Summary

Over the past decades, measurement technology has been rapidly developing, allowing for the large-scale collection of several 'omics' data types. A statistical analysis of all data simultaneously can lead to a better understanding of the underlying biological mechanism. However, the analysis of these data poses several statistical challenges. Most importantly, the datasets are characterized by a complex correlation structure within and between each type, and often have many more variables than samples. These characteristics render approaches such as the multivariate linear model infeasible. Other challenges are systematic differences in terms of size, scale, distribution and measurement platform among the datasets. This dissertation focuses on omics data integration as a holistic approach to jointly analyze multiple omics datasets. In the first part, we study data integration methods and their applicability to multiple omics data from population cohorts. In the second part, we propose a probabilistic framework to model the relations and systematic differences between two omics datasets.

The first chapter gives an introduction to current data integration methodology and describes some challenges when estimating relations between datasets. First, characteristics shared between and specific to the omics data are presented. Then, omics data integration approaches are introduced and unified in the Structural Equation Modeling (SEM) framework. In this framework, properties as strong correlations and high dimensionality can be incorporated. Standard SEMs do not model heterogeneity between omics datasets, therefore, extensions to the SEM framework that include data-specific parts in the model are described. The current data integration approaches mainly differ in the parametrization of the residual variance terms. Strategies to estimate the parameters in an SEM can be (roughly) categorized into two: Least Squares and Maximum Likelihood. The latter approach is preferred to reduce the risk of overfitting and facilitates incorporating complex dependency structures and extending to supervised omics data integration (as in Chapter 6). In both strategies, high dimensionality poses challenges to the estimation procedure. Identifiability of the SEM model parameters is described and, to this end, natural constraints are proposed. Finally, in the field of omics data integration, statistical software is needed to facilitate the use and development of novel methodology. The chapter finishes with challenges associated with developing open-source software packages.

Chapter 2 is the first of two chapters that aim to study existing data integration methods and their applicability to omics data from population cohorts. As a motivation, transcriptomics and metabolomics data from the DILGOM[1] study are integrated. A particular method based on least squares, Two-way Orthogonal Partial Least Squares (O2PLS), is applied to estimate both joint and specific parts. The components in these parts are sequentially obtained, using a singular value decomposition in each step. To determine the optimal number of components, a fast alternative to standard cross-validation is proposed. Via a simulation study, the performance of O2PLS is evaluated. It appeared that a high noise level affects the estimation of the specific parts, while the joint parts are mainly unaltered. Application to the omics data in the DILGOM study showed that O2PLS structured the joint parts into interpretable genetic and metabolic components. The top genes were linked to anti-inflammation, while the metabolic components revealed the clustering of several lipoprotein subtypes. These findings include the results of previous standard analyses.

[1] Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome

In the third chapter, a free and open-source software package, OmicsPLS, is introduced that implements O2PLS in R together with various tools, such as cross-validation. The memory-efficient algorithm can handle low- and high-dimensional data. To enhance the use of the software package, interpretation of the model and parameters is discussed, and to facilitate the application of omics data integration, several post-analysis functions are implemented: plot generators using the ggplot2 package, text summaries of the O2PLS fit, and wrappers to easily access the estimates. A simulation study is conducted to compare OmicsPLS with an alternative method in terms of accuracy and speed. OmicsPLS was much faster and the alternative often failed to converge when data-specific parts were present. OmicsPLS was applied to genetic and glycomic data from the Croatian Korcula cohort. The joint components revealed a clear clustering in terms of the molecular structure of the glycans, and the top genes were related to the manipulation and location of glycoproteins.

In the chapters 4 and 5, a probabilistic framework for (high dimensional) omics data integration is proposed and studied. Chapter 4 presents the Probabilistic PLS (PPLS) model for the relation between two omics datasets. The model is based on the standard SEM framework, using a normal distribution for the latent variables. Identifiability of the model is established using natural constraints on the parameters. These parameters are estimated with maximum likelihood, using an EM algorithm, where the constraints are properly incorporated in the M step. The EM steps can be calculated separately for the parameters involving the first and second dataset, respectively. The EM algorithm can be implemented efficiently in terms of memory usage. An explicit expression for the standard errors can be obtained by calculating the observed Fisher information matrix. An extensive simulation study was conducted to evaluate the performance of PPLS in terms of bias and variance of the estimators and to compare it with other methods. Most notably, the PPLS estimates were robust against departures from normality and performed well in both low and high dimensional data. The PPLS model was used to integrate the IgG1 and IgG2 parts of glycomics datasets from two Croatian cohorts, Korcula and Vis. The joint components in the Korcula cohort revealed a grouping of glycan subtypes; these components were replicated in the Vis cohort.

Chapter 5 presents a probabilistic reformulation of O2PLS and an extension of PPLS to model joint and specific characteristics in two sets of variables. The proposed model, Probabilistic O2PLS (PO2PLS), includes data-specific parts. Similar to PPLS, by using natural constraints, the parameters are identifiable. They are estimated with maximum likelihood using a memory-efficient EM algorithm, and standard errors can be calculated. The performance of PO2PLS was evaluated with a comprehensive simulation study, focusing on predictive performance and interpretability. Especially in the setting where data are heterogeneous or high dimensional, PO2PLS performed better than alternative methods. Notably, overfitting was considerably reduced compared to O2PLS. The PO2PLS model was fitted to omics data from two studies: transcriptomics and metabolomics data from the DILGOM cohort (as in Chapter 2), and genetics and glycomics data from the Korcula cohort (as in Chapter 3). In the DILGOM study, the top genes in each PO2PLS joint component were clustered in distinct genetic pathways without overlap across components, reflecting current biological knowledge, while the pathways derived from the O2PLS top genes were found in both O2PLS joint components. By using the Korcula cohort as training and the

Vis cohort as test data, the prediction error of PO2PLS was smaller compared to O2PLS.

The sixth chapter discusses extensions of the probabilistic data integration framework to incorporate complex study designs. In particular, a supervised framework is considered to incorporate an outcome in omics data integration. As an illustration, in the first part of the chapter, O2PLS is applied to genetic and epigenetic data from the Genetic Analysis Workshop (GAW). These data are measured using a longitudinal family study design. We aim to model the effect of these datasets on triglyceride levels. A three-step approach was used to first capture time-varying parts in the epigenetic data with the O2PLS specific components. The top genes in the joint parts were linked to immunological pathways, while in the genetic-specific parts clustering of families was observed. In the second step, joint genetic and time-varying components were captured with O2PLS. Finally, the outcome was regressed on these components using linear mixed models where familial correlations are taken into account. The second part of this chapter sketches a novel supervised probabilistic model, as an extension of PO2PLS, to simultaneously model the relationship between the genetic, epigenetic and triglyceride data. In this model, epigenetic correlations across time points are incorporated, as well as familial correlations across subjects. Each contribution to the overall likelihood can again be decoupled in several steps, allowing for separate optimizations per part related to the genetic, epigenetic and triglyceride parameters, respectively.

To conclude, the probabilistic modeling framework can be extended to analyze omics data from various study designs, in the presence of several types of outcomes. However, further research is needed in this direction, such as including non-linear relations or non-normal outcomes in the joint model. Also, to combine multiple omics data types, the framework can be extended to model the relation between more than two datasets simultaneously. Together with free and open-source software, this will aid the development of novel methodology in the field of omics data integration.