm{RMSE_Y}$ calculation is described in Section 2.3.2.

$\mathrm{RMSE_X}$ and $\mathrm{RMSE_Y}$ are calculated by cross-validation [19]: Segments (indexed by $m = 1, \ldots, M$) are formed by leaving, in turn, one or several samples out, resulting in left-out segments of explanatory data ($\mathbf{X}_m$) and response data ($\mathbf{Y}_m$), and leave-in segments of explanatory data ($\mathbf{X}_{-m}$) and response data ($\mathbf{Y}_{-m}$). During each round of cross-validation MBPLSR models are calculated for the leave-in data ($\mathbf{X}_{-m}$ and $\mathbf{Y}_{-m}$). These models are afterwards fitted to the left-out samples in order to calculate $\mathrm{RMSE_X}$ (or $\mathrm{RMSE_Y}$) of the left-out segments.

2.3.1. $RMSE_X$: RMSE calculation for X

The predicted values for the left-out explanatory data ($\hat{\mathbf{X}}_m = \left[\hat{\mathbf{X}}_m^1\;\hat{\mathbf{X}}_m^2\ldots\hat{\mathbf{X}}_m^b\ldots\hat{\mathbf{X}}_m^B\right]$) are assessed by applying the block loading weights for an A-dimensional model of the leave-in samples ($\tilde{\mathbf{W}}'_{-m,A} = \left[\tilde{\mathbf{W}}'^1_{-m,A}\,\tilde{\mathbf{W}}'^2_{-m,A}\ldots\tilde{\mathbf{W}}'^b_{-m,A}\ldots\tilde{\mathbf{W}}'^B_{-m,A}\right]$ calculated according to step 1.2 of the MBPLSR algorithm in Section 2.2.3) to the left-out data. The difference between the true values of the left-out data ($\mathbf{X}_m$) and their predicted values ($\hat{\mathbf{X}}_m$) is collected in the segment m of the residual matrix $\mathbf{E}$ ($\mathbf{E}_m$). This procedure is repeated for all of the segments of the data and for every $A = 0, 1, \ldots, A_{\max}$ which results in $A_{\max}$ residual matrices ($\mathbf{E}_A$).

In Fig. 1 the different steps of the validation procedure are shown and the calculations are coming in the following:

Prediction of $\hat{\mathbf{X}}_m$ requires the calculation of $\mathbf{X}$ block weights, $\tilde{\mathbf{W}}'^b_{-m,A}, b = 1, \ldots, B$ from the MBPLSR model of leave-in samples. Leave-in data is mean-centered prior to MBPLSR calculation according to

$$\begin{aligned}\mathbf{X}_{\mathbf{C}_{-m}} &= \mathbf{X}_{-\mathbf{m}} - \mathbf{1}\cdot\bar{\mathbf{x}}'_{-m} \\ \mathbf{Y}_{\mathbf{C}_{-m}} &= \mathbf{Y}_{-\mathbf{m}} - \mathbf{1}\cdot\bar{\mathbf{y}}'_{-m}\end{aligned} \tag{4}$$

where $\mathbf{X}_{\mathbf{C}_{-m}}$ and $\mathbf{Y}_{\mathbf{C}_{-m}}$ represent the mean-centered leave-in values of explanatory and response data sets respectively. $\mathbf{X}_{-m} = \left[\mathbf{X}^1_{-m}\,\mathbf{X}^2_{-m}\ldots\mathbf{X}^b_{-m}\ldots\mathbf{X}^B_{-m}\right]$ and $\mathbf{Y}_{-m}$ are leave-in samples before mean-centering. $\bar{\mathbf{x}}_{-m}$ is a $K \times 1$ vector of the means along the samples of $\mathbf{X}_{-m}$ and $\bar{\mathbf{y}}_{-m}$ is a $J \times 1$ vector of the means along the samples of $\mathbf{Y}_{-m}$.

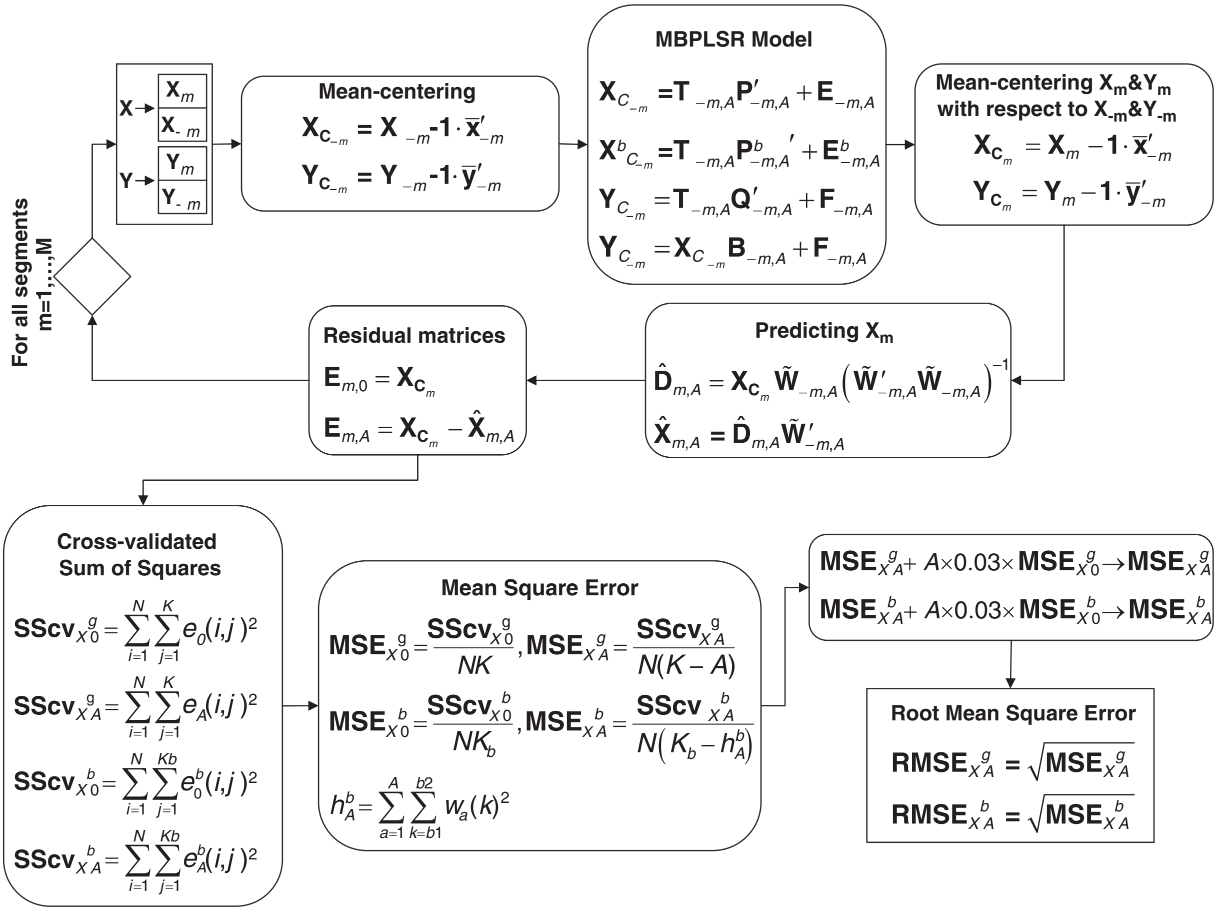

For all segments
m=1,...,M
X
$\mathbf{X}_m$
$\mathbf{X}_{-m}$
Y
$\mathbf{Y}_m$
$\mathbf{Y}_{-m}$
Mean-centering
$\mathbf{X}_{C_{-m}} = \mathbf{X}_{-m} - \mathbf{1}\cdot\bar{\mathbf{x}}'_{-m}$
$\mathbf{Y}_{C_{-m}} = \mathbf{Y}_{-m} - \mathbf{1}\cdot\bar{\mathbf{y}}'_{-m}$
MBPLSR Model
$\mathbf{X}_{C_{-m}} = \mathbf{T}_{-m,A}\mathbf{P}'_{-m,A} + \mathbf{E}_{-m,A}$
$\mathbf{X}^b_{C_{-m}} = \mathbf{T}_{-m,A}\mathbf{P}^{b\,\prime}_{-m,A} + \mathbf{E}^b_{-m,A}$
$\mathbf{Y}_{C_{-m}} = \mathbf{T}_{-m,A}\mathbf{Q}'_{-m,A} + \mathbf{F}_{-m,A}$
$\mathbf{Y}_{C_{-m}} = \mathbf{X}_{C_{-m}}\mathbf{B}_{-m,A} + \mathbf{F}_{-m,A}$
Mean-centering $\mathbf{X}_m$&$\mathbf{Y}_m$ with respect to $\mathbf{X}_{-m}$&$\mathbf{Y}_{-m}$
$\mathbf{X}_{C_m} = \mathbf{X}_m - \mathbf{1}\cdot\bar{\mathbf{x}}'_{-m}$
$\mathbf{Y}_{C_m} = \mathbf{Y}_m - \mathbf{1}\cdot\bar{\mathbf{y}}'_{-m}$
Predicting $\mathbf{X}_m$
$\hat{\mathbf{D}}_{m,A} = \mathbf{X}_{C_m}\tilde{\mathbf{W}}_{-m,A}\left(\tilde{\mathbf{W}}'_{-m,A}\tilde{\mathbf{W}}_{-m,A}\right)^{-1}$
$\hat{\mathbf{X}}_{m,A} = \hat{\mathbf{D}}_{m,A}\tilde{\mathbf{W}}'_{-m,A}$
Residual matrices
$\mathbf{E}_{m,0} = \mathbf{X}_{C_m}$
$\mathbf{E}_{m,A} = \mathbf{X}_{C_m} - \hat{\mathbf{X}}_{m,A}$
Cross-validated Sum of Squares
$\mathbf{SScv}^g_{X0} = \sum_{i=1}^{N}\sum_{j=1}^{K} e_0(i,j)^2$
$\mathbf{SScv}^g_{XA} = \sum_{i=1}^{N}\sum_{j=1}^{K} e_A(i,j)^2$
$\mathbf{SScv}^b_{X0} = \sum_{i=1}^{N}\sum_{j=1}^{Kb} e^b_0(i,j)^2$
$\mathbf{SScv}^b_{XA} = \sum_{i=1}^{N}\sum_{j=1}^{Kb} e^b_A(i,j)^2$
Mean Square Error
$\mathbf{MSE}^g_{X0} = \frac{\mathbf{SScv}^g_{X0}}{NK}, \mathbf{MSE}^g_{XA} = \frac{\mathbf{SScv}^g_{XA}}{N(K-A)}$
$\mathbf{MSE}^b_{X0} = \frac{\mathbf{SScv}^b_{X0}}{NK_b}, \mathbf{MSE}^b_{XA} = \frac{\mathbf{SScv}^b_{XA}}{N\left(K_b - h^b_A\right)}$
$h^b_A = \sum_{a=1}^{A}\sum_{k=b1}^{b2} w_a(k)^2$
$\mathbf{MSE}^g_{XA} + A\times 0.03\times\mathbf{MSE}^g_{X0} \rightarrow \mathbf{MSE}^g_{XA}$
$\mathbf{MSE}^b_{XA} + A\times 0.03\times\mathbf{MSE}^b_{X0} \rightarrow \mathbf{MSE}^b_{XA}$
Root Mean Square Error
$\mathbf{RMSE}^g_{XA} = \sqrt{\mathbf{MSE}^g_{XA}}$
$\mathbf{RMSE}^b_{XA} = \sqrt{\mathbf{MSE}^b_{XA}}$

Fig. 1. Flow chart of the $\mathrm{RMSE_X}$ calculation. Square boxes are input and output, diamond is for-loop and rounded boxes are instructions. The following indices are used: (i,j) for the (i,j)-th entry of the respective matrix, N for the total number of samples, K for the total number of variables in explanatory data set $\mathbf{X}$, K_b for the number of variables in block b of explanatory data set $\mathbf{X}^b$, A for the number of PLSR components, $m = 1,\ldots,M$ for the cross-validation segments of samples, $b = 1,\ldots,B$ for the blocks of data and 'g' for the global.

Eq. (5) shows the MBPLSR model of $\mathbf{X}_{\mathbf{C}_{-m}}$ and $\mathbf{Y}_{\mathbf{C}_{-m}}$:

$$\begin{aligned}\mathbf{X}_{C_{-m}} &= \mathbf{T}_{-m,A}\mathbf{P}'_{-m,A} + \mathbf{E}_{-m,A}\\ \mathbf{X}^b_{C_{-m}} &= \mathbf{T}_{-m,A}\mathbf{P}^{b\,\prime}_{-m,A} + \mathbf{E}^b_{-m,A}\\ \mathbf{Y}_{C_{-m}} &= \mathbf{T}_{-m,A}\mathbf{Q}'_{-m,A} + \mathbf{F}_{-m,A}\end{aligned} \tag{5}$$

where $\mathbf{X}_{\mathbf{C}_{-m}} = \left[\mathbf{X}^1_{\mathbf{C}_{-m}}\ \mathbf{X}^2_{\mathbf{C}_{-m}}\ldots\mathbf{X}^b_{\mathbf{C}_{-m}}\ldots\mathbf{X}^{\mathrm{B}}_{\mathbf{C}_{-m}}\right]$; $\mathbf{T}_{-m,A}$, $\mathbf{P}'_{-m,A} = \left[\mathbf{P}^1_{-m,A}{}'\ \mathbf{P}^2_{-m,A}{}'\ldots\mathbf{P}^b_{-m,A}{}'\ldots\mathbf{P}^B_{-m,A}{}'\right]$ and $\mathbf{Q}'_{-m,A}$ contain X-scores, X-loadings and Y-weights (having A PLSR components in the model) of the leave-in samples respectively. $\mathbf{E}_{-m,A} = \left[\mathbf{E}^1_{-m,A}\ \mathbf{E}^2_{-m,A}\ldots\mathbf{E}^b_{-m,A}\ldots\mathbf{E}^B_{-m,A}\right]$ and $\mathbf{F}_{-m,A}$ are segments m of the residual matrices $\mathbf{E}_A$ and $\mathbf{F}_A$ respectively.

The predicted values for the left-out segment of $\mathbf{X}$ are calculated in the following way. At first the left-out data is centered using the means of the leave-in samples according to Eq. (6):

$$\begin{aligned}\mathbf{X}_{\mathbf{C}_m} &= \mathbf{X}_m - \mathbf{1}\cdot\bar{\mathbf{x}}'_{-m}\\ \mathbf{Y}_{\mathbf{C}_m} &= \mathbf{Y}_m - \mathbf{1}\cdot\bar{\mathbf{y}}'_{-m}\end{aligned} \tag{6}$$

The projection scores for the left-out samples are then predicted by applying the loading weights of $\mathbf{X}$ ($\tilde{\mathbf{W}}'_{-m,A} = [\tilde{\mathbf{W}}'^1_{-m,A}\ \tilde{\mathbf{W}}'^2_{-m,A}\ldots\tilde{\mathbf{W}}'^b_{-m,A}\ldots\tilde{\mathbf{W}}'^B_{-m,A}]$), which are estimated from the leave-in samples, on the mean-centered left-out data ($\mathbf{X}_{\mathbf{C}_m}$):

$$\hat{\mathbf{D}}_{m,A} = \mathbf{X}_{\mathbf{C}_m}\tilde{\mathbf{W}}_{-m,A}\left(\tilde{\mathbf{W}}'_{-m,A}\tilde{\mathbf{W}}_{-m,A}\right)^{-1} \tag{7}$$

Projecting the left-out samples on all the A loading weights in $\tilde{\mathbf{W}}_{-m,A}$ simultaneously, by a sequence of individual loading weights, followed by deflation, results in simple projection scores $\hat{\mathbf{D}}_{m,A}$ in the sample space of the left-out data $\mathbf{X}_{\mathbf{C}_m}$. These are equivalent to how the scores are calculated and presented in Martens' PLSR algorithm [20]. Finally, estimation for the left-out data ($\hat{\mathbf{X}}_{m,A}$) is calculated from the projection scores and loading weights according to Eq. (8):

$$\hat{\mathbf{X}}_{m,A} = \hat{\mathbf{D}}_{m,A}\tilde{\mathbf{W}}'_{-m,A} \tag{8}$$

where $\hat{\mathbf{D}}_{m,A} = \left[\hat{\mathbf{d}}_{m,1}, \hat{\mathbf{d}}_{m,2}, \ldots, \hat{\mathbf{d}}_{m,a}, \ldots, \hat{\mathbf{d}}_{m,A}\right]$ contains A projection score vectors for the left-out samples and $\tilde{\mathbf{W}}'_{-m,A} = [\tilde{\mathbf{W}}'^1_{-m,A}\ \tilde{\mathbf{W}}'^2_{-m,A}\ldots\tilde{\mathbf{W}}'^b_{-m,A}\ldots\tilde{\mathbf{W}}'^B_{-m,A}]$ is the matrix of concatenated block loading weights of the leave-in samples.

Residual matrices for the left-out samples of the explanatory data ($\mathbf{E}_m$) are calculated as

$$\begin{aligned}\mathbf{E}_{m,0} &= \mathbf{X}_{\mathbf{C}_m}\\ \mathbf{E}_{m,A} &= \mathbf{X}_{\mathbf{C}_m} - \hat{\mathbf{X}}_{m,A}\end{aligned} \tag{9}$$

where $\mathbf{X}_{\mathbf{C}_m}$ is the mean-centered segment m of the explanatory data which was calculated by Eq. (6), $\hat{\mathbf{X}}_{m,A}$ is the prediction for that segment based on A latent variables, $\mathbf{E}_{m,0} = \left[\mathbf{E}^1_{m,0}\ \mathbf{E}^2_{m,0}\ldots\mathbf{E}^b_{m,0}\ldots\mathbf{E}^B_{m,0}\right]$ is the initial residual matrix for segment m and $\mathbf{E}_{m,A} = \left[\mathbf{E}^1_{m,A}\ \mathbf{E}^2_{m,A}\ldots\mathbf{E}^b_{m,A}\ldots\mathbf{E}^B_{m,A}\right]$ is the corresponding residual matrix for segment m.

The predictions for the left-out samples ($\hat{\mathbf{X}}_{m,A}$) were estimated for various values of $A = 0, 1, \ldots, A_{\max}$ ($A_{\max}$ was chosen sufficiently large) resulting in $A_{\max}$ different residual matrices for each segment (i.e.

$\mathbf{E}_{m,1}, \mathbf{E}_{m,2}, \ldots, \mathbf{E}_{m,A}, \ldots, \mathbf{E}_{m,A\,max}$). These matrices $\mathbf{E}_{m,A}$ were concatenated vertically for $m=1,\ldots,M$ resulting in a residual matrix $\mathbf{E}_A = \left[\mathbf{E}_A^1\,\mathbf{E}_A^2\ldots\mathbf{E}_A^b\ldots\mathbf{E}_A^B\right]$ for every value of A. The residual matrices $\mathbf{E}_{m,0}$ were concatenated in the same way yielding the initial residual matrix $\mathbf{E}_0 = \left[\mathbf{E}_0^1\,\mathbf{E}_0^2\ldots\mathbf{E}_0^b\ldots\mathbf{E}_0^B\right]$.

The reason why we project on loading weights in Eq. (7) and not on loadings is because loading weights have – as can be seen from the NIPALS algorithm for multi-block PLSR in Section 2.2.3 – a direct relation to the block scores which are used for visualization in block score plots. While the block score plots visualize the sample variation pattern in each block which is related to the variable variation pattern in the block loading weights, the block model residuals represent the cross-validated residuals using the same loading weights as model parameters.

2.3.1.1. Global errors for ***X***. Global errors calculation is performed as in [2]. This is briefly recalled in the following.

Cross-validated sum of squares for **X** was calculated from the residual matrices according to

$$\begin{aligned} SScv^g_{X\,0} &= \sum_{i=1}^{N}\sum_{j=1}^{K} e_0(i,j)^2 \\ SScv^g_{XA} &= \sum_{i=1}^{N}\sum_{j=1}^{K} e_A(i,j)^2 \end{aligned} \tag{10}$$

where $e_0(i,j)$ and $e_A(i,j)$ are the (i,j)-*th* entry of residual matrices $\mathbf{E}_0$ and $\mathbf{E}_A$ respectively.

Thereafter, mean square errors were calculated by correcting the sum of squares for their approximate degrees of freedom according to

$$\begin{aligned} MSE^g_{X\,0} &= \frac{SScv^g_{X0}}{NK} \\ MSE^g_{XA} &= \frac{SScv^g_{XA}}{N(K-A)} \end{aligned} \tag{11}$$

Eq. (11) relies on the assumptions of independent sampling of the N objects and independent measurement errors for all K variables. Since these assumptions are not satisfied in practice, MSE$_X$ values can be regarded as approximations.

In order to display RMSE$_X$ plots that are used for estimating the optimal model rank, the mean square errors corresponding to the global model were augmented by 3% of the initial mean square error MSE^g_{X0} for each latent variable added to the model [20]:

$$MSE^g_{XA} + A \times 0.03 \times MSE^g_{X0} \rightarrow MSE^g_{XA} \tag{12}$$

This 3% rule is a rule of thumb strategy to estimate the number of latent variables in the model and is shown to be statistically stable. The resulting RMSE plots, give conservative estimations for the optimal rank of the model.

Finally, cross-validated RMSE$_X$ was determined according to Eq. (13):

$$RMSE^g_{XA} = \sqrt{MSE^g_{XA}} \tag{13}$$

2.3.1.2. Block errors for ***X***. Calculating block errors for **X** is done in a similar way as it was done globally. It just varies in the sense that every calculation is performed block-wise and repeated for all of the blocks. Cross-validated sum of squares for every block of **X** is calculated according to:

$$\begin{aligned} SScv^b_{X0} &= \sum_{i=1}^{N}\sum_{j=1}^{Kb} e^b_0(i,j)^2 \\ SScv^b_{XA} &= \sum_{i=1}^{N}\sum_{j=1}^{Kb} e^b_A(i,j)^2 \end{aligned} \tag{14}$$

where $e^b_0(i,j)$ and $e^b_A(i,j)$ are the (i,j)-*th* entry of block b of the residual matrices $\mathbf{E}^b_0$ and $\mathbf{E}^b_A$ respectively.

Mean square errors of **X** for each block b are then calculated from correcting the sum of squares of each block by the approximate degrees of freedom consumed by that block:

$$\begin{aligned} MSE^b_{X0} &= \frac{SScv^b_{X0}}{NK_b} \\ MSE^b_{XA} &= \frac{SScv^b_{XA}}{N(K_b-h^b_A)} \end{aligned} \tag{15}$$

where h^b_A is the partial block leverage intending to represent the contribution of block b in consuming A degrees of freedom for predicting A global scores $\mathbf{t}_a$. h^b_A is calculated according to Eq. (16):

$$\begin{aligned} \tilde{\mathbf{w}}'_a &= \left[\tilde{\mathbf{w}}'^1_a\,\tilde{\mathbf{w}}'^2_a\ldots\tilde{\mathbf{w}}'^b_a\ldots\tilde{\mathbf{w}}'^B_a\right] \\ \mathbf{w}_a &= \frac{\tilde{\mathbf{w}}_a}{||\tilde{\mathbf{w}}_a||} \\ h^b_A &= \sum_{a=1}^{A}\sum_{k=b1}^{b2} w_a(k)^2 \end{aligned} \tag{16}$$

where $\tilde{\mathbf{w}}'^b_a$ are block loading weights which were calculated in step 1.2 of the MBPLSR algorithm given in Section 2.2.3, $w_a(k)$ is the (k)-*th* entry of the normalized loading weight matrix $\mathbf{w}_a$. b_1 and b_2 are the numbers of the first and the last variables in block b respectively in the concatenated set of explanatory variables.

For the rank selection as in the global error, 3% of the initial variance in each block is added to the mean square error of that block for each new latent variable:

$$MSE^b_{XA} + A \times 0.03 \times MSE^b_{X0} \rightarrow MSE^b_{XA}. \tag{17}$$

The cross-validated RMSE$_X$ corresponding to each block is then calculated as:

$$RMSE^b_{XA} = \sqrt{MSE^b_{XA}}. \tag{18}$$

Plotting approximate global and block RMSE$_X$s against the number of latent variables, visualizes the contribution of each block to the MBPLSR model. The block patterns can also be compared to the global pattern. In addition, one can decide the number of relevant PLSR components to be retained in the MBPLSR model (A_{opt}), in the whole data set and also in each block individually. Fig. 1 shows the flow chart of the approximate RMSE$_X$ calculations.

2.3.2. RMSE$_Y$: RMSE calculation for ***Y***

Response values of the left-out samples $\hat{\mathbf{Y}}_m$ can be predicted by applying the prediction model for the leave-in samples on the left-out samples. The difference between the true response value for the left-out data ($\mathbf{Y}_m$) and the predicted value ($\hat{\mathbf{Y}}_m$) is stored in the appropriate segment of the residual matrix ($\mathbf{F}_m$). This procedure is repeated for all of the cross-validation segments and for every $A=0,1,\ldots,A_{\max}$ resulting in $A_{\max}$ residual matrices ($\mathbf{F}_A$). Fig. 2 visualizes the procedure. The detailed calculations are explained in the following.

Predicting response values for the left-out segment ($\hat{\mathbf{Y}}_m$) requires using the regression coefficients of the model built on the leave-in samples. MBPLSR model for the mean-centered leave-in data ($\mathbf{X}_{\mathbf{C}_{-m}}$

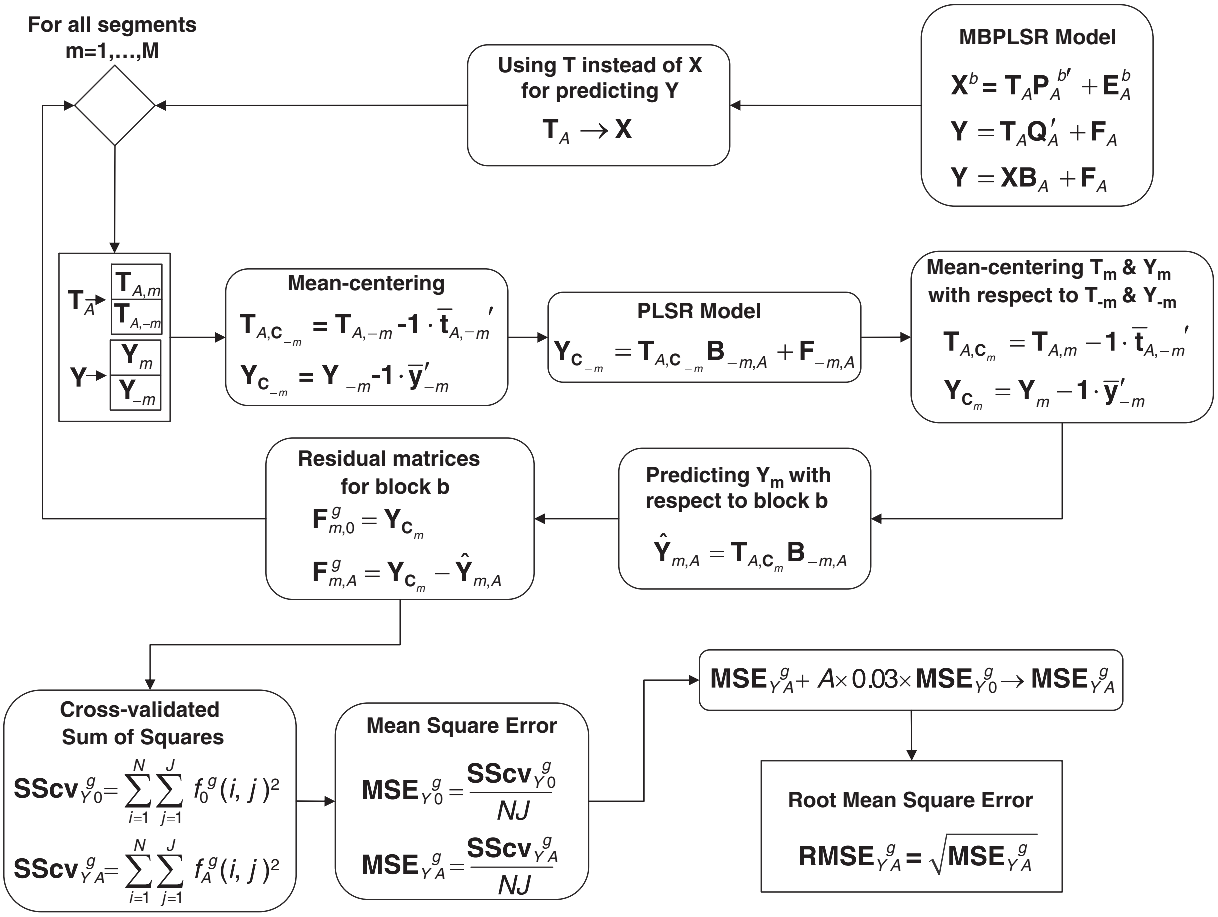

For all segments m=1,…,M
Using T instead of X for predicting Y
$\mathbf{T}_A \rightarrow \mathbf{X}$
MBPLSR Model
$\mathbf{X}^b = \mathbf{T}_A\mathbf{P}_A^{b\prime} + \mathbf{E}_A^b$
$\mathbf{Y} = \mathbf{T}_A\mathbf{Q}_A' + \mathbf{F}_A$
$\mathbf{Y} = \mathbf{X}\mathbf{B}_A + \mathbf{F}_A$
$\mathbf{T}_A$ $\mathbf{T}_{A,m}$ $\mathbf{T}_{A,-m}$
$\mathbf{Y}$ $\mathbf{Y}_m$ $\mathbf{Y}_{-m}$
Mean-centering
$\mathbf{T}_{A,\mathbf{C}_{-m}} = \mathbf{T}_{A,-m} - \mathbf{1}\cdot\bar{\mathbf{t}}_{A,-m}'$
$\mathbf{Y}_{\mathbf{C}_{-m}} = \mathbf{Y}_{-m} - \mathbf{1}\cdot\bar{\mathbf{y}}_{-m}'$
PLSR Model
$\mathbf{Y}_{\mathbf{C}_{-m}} = \mathbf{T}_{A,\mathbf{C}_{-m}}\mathbf{B}_{-m,A} + \mathbf{F}_{-m,A}$
Mean-centering $\mathbf{T}_m$ & $\mathbf{Y}_m$ with respect to $\mathbf{T}_{-m}$ & $\mathbf{Y}_{-m}$
$\mathbf{T}_{A,\mathbf{C}_m} = \mathbf{T}_{A,m} - \mathbf{1}\cdot\bar{\mathbf{t}}_{A,-m}'$
$\mathbf{Y}_{\mathbf{C}_m} = \mathbf{Y}_m - \mathbf{1}\cdot\bar{\mathbf{y}}_{-m}'$
Residual matrices for block b
$\mathbf{F}_{m,0}^g = \mathbf{Y}_{\mathbf{C}_m}$
$\mathbf{F}_{m,A}^g = \mathbf{Y}_{\mathbf{C}_m} - \hat{\mathbf{Y}}_{m,A}$
Predicting $\mathbf{Y}_m$ with respect to block b
$\hat{\mathbf{Y}}_{m,A} = \mathbf{T}_{A,\mathbf{C}_m}\mathbf{B}_{-m,A}$
$\mathbf{MSE}_{YA}^g + A\times 0.03\times\mathbf{MSE}_{Y0}^g \rightarrow \mathbf{MSE}_{YA}^g$
Cross-validated Sum of Squares
$\mathbf{SScv}_{Y0}^g = \sum_{i=1}^{N}\sum_{j=1}^{J} f_0^g(i,j)^2$
$\mathbf{SScv}_{YA}^g = \sum_{i=1}^{N}\sum_{j=1}^{J} f_A^g(i,j)^2$
Mean Square Error
$\mathbf{MSE}_{Y0}^g = \frac{\mathbf{SScv}_{Y0}^g}{NJ}$
$\mathbf{MSE}_{YA}^g = \frac{\mathbf{SScv}_{YA}^g}{NJ}$
Root Mean Square Error
$\mathbf{RMSE}_{YA}^g = \sqrt{\mathbf{MSE}_{YA}^g}$

Fig. 2. Flow chart of the global RMSE_Y calculation. Square boxes are input and output, diamond is for-loop and rounded boxes are instructions. The following indices are used: (i,j) for the (i,j)-th entry of the respective matrix, J for the number of variables in response data set $\mathbf{Y}$, N for the total number of samples, $m = 1,\ldots,M$ for the cross-validation segments of samples, A for the number of PLSR components and 'g' for the global.

and $\mathbf{Y}_{\mathbf{C}_{-m}}$) was given in Eq. (5). Regression coefficients calculated from that MBPLSR model satisfy Eq. (19):

$$\mathbf{Y}_{\mathbf{C}_{-m}} = \mathbf{X}_{\mathbf{C}_{-m}}\mathbf{B}_{-m,A} + \mathbf{F}_{-m,A} \tag{19}$$

where $\mathbf{X}_{\mathbf{C}_{-m}} = \left[\mathbf{X}^1_{\mathbf{C}_{-m}}\mathbf{X}^2_{\mathbf{C}_{-m}}\ldots\mathbf{X}^b_{\mathbf{C}_{-m}}\ldots\mathbf{X}^B_{\mathbf{C}_{-m}}\right]$ and $\mathbf{Y}_{\mathbf{C}_{-m}}$ are leave-in data which are mean-centered according to Eq. (4), $\mathbf{B}_{-m,A}$ is a $K\times J$ matrix which contains regression coefficients for the MBPLSR model of the leave-in data based on A latent variables.

Predicting the response values for the left-out data is done in the following way: First left-out explanatory data are centered using the mean estimated from the leave-in samples (Eq. (6)). After this we obtained a mean centered left-out segment $\mathbf{X}_{\mathbf{C}_m}$. In order to estimate the response values for the left-out samples ($\hat{\mathbf{Y}}_m$), regression coefficients of the MBPLSR model of leave-in data ($\mathbf{B}_{-m,A}$) are applied to the mean-centered explanatory left-out data ($\mathbf{X}_{\mathbf{C}_m}$) according to Eq. (20):

$$\hat{\mathbf{Y}}_m = \mathbf{X}_{\mathbf{C}_m}\mathbf{B}_{-m,A}. \tag{20}$$

Segment m of residual matrix of the response data ($\mathbf{F}_{m,A}$) is calculated according to

$$\begin{aligned} \hat{\mathbf{Y}}_{m,A} &= \mathbf{X}_{\mathbf{C}_m}\mathbf{B}_{-m,A} \\ \mathbf{F}_{m,0} &= \mathbf{Y}_{\mathbf{C}_m} \\ \mathbf{F}_{m,A} &= \mathbf{Y}_{\mathbf{C}_m} - \hat{\mathbf{Y}}_{m,A} \end{aligned} \tag{21}$$

where $\hat{\mathbf{Y}}_{m,A}$ is the predicted response values for the left-out samples based on A MBPLSR components, $\mathbf{B}_{-m,A}$ contains regression coefficients for the MBPLSR model of leave-in data having A latent variables in the model, $\mathbf{X}_{\mathbf{C}_m}$ and $\mathbf{Y}_{\mathbf{C}_m}$ are segments m of the explanatory and response data tables respectively calculated in Eq. (6), $\mathbf{F}_{m,0}$ is the initial residual matrix for segment m and $\mathbf{F}_{m,A}$ is the corresponding residual matrix for segment m using A latent variables.

Predicting response values for the left-out samples ($\hat{\mathbf{Y}}_{m,A}$) was performed for various values of $A = 0,1,\ldots,A_{\max}$ ($A_{\max}$ was chosen sufficiently large) resulting in $A_{\max}$ different residual matrices for each segment i.e. $\mathbf{F}_{m,1}, \mathbf{F}_{m,2}, \ldots, \mathbf{F}_{m,A}, \ldots, \mathbf{F}_{m,A_{max}}$. In order to have one residual matrix for the response data for each value of A, $\mathbf{F}_{m,A}$ s were concatenated vertically for $m = 1,\ldots,M$ resulting in a residual matrix $\mathbf{F}_A$ for every value of A. The residual matrices $\mathbf{F}_{m,0}$ were concatenated in the same way in order to acquire the initial residual matrix $\mathbf{F}_0$.

Estimating the errors by cross-validation in the way that is described above was explained within PLSR's framework [20]. Our aim herein is to extend the procedure to the multi-block setting and propose graphical tools to enhance their interpretation. Investigating the contribution of each block to the prediction of $\mathbf{Y}$ can be done in two ways: Either to establish a PLSR model for each block separately or to use the block parameters of MBPLSR model. As our aim herein is not to test different PLSR models based on separate block analyses, we will consider the block parameters derived from the MBPLSR model. Section 2.3.2.1 describes the common method for calculating the cross-validation errors. Sections 2.3.2.2 and 2.3.2.3 describe our proposed method.

2.3.2.1. Global errors for **Y** *(Common method).* Calculating the Root Mean Square Error of **Y** in a multi-block setting is the same as is usually done in PLSR. It is briefly recalled in the following:

Cross-validated sum of squares for **Y** is calculated from the residual matrices according to Eq. (22)

$$\begin{aligned} SScv^g_{Y0} &= \sum_{i=1}^{N}\sum_{j=1}^{J} f_0(i,j)^2 \\ SScv^g_{YA} &= \sum_{i=1}^{N}\sum_{j=1}^{J} f_A(i,j)^2 \end{aligned} \tag{22}$$

where $f_0(i,j)$ and $f_A(i,j)$ are the (i,j)-*th* entry of the residual matrices $\mathbf{F}_0$ and $\mathbf{F}_A$ respectively. Calculations of $\mathbf{F}_0$ and $\mathbf{F}_A$ were described in Section 2.3.2.

Mean square error of **Y** is then calculated from correcting the sum of squares by the approximate degrees of freedom being consumed:

$$\begin{aligned} MSE^g_{Y0} &= \frac{SScv^g_{Y0}}{NJ} \\ MSE^g_{YA} &= \frac{SScv^g_{YA}}{NJ}. \end{aligned} \tag{23}$$

MSE_Y values are regarded as approximations for the same reasons as advocated for Eq. (11).

As it is done in the previous sections of this study, mean square errors are augmented by 3% of the initial variance for each new latent variable introduced in the model:

$$MSE^g_{YA} + A \times 0.03 \times MSE^g_{Y0} \rightarrow MSE^g_{YA} \tag{24}$$

Finally, the cross-validated global $RMSE_Y$ is determined as:

$$RMSE^g_{YA} = \sqrt{MSE^g_{YA}}\ . \tag{25}$$

2.3.2.2. Global errors for **Y** *(Proposed method).* For the estimation of global error, we suggest to predict **Y** from global scores $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_a, \ldots, \mathbf{t}_A]$ for $A = 1,\ldots,A_{\max}$ components. The calculations of residual matrices $\mathbf{F}^g_A$ and $\mathbf{F}^g_0$ for every different value of A are the same as described in Section 2.3.2 while every **X** is replaced by $\mathbf{T}_A$. Root mean square error calculations are done in the same way as was described in Section 2.3.2.1. The flow chart of the approximate $RMSE^g_{YA}$ calculation for this method is given in Fig. 2.

2.3.2.3. Block errors for **Y** *(Proposed method).* For the estimation of block errors, we suggest to predict **Y** from block scores $\mathbf{T}^b_A = \left[\mathbf{t}^b_1, \mathbf{t}^b_2, \ldots, \mathbf{t}^b_a, \ldots, \mathbf{t}^b_A\right]$ for $A = 1,\ldots,A_{\max}$ components. The calculations of residual matrices $\mathbf{F}^b_A$ and $\mathbf{F}^b_0$ for each block b and every different value of A are the same as described in Section 2.3.2 while every **X** is replaced by $\mathbf{T}^b_A$ and the procedure is repeated for $b = 1,\ldots,B$ and $A = 0, 1,\ldots,A_{\max}$.

The cross-validated sum of squares of the residual matrices for every block $b = 1,\ldots,B$ and for all model ranks $A = 0, 1,\ldots,A_{\max}$ is then calculated according to Eq. (26)

$$\begin{aligned} SScv^b_{Y0} &= \sum_{i=1}^{N}\sum_{j=1}^{J} f^b_0(i,j)^2 \\ SScv^b_{YA} &= \sum_{i=1}^{N}\sum_{j=1}^{J} f^b_A(i,j)^2 \end{aligned} \tag{26}$$

where $f^b_0(i,j)$ and $f^b_A(i,j)$ are the (i,j)-*th* entries of residual block matrices $\mathbf{F}^b_0$ and $\mathbf{F}^b_A$; $\mathbf{F}^b_A$ is the residual matrix for block b associated with an MBPLSR model of $\mathbf{T}^b_A$ and **Y**, with A latent variables introduced in the model.

Cross-validated mean square errors of estimating $\hat{\mathbf{Y}}$ from block scores are then calculated according to:

$$\begin{aligned} MSE^b_{Y0} &= \frac{SScv^b_{Y0}}{NJ} \\ MSE^b_{YA} &= \frac{SScv^b_{YA}}{NJ}. \end{aligned} \tag{27}$$

Following the same procedure in this paper mean square errors associated with each block are augmented by 3% of the initial variance for each new latent variable introduced in the model:

$$MSE^b_{YA} + A \times 0.03 \times MSE^b_{Y0} \rightarrow MSE^b_{YA}\ . \tag{28}$$

Finally the cross-validated RMSE of estimating $\hat{\mathbf{Y}}$ from block scores is determined by calculating the square root of MSE^b_{YA} for every block and for all various values of A:

$$RMSE^b_{YA} = \sqrt{MSE^b_{YA}}\ . \tag{29}$$

By examining the plot where the approximate global and block $RMSE_Y$s are plotted against the number of PLSR components one can investigate the predictability of every block i.e. how much each block is contributing to the prediction of **Y**. It is worth mentioning that if the design matrix is used as **Y** then it will not be meaningful to calculate $RMSE_Y$ as it is described in this section for the whole **Y** matrix. In that case, we suggest that cross-validated sum of squares $SScv^b_Y$, mean square errors MSE^b_Y and therefore root mean square errors $RMSE^b_Y$ are calculated for every **Y**-variable separately.

3. Multi-block data set

The multi-block data set which is used as the explanatory data in this study consists of four data blocks with different number of variables in each block; all of the variables are measured on the same 88 microbiological samples. The original multi-block data set is described in detail in the references [1,21]. Fig. 3 illustrates the multi-block data set. The multi-block explanatory data set contains Fourier Transform Infrared (FTIR) spectra and Amplified Fragment Length Polymorphism (AFLP) data (genetic fingerprinting) of 88 *L. monocytogenes* strains. FTIR spectroscopy is a rapid technique for metabolic fingerprinting of microorganisms [22]. As it can be seen in Fig. 3 the FTIR data are divided into three blocks of different spectral regions: polysaccharide region and fingerprint region (1200–720 cm^{-1}) define block $\mathbf{X}^1$ (498 variables), the protein region (1700–1500 cm^{-1}) defines block $\mathbf{X}^2$ (209 variables) and the fatty acid region (3000–2800 cm^{-1}) defines block $\mathbf{X}^3$ (208 variables). Block $\mathbf{X}^4$ (1701 variables) contains AFLP data. It should be noted that the spectral data is pre-processed by EMSC [23,24] prior to data analysis. Analysis of the original data set was done in the reference [1] and further information about the background behind the *L. monocytogenes* strain to strain variation in general and especially the variation in susceptibility to bacteriocins was obtained. Different ways of grouping the 88 strains phenotypically were studied: (1) According to susceptibility to Sakacin P, since the strains form two distinct sensitivity groups. Half of the strains lie below and half lie above a sensitivity threshold. (2) According to serotype. (3) According to the polysaccharide-fingerprint region in FTIR: The polysaccharide-fingerprint region of FTIR shows three distinct groups which we named *FTIR groups* [1,21]. The three polysaccharide-fingerprint groups can be obtained by running a principal component analysis (PCA) on the polysaccharide-fingerprint region (1200–720 cm^{-1}) and considering the first two components. In this study we use FTIR groups for graphical illustration.

Please cite this article as: S. Hassani, et al., Model validation and error estimation in multi-block partial least squares regression, Chemometr. Intell. Lab. Syst. (2011), doi:10.1016/j.chemolab.2011.06.001

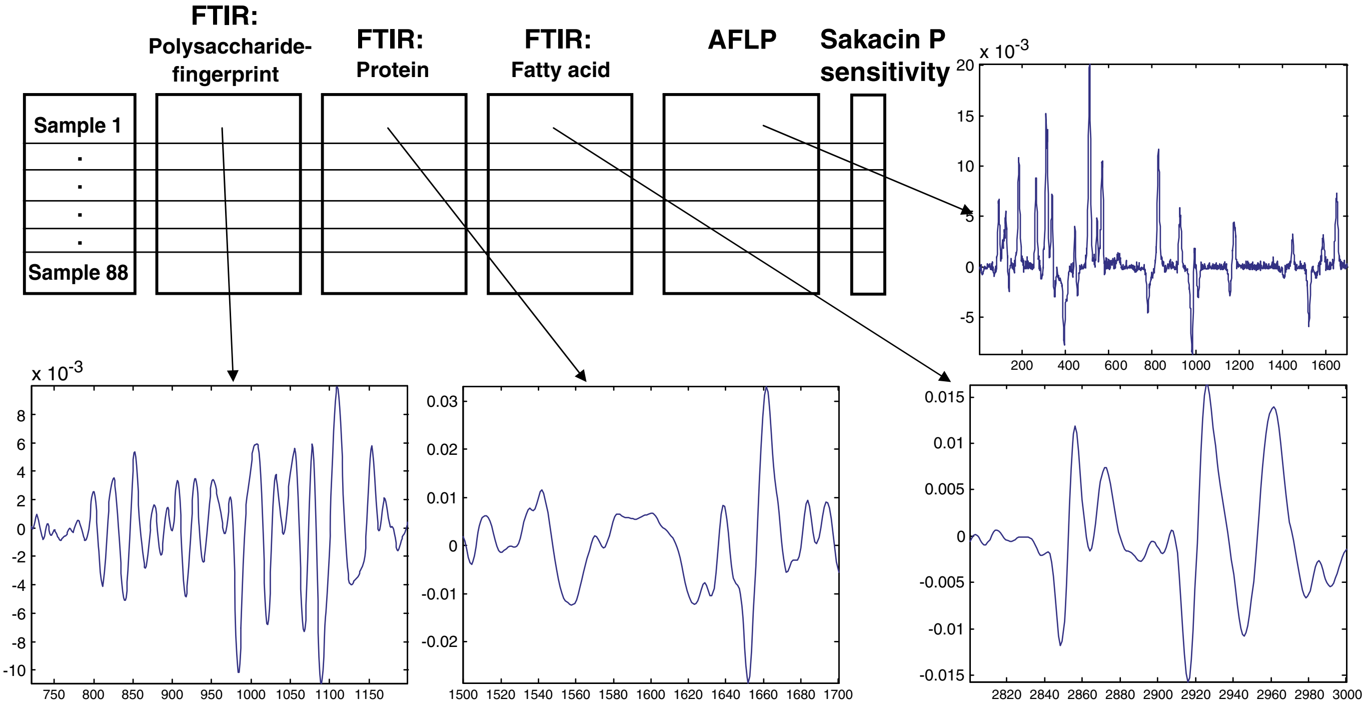


Fig. 3. Structure of the data set. Three different ranges of the FT-IR spectra (720–1200 cm^{-1}, fingerprint region and polysaccharide region; 1500–1700 cm^{-1}, protein region; 2800–3000 cm^{-1}, fatty acid region) are used in order to produce three different blocks. In addition, AFLP data is used as another data block. A phenotype variable i.e. sensitivity to Sakacin P is used as **Y** [1,21].

4. Results and discussion

4.1. MBPLSR of a multi-block data set

In order to find the common variation pattern in the explanatory multi-block data set which can predict the Sakacin P sensitivity of the strains in the study, MBPLSR was performed. For the MBPLSR algorithm we used super score deflation of **X** and **Y**. We have also tested the MBPLSR algorithm were only **Y** is deflated. The algorithms lead to different block scores and consequently the obtained RMSE plots are expected to be different. The obtained validation results were slightly different (results not shown). Since a comparison of both algorithms is beyond the scope of this paper we only present the results from the MBPLSR algorithm were both **X** and **Y** are deflated. Multi-block data set described in Section 3 was used as explanatory data (**X**) and Sakacin P sensitivity was used as the response variable (**Y**). The structure of the data set is described above and illustrated in Fig. 3. Fig. 4 shows the score plots (blocks and global plots) for the first and second latent variables. Samples are labeled “1”(red), “2”(blue) or “3” (green) according to different FTIR groups that they belong to. Different FTIR groups can be detected easily in the global pattern in Fig. 4e. Unvalidated explained variances for every explanatory block and for the global pattern are shown by the respective axes. The first two PLSR components cover 43% of the total variation in explanatory data set and 57% of the total variation in the response data. It is seen in the global pattern that the first latent variable separates group 3 from groups 1 and 2 while the second latent variable is responsible for separation of groups 1 and 2 from each other. Blocks one, three and four show tendencies toward the same grouping pattern in the global score plot while the second block has a different pattern. Although the grouping pattern seen in blocks one, three and four are different from each other, similar patterns can still be identified. An important question arises here: are there similar co-variation patterns in these three blocks or are there few dominant blocks which are imposing their patterns to the other blocks. The (unvalidated) explained variance of block 1, already gives an indication that block 1 is dominant. Nevertheless, all the other blocks show a relatively large explained variance. The second block has the second largest explained variance among the others while its sample pattern does not follow the pattern of the samples in the other blocks. The score plots (blocks and global plots) for the third and fourth latent variables are shown in Fig. 5. The third and the forth components do not explain a large variation in the data as they cover only 18% of the variation in the explanatory data set and 10% of that of the response data.

4.2. Root mean square error of **X** (RMSE$_X$): a cross-validation approach

In order to validate variation patterns that have been visually detected from the outcomes of MBPLSR, we propose to study global RMSE$_X$ and block RMSE$_X$s, calculated by Eq. (13) and (18) respectively. Since scientist's mind is always looking for grouping patterns when inspecting the score plots, he or she can be misled by visual perception. The use of colors in the score plots is likely to help in identifying the underlying patterns and clusters while increasing the risk of false discovery. When we have different blocks of variables containing different types of data and we want these blocks to predict common response data, one way to meet our goal is to implement MBPLSR and study the global score and block score displays. Inspecting the score plots necessitates the user to be provided with a tool to validate the contribution of each block in the global pattern. RMSE$_X$ plots give a validated image of block score plots which reveals the contribution of every block to the pattern that is seen in the global score plots. RMSE$_X$ plots for four blocks together with the global RMSE$_X$ are shown in Fig. 6. It is observed from the plots that blocks 1 and 2 have the most important contribution to the first and second latent variables; this was indicated by the unvalidated explained variances for these blocks. The remarkable result here is the significant contribution of block 2 to the global pattern although the block score plot (Fig. 4b) does not show a pattern similar to the global one.

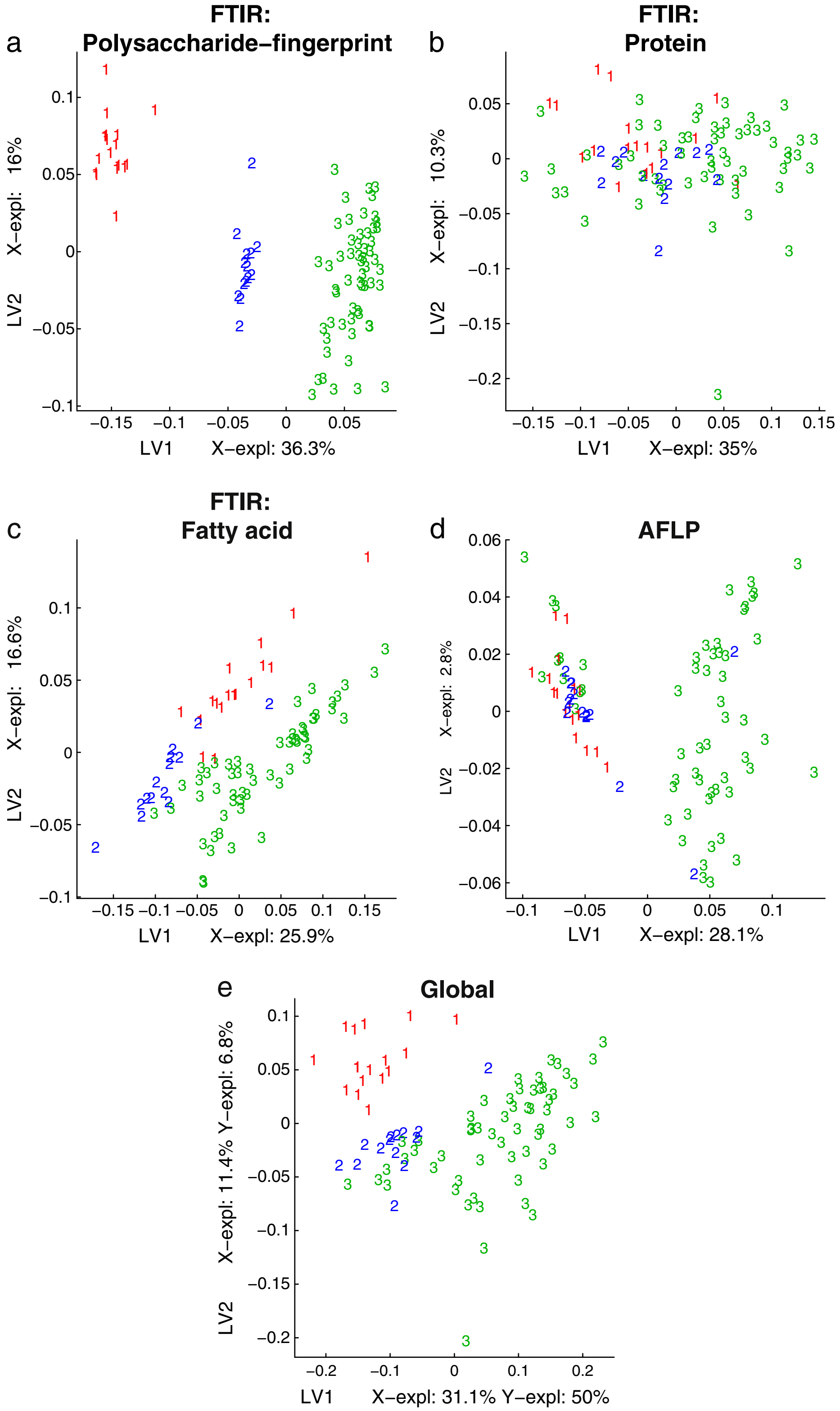


Fig. 4. Multi-block Partial Least Squares Regression (MBPLSR) analysis of the data. The samples are labeled “1”(red), “2”(blue) or “3” (green) according to FTIR polysaccharide-fingerprint groups. (a–d) First and second PLSR components of block scores. (e) First and second PLSR components of Global scores. The (un-validated) explained variances are shown on the axes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

$RMSE_X$ plots can also be used for determining the number of relevant latent variables in the model. As it is seen in Fig. 6 the global $RMSE_X$ reaches a local minimum having six latent variables in the model. Studying the $RMSE_X$ plots for every block gives an indication for the number of important latent variables in that block.

From the $RMSE_X$ plot, it is seen that the fourth block is different from the other blocks, as it shares only the first and the sixth PLSR component with the other blocks and it does not fit the common pattern in the data set. This important finding would not have been discovered by only looking at the patterns of the score plots (Fig. 4).

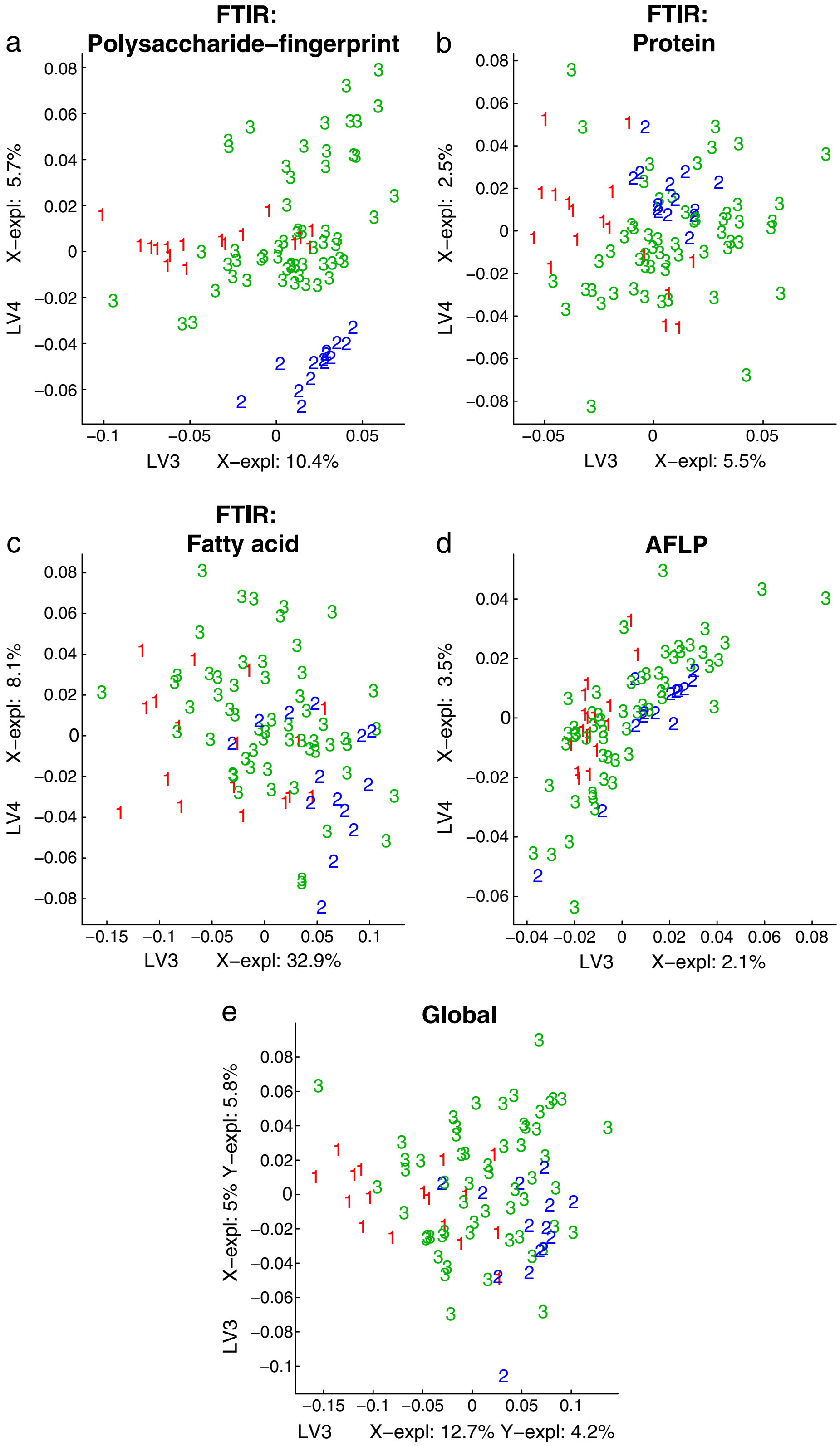


Fig. 5. Multi-block Partial Least Squares Regression (MBPLSR) analysis of the data. The samples are labeled "1"(red), "2"(blue) or "3" (green) according to FTIR polysaccharide-fingerprint groups. (a–d) Third and fourth PLSR components of block scores. (e) Third and fourth PLSR components of Global scores. The (un-validated) explained variances are shown on the axes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Looking at the score plots before validating the patterns could have given the wrong impression that block 2 has a very different pattern from the global pattern in the data.

It worth mentioning that the rank of the data blocks can be visualized by the $RMSE_X$ plots. Low rank data block will not pose any problem to the modeling and visualization. E.g. the very low rank of

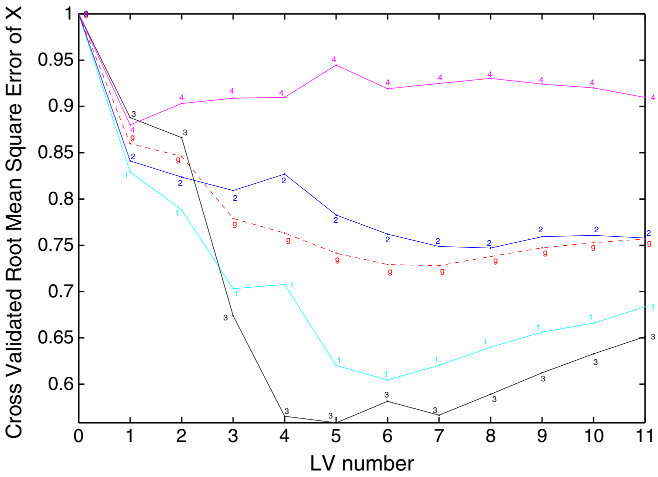


Fig. 6. Root Mean Square Error ($RMSE_X$) plot for the data set. (1–4) $RMSE_X$ plots for blocks 1–4. (Red dotted) Global $RMSE_X$ plot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

one or several of the blocks will be seen in the $RMSE_X$ plot as: (1) If the variation of a low-rank block is captured by the global model, the $RMSE_X$ of this block will go to zero after few components. (2) If the low-rank block contains a variation pattern that does not exist in other blocks and the other blocks have a high rank with a strong co-variation pattern, $RMSE_X$ of this block will not go to zero before the common variation pattern of the other blocks is modeled.

4.3. Root mean square error of *Y* ($RMSE_Y$): a cross-validation approach

In order to validate the ability of the MBPLSR model for predicting the response variable we propose to study global $RMSE_Y$ and block $RMSE_Y$s, calculated by Eq. (25) and (29) respectively. The block $RMSE_Y$ plots give a validated image of the contribution of every block to the prediction of response data in the MBPLSR model. Global and block $RMSE_Y$ plots are shown in Fig. 7. It is observed that Sakacin P sensitivity is well predicted by the MBPLSR model from our set of explanatory variables, specifically after the first PLSR component. We can also observe that all of the blocks are able to predict the Sakacin P sensitivity on the basis of the first two latent variables. It is worth mentioning that the contribution of blocks 1 and 4 to the prediction (in the first two components) is more significant in comparison to blocks 2 and 3.

4.4. Choosing the number of latent variables in the model: a comparison of different approaches

After introducing two different approaches for validating the MBPLSR model, we still have not answered a fundamental question: How many latent variables should be included in an MBPLSR model? The first and foremost thing that should be kept in mind while looking for the sufficient number of components is the purpose of the analysis. The aim of the data analysis is an important aspect which gives different weights to the latent variables i.e. every latent variable can be more or less important depending on the point of view, which is the aim of the study. In this study, we have proposed two different types of RMSE plots for the same MBPLSR model. Comparing Fig. 6 and 7 reveals a significant difference between the two validation plots which raises the question as to which one of them should be used for choosing the appropriate number of latent variables.

We can look at our MBPLSR model from two different points of view: (1) Do we want our model to be stable and not to change significantly after having a certain number of latent variables? or (2) Do we emphasize the ability of our model for predicting the Sakacin P sensitivity and we would like to use our model for future predictions? Fig. 6 gives the answer to the first question. If one is looking for a stable model with trustworthy grouping patterns, he or she should keep six latent variables in the model. Especially if one is interested in studying the patterns in the different blocks, interesting information is still available in the third to sixth components (e.g. the third and forth latent variables of block 3 are very informative, the fifth and sixth latent variables of block 1 also contain important information). Alternatively, if one is not interested in revealing the different patterns of different blocks and if setting up a stable model is not the main goal, but the main purpose is to set up a global model with good prediction ability, then Fig. 7 is a very relevant tool. Indeed, this figure suggests selecting less than four components and, moreover, the global RMSE does not significantly change from the second to the fourth latent variables. Therefore an MBPLSR model containing only two latent variables could be a good choice for the prediction purposes. However if one would like to have the most precise model, he or she should use the number of PLSR components where global plot reaches its minimum and that would be seven (or even eight) components in the model which is relatively the same number suggested by Fig. 6. It is concluded that having seven MBPLSR

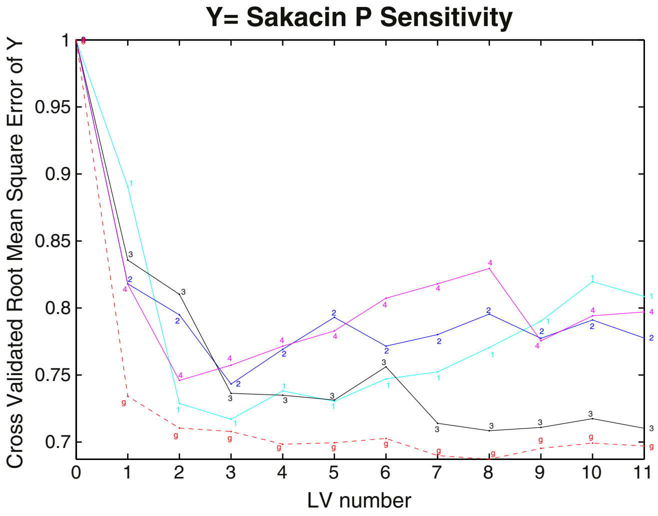


Fig. 7. Root Mean Square Error ($RMSE_Y$) plot. (1–4) $RMSE_Y$ plots for blocks 1–4. (Red dotted) Global $RMSE_Y$ plot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Please cite this article as: S. Hassani, et al., Model validation and error estimation in multi-block partial least squares regression, Chemometr. Intell. Lab. Syst. (2011), doi:10.1016/j.chemolab.2011.06.001

components in the model results in a stable model with a great ability for predicting sensitivity to Sakacin P.

5. Conclusion

Multi-block techniques such as MBPLSR provide the user with powerful visualization tools that aim at a better understanding of the data. Sample- and variable- variation patterns are detected between and within the data blocks. However the patterns that are identified visually should not be taken for granted and should be validated for interpretation purposes. Indeed, the identification of patterns by visual inspection can be misleading. In order to avoid misinterpreting the identified block and global patterns, we have proposed different validation techniques. These techniques also make it possible to assess the stability of the model. Moreover, the user has the possibility to assess the prediction ability of the model. We have also tackled another important aspect in explorative data analysis methods such as MBPLSR which is the selection of appropriate number of latent variables to be introduced in the model.

Acknowledgments

The authors are grateful for financial support by the Nordic Centre of Excellence on Food, Nutrition and Health “Systems biology in controlled dietary interventions and cohort studies” (SYSDIET) funded by NordForsk, by Foundation for Research Levy on Agricultural Products in Norway and by the grant 203699 (New statistical tools for integrating and exploiting complex genomic and phenotypic data sets) financed by the Research Council of Norway.

References

[1] A. Kohler, M. Hanafi, D. Bertrand, A. Oust Janbu, T. Møretrø, K. Naderstad, M. Qannari, H. Martens, Interpreting several types of measurements in bioscience, in: P. Lasch, J. Kneipp (Eds.), Modern Concepts in Biomedical Vibrational Spectroscopy, John Wiley & Sons, USA, 2008.

[2] S. Hassani, H. Martens, E.M. Qannari, M. Hanafi, G.I. Borge, A. Kohler, Analysis of -omics data: graphical interpretation- and validation tools in multi-block methods, Chemometrics and Intelligent Laboratory Systems 104 (1) (2010) 140–153.

[3] M.Z. Jaafar, A.H. Khan, S. Adnan, A. Markwitz, N. Siddique, S. Waheed, R.G. Brereton, Multiblock analysis of environmental measurements: A case study of using Proton Induced X-ray Emission and meteorology dataset obtained from Islamabad Pakistan, Chemometrics and Intelligent Laboratory Systems In Press, Corrected Proof.

[4] L.P. Brás, S.A. Bernardino, J.A. Lopes, J.C. Menezes, Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour, Chemometrics and Intelligent Laboratory Systems 75 (1) (2005) 91–99.

[5] M. Jing, W. Cai, X. Shao, Multiblock partial least squares regression based on wavelet transform for quantitative analysis of near infrared spectra, Chemometrics and Intelligent Laboratory Systems 100 (1) (2010) 22–27.

[6] L.P. Brás, J.A. Lopes, C.R. Santos, J.P. Cardoso, J.C. Menezes, A. Barbosa-Póvoa, H. Matos, Modelling and identification of individual stage contributions in an industrial pharmaceutical process by multiblock PLS, Computer Aided Chemical Engineering, Elsevier, 2004, pp. 601–606.

[7] S.W. Choi, I.-B. Lee, Multiblock PLS-based localized process diagnosis, Journal of Process Control 15 (3) (2005) 295–306.

[8] H. Martens, M. Martens, Multivariate Analysis of Quality: An Introduction, Wiley, Chichester, UK, 2001.

[9] L.E. Wangen, B.R. Kowalski, A multiblock partial least squares algorithm for investigating complex chemical systems, Journal of Chemometrics 3 (1) (1989) 3–20.

[10] S. Wold, H. Martens, H. Wold, The multivariate calibration problem in chemistry solved by the PLS method, in: B. Kågström, A. Ruhe (Eds.), Matrix Pencils, Springer Berlin/Heidelberg, 1983, pp. 286–293.

[11] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, Journal of Chemometrics 12 (5) (1998) 301–321.

[12] R.W. Gerlach, B.R. Kowalski, H.O.A. Wold, Partial least-squares path modelling with latent variables, Analytica Chimica Acta 112 (4) (1979) 417–421.

[13] I.E. Frank, B.R. Kowalski, Prediction of wine quality and geographic origin from chemical measurements by parital least-squares regression modeling, Analytica Chimica Acta 162 (1984) 241–251.

[14] I. Frank, J. Feikema, N. Constantine, B. Kowalski, Prediction of product quality from spectral data using the partial least-squares method, Journal of Chemical Information and Computer Sciences 24 (1) (1984) 20–24.

[15] I.E. Frank, B.R. Kowalski, A multivariate method for relating groups of measurements connected by a causal pathway, Analytica Chimica Acta 167 (1985) 51–63.

[16] S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom, H. Wold, PLS Model Building Theory and Application, Frankfurt am Main, Frankfurt, Germany, 1987.

[17] S.J. Qin, S. Valle, M.J. Piovoso, On unifying multiblock analysis with application to decentralized process monitoring, Journal of Chemometrics 15 (2001) 715–742.

[18] J.A. Westerhuis, A.K. Smilde, Deflation in multiblock PLS, Journal of Chemometrics 15 (2001) 485–493.

[19] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society. Series B (Methodological) 36 (2) (1974) 111–147.

[20] H. Martens, T. Næs, Multivariate Calibration, Wiley & Sons, Chichester, 1989.

[21] A. Oust, T. Moretro, K. Naterstad, G.D. Sockalingum, I. Adt, M. Manfait, A. Kohler, Fourier transform infrared and Raman spectroscopy for characterization of Listeria monocytogenes strains, Applied and Environmental Microbiology 72 (1) (2006) 228–232.

[22] C.L. Winder, S.V. Gordon, J. Dale, R.G. Hewinson, R. Goodacre, Metabolic fingerprints of Mycobacterium bovis cluster with molecular type: implications for genotype–phenotype links, Microbiology 152 (9) (2006) 2757–2765.

[23] A. Kohler, C. Kirschner, A. Oust, H. Martens, Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin, Applied Spectroscopy 59 (6) (2005) 707–716.

[24] A. Kohler, M. Zimonja, V. Segtnan, H. Martens, Data preprocessing: SNV, MSC and EMSC pre-processing in biospectroscopy, in: S. Brown, R. Tauler, R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 139–162.

Please cite this article as: S. Hassani, et al., Model validation and error estimation in multi-block partial least squares regression, Chemometr. Intell. Lab. Syst. (2011), doi:10.1016/j.chemolab.2011.06.001

Paper III

Degrees of freedom estimation in Principal Component Analysis and Consensus Principal Component Analysis

Authors

Sahar Hassani[1,2], Harald Martens[1,2], El Mostafa Qannari[3] and Achim Kohler[1,2]

[1] Nofima AS, Osloveien 1, N-1430 Ås, Norway

[2] Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway

[3] Unité de Sensométrie et de Chimiométrie, ONIRIS/INRA, BP 82225, 44322 Nantes Cedex 3, France

Corresponding author:
Sahar Hassani, Nofima AS, Osloveien 1, N-1430 Ås, Norway,
e-mail: sahar.hassani@nofima.no

Abstract

The concept of Degree of Freedom (DF) is an important issue in statistical model assessment and parameter estimation. In this paper, we investigate this concept within the context of data modeling by Principal Component Analysis (PCA) and its multi-block extension, the Consensus Principal Component Analysis (CPCA). We run simulation studies and assess the degrees of freedom by comparing cross-validated error estimates with error estimates from uncorrected model fits. These simulation studies reveal that the DF consumption in PCA and CPCA depends on the eigenvalue structure of the data at hand. We also show that the obtained DF estimates can be used to obtain realistic error estimations without performing cross-validation. Furthermore, it is shown how different strategies of cross-validation and the use of an independent test set affect the estimate of the degrees of freedom and the estimate of the model error.

Keywords

Principal Component Analysis (PCA), Consensus Principal Component Analysis (CPCA), Cross-validation (CV), Degree of Freedom (DF).

1. Introduction

1.1 The concept of DF

The term "Degree of Freedom" (DF) is widely used in mechanics, physics, chemistry, statistics and chemometrics, and refers to different, but related concepts. The concepts range from "independent displacements and/or rotations that specify the orientation of the body or system" in mechanics [1] to the "number of values in the final calculation of a statistic that are free to vary" in statistics [2] [3]. In parameter estimation, e.g. estimating the mean or variance for a given response variable, the DFs consumed is determined by the number of parameters that are estimated independently, so the available DFs after the estimation equals the number of independent observations ('samples') minus the number of independent parameters. The DF quantification plays an important role in classical statistics and is reported together with the results, e.g. when parameters are estimated in statistical hypothesis tests as in the F-test, Student's t-test and in linear modeling. It is important to apply a correct estimate of the consumed DFs, otherwise the assessment (confidence limits, p-values etc) will be over-optimistic or too pessimistic. In multivariate data modeling the DF concept is equally important but slightly more complicated; both samples (rows) and variables (columns) contribute to DFs, and since both relations between samples and variables are modeled, the DF availability also has to take patterns of variable co-variation into account. By "the number of DFs consumed during modeling" in this paper we mean the number of pieces of independent, useful information in the data set that are used for the modeling process.

DF is important in PCA in order to assess the level of random noise in the data set at hand, in order to optimize the number of PCs to trust as valid information. The more PCs are included in the PCA model, the more DFs are consumed and the smaller the apparent lack-of-fit residuals between the data and the fitted model is, irrespective whether the PCs pick up valid signal or random noise. Therefore, a model assessment based on naively averaging the squared lack-of-fit residuals will lead to over-optimism with respect to the model rank i.e. how many interesting phenomena can be validly estimated from the data, as well as to the model's predictive ability, which for PCA means the expected fit of future data vectors to the bilinear model. To guard against such over-fitting, model assessment can be done by cross-validation/jack-knifing or with the use of an independent test set of samples. However, the former can be time-consuming since the model has to be recalculated repeatedly, and the latter can be expensive because a large, representative test set is required, otherwise the results will be unreliable or irrelevant. The purpose of the present paper is to study the role of DF consumption in PCA and to assess the potential for a simple and realistic estimate of DF consumption in such models.

1.2 DF estimation in PCA

Traditionally, for models based on estimated latent variables (“components” or “factors”), such as Principal Component Analysis (PCA) or Partial Least Squares Regression (PLSR), it has not been clear how many DFs are consumed by the individual components. Faber [4] studied DFs for the residuals of a PCA model. The method that he used for calculating DFs was based on standard statistical formula for DF calculation consisting of two steps: a) Identifying the projection matrix. b) Determining the trace of the projection matrix which gives the DF for the residuals of the model. However, while this formula accounts for the number of independent parameters estimated, it does not take into account the DFs consumption caused by the PCA search for maximal covariance.

Assessing the bi-linear data approximation model from PCA can be compared to the corresponding assessment of bi-linear regressions by PLSR. In both cases variance estimation, rank optimization and prediction assessment would be simplified if a simple estimate of the number of DFs consumed existed. Van Der Voet in 1999 estimated DFs consumed from **Y** in predictive models ($\mathbf{Y} = f(\mathbf{X})$) by introducing pseudo-DF (PDF). He calculated and applied PDFs for a Multiple Linear Regression (MLR) model as well as for a PLSR model on a real data set [5]. His method consists of comparing error variances obtained by direct model fitting and by cross-validation.

In 2007 Krämer and Braun calculated DF in **Y** for kernel PLS. They implemented methods based on either DF or cross-validation for addressing important issues such as model selection [6]. Kengo Kato [7] studied DF consumption in estimating regression coefficients. He calculated an unbiased estimation for DFs related to those methods. Recently, Krämer and Sugiyama [8] calculated DFs in **Y** for PLSR. They derived an unbiased estimator for DF and applied it for choosing the number of relevant PLSR components. In the PLSR literature there has been little focus on estimating DF consumption in **X**.

The procedure of modeling a data set by means of PCA or CPCA consumes DFs in two respects: 1) The search process that leads to the identification of the Principal Component (PC) directions in the

X-space – this may be seen as a kind of variable selection. 2) The estimation of the parameters (loadings and scores) for the chosen directions – this may be seen as a traditional OLS regressions on orthogonal regressors in both loading and score directions, but with an increasing number of orthogonality constraints for each PC. Since CPCA is a multi-block extension of PCA, the latter will be assessed by a block-wise redistribution of the DFs consumed by each PCA PC.

The assessment of the correct number of consumed DFs is of particular importance when dealing with issues such as calculating Mean Squared Error (MSE) after 0,1,2,… PCs. Tentative formulas have previously been proposed by us [9] [10]. In this work, we set up simulation studies in order to corroborate the validity of those tentative formulas. The work has two aspects: a) Studying how DFs are consumed as PCA and CPCA models are developed from data with different structures, based on comparing fitted and cross-validated error estimates. b) Deriving a metamodel that can replace cross-validation in the sense that it gives a realistic prediction of DF consumption in new data sets from the characterization of individual data sets. If successful, this would allow a fast and cheap determination of predictive MSE and optimal model rank. For PCA and CPCA modeling we expect that the eigenvalue structure of the data at hand affects the search process for these models. Therefore we take this into account when estimating the DF consumption in a wide range of data types: In our simulation study, thousands of data sets are generated with differing singular value structures in order to cover as a large range of data qualities as possible.

Our proposed method for determining the number of DF requires validating the models – i.e. estimating the residual variances by external means that do not over-fit. This can be performed either through running cross-validation or using an independent test set. We will apply both approaches and compare the validated errors and the DF consumption by CV with that for an independent test set. Since the number of CV segments may affect the results, we will also run the same study using different CV segments in order to find an optimal number of CV segments to be used.

The paper is organized in the following way: The theory of the paper (i.e. notations, data simulation procedures, preprocessing of data, data modeling and DF estimations) are given in Section 2. The results of the paper are given in Section 3: the correctness of the intuitive MSE formulas that we used in our previous study is checked in Section 3.2 and 3.3. DFs associated with PCA calculated for 500 simulated data sets are shown in Section 3.4. Section 3.5 contains the results for DF calculations in CPCA. The number of cross-validation segments to be used for calculations are studied in Section 3.6. Application of DF calculations for real data sets is shown in Section 3.7. We finish the paper by a conclusion in Section 4.

2. Theory

2.1 Notation

We follow the notations commonly used in chemometrics, e.g. Martens & Martens in [11]: Matrices and vectors are written as bold-face, matrices as upper-case letters and vectors as lower-case letters. The indices $i = 1, 2, ..., N$ and $k = 1, 2, ..., K$ denote samples (rows) and variables (columns) in a data table **X**, respectively. We denote blocks of variables by $b = 1, 2, ..., B$, by $m = 1, 2, ..., M$ cross-validation segments of samples and by $a = 1, 2, ..., A$ the number of PCA components. By $\mathbf{X} = \left[\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^b, ..., \mathbf{X}^B\right]$ we denote the multi-block data set consisting of *B* blocks, each of them containing K_b variables. Data pertaining to the same measurement technique, e.g. spectroscopy, chromatography and gene fragment analysis profiling are typically collected in the same descriptor block $\mathbf{X}^b$. In omics experiments different measurement techniques are applied to the same samples and, for the multi-block analysis, data need to be ordered in a way that a sample-to-sample (row-to-row) correspondence between the blocks is achieved. Only measurements originating from the same biological replicate can be related to each other by a row-to-row correspondence. If, for example, for different methods several and different biological replicates are used, only means of biological replicates can be related to each others. The total number of samples in each data set is represented by *N*, the total number of variables in a given descriptor block *b* by K_b and the total number of variables in the whole data set (**X**) by *K* ($K = \sum_{b=1}^{B} K_b$).

2.2 Data simulation

2.2.1 Single-block data set

A data set with known characteristics is generated for the purpose of estimating the degrees of freedom consumed during the calculations of the model parameters by PCA. An important aspect in simulating the data is to make sure that we can control different features of the data sets that are generated. In particular, we want to control the latent structure of the simulated data sets by monitoring the eigenvalue structure of their covariance matrices. Simulating data sets with specified structures was performed according to the algorithm given by Arteaga and Ferrer in [12] with a slight simplification. Data was simulated as follows: First a random normally distributed data set was generated with the desired size:

$$\mathbf{G}^1_{\text{Input}} = randn(N, K_1) \quad (1)$$

where $randn(N, K_1)$ is a random normally distributed matrix of size $N \times K_1$; N is the number of samples and K_1 is the number of variables.

The matrix $\mathbf{G}^1_{\text{Input}}$ was mean-centered according to:

$$\mathbf{G}^1 = \mathbf{G}^1_{\text{Input}} - \mathbf{1} \cdot \bar{\mathbf{g}}^{1\,\prime}_{\text{Input}} \quad (2)$$

where $\mathbf{G}^1$ is the mean-centered random data set, $\mathbf{1}$ is an $N \times 1$ vector of 1s and $\bar{\mathbf{g}}^{1\,\prime}_{\text{Input}}$ is a $1 \times K_1$ vector of the mean values of K_1 variables in $\mathbf{G}^1$ calculated along the N samples. Then the Singular Value Decomposition (SVD) of mean-centered $\mathbf{G}^1$ was calculated as:

$$\mathbf{G}^1 = \mathbf{U}^1\mathbf{S}^1\mathbf{V}^{1\prime} \quad (3)$$

where $\mathbf{U}^1$ and $\mathbf{V}^1$ are unitary matrices of sizes $N \times N$ and $K_1 \times K_1$ respectively and $\mathbf{S}^1$ is a diagonal matrix of the size $N \times K_1$ containing the singular values of $\mathbf{G}^1$ i.e. square root of the eigenvalues of the covariance matrix $\mathbf{G}^{1\prime}\mathbf{G}^1$.

The desired singular value structure was produced in the following way: We generated a vector of eigenvalues between 0 and 1, in descending order as:

$$\boldsymbol{\lambda}_1 = \left(\left(\frac{A_1 - 1}{A_1} \right)^{m1}, \left(\frac{A_1 - 2}{A_1} \right)^{m1}, ..., \left(\frac{1}{A_1} \right)^{m1}, 0 \right) \quad (4)$$

where A_1 is the number of PCs in the simulated data block which is chosen herein as the minimum of N and K_1; $m1$ is a tuning parameter which controls the eigenvalue structure.

Simulated singular values are computed as the square root of the generated eigenvalues:

$$\mathbf{S}^1_{\mathrm{Sim}} = diag(\sqrt{\boldsymbol{\lambda}_1})$$

$$\mathbf{S}^1_{\mathrm{Sim}} = \begin{pmatrix} \sqrt{\left(\frac{A_1-1}{A_1}\right)^{m1}} & 0 & 0 & \cdots\cdots & 0 & 0 \\ 0 & \sqrt{\left(\frac{A_1-2}{A_1}\right)^{m1}} & 0 & \cdots\cdots & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & \cdots\cdots & \sqrt{\left(\frac{1}{A_1}\right)^{m1}} & 0 \\ 0 & . & . & . & . & 0 \end{pmatrix} \qquad (5)$$

We replace the singular values of the mean-centered random data set $\mathbf{G}^1$ by the simulated singular values in order to generate a data set with specified singular values:

$$\mathbf{X}^1_{\mathrm{True}} = \mathbf{U}^1\mathbf{S}^1_{\mathrm{Sim}}\mathbf{V}^{1\prime} \qquad (6)$$

where $\mathbf{U}^1$ and $\mathbf{V}^{1\prime}$ are the matrices which were calculated by Eq. 3. $\mathbf{S}^1_{\mathrm{Sim}}$ contains the simulated singular values calculated in Eq. 5 and $\mathbf{X}^1_{\mathrm{True}}$ is the generated data set which has the desired singular value structure.

Finally, random normally distributed errors were added to the generated data matrix $\mathbf{X}^1_{\mathrm{True}}$:

$$\mathbf{X}^1_{\mathrm{Input}} = \mathbf{X}^1_{\mathrm{True}} + randn(N, K_1) \times \sigma_1^2 \qquad (7)$$

where $randn(N, K_1)$ is an $N \times K_1$ matrix of random normally distributed numbers, σ_1^2 is the parameter which controls the noise level added to the data block and $\mathbf{X}^1_{\mathrm{Input}}$ is a single-block simulated data set with the desired singular value structure and noise level.

2.2.2 Multi-block data set

The procedure described in the previous section needs to be slightly modified in order to be used for simulating different blocks in a multi-block data set. Repeating the same procedure as in Section 2.2.1 produces a multi-block set of data without any connection among the different blocks whereas in a real multi-block data set one would expect to find a row-to-row relationship between different blocks. The only modification which is applied on the procedure in Section

2.2.1 for simulating a multi-block data set is to introduce such a relationship. We have accomplished this task by imposing the same sample variation pattern to all of the blocks. This is done by replacing Eq. 6 by the following Eq. for $b = 1,...,B$:

$$\mathbf{X}_{\mathrm{True}}^{b} = \mathbf{U}^{1}\mathbf{S}_{\mathrm{Sim}}^{b}\mathbf{V}^{b\prime} \qquad (8)$$

where $\mathbf{U}^{1}$ is the unitary matrix of size $N \times N$ calculated for the first data block by Eq. 3, $\mathbf{V}^{b\prime}$ is the unitary matrix of size $K_b \times K_b$ calculated for simulating block *b*, $\mathbf{S}_{\mathrm{Sim}}^{b}$ is the simulated singular values belonging to block *b* and $\mathbf{X}_{\mathrm{True}}^{b}$ is the block *b* of the multi-block simulated data set with the desired singular value structure.

2.3 Preprocessing of the data

2.3.1 Mean-centering

It is common to mean-center the data prior to PCA and CPCA. Therefore we mean-center each block in the simulated data set by subtracting the mean value of each variable according to Eq. 9:

$$\mathbf{X}_{\mathrm{Unscaled}}^{b} = \mathbf{X}_{\mathrm{Input}}^{b} - \mathbf{1}\cdot\bar{\mathbf{x}}_{\mathrm{Input}}^{\prime} \qquad (9)$$

where $\mathbf{X}_{\mathrm{Input}}^{b}$ is the block *b* of the data set at hand, $\mathbf{1}$ is an $N \times 1$ vector of 1s, $\bar{\mathbf{x}}_{\mathrm{Input}}^{\prime}$ is a $1 \times K$ vector of the means of the variables in the data block calculated over the *N* samples and $\mathbf{X}_{\mathrm{Unscaled}}^{b}$ is the unscaled mean-centered data block.

2.3.2 Scaling

In order to put the data tables on the same footing (i.e. the same total variance) when running PCA or CPCA, variables may be scaled. When the data is used without scaling, the measurement units or number of variables in each block may influence the multi-block model. When scaling the data in a multi-block setting the user has also the possibility to control the influence of some blocks on the multi-block model. For instance, the user may want to assign relatively large weights to some blocks in order to let them dictate the variation pattern or, contrariwise, he may down-weight some blocks in order to minimize their influence on the multi-block model. In the present study, we scaled all the blocks in the same way by dividing each block by its norm according to:

$$\mathbf{X}^b = \frac{\mathbf{X}^b_{\text{Unscaled}}}{\sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K_b}(\mathbf{X}^b_{\text{Unscaled}}(i,k))^2}} \quad (10)$$

where $\mathbf{X}^b_{\text{Unscaled}}$ is the mean-centered unscaled data, $\mathbf{X}^b_{\text{Unscaled}}(i,k)$ is the (i,k)-th entry of $\mathbf{X}^b_{\text{Unscaled}}$ and $\mathbf{X}^b$ is the mean-centered and scaled data block which will subsequently be used in the following model calculations.

2.4 Data modeling

2.4.1 PCA

PCA aims at investigating the covariation patterns within a data table. In this method of analysis, the mean-centered data table **X** is modeled as a sum of *A* relevant PCs plus a residual matrix **E** as:

$$\mathbf{X} = \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A \quad (11)$$

where $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_a, ..., \mathbf{t}_A]$ contains *A* score vectors $\mathbf{t}_a$, and $\mathbf{P}_A$ is the corresponding matrix of loading vectors $\mathbf{p}_a$, $\mathbf{P}_A = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_a, ..., \mathbf{p}_A]$.

Calculating the parameters of a PCA model can be performed through several equivalent algorithms e.g. NIPALS in [11].

2.4.2 CPCA

CPCA can be used in order to explore the common variation pattern within and between the data blocks in a multi-block data set. The CPCA model for the mean-centered multi-block data set $\mathbf{X} = \left[\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^B\right]$ is given by:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A \\ \mathbf{X}^b &= \mathbf{T}^b_A\mathbf{P}^{b\prime}_A + \mathbf{E}^b_A \end{aligned} \quad (12)$$

where $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_a, ..., \mathbf{t}_A]$ contains *A* global score vectors $\mathbf{t}_a$, and $\mathbf{P}_A$ is the corresponding matrix of global loading vectors $\mathbf{p}_a$, $\mathbf{P}_A = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_a, ..., \mathbf{p}_A]$. The global loading matrix $\mathbf{P}_A$ can also be written as the matrix of concatenated block loading matrices $\mathbf{P}^b_A$ as: $\mathbf{P}'_A = \left[\mathbf{P}^{1\prime}_A \ \mathbf{P}^{2\prime}_A \ \ldots \ \mathbf{P}^{b\prime}_A \ \ldots \ \mathbf{P}^{B\prime}_A\right]$.

There are several algorithms which can be used for the calculation of the parameters of a CPCA model. NIPALS which explicitly shows how to calculate the block parameters (i.e. block scores and block loadings) in addition to the global parameters (i.e. global scores and global loadings) is the most popular one. A presentation of the NIPALS algorithm for CPCA can be found in [10].

2.5 Degrees of Freedom (DFs) estimation

Modeling a data set by means of multivariate techniques, such as PCA and CPCA, consumes useful information and consequently DFs in the data set. Finding latent variables requires searching for the direction of the main variation in the data. The search process obviously consumes DFs that need to be taken into account. The easier the main variation patterns are found, the less DFs are consumed in the process. The methods used for determining PCs in PCA and CPCA are based on an eigenvector analysis. Since the performance of the algorithm searching for eigenvectors depends on the eigenvector structure in the data set, the DF consumption for modeling a data set by PCA and CPCA is tightly related to its eigenvalue structure. Therefore it is desirable that a simulated data set has a defined eigenvalue structure.

For each data set, we estimated the DF consumption during the modeling process in the following way: 1) We model a data set by PCA (or CPCA) and we calculate the “calibration” sum of squares of the errors ($SS_{Cal,A}$) over the *N* samples, for the calibration model after $A = 0,1,2,...$ PCs. 2) We also calculate the corresponding cross-validated sum of squares ($SS_{CV,A}$) for the same data set, after $A = 0,1,2,...$ PCs. 3) By comparing $SS_{Cal,A}$ and $SS_{CV,A}$ we find the number of DFs that makes the calibration based error estimate $MSE_{Cal,A}$ equal to corresponding cross-validated, and presumably “true” error estimate $MSE_{CV,A}$. The detailed calculations for these three steps will be explained in the following sections.

2.5.1 Estimated Degree of Freedom (EDF) in PCA

2.5.1.1 Calibration Mean Squared Error (MSE$_{Cal}$)

PCA parameters are calculated by means of SVD as follows:

$$\begin{aligned}\mathbf{X} &= \mathbf{USV}' \\ \mathbf{T} &= \mathbf{US} \qquad (13) \\ \mathbf{P} &= \mathbf{V}\end{aligned}$$

where $\mathbf{X}$ is the simulated data set of size $N \times K$, $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices of sizes $N \times N$ and $K \times K$ respectively, $\mathbf{S}$ is a diagonal matrix of the size $N \times K$ containing the singular values of $\mathbf{X}$, $\mathbf{T}$ is the $N \times K$ matrix of all the K score vectors for the PCA model of $\mathbf{X}$ and $\mathbf{P}$ is the $K \times K$ matrix of all the K loading vectors for the PCA model of $\mathbf{X}$.

The predicted values for the data are then estimated from the scores and loadings for different values of $A = 1,...,A_{\max}$ according to:

$$\hat{\mathbf{X}}_A = \mathbf{T}_A \mathbf{P}'_A \qquad (14)$$

where $\mathbf{T}_A = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_a, ..., \mathbf{t}_A]$ is the $N \times A$ matrix of the first A score vectors of the model, $\mathbf{P}_A = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_a, ..., \mathbf{p}_A]$ is the $K \times A$ matrix of the first A loading vectors of the model and $\hat{\mathbf{X}}_A$ is the prediction values for the data set based on A PCs.

The residual matrix for the PCA model for the total number of components $A = 1,...,A_{\max}$ is calculated according to:

$$\mathbf{E}_A = \mathbf{X} - \hat{\mathbf{X}}_A \qquad (15)$$

where $\mathbf{X}$ is the mean-centered data set, $\hat{\mathbf{X}}_A$ is the predicted values of data estimated by Eq. 14 and $\mathbf{E}_A$ is the residual matrix for A PCs.

The sum of squares for $A = 1,...,A_{\max}$ components is obtained as:

$$SScal_A = \sum_{i=1}^{N} \sum_{k=1}^{K} \left(e_A(i,k) \right)^2 \qquad (16)$$

where $e_A(i,k)$ is the (i, k)-th entry of residual matrix $\mathbf{E}_A$ calculated in Eq. 15.

Finally, the Mean Squared Errors for the calibration model is calculated according to

$$MSEcal_A = \frac{SScal_A}{N(K-A) - DF_A} \; (17)$$

where DF_A is the number of DFs that are consumed for finding the A PCs of the PCA model and which will be estimated in the following.

Eq. 17 can also be reformulated as:

$$MSEcal_A = \frac{SScal_A}{N\left(K-\left(A+\frac{DF_A}{N}\right)\right)} = \frac{SScal_A}{N\left(K-EDF_A\right)} \qquad (18)$$

where DF_A is the number of DFs consumed during the search process and EDF_A represents the total number of DFs consumed for modeling the data set with *A* PCs.

2.5.1.2 Cross-validated Mean Squared Errors (MSE$_{CV}$)

In order to calculate MSE$_{\mathrm{CV,A}}$, segments of data which are indexed by $m=1,...,M$ are left-out, in turn, resulting in a leave-in segment and a left-out segment of the data. During each round of cross-validation, a PCA model is established for the leave-in data. The left-out segment is then fitted to the model in order to calculate MSE$_{\mathrm{CV,A}}$ for the left-out data segment. A detailed description for calculating MSE$_{\mathrm{CV,A}}$ in a multi-block situation is given in [10] which also covers the case of one block which amounts to PCA. Thus, MSE$_{\mathrm{CV,A}}$ in PCA is given by:

$$MSEcv_A = \frac{SScv_A}{N\left(K-A\right)} \quad (19)\ [9]$$

where $SScv_A$ is the cross-validated sum of squares for $A=1,...,A_{\max}$. The parameter DF, which was included in Eq. 17, does not appear when cross-validated mean squared errors since there is no consumption of DFs in fitting the model except for *A*, which reflects the number of PC scores estimated in each local “test”-sample during cross-validation.

We have here chosen to use cross-validation for estimating MSE$_{\mathrm{CV,A}}$. The motive for this is to ensure that the same data set is used both for calibration and for validation in the simulations. Moreover, we wanted to study the performance of MSE$_{\mathrm{CV,A}}$ with different number of cross validation segments ($m=1,2,...,M$). It is worth noting that it is possible to replace cross-validation calculations by an independent test set. An independent test set would in that case be treated like a leave-out segment of data i.e. MSE$_{\mathrm{Test}}$ calculations would be similar to those for a leave-out segment of data.

2.5.1.3 Calculating the number of DFs in PCA

In order to estimate the DFs, we require that the calibration Mean Squared Error and the cross-validated Mean Squared Error are equal. This condition is justified since they are calculated for the same data set. We obtain

$$MSEcal_A = MSEcv_A \quad (20)$$

where $MSEcal_A$ is the calibration mean squared error calculated by Eq. 17 and $MSEcv_A$ is the cross-validated mean squared error calculated in Eq. 19. By solving Eq. 20 we obtain an estimate of the DF:

$$DF_A = N\left(K - A\right)\left(1 - \frac{SScal_A}{SScv_A}\right) \quad (21)$$

If an independent test set is used instead of cross-validation, $SScv_A$ in Eq. 21 will be replaced by $SStest_A$. However, it is important to note that when using an independent test set instead of cross-validation, equality between MSE$_{\text{Cal}}$ and MSE$_{\text{Test}}$ is valid if we assume that the test set arises from the same population as the training set.

This formula, based on our choice of simple variance estimation principles, is slightly different from that used by Van Der Voet [5], who used a formula where DF is a function of the square root of the variances.

2.5.2 Estimated Degree of Freedom (EDF) in CPCA

2.5.2.1 Calibration Mean Squared Error for every block (MSE$_{Cal}$)

Mean squared error calculations for the multi-block setting are very similar to those for a single-block data set. The calculation of errors is done according to Eq. 15 for the matrix of concatenated data blocks ($\mathbf{X} = \left[\mathbf{X}^1 \ \mathbf{X}^2 ... \ \mathbf{X}^b ... \ \mathbf{X}^B\right]$). The sum of squared errors is then calculated for each block separately according to:

$$SScal_A^b = \sum_{i=1}^{N} \sum_{k=1}^{Kb} \left(e_A^b(i,k)\right)^2 \qquad (22)$$

where N is the number of samples, K_b is the number of variables in block *b* of the multi-block data set, $e_A^b(i,k)$ is the (*i*,*k*)-th entry of residual matrix $\mathbf{E}_A^b$ and $\mathbf{E}_A^b$ is the block *b* of the residual matrix $\mathbf{E}_A = \left[\mathbf{E}_A^1 \ \mathbf{E}_A^2 ... \ \mathbf{E}_A^b ... \ \mathbf{E}_A^B\right]$ calculated for the concatenated multi-block data set by Eq. 15. The global sum of squared errors $SScal_A^g$ is calculated according to Eq. 16.

The global Calibration Mean Squared Errors ($MSEcal_A^g$) is calculated according to Eq. 17. While the calculation of Calibration Mean Squared Errors for every block ($MSEcal_A^b$) in the multi-block setting is slightly different from that of the single-block data. The term A in Eq. 17, representing the A degrees of freedom consumed in predicting A global scores, needs to be modified for a multi-block setting. Instead of A we suggested to use the partial block leverage h_A^b [10]. For the block $MSEcal$ we then obtain:

$$MSEcal_A^b = \frac{SScal_A^b}{N\left(K_b - h_A^b\right) - DF_A^b} \qquad (23)$$

where DF_A^b is the number of DFs that are consumed in block b of the multi-block data set for finding the A PCs of the multi-block model and the quantity h_A^b is the partial block leverage, intended to represent block b contribution in the A degrees of freedom consumed in predicting A global scores $\mathbf{t}_a$. The partial block leverages h_A^b for A components and every block are calculated according to Eq. 24:

$$h_A^b = \sum_{a=1}^{A} \sum_{k=b1}^{b2} p_{(a,k)}^2 \qquad (24)$$

where $\sum_{k=b1}^{b2} p_{(a,k)}^2$ calculates the sums of squared values in the loading matrix pertaining to block b of the multi-block data set, $p_{(a,k)}^2$ is the (a,k)-th squared entry of the loading matrix $\mathbf{P}_A'$ and $\mathbf{P}_A = \left[\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_a, ..., \mathbf{p}_A\right]$ is the loading matrix for the first A PCs defined in Eq. 14. b_1 and b_2 are the column numbers of the first and the last variables in block b, respectively, considering all blocks. It is worth noting that the sum of the partial leverages equals A:

$$\sum_{b=1}^{B} h_A^b = \sum_{a=1}^{A} \sum_{k=1}^{K} p_{(a,k)}^2 = A \qquad (25)$$

Finally, similarly to single-block data, Eq. 23 is reformulated for calculating the total number of DFs consumed in each block (EDF_A^b):

$$\begin{aligned} MSEcal_A^b &= \frac{SScal_A^b}{N\left(K_b - \left(h_A^b + \frac{DF_A^b}{N}\right)\right)} \\ &= \frac{SScal_A^b}{N\left(K_b - EDF_A^b\right)} \end{aligned} \qquad (26)$$

where DF_A^b is the number of DFs consumed in block *b* during the process of determining component *A* and EDF_A^b represents the total number of DFs consumed in block *b* for modeling the data set with *A* PCs, reflecting both the data-driven choices leading to the PC directions and the actual estimation of the component parameters (its loadings and scores).

2.5.2.2 Cross-validated Mean Squared Error for every block (MSE$_{CV}$)

Cross-validated mean squared errors calculations in the multi-block situation is done according to [10]. Since MSE$_{\text{CV}}$ does not involve the estimation of any parameter, the formula for MSE$_{\text{CV}}$ calculation for every block $b = 1,...,B$ is written without the number of DF as:

$$MSEcv_A^b = \frac{SScv_A^b}{N\left(K_b - h_A^b\right)} \qquad (27)\ [10]$$

where K_b is the number of variables in block *b* and the quantity h_A^b is the partial block leverage, intended to represent block *b* contribution in the *A* degrees of freedom consumed in predicting *A* global scores $\mathbf{t}_a$. Partial block leverage calculation was given in Eq. 24.

2.5.2.3 Calculating EDF in CPCA

In order to calculate DFs in a multi-block setting similar considerations as for one block can be made: The Mean Squared Error of every block are assumed to be equal when estimated in calibration and cross-validation. This assumption leads to:

$$MSEcal_A^b = MSEcv_A^b \quad (28)$$

where $MSEcal_A^b$ is the calibration Mean Squared Error for block *b* calculated by Eq. 23 and $MSEcv_A^b$ is the block cross-validated Mean Squared Error calculated by Eq. 27. As in the PCA we can solve Eq. 28 for the DF_A^b and obtain:

$$DF_A^b = N\left(K_b - h_A^b\right)\left(1 - \frac{SScal_A^b}{SScv_A^b}\right) \qquad (29)$$

3. Results and discussions

3.1 Simulated data sets

Since the relative distribution of singular value structure of a data set defines its correct PCA solution and also influences the process of determining these PCs, we undertook a simulation study where this structure is controlled by the tuning parameter m_1 in Eq. 4. Fig. 1a shows different singular value structures produced by different tuning parameters m_1. The simulated data sets in this figure contain 500 samples and 200 variables (note that no noise is added to the data sets). It can be seen that $m_1 = 2$ gives linear decrease in the singular values. Increasing m_1, from 2 to 500, shifts the structure toward having few large singular values and many small ones whereas decreasing m_1, from 2 to 0.01, influences the structure in the opposite way (i.e. many large singular values and few small ones). The figure also shows that the formula given in Eq. 4 encompasses a wide range of possible singular value structures.

The noise level in the simulated data set is controlled by the parameter σ_1^2 of Eq. 7. Fig. 1b shows the effect of different levels of noise on the singular value structure of a simulated data set. The simulated data set for this figure contain 500 samples and 200 variables. The tuning parameter $m_1 = 200$ is used for the simulations in Fig. 1b. It can be seen that increasing the noise level from 0 to 1 can tremendously change the singular value structure of the data. Since we wanted to deal with a defined singular value structure we did not use noise levels above 0.005 in our simulations. However, the level of the added noise was close to what we might encounter in real situations.

We compared the simulated data to four different real data sets containing 88 sample spectra from FTIR [13] [14] . Singular value structures of real data sets are shown in Fig. 1c. Simulated data sets with similar structures to our real data sets are shown by dashed lines. The noise level of the simulated data sets was set to 0.0005. As shown in the figure, different values of m_1 were used in order to control the singular value structures for the simulated data sets. Although the real data sets (from spectroscopy) used here in pertain to very different FTIR sampling techniques and biological materials (i.e. listeria, meat tissue samples and yeast samples) they still show similar behavior: FTIR data sets show a tendency to having few large singular values and many small ones. Fig. 1c leads to the conclusion that m_1 value chosen between 30 and 70 can be a realistic choice when we want the simulated data to have a similar structure as real data sets in FTIR spectroscopy.

3.2 MSEcv in PCA

In order to assess the validity of the tentative formula for MSE_{CV} in Eq. 19 we have simulated 500 data sets, each of them containing 500 samples and 200 variables, according to Section 2.2.1 (Eq. 7) where the parameters $m_1 = 60$ and $\sigma_1^2 = 0.005$ were used for simulations. Fig. 2a shows the singular value structures of the simulated data sets before adding noise (in green) and after adding noise (in red). For the cross-validation 5 randomly chosen segments (i.e. 100 samples in each segment) were used. Thereafter MSE_{CV} was calculated using two different formulas: firstly by Eq. 19, where *A* is subtracted in the denominator and secondly without subtracting *A*. Cross-validated Root Mean Squared Errors (RMSE_{CV} i.e. the root of the MSE_{CV}) plots calculated both with and without subtracting *A* together with RMSE_{Cal} without DF-correction (i.e. root of MSE_{Cal} in Eq. 17 when $DF_A = 0$) are plotted in Fig. 2b for all PCs, with the first PCs enlarged in Fig. 2c. RMSE_{CV} with rank corrected DFs (Eq. 19) drops drastically during the description of the first few components, expected to contain systematic information, while it flattens out for the subsequent PCs, expected to reflect random noise. In contrast, RMSE_{CV} without the rank correction does not flatten out, and neither does the RMSE_{Cal} without DF-correction. The horizontal dashed line plotted in Fig. 2c highlights the fact that RMSE_{CV} calculated from Eq. 19 clearly flattens out. Flattening out of the plots is important here since it shows that residual reduction caused by adding more components to the model is compensated by the DF-correction, and demonstrates that adding more PCs beyond the interesting ones does not improve the model. This property of Eq. 19 is useful for choosing the number of PCs to include in a PCA model.

3.3 MSEcv in CPCA

In order to corroborate the tentative formula in Eq. 27 for CPCA we proceed similarly to the case of PCA in Section 3.2. We have simulated 500 multi-block data sets. Each data set contains 4 blocks of data. There are 500 samples and 200 variables in every block. The method that is used for data simulation was described in Section 2.2.2. The noise level was set to be $\sigma_b^2 = 0.005$ for $b = 1,...,4$. Again, 5 random segments of rows (i.e. 100 samples in each segment) were used for cross-validation. Tuning parameters m_b that control the singular value structures of the data blocks are as follow: $m_1 = 55$, $m_2 = 60$, $m_3 = 65$ and $m_4 = 70$. RMSE_{CV}s are firstly calculated using Eq. 27 and, thereafter, without subtracting the partial block leverages (h_A^b). RMSE_{CV}s for every block and global $\mathrm{RMSE_{CV}}$s are plotted in Fig 3. Again, correcting the $\mathrm{MSE_{CV}}$ with a term for model rank, in this case distributed as partial block leverages, $\mathrm{RMSE_{CV}}$ plots flatten out when enough number of PCs have been included in the multi-block model.

3.4 EDF in PCA

We have simulated 500 data sets for calculating the DFs being consumed during the PCA process. DFs in this section are estimated by two different methods: (1) Using cross-validation. (2) Using a separate test set. The results from both methods are compared to DFs calculated by the formula proposed by Faber in [4] (i.e. $(N-a-1)(K-a)$ degrees of freedom for the residuals of a PCA model). The simulated data sets are generated according to the procedure described in Section 2.2.1, each of them containing 2500 samples and 200 variables. For these simulations, we used the parameters $m_1 = 60$ and $\sigma_1^2 = 0.005$. Five segments of samples (i.e. 100 samples in each segment) were used for cross-validation. Each simulated data set is divided randomly into two sets of test set (2000 samples) and training set (500 samples) resulting in 500 simulated test sets and training sets. Fig. 4a shows the singular value structure for 500 sets of 2500 samples together with the structure of the test sets and training sets. The former, with a lower aspect ratio (# samples / # of variables), shows a steeper decline in singular values than the latter. This was expected, since for singular value decomposition of random data tables, an aspect ratio of 1 gives singular values decreasing linearly to zero, while a very high or very low aspect ratio yields almost constant singular values.

RMSE according to Faber, $\mathrm{RMSE_{Cal}}$ without DF-correction (calculated from Eq. 17 when $DF_A = 0$) and $\mathrm{RMSE_{CV}}$ (Eq. 19) were calculated for the training sets while $\mathrm{RMSE_{Test}}$ were calculated for the test sets. RMSE plots ($\mathrm{RMSE_{Cal}}$ not DF-corrected, $\mathrm{RMSE_{CV}}$, $\mathrm{RMSE_{Test}}$ and RMSE according to Faber) are shown in Fig. 4b (enlarged in Fig. 4c). It can be seen that $\mathrm{RMSE_{Test}}$ calculated for the independent test sets (plotted in red) are very similar to the mean of $\mathrm{RMSE_{CV}}$ calculated for the training sets (plotted in blue). However, they are very different from RMSE calculated according to Faber's formula. RMSE calculated according to Faber are more similar to $\mathrm{RMSE_{Cal}}$ (not DF-corrected) indicating that the DFs consumed in the search process are not taken into account. DF was then estimated by cross-validation and by independent test set. The EDF results from individual simulations are plotted in Fig. 4d and averaged in Fig. 4e (red and blue plots). Some simulation runs got negative EDFs in their very first components, probably due to rotational differences or PC ordering differences between calibration and validation models (it is not possible to identify them in Fig. 4d since the plots are truncated at EDF equal to zero). But the average EDFs calculated for those components (Fig. 4e) are far from being negative. This figure reveals that over-all, EDFs calculated by both methods have the same structure on average, although they may look different for their very first components in Fig. 4d. Fig. 4f shows EDFs for the first 30 components. It can be seen that both methods, lead to EDFs which are, on average, very close to each other. EDFs according to Faber are also plotted in Fig. 4d-f (in green). Faber's formula underestimates the consumed DFs in the first PCs and overestimates them in the last PCs. This highlights the fact that the search process in PCA (for the given singular value structure) consumes much more DFs for finding the first PCs compared to the last ones.

The reason for calculating EDF by two different methods was to see the effect of using a separate test set on the EDF results versus using cross-validation. We did not detect any significant difference between the two methods (comparing the blue and red plots in Fig. 4e-f). However, since the number of cross-validation segments may affect the EDF results, we will investigate this issue in detail in Section 3.6.

3.5 EDF in CPCA

For calculating EDF in CPCA, we used cross-validation with the same number of segments as in Section 3.4 and ran the simulations 500 times. Each of the 500 multi-block sets are simulated with the following properties: Four blocks of data with 500 samples and 200 variables in each block. We followed the simulation procedure described in Section 2.2.2. The noise level was $\sigma_b^2 = 0.005$ for $b = 1,...,4$. Five segments of data (i.e. 100 samples in each segment) were used for cross-validation. The tuning parameters m_b are as follow: $m_1 = 55$, $m_2 = 60$, $m_3 = 65$ and $m_4 = 70$. EDFs calculated for each block and the global EDF, for all of the 500 simulations, are plotted in Fig. 5. Fig. 6 shows a zoom out of the average EDFs calculated for every block and globally for the first 30 components. It can be seen that the average EDF in different blocks are similar to each other. This is probably due to the fact that all the four blocks share the same sample patterns in this illustration. The simulated multi-block data sets are, as a matter of fact, simple examples which are used for visualization purposes. EDF for more complicated cases can be calculated exactly in the same way.

3.6 Number of cross-validation (CV) segments

In cross-validation the user has to specify how to split the samples into segments. In order to assess the effect due to the number of CV segments we compared results for different numbers of CV segments and results from using an independent test set. For this purpose we have simulated 5 data sets containing 2500 samples and 200 variables. 500 samples (from 2500 samples in each data set) are used as training sets while the rest 2000 samples are set aside to be used as independent test sets. $m_1 = 60$ and $\sigma_1^2 = 0.005$ are used as the parameters for data simulations. RMSE_{Cal} (not DF-corrected) and $\text{RMSE}_{\text{Test}}$ are calculated for training sets and test sets respectively. Then, we calculated RMSE_{CV} for training sets using different number of CV segments. Similar procedures as in Section 3.4 were applied here for calculating EDFs for test sets versus cross-validation using different choices of CV segments: 2, 5, 20, 100 and finally 500 (i.e. leave-one-out CV).

Fig 7a shows RMSE plots when samples are divided into two segments (i.e. 250 samples in each segment) for CV calculations. It can be seen that RMSE_{CV} (plotted in blue) and $\text{RMSE}_{\text{Test}}$ (plotted in red) are slightly different from each other. EDF calculated from test set and CV are plotted in Fig. 7b-c. It can be seen that EDF plots based on test sets and 2-fold cross-validation are not very different from each other.

The same calculations that were applied for 2 segments were repeated for 5, 20, 100 and 500 segments on the same simulated data sets. RMSE plots for leave-one-out CV are shown in Fig. 7d. It can be seen that the leave-one-out CV results are noisier than those from split-half CV. EDF plots for the leave-one-out CV are shown in Fig. 7e-f and appear to be very noisy compared to those from split-half CV. Even the average plot (blue plot in Fig. 7f) seems to be extremely noisy compare to the blue plot in Fig. 7c. The plots for intermediate segmentation schemes (5-fold, 20-fold and 100-fold CV) gave intermediate results; they are not shown here, for simplicity.

As a rule of thumb, scientists have always believed that running cross-validation with more CV segments gives better and more robust results. Especially, the leave-one-out CV is supposed to give the most stable results due to the fact that the models fitted to the data are not expected to change very much since only one sample is set aside at a time. One can see that the results from our simulation study indicate that a price is thereby paid in estimation precision.

3.7 DF's application for real data sets

In order to estimate the DF for real data sets in FTIR spectroscopy we ran a simulation study where we simulated 500 data sets using the eigenvalue structure of a real dataset. Data sets were simulated using the eigenvalue structure of FTIR data of 88 listeria samples with 2282 absorbance variables. For the simulation study, the procedure described in Section 2.2.1 was followed and the eigenvalue structure shown in Fig. 1c (the black curve) was used. An error was added with $\sigma_1^2 = 0.0001$. For the cross validation four segments were used with 22 samples in each. The DFs were estimated as before and results were averaged using 500 simulations.

In order to check if the DF estimate could be used to correct the RMSE of a new independent data set, we used different data sets from FTIR spectroscopy for the calculation of RMSE and corrected them by the DF obtained from the Listeria data set. The three new datasets were two FTIR data sets of yeast strains and one data set from FTIR microspectroscopy of meat tissues. All sets were obtained by different instruments. Their eigenvalue structures are shown in Fig. 1c. All three sets contained several thousand of samples but with the same number of variables as in the listeria set, making it possible to sample 100 times randomly 88 samples from each of them for calculating RMSE. Fig. 8 shows average RMSE plots for the Listeria set and the three other sets: RMSE_{Cal} not DF-corrected (plotted in green), RMSE_{CV} (plotted in red) and RMSE_{Cal} DF-corrected (plotted in blue). The figure shows that correcting RMSE_{Cal} for the number of DFs being consumed gives a fairly good estimation of RMSE_{CV} and hence the optimal rank, without running any CV on the data sets. Considering the correct number of DFs that are being consumed for modeling a data set and applying them correctly has enabled us replacing the time- and memory-consuming procedure of cross-validation by easy-to-calculate formulas. As it can be seen in Fig. 8 RMSE plots calculated by DF correction (plotted in blue) are fairly close to the RMSE_{CV} plots calculated by cross-validation (plotted in red) but their differences get larger as the number of PCs increases. Unlike what we observed on the plots based on simulated data sets the DF corrected plots are not perfectly flattening herein. The reason could be that the samples and/or the variables in the input data are not completely independent, as assumed by the present theory. For instance, as part of the Fourier transform, the noise in the FTIR variables may have become somewhat intercorrelated. The lack of independence may be compensated for by modifications of the proposed MSE and EDF formulas.

4. Conclusion

Degrees of freedom (DFs) are consumed when parameters are estimated in PCA and CPCA. We have shown that the search process itself consumes some DFs and therefore the total DF consumption during the modeling process is more than (only) the number of independently estimated parameters. However when cross-validating CPCA models there is no DF consumption due to the search process, the obtained Mean Squared Errors (MSE) still need to be corrected for the DFs consumed by the estimation of the model parameters for the leave-out segments. Since block loadings in CPCA are not independent, we previously suggested that the MSE estimated by cross-validation for each block should be corrected by the leverage of the block loadings. In the present paper we have simulated data sets to show that the cross-validated Root Mean Squared Error (RMSE) plots for each block flatten out when corrected by our previously proposed formula. This proves the validity of the proposed formula.

When errors are calculated without cross-validation, simply by fitting, the number of DFs needed for correcting the errors is higher. This extended number of DFs is due to the search process and depends on the eigenvalue structure of the data. The process of searching a principal component is faster, when eigenvalues are decreasing rapidly. In this study we have proposed an easy and straightforward method for estimating the number of consumed DFs while looking for the PC directions. We have also estimated DFs for a real data set from FTIR spectroscopy of biological material and afterwards applied it to the other data sets from the same field. The results reveal the fact that we do not necessarily need to estimate DF for every individual data set. Instead it is possible to calculate DFs for one data set from a specific field and apply the estimated DFs to future data sets within the same field. Therefore, we believe that estimating cross-validated errors for a PCA or CPCA model can be obtained without running cross-validation, when a characteristic DF is already known.

Another important challenge in PCA and CPCA data analysis is to assess the reliability of a model. Using an independent test set seems to be a good choice for this purpose. However, because of a lack of sufficient samples, it is not always possible to set aside a subset of the observations to serve as a validation set. Cross-validation has been in focus as an alternative method. However, it has been criticized since the same samples are used both for the modeling procedure as well as for validating the model. Our simulation study reveals an interesting finding: the outcomes from a cross-validation study agree to a large extent with those from an independent test set. One should note that the choice of the number of segments to be used for running cross-validation affects significantly the results. In this article, we have shown that an increased number of CV-segments do not necessarily lead to better results. In our case, it turned out that a 10-fold cross-validation gives results which are close to those from an independent test set.

Acknowledgements

The authors are grateful for financial support by the Nordic Centre of Excellence on Food, Nutrition and Health “Systems biology in controlled dietary interventions and cohort studies” (SYSDIET) funded by NordForsk, by Foundation for Research Levy on Agricultural Products in Norway and the grant 203699 (New statistical tools for integrating and exploiting complex genomic and phenotypic data sets) financed by the Research Council of Norway.

References

[1] E. Pennestri, M. Cavacece, L. Vita, On the Computation of Degrees-of-Freedom: A Didactic Perspective, ASME Conference Proceedings 2005 (2005) 1733-1741.

[2] J. Jaccard, M.A. Becker, Statistics for the behavioral sciences., Belmont, CA: Wadsworth., 1990.

[3] C.H. Yu, M. Lovric, Degrees of Freedom, International Encyclopedia of Statistical Science, Springer Berlin Heidelberg, 2011, pp. 363-365.

[4] N.M. Faber, Degrees of freedom for the residuals of a principal component analysis -- A clarification, Chemometrics and Intelligent Laboratory Systems 93 (2008) 80-86.

[5] H. van der Voet, Pseudo-degrees of freedom for complex predictive models: the example of partial least squares, 13 (1999) 195-208.

[6] N. Krämer, M.L. Braun, Kernelizing PLS, degrees of freedom, and efficient model selection, Proceedings of the 24th international conference on Machine learning, ACM, Corvalis, Oregon, 2007.

[7] K. Kato, On the degrees of freedom in shrinkage estimation, Journal of Multivariate Analysis 100 (2009) 1338-1352.

[8] N. Krämer, M. Sugiyama, The Degrees of Freedom of Partial Least Squares Regression, Journal of the American Statistical Association 106 (2011) 697-705.

[9] H. Martens, T. Næs, Multivariate Calibration, John Wiley & Sons, Chichester, 1989.

[10] S. Hassani, H. Martens, E.M. Qannari, M. Hanafi, G.I. Borge, A. Kohler, Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods, Chemometrics and Intelligent Laboratory Systems 104 (2010) 140-153.

[11] H. Martens, M. Martens, Multivariate Analysis of Quality: An Introduction, Wiley, Chichester, UK, 2001.

[12] F. Arteaga, A. Ferrer, How to simulate normal data sets with the desired correlation structure, Chemometrics and Intelligent Laboratory Systems 101 (2010) 38-42.

[13] A. Oust, T. Moretro, K. Naterstad, G.D. Sockalingum, I. Adt, M. Manfait, A. Kohler, Fourier transform infrared and Raman spectroscopy for characterization of Listeria monocytogenes strains, Appl Environ Microb 72 (2006) 228-232.

[14] N. Perisic, N.K. Afseth, R. Ofstad, A. Kohler, Monitoring Protein Structural Changes and Hydration in Bovine Meat Tissue Due to Salt Substitutes by Fourier Transform Infrared (FTIR) Microspectroscopy, Journal of Agricultural and Food Chemistry 59 (2011) 10052-10061.

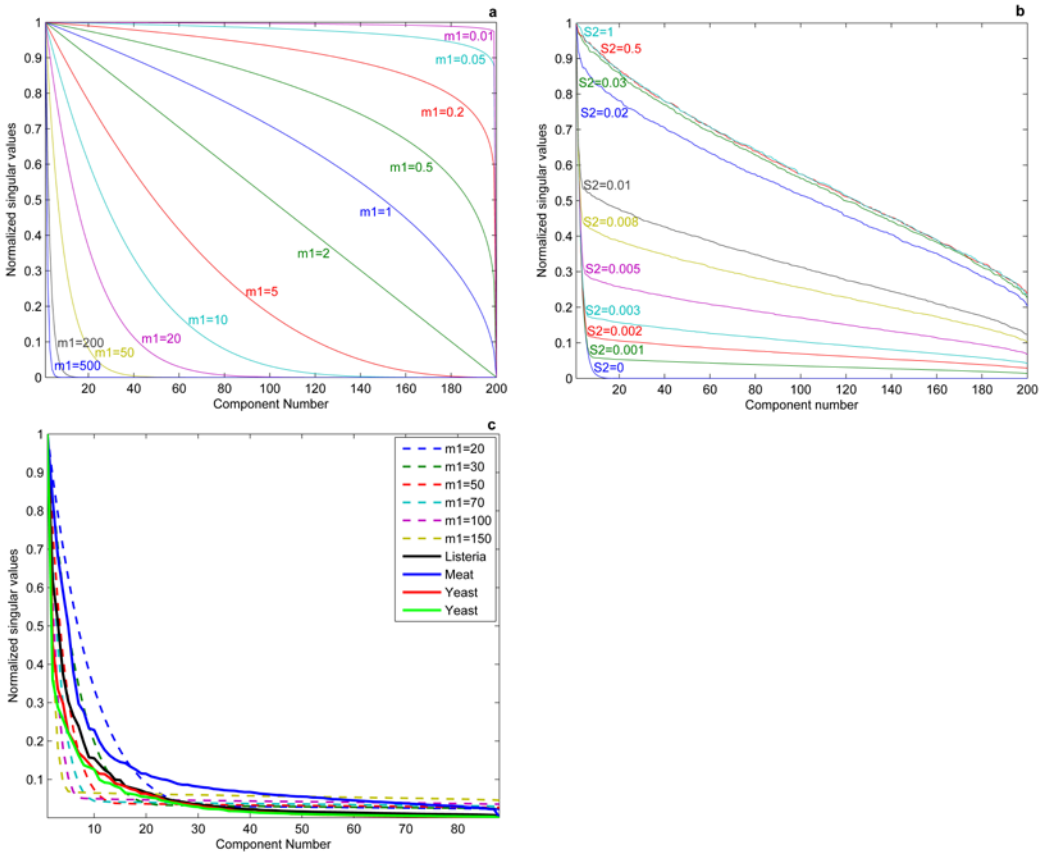


Figure 1: Normalized singular values. Singular value structures for simulated data sets containing 500 samples and 200 variables are shown in (a) and (b): (a) Singular value structures for data sets generated by different tuning parameter (m_1) (without adding any noise) are plotted. (b) Singular value structures for data sets generated by adding different noise levels to the same data structure, are plotted. Tuning parameter $m_1 = 200$ is used here. (c) Normalized singular values for different simulated data sets, with different tuning parameter (m_1) values, are plotted together with normalized singular values for real data sets containing FTIR spectra. The same noise level (i.e. 0.0005) is used for all of the simulations.

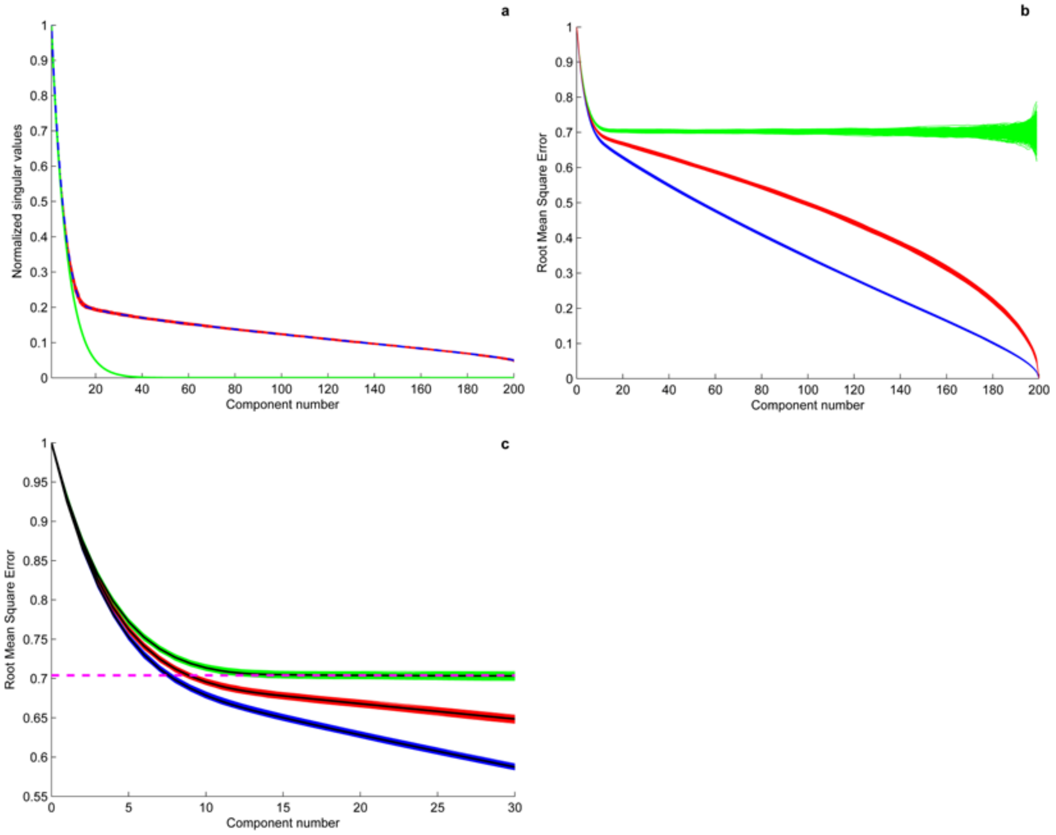


Figure 2: 500 simulated single-block data sets containing 500 samples and 200 variables. (a) Normalized singular value structures of the simulated data sets (parameters $m_1 = 60$ and $\sigma_1^2 = 0.005$). The singular values structure before adding noise is plotted in green. Red plots show the structures after adding noise. The dashed blue line is the average of the red plots. **(b,c) Single-block Root Mean Squared Error (RMSE) plots.** RMSE$_{\text{Cal}}$ (without DF-correction) plots are plotted in blue. RMSE$_{\text{CV}}$ not corrected for *A* degrees of freedom consumptions are plotted in red. RMSE$_{\text{CV}}$ corrected for *A* degrees of freedom consumptions are plotted in green. (b) RMSE plots plotted for 200 components. (c) A zoom out of plots in (b) for the first 30 components together with the average lines in black. A horizontal dashed line is plotted in pink for visualizing that RMSE$_{\text{CV}}$ corrected for DF consumption flattens out for PCs presumably dominated by random noise.

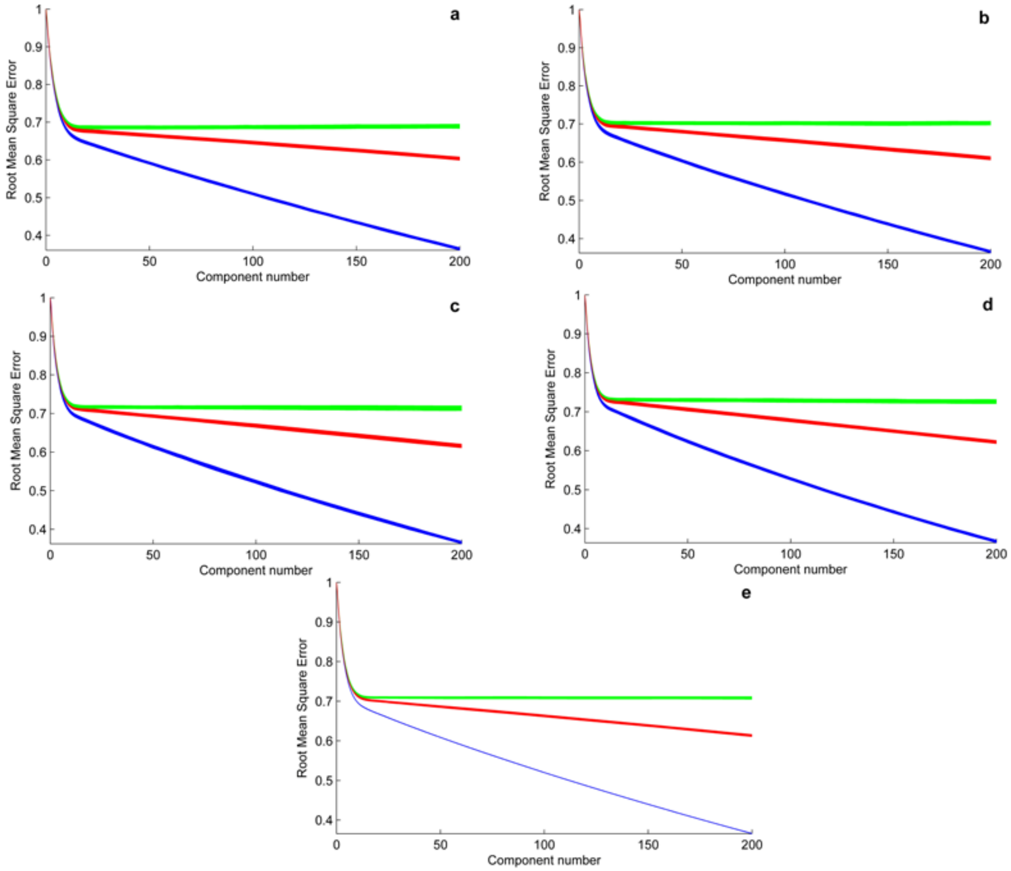


Figure 3: Multi-block Root Mean Squared Error (RMSE) plots for 500 simulated data sets, using 4 blocks of data, each of them containing 200 variables. $\mathrm{RMSE}_{\mathrm{Cal}}$ (without DF-correction) plots are plotted in blue. $\mathrm{RMSE}_{\mathrm{CV}}$ not corrected for the degrees of freedom consumptions are plotted in red. $\mathrm{RMSE}_{\mathrm{CV}}$ corrected for the degrees of freedom consumptions are plotted in green. (a-d) RMSE plots for different blocks. (e) Global RMSE plots.

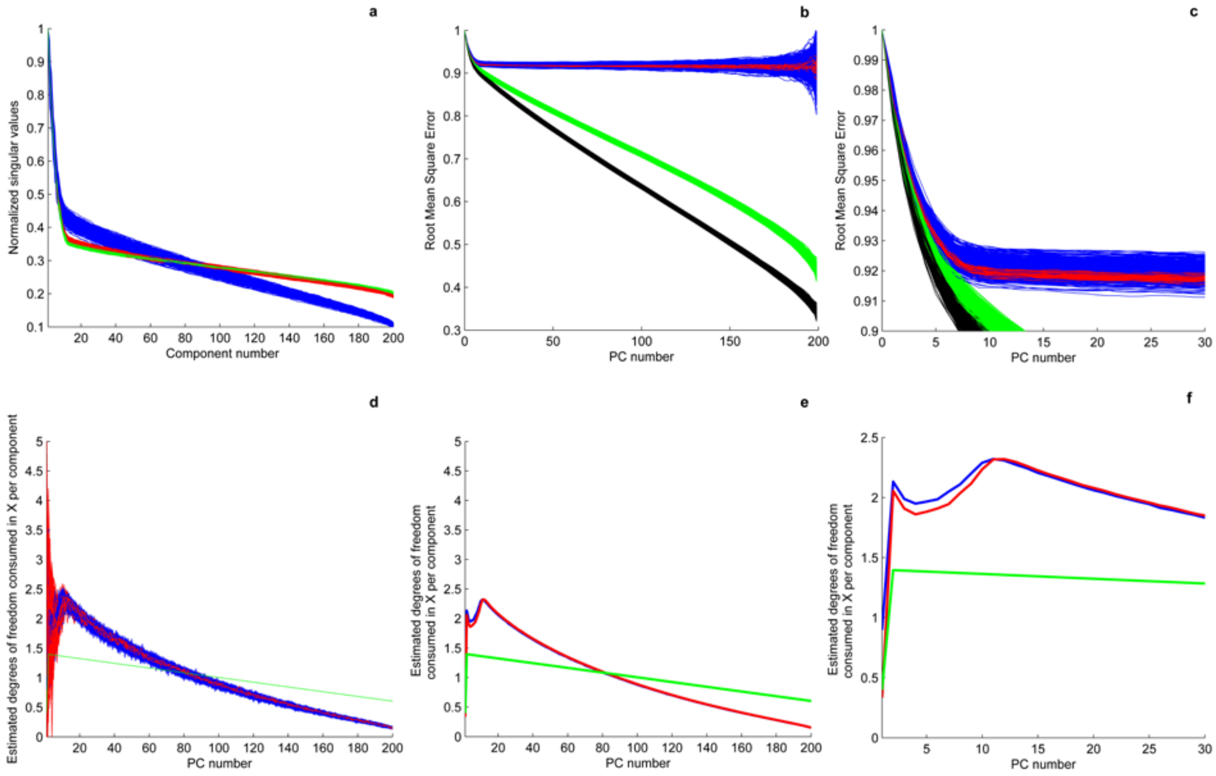


Figure 4: 500 simulated training and test data sets. (a) Normalized singular values for 500 simulated data sets, each of them with 2500 samples, are plotted in green. Normalized singular values for 500 simulated training sets, each of them with 500 samples, are plotted in blue. Normalized singular values for 500 simulated test sets, each of them with 2000 samples, are plotted in red. Every data set contains 200 variables. (b) $\mathrm{RMSE}_{\mathrm{Cal}}$ (without DF-correction) plots are plotted in black. $\mathrm{RMSE}_{\mathrm{CV}}$ plots for the training sets are plotted in blue. $\mathrm{RMSE}_{\mathrm{Test}}$ plots are plotted in red. RMSE plots corrected for DF according to Faber's formula are plotted in green. Training sets and test sets have 500 and 2000 samples respectively. (c) A zoom out of the plots in (a) for the first 30 components. (d) Estimated degrees of freedom (EDF) consumption using cross-validation are plotted in blue. EDF consumption using test set are plotted in red. EDF consumption according to Faber's formula are plotted in green. (e) The average of the 500 plots in (d). (f) A zoom out of plots in (e) for the first 30 components.

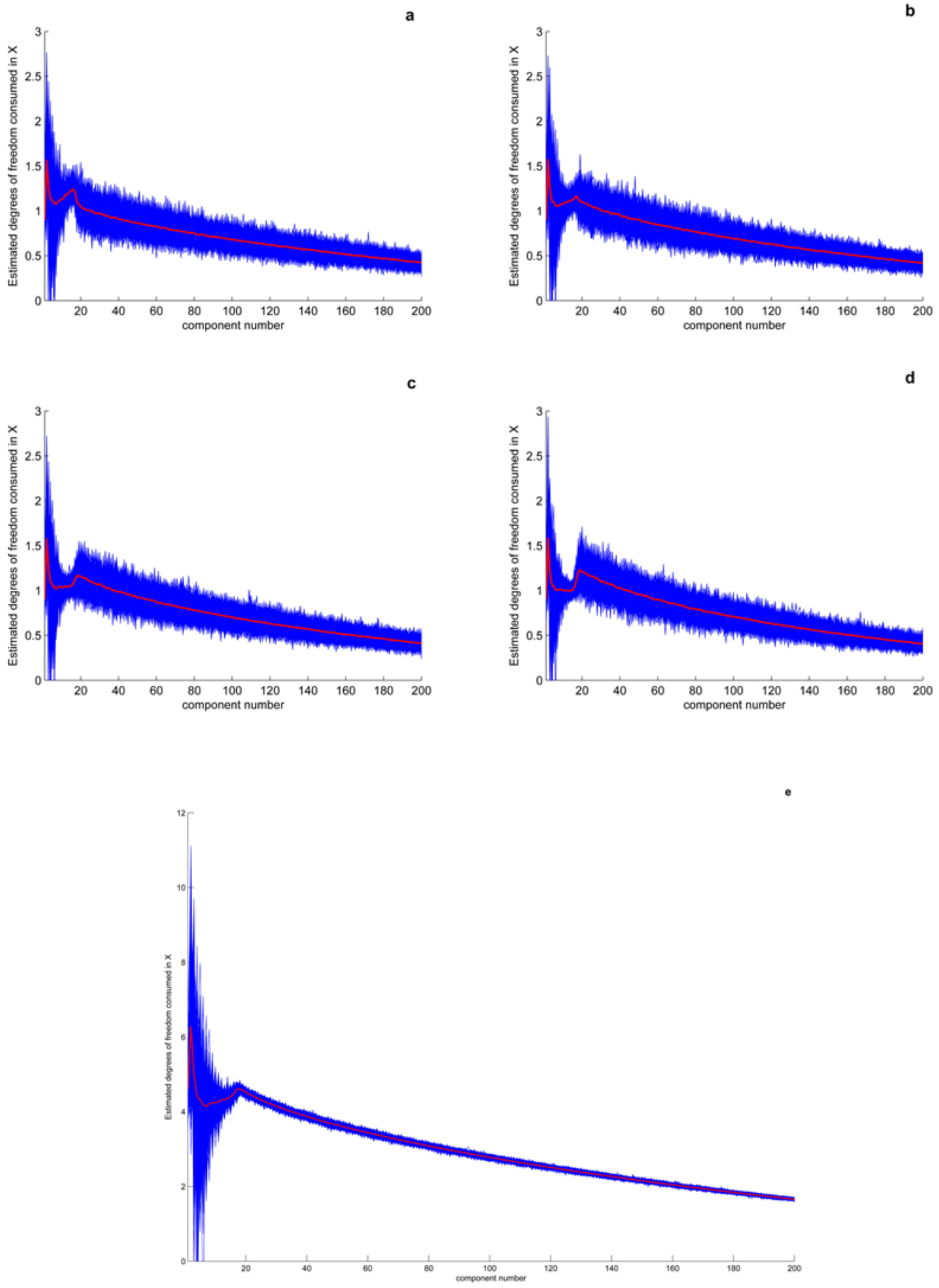


Figure 5: Estimated Degrees of Freedom (EDF) for 500 multi-block simulated data sets. EDF consumptions using cross-validation are plotted in blue. The average EDF consumptions are plotted in red. (a-d) EDF for different blocks. (e) Global EDF.

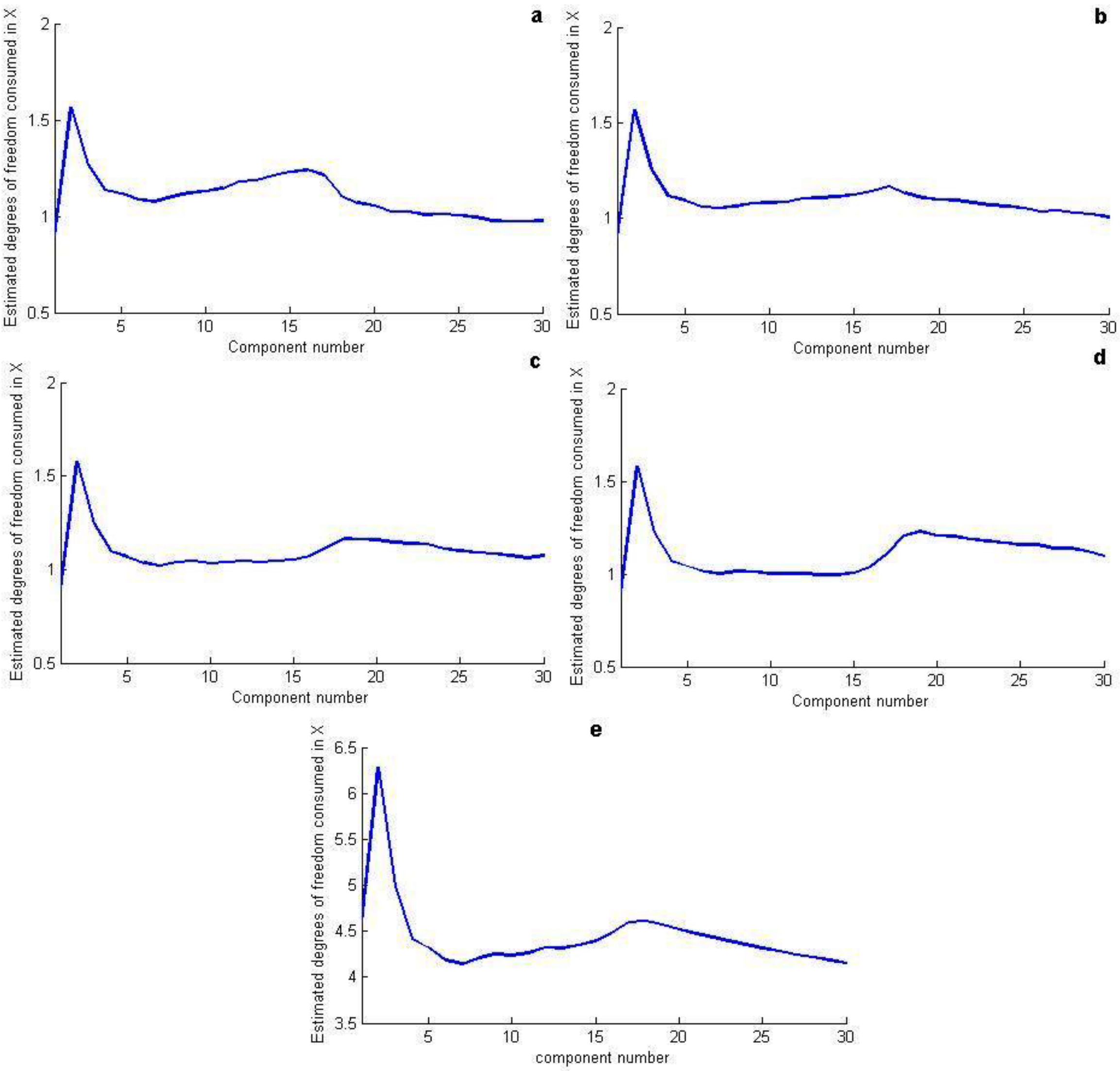


Figure 6: Zoom out of the plots in Fig. 5. A zoom out of the average plots in Fig. 5 are plotted for the first 30 components. (a-d) EDF in different blocks. (e) Global EDF.

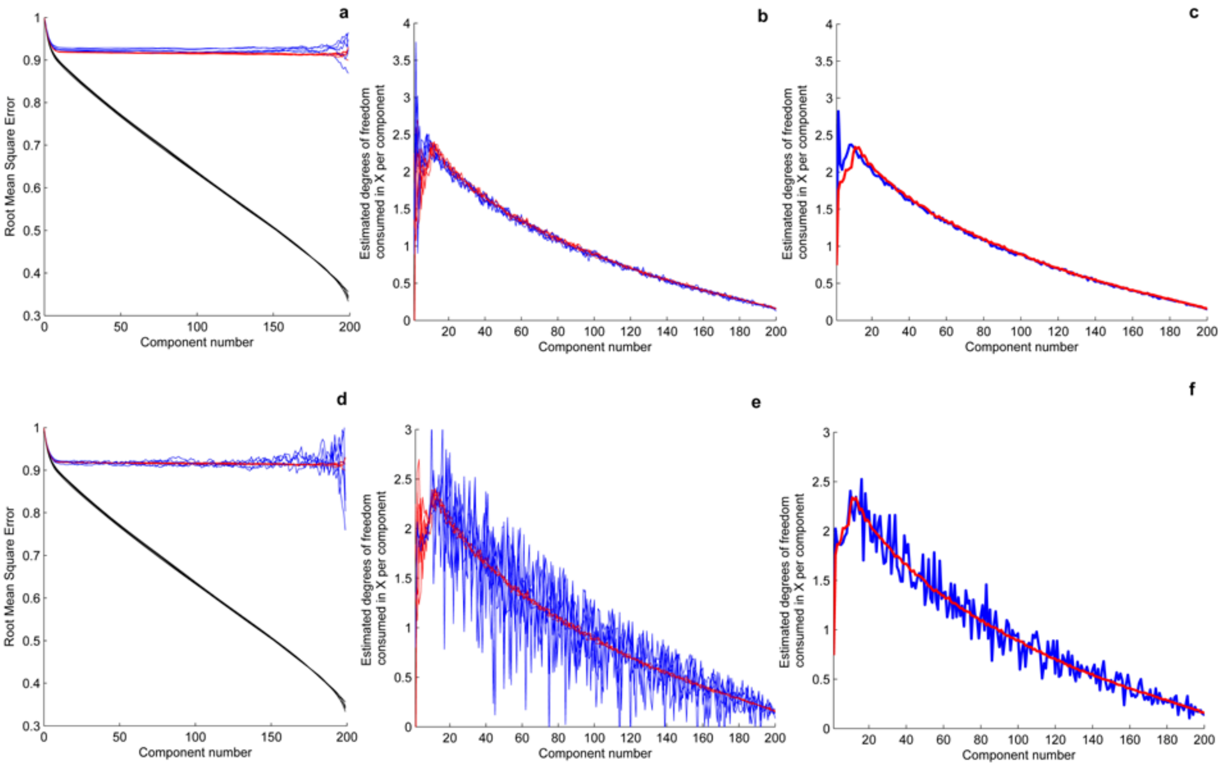


Figure 7: Simulated training and test sets with 500 and 2000 samples respectively containing 200 variables. (a) RMSE_{Cal} (without DF-correction) and RMSE_{CV} (2-fold cross-validation) for training sets are plotted in black and blue respectively. $\text{RMSE}_{\text{Test}}$ plots are plotted in red. (b) Estimated Degrees of Freedom (EDF) calculated by cross-validation are plotted in blue. EDF calculated by test set are plotted in red. (c) The average plot for the plots in (b). (d) RMSE_{Cal} (without DF-correction) and RMSE_{CV} (leave-one-out cross-validation) for training sets are plotted in black and blue respectively. $\text{RMSE}_{\text{Test}}$ plots are plotted in red. (e) EDF calculated by leave-one-out cross-validation are plotted in blue. EDF calculated by test set are plotted in red. (f) The average plot for the plots in (e).

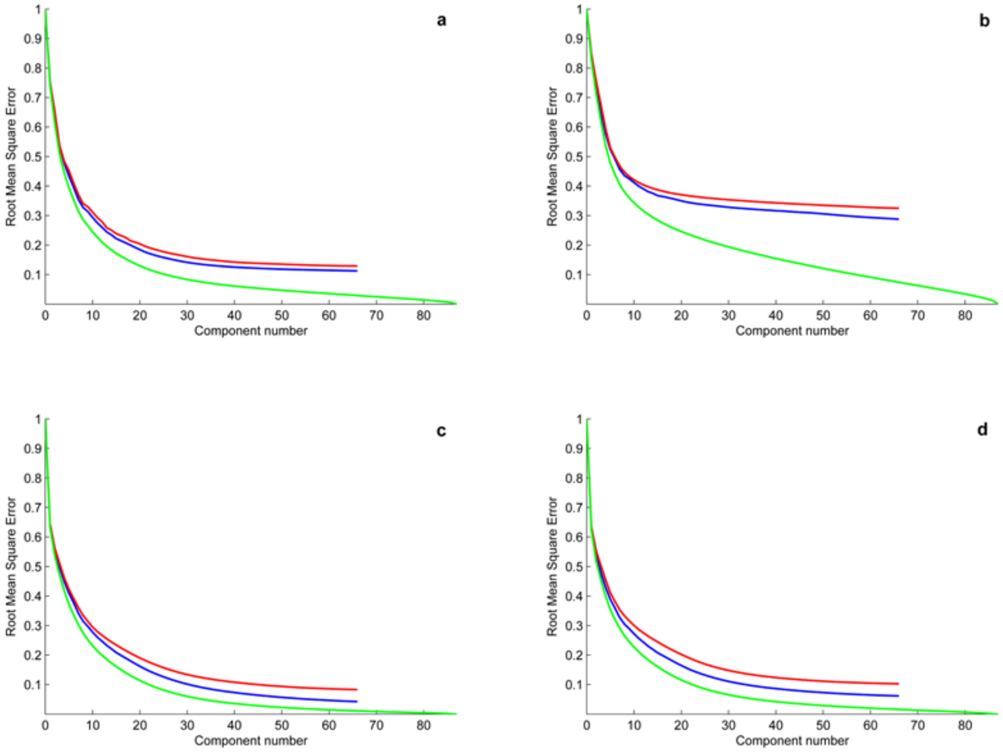


Figure 8: Root Mean Squared Error (RMSE) plots for four real data sets. RMSE plots for real data sets containing FTIR spectra are plotted. $RMSE_{Cal}$ (without DF-correction) are plotted in green. $RMSE_{CV}$ (4-fold cross-validation i.e. 22 samples per segment) are plotted in red. $RMSE_{Cal}$ corrected for the number of consumed DFs are plotted in blue. (a) RMSE plots for the listeria data set. (b) RMSE plots for the meat data set. (c-d) RMSE plots for yeast data sets.

Paper IV

Deflation strategies for multi-block principal component analysis revisited

Authors

Sahar Hassani[1,2], Mohamed Hanafi[3], El Mostafa Qannari[3], and Achim Kohler[2,1]

[1] Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, N 1430 Ås, Norway

[2] Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway

[3] LUNAM University, ONIRIS, Sensometrics and Chemometrics Laboratory, Nantes, F-44307, France; INRA, Nantes, F-44307, France

Corresponding author:
Sahar Hassani, Norwegian Institute of Food, Fisheries and Aquaculture Research,
Osloveien 1, N 1430 Ås, Norway
e-mail: sahar.hassani@nofima.no

Abstract

Within the framework of multi-block data sets, multi-block principal component analysis has been successfully used as a tool to investigate the structure of spectroscopy, -omics and sensory data. The determination of the successive principal components involves a deflation procedure which can be performed according to several strategies. We discuss the respective interest of these strategies and show orthogonality properties related to the vectors of loadings or to the scores. Reconstruction formulas for the data blocks are established for each deflation strategy. Interpretational aspects of the different deflation strategies are discussed and illustrated on the basis of a real and a simulated data set.

1. Introduction

The interest in multi-block methods has gained ground during the recent years particularly in the field of biology, where scientists aim at integrating biological data acquired from the same samples by different instruments [1][2]. Several multi-block methods can be found in the scientific literature together with mathematical and statistical properties which highlight some similarities and differences between these methods. Nonetheless, a biologist may be lost to decide which method should be used in a particular situation and how to interpret the results. This is one of the reasons why further investigations are needed to compare different types of multi-block methods and to point out the extent to which they are similar or different and the implications of these differences, if any, on the interpretation of the outcomes. Different multi-block principal component analysis techniques have been proposed. In all these situations, several (and sometimes very many) variables are measured on the same set of (*N*) samples using different techniques or different assessors. These data could be presented in blocks $\mathbf{X}^b$ with ($b = 1, 2, ..., B$). The total set of measurements can be described by the global ($N \times K$) matrix $\mathbf{X} = \left[\mathbf{X}^1, ..., \mathbf{X}^b, ..., \mathbf{X}^B\right]$, where $K = K_1 + ... + K_b + ... + K_B$ and K_b is the number of columns (variables) of the block $\mathbf{X}^b$.

The aim of multi-block analysis is to investigate the relationships between data blocks $\mathbf{X}^b$ which are supposed to have similar underlying patterns. In order to analyze multi-block data sets, Multi-block Principal Component Analysis (MBPCA) has been proposed. MBPCA is a Multi-block components model, where components or latent variables are constructed and used to summarize simultaneously the relevant information between and within the blocks. The computation of the components in MBPCA involves a sequential process in the course of which a principal component is computed at each step. Thereafter, the variation explained by this component is removed from the data sets (deflation) and the subsequent component is calculated from the residuals. More precisely, MBPCA consists in iterating the following two steps: (i) The computation of a global score vector, block loading vectors and block score vectors of the matrix $\mathbf{X} = \left[\mathbf{X}^1, ..., \mathbf{X}^b, ..., \mathbf{X}^B\right]$. (ii) A deflation step consists in replacing $\mathbf{X} = \left[\mathbf{X}^1, ..., \mathbf{X}^b, ..., \mathbf{X}^B\right]$ by a matrix of residuals.

In the first step, the computation of the global score vector, block score and block loading vectors is performed using an iterative procedure. This procedure can be considered as an extension of the NIPALS procedure [3][4] to more than one block. It has been subsequently shown that the iterative procedure used in MBPCA gives the same global score vector (principal component) as the NIPALS procedure applied to the concatenate matrix $\mathbf{X} = \left[\frac{\mathbf{X}^1}{\sqrt{K_1}} \dots \frac{\mathbf{X}^b}{\sqrt{K_b}} \dots \frac{\mathbf{X}^B}{\sqrt{K_B}} \right]$ [5]. Although this result shows that the global scores of MBPCA can be calculated from $\mathbf{X}$ by the NIPALS algorithm, the rationale behind the iterative procedure of MBPCA is important since it exhibits, on the one hand, block scores and block loadings, which allow us to investigate sample and variable variation patterns within the blocks (block analysis), and, on the other hand, global scores and global loadings which highlight the global variation patterns (global analysis) [1].

Three different deflation strategies (step ii) have been introduced for MBPCA: 1) Deflation by global scores where the variation explained by the global score vector is subtracted by regressing all the variables at hand onto the global score vector. This variant of MBPCA is called Consensus PCA (CPCA) [6]. 2) Deflation by block scores (introduced by Chen and McAvoy) where the deflation is performed on every block matrix $\mathbf{X}^b$ using block scores [7]. 3) A third alternative deflation strategy which consists in deflating with respect to block loadings is very popular in the French literature [8][9]. This latter variant of MBPCA has been called Multiple Co-inertia Analysis (MCoA). It is clear that the deflation is a crucial step in MBPCA: The obtained results in terms of global/block loading and score vectors are different and the outcome studied by the scientist in terms of graphical representation (score plots or different loadings plots) may completely depend on which deflation procedure is chosen.

The three different deflation strategies have not been studied and compared so far and it remains unclear to which extend they are similar or different. To decide upon which deflation strategy to be used, orthogonality, reconstruction properties and interpretational aspects need to be clarified for the three methods.

The objective of this paper is to discuss different aspects of the three mentioned deflation possibilities for MBPCA. (1) The paper gives a complete presentation of orthogonality properties for (block) scores and (block) loadings, which are different for the three deflation possibilities. (2) The reconstruction of block matrices by the (block) scores and (block) loadings is discussed for the respective methods. (3) The differences of the three deflation possibilities with respect to the interpretation of the results in terms of global and block scores are highlighted and illustrated by examples.

The paper is organised as follows: In section 2, the iterative procedure for MBPCA is presented and the three different deflation possibilities, deflation by global scores, deflation by block loadings and deflation by block scores respectively are described. In section 3 orthogonality properties of global scores and loadings and block scores and loadings are proven. In section 4 the reconstruction of the data blocks $\mathbf{X}^b$ for the three different deflation strategies are presented. In section 5 the interpretational aspects of different deflation strategies are discussed. In section 6 some of the theoretical findings are illustrated by an example. In section 7, we will give a conclusion.

2. Multi-block Principal Component Analysis and its alternative deflations strategies.

The NIPALS algorithm for MBPCA

The iterative procedure of MBPCA is given below, where the notation according to Westerhuis et al. (1998) [5] is used. Throughout the paper we assume that the block matrices $\mathbf{X}^b$ are divided by $\sqrt{K_b}$, in order to have equivalence with the procedure described by Westerhuis et al. (1998) [5].

A. Initialization

1.1 Choose an arbitrary starting global score vector $\mathbf{t}$

B. Computation of block scores and block loadings (for $b = 1, 2, ..., B$)

1.2 $\tilde{\mathbf{p}}^b = \dfrac{\mathbf{X}^{b\prime}\mathbf{t}}{\mathbf{t}'\mathbf{t}}$ Preliminary block loadings

1.3 $\mathbf{p}^b = \dfrac{\tilde{\mathbf{p}}^b}{\left\|\tilde{\mathbf{p}}^b\right\|}$ Block loadings, scaled to length 1 in every block

1.4 $\mathbf{t}^b = \mathbf{X}^b\mathbf{p}^b$ Block scores

C. Computation of global scores and global loadings

1.5 $\mathbf{T} = \left[\mathbf{t}^1\ \mathbf{t}^2\ \ldots\ \mathbf{t}^B\right]$

1.6 $\mathbf{w} = \dfrac{\mathbf{T}'\mathbf{t}}{\mathbf{t}'\mathbf{t}}$ Block weights

1.7 Normalize $\mathbf{w}$ to $\|\mathbf{w}\| = 1$

1.8 $\mathbf{t} = \mathbf{T}\mathbf{w}$ Global scores

This procedure is graphically illustrated in Fig. 1. To start the iterative procedure, an arbitrary starting global score $\mathbf{t}$ is chosen. All blocks $\mathbf{X}^b$ are regressed on global score $\mathbf{t}$ in order to obtain the block loadings $\tilde{\mathbf{p}}^b$. From the normalized block loadings $\mathbf{p}^b$, the block scores $\mathbf{t}^b = \mathbf{X}^b\mathbf{p}^b$ for all blocks are calculated. All block scores are combined to a global score matrix $\mathbf{T}$. The global score matrix $\mathbf{T}$ is then regressed on the global score vector $\mathbf{t}$ resulting in the global weights $\mathbf{w}$ as regression coefficients. The global weights are normalized to length one and a new global score vector $\mathbf{t}$ is then calculated. The algorithm is iterated until convergence. Convergence is guaranteed as shown in [10].

Once the parameters of the first component (i.e. $\mathbf{t}_1$, $\mathbf{w}_1$, $\mathbf{t}_1^b$ and $\mathbf{p}_1^b$) are calculated, the matrix $\mathbf{X}_1 = \mathbf{X} = \left[\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^B\right]$ is deflated to $\mathbf{X}_2 = \left[\mathbf{X}_2^1, \mathbf{X}_2^2, \ldots, \mathbf{X}_2^B\right]$. The deflation can be performed according to three different strategies given in table 1. The second components are calculated by applying the iterative procedure to $\mathbf{X}_2$ and the same procedure is repeated for the calculation of the subsequent components. The parameters of the *a*th components (i.e. $\mathbf{t}_a$, $\mathbf{w}_a$, $\mathbf{t}_a^b$ and $\mathbf{p}_a^b$) are calculated by applying the iterative procedure to $\mathbf{X}_a = \left[\mathbf{X}_a^1, \mathbf{X}_a^2, \ldots, \mathbf{X}_a^B\right]$ and $\mathbf{X}_{a+1} = \left[\mathbf{X}_{a+1}^1, \mathbf{X}_{a+1}^2, \ldots, \mathbf{X}_{a+1}^B\right]$ is then calculated by deflating $\mathbf{X}_a$ according to one of the three different strategies given in table 1.

Deflation strategies in MBPCA

The residuals $\mathbf{X}_{a+1}^b$ in table 1 correspond to different regression models that are derived in the following. It is important to grasp the rationale behind the various strategies in order to gain insight into the differences between the three deflation strategies and assert their implications in terms of the interpretation of the outcomes.

For the deflation by global scores (employed by CPCA), the best rank one approximation (prediction) of $\mathbf{X}_a^b$ by the global scores $\mathbf{t}_a$ is achieved, as a solution of the following minimization problem:

$$\left\| \mathbf{X}_a^b - \mathbf{t}_a \tilde{\mathbf{p}}_a^{b\prime} \right\|^2 \quad (1)$$

The solution of (1) is given by the non-normalized block loadings defined by the MBPCA algorithm:

$$\tilde{\mathbf{p}}_a^b = \frac{\mathbf{X}_a^{b\prime} \mathbf{t}_a}{\mathbf{t}_a' \mathbf{t}_a} \quad (2)$$

It is clear that $\mathbf{X}_{a+1}^b = \mathbf{X}_a^b - \frac{\mathbf{t}_a \mathbf{t}_a'}{\mathbf{t}_a' \mathbf{t}_a} \mathbf{X}_a^b$ is the residual part of the regression of block $\mathbf{X}_a^b$ on global score $\mathbf{t}_a$. It is important to note that the deflation on global scores can either be performed block-wise or on the global matrix $\mathbf{X}$ according to:

$$\mathbf{X}_{a+1} = \mathbf{X}_a - \mathbf{t}_a \tilde{\mathbf{p}}_a' = \mathbf{X}_a - \frac{\mathbf{t}_a \mathbf{t}_a'}{\mathbf{t}_a' \mathbf{t}_a} \mathbf{X}_a \qquad (3)$$

Where $\tilde{\mathbf{p}}_a$ is the vector obtained by concatenating the non-normalized block loadings $\tilde{\mathbf{p}}_a^b$ calculated by Eq. 2.

For the deflation by block loadings the best rank one approximation of $\mathbf{X}_a^b$ by the block loadings $\mathbf{p}_a^b$ is achieved, as a solution of the following minimization problem:

$$\left\| \mathbf{X}_a^b - \mathbf{t} \mathbf{p}_a^{b\prime} \right\|^2 \quad (4)$$

Since $\mathbf{p}_a^b$ is normalised to length one, the solution of the minimization of (4) is given by the block scores that are obtained by the iterative procedure for MBPCA: $\mathbf{t}_a^b = \mathbf{X}_a^b \mathbf{p}_a^b$. For the deflation by block loadings, the deflation is based on the residual part of (4), which is given by:

$$\mathbf{X}_{a+1}^b = \mathbf{X}_a^b - \mathbf{t}_a^b \mathbf{p}_a^{b\prime} = \mathbf{X}_a^b - \mathbf{X}_a \mathbf{p}_a^b \mathbf{p}_a^{b\prime} \qquad (5)$$

For the deflation by block scores the best rank one approximation (prediction) of $\mathbf{X}_a^b$ by $\mathbf{t}_a^b$ is achieved, as a solution of the minimization problem:

$$\left\|\mathbf{X}_a^b - \mathbf{t}_a^b\mathbf{q}'\right\|^2 \quad (6)$$

The solution is given by

$$\mathbf{q}_a^b = \frac{\mathbf{X}_a^{b\prime}\mathbf{t}_a^b}{\mathbf{t}_a^{b\prime}\mathbf{t}_a^b} \quad (7)$$

The residual part of (6) is given by:

$$\mathbf{X}_{a+1}^b = \mathbf{X}_a^b - \frac{\mathbf{t}_a^b\mathbf{t}_a^{b\prime}}{\mathbf{t}_a^{b\prime}\mathbf{t}_a^b}\mathbf{X}_a^b \quad (8)$$

We see that the deflation by block scores is based on a new underlying block loadings $\mathbf{q}_a^b$. In order to obtain the block loadings $\mathbf{p}^b$, the block $\mathbf{X}^b$ is projected column-wise on the normalized global score vector $\mathbf{t}$. Then, for the calculation of the block scores $\mathbf{t}^b$, the blocks $\mathbf{X}^b$ are projected on the normalized block loadings $\mathbf{p}^b$. Finally, to find the block loadings $\mathbf{q}^b$, the blocks $\mathbf{X}^b$ are projected on the normalized block scores $\mathbf{t}^b$.

The deflation by global scores is similar to the deflation of the **X** matrix in Partial Least Squares Regression (PLSR) [11] as illustrated in Fig. 2. The deflation in PLSR is usually performed with respect to the same score vector $\mathbf{t}$ for **X** and **Y**, i.e. both matrices are deflated with respect to the **X** scores. According to the NIPALS algorithm for PLSR, the score vector **t** is obtained by the projection of the matrix **X** onto the loading weights **w**. The Y-loadings are obtained by the projection of the matrix **Y** onto the global score vector **t** and the X-loadings by the projection of the matrix **X** onto the score vector **t.** The deflation of **X** and **Y** is then performed according to

$$\mathbf{X}_{a+1} = \mathbf{X}_a - \mathbf{t}\mathbf{p}' \quad (9)$$

$$\mathbf{Y}_{a+1} = \mathbf{Y}_a - \mathbf{t}\mathbf{q}' \quad (10)$$

Deflating on block scores attributes therefore the MBPCA block loading vector $\mathbf{p}_a^b$ the role of PLSR loading weights (**w**), and the MBPCA block loading vector $\mathbf{q}_a^b$ the role of PLSR loadings ($\mathbf{p}$).

3. Orthogonality properties

For the different deflation strategies different orthogonality properties follow. These properties are discussed in the following and summarised in table 2.

The orthogonality of the global score vectors $\mathbf{t}_a$, on the one hand, and the global loadings $\mathbf{p}_a$, on the other hand for the deflation by global scores stem from the fact that CPCA amounts to a PCA performed on matrix **X** [5, 12]. Block loading vectors and block score vectors are, in general, not orthogonal.

For MBPCA with deflation by block loadings, the orthogonality of block loadings (within each block) follows, i.e. $\mathbf{p}_a^b$ is orthogonal to $\mathbf{p}_k^b$ for $a \neq k$. The global scores $\mathbf{t}_a$ are also orthogonal [8]. The orthogonality of block loadings and the global scores in MBPCA with deflation by block loadings is shown in appendix A. The corresponding block scores $\mathbf{t}_a^b$ are, in general, not orthogonal, but it can be proven that block scores $\mathbf{t}_k^b$ are orthogonal to global scores $\mathbf{t}_a$ for $k > a$ (see appendix A, property A3). From the orthogonality of block loadings, the orthogonality of global loadings $\mathbf{p}_a$ follows immediately, since $\mathbf{p}_a$ can be written as a linear combination of the block loadings $\mathbf{p}_a^b$ as follows:

$$\mathbf{p}_a = \left\|\mathbf{X}_a^{1\prime}\mathbf{t}_a\right\| \begin{bmatrix} \mathbf{p}_a^1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \left\|\mathbf{X}_a^{2\prime}\mathbf{t}_a\right\| \begin{bmatrix} 0 \\ \mathbf{p}_a^2 \\ \vdots \\ 0 \end{bmatrix} + \ldots + \left\|\mathbf{X}_a^{B\prime}\mathbf{t}_a\right\| \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \mathbf{p}_a^B \end{bmatrix} \quad (11)$$

For MBPCA with deflation by block scores, we can show that the block loadings $\mathbf{p}_a^b$ within each block are orthogonal. This property has been mentioned without proof by [12]. We present a proof in appendix B. The global scores $\mathbf{t}_a$ are in general not orthogonal for the deflation by block scores, while block scores referring to the same block are orthogonal. This has been shown by [12]. For the sake of completeness this is also proven in appendix B. From the orthogonality of block loadings, the orthogonality of global loadings $\mathbf{p}_a$ follows immediately from Eq. 11.

Because of the orthogonality of the block loadings in MBPCA with deflation by block loadings, the block scores can be considered as a weighted sum of the original variables $\mathbf{X}^b$:

$$\mathbf{t}_a^b = \mathbf{X}_a^b \mathbf{p}_a^b = \mathbf{X}^b \mathbf{p}_a^b \qquad (12)$$

However, this is, in general, not the case for deflation by block scores:

$$\mathbf{t}_a^b = \mathbf{X}_a^b \mathbf{p}_a^b = \mathbf{X}^b \tilde{\mathbf{q}}_a^b \text{ with } \tilde{\mathbf{q}}_a^b \neq \mathbf{p}_a^b \qquad (13)$$

In addition, we can show that the hereby defined $\tilde{\mathbf{q}}_a^b$ are not orthogonal.

4. Reconstruction of block matrices by block scores and block loadings

The reconstruction of the block matrices $\mathbf{X}^b$ depends on the deflation method being chosen. The three possibilities for reconstruction are shown in table 3, where r_b denotes the rank of block $\mathbf{X}^b$ and r the rank of $\mathbf{X} = \left[\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^B\right]$.

For the deflation by global scores ($\mathbf{t}_a$) the underlying block loadings used for reconstruction are the non-normalised block loadings defined by the NIPALS algorithm. For the deflation by block loadings the block scores and block loadings defined by the NIPALS algorithm are used for reconstruction. For the deflation by block scores new underlying block loadings $\mathbf{q}_a^b$ are used for reconstruction.

5. Interpretational aspects

When studying score and loading plots we are visually inspecting how samples and variables are clustered and related to each others, i.e. we study scores and loadings visually to discover *sample variation patterns* and *variable variation patterns*, respectively. To be able to interpret these sample and variable variation patterns, it is important to be aware of the relations between global and block scores and global and block loadings. As stated above, how the global and block parameters are related to each others strongly depends on the deflation procedure. Indeed, in each deflation step, depending on which deflation procedure is performed, we either subtract a sample or variable variation pattern. For the deflation by global scores, it follows from (1) that the pattern subtracted at every deflation step is given by the global scores $\mathbf{t}_a$. According to (1), these scores are also subtracted from every block. For the deflation by block loadings, the block variable variation pattern is subtracted from each block during deflation. This is obvious when considering the minimization problem in equation (4) which corresponds to the deflation by block loadings. We see that the variation in variable space presented by the block loadings $\mathbf{p}^b$ is subtracted completely from every block. This will lead to the particular situation where the variable variation pattern in the second component for the same block will be independent from the first block and so on. This is in general not the case for the deflation on global scores. It is difficult to state which variation pattern is subtracted for the deflation by block scores. The additional step, that is necessary in order to define the underlying block loadings $\mathbf{q}_a^b$ for the deflation with respect to block scores can also be seen as an additional NIPALS step per block which bears a resemblance to setting up a PCA model per block matrix and, therefore, is not really aiming at finding common underlying patterns. Through this additional NIPALS step which is performed for each block, the block sample variation pattern subtracted in every deflation step is not any longer related to the global variation pattern in a direct way and makes the relation between deflation and score patterns unclear. This has important implications for the interpretation of the multi-block results as explained in the following:

(a) Deflation by global scores: The block matrices are reconstructed by the orthogonal global scores $\mathbf{t}_a$. The global sample variation pattern can be visualised by score plots based on global scores. Notwithstanding, care should be taken in the interpretation of score plots of block scores. The block scores $\mathbf{t}_a^b$ are defined on the basis of normalised block loadings $\mathbf{p}_a^b$ (the $\mathbf{t}_a^b$ are defined by means of $\mathbf{p}_a^b$) and represent the block sample variation pattern with respect to $\mathbf{p}_a^b$. Therefore, they reflect the block sample variation pattern that corresponds to the variable variation pattern $\mathbf{p}_a^b$, which, in general, is not the same as the global variation pattern $\mathbf{t}_a$. In other words, the global sample variation pattern $\mathbf{t}_a$ is related to a certain variable variation pattern $\mathbf{p}_a^b$ in each block. This block variable variation pattern may in general also be related to other sample variation patterns

in the same block, i.e. the same block variable variation pattern may be related to different block sample variation patterns and therefore be present in several block loadings. This will be illustrated by an example.

(b) Deflation by block loadings: For the deflation by block loadings, the block variation pattern $\mathbf{t}_a^b$ which is related to the variable variation $\mathbf{p}_a^b$ is subtracted at each deflation step. Consequently, score plots of block scores $\mathbf{t}_a^b$ represent the block variation patterns subtracted in each deflation step. Therefore, in the deflation by block loadings, the corresponding variable variation pattern $\mathbf{p}_a^b$ is completely subtracted from each block. For the interpretation of the global variation pattern as expressed by score plots of $\mathbf{t}_a$ we have to keep in mind, that the global variation pattern is not the variation pattern that is subtracted in every deflation step, but it represents the calculated 'consensus'. For the deflation by block loadings the property that block loadings are orthogonal follows, which guarantees that the 'information' contained in the block components is independent from one component to another. This may be helpful in situations where interpretative signals are considered. An example will be discussed in the next section.

(c) Deflation by block scores: For the deflation by block scores new underlying block loadings $\mathbf{q}_a^b$ are defined and used for reconstruction of the original data. Among the three deflation methods considered, it is therefore the only strategy where the loading weights and loadings are different. Since the orthogonal block scores used for the reconstruction of the data are not directly related to the global scores, they are difficult to interpret in terms of the global variation pattern $\mathbf{t}_a$: The variation that is subtracted at each deflation step for each block is not directly related to the global variation pattern. Therefore, the authors do not recommend using this type of deflation, since common and block variation patterns can not easily be interpreted.

From the reconstruction formula given in table 3, the computation of the explained variances for each block $\mathbf{X}^b$ and at step *a* follows and is given in table 4 for the three different deflation methods. The proof is given in the appendix C.

6. Illustration

The multi-block data set used for illustrating the different strategies of deflation is described in detail in the references [1, 13] and consists of 5 data blocks with different number of variables in each block measured on the same set of 88 samples. This multi-block data set contains amplified fragment length polymorphism (AFLP) data (genetic fingerprinting), Fourier Transform Infrared (FTIR) spectra, and other phenotypes (sero grouping, susceptibility to sakacin P, nisin and the antibacterial agent benzalkonium chloride) of 88 *L. monocytogenes* strains. As in [1] the AFLP data block defines block $\mathbf{X}^1$ (1701 variables), the FTIR data block is subdivided into the following spectra regions: polysaccharide region and fingerprint region (1200-720 cm^{-1}) defining block $\mathbf{X}^2$ (498 variables), the protein region (1700-1500cm^{-1}) defining block $\mathbf{X}^3$ (209 variables) and the fatty acid region (3000-2800cm^{-1}) defining block $\mathbf{X}^4$ (208 variables). The phenotypes are collected in data block $\mathbf{X}^5$ (10 variables). Previous to MBPCA the spectral data was pre-processed by EMSC [14, 15].

The block score plots and the global score plots for the deflation by global scores, block loadings and block scores are shown in Figs. 3-5, respectively ($\mathbf{X}^1$ is shown in (a), $\mathbf{X}^2$ in (b), …, $\mathbf{X}^5$ in (e) and the global scores in (f)). The outcomes of PCA performed on each block ($\mathbf{X}^1$ to $\mathbf{X}^5$) separately is shown in Fig. 6a-e. The explained variances are calculated according to table 4 and are indicated on the axes for each score plot (block scores, global scores and score plots of PCAs) in the Figs 3-6.

Obviously, the global/block scores and loadings are always identical for the first components of all three deflation strategies, since no deflation has been applied yet. Therefore, the score values for the first components are identical for the same blocks in Fig. 3-5. The score values for all subsequent components are in general different. The percentages of explained variances calculated according to table 4 are different even for the first component, since they depend on which deflation procedure is used. We observe that the percentage of explained variance by the first component of each block increases from Fig. 3 to Fig. 6: For example for block $\mathbf{X}^1$ the explained variance of the first component is 27% for the deflation on global scores, 36.2% for the deflation on block loadings, 36.8% for the deflation on block scores and 36.9% for PCA of $\mathbf{X}^1$. The increase of explained variance of the first components from Fig. 3 to Fig. 6 can be explained by the fact that the calculation of block loadings defines the first NIPALS step of PCA for every single block $\mathbf{X}^b$, while the definition of block scores defines the second NIPALS step. We see that the variances explained by the first component greatly change from Fig. 3 to Fig. 6 for the different blocks. For example, the result for the deflation by block loadings for the first block $\mathbf{X}^1$ (Fig. 4a) shows that the explained variance has almost approached the result of PCA of $\mathbf{X}^1$ (Fig. 6a). This is due to the fact that the main variation in block $\mathbf{X}^1$ is very similar to the main consensus variation. For block $\mathbf{X}^4$ the behaviour is completely different. The explained variance increases from 17.5% for the deflation on global scores to 31.9% for the deflation on block loadings, 44.1% for the deflation on block scores and reaches 62.3% for PCA of $\mathbf{X}^4$. This indicates that the main variation in block $\mathbf{X}^4$ is very different from the main consensus variation. Therefore, the explained variances for $\mathbf{X}^4$ considerably change from Fig. 3 to Fig. 6.

In order to further elucidate the differences between the different deflation procedures we simulated an artificial multi-block dataset using the blocks $\mathbf{X}^1,...,\mathbf{X}^5$ from the above example. The simulation procedure for generating the artificial data blocks is described in the following: First, a PCA of each block was performed in order to obtain scores and loadings for each block separately (i.e. $\mathbf{X}^b = \mathbf{T}^b\mathbf{P}^{b\prime} + \mathbf{E}^b$ where $\mathbf{T}^b$ and $\mathbf{P}^b$ are respectively the matrices of scores and loadings of PCA for block *b* using A_b principal components). Then, a multi-block matrix (containing 5 data blocks) was simulated according to: $\mathbf{X}_{\mathrm{Sim}} = \left[\mathbf{X}^1_{\mathrm{Sim}},...,\mathbf{X}^5_{\mathrm{Sim}}\right] = \left[\mathbf{T}^2\mathbf{P}^{1\prime},\mathbf{T}^2\mathbf{P}^{2\prime},\mathbf{T}^2\mathbf{P}^{3\prime},\mathbf{T}^2\mathbf{P}^{4\prime},\mathbf{T}^2\mathbf{P}^{5\prime}\right]$ using $A_b = 3$ components for each block (i.e. $\mathbf{T}^b = \mathbf{T}^2 = \left[\mathbf{t}^2_1\ \mathbf{t}^2_2\ \mathbf{t}^2_3\right]$ and $\mathbf{P}^b = \left[\mathbf{p}^b_1\ \mathbf{p}^b_2\ \mathbf{p}^b_3\right]$ for $b = 1,...,5$). Performing MBPCA on $\mathbf{X}_{\mathrm{Sim}}$, results in completely identical block patterns for block scores and global scores, regardless of the deflation method used (Fig. 7). The pattern that is seen in the block score plots as well as in the global score plot in Fig. 7 is the same pattern seen in Fig. 6b. This is obvious since the sample variation pattern belonging to block $\mathbf{X}^2$ (given by $\mathbf{T}^2 = \left[\mathbf{t}^2_1\ \mathbf{t}^2_2\ \mathbf{t}^2_3\right]$) was introduced in all the blocks.

A second multi-block data set was simulated by modifying the second block of the previously simulated multi-block data set ($\mathbf{X}^2_{\text{Sim}}$). The modification was as follows: the scores and loadings used for the simulation of $\mathbf{X}^2_{\text{Sim}}$ (i.e. $\mathbf{T}^2 = \left[\mathbf{t}^2_1\ \mathbf{t}^2_2\ \mathbf{t}^2_3\right]$ and $\mathbf{P}^2 = [\mathbf{p}^2_1\ \mathbf{p}^2_2\ \mathbf{p}^2_3]$) were replaced by $\mathbf{T}^2_{\text{Sim2}} = [\mathbf{t}^2_1\ \mathbf{t}^2_2\ \mathbf{t}^2_3\ \mathbf{t}^2_4]$ and $\mathbf{P}^2_{\text{Sim2}} = [\mathbf{p}^2_1\ \mathbf{p}^2_1\ \mathbf{p}^2_2\ \mathbf{p}^2_3]$ respectively, resulting in a second multi-block simulated data set: $\mathbf{X}_{\text{Sim2}} = \left[\mathbf{X}^1_{\text{Sim2}},\ldots,\mathbf{X}^5_{\text{Sim2}}\right] = \left[\mathbf{T}^2\mathbf{P}^{1\prime},\mathbf{T}^2_{\text{Sim2}}\mathbf{P}^2_{\text{Sim2}}{}^{\prime},\mathbf{T}^2\mathbf{P}^{3\prime},\mathbf{T}^2\mathbf{P}^{4\prime},\mathbf{T}^2\mathbf{P}^{5\prime}\right]$ where $\mathbf{X}^b_{\text{Sim2}} = \mathbf{X}^b_{\text{Sim}}$ for $b = 1,3,4,5$. This means $\mathbf{X}^2_{\text{Sim2}}$ is simulated such that two independent sample variation patterns (i.e. $\mathbf{t}^2_1$ and $\mathbf{t}^2_2$) correspond to the same variable variation pattern ($\mathbf{p}^2_1$). Performing CPCA on $\mathbf{X}_{\text{Sim2}}$ (i.e. MBPCA with deflation by global scores) resulted in block and global score plots shown in Fig. 8. We can see that the score patterns are almost identical in all the plots, although the block score pattern in the first and second component in block $b = 2$ corresponds to the same variable pattern. The corresponding block loadings for block two are shown in Fig. 9. We can see that the block loadings for the first and second components are identical. Performing MBPCA with deflation by block loadings resulted in block and global score plots shown in Fig. 10. We can see that the block score pattern of block two has completely changed. This is due to the fact that by deflating with respect to block loadings, the next block loading is constrained to be orthogonal to the previous ones. Therefore the sample variation related $\mathbf{p}^2_1$ (i.e. $\mathbf{t}^2_1$ and $\mathbf{t}^2_2$) is captured at once by calculating $\mathbf{t}^b_1 = \mathbf{X}^b_1\mathbf{p}^b_1$ and already subtracted in the first deflation step. The corresponding first and second block loadings are shown in Fig. 11. The first block loading in Fig. 11 is identical to the block loadings in Fig. 9. However, the second block loading in Fig. 11 is different.

The situation in this simulated data set is likely to occur in real data sets where for example two different design factors lead to the same or similar variable variation patterns in one data block, while the two design factors may be independent or may have a different effect on the variable variation pattern in a different block.

7. Conclusion

Three main deflation strategies for MBPCA have been studied, namely the deflation by global scores, the deflation by block loadings and the deflation by block scores. It has been shown that the three deflation possibilities lead to completely different orthogonality properties for global/block scores and loadings, to different reconstruction procedures of the block matrices and to differences in the calculation of the explained variances. These differences were illustrated using a real data set and simulated data sets. It turned out that the different deflation strategies yield very different graphical displays. We have shown that the deflation methods are also different from an interpretational point of view. Indeed, whereas, for the deflation by global scores, the global variation pattern is subtracted from every block, for the deflation by block loadings the block variable variation pattern is subtracted in every deflation step. Since both deflation methods (i.e. deflation on global scores and deflation on block loadings) clearly have different advantages, they may be performed successively in order to gain more insight into the data at hand. It was also pointed out that for the deflation by block scores new underlying block loadings are defined as in PLSR. Therefore, the global scores and loadings are not directly related to the block loadings used for deflation by block scores. As a result, it may be more difficult to interpret block results in connection with global results. We have illustrated the fact that the definition of the new underling block loadings can be considered as a further NIPALS step towards a PCA of each block $\mathbf{X}^b$.

The choice of the deflation procedure is a generic problem in multi-block analysis and the results obtained in this paper can easily be transferred to, e.g., multi-block partial least squares regression (MBPLS) [16].

Acknowledgements

The authors are grateful for financial support by Foundation for Research Levy on Agricultural Products in Norway, the grant 203699 (New statistical tools for integrating and exploiting complex genomic and phenotypic data sets) financed by the Research Council of Norway and for financial support by the Nordic Centre of Excellence on Food, Nutrition and Health “Systems biology in controlled dietary interventions and cohort studies” (SYSDIET) funded by NordForsk.

References

[1] A. Kohler, M. Hanafi, D. Bertrand, A. Oust Janbu, T. Møretrø, K. Naderstad, M. Qannari, H. Martens, Interpreting several types of measurements in bioscience, in: P. Lasch, J. Kneipp (Eds.), Modern concepts in biomedical vibrational spectroscopy, John Wiley & Sons, USA, 2008.

[2] S. Hassani, H. Martens, E.M. Qannari, M. Hanafi, G.I. Borge, A. Kohler, Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods, Chemometrics and Intelligent Laboratory Systems 104 (2010) 140-153.

[3] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, 1966, pp. 391-420.

[4] Y. Miyashita, T. Itozawa, H. Katsumi, S.-I. Sasaki, Comments on the NIPALS algorithm, Journal of Chemometrics 4 (1990) 97-100.

[5] J.A. Westerhuis, T. Kourti, J.F. Macgregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemometrics 12 (1998) 301-321.

[6] S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom, W. H., PLS Model Building Theory and Application, Frankfurt am Main, Frankfurt, Germany, 1987.

[7] G. Chen, T.J. McAvoy, Multi-block predictive monitoring of continuous processes, IFAC ADCHEM'97, Banff, 1997.

[8] D. Chessel, M. Hanafi, Analyses de la co-inertie de *K* nuages de points, Rev. Statistique Appliquée 44 (1996) 35-60.

[9] M. Hanafi, G. Mazerolles, E. Dufour, E.M. Qannari, Common components and specific weight analysis and multiple Co-inertia analysis applied to the coupling of several measurement techniques, Journal of Chemometrics 20 (2006) 172-183.

[10] M. Hanafi, A. Kohler, E.-M. Qannari, Connections between multiple co-inertia analysis and consensus principal component analysis, Chemometrics and Intelligent Laboratory Systems 106 (2011) 37-40.

[11] H. Martens, T. Næs, Multivariate calibration, Wiley & Sons, Chichester, 1989.

[12] S.J. Qin, S. Valle, M.J. Piovoso, On unifying multiblock analysis with application to decentralized process monitoring, J. Chemometrics 15 (2001) 715-742.

[13] A. Oust, T. Moretro, K. Naterstad, G.D. Sockalingum, I. Adt, M. Manfait, A. Kohler, Fourier transform infrared and Raman spectroscopy for characterization of Listeria monocytogenes strains, Applied and Environmental Microbiology 72 (2006) 228-232.

[14] A. Kohler, C. Kirschner, A. Oust, H. Martens, Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin, Applied Spectroscopy 59 (2005) 707-716.

[15] A. Kohler, M. Zimonja, V. Segtnan, H. Martens, Data preprocessing: SNV, MSC and EMSC pre-processing in biospectroscopy, in: S. Brown, R. Tauler, R. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 139-162.

[16] J.A. Westerhuis, A.K. Smilde, Deflation in multiblock PLS, J. Chemometrics 15 (2001) 485-493.

[17] S. Wold, S. Hellberg, Y. Lundstedt, M. Sjostrom, H. Wold, Proc. Symp. on PLS Model Building: Theory and Application, Frankfurt am Main., 1987.

Appendix A. Orthogonality properties in MBPCA with deflation by block loadings

$$\mathbf{X}_{a+1}^{b} = \mathbf{X}_{a}^{b} - \mathbf{t}_{a}^{b}\mathbf{p}_{a}^{b\prime} = \mathbf{X}_{a}^{b} - \mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{a}^{b\prime}\ ,\ b = 1, 2, ..., B$$

A1. Property 1. (Orthognality of block loadings)

For $1 \le k, a \le A$ and $k \ne a \quad \mathbf{p}_{k}^{b\prime}\mathbf{p}_{a}^{b} = 0$

A2. Property 2. (Orthognality of global scores)

For $1 \le k, a \le A$ and $k \ne a \quad \mathbf{t}_{k}^{\prime}\mathbf{t}_{a} = 0$

A3. Property 3. (Orthognality between global scores and block scores)

For $1 \le k < a \le A \quad \mathbf{t}_{a}^{b\prime}\mathbf{t}_{k} = 0$

Preliminary lemma.

for $a = 1, 2, ..., A-1$. $\mathbf{X}_{A}^{b}\mathbf{X}_{A}^{b\prime} = \mathbf{X}_{a}^{b}\mathbf{X}_{A}^{b\prime}$ (a)

for $a > k$, $\mathbf{X}_{a}^{b} = \mathbf{X}_{k}^{b}\left(\prod_{l=k}^{a-1}\left(\mathbf{I}_{m_b} - \mathbf{p}_{l}^{b}\mathbf{p}_{l}^{b\prime}\right)\right)$ (b)

Proof of property 1.

The proof of property 1 will be given by recurrence. First, we prove the property 1 for $A = 2$, i.e. $\left(\mathbf{p}_{1}^{b\prime}\mathbf{p}_{2}^{b}\right) = 0$. Note that the $\mathbf{p}_{a}^{b}$ for $a = 1, 2, ...$; are given by

$$\mathbf{p}_{1}^{b} = \frac{\mathbf{X}_{1}^{b\prime}\mathbf{t}_{1}}{\left\|\mathbf{X}_{1}^{b\prime}\mathbf{t}_{1}\right\|} \text{ and } \mathbf{p}_{2}^{b} = \frac{\mathbf{X}_{2}^{b\prime}\mathbf{t}_{2}}{\left\|\mathbf{X}_{2}^{b\prime}\mathbf{t}_{2}\right\|}\ ,\ b = 1, 2, ..., B \quad \text{(A1.1)}$$

where in (A1.1) $\mathbf{X}_{1}^{b} = \mathbf{X}^{b}$ and $\mathbf{X}_{2}^{b} = \mathbf{X}_{1}^{b} - \mathbf{X}_{1}^{b}\mathbf{p}_{1}^{b}\mathbf{p}_{1}^{b\prime}$.

Using $\mathbf{X}_{2}^{b} = \mathbf{X}_{1}^{b}\left(\mathbf{I}_{m_b} - \mathbf{p}_{1}^{b}\mathbf{p}_{1}^{b\prime}\right)$ then $\mathbf{p}_{2}^{b}$ can be written as:

$$\mathbf{p}_{2}^{b} = \frac{\left(\mathbf{I}_{m_b} - \mathbf{p}_{1}^{b}\mathbf{p}_{1}^{b\prime}\right)\mathbf{X}_{1}^{b\prime}\mathbf{t}_{2}}{\left\|\left(\mathbf{I}_{m_b} - \mathbf{p}_{1}^{b}\mathbf{p}_{1}^{b\prime}\right)\mathbf{X}_{1}^{b\prime}\mathbf{t}_{2}\right\|} \quad \text{(A1.2)}$$

Multiplying (A1.2) by $\left(\mathbf{I}_{m_b} - \mathbf{p}_1^b \mathbf{p}_1^{b\prime}\right)$ and using the equality $\left(\mathbf{I}_{m_b} - \mathbf{p}_1^b \mathbf{p}_1^{b\prime}\right)^2 = \left(\mathbf{I}_{m_b} - \mathbf{p}_1^b \mathbf{p}_1^{b\prime}\right)$ which stems from the fact that $\left(\mathbf{I}_{m_b} - \mathbf{p}_1^b \mathbf{p}_1^{b\prime}\right)$ is a projector the right hand side of (A1.2) will remain unchanged and it therefore follows:

$$\left(\mathbf{I}_{m_b} - \mathbf{p}_1^b \mathbf{p}_1^{b\prime}\right)\mathbf{p}_2^b = \mathbf{p}_2^b \quad \text{(A1.3)}$$

Expanding (A.3) we obtain $\left(\mathbf{p}_1^{b\prime} \mathbf{p}_2^b\right) = 0$ for $\mathbf{p}_1^b \neq \mathbf{0}$.

We suppose now that property 1 is valid for $a = 1, 2, ..., A-1$, i.e. for $1 \leq k, a \leq A-1$ and $k \neq a$ we assume that $\mathbf{p}_k^{b\prime} \mathbf{p}_a^b = 0$. From this it follows that $\mathbf{X}_A^b = \mathbf{X}_{A-1}^b - \mathbf{X}_{A-1}^b \mathbf{p}_{A-1}^b \mathbf{p}_{A-1}^{b}{}'$ can be written as

$$\mathbf{X}_A^b = \mathbf{X}_1^b - \sum_{a=1}^{A-1} \mathbf{X}_1^b \mathbf{p}_a^b \mathbf{p}_a^{b\prime} \quad \text{(A1.4)}$$

or equivalently

$$\mathbf{X}_A^b = \mathbf{X}_1^b \left(\mathbf{I}_{m_b} - \sum_{a=1}^{A-1} \mathbf{p}_a^b \mathbf{p}_a^{b\prime}\right) \quad \text{(A1.5)}$$

We will now show that $\mathbf{p}_a^{b\prime} \mathbf{p}_A^b = 0$ for $a = 1, 2, ..., A-1$, where

$$\mathbf{p}_A^b = \frac{\mathbf{X}_A^{b\prime} \mathbf{t}_A}{\left\| \mathbf{X}_A^{b\prime} \mathbf{t}_A \right\|} \quad \text{(A1.6)}$$

Using (A1.5) we can write $\mathbf{p}_A^b$ as :

$$\mathbf{p}_A^b = \frac{\left(\mathbf{I}_{m_b} - \sum_{a=1}^{A-1} \mathbf{p}_a^b \mathbf{p}_a^{b\prime}\right) \mathbf{X}_1^{b\prime} \mathbf{t}_A}{\left\| \left(\mathbf{I}_{m_b} - \sum_{a=1}^{A-1} \mathbf{p}_a^b \mathbf{p}_a^{b\prime}\right) \mathbf{X}_a^{b\prime} \mathbf{t}_A \right\|} \quad \text{(A1.7)}$$

Multiplying (A1.7) by $\left(\mathbf{I}_{m_b} - \sum_{a=1}^{A-1} \mathbf{p}_a^b \mathbf{p}_a^{b\prime}\right)$ and using the equality

$\left(\mathbf{I}_{m_b} - \sum_{a=1}^{A-1} \mathbf{p}_a^b \mathbf{p}_a^{b\prime}\right)^2 = \left(\mathbf{I}_{m_b} - \sum_{a=1}^{A-1} \mathbf{p}_a^b \mathbf{p}_a^{b\prime}\right)$, the right hand side of (A1.7) will remain unchanged and it therefore follows:

$$\left(\mathbf{I}_{m_b}-\sum_{a=1}^{A-1}\mathbf{p}_a^b\mathbf{p}_a^{b\prime}\right)\mathbf{p}_A^b=\mathbf{p}_A^b \text{ (A1.8)}$$

Expanding (A1.8) we obtain $\sum_{a=1}^{A-1}\left(\mathbf{p}_a^{b\prime}\mathbf{p}_A^b\right)\mathbf{p}_a^b=\mathbf{0}$. Because the vectors $\mathbf{p}_a^b$ are mutually orthogonal for $a=1,2,...,A-1$, it follows that $\left(\mathbf{p}_a^{b\prime}\mathbf{p}_A^b\right)=0$ for $a=1,2,...,A-1$. ☺

Proof of property 2.

As above the proof will be given by recurrence. First, we proof the property 2 for $A=2$. i.e, $\mathbf{t}_2^{b\prime}\mathbf{t}_1=0$. We consider:

$$\mathbf{X}_2^b\mathbf{X}_2^{b\prime}=\mathbf{X}_1^b\left(\mathbf{I}-\mathbf{p}_1^b\mathbf{p}_1^{b\prime}\right)\left(\mathbf{I}-\mathbf{p}_1^b\mathbf{p}_1^{b\prime}\right)\mathbf{X}_1^{b\prime}=\mathbf{X}_1^b\mathbf{X}_2^{b\prime} \text{ (A2.1)}$$

By summing over b we obtain

$$\sum_{b=1}^{B}\mathbf{X}_2^b\mathbf{X}_2^{b\prime}=\sum_{b=1}^{B}\mathbf{X}_1^b\mathbf{X}_2^{b\prime} \text{ (A2.2)}$$

By multiplying (A2.2) by $\mathbf{t}_1'$ and $\mathbf{t}_2$ we find:

$$\mathbf{t}_1'\left(\sum_{b=1}^{B}\mathbf{X}_2^b\mathbf{X}_2^{b\prime}\right)\mathbf{t}_2=\mathbf{t}_1'\left(\sum_{b=1}^{B}\mathbf{X}_1^b\mathbf{X}_2^{b\prime}\right)\mathbf{t}_2 \text{ (A2.3)}$$

As $\mathbf{t}_2$ is the first eigenvector of $\mathbf{X}_2\mathbf{X}_2'=\sum_{b=1}^{B}\mathbf{X}_2^b\mathbf{X}_2^{b\prime}$ associated to the largest eigenvalue denoted by λ_2, the left side of (A2.3) results in

$$\mathbf{t}_1'\left(\sum_{b=1}^{B}\mathbf{X}_2^b\mathbf{X}_2^{b\prime}\right)\mathbf{t}_2=\lambda_2\mathbf{t}_1'\mathbf{t}_2, \quad \text{(A2.4)}$$

By using the definition of $\mathbf{p}_1^b$ and $\mathbf{p}_2^b$ in (A.1), for the right hand side of (A2.3) becomes

$$\mathbf{t}_1'\left(\sum_{b=1}^{B}\mathbf{X}_1^b\mathbf{X}_2^{b\prime}\right)\mathbf{t}_2=\sum_{b=1}^{B}\mathbf{t}_1'\mathbf{X}_1^b\mathbf{X}_2^{b\prime}\mathbf{t}_2=\sum_{b=1}^{B}\left\|\mathbf{t}_1'\mathbf{X}_1^b\right\|\cdot\left\|\mathbf{X}_2^{b\prime}\mathbf{t}_2\right\|\mathbf{p}_2^{b\prime}\mathbf{p}_1^b \quad \text{(A2.5)}$$

Combining (A2.4) and (A2.5) we get

$$\lambda_2\mathbf{t}_1'\mathbf{t}_2=\sum_{b=1}^{B}\left\|\mathbf{t}_1'\mathbf{X}_1^b\right\|\cdot\left\|\mathbf{X}_2^{b\prime}\mathbf{t}_2\right\|\mathbf{p}_2^{b\prime}\mathbf{p}_1^b \quad \text{(A2.6)}$$

Because of property 1 the right hand side of (A2.6) is zero and therefore $\mathbf{t}_1'\mathbf{t}_2=0$.

We suppose that property 2 is valid for $a = 1, 2, ..., A-1$, i.e. for $1 \le k, a \le A-1$ and $k \ne a$ it is $\mathbf{t}'_k \mathbf{t}_a = 0$. We will show now that $\mathbf{t}'_a \mathbf{t}_A = 0$ for $a = 1, 2, ..., A-1$.

By summing lemma 1a over b we obtain

$$\sum_{b=1}^{B} \mathbf{X}_A^b \mathbf{X}_A^{b\prime} = \sum_{b=1}^{B} \mathbf{X}_a^b \mathbf{X}_A^{b\prime} \text{ (A2.7)}$$

Multiplying (A2.7) by $\mathbf{t}'_a$ and $\mathbf{t}_A$ it follows:

$$\mathbf{t}'_a \left(\sum_{b=1}^{B} \mathbf{X}_A^b \mathbf{X}_A^{b\prime} \right) \mathbf{t}_A = \mathbf{t}'_a \left(\sum_{b=1}^{B} \mathbf{X}_a^b \mathbf{X}_A^{b\prime} \right) \mathbf{t}_A \quad \text{(A2.8)}$$

As $\mathbf{t}_A$ is the first eigenvector of $\mathbf{X}_A \mathbf{X}'_A = \sum_{b=1}^{B} \mathbf{X}_A^b \mathbf{X}_A^{b\prime}$ associated to the largest eigenvalue denoted by λ_A, the left side of (A2.8) results in

$$\mathbf{t}'_a \left(\sum_{b=1}^{B} \mathbf{X}_A^b \mathbf{X}_A^{b\prime} \right) \mathbf{t}_A = \lambda_A \mathbf{t}'_a \mathbf{t}_A, \text{ for } a = 1, 2, ..., A-1 \quad \text{(A2.9)}$$

By using the definition of $\mathbf{p}_a^b = \dfrac{\mathbf{X}_a^{b\prime} \mathbf{t}_a}{\left\| \mathbf{X}_a^{b\prime} \mathbf{t}_a \right\|}$, $a = 1, 2, ..., A$, for the right hand side of (A2.8) becomes

$$\mathbf{t}'_a \left(\sum_{b=1}^{B} \mathbf{X}_a^b \mathbf{X}_A^{b\prime} \right) \mathbf{t}_A = \sum_{b=1}^{B} \mathbf{t}'_a \mathbf{X}_a^b \mathbf{X}_A^{b\prime} \mathbf{t}_A = \sum_{b=1}^{B} \left\| \mathbf{t}'_a \mathbf{X}_a^b \right\| \cdot \left\| \mathbf{X}_A^{b\prime} \mathbf{t}_A \right\| \mathbf{p}_a^{b\prime} \mathbf{p}_A^b \text{ for } a = 1, 2, ..., A-1 \text{ (A2.10)}$$

Combining (A2.9) and (A2.10) we get

$$\lambda_A \mathbf{t}'_a \mathbf{t}_A = \sum_{b=1}^{B} \left\| \mathbf{t}'_a \mathbf{X}_1^b \right\| \cdot \left\| \mathbf{X}_A^{b\prime} \mathbf{t}_A \right\| \mathbf{p}_A^{b\prime} \mathbf{p}_a^b \quad \text{for } a = 1, 2, ..., A-1 \quad \text{(A2.11)}$$

Because of property 1 the right hand side of (A2.11) is zero and therefore $\mathbf{t}'_a \mathbf{t}_A = 0$ for $a = 1, 2, ..., A-1$. ☺

Proof of property 3.

$$\mathbf{t}_a^{b\prime} \mathbf{t}_k = \mathbf{p}_a^{b\prime} \mathbf{X}_a^{b\prime} \mathbf{t}_k \quad \text{(A3.1)}$$

Using lemma 1b we obtain

$$\mathbf{p}_a^{b\prime} \mathbf{X}_a^{b\prime} \mathbf{t}_k = \mathbf{p}_a^{b\prime} \prod_{l=k}^{a-1} \left(\mathbf{I}_{m_b} - \mathbf{p}_l^b \mathbf{p}_l^{b\prime} \right) \mathbf{X}_k^{b\prime} \mathbf{t}_k \quad \text{(A3.2)}$$

Since the block loadings are orthogonal according to property 1, it follows

$$\mathbf{p}_a^{b\prime}\prod_{l=k}^{a-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)\mathbf{X}_k^{b\prime}\mathbf{t}_k=\mathbf{p}_a^{b\prime}\mathbf{X}_k^{b\prime}\mathbf{t}_k \quad \text{(A3.3)}$$

Replacing $\mathbf{X}_k^{b\prime}\mathbf{t}_k=\left\|\mathbf{X}_k^{b\prime}\mathbf{t}_k\right\|\mathbf{p}_k^b$ according to (A1.6) we obtain

$$\mathbf{t}_a^{b\prime}\mathbf{t}_k=\mathbf{p}_a^{b\prime}\mathbf{X}_k^{b\prime}\mathbf{t}_k=\left\|\mathbf{X}_k^{b\prime}\mathbf{t}_k\right\|\mathbf{p}_a^{b\prime}\mathbf{p}_k^b=0 \quad \text{(A3.4)} \quad ☺$$

Proof of the preliminary lemma.

(a) The proof is based on the following equality which results from property 1:

$$\left(\mathbf{I}_{m_b}-\sum_{l=1}^{A-1}\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)=\prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)$$

It follows

$$\mathbf{X}_a^b=\mathbf{X}_1^b\left(\mathbf{I}-\sum_{l=1}^{a-1}\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)=\mathbf{X}_1^b\left(\prod_{l=1}^{a-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)\right)$$

$$\mathbf{X}_A^b=\mathbf{X}_1^b\left(\mathbf{I}-\sum_{l=1}^{A-1}\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)=\mathbf{X}_1^b\left(\prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)\right)$$

hence

$$\mathbf{X}_a^b\mathbf{X}_A^{b\prime}=\mathbf{X}_1^b\left(\prod_{l=1}^{a-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)\prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)\right)\mathbf{X}_1^{b\prime}$$

By using the equality $\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)^2=\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)$ we have

$$\prod_{l=1}^{a-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)\prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right)=\prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b}-\mathbf{p}_l^b\mathbf{p}_l^{b\prime}\right) \quad ☺$$

(b) Lemma 1b follows directly the previous equality. ☺

Appendix B. Orthogonality properties in MBPCA with deflation by block scores

$$\mathbf{X}_{a+1}^b=\mathbf{X}_a^b-\mathbf{t}_a^b\mathbf{q}_a^{b\prime}=\mathbf{X}_a^b-\frac{\mathbf{t}_a^b\mathbf{t}_a^{b\prime}}{\mathbf{t}_a^{b\prime}\mathbf{t}_a^b}\mathbf{X}_a^b,\ b=1,2,\ldots,B$$

Preliminary lemma.

$$\text{for } a>k \qquad \mathbf{X}_a^b=\prod_{l=k}^{a-1}\left(\mathbf{I}_{m_b}-\frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{X}_k^b$$

B1. Property 1. (Orthognality of block loadings)

For $1 \le k, a \le A$ and $k \ne a$ $\quad \mathbf{p}_k^{b\prime}\mathbf{p}_a^b = 0$

B2. Property 2. (Orthognality of block scores)

For $1 \le k, a \le A$ and $k \ne a$ $\quad \mathbf{t}_k^{b\prime}\mathbf{t}_a^b = 0$

Proof of property 1.

The proof of property 1 will be given by recurrence. First, we proof the property 1 for $A = 2$, i.e. $\left(\mathbf{p}_1^{b\prime}\mathbf{p}_2^b\right) = 0$. Note that the $\mathbf{p}_a^b$ for $a = 1, 2, \ldots$; are given by

$$\mathbf{p}_1^b = \frac{\mathbf{X}_1^{b\prime}\mathbf{t}_1}{\left\|\mathbf{X}_1^{b\prime}\mathbf{t}_1\right\|} \quad b = 1, 2, \ldots, B \quad \text{(B1.1)}$$

and

$$\mathbf{p}_2^b = \frac{\mathbf{X}_2^{b\prime}\mathbf{t}_2}{\left\|\mathbf{X}_2^{b\prime}\mathbf{t}_2\right\|}, \quad b = 1, 2, \ldots, B \quad \text{(B1.2)}$$

where in (B1.1) $\mathbf{X}_1^b = \mathbf{X}^b$ and in (B1.2) $\mathbf{X}_2^b = \mathbf{X}_1^b - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\mathbf{X}_1^b$.

Using $\mathbf{X}_2^b = \left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{X}_1^b$ then $\mathbf{p}_2^{b\prime}$ can be written as :

$$\mathbf{p}_2^{b\prime} = \frac{\mathbf{t}_2'\left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{X}_1^b}{\left\|\mathbf{X}_1^{b\prime}\left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{t}_2\right\|} \quad \text{(B1.3)}$$

Multiplying (B1.3) by $\mathbf{p}_1^b$

$$\mathbf{p}_2^{b\prime}\mathbf{p}_1^b = \frac{\mathbf{t}_2\left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{X}_1^b\mathbf{p}_1^b}{\left\|\mathbf{X}_1^{b\prime}\mathbf{t}_2\right\|\left\|\mathbf{X}_1^{b\prime}\left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{t}_2\right\|} \quad \text{(B1.4)}$$

Using the fact that $\mathbf{t}_1^b = \mathbf{X}_1^b\mathbf{p}_1^b$ in (B1.5) we obtain:

$$\mathbf{p}_2^{b\prime}\mathbf{p}_1^b = \frac{\mathbf{t}_2\left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{t}_1^b}{\left\|\mathbf{X}_1^{b\prime}\mathbf{t}_2\right\|\left\|\mathbf{X}_1^{b\prime}\left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{t}_2\right\|} \quad \text{(B1.5)}$$

Remark that $\left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{t}_1^b = 0$ and the right hand (B1.4) is equal 0, we obtain

$\left(\mathbf{p}_2^{b\prime}\mathbf{p}_1^b\right) = 0$.

We suppose now that property 1 is valid for $a = 1, 2, ..., A-1$, i.e. for $1 \le k, a \le A-1$ and $k \ne a$ we assume that $\mathbf{p}_k^{b\prime}\mathbf{p}_a^b = 0$. We will now show that $\mathbf{p}_a^{b\prime}\mathbf{p}_A^b = 0$ for $a = 1, 2, ..., A-1$, where

$$\mathbf{p}_A^b = \frac{\mathbf{X}_A^{b\prime}\mathbf{t}_A}{\left\|\mathbf{X}_A^{b\prime}\mathbf{t}_A\right\|} \quad \text{(B1.6)}$$

It follows that $\mathbf{X}_A^b = \mathbf{X}_{A-1}^b - \frac{\mathbf{t}_{A-1}^b\mathbf{t}_{A-1}^b{}^\prime}{\mathbf{t}_{A-1}^b{}^\prime\mathbf{t}_{A-1}^b}\mathbf{X}_{A-1}^b$ can be written as

$$\mathbf{X}_A^b = \prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{X}_1^b \quad \text{(B1.7)}$$

And using (B1.7), the equation (B1.6) can be written as the following.

$$\mathbf{p}_A^{b\prime} = \frac{\mathbf{t}_A^\prime\prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{X}_1^b}{\left\|\prod_{l=1}^{A-1}\left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{X}_1^b\mathbf{t}_A\right\|} \quad \text{(B1.8)}$$

Using the obvious fact that $\mathbf{X}_A^b = \prod_{l=a_0}^{A-1}\left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{X}_{a_0}^b$ and multiplying equation (B1.8) by $\mathbf{p}_{a_0}^b$ we obtain

$$\mathbf{p}_A^{b\prime}\mathbf{p}_{a_0}^b = \frac{\mathbf{t}_A'\left[\prod_{l=a_0}^{A-1}\left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{X}_{a_0}^b\mathbf{p}_{a_0}^b\right]}{\left\|\mathbf{X}_1^{b\prime}\left(\mathbf{I}_n - \sum_{a=1}^{A-1}\frac{\mathbf{t}_a^b\mathbf{t}_a^{b\prime}}{\mathbf{t}_a^{b\prime}\mathbf{t}_a^b}\right)\mathbf{t}_A\right\|} \quad \text{(B1.9)}$$

Using the fact that $\mathbf{t}_{a_0}^b = \mathbf{X}_{a_0}^b\mathbf{p}_{a_0}^b$ in (B1.9) we obtain:

$$\mathbf{p}_A^{b\prime}\mathbf{p}_{a_0}^b = \frac{\mathbf{t}_A'\left[\prod_{l=a_0}^{A-1}\left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{t}_{a_0}^b\right]}{\left\|\mathbf{X}_1^{b\prime}\left(\mathbf{I}_n - \sum_{a=1}^{A-1}\frac{\mathbf{t}_a^b\mathbf{t}_a^{b\prime}}{\mathbf{t}_a^{b\prime}\mathbf{t}_a^b}\right)\mathbf{t}_A\right\|}$$

Using the obvious fact that $\prod_{l=a_0}^{A-1}\left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\mathbf{t}_{a_0}^b = \prod_{l=a_0+1}^{A-1}\left(\mathbf{I}_n - \frac{\mathbf{t}_l^b\mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime}\mathbf{t}_l^b}\right)\left(\mathbf{I}_n - \frac{\mathbf{t}_{a_0}^b\mathbf{t}_{a_0}^{b\,\prime}}{\mathbf{t}_{a_0}^{b\,\prime}\mathbf{t}_{a_0}^b}\right)\mathbf{t}_{a_0}^b$ and by remarking that $\left(\mathbf{I}_n - \frac{\mathbf{t}_{a_0}^b\mathbf{t}_{a_0}^{b\,\prime}}{\mathbf{t}_{a_0}^{b\,\prime}\mathbf{t}_{a_0}^b}\right)\mathbf{t}_{a_0}^b = 0$ we obtain the hope result.

Proof of property 2.

The proof of property 2 will be given by recurrence. First, we proof the property 1 for $A = 2$, i.e. $\left(\mathbf{t}_1^{b\prime}\mathbf{t}_2^b\right) = 0$. Note that the $\mathbf{t}_a^b$ for $a = 1,2$; are given by

$$\mathbf{t}_1^b = \mathbf{X}_1^b\mathbf{p}_1^b \text{ and } \mathbf{p}_1^b = \frac{\mathbf{X}_1^{b\prime}\mathbf{t}_1}{\left\|\mathbf{X}_1^{b\prime}\mathbf{t}_1\right\|} \quad b = 1,2,...,B \quad \text{(B2.1)}$$

and

$$\mathbf{t}_1^b = \mathbf{X}_2^b\mathbf{p}_1^b\ \mathbf{p}_2^b = \frac{\mathbf{X}_2^{b\prime}\mathbf{t}_2}{\left\|\mathbf{X}_2^{b\prime}\mathbf{t}_2\right\|},\ b = 1,2,...,B \quad \text{(B2.2)}$$

where in (B2.1) $\mathbf{X}_1^b = \mathbf{X}^b$ and in (B2.2) $\mathbf{X}_2^b = \mathbf{X}_1^b - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\mathbf{X}_1^b$.

Using $\mathbf{X}_2^b = \left(\mathbf{I}_n - \frac{\mathbf{t}_1^b\mathbf{t}_1^{b'}}{\mathbf{t}_1^{b\prime}\mathbf{t}_1^b}\right)\mathbf{X}_1^b$ then $\mathbf{t}_2^b$ can be written as:

$$\mathbf{t}_2^b = \left(\mathbf{I}_n - \frac{\mathbf{t}_1^b \mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime} \mathbf{t}_1^b} \right) \mathbf{X}_1^b \mathbf{p}_2^b = \left(\mathbf{X}_1^b - \frac{\mathbf{t}_1^b \mathbf{t}_1^{b\prime}}{\mathbf{t}_1^{b\prime} \mathbf{t}_1^b} \mathbf{X}_1^b \right) \mathbf{p}_2^b \quad \text{(B2.3)}$$

Multiplying (B2.3) by $\mathbf{t}_1^{b\prime}$

$$\mathbf{t}_1^{b\prime} \mathbf{t}_2^b = \left(\mathbf{t}_1^{b\prime} \mathbf{X}_1^b - \mathbf{t}_1^{b\prime} \mathbf{X}_1^b \right) \mathbf{p}_2^b = 0 \ \text{(B2.4)}$$

We suppose now that property 1 is valid for $a = 1, 2, ..., A-1$, i.e. for $1 \le k, a \le A-1$ and $k \ne a$ we assume that $\mathbf{t}_k^{b\prime} \mathbf{t}_a^b = 0$. We will now show that $\mathbf{t}_a^{b\prime} \mathbf{t}_A^b = 0$ for $a = 1, 2, ..., A-1$, where

$$\mathbf{t}_A^b = \mathbf{X}_A^b \mathbf{p}_A^b \quad \text{(B2.5)}$$

It follows that $\mathbf{X}_A^b = \mathbf{X}_{A-1}^b - \dfrac{\mathbf{t}_{A-1}^b \mathbf{t}_{A-1}^b{}'}{\mathbf{t}_{A-1}^b{}' \mathbf{t}_{A-1}^b} \mathbf{X}_{A-1}^b$ can be written as

$$\mathbf{X}_A^b = \prod_{l=1}^{A-1} \left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right) \mathbf{X}_1^b \text{(B2.6)}$$

And using (B2.6), the equation (B2.5) can be written as the following.

Property 2.

$$\prod_{l=1}^{A-1} \left(\mathbf{I}_{m_b} - \frac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right) = \left(\mathbf{I}_{m_b} - \sum_{l=1}^{A-1} \frac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right)$$

$$\mathbf{t}_A^b = \left(\mathbf{I}_{m_b} - \sum_{l=1}^{A-1} \frac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right) \mathbf{p}_A^b \ \text{(B2.7)}$$

Using the obvious fact that $\mathbf{X}_A^b = \prod_{l=a_0}^{A-1} \left(\mathbf{I}_{m_b} - \dfrac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right) \mathbf{X}_{a_0}^b = \left(\mathbf{I}_{m_b} - \sum_{l=a}^{a-1} \dfrac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right) \mathbf{X}_{a_0}^b$ and multiplying equation (B2.7) by $\mathbf{t}_a^b$ we obtain

$$\mathbf{t}_{a_0}^b{}' \mathbf{t}_A^b = \mathbf{t}_{a_0}^b \left(\mathbf{I}_{m_b} - \sum_{l=a}^{A-1} \frac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right) \mathbf{X}_{a_0}^b \mathbf{p}_{a_0}^b \quad \text{(B2.8)}$$

$$\mathbf{t}_{a_0}^b{}' \mathbf{t}_A^b = \mathbf{t}_{a_0}^b{}' \left(\mathbf{I}_{m_b} - \sum_{l=a}^{A-1} \frac{\mathbf{t}_l^b \mathbf{t}_l^{b\prime}}{\mathbf{t}_l^{b\prime} \mathbf{t}_l^b} \right) = 0$$

Appendix C. Computation of the explained variances for each block

C1: Deflation by global scores:

$$\operatorname{var}\left(\mathbf{X}^{b}\right)=\left\|\sum_{a=1}^{R}\frac{\mathbf{t}_{a}\mathbf{t}_{a}'}{\mathbf{t}_{a}'\mathbf{t}_{a}}\mathbf{X}_{a}^{b\prime}\right\|^{2}=\sum_{a=1}^{R}\left\|\frac{\mathbf{t}_{a}\mathbf{t}_{a}'}{\mathbf{t}_{a}'\mathbf{t}_{a}}\mathbf{X}_{a}^{b\prime}\right\|^{2}$$

$$\operatorname{var}\left(\mathbf{X}^{b}\right)=\left\|\sum_{a=1}^{R}\frac{\mathbf{t}_{a}\mathbf{t}_{a}'}{\mathbf{t}_{a}'\mathbf{t}_{a}}\mathbf{X}_{a}^{b\prime}\right\|^{2}=\sum_{a,k=1}^{R}trace\left(\mathbf{X}_{a}^{b}\frac{\mathbf{t}_{a}\mathbf{t}_{a}'}{\mathbf{t}_{a}'\mathbf{t}_{a}}\frac{\mathbf{t}_{k}\mathbf{t}_{k}'}{\mathbf{t}_{k}'\mathbf{t}_{k}}\mathbf{X}_{a}^{b\prime}\right)=\sum_{a,k=1}^{R}\left(\mathbf{t}_{a}'\mathbf{t}_{k}\right)trace\left(\mathbf{X}_{a}^{b}\frac{\mathbf{t}_{a}}{\mathbf{t}_{a}'\mathbf{t}_{a}}\frac{\mathbf{t}_{k}'}{\mathbf{t}_{k}'\mathbf{t}_{k}}\mathbf{X}_{a}^{b\prime}\right)$$

$$\sum_{a=1}^{R}trace\left(\mathbf{X}_{a}^{b}\frac{\mathbf{t}_{a}\mathbf{t}_{a}'}{\mathbf{t}_{k}'\mathbf{t}_{k}}\mathbf{X}_{a}^{b\prime}\right)=\sum_{a=1}^{R}\left\|\frac{\mathbf{t}_{a}\mathbf{t}_{a}'}{\mathbf{t}_{a}'\mathbf{t}_{a}}\mathbf{X}_{a}^{b\prime}\right\|^{2}$$

C2: Deflation by block loadings:

$$\operatorname{var}\left(\mathbf{X}^{b}\right)=\left\|\sum_{a=1}^{R}\mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{a}^{b\prime}\right\|^{2}=\sum_{a=1}^{R}\left\|\mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{a}^{b\prime}\right\|^{2}=$$

$$\operatorname{var}\left(\mathbf{X}^{b}\right)=\left\|\sum_{a=1}^{R}\mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{a}^{b\prime}\right\|^{2}=\sum_{a,k=1}^{R}trace\left(\mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{a}^{b\prime}\mathbf{p}_{k}^{b}\mathbf{p}_{k}^{b\prime}\mathbf{X}_{a}^{b\prime}\right)=\sum_{a,k=1}^{R}\mathbf{p}_{a}^{b\prime}\mathbf{p}_{k}^{b}trace\left(\mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{k}^{b\prime}\mathbf{X}_{k}^{b\prime}\right)$$

$$=\sum_{a=1}^{R}trace\left(\mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{a}^{b\prime}\mathbf{X}_{a}^{b\prime}\right)=\sum_{a=1}^{R}\left\|\mathbf{X}_{a}^{b}\mathbf{p}_{a}^{b}\mathbf{p}_{a}^{b\prime}\right\|^{2}$$

C3: Deflation by block scores:

$$\operatorname{var}\left(\mathbf{X}^{b}\right)=\left\|\sum_{a=1}^{R}\frac{\mathbf{t}_{a}^{b}\mathbf{t}_{a}^{b\prime}}{\mathbf{t}_{a}^{b\prime}\mathbf{t}_{a}^{b}}\mathbf{X}_{a}^{b\prime}\right\|^{2}=\sum_{a=1}^{R}\left\|\frac{\mathbf{t}_{a}^{b}\mathbf{t}_{a}^{b\prime}}{\mathbf{t}_{a}^{b\prime}\mathbf{t}_{a}^{b}}\mathbf{X}_{a}^{b\prime}\right\|^{2}$$

$$\operatorname{var}\left(\mathbf{X}^{b}\right)=\left\|\sum_{a=1}^{R}\frac{\mathbf{t}_{a}^{b}\mathbf{t}_{a}^{b\prime}}{\mathbf{t}_{a}^{b\prime}\mathbf{t}_{a}^{b}}\mathbf{X}_{a}^{b\prime}\right\|^{2}=\sum_{a,k=1}^{R}trace\left(\mathbf{X}_{a}^{b}\frac{\mathbf{t}_{a}^{b}\mathbf{t}_{a}^{b\prime}}{\mathbf{t}_{a}^{b\prime}\mathbf{t}_{a}^{b}}\frac{\mathbf{t}_{k}^{b}\mathbf{t}_{k}^{b\prime}}{\mathbf{t}_{k}^{b\prime}\mathbf{t}_{k}^{b}}\mathbf{X}_{a}^{b\prime}\right)=\sum_{a,k=1}^{R}\left(\mathbf{t}_{a}^{b\prime}\mathbf{t}_{k}^{b}\right)trace\left(\mathbf{X}_{a}^{b}\frac{\mathbf{t}_{a}^{b}}{\mathbf{t}_{a}^{b\prime}\mathbf{t}_{a}^{b}}\frac{\mathbf{t}_{k}^{b\prime}}{\mathbf{t}_{k}^{b\prime}\mathbf{t}_{k}^{b}}\mathbf{X}_{a}^{b\prime}\right)$$

$$=\sum_{a=1}^{R}trace\left(\mathbf{X}_{a}^{b}\frac{\mathbf{t}_{a}^{b}\mathbf{t}_{a}^{b\prime}}{\mathbf{t}_{k}^{b\prime}\mathbf{t}_{k}^{b}}\mathbf{X}_{a}^{b\prime}\right)=\sum_{a=1}^{R}\left\|\frac{\mathbf{t}_{a}^{b}\mathbf{t}_{a}^{b\prime}}{\mathbf{t}_{a}^{b\prime}\mathbf{t}_{a}^{b}}\mathbf{X}_{a}^{b\prime}\right\|^{2}$$

Deflation strategies	$\mathbf{X}_{a+1}^{b}$
Deflation by Global scores [17]	$\mathbf{X}_a^b - \frac{\mathbf{t}_a\mathbf{t}_a'}{\mathbf{t}_a'\mathbf{t}_a}\mathbf{X}_a^{b\prime}$
Deflation by Block loadings [8]	$\mathbf{X}_a^b - \mathbf{X}_a^b\mathbf{p}_a^b\mathbf{p}_a^{b\prime}$
Deflation by Block scores [12]	$\mathbf{X}_a^b - \frac{\mathbf{t}_a^b\mathbf{t}_a^{b\prime}}{\mathbf{t}_a^{b\prime}\mathbf{t}_a^b}\mathbf{X}_a^b$

Table 1: Deflation in MBPCA.

Deflation strategies	$\mathbf{p}_a^b$	$\mathbf{t}_a^b$	$\mathbf{p}_a$	$\mathbf{t}_a$
Deflation by Global scores	*Not Orthogonal*	*Not Orthogonal*	*Orthogonal*	*Orthogonal*
Deflation by Block loadings	*Orthogonal*	*Not Orthogonal*	*Orthogonal*	*Orthogonal*
Deflation by Block scores	*Orthogonal*	*Orthogonal*	*Orthogonal*	*Not Orthogonal*

Table 2: Orthogonality properties for block loadings, block scores, global loadings and global scores for the different deflation methods.

Deflation strategies	***Reconstruction formulae of Block***
Deflation by Global scores	$\mathbf{X}^b = \sum_{a=1}^{r}\mathbf{t}_a\left(\frac{\mathbf{X}_a^{b\prime}\mathbf{t}_a}{\mathbf{t}_a'\mathbf{t}_a}\right)'$
Deflation by Block loadings	$\mathbf{X}^b = \sum_{a=1}^{r_b}\mathbf{t}_a^b\mathbf{p}_a^{b\prime}$
Deflation by Block scores	$\mathbf{X}^b = \sum_{a=1}^{r_b}\mathbf{t}_a^b\mathbf{q}_a^{b\prime}$

Table 3: Reconstruction of block matrices for the different deflation methods

Deflation strategies	***Decomposition the total variance***
Deflation by Global scores	$\mathrm{var}\left(\mathbf{X}^b\right) = \sum_{a=1}^{r}\left\Vert\mathbf{t}_a\left(\frac{\mathbf{X}_a^{b\prime}\mathbf{t}_a}{\mathbf{t}_a'\mathbf{t}_a}\right)'\right\Vert^2$
Deflation by Block loadings	$\mathrm{var}\left(\mathbf{X}^b\right) = \sum_{a=1}^{r_b}\left\Vert\mathbf{t}_a^b\mathbf{p}_a^{b\prime}\right\Vert^2$
Deflation by Block scores	$\mathrm{var}\left(\mathbf{X}^b\right) = \sum_{a=1}^{r_b}\left\Vert\mathbf{t}_a^b\mathbf{q}_a^{b\prime}\right\Vert^2$

Table 4: The total explained variance for each deflation step for MBPCA

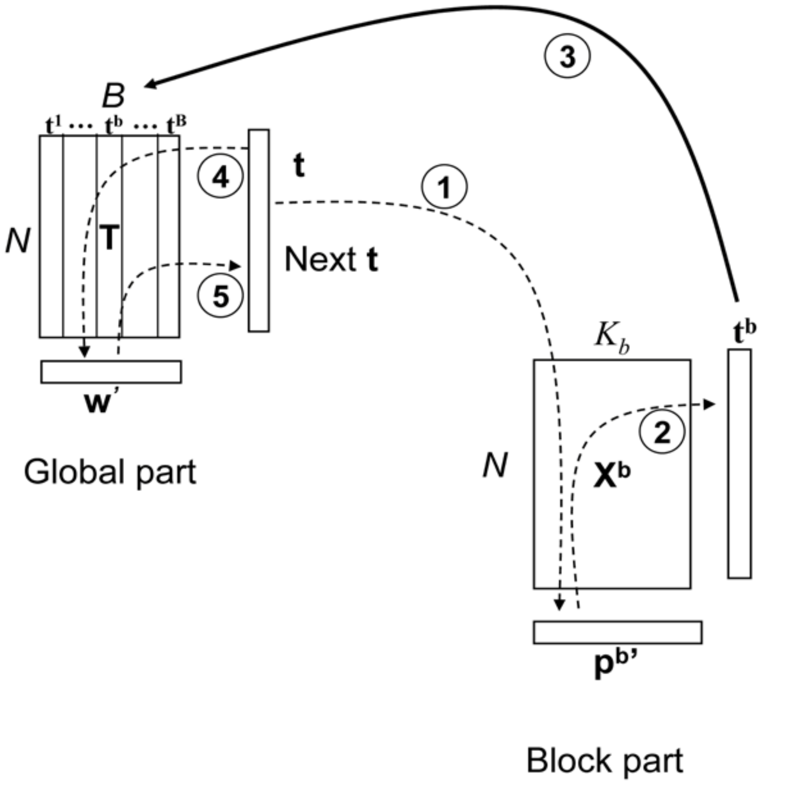


Figure 1: The iterative algorithm for MBPCA.

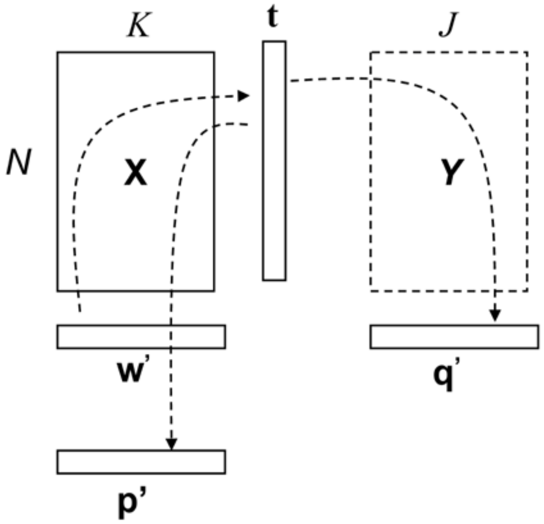


Figure 2: The deflation strategy in PLSR.

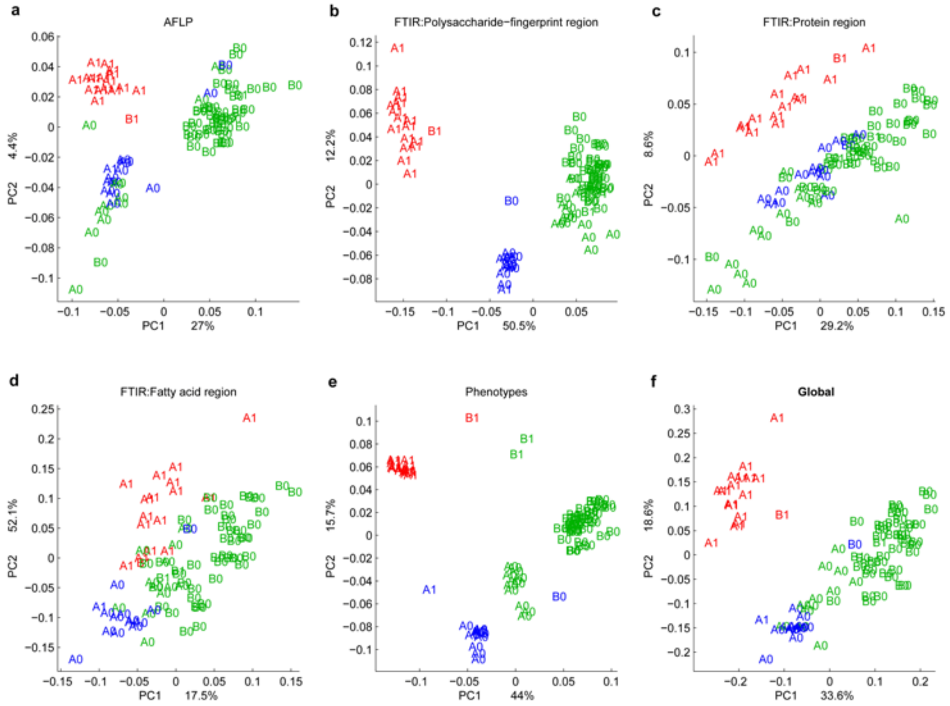


Figure 3: MBPCA with deflation by global scores (CPCA): Block score plots (a-e) and global score plot (f) for CPCA for an example from biospectroscopy.

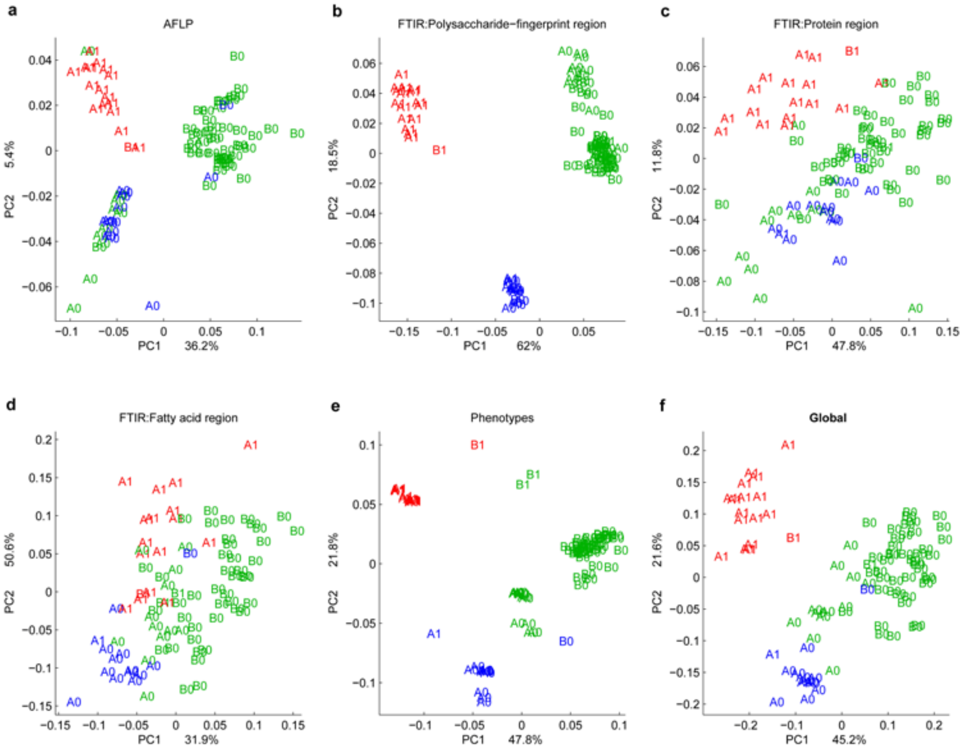


Figure 4: MBPCA with deflation by block loadings: Block score plots (a-e) and global score plot (f) for MBPCA with deflation by block loadings for the same example as in Fig. 3.

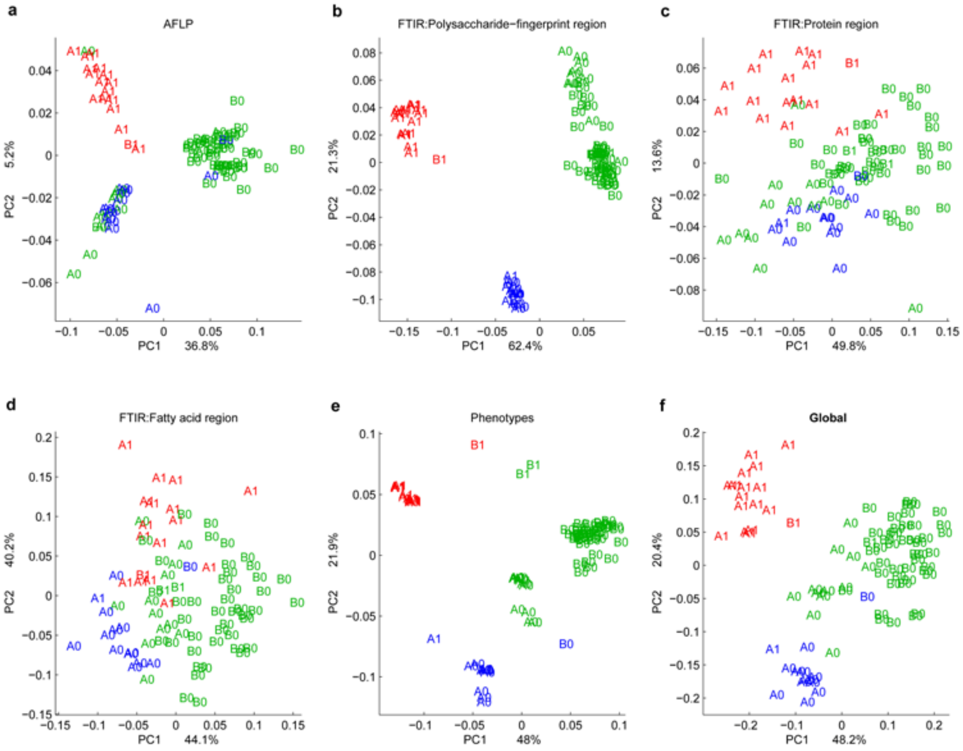


Figure 5: MBPCA with deflation by block scores: Block score plots (a-e) and global score plot (f) for MBPCA with deflation by block scores for the same example as in Fig. 3.

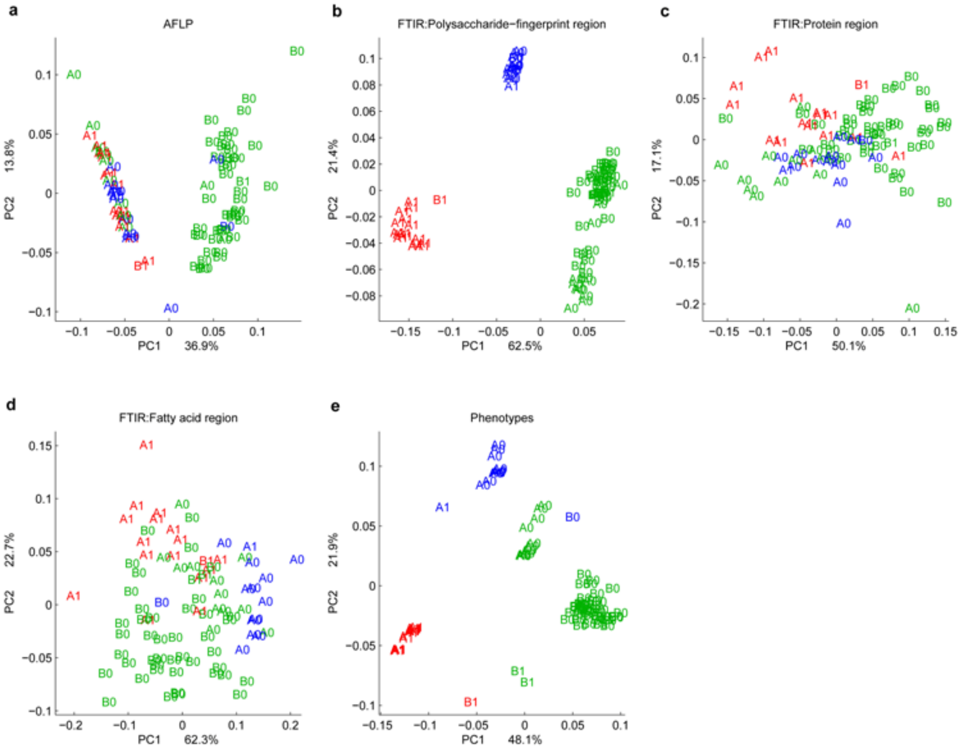


Figure 6: PCA of the single blocks: Score plots (a-e) for PCA of the single blocks for the same example as in Fig. 3.

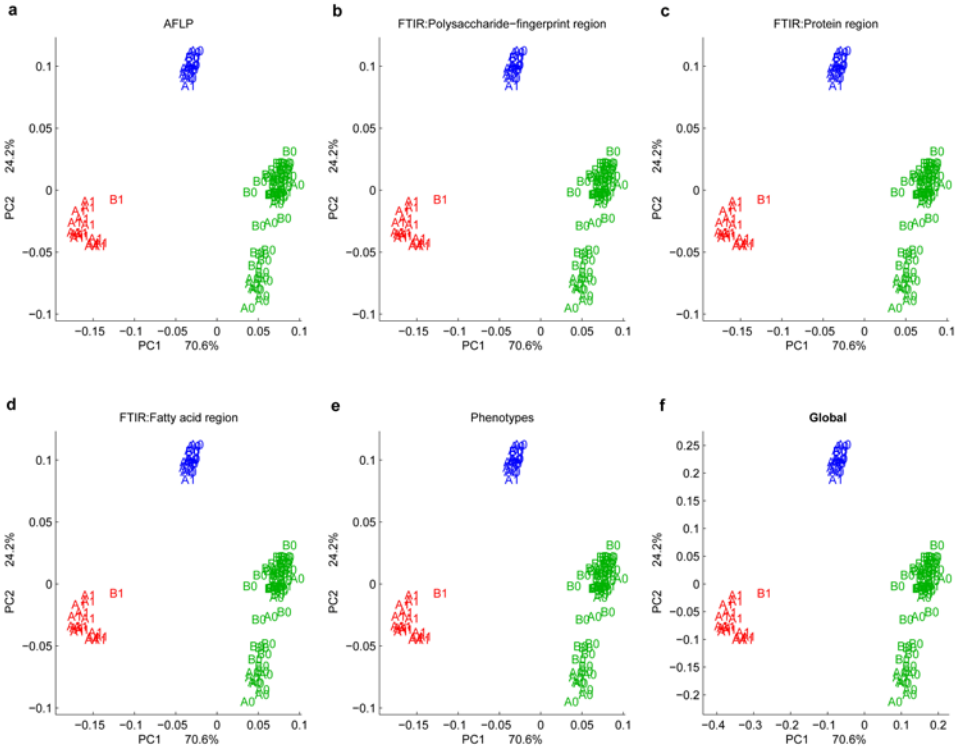


Figure 7: MBPCA with any of the three deflation strategies: Block score plots (a-e) and global score plot (f) for MBPCA for a simulated example as explained in the text.

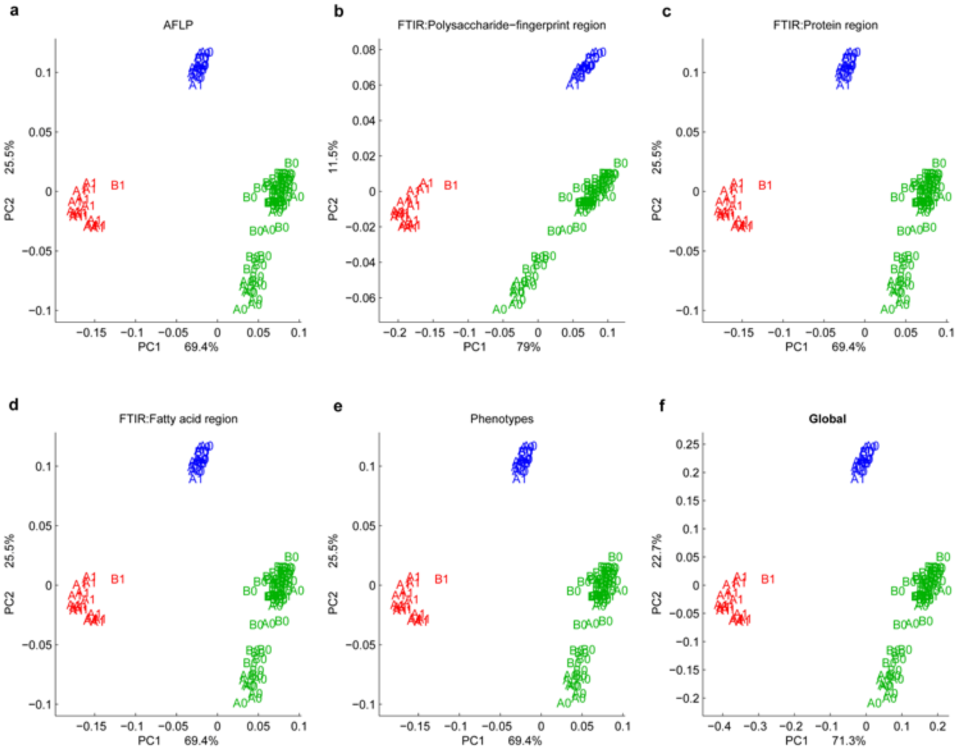


Figure 8: MBPCA with deflation by block loadings: Block score plots (a-e) and global score plot (f) for MBPCA with deflation by block loadings for a simulated example.

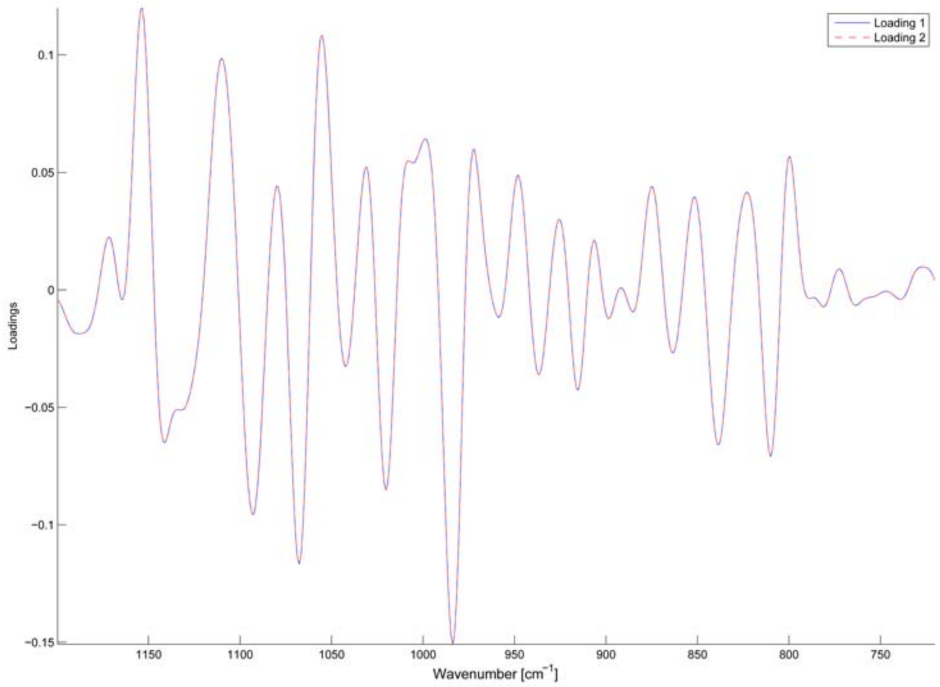


Figure 9: MBPCA with deflation by global scores (CPCA): First and second block loadings for CPCA for the second block for the same example as in Fig. 8.

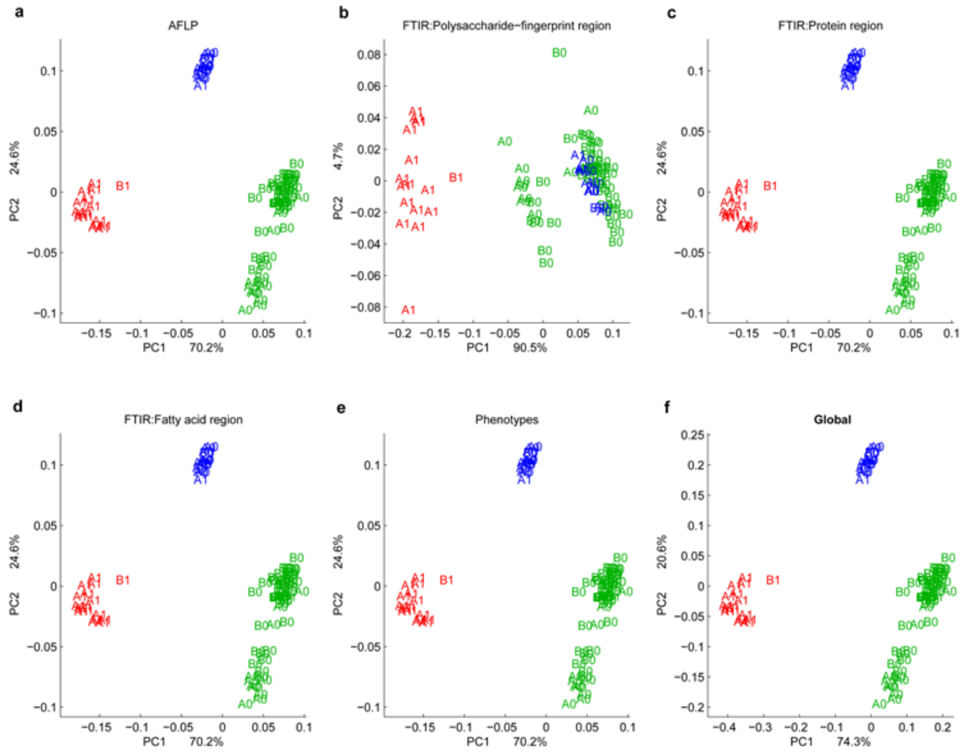


Figure 10: MBPCA with deflation by block loadings: Block score plots (a-e) and global score plot (f) for MBPCA with deflation by block loadings for the same example as in Fig. 8.

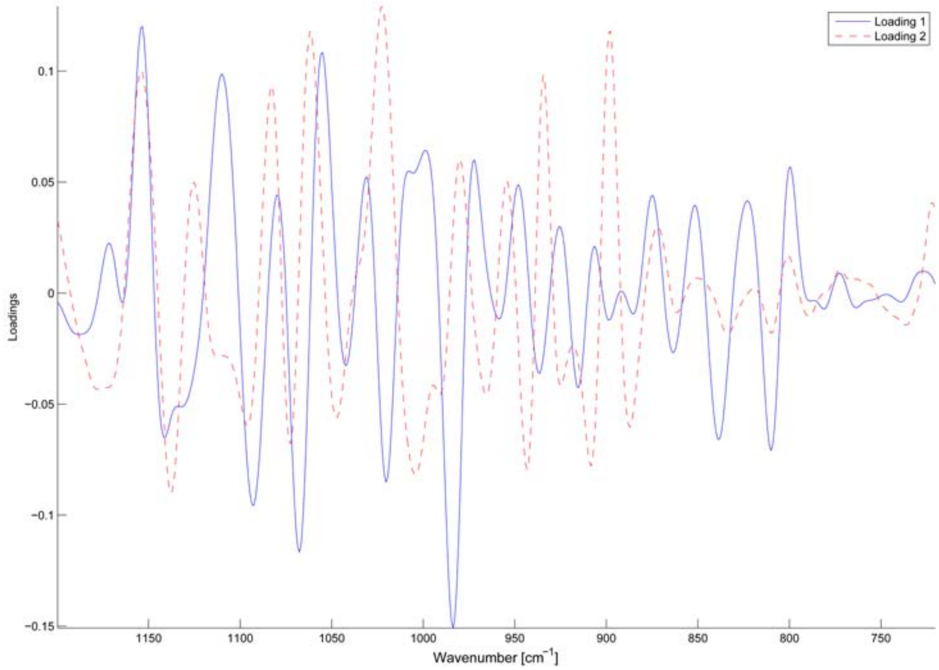


Figure 11: MBPCA with deflation by block loadings: First and second block loadings for MBPCA with deflation by block loadings for the second block for the same example as in Fig. 8.

Paper V

Simultaneous analysis of inter- and intra-class lipid changes in lipidomics studies

Sahar Hassani [1,2§], Harald Martens [1,2], Inger Ottestad [3,4], Grethe I. Borge [1], Mari CW. Myhrstad [3] and Achim Kohler [2,1]

[1] Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, 1430 Ås, Norway

[2] Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology (IMT), Norwegian University of Life Sciences, 1432 Ås, Norway

[3] Department of Health, Nutrition and Management, Faculty of Health Sciences, Oslo and Akershus University College of Applied Sciences, P.O. Box 4, St Olavs plass, 0130 Oslo, Norway

[4] Department of Nutrition, Institute for Basic Medical Sciences, University of Oslo, P.O. Box 1046 Blindern, 0317 Oslo, Norway

[§] Corresponding author

Email addresses:

SH: sahar.hassani@nofima.no

HM: harald.martens@nofima.no

IO: inger.ottestad@hioa.no

GB: grethe.iren.borge@nofima.no

MM: mari.myhrstad@hioa.no

AK: achim.kohler@nofima.no

Abstract

A framework is presented that allows simultaneous analysis of metabolic shifts between lipid classes and remodelling shifts within lipid classes. The framework consists of a particular within/between-class pre-processing, followed by multi-block multivariate data analysis. Particular features are illustrated using simulated data, and the framework is demonstrated for a lipidomics data set from a human intervention study.

Keywords

Lipidomics; lipid classes; lipid remodelling; multi-block data analysis; pre-processing; bilinear data modelling;

Background

Lipidomics is an emerging –omics technology aiming at investigating the diverse nature and roles of lipids in biological systems. *Lipids* are a diverse group of chemical species, often grouped into a low number of more homogeneous *classes* (chemical groups), each containing a range of chemical *species* (molecular types). Biologically, lipids are involved in many different cellular functions and metabolic pathways. Therefore the "*lipidome*" (the profile consisting of "all" relevant lipid species) can vary in many different ways between different tissue types or at different points in time, as well as between different individuals, e.g. due to different food habits, medical syndromes, treatments, etc. As science's ability to quantify the details of the lipidome increases, the challenge of interpreting the lipidomic data also increases.

Lipidomics studies are carried out in many fields of science. In nutrition and food science [1-3], scientists are interested in studying effects of different diets on the lipid profile in body liquids and cells in order to estimate health effects of the diets. In pharmaceutical and medical sciences [4, 5] lipidomics studies are undertaken to test effects of drugs on lipid profiles. In healthcare studies [6-9] investigating the lipid profiles is an important step towards a better healthcare management. For instance, the cholesterol level and composition has been used for many years as biomarker for the estimation of the risk of heart disease, while the triglyceride level and composition are important biomarkers for estimating the risk of diabetes [10].

Eight major, more or less homogenous lipid classes are usually recognized: *Fatty acids* (free, mono-di and tri-glycerides), *Glycerolipids, Sphingolipids, Sterols, Glycerophospholipids, Prenol lipids, Saccharolipids and Polyketides* [11]. Each of these classes has its own sub-classification hierarchy. For instance, Sphingolipids are classified into the following three sub-classes: *Ceramides* (Cer), *Sphingomyelins* (SM) and *Glycosphingolipids* ([glycan]-Cer), while *Glycerophospholipids* are classified into several sub-classes among them are *Glycerophosphocholines* (PC), *Glycerophosphoethanolamines* (PE), *Glycerophosphates* (PA), *Glycerophosphoserines* (PS) and *Glycerophosphoglycerols* (PG).

Shifts in the concentrations *between* the different lipid classes may reflect important metabolic processes that transform lipids in one lipid class into lipids in a different lipid class. On the other hand, concentration changes of lipid species *within a given lipid class* (or sub-class) may also take place – this is in lipidomics termed *remodelling*. The two types of processes can take place simultaneously or independently, being due to the same cause or different causes. To discover the causal patterns behind complex lipidomic changes, both types of changes need to be experimentally perturbed and data-analytically detected and visualized. For instance, when studying remodelling effects it is important to monitor the relative changes within each of the lipid classes.

Data analytical challenges in lipidomics. Several studies [12-15] have mentioned the need for modern statistical and computational tools in lipidomics. One of the reasons is that modern instrumentation offers new opportunities to understand causality and predict developments, but also poses new challenges – in lipidomics as elsewhere. High-throughput high-dimensional instrumentation (chromatography, mass spectrometry etc.) allows identification and quantification of hundreds or thousands of lipid species in each biological *sample*, at widely different concentrations. This large and heterogeneous set of measured *variables* presents cognitive as well as statistical problems: How can we find the essential changes in the lipidome, and how can we test if the observed effects are statistically valid? With a high number of input variables (lipid constituents) from a low number of samples (e.g. patients), traditional univariate analysis methods are easily misinterpreted, for two reasons. The first one is the *multiple testing* problem: When e.g. conventional t-tests are performed *independently* on a high number of variables in a low number of independent samples, the chance for false discovery is high. To guard against this, penalizing the p-values from the many t-tests is often used – e.g. by Bonferroni correction [16, 17]. But this can lead to false negatives – even causally valid relationships in the data may then be deemed "non-significant". For instance, detecting potential lipid biomarkers is often one of the major goals of the lipidomics studies. Univariate statistical tests (such as Student t-test) are widely used for this purpose. However, a multiple-testing problem is often likely, due to the large number of lipids and low number of patients in these studies.

But traditional univariate analyses pose a second problem as well: Changes in the different lipids reflect systematic biological mechanisms. Therefore lipids within a lipid class are often inter-correlated, and so are the different lipid classes. These natural *inter-correlations* between different lipids can lead to misguided interpretation, if traditional, but commonly used full-rank statistical regression and discrimination tools are used: Because these natural inter-correlations then pose estimation problems ("the multi-collinearity problem'), only a small subset of the available input variables is then reported, and this variable selection process may be misleading.

One-block multivariate data modelling. The risk of over-fitting is lowered when using subspace-methods like Principal Component Analysis (PCA) [18]. Recently, various studies have employed PCA for investigating and interpreting lipidomics data sets [19-22]. PCA is an un-supervised subspace method for multivariate data analysis. This method can detect and display patterns of samples and variables, revealing their actual relationship in the data and enabling new hypothesis generation. The subspace-method collects the naturally occurring patterns of co-variation between the many input lipid variables into a lower number of "super-variables" (estimated latent variables). And since strong co-variation patterns are thus detected in an unsupervised manner, and only few latent variables are usually needed, the problem of multiple testing can be greatly reduced. In fact, methods such as PCA utilize such natural inter-correlation patterns as a stabilizing advantage (rather than a "collinearity problem" as in many traditional statistical methods), and the interpretation of the low-dimensional solutions is far simpler than assessing all the input variables individually. In most data sets, there are strong inter-correlation patterns between single variables, desirable ones due to common causal mechanisms, and undesirable ones due to inadequate sampling. The analysis of a given data set can not resolve these ambiguities, because the measured data set at hand simply does not contain enough information. However, when all the inter-correlated variables are visualized together in a sub-space that reveals the co-variation patterns, the scientist may come up with causal hypotheses, based on his or her background knowledge. That is one of the main strengths of graphically oriented multivariate data modelling by subspace analysis.

Another important challenge for the data analysis of lipidomics data is the need to *integrate information from different classes of lipids* and analyze them simultaneously, in order to explore lipid-lipid remodelling within classes as well as the dynamics of metabolism between classes. Above all, there is a need for integrating lipidomics data together with *other types of variables* in order to study how the co-variation patterns within and between the lipid classes relate to other measured variables, and/or background information such as personal data and medical or nutritional treatment. In order to give answers to these questions, data analytical tools that are capable of integrating lipidomics data into one interpretable data model are needed.

Two-block multivariate data modelling. The PCA method, which extracts the internal patterns of co-variation within *one block* or class of input variables at a time (by maximizing the amount of variance described *within* a block of variables), has been extended to *two-block* subspace regression methods. Partial Least Squares Regression (PLSR) [23] is a supervised subspace method that maximizes covariance *between two* blocks of variables. PLSR is now widely used in lipidomics to relate lipid variables to other types of variables, e.g. when the researcher is interested in interpreting the lipid measurement profile in light of the design information [24-27]. However, when more than two blocks of variables are to be interrelated (e.g. several lipid classes, gene expression, external patient data and experiments design descriptors), the two-block PLSR is not sufficiently informative.

Multi-block multivariate data modelling. Data modelling methods that are capable of maintaining block structure, as for example the block structure of the blocks of lipid classes, are called *multi-block methods*. Multi-block methods were originally introduced for integrating data in sensometrics and psychometrics [28-31]. In recent years these methods have received attention in the field of data modelling of -omics data [32-34].

Pre-processing to change the units of the data. A promising use of multi-block analysis of lipidomics data is the study of *remodelling* effects. When lipids are said to be remodelled within specific lipid classes, it means that some kind of lipid dynamics is taking place within those specific lipid classes. This is opposed to lipid *metabolisms*, which transfers lipids form one lipid class into lipids from a different lipid class.

Detecting the remodelling effects and revealing the mechanisms between different lipid classes are some of the major challenges in the field of lipidomics. The reason is that, in order to reveal these two effect types clearly, the data modelling processes need to keep both the lipid class and species structures in the data modelling process, and they are not necessarily easy to represent with variables given in the same *unit*.

For instance, metabolic processes leading to conversion of molecules from one class into another may most easily be studied in a unit that reflects mass balance (e.g. mg/cell or microgram/gram tissue). But the remodelling within a lipid class may most easily be interpreted in a unit that is insensitive to the total concentration of each lipid class, e.g. in % of the lipids within each class. There is a need for data analysis methods that are able to analyze data sets on the lipid species level and on the lipid class level *simultaneously* and to make the mechanisms visible. This calls for conscious *pre-processing* of the variables to ensure that they are represented in suitable units: otherwise they will not correlate across the samples in a way that reveal the difference between metabolic changes and class remodelling.

The aim of this study is to present a framework for integrating lipidomics data that is capable of detect and reveal both between- and within-class variations, detecting both metabolic and remodelling processes and showing how they relate to external variables such as feeding descriptions. This is accomplished by applying a pre-processing strategy that keeps the lipid class and species structure in the data modelling process. Two different versions of the data are then submitted to multi-block techniques such as Consensus PCA (CPCA) and Multi-block PLSR (MBPLSR).

These methods are intended to provide the users with efficient methods of information extraction, compact and statistically stable information representation, simple statistical validation tools, and good graphical functionalities for interpretation in terms of both overview and detail. In this article we present statistical validation tools based on cross-validation (assessing optimal subspace model rank and model stability) and Monte Carlo permutation tests (assessing the distribution of the null hypotheses of no valid relationships in the data). This is applied for testing the contribution of lipid classes and species to the detected patterns. The pre-processing and subspace data modelling is demonstrated on a real lipidomic data set from a human nutrition study. But first, some essential phenomena are illustrated by slightly modifying this data set artificially.

Results

Structuring lipid classes prior to data integration

In order to analyze lipidomics data we suggest arranging lipidomics data as a multi-block data set, e.g. according to the major classes of lipids mentioned above. This is done by aligning the data in such a way that samples (e.g. patients) are represented by rows and variables (lipid species, background variables etc.) by columns. The lipid species are grouped into different data blocks according to their lipid classes, each data block containing lipids belonging to the same lipid class. Each individual lipid variable is given a short name, to give informative plots. Fig. 1a illustrates an example for arranging the lipidomics data set as a multi-block data set. We assign one additional lipid class to possibly *unidentified* lipids. This predefinition of lipid classes is a structure, which the scientist imposes. The grouping of lipids into any different lipid classes or sub-classes is in principle also possible.

Pre-processing of structured data with respect to the lipid classes

In order to highlight remodelling effects we suggest using two alternative normalization strategies:

1) To represent all variables in, or proportionally to the same basic unit, to maintain over-all mass balance, i.e. to normalize each row containing the total lipid profile for one sample with respect to the total amount of lipids.

2) To normalize all the rows in a block (lipid class) relative to the total amount of lipids within this class. First, the concentration of lipids of each lipid class is summed and defined as a new variable within that class, referred to as its ‘lipid class sum’. Then the individual lipid variables are normalized relative to this lipid class sum. This means that for each sample (row), all lipids in each class are divided by its lipid class sum. Thus, each class consists of a set of variables reflecting the relative concentration (e.g. percentage) of its individual lipids, plus one variable reflecting its over-all contribution to the total amount of lipid.

The second lipid class normalization procedure is illustrated in Fig. 1b for an example lipid class. Detailed theory for implementation is provided in Methods section. Thus, metabolic changes in the total amount of the different lipid classes are to be revealed by patterns of co-variation in the lipid class sums in the different blocks, while within-class remodelling processes are to be revealed by the relative fraction (or percentages) of lipids within each lipid class. If both metabolic and remodelling processes are controlled by the same external phenomena, they will be revealed by the *same* subspace dimensions, while if they are independent of each other, they will be seen to span *different* subspace dimensions.

CPCA: An explorative unsupervised approach for integrating lipid classes

CPCA [35] is an un-supervised multivariate exploratory technique that allows integrating several data blocks with multivariate measured variables for the same samples. As in PCA and PLSR, every variable is usually mean-centred and standardized to a total initial variance of 1 before the data modelling process. Moreover, as a standard procedure, the different data blocks in CPCA (and MBPLSR) are then scaled in order to set them on an equal footing, i.e. so that the total initial sum-of-squares is the same for all blocks. The rationale behind this is to give every data block the same importance in the data modelling process. This scaling of the variables enhances the probability of the CPCA detecting all interesting co-variance structures before being hit by noise. But it does not affect the relationship between them, provided that the model loading parameters are plotted in a descaled or scale free version. Details about mean-centring and scaling as well as a brief description of CPCA are given in the Methods section. A more detailed description on the implementation of CPCA together with algorithms for validation can be found in [32]. One of the strengths of CPCA is that visualization tools of sample variation patterns and variable variation patterns are readily available. As in PCA and PLSR, the patterns of covariation between the variables are shown by their so-called loadings, usually pair-wise for the first few subspace dimensions. For visual simplicity, these loadings are often represented as unit-free correlation loadings (correlations between input variables and latent variables). The corresponding patterns of samples are plotted as so-called scores. Together, the plots of loadings and scores allow the interpretation of the main variation patterns present in each block and relations between the data blocks. For visualization of sample variation patterns, global and block score plots are used, while correlation loading plots are employed to discover relations between variables in each data block and between the data blocks. The global score plot offers an overview over sample variation patterns and sample clusters as a consensus for all the different lipid classes. Thus, it gives a common compromise picture of the samples' groupings, more or less shared between all of the lipid classes. For data in the first pre-processing, it shows the main patterns of the whole lipid profile. In addition to the global score plot, a block score plot is produced for each lipid class, visualizing how these patterns are manifested in each individual block.

An overview of the lipid-lipid interactions is shown in the correlation loading plot. This plot displays all the lipids within the space of few latent variables and helps identifying correlations among all lipids, thus providing an opportunity for studying the causal pathways and their dynamics within and between different lipid species and lipid classes, under the different pre-processing regimes applied.

Identifying the grouping patterns in the score plots is somewhat subjective. Therefore, it is crucial to validate the detected patterns. In a previous article we proposed studying the global and block Root Mean Squared Error (RMSE) plots as a tool for validating the CPCA results [32]. Investigating these plots helps identifying the lipid classes that are significantly contributing to the detected grouping patterns. Here, we propose in addition to assess significance by Monte Carlo testing. Details are in the Methods section.

The above mentioned permutation tests provide p-values for the contribution of each lipid class to the grouping of samples which is then used for detecting the important lipid classes. Investigating the important lipids may be done by running an uncertainty t-test which results in a p-value for each lipid. We previously described the details for implementing the uncertainty t-test in [32].

Integrating two lipid classes with simulated remodelling effect

In order to investigate the effect of the two different normalization procedures (described in Fig. 1b) a data set was simulated. The data set mimics a two-block lipidomics data set consisting of lipids belonging to the lipid classes LycoPC and Ceramides (Cer). The data set was simulated based on a real lipidomics data set where LycoPC and Ceramides were measured for three different intervention groups at the baseline and after the intervention period. The complete data set is described in more detail in the application example part. In the simulated data set, remodelling was predefined by increasing the concentrations of the compound named Cer(d18:1/24:0) by a small amount, while correspondingly decreasing the concentration of another compound, Cer(d18:1/22:6) by the same small amount in Group 1 samples during the intervention period. This was done by adding a small number to the baseline amount of Cer(d18:1/24:0) and then using the increased amount as the level of Cer(d18:1/24:0) after the intervention period. A similar process (in the opposite direction) was performed by reducing the amount of Cer(d18:1/22:6) from the baseline and using the results for the level of Cer(d18:1/22:6) after the intervention period. The rest of the real data set remained untouched. Correlation loading plots for the CPCA modelling of the simulated data set after normalization with respect to total amount of lipids are shown in Fig. 2a and after normalization with respect to the lipid classes in Fig. 2b. The remodelling effect of Cer(d18:1/22:6) into Cer(d18:1/24:0) is depicted most clearly by Fig. 2b: they are located opposite of each other and they are related to Group 1 samples. Cer(d18:1/24:0) is positively correlated to Group 1 which leads to the conclusion that the relative amount of this lipid in Group 1 has increased during the intervention. The negative correlation of Cer(d18:1/22:6) with Group 1 is an indication for a reduced percentage of this lipid during the intervention period. The total amount of Ceramides (TotalCer) falls inside the inner circle and is therefore not significantly changing during the intervention period. In the correlation loading plot in Fig. 2a we cannot observe this remodelling process within the Ceramides lipid class, since Cer(d18:1/24:0) and Cer(d18:1/22:6) are not located on opposite sides of the correlation loading plot.

When investigating the significance of the observed changes in the Ceramides lipid class (using Monte-Carlo permutation tests) we found it to be not significant when we normalize with respect to the total amount of lipids (by method 1, Fig. 2a, p-value=0.14) while it is found to be significant when the normalization procedure is performed with respect to the lipid classes (by method 2, Fig. 2b, p-value=0.05).

We also simulated a different lipidomics data set where the lipids of a lipid class are transformed into the lipids from another lipid class. As in the data set used before, the simulated data set is a modified real data set already containing a distinct variation pattern. We used a four-block data set containing Ceramides, LycoPC, LycoPE and PA. The data set was modified by increasing all lipids in the lipid class Ceramides for Group 1 samples during the intervention period while lipids in the LycoPC class were decreased for these samples. Correlation loading plots for the two different normalization procedures are shown in Fig. 2c-d. Fig. 2d illustrates the results when the data is normalized by method 2 within the lipid classes. One can see that the total amount of Ceramides and LycoPC are found to be clearly changing (they are positioned near the outer circle) while the total amounts of lipids in the other two classes are not significantly changing. Since the total amount of Ceramides stands close to the Group 1 samples we can conclude that this lipid class is increased in this intervention group. The total amount of LycoPCs is on the top of the correlation plot, in the positive direction of component two, while the design factors for group 2 and 3 are in the positive direction of the first component. This separation can be explained by a variation pattern that was originally present in this data set, before the data were modified. What is striking is that the correlation plot in Fig. 2d suggests a remodelling in the Ceramides class, since the three lipids in this class are separated along the first component. Cer(d18:1/24:0) and Cer(d18:1/16:0) are now on the opposite side as Cer(d18:1/22:6). The reason for this is the normalization within each lipid class: Since each lipid class is normalized to contain the same amount of lipids, we now focus on relative changes among the lipids in each class, which is also present only if all lipids within one class increase accordingly, therefore no remodelling is detected.

We can conclude that relative minor remodelling shifts, already present in the input data, were only weakly evident after pre-processing method 1 (Fig. 2c). But they became much more evident after the normalization within each lipid class (pre-processing method 2).

When increasing all individual lipids in the Cermides class for Group 1, Fig. 2d might be misunderstood as indicating that the some of the LysoPC lipids were decreased in this class. This illustrates that when interpreting the correlation loading plots, the unit in which the variables are represented must always be kept in mind. In the present illustration, after pre-processing method 2 the total Cermides was given in one unit (here: micromol/liter), and the individual LysoPC lipids in another (fraction of the total LysoPC concentration). But incidentally, they happened to vary in similar ways here, which could be interpreted as a common causality behind both a metabolic between-class process and a within-class remodelling.

MBPLSR: An explorative supervised approach

The multi-block extension of PLSR (MBPLSR) is useful when lipidomics data are to be regressed on or compared explicitly to other types of variables, e.g. describing the experimental design. MBPLSR may be used as an explorative technique that tries to find the main patterns in the descriptor data that at the same time can predict the main patterns in the response data. But it may also be used for confirmative analysis, as a generalization of e.g. multi-response ANOVA. Thus the MBPLSR can be used in two different ways, depending on whether the lipidomics variables are used for predicting the other variables or vice versa. These will here be named MBPLSR-DA and ANOVA-MBPLSR, following the naming conventions in two-block PLSR [36].

MBPLSR-Discriminant Analysis (MBPLSR-DA). MBPLSR is called MBPLSR-DA when the pre-processed data of multi-block lipid variables are used as the descriptor variable set ($\mathbf{X}$) and a matrix containing the treatment grouping information is used as the response variable set ($\mathbf{Y}$). The MBPLSR-DA can then be used for revealing patterns and groupings in the lipidomic data (similar to CPCA), while at the same time investigating how these patterns relate to e.g. the design factors of the experiment.

The response data matrix (i.e. a matrix of 0s and 1s) is usually built in a way that one variable is assigned to every treatment group in the experiment. The samples belonging to the assigned group get a value of 1 while the rest of the samples get 0. Assuming that the N samples (people or people times time points) in the experiment are grouped into L different intervention groups, $\mathbf{Y}$ will be a matrix of size $N \times L$ of 1s and 0s. It is worth mentioning that we do not necessarily need to have as many variables in $\mathbf{Y}$ as the number of groups since the samples that do not belong to the previous $L-1$ groups are obviously belonging to the last one. Therefore a matrix of size $N \times (L-1)$ can equally be used in most of the situations. Still using one column for every group is recommended for graphical clarity, and does not pose any rank problems, since PLSR and MPLSR are designed to handle collinearity.

Similar to CPCA, the data sets (i.e. both $\mathbf{X}$ and $\mathbf{Y}$) should be pre-processed in terms of mean-centring, standardizing and (optionally) block-normalizing prior to the data MBPLSR modelling. The modelling is described briefly in Methods section while the detailed algorithms were described by us in [37].

Different parameters (e.g. global/block scores and loadings) are generated during the implementation of MBPLSR-DA on a lipidomics data set. The same powerful visualization tools are available as for CPCA. For instance, the grouping patterns that are common between different lipid classes are investigated by means of a global score plot. The grouping patterns within each lipid class then can be studied through corresponding block score plots. We generally expect to detect similar patterns as were seen by CPCA. However, the grouping patterns that are related to the treatment may be more emphasized, since PLSR is a supervised method, and patterns in the descriptor variables $\mathbf{Y}$ that have strong covariance with $\mathbf{X}$ will be favoured.

Similar to CPCA, validating the detected patterns is required.

Statistical validation. Validation may be performed by investigating the $RMSE_Y$ plots, calculating prediction error for each lipid class block [37]. The number of statistically valid PLS components must be determined. Like for ordinary PLSR, this may be attained by initially computing more components than conceivably necessary, and use cross-validation to determine the optimal model rank – which is normally defined as the number of PLS components from **X** that have clear predictive ability for **Y**. The cross-validation information is assessed block-wise to assess the importance of the different PLS components in the different blocks. In addition we suggest Monte Carlo permutation tests in order to estimate the significance of each lipid class for predicting each of the intervention groups. The detail for how to implement the Monte Carlo permutation tests for MBPLSR-DA models is given in Methods section.

As in CPCA, lipid-lipid interactions within the lipid classes as well as their relationship to each lipid class as a whole can be investigated by studying the correlation loading plots. Significance tests on regression coefficients of the MBPLSR-DA models are performed to estimate the significance of single lipids.

ANOVA-MBPLSR. The MBPLSR may also be used for analysis of variance with a high number of response variables. In order to perform ANOVA-MBPLSR, the variables with design information about the rows (e.g. the patients) are used as descriptor variables ($\mathbf{X}$) while the multi-block set of lipid variables in the different classes is used as the response variables ($\mathbf{Y} = \left[\mathbf{Y}^1,...,\mathbf{Y}^b,...,\mathbf{Y}^B\right]$). MBPLSR is then employed in order to make a linear, possibly reduced-rank predictive model where the design information in **X** is used for predicting the lipidomic variable set in **Y**. The detail for running MBPLSR when the response data ($\mathbf{Y}$) is a multi-block data set does not yet exist in the literature. Details are therefore given in the Methods section. The results of the analysis can be investigated through the global/block score plots for $\mathbf{Y}$ where the grouping patterns of the samples can be studied both between the lipid classes as well as within a lipid class. We propose to validate the model by studying the $RMSE_X$ plots. The detail for how to calculate $RMSE_X$ and generate the error plots when dealing with a multi-block response data set is given in the Methods section. The global/block $RMSE_X$ plots are plotted for each different factor of the background information separately. These plots validate the effect of each factor in the matrix containing group indicator variables on the detected grouping patterns. As before, we propose running Monte Carlo permutation tests and estimating the p-values for getting estimates for how strongly the design factors affected the lipid classes. Details are given in the Methods section.

Application example

In this section a lipidomics data set from an intervention study is used for demonstrating the above described methods. The intervention study was a double-blinded randomized controlled parallel-group study on healthy subjects where each person was assigned to one of the three intervention groups: (a) fish oil group, (b) oxidised fish oil group or (c) high-oleic sunflower oil group. Fifty people completed a fully controlled diet period of three weeks. The intervention study was previously described in [38]. The lipid profiles of subjects for the baseline and after the three weeks of intervention period were measured where 568 lipids were characterized and 260 were identified. The 568 lipids were split into the following 11 blocks of variables, representing the lipid classes and subclasses deemed most relevant for the present case [3]: Ceramides, lysophosphatidylcholines (lysoPC), lysophosphatidylethanolamines (lysoPE), phosphatidic acid (PA), phosphatidylcholines (PC), phosphatidylethanolamines (PE), phosphatidylglycerols (PG), phosphatidylserines, sphingomyelins (SM), triglycerides (TG). For illustration we show results of an integration of the classes, Ceramides, lysoPC, PC and TG where the classes contain 3, 12, 57 and 87 measured lipids respectively. In the following we will refer to the three intervention groups as group 1, group 2 and group 3, respectively

Data pre-processing

As described before, the lipids were measured at two different time points: once at the baseline and then after three weeks of intervention period. The data blocks of lipid classes (both for baseline and after three weeks) were first pre-processed according to the second normalization procedure described in the Results section. Secondly, in order to correct for baseline effects, log2 ratios of the data from two visits were calculated. This is done because the fold changes of the lipids after the intervention period are often of more interest than the original measured concentrations at every visit. This enabled us to investigate the proportional effect of the intervention diet on the lipidomic profile regardless of the baseline amounts of lipids. The structure of the pre-processed data is shown in Fig. 3b and will be used for CPCA and MBPLSR data modelling:

CPCA modelling

The grouping patterns of the samples are studied by block/global score plots of CPCA. Fig. 4 illustrates the block and global score plots for the CPCA model. The grouping pattern of the subjects that are in common between different lipid classes (i.e. the global pattern of all the lipids in the analysis) is seen in the global score plot (Fig. 4f). The global score plot in Fig. 4f shows a clear separation between the intervention group 1 and the other two intervention groups along the first principal component. A similar pattern as is seen in the global score plot is also detected in some of the lipid classes e.g. lycoPC, PC and TG while the other two blocks (i.e. Ceramides and Sum Lipids) do not show clear grouping patterns. Higher components did not show any informative groupings (results are not shown).

In order to validate the detected patterns, RMSEs for each lipid class are estimated as described in the methods section. RMSEs for the different lipid classes are shown in Fig. 5a as a function of the number of components included in the model. Inspecting the RSME plots we can see that blocks 2, 3 and 4 (i.e. lycoPC, PC and TG) are contributing mostly to the global patterns as shown in Fig. 4f. Validated explained variances are shown in Fig. 5b, revealing that validated explained variances for the Ceramides and “Sum Lipids” are negative for the first component indicating that these blocks are not contributing to the separation of the intervention groups. One can see that Ceramides are contributing considerably to the second component, however the intervention groups were only separated with respect to the first component and therefore this class is not relevant for the separation of the intervention groups.

Further we estimated p-values for the significance of the contributions of each block to the global pattern by Monte Carlo permutation tests. As an example, the calculation of the p-value for Ceramides data block for a CPCA model containing one component is described in the following: First, $RMSE_{A=1}^{b=1}$ is calculated for the multi-block data set (we call it true $RMSE_{A=1}^{b=1}$ here). Then, 1000 permutations are run (according to the

Methods section) and $RMSE_{A=1}^{b=1}$ is calculated in each permutation run. The p-value for the test is given by dividing the number of the permutations whose calculated $RMSE_{A=1}^{b=1}$s are smaller than the true $RMSE_{A=1}^{b=1}$ by the total number of permutations (i.e. 1000 here). The calculated p-values for the lipid classes are as following: Ceramides (0.996), lycoPC (0.000), PC (0.000), TG (0.007) and Sum Lipids (0.672), i.e. the lycoPC, PC and TG class are considered to be highly significant We notice that group 2 and group 3 are not well separated in the CPCA model.

MBPLSR-DA

When analyzing the data by MBPLSR-DA the pre-processed, mean-centred and scaled five-block lipidomic data set is used as descriptor data ($\mathbf{X} = \left[\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, \mathbf{X}^4, \mathbf{X}^5\right]$ where $\mathbf{X}^1,...,\mathbf{X}^5$ are Ceramides, lycoPC, PC, TG and Sum Lipids respectively) and the intervention groups of the samples are used as the response data block $\mathbf{Y}$ of size (48×3). Block and global score plots for the MBPLSR-DA model are plotted in Fig. 6. The global score plot (Fig. 6f) shows a distinct grouping for the samples belonging to group 1. One can see that the grouping here is more clear compare to Fig. 6f. Similar to the results from CPCA, Ceramides and the Sum Lipids do not show a clear grouping pattern. However, the other three classes show clear grouping patterns. Here even a separation between group 2 and 3 is visible, especially for the lipid class PC.

The correlation loading plot for the MBPLSR-DA model is shown in Fig. 7. This plot shows that the first component is able to separate the group 1 samples from the other samples pretty well. We also can see that group 2 and 3 are fairly separated from each other by the second component. The lipids that are significantly contributing to the separation between the different groups: lycoPC(22:6), lycoPC(20:5), PC(36:5), PC(40:6) and several TGs (e.g. TG(56:9)) are contributing to the separation of the group 1 from the other intervention groups. All of these lipids are located close to each other leading to the conclusion that they increase and decrease together. Three lipids of the PC class (i.e. PC(40:4e), PC(37:4)/PE(40:4) and PC(38:5e)) are located on the opposite side of the PC(36:5) showing that an increase of these three lipids leads to the decrease of PC(36:5) and vice versa. PC(36:5) is located on the same side of the plot as the intervention group 1. Therefore, the amount of this lipid in the subjects belonging to the intervention group 1 is high while the amount of the three lipids which are located in the opposite direction (i.e. PC(40:4e), PC(37:4)/PE(40:4) and PC(38:5e)) is low in group 1.

The detected grouping patterns are validated by studying the error plots for every intervention group. These plots validate the ability of the lipid classes discriminating the respective intervention group from the other groups. Three RMSE_Y plots for the three intervention groups are shown in Fig. 8. Comparing the plots with each other, one can see that the error plots for intervention group 1 reduces much more than the other two groups. Therefore, the separation of group 1 is much more significant compare to that for the other groups. Fig. 8a illustrates that this separation can be detected just by the first component and lycoPC, PC and TG are responsible lipid classes for this grouping. The RMSE_Y plots for Ceramides and Sum Lipids (blocks 1 and 5 in Fig. 8a-c) show that these lipid classes do not contribute to the separation of any of the intervention groups. Fig. 8b illustrates also an important aspect of PC since it shows that PC can distinguish the group 2 samples (by the second component) from the other intervention groups pretty well compared to the other lipid classes.

The contribution of different lipid classes to the MBPLSR-DA model can be further examined by studying the explained variances (for details see methods section). The validated explained variances are plotted in Fig. 9. The validated explained variance for the global model for separating group 1 samples (Fig. 9a) by the first component is more than 80%. Therefore, the global MBPLSR-DA model can separate the group 1 samples from the rest by means of the first component to a large extent. LycoPC, PC and TG are the lipid classes which are contributing to this separation (by explaining almost 80%, 70% and 60% variances) while Ceramides and Sum Lipids are not contributing at all (these two class are showing negative validated explained variance). Fig. 9b and Fig. 9c show that lysoPC, PC and TG are also the lipid classes that contribute most to the separation of both groups 2 and 3. However, this separation is week, since the validated explained variances in **Y** less than 20% for allof these classes. Fig. 8 and Fig. 9 show that there is a significant difference between the separation of group 1 samples compared to the other two groups.

In order to estimate if blocks are contributing significantly to the discrimination of group 1 from the other groups, we ran Monte Carlo permutation tests (1000 permutation runs) and obtained p-values for each lipid class. The calculated p-values for the lipid classes for the separation of group 1 for a 1-component model are as follows: Ceramides (0.328), lycoPC (0.000), PC (0.000), TG (0.000) and Sum Lipids

(0.453). The p-values confirm that the lycoPC, PC and TG lipid classes are clearly separating group 1 from the other groups by the first component.

Table 1 shows the p-values for the significant lipids for the separation of group 1 samples obtained by a significance test on the regression as explained in the method section.

The fact that the Sum Lipids data does not show any significant contribution to the grouping patterns shows that the total amount of the lipids in the lipid classes is not changing significantly during the intervention. However, few lipids are identified that are significantly contributing to the grouping patterns meaning that these lipid species are changing significantly during the intervention (Table 1). Therefore, we can conclude that the intervention resulted in the remodelling of these detected lipids.

Including background information

In order to include additional background information we use ANOVA-MBPLSR for investigating the effect of the collected background information on the detected patterns in the data. For the present lipidomic data set we have three additional background variables: gender, age, BMI. As described in the Results section, we use the intervention grouping information and these three variables as the descriptor data (one data matrix) while the lipid classes are used for the response data set (five data matrices). This means that we estimate by ANOVA-MBPLSR modelling if grouping and background information can predict the lipid profile. The global and block score plots for the first and second components are shown in Fig. 10. We see that the global score plot separated group 1 samples nicely from the rest of the samples. As before, PC, lycoPC and TG are the lipid classes that are contributing mostly to this separation. The score plots of the third and fourth components did not reveal any clear grouping patterns (plots are not shown here).

In order to investigating the effect of the background information on the lipid profile error plots are studied. The error plots for a 2-component model are shown in this Fig 11. BMI is the factor which is contributing mostly followed by Age. Gender does not seem to have any effect at all. P-values for estimating the contribution of the BMI and Age to the grouping pattern (i.e. the grouping pattern in Fig. 10f) were calculated by running Monte Carlo simulations. The calculated p-values were 0.178 and 0.180 for BMI and Age, respectively, again showing that BMI and Age do not have a significant effect on the lipid profile.

Software

The above mentioned steps of pre-processing and multivariate analyses were performed using in-house-written and standard MATLAB routines (MATLAB Version 7.8). The MATLAB routines for the whole framework are available at http://arken.umb.no/~achik/algorithms.html.

Conclusions

Due to the fast development of the lipidomics field there is a growing need for statistical methods that can integrate, analyze, understand and interpret such massive data sets. In this study we have presented a strategy for analyzing lipidomics data. The strategy includes a pre-processing process for the lipidomics data sets that reveals remodelling within lipid classes and lipid dynamics between lipid classes in the subsequent data modelling. By applying two different pre-processing strategies, remodelling within lipid classes and lipid dynamics between lipid classes could be clearly identified. The pre-processing strategies are based on presenting the data in different units, where either all data is presented in the same basic unit, to keep the over-all mass balance, or to normalize each lipid class relative to the total amount of lipids within this class. The data modelling strategy presented keeps the lipid class and lipid species structure in the data modelling process and brings forward the interplay and inter-correlations between the many lipidomics variables and their relation to other variables such as background variables, design variables and other omics measurement.

Methods

Pre-processing of data with respect to the lipid classes

Assuming that $\mathbf{X}_{Raw}$ (of size $N \times K$) is the raw lipidomic data set for N samples and K variables (i.e. measured lipids). The first step is to split data into B different data blocks with respect to the lipid classes as:

$$\mathbf{X}_{Raw} = [\mathbf{X}_{Raw}^{1}, \mathbf{X}_{Raw}^{2}, \ldots, \mathbf{X}_{Raw}^{b}, \ldots, \mathbf{X}_{Raw}^{B}] \quad (1)$$

where $\mathbf{X}_{Raw}$ is the raw lipidomic data and $\mathbf{X}_{Raw}^{b}$ (of size $N \times K_b$) consists of the lipids in *b*th class. The second step is to replace the amount of each lipid in every data block by its relative amount within its respective class. This is done by dividing each data block by the total amount of lipids it contains, according to the following:

$$\mathbf{X}^{b}_{\text{Preprocessed}}(i,:) = \frac{\mathbf{X}^{b}_{\text{Raw}}(i,:)}{\sum_{k=1}^{K_b} \mathbf{X}^{b}_{\text{Raw}}(i,k)} \qquad (2)$$

where $i = 1,...,N$ indicates the samples, $k = 1,...,K_b$ indicates the variables (i.e. lipids), $\mathbf{X}^{b}_{\text{Raw}}(i,k)$ stands for the (*i*,*k*)th entry of data block $\mathbf{X}^{b}_{\text{Raw}}$ and $\mathbf{X}^{b}_{\text{Preprocessed}}(i,:)$ and $\mathbf{X}^{b}_{\text{Raw}}(i,:)$ are the *i*th row of the data block *b* of the pre-processed data and raw data respectively.

Mean-centring

Mean-centring is an essential step which is usually performed on data blocks prior to CPCA and MBPLSR. Mean-centring is performed by reducing the mean of the variables over all samples according to

$$\mathbf{X}^{b}_{\text{Mean-centred}} = \mathbf{X}^{b}_{\text{Preprocessed}} - \mathbf{1}\cdot\bar{\mathbf{x}}^{b}_{\text{Preprocessed}}{}' \qquad (3)$$

where $\mathbf{X}^{b}_{\text{Preprocessed}}$ is the pre-processed data set calculated by Eq. 2, $\bar{\mathbf{x}}^{b}_{\text{Preprocessed}}$ is the vector of means of the variables in block *b* over all samples and $\mathbf{X}^{b}_{\text{Mean-centred}}$ is the resulting mean-centred data block. Eq. 3 is repeated for $b = 1,2,...,B$.

Mean-centring for response data matrix $\mathbf{Y}$ is performed in a similar way, according to:

$$\mathbf{Y}_{\text{Mean-centred}} = \mathbf{Y}_{\text{Raw}} - \mathbf{1}\cdot\bar{\mathbf{y}}' \qquad (4)$$

where $\mathbf{Y}_{\text{Mean-centred}}$ is the mean-centred data matrix, $\mathbf{Y}_{\text{Raw}}$ is the raw data matrix and $\bar{\mathbf{y}}$ is the mean vector of the variables in $\mathbf{Y}_{\text{Raw}}$ calculated over all samples.

Scaling

All the input variables may be first standardized to a total initial variance of 1, if the different lipids within each block have very different initial variance (this was not deemed necessary in the present dataset). Then the different blocks of variables were scaled block-wise, to ensure equal block sum of squares, since the data blocks are different with respect to their number of variables and also their measurement units. A neutral scaling can be performed by dividing each data block by its Frobenius norm according to

$$\mathbf{X}^b = \frac{\mathbf{X}^b_{\text{Mean-centred}}}{\sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K_b}(\mathbf{X}^b_{\text{Mean-centred}}(i,k))^2}} \qquad (5)$$

where $\mathbf{X}^b_{\text{Mean-centred}}$ is data block calculated by Eq. 3, $\mathbf{X}^b_{\text{Mean-centred}}(i,k)$ is the (i,k)th entry of data block $\mathbf{X}^b_{\text{Mean-centred}}$ and $\mathbf{X}^b$ is the lipidomic data block that is pre-processed, mean-centred and scaled and it ready to be used by the analysis techniques e.g. CPCA and MBPLSR.

When running MBPLSR, the response data matrix should also be scaled in order to be on the same footing as the descriptor data blocks. This can be done in the same as scaling was performed for data blocks in $\mathbf{X}$, according to:

$$\mathbf{Y} = \frac{\mathbf{Y}_{\text{Mean-centred}}}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{J}(\mathbf{Y}_{\text{Mean-centred}}(i,j))^2}} \qquad (6)$$

where $\mathbf{Y}_{\text{Mean-centred}}$ is the mean-centred response data matrix calculated by Eq. 4, $\mathbf{Y}_{\text{Mean-centred}}(i,j)$ is the (i,j)th entry of data matrix $\mathbf{Y}_{\text{Mean-centred}}$ and $\mathbf{Y}$ is the mean-centred and scaled response data matrix.

Data modelling by CPCA

It is important to know that the multi-block data set should be mean-centred and scaled prior to CPCA. CPCA models the multi-block data set as sums of *A* Principal Components (PCs) plus residual matrices according to:

$$\begin{aligned}\mathbf{X} &= \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A \\ \mathbf{X}^b &= \mathbf{T}^b_A\mathbf{P}^{b\prime}_A + \mathbf{E}^b_A\end{aligned} \qquad (7)$$

where $\mathbf{X} = \left[\mathbf{X}^1\mathbf{X}^2...\mathbf{X}^b...\mathbf{X}^B\right]$ is the multi-block mean-centred and scaled data set where the blocks are calculated by Eq. 3, $\mathbf{T}_A$ and $\mathbf{T}^b_A$ are global and block scores respectively, $\mathbf{P}_A$ and $\mathbf{P}^b_A$ are global and block loadings respectively, $\mathbf{E}_A$ and $\mathbf{E}^b_A$ are the global and block residual matrices for an *A*-PC model. Nonlinear Iterative Partial Least Squares (NIPALS) [39, 40] is commonly used for the calculation of the CPCA parameters. The detailed algorithm of NIPALS for CPCA may be found in [41].

Data modelling by MBPLSR-DA

Similar to CPCA, the data blocks in $\mathbf{X}$ (so called the descriptor data set) and the data matrix $\mathbf{Y}$ (so called the response data set) should be mean-centred and scaled prior to the analysis. The predictive MBPLSR model is as the following:

$$\mathbf{Y} = \mathbf{X}\mathbf{B}_A + \mathbf{F}_A \quad (8)$$

where $\mathbf{X} = \left[\mathbf{X}^1\mathbf{X}^2...\mathbf{X}^b...\mathbf{X}^B\right]$ is the multi-block descriptor data set which is mean-centred and scaled, $\mathbf{Y}$ is the mean-centred and scaled response data matrix, $\mathbf{B}_A$ is the regression coefficients for a model including *A*-PCs and $\mathbf{F}_A$ is the residual matrix for the corresponding model. Moreover, employing MBPLSR-DA models the data sets $\mathbf{X}$ and $\mathbf{Y}$ as:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A \\ \mathbf{X}^b &= \mathbf{T}_A^b\mathbf{P}_A^{b\prime} + \mathbf{E}_A^b \qquad (9) \\ \mathbf{Y} &= \mathbf{T}_A\mathbf{Q}'_A + \mathbf{F}_A \end{aligned}$$

where $\mathbf{T}_A$ and $\mathbf{T}_A^b$ are the global scores and X-block scores respectively, $\mathbf{P}_A$ and $\mathbf{P}_A^b$ are global loadings and X-block loadings respectively, $\mathbf{Q}_A$ is the loadings of $\mathbf{Y}$ and $\mathbf{E}_A = \left[\mathbf{E}_A^1...\mathbf{E}_A^b...\mathbf{E}_A^B\right]$ and $\mathbf{F}_A$ are the residuals of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

Several methods for calculating the parameters of an MBPLSR-model can be found. The MBPLSR algorithm of Wangen and Kowalski [42] which handles most types of the relationships between the data blocks was described by us in detail in [37].

Data modelling by ANOVA-MBPLSR

In the ANOVA-MBPLSR the grouping or design information is used as the descriptor data block ($\mathbf{X}$) and the measured variables are used as response data ($\mathbf{Y}$). Ordering the lipidomic data set as a multi-block data leads to the situation where the response data set is a multi-block data. In order to get block parameters (i.e. block scores and loadings) as we had in the previous analysis, we need to modify the usual MBPLSR algorithm in a way that we are able to calculate block parameters for $\mathbf{Y}$. The MBPLSR model when the response data set ($\mathbf{Y}$) is a multi-block set is as the following:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{B}_A + \mathbf{F}_A \\ \mathbf{X} &= \mathbf{T}_A\mathbf{P}'_A + \mathbf{E}_A \qquad (10) \\ \mathbf{Y} &= \mathbf{T}_A\mathbf{Q}'_A + \mathbf{F}_A \end{aligned}$$

While the global scores ($\mathbf{T}_A$) from **X** form the basis for the data modelling in PLSR and MBPLSR, the parameters Y-scores ($\mathbf{U}_A$) may also be estimate [37]; they may be used for studying the grouping patterns in the response data set.

There are several equivalent algorithms to estimate the parameters in the PLSR, and hence also in its multi-block version. Here we employ a multi-block extension of the NIPALS algorithm: In cases where the response data is a multi-block set, we calculate block scores for $\mathbf{Y}$ from the Y-scores ($\mathbf{u}_A$). In order to calculate the corresponding parameters, the following procedure is performed for each PLSR component ($a = 1, 2, \ldots$):

A. Initializing

1.1 Choose an arbitrary starting **Y** score vector, $\mathbf{u}$

B. Computing the scores and loading weights

1.2 $\tilde{\mathbf{w}} = \dfrac{\mathbf{X}'\mathbf{u}}{\mathbf{u}'\mathbf{u}}$ **X** loading weights

1.3 $\mathbf{w} = \tilde{\mathbf{w}}(\sqrt{\tilde{\mathbf{w}}'\tilde{\mathbf{w}}})^{-1}$

1.4 $\mathbf{t} = \mathbf{X}\mathbf{w}(\mathbf{w}'\mathbf{w})^{-1}$ **X** scores

1.5 $\mathbf{q} = \dfrac{\mathbf{Y}'\mathbf{t}}{\mathbf{t}'\mathbf{t}}$ **Y** loading weights

1.6 $\mathbf{u} = \dfrac{\mathbf{Y}\mathbf{q}}{\mathbf{q}'\mathbf{q}}$ **Y** scores

1.7 $\mathbf{q} = \left[\mathbf{q}^1 \ldots \mathbf{q}^b \ldots \mathbf{q}^B\right]$ Partitioning **Y** loading weights into respective segments for every block in **Y**

1.8 $\mathbf{u}^b = \dfrac{\mathbf{Y}^b\mathbf{q}^b}{\mathbf{q}^{b\prime}\mathbf{q}^b}$ **Y** block scores

C. Replacing the **Y** score vector **u** by the updated vector of **Y** scores in 1.6 and iterating until convergence (i.e. no significant change in scores **t).**

D. Deflating the data on global scores

1.9 $\mathbf{p}_a = \dfrac{\mathbf{X}'\mathbf{t}}{\mathbf{t}'\mathbf{t}}$ **X** loadings

1.10 $\mathbf{X}_a = \mathbf{X}$ and $\mathbf{Y}_a = \mathbf{Y}$

1.11 $\mathbf{X} = \mathbf{X}_a - \mathbf{t}\mathbf{p}_a'$ **X** deflation

1.12 $\mathbf{Y} = \mathbf{Y}_a - \mathbf{t}\mathbf{q}'$ **Y** deflation

Monte Carlo permutation tests

Permutation tests in CPCA

Here, we propose a method for investigating the contribution of every data block (i.e. every lipid class) to the CPCA model in order to have an estimation of the importance of the role of each block (i.e. lipid class) for modelling the lipidomic data. For this purpose we generate a reference distribution for $RMSE_A^b$ by Monte Carlo sampling [43, 44]. Monte Carlo sampling is performed by permuting the samples for the given block and calculating the RMSE for that block for a given number of PC (i.e. $RMSE_A^b$). The choice for the number of PCs is made based on the RMSE plots calculated for the multi-block data set prior to performing any permutations. Calculation of the RMSE and the respective plots were described by us in detail in [32]. Here, we describe how to perform the permutation tests for the given block $\mathbf{X}^b$, assuming that $\mathbf{X}^b$ is an $N \times K_b$ data matrix where each row corresponds to a sample (i.e. a subject in the experiment) and each column corresponds to a variable (i.e. a measured lipid). $\mathbf{X}^b$ can be written as following:

$$\mathbf{X}^b = \left[\mathbf{x}_1^b, ..., \mathbf{x}_i^b, ..., \mathbf{x}_N^b\right]' \qquad (11)$$

where $i = 1, 2, ..., N$ are the row numbers of the data matrix and $\mathbf{x}_i^b$ (a vector of size $1 \times K_b$) is the *i*-th row in the data block $\mathbf{X}^b$.

Firstly, we calculate the $RMSE_A^b$ for block *b* in the multi-block data set $\mathbf{X} = \left[\mathbf{X}^1, ..., \mathbf{X}^b, ..., \mathbf{X}^B\right]$ (we call this $RMSE_A^b$: observed $RMSE_A^b$). Secondly, we randomly permute the rows in the data block $\mathbf{X}^b$ which results in different combination of the rows e.g.:

$$\mathbf{X}^b = \left[\mathbf{x}_e^b, ..., \mathbf{x}_f^b, ..., \mathbf{x}_g^b\right]' \qquad (12)$$

where $1 \le e, f, g \le N$. $RMSE_A^b$ is calculated for $\mathbf{X} = \left[\mathbf{X}^1, ..., \mathbf{X}^b, ..., \mathbf{X}^B\right]$ every time the rows of the data block *b* are permuted. Calculating $RMSE_A^b$ for every permutation round leads to the generation of different $RMSE_A^b$ which can then be put together in order to generate a distribution for $RMSE_A^b$ for the given block for the given number of PCs. A p-value is then assigned to the given block b by testing the observed $RMSE_A^b$ in the generated distribution.

Permutation tests in MBPLSR-DA

Similar to CPCA, we propose to perform Monte Carlo permutations and calculate p-values for the contribution of each block to the separation of the groups. Observed $RMSE_{Y}{}_A^b$ is firstly calculated for the data set prior to the permutations. The details for how to calculate $RMSE_{Y}{}_A^b$ was described by us in [37]. Each block is then permuted in the same way that it was described for CPCA in the previous section. $RMSE_{Y}{}_A^b$ is calculated in every permutation round which results in generating a reference distribution for $RMSE_{Y}{}_A^b$ by Monte Carlo sampling for every block having the chosen number of PCs in the MBPLSR model. Finally, a p-value for the observed $RMSE_{Y}{}_A^b$ in the generated distribution is calculated.

Permutation tests in ANOVA-MBPLSR

The procedure of running the Monte Carlo permutation tests and calculating the p-values for an ANOVA-MBPLSR model is very similar to that for the MBPLSR-DA. Observed $RMSE_{X}{}_A^g$ and $RMSE_{X}{}_A^b$s are calculated prior to the permutations (the calculation details are described in the respective section). Thereafter, reference distributions for the contribution of each background factor to the ANOVA-MBPLSR model are generated. This is done by permuting the values in the respective factor and calculating $RMSE_{X}{}_A^g$ and $RMSE_{X}{}_A^b$ for every round. Finally, p-values for the observed $RMSE_{X}{}_A^g$ and $RMSE_{X}{}_A^b$s are calculated with respect to the generated distribution.

Uncertainty t-tests for MBPLSR

Assessing the contribution of the variables (i.e. lipids) to the MBPLSR model can be done by running uncertainty t-tests on the regression coefficients of the MBPLSR model. For this purpose we run cross-validation and re-calculate the regression coefficients in every cross-validation round. Firstly, both $\mathbf{X}$ and $\mathbf{Y}$ are divided into M segments resulting in leave-in data segments (i.e. $\mathbf{X}_{-m}$ and $\mathbf{Y}_{-m}$) and left-out data segments (i.e. $\mathbf{X}_m$ and $\mathbf{Y}_m$) for $m = 1, 2, ..., M$. Regression coefficients ($\mathbf{B}_{-m,A}$) are calculated for the MBPLSR models of the leave-in data segments (for the given number of components *A*). We then calculate the Jack-knife estimate for the standard deviation of the regression coefficients according to

$$s_{k,j,A} = \sqrt{\left(\frac{M-1}{M}\right)\sum_{m=1}^{M}\left(b_{k,j,A} - b_{-m,(k,j,A)}\right)^{2}} \qquad (13)$$

where $b_{k,j,A}$ is the (*k,j*)-th entry of the matrix of the regression coefficients ($\mathbf{B}_A$ of size ($K \times J$)) for an *A*-component MBPLSR model of $\mathbf{X}$ and $\mathbf{Y}$, $b_{-m,(k,j,A)}$ is the (*k,j*)-th entry of the regression coefficient matrix ($\mathbf{B}_{-m,A}$ of size ($K \times J$)) for an *A*-component MBPLSR model of $\mathbf{X}_{-m}$ and $\mathbf{Y}_{-m}$, *M* is the number of cross-validation segments and $s_{k,j,A}$ is the standard deviation for the regression coefficients having *A* components in the MBPLSR model.

The t-statistic is then calculated according to

$$t_{k,j,A} = \frac{b_{k,j,A}}{s_{k,j,A}} \qquad (14)$$

where $b_{k,j,A}$ is the (*k,j*)-th entry of the matrix of the regression coefficients for an *A*-component MBPLSR model and $s_{k,j,A}$ is the standard deviation for the regression coefficients calculated by Eq. 13. From the t-statistic in Eq. 14, p-values are calculated which shows how significant variable *k* in $\mathbf{X}$ is for predicting variable *j* in $\mathbf{Y}$ by an *A*-component MBPLSR model. It is worth mentioning that although we are testing a large number of variables the calculated p-values do not need any correction such as Bonferroni. The reason is that using the *A*-component MBPLSR model reduces the number of original variables (*K*) to *A* latent variables. Since we are testing the MBPLSR model we are in fact implementing tests on few latent variables which do not require any correction.

RMSE$_X$ calculations for ANOVA-MBPLSR

For assessing the predictability of an ANOVA-MBPLSR model, we suggest a cross-validation based method: First, the data (i.e. both $\mathbf{X}$ and $\mathbf{Y}$) are divided into leave-in segments (i.e. $\mathbf{X}_{-m}$ and $\mathbf{Y}_{-m}$) and leave-out segments (i.e. $\mathbf{X}_m$ and $\mathbf{Y}_m$). ANOVA-MBPLSR models are built for the leave-in data. The models are then used for predicting the descriptor leave-out data segments ($\hat{\mathbf{X}}_m$). The differences between the true values ($\mathbf{X}_m$) and the predicted ones ($\hat{\mathbf{X}}_m$) gives the error for the ANOVA-

MBPLSR model for the given number of components. The errors are then studied by plotting $RMSE_X$ plots. We have previously described the calculations of $RMSE_X$ for the case where the descriptor data ($\mathbf{X}$) was a multi-block data set in [37]. Now, we have a different situation where the response data set ($\mathbf{Y}$) is multi-block. Here, the calculation of the global errors ($RMSE_X^g$) and block errors ($RMSE_X^b$) is done in a similar way to what we described in that article. We do not explain all details here; instead we mention the modifications that need to be implemented on the previously described method in order to calculate $RMSE_X$ for a multi-block response data.

The global errors $RMSE_X^g$ are calculated according to the method described in Section 2.3.1.1 in [37]. One should just note that the descriptor data set ($\mathbf{X}$) contains only one data block. This does not affect the description of the method.

The block errors $RMSE_X^b$ are also calculated by the same method described in Section 2.3.1.1 in [37]. However, some modifications are necessary for the calculations here: the loading weights ($\tilde{\mathbf{W}}_{-m,A}$) that were used in Eq. 7 and Eq. 8 (in [37]) are here replaced by the block loading weights ($\tilde{\mathbf{W}}^b_{-m,A}$). The block Y-scores ($\mathbf{u}^b$) (calculated in step 1.8 of the Data modelling by ANOVA-MBPLSR section) are contributing to the calculation of the block loading weights ($\tilde{\mathbf{W}}^b_{-m,A}$) according to the following Eq.:

$$\tilde{\mathbf{w}}^b_{-m,a} = \frac{\mathbf{X}'_{-m}\mathbf{u}^b_{-m,a}}{\mathbf{u}^b_{-m,a}{}'\mathbf{u}^b_{-m,a}} \qquad (15)$$

The rest of the calculations are similar to those for global $RMSE_X^g$.

List of abbreviations

PCA: Principal Component Analysis; PLSR: Partial Least Squares Regression; CPCA: Consensus PCA; RMSE: Root Mean Squared Errors; MBPLSR: Multi-block PLSR; MBPLSR-DA: MBPLSR-Discriminant Analysis; ANOVA: Analysis of Variance; BMI: Body Mass Index; lysoPC: lysophosphatidylcholines; PC: phosphatidylcholines; TG: triglycerides;

Authors' contributions

All authors contributed to the development of the ideas. SH and AK contributed to the development of the methods, SH wrote the Matlab codes and performed the data analysis, AK debugged the Matlab codes, SH wrote the paper, HM and AK revised the manuscript critically, all authors contributed to and approved the final manuscript.

Acknowledgements

The authors are grateful for financial support by the Nordic Centre of Excellence on Food, Nutrition and Health “Systems biology in controlled dietary interventions and cohort studies” (SYSDIET) funded by NordForsk, Foundation for Research Levy on Agricultural Products in Norway and the grant 203699 (New statistical tools for integrating and exploiting complex genomic and phenotypic data sets) financed by the Research Council of Norway.

References

1. Little SJ, Lynch MA, Manku M, Nicolaou A: **Docosahexaenoic acid-induced changes in phospholipids in cortex of young and aged rats: A lipidomic analysis**. *Prostaglandins, Leukotrienes and Essential Fatty Acids* 2007, **77**(3-4):155-162.
2. M Oi: **Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction**. *Nutrition, Metabolism and Cardiovascular Diseases* 2009, **19**(11):816-824.
3. Ottestad I, Hassani S, Borge GI, Kohler A, Vogt G, Hyötyläinen T, Orešič M, Brønner KW, Holven KB, Ulven SM *et al*: **Fish Oil Supplementation Alters the Plasma Lipidomic Profile and Increases Long-Chain PUFAs of Phospholipids and Triglycerides in Healthy Subjects**. *submitted to PloSOne*.
4. Hunt AN, Macken M, Koster G, Kohler JA, Postle AD: **Diclofenac mediated derangement of neuroblastoma cell lipidomic profiles is accompanied by increased phosphatidylcholine biosynthesis**. *Advances in Enzyme Regulation* 2008, **48**(1):74-87.
5. Meikle P, Tsorotes D, Barlow C, Weir J, MacIntosh G, Barber M, Goudey B, Bedo J, Stern L, Kowalczyk A *et al*: **P36 PLASMA LIPIDOMIC ANALYSIS OF STABLE AND UNSTABLE CORONARY ARTERY DISEASE**. *Atherosclerosis Supplements*, **11**(2):24.
6. Lankinen M, Schwab U, Gopalacharyulu PV, SeppÃ¤nen-Laakso T, Yetukuri L, Sysi-Aho M, Kallio P, Suortti T, Laaksonen DE, Gylling H *et al*: **Dietary carbohydrate modification alters serum metabolic profiles in individuals with the metabolic syndrome**. *Nutrition, Metabolism and Cardiovascular Diseases*, **20**(4):249-257.
7. Kelishadi R, Pour MH, Zadegan NS, Kahbazi M, Sadry G, Amani A, Ansari R, Alikhassy H, Bashardoust N: **Dietary fat intake and lipid profiles of Iranian adolescents: Isfahan Healthy Heart Programâ€"Heart Health Promotion from Childhood**. *Preventive Medicine* 2004, **39**(4):760-766.
8. Laclaustra M, Stranges S, Navas-Acien A, Ordovas JM, Guallar E: **Serum selenium and serum lipids in US adults: National Health and Nutrition Examination Survey (NHANES) 2003â€"2004**. *Atherosclerosis* 2010, **210**(2):643-648.
9. Jorde R, Grimnes G: **Vitamin D and metabolic health with special reference to the effect of vitamin D on serum lipids**. *Progress in Lipid Research* 2011, **50**(4):303-312.
10. Peter M, Christopher B, Jacqui W: **Lipidomics and Lipid Biomarker Discovery**. *Australian Biochemist* 2009, **40**(3):12-16.
11. Yetukuri LR: **Bioinformatics approaches for the analysis of lipidomics data**. *PhD dissertation*. The Aalto University School of Science and Technology; 2010.
12. Niemelä PS, Castillo S, Sysi-Aho M, Orešič M: **Bioinformatics and computational methods for lipidomics**. *Journal of Chromatography B* 2009, **877**(26):2855-2862.
13. Matej O: **Informatics and computational strategies for the study of lipids**. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2011, **1811**(11):991-999.

14. Wenk MR: **Lipidomics: New Tools and Applications**. *Cell* 2010, **143**(6):888-895.

15. Navas-Iglesias N, Carrasco-Pancorbo A, Cuadros-Rodríguez L: **From lipids analysis towards lipidomics, a new challenge for the analytical chemistry of the 21st century. Part II: Analytical lipidomics**. *TrAC Trends in Analytical Chemistry* 2009, **28**(4):393-403.

16. Holm S: **A Simple Sequentially Rejective Multiple Test Procedure**. *Scandinavian Journal of Statistics* 1979, **6**(2):65-70.

17. Hochberg Y: **A sharper Bonferroni procedure for multiple tests of significance**. *Biometrika* 1988, **75**(4):800-802.

18. Pearson K: **On lines and planes of closest fit to systems of points in space**. *Philosophical Magazine* 1901, **2**(6):559-572.

19. Balogh G, Péter M, Liebisch G, HorvÃ¡th I, Török Z, Nagy E, Maslyanko A, Benkő S, Schmitz G, Harwood JL *et al*: **Lipidomics reveals membrane lipid remodelling and release of potential lipid mediators during early stress responses in a murine melanoma cell line**. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2010, **1801**(9):1036-1047.

20. Yang S, Lu S-H, Yuan Y-J: **Lipidomic analysis reveals differential defense responses of Taxus cuspidata cells to two elicitors, methyl jasmonate and cerium (Ce4+)**. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 2008, **1781**(3):123-134.

21. Wetzel DL, Reynolds Iii JE, Sprinkel JM, Schwacke L, Mercurio P, Rommel SA: **Fatty acid profiles as a potential lipidomic biomarker of exposure to brevetoxin for endangered Florida manatees (Trichechus manatus latirostris)**. *Science of The Total Environment* 2010, **408**(24):6124-6133.

22. Fernando H, Bhopale KK, Kondraganti S, Kaphalia BS, Shakeel Ansari GA: **Lipidomic changes in rat liver after long-term exposure to ethanol**. *Toxicology and Applied Pharmacology* 2011, **255**(2):127-137.

23. Wold S, Martens H, Wold H: **The multivariate calibration problem in chemistry solved by the PLS method**. In: *Matrix Pencils.* Edited by Kågström B, Ruhe A, vol. 973: Springer Berlin / Heidelberg; 1983: 286-293.

24. Thissen U, Wopereis S, van den Berg S, Bobeldijk I, Kleemann R, Kooistra T, Willems van Dijk K, van Ommen B, Smilde AK: **Improving the analysis of designed studies by combining statistical modelling with study design information**. *BMC Bioinformatics* 2009, **10**(1):52.

25. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, van Ommen B, Smilde AK: **Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation**. *Analytical Chemistry* 2005, **78**(2):567-574.

26. Yetukuri L, Katajamaa M, Medina-Gomez G, Seppanen-Laakso T, Vidal-Puig A, Oresic M: **Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis**. *BMC Systems Biology* 2007, **1**(1):12.

27. Del Boccio P, Pieragostino D, Di Ioia M, Petrucci F, Lugaresi A, De Luca G, Gambi D, Onofrj M, Di Ilio C, Sacchetta P *et al*: **Lipidomic investigations for the characterization of circulating serum lipids in multiple sclerosis**. *Journal of Proteomics* 2011, **74**(12):2826-2836.

28. Hotelling H: **The most predictable criterion**. *Journal of Educational Psychology* 1935, **26**:139-142.

29. Hotelling H: **Relations between two blocks of variates.** *Biometrika* 1936, **28**:321-377.

30. Gower JC: **Generalized procrustes analysis**. *Psychometrika* 1975, **40**(1):33-51.

31. Qannari EM, Wakeling I, Courcoux P, MacFie HJH: **Defining the underlying sensory dimensions**. *Food Quality and Preference* 2000, **11**(1-2):151-154.

32. Hassani S, Martens H, Qannari EM, Hanafi M, Borge GI, Kohler A: **Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods**. *Chemometrics and Intelligent Laboratory Systems* 2010, **104**(1):140-153.

33. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers R-JAN, van der Greef J, Timmerman ME: **ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data**. *Bioinformatics* 2005, **21**(13):3043-3048.

34. Conesa A, Prats-MontalbÃ¡n JM, Tarazona S, Nueda MJ, Ferrer A: **A multiway approach to data integration in systems biology based on Tucker3 and N-PLS**. *Chemometrics and Intelligent Laboratory Systems* 2010, **104**(1):101-111.

35. Wold S, Hellberg S, Lundstedt Y, Sjostrom M, Wold H: In: *Proc Symp on PLS Model Building: Theory and Application: 1987; Frankfurt am Main.*; 1987.

36. Martens H, Martens M: **Multivariate Analysis of Quality: An Introduction**. Chichester, UK: Wiley; 2001.

37. Hassani S, Martens H, Qannari EM, Hanafi M, Kohler A: **Model validation and error estimation in multi-block partial least squares regression**. *Chemometrics and Intelligent Laboratory Systems* 2011, **In press**.

38. Ottestad I, Vogt G, Retterstøl K, Myhrstad MC, Haugen J-E, Nilsson A, Ravn-Haren G, Nordvi B, Brønner KW, Andersen LF *et al*: **Oxidised fish oil does not influence established markers of oxidative stress in healthy human subjects: a randomised controlled trial**. *British Journal of Nutrition* 2011, **FirstView Article**:1-12.

39. Wold H: **Estimation of principal components and related models by iterative least squares**. In: *Multivariate Analysis.* Edited by Krishnaiah PR. New York: Academic Press; 1966: 391-420.

40. Miyashita Y, Itozawa T, Katsumi H, Sasaki S-I: **Comments on the NIPALS algorithm**. *Journal of Chemometrics* 1990, **4**(1).

41. Qin SJ, Valle S, Piovoso MJ: **On unifying multiblock analysis with application to decentralized process monitoring**. *J Chemometrics* 2001, **15**:715-742.

42. Wangen LE, Kowalski BR: **A multiblock partial least squares algorithm for investigating complex chemical systems**. *Journal of Chemometrics* 1989, **3**(1):3-20.

43. Dwass M: **Modified Randomization Tests for Nonparametric Hypotheses**. *The Annals of Mathematical Statistics* 1957, **28**:181-187.

44. Nichols TE, Holmes AP: **Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples**. *Human Brain Mapping* 2001, **15**(1):1-25.

Lipid	**p-value for 1-component MBPLSR-DA model**
LysoPC(20:5)	0,000
LysoPC(22:5)	0,001
LysoPC(22:6)	0,000
PC(30:3)	0,001
PC(32:5)	0,000
PC(33:2)+PE(36:2)	0,024
PC(33:2)+PE(36:2)	0,000
PC(34:0e)	0,001
PC(34:1)	0,018
PC(34:1e)+PE(37:1e)	0,000
PC(34:2)	0,000
PC(34:3)	0,046
PC(34:3e)+PE(37:3e)	0,030
PC(35:2)	0,010
PC(36:2)	0,001
PC(36:3)	0,000
PC(36:4)	0,045
PC(36:4e)	0,013
PC(36:5)	0,000
PC(36:5)	0,003
PC(36:5e)+PE(38:5e)	0,002
PC(37:4)/PE(40:4)	0,000
PC(38:1)	0,042
PC(38:3)	0,001
PC(38:3e)	0,010
PC(38:4)	0,000
PC(38:4)	0,000
PC(38:5)	0,000
PC(38:5e)	0,000
PC(38:5e)	0,002
PC(38:6)	0,000
PC(38:6)	0,000
PC(38:7)	0,000
PC(38:7)	0,000
PC(38:7)	0,017
PC(40:2)	0,000
PC(40:3)	0,001
PC(40:4)	0,001
PC(40:4e)	0,000
PC(40:5)	0,000
PC(40:5e)	0,023
PC(40:6)	0,000
PC(40:7)	0,003
TG(16:0/18:2/18:1)	0,049
TG(17:0/18:1/18:1)*	0,017
TG(18:1/18:1/18:1)	0,000
TG(18:1/18:1/22:1)+ TG(20:1/20:1/18:1)	0,027
TG(18:1/18:2/18:1)	0,044

Lipid	p-value for 1-component MBPLSR-DA model
TG(40:0)*	0,006
TG(42:0)*	0,039
TG(45:1)*	0,026
TG(47:2)*	0,042
TG(50:4)	0,000
TG(50:5)*	0,040
TG(52:0)	0,000
TG(52:6)	0,000
TG(52:7)*	0,000
TG(53:3)*	0,031
TG(54:2)	0,006
TG(54:3)	0,019
TG(54:4)	0,000
TG(54:4)	0,000
TG(54:5)*	0,020
TG(54:8)*	0,000
TG(56:2)*	0,001
TG(56:4)	0,016
TG(56:7)*	0,000
TG(56:8)*	0,000
TG(56:9)*	0,000
TG(58:10)*	0,000
TG(58:6)	0,002
TG(58:6)	0,000
TG(58:8)*	0,000
TG(58:9)*	0,000
TG(59:2)*	0,000

Table 1 - P-values for 1-component MBPLSR-DA model

The lipids that were contributing significantly to the first component of the MBPLSR-DA model (i.e. the lipids that changed significantly in response to the intervention diet) are listed with their p-values.

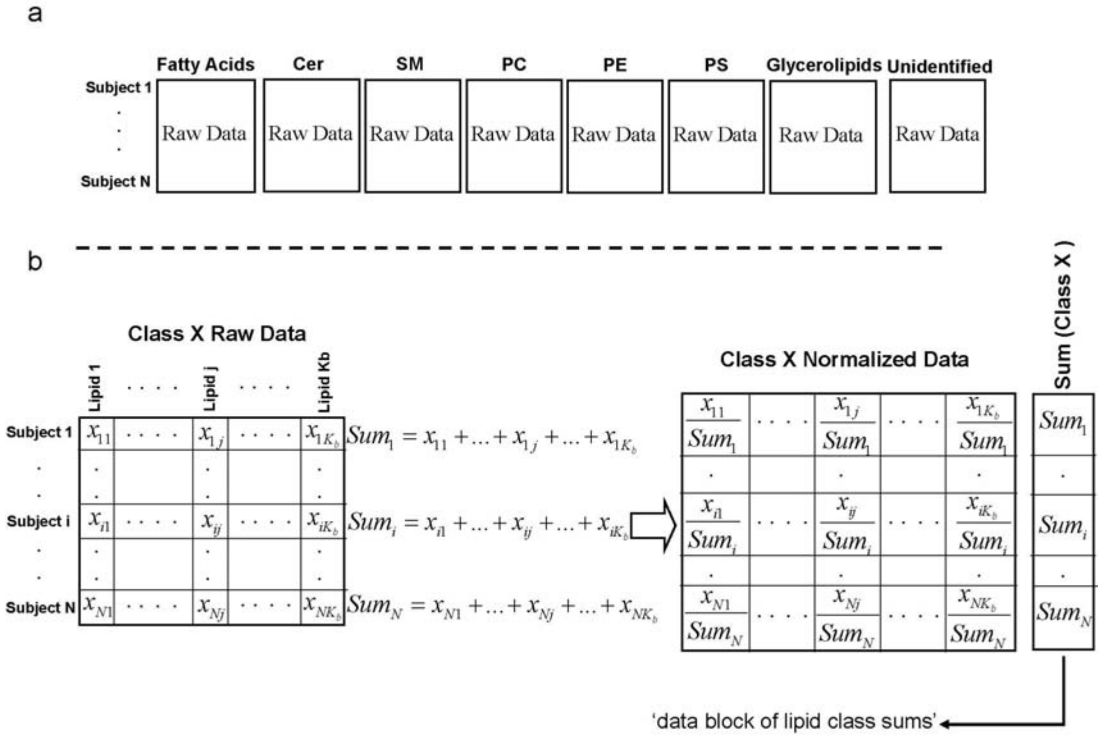

a
Fatty Acids
Cer
SM
PC
PE
PS
Glycerolipids
Unidentified
Subject 1
Subject N
Raw Data
b
Class X Raw Data
Lipid 1
Lipid j
Lipid Kb
Subject 1
Subject i
Subject N
x_{11} x_{1j} x_{1K_b}
x_{i1} x_{ij} x_{iK_b}
x_{N1} x_{Nj} x_{NK_b}
$Sum_1 = x_{11} + \ldots + x_{1j} + \ldots + x_{1K_b}$
$Sum_i = x_{i1} + \ldots + x_{ij} + \ldots + x_{iK_b}$
$Sum_N = x_{N1} + \ldots + x_{Nj} + \ldots + x_{NK_b}$
Class X Normalized Data
$\frac{x_{11}}{Sum_1}$ $\frac{x_{1j}}{Sum_1}$ $\frac{x_{1K_b}}{Sum_1}$
$\frac{x_{i1}}{Sum_i}$ $\frac{x_{ij}}{Sum_i}$ $\frac{x_{iK_b}}{Sum_i}$
$\frac{x_{N1}}{Sum_N}$ $\frac{x_{Nj}}{Sum_N}$ $\frac{x_{NK_b}}{Sum_N}$
Sum (Class X)
Sum_1
Sum_i
Sum_N
'data block of lipid class sums'

Figure 1 – Multi-block structure and normalization procedures of an example lipidomics data set

a) The structure of an example multi-block lipidomics data set is illustrated. b) The normalization procedure for an example lipid class is shown.

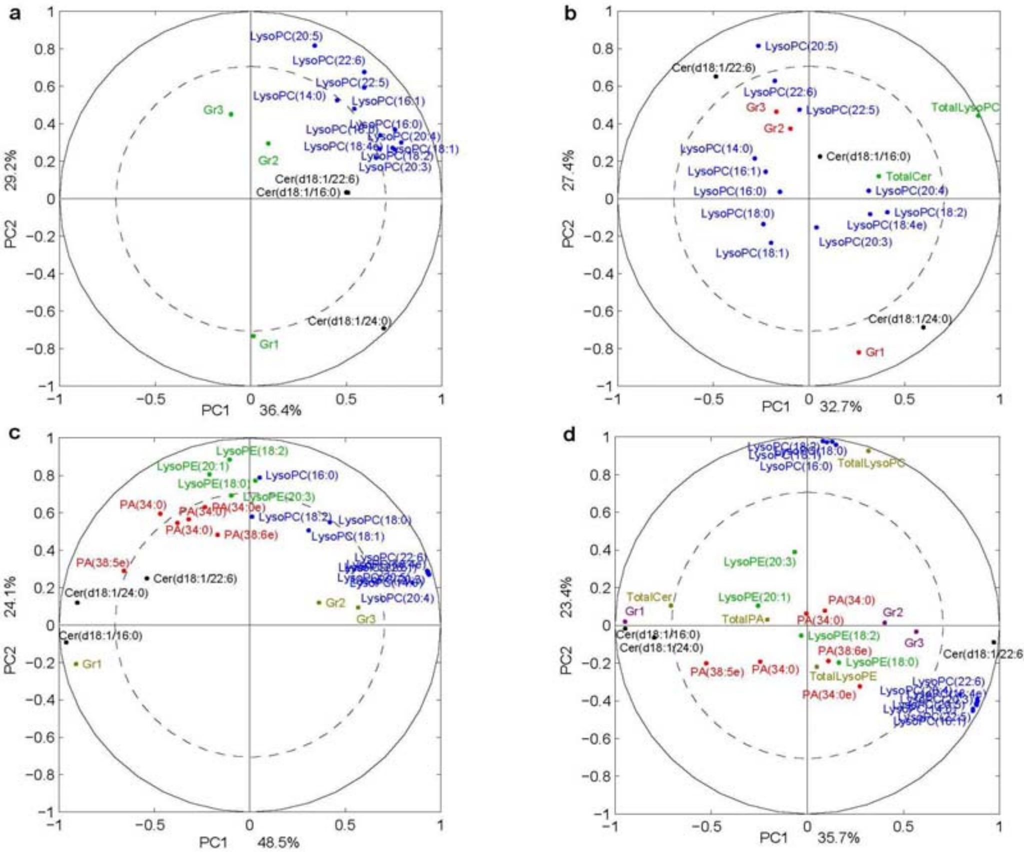


Figure 2 – The effect of different normalization procedures on the multi-block analysis results of simulated data sets

a) Correlation loading plot for the multi-block analysis when the data is normalized according to method 1. b) Correlation loading plot for the multi-block analysis when the data is normalized according to method 2. The simulated data set mimics a two-block lipidomics data set consisting of lipid classes LycoPC and Ceramides (Cer).

c) Correlation loading plot for the multi-block analysis when the data is normalized according to method 1. d) Correlation loading plot for the multi-block analysis when the data is normalized according to method 2. The data is a four-block simulated data set containing Ceramides (Cer), LycoPC, LycoPE and PA.

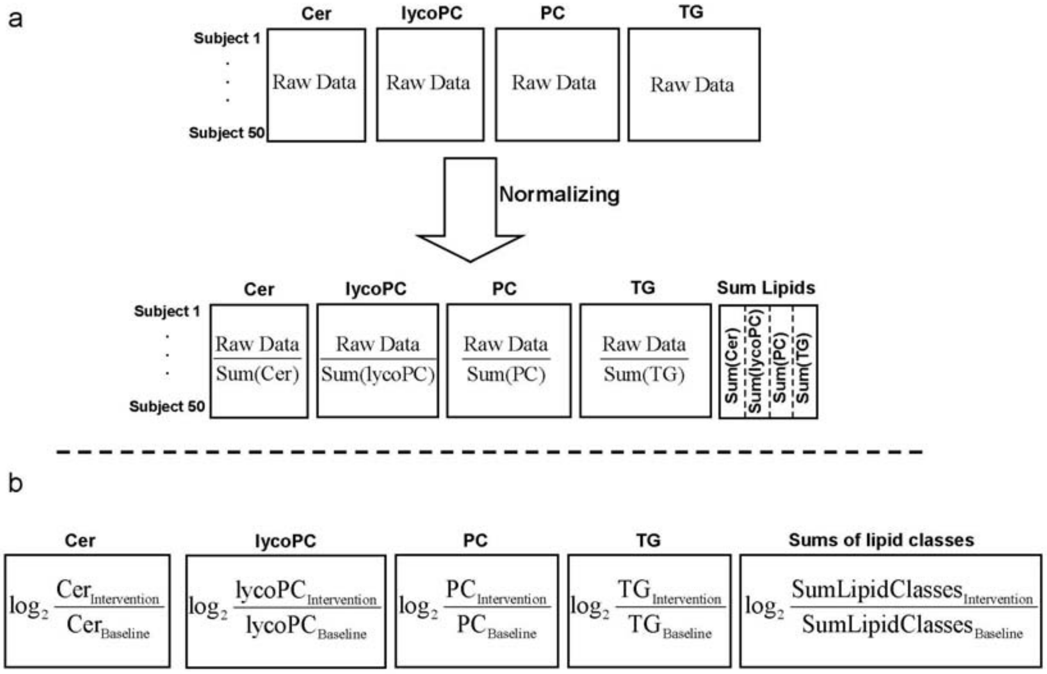

a
Cer
lycoPC
PC
TG
Subject 1
Subject 50
Raw Data
Raw Data
Raw Data
Raw Data
Normalizing
Cer
lycoPC
PC
TG
Sum Lipids
Subject 1
Subject 50
Raw Data / Sum(Cer)
Raw Data / Sum(lycoPC)
Raw Data / Sum(PC)
Raw Data / Sum(TG)
Sum(Cer)
Sum(lycoPC)
Sum(PC)
Sum(TG)
b
Cer
lycoPC
PC
TG
Sums of lipid classes
$\log_2 \frac{Cer_{Intervention}}{Cer_{Baseline}}$
$\log_2 \frac{lycoPC_{Intervention}}{lycoPC_{Baseline}}$
$\log_2 \frac{PC_{Intervention}}{PC_{Baseline}}$
$\log_2 \frac{TG_{Intervention}}{TG_{Baseline}}$
$\log_2 \frac{SumLipidClasses_{Intervention}}{SumLipidClasses_{Baseline}}$

Figure 3 – Pre-processing of an example lipidomics data set

a) The multi-block structure and normalization procedure for a lipidomics data set from an intervention study is illustrated. b) The baseline correction procedure is shown for the multi-block lipidomics data set.

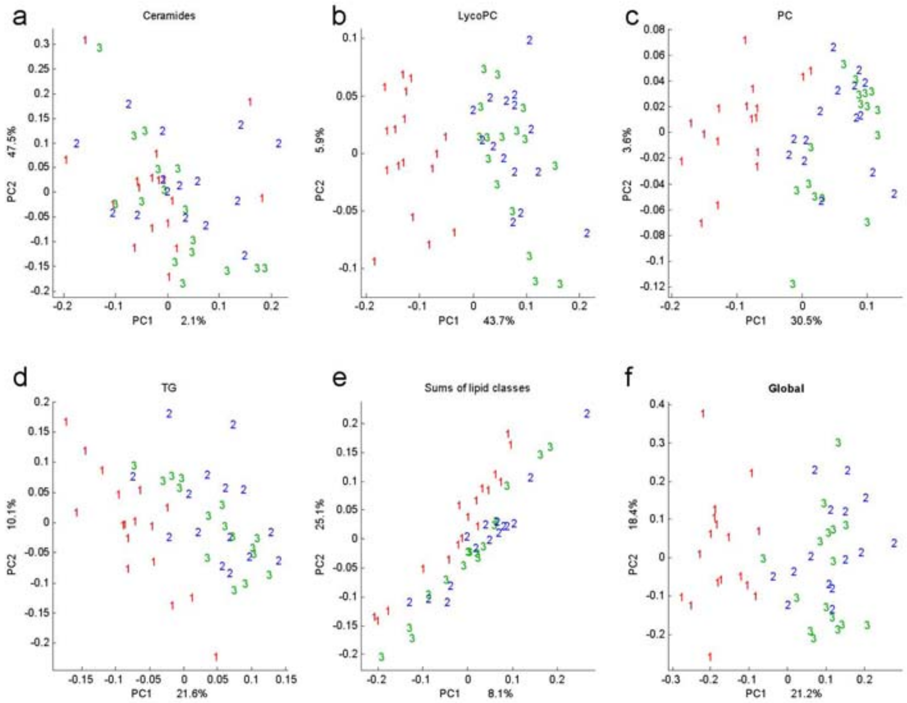


Figure 4 – Global and block score plots for CPCA

The samples are labelled "1" (red), "2" (blue) and "3" (green) according to the intervention groups. The (un-validated) explained variances are written by the axes. (a-e) Block score plots and (f) global score plot are shown for the first and second components.

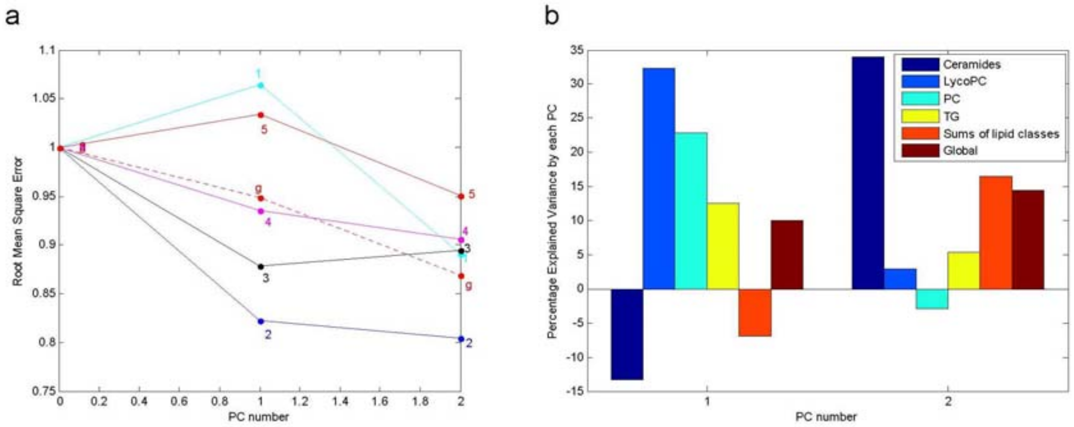


Figure 5 – Validation plots for CPCA

a) Error plots (i.e. RMSE plots) for the global model and for the block models of the different lipid classes are illustrated for the first and second components. b) Bar plots of the percentage cross-validated explained variances are illustrated for the first and second components.

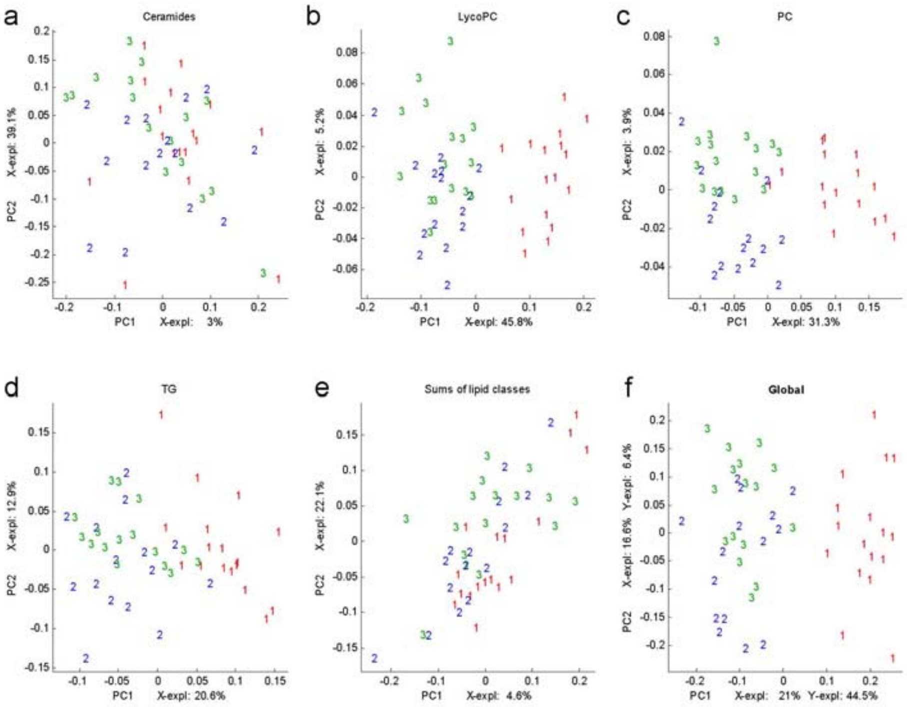


Figure 6 – Global and block score plots for MBPLSR-DA

The samples are labelled “1” (red), “2” (blue) and “3” (green) according to the intervention groups. The (un-validated) explained variances are written by the axes. (a-e) Block score plots and (f) global score plot are shown for the first and second components.

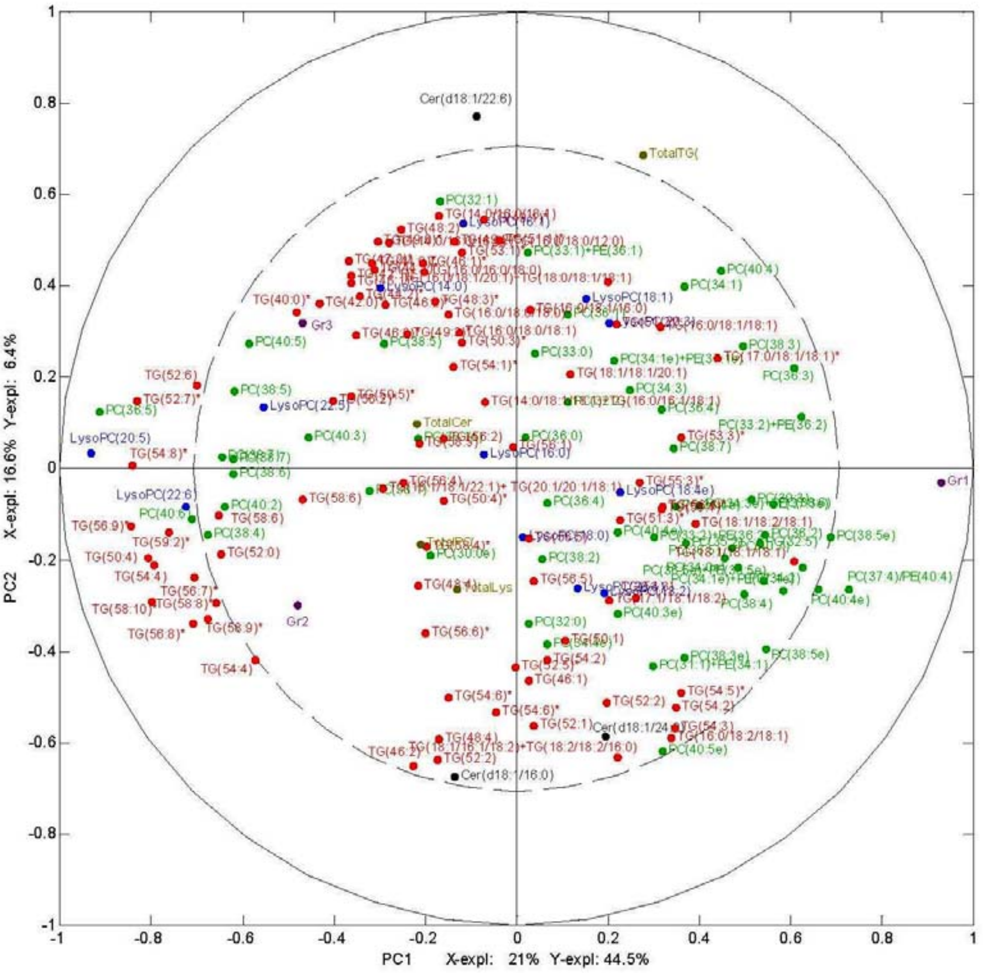


Figure 7– Correlation loading plot for MBPLSR-DA

The lipid-lipid variation and co-variation patterns are shown in the correlation loading plot. The lipids are illustrated by different colours according to their lipid classes. The (un-validated) explained variances (for both **X** and **Y**) are written by the axes.

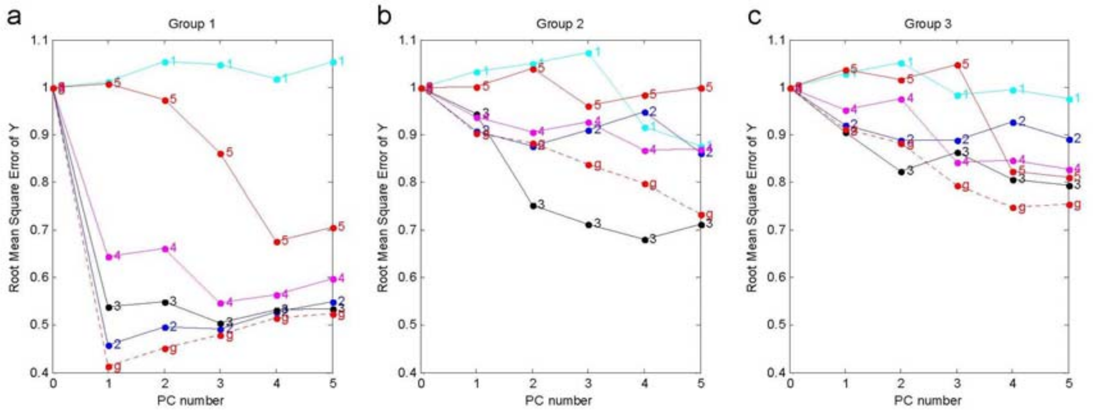


Figure 8 – Validation plots for MBPLSR-DA

a) Error plots (i.e. RMSE plots) for the global model and for the block models of the different lipid classes are illustrated for five components. a) The error plots for validating the separation of group 1 samples from the rest of the samples are illustrated. b) The error plots for validating the separation of group 2 samples from the rest of the samples are illustrated. c) The error plots for validating the separation of group 3 samples from the rest of the samples are illustrated.

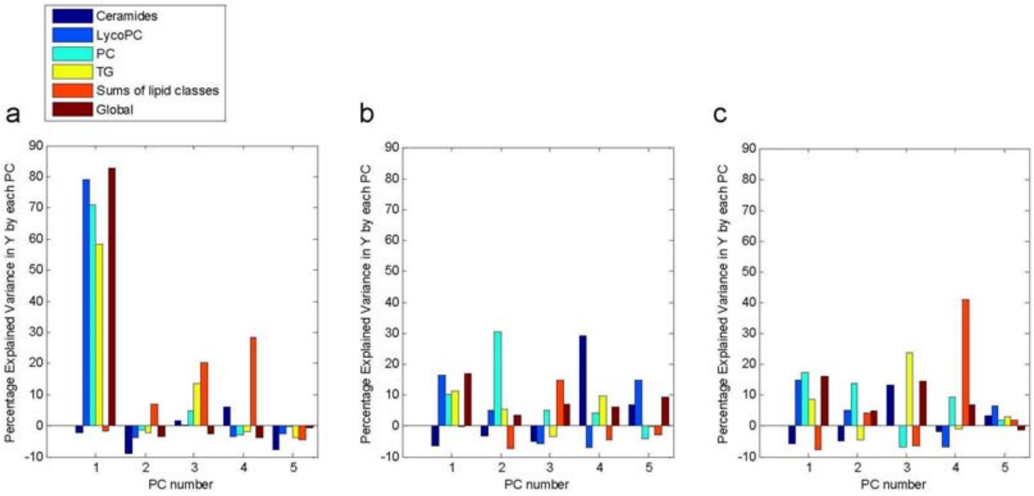


Figure 9 – Validation plots for MBPLSR-DA

Bar plots of the percentage cross-validated explained variances are illustrated for five components. a) Bar plots showing the contribution of the different lipid classes and the global model to the separation of group 1 samples from the rest of the samples are illustrated. b) Bar plots showing the contribution of the different lipid classes and the global model to the separation of group 2 samples from the rest of the samples are illustrated. c) Bar plots showing the contribution of the different lipid classes and the global model to the separation of group 3 samples from the rest of the samples are illustrated.

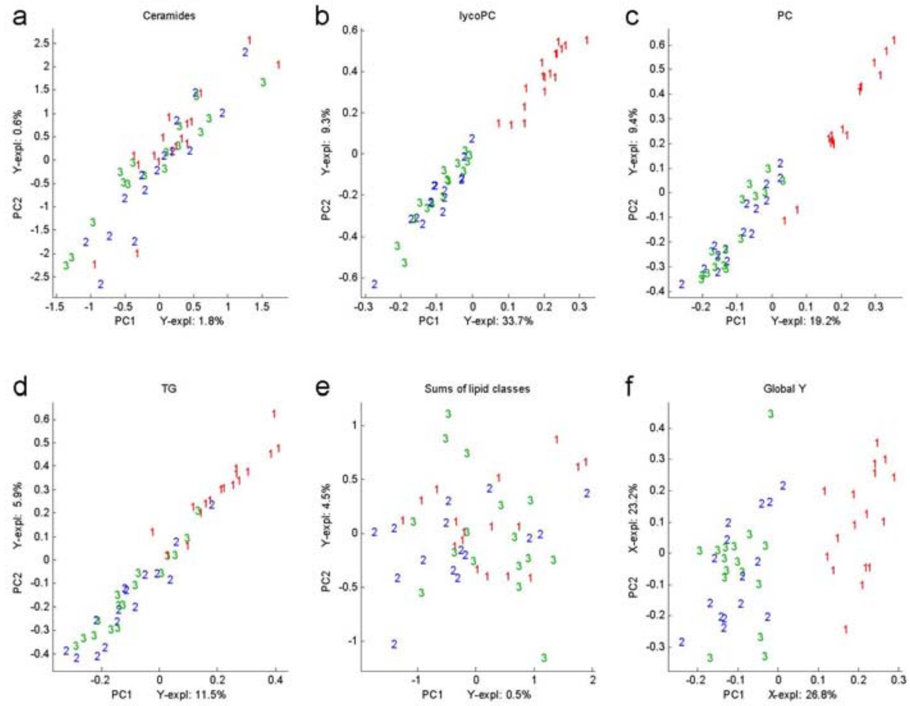


Figure 10 – Global and block score plots for ANOVA-MBPLSR

The samples are labelled “1” (red), “2” (blue) and “3” (green) according to the intervention groups. The (un-validated) explained variances are written by the axes. (a-e) Block score plots and (f) global score plot are shown for the first and second components.

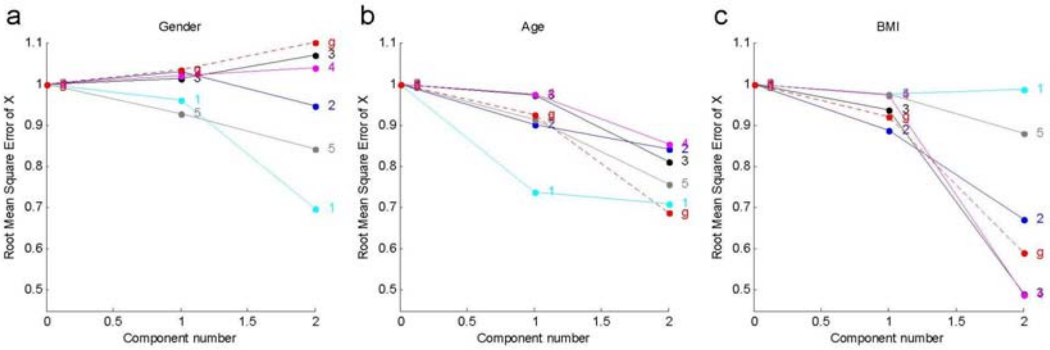


Figure 11 – Validation plots for ANOVA-MBPLSR

a) Error plots (i.e. RMSE plots) for the global model and for the block models of the different lipid classes are illustrated for two components. a) The error plots for validating the contribution of “Gender” to the patterns in Fig. 12 are illustrated. b) The error plots for validating the contribution of “Age” to the patterns in Fig. 12 are illustrated. c) The error plots for validating the contribution of “BMI” to the patterns in Fig. 12 are illustrated.

Paper VI

Fish Oil Supplementation Alters the Plasma Lipidomic Profile and Increases Long-Chain PUFAs of Phospholipids and Triglycerides in Healthy Subjects

Inger Ottestad[1,2], Sahar Hassani[3,4], Grethe I. Borge[3], Achim Kohler[4,3], Gjermund Vogt[3], Tuulia Hyötyläinen[5], Matej Orešič[5], Kirsti W. Brønner[6], Kirsten B. Holven[2], Stine M. Ulven[1] and Mari CW. Myhrstad[1]

[1]Department of Health, Nutrition and Management, Faculty of Health Sciences, Oslo and Akershus University College of Applied Sciences, P.O. Box 4, St Olavs plass, 0130 Oslo, Norway

[2]Department of Nutrition, Institute for Basic Medical Sciences, University of Oslo, P.O. Box 1046 Blindern, 0317 Oslo, Norway

[3] Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, 1430 Ås, Norway

[4]Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology, Norwegian University of Life Science, 1432 Ås, Norway

[5] VTT Technical Research Centre of Finland, Tietotie 2, FI-02044 VTT, Espoo, Finland

[6]TINE SA, Centre for Research and Development, P.O. Box 7, Kalbakken, N-0902 Oslo, Norway

Trial registration: ClinicalTrials.gov NCT01034423

Corresponding author: Myhrstad MCW, Fax no +47-64849001, email: mari.myhrstad@hioa.no

Funding: This study was supported by the Norwegian Research Council and the Nordic Centre of Excellence on Food, Nutrition and Health "Systems biology in controlled dietary interventions and cohort studies" (SYSDIET nr. 070014).

Abstract

Background: While beneficial health effects of fish and fish oil consumption are well documented, the incorporation of n-3 polyunsaturated fatty acids in plasma lipid classes is not completely understood. The aim of this study was to investigate the effect of fish oil supplementation on the plasma lipidomic profile in healthy subjects.

***Methodology/Principle Findings*:** In a double-blinded randomized controlled parallel-group study, healthy subjects received capsules containing either 8 g/d of fish oil (FO) (1.6 g/d EPA+DHA) (n=16) or 8 g/d of high oleic sunflower oil (HOSO) (n=17) for seven weeks. During the first three weeks of intervention, the subjects completed a fully controlled diet period. BMI and serum triglycerides, total-, LDL- and HDL-cholesterol were unchanged during the intervention period. Lipidomic analyses were performed using Ultra Performance Liquid Chromatography (UPLC) coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (QTOFMS), where 568 lipids were characterized and 260 identified. Both t-tests and Multi-Block Partial Least Square Regression (MBPLSR) analysis were performed for analysing differences between the intervention groups. The intervention groups were well separated by the lipidomic data after three weeks of intervention. Several lipid classes such as phosphatidylcholine, phosphatidylethanolamine, lysophosphatidylcholine, sphingomyelin, phosphatidylserine, phosphatidylglycerol, and triglycerides contributed strongly to this separation. Twenty-three lipids were significantly decreased (FDR< 0.05) in the FO group after three weeks compared with the HOSO group, whereas fifty-one were increased including selected phospholipids and triglycerides of long-chain polyunsaturated fatty acids. After seven weeks of intervention the two intervention groups showed similar grouping.

Conclusions/Significance: In healthy subjects, fish oil supplementation alters lipid metabolism and increases the proportion of phospholipids and triglycerides containing long-chain polyunsaturated fatty acids. Whether the beneficial effects of fish oil supplementation may be explained by a remodeling of the plasma lipids into phospholipids and triglycerides of long-chain polyunsaturated fatty acids needs to be further investigated.

Introduction

Intake of fish and fish oil, containing n-3 fatty acids; eicosapentaenoic acid (EPA; 20:5) and docosahexaenoic acid (DHA; 22:6), is associated with beneficial health effects such as reduced risk of cardiovascular disease and sudden cardiac death [1-4]. The beneficial effects of marine n-3 fatty acids have been explained by decreased plasma triglycerides (TGs) [5,6], moderate reduction in blood pressure [7], reduced platelet aggregation [8,9], and protection against cardiac arrhythmias [10,11]. It has been suggested that bioactive lipid components may be important in mediating these effects, but the molecular mechanisms is still to a large extent unknown.

Cells, tissues and biological fluids contain tens of thousands of structurally different lipids, that fulfil multiple roles in cellular signalling, in membrane structure, and as fuel sources for many cell types [12]. The entire spectrum of lipids in a biological system, can be defined as the lipidome [13], which combines mass spectrometry technology and bioinformatics methods with traditional methods such as sample preparation, lipid extraction and separation. Lipidome analyses have revealed a diversity of lipid compounds in human plasma, which can be classified into six main lipid categories including fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids and prenol lipids [14]. The major plasma lipids are the glycerolipids (TGs), glycerophospholipids (phospholipids) and sterol lipids which are transported in the lipoprotein particles [14,15].

In n-3 FAs intervention studies fatty acids have been measured in different blood compartments such as in platelets and red blood cells, and in plasma cholesteryl esters, triglycerides and phospholipids. Lipidomic analysis now offers the opportunity to detect exact fatty acid composition of these individual lipids. [16]. Recently it was shown that the plasma lipidomic profile was altered in subjects with coronary heart disease after intake of fatty fish, and in subjects with metabolic syndrome after consumption of a healthy diet containing fatty fish, wholegrain products and bilberries [17,18]. Furthermore, profiling of the plasma lipids suggests a relationship between the composition of plasma lipids and diet [19,20], with diet-induced weight loss [21] and to diet-related diseases such as diabetes mellitus [22]. This opens up the opportunity to identify new functional lipid biomarkers to detect and prevent diet-related diseases. We have however not been able to find studies showing the plasma lipidomic profile in healthy subjects after intake of fish oil.

We have previously reported that in the present study a daily intake of fish oil (1.6 g EPA + DHA/d) did not change the level of serum lipids, markers of oxidative stress, lipid oxidation or inflammation, whereas an increase in plasma EPA, DPA and DHA was observed after three and seven weeks of intervention in a randomized controlled study in healthy subjects [23]. The aim of this study was to apply a lipidomic strategy to further describe the effect of fish oil supplementation in healthy subjects.

Materials and Methods

Subjects

Healthy men and women between 18-50 years were recruited into this study. Detailed description of the protocol, participant recruitment and enrolment, inclusion and exclusion criteria, and compliance are described in details elsewhere [23] . In brief, exclusion criteria were total cholesterol > 7.5 mmol/l, triglycerides > 4 mmol/l, glucose > 6.0 mmol/l, C-reactive protein (CRP) > 10 mg/l, body mass index (BMI) ≥ 30 kg/m^2 and blood pressure (≥ 160/100). The study was performed at the Akershus University College, Norway between September and December 2009.

Ethics Statement

Written informed consent was obtained from all participants and the protocol was approved by the Regional Committee of Medical Ethics (approval no.6.2008.2215) and by the Norwegian Social Science Data Services (approval no.21924), and was conducted in accordance with the Declaration of Helsinki.

Study Design

This study was a part of a randomized controlled double-blinded three-arm parallel group study, designed to investigate health effects from intake of fish oil [23]. In the present study, data from two of the intervention groups are included, as shown in Figure 1. Subjects in the present study were given 8 g/d of either fish oil (FO) or high oleic sunflower oil (HOSO), and each subject was taking 16 capsules/d minimum twice each day for seven weeks. Subjects in the fish oil group received capsules containing 0.7 g/d EPA + 0.9 g/d DHA from cod liver oil (Gadidae sp., TINE EPADHA Oil 1200) provided by TINE SA (Oslo, Norway) and subjects in the control group received high oleic sunflower oil purchased from AarhusKarlshamn AB (Malmø, Sweden). The subjects were instructed to take the capsules with food (minimum two meals). The fatty acid composition in the oils has been described elsewhere [23].

The subjects met for visits and blood samples for the lipidome analyses were collected at 0, 3 and 7 weeks. Between the screening and baseline visit (week 0), the subjects conducted a four-week washout period, where foods containing marine n-3 fatty acids were avoided.

During the first three weeks of the intervention the subjects conducted a fully-controlled isocaloric diet, provided with all food and beverages at Akershus University College, Norway. The last four weeks of the intervention period the subjects returned to their habitual diet. The food items provided in this study and the energy provided from diet have previously been described [23]. Intake of fish, fish products, marine n-3 enriched food or dietary supplements was not allowed during the entire study period of 11 weeks. The study was registered at www.clinicaltrial.gov (IDno. NCT01034423).

Blood sampling

Subjects were told to refrain from alcohol consumption and vigorous physical activity the day prior to blood sampling. Venous blood samples were drawn after an overnight fast (≥12 hours) at the same time (± 2h) and serum were kept at room temperature at 30 min before centrifuged (1500*g* 12 min). EDTA-plasma was immediately placed on ice and centrifuged within 10 min (1500*g*, 4°C, 10 min). N_2 flushed plasma samples were snap frozen and stored at -80°C until further analysis.

Routine laboratory analysis

Fasting serum hsCRP, total- cholesterol, LDL-cholesterol, HDL-cholesterol, triglycerides and glucose were measured by standard methods at a routine laboratory (Fürst Medical Laboratory, Oslo, Norway).

Lipidomic analyses

An aliquot (10 µL) of plasma sample was diluted with 10 µL of 0.15 M (0.9%) sodium chloride and 10 µL of internal standard mixture containing PC(17:0/0:0), PC(17:0/17:0), PE(17:0/17:0), PG(17:0/17:0)[rac], Cer(d18:1/17:0), PS(17:0/17:0) and PA(17:0/17:0) (Avanti Polar Lipids, Inc., Alabaster, AL, USA) and TG(17:0/17:0/17:0) and MG(17:0/0:0/0:0)[rac], DG(17:0/17:0/0:0)[rac] (Larodan Fine Chemicals) was added . The lipids were extracted using the mixture of HPLC-grade chloroform and methanol (2:1; 100 µL). The lower phase was collected (60 µL) and 10 µL internal standard mixture containing labeled PC(16:1/0:0-D_3), PC(16:1/16:1-D_6) and TG(16:0/16:0/16:0-^{13}C3) was added.

The extracts were analyzed on a Waters Q-Tof Premier mass spectrometer combined with an Acquity Ultra Performance LC^{TM} (UPLC). The column (at 50 °C) was an Acquity UPLC^{TM} BEH C18 2.1 × 100 mm with 1.7 µm particles. The solvent system included A: ultrapure water with 1% 1 M NH_4Ac and 0.1% HCOOH, and B: LC/MS grade acetonitrile/isopropanol (1:1) with 1% 1M NH_4Ac and 0.1% HCOOH. The gradient started from 65% A / 35% B, reached 80% B in 2 min, 100% B in 7 min and remained there for 7 min. The flow rate was 0.400 ml/min and the injected amount was 2.0 µl (Acquity Sample Organizer, at 10 °C). Reserpine was used as the lock spray reference compound. The lipid profiling was carried out using ESI+ mode and the data was collected at mass range of m/z 300-1200 with scan duration of 0.2 sec. The data was processed by using MZmine2 software [24] and the lipid identification was based on an internal spectral library.

The data processing included alignment of peaks, peak integration, normalization and identification. Lipids were identified using an internal spectral library. The data was normalized using one or more internal standards representative of each class of lipid present in the samples: the intensity of each identified lipid was normalized by dividing it with the intensity of its corresponding standard and multiplying it by the concentration of the standard. All monoacyl lipids except cholesterol esters, such as monoacylglycerols and monoacylglycerophospholipids, are normalized with PC(17:0/0:0), all diacyl lipids except ethanolamine phospholipids are normalized with PC(17:0/17:0), all ceramides with Cer(d18:1/17:0), all diacyl ethanolamine phospholipids with PE(17:0/17:0), and TG and cholesterol esters with TG(17:0/17:0/17:0). Other (unidentified) molecular species were normalized with PC(17:0/0:0) for retention time < 300 s, PC(17:0/17:0) for retention time between 300 s and 410 s, and TG(17:0/17:0/17:0) for higher retention times.

Statistical analyses

Sample size was calculated using expected change in plasma n-3 fatty acids as described as previously described [23]. Multi-Block Partial Least Squares Regression (MBPLSR) analysis was used for exploring the sample and variable variation patterns in the data [25] where each lipid class was defined as one individual block [26] resulting in 11 blocks of descriptor variables in total (i.e. $[\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^{11}]$). The multi-block set of descriptor variables were organized in the following order: Ceramides as block one ($\mathbf{X}^1$), lysophosphatidylcholines (lysoPC) as block two ($\mathbf{X}^2$), lysophosphatidylethanolamines (lysoPE) as block three ($\mathbf{X}^3$), phosphatidic acid (PA) as block four ($\mathbf{X}^4$), phosphatidylcholines (PC) as block five ($\mathbf{X}^5$), phosphatidylethanolamines (PE) as block six ($\mathbf{X}^6$), phosphatidylglycerols (PG) as block seven ($\mathbf{X}^7$), phosphatidylserines (PS) as block eight ($\mathbf{X}^8$), sphingomyelins (SM) as block nine ($\mathbf{X}^9$), triglycerides (TG) as block ten ($\mathbf{X}^{10}$) and sums of lipid classes together with phosphatidylinositol (PI) as block eleven ($\mathbf{X}^{11}$) (a separate block was not assigned to PI class since it contained only onelipid). An intervention group indicator variable was used as response variable (*y*-variable). In order to estimate both the influence of the total amount of lipids in each lipid class and simultaneously the influence of the relative variation within each lipid class, -each block was normalized by a division by the total amount of lipids in the corresponding class. The total amounts of lipids of each lipid classes were then used as an additional block and named "the sums of lipids". Subsequently, plasma lipids were transformed by taking the log2 ratio (baseline adjusted log2 values) after three and seven weeks in both the FO and the HOSO group. After this, each block (lipid class) was set on the same footing prior to MBPLSR analysis as described in [26]. Model-validation and testing of the influence of each lipid class to the global model was done by cross-validation as described in [26].Variable significance testing for the difference between the groups at baseline and after the intervention (baseline adjusted values) was done by cross-validation [25,26] of the multivariate MBPLSR model and by univariate testing using Student`s- *t* test. For the univariate testing the log2 ratios were used and False Discovery Rate (FDR) corrected *q*-values were computed using the R package 'qvalue'. Two subjects were detected as outliers by the MBPLSR models and were therefore excluded from the further analysis. Baseline characteristics were analyzed using (baseline adjusted values) Student`s- *t* test and Mann Whitney U test (serum triglycerides) when data was normally and not normally distributed, respectively. The significance level was set to 5% (two-sided) and the power of the test was chosen to be 80%. Data in Table 1 are presented as mean ± SD. All univariate analyses were performed using SPSS for windows (SPSS, version 19.0) and multivariate analyses were performed using in-house-written and standard MATLAB routines (MATLAB version 7.8).

Results

Characteristics of the subjects

A total of 33 normal weight healthy subjects (n=8 men and n=25 women) completed this study. The subjects were young (28 ± 8 years), and with serum lipids within the normal range as shown in Table 1. No differences in age, BMI or serum lipids were observed between the FO group (n=16) and the HOSO group (n=17) at baseline (Table 1). Serum lipids and BMI were not significantly changed between the groups after three (Table 1) or seven weeks of intervention (data not shown).

Plasma lipidomic profile

A total of 568 lipids were detected and quantified in the plasma samples of the two intervention groups. Of these lipids, 260 were identified, including the following lipid classes;

ceramides, sphingomyelins (SM), lysophosphatidylcholines (lysoPC), lysophosphatidylethanolamines (lysoPE), phosphatidic acids (PA), phosphatidylcholines (PC), phosphatidyethanolamines (PE), phosphatidylglycerols (PG), phosphatidylinositols (PI), phosphatidyserines (PS) and triglycerides (TG). In the present study, the identified lipids were included in the statistical analysis.

MBPLSR was performed with lipid class blocks as a multi-block **X** and intervention group indicator variable as y-variable. The lipid class block variation patterns after three weeks of intervention are shown in Figure 2 and Figure 3A-D. The lipidomic profiles of the two intervention groups were well separated in the global sample variation pattern (global score plot) (Figure 3F). The first principal component accounted for most of the separation of the two groups and explained 91.5% of the total variance in the y-variable. The explained block variances are shown on the respective axes. Several lipid class blocks (lysoPC, PC, PE, PG, PS, SM and TG) showed a clear separation of the FO and the HOSO group, whereas ceramides, lysoPE, PA lipids did not separate the FO and the HOSO group (Figure 2 and Figure 3A-D). In addition, the sums of lipid classes did not separate the FO and the HOSO group (Figure 3E) showing that the differences between FO and HOSO group can be explained by remodelling within lipid classes rather than by changes in the total amount of lipids in each class. A similar separation of the groups and patterns in block score plots were observed also after seven weeks of intervention (data not shown). To further analyze the plasma lipid profile and to identify specific lipids contributing to the separation, we decided to use the data obtained after completing a three weeks fully controlled diet period. After three weeks of intervention, the contribution of each lipid class block to the prediction of the group indicator variable was validated by calculating root mean squared errors of cross-validation per block. The validated explained variances for the first two components are shown in Figure 4. The first principal component of the lysoPC, PC, PE and SM lipid classes described most of the group separation and explained more than 70% of the y-variance. The second component of the PG, PS and TG lipids explained the separation of the two intervention groups further accounting for 19-37% of the y-variance. By significance testing using cross-validation and Jack-knifing [26] a total of 75 lipids were identified as significant for the separation of the two intervention groups after three weeks using a two principal component model (Supplementary Table S1). By investigating the validated root mean squared error as a function of components a two component model was selected (Supplementary Figure S1).

To further identify the specific lipids that contributed to the distinction of the FO and the HOSO group correlation loading plots were studied [26]. The correlation loading plots in

Figure 5 and 6 show that several phospholipids and TGs containing long-chain PUFAs including lysoPC(20:5), lysoPC(22:6), PC(36:5), PC(40:6), PE(38:5), TG(50:4), TG(52:6), TG(52:7), TG(54:8), TG(56:7), TG(56:8), TG(56:9), TG(58:6), TG(58:8), TG(58:9), and TG(58:10) were strongly positively correlated to intake of FO supplementation.

A strong positive correlation was also observed between intake of FO supplementation and lipids of long-chain and lower double bond content such as SM(18:0/24:0), TG(59:2) and TG(52:0). Only PC(40:4e), PC(37:4)/PE(40:4), PC(38:5) and PC(34:2) were found negatively correlated to the FO group.

In order to describe altered lipids in the FO group compared to the HOSO group unpaired t-test was performed. In the FO group, 74 lipids were significantly altered (FDR< 0.05) compared with the HOSO group after three weeks of intervention, and 51 out of these 74 lipids were significantly increased. Several phospholipids and TGs containing long-chain PUFAs were increased in the FO group, compared to the HOSO group. Significantly altered lysoPC, PC, PE, PA, PG, PS, PI, SM and TG lipids are shown in Table 2 and 3. Furthermore, 49 lipids were identified as significantly altered in the FO group compared to the HOSO group by both unpaired t-test and MBPLSR (Supplementary Table S1). After seven weeks of intervention 58 significant altered lipids were identified in the FO group compared to the HOSO group, and 33 out of these 58 lipids were significantly altered after both three and seven weeks (data not shown).

Discussion

We have investigated the effect of fish oil supplementation on the plasma lipidomic profile in healthy subjects. A clear distinction of the lipidomic profile was obtained between the FO and the HOSO group after three and seven weeks of intervention. The lipid classes that contributed to the separation of the intervention groups were LysoPC, PC, PE, PG, PS, SM and TG. FO supplementation especially increased phospholipids and TGs of long-chain PUFAs, but the total concentration of the lipids within each lipid classes remained unchanged and did not differ in the FO compared to the HOSO group. The clear distinction between the FO and the HOSO group was observed after a fully-controlled isocaloric diet period for three weeks and it was also evident after the subjects had continued on their habitual diet for additional four weeks.

By using MBPLSR, co-variation patterns in sample and variable space for the different lipid classes was studied. MBPLSR is a method based on latent variables, where by using only few latent variables the problem of over-fitting and false discovery is minimized. For MBPLSR analysis data blocks were organized and normalized, such that remodelling effects in each lipid class and changes in total amounts of lipids per class could be studied separately.

Recent results from a dietary intervention study have shown that fish intake increased TGs of long-chain PUFA similar to our results, and that fish consumption for eight weeks increased plasma long-chain TGs in subjects with coronary heart disease [18]. Interestingly, this effect was significant after intake of lean fish and not fatty fish [18]. A healthy diet rich in whole grain products, fish and bilberries significantly changed multiple TGs incorporating long-chain PUFAs after 12 weeks intervention in subjects with impaired glucose metabolism [17]. Fish oil supplementation was previously found to reduce the total plasma TG concentration by selective reducing short chain fatty acids and to increase various phospholipids [27].Thus, it is reasonable to assume that intake of fish and fish oil causes a remodulation of plasma TG species towards more long-chained fatty acids. Our results demonstrate that this remodeling occurs in healthy subjects where the total serum TG level and the BMI are unchanged.

We observed an increase in several phospholipids incorporating n-3 PUFAs, such as lysoPC(20:5) and lysoPC(22:6) in the FO group compared to the HOSO group. An association between n-3 FA intake and changes in lysoPC has previously been described [18,28], and in accordance with our results, lysoPC(20:5) was significantly increased in subjects with impaired glucose metabolism after a healthy diet containing fatty fish [17]. In contrast, fatty fish consumption for eight weeks in subjects with CVD decreased the total concentration of lysoPC [18]. In addition, Block and colleagues found that FO supplementation increased the EPA and DHA species of lysoPC in healthy individuals [28]. The potential health effect of altering the blood plasma concentration of EPA and DHA lysoPC compounds is uncertain. However, the biological functions of lysoPC compounds are assumed to vary with the degree of saturation and acyl length [28] and LysoPC has been suggested as the major carrier of DHA to brain tissues [29].

Three out of four significantly altered SM lipids were increased in the FO group compared to the HOSO group. SM lipids are by far the most dominant circulating sphingolipid representing 88 % of the total concentration compared to ceramides that account for approximately 3 % [30]. SM in blood is key components and exists predominantly in the hydrophobic outer layer of lipoprotein particles. Of the lipoprotein particles, the VLDL particle contains the highest amount of SM lipids [31]. However, the localization, distribution and role SM lipid species among the lipoprotein particles is still obscure.

In the present study, FO supplementation was not associated with changes in plasma PA, lysoPE and ceramides, indicating that n-3 PUFA is selectively incorporated into other lipid classes. Ceramides have been associated with inflammation and cardiovascular disease [32,33]. However, high content of specific C_{24} ceramides have been linked to less atherogenic lipoprotein particles in healthy subjects [31]. Lankinen et al. observed that the total concentration of ceramides decreased after fatty fish consumption for eight weeks [18]. The discrepancies observed between these studies may be due to differences in the study population and design, or due to lack specific bioactive components in fish oil which are normally present in fish.

Lipid profiling has identified a relation between lipid acyl chain structure and risk of disease [22]. The present study shows that fish oil supplement increases the level of lipids such as TG(56:9), TG(58:10), LysoPC(22:6) and PC(38:6). These lipid species were recently associated with decreased risk of diabetes, when lipidome analyses were applied to plasma obtained from participants in the Framingham heart cohort study [22]. In that study a higher carbon number and higher double bond content were associated with decreased risk of diabetes. Thus, long-chain highly unsaturated TGs that have been associated with diabetes risk reduction were increased after intake of fish oil in the present study.

Whether the beneficial effects of fish oil supplementation may be explained by a remodeling of the plasma lipids into TGs and PLs of long-chain PUFAs, needs to be further investigated. However, PUFAs incorporated into TGs and PLs may reach tissues, cells and lipoproteins by an selective lipid exchange [34]. In the tissues, EPA and DHA can be incorporated into membranes and cause alterations in signaling pathways and the formation of lipid mediators that are important in inflammation [35,36]. In addition, EPA and DHA or their oxidation products have the ability to activate transcription factors both in the liver and in other metabolic active tissues and increase the expression of target genes involved in lipid metabolism and inflammation [37-40] . Altering the lipid composition of lipoprotein particles can also contribute to modulation of the lipoprotein particles [15], including altered spatial distribution of lipids and therefore also alternation of the function [41,42].

In conclusion, fish oil supplementation for three and seven weeks alter the plasma lipidomic profile markedly compared to intake of high-oleic sunflower oil. The selective elevation of TGs and phospholipids of high carbon number and double bond content may represent beneficial effects of fish oil supplementation in healthy subjects. Future studies are needed in order to elucidate the health benefits of incorporation of long-chain PUFAs into selective phospholipids classes and TGs.

Conflicts of interest:

Inger Ottestad, Sahar Hassani, Grethe I. Borge, Achim Kohler, Gjermund Vogt, Tuulia Hyötyläinen, Matej Oresic, Kirsten B. Holven, Stine M. Ulven and Mari CW Myhrstad have "no conflicts of interest", or any financial or personal interest.

Kirsti W. Brønner is clinical nutritionist/Project manager in TINE SA R&D Center, Norway, with no financial interest.

***Contributions of authors*:**

Conception and design: IO, GIB, GV, KWB, KBH, SMU, MCWM

Data analysis and interpretation: IO, SH, GIB, AK, TH, MO, KBH, SMU, MCWM

Drafting the article: IO, SH, AK, MCWM

Revising the manuscript critically: IO, SH, GIB, AK, GV, TH, MO, KWB, KBH, SMU, MCWM

Final approval of the version to be published: IO, SH, GIB, AK, GV, TH, MO, KWB, KBH, SMU, MCWM

References

1. Skeaff CM, Miller J (2009) Dietary fat and coronary heart disease: summary of evidence from prospective cohort and randomised controlled trials. Ann Nutr Metab 55: 173-201.
2. Kris-Etherton PM, Harris WS, Appel LJ (2003) Fish consumption, fish oil, omega-3 fatty acids, and cardiovascular disease. Arterioscler Thromb Vasc Biol 23: e20-30.
3. Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico (1999) Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-Prevenzione trial. Lancet 354: 447-455.
4. Yokoyama M, Origasa H, Matsuzaki M, Matsuzawa Y, Saito Y, et al. (2007) Effects of eicosapentaenoic acid on major coronary events in hypercholesterolaemic patients (JELIS): a randomised open-label, blinded endpoint analysis. Lancet 369: 1090-1098.
5. Hartweg J, Perera R, Montori V, Dinneen S, Neil HA, et al. (2008) Omega-3 polyunsaturated fatty acids (PUFA) for type 2 diabetes mellitus. Cochrane Database Syst Rev: CD003205.
6. Harris WS (1997) n-3 fatty acids and serum lipoproteins: human studies. Am J Clin Nutr 65: 1645S-1654S.
7. Geleijnse JM, Giltay EJ, Grobbee DE, Donders AR, Kok FJ (2002) Blood pressure response to fish oil supplementation: metaregression analysis of randomized trials. J Hypertens 20: 1493-1499.
8. Knapp HR (1997) Dietary fatty acids in human thrombosis and hemostasis. Am J Clin Nutr 65: 1687S-1698S.
9. Hornstra G (2001) Influence of dietary fat type on arterial thrombosis tendency. J Nutr Health Aging 5: 160-166.
10. Christensen JH, Gustenhoff P, Korup E, Aaroe J, Toft E, et al. (1997) n-3 polyunsaturated fatty acids, heart rate variability and ventricular arrhythmias in post-AMI-patients. A clinical controlled trial. Ugeskr Laeger 159: 5525-5529.
11. Nodari S, Metra M, Milesi G, Manerba A, Cesana BM, et al. (2009) The role of n-3 PUFAs in preventing the arrhythmic risk in patients with idiopathic dilated cardiomyopathy. Cardiovasc Drugs Ther 23: 5-15.
12. Gross RW, Han X (2011) Lipidomics at the interface of structure and function in systems biology. Chem Biol 18: 284-291.
13. Seppanen-Laakso T, Oresic M (2009) How to study lipidomes. J Mol Endocrinol 42: 185-190.
14. Quehenberger O, Armando AM, Brown AH, Milne SB, Myers DS, et al. (2010) Lipidomics reveals a remarkable diversity of lipids in human plasma. J Lipid Res 51: 3299-3305.
15. Kontush A, Chapman MJ (2010) Lipidomics as a tool for the study of lipoprotein metabolism. Curr Atheroscler Rep 12: 194-201.
16. Fekete K, Marosvolgyi T, Jakobik V, Decsi T (2009) Methods of assessment of n-3 long-chain polyunsaturated fatty acid status in humans: a systematic review. Am J Clin Nutr 89: 2070S-2084S.
17. Lankinen M, Schwab U, Kolehmainen M, Paananen J, Poutanen K, et al. (2011) Whole grain products, fish and bilberries alter glucose and lipid metabolism in a randomized, controlled trial: the sysdimet study. PLoS ONE 6: e22646.
18. Lankinen M, Schwab U, Erkkila A, Seppanen-Laakso T, Hannila ML, et al. (2009) Fatty fish intake decreases lipids related to inflammation and insulin signaling--a lipidomics approach. PLoS ONE 4: e5258.
19. Harris WS (1989) Fish oils and plasma lipid and lipoprotein metabolism in humans: a critical review. J Lipid Res 30: 785-807.

20. Hodge AM, Simpson JA, Gibson RA, Sinclair AJ, Makrides M, et al. (2007) Plasma phospholipid fatty acid composition as a biomarker of habitual dietary fat intake in an ethnically diverse cohort. Nutr Metab Cardiovasc Dis 17: 415-426.
21. Schwab U, Seppanen-Laakso T, Yetukuri L, Agren J, Kolehmainen M, et al. (2008) Triacylglycerol fatty acid composition in diet-induced weight loss in subjects with abnormal glucose metabolism--the GENOBIN study. PLoS ONE 3: e2630.
22. Rhee EP, Cheng S, Larson MG, Walford GA, Lewis GD, et al. (2011) Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. J Clin Invest 121: 1402-1411.
23. Ottestad I, Vogt G, Retterstol K, Myhrstad MC, Haugen JE, et al. (2011) Oxidised fish oil does not influence established markers of oxidative stress in healthy human subjects: a randomised controlled trial. Br J Nutr: doi:10.1017/S0007114511005484
24. Pluskal T, Castillo S, Villar-Briones A, Oresic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11: 395.
25. Wangen LE, Kowalski BR (1989) A multiblock partial least squares algorithm for investigating complex chemical systems. Journal of Chemometrics 3: 3-20.
26. Hassani S, Martens H, Qannari M, Hanafi M, Kohler A (2011) Model validation and error estimation in multi-block partial least squares regression Chemometrics and Intelligent Laboratory Systems 10.1016/j.chemolab.2011.06.001.
27. McCombie G, Browning LM, Titman CM, Song M, Shockcor J, et al. (2009) Omega-3 oil intake during weight loss in obese women results in remodelling of plasma triglyceride and fatty acids. Metabolomics 5: 363-374.
28. Block RC, Duff R, Lawrence P, Kakinami L, Brenna JT, et al. (2010) The effects of EPA, DHA, and aspirin ingestion on plasma lysophospholipids and autotaxin. Prostaglandins Leukot Essent Fatty Acids 82: 87-95.
29. Lagarde M, Bernoud N, Brossard N, Lemaitre-Delaunay D, Thies F, et al. (2001) Lysophosphatidylcholine as a preferred carrier form of docosahexaenoic acid to the brain. J Mol Neurosci 16: 201-204; discussion 215-221.
30. Hammad SM (2011) Blood sphingolipids in homeostasis and pathobiology. Adv Exp Med Biol 721: 57-66.
31. Hammad SM, Pierce JS, Soodavar F, Smith KJ, Al Gadban MM, et al. (2010) Blood sphingolipidomics in healthy humans: impact of sample collection methodology. J Lipid Res 51: 3074-3087.
32. Pfeiffer A, Bottcher A, Orso E, Kapinsky M, Nagy P, et al. (2001) Lipopolysaccharide and ceramide docking to CD14 provokes ligand-specific receptor clustering in rafts. Eur J Immunol 31: 3153-3164.
33. de Mello VD, Lankinen M, Schwab U, Kolehmainen M, Lehto S, et al. (2009) Link between plasma ceramides, inflammation and insulin resistance: association with serum IL-6 concentration in patients with coronary heart disease. Diabetologia 52: 2612-2615.
34. Shearer GC, Savinova OV, Harris WS (2011) Fish oil - How does it reduce plasma triglycerides? Biochim Biophys Acta. doi:10.1016/j.bbalip.2011.10.011
35. Calder PC (2006) n-3 polyunsaturated fatty acids, inflammation, and inflammatory diseases. Am J Clin Nutr 83: 1505S-1519S.
36. Sijben JW, Calder PC (2007) Differential immunomodulation with long-chain n-3 PUFA in health and chronic disease. Proc Nutr Soc 66: 237-259.
37. Jump DB (2008) n-3 polyunsaturated fatty acid regulation of hepatic gene transcription. Curr Opin Lipidol 19: 242-247.

38. Tontonoz P, Spiegelman BM (2008) Fat and beyond: the diverse biology of PPARgamma. Annu Rev Biochem 77: 289-312.

39. Kliewer SA, Sundseth SS, Jones SA, Brown PJ, Wisely GB, et al. (1997) Fatty acids and eicosanoids regulate gene expression through direct interactions with peroxisome proliferator-activated receptors alpha and gamma. Proc Natl Acad Sci USA 94: 4318-4323.

40. Hong C, Tontonoz P (2008) Coordination of inflammation and metabolism by PPAR and LXR nuclear receptors. Curr Opin Genet Dev 18: 461-467.

41. Yetukuri L, Soderlund S, Koivuniemi A, Seppanen-Laakso T, Niemela PS, et al. (2010) Composition and lipid spatial distribution of HDL particles in subjects with low and high HDL-cholesterol. J Lipid Res 51: 2341-2351.

42. Yetukuri L, Huopaniemi I, Koivuniemi A, Maranghi M, Hiukka A, et al. (2011) High density lipoprotein structural changes and drug response in lipidomic profiles following the long-term fenofibrate therapy in the FIELD substudy. PLoS ONE 6: e23589.

Table 1 BMI and serum lipids at baseline and after three weeks of intervention with fish oil (n=16) and high oleic sunflower oil (n=17)

	Fish oil		Sunflower oil			
Parameter	Baseline	3 wk	Baseline	3 wk	P-value*	P-value**
BMI (kg/m^2)	22 ± 3	22 ± 3	23 ± 3	23 ± 3	0.25	0.53
Triglycerides (mmol/l)	0.9 ± 0.4	0.9 ± 0.3	1.1 ± 0.7	1.1 ± 0.4	0.46	0.68
Total-cholesterol (mmol/l)	4.6 ± 0.8	4.4 ± 0.6	4.9 ± 0.9	4.6 ± 1.0	0.27	0.36
LDL-cholesterol (mmol/l)	2.5 ± 0.8	2.4 ± 0.8	2.7 ± 0.6	2.6 ± 0.6	0.35	0.64
HDL-cholesterol (mmol/l)	1.5 ± 0.3	1.4 ± 0.4	1.5 ± 0.4	1.4 ± 0.4	1	0.86

* Independent t- test for between groups at baseline

** Independent t- test for changes between groups after three weeks

Table 2 Significantly altered lipids (FDR < 0.05) in the fish oil (FO) group compared to the sunflower oil group (HOSO) after three weeks of intervention

		Fold change from baseline	
Lipid Name	**q-value**	**HOSO**	**FO**
LysoPC(20:5)	<0.001	0.80	4.35
LysoPC(22:5)	0.025	0.91	1.67
LysoPC(22:6)	0.003	0.94	1.89
PA(38:5e)	0.006	1.05	0.75
PE(38:4)	0.026	1.11	0.88
PE(38:4)+PC(35:4)	0.029	1.27	0.90
PE(38:5)	<0.001	0.92	3.29
PE(38:5e)	0.042	1.09	0.75
PE(38:7e)	<0.001	1.17	2.58
PE(38:7e)	0.013	1.11	1.40
PE(40:4)	0.027	1.10	0.78
PE(40:6)	0.013	1.10	1.75
PE(40:7e)	0.010	1.08	1.51
PG(36:2)	0.015	0.96	0.68
PG(36:5e)	0.028	1.02	0.80
PG(38:4)	0.013	1.24	0.89
PG(40:6)	0.003	1.11	1.91
PI(40:7)	0.024	1.25	0.96
PS(36:1)	0.001	1.14	1.84
PS(38:0)	0.031	1.18	1.67
PS(38:1)	0.010	1.11	1.58
PS(38:1)	0.019	1.26	1.83
PS(41:5)	0.031	1.17	0.96
PS(42:1)	0.003	1.09	1.86
PS(42:6)	0.001	0.96	1.54
PS(42:7)	0.001	0.92	1.39
PS(42:8)	0.001	1.36	2.51
PS(44:1)	0.001	1.08	1.86
SM(d18:0/20:0)	0.029	1.08	0.81
SM(d18:0/22:6)	0.015	1.01	1.44
SM(d18:0/24:0)	<0.001	0.93	2.62
SM(d18:1/26:2)	0.015	1.10	1.43

Table 3 Significantly altered lipids (FDR < 0.05) in the fish oil (FO) group compared to the sunflower oil (HOSO) group after three weeks of intervention

		Fold change from baseline	
Lipid Name	**q-value**	**HOSO**	**FO**
PC(30:3)	0.015	0.85	0.6
PC(32:5)	0.024	1.03	0.77
PC(36:3)	0.031	0.97	0.73
PC(36:5)	<0.001	0.80	4.00
PC(37:4)/PE(40:4)	0.021	1.07	0.83
PC(38:1)	0.025	1.05	1.60
PC(38:1e)	0.026	0.91	1.39
PC(38:4)	0.007	2.35	9.08
PC(38:5)	<0.001	1.12	3.97
PC(38:5e)	0.006	1.2	0.96
PC(38:6)	0.006	1.15	1.52
PC(38:6)	0.037	0.89	0.69
PC(38:7)	0.001	1.09	2.27
PC(38:7)	0.001	1.09	2.25
PC(40:2)	<0.001	0.99	3.54
PC(40:3)	0.001	0.93	1.74
PC(40:4)	0.029	1.09	0.79
PC(40:4e)	0.015	1.01	0.78
PC(40:5)	<0.001	0.89	1.57
PC(40:6)	<0.001	1.04	1.71
TG(50:4)	<0.001	1.04	3.57
TG(52:0)	0.001	1.03	2.52
TG(52:2)	0.029	1.09	0.79
TG(52:6)	0.004	0.96	2.86
TG(52:7)	<0.001	0.80	4.55
TG(54:2)	0.001	1.27	0.59
TG(54:3)	0.029	1.26	0.83
TG(54:4)	<0.001	0.96	2.34
TG(54:4)	0.003	0.97	1.73
TG(54:5)	0.022	1.51	0.82
TG(54:8)	<0.001	0.98	5.02
TG(56:2)	0.026	1.02	1.93
TG(56:4)	0.038	4.50	1.47
TG(56:7)	<0.001	1.09	3.02
TG(56:8)	<0.001	1.27	2.71
TG(56:9)	<0.001	1.01	4.76
TG(58:10)	<0.001	1.17	4.14
TG(58:6)	0.003	1.17	2.10
TG(58:6)	0.023	0.99	1.70
TG(58:8)	<0.001	1.36	4.07
TG(58:9)	0.001	1.48	4.05
TG(59:2)	<0.001	1.08	2.73

Supplementary Table 1 Significantly altered lipids (Multivariate analyses, p<0.05) in the fish oil (FO) group compared to the sunflower oil group (HOSO) after three weeks intervention. The corresponding q-values from univariate analyses are also given.

	Multivariate analyses	**Univariate analyses**	**Fold change**	**Fold change**
lipid	**p-value for 2PCs model**	**q-value**	**HOSO group**	**FO group**
PE(38:5)	0,0000	0,0000	0,92	3,29
PC(36:5)	0,0000	0,0000	0,80	4,00
LysoPC(20:5)	0,0000	0,0000	0,80	4,35
TG(54:8)	0,0000	0,0000	0,98	5,02
TG(52:7)	0,0000	0,0000	0,80	4,55
TG(56:9)	0,0000	0,0000	1,01	4,76
SM(d18:0/24:0)	0,0000	0,0000	0,93	2,62
TG(54:4)	0,0000	0,0031	0,97	1,73
TG(50:4)	0,0000	0,0000	1,04	3,57
TG(58:10)	0,0000	0,0000	1,17	4,14
TG(59:2)	0,0001	0,0001	1,08	2,73
PC(40:2)	0,0001	0,0001	0,99	3,54
PS(41:5)	0,0002	0,0305	1,17	0,96
PE(40:7e)	0,0003	0,4802	1,09	1,05
LysoPC(22:6)	0,0004	0,0031	0,94	1,89
PS(40:0)	0,0005	0,2308	1,39	1,19
PC(37:4)/PE(40:4)	0,0005	0,0205	1,07	0,83
TG(58:8)	0,0005	0,0001	1,36	4,07
TG(56:8)	0,0006	0,0004	1,27	2,71
PC(40:5)	0,0006	0,0001	0,89	1,57
PC(40:4e)	0,0007	0,1938	1,39	1,17
PC(38:5)	0,0008	0,1315	1,01	1,17
PE(38:7e)	0,0013	0,0004	1,17	2,58
TG(52:0)	0,0014	0,0006	1,03	2,52
PC(38:5e)	0,0017	0,0063	1,20	0,96
PC(34:2)	0,0017	0,1374	0,98	0,88
TG(56:7)	0,0019	0,0000	1,09	3,02
TG(58:9)	0,0019	0,0005	1,48	4,05
PC(38:7)	0,0022	0,0009	1,09	2,27
TG(58:6)	0,0024	0,0226	0,99	1,70
PE(38:4)	0,0025	0,0255	1,11	0,88
PC(38:7)	0,0027	0,0011	1,09	2,25
PE(38:7e)	0,0028	0,0126	1,11	1,40
PC(40:3)	0,0028	0,0005	0,93	1,74
PC(40:6)	0,0029	0,0000	1,04	1,71
PC(36:3)	0,0029	0,0305	0,97	0,73

	Multivariate analyses	Univariate analyses	Fold change	Fold change
lipid	p-value for 2PCs model	q-value	HOSO group	FO group
TG(56:2)	0,0033	0,0255	1,02	1,93
TG(52:6)	0,0036	0,0037	0,96	2,86
PE(40:6)	0,0038	0,1359	1,39	1,92
PE(40:4)	0,0042	0,0267	1,10	0,78
TG(52:2)	0,0062	0,2738	1,13	1,26
PC(38:4)	0,0063	0,0068	2,35	9,08
SM(d18:0/22:6)	0,0066	0,0152	1,01	1,44
PA(34:0e)	0,0068	0,4817	1,05	1,05
PC(40:4)	0,0087	0,0292	1,09	0,79
PE(36:3e)	0,0100	0,0508	0,85	0,52
TG(54:3)	0,0102	0,0290	1,26	0,83
PC(36:2)	0,0125	0,0706	0,94	0,83
SM(d18:1/24:1)	0,0125	0,1095	1,00	1,20
PG(40:6)	0,0147	0,0032	1,11	1,91
PC(38:5e)	0,0171	0,0610	1,38	1,06
PC(38:6)	0,0177	0,0372	0,89	0,69
SM(d18:0/20:0)	0,0194	0,0292	1,08	0,81
PC(38:3e)	0,0213	0,0766	0,98	0,75
PC(34:3e)+PE(37:3e)	0,0218	0,1910	1,03	0,92
PC(38:6)	0,0237	0,0060	1,15	1,52
TG(18:1/18:1/22:1)+ TG(20:1/20:1/18:1)	0,0251	0,1983	1,23	1,99
PE(40:3)	0,0251	0,1076	1,12	0,81
TG(56:4)	0,0259	0,0379	4,50	1,47
PG(36:2)	0,0263	0,0145	0,96	0,68
PC(30:3)	0,0264	0,0152	0,85	0,60
TG(54:5)	0,0287	0,0222	1,51	0,82
SM(d18:1/26:2)	0,0342	0,0152	1,10	1,43
SM(d18:1/16:1)	0,0346	0,2927	0,97	0,91
PE(38:3)	0,0363	0,0856	1,14	0,82
PC(38:4)	0,0363	0,1333	1,13	0,99
PE(38:4e)	0,0365	0,1032	1,15	0,93
PC(36:4e)	0,0386	0,2308	1,21	1,03
PE(40:6)	0,0409	0,1486	1,27	1,56
PC(38:3)	0,0413	0,0779	0,97	0,75
TG(16:0/18:2/18:1)	0,0465	0,1614	1,22	0,99
PS(40:0)	0,0466	0,1614	1,12	1,30
TG(54:4)	0,0475	0,0000	0,96	2,34
SM(d18:1/22:0)	0,0495	0,2535	0,96	1,12

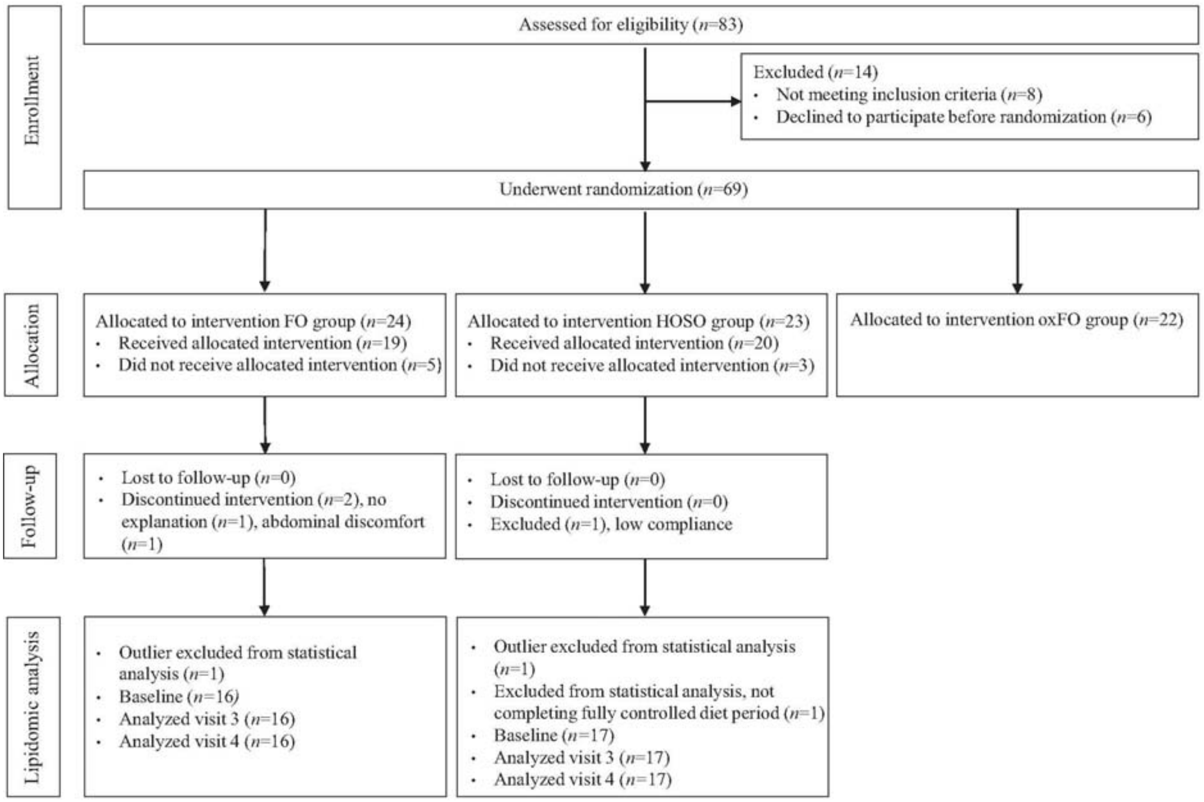

Enrollment
Assessed for eligibility (n=83)
Excluded (n=14)
Not meeting inclusion criteria (n=8)
Declined to participate before randomization (n=6)
Underwent randomization (n=69)
Allocation
Allocated to intervention FO group (n=24)
Received allocated intervention (n=19)
Did not receive allocated intervention (n=5)
Allocated to intervention HOSO group (n=23)
Received allocated intervention (n=20)
Did not receive allocated intervention (n=3)
Allocated to intervention oxFO group (n=22)
Follow-up
Lost to follow-up (n=0)
Discontinued intervention (n=2), no explanation (n=1), abdominal discomfort (n=1)
Lost to follow-up (n=0)
Discontinued intervention (n=0)
Excluded (n=1), low compliance
Lipidomic analysis
Outlier excluded from statistical analysis (n=1)
Baseline (n=16)
Analyzed visit 3 (n=16)
Analyzed visit 4 (n=16)
Outlier excluded from statistical analysis (n=1)
Excluded from statistical analysis, not completing fully controlled diet period (n=1)
Baseline (n=17)
Analyzed visit 3 (n=17)
Analyzed visit 4 (n=17)

Figure 1. Flow chart of the study showing subjects enrolled, lost during follow-up and number of subjects included in the statistical analysis at baseline and after three and seven weeks of fish oil supplementation. FO group, fish oil group; HOSO, high oleic sunflower oil group; oxFO, oxidized fish oil group (not included in the present study).

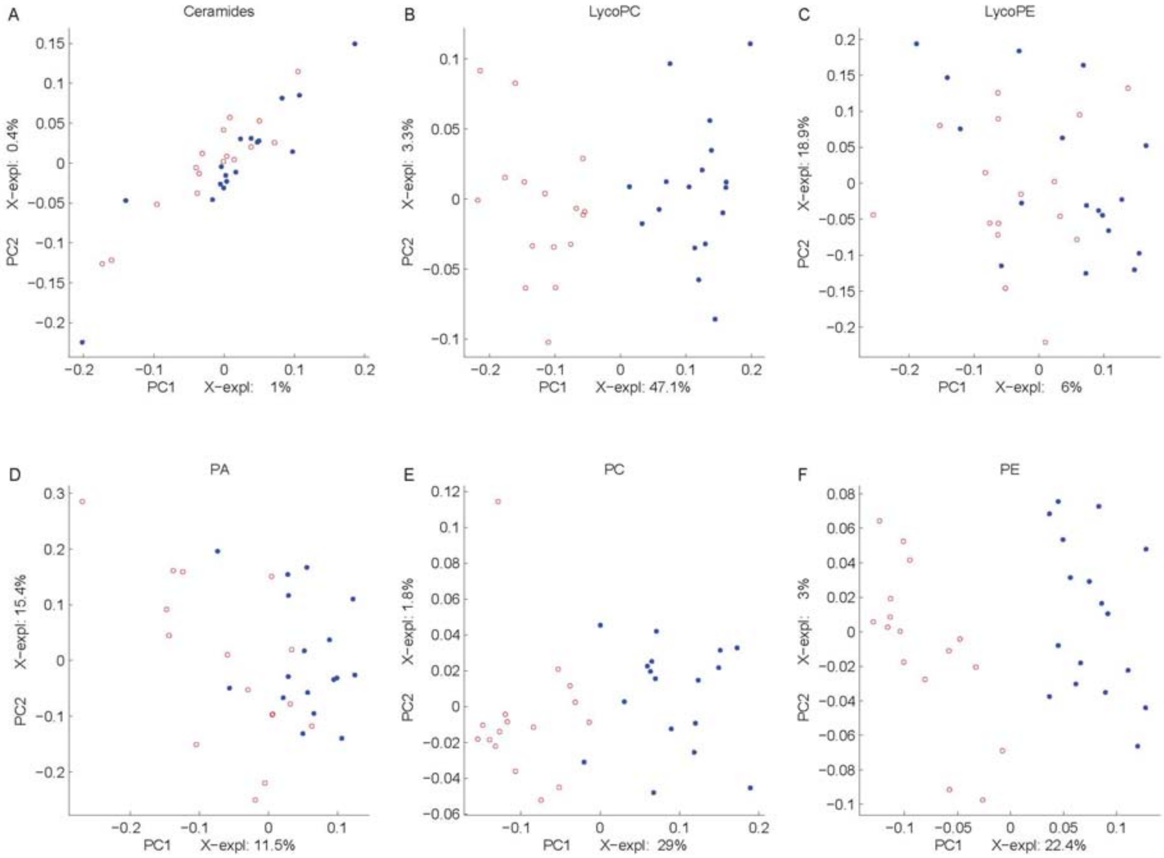


Figure 2. **Multi-Block Partial Least Squares Regression (MBPLSR) analysis of the data after three weeks of intervention.** First and second PLSR components of block scores of ceramides, lysoPC, lysoPE, PA, PC and PE are shown (A-F). The samples of each intervention group are presented as blue (HOSO group) or red (FO group) circles. The (un-validated) explained variances are shown on the axes.

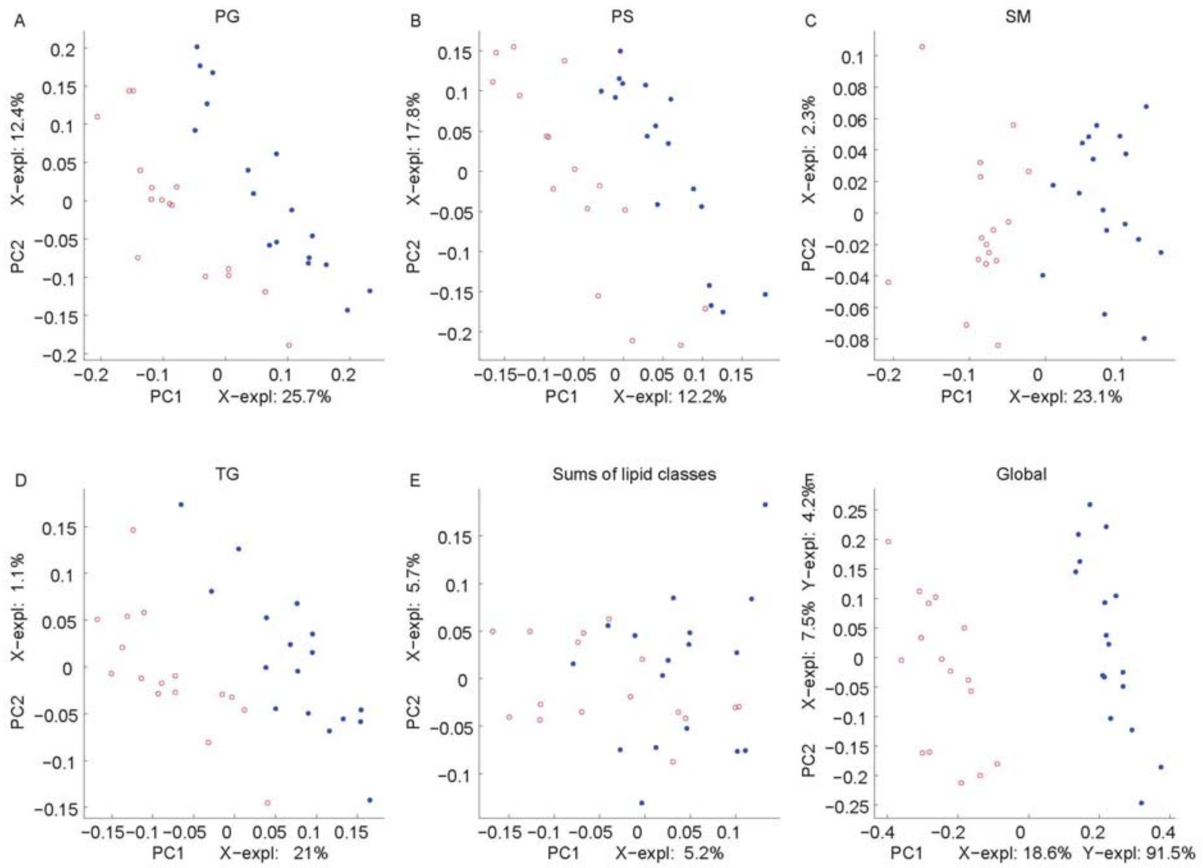


Figure 3. **Multi-Block Partial Least Squares Regression (MBPLSR) analysis of the data after three weeks of intervention.** First and second PLSR components of block scores of PG, PS, SM, TG, the sums of lipid classes and global scores are shown (A-F). The samples of each intervention group are presented as blue (HOSO group) or red (FO group) circles. The (un-validated) explained variances are shown on the axes.

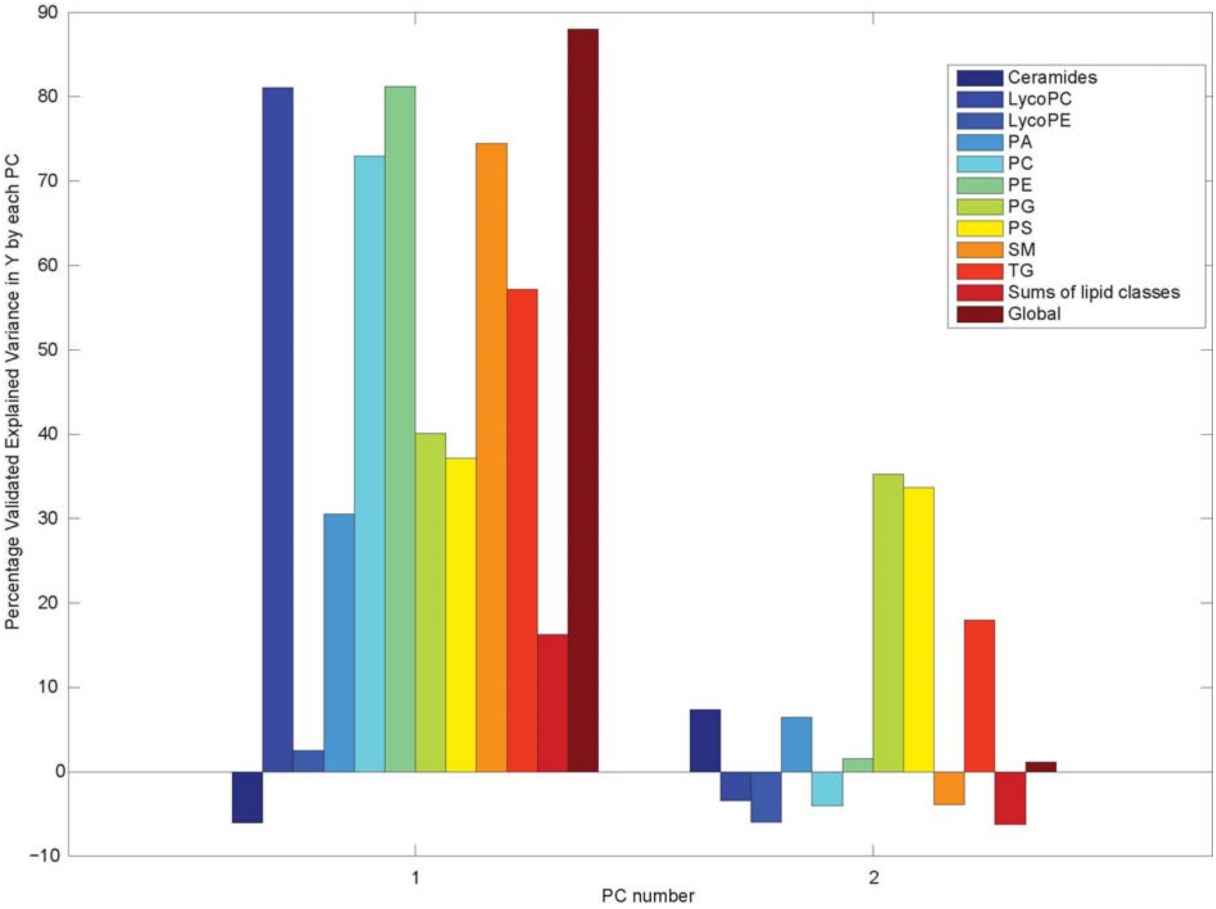


Figure 4. Cross-validated explained variance in Y. Bar plots of the validated explained variances in **Y** for each block and for the global model using data obtained after three weeks of intervention are presented.

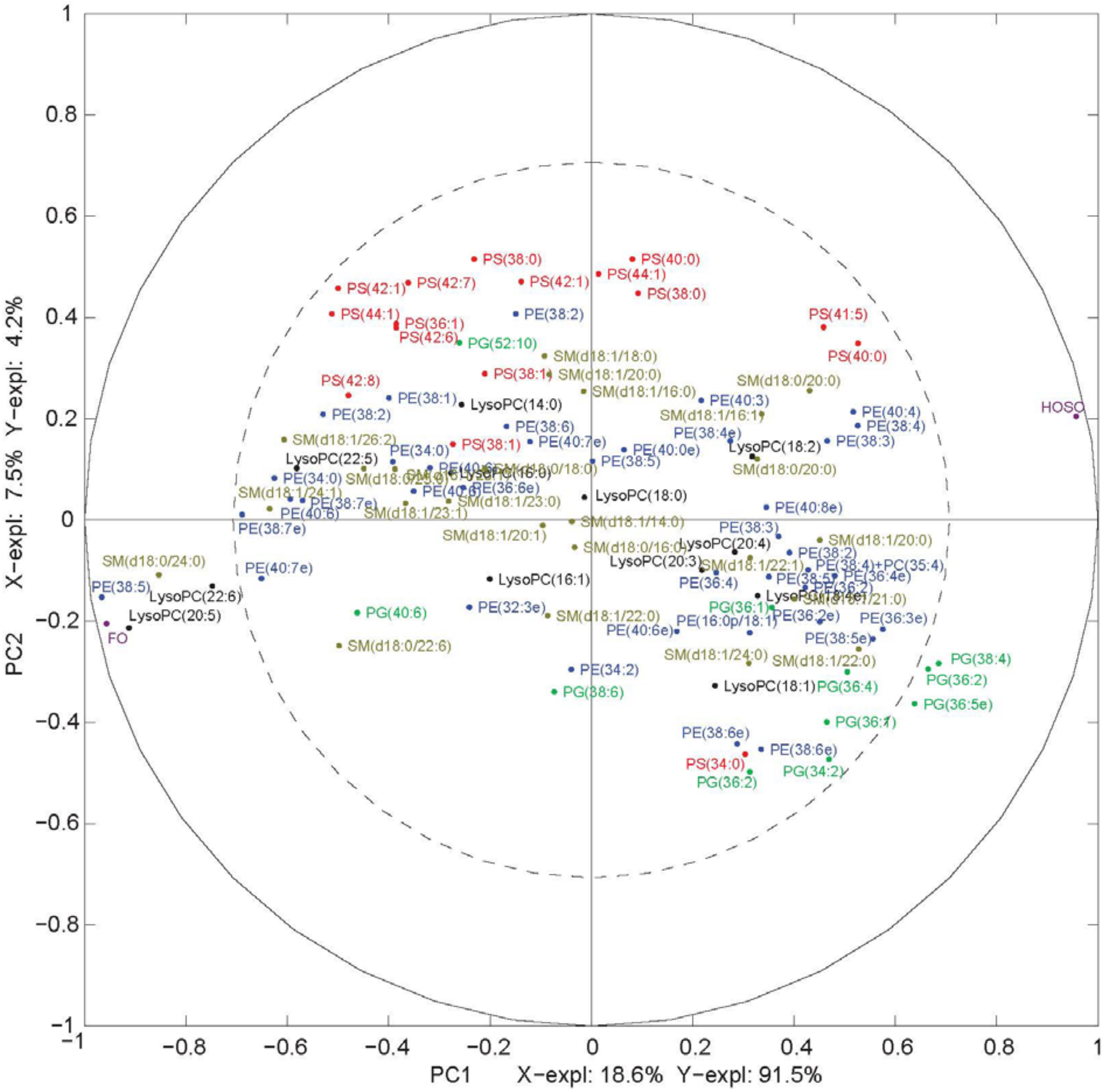


Figure 5. Multi-Block Partial Least Squares Regression (MBPLSR) analysis of the data after three weeks of intervention. Correlation loading plot for the variables contributing to the separation of the FO and the HOSO group after three weeks are shown for LycoPC, PE, PG, PS and SM. The (un-validated) explained variances in **X** and **Y** are shown on the axes.

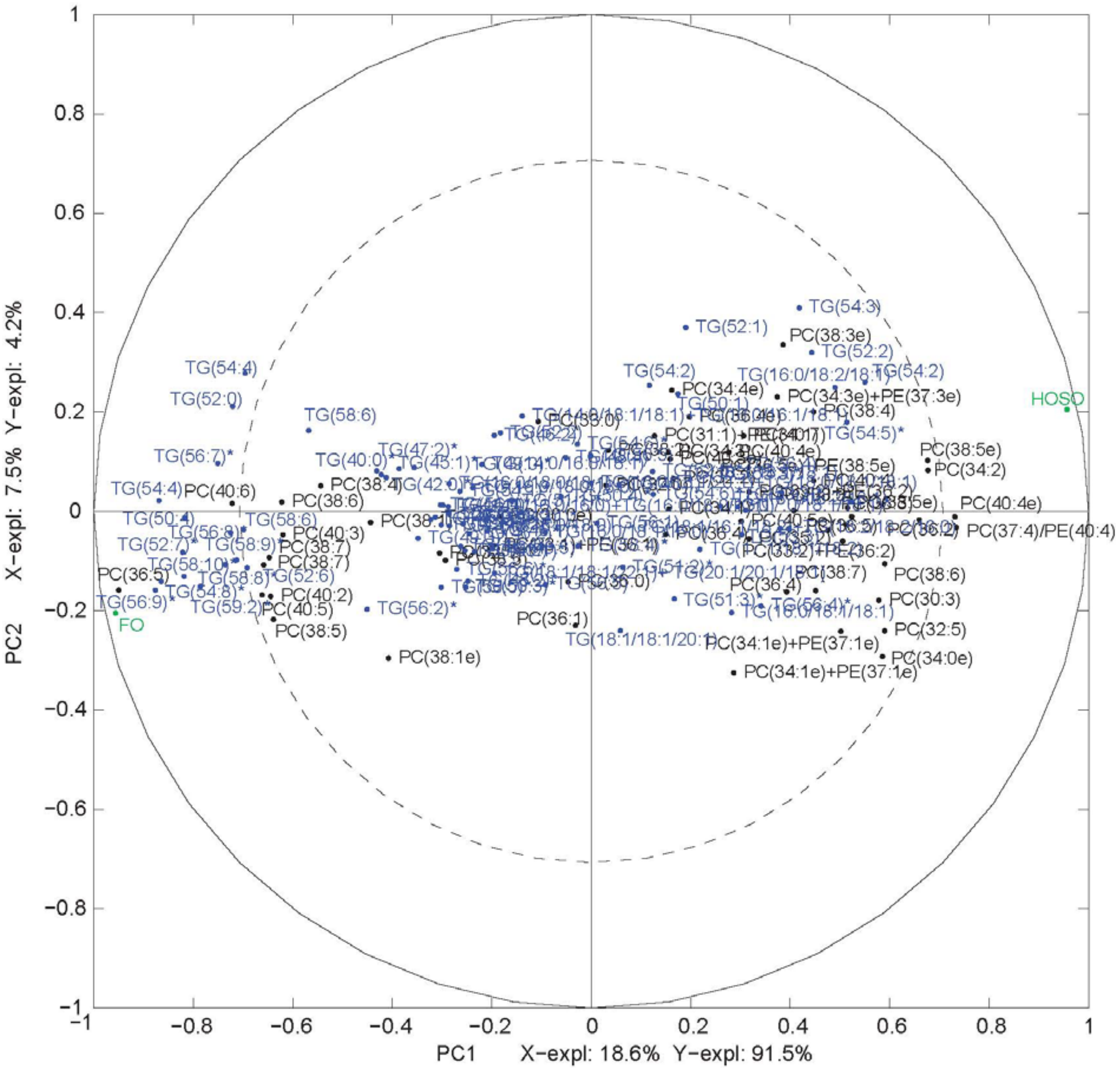
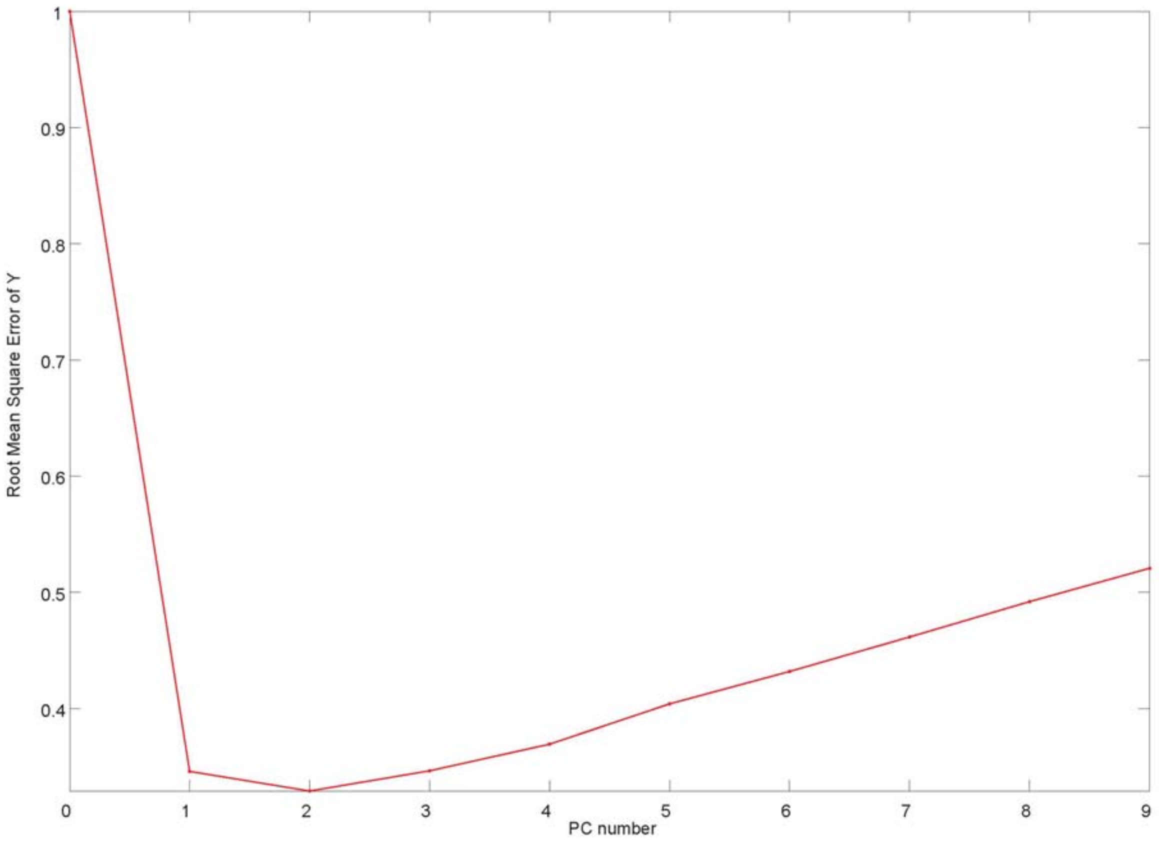


Figure 6. Multi-Block Partial Least Squares Regression (MBPLSR) analysis of the data after three weeks of intervention. Correlation loading plot for the variables contributing to the separation of the FO and the HOSO group after three weeks are shown for TG and PC. The (un-validated) explained variances in **X** and **Y** are shown on the axes.



Supplementary material Figure S1. Global Root Mean Square Error plot of Y (RMSE$_Y$). RMSE of Y for the global model is plotted as a function of the number of components. Detailed explanation for the plot is given in [26].

Erratum

Dedication page added.
Page 6: Status of papers IV and VI changed to “*Under revision*”.
Paper III, page 24: “the grant 203699 (New statistical tools for integrating and exploiting complex genomic and phenotypic data sets) financed by the Research Council of Norway” added to the Acknowledgements.
Paper IV, page 1: Mohamed Hanafi’s address changed from “1,2” to “3”.
Paper IV, page 16: “and for financial support by the Nordic Centre of Excellence on Food, Nutrition and Health “Systems biology in controlled dietary interventions and cohort studies” (SYSDIET) funded by NordForsk” added to the Acknowledgements.
Paper V, page 7, paragraph 5: “columns” changed to “rows”.
Paper V, page 35, line 8: “b” changed to “d”.
Paper V and VI: Text justified.
.