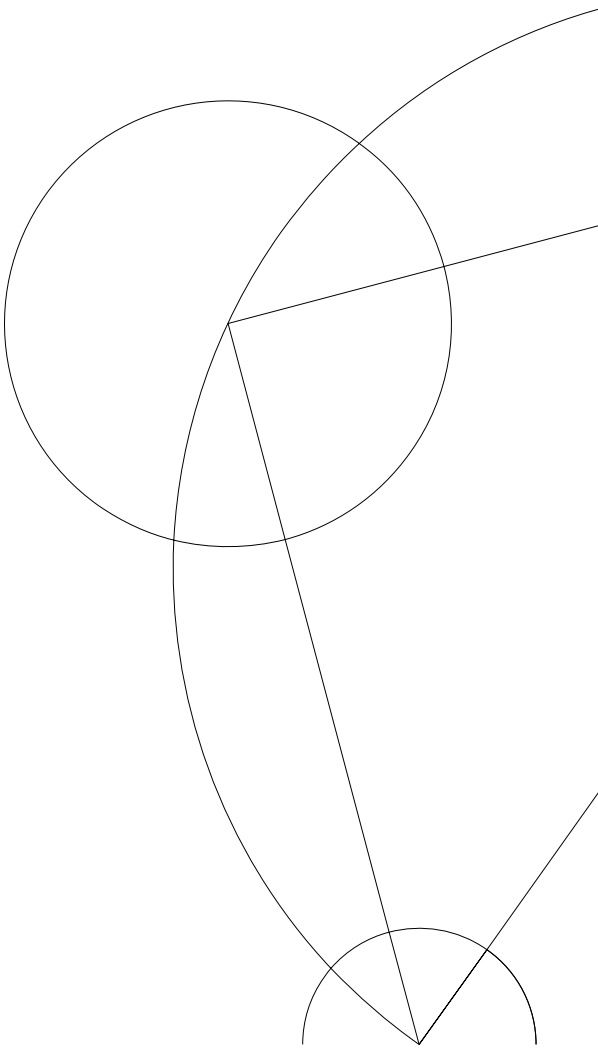
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Towards improved biomarker research

Data analytical challenges of high-dimensional biological data



PhD thesis

Karin Kjeldahl

Supervisor: Prof Rasmus Bro

2013

Abstract

Towards improved biomarker research: Data analytical challenges of large-scale high-dimensional biological data

This thesis takes a look at the data analytical challenges associated with the search for biomarkers in large-scale biological data such as transcriptomics, proteomics and metabolomics data. These studies aim to identify genes, proteins or metabolites which can be associated with e.g. a diet, disease (e.g. cancer), drug response or physiological status.

The value of these *omics* studies has to some extent been questioned as it is often observed that the validity of claimed biomarkers has been very difficult to verify in other studies. On the other hand, in many studies it is difficult to actually identify strong biomarkers when strict validation is applied; the latter phenomenon is to some extent masked by a publication bias, but has been widely observed among researchers working with *omics* data.

In this thesis, the background of this apparent small effect size of the biomarkers is investigated and followed by some suggestions which can potentially improve the chances of a successful outcome of an omics study. A method widely applied in the analysis of omics studies is Partial Least Squares (PLS) regression which is one of the work horses within the chemometrics tool box; a method which is used both for regression and classification purposes. This

method has proven its strong worth in the multivariate data analysis throughout an enormous range of applications; a very classic data type is near infrared (NIR) data, but many similar data types have also be very successful.

On that background, the general characteristics of omics data are described and related to the characteristics of classical NIR-type data. This shows that omics data, which are generally much bigger data sets than classical data, are not just simple extensions of NIR data. The sample type, analytical method and the application types are different and introduce a larger complexity, weaker signals and many potential sources of experimental and analytical bias and errors. The risk of the latter is further increased by the complexity of the entire omics experimental setup which often involves various project partners with very specific competencies.

In order to optimize the basis of a sound and fruitful data analysis, suggestions are given which focus on (1) collection of good data, (2) preparation of data for the data analysis and (3) a sound data analysis. If these steps are optimized, PLS is a also a very good method for the analysis of *omics* data.

The five research papers included in the thesis touch upon different aspects of the issues discussed in the thesis.

Resume

Mod bedre biomarkørforskning: Dataanalytiske udfordringer i højdimensionelle biologiske data

Denne afhandling ser på nogle af de dataanalytiske udfordringer, der findes i forbindelse med jagten på biomarkører i store biologiske datasæt såsom transcriptomics, proteomics og metabolomicsdata. Denne forskning har til formål at identificere gener, proteiner eller metabolitter, som kan sige noget om f.eks. kost, sygdom (f.eks. kræft), lægemiddelrespons eller en fysiologisk status.

Værdien af disse *omics*-studier er der i nogen grad blevet sat spørgsmålstegn ved, da det ofte er konstateret, at gyldigheden af påståede biomarkører har været meget vanskeligt at verificere i andre undersøgelser. På den anden side er det i mange studier svært rent faktisk at identificere stærke biomarkører, når god validering er anvendt. Det sidstnævnte fænomen er til en vis grad maskeret af en publikationsbias, men er ofte observeret blandt forskere, der arbejder med *omics*-data.

I denne afhandling er baggrunden for dette undersøgt, efterfulgt af nogle forslag, som potentielt kan forbedre chancerne for et vellykket resultat af et *omics* studie. En udbredt metode til analyse af omics data er Partial Least Squares (PLS) regression, som er en af arbejdshestene i kemometri - en metode, som bruges både til regression og klassifikation. Denne metode har bevist sit værd i den multivariate dataanalyse i form af en nærmest uendelig række af app-

likationer. En meget klassisk datatype er nærinfrarøde (NIR) data, men mange lignende datatyper har med succes været kombineret med PLS.

På den baggrund er de generelle kendetegn ved *omics* data beskrevet og relateret til de typiske træk af den klassiske NIR-type data. Dette viser, at *omics* data, som generelt inkluderer langt større datamængder end klassiske data, ikke bare er simple udvidelser af NIR data. Prøverne, analysemetoderne og applikationerne er forskellige og bidrager til en større kompleksitet og svagere signaler, og det betyder, at der er mange potentielle kilder til eksperimentelle og analytiske skævheder og fejl. Risikoen for sidstnævnte er yderligere forstærket af kompleksiteten i hele det eksperimentelle setup i *omics*-studier som ofte involverer forskellige projektpartnere med meget specifikke kompetencer.

I afhandlingen er der en række forslag som yder basis for en sund og frugtbar dataanalyse; disse fokuserer på(1) indsamling af gode data, (2) forberedelse af data til dataanalyse og (3) en sund dataanalyse. Hvis disse trin er optimerede er PLS også en meget velegnet metode til analyse af *omics* data.

De fem forskningsartikler, der indgår i afhandlingen, berører forskellige aspekter af de områder, der drøftes i afhandlingen.

Acknowledgements

I owe thanks to a lot of people who have helped and supported me while writing my dissertation. First and foremost, my supervisor Rasmus Bro whose help and inspiration has been priceless. Thanks for always having time for me, for constructive criticism, discussions and support and thanks for tremendous inspiration during the years.

Thanks to all my lovely colleagues who have made Q & T a warm and inspiring work-place. Peter Hansen, Hanne Winning, Morten Rasmussen, Mette Skau, Francesco Savorani and my office mate Thomas Skov, you contributed substantially to the social and academic dimension of my office life and I am very grateful for the endless discussions and your fabulous sense of humor. Lunch breaks will never be the same again.

Finally a special thanks to those nearest and dearest to me, Michael and Mads for endless support and for a priceless contribution in keeping the family up and running through the final stretches of my PhD. I am also deeply indebted to my sister for competent linguistic assistance throughout my PhD time.

Thanks to FOSS for providing near infrared feed data.

Contents

CHAPTER

1 Omics is difficult

1.1 What is omics?

The rapid development of omics technologies has created the possibility of utilizing these approaches to investigate the molecular complexity of biological systems and the effects induced in them by perturbations like disease, toxic substances, drugs, nutrition or other external intervening factors. The advances in high-throughput omics analytical technologies such as DNA microarrays, LC-MS, GC-MS, NMR or LC-NMR etc. have enabled the identification and quantification of many of the components of a particular biological system in a single experiment [55].

Conceptually, doing omics involves the study of "all" components of a biological system measured simultaneously. In practice though, "all" translates to "many", because the choice of analytical platform and of methodology dictate some degree of filtering. Many different omics disciplines exist, all referring to a study of some kind of totality within the specified biological domain. The major omics areas within the life sciences are:

- *Genomics*. Involves areas such as (1) DNA sequencing, (2) comparative genomics studies across organisms, or (3) functional genomics which seeks to assign functional terms to genes.
- *Transcriptomics*. mRNA is the object of study because it reflects which genes are expressed in a given situation.
- *Proteomics*. Involves the study of all the proteins present.
- *Metabolomics*. Deals with the available metabolites in a given body fluid or tissue. Metabolomics represents the omics level which is closest to phenotype. Potentially, this should provide good opportunities of identifying metabolites correlated with given phenotypes.

The four system levels and the associated omes are visualized in Figure 1.1. Genomics deals with sequencing and is fundamentally different in approach from the remaining three, which are functional studies of the principal biomolecules involved in proper functioning of the cell.

In systems biology, a relatively new biological research field, the interactions between different biological levels are studied by combining omics techniques. One example is the field of nutrigenomics which combines genomics, transcriptomics, proteomics and metabolomics with nutritional information with the purpose of obtaining an improved understanding of the interaction between genetics and the metabolism of foods. A hot topic in that area is personalized nutrition.

Terminology

Different definitions exist on the different omics; specifically, the use of the terms metabo*l*omics and metabo*n*omics is inconsistent in the literature, but the understanding of the two terms seems to converge [18]. In the present thesis, the term *omics* is used for large-scale, high rank biological data and I will stick to the term metabolomics when referring to omics studies on metabolites present in biofluids or tissue.

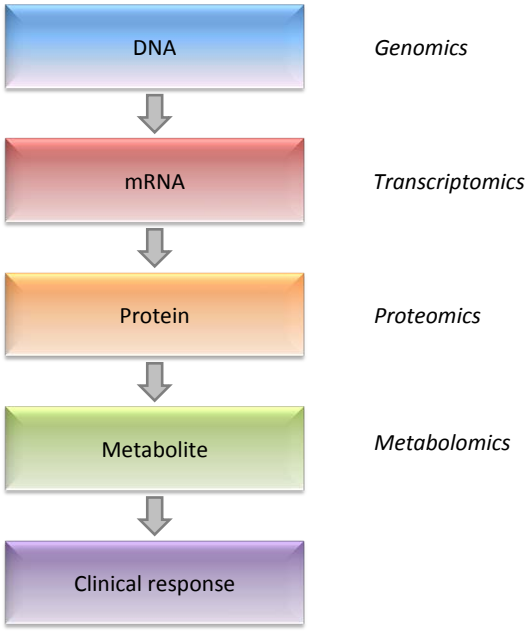


Figure 1.1: The levels of the bio system and the associated omics domains

Usually omics studies involve expertise from different scientific fields; in Figure 1.2 some key actors in the process are shown. Obviously, it is study dependent which people are involved. A generalized setting could be the following: THE BIOLOGIST (a microbiologist, a medical researcher, a molecular biologist or any scientist working within biology) got the idea to start the study and formulated the research questions. THE GENERAL PRACTITIONER (GP) checks questionnaires and takes care of sample collection storage and shipping. THE ANALYTICAL CHEMIST analyzes the sample at his state-of-the-art analytical facility, and THE DATA ANALYST analyzes the data. All these project partners

have a number of technicians, PhD students etc. associated. Throughout this thesis we will meet these members of the omics experiment again.

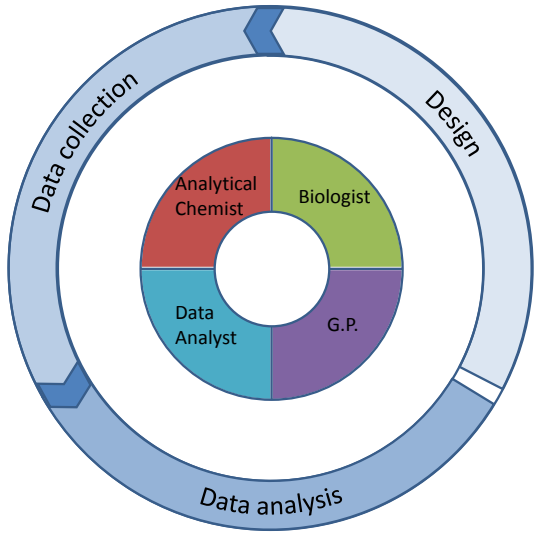


Figure 1.2: The omics experiment - processes and people involved. Each actor in the inner circle represents a whole group of people including e.g. (lab) technicians and a number of PhD or other students

1.2 Omics from the data analyst's perspective

In the following, four cases are used for illustration of some major challenges related to omics experiments that I have experienced during my work as a chemometrician. These cases are based on real ones. It is by no means the intention to criticize these particular studies; rather I believe the issues addressed in these cases can be encountered in many other omics applications, and it is there-

fore relevant to highlight them. I hope the people involved in these studies will recognize that my purpose is merely to illustrate potential pitfalls.

Case I: Transcriptomics

A molecular biologist investigating the human immune response system contacts the university chemometrics group for assistance with his transcriptomics data. His study is about allergic contact dermatitis and in the experiment, 7 patients allergic to nickel and 5 controls were exposed to nickel at the skin surface three times with a duration of 7, 48 and 96 hours respectively. In addition, all participants were exposed to the same type of plaster but without nickel for a period of 48 hours, termed "0 hours". After each experiment, skin biopsies were taken to evaluate transcription levels (mRNA) of practically the entire human genome. The design was a fractional full-factor design with a total of 34 points (full-factor: 4 x 12 = 48). The biologist wants to know about the genetic response in the patients when they are exposed to nickel.

The chemometrician's standard procedure in such cases is to do a Principal Component Analysis (PCA), and at first, the results are quite appealing. The patients and controls are quite well separated by the first principal component (PC) (Figure 1.3).

The interpretations come from the loadings, but plotting these does not give any clear answers (Figure 1.4).

It appears that the "0 hours" samples are also separated between the two groups and as a result it may be possible to identify genetic markers which can be used to diagnose nickel allergy without having to expose the patient to nickel. Consequently, a PLS-DA model is built and using one component, the two classes are completely separated in leave-one-out cross-validation. The regression vector is used to identify the top-ten of up- or down regulated transcripts; and by use of a database, functional terms associated with the identified transcripts are given. The chemometrician hands these results to the biologist who interprets the findings and together they publish the results. These appear

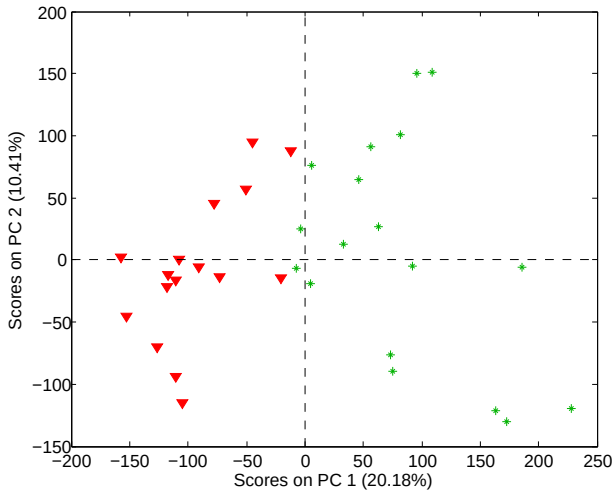


Figure 1.3: Scores plot from transcriptomics data, where the patients (green) and controls (red) are well separated

to be great findings but in a later experiment, it was not possible to verify these markers.

Case II: Proteomics for ovary cancer diagnostics

Case II is an example of a proteomics study dedicated to the discovery of biomarkers of ovarian cancer. A total of 256 serum samples from patients with ovarian cancer or benign pelvic conditions were subjected to proteome analysis by means of MALDI-TOF mass spectrometry (MS) following pre-fractionation.

After all the samples had been measured it became clear that reproducibility of the analytical procedure was too poor. The methodology was optimized and in the second round, this was improved considerably through a simplified approach.

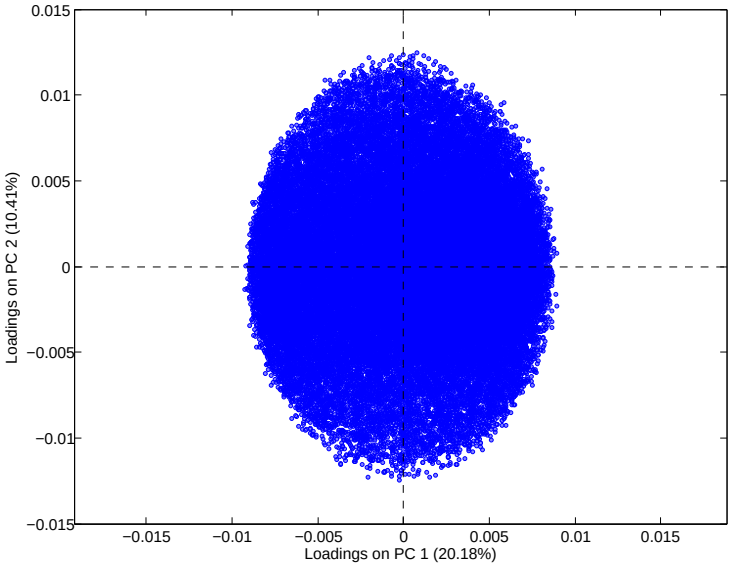


Figure 1.4: Loadings plot from transcriptomics data

When the data was subjected to chemometric analysis, the initial analysis invoked optimism as benign and malign samples were reasonably well separated. However, during the analysis, an important interfering storage effect became clear: The samples were transferred batch-wise from the clinical facilities to the mass spectrometry facilities. Whereas the samples were stored at -80°C at the hospital, they were stored at -20°C at the MS facilities because this was most convenient and it was assumed that the samples were safely preserved at -20°C.

Due to the diagnostic procedure in the study, sample batches transferred from the hospital were not randomized. As a result, storage time at the analytical facility was largely confounded with class membership and hence the classification model was mainly depicting sample aging.

Case III: Metabolomics - Twin data

Case III is based on the study published in PAPER IV. In this study, ^{1}H NMR metabolomic profiles of healthy Danish twins were combined with genotypic information about the same subjects to see if any associations between the two sets of information could be found.

The reader is referred to the paper (and references herein) for more details about the study. No significant associations could be found and this study is an example of a study where the starting point is difficult: The distance across several system levels (from genome to metabolome) is large in terms of mechanistic linking. The number of plausible interfering genetic and environmental effects in a healthy free-living population is very high, which shrinks possible underlying effect sizes to a level below the limit of detection. Many other variations are probably much more dominating.

The first quick-and-dirty data analysis did not point to new positive discoveries. However, the intense hope of identifying new interesting biomarkers drove the data analysis through a long course of data gymnastics where it was very difficult to establish a finish line.

Case IV: Metabolomics - intervention study

A human food intervention study was initiated by a university group with great expertise in human nutrition. The study was a full-factorial design with two factors (amount of protein in diet and glycemic index of diet, each with two levels (high/low). The subjects were told to follow the diet and were supplied with all ingredients. The participants were instructed to collect all their urine during 24 hours at 4 specified time points during the 8 week intervention. These urine samples were then subjected to ^{1}H NMR metabolomic analysis. The purpose was to identify bio-markers for the given dietary patterns as outlined in the two treatments.

Initially, no effect of diet could be documented. Speculations on the cause

of this initiated further investigations on intra-individual variations in urine composition and it became clear that these were rather large relative to the effect of treatment, so only after pooling of urine collection over time, an effect could be observed. Furthermore it was concluded that the treatment variation was too little to significantly accommodate for all the uncertainties present in the data. One major uncertainty was the actual intake by the participants.

1.3 The omics problems

The four cases described above touch upon a number of problems associated with studies of omics data, which will be further enlightened in this thesis. However, the common purpose for all four cases is to identify valid bio-markers, and all cases fail this for different reasons.

Apparently, there are two ways in which they fail:

1. The identified bio-markers turn out to be false. or
2. No biomarkers can be identified.

The scientific literature documents that (1) is a problem which has received increasing attention within all omics fields through the years and which has invoked some skepticism towards the value of omics studies [54, 53, 9, 50, 30, 29, 40]. Various examples of (2) can be found in literature also, but presumably the balance between these two problems in literature is largely influenced by a publication bias towards more of (1). Although negative results can be very valuable, true negative results are rather difficult to establish and hence (2) may end out more or less inconclusive with respect to the question of whether biomarkers are actually present or if they have just not been uncovered due to limitations in the given experimental and data analytical setup.

Omics studies are often large studies involving several parties over longer periods of time and state-of-the-art technologies which make them very costly

scientific projects.

The high costs and massive efforts associated with omics studies make it reasonable to investigate if the chances of successful outcome in omics studies can be optimized.

Hypothesis

In many omics projects the possible true bio-markers are not strong enough to manifest in the given experimental setup when standard data analytical methods are applied, whereas false markers may appear to be real.

The root causes of this must be found either in data, in the data analytical methods or in the combination of these. By gaining a deeper understanding of the properties of omics data and the standard methods used for the analysis of these data, we might be able to optimize our procedures and achieve better and more reliable results from omics experiments.

Thesis structure

A data analytical method widely applied in omics data analysis is Partial Least Squares (PLS) regression. This standard method, which is the focus of this thesis, has proven very successful in combination with well-known data types such as Near Infrared (NIR) spectral data and various other types with similar characteristics. It is therefore natural to evaluate the properties of omics data in light of the reasons why NIR data and the associated well-established methods are so successful.

Through the following thesis structure the characteristics of NIR, PLS and omics data are visited, followed by suggestions for obtaining better omics experiments.

Thesis outline

- The historical background for the development of PLS and the type of data it was developed for.
- Characteristics of NIR type data and why PLS is well suited for NIR.
- The role of software in the success of PLS and some issues associated with this.
- Characteristics of omics data.
- Possible solutions for more successful omics experiments.
  - Data collection and experimental design.
  - Data cleaning and pre-treatment.
  - Data analysis and validation.
- Discussion and conclusion.

The thesis focuses on PLS and on omics studies with the objective of identifying biomarkers. The cases outlined previously in this chapter and the papers included define the type of considered omics applications. The following research papers are included in this thesis:

Included papers

PAPER I
Cross-validation of component models: A critical look at current methods. [8]

PAPER II
Some common misunderstandings in chemometrics. [35]

PAPER III
Direct functional assessment of the composite phenotype through multivariate projection strategies. [14].

PAPER IV
No genetic footprints of the fat mass and obesity associated (FTO) gene in human plasma ^{1}H CPMG NMR metabolic profiles. [36]

PAPER V
A simplified approach for identifying and separating unique and bulk variations in microarray data. [34]

PAPER I-IV have been accepted. The submission of PAPER V was delayed for political reasons and meanwhile another paper with similar content was published [55]. Consequently the paper will not be submitted.

CHAPTER

2 The near infrared revolution

This chapter takes a brief look into the history of chemometrics, with particular emphasis on the contribution of one of the key developments, Partial Least Squares Regression (PLS), to what might be called the Near-Infrared (NIR) revolution.

2.1 Early chemometrics

Multivariate pattern recognition is an old field, but the branch of chemometrics evolved from the early 1970s. Papers like those of Bruce Kowalski [37, 38], Svante Wold [67] and Luc Massart [43] were some of the first where multivariate soft modeling methods were applied within analytical chemistry [17, 20]. Wold coined the Swedish word *kemometri* in 1972 [67], and together with Kowalski he founded the Chemometrics Society in 1974. After some years (around 1980) The Chemometrics Society was split into two groups; one working primarily with structure-activity correlation (QSAR) studies and one working in analytical chemistry. The news bulletins from the early years of the Chemometrics Society provide a good insight into the pioneering spirit and broad range of visited research areas that made up the research field from the start. Advanced analytical

chemistry is an integral part of large-scale omics studies and in the remainder of this thesis focus will therefore be on this branch of chemometrics.

At the time when the Chemometrics Society was founded, spectra and chromatograms became increasingly common in analytical chemistry, to some extent replacing univariate observations from test-tube experiments. Instrumentation simply produced much more data than could be handled reasonably by existing methods and the complexity exceeded the capabilities of the human mind. Therefore, new methods had to be developed to avoid that substantial information remain unrevealed. The fact that (micro) computers capable of dealing with the numerical and quantitative issues became ubiquitous was an obvious fundamental co-trigger for the development of the field of chemometrics [20, 68, 28].

The chemometric methodology is often pragmatic and focused on problem-solving rather than defining natural "laws" and chemical "truth". This is fundamentally different to the traditional methodology of many scientific fields, for example physics, which relies strongly on the fundamental laws and hard-modeling. Analytical chemistry is a field with complex, noisy data which are difficult to model 100% by hard-modeling. That leaves an obvious room for the soft modeling tools of chemometrics [68, 69, 20, 17]. The soft modeling basically follows a well-established concept from statistics which seeks to separate the observed phenomena into a systematic (chemical) part, M, and a residual part, E. [69]:

$$X = M + E$$
or
$$data = chemistry + noise$$

The model is intrinsically approximate and has limited validity in terms of range and domain; however "All models are wrong, but some are still useful" as stated by G.E.P Box [70]. The appropriate model may therefore serve as a good basis for new theoretical insights.

Chemometrics facilitated a change in the chemical doctrine: the desired information is not necessarily obtained by measuring the most selective signal with the utmost precision; rather, combining several signals may be more informative [68].

One of most important influential modeling developments within chemometrics has been that of Partial Least Squares (PLS) regression. It provided a solution to a serious limitation of classical regression analysis, as will be illustrated below.

The limitations of existing methods - MLR

A fundamental mathematical problem throughout science is regression, $Y = bX + a$, but the classical method for multivariate regression, multiple linear regression (MLR), possesses some limitations; (a) it does not solve problems with more variables than samples and (b) $\mathbf{X}$ cannot be close to singular, i.e. rows or columns cannot be collinear. The reason for this shall be illustrated briefly:

MLR solves the regression problem $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ for a set of independent variables $\mathbf{X}$ and one or more response variables $\mathbf{Y}$. With n samples and m variables in $\mathbf{X}$, MLR has the following properties:

1. There is a unique solution if $m = n$, given that $\mathbf{X}$ is full rank.
2. There is an infinite number of solutions when $m > n$.
3. For $m < n$ a solution can be found by minimizing $\mathbf{E} = \mathbf{Y} - \mathbf{XB}$. The least squares solution is then $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Thus, for a solution to be obtained in (3) the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ must exist. When $\mathbf{X}$ is highly collinear the inverse is either highly unstable or non-existing.

As a consequence, MLR is not well suited for spectral datasets such as Near Infrared Spectroscopy (NIRS) data where the variables are highly collinear and the number of samples may be lower than the number of variables.

PLS Regression

Partial Least Squares (PLS) regression was originated by Herman Wold within the field of econometrics around 1975 and via Svante Wold, the method became a cornerstone of chemometrics. PLS was useful for its ability to model many, noisy, collinear data and thus circumventing the limitations of traditional regression methods like MLR. Quoting Höskuldsson [27]:

> "An important question is 'What situations are typical of those where PLS methods can be expected to be good for modelling purposes?'. They are the ones where there are many variables but not necessarily many samples or observations."

Description of PLS and the relevant algorithms (e.g. the original NIPALS algorithm) can be found many places, e.g. [71, 21, 27, 72].

PLS assumes that a low-dimensional underlying latent structure, *latent variables*, is present and that both X and Y are realizations of this [72]. At the same time there are parts of X which are not primarily related to Y which are also explained by the model and this might hamper interpretation. With a further development of PLS, O-PLS, improved interpretability is sought by stripping off the parts of W which are not primarily related to Y. [59].

PLS for classification

Discriminant PLS (PLS-DA) is a special use of PLS used for classification purposes. It is the "PLS version" of classical linear discriminant analysis (LDA), i.e. it also works for rank-deficient problems [72].

In PLS-DA regression, a model is built between the multi-dimensional dataset **X** and a "dummy" class variable **Y**. **Y** is constructed so that each column $\mathbf{y}_g$ states the membership of objects to the g'th class, (normally) indicated by logical 0 and 1 values. Thus, with N samples belonging to one of G classes, **Y** is an [N x G] matrix. In practice, closure is usually assumed so **Y**: [N x $G-1$]

PLS-DA works well when classes are “tight’, i.e. they each occupy a small and separated area in space. In these cases a discriminant plane can be found. If there are large inhomogeneities within the classes other methods may be more suitable [72]. PLS-DA is also not suitable for multi-class problems where the classes are clusters along an underlying axis (Figure 2.1), since a discriminant plane cannot be found; hierarchical methods such as CART may be better at handling such conditions.

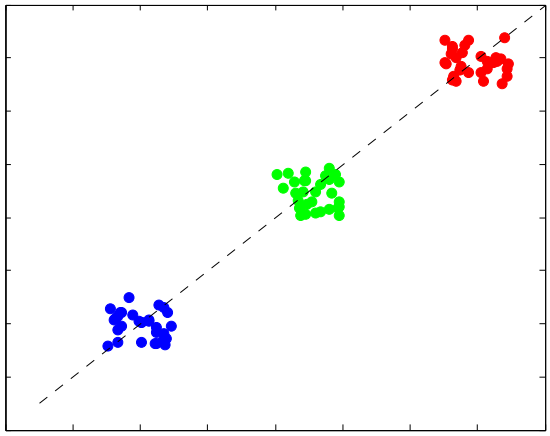


Figure 2.1: PLS-DA is not suitable for multi-class problems where the classes are separated as clusters along an underlying axis

2.2 Near infrared spectroscopy and chemometrics

From the time of its emergence, PLS has been used for a wide array of applications within several areas of chemistry. Near Infrared (NIR) spectroscopy is an analytical platform which has become very successful due to the introduction of chemometrics. A search on Web of Science (WoS) and Google Scholar on the combination of the terms *"Near infrared"* and *PLS* underlines this - WoS

provides more than 3000 hits and Google Scholar gives nearly 15000 (including citations).

The NIR spectrum ranges from approx. 780-2500 nm and consists of overtones and combination bands of the fundamental stretching and bending vibrations found in the IR region. Any molecule containing hydrogen absorbs in the near-infrared region. The complexity of the spectra is very high and until the advent of chemometrics and the concurrent availability of necessary computational power, NIR spectra were generally considered impossible to utilize. MLR (and Principal Component Regression, PCR) formed the starting point but the introduction of PLS provided a major breakthrough. PLS facilitates full-range spectra to be processed directly using PLS without prior compression or selection of variables, and the response variable can be modeled by a less complex model.

The fact that C-H, O-H and N-H molecular vibrations are so well represented by IR and NIR spectroscopy make the two spectral regions extremely well suited for analysis of the major constituents of biological samples: fat, protein, carbohydrate and water. The NIR region has the advantage over IR that the spectral energies are higher (shorter wavelengths), therefore transmission mode is feasible for amorphous and even solid samples of a reasonable thickness. This is highly advantageous in terms of sample preparation.

Within the life sciences, all sorts of agricultural commodities and food samples including process intermediates have been subjected to NIR spectroscopy. Grain is an all-time classic where Karl Norris did pioneering work [1, 19], and every day grain is analyzed in versatile applications using NIR instruments. Further work extended NIR spectroscopy to the analysis of analytes in food, animal feeds, polymers, wool, natural and synthetic textiles, pharmaceuticals, chemicals, and petroleum. It is clear that there are few areas of analytical chemistry in which NIR spectroscopy has not been or cannot be utilized [15].

Multivariate modeling with NIR spectra has been used for all kinds of purposes belonging to the three main categories of chemometrics: exploratory

analysis, calibration and classification.

The exploratory analysis is useful for investigations of variations in data. Expected clustering and trends can be visualized and analyzed, and most importantly, unexpected phenomena may emerge. The workhorse of the unsupervised exploratory analysis is Principal Component Analysis (PCA).

Calibration is regression of one or more dependent variables $\mathbf{Y}$ (e.g. protein, water) on a set of independent variables (e.g. NIR spectra of grain). The resulting model may then be very suited for rapid routine analysis of e.g. raw-, intermediate- or end products in a production setting.

Classification is the art of assigning class membership to samples based on a set of attributes. An example could be classification of wheat grain as baking wheat or wheat for animal feed.

All three categories of applications are relevant for NIR data, but calibration is probably the more widely used of them - at least with respect to commercial applications.

The massive amount of applications within NIR and related areas (IR, HPLC, GC etc) developed through the last thirty years forms a considerable part of the knowledge and experience which has been built up in the various chemometrics groups.

2.3 Summary

In summary, PLS - one of the core methods of chemometrics - was developed in a context where it was a true revolution to be able to handle 100 or even 1000 variables obtained on 20 or 50 samples. PLS facilitated the analysis of rank-deficient problems, typical of e.g. spectral data, and played a major role in the success of NIR spectroscopy.

NIR spectroscopy is an analytical technique whose potential could not be utilized fully until the advent of chemometrics due to the high complexity of the spectral information. The NIR revolution was largely facilitated by chemo-

metrics, and for many years NIR technology has succeeded in seemingly endless numbers of applications within analytical chemistry, not least within the agribusiness. Analysis of NIR data and related data types has formed a substantial part of the work undertaken by chemometricians and as such forms the basis for the knowledge and frame of data analytical understanding present in the various chemometrics groups today.

CHAPTER

3 Chemometrics on NIR data is simple

In this chapter the properties of typical NIR data are examined and it is clarified why the PLS method and the NIR data make a good match. NIR data are chosen as a typical example of a "classical" type of data, which has been the basis of a large part of the data analytical knowledge built up through the years.

3.1 NIR samples are uniform

Samples

The available number of samples in a well-designed typical NIR dataset is of course application dependent; but 50 to 500 is not unusual. An important characteristic of samples typically subjected to NIR spectroscopy is that often they are available in ample numbers at a reasonable price. Sample preparation is minimal and the scanning procedure is generally fast and un-destructive, allowing many samples to be easily included. The limiting factor is often the reference analyses, which are far more costly and time-consuming.

An important characteristic of samples from NIR PLS applications is that

they are usually sampled from a very specific group of samples. Uniformity is a quality parameter of agricultural and industrial products and consequently complex interference is limited.

Often samples are obtained from random sampling which results in near-normal distributions, but it can be beneficial for the regression to sample flat [31]. It is not always straight-forward to sample flat because it may be very difficult to guess what the reference value is, but Isaksson and Næs [31] have suggested a method where the samples for reference analysis are selected based on the spectra. Alternatively, design such as D-optimal design provides a strong means for obtaining good parameter estimates during modeling.

Signals

Depending on the spectral width and resolution, an NIR spectrum may typically consist of 50-500 variables. As an example, the Infratec 1241 Grain Analyzer (FOSS Analytical) which is used worldwide for analysis of whole grain records 265 data points per scan.

NIR spectra are sensitive to temperature variations, mainly due to the state of hydrogen bonds in the water fraction [10]. Minor fluctuations in temperature are of less importance, in particular if a reasonable temperature variation is included in the modeling.

The fact that NIR spectra consist of overtones and combination bands result in spectra with broad and many overlapping peaks that are difficult to interpret in detail visually.

Noise

NIR instruments have a very high signal to noise ratio which is typically 10000:1. Sensitivity is around 0.1% for most constituents [10], thus, the technology is not well suited for detection and quantification of low-concentration analytes.

Scatter may be an important issue in NIR spectroscopy depending on the nature of the samples. Frequently, NIR spectroscopy is carried out on highly scattering samples where full adherence to Beer's Law is not expected.

Instrument noise is low but biological variation (complexity of the sample matrix) may challenge the data analysis. Again, using grain analysis as an example, genetic background, storage conditions, harvest year, location, disease and adverse weather conditions are factors which may introduce variations to an extent which should be considered with respect to the purpose (e.g. accuracy) and intended dynamic range for the model.

Nevertheless, the biological variation (within a given product type) is generally relatively low for agricultural and food samples in the sense that the genetic and production variation are usually factors which are minimized in a production setting.

Rank

As just described in chapter 2, a great expectation of chemometrics regarded its ability to handle the "fat" matrices, i.e. matrices with much more variables than samples.

With 100 samples and 265 variables, apparently this dataset is a fat matrix. However, for several reasons, the "biological rank", understood as the number of biological phenomena at play in the dataset is much lower than this.

First, an inherent property of spectral data is collinearity between neighboring wavelengths, which reduces the rank considerably. Secondly, primarily the major components (lipids, carbohydrates, water, protein) are present in concentrations suitable for NIR, thus the number of analytes giving signal is relatively limited. Furthermore, the same chemical compound may be represented several times in the spectrum as a result of several NIR-active functional groups in the same compound and through the resonances from overtones and combinations of these. Eventually signals from several compounds may often be very correlated because the samples are very alike in their chemical composition -

e.g in a production setting or as a result of shared biology such as genetics and environment.

As a result, chemical/biological rank of NIR samples is typically below 15 and such a NIR dataset is really a good old "tall" dataset with 100 samples and few phenomena (latent variables) that vary.

Curse of dimensionality is canceled for NIR data

It is a fundamental property of bound spaces that the volume increases exponentially with the increase in dimensionality. This "curse of dimensionality" [5] dictates that the number of samples needed to describe a multivariate system accurately (i.e. maintain equal Euclidian distance between samples) grows exponentially with the number of variables [66], so we would need many samples to describe a 265 dimensional space. However, as just described, the dimensionality of the underlying structure (rank) is much lower and as a consequence, the "curse of dimensionality" does not have practical implications for low-rank data like NIR data.

3.2 Preprocessing of NIR data is straightforward

The critical difference between inadequate and successful chemometric models is often data pre-treatment, i.e. what is done to the data before using PCA, PLS etc. The goal of preprocessing is to remove variation not related to the problem of interest so that the variation of interest is more evident and can be modeled more easily.

An important interfering phenomenon in NIR data which must generally be taken care of, is light scattering from particles. This is a physical phenomenon related to particle size and it is very often successfully corrected for using standard pre-processing methods like Multiplicative Scatter Correction (MSC) or Standard Normal Variate (SNV). Some spectral areas may also be obviously noisy and hence qualify for elimination. The number of variables in an NIR spectrum is computationally manageable and the uniformity of the samples usually make sample preprocessing reasonable straight-forward.

3.3 PLS is perfect for NIR

If the purpose is to quantify the content of moisture, protein, carbohydrates or lipids, NIR is generally a method of excellence, and very accurate determinations can be made. The model prediction error is often much lower than the error on the reference methods, thus the NIR technology and the chemometrics perform at their best with this kind of application.

An essential part of the background for this success is that very often, the predicted attribute is directly linked to the NIR spectrum via Beer's law:

$$A = \epsilon \, l \, c \tag{3.1}$$

where A is the absorption by the analyte, ϵ is the molar absorption of the analyte, l is the path length and c is the concentration of the analyte.

Thus, determining protein from an NIR spectrum is in theory straightforward because NIRS is sensitive to the amine N-H bonds, which are abundant in proteins. There is a very direct connection between the signals and the parameter of interest, although this is to some extent disturbed by scatter and other interferences mentioned above. Figure 3.1 shows an example of protein determination in cattle feed by use of NIR in reflectance mode and PLS modeling.

3.4 IR, GC, HPLC is also NIR

As already mentioned, this thesis is not about NIR, but rather about how omics data are different from more classical data types, such as *for example* NIR data. I would like to stress, that what is true for NIR regarding the general properties of NIR data generally also applies to similar data types like IR, GC, HPLC etc. Each method (and application) has its specific characteristics, but the central elements such as rank, number of variables and samples, sample types and the fact that a direct link between X and Y exists, is generally valid for all the mentioned data types.

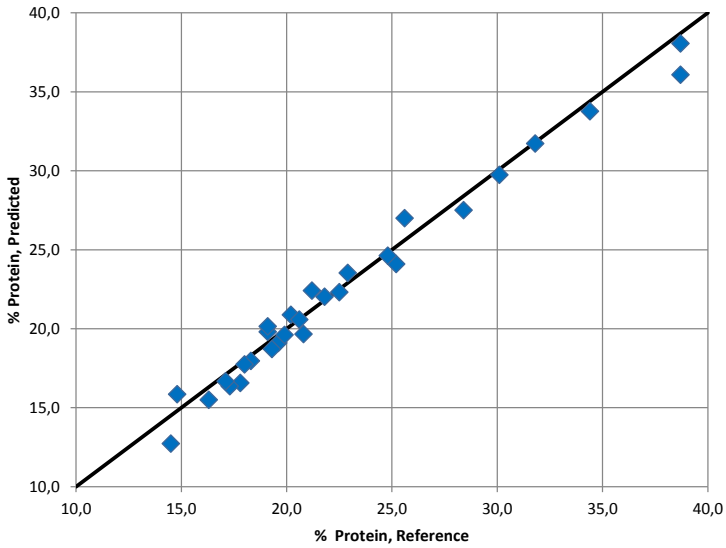


Figure 3.1: Determination of protein in cattle feed by use of NIR and PLS. Data from FOSS.

3.5 Software - a seductive necessity

A wide range of commercial software packages (e.g. The Unscrambler (CAMO), SIMCA (Umetrics), PLS_toolbox (Eigenvector)) are available which provide good tools for exploratory analysis and calibration of multivariate data in general, and which are very well suited for NIR data. The packages differ slightly in terms of structure and graphics and consequently possess different strengths and weaknesses but as a minimum they all contain a PCA module, a PLS module and some possibilities for preprocessing. The available software packages are generally easy to use and can be applied after a relatively short introduction to the concepts of chemometrics.

The availability of good software packages has been central for the massive

success of NIR applications. They have made multivariate data analysis and chemometrics possible for a large number of people with strong expertise in the domain of their data but with limited insight into the complexity of multivariate mathematical and statistical modeling. By means of the software, chemometrics and multivariate data analysis have found their way to new academic areas and important corners of the industry - mainly the agricultural-, food- and pharmaceutical industry.

With the relevant data at hand and basic knowledge about chemometrics it is reasonably straightforward to make a PLS model which predicts the (quality) parameter y from a NIR spectrum. All types of data require their specific preprocessing, and the optimal model may require more advanced efforts and of course some applications are more tricky than others, but some sort of standard recipe generally works well.

The seamy side of the "ready-to-eat" software is that it promotes uncritical "push-the-button" automated analysis. When the software automatically serves given plots and figures of merits to the user, it is convenient for the inexperienced user to uncritically assume these are the most relevant. This may mislead the user in several ways:

- Relevant plots which are not directly served are not inspected although they may contain important information.
- Some plots are interpreted wrongly because the graphical representation (e.g. axis scaling) is inappropriate for the given purpose.
- The presented figures of merit are given way too much emphasis although they may be irrelevant for the purpose at hand.

Many common misunderstandings and bad habits in chemometrics are related to these points. A selection of such problems is dealt with in detail in PAPER II.

Another software-related issue of more specific character is how to cross-validate PCA models. Several cross-validation regimes have been proposed, and the various software packages have different cross-validation routines implemented. These routines were tested using both simulated and real data of a spectral type and the results are shown in PAPER I.

3.6 Summary

In this chapter the general properties of NIR data and related spectral type data has been examined. In summary, this type of data is high-dimensional, has more variables than samples (ratio possibly 2-10:1) but the chemical rank is relatively low. The type of products under investigation are often samples related to some kind of production where uniformity of samples is a fundamental quality parameter. The applications under study often rely on a direct link between X and Y, such as Beer's law, thus a clear signal from specific spectral regions are generally observed. PLS handles this type of data very well.

It is beyond any doubt that chemometrics software has contributed considerably to the spreading of chemometrics in general - not least with respect to the use of PLS on NIR type data. Software assists the data analytical workflow and readily provides figures and diagnostics to the user. It is however crucial that these are interpreted with responsibility and there is a risk that the inexperienced user may be mislead.

CHAPTER

4 Why are omics data difficult to handle?

In this chapter the properties of omics data are reviewed in detail. I will address how omics data differ from more traditional NIR type data, and why this may present more challenges to the data analysis. Some central topics are the variations present in omics data, including both biological variations and analytical errors, and the types of scientific questions which are often addressed with an omics approach. The range of published applications and designs is obviously very diverse; nevertheless, I believe some general points can be made and the cases outlined in the beginning of this thesis together with the included omics papers (PAPER III, PAPER IV) will serve as the basis for this quest.

4.1 Omics data are true fat matrices

A clear characteristic of omics data is the high number of variables. Exactly how many obviously varies with the application, but numbers in the range of 20000-50000 are often seen. The number of samples also vary across applications, but is generally substantially lower than the number of variables. The twin case in PAPER IV represents an omics application with very many samples (>1000), whereas the microarray nickel case has 34 samples. The second case is quite

typical for micro array experiments. Depending on the experimental design, studies of human subjects are often labor-intensive, expensive and recruitment may be difficult.

NIR data are also characterized by having more variables than samples, and one might regard omics data as "extended NIR". However, whereas the relatively low chemical rank of NIR data means that they are not true "fat matrices", omics data are generally high-rank data and thus for most practical purposes underdetermined systems. The high rank of omics is a result of the complex biological and environmental interactions at play (see more below).

It is difficult to estimate the biological rank of an omics dataset - not least due to the fact the systems are generally underdetermined - at least when we consider the noise level. The rank obviously depends on the biofluid, the analytical platform and the design of the experiment. Due to the relatively low sensitivity of ^{1}H NMR, such profiles will probably only contain signals from about 10% of the metabolome, i.e. a few hundred signals [62].

The fact that omics data are true fat matrices means that the number of samples becomes a limiting factor. With a high noise level, many samples are needed to accurately describe a multivariate system [66].

4.2 Biological variation is large in human omics data

In samples from a free-living population of human beings, it is expected that a very complex biological variation is present which is not related to the attributes of particular interest in the study. Some of the most important ones are highlighted here.

Studies of human beings will generally contain a diverse genetic background variation relative to studies involving for example laboratory rats or cultivated plants. An obvious genetic variation is the inclusion of both genders.

One might argue, that due to the pathway-regulated nature of our metabolism, biological rank cannot be very high. Levels of elements from the same pathway could be expected to be correlated, but this is very often not the case.

Metabolism runs in all cells in a network of pathways within and across cells, and bio-fluids such as blood or urine hence reflect what is going on in the system as a whole at a given time point. Metabolites are synthesized from other metabolites in a complex network of biochemical reactions [53] and phenomena such as chemical equilibrium interfere [11].

The fact that each individual has a unique genetic profile and in particular that the physiological state is very different between individuals as a result of current and accumulated physiological situation means that the regulation varies between individuals and that "normal" states may vary considerably in level among individuals. This reduces effect size and hence challenges both the identification and threshold determination of univariate quantitative markers.

Our primary metabolism is largely influenced by the homeostatic regulation (see section 4.5) which seeks to keep the system at an operational state. However, environmental factors such as food intake, physical exercise or drug intake result in relatively large fluctuations in the individual. Other fluctuations are governed by well-regulated cycles such as the female hormonal system or diurnal regulation i.e. levels are controlled by time of day. Some metabolites show large variations under identical experimental conditions. CASE IV (metabolomics/dietary intervention study) underlines that intra-subject variation may be considerable. Urine samples were collected during 24 hours, but still the intra-individual differences were significant and further data-pooling was necessary in order to see any effect of diet.

Age is a factor which exhibits both some biological impact, e.g. in the body's responses to metabolic or environmental changes, but also some impact on the lifestyle due to differences in "life situations" and influence by the historical context the subjects have been part of. These may greatly influence for example dietary patterns.

Most studies include some exclusion criteria to prevent too strong interfer-

ences with the phenotype of interest. In the GEMINAKAR study for example, which is the basis for PAPER IV, exclusion criteria were pregnancy, breastfeeding, known diabetes or cardiovascular disease, and any condition precluding a bicycle test [52]. Thus the subjects might suffer from other conditions which may influence the NMR spectra.

Because of the biological complexity it can be difficult to align individuals biologically and hence obtain uniform replicates. Ethical aspects will generally restrict this for studies of humans, and even the study of bacteria during fermentation can be difficult to standardize. An example of a transcriptomics study of Lactococcus Lactis during milk fermentation is given in section 6.2. The genetic response of the individual cells at a given time point will vary across the population due to differences in micro environment.

4.3 Measurements are not merely biological signals

Omics signals do not always reflect the biological state they were intended to reflect. In addition to the biological variations mentioned above, sample work-up and analytical errors may interfere with the signals.

Conceptually, omics experiments in general aim at getting signals from all the specimens in the ome under study, but it is almost impossible to have a method which can cover this across the wide dynamic ranges of concentration, molecular size, and chemical characteristics. Thus, some extent of filtering is associated with the choice of methodology. Furthermore, reproducibility of the method - "robustness" - may considerably influence the variation of signals present in the dataset, and this way add phenomena which are not related to the state of the samples the moment they were taken.

These issues are general for analytical biology/chemistry but whereas classical techniques and applications can be optimized for the few specific signals they are used for, this is much less feasible for omics data. The problems are therefore much more pronounced in omics applications.

The underlying assumption that the information extracted from a measure-

ment of a sample reflects the composition/state of the subject/object it is taken from, may be violated in a number of ways:

The measurement may not reflect the sample subject

- Sampling may not be representative
- Sample composition may change as a result of sample handling (storage, addition of stabilizers etc.)
- Data from different samples may not be directly comparable due to the introduction of analytical artifacts (different sample amount, shimming, temperature effects)
- Sensitivity may vary for specimen.

The sampling issue does not differ remarkably from other applications e.g. NIR applications and is not discussed further here. Sample handling may be more critical for omics data because the various constituents may interact differently with the handling procedures, which is illustrated below. Some analytical artifacts are also highlighted in the following.

Changes in sample composition

Sample handling is a factor which may disturb the sample composition between sampling and measurement time. When a blood or urine sample has been taken from the subject or patient it is usually not instantly chemically analyzed. In omics studies there is often a need to transfer samples between facilities for data acquisition. Responsibility of sample care may then move with the samples whereby knowledge and awareness of critical issues regarding sample handling may be changed.

For both urine and blood samples some constituents are very sensitive to

e.g. temperature, whereas others are not [39]. The proteomics case (CASE II) is a very good example of how influential sample handling can be. In that case, blood samples were stored at minus 80 °C after collection, and then transferred batch-wise to another facility where minus 20 °C was the most feasible storage possibility. After measurements were performed, the predominant variation reflected for how long time the samples had been stored at -20 °C.

Some constituents are much more prone to transformation during storage than others. Clark et al. [12] found that a wide range of constituents in whole blood; albumin, apolipoproteins, cholesterol and triglyceride concentrations were very stable and changed very little even by storage at room temperature, but as the proteomics case shows, this is not the case for all constituents.

Stabilizing agents can be added to the sample to minimize deterioration. These may affect chemical equilibria such as complexation or pH dependent equilibria.

Analytical filtration, errors and biases

The analytical technique presents some limitations to the output data. In metabolomics for example, the choice is often between a hyphenated mass spectrometry based or an NMR based approach. Overall, the MS based methodology has the advantage of high sensitivity but at the prize of a relatively poor repeatability. The opposite is valid for NMR.

Below, some specific issues regarding the following analytical techniques will be mentioned

- MALDI-TOF Mass Spectrometry for proteomics.
- ^{1}H NMR for metabolomics analysis

MALDI-TOF MS for proteomics

Mass spectrometry is very sensitive and is superior with respect to detection of low concentration analytes, but when applied in proteomics (and metabo-

lomics) the wide dynamic ranges is a serious challenge. Varios concepts for sample preparation have been suggested, but most of these lack thorough investigations and documentation on reproducibility and sensitivity to experimental artifacts [65, 63].

Specifically in CASE II, IMAC beads were used for sample separation [64], and the selectivity of these will affect the resulting MS data. Temperature might also affect crystallization prior to to the MALDI ionization.

^{1}H NMR for metabolomics

^{1}H NMR has the advantage that it is adequately robust to present a good repeatability. On the other hand, sensitivity is not very high and probably only around 10% of the metabolites present in a plasma sample will be present in the NMR spectrum [62].

The choice of pulse program affects the signals. The Carr Purcell Meiboom Gill (CPMG) sequence suppresses the signals of macromolecules such as lipoproteins, hence enhancing the signals of small metabolites. Overhauser Enhancement SpectroscopY (NOESY) recordings provide a good overview of all the types of molecules present in the sample matrix [4]. Still, pre-saturation of the water resonance is indeed desirable. Sometimes water suppression is not successful under automated conditions; in such cases the quantitative output of neighboring peaks is of very little use (Figure 4.1).

pH variations in the sample induce peak shifts of pH-sensitive functional groups such as carboxylic acid groups. This is shown for citric acid in plasma in Figure 4.2, bearing in mind that plasma is quite pH stable.

4.4 Most measurements are irrelevant

The traditional hypothesis driven study is the strongly focused work which only investigates phenomena hypothesized *a priori.* Ideally, it is designed to provide

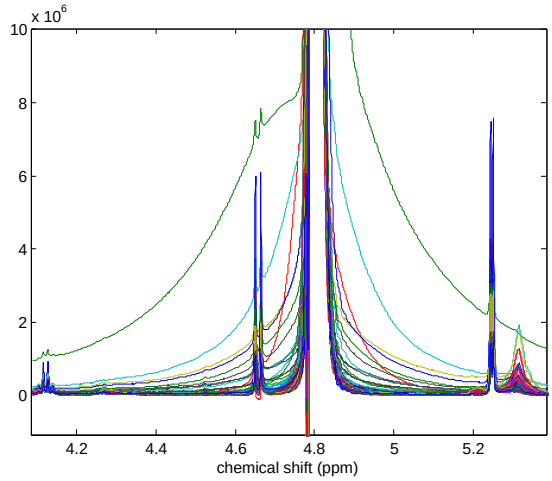


Figure 4.1: Poor water suppression affects neighboring regions

the strongest statistical power to detect the effect size in exactly the factors under study.

The exploratory spirit which is an integral part of many chemometrics environments represents another corner stone with mantra like “let data talk through unsupervised modeling”. It is a beautiful concept that simple visualizations (like those of PCA) can generate new hypotheses about biological phenomena at play and as such cross the barrier formed by our intellectual capacity to hypothesize. However, the exploratory spirit comes at a cost.

The fascination of the omics technology is often associated with the large exploratory component. It is conceptually very attractive to measure “everything” and extract the relevant information. Consequently, most omics studies produce data where by far most of the measured variables are irrelevant.

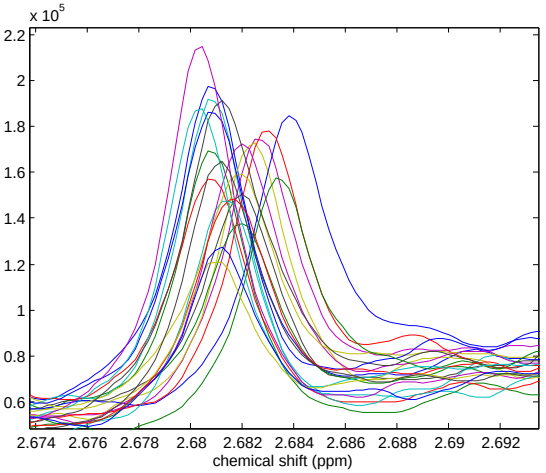


Figure 4.2: pH variation induces peak shifts. One finger of the citrate doublet of doublets in plasma ^{1}H NMR profile

In an omics experiment where we measure say 50.000 variables, we aim to find 1,2, 10 or maybe 50 biomarkers. That is < 0.1% of the variables and the remaining are more or less irrelevant. With M variables in a dataset, it is possible to combine these in $2^M - 1$ ways. For $M = 1000$, which is a quite low number in an omics context, this means $1.0715 * 10^{301}$ combinations and by far the most of these would be completely irrelevant. Trying out all these combinations would be computationally impossible but furthermore, considering the statistical power of such an approach would stop the initiative immediately.

For a short moment, consider all the irrelevant variables as completely random numbers. With so many variables and relatively few samples, it is very likely that random correlations will be found. In Figure 4.3 the percentage of variables that obtain Pearson's correlation $r^2 > 0.5$, 0.6, 0.7 and 0.8 respectively is shown. With 20 samples and 50.000 random variables, then 1458 variables will have $r^2 > 0.5$, 442 variables will have $r^2 > 0.6$, 112 variables will have $r^2 > 0.7$ and 68

variables will have $r^2 > 0.8$! It is no wonder it is possible to find biomarkers if proper caution is not paid.

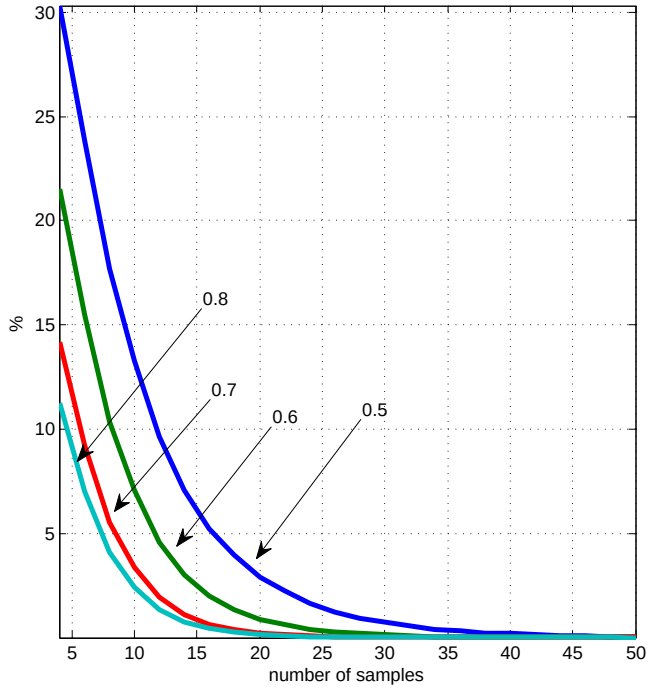


Figure 4.3: Random correlations. The percentage of random-number variables with r^2 larger than 0.5 - 0.8 as a function of the number of samples in the dataset

PLS can not handle many irrelevant variables

The PLS algorithm seeks to project **X** onto a hyper plane so that the projection of **X** correlates well with y. In that way, PLS models both **X** and y. As a consequence, modeling of datasets with a high degree of noisy and irrelevant variables will be

largely influenced by the noise, in particular in the absence of strong signals.

This is illustrated in Figure 4.4. The figures shows how the cross-validation prediction error (RMSECV) increases with the amount of irrelevant noisy variables for simulated datasets.

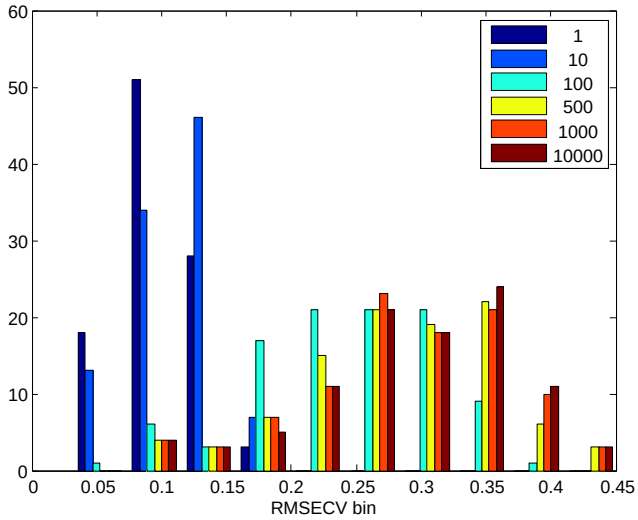


Figure 4.4: Distribution of RMSECV of PLS models as a function of the number of irrelevant variables.

This example was created with the following simple setup:

- $\mathbf{X}_{rel}$: [50 samples x 3 relevant random variables]
- $\mathbf{y}$ is a [50 x 1] vector formed as a random linear combination of $\mathbf{X}_{rel}$.
- $\mathbf{X}_{irrel}$ [50 samples x N irrelevant random variables]
- $\mathbf{X}$ is a concatenated matrix of $\mathbf{X}_{rel}$ and $\mathbf{X}_{irrel}$: [50 x 3 +N] + 10% homoscedastic noise.

- $N \in \{50, 100, 500, 1000, 10000\}$

For each value of N:

- Cross-validate an up to 25 LV PLS model using leave-one-out.
- Determine RMSECV and number of PLS components by minimum RMSECV.
- Repeat experiment 100 times.

The noise effect is adversed by application of scaling to unit variance (autoscaling) which is intended to compensate for the mismatch between absolute signal size (peak height) and relevance of the analyte. This is due to the noisy variables being amplified.

4.5 Weak links

Above, various issues regarding the enormous data matrix ($\mathbf{X}$) produced by the high-throughput analytical platform have been considered. In this section two other elements, namely the response variable ($\mathbf{y}$) and the link between $\mathbf{X}$ and $\mathbf{y}$ are addressed.

The uncertain response variable

The response variables used in omics studies are often quite complex phenotypes, such as *conditions*. Obese/lean and cancer/no-cancer are typical examples.

First of all, these conditions may be difficult to define and represent by one response variable (as is often the case). A widely used obesity measure could be simple BMI measurements although obesity as a condition can manifest in various ways affecting BMI differently. Hence we search for obesity markers but find BMI markers.

A basic condition in most disease studies is that the group of patients is much more rare than the control group. As a consequence, the design may

be poorly balanced. A means to circumvent this problem is to widen up the included pathological pictures, i.e. include more subtypes and disease stages, as was the case in CASE III (cancer proteomics). These subclasses of the classes may have very distinct responses in the biological snapshot which was taken with the urine or blood sample.

As such, these conditions probably represent a mixture of phenomena which are continuous and phenomena which are characteristic for distinct stages. As a result there may be a considerable uncertainty in the phenotype.

Intervention studies have an uncertainty with respect to the actualized treatment regimes. Drug intervention studies are generally well off with respect to drug administration, but may also have the problem just described with a variation in the patients' disease pattern. Food intervention studies have a fundamental problem of not knowing the exact food intake. Self-reporting is uncertain (and difficult) and e.g. the problem of under-reporting for obese subjects is a well-known phenomenon [25, 45].

Small effect size

It is a characteristic of many studies that the link between the omics data and the phenotype may not be very direct - and is generally quite complex. A direct effect of this is that observed effect sizes are shrunk. Figure 1.1 illustrates one aspect of this problem, namely that the path from gene to clinical response (i.e. phenotype) goes through several biological domains each of which interacts both internally and with external (environmental) factors. Hence, possessing a risk allele of a given gene for a given disease does not mean you have the disease. Other genes may enhance or cancel the effect of the gene and various environmental conditions (triggers) may also be required for disease development. As the figure illustrates, this problem increases as the distance in system level increases between response and predictors, and genome-wide-association studies are generally prone to this problem.

An important factor to consider here is the phenomenon of *homeostasis*. We are still on this planet because we are very robust to many environmental

variations. A complex network of feedback regulation systems ensures that the necessary fundamental body functions keep running, i.e. elements such as osmotic pressure, pH, water balance and temperature are kept within functional limits. Traces of drug treatments are generally not too difficult to observe, but as a consequence of the homeostatic control, dietary intervention studies which in general involve healthy individuals might need rather extreme levels to leave an effect. This was one of the problems in CASE IV (Food intervention study) where the administered diets were not very extreme, so the fundamental effect size, i.e. ignoring shrinkage by other effects, was relatively small.

Thus, in omics applications we are often looking for small effect sizes (Figure 4.5, in particular when we are dealing with healthy individuals. Large effect sizes will in many cases not be new discoveries as they are well known beforehand. The marker we look for may well be of a multivariate character or several univariate markers may be present each with small effect sizes.

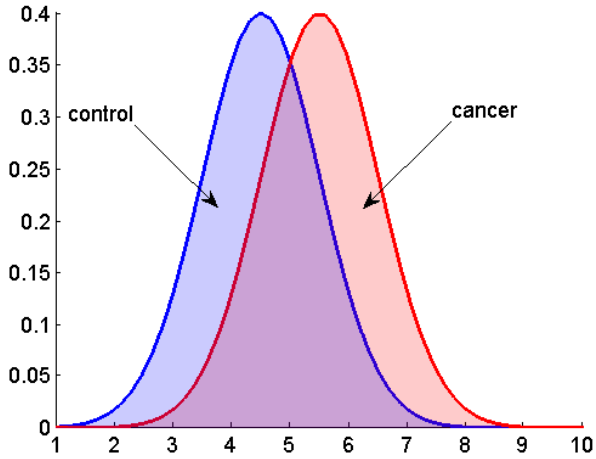


Figure 4.5: Small effect sizes is a basic condition in many omics dataset

4.6 Gathering of experts

It is another characteristic of omics studies that a wide range of competences are needed for a successful outcome. A good biologist is required to put up the good research question and make sure that relevant data is collected. The high-throughput omics platforms are generally very advanced and require highly skilled personnel in order to acquire high-quality data. Eventually, experienced data analytical competences are needed to analyze the data.

CASE II (proteomics) illustrates how this expertise is not always integrated throughout the process. Someone in the project might have known that samples must be stored at minus 80°C until analysis but the experts at the proteomics facility were not aware of this and intuitively felt that -20°C was adequate.

All these resources must be gathered and integrated appropriately in the project. As these experts are generally found in different research environments, it takes time and effort to establish good communication and collaboration overcoming differences in scientific culture and vocabulary. This subject is further commented in the end of this thesis (Chapter 9).

4.7 Summary of omics properties

This chapter shows that omics data are very complex data and not just simple extensions of NIR. The (human) omics samples represent a much broader biological variation than the typical production samples which are usually the focus of NIR experiments, and they are also difficult to obtain in high numbers. The number of variables is comparable or higher in omics samples but the fraction of relevant data (information) is little and difficult for PLS to extract from the ocean of noisy data. A direct result of the massive biology encompassed by omics technologies is data with a high biological rank. The limiting number of samples makes it uncertain to extract the real underlying structure of the data.

The link between predictor and response variables is often much more com-

plex, indirect and hence weaker in omics data than in NIR data, as it does rarely rely on a direct physical/chemical fundamental like Beer's law.

The intention of omics data is generally to obtain signals from as much of the ome as possible, which presents some analytical challenges due to the wide dynamic range. There is thus a compromise between getting a wide range of signals (qualitative requirement) and (1) reproducibility due to complexed analytical method or (2) non-linearities in signal gain across the range.

CHAPTER

5 Collect good data

It is now clear that the automated acquisition of large-scale data from the omics domain results in substantial exploratory and regression related problems. The properties of high-dimensional data can affect the ability of statistical models to extract meaningful information, and having more complex measurements means having more possibilities for errors and biases [30].

The path to useful results from omics experiments goes through three core areas, which must all be addressed:

- How to obtain good omics data.
- How to prepare acquired data for data analysis.
- How to obtain useful results from omics data.

This chapter deals with the first area. First, by looking into experimental design matters, and afterwards by contemplating analytical considerations. Data analysis is the overall focus of this thesis, therefore this chapter deals primarily with design issues which affect data quality in general, rather than on platform specific details.

In the next two chapters (Chapter 6 & 7), data analytical issues are addressed; chapter 6 focuses on proper cleaning of data prior to analysis, and chapter 7 deals with the actual data analysis, including validation.

5.1 Design

Biology is complex and human biology is by no means an exception. When on top of this we add analytical error and bias, it is therefore crucial that we seek to span the variation which is relevant to our study and minimize other influences.

From a data analytical point of view, a good design fundamentally seeks to ensure a strong link between $\mathbf{X}$ and $\mathbf{y}$. This is obtained by trying to fulfill the following points:

- a well-defined response variable $\mathbf{y}$ (phenotype).
- a targeted descriptor matrix $\mathbf{X}$, i.e. data which are relevant for the given purpose.
- a fundamental sound biological relation between $\mathbf{X}$ and $\mathbf{y}$.
- minimized analytical error and bias.

In other words, the design is a means to optimize the match of $\mathbf{X}$ and $\mathbf{y}$.

Reduce ambitions, retain power

The previous chapter underlined that a fundamental challenge in omics studies is the small effect size, hence low power.

When designing an omics experiment it is crucial to define the research questions. This may sound trivial but as the case stories illustrate, the research questions are often quite vague á la "we would like to know what is going on". Many studies will have a design which respects the basic hypothesis rather well like in the transcriptomics case CASE I. The idea that nickel allergic patients

will respond to nickel exposure is reflected in the design, but the exploratory element is much more unsubstantiated, and the cost of this has not been realized.

Apparently, the fact that (almost) the entire ome can be analyzed misleads the researcher to believe we can tell the full story by measuring 20 samples. The more focused the questions are, the easier it is to build a strong design which will focus upon this. Too often we end up in a situation where the biologist wants to know "everything" based on very few samples. The result is very low power and high risk of over-fitting. Focused research questions assist defining relevant descriptor and response variables.

Targeting of the descriptor matrix to contain a high fraction of relevant variables improves the statistical power of the experiment. It is highly relevant to consider how exploratory the experiment "can afford" to be; pilot studies can indicate this.

Uncertainty in the response variable should be limited as has been pointed out in the literature several times, e.g. [32, 13]. Obviously, a case-control search for susceptibility genes for a certain disease suffers serious power reduction if say, a group of the controls were actually genetically susceptible to the disease but were not exposed to the disease. Moreover, including many subtypes in the class variable may reduce the effect size.

It is valuable if the phenotypes (classes) to be modeled represent extremes; the larger the variation between classes, the larger the effect size.

In the previous chapter it was mentioned that one of the classes is often much less available (e.g. the patient class). This is a fundamental problem, but it is important to realize beforehand that it may come at a cost to merge several subclasses, although sometimes it may be successful. In these situations it should be emphasized even more to focus the study towards expected relevant biomarkers.

In Pere-Trepat *et al* (2010) [48] the same NMR dataset as in PAPER IV is analyzed for a different purpose; here the intention is to relate ^{1}H NMR CPMG plasma profiles to reported dietary habits. This is an example where the classes are not necessarily well-defined, and it may be difficult beforehand to define

them. This was solved by calculating a PCA model based on the dietary reportings ("how often do you eat apples?", "How often do you eat fish?" etc). The participants seemed to cluster according to some dietary pattern, which could form the basis of dietary classes which were afterwards related to the NMR spectra through PLSDA.

Know your background

It is an essential part of data analysis to be able to assess methodological repeatability and variations within individuals, such as diurnal and day to day variations. It is usually difficult to obtain true biological replicates, i.e. repeated experimental measurements from the same individual, but efforts should be made to repeat as much of the process as possible, i.e. more than one blood sample should be taken from the same individual, it should be measured on different days, samples should be taken from similar subjects etc.

One way to get to know the background variation is through pilot studies, which are excellent means to start mapping inter- and intra-individual variations. Pilot studies are often performed in order to make sure the assays and instrumental part work satisfactory. What is sometimes missing is the pilot study which gives an idea about effect sizes and so assists in dimensioning of the study. Or the pilot study which investigates the underlying variations such as clustering according to storage time or other experimental artifacts.

Essentially, many omics studies originally intended to be full-scale experiments end up as pilot studies because fundamental parts were not adequately clarified initially with the result that findings were negative or very weak. A lot of insight and experience is gained through such a study, but it is a very costly learning process. To some extent this is true for all the case stories outlined in the beginning of this thesis.

Control variation

As described in the previous chapter, biological variation in omics is massive and this makes it difficult to discover and quantify the systematic parts, in par-

ticular because we often look for subtle effects. One way around this is to try to reduce the included variation.

We must have variation but we do not want too much. On the one hand we like to span variation in order to be able to draw conclusions which are generally valid. It is of little interest to make conclusions which are so specific that they are only valid for women aged 27 with red hair, who never smoked, shoe size 39 and born in a specific town in October. However, sometimes this may to some extent be a sound road to follow. Start out simple, find candidate markers in such a narrow study and expand the variation in the next study with focus on the candidate markers instead of the full set. To THE BIOLOGIST such a study may appear under-ambitious, and the possibility of short-cutting is probably more appealing in many situations. However, this is the price of including so much variation as opposed to much more targeted approaches.

5.2 Reduce analytical error and bias

It is important to realize that the choice of analytical methodology has some important implications for the obtained omics data. Obtaining high signal to noise ratio (S/N) for relevant signals and at the same time avoid introduction of bias is the task. Sample handling may be critical for some applications, the analytical setup represents choices in terms of selectivity and sensitivity for certain types of signals, and analytical artifacts are always present to some extent.

As pointed out, it is not the intention to go into details regarding the analytical setup. However, a few general issues are highlighted below.

Proper sample handling

As illustrated with the proteomics case story (CASE II, section 1.2), improper sample handling may have detrimental effects to the samples or introduce bias. This should be considered in relation to training of clinical staff, including the PhD student who will do a lot of the practical work across the experimental process. Make sure the necessary knowledge concerning stability of the specific

bio-fluid is available. Various papers address sample collection and handling issues for bio-marker studies, e.g. [26, 60]. Elements to consider are volatiles (e.g in urine), use of stabilizing and preservative agents - how do they interfere with the sample matrix?

This topic is highly relevant, nevertheless quite difficult to deal with in practice. Collecting 24 h urine from human beings in a normal daily life setting (CASE IV, section 1.2) obviously enforces a lot of practical challenges. Thus, collecting "everything" disregarding the situation and ensuring well-controlled storage are not easily taken care of.

Handle batch variations

It is often impossible to measure all the samples in one batch and as a consequence, batch to batch variations may be introduced. One way to assess this effect is through the use of quality control (QC) samples. Bijlsma *et al* ([7]) pool blood samples from case and control samples separately and measure these with each batch. This way batch variations are not as such taken care of, but they are at least assessed and a potential source of bias has been addressed.

Choose appropriate analytical platform

Choosing the best suited method can reduce problems significantly. Choose a platform which will give strong signals for the analytes expected to be relevant in order to optimize S/N. If you expect lipo-proteins in a blood sample to be important, do not measure it using a CPMG pulse sequence, which suppresses macro-molecules in order to enhance signals from smaller molecules. Beware of the limitations of the chosen method; ^{1}H NMR for example, has a limit of detection around 10 μM (down to nM in setups with small-diameter cryoprobes and high field strength) [47].

A number of papers deal with the details of sample handling and analytical procedures for optimization of S/N and minimization of analytical bias in omics studies. Several of these are relevant across the different omics.

transcriptomics	[6]
proteomics	[44, 49, 63, 33]
metabolomics	[4, 57, 39, 42]

The reader is referred to these for details on this highly important analytical area.

CHAPTER

6 Prepare data for analysis

6.1 Data preparation is the most important part of the data analysis

In order to utilize the advantages of PLS, data must be in a proper form when parsed to the PLS processing. Fundamentally, there should be an approximately linear relationship between $\mathbf{X}$ and $\mathbf{y}$ and $\mathbf{X}$ and $\mathbf{y}$ should be multivariately normally distributed.

The best results are obtained if the fraction of relevant variables is high, noise is low and the relevant signals are large. This is usually not the apparent characteristics of omics data as shown previously (chapter 4). In the data analytical process, it may thus be worth spending a major effort reducing and restructuring data to a form which is more suitable for e.g. PLS. This is the process of *data preparation*.

Omics data contain too many variables

As described in Chapter 4, with M variables in a dataset, it is possible to combine these in $2^M - 1$ ways. For $M = 1000$, which is a quite low number in an omics context, this means $1.0715 * 10^{301}$ combinations, and if we were completely ignorant about these data, we would have to try all combinations to find the optimal solution, which is problematic from a computational point of view and undesirable in terms of statistical power. In other words, even if we possessed the computational power to try all combinations the result would need very careful validation because the risk of chance correlations, and thereby over-fitting, is extremely high with these "fat" matrices.

However, this is where the use of *a priori* knowledge comes to play a major role. Generally, we do have some knowledge about the underlying structure of data which can help us analyze data in a way which we expect to be meaningful. In the following section I will shed some light on how knowledge about data (i.e. about design, analytical technique, biology) can be used to prepare very complex high-dimensional omics data for PLS modeling by reducing them to much more manageable datasets without a major loss of information and with biological interpretation in focus.

6.2 Informed data preparation

Most of the variables in an omics dataset are irrelevant for our purpose and this harms performance of the (PLS) data analysis as shown previously (Section 4.4). The high level of noise and the low number of samples makes it very difficult to reliably estimate the latent structures in data. However, we do know a lot about data and if we make use of this knowledge in the data analysis to (1) get rid of a lot of the noise and (2) structure data in a way which enables the data analysis to give results that are directly biologically interpretable, then it is much more likely that the data analysis will produce useful results. In this way, we try to compensate for the low number of samples by guiding the modeling process by our knowledge.

The actual steps involved in this *informed data preparation* are data specific and obviously also depend on the intended processing method, but for PLS which is the focus of the present work, some relevant elements could be subset selection (variables and/or samples), alignment, averaging, and compression to fewer but informative variables which represent data at a level which is suitable for interpretation. In the following this will be exemplified using data from different omics domains.

A priori, we know signals are related

Depending on the type of data, we may have information about the way signals are related chemically. This may be at various levels, e.g.:

- Neighboring spectral variables are members of the same peak (i.e. we know about peak shape).
- One analyte may give rise to several signals; for example:
  - Spin-spin coupling patterns in NMR which result in peak splitting (this is actually one signal, but resembles several peaks).
  - Each unique H in a molecule gives rise to one signal.
  - An mRNA transcript may bind to several probes of a microarray chip.
  - In mass spectrometry, a molecule is split into several fragments each giving one or more signals (+ adducts).
- The analytes may be related in various ways, they can e.g.
  - belong to the same type of response.
  - belong to the same pathway.
  - belong to the same cellular component.
  - belong to the same chemically functional group.

On top of this, any knowledge regarding uninformative areas, e.g. the water suppressed region in ^{1}H NMR, signals conflicting with detection limits or transcripts which have not been annotated with functional terms, can be used to eliminate irrelevant variables.

There is a lot of redundancy in data, and therefore a great potential for compression. By a priori knowledge this compression can be guided to lead to meaningful new variables (e.g. functional genetic terms or specific metabolites).

The major benefits of data compression

- Data reduction/compression - computationally important!
- An obtained data structure which is much lower in noise level, thus scaling is less prone to boost the influence of noisy variables.
- A dataset with a substantially higher ratio of relevant to irrelevant variables.
- A set of chemically/biologically meaningful variables which may possibly be examined at various functional levels.
- A reduced dimensionality of the model space, whereby model uncertainty is decreased.
- Improved power.

The way a given dataset is meaningfully compressed is very dataset specific, and experience must be built up for each type of data. Following a brief introduction to the subject of ontologies, some examples of how informed data preparation can be applied to transcriptomics data and ^{1}H NMR metabolomics data are given below.

Ontologies

Ontologies represent a means of structuring knowledge into a system suited for computer-assisted high-throughput analysis. Examples of such formal represen-

tation of knowledge can be found through the *Open Biological and Biomedical Ontologies* (OBO) - an initiative establishing principles for the development of ontologies within the biomedical domain. Several ontologies relevant for omics data are well-established OBOs including the Gene Ontology (GO) and the PRotein Ontology (PRO).

The ontologies are very comprehensive and complex database systems which are inter-operable with other OBO ontologies. In the GO database, available information has been coded into three domains: molecular function, cellular component and biological process. The GO-terms associated with a given genetic sequence (e.g. a probe on an Affymetrix microarray chip) are termed annotations and can be used to group genes according to selected properties.

Examples of informed data preparation

In PAPER III the GO database is used for a functional genomics analysis. Subsets of variables were selected on the basis of functional annotations and modeled individually. The outcome of these models were then input to a joint model.

PAPER V is also a transcriptomics application, but the design and data analytical approach is different from the type treated in this thesis, as it does not involve PLS or comparison between two classes. Rather an unsupervised MCR approach is applied to identify which genes in *Lactoccus Lactis* dominate the different parts of a fermentation process. Nevertheless, it is relevant to mention the data preparation process here.

In this application, data were cleaned by comparison of replicates to eliminate from further analysis genes whose expression was not consistent. Hereby an important element of uncertainty was eliminated. In this way replicates can be used to qualify data and to estimate the uncertainty.

Filtering by putting thresholds on reproducibility, signal amplitude etc. is a means of avoiding false discoveries. Other simple means are univariate t-tests in particular if the consistency of such variable selection is assessed e.g through bootstrapping as in PAPER IV.

^{1}H NMR metabolomics

In PAPER IV informed data preparation is incorporated for the analysis of ^{1}H NMR plasma profiles in a metabolomics context. The data preparation which compressed the spectra from 32K data points to 171 variables is outlined in the paper, but in the following further details about the procedure are given.

The NMR data

The dataset consisted initially of 1116 NMR spectra of human plasma samples as described in the paper. The spectra were already phase-corrected, Fourier transformed using a line broadening of 1 Hz, and an initial baseline correction had been applied. The spectra were imported into MATLAB for the data analysis.

Spectral concatenation and data cleaning

Each spectrum was referenced against the alpha-glucose doublet peak at 5.23 ppm using a correlation optimized shifting (coshift) procedure [58]. After that, the spectra were interpolated to a common axis with 32K data points between 0 and 8.5 ppm and concatenated to form a matrix. The spectral tails were trimmed to 0.1-8.25 ppm and the water region 4.55-5.17 was removed. The spectra including the water region are shown in Figure 6.1.

Subsequently, the spectra were inspected for obvious non-conformities, whereby a couple of outlying samples were identified. A number of samples were not well water suppressed as can be seen from the broadened water peak (approx 4.78 ppm) in Figure 6.1 and Figure 4.1.

Peak integration

The scope of the next step in the data preparation was to compress data by peak integration. The strategy was to define each peak and model it by a PCA or an MCR model. In theory, if peak regions could be defined that contained only one peak, then a one-component model should be able to model the peak, and the

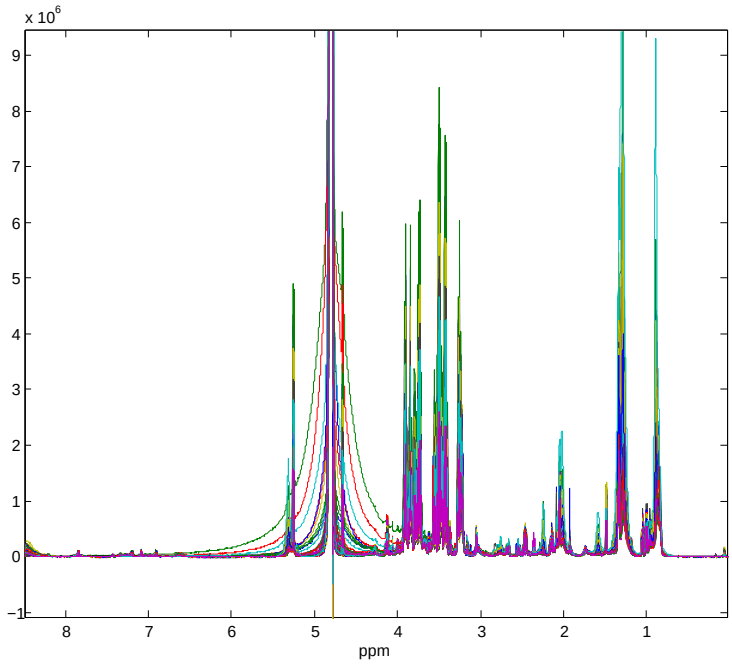


Figure 6.1: Plasma NMR profiles (sample subset)

score value would then correspond to the peak integral.

As shown previously, (Figure 4.2) some peaks are shifted as a result of pH variation in the samples, although plasma is actually a rather strong pH buffer. Consequently alignment is required prior to peak modeling.

A peak region was defined manually for each peak along the chemical shift axis. Whenever possible, doublets, triplets etc were put together in the same region, but often overlapping peaks would hinder this. Focus was to represent all peaks, rather than retaining structural information as this could be restored at a later stage for selected compounds. During this process, only peaks which

were significantly above the noise level were included. For each peak it was evaluated whether it was convincing enough to be shown in a publication if it turned out to be a biomarker, and excluded it was not.

Some peaks would be so tightly and complexly overlapping that it was not possible to define separate regions for all these areas, and some intervals would therefore contain more than one peak in practice.

The spectra were then aligned using the *I*coshift algorithm [51], which shifts peak intervals individually along the axis while preserving peak shape. Prior to this each interval was inspected and it was decided whether alignment was required. The effect of alignment was examined for each peak. Some peaks were rather straight-forward to align whereas others required more parameter optimization in order for good alignment to be obtained. In some cases alignment of the Savitzky-Golay second derivative of a peak provided very good results, when the hereby obtained spectral shifts were applied to the original (non-derivative) spectra afterwards. A before-after alignment example is shown i Figure 6.2

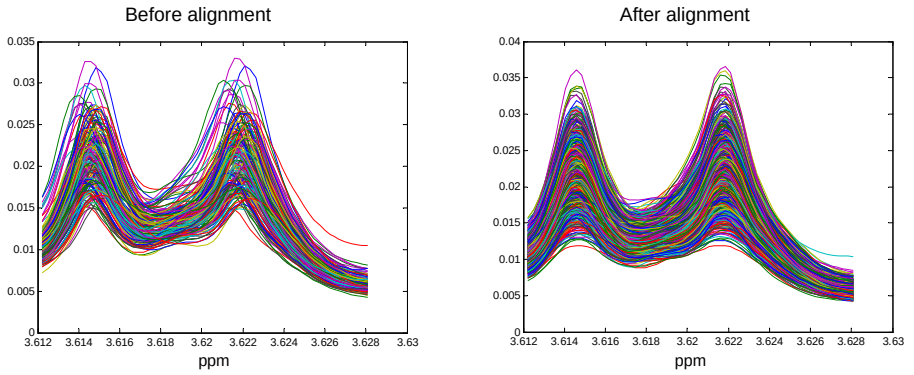


Figure 6.2: Effect of alignment shown for one doublet. Left: before alignment, Right: after alignment

Some peaks were small shoulders on a broader peak and integration would therefore mainly reflect the large background peaks. In these cases baseline correction was performed where the contribution of the large peak was minimized; this procedure required a temporary broadening of the interval.

The next step was to develop a PCA model for each interval. Following, each of these was inspected and compared to the original peak to make sure the obtained loading actually reflected the peak it was intended to model. In some cases the intervals had to be resized to obtain this, in a few cases two components were needed and in a few cases MCR models provided more chemically meaningful loadings.

Eventually, collecting the scores from each of 164 defined peak regions yielded 171 new variables. Thus by this procedure, the number of variables was reduced to a much more manageable size and the noise level was lowered substantially.

All these steps were quite cumbersome and not very appealing if one who generally seeks automation of such routines. However, considering the time it took to collect the data, this process was really insignificant and concurrently a lot of insight into the details in the data was acquired.

Summary

Omics data contains massive data amounts that are not easy to manage on well-equipped standard PCs. Moreover - and most importantly - most of the data is irrelevant and possibly harmful to the modeling, in particular in cases of low effect sizes. However, if efforts are spent on cleaning and compression of the complex **X** matrix, PLSDA may be very well suited for omics classification purposes.

Applying a priori knowledge about the structure in data is a very useful way to compensate for the general problem of too few samples by constraining the possible model outcome. Knowledge about the analytical method and the biology behind can be used for *informed data preparation* whereby data is put in a form which is well suited for PLSDA, and which leads to more readily interpretable results. How to perform this data preparation is highly data- and

application specific and requires that THE DATA ANALYST consults both THE ANALYTICAL CHEMIST and THE BIOLOGIST. This part of the data analysis may be rather time-consuming but considering the time frame of the data collection, it is generally absolutely worth the effort.

CHAPTER

7 Analyze data properly

The previous two chapters dealt with collection of high-quality data and preparation of data for data analysis. Hereby the scene is set for the actual data analysis.

A very general objective of the data analysis within the omics area is classification and identification of biomarkers. This chapter is centered around this quest. Focus will be on PLS used for discriminant analysis, PLSDA, which is widely used within the omics area. It appears that as omics has processed through the system levels from genomics to metabolomics, the influence of chemometrics has increased, and as a consequence, PLSDA and its relatives (e.g. OPLS) are particularly well-spread methods within the field of metabolomics.

7.1 The work-flow of PLS for Discriminant Analysis (PLSDA)

Discriminant PLS (PLSDA) is a special case of PLS used for classification purposes. A regression model is built between the multi-dimensional dataset $\mathbf{X}$

and a "dummy" class variable **Y** as described in section 2.1.

The PLSDA workflow

After data cleaning (as described in Chapter 6) the PLSDA workflow consists of the following primary steps:

1. Scaling
2. Modeling
    a) Outlier detection
    b) Variable selection
3. Identification of discriminatory variables (biomarkers)
4. Validation

This workflow is very generic for any PLS based modeling, and this is a main point of the present work. If data has been properly prepared for the modeling, then standard procedures can usually be applied with good chances of success. In other words, the majority of the data analytical workload should be put into the data preparation phase.

As a consequence, this chapter will be rather short; a few comments are given on some issues, viz. scaling, variable selection and selection of biomarkers. Eventually, the need for validation is emphasized.

7.2 Identification of biomarkers

Scaling

Scaling is a pre-processing step, which aims at bringing data to a form which makes the processing method suited for the purpose at hand. Projection tech-

niques such as PCA and PLS work on covariance which is highly dependent on numerical levels.

For data measured in different units, scaling is generally required in order to avoid that the projection only focuses on the variation spanned by variables measured in small units. The variance of a variable measured in millimeters might dominate a model completely, whereas if the same variable was measured in kilometers its variation would be insignificant and thus not reflected in a projection model. A popular means of scaling is scaling to unit variance; when combined with a mean centering operation this is called auto-scaling. With auto-scaling all variables are given the same numerical influence to the PLS model.

For data measured on the same scale, e.g. spectral data, scaling is not always required. However, peak intensities generally reflect concentrations and these may not at all be proportional to the biological interest of the peak. Rather, the small variance in a small peak may be just as interesting as variations in a large peak, and it is not unlikely that many new discoveries are to be found in the smaller peaks which are more difficult to handle than the "easy" analytes in higher concentrations. Consequently, scaling may also be desirable for spectral data.

The downside of auto-scaling is that the influence of noisy variables with low intensities is increased dramatically, which is undesirable. The result may be false conclusions based on random correlations with noisy variables. This is a major reason why the noise level in data should be reduced in an initial data preparation process (chapter 6).

As a result, various other types of scaling have been suggested; a popular type of scaling within metabolomics is pareto-scaling [16] in which the intensity of a variable is transformed by subtraction of the mean value and scaling by the square-root of the standard deviation. The square-root operation on the dispersion is the only deviation from auto-scaling.

Van den berg et al [61] has made a thorough investigation of the effect of

various scaling types on the outcome of metabolomics data analysis. Table 1 in this paper gives an excellent overview of the properties of selected scaling types.

A major benefit of the concept of informed data preparation as described in the previous chapter is that the noise level is considerably reduced. Representing NMR data as peak integrals obtained through PCA modeling of each peak produces a dataset which is almost noise free, and thus enables auto-scaling.

It is important to be aware that as scaling governs how the model inputs are allowed to influence the model, the manifestation of potential biomarker candidates will depend on the scaling method. Sometimes, this phenomenon will invoke some skepticism in THE BIOLOGIST'S trust in the modeling. Validation is the only way to test if the scaling (and remaining data analytical setup) is reasonable.

Selection of biomarker candidates

Various diagnostics and methods exist which can be used to select the variables with the best discriminative power. It is outside the scope of this work to make a thorough examination of these; rather I shall refer to [2] for a good overview.

An additional principle of variable selection not mentioned in the reference above and which has gained increasing attention in the past years is the use of regularization, which comes from the fields of machine learning. With regularization, a penalty is put on the parameters, whereby the model is pushed in certain directions. Various types of regularizations exist; in particular the L1 norm regularization has gained attention.

Whereas L2 norm regularization of a regression vector penalizes large elements, L1 norm regularization penalizes all values equally. The result is a *sparse* regression vector, i.e a vector with few non-zero elements, hence interpretation is emphasized.

7.3 Validation, validation, validation!

As has been shown in the previous chapters, omics analysis includes a high risk of chance correlations, and the statistical basis of finding strong evidence is weak. Because of this, when the list of biomarkers has been identified, it is an

absolute requirement that these be validated appropriately before it is relevant to conclude anything about their existence. Most likely, many false discoveries encountered within omics would have been avoided with adequate validation.

Validation is a process working at several levels (see e.g. [24] for an extremely thorough elaboration on this). Assessing the statistical validity of (1) the method and (2) the findings represent two core elements, but another level which is just as important is the biological validation; the process of assessing the results in a biological context based on the available knowledge of the area.

It is difficult to give a clear definition of adequate validation. Optimally, the full experiment (sample collection - data acquisition - data analysis) should be repeated. However, this is rarely feasible to do in full scale within the structural framework of the research; for example, this is most likely not possible to do within the time-frame of a research donation or a PhD education. Hence, a more pragmatic path may be suggested.

As mentioned previously (section 5.1) the use of pilot studies is encouraged. These sub-studies validate the reproducibility and assess uncertainties of the procedures and background variations in the experiment and are thus extremely valuable in terms of consolidation of conclusions. It may be beneficial to use smaller substudies iteratively in the procedure of optimizing the experiment.

Often, samples and data move physically during the experimental process, and these transfer points represent natural borders of validation "boxes". Along with the final data, the data analyst receives the knowledge about the uncertainties of the data collection and analytical processes obtained through the previous validation steps. The data analyst can then focus on the data analysis and the validation of this.

Validation of the data analysis can be performed at several levels. Again, a representative independent test set is ideal but as this is not always available, other approaches may be used. Cross-validation tends to overfit data, but used with sound reason cross-validation may provide a reasonable means to evaluate

parameter estimates throughout the process. Below, a nice implementation of cross-validation named *cross-model-validation* is illustrated.

However, first the importance of validation (and the proper use of scores plots) is underlined by an example showing the capability of PLSDA to separate groups on the basis of random numbers.

Example: PLSDA

Let us form two datasets **X** and **y**. **X** is a matrix of random numbers of dimensions [50 x 100], i.e. consisting of 50 samples and 100 variables. **y** is a [50 x 1] vector, where $y_i = 1$ for $i = 1,...,25$ and $y_i = 2$ for $i = 26,...,50$. A PLSDA model is built upon these (non-sense) data. The resulting scores plot for PLS components one and two is shown in Figure 7.1.

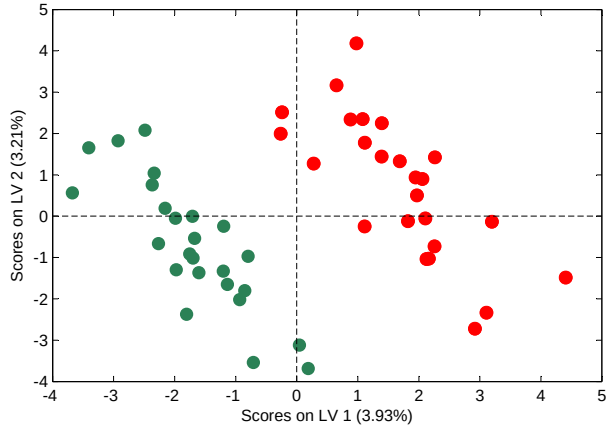


Figure 7.1: The PLS-DA scores plot is appealing but dangerous. Scores plot of PLSDA model based on 50 samples of 100 random variables randomly assigned $(50-50)$ to two classes shows perfect separation

Apparently, the two classes can be separated perfectly using PLSDA! How-

ever, as would be expected, cross-validation underlines that there is no model (Figure 7.2).

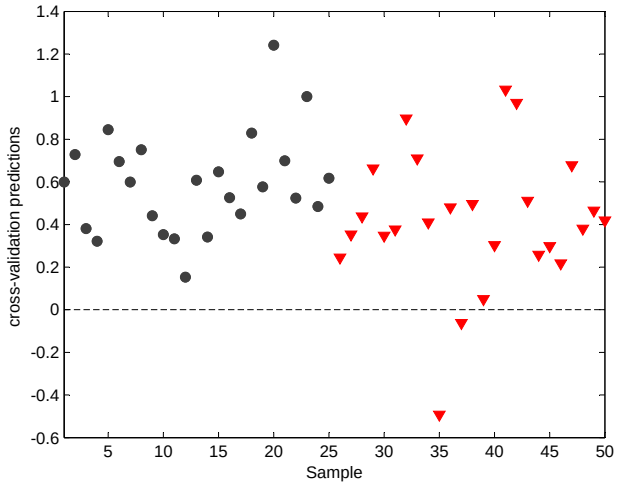


Figure 7.2: Validation of PLS-DA model built on random numbers shows that there is no model.

Cross model validation (CMV)

In the very clear non sense PLSDA example shown above, cross-validation is of course absolutely adequate to enlighten model problems. In general however, ordinary cross-validation (CV) results in overfitting, in particular because the cross-validation normally applied does not handle the full modeling procedure. Optimally, preprocessing, scaling, PLS parameters and variable selection should also be validated. A way to do this is through so-called Cross Model Validation (CMV) [3, 66, 22], a procedure which has been developed independently a couple of times and hence given different names: “leave one object completely out”

(LOCO) [46] and "double-cross validation" [3].

CMV follows the same fundamental principles as CV, except that it works across more or less the full modeling procedure rather than just the PLS modeling itself. That is, prior to any modeling, data is split into segments. One segment is then left on hold while the remaining segments are used to build a PLS model, including pre-processing, modeling, variable selection etc. These steps may involve CV for parameter estimation. Following this, the left out segment is manipulated using the obtained model to yield a prediction for each sample in the segment. This prediction can be compared to the reference in the usual CV way. The procedure is repeated with each segment left out sequentially, resulting in as many sub-model-parameter sets as there were segments. It has been shown that the error estimates obtained by this procedure are much more realistic than those obtained by CV.

It should be emphasized that CMV does not necessarily lead to an optimal solution, but it estimates the validity of the found ensemble of sub-models.

Biological validation

A core element of validation of results is the biological validation. The biological validation is the sound reasoning of the biologist to speculate if the results make sense in a biological context.

In order to be able to perform this validation, it is necessary that the results are in a form which makes sense to the biologist. Very complex transformations such as kernel methods or neural network models bring the variables to a form, where interpretability is widely lost.

7.4 The iterative process

In order to obtain interesting and meaningful results from the data analysis, it is important that THE DATA ANALYST take part in a iterative process with THE BIOLOGIST and THE ANALYTICAL CHEMIST. The data analyst should present data to the other two parties in a way which makes them (1) identify

possible artifacts that should be dealt with, and (2) generate new ideas or lead the data analyst to focus on specific areas of the results. The ability to visualize such complex data in various compressed ways is a required competence from the data analyst. Sharing of knowledge is a key element in successful data analysis.

CHAPTER

8 Conclusion

Numerous high-dimensional so-called 'omics' studies have been carried out throughout the past 10-15 years with the purpose of identifying biomarkers and obtaining a better understanding of the complex biology of our organisms. These experiments are generally very costly and time-consuming and the results have not been overwhelming. As a result, some skepticism with respect to the potential of the omics technologies can be observed.

The main objective of this thesis was to characterize omics data and thereby try to understand why they are difficult to handle with the standard chemometrical approach which is well-established for many other data types such as NIR data. On this basis, suggestions for how to improve the outcome of omics experiments should be given.

Characteristics of classical applications

The dominating chemometrical regression method is Partial Least Squares (PLS) regression and this method has also been widely used within the omics fields, in

particular the special use for discriminant analysis, PLSDA. PLS was developed and gained its popularity in a historical context where instruments produced data amounts which traditional methods could not handle.

Near infrared (NIR) data was in this thesis characterized as a representative of the classical data types. NIR data are multi-dimensional but low rank, and the typical PLS application on such data would involve a hundred samples of some industrial biologically based intermediate or end product. The samples would be rather homogeneous as the production setting is optimized with this as a quality parameter. A classical response variable would be one of the major components fat, protein or moisture which is directly linked to the information in the spectrum. The reference values might have been obtained by a well-established wet chemistry method with low uncertainty.

The low rank means that a hundred samples is sufficient to describe the system with a low uncertainty, so systematic and random variations can be well separated by the PLS model. The relevant information is present in data and the signal to noise ratio is high.

Characteristics of omics applications

Although the description above tends to over-simplify the data analytical challenges associated with NIR type applications it serves to underline some fundamental differences between NIR and omics data. Omics data are also high-dimensional and usually with many more variables than NIR data. However, they are not simple extensions of NIR data.

This thesis has focused on omics experiments involving human subjects, and such studies introduce some important differences relative to NIR applications.

First of all the variation in data is much larger and more complex for several reasons. Omics analytical technologies focus at measuring a much wider palette of analytes and as such potentially provide a much more detailed picture. More-

over, the biological variations, i.e. the genetic background, the environmental influences and their interactions, are diverse and complex and difficult to control in experimental setups.

In omics experiments the response variable is often not particularly well-defined and the link to the descriptor matrix may be less direct. On top of this, analytical bias or errors are easily introduced.

The net result is that the effect size (or signal to noise ratio at application level) is lower than for a typical NIR application. And whereas NIR data are high-dimensional but low rank, omics data are high-dimensional high rank data. In other words, the systems being analyzed in omics studies are much more complex, and as a consequence, many more samples are needed to describe them well.

The access to samples is generally more problematic and the number of samples is therefore comparable to those of NIR experiments. Power is therefore low for these experiments as a whole and obviously, the risk of overlooking small effects or making false discoveries is much higher.

This effect is also reflected in the performance of PLS as this method also needs high signal to noise to extract the true systematic pattern. With many irrelevant variables PLS does not perform well; since the PLS model maximizes covariance between descriptor matrix and the response variable, random correlations will harm the model. As a result, weak signals are not detected and random correlations may appear to be real.

It is a characteristic of omics studies that the experiments are large and involve several experts across different scientific environments. Some central elements of the practical part of the experiment include recruitment, training of staff and subjects, clinical sample and data collection, and application of advanced analytical techniques. Other central players include the biologist who originated the project and the data analyst who is supposed to extract the results.

How to obtain better omics results

For an omics experiment intended to identify biomarkers of a given *condition* the successful outcome is one or more biomarkers which can be reproduced in a similar experiment. Getting better results from an omics experiments is about realizing the issues above and try to minimize their influence through (1) collection of high quality data and (2) good data analysis.

Good data collection

Characteristics of well collected data

- a well-defined response variable $\mathbf{y}$.
- a targeted descriptor matrix $\mathbf{X}$, i.e. data which are relevant for the given purpose.
- a fundamental sound biological relation between $\mathbf{X}$ and $\mathbf{y}$.
- minimized analytical error and bias.

The design, planning phase of the experiment should optimize this, and in the practical phase skilled personnel should ensure the realization. An important question in the design phase is to set the balance between targeted and exploratory analysis. This choice represents a trade-off between on one side high power and signal-to-noise and on the other side the fascinating exploratory possibility of discovering new non-hypothesized phenomena.

The value of pilot studies is emphasized as not only the data but also the experimental setup is complex and thus involves a high risk of introducing analytical errors and biases. Moreover, pilot studies can assist targeting the data acquisition towards the most relevant signals and also indicate how exploratory a given experiment “can afford” to be.

Good data analysis

The data analysis should ensure that present biomarkers be identified and that these are generally valid and not a result of chance correlations and over-fitting. This is about applying a good modeling method, optimizing the input to the modeling process and validation of the results.

Applying PLSDA as the modeling engine, optimizing model input involves reduction of the fraction of irrelevant (noisy) variables to increase the chance of identifying the relevant biomarkers. This is signal-to-noise enhancement and will for any model improve chances of success if it is performed in a sound manner.

Model validation should be performed at several levels. Both in a strict technical meaning by testing the model with ideally an independent test set, but also validation by relating the outcome to prior knowledge and expected results ("biological validation"). In order to perform the latter, the obtained results must be presented so that they are directly interpretable by the biologist in collaboration with the data analyst. This may require some kind of data preparation which transforms the data into other more meaningful variables.

Consequently, intelligent preparation of data for the actual classification may be a key element in successful outcome and to some extent compensates for the insufficient amount of available samples. Such data preparation is very dataset specific; in this thesis some examples are given. PAPER III deals with transcriptomics data, and PAPER IV compresses ^{1}H NMR CPMG metabolic profiles into a very manageable number of input variables. Data preparation can be very time-consuming and very manual but considering the time it has taken to collect data, I believe it is worth the effort.

When data has been properly prepared for the data analysis, the actual PLSDA modeling is relatively straight-forward and does not differ remarkably from the data analysis of NIR data. It is a main conclusion of this work that if data has been properly collected and prepared, then PLS is generally suitable for analysis of omics data. The problems with low signal-to-noise levels in omics data are fundamental and will challenge any classification method, not only PLS.

The data preparation based on external knowledge is therefore highly relevant irrespective of the choice of classification method.

The complexity of the data and the limited number of samples, hence higher risk of over-fitting enforces the need for proper validation; biological validation should be a natural part of this.

The clarity of hind sight

It might appear that this thesis is largely built on criticism of hard work performed by otherwise respected scientists. As stated previously, it is not the intention of this work to criticize anyone in particular; rather I think the cases presented in this thesis represent aspects of omics studies which are found to some extent in many studies.

It seems easy to criticize in retrospect, but I fully respect that these experiments were performed utilizing the knowledge present at the time being. Familiarizing with the omics technologies is a process for all parties involved.

8.1 Conclusions in brief

Conclusions in brief

1. The standard operating procedures (SOP) for chemometrics analysis are developed in an era of NIR type data and works well with this type of data.
2. Omics data are not just wheat NIR data with more variables.
    - There are generally much more variables, whereas the number of samples is comparable or smaller.
    - The rank is higher.
    - The amount of irrelevant data is substantial.
    - The biological background diversity is much broader.
    - The causal link between the large-scale dataset and the response variable is often weaker and much more complex.
3. Experimental design and data collection deserve special attention.
    - All relevant competencies must be involved in the design phase and protocols approved across the involved scientific units.
    - Pilot-scale experiments are necessary.
    - High-quality data require great expertise.
    - Targeted analysis has clear advantages.

continued on next page...

Conclusions in brief ...continued

4. Preparation of data should be emphasized in the SOP for data analysis of omics data.
    - A priori knowledge is essential for a meaningful data compression and structure
5. PLS is suitable for analysis of omics data with proper data presentation.
    - Omics data analysis is prone to over-fitting.
    - Scaling and selection criteria influence identification of biomarkers.
    - Data knowledge is fundamental for successful data analysis.
6. Validation cannot be over-emphasized.
    - Key elements of the experiment should be validated (measuring chain validation).
    - Over-fitting should be counter-acted by proper validation.
    - Biological validation is essential.

CHAPTER

9 Some perspectives

The present analysis of the properties of omics data relative to more traditional data types leads me to ask and comment on a few questions regarding some very general aspects of omics experiments.

1. Focusing specifically on the data analytical part, it is obvious that the starting point of the analysis is mathematically and biologically more complex than that of NIR data analysis, and it is evident that the number of samples is a limiting factor. Does it make any sense to perform the experiments at all?

2. Automation through implementation of methods into commercial software has been an important factor in the successful spreading of chemometrics previously. To what extent is omics data analysis suitable for automation and implementation of such routines into software. Can the data analyst participation be excluded with the right software at hand?

3. Some of the suggestions in this thesis invoke some changes in the traditional chemometric mindset, notably with respect to the exploratory spirit.

4. It is clear that omics experiments as a whole - not just data - are complex: (1) They require expert knowledge across several scientific areas, and (2) these competencies must be strongly coordinated in order to ensure sufficient exchange of knowledge across the chain. In a traditional university setting, (1) is probably the easiest to recognize, whereas the importance of project management is much more overlooked. This topic is discussed briefly.

Worth the effort?

For THE BIOLOGIST the high throughput nature of omics technologies is very attractive, whereas the (conservative) data analyst may be considerably more skeptical due to the limited number of samples. As has been shown in this thesis, the combination of an un-targeted approach and few samples is difficult. As a result the trade-off between an exploratory element and the statistical power must necessarily result in a compromise.

I think it is important to realize this compromise, and in the design phase identify where the given experiment should be located on this axis. This is also important to communicate when publishing results.

Some disappointment can be felt in the omics fields, mainly due to lack of general validity of discovered biomarkers. Great work has been done trying to set publication standards for the various omics areas ([23, 56]) and I greatly welcome this work, particularly the requirements to state validation details, hence forcing the focus upon this.

However, most likely the situation of having too few samples is not going to change dramatically within the coming years. Nevertheless I do believe that valuable findings can be made from omics experiments if they are optimized. I think we have to accept that omics studies only rarely will provide very strong conclusions; it is a basic condition but it does not mean we should not publish useful results. I advocate a culture where authors do not oversell their findings as valid markers if they are de facto only indications of such and a culture, where the reader should not expect to find the ultimate *truth* in one paper claiming the

discovery of a biomarker. Rather the repeated finding of a given marker across experiments serves as the validation of it.

Automation?

The work flow of routine PLS data analysis has to a large extent been implemented in commercially available software (see PAPER I). By such software, non-chemometrics specialists are capable of performing standard analysis.

However as PAPER II shows, the fact that non-skilled users apply the chemometrical methods does result in some commonly occurring mistakes.

The question is to what extent omics analysis can be automated, as requested e.g. in [41, 55]. Meaningful automation would both save time for chemometricians and could enable non-chemometricians (i.e. THE BIOLOGIST or THE ANALYTICAL CHEMIST) to perform the analysis themselves.

As shown in this work, the process of data preparation prior to analysis can be a key factor for successful outcome of the experiment. This process is to a large extent ad-hoc in the sense that every problem requires its own procedure. However, some general work-flows and procedures can probably be established for various types of data and problems and this way procedures can be semi-automated.

As shown in this thesis (particularly PAPER II) the risk of automation is that it tends to make the user forget his responsibility. Automation requires special attention be paid to verification that the automatic routine produced useful results. It is questionable whether a non-data analyst would be able to do something useful if the automation turned out to produce poor results.

In essence, I believe it is possible to make graphical user interfaces with some sets of standard routines relevant for specific kinds of data types and to build up some experience outside the data analytical world so that some of the data analysis can be performed by e.g. the analytical chemist or the biologist.

However in my view, omics data are generally too complex to be analyzed completely without the competencies of the data analyst.

Changes in the chemometrician's set of mind

The exploratory spirit present within many chemometric environments has often resulted in a welcoming of many variables from experiments - to some extent in a "the more the merrier" fashion.

The fact that the number of samples represent a serious limitation - due to the number of variables and due to the complexity of the data - is rather new within chemometrics, and it has been a process to realize that we cannot be very exploratory when we do not have enough samples to cover the large variation present in human samples. Advocating strictly targeted analysis is a rather new element within chemometrics.

Project management

All projects are of course different as dictated by their scope and combination of involved people. Nevertheless, as a chemometrician, I have too often experienced not to be involved in an omics project until the project initiator needed someone to analyze his data. It is far from optimal not to be involved in the design phase and in the pilot study phase, where the chemometrician's input may guide the process.

As illustrated in Chapter 4, an omics experiment usually involves partners from several scientific areas, generally settled in different organizational structures and representing different scientific cultures. Such collaborations are complex to manage, but in my point of view, it is necessary to put a higher emphasis on this aspect in future omics projects.

One person (possibly a partner in the project) should be responsible for the coordination and make sure the relevant competencies would gather at given points in the process. This person should take care of communication and in general be responsible for the overview of the project. A major challenge is that

usually none of the partners in an omics experiment is in a position to prioritize the tasks for the other partners.

Figure 9.1 illustrates how the main actors in the omics experiment work closely together during the experimental process of design, data collection and data analysis. The project manager in the center should ensure that the relevant parties communicate when appropriate.

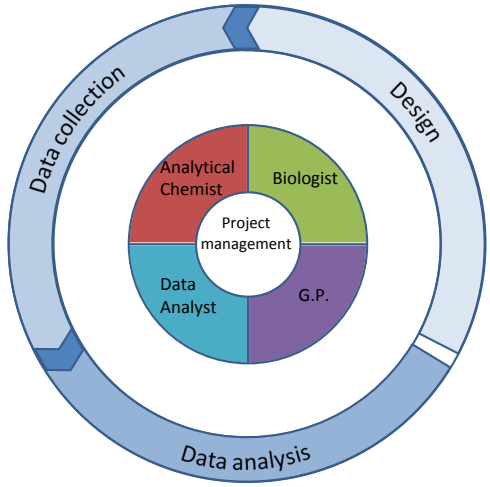


Figure 9.1: The omics experiment - processes and people

Bibliography

[1] L. E. Agelet and C. R. Hurburgh, Jr. A tutorial on near infrared spectroscopy and its calibration. *Crit. Rev. Anal. Chem.*, 40(4):246–260, 2010.

[2] C. M. Andersen and R. Bro. Variable selection in regression-a tutorial. *Journal of Chemometrics*, 24(11-12):728–737, Nov. 2010.

[3] E. Anderssen, K. Dyrstad, F. Westad, and H. Martens. Reducing over-optimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst.*, 84(1-2):69–74, 2006.

[4] O. Beckonert, H. Keun, T. Ebbels, J. Bundy, E. Holmes, J. Lindon, and J. Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*, 2(11):2692–2703, 2007.

[5] R. Bellman. Adaptive control processes: a guided tour. *Princeton University Press*, 1:2, 1961.

[6] F. Betsou, R. Barnes, T. Burke, D. Coppola, Y. DeSouza, J. Eliason, B. Glazer, D. Horsfall, C. Kleeberger, S. Lehmann, A. Prasad, A. Skubitz, S. Somiari,

and E. Gunter. Human biospecimen research: Experimental protocol and quality control tools. *Cancer Epidemiology Biomarkers & Prevention*, 18(4):1017–1025, 2009.

[7] S. Bijlsma, I. Bobeldijk, E. Verheij, R. Ramaker, S. Kochhar, I. Macdonald, B. van Ommen, and A. Smilde. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal. Chem*, 78(2):567–574, 2006.

[8] R. Bro, K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers. Cross-validation of component models: A critical look at current methods. *Anal. Bioanal. Chem.*, 390(5):1241–1251, MAR 2008.

[9] D. Broadhurst and D. Kell. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2:171–196, 2006. 10.1007/s11306-006-0037-z.

[10] H. Büning-Pfaue. Analysis of water in food by near infrared spectroscopy. *Food Chem.*, 82(1):107 – 115, 2003. 2nd International Workshop on Water in Foods.

[11] D. Camacho, A. de la Fuente, and P. Mendes. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53–63, 2005.

[12] S. Clark, L. D. Youngman, A. Palmer, S. Parish, R. Peto, and R. Collins. Stability of plasma analytes after delayed separation of whole blood: implications for epidemiological studies. *Int. J. Epidemiol.*, 32(1):125–130, 2003.

[13] A. Collins. Approaches to the identification of susceptibility genes. *Parasite Immunol.*, 31(5):225–233, 2009.

[14] A. Conesa, R. Bro, F. Garcia-Garcia, J. M. Prats, S. Gotz, K. Kjeldahl, D. Montaner, and J. Dopazo. Direct functional assessment of the composite phenotype through multivariate projection strategies. *Genomics*, 92(6):373–383, DEC 2008.

[15] T. Davies. The history of near infrared spectroscopic analysis: Past, present and future" From sleeping technique to the morning star of spectroscopy". *Analusis*, 26(4):17–19, 1998.

[16] L. Eriksson. *Introduction to multi-and megavariate data analysis using projection methods (PCA & PLS)*. Umetrics AB, 1999.

[17] K. Esbensen and P. Geladi. The start and early history of chemometrics: Selected interviews. Part 2. *J. Chemom.*, 4(6):389–412, 1990.

[18] O. Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48(1):155–171, 2002.

[19] E. Finney and K. Norris. Determination of moisture in corn kernels by near-infrared transmittance measurements. *Trans. ASAE*, 21(3):581, 1978.

[20] P. Geladi and K. Esbensen. The start and early history of chemometrics: Selected interviews. Part 1. *J. Chemom.*, 4(5):337–354, 1990.

[21] P. Geladi and B. Kowalski. Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, 185:1–17, 1986.

[22] L. Gidskehaug, E. Anderssen, and B. Alsberg. Cross model validation and optimisation of bilinear regression models. *Chemom. Intell. Lab. Syst.*, 93(1):1–10, 2008.

[23] R. Goodacre, D. Broadhurst, A. Smilde, B. Kristal, J. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, et al. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3):231–241, 2007.

[24] R. Harshman. "How can I know if it's real?" A catalog of diagnostics for use with three-mode factor analysis and multidimensional scaling. *Research Methods for Multimode Data Analysis. Praeger, New York*, pages 566–591, 1984.

[25] B. L. Heitmann and L. Lissner. Dietary underreporting by obese individuals - is it specific or nonspecific. *Br. Med. J.*, 311(7011):986–989, Oct. 1995.

[26] N. T. Holland, M. T. Smith, B. Eskenazi, and M. Bastaki. Biological sample collection and processing for molecular epidemiological studies. *Mutation Research/Reviews in Mutation Research*, 543(3):217 – 234, 2003.

[27] A. Höskuldsson. PLS regression methods. *J. Chemom.*, 2(3):211–228, 1988.

[28] D. Howery and R. Hirsch. Chemometrics in the chemistry curriculum. *J. Chem. Educ.*, 60(8):656, 1983.

[29] J. P. A. Ioannidis. Limits to forecasting in personalized medicine: An overview. *INTERNATIONAL JOURNAL OF FORECASTING*, 25(4):773–783, OCT-DEC 2009.

[30] J. P. A. Ioannidis. Expectations, validity, and reality in omics. *J. Clin. Epidemiol.*, 63(9):945–949, SEP 2010.

[31] T. Isaksson and T. Næs. Selection of samples for calibration in near-infrared spectroscopy. Part II: Selection based on spectral measurements. *Appl. Spectrosc.*, 44(7):1152–1158, 1990.

[32] R. Kaddurah-Daouk, B. S. Kristal, and R. M. Weinshilboum. Metabolomics: A global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.*, 48(1):653–683, 2008.

[33] A. Karsan, B. Eigl, S. Flibotte, K. Gelmon, P. Switzer, P. Hassell, D. Harrison, J. Law, M. Hayes, M. Stillwell, et al. Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis. *Clinical chemistry*, 51(8):1525, 2005.

[34] K. Kjeldahl, C. M. Andersen, B. Schmidt, K. Kjeldahl, J. Kok, A. de Jong, and R. Bro. A simplified approach for identifying and separating unique and bulk variations in microarray data. will not be published.

[35] K. Kjeldahl and R. Bro. Some common misunderstandings in chemometrics. *J. Chemom.*, 24(7-8):558–564, JUL-AUG 2010.

[36] K. Kjeldahl, M. A. Rasmussen, A. Hasselbalch, K. O. Kyvik, L. Christiansen, S. Rezzi, S. Kochhar, T. I. A. Sørensen, and R. Bro. No genetic footprints of the fat mass and obesity associated (fto) gene in human plasma 1h cpmg nmr metabolic profiles. *Metabolomics*, 2013. submitted.

[37] B. Kowalski and C. Bender. Pattern recognition. Powerful approach to interpreting chemical data. *J. Am. Chem. Soc.*, 94(16):5632–5639, 1972.

[38] B. Kowalski, T. Schatzki, and F. Stross. Classification of archaeological artifacts by applying pattern recognition to trace element data. *Anal. Chem.*, 44(13):2176–2180, 1972.

[39] M. Lauridsen, S. H. Hansen, J. W. Jaroszewski, and C. Cornett. Human urine as test material in 1h nmr-based metabonomics: Recommendations for sample preparation and storage. *Analytical Chemistry*, 79(3):1181–1186, 2007. PMID: 17263352.

[40] J. Lay Jr, S. Borgmann, R. Liyanage, and C. Wilkins. Problems with the "omicss". *Trends Anal Chem*, 25:1046–1056, 2006.

[41] R. Madsen, T. Lundstedt, and J. Trygg. Chemometrics in metabolomics-a review in human disease diagnosis. *Anal. Chim. Acta*, 659(1-2):23–33, FEB 5 2010.

[42] A. D. Maher, S. F. M. Zirah, E. Holmes, and J. K. Nicholson. Experimental and analytical variation in human urine in 1h nmr spectroscopy-based metabolic phenotyping studies. *Analytical Chemistry*, 79(14):5204–5211, 2007.

[43] D. Massart, C. Janssens, L. Kaufman, and R. Smits. Application of the theory of graphs to the optimization of chromatographic separation schemes for multicomponent samples. *Anal. Chem.*, 44(14):2390–2393, 1972.

[44] D. McLerran, W. Grizzle, Z. Feng, W. Bigbee, L. Banez, L. Cazares, D. Chan, J. Diaz, E. Izbicka, J. Kagan, et al. Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias. *Clinical chemistry*, 54(1):44, 2008.

[45] B. M. Nielsen, M. M. Nielsen, S. Toubro, O. Pedersen, A. Astrup, T. I. A. Sørensen, T. Jess, and B. L. Heitmann. Past and current body size affect validity of reported energy intake among middle-aged danish men. *J. Nutr.*, 139(12):2337–2343, Dec. 2009.

[46] L. Nørgaard and R. Bro. Pls regression in the food industry. a study of n-pls regression and variable selection for improving prediction errors and interpretation. In M. Tenenhaus and A. Morineau, editors, *Les Methodes PLS. Symposium International PLS'99*, volume 99, pages 187–202, Cisia-Ceresta, France, 1999.

[47] Z. Pan and D. Raftery. Comparing and combining nmr spectroscopy and mass spectrometry in metabolomics. *Analytical and Bioanalytical Chemistry*, 387(2):525–527, 2007.

[48] E. Peré-Trepat, A. B. Ross, F.-P. Martin, S. Rezzi, S. Kochhar, A. L. Hasselbalch, K. O. Kyvik, and T. I. Sørensen. Chemometric strategies to assess metabonomic imprinting of food habits in epidemiological studies. *Chemom. Intell. Lab. Syst.*, 104(1):95 – 100, 2010. OMICS.

[49] A. J. Rai, C. A. Gelfand, B. C. Haywood, D. J. Warunek, J. Z. Yi, M. D. Schuchard, R. J. Mehigh, S. L. Cockrill, G. B. I. Scott, H. Tammen, P. Schulz-Knappe, D. W. Speicher, F. Vitzthum, B. B. Haab, G. Siest, and D. W. Chan. Hupo plasma proteome project specimen collection and handling: Towards the standardization of parameters for plasma proteome samples. *Proteomics*, 5(13):3262–3277, Aug. 2005.

[50] D. F. Ransohoff. Lessons from controversy: Ovarian cancer screening and serum proteomics. *Journal of the National Cancer Institute*, 97(4):315–319, 2005.

[51] F. Savorani, G. Tomasi, and S. Engelsen. *i*coshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[52] K. Schousboe, P. Visscher, B. Erbas, K. Kyvik, J. Hopper, J. Henriksen, B. Heitmann, and T. Sørensen. Twin study of genetic and environmental influences on adult body size, shape, and composition. *Int. J. Obes.*, 28(1):39–48, 2003.

[53] R. Steuer. Review: On the analysis and interpretation of correlations in metabolomic data. *Briefings in Bioinformatics*, 7(2):151–158, 2006.

[54] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019, 2003.

[55] R. Tauler. 'omics' special issue for chemometrics and intelligent laboratory systems preface. *Chemom. Intell. Lab. Syst.*, 104(1, Sp. Iss. SI):1, NOV 15 2010.

[56] C. Taylor, N. Paton, K. Lilley, P. Binz, R. Julian, A. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. Deutsch, et al. The minimum information about a proteomics experiment (MIAPE). *Nature biotechnology*, 25(8):887–893, 2007.

[57] O. Teahan, S. Gamble, E. Holmes, J. Waxman, J. K. Nicholson, C. Bevan, and H. C. Keun. Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Analytical Chemistry*, 78(13):4307–4318, 2006.

[58] G. Tomasi, F. van den Berg, and C. Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, 18(5):231–241, 2004.

[59] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *J. Chemom.*, 16(3):119–128, 2002.

[60] M. Tuck, D. Chan, D. Chia, A. Godwin, W. Grizzle, K. Krueger, W. Rom, M. Sanda, L. Sorbara, S. Stass, et al. Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J. Proteome Res.*, 8(1):113–117, 2008.

[61] R. van den Berg, H. Hoefsloot, J. Westerhuis, A. Smilde, and M. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(1):142, 2006.

[62] M. Viant, C. Ludwig, and U. Gunther. 1D and 2D NMR Spectroscopy: From Metabolic Fingerprinting to Profiling. *Metabolomics, Metabonomics Metab. Profiling*, page 44, 2008.

[63] M. West-Nielsen, E. Høgdall, E. Marchiori, C. Høgdall, C. Schou, and N. Heegaard. Sample handling for mass spectrometric proteomic investigations of human sera. *Anal. Chem.*, 77(16):5114–5123, 2005.

[64] M. West-Nørager, R. Bro, F. Marini, E. Høgdall, C. Høgdall, L. Nedergaard, and N. Heegaard. Feasibility of Serodiagnosis of Ovarian Cancer by Mass Spectrometry. *Anal. Chem.*, 81(5):1907–1913, 2009.

[65] M. West-Nørager, C. Kelstrup, C. Schou, E. Hogdall, C. Hogdall, and N. Heegaard. Unravelling in vitro variables of major importance for the outcome of mass spectrometry-based serum proteomics. *J. Chromatogr. B*, 847(1):30–37, 2007.

[66] J. Westerhuis, H. Hoefsloot, S. Smit, D. Vis, A. Smilde, E. van Velzen, J. van Duijnhoven, and F. van Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, 2008.

[67] S. Wold. Spline-funktioner - ett nytt verktyg i data-analysen. *Kem. Tidskr.*, 3:34–37, 1972.

[68] S. Wold. Chemometrics, why, what and where to next. *J. Pharm. Biomed. Anal.*, 9(8):589–596, 1991.

[69] S. Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemom. Intell. Lab. Syst.*, 30(1):109–115, NOV 1995.

[70] S. Wold, A. Berglund, and N. Kettaneh. New and old trends in chemometrics. how to deal with the increasing data volumes in r&d&p (research,

development and production) - with examples from pharmaceutical research and process modeling. *J. Chemom.*, 16(8-10, Sp. Iss. SI):377–386, AUG-OCT 2002.

[71] S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the pls method. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, volume 973 of *Lecture Notes in Mathematics*, pages 286–293. Springer Berlin / Heidelberg, 1983. 10.1007/BFb0062108.

[72] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, 58(2):109 – 130, 2001.

PAPER I

Anal Bioanal Chem (2008) 390:1241–1251
DOI 10.1007/s00216-007-1790-1

REVIEW

Cross-validation of component models: A critical look at current methods

R. Bro · K. Kjeldahl · A. K. Smilde · H. A. L. Kiers

Received: 24 September 2007 / Revised: 28 November 2007 / Accepted: 4 December 2007 / Published online: 24 January 2008

Abstract In regression, cross-validation is an effective and popular approach that is used to decide, for example, the number of underlying features, and to estimate the average prediction error. The basic principle of cross-validation is to leave out part of the data, build a model, and then predict the left-out samples. While such an approach can also be envisioned for component models such as principal component analysis (PCA), most current implementations do not comply with the essential requirement that the predictions should be independent of the entity being predicted. Further, these methods have not been properly reviewed in the literature. In this paper, we review the most commonly used generic PCA cross-validation schemes and assess how well they work in various scenarios.

Keywords Overfitting · PRESS · Cross-validation · PCA · Rank estimation

R. Bro (✉) · K. Kjeldahl
Chemometrics Group, Faculty of Life Sciences,
University of Copenhagen,
1958 Frederiksberg C, Denmark
e-mail: rb@life.ku.dk

A. K. Smilde
Biosystems Data Analysis (BDA),
Swammerdam Institute for Life Sciences,
Nieuwe Achtergracht 166,
1018 WV Amsterdam, The Netherlands

H. A. L. Kiers
Heymans Institute (DPMG), University of Groningen,
Grote Kruisstraat 2/1,
9712 TS Groningen, The Netherlands

Introduction

Cross-validation is a standard resampling technique used in many chemometric applications. Results from cross-validation often simplify the selection of meta-parameters, such as the number of components, and also provide a more realistic basis for residual and influence analysis. However, most of the cross-validation techniques currently used in component analysis have some built-in yet seldom mentioned drawbacks, which can hamper the interpretation of the results.

The concept of cross-validation was initially proposed by Mosier [1] as a "design" for assessing the effectiveness of model weights, and it has been explored mainly by Stone [2], Geisser [3] and Allen [4]. In 1976 [5] and 1978 [6] Wold laid the foundations for principal component analysis (PCA) cross-validation, a method used to identify the dimensions that best describe the systematic variations in data. Eastment and Krzanowski [7] and Osten [8] later developed an alternative method. Recently Louwerse and co-workers [9] developed several cross-validation procedures, based on the work of Eastment and Krzanowski, for multiway models such as PARAFAC and Tucker.

In this paper a critical review of current cross-validation techniques for component models is provided, focusing on PCA. Some of the methods described have not been described previously in the literature but have been implemented in commercial software. This paper provides a detailed description of the different implementations of cross-validation. It is important, however, to note that many software companies have included various ad hoc rules for simplifying the decision about the number of components to suggest. The implementations discussed in this paper may differ in these rules. Therefore, it is not a review of

specific software implementations, but rather a review of generic types of cross-validation schemes.

The purpose of cross-validation, as described here, is to find a suitable number of components for a PCA model. Suitable implies that the model describes the systematic variation in the data and preferably not the noise. The noise can be loosely defined as any specific variation in a measurement which is not correlated with any other variation in the measured data. Thus, the aim is to find the number of components for which adding more components does not provide a better (in an overall least squares sense) description of the data not previously included. In this paper, as well as in the actual implementations of cross-validation, it will be assumed that the residuals are not correlated across samples or variables.

Theory

In this section six cross-validation methods are described. Four of these are commonly used in software, while two additional ones are new proposals aimed at overcoming some of the potential problems with the currently used techniques. Some of the current methods have been described in the literature previously, but none of them have been compared to each other in detail. In the following, no mention will be made of preprocessing, but it is assumed that an appropriate preprocessing process is applied before the analysis. In fact, preprocessing should ideally be incorporated into the cross-validation scheme, but to simplify the descriptions we have deliberately chosen to look at models where no preprocessing is required. The results presented will generalize to include preprocessing.

Row-wise cross-validation

This approach illustrates a cross-validation scheme similar to the one used in the UNSCRAMBLER software (CAMO; [10, 11]). A PCA model is sought for a data matrix $\mathbf{X}$, and in this approach each individual segment, consisting of a defined number of whole samples, is left out in turn. In this investigation, the segment will be restricted to one object (one row of $\mathbf{X}$); hence this is termed "leave-one-out cross-validation." For a maximum number of components F, the following procedure is applied:

For number of factors, $f=1,...,F$

(1) for left-out sample(s) (or rows when more than one row is left out) $i=1,...,I$

(a) Split $\mathbf{X}$ ($I\times J$) into $\mathbf{X}^{(-i)}$ and $(\mathbf{x}^i)^T$, where $\mathbf{X}^{(-i)}$ holds all rows except the ith, and $(\mathbf{x}^i)^T$ is a row vector containing only the ith row.

(b) Fit a PCA model to $\mathbf{X}^{(-i)}$ by solving

$$\min \left\| \mathbf{X}^{(-i)} - \mathbf{T}^{(-i)}\mathbf{P}^{(-i)T} \right\|_f^2 \quad (1)$$

where $\|\bullet\|_f^2$ defines the squared Frobenius norm, $\mathbf{P}^{(-i)T}\mathbf{P}^{(-i)} = \mathbf{I}$, and $\mathbf{T}^{(-i)}, \mathbf{P}^{(-i)}$ are of dimension $(I-1) \times f$ and $J\times f$, respectively, where J is the number of variables.

(c) Project $\mathbf{x}^i$ onto the loadings and find the scores of the left-out sample as

$$\left(\mathbf{t}^i\right)^T = \left(\mathbf{x}^i\right)^T \mathbf{P}^{(-i)} \quad (2)$$

(d) Determine the residual variation of $\mathbf{x}^i$ as

$$\begin{aligned} \left(\mathbf{e}^i\right)^T &= \left(\mathbf{x}^i\right)^T - \left(\hat{\mathbf{x}}^i\right)^T \\ &= \left(\mathbf{x}^i\right)^T - \left(\mathbf{t}^i\right)^T \mathbf{P}^{(-i)T} \\ &= \left(\mathbf{x}^i\right)^T - \left(\mathbf{x}^i\right)^T \mathbf{P}^{(-i)}\mathbf{P}^{(-i)T} \end{aligned} \quad (3)$$

(2) Collect all residuals in $\mathbf{E}$ ($I\times J\times F$); one $I\times J$ matrix for each number of components

(3) Calculate mean residual validation variance and correct for the degrees of freedom, resulting in the mean predicted residual sum of squares (MPRESS):

$$MPRESS(f) = \frac{1}{I(J-f)} \sum^{I} \sum^{J} e_{ijf}^2 \quad (4)$$

Characteristics of row-wise cross-validation

Equations 2 and 3 show that the left-out data $\mathbf{x}^i$ are used to find the model of $\mathbf{x}^i$. Consequently, the residuals from the model of $\mathbf{x}^i$ are not independent of $\mathbf{x}^i$, which will result in overfitting; i.e., the more components there are, the smaller the residuals. This is not appropriate because the whole idea of cross-validation is to avoid overfitting by estimating the model independently from the data to be modeled. The numerator in Eq. 4 is meant to correct for this overfitting, but the underlying assumptions and hence validity of this correction are not clear.

Another characteristic of the row-wise cross-validation is that it enables the assessment of different types of sample-specific errors. For example, by leaving out only one replicate in each segment the repeatability can be assessed, and by leaving out all replicates or possibly samples of, say, a given day of analysis, the reproducibility can be assessed.

Cross-validation of Wold

The cross-validation method for PCA proposed by Wold [6] relies on the special property of the NIPALS [12] algorithm to cope with a moderate amount of randomly missing data. In this cross-validation method a selected sequence of

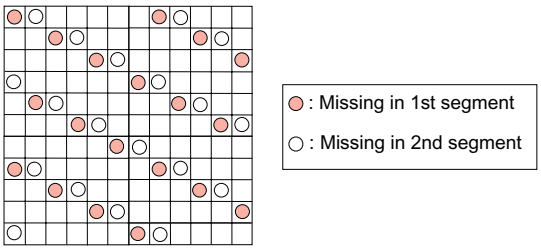


Fig. 1 The pattern of missing values used in Wold cross-validation for K=7

individual elements x_{ij} are left out in a diagonal pattern and are regarded as "missing values." A model is fitted to the remaining data, and the fit of the model to the left-out elements is calculated. The user has to choose the number of segments, K. For the kth (k=1,...,K) segment, the element numbers k, k+K, k+2K, etc., of the complete data set are set to missing. The counting of elements is performed row-wise (see Fig. 1). In this way, upon completing K segments, all elements will have been left out once. The default number of segments in commercial software is often seven, and this number is also used here. This cross-validation approach used to be implemented in the SIMCA software, but a different proprietary method is currently used.

In this method one component is validated at a time. If the first component is judged to be valid it is subtracted from the full data set, and only the residuals are then used to test whether the next component is valid.

The cross-validation method can formally be described as follows. For a data matrix $\mathbf{X}$ (I×J) a PCA model is sought. For the component, f, to be validated, the following procedure is applied. For factor f

(1) Let $\mathbf{X}(f)$ be the residuals after f–1 components, i.e., initially $\mathbf{X}(1)=\mathbf{X}$ and subsequently $\mathbf{X}(f)$ is the residual matrix $\mathbf{E}(f-1)$ found in step 6.
(2) Calculate the sum of squared elements of $\mathbf{X}(f)$:

$$SS_{\mathbf{x}(f)} = \sum^{I}\sum^{J} x_{ij}^2(f) \tag{5}$$

(3) For the left-out segment k=1,...,K
   (A) Split data $\mathbf{X}(f)$ into $\mathbf{X}^{(-k)}(f)$ and $\mathbf{X}^{(k)}(f)$, where $\mathbf{X}^{(-k)}(f)$ holds all observations except the elements in the kth segment and $\mathbf{X}^{(k)}(f)$ holds only those.
   (B) Estimate the next principal component ($\boldsymbol{t}^{(-k)}$, $\boldsymbol{p}^{(-k)}$) by fitting to $\mathbf{X}^{(-k)}(f)$ with the NIPALS algorithm.
   (C) Calculate the model of $\mathbf{X}^{(k)}(f)$ using

$$\hat{x}_{ij}(f) = t_i^{(-k)} p_j^{(-k)} \tag{6}$$

(4) Find the total *PRESS*,

$$PRESS(f) = \sum^{I}\sum^{J} e_{ij}^2(f) = \sum^{I}\sum^{J} \left(x_{ij}(f) - \hat{x}_{ij}(f)\right)^2 \tag{7}$$

(5) Form the ratio R

$$R(f) = \frac{PRESS(f)}{SS_{\mathbf{x}(f)}} \tag{8}$$

R<1 indicates that the predictions are improved with the inclusion of the last component (f). If R>1, the component did not improve the prediction errors. Hence, the appropriate number of components is f–1. Note that instead of using this specific rule, it is also possible to use alternatives, such as looking for the minimal *PRESS*, etc.
(6) Fit a PCA model ($\mathbf{t}$, $\mathbf{p}$) to the complete data set $\mathbf{X}(f)$ with one component. Determine the residual variation as $\mathbf{E}(f)=\mathbf{X}(f)-\mathbf{t}\mathbf{p}^T$
(7) Increase f by 1 and go back to step (1).

Characteristics of Wold's cross-validation

Wold's cross-validation scheme provides both a way to calculate *PRESS* by a specific leave-out pattern and a criterion for the selection of the number of components.

From step (3) it can be seen that the cross-validation of component f is based on a model where the first f–1 components are estimated from the complete data set and only component f is estimated on a segmented basis. This means that, when leaving out elements, the estimates of these left-out elements $\hat{x}_{ij}$ depend on the f–1 component PCA model, where x_{ij} was included. Hence, the left-out elements and their predicted values are not independent and overfitting can be expected.

Cross-validation of Eastment and Krzanowski (EK)

In 1982, Eastment and Krzanowski [7] suggested an approach that could be used to choose a feasible number of components in PCA. The incentive was to create a cross-validation approach that ensured that each data point was not used at both the prediction and the assessment stages, thus avoiding problems with overfitting. The approach intends to use as much of the original data as possible in order to predict a left-out element for each possible choice of factors f=1,...,F (i=1,...,I and j=1,...,J).

In order to ensure that the residuals are independent of the data element being predicted, the following scheme is adopted. For each combination of i (1,...,I) and j (1,...,J), one PCA model is fitted to $\mathbf{X}^{(-i)}$ and another PCA model is

fitted to $\mathbf{X}^{(-j)}$; that is, they are fitted to the data with row i left out and to the data with column j left out. The PCA models are represented as singular value decompositions:

$$\mathbf{x}^{(-i)} = \mathbf{U}^{(-i)}\mathbf{S}^{(-i)}\mathbf{V}^{(-i)T} \tag{9}$$

and

$$\mathbf{x}^{(-j)} = \mathbf{U}^{(-j)}\mathbf{S}^{(-j)}\mathbf{V}^{(-j)T} \tag{10}$$

In Fig. 2, a graphical representation of the data for the PCA model of $\mathbf{X}^{(-i)}$ is provided, where the gray area designates the part of the data (row i), not the part of the model. Similarly, the model of $\mathbf{X}^{(-j)}$ is created by excluding the elements in column j.

It is apparent from Fig. 2 that the parameters in $\mathbf{V}^{(-i)}$ and $\mathbf{S}^{(-i)}$ are not influenced by sample i. Likewise $\mathbf{U}^{(-j)}$ is not influenced by the jth column of $\mathbf{X}$. From the model in the figure, the "loadings" in $\mathbf{V}^{(-i)}$ do not provide the means to predict sample i. However, the scores in $\mathbf{U}^{(-j)}$ which are independent of variable j also contain scores for the ith sample. Hence, by combining the two models, an estimate of element x_{ij} can be obtained which is independent of the value in x_{ij}. The crucial aspect of this approach is deciding how to combine the two models into one estimate. Eastment and Krzanowski [7] suggest using the following combination of the two models

$$\hat{x}_{ij}(f) = \sum^{f} u_i^{(-1)}(f)\sqrt{s^{(-j)}(f)}\sqrt{s^{(-i)}(f)}v_j^{(-i)}(f) \tag{11}$$

The square roots of the two sets of singular values are used to accommodate for the possible differences in their magnitudes. Some attention is required to avoid problems with sign indeterminacies that can occur when matching two different models. This sign incompatibility is mended by what Eastment and Krzanowski called a "parity check," where the products of component scores and loadings, per component, are given the same sign as the corresponding product of the component scores and loadings found for the full data.

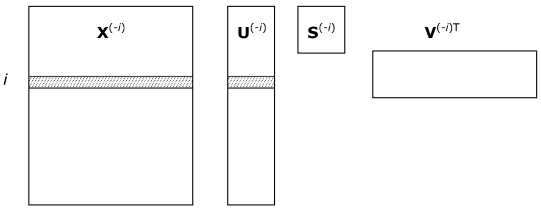


Fig. 2 Eastment–Krzanowski (EK) cross-validation. Data for one of two submodels is obtained by leaving out element x_{ij}. Likewise, the jth column of $\mathbf{X}$ is left out, whereby $\mathbf{X}^{(-j)}$, $\mathbf{U}^{(-j)}$, $\mathbf{S}^{(-j)}$ and $\mathbf{V}^{(-j)}$ are obtained

The *MPRESS*(f) expresses the mean difference between the actual and the predicted value:

$$MPRESS(f) = \frac{1}{IJ}\sum^{I}\sum^{J}\left(\hat{x}_{ij}(f) - x_{ij}\right)^2 \tag{12}$$

In order to determine the number of components, Eastment and Krzanowski introduced a diagnostic called W, which compares the gain in *MPRESS* between two levels of model complexity:

$$W(f) = \frac{MPRESS(f-1) - MPRESS(f)}{D_{fit}(f)} \div \frac{MPRESS(f)}{D_r(f)} \tag{13}$$

where $D_{fit}(f) = I + J - 2f$ is the number of degrees of freedom lost in fitting the fth component and $D_r(f)$ is the number of degrees of freedom remaining after fitting f components. Prior to modeling, and with no centering, $D_r(f)=IJ$, so after fitting f factors $D_r(f) = IJ - \sum_1^f (I + J - 2f)$. Originally [7], Eastment and Krzanowski suggested that the factor f should be included as long as $W(f)>1$; this threshold was later modified to a value of 0.9 by Krzanowski [13].

Characteristics of the Eastment and Krzanowski cross-validation

The rationale behind the Eastment and Krzanowski (EK) approach is that the two PCA models can be combined. The authors [7] argue that this is possible because the PCA model is unique. However, the conclusion that the models are comparable due to this uniqueness is based on a misunderstanding. It is true that the PCA model is unique and hence the same solution is obtained every time the model parameters are estimated from one particular data set. However, if the parameters are found from different subsets, then different parameters are generally found.

As an example, consider a situation where there are two components in a data set. In one set of samples, it is mainly one phenomenon that is present; hence a PCA model will primarily reflect this phenomenon in its first loading. Another subset primarily reflects the second phenomenon, and hence a PCA model would primarily reflect the second phenomenon in its first component. Therefore, even in the absolutely noise-free case, the uniqueness of the PCA model does not imply anything at all with respect to whether the same components of different models can be compared. The components in a bilinear model have rotational freedom. Uniqueness is obtained in PCA by adding restrictions (orthogonality, maximum variance per component). These additional constraints will affect PCA models of different subsets differently and will not yield, for example, the same first

loading vector when models are fitted to comparable but different subsets of data. As expected, and as seen in the simulation study below, this problem is most pronounced for small amounts of data.

An additional, possibly minor, problem relates to the parity check.The parity check is simple, but it is a source of overfitting because it implies that the sign of the prediction is chosen using information from the data element that was excluded.

Louwerse et al. [14] note that EK's estimates of x_{ij}, $\widehat{x}_{ij}$, are biased because the matrices $\mathbf{S}^{(-i)}$ and $\mathbf{S}^{(-j)}$ systematically underestimate $\mathbf{S}$. On average, this bias is eliminated by correcting $\mathbf{S}^{(-i)}$ by a factor of $\sqrt{I/(I-1)}$ and $\mathbf{S}^{(-j)}$ by a factor of $\sqrt{J/(J-1)}$ [19].

Cross-validation by Eigenvector

This approach illustrates a cross-validation scheme similar to the one used in the PLS_Toolbox software (Eigenvector) [15]. In this approach, PCA models are calculated with one or several samples left out and then the model is used to predict estimates of the left-out samples [16]. Assume, without loss of generality, that leave-one-out cross-validation is used. A PCA model is determined from the $I-1$ remaining samples. For the left out sample, each variable is predicted independently through the following procedure. First a score value for the left-out sample is estimated in a least squares sense using the $J-1$ remaining variables of that sample and the PCA model where the jth row of the loading matrix has been excluded. This score value combined with the PCA loading matrix with all variable loadings included gives an estimate of the kept-out element x_{ij}. In essence, this is a missing data problem where the missing variable is predicted from the model and the sample observation excluding the one variable [16].

(1) Leave out one or several samples and calculate a PCA model (**T**, **P**) on the remainder.
(2) For each factor $f=1,...,F$,
for the left-out sample(s) $i=1,...,I$, then
(a) for the left-out variable(s) $j=1,...,J$:

(i) Estimate the score as

$$\mathbf{t}^{(-j)T} = \mathbf{x}_i^{(-j)T}\mathbf{P}^{(-j)}\left(\mathbf{P}^{(-j)T}\mathbf{P}^{(-j)}\right)^{-1},$$

where $\mathbf{P}^{(-j)}$ is the loading matrix $\mathbf{P}$ found in step 1 with the jth row excluded. $\mathbf{x}_i^{(-j)T}$ is a row vector containing the ith row of $\mathbf{x}$ except for the jth element.

(ii) Estimate the element $x^{(ij)}$ as $\widehat{x}_{ij}(f) = \mathbf{t}^{(-j)}\mathbf{p}_j^T$, where $\mathbf{p}_j$ is the jth row of $\mathbf{p}$

(iii) Find the prediction error of the (i,j)th element $e_{ij}(f) = x_{ij} - \widehat{x}_{ij}(f)$

(b) Estimate

$$PRESS(f) = \sum_{I}\sum_{J}\left(e_{ij}(f)\right)^2 \tag{14}$$

Characteristics of Eigenvector's cross-validation

In contrast to the other methods, the *PRESS* values estimated with Eigenvector's method are actually independent from the predicted elements. Furthermore, the Eigenvector method, like the row-wise method, possesses the advantage that sample-specific error measures such as repeatability and reproducibility can be calculated based on leaving out samples in different patterns. The least squares element of the method requires that variables are correlated, and thus it is particularly suited to "spectral type" data.

EM cross-validations

Some of the cross-validation techniques mentioned so far display two significant problems. Either overfitting is introduced, because the model with which left-out elements are predicted is not independent of the left-out elements, or an unintended additional error is introduced because the rationale behind the method is not correct. Both of these problems can be eliminated though, by properly designing the cross-validation procedure as outlined below. The two suggested methods are called expectation maximization (EM) approaches in the following, because the remedies suggested are special cases of EM [17].

Cross-validation based on an improved Wold procedure (EM-Wold)

A revised cross-validation approach based on the Wold [6] procedure is presented that removes the problem with the original method by estimating all components simultaneously for each segment. Then, however, the NIPALS approach for handling missing values is no longer feasible, as this is only optimal for one-component models. Instead, the full PCA models are calculated for each number of components using imputation, as described by Kiers and Bro [18, 19]. Unlike NIPALS, such an approach can calculate several-component PCA models in a true least squares sense even when data elements are missing. The price paid for this is that the estimation depends on the number of components, and hence solutions must be calculated anew for each number of principal components. The procedure is as follows. For factors $f=1,...,F$

(1) For left-out element $k=1,...,IJ$

(A) Split data into $\mathbf{X}^{(-k)}$ and $x^{(k)}$, where $\mathbf{X}^{(-k)}$ holds all data except the kth element x_k.

(B) Fit a PCA model to $\mathbf{X}^{(-k)}$ by solving

$$\min \left\| \mathbf{X}^{(-k)} - \mathbf{T}\mathbf{P}^T \right\|_f^2 \quad (15)$$

(C) Find the model of the whole data set as $\widehat{\mathbf{X}}(f) = \mathbf{T}\mathbf{P}^T$

(D) Determine the residual of the left-out element as $e_k(f) = x_k - \widehat{x}_k(f)$

(E) Calculate $PRESS(f)$

$$PRESS(f) = \frac{1}{IJ}\sum^{k=IJ} (e_k(f))^2 \quad (16)$$

As can be seen, the prediction of the left-out element is independent of the left-out element, and no additional error is introduced by the estimation method because the PCA model is estimated in a least squares sense (Eq. 15). Algorithms for estimating the model with missing data in a least squares sense are readily available, as described in the literature [18, 19].

Cross-validation based on an improved Eastment and Krzanowski procedure (EM-EK)

It is also possible to correct the procedure of Eastment and Krzanowski [7] to take the rotation problems into account. This correction consists of a simple remedy before matching the scores and loadings. The assumption in the Eastment and Krzanowski [7] approach is that for a given left-out element, x_{ij}, the model of the element is defined by the subspaces given by $\mathbf{U}^{(-j)}$ and $\mathbf{V}^{(-i)}$. However, the scale and rotation is not defined immediately. In order to define those, a model based on $\mathbf{U}^{(-j)}$ and $\mathbf{V}^{(-i)}$ is found to maximize the fit to the data that have not been left out. This is done through an iterative procedure where the model of $\mathbf{X}^{(-ij)}$ is determined from

$$\widehat{\mathbf{X}} = \left(\mathbf{U}^{(-j)}\left(\mathbf{U}^{(-j)}\right)^+\right)\mathbf{X}^{(-ij)}\left(\mathbf{V}^{(-i)}\left(\mathbf{V}^{(-i)}\right)^+\right)^T \quad (17)$$

Hence, we obtain a projection onto the spaces spanned by the given scores and loadings without assuming that the columns of these are pairwise-matched. As the element x_{ij} in $\mathbf{X}^{(-ij)}$ is missing, the above cannot be calculated immediately. Rather, iterative imputation is used to deal with the missing data in the following way. The missing element is initially replaced with its original value. Other values can also be used, but convergence is typically fastest when the original value is used. Note though that the original value does not affect the actual solution; it only provides an initial estimate. Subsequently, the missing element is replaced with the estimate from Eq. 17 and the procedure is repeated until the missing element does not change significantly. Upon convergence, an estimate of the missing element is obtained which is in the same subspace as that dictated by the Eastment and Krzanowski [17] approach, but where the scaling and rotation ambiguities have been resolved. In essence, this is missing data imputation by expectation maximization [17].

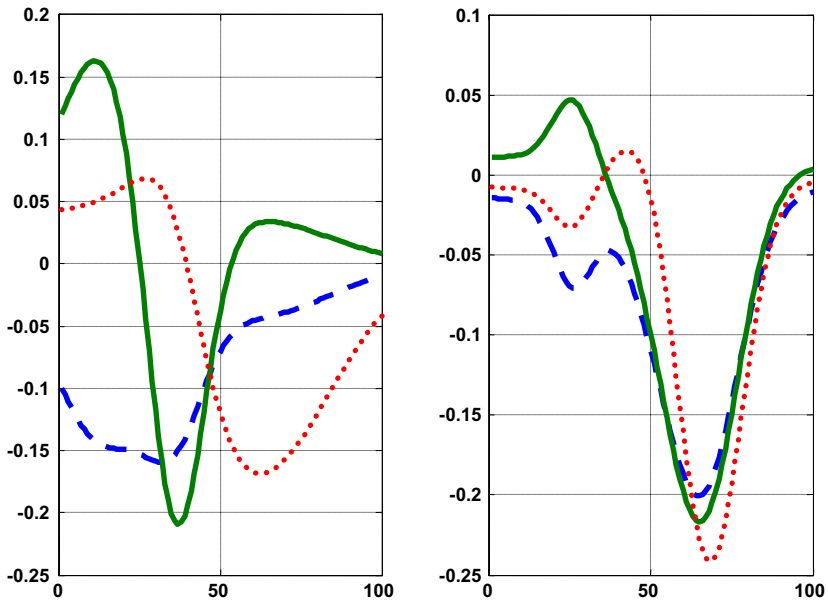


Fig. 3 Low and high correlation between underlying features in two rank-three data sets. *Left*: corr=0.0, *right*: corr=0.9

Selection of the number of components

The series of *f*-dependent *PRESS* values obtained from the cross-validation form the basis for choosing the optimal number of components. The row-wise and Eigenvector cross-validations leave it up to the user to choose the number of components, whereas Wold and EK provide explicit ways to select components through the R and W statistics. Alternatively, one may simply choose the number of components representing either the global minimum *PRESS* value or the first local minimum *PRESS*. The latter approach is often used in practice. In the following we have applied all of these four different criteria to all methods in order to assess the influence of the selection method. It is worth noting that in many real applications of PCA, visual interpretation of the cross-validation results is used as an important guide when selecting the right number of components. However, this is not important in this study, where the focus is on the extent to which cross-validation provides consistent results in general. Hence automated selection of the number of components is the only criterion used.

Simulation and discussion

The capability of the six cross-validation methods described above was examined by performing a simulation study. Datasets of rank one, three or five containing either 10 or 100 variables and either 10, 100 or 300 samples were created and homoscedastic noise was added at a level of either 1 or 10%. For the rank three and five datasets, correlation between underlying loadings was set to either 0, 0.3 or 0.9. Figure 3 illustrates how low and high correlations are reflected in the data structure. All combinations of these parameters sum to 84 different dataset configurations. The rank-one data cannot be set to have different correlations between loading vectors as there is only one loading vector, and so correlations were not varied

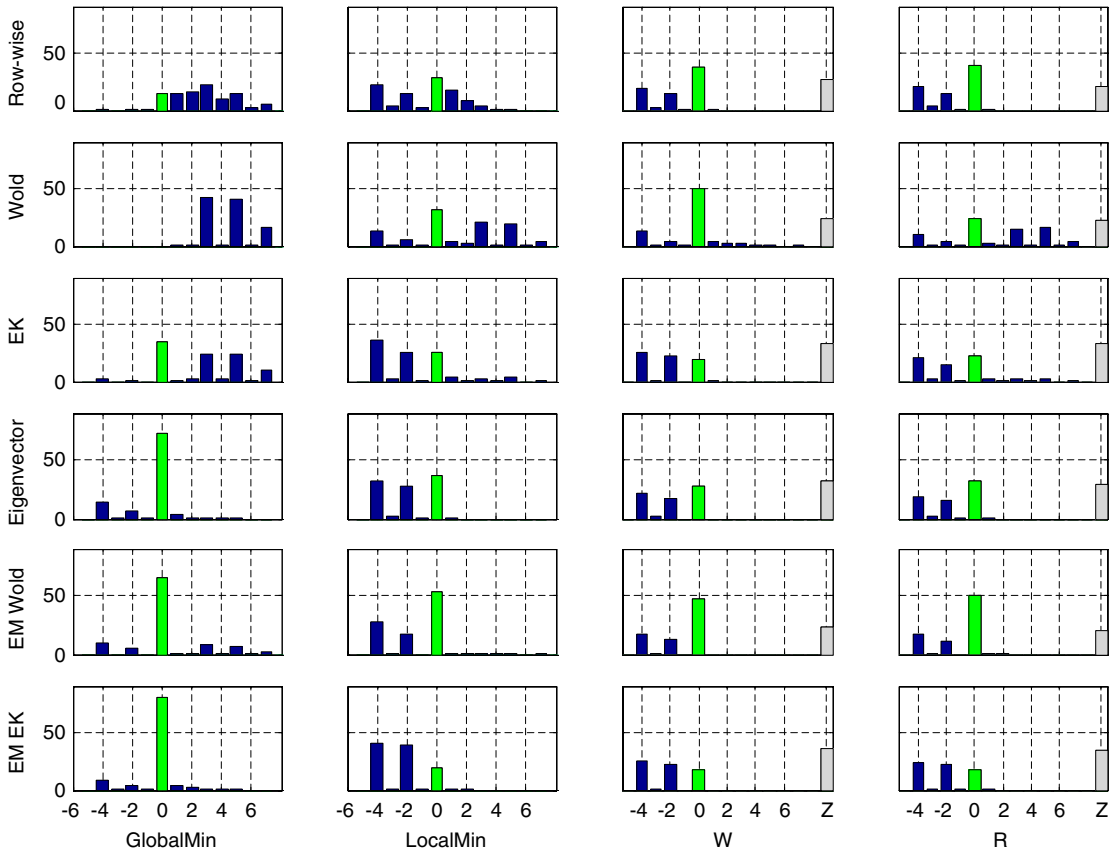


Fig. 4 Distributions of deviations from reference rank for method–criterion combinations. The reported numbers are the percentage of models with a given deviation from the reference rank. Negative values correspond to underfitted and Z= zero components assigned

for rank-one data. Ten different datasets were created for each configuration, and all 840 datasets were subjected to rank estimation by each of the six PCA cross-validation approaches, resulting in a total of 5040 runs.

Since no offset was included in the data, all PCA modeling was performed without the use of mean centering (or any other preprocessing). Eight principal components were calculated for each dataset and *PRESS* or *MPRESS* values calculated as described above. The cross-validation algorithms were implemented according to the descriptions above; Wold's using seven segments, row-wise and Eigenvector using leave-one-out cross-validation. For comparative purposes, success rates were calculated on the basis of a binary (correct/incorrect) comparison of the selected number of components and the known rank of the dataset. The number of components was selected by the following four criteria:

(A) Minimum (M)PRESS
(B) Wold's R criterion (see Eq. 8). If $R(F=1)>1$ the result is zero components.
(C) Krzanowski's W criterion (see Eq. 13). If no W values were found to exceed 0.9 the result is zero components.
(D) The choice of the first local minimum value of (M) PRESS was also investigated, but a clear result was not obtained and so this was kept out of the following.

Result of simulations

Figure 4 evaluates the cross-validation methods and criteria across noise, correlation, and number of samples and variables. It shows that W and R are the criteria best suited to the Wold and Row-wise methods, but they tend to provide conservative rank estimates. The simplest criterion, GlobalMin, is not useful for Wold's method, but it is very well suited to the EK, Eigenvector and EM methods, although it has a tendency to result in overfitting in the EK case. In the following, the row-wise and Wold methods are used in combination with the W criterion, whereas Global-Min is used with the remaining methods.

It was expected that higher noise and higher correlation would present greater challenges to the cross-validatory rank estimation. Figure 5 suggests that this may be valid for the correlation part, whereas—with the noise levels at play here—a high noise level is only critical if correlation is also

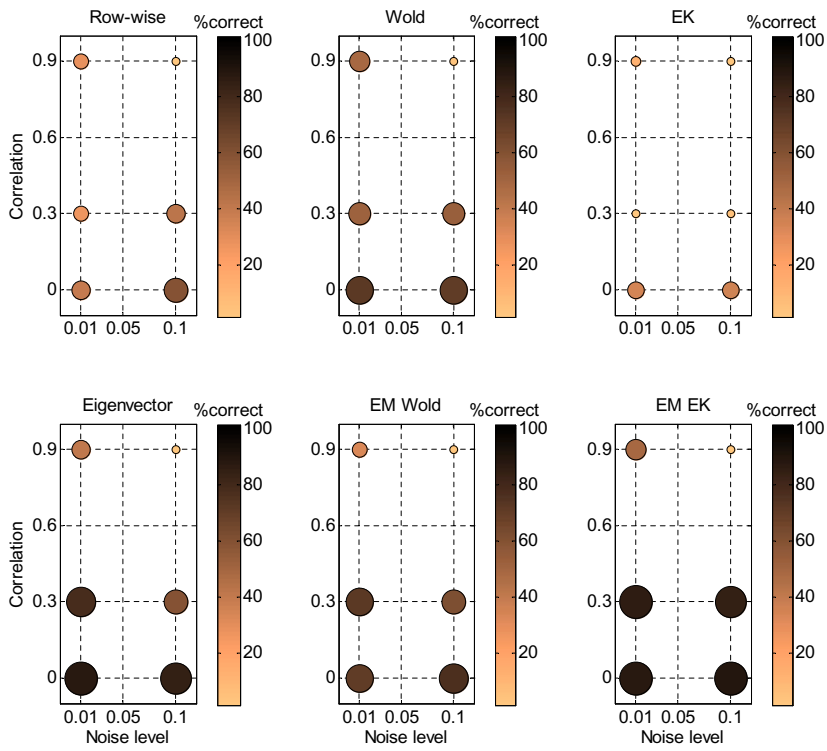


Fig. 5 The influence of noise and correlation level on the success rate in rank estimation. Marker size and color reflect success rate (%), which was determined using the best criteria for each method

Table 1 Overall success rates for each tested cross-validation method using the optimal criterion and excluding extreme noise and correlation combinations

Method	Success rate (%)
Row-wise	38
Wold	49
EK	34
Eigenvector	72
EM Wold	65
EM EK	81

high, in which case all methods fail. In Table 1, Fig. 4 and Fig. 6 the (noise,correlation)=(10%, 0.9) cases have been excluded. Intrinsically, the rank-one configurations did not contain these critical combinations, and all methods performed well for rank-one data, irrespective of the noise level and the number of samples and variables.

According to Fig. 6, a higher number of samples improves the results slightly, whereas the number of variables seems to have a considerably greater impact. The EK method in particular performs poorly for a low number of samples and/or variables, but the row-wise and Wold methods also require many variables and samples to provide reliable results. For an insufficient number of variables (and samples) it is no longer clear which criterion to use. The percentages in Figs. 5 and 6 and in Table 1 are of course based on sample estimates. Ninety-five percent confidence intervals will be at most ±3.7% around the results in Table 1, and ±9.0% around the results in Figs. 5 and 6. Hence our main findings remain if we take into account the ranges of these confidence intervals.

Although also influenced by the matrix size, the EM EK, EM Wold and Eigenvector methods are clearly the best at handling relatively small datasets, and GlobalMin is consistently (across settings) the best criterion to use for these methods. Each cross-validation method obtained an overall success rate as listed in Table 1.

A short comment is given on computational aspects. These were not the focus of this study, and so intensive code optimization was not undertaken. Figure 7 shows how the computer time was spent during the simulations; clearly, the EM methods are very computationally intensive. Inherently calculation time increases rapidly with number of elements for these methods. It should be noted that eight principal components were calculated for all datasets, which disagrees slightly with the original outline of Wold's method.

Real data: UV-VIS

A set of real process data was subjected to cross-validation by the six methods. The data consist of UV-VIS spectra for eight batches of the two-step chemical reaction of 3-chlorophenylhydrazonopropane dinitrile

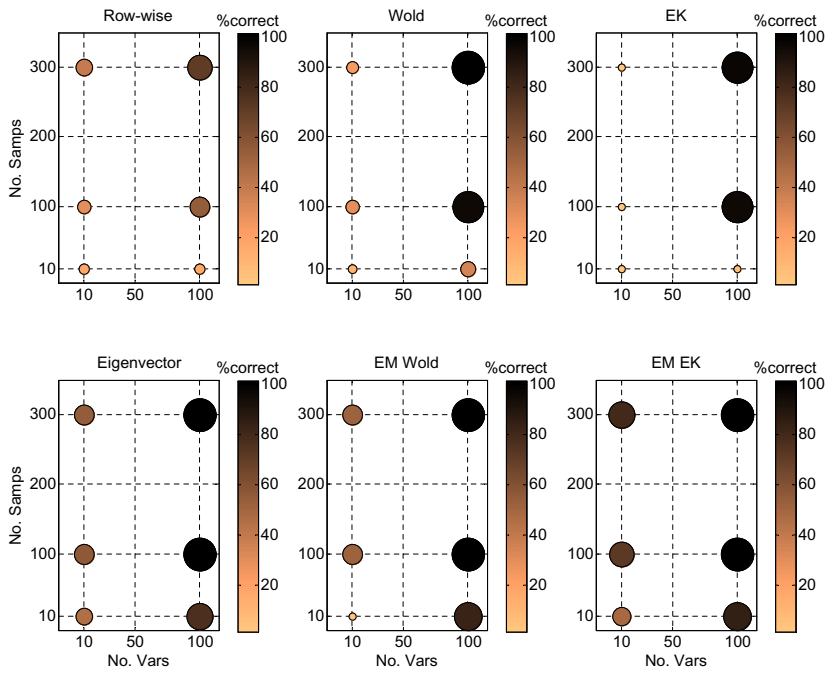


Fig. 6 The influence of the number of variables and samples on the success rate in rank estimation. Marker size and color reflect success rate (%)

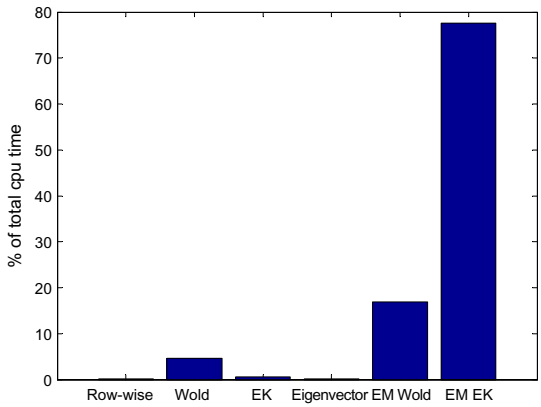


Fig. 7 Computational times for the six methods

with 2-mercaptoethanol to form 3-chlorophenyl-hydrazonocyanoacetamide and the byproduct ethylene sulfide [20]. Prior to analysis, the original dataset was reduced by extracting every ninth object and every second variable, resulting in 31 samples and 201 variables.

Figure 8 shows the PRESS patterns obtained for cross-validations of the batch 4 dataset, and since little variation was found between batches, this is representative of the differences between the methods. All PRESS patterns show a decrease for three principal components, but after that the picture differs. The row-wise method displays a gradual decrease in PRESS as more components are added. The result for Wold's method looks similar, but for most of the batches (five of eight) a very small PRESS increase was detected for four principal components. Wold's method is the only one which had some noticeable batch differences (not shown).

The EK method evaluated by the GlobalMin diagnostic consistently suggests eight components. However, by using the W criterion which was originally associated with the method, we find that three components are suggested for all batches.

Eigenvector's method suggests three principal components in six of eight batches and four in the remainder. Again, a steep decrease in PRESS is observed for three components, but a clear increase in PRESS is also observed following the suggested number of components, making the manual choice of model complexity relatively simple. This quality is even more pronounced for the EM Wold method, which consistently suggests three principal components throughout the eight batches.

The other EM method, EM EK, displays a local PRESS minimum for three components, but the global minimum is observed at six components. This slightly "bumpy" behavior of the PRESS pattern after three components is found to a greater or lesser extent in the results of all eight batches. Nevertheless, three components is suggested for four of the batches, and only the example in Fig. 8 suggests that as many as six components should be included.

The figure illustrates the fact that rank estimation by cross-validation consists of two parts: (A) formation of a PRESS sequence and (B) selection of the optimal model based on (A). The main focus of this study has been (A); however, it has been necessary to automate (B) through the use of the criteria described in this paper in order to summarize the results of the

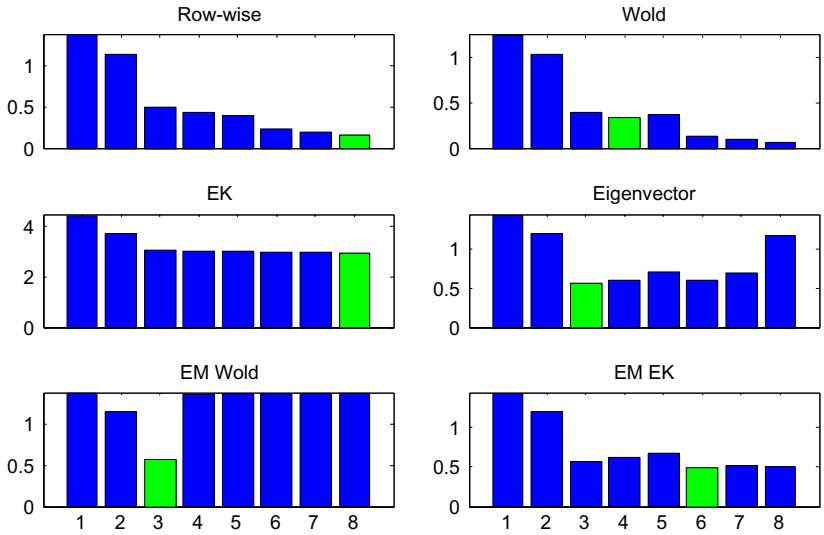


Fig. 8 Real data: PRESS patterns obtained by cross-validation of UV-VIS batch-4 data. The abscissa represents the number of principal components

simulation. Many experienced users would probably choose to select three components based on any (or most) of the bar plots in Fig. 8, but with greater confidence for the EM Wold than for, say, the row-wise pattern. As for the simulations here, it is clear that for the less experienced user or for an automated selection, the choice of method greatly affects the resulting PRESS values.

Conclusion

There are many approaches that can be used to perform cross-validation of PCA models. However, few of them are trivial extensions of the original idea of cross-validation, and therefore the question of whether they provide meaningful results remains to be shown thoroughly in practice. In this paper, the different approaches have been described and their seldom-mentioned deficiencies have been highlighted. When tested on simulated "spectral type" data, theoretical weaknesses and method performance could be linked. In terms of accuracy in rank estimation, the Eigenvector and EM methods in combination with the GlobalMin criterion outperform the other methods under most circumstances, and were particularly superior in cases involving few variables and samples. Considering computational effort too, the Eigenvector method represents a very good choice based on the results of this study.

References

1. Mosier C (1951) Educ Psychol Meas 11:5–11
2. Stone M (1974) J Roy Stat Soc B 36:111–148
3. Geisser S (2000) Biometrika 61:101–107
4. Allen D (1974) Technometrics 16:125–127
5. Wold S (1976) Pattern Recogn 8:127–139
6. Wold S (1978) Technometrics 20:397–405
7. Eastment HT, Krzanowski WJ (1982) Technometrics 24:73–77
8. Osten D (1988) J Chemom 2:39–48
9. Louwerse D, Kiers H, Smilde A (1999) J Chemom 13:491–510
10. Martens H, Martens M (2001) Multivariate analysis of quality: an introduction. Wiley, Chichester, UK
11. Martens H, Næs T (1989) Multivariate calibration. Wiley, Chichester, UK
12. Wold H (1975) Quantitative sociology. In: Blalock H, Aganbegian A, Borodkin F, Boudon R, Capecchi V (eds) International perspectives on mathematical and statistical modeling. Academic Press, New York, pp 307–357
13. Krzanowski WJ (1983) J Stat Comput Simul 18:299–314
14. Louwerse D, Kiers H, Smilde A (1997) Internal Report 8:1–6
15. Wise B, Gallagher N, Bro R, Shaver J (2003) PLS Toolbox 3.0. Manson, WA
16. Wise B, Ricker N (1991) In: Najim K, Dufour E (eds) IFAC Symp on Advanced Control of Chemical Processes, Toulouse, France, 14–16 October 1991, pp 125–130
17. Dempster A, Laird N, Rubin D (1977) J Roy Stat Soc B 39:1–38
18. Bro R (1998) Multi-way analysis in the food industry. Models, algorithms, and applications. Ph.D. Thesis, University of Amsterdam, Amsterdam (see http://www.models.life.ku.dk/research/theses. Accessed 2 Jan 2007)
19. Kiers H (1997) Psychometrika 62:251–266
20. Bijlsma S, Boelens H, Smilde A (2001) Appl Spectrosc 55:77–83

PAPER II

Special Issue Article

CHEMOMETRICS

Received: 11 April 2010, Revised: 17 July 2010, Accepted: 19 July 2010, Published online in Wiley InterScience: 2010

(www.interscience.wiley.com) DOI: 10.1002/cem.1346

Some common misunderstandings in chemometrics

Karin Kjeldahl[a] and Rasmus Bro[a]*

This paper describes a number of issues and tools in practical chemometric data analysis that are often either misunderstood or misused. Deciding what are relevant samples and variables, (mis-)use of common model diagnostics, and interpretational issues are addressed in relation to component models such as PCA and PLS models. Along with simple misunderstandings, the use of chemometric software packages may contribute to the mistakes if not used critically, and it is thus a main conclusion that good data analysis practice requires the analyst to take responsibility and do what is relevant for the given purpose. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: PCA; PLS; misunderstandings; model interpretation; data analysis

1. INTRODUCTION

The field of chemometrics is successfully being applied within many scientific branches. The chemometric tools enable the analysis of complex multivariate data, whereby the extraction of the relevant information is facilitated. However, getting meaningful results requires not only meaningful data but also meaningful analysis and understanding of the purpose of the analysis.

During the process of solving problems with data analysis, a number of statistical numbers and more or less standardized visualizations may assist the analyst in (1) selecting the right experiments and measurements, (2) building the optimal model and (3) interpreting the model. The appropriate usage of these aids is important for the result, but years of extensive teaching experience and data analytical guidance within many research areas show that a number of common misunderstandings exist that may hamper appropriate multivariate data exploration and modeling.

By highlighting some of the important problems, it is the hope to eliminate some erroneous conclusions and improve the performed data analysis in general. This paper focuses on some common problems seen for component models such as those obtained with PCA and PLS regression, and both optimization and interpretation issues are addressed. First, some points about which samples and variables to use are commented, then the usage of model diagnostics for optimization and interpretation is addressed, and eventually issues regarding model interpretation are discussed. In essence, it is all about the use of sound, rational thinking and responsibility instead of 'push-the-button' automated data analysis. Most of what is explained in this paper is far from new, but it seems that it is necessary to reiterate these issues from time to time. A number of papers have been written that highlight similar issues and these are highly recommended for further insight [1–5].

2. USING THE RIGHT SAMPLES AND VARIABLES

The selection of samples and variables to include in the modeling is of major importance for the results and for the applicability of the model for future use. The *purpose* of the modeling should be in strong focus during the phases of experimental design and data modeling, so that the collected samples and variables reflect the relevant variation for the given purpose—not necessarily the samples most easily at hand or the full spectrum simply provided by the instrument.

2.1. Samples

If the aim of an analysis is to build an MSPC (multivariate statistical process control) model that should detect abnormal samples in a production setting, the best sensitivity is obtained when the samples used for building the model are as close to normal as possible. Typically, this is what could be called a one-class classification model, used for detecting whether a new sample is similar to the defined normal class [6]. This is also relevant in settings similar to MSPC such as diagnostic models or models meant for monitoring adulteration.

Samples from Design of Experiments (DoE) represent the extremes of the variation and should not be used for that type of application. Rephrased using an oversimplified example: if a model is meant to detect whether a new sample represents normal-sized people, the model should be made from normal-sized people, not the extremes.

Remarks:

- In an ideal case, the normal samples in an MSPC model would follow the same underlying variation and hence be efficiently modeled with, e.g. a PCA model if the samples are approximately following a multi-normal distribution [6].
- A well-controlled process exhibits purely random variation in its quality parameters because the process control is handling and removing the influence of all systematic effects. If this is not the case, it is implicitly an indication that there is still a job

* *Correspondence to: R. Bro, Department of Food Science, University of Copenhagen, Denmark.*
E-mail: rb@life.ku.dk

a *K. Kjeldahl, R. Bro*
Department of Food Science, University of Copenhagen, Denmark

for the engineers in controlling the process. This also means that ideally, an MSPC model will describe a very low percentage of the variation and be difficult to validate in a traditional chemometric way. This is not necessarily the case, when the MSPC approach is used, e.g. in medical diagnostics where individual systematic differences are usually significant.

DoE samples can be strong players for building a good regression model when focus is on accurate predictions. It is here important that the design reflects the variation that is expected in the future and within the range where good predictions are important for the application. That is, there is no need to build a model that predicts alcohol content with training samples between 0 and 100% if only the range between 3 and 10% is important. On the other hand, also keep in mind that extrapolation outside the validated range is generally risky business.

2.2. Variables

Sometimes an enormous number of variables are at hand, e.g. in spectroscopic applications. Typically, the individual variables are not chosen based on relevance for the problem but simply because the instrument provides them. A resulting model and its interpretations can depend highly on the variables included in the model. Therefore, only the variables relevant for the given purpose should be included but that does not necessarily imply intensive variable selection down to the very few most important variables. Robustness and detection of relevant outliers are also factors to consider in this matter.

Outliers detected from irrelevant data are irrelevant outliers as shown with an example from spectroscopy (Figure 1), where VIS-NIR spectra of beer samples are subjected to PCA.

Here, some samples are identified as outliers using the full spectrum in a PCA model. However, PCA reflects the main variations in data, and may as such be misleading if the main variations are not relevant for the given purpose. In the example in Figure 1, only a part of the spectrum is relevant for the quality parameter of interest. Basing the PCA model on only this part results in detection of completely different and much more relevant outliers, those that are outliers in relation to the quality parameter.

In untargeted analysis of, for example, omics data, the problem is even more pronounced. In such a case, the majority of the variables may be irrelevant for a stated purpose and hence

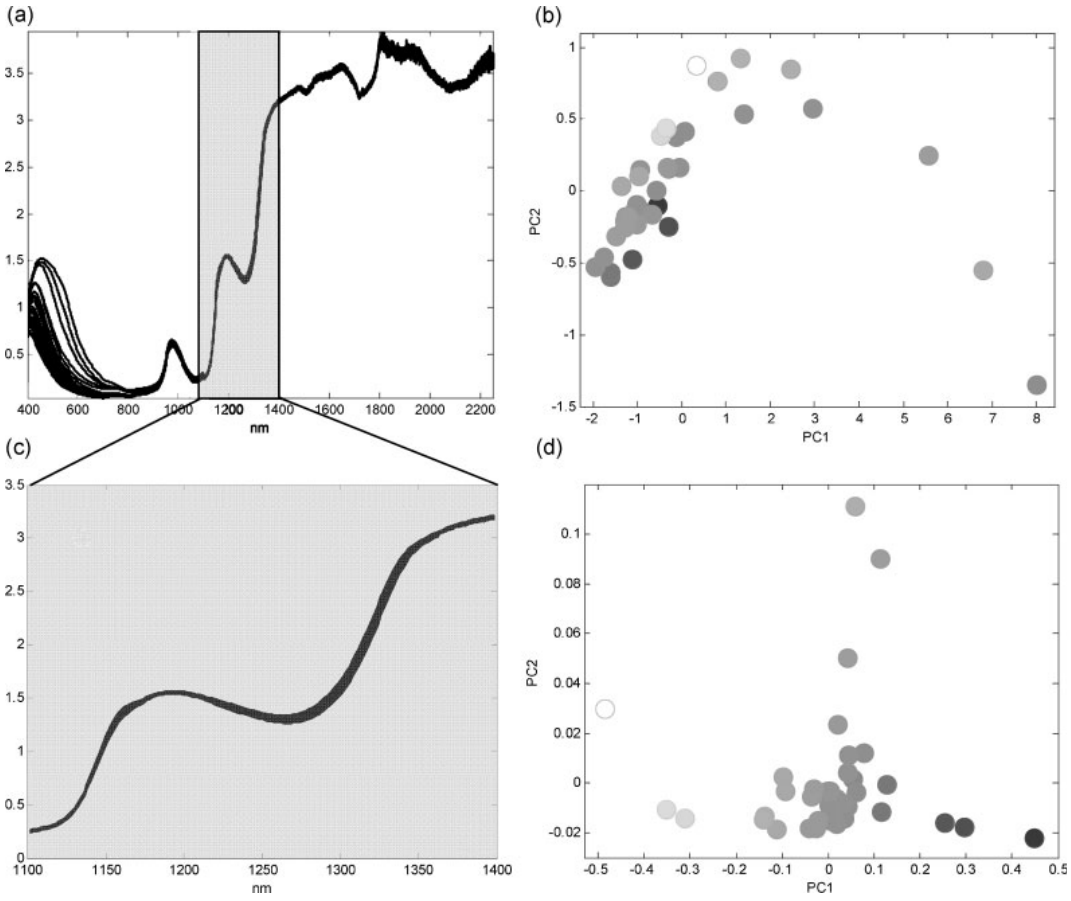


Figure 1. Use the relevant variables. (a) Full spectrum and (b) corresponding score plot. (c) Selected spectral window and (d) corresponding score plot. The score plots are colored according to the value of the quality parameter of interest, and it is clear that the scores in (d) directly reflect this parameter.

interpretations based on the total datasets should be performed with utmost care.

In conclusion, using the right samples and variables is all about spanning the *relevant variation*. Relevance can only be determined based on an agreed understanding of what the specific purpose of analysis is and what the relevance of the measured samples and variables is in that context. This becomes highly relevant when using model diagnostics such as 'explained variance' as shown below.

3. MODEL DIAGNOSTICS

For assessment of overall model performance or for model optimization such as deciding how many components to use or selection of variables, model diagnostics form a good support when applied properly. Software packages readily produce and present a range of diagnostic numbers, which sometimes may mislead the inexperienced user.

A great deal of model diagnostics is based on cross-validation; it is thus essential that the validation is performed adequately and produces reliable results. It has been stated several times that cross-validation, and in particular leave-one-out cross-validation, is optimistic and that care should be taken to reduce this bias through reasonable segmentation, combination with other techniques such as permutation tests, and cross-model validation [7–10]. Such issues will not be described in detail here.

3.1. What explained variance explains

It is a commonly met misunderstanding in both PCA and PLS modeling that higher explained variance or higher correlation means a better model. The two measures are, in many situations, directly linked, so that for a PLS model, the squared correlation coefficient R^2 can be interpreted as the proportion of the variance of the reference values, which can be explained by the fitted line [11].

One specific PCA model may explain 95% of the variation, and upon removing an outlier, the model explains 57%, so the new model is a poorer model? Percentages are relative numbers and such are *not* meaningful to compare *at all* in this case because they refer to two different datasets and hence the percentages are relative to two different sets of samples. One or the other model may be better, but this is impossible to state from a relative measure such as a percentage or a correlation. Instead an absolute measure has to be used such as the prediction error; preferably a validated error. This is also illustrated by an example from univariate regression (Figure 2). In this situation, removing an extreme sample reduces the percentage variance explained but actually decreases the error. This reflects the fundamental property of the explained variance that it depends on not only the performance of the model but also the distribution of the data. One extreme sample may influence the model's overall explained variance considerably, thus masking a much poorer fit to the remaining 'normal' samples.

Remarks:

- Similar arguments can be made about comparing different approaches to preprocessing. It is not possible to assess and choose between different preprocessing methods solely by how much variance a subsequent PCA model describes of the preprocessed data.
- The same arguments hold for using correlations as for percentage variance explained. Correlation is a relative measure and cannot be meaningfully used for comparing different datasets.

An additional point to consider is the *relevance* of the explained variance. It is of minimal interest that a model explains 90% of the variation if the interesting information is within the remaining 10%. Figure 3 shows an example where a score plot of a model that explains 73% of the variance is colored according to two different external variables (male/female and ill/healthy). Clearly, 73% explains the major part of the male/female information, but the ill/healthy information is not explained within these 73%.

In conclusion, explained variance is a measure that only with care should be used as a figure of merit during model optimization. On the other hand, percentage of variance is indeed a useful measure for conveying the quality of a final model once modeling decisions have been taken. For example, it may be useful for application specialists to know that the presented production model explains 44% variation or that the model of the spectroscopic data only explains 79% variation.

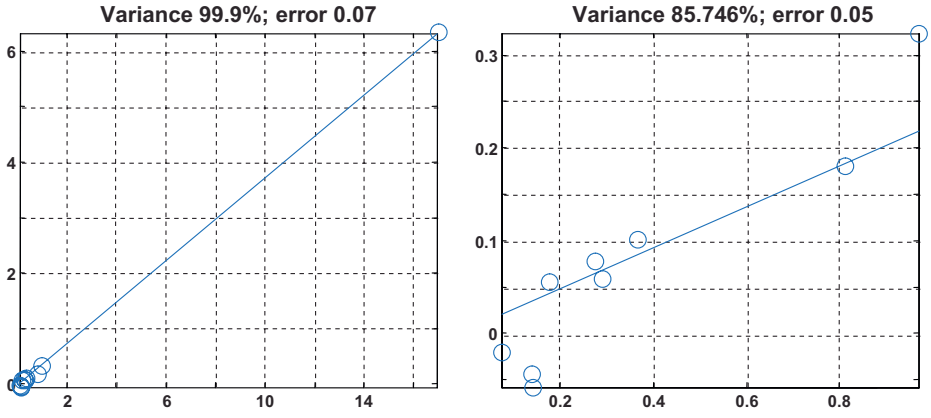


Figure 2. Example of the noncomparability of percentages across different datasets. To the left, a univariate regression describes almost 100% of the variation giving an error of 0.07. Upon removing the extreme upper-right sample, the explained variance is reduced, but so is the error.

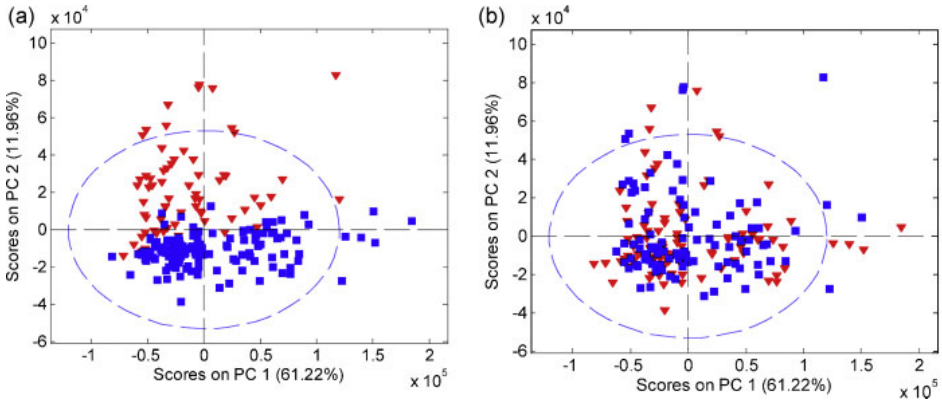


Figure 3. Two identical score plots (PC1, PC2) colored according to two different nonincluded variables: (a) male/female and (b) ill/healthy. Explained variance = 73%.

3.2. PLSDA: RMSECV is not useful

PLS used for classification, PLS-DA, is essentially just ordinary PLS regression with a special binary 'dummy' *y*-variable. It may therefore be tempting to optimize the model with respect to the root mean square error of cross-validation (RMSECV) 'as usual', but this is most often not optimal for classification purposes. RMSECV measures the model error in terms of deviations from the dummy reference, and larger deviations contribute more to the RMSECV than lower deviations. However, this approach does not include any considerations about the actual class borders or the classification rules, and basically, the predictions are rather irrelevant.

The interesting thing is how the predictions or scores are used for classification. The scores are the projections of the samples on the basis given by the training samples, and this is the space that in multiple, yet manageable dimensions best reveals the similarity of the new sample compared to the training samples. Here classification rules can be set and, on the basis of these, misclassification rates can be calculated as a much better measure of model performance.

Remark:

- The validation is always a critical point. By small sample sizes, validated misclassification rates can be very sensitive to the choice of segmentation; hence, it can be difficult to assess whether an obtained rate of misclassification is substantially biased relative to random results. If you toss a coin ten times, and obtain 8 heads and 2 tails, does that imply the coin is biased? [12].

3.3. PLSDA: Is the score plot validated?

Quite often—also in many published papers—scores and loadings from PLS-DA are shown and interpreted for a model that is not validated! Sometimes it is argued that validation is not needed because interest is only in visualization not classification, but this is nonsense. If validation shows that the model is not valid, it explicitly means that parameters such as scores and loadings cannot be trusted. For illustration, a PLSDA model is made on a random data matrix of 100 samples × 100 variables, assigned 50/50 to two classes, hence a completely meaningless datasets. As can be seen in Figure 4, the score plot from PLS-DA shows excellent separation between the classes. Validation of course shows that the model is completely invalid.

4. MODEL INTERPRETATION

4.1. Loadings

It is sometimes heard that in a loading plot, two variables close to each other are highly correlated. This is particularly relevant for PCA loading plots, like the one in Figure 5, but goes for PLS loading plots as well. This is not possible to claim unless the explained variance of the individual variables is assessed. In Figure 5 the two variables C-Glucose and H-MCHC are close to each other but also close to (0, 0), which means that they are (probably) not well explained by the two displayed principal components. Consequently, very little can sensibly be concluded about these two variables, and hence nothing about their similarity/correlation.

Remark:

- It could actually be that C-Glucose and H-MCHC, are highly correlated, but if they are measured in very low numerical values and not appropriately scaled, this cannot be seen from the loading plot directly. A correlation loading plot would reveal that the variables are well-explained and that they are then correlated.

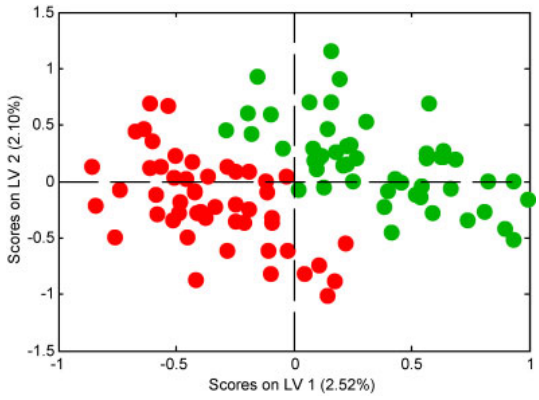


Figure 4. Score plot for a PLSDA model of random data. Excellent—but meaningless—class separation is obtained.

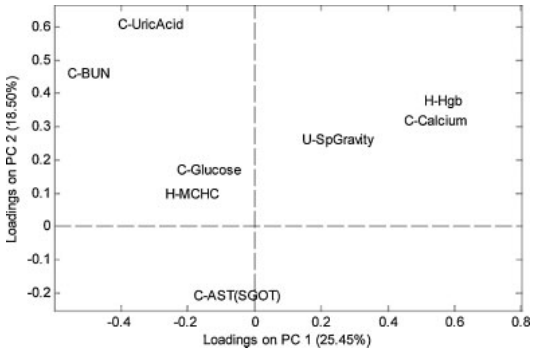


Figure 5. PCA loadings plot. H-Hgb and C-Calcium are highly correlated within the variation explained by PC1 and PC2, whereas nothing similar is certain about C-Glucose and H-MCHC regardless of their close position. Data from Reference [13].

Another set of variables, H-Hgb and C-Calcium are also closely positioned in the loading plot, but their more peripheral position reflects that they are far better explained in the plotted components. It is therefore appropriate to assume that they are correlated. However, even with H-Hgb and C-Calcium one should be careful about conclusions until the explained variability of these variables has been checked. This two-component model explains 72 and 64% of H-Hgb and C-Calcium, respectively, and there is no way to know what holds for the remaining variation. A scatter plot of H-Hgb and C-Calcium does therefore not show an impressive correlation as could possibly erroneously be deduced from looking only at the loading plot.

5. SCATTER AND BIPLOTS

It is common practice in chemometrics to make scatter plots of e.g. scores in order to spot groupings, outliers or other types of patterns in data. Scatter plots are excellent tools for doing so, but there are some very simple premises that have to be known in order not to misinterpret such plots. Geladi *et al.* [14] have written a tutorial on this subject, the essence of which is paraphrased in the following. Most of all, the scales on such a plot are crucial to be aware of when making interpretations. Consider the map shown in Figure 6 (left). This is a (fairly) distance-preserving plot of part of the world. Three cities are highlighted: New York in USA, Copenhagen (northmost in Europe) and Rome (southmost in Europe). As can be easily seen, Rome and Copenhagen are closer to each other than to New York. Consider instead the alternative plot in Figure 6 (right). In this plot, it seems that New York is not much further away from Rome than Copenhagen is. Clearly this interpretation is wrong and clearly the reason is that the horizontal and vertical axes cannot be compared. The scale is different on the two axes.

While the above example is quite simple, it is interesting that when we move to a scatter plot of scores, many users will be just as happy with assessing distances in any of the above plots. Both of the above plots are correct, but they allow for different types of evaluations. If a user wants to look at a score plot and say that two samples are similar (implicitly meaning similar with respect to the data input to the model) because the samples are closely positioned, then the axes that the scores are plotted upon *must* be comparable. Most programs will not be able to produce such a plot because the plotting mostly just aims at filling the window, hence stretching the axes and completely destroying the opportunity to assess distances. There is a simple solution though: just avoid stretching the axes and plot on a scale where the loadings are normalized. Then the scores *do* reflect distances in the data space.

Remarks:

- If distances are to be assessed in a loading plot, then again, the scale of the two axes should be comparable. Therefore, the loadings should be re-scaled so that they are given in terms of normalized scores.
- When it comes to bi-plots, the situation is a bit complicated because it is actually not possible to make a plot that maintains distances in both row and column space. The theory of bi-plots [15] provides means for reasonable compromise plots where the incorrectness is equally spread on scores and loadings. Apart from this, the main fault, in bi-plots though, is still that many software packages simply plot the scores and loadings on top of each other filling out the entire window of the plot. Clearly, this completely ruins the possibility for making detailed meaningful interpretations of the plot.

Correlation loadings (or scores) can, to some extent, alleviate some of the problems mentioned above, but while such are useful, they also destroy some of the intrinsic interpretational possibilities of scatter and bi-plots.

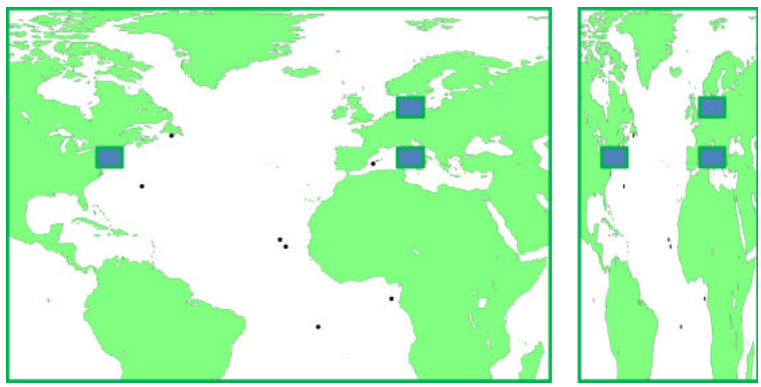


Figure 6. Two maps showing the same part of the world, either true to the surface distances (left), or with different scaling on the axes (right).

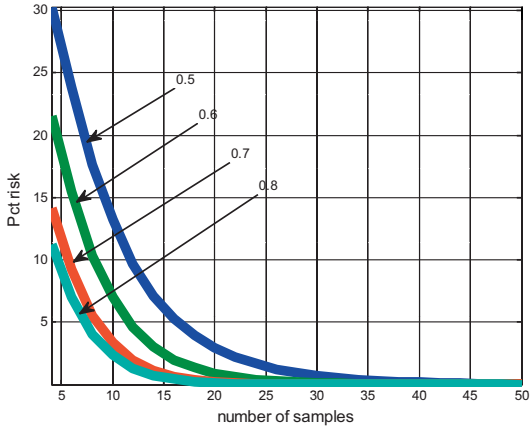


Figure 7. Effect of sample size on the risk of obtaining a Pearson's correlation r of $\pm$ 0.5–0.8 from random numbers.

5.1. Correlation and regression coefficients

An essential element of regression modeling is correlation. Various correlations are used for assessment of model performance, variable selection and form the basis of model interpretation. However, as already shown above, correlations are not perfect measures for all purposes and should not be trusted and used blindly. There are more pitfalls to be aware of; some of these shall be highlighted here.

First of all, a correlation is basically a mathematical construction, which in many cases has a meaningful interpretation useful for real-life data. This interpretation is not always as straightforward as it may appear. Correlations may be coincidental, so that nonrelevant data by chance happen to display a good correlation with a variable of interest. This is particularly relevant for datasets of few samples and many variables as is mostly the case for *omics* data, for example.

Figure 7 shows how the number of samples affects the probability of obtaining a correlation of $\pm$ 0.5–0.8 or higher when comparing two variables consisting of only random numbers. If we have a dataset of 30 samples and 50 000 variables (e.g. microarray data), and we wish to see how each of the 50 000 variables correlate with another variable y (e.g. glucose concentration), then if none of the 50 000 variables are actually causally or indirectly related to y, we will still expect to find 8–9 variables that have a Pearson's correlation r higher than $\pm$0.7. We may then mistakenly suggest these as biomarkers if adequate validation is not implemented. The latter is not a trivial task for such data, as cross-validation is generally of limited use [8].

Correlation does not imply causation! This is fundamental, but it still seems to be forgotten sometimes. In some cases, correlations are directly causal, such as the relation between absorbance and concentration of an analyte where Beer's law applies. Such a direct relation is necessary for causal explanations.

In other cases, the correlation is only an observed correlation as the relation between sales of ice cream and the number of drowning accidents, where an underlying factor (the sun) is the direct causal link. Such relations are true and absolutely valid to apply for predictions, e.g. for health diagnostic purposes; one should just be aware of the limitations with respect to interpretation.

It is consequently highly important to try to identify whether the correlation is causal before explanatory conclusions are made. Often confounding factors are present that are not easily revealed or unexpected and hence not taken care of by the experimental design. Below (Figure 8) is a classification example, where the search for mass spectrometry biomarkers for ovarian cancer would be seriously misled if the effect of sample storage time (Figure 8a) is not recognized. Imagine that healthy samples are collected first and samples from ill patients later. An apparently near-perfect diagnostic model could then be developed (Figure 8b), and the $m/z = 6638.41$ variable claimed a cancer biomarker, but it merely reflects sample age [16].

Remark:

- Sometimes, thorough variable selection is performed with the purpose of reducing the dataset to a very low number of variables, which are then believed to be candidates as causal markers. It may however be the case that noncausal variables are stronger markers, and by this procedure they will thus be among the few selected variables whereas the causal variables may have been removed.

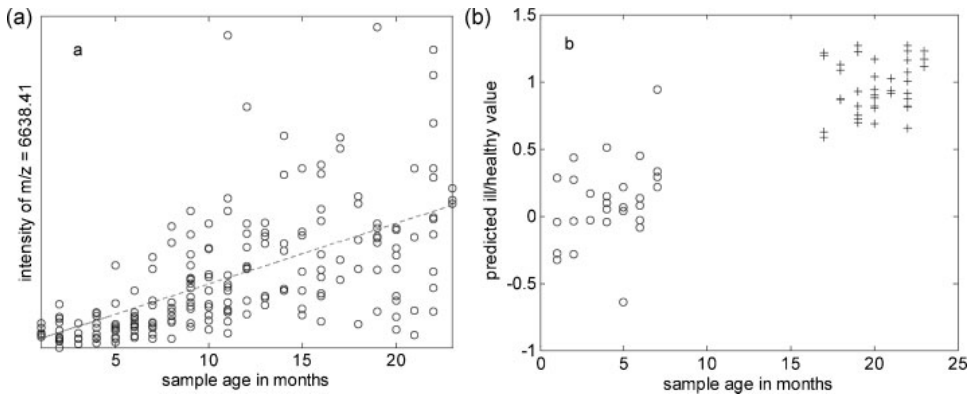


Figure 8. High degree of confounding of storage factor with class membership gives optimistic classification results and misleads biomarker search. (a) Effect of sample age on selected m/z variable. (b) With poor sampling due to unexpected confounding, a model on the full m/z-spectrum separates samples from healthy (o) and ill (+) patients well [16].

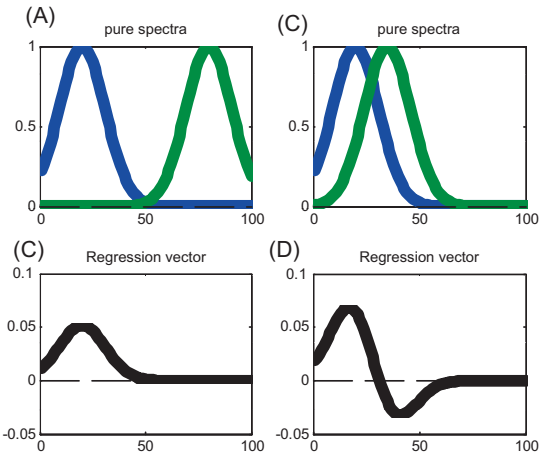


Figure 9. The presence of overlapping signals makes interpretation complicated. In (a) and (b) the blue peak is the analyte and the green peak is the interferent. (c) and (d) show the resulting PLS regression vector.

Regression coefficients may be used for interpretation and variable selection. The units of these match both the independent variables, **X**, and the dependent variable **y**; consequently, the scaling of **X** must be kept in mind when interpreting regression coefficients. Interpretation is most straightforward if all variables of **X** are brought to the same numerical level by scaling to unit variance ('auto-scaling') or similar. In this case the regression vector may reflect the relative importance of the individual variables, but it may not be that simple. With spectral data, overlapping signals introduce phenomena in the regression vector that disturb interpretation. Seasholtz and Kowalski [17] elaborate neatly on this; in essence, what happens to the regression vector in case of overlapping (= nonorthogonal) signals is shown in Figure 9. As the pure spectra of two analytes begin to overlap, the ideal regression vector no longer looks like the pure spectrum because negative parts and shifts in position of peak maximum are introduced. The presence of such phenomena may blur interpretation seriously. Similar problems occur for nonspectral data. For example a correctly estimated negative regression coefficient can easily be obtained for a variable that is positively correlated to the response.

6. CONCLUSION

In this paper we looked at a number of commonly occurring mistakes in the use of chemometrics. Generally, the problems often appear to be a result of a combination of misunderstandings and noncritical push-the-button analysis. What often happens is that the software readily throws plots and diagnostics in the face of the user, and the inexperienced user is inclined to apply these rather uncritically. Using well-known, widely used diagnostics seems safer and more 'correct' than sound reasoning, although the latter is often preferable. The only way to go is to take responsibility: decide what is relevant by support of biological/chemical knowledge and sound reasoning and always keep the purpose of the modeling in focus!

REFERENCES

1. Daszykowski M, Walczak B. Use and abuse of chemometrics in chromatography. *Trends Anal. Chem.* 2006; **25**: 1081–1096.
2. Difoggio R. Guidelines for applying chemometrics to spectra: feasibility and error propagation. *Appl. Spectrosc.* 2000; **54**: 94–113.
3. DiFoggio R. Examination of some misconceptions about near-infrared analysis. *Appl. Spectrosc.* 1995; **49**: 67–75.
4. Kourti T. Multivariate statistical process control and process control, using latent variables. In *Comprehensive Chemometrics*, Vol. 4, Brown S, Tauler R, Walczak R (eds). Elsevier: Oxford, 2009; 21–54.
5. Kourti T. Process analysis and abnormal situation detection: from theory to practice. *IEEE Contr. Syst. Mag.* 2002; **22**: 10–25.
6. Wise B, Ricker N, Veltkamp D, Kowalski B. A theoretical basis for the use of principal component models for monitoring multivariate processes. *Process Contr. Qual.* 1990; **1**: 41–51.
7. Breiman L, Spector P. Submodel selection and evaluation in regression. The X-random case. *Int. Stat. Rev./Revue Internationale de Statistique* 1992; **60**: 291–319.
8. Rubingh C, Bijlsma S, Derks E, Bobeldijk I, Verheij E, Kochhar S, Smilde A. Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* 2006; **2**: 53–61.
9. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst.* 2006; **84**: 69–74.
10. 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**: 81–89.
11. Næs T, Isaksson T, Fearn T, Davies T (eds). *A User-friendly Guide to Multivariate Calibration and Classification*. NIR publications: Chichester, UK, 2002.
12. Brereton R. Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. *Trends Anal. Chem.* 2006; **25**: 1103–1111.
13. Wise B, Gallagher N, Bro R, Shaver J, Windig W, Koch R. *PLS_Toolbox for Use with MATLAB, Version 5.5.1*. Eigenvector Research, Inc.: Wenatchee, USA, 2009.
14. Geladi P, Manley M, Lestander T. Scatter plotting in multivariate data analysis. *J. Chemom.* 2003; **17**: 503–511.
15. Gower J, Krzanowski WJ (eds). A general theory of biplots. In *Recent Advances in Descriptive Multivariate Analysis*. Clarendon Press: Oxford, 1995; 283–303.
16. West-Nørager M, Bro R, Marini F, Høgdall E, Høgdall C, Nedergaard L, Heegaard N. Feasibility of serodiagnosis of ovarian cancer by mass spectrometry. *Anal. Chem.* 2009; **81**: 1907–1913 .
17. Seasholtz M, Kowalski B. Qualitative information from multivariate calibration models. *Appl. Spectrosc.* 1990; **44**: 1337–1348.

PAPER III

Genomics 92 (2008) 373–383

Contents lists available at ScienceDirect

Genomics

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

Direct functional assessment of the composite phenotype through multivariate projection strategies

Ana Conesa [a,*], Rasmus Bro [b], Francisco García-García [a], José Manuel Prats [c], Stefan Götz [a,d], Karin Kjeldahl [b], David Montaner [a], Joaquín Dopazo [a,d,e]

[a] Bioinformatics Department, Centro de Investigacin Principe Felipe, Valencia, Spain
[b] Department of Dairy and Food Science, Faculty of Life Sciences, Copenhagen University, Denmark
[c] Department of Applied Statistics, Technical University of Valencia, Valencia, Spain
[d] Center for Biomedical Research on Rare Diseases (CIBERER)
[e] Functional Genomics Node (National Institute for Bioinformatics, INB), Valencia, Spain

ARTICLE INFO

Article history:
Received 22 February 2008
Accepted 28 May 2008
Available online 13 September 2008

Keywords:
Data integration
Functional genomics
Multivariate regression
Gene ontology
Phenotype
Gene annotation
Partial least squares
Principal component analysis

ABSTRACT

We present a novel approach for the analysis of transcriptomics data that integrates functional annotation of gene sets with expression values in a multivariate fashion, and directly assesses the relation of functional features to a multivariate space of response phenotypical variables. Multivariate projection methods are used to obtain new correlated variables for a set of genes that share a given function. These new functional variables are then related to the response variables of interest. The analysis of the principal directions of the multivariate regression allows for the identification of gene function features correlated with the phenotype. Two different transcriptomics studies are used to illustrate the statistical and interpretative aspects of the methodology. We demonstrate the superiority of the proposed method over equivalent approaches.

Introduction

Gene expression profiling is used to study the gene regulatory basis of phenotypic or developmental characteristics. Statistical analysis of transcriptomics data is normally addressed through a two-step process: first, a statistical test is performed to derive a *P* value for the association of individual gene expression values to the phenotype or experimental condition(s), and a number of "significant genes" are selected on the basis of an arbitrary *P* value threshold. Most commonly used methods apply modifications of the *t* statistics or ANOVA to generate hypothesis testing of differential expression [1–4]. Secondly, selected genes are further analyzed to identify their relevant association to cellular functionalities [5,6]. Fisher's exact test, the Kolmogorov-Smirnov test, or the chi-squared are common statistics to identify functional classes with a significant enrichment within the pool of differentially expressed genes [7]. This widely used approach presents a number of drawbacks. On one hand, the univariate nature of the by-gene statistical assessments implies that any informative correlation pattern within gene expression will be ignored. On the other hand, strong *P* value corrections need to be applied to deal with the concomitant multiple testing scenarios and this can seriously hamper the identification of significant features on large datasets [8]. Furthermore, as functional assessments—which paradoxically have their foundation on the correlated nature of gene activity—are performed after univariate gene selection, results are dependent on the *P* value cutoff of choice, which can be problematic. Thus, too strict *P* value thresholds may lead to univariately nonsignificant genes (that are in fact significant in the multivariate space, but remain undetected) while too permisive cutoffs may result in multivariate important features getting lost among irrelevant information. Finally, when the target phenotype is not composed by a single variable but a space of different measurements (e.g., age, gender, different clinical parameters), the evaluation of differential expression under a univariate strategy can imply multiple and difficult assessments.

Multivariate approaches to gene expression analysis try to overcome these limitations. Principal component analysis (PCA), factor analysis, and multiple correspondence analysis are multivariate space reduction methodologies that exploit the correlation structure in the data to identify relevant patterns of variation [9,10]. These approaches have been applied to the analysis of transcriptomics data and have showed their potential in capturing relevant associations in the multivariate expression space that would escape to univariate analysis [11–13]. Several authors have proposed different strategies for deriving gene-associated

* Corresponding author. Fax: +34 96 328 97 01.
E-mail address: aconesa@cipf.es (A. Conesa).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.05.015

significance values of differential expression in this multivariate analysis context. Lu et al. [13] used the Hotelling T^2—a multivariate extension of the univariate *t* statistics—to select significant genes, and proposed a recursive method to deal with the singular data structures that appear when the number of variables greatly exceeds the number of observations. Landgrebe et al. [14] applied ANOVA on PCA results to identify components with a significant difference between experimental groups and applied a VIP-like statistics to evaluate component significance. Nueda et al. [15] employed the gene leverage along with a permutation test to find significant contributions to the multivariate projection. In all these examples data analysis focuses exclusively on expression values and does not incorporate a priori knowledge. Approaches that consider the functional role of genes while trying to capture the cooperative acting of the set of genes as a whole are, e.g., the so-called gene set methods, such as the GSEA [16] and FatiScan [17]. In these methodologies, genes are ranked according to a measure of differential expression and the enrichment of functional classes towards the extremes of this ranking, rather than a single group of genes, is tested. These methods have proven

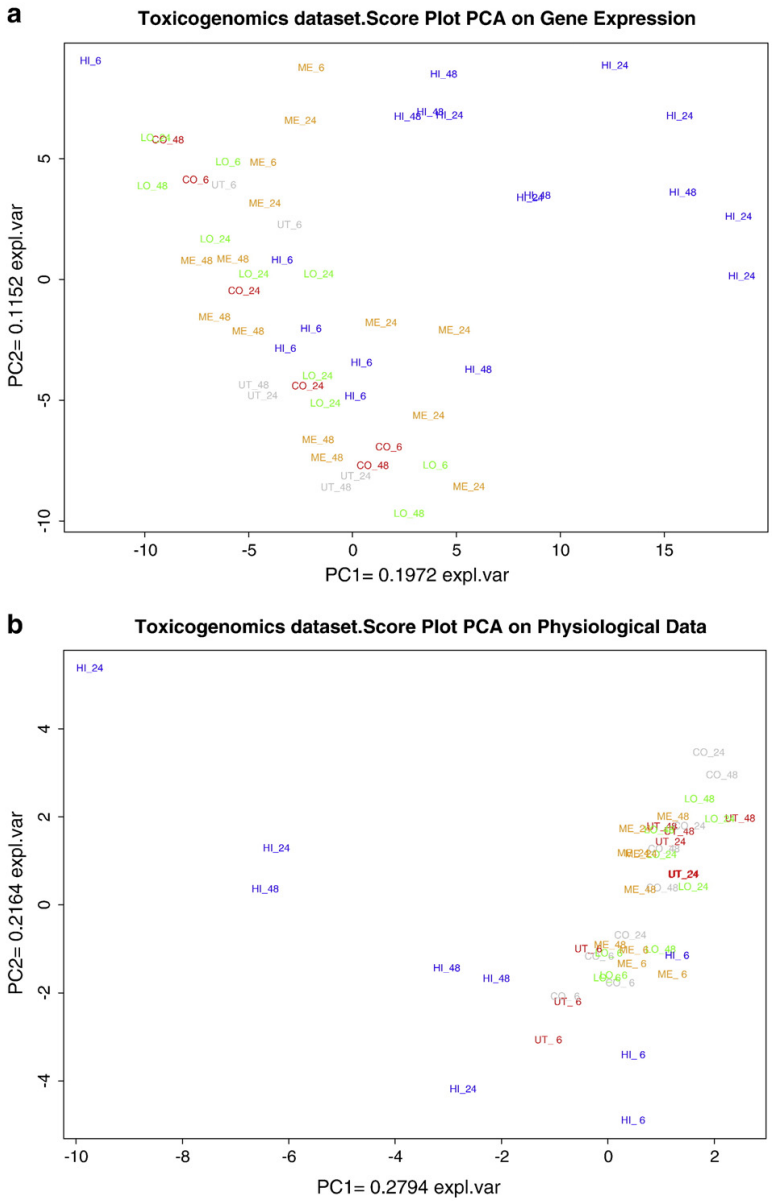


Fig. 1. PCA analysis of toxicogenomics data. Samples are labeled by the treatment group: HI, high bromobenzene dose; ME, medium bromobenzene dose; LO, low bromobenzene dose; CO, placebo; UT, untreated control. _6, _24 and _48 denote hours of administration. Closeness in the projected space indicates similarity between samples. (a) PCA score plot with gene expression data. (b) PCA score plot on physiological variables. (c) PCA score plot on functional variables.

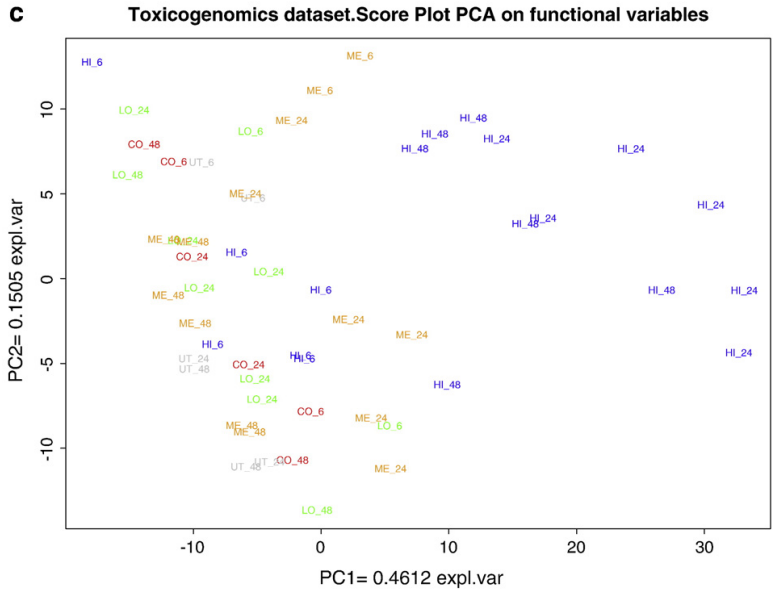


Fig. 1 (*continued*).

to be very sensitive as they do not require genes to be significantly declared, but still follow a prior univariate test to generate the gene rank which, on the other hand, can normally only be derived for two-class problems. Recent work has tried to combine multivariate statistics with functional assessment to overcome some of the limitations of the gene set methods. Kong et al. [18] proposed a methodology in which functional expression submatrices—i.e., the set of gene sharing the same function—were used to create projected functional subspaces which subsequently were tested by Hotelling T^2 statistics for their ability to separate between treatments. The same basic idea, but with a different statistical approach, is followed by Nettleton et al. [19]. These authors apply the multiresponse permutation procedure (MRPP) [20] to test against the null hypothesis of invariant distribution of gene expression among different treatment classes. Different in their purpose are the methods which combine multivariate expression with gene function to identify significant co-occurrences of functional classes [21,22] or predict gene function from transcriptomics data [23,24].

Still, all these methodologies deal with a single response variable—at two or more levels—if any, to evaluate significant functional associations. Very little has been done in the multivariate analysis of complex phenotypic determinations that follow transcriptomics profiling. Such data are of special relevance in biomedical research where composite clinical frames need to be understood in the light of gene activity [25,26].

In this work we present a novel approach for the analysis of transcriptomics data that integrates functional annotation and expression values in a multivariate fashion and directly assesses the relation of functional features to a multivariate space of response variables. Our method benefits from the correlation patterns between both gene and gene functions to identify a functional signature that best predicts the phenotypic outcome.

Results

Troxicogenomics dataset

PCA analysis of both gene expression data (Fig. 1a) and phenotypic variables (Fig. 1b) revealed a first component of variability that basically differentiates the high bromobenzene dose treatment at 24 and 48 h from the rest of the conditions. A pretty similar PCA score plot was found for gene expression and phenotypic variables, indicating that the major pattern of the variability in both datasets had similar structures and related to the intensive administration of the drug.

The initial GO term filter procedure generated a total of 1140 GO terms from the three main GO branches. After PCA-based transformations 823 functional components were created with an average explained variance of 40%. In most cases GO terms were represented by one functional variable and only in a few cases up to two variables were derived by a single functional class. Table 1 summarizes the results of the analysis procedure.

PCA analysis of the new matrix of functional variables showed a projected space similar to that obtained previously with gene expression and clinical data (Fig. 1c), but explained variance for the first and most discriminating component was clearly higher (46% with functional variables versus 20% with gene expression data), indicating a more compact signal in the transformed data. PLS (partial least squares) regression was then applied to relate the measured physiological parameters to the new space of functional variables. The number of components was selected by leave-one-out cross-validation, resulting in a 7-component model with maximal overall predicting value. Analysis of the R^2, Q^2, and VIP parameters for individual physiological variables permitted the identification of differences in relation to the functional data (Table 2). Variables such as ASAT, bilirubin, LDH, and ALAT showed important contributions (high VIP) and were well predicted by the model (high Q^2), which means that these parameters are highly related to the gene functional response triggered by the toxic compound. Other variables such as total protein or albumin were low contributors to the model and failed to be predicted, indicating their poor association to the gene expression pattern revealed in the analysis. These variables were removed from the final PLS model which obtained an average prediction error of 0.58 and a determination coefficient R^2 of 0.75, both highly significant (P value ~ 0). It is worth noting that R^2 values were in general high, also for poorly predicted variables, which

Table 1
Quantitive figures in the analysis procedure of the toxicogenomics and breast cancer datasets

	Origina data		PCA transformation to funcional data				PLS model			
	Probes	Annotated GO terms	GO term selection	Functional variables	Mean expl. var.	Mean GO level	No. comp.	Average R^2	Average Q^2	Significant funct. vars.
Toxicogenomics	2665	7411	1140	823	0.4	6.7	7	0.75	0.58	50
Breast cancer	22283	10940	3129	1901	0.44	6.2	4	0.45	0.3	65

illustrates the optimistic nature of this parameter to evaluate the validity of a regression model.

Graphical analysis of the PLS model discovers further aspects of the data. The score plots associated with the functional and physiological data matrices showed a stronger differentiation between the high bromobenzene doses at 24 and 48 h and the remaining samples (Figs. 2a and 2b), indicating that the physiological response to the gene expression effects of toxic compound is mainly concentrated under these conditions. Additionally, the **Y** biplot of the PLS model (Fig. 2c) revealed a positive correlation (same orientation in the projected space) of these high toxicity levels with the most responsive cellular compounds, while other parameters, such as glucose and kidney weight, showed a negative relationship. In fact, increased levels of ASAT, bilirubin, LDH, ALAT, and phospholipids have been associated with the response to xenobiotics and are considered as markers of toxicity [27], while plasma glucose concentrations tend to decrease for the imbalance in energy requirements [27].

Selection of significant functional variables in the PLS model was done by resampling methods. Fifty functional classes were selected at a P value < 0.05. Functional variables are represented in the **Y** biplot by open dots (Fig. 2c). Significant variables are depicted colored. Significant GO terms include *response to stimulus*, *heme binding*, *fatty acid metabolic process*, *oxidoreductase activity*, *glutathione transferase*, *apoptosis*, *ribosomal unit,* and *cytoskeleton* (see Supplemental Material T1 for a complete list). Fig. 3 shows the DAG of the significant functional terms corresponding to the Biological Process GO branch. Significant functional categories extensively explain the cellular adaptative response to drug administration which includes conjugation to glutathione by glutathione transferase, modification in oxidative, heme containing enzymes, activation of the ribosome machinery for protein synthesis, and cytoskeleton reorganization [27].

Table 2
PLS model parameters for the **Y** data structure (physiological variables) of the toxicogenomics dataset

Physiological variable	R^2	Q^2	VIP
ASAT	0.94	0.81	5.54
Bilirubin.tot	0.89	0.67	5.02
LDH	0.92	0.67	5.66
ALAT	0.90	0.66	5.27
Phospholipids	0.81	0.61	4.16
Liver.BW	0.79	0.51	3.23
Liver	0.72	0.48	2.38
Body.Weight	0.59	0.43	2.14
Glucose	0.58	0.43	2.69
Creatin	0.72	0.43	4.41
Kidney.BW	0.61	0.39	2.76
GSH.corr	0.57	0.37	3.03
Triglycerides	0.63	0.36	1.77
Cholesterol	0.67	0.35	3.64
Urea	0.54	0.27	1.97
ALP	0.72	0.27	3.05
AG.ratio	0.59	0.15	3.05
Tot.Protein	0.53	−0.03	2.85
Kidneys.weight	0.20	−0.09	0.78
Albumin	0.38	−0.12	1.98

Breast cancer dataset

The breast cancer dataset contained nearly 10 times the number of probes of the toxicogenomics dataset. Still data transformation by PCA on functional classes rendered a not very different compression result: 1901 functional variables with an averaged explained variance of 44% (Table 1). PCA and PLS score plots of both gene expression and functional data showed a different distribution of p53+ and p53- samples along the first component, which was more pronounced when samples were labeled by their ER status (Figs. 4a and 4b). This is in agreement with observations in the original work on the incompleteness of the p53 sequence determinations to establish the p53 deficiency status in breast tumors [28]. The PLS analysis with functional variables resulted in a 3-component model with significant Q^2 and R^2 parameters. Again, only a subset of clinical variables was well predicted by the model, namely the p53seq, ER status, histological grade, and PgR status for which the mean R^2 and Q^2 were 0.45 and 0.30, respectively (Supplemental Table T2). Furthermore, the **Y** loading plot of the PLS model showed a negative correlation between the p53 genotype and the ER status and histological grade (Fig. 5), which has been described in previous reports [29]. Sixty-five significant functional variables were detected by resampling. The corresponding Gene Ontology terms pointed to functions related to the immune response, cell division and proliferation, cytoskeleton organization, estrogen receptor signaling—already highlighted by Miller and co-workers [28]—and also to novel functional activities such as *activation of JNK activity*, *fiber development,* and *chemokine activity*. The complete list of significant functional terms in the breast cancer study is provided as supplemental material T1.

Comparison with other functional assessment methods

We compared the functional class results in the toxicogenomics and breast cancer examples, respectively, to two traditional univariate pathway analysis methods, namely the enrichment analysis by the Fisher exact test [30] and the Gene Set Enrichment Analysis provided by the FatiScan [17]. Additionally, we compared our results in both data examples to those obtained with the multivariate approach proposed by Kong et al. [18] where the Hotelling T^2 statistics is used to find treatment-associated significant differences between functional class-defined gene expression submatrices. GO term comparisons were done using the Blast2GO software [31]. In contrast to our strategy, all comparing methodologies required the selection of two contrasting conditions—HI bromobenze treatment vs control in the toxicogenomics example, and p53seq label for the breast cancer study—to define the analysis. In both study cases, traditional univariate approaches provided a reduced and semantically less rich, i.e., consisting of more general terms, set of significant functional classes (see Supplemental Table T1). On the contrary, the Hotelling T^2 method by Kong et al. consistently generated a far too large selection of GO classes (256 and 1520 GO terms for toxicogenomics and breast cancer datasets, respectively) which included most of the functions detected as significant by our method and many others suspiciously false positives, such as neural activity-associated processes in the case of the toxicogenomics liver samples and eye and bone specific functions in the case of the breast cancer data. Detailed information in functional results is provided in Supplemental Table T1.

Discussion

The proposed method integrates in one analysis three basic elements of transcriptomics studies: gene expression data, functional annotation, and phenotype characteristics, providing a direct relationships between gene function and response variables. The integrative analysis of transcriptomics data has been the subject of recent statistical developments [18,32–34]. Our method differs from other approaches in that it translates gene expression to a distinct expression signature of the functional class. By applying PCA on gene sets that share a function, the major expression patterns associated to the functional class can be identified and used as novel variables to study the phenotype. With this approach, two potentially critical problems can be overcome based on the assumption that important genes are correlated to similar genes. First of all, the unimportant genes are dramatically reduced in numbers which can be decisive to be able to detect significant variations. Secondly, the important (as well as unimportant) variation is expressed in a reduced form by scores from principal component analysis. Hence, ideally, each phenomenon appears only once and therefore has a better chance of influencing the further analysis. A key element to achieve this is the criterion for selecting Gene Ontology terms and functional components. We applied a simple filtering procedure on the set of initial GO terms to avoid annotation redundancies that arise from the hierarchical structure of the Gene Ontology. In this way candidate GO terms are guaranteed to collect at least partially different annotation sets. More important even is the criterion for selecting functional components. Component selection in dimension reduction approaches are habitually based on cross-validation or scree-plot analysis [35]. These procedures consist of building and evaluating different models by leaving out one or more observations that are then predicted by the model built, or using as many components as needed to reach a given amount of explained variance. In our case, the purpose of component selection is to identify a relevant expression features of the functional class, rather than to test prediction ability or sufficiently explain the functional submatrix. Therefore we choose a criterion that would select functional variables when they collect an amount of variance above what could be considered random noise. The effect is an important reduction in functional classes from the original GO set and a selection of terms of medium hierarchy depth level (mean value around 6.5) with a sufficient explanatory capacity (~40% on average).

Compared with common univariate statistical approaches for the assessment of gene functional enrichments [5,16,17], the method proposed in this work differentiates for its consideration of the coordinative behavior between functional classes—not only within—and therefore potentially capturing the cooperative activity of functional processes. This implies that covariance between genes is particularly stressed in our approach, since a functional class of differentially expressed genes but not correlated gene members might not be detected by our method but could be identified by a univariate strategy. Compared to another published multivariate method, our approach seems to achieve a good trade-off between sensitivity and selectivity in the selection of significant functional classes. We postulate that the two-step strategy of our method—creation of functional variables followed by PLS inference—and significance criterion based on the distribution of VIP values of the randomized PLS models are key for obtaining a sensible selection of functional variables. The simple randomization of expression values in functional submatrices would tend to create in too compacted Hotelling T^2 null distributions that would declare as significant an excessive number of variables in the Kong et al. method.

Furthermore an additional aspect in our approach is that the analysis is not restricted to pairwise comparisons between conditions, but it can evaluate the composite phenotype dynamically and for the relationships within outcome parameters. This last consideration of multiple phenotypic characterizations in microarray datasets was likewise addressed by Fellenberg et al. [36]. In this work, Correspondence Analysis was used to study relationships between transcriptomics data and extensive sample annotations. The authors developed an interesting method to extract relevant phenotypic characteristics and map them to gene expression features by multivariate projection methods. However, this work does not incorporate the gene functional information which can provide a more interpretable result, in terms of biological processes, to the relationship phenotype-transcriptome, and also does not exploit an inferential relationship, such as PLS does, to achieve an optimized projection of the gene expression and the phenotypic spaces.

All together, our results indicated that the proposed method is effective in extracting informative functional signatures that differentially correlate with diverse aspects of the phenotype. We believe that this approach will be of great help in the study of the molecular mechanisms behind the observed characteristics of organisms and to unravel genotype-phenotype functional relationships.

Material and methods

The proposed method

Schematically, our proposal uses multivariate projection methods to obtain new correlated variables for gene sets which share a given function. These new "functional variables" are then used to perform a multivariate regression on a set of response variables. The analysis of the principal directions of the multivariate regression allows for the identification of gene function features correlated with the phenotype. The proposed method consists of the following steps:

1. Find the functional annotation of the genes in the transcriptomics dataset. For each functional term, create a "subexpression matrix" with all associated genes.
2. Perform principal component analysis in each of the new expression matrices and select a number of components that collect nonrandom variation.
3. Collect the PCA scores of the selected components into a new matrix of "functional variables." These functional variables represent coordinative expression patterns of genes associated by a functional label.
4. Use this new matrix to perform partial least square (PLS) regression on the response variables.
5. Select significant functional variables in the PLS by bootstrap.

In principle, any functional vocabulary can be used to elaborate functional variables. In this work we have taken the Gene Ontology scheme (http://www.geneontology.org) as it is the most extensive vocabulary for the description of gene function. We considered all terms present in the Direct Acyclic Graph (DAG) encompassed by the gene collection of the transcriptomics datasets but removing all annotation redundant terms. A term is considered annotation redundant within a given gene collection if it has a child term with identical gene annotation set. For example, if GO:0006915 (*apoptosis*) has 15 annotated genes and parent term GO:0012501 (*programmed cell death*) back-inherits these and only these 15 genes, then *programmed cell death* is considered annotation redundant and removed from the initial set of functional classes.

Principal component analysis projects a data matrix into a space of lower dimension while keeping most of the variability in the original data [9].

The PCA model for each functional class can be expressed in matrix notation as

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{R},$$

where $\mathbf{A}$ ($\mathbf{I}\times\mathbf{F}$) is the matrix collecting the F functional variables, $\mathbf{B}$ ($\mathbf{J}\times\mathbf{F}$) is the loading matrix that indicates the importance of each gene on each functional variable, and $\mathbf{R}$ is the residual matrix. Dimension reduction is possible when there exists a correlation structure in the dataset, i.e., where there is a sufficient number of genes with

correlated expressions. In this sense, PCA can be considered as a summarizing method and in our approach the profile given by the observations scores of each principal component reflects a coordinated behavior of a group of genes within the functional class and defines the so-called functional variables. Selection of functional variables is done based on the amount of variance explained by the corresponding component, normalized by the number of genes in the functional class. Components—i.e., functional variables—are selected in this case if their normalized variance is greater than the average gene variance of the complete dataset.

The relationship between functional variables and phenotypic variables is analyzed by partial least squares [10]. PLS is a dimension reduction regression approach which finds a projected space that maximizes the correlation between independent and dependent data

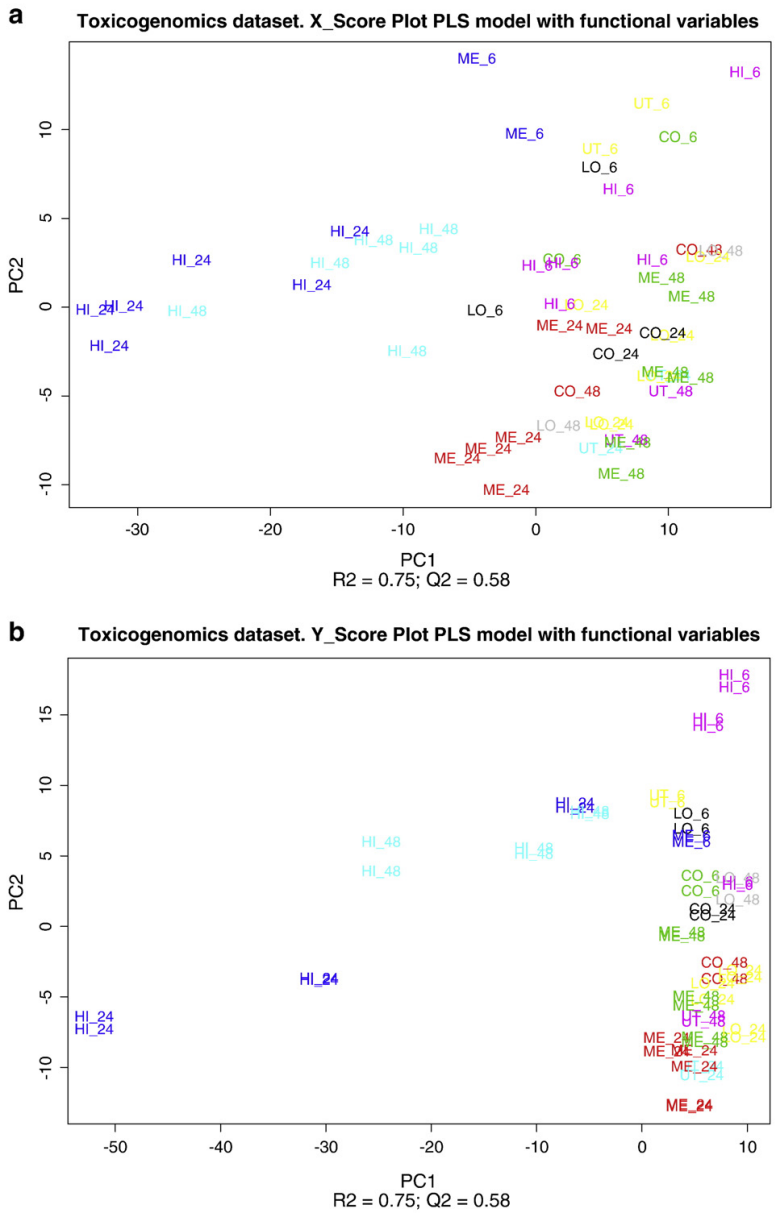


Fig. 2. PLS analysis of toxicogenomics data. Samples are labeled by the treatment group: HI, high bromobenzene dose; ME, medium bromobenzene dose; LO, low bromobenzene dose; CO, placebo; UT, untreated control. _6, _24 and _48 denote hours of administration. (a) **X**_score plot showing the relationships between treatments according to the dimension reduction of the functional data. (b) **Y**_score plot showing the relationships between treatments according to the dimension reduction of the physiological variables. (c) **Y**_ biplot shows the projection of both functional and physiological variables. Variables poorly explained by the model are given in gray. Functional variables are represented by dots and colored when significant.

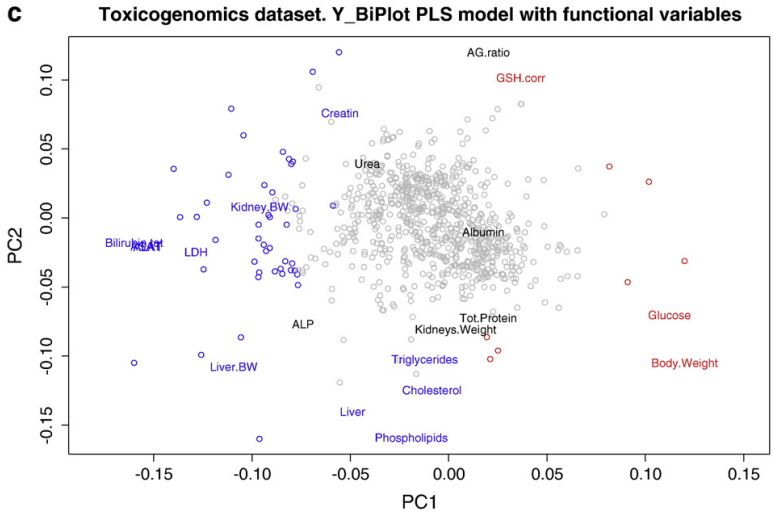


Fig. 2 (*continued*).

structures, as well as the explained variability within both data matrices.

The PLS models the data through the use of the following expressions,

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^{\mathrm{T}} + \mathbf{F}$$

where **X** and **Y** are the matrices of functional and physiological variables, respectively. **T** is the matrix that maximizes the covariance between **X** and **Y**, **P** the loading matrix for **X**, **C** the loading matrix for **Y**, **W** and **W*** are weighting matrices that indicate the importance of each functional variable in the new projected space, and **E** and **F** the residual matrices for **X** and **Y**, respectively.

Each component of the PLS model represents a pattern of variation that relates independent and dependent variables. Therefore, by analyzing the weights of functional and response variables in the PLS model we can identify gene function features that are associated with the observed phenotype. The significance of the PLS model is habitually given by the R^2 and Q^2 statistics, which indicate respectively the explanatory and predictive power of the model.

The R^2 is defined as the fraction of the total sum of squares which is captured by the model. For a model with F components,

$$\mathrm{R}^2 = \frac{\mathrm{SSM_F}}{\mathrm{SST}},$$

where SSM is the sum of squares of the model with F components and SST is the total sum of squares.

Q^2 parameter is given by

$$Q^2_{cum}(F) = 1 - \frac{\mathrm{PRESS}(F)}{\mathrm{SST}}$$

$$\mathrm{PRESS}_F = \sum_{i=1}^{I} r_i^2,$$

which indicates the sum of squares of the prediction errors "r" for the observations not included in the model during the cross-validation procedure.

Furthermore, the importance of each functional variable in the model can be computed by the VIP parameter, which is the sum of the contributions of the variable to the model components moderated by the weight of the component. This VIP parameter computes the influence on **Y** of every term x_k in the model, according to

$$\mathrm{VIP}_{FK} = \sqrt{\sum_{f=1}^{F} \left(w_{fk}^2 * \left(\mathrm{SSY}_{f-1} - \mathrm{SSY}_f\right)\right) * \frac{K}{\mathrm{SSY}_0 - \mathrm{SSY}_F}}.$$

Finally, we include a permutation test to determine the probability of the computed model parameters to occur by chance and to select significant functional terms. This permutation is performed on the original data matrix and therefore affects the steps of generation (PCA) and selection (PLS) of functional variables.

Datasets

We have applied the proposed method to two different datasets

The first dataset corresponds to a toxicogenomics study in which the effect of bromobenzene in liver toxicity in rats is analyzed [27]. In this experiment, rats are administrated the drug bromobenzene at three different doses (high, medium, and low) and liver/blood/urine samples are taken after 6, 24, and 48 h of treatment. There are control (no administration) and placebo (only drug vehicle administration) rat groups. For each experimental condition one to three rats were taken for gene expression profiling and microarray experiments were done with a dye-swap design on a custom cDNA microarray. Gene expression information is available for 2665 genes. Additionally, physiological and morphological determinations were conducted on the same rats, including body weight (g), kidneys weight (g), kidney/BW (g/kg), liver (g), liver/BW, bilirubin tot, ASAT, ALAT, LDH, albumin g/L, ALP (U/L), creatin umol/L, cholesterol (mmol/L), glucose (mmol/L), phospholipids (mmol/L), triglycerides (mmol/L), tot.protein (g/L), urea (mmol/L, A/G ratio, GSH corr. (M) [27].

The second dataset is a breast cancer study by Miller and co-workers [28]. This work explores the relationship between the p53 (TP53) pathway and breast tumor severity. The Affymetrix U133 A and B human GeneChips (~25,000 probes) were used to assess the genome-wide transcriptome profile of 251 primary invasive breast tumors for which detailed information on p53 status (p53+, mutant;

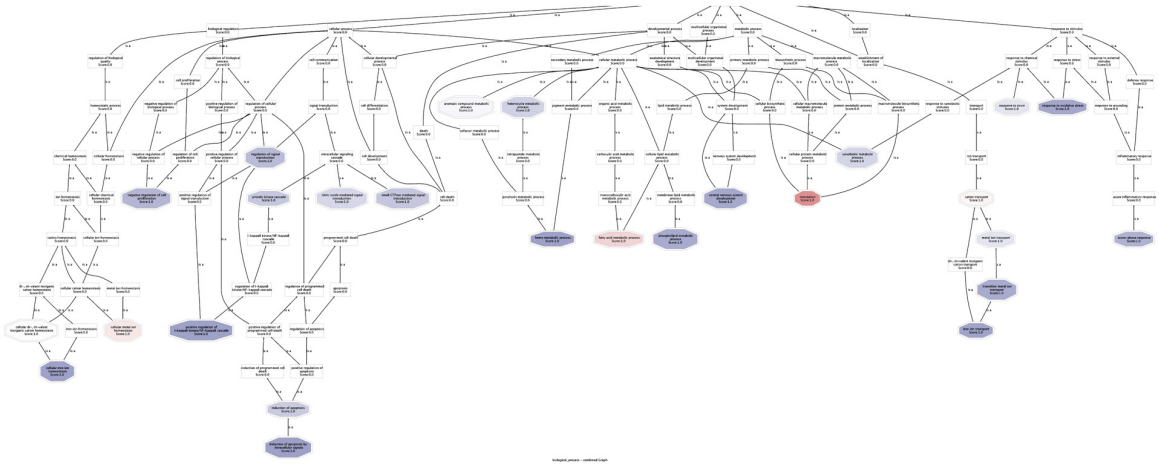


Fig. 3. Gene Ontology Direct Acyclic graph of the pool of Biological Process significant terms detected by the PLS model on functional variables. Term color intensity is proportional to the importance of the functional class in the PLS model. Hexagonal nodes are the actual selected GO terms. For a fuller view of this figure, please see Appendix A.

p53-, wt) was available. Additionally tumors were characterized by their estrogen-receptor (ER) status, Elston histological grade, PgR status, age at diagnosis, tumor size (mm), lymph node status, DSS TIME (disease-specific survival time in years), and DSS EVENT (disease-specific survival event; 1=death from breast cancer, 0=alive or censored).

These two datasets represent two analysis scenarios. The toxicogenomics dataset contains a multifactorial experimental design with strong gene expression signals associated to the treatments. A relative low number of genes and a wide array of response variables are present. The breast cancer dataset illustrates a typical cancer study with a large number of cases and a genome-wide transcriptomics profiling. A few clinical parameters were evaluated for each patient and gene expression signals are expected to be more diluted.

Data preprocessing and analysis

The toxicogenomics dataset was obtained directly from the authors, normalized by lowess, and centered genewise for each dye-swap pair as in [37]. Breast Cancer Affymetrix data were downloaded from the GEO database as global mean normalized data. Physiological/clinical variables were scaled in all cases and missing values were imputed by the *k*th nearest neighbors algorithm [38]. Gene Ontology functional annotations were obtained from public repositories. Annotated Gene Ontology DAG structures were generated with the Blast2GO software [31]. Noninformative reference distributions for the toxicogenomics and breast cancer dataset were generated by bootstrap. One thousand bootstrap runs were executed, in each case resampling both column- (samples) and row-wise (genes). Resampling by columns

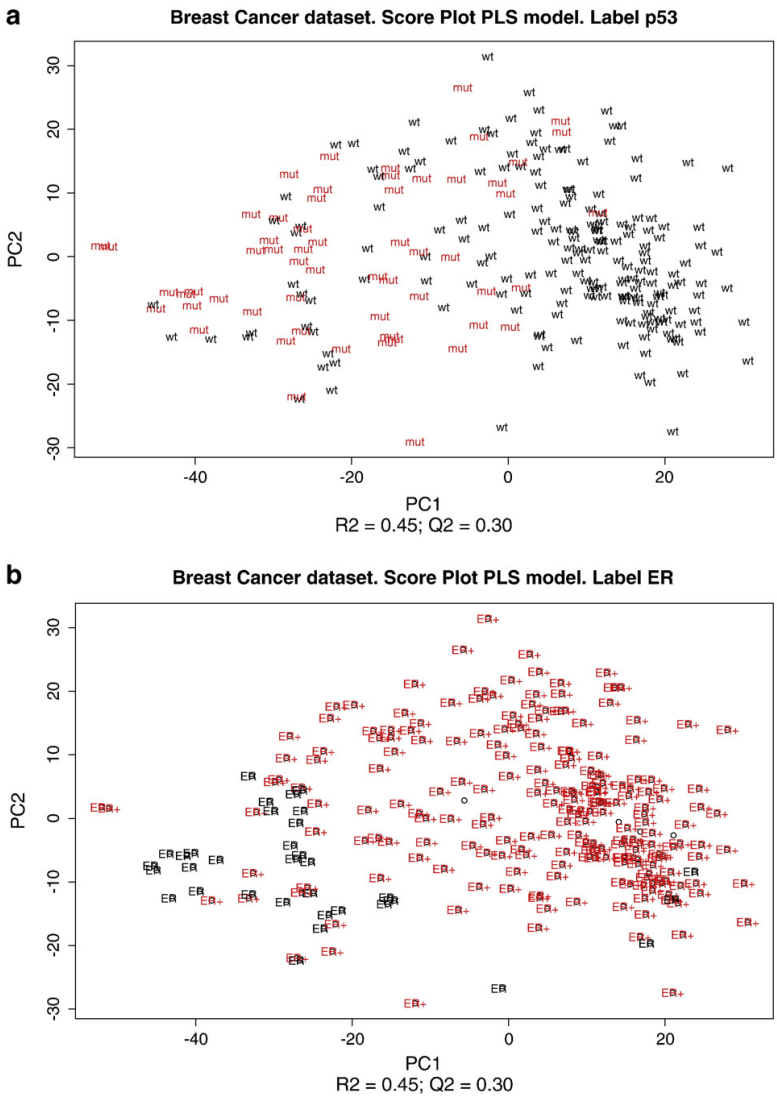


Fig. 4. X_score plot PLS model for breast cancer data. PLS model computed with functional variables. Tumor samples are labeled either for their p53 genotype (a) or ER status (b).

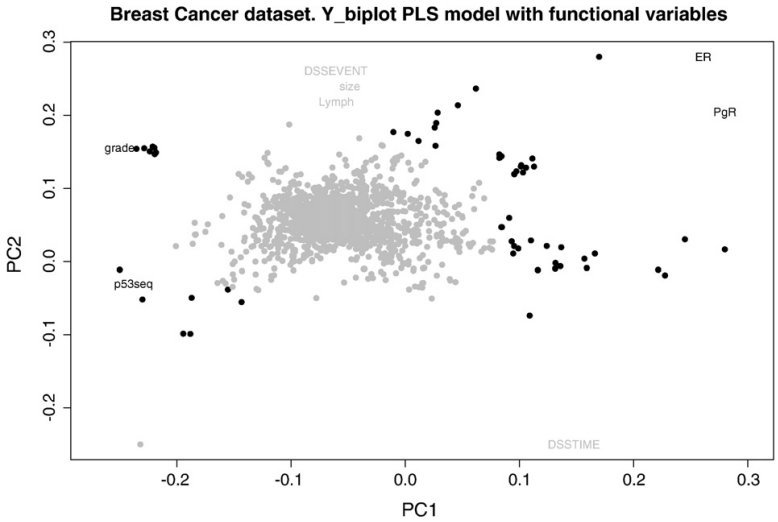


Fig. 5. Breast Cancer PLS **Y**_biplot. Projection of functional and clinical parameters into the first two components of the PLS model. Variables poorly explained by the model given in gray. Functional variables are represented by dots and colored dark when significant.

eliminates the relationship between the gene expression and the phenotype, while rearrangements by rows will destroy the coordinative structures within each functional class. The *P* value corresponding to the PLS model parameters (R^2 and Q^2) and the importance of functional variables (VIP) were computed as the frequency of occurrence of true data values in the respective reference null distributions. Significance threshold was set to 0.05.

All computations were performed in *R*, using limma [3], pls [39] and EMV packages. Scripts are available on request to the authors.

Acknowledgments

This work was funded by the Spanish Ramon y Cajal Program, grants from the Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER) ISCIII, and grant BIO2005-01078 from the Spanish Ministry of Education and Science. Publication costs were granted by the National Institute of Bioinformatics (www.inab.org) a platform of Genoma España.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.05.015.

References

[1] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. USA 98 (2001) 5116–5121.
[2] M.K. Kerr, M. Martin, G.A. Churchill, Analysis of variance for gene expression microarray data, J. Comput. Biol. 7 (2000) 819–837.
[3] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004) 3.
[4] T. Speed, Statistical Analysis of Gene Expression Microarray Data, Chapman and Hall/CRC, London, 2003.
[5] F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo, Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information, Bioinformatics 21 (2005) 2988–2993.
[6] J. Dopazo, Functional interpretation of microarray experiments, Omics 10 (2006) 398–410.
[7] I. Rivals, L. Personnaz, L. Taing, M.C. Potier, Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23 (2007) 401–407.
[8] S.B. Pounds, Estimation and control of multiple testing error rates for microarray studies, Brief Bioinform. 7 (2006) 25–36.
[9] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[10] A. Smilde, R. Bro, P. Geladi, Multiway Analysis: Applications to the Chemical Sciences, Wiley, Chichester, England, 2004.
[11] R. Kustra, R. Shioda, M. Zhu, A factor analysis model for functional genomics, BMC Bioinform. 7 (2006) 216.
[12] P. Carmona-Saez, M. Chagoyen, F. Tirado F, J.M. Carazo, A. Pascual-Montano, GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists, Genome Biol. 8 (2007) R3.
[13] Y. Lu, P.Y. Liu, P. Xiao, H.W. Deng, Hotelling's T2 multivariate profiling for detecting differential expression in microarrays, Bioinformatics 21 (2005) 3105–3113.
[14] J. Landgrebe, W. Wolfgang, G. Welzl, Permutation-validated principal components analysis of microarray data, Genome Biol. 3 (2002) 19.1-19.1.
[15] M.J. Nueda, A. Conesa, J.A. Westerhuis, H.C. Hoefsloot, A.K. Smilde, M. Taln, A. Ferrer, Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA, Bioinformatics 23 (2007) 1792–1800.
[16] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (2005) 15545–15550.
[17] F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Minguez, D. Montaner, J. Dopazo, From genes to functional classes in the study of biological systems, BMC Bioinform. 8 (2007) 114.
[18] S.W. Kong, W.T. Pu, P.J. Park, A multivariate approach for integrating genome-wide expression data and biological knowledge, Bioinformatics 22 (2006) 2373–2380.
[19] D. Nettleton, J. Recknor, J.M. Reecy. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis, Bioinformatics 24 (2008) 192–201.
[20] P. Mielke Jr, K. Berry, Permutation methods: a distance function approach, Springer-Verlag, New York, 2001.
[21] H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin, P. Pavlidis, Coexpression analysis of human genes across many microarray data sets, BMC Bioinformatics 25 (5) (2004) 18.
[22] D.J. Allocco, I.S. Kohane, A.J. Butte, Quantifying the relationship between co-expression, co-regulation and gene function, Genome Res. 14 (2004) 1085–1094.
[23] F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D.K. Thompson, J. Zhou, Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory, BMC Bioinform (8) (2007) 299.
[24] T.M. Murali, C.J. Wu, S. Kasif. The art of gene function prediction. Nat. Biotechnol. 24 (2006) 1474–1475.
[25] T.O. Khor, Toxicogenomics in drug discovery and drug development: potential applications and future challenges, Pharm. Res. 23 (2006) 1659–1664.
[26] R. Clarke, H.W. Ressom, A. Wang, J. Xuan, M.C. Liu, E.A. Gehan, Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nat. Rev. Cancer 8 (2008) 37–49.
[27] W.H. Heijne, R.H. Stierum, M. Slijper, P.J. van Bladeren, B. van Ommen, Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach, Biochem. Pharmacol. 65 (2003) 857–875.
[28] L.D. Miller, J. Smeds, J. George, V.B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E.T. Liu, J. Bergh, An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival, Proc. Natl. Acad. Sci. USA 102 (2005) 13550–13555.

[29] G. Cattoretti, F. Rilke, S. Andreola, L. D'Amato, D. Delia, P53 expression in breast cancer, Int. J. Cancer 41 (1988) 178–183.
[30] N. Blthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel, D. Beule, Biological profiling of gene groups utilizing Gene Ontology, Genome Inform. 16 (2005) 106–115.
[31] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing Value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.
[32] A. Fagan, A.C. Culhane, D.G. Higgins, A multivariate analysis approach to the integration of proteomic and gene expression data, Proteomics 7 (2007) 2162–2171.
[33] P. Wei, W. Pan, Incorporating gene networks into statistical tests for genomics data via spatially correlated mixture model, Bioinformatics 24 (2008) 404–411.
[34] S. Matsui, M. Ito, H. Nishiyama, H. Uno, H. Kotani, J. Watanabe, P. Guilford, A. Reeve, M. Fukushima, O. Ogawa, Genomic characterization of multiple clinical phenotypes of cancer using multivariate linear regression models, Bioinformatics 15 (2007) 732–738.
[35] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis, UMETRICS AB, Umea, 2001.
[36] K. Fellenberg, C.H. Busold, O. Witt, A. Bauer, B. Beckmann, N.C. Hauser, M. Frohme, S. Winter, J. Dippon, J.D. Hoheisel. Systematic interpretation of microarray data using experiment annotations, BMC Genomics 7 (2006) 319.
[37] A. Conesa, M.J. Nueda, A. Ferrer, M. Talon, maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments, Bioinformatics 22 (2006) 1096–1102.
[38] A. Conesa, S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676.
[39] B.H. Mevik, R. Wehrens, The pls Package: Principal Component and Partial Least Squares Regression in R, J. Stat. Software 18 (2007) 1–24.

PAPER IV

Metabolomics
DOI 10.1007/s11306-013-0560-7

ORIGINAL ARTICLE

No genetic footprints of the fat mass and obesity associated (FTO) gene in human plasma ^{1}H CPMG NMR metabolic profiles

K. Kjeldahl · M. A. Rasmussen · A. L. Hasselbalch · K. O. Kyvik · L. Christiansen · S. Rezzi · S. Kochhar · T. I. A. Sørensen · R. Bro

Received: 25 January 2013 / Accepted: 14 June 2013

Abstract In this paper it was investigated if any genotypic footprints from the fat mass and obesity associated (FTO) SNP could be found in 600 MHz ^{1}H CPMG NMR profiles of around 1,000 human plasma samples from healthy Danish twins. The problem was addressed with a combination of univariate and multivariate methods. The NMR data was substantially compressed using principal component analysis or multivariate curve resolution-alternating least squares with focus on chemically meaningful feature selection reflecting the nature of chemical signals in an NMR spectrum. The possible existence of an *FTO* signature in the plasma samples was investigated at the subject level using supervised multivariate classification in the form of extended canonical variate analysis, classification tree modeling and Lasso (L1) regularized linear logistic regression model (GLMNET). Univariate hypothesis testing of peak intensities was used to explore the genotypic effect on the plasma at the population level. The multivariate classification approaches indicated poor discriminative power of the metabolic profiles whereas univariate hypothesis testing provided seven spectral regions with $p < 0.05$. Applying false discovery rate control, no reliable markers could be identified, which was confirmed by test set validation. We conclude that it is very unlikely that an *FTO*-correlated signal can be identified in these ^{1}H CPMG NMR plasma metabolic profiles and speculate that high-throughput untargeted genotype-metabolic correlations will in many cases be a difficult path to follow.

Keywords FTO · NMR · CPMG · Data compression · ECVA · MCR-ALS

K. Kjeldahl · M. A. Rasmussen (✉) · R. Bro
Univ Copenhagen, 1958 Frederiksberg C, Denmark
e-mail: mortenr@life.ku.dk

A. L. Hasselbalch · T. I. A. Sørensen
Institute of Preventive Medicine, Frederiksberg and Bispebjerg University Hospital, Capital Region, Copenhagen, Denmark

K. O. Kyvik · L. Christiansen
Institute of Regional Health Services Research, University of Southern Denmark, Winsløwparken 19, 3, 5000 Odense C, Denmark

S. Rezzi · S. Kochhar
Nestlé Research Center, PO Box 44, Vers-chez-les-Blanc, 1000 Lausanne 26, Switzerland

T. I. A. Sørensen
Novo Nordisk Centre for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Capital Region, Copenhagen, Denmark

1 Introduction

The quantitative genetic contribution to body mass index variation and hence to obesity is well established (Walley et al. 2009; Yang et al. 2007) and has been confirmed in the population used for the present study (Schousboe et al. 2003). Many obesity candidate genes have been discovered, among which the most consistent associations have been found between body weight and single nucleotide polymorphisms (SNP) in the fat mass and obesity associated *FTO* gene (Frayling et al. 2007; Jess et al. 2008; Kring et al. 2008), confirmed in a recent meta-analysis (Peng et al. 2011).

The *FTO* gene is located on chromosome 16. To date, the locus with the strongest association with obesity is the rs9939609 (Walley et al. 2009), which is one of a cluster of several SNPs located in the first intron of the gene. The verification of the association between the *FTO* gene and

Published online: 02 July 2013

body mass strongly supports the suggestion that this gene has common variants (the AA and AT genotypes) that predispose to obesity, relative to the wild (TT) genotype (Peng et al. 2011). Varying effect sizes of the *FTO* locus have been reported but a number of studies states an effect size of about 30 % increase in the risk of obesity or an average of 0.35 kg/m^2 (0.1 *z*-score units for BMI) per susceptibility allele (Peng et al. 2011).

Expression studies indicate that *FTO* is widely expressed in many tissues, but has its highest expression in the brain, particularly the arcuate nucleus of the hypothalamus (Frayling et al. 2007), where it is believed to be involved in energy uptake rather than energy expenditure (Berentzen et al. 2008, Haupt et al. 2009). The findings of Wardle (Wardle et al. 2008) that the *FTO* risk allele was associated with reduced satiety responsiveness in children support this putative functional role, whereas another study (Hasselbalch et al.2010) found no association between *FTO* and increased energy intake or food preferences. The study by Wahlen et al. (2008) indicates a role of *FTO* in fat cell lipolysis, and a study by Pitman et al (2012) indicates an important role in cellular energy balance. A meta-analysis by Wang et al. (2012) suggests that several SNPs in *FTO* is associated with the metabolic syndrome.

Gerken et al. (2007) suggested that *FTO* catalyzes demethylation of 3-methylthymine in DNA, with concomitant production of succinate, formaldehyde, and carbon dioxide, but a direct functional role of the *FTO* gene in obesity development remains unsolved. It is likely that the *FTO* variants are in linkage disequilibrium with the true causative variant (Saunders et al. 2007).

Since the mechanism of how the *FTO* variants influence the size of the body and especially the fat mass is essentially unknown, we undertook an exploratory analysis of the possible association between these gene variants and the metabolomic profile in blood. As part of the Danish nationwide GEMINAKAR study which took part 1997–2000 fasting blood samples were collected from healthy Danish twins and analyzed for a number of constituents. Genotyping with respect to the *FTO* locus rs9939609 was also conducted for a subset of the twins, and at a later stage, Nestlé Research Centre (Lausanne, Switzerland) subjected the plasma to ^{1}H CPMG NMR analysis for metabolomics studies (Peré-Trepat et al. 2010).

The purpose of the present study was to combine these two datasets and investigate whether the *FTO* (rs9939609) genotype is reflected in the blood composition as measured with 600 MHz ^{1}H CPMG NMR.

In this paper, we attempted to identify metabolic associations between *FTO* polymorphism and metabolic signatures through the interrogation of ^{1}H CPMG NMR generated metabolic profiles with a combination of multivariate and univariate statistics.

2 Materials and Methods

The GEMINAKAR study regarded the relative influence of environmental and genetic factors on especially the metabolic syndrome, and was based on data from 756 healthy Danish twin pairs. The details of the GEMINAKAR study are described elsewhere (Benyamin et al. 2007; Schousboe et al. 2003). The blood samples were collected during the years 1997–2000, added NaF as an anti-coagulant and preservative and stored at −80°C until 2005 when the plasma was subjected to NMR analysis, whereby individual fasting metabolic profiles were obtained. SNP data was obtained from the same blood samples. Combined NMR and SNP data was available for 1116 individuals.

Data handling and analysis was performed using the commercial software package MATLAB®, ver. 7.10.0 including PLS_Toolbox ver. 5.5.1.

2.1 SNP data

The *FTO* SNP (rs9939609) was genotyped by conducting allelic discrimination using pre-designed Taqman® SNP genotyping assays (Applied Biosystems). The conditions described by the manufacturer were applied. PCR was performed in the ABI Prism 7700 and analyzed using the Sequence Detection System software (Applied Biosystems). The distribution of the three *FTO* genotypes is shown in Table 1.

2.2 NMR acquisition and preprocessing

Metabolic profiles of fasting blood plasma were measured at 298 K on a Bruker DRX 600 MHz spectrometer equipped with a 5 mm probe (Bruker Biospin, Rheinstetten, Germany). ^{1}H NMR spectra were registered for each blood plasma sample using a standard Carr–Purcell–Meiboom–Gill (CPMG) spin echo pulse sequence with water suppression. For each sample, 256 scans were collected into 32 K data points using a spectral width of 13.9790 Hz, corresponding to an acquisition time of 1.95 s.

Table 1 Distribution of *FTO* genotypes in this study

Genotype	Number of subjects	(%)
TT	349	34
AT	482	47
AA	197	19

Prior to Fourier transformation, NMR data were multiplied by an exponential weighting function corresponding to a line broadening of 1.0 Hz. The acquired NMR spectra were manually corrected for phase and baseline distortions, and referenced to the chemical shift of the alpha-glucose doublet peak at 5.23 ppm using TOPSPIN software (version 2.1, Bruker Biospin, Rheinstetten, Germany). The CPMG plasma spectra included 22 K data points over the range of δ 0.1 to 8.25 (Rezzi et al. 2007). CPMG NMR acquisition is more suited for the detection of small molecules by providing an attenuation of broad signals arising from macromolecular species (proteins and lipoproteins mainly). Interpolation of all the spectra to the same chemical shift was performed. The water peak at δ 4.55–5.17 ppm was removed and NMR spectra normalized to a constant total sum of all intensities within the specified range prior to the chemometrics analysis. 85 spectra out of 1,116 were excluded due to poor data quality or obvious outlying sample nature such as the presence of e.g. ethanol or drugs.

The very high number of variables represents a computational challenge, but may potentially also hamper the data analysis adversely by substantial power reduction and by presence of spurious correlations in data. It is therefore desirable to reduce the data in a reasonable way. An integration approach was applied in the following way:

1. Define K spectral regions, if possible so that each region contains only one peak. Omit peaks which one would not trust if they turn out to be significant, i.e. peaks with very weak intensities or regions with many overlapping small signals. Obvious spin-spin coupling splits (i.e. doublets, triplets etc), with no interfering peaks should be put in same region, but emphasis is rather on representing all informative signals than taking care of structural information. For each peak region, j:
   (a) Align peaks using the *i*coshift tool (Savorani et al. 2010). *i*coshift shifts the peaks of each region individually and preserves the peak shape. Possibly apply Savitzky–Golay differentiation to the spectra prior to alignment for improved alignment and apply the resulting spectral adjustment to the original spectra.
   (b) If the peak is a small signal on a shoulder of large broad parent peak, model the parent peak as baseline and subtract this from the small peak.
   (c) Decompose region by Principal Component Analysis (PCA) (Hotelling 1933) or Multivariate Curve Resolution (MCR) (Tauler and Barceló 1993) using a reasonable number of components $l, l \in \{1, 2\}$ determined manually for each region. For improved robustness, include only the approx. 80 % most normal samples as determined by initial PCA submodels for each class (AA, AT, TT) for the modeling, but eventually project all samples onto the regional PCA or MCR model. The choice of method (PCA or MCR) was made by comparison of the spectra with the obtained loadings; the shape of the loadings should ideally only model the peak of interest in the region and not artefacts arising from interfering signals. After this step, each region is represented by l score variables.
2. Collect scores to form the new matrix $\mathbf{X}$.

Eventually, three extreme outliers were removed by use of PCA diagnostics from a model on $\mathbf{X}$, resulting in a dataset consisting of 1,028 samples and 171 variables originating from 164 peak regions.

2.3 Data analysis

Immediately after the preprocessing, the dataset was split in a training set of 800 samples and a test set of 228 samples by Kennard–Stone subset selection (Daszykowski et al. 2002; Kennard and Stone 1969) with the constraint that (twin) siblings should be put in the same set. The two datasets had comparable composition with respect to *FTO* classes.

2.4 Distribution of variation

The distribution of variation sources was estimated for each variable using simple analysis of variance (ANOVA), with gender as categorical and age as linear effects. For the total variable space simple means of the contributions to the individual variables were used.

2.4.1 Data quality

An initial confirmation of data quality of both full spectra and the compressed data was performed by (a) modeling of age and (b) visual verification that multivariate gender differences were present. Age was modeled by Partial Least Squares (PLS) regression and gender differences were visualized via PCA. Where PCA captures main variation that in turn also relates to gender differences, it does not clearly elucidate signals in relation to age. PLS was applied to unravel patterns related to this external information.

2.5 Modeling

The relation between the metabolomic and the genotypic profiles was investigated both uni- and multivariately on the basis of the compressed variables:

1. Multivariate classification.
2. Univariate significance testing.

2.5.1 Multivariate modeling

Multivariate classification was performed using extended canonical variate analysis (ECVA) (Nørgaard et al. 2006). Furthermore a classification tree model (CART) (Breiman et al. 1984) and a Lasso (L1) regularized linear logistic regression model (GLMNET) (Friedman et al. 2007, 2010) were fit. The Lasso approach represents a linear method which incorporates variable selection, i.e. this approach will be successful if the signal is only in a minor part of the variables. The CART approach is assumption free in the sense that non-linearity and even non-monotonicity and interactions can be captured by this approach, however, with the price of having high variance. In addition, ECVA is a linear method in a truncated subspace. All of these methods represent different strategies where the chance of success depends on the distribution of the true signal, i.e linear or non-linear, latent variables, few variables, interactions etc. The Lasso and CART models were not able to describe the validation data; it is therefore concluded that these models do not describe a true systematic variation in data, hence they are not meaningful and results will not be shown for these.

The ECVA modeling aims to separate the two classes by seeking new directions which separate the variation between the classes relative to the variation within the classes. This is a supervised method, and consequently there is a risk of over-fitting in the sense that optimistically good separation may be found. For the ECVA modeling, AA and TT samples were included to simplify the problem to a two-class problem with only the most extreme groups. The apparent drawback of this is the reduced sample size. Consequently, the experiment was repeated observing AA–AT versus TT and AA versus AT–TT. This extension confirmed conclusions from the simplified approach, therefore only the AA versus TT results are shown in this paper.

A major advantage of the applied data reduction is that it produces essentially noise free data, and thus up-weighting of small peaks is not associated with noise boosting. Consequently, data for the ECVA models was autoscaled, i.e. each variable was mean-centered and scaled to unit variance. Prior to this, the compressed spectral data was adjusted for age and gender variation as described below.

The performance of the ECVA modeling was assessed by a combination of cross-validation and permutation test. Specificity and sensitivity were first assessed by cross-validation where the samples were split randomly into 16 segments where each segment conserved the class distribution. This was repeated 200 times, whereby 200 (specificity, sensitivity)-pairs were obtained. To assesses if the obtained performance was better than random, modeling and cross-validation were repeated 1,000 times with a shuffled class membership.

2.5.2 Univariate significance

Each of the 164 individual spectral regions was tested in parallel. A test statistic, p_j $(j = 1, .., 164)$, was obtained for each spectral region by comparing two nested generalized linear models predicting the occurrence of A in the *FTO* locus via logistical regression.

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_{ij}\beta + \epsilon_{ij} \quad (1)$$

with a pure intercept spectral independent model:

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \epsilon_i \quad (2)$$

where p_i refers to the probability of locus A for the i'th person $(i = 1, \ldots, n)$. x_{ij} is the derived scores for person i spectral region j corrected for age and gender information (see handling of covariates below).

The deviance $(-2{*}log$ likelihood) between the two models $dev_{0j} - dev_{1j}$ is assumed $\chi^2(df_j)$ with $df_j = df_0 - df_{1j}$.

Thus a significance test was produced for each spectral region $(p_1, .., p_{164})$, such that the minimum p value refers to the most interesting spectral region. The 164 spectral regions were split into a set of significant regions and a set of non significant regions using the method of Benjamini and Hochberg (1995) for control of false discovery rate (FDR).

The selected spectral regions suffer from selection bias, that is; the most significant regions are selected due to a combination of true effect size, but also by chance. This bias is known as winner's curse (Zöllner and Pritchard 2007). The spectral regions were investigated for selection consistency by applying a non parametric bootstrap procedure evaluating the frequency of selection for individual regions at different FDR settings.

2.5.3 Handling of covariates

The plasma profiles systematically reflect gender and age, hence a priori correction for those was appropriate. Let **D** be a design matrix corresponding to gender and age, such that the first and second column are dummy vectors for male and female respectively and the third column is age. Linear correction for both age and gender can be done by orthogonalization with respect to **D**:

$$\mathbf{X}_{corr} = \left(\mathbf{I} - \mathbf{D}(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}\right)\mathbf{X} \quad (3)$$

It is assumed that the *FTO* is independent of gender and age. Nevertheless the collected data might exhibit a small partial confounding with either gender or age. Under such circumstances (3) is an overcorrection, and subsequent models will suffer from that. In order to remove information *only* related to gender and age, while retaining all information related to *FTO*, the design matrix **D** is projected onto the null space of *FTO*. Let **F** be an *FTO* design matrix with three columns corresponding to TT, AT and AA. For the ECVA models only TT and AA columns were included.

$$\mathbf{D}^{*} = (\mathbf{I} - \mathbf{F}(\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}})\mathbf{D} \quad (4)$$

$\mathbf{D}^{*}$ is then used for correction of **X** in Eq. (3).

3 Results

3.1 Data source

PLS models of subject age based on (a) raw spectra (without water peak) and (b) compressed data showed good performance in both cases. Fig. 1 shows the performance of the model with the compressed data, similar results were found for the full-spectrum model. Model complexity was estimated using six-fold cross-validation on the training set in both models, and performance was assessed using the test set. The two models showed comparable root mean squared error of prediction (RMSEP) at 6.5 and 6.8 years respectively.

Gender differences could also be recognized using PCA for both data sets. Figure 2 shows a scores plot for the compressed data set, where it is obvious that gender differences are found in this data set. PCA is an unsupervised method, i.e the model has not been "'requested"' to model gender differences, rather it extracts underlying features. The largest variations manifest in the first principal components, and in this case, gender differences were best revealed in the combination of PC2 versus PC6.

Examination of the distribution of variation sources across all variables reveals that both age and gender at most contribute with 1 % variance for single variables, and 0.1 % across all variables (results not shown). The performance of a model without correction of age and gender are practically identical, and hence omitted.

3.2 Multivariate classification

ECVA classification models were built to separate the two extreme genotypes (TT, AA). Cross-validation estimated a mean sensitivity of 0.42 and specificity of 0.70. These values are quite poor, and it is not obvious whether they are better than random. The permutation test (Fig. 3) showed that the obtained performance appear to be slightly better than random, although not very robust.

3.3 Univariate significance

Each of the 164 spectral regions was tested one by one, resulting in p values ranging from 0.028 up to 0.991. Correction for multiple testing by FDR control by Benjamini and Hochberg at level 0.05 leads to a significance threshold of 0.0003. No variables survived this level.

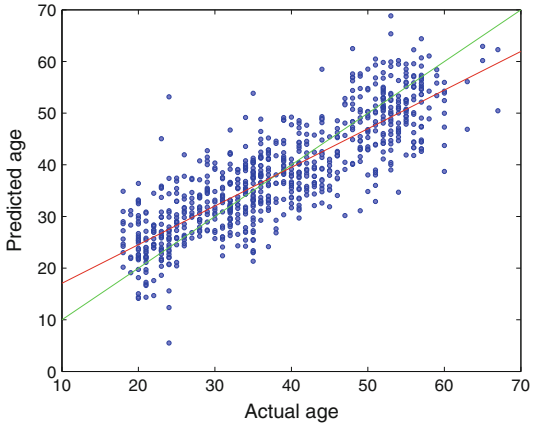


Fig. 1 Prediction of subject age using compressed data. *Green line* ideal prediction fit, *red line* obtained cross-validation prediction fit. RMSEP = 6.8 years, $R^2 = 0.77$)

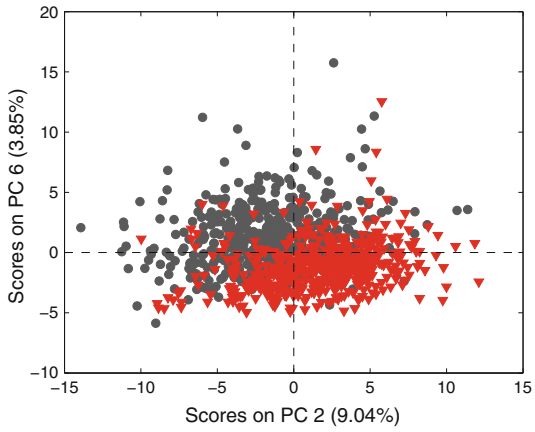


Fig. 2 Gender differences are present in compressed data. Visualized by PCA score plot (PC2 vs. PC6)

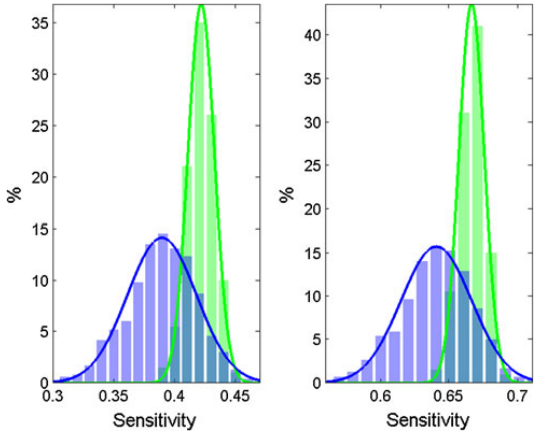


Fig. 3 Sensitivity (*left*) and specificity (*right*) obtained by multivariate ECVA models (*green*), compared to permutation test (*blue*)

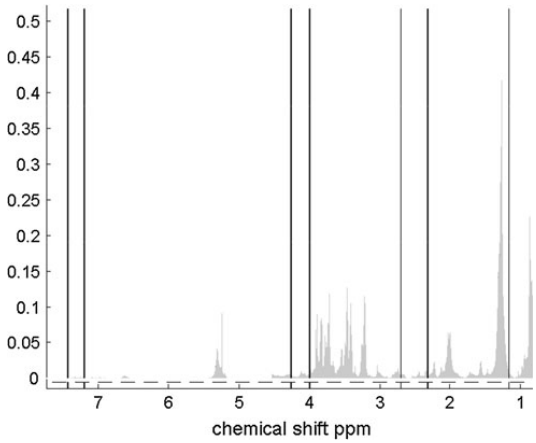


Fig. 4 NMR regions with $p \leq 0.05$

Ignoring the multiple testing correction, seven regions were found to have $p \leq 0.05$. The positions of these in the NMR spectrum are shown in Fig. 4. The p values and the FDR control were bootstrapped and by this procedure, the frequency of "winning" (i.e. being selected as significant) was assessed for each of the variables. The seven variables were selected as significant at a frequency between 44 and 56 %, which is not overwhelming, but considerably higher than the average of 20 %.

The validity of the seven regions was evaluated with the test set of 228 samples. The mean values of each of the seven regions are shown in Fig. 5, and it is obvious that the results from the training set are not valid for the test set. The result of region 44 is the only one where effect size and pattern across genotypes are comparable between the

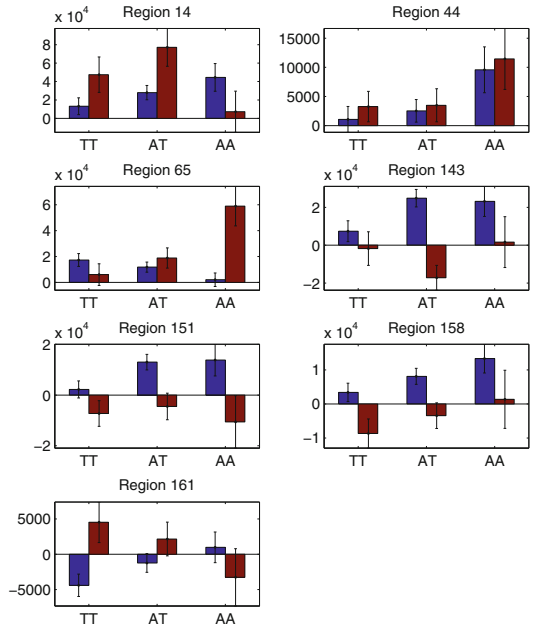


Fig. 5 Mean values (with standard errors of the mean) of the seven regions with $p \leq 0.05$ for adjusted training (*blue*) and test set (*red*) data

training and the test set. In fact region 65, 151 and 161 display completely opposite effect for the two sets. Closer investigation of the distribution of intensities of region 44 showed that differences were so small that they were of no value for any purpose (not shown).

As a final attempt, the seven selected regions were used as input to a multivariate ECVA model following the same procedure as described above. The resulting model is expected to give too optimistic results for the training performance relative to validation performance because the seven variables have been pre-selected to be the most discriminatory for the training set. AUC for the associated training ROC curve was 0.59, which is a very poor classifier. Both test set validation and permutation test confirmed that the combination of the seven selected variables could not be associated with FTO genotype (not shown).

4 Discussion

Possible markers of the *FTO* gene in NMR based plasma profiles could be interesting candidates for further investigations of how the *FTO* influences body mass and fat mass. The results show that we have not been able to identify reliable plasma markers of the *FTO* genotype when we took care of multiple testing by Benjamini–

Hochberg FDR control. The use of a test set illustrated that candidate markers with $p < 0.05$ had very little validity. This underlines that multiple testing situations require careful handling of p values and that the use of test sets is recommendable.

Does that mean that the *FTO* gene does not leave any footprint in a blood sample? There may be several reasons why we did not make any positive findings in this study. We shall here address (a) the quality of the data, (b) the data analysis and (c) the overall research questions.

The data quality check indicated that the substantial data compression from 29,027 to 171 variables did not result in massive information loss since the information in the compressed data was adequate to model a highly complex attribute such as age, which was found to contribute with less than 1 % of the variation in the 171 variables. One could consider to include more than 164 spectral regions as many more are definitely present, but visual inspection of the remaining spectral regions made us very uncertain about concluding anything about these as most of them were close to the noise limit or weak shoulder peaks difficult to align etc.

Low sensitivity is an intrinsic problem in NMR and as a result only around hundred metabolites give signal in the NMR spectrum, which is about 10 % of the total metabolome (Viant et al. 2008). On top of this, the CPMG pulse sequence filters the signals by suppressing signals from macromolecules. Lipo-proteins are thus suppressed in these spectra, which potentially could be an important loss of information. It might thus have been useful to combine the CPMG recordings with e.g Overhauser enhancement spectroscopy (NOESY) recordings which provide a good overview of all the types of molecules present in the sample matrix (Beckonert et al. 2007). LC-MS is an alternative platform with much higher sensitivity but this technique certainly represents other challenges. In any case recording of more information consequently lowers the statistical power.

In the data analysis, four (three multivariate and one univariate) classification paths were investigated. The ECVA modeling investigates multivariate solutions where all variables are active. The LASSO search path recovers sparse solutions and it is hence the assumption that a combination of a subset of the variables has discriminatory power. Classification trees are scale invariant and superiority of such a model relies on non linearity compared to the linear ECVA and LASSO models. If the different regions are independent, multiple univariate tests sorts the regions in terms of association with *FTO*. The different models reveal different representations of the biological system and this exhaustive search confirms lack of consistent information opposed to wrong modeling choices.

In this study we searched for any kind of metabolic response in a plasma sample correlated with the *FTO* genotype. The metabolic profile is a result of a very complex network of interactions between genetic and environmental factors and at system level there is a long path from genotype to metabolic profile. Epistatic effects, either at the genomic or phenotypic level may mask the signal, interactions with the environment and the fact that the whole homeostatic system is very robust shrink the correlation between genotype and metabolic profile.

In particular, we have previously investigated the influence of dietary preferences on the ^{1}H CPMG NMR-based metabolic profiles of blood plasma (Peré-Trepat et al. 2010). The dual nature of the information generated from blood plasma compositional analysis reflects individual metabolic adaption to specific lifestyle, as well as the results of tight homeostatic regulatory processes. We described that five major dietary patterns were significantly reflected in the biochemical composition of blood plasma, as per variations in the concentration of circulating lipids and amino acids. This source of metabolic variation is a major contributor to inter-individual metabolic differences, which often represent a larger source of variance when compared to the one associated to gender and age. These metabonomic applications therefore illustrate the complexity of associating specific genotypes to metabolic profiles of a systemic biofluid. Also, the relative lack of sensitivity of NMR spectroscopy combined with the use of acquisition conditions depleting the signals from macromolecules, can be considered as an additional challenge to identify very subtle metabolic differences associated with specific genetic polymorphism.

Despite the reported approximate effect size of 0.1 z-score units (Hennig et al. 2009) for BMI per susceptibility allele we did not find metabolic signatures suggesting how the *FTO* gene variant may operate. The variation related to age and gender is only around 0.1 % across all 171 variables, and at most 1 % for single variables, revealing a high degree of unexplained variation. Targeted analysis opposed to un-targeted metabolic profiling examining a narrow range of biological relevant metabolites in connection with *FTO* expression and genetic and environmental confounders would increase the probability of discovery. This is supported by a range of recent works (An et al. 2010) which suggest data driven analysis in combination with mechanistic understanding as a strategic path to uncovering associations from high throughput methods. In this way, narrowing the model range by a priori parameter restriction leads to more statistical power, which is critical for data with large degree of unexplained variation.

5 Conclusion

This study shows that ^{1}H CPMG NMR plasma profiles do not contain signals which can be strongly associated with

FTO genotype. We speculate that high-throughput un-targeted genotype-metabolic correlations in many cases is a difficult path to follow due to correlation shrinkage by multiple interacting factors and inadequate power. Nevertheless, we believe it is worthwhile to explore also this possibility for getting knowledge about how the genotypes work in cases, where we have no information guiding us to the mechanisms.

Acknowledgments The GEMINAKAR study was supported by grants from the Danish Medical Research Fund, the Danish Diabetes Association, the NOVO Foundation, the Danish Heart Foundation, and Apotekerfonden. The present study was supported by the Diogenes study which is the acronym for "Diet, Obesity and Genes" supported by the European Community (Contract no. FP6-513946), http://www.diogenes-eu.org/. The study was part of the research in The Danish Obesity Research Centre, DanORC (http://www.danorc.dk).

References

An, G., Bartels, J., & Vodovotz, Y. (2010). In silico augmentation of the drug development pipeline: Examples from the study of acute inflammation. *Drug Development Research, 72*(2), 187-200

Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols, 2*(11), 2692–2703.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Benyamin, B., Sørensen, T., Schousboe, K., Fenger, M., Visscher, P., & Kyvik, K. (2007). Are there common genetic and environmental factors behind the endophenotypes associated with the metabolic syndrome?. *Diabetologia, 50*, 1880–1888. doi: 10.1007/s00125-007-0758-1.

Berentzen, T., Kring, S. I. I., Holst, C., Zimmermann, E., Jess, T., Hansen, T., Pedersen, O., Toubro, S., Astrup, A., & Sørensen, T. I. A. (2008). Lack of association of fatness-related FTO gene variants with energy expenditure or physical activity. *Journal of Clinical Endocrinology & Metabolism, 93*(7), 2904–2908.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software.

Daszykowski, M., Walczak, B., & Massart, D. L. (2002). Representative subset selection. *Analytica Chimica Acta, 468*(1), 91–103.

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science, 316*(5826), 889.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals, 1*(2), 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1.

Gerken, T., Girard, C. A., Tung, Y. C. L., Webby, C. J., Saudek, V., Hewitson, K. S., Yeo, G. S. H., McDonough, M. A., Cunliffe, S., McNeill, L. A., et al. (2007). The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science, 318*(5855), 1469.

Hasselbalch, A. L., Angquist, L., Christiansen, L., Heitmann, B. L., Kyvik, K. O. & Sørensen, T. I. A. (2010). A variant in the fat mass and obesity-associated gene (FTO) and variants near the melanocortin-4 receptor gene (MC4R) do not influence dietary intake. *Journal of Nutrition, 140*(4), 831.

Haupt, A., Thamer, C., Staiger, H., Tschritter, O., Kirchhoff, K., Machicao, F., Haring, H. U., Stefan, N., & Fritsche, A. (2009). Variation in the FTO gene influences food intake but not energy expenditure. *Exp Clin Endocrinol Diabetes, 117*, 194–197.

Hennig, B., Fulford, A., Sirugo, G., Rayco-Solon, P., Hattersley, A., Frayling, T., & Prentice, A. (2009). Fto gene variation and measures of body mass in an african population. *BMC Medical Genetics, 10*(1), 21.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology, 24*(6), 417–441.

Jess, T., Zimmermann, E., Kring, S. I. I., Berentzen, T., Holst, C., Toubro, S., Astrup, A., Hansen, T., Pedersen, O., & Sørensen, T. I. A. (2008). Impact on weight dynamics and general growth of the common fto rs9939609: A longitudinal Danish cohort study. *International Journal of Obesity, 32*(9), 1388–1394.

Kennard, R.W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics, 11*(1), 137–148.

Kring, S. I. I., Holst, C., Zimmermann, E., Jess, T., Berentzen, T., Toubro, S., Hansen, T., Astrup, A., Pedersen, O., & Sørensen, T. I. A. (2008). Fto gene associated fatness in relation to body fat distribution and metabolic traits throughout a broad range of fatness. *PLoS One, 3*(8), e2958.

Nørgaard, L., Bro, R., Westad, F., & Engelsen, S. B. (2006). A modification of canonical variates analysis to handle highly collinear multivariate data. *Journal of Chemometrics, 20*, 425–435.

Peng, S., Zhu, Y., Xu, F., Ren, X., Li, X., & Lai, M. (2011). Fto gene polymorphisms and obesity risk: A meta-analysis. *BMC medicine, 9*(1), 71.

Peré-Trepat, E., Ross, A. B., Martin, F. P., Rezzi, S., Kochhar, S., Hasselbalch, A. L., Kyvik, K. O., & Sørensen, T. I. A. (2010). Chemometric strategies to assess metabonomic imprinting of food habits in epidemiological studies. *Chemometrics and Intelligent Laboratory Systems, 104*(1), 95–100.

Pitman, R. T., Fong, J. T., Billman, P., & Puri, N. (2012). Knockdown of the fat mass and obesity gene disrupts cellular energy balance in a cell-type specific manner. *PloS one, 7*(6), e38444.

Rezzi, S., Ramadan, Z., Martin, F. P. J., Fay, L. B., van Bladeren, P., Lindon, J. C., Nicholson, J. K., & Kochhar, S. (2007). Human metabolic phenotypes link directly to specific dietary preferences in healthy individuals. *Journal of proteome research, 6*(11), 4469–4477.

Saunders, C. L., Chiodini, B. D., Sham, P., Lewis, C. M., Abkevich, V., Adeyemo, A. A., de Andrade, M., Arya, R., Berenson, G. S., Blangero, J., Boehnke, M., Borecki, I. B., Chagnon, Y. C., Chen, W., Comuzzie, A. G., Deng, H.-W., Duggirala, R., Feitosa, M. F., Froguel, P., Hanson, R. L., Hebebrand, J., Huezo-Dias, P., Kissebah, A. H., Li, W., Luke, A., Martin, L. J., Nash, M., Ohman, M., Palmer, L. J., Peltonen, L., Perola, M., Price, R. A., Redline, S., Srinivasan, S. R., Stern, M. P., Stone, S., Stringham, H., Turner, S., Wijmenga, C., & Collier, D. A. (2007). Meta-analysis of genome-wide linkage studies in BMI and obesity[ast]. *Obesity, 15*(9), 2263–2275.

Savorani, F., Tomasi, G., & Engelsen, S.B. (2010). *i*coshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance, 202*(2), 190–202.

Schousboe, K., Visscher, P. M., Erbas, B., Kyvik, K. O., Hopper, J. L., Henriksen, J. E., Heitmann, B. L., & Sørensen, T. I. A.

(2003). Twin study of genetic and environmental influences on adult body size, shape, and composition. *International Journal of Obesity, 28*(1), 39–48.

Tauler, R., & Barceló, D. (1993). Multivariate curve resolution applied to liquid chromatography–diode array detection. *TrAC Trends in Analytical Chemistry, 12*(8), 319–327.

Viant, M. R., Ludwig, C., & Gunther, U. L. (2008). 1D and 2D NMR Spectroscopy: From Metabolic Fingerprinting to Profiling. Metabolomics, metabonomics and metabolite profiling, page 44.

Wahlen, K., Sjolin, E., & Hoffstedt, J. (2008). The common rs9939609 gene variant of the fat mass-and obesity-associated gene FTO is related to fat cell lipolysis. *The Journal of Lipid Research, 49*(3), 607.

Walley, A. J., Asher, J. E., & Froguel, P. (2009). The genetic contribution to non-syndromic human obesity. *Nat Rev Genet, 10*(7), 431–442.

Wang, H., Dong, S., Xu, H., Qian, J., & Yang, J. (2012). Genetic variants in fto associated with metabolic syndrome: A meta-and gene-based analysis. *Molecular biology reports, 39*(5), 5691–5698.

Wardle, J., Carnell, S., Haworth, C. M. A., Farooqi, I. S., O'Rahilly, S., & Plomin, R. (2008). Obesity associated genetic variation in fto is associated with diminished satiety. *J Clin Endocrinol Metab, 93*(9), 3640–3643.

Yang, W., Kelly, T., & He, J. (2007). Genetic epidemiology of obesity. *Epidemiologic reviews, 29*(1), 49.

Zöllner, S., & Pritchard, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *The American Journal of Human Genetics, 80*(4), 605–615.

PAPER V

A simplified approach for identifying and separating unique and bulk variations in microarray data

K. Kjeldahl[a], Andersen, C.M.[a] , B. Schmidt, A. de Jong[b], J. Kok and R. Bro[a*]

[a] Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C

[b] Department of Molecular Genetics, University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute, NL-9750 AA Haren

*corresponding author: email: rb@life.ku.dk

ABSTRACT

This paper presents a robust analysis of time course DNA microarray data using principal component analysis (PCA) and multivariate curve resolution (MCR). The method identifies a selection of unique genes, which are genes with a distinct characteristic profile, and identifies of a larger group of genes, which have a common underlying structure. It is a step-wise method, which identifies unique genes as outliers in an initial PCA model. These genes are removed with the knowledge that they are relevant for the process studied. The rest of the genes contain a common systematic variation, which is identified and visualized by MCR analysis. The MCR analysis leads to a set of distinct phenomena which are characteristic of a large number of genes. The genes with the highest score values in this model are evaluated further and are the ones, which together with the unique genes are most important candidates for describing the process studied. Data from a milk fermentation process by the lactic acid bacteria *Lactococcus lactis* are used as an example. The results obtained supports the conclusions from previous studies but the approach illustrates a simple, easy and robust way of obtaining information from complex data structures.

KEYWORDS: Multivariate data analysis, DNA microarrays

INTRODUCTION

Within life sciences a lot of effort is dedicated to the discovery and understanding of fundamental physiological processes within cells. At the genetic level, genes are up- or down-regulated to accommodate to the altered needs in the cells' protein composition due to cellular homeostatic responses or other influences, such as disease or treatment. Mapping of the gene expression responses to specific conditions is a key to understanding and hence control of the metabolic pathways and is therefore of great interest (Sorek and Cossart 2009).

The DNA microarray technology represents an attractive tool in this area because the expression level of thousands of genes within a cell of any organism can be measured simultaneously at a given time point. This facilitates a wide range of gene expression profiling experiments such as investigation of the genetic response to given treatments or tissue-specific monitoring of physiological state over time. Cancer diagnostics, drug development and optimization of fermentation processes are examples of areas, which benefit from the use of microarrays. The subject of this paper is a time course experiment where the gene expression of a culture is monitored during a fermentation process.

Microarray data are typically large data sets containing information about many genes for a few experiments. Thus, there is a need for data analytical methods, which can extract the relevant information from these large and complex data structures. Some challenges in the analysis of microarray data is a large number of missing values, noisy data and few replicates. Furthermore, the timing of many biological processes are not well-defined, making replicates even more difficult to obtain. There are several ways of analyzing data from DNA microarrays such as classical statistical hypothesis testing and cluster analysis techniques. Other methods are the various multivariate data analytical techniques such as principal component analysis (PCA) (Alter et al. 2000; Holter et al. 2000; Jonnalagadda, and Srinivasan 2008) and multivariate curve resolution (MCR) (Jaumot et al. 2006; Wentzell et al. 2008; Jaumot et al. 2010). Such methods are especially relevant for time course data since they enable a continuous representation of all genes in a time series experiment even when measurements are only performed at a few time points.

This paper presents a robust and simple analysis of time course microarray data using a combination of PCA and MCR. These are both well-established methods but here they are used in a slightly different way than normal. The data analysis approach enables a separation of the individual genes into two groups: one group consisting of distinct genes showing unique individual variation that needs to be analyzed separately and another large group of genes that are characterized by showing the same *underlying* patterns. This group is further simplified by a selection of the most characteristic genes, which are used in MCR to determine the underlying fermentation profiles of the genes and the distribution of the genes in relation to these profiles. Microarray data of the bacteria *Lactococcus lactis* obtained during the fermentation of milk is used as an example. *Lactococcus lactis* is a bacteria typically used in milk fermentations and is also genetically well characterized (Kok et al. 2005). During fermentation of milk, a change in pH, chemical composition, etc. takes place in the sample matrix surrounding the cells and the purpose is to identify how the bacteria sense these changes. If for example the bacteria are exposed to stress, it would be seen as an increased expression of stress related genes. This information can be used to remove the stress factors, thus increasing the effectiveness within the industry.

Methods

Experimental

Reconstituted skim milk was inoculated with *Lactococcus lactis* whereby the fermentation started. Samples were taken twelve times during the fermentation process (100, 150, 200, 250, 300, 350, 400, 500, 600, 800, 1000 and 1100 min) and subjected to microarray analysis. The whole experiment was made in duplicate.

Data analysis

The microarray analysis contained information from 1204 annotated genes in duplicates and only these were used in the data analysis. Since the main changes in the bacterial growth took place around 400 min, this level was defined as reference (=1). All mRNA levels were normalized relative to this.

The method developed is based on the chemometric techniques principal component analysis (PCA) and multivariate curve resolution (MCR). The microarray data were arranged in a matrix, **X** (*i*,*j*), with the genes in the rows (*i*) and the measured time points in the columns (*j*).

PCA extracts the main systematic variation of **X** by resolving the information into principal components (PCs) (Wold et al. 1987):

$$\mathbf{X} = \mathbf{TP'} + \mathbf{E}$$

where **T** is the score matrix, containing information about the amount of each component in every sample (gene). **P** is the loading matrix and contains information about the contribution of the variables (time points) to each PC. The PCA model describes the underlying latent variables, and as consequence, genes which do not follow the common underlying structures can be identified as unique genes and can be removed for separate observation. The remaining genes then all follow the underlying structure and the PCA model can then be used to describe this.

An alternative to PCA is MCR which can be used for more natural description of time domain phemonena.MCR resolves the data into a similar model consisting of scores and loadings that are nonnegative as opposed to PCA that provides orthogonal scores and loadings. Using MCR as an alternative final representation instead of PCA can lead to a model which is easier to interpret and visualize. The MCR model can be written as (de Juan and Tauler 2003):

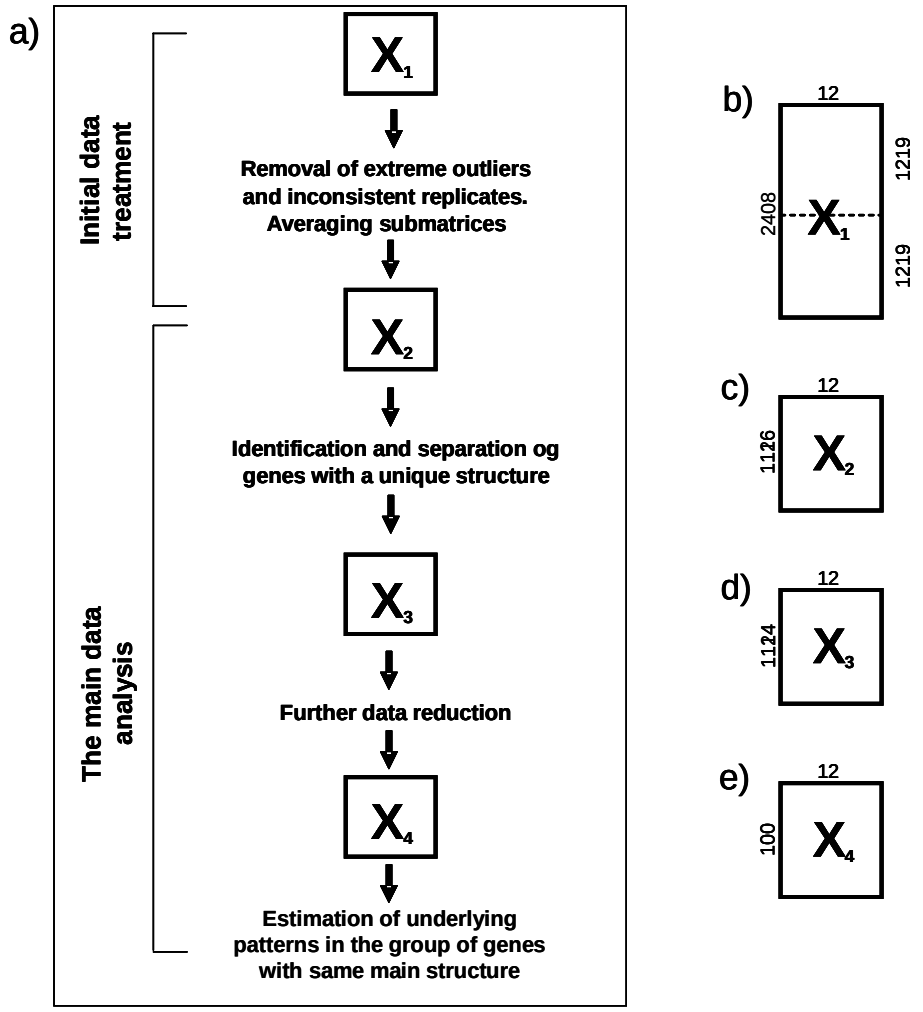
$$\mathbf{X} = \mathbf{CS}^{\mathbf{T}} + \mathbf{E}$$

The data, **X**, is decomposed into **C** and $\mathbf{S}^{\mathbf{T}}$, which contain the responses of the mRNA levels and the fermentation profiles, respectively. The parameters are constrained to be non-negative. For both models, **E** denotes the error matrix, which contains the variation in data not explained by the given model. These residuals can be different in the two models. In both PCA and MCR, the number of model components to include has to be chosen by the user. In this case, cross-validation has been used as the main tool together with visual assessment of the models (Bro and Kjeldahl 2008).

RESULTS AND DISCUSSION

Initially, clearly incorrect outliers and inconsistent replicates are removed from the data. Thus, only the reliable measurements are used in the main data analysis where the unique samples are identified first as outliers in PCA. This is followed by MCR modeling of the underlying structure in the larger group of data that follows a common trend. Figure 1 shows a flow chart of the method as well as the sizes of the data matrices that change during the consecutive analysis steps.

The data is arranged into a matrix, $\mathbf{X}_1$, with the genes as rows and the measured time points as columns. Concatenating the data from the two experimental runs gives a data matrix of the dimension 2438 × 12 (Figure 1.b). A closer look into the data shows that some samples have a considerable amount of missing values, which makes the modeling difficult . They are removed from the data set together with their replicate. In total, the data from 15 genes were removed.



140 Figure 1 a) Flow chart of the method developed for the analysis of DNA microarray data set, b) – e) dimensions of the data matrices X_1 to X_4.

A new PCA is made on the reduced data set. The scores of principal component one and two show that almost all genes are placed in a group near the origin (Figure 2). In
145 addition, there are few genes spread out along the first two PCs. The projected measurements from the two experiments of some of these are placed close together such as *uspA2* and *osmC*, whereas a few others display high variance between replicates, e.g.

trmA and *yfiA*. For these data, it is essential that the results obtained from the two replicate experiments are similar. Thus, the PCA indicates that further data reduction is necessary to ensure reliable and valid conclusions including consistency across the two experiments.

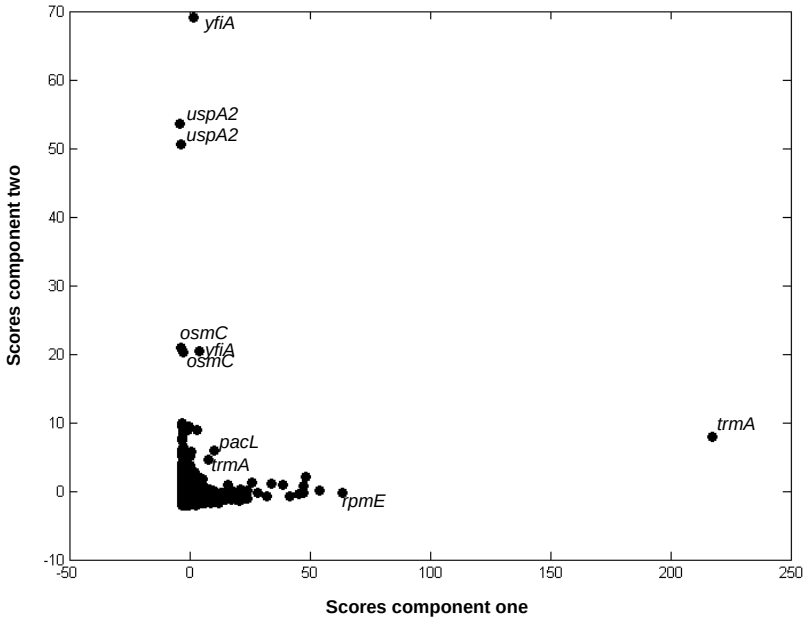


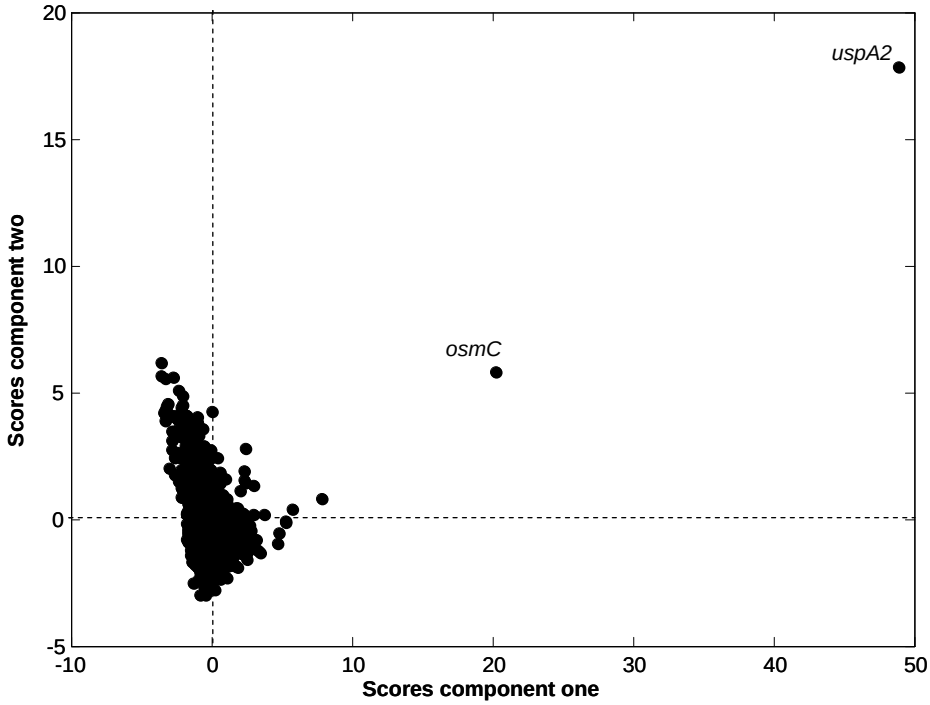
Figure 2 Scores of principal components 1 and 2 for a PCA made on the data matrix with the 2408 genes as rows and the 12 measured points as variables. Genes are marked by their names.

The genes with the largest variation between the two experiments are removed from the data set. These genes are found by the pooled standard deviation between the two experiments calculated for each gene. It is obtained as:

$$s_p = \sqrt{\frac{\sum (x_{i1} - x_{i2})^2}{2k}}$$

where k denotes the number of measurements obtained for each gene. In this case, it is 12. x_{i1} and x_{i2} denote the relative mRNA levels for the two experiments, respectively. The subscript i (1,2,….,12) indicates the 12 time points where the mRNA levels have been measured. All genes with a pooled standard deviation of less than 1.5 are retained in the data set. In total, 1126 genes are found to follow this criterion. The limit of 1.5 for the pooled standard deviation has to be chosen carefully. If it is too high, genes with inconsistent replicates may be included in the data set. If it is too low, some of the unique genes or genes that are of utmost relevance for the main variation may be excluded.

Since it is found that the measurements of the remaining genes are precise, the following data analysis is made on average values obtained from the two experiments. Thus, the size of the data matrix, $\mathbf{X}_2$, is 1126 × 12 (Figure 1.c). Again, a PCA is made. The scores of the first two components show that two genes are separated from the others (Figure 3). It is the same two genes as were identified in the previous PCA as having extreme score values, *osmC* and *uspA2*. These two genes have fermentation profiles differing from the others. They can be considered as unique genes whose expression is important for the bacteria during the fermentation. The expression of those two genes varies considerably from the others. It is seen that the mRNA levels increase dramatically after 500 min of the fermentation and have the highest concentrations at the end of the fermentation of all genes (Figure 4.a). Both genes are related to stress and code for universal and osmotic stress, respectively.



185 Figure 3 Scores of PC1 and PC2 for a model made on the matrix with the size 1126 × 12. The two outliers are marked by their name.

The two genes with the characteristic profiles are then removed from the data set, which then obtains the dimension 1124 × 12 (Figure 1.d). The raw fermentation profiles of the
190 genes left in the data set are in the same level, indicating that there are no more extreme genes or outliers (4.b).

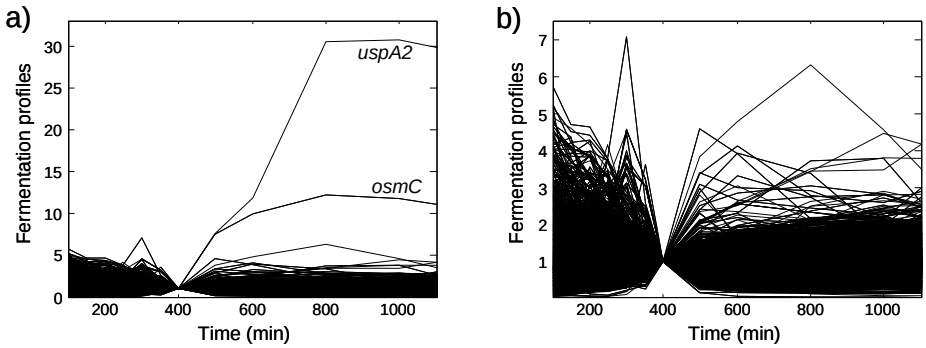


Figure 4 Fermentation profile of the genes, a) the two unique genes: *uspA2* and *osmC* are marked and b) the data set with the two unique genes removed. Averages over the results from the two experiments are shown.

There are still more than 1000 genes left in the data and these genes together behave according to the same underlying latent variation as expressed by the PCA model. In order to simplify the interpretation and understanding, the genes with the most distinct profiles are selected using the scores from a PCA. The 100 genes with the largest Mahalanobis distance to the origin are found. It is these genes that contribute most to the model and thus the genes with the most archetypical fermentation profiles. The Mahalanobis distance of gene (*i*) is calculated as (Maesschalck et al. 2000):

$$\mathbf{D_{mahal}(i)} = \sqrt{\mathbf{(origin - T_i) * inv(cov(T) * (origin - T_i)'}}$$

where **T** denotes the score matrix and $\mathbf{T_i}$ is the scores of the *i*'th sample (gene). The 100 genes with the highest Mahalanobis distances are retained in a new data matrix, $\mathbf{X}_4$, with the dimension 100 x 12 (Figure 1.e). These genes are evaluated by multivariate curve resolution (MCR), giving the underlying fermentation profiles and the relative mRNA levels of the 100 chosen genes. The number of components in this model is related to the number of characteristic fermentation profiles. Three components are found to be present in the current data set (Figure 5). They explain 8%, 34% and 53% of the variation in the data, respectively. As evidenced by the loadings in Figure 5, the first component describes genes that are almost not expressed in the beginning of the fermentation and whose expression increase during the exponential growth and ends at a high level. Component two explains genes that are first up regulated but are down regulated as the bacteria experiences exponential growth. The third component contains information about genes that are highly expressed at the initiation of the fermentation but which decrease sharply as the fermentation runs.

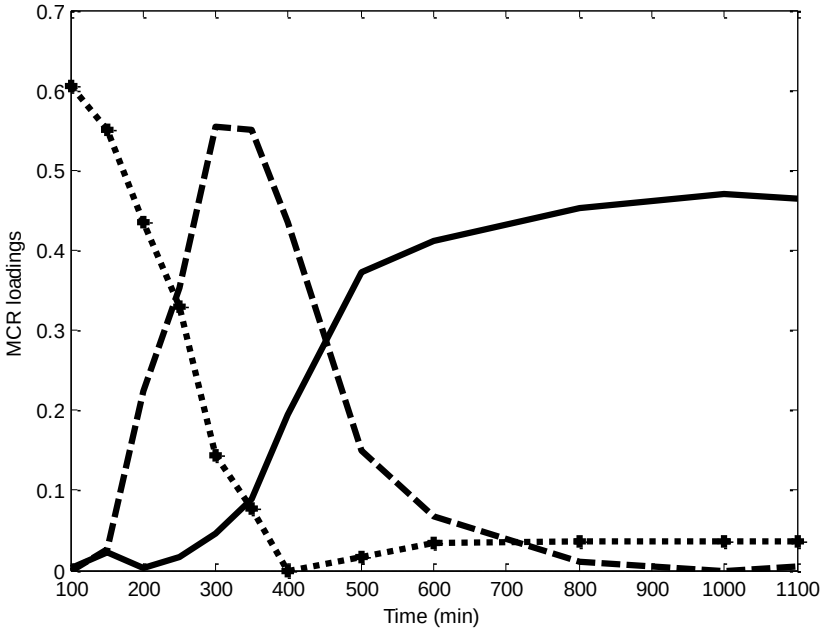


Figure 5 Loadings of the MCR model. The lines indicate component one (dotted), component two (solid) and component three (dashed).

The fermentation profiles of the genes with the highest scores in the three components show the similarity with the profiles estimated by MCR (Figure 6). Also, the unique genes excluded previously, *uspA2* and *osmC* (Figure 4.a), display profiles similar to the profiles of the genes described by component one.

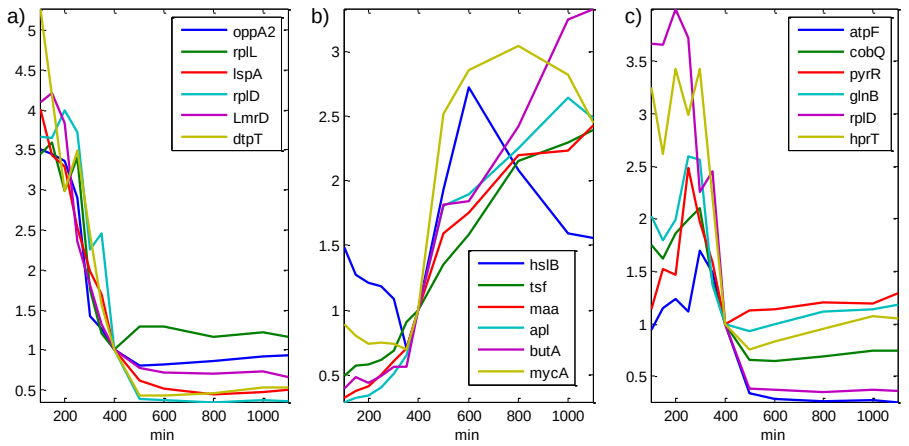


Figure 6 Fermentation profiles of the genes with the highest score in component one (a), component two (b) and component three (c)

For the results shown, 100 genes were chosen from the PCA scores and the calculated Mahalanobis distances. It was also tried to make the MCR model with other numbers of genes. This gave similar estimated fermentation profiles and only a little variation in genes identified as important for the three MCR components. Thus, it was found that the number of genes included in the MCR model is not critical for the interpretation of the final results.

From previous studies, important genes were identified together with their relationship to the fermentation process. These findings are confirmed in this analysis, and the approach suggested here ensures a robust and relatively easy identification of the fermentation profiles, unique genes as well as the group of genes that displays a common underlying structure. The present data contains information from approximately 1200 genes. As shown this is too many to give a full understanding of all of these from a visual inspection of scores and loadings. However, the fact that most of the genes can be described by the same underlying variation evidenced through the PCA/MCR model makes it feasible to perform a reduction based on selecting the most distinct of the bulk of genes behaving similarly. This is not a large data set compared with other microarray

data where data reduction is even more important. The method developed could be an important tool in situations with larger data sets and when these are not as well known. The method may not only apply to microarray data, but shows possibilities in the analysis of other large and complex data structures such as NMR and LC-MS based metabonomics data sets. Furthermore, the principle in separating irrelevant information, unique genes and the main underlying structure can be adopted using other multivariate or multi-way techniques as well as other statistical calculations.

There are other applications of MCR in the study of DNA microarrays. However, none of these use a stepwise procedure including removal of all irrelevant data, identification of the unique genes and finally an evaluation of the main systematic variation in the group of data containing a common underlying structure. Jaumot et al. (2006) turns the data matrix such as the experiments make up the rows and the genes make up the columns. The results of their analysis are pure gene expression profiles and the relative contribution of each experimental condition. This makes sense in their study since the purpose is to discriminate between the experiments using the microarray data. Wentzell et al. (2006) studies time course microarray data. They use a weighted approach where the non-uniform error distribution, missing data, etc. are handled through the model algorithm. For both studies, information about the behavior of the individual genes has to be identified after the modeling step. For example, Wentzell et al (2006) finds the most important genes as the genes with the largest correlation to the profiles extracted by MCR in contrast to the use of Mahalanobis distances and score values. Jaumot et al (2010) compares two MCR algorithms for microarray time series analysis; the present paper focuses on simplicity and robustness.

Conclusion

The paper shows a robust multivariate approach to the study of DNA microarray data. The method identifies individual genes with a unique fermentation profile and a larger group of genes with a common systematic variation. It is a step-wise procedure, which first removes outliers and inconsistent replicates. It is followed by the main data analysis. This secures a robust identification of the most important information present in the data.

References

Alter, O.; Brown P.O.; Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97, 10101-10106.

Bro, R.; Kjeldahl, K. (2008): Cross-validation of component models: A critical look at current methods. *Anal Bioanal Chem*, 390:1241–1251

de Juan, A.; Tauler, R. (2003) Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta*, 500, 195-210.

Holter, N.S.; Mitra, M.; Maritan, A.; Cieplak, M.; Banavar, J.R.; Fedoroff, N.V. (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl. Acad. Sci. USA*, 97, 8409-8414.

Jaumot, J.; Piña, B.; Tauler, R. (2010) Application of multivariate curve resolution to the analysis of yeast genome-wide screens. *Chemometrics and Intelligent Laboratory Systems* 104 (2010) 53–64.

Jaumot, J.; Tauler, R.; Gargallo, R. (2006) Exploratory data analysis of DNA microarrays by multivariate curve resolution. *Analytical Biochemistry*, 358, 76-89.

Jonnalagadda, S.; Srinivasan, R. (2008) Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. *BMC Bioinformatics*, 9: 267.

Kok, J.; Buist, G.; Zomer, A.L., van Hijum, S.A.F.T.; Kuipers, O.P. (2005) Comparative and functional genomics of lactococci. *FEMS Microbiological Reviews*, 29, 411-433.

de Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. (2000) Tutorial. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50, 1-18.

Sorek, R.; Cossart, B. (2009). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics*, 11, 9-16

Wentzell, P.D.; Karakach, T.K.; Roy, S.; Martinez, M.J.; Allen, C.P.; Werner-Washburne, M. (2006) Multivariate curve resolution of time course microarray data. *BMC Bioinformatics*, 7, 343-

Wold, S.; Esbensen, K.; Geladi, P. (1987) Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*. 2, 37-52.

Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20, 397-405.