# Statistical integration of diverse omics data

Bouhaddani, S. el

# 1

# Introduction

## 1.1 Background

One of the aims in statistics is to describe the relationship between two sets of variables, for which the multivariate linear regression model is widely used. In this model, the relation between the underlying random vectors $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ is given as [20]

$$y = x\beta + \epsilon. \tag{1.1}$$

The $p \times q$ coefficient matrix $\beta$ describes how the variables in $x$ relate to those in $y$. The random vector $\epsilon \in \mathbb{R}^q$ is the residual error; its variance indicates a lack of a linear relation between $x$ and $y$.

The most frequent method to estimate $\beta$ from data matrices $X$ and $Y$, with $N$ rows, minimizes the squared residual error $\sum_i ||Y_i - X_i\beta||^2$ over $\{\beta \in \mathbb{R}^{p\times q}\}$, where $i = 1, \ldots, N$. Under some assumptions [31], the solution is given by

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y. \tag{1.2}$$

This estimator is unbiased, and optimal in the sense that any linear combination of the elements of $\hat{\beta}$ has a lower variance than the same linear combination of any other linear unbiased estimator [31].

The linear regression approach has some drawbacks. Firstly, when $x$ is high dimensional, i.e. $p > N$, the inverse of $X^{\mathrm{T}}X$ does not exist. Secondly, when the columns of $X$ are highly correlated, $X^{\mathrm{T}}X$ will be nearly singular. Consequently, the covariance matrix of $\hat{\beta}$, given by $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{Var}(\epsilon)$, will have large eigenvalues. Unless the number of rows of $X$ is also large, this leads to an inflation of the mean squared error of $\hat{\beta}$ and imprecise estimates ([6, 32]). Finally, each column of $\hat{\beta}$ can be derived by regressing the columns of $Y$ on $X$ separately, since $\hat{\beta}$ is equivalent to $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}[Y_1, \ldots, Y_q]$. Therefore, $\hat{\beta}$ does not take into account the correlation structure of $Y$.

The described scenarios, high dimensional and highly correlated data, are becoming common in many research areas, especially in the field of biomedical and life sciences. The ongoing technological developments have led to an unprecedented increase in the amount of available data. These data contain several types of molecular measurements for the same samples, and are often suffixed with *-omics*. Examples include genomics, transcriptomics and glycomics, see Figure 1.1. These omics datasets are typically high dimensional (e.g. more than $10^7$ genetic variants, $10^4$ transcripts measured on fewer subjects) and highly correlated, where relationship exist both within and between the different 'omics' levels. Since these data are measured on the same samples, statistical research also focuses on investigating relationships between these omics levels to better understand the underlying biology.

The research presented in this thesis is part of the European FP7 project, Methods for Integrated analysis of Multiple Omics datasets (MIMOmics), see `mimomics.eu`. One of the objectives of MIMOmics is "*to integrate data derived from multiple omics platforms across several study designs and populations*". This thesis addresses the aim by firstly evaluating existing 'data integration' methods, and secondly developing a statistical framework for integrating multi-omics data.

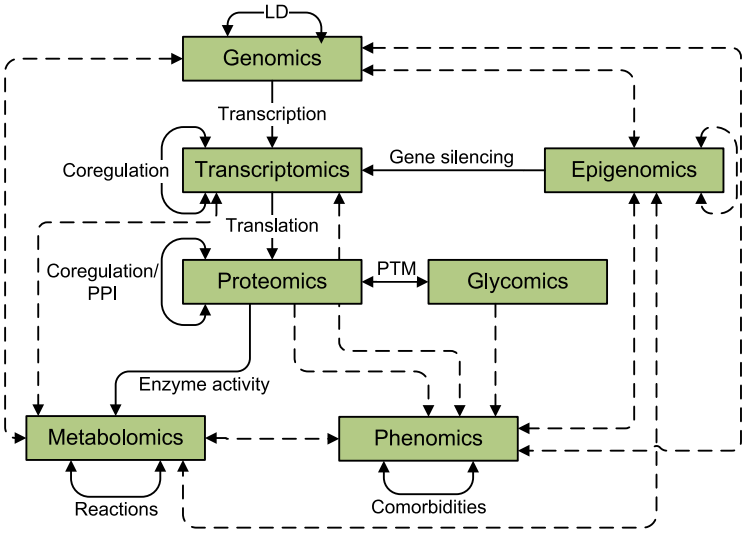


Figure 1.1: **Overview of several types of omics data.** Each rectangle represents an omics domain, where arrows depict possible relations between the domains. For example, a simple model for the relation between four 'main' omics domains goes from genomics to transcriptomics to proteomics to metabolomics. Relations among the measurements in one domain also exist. Figure taken from [44].

### 1.1.1 Omics data characteristics

**Shared characteristics.** In this thesis, several omics datasets are analyzed. A prominent characteristic present among these data is the complex dependence structure, both within as well as between the datasets [34]. An explanation for these dependencies is that, typically, the features $x_j$ in each dataset $x$ are organized such that several sets of molecules (e.g. genes or proteins) $\{x_j, j \in \mathcal{P}\}$ are involved in the same biological "pathway" $\mathcal{P} \subseteq \mathbb{N}$, see e.g. [8]. Pathways from different datasets can be connected, in the sense that the molecules in these pathways share the same goal. Measurements on molecules in connected pathways should then be statistically correlated [28], i.e. for two connected pathways $\mathcal{P}_1$ and $\mathcal{P}_2$ the off-diagonal elements of the correlation matrix of $(x_{\mathcal{P}_1}, y_{\mathcal{P}_2})$, are not zero. Interpretation of currently known pathways is documented in several bioinformatics databases that provide high-level information about their functions [13, 23]. Furthermore, a direction of the dependence between several omics is sometimes assumed. For example information can explained to flow from genomics to transcriptomics to proteomics [4]. For developing methodology for analyzing omics data, these concepts can be utilized to build models that describe the relation between these datasets.

**Data-specific characteristics.** Omics data are also different in several aspects, in particular with respect to dimensionality, distribution and measurement platform. Firstly, as example, the dimensionality of genetic data is of order $10^6$, while in glycomics, the number of features is of order $10^1$. Secondly, the distributions of the genetic

measurements often have support on two or three points, while other omics measurements are known to be non-negative with few or many zeros, skewed or symmetric and discrete or continuous. Finally, the platforms to measure the data are by design very different, each with another type of measurement error distribution (see [43], Table 1). These data-specific characteristics need to be considered when performing statistical analysis of omics data [37]. Furthermore, in a model where data-specific parts are included, this type of variation can be inspected. In particular, it can provide information about technical or biological artifacts in the data. The aim of this thesis is to model the relation between all features in $x$ and $y$ and incorporate data-specific characteristics.

**Aims and datasets in MIMOmics.** The data used in this thesis are obtained from two studies. The DILGOM (DIetary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome) study [9] is conducted in 2007, and contains data on, in total, 4974 participants in the working age population, in 3 to 5 large study areas of Finland. In particular, transcriptomic and metabolomic data are available for the same 466 individuals [10]. The transcriptomics dataset contain 35419 expression levels, measured with the Illumina HT-12 expression array. Regarding the metabolomics data, 137 serum metabolite levels were determined with $^1$H NMR. In this study, a sequential approach was used to investigate how lipid metabolism relate to gene expression. First, the individual genes were summarized to features representing sets of correlated genes. Then, these features were associated with the metabolites using pair-wise correlation tests [9]. An improved statistical framework to test these associations using a subset of the metabolites was also proposed [27]. Here, we consider statistical approaches to simultaneously estimate relationships between all gene expression and metabolite variables, taking into account the correlations within and between each dataset, as well as data-specific variation.

The Croatian study consists of two cohorts, Korcula and Vis [15]. In the Korcula cohort, 969 participants of adult age were recruited in 2007. The Croatian Vis cohort consists of, in total, 1008 participants recruited in 2003 and 2004. Within these studies, 333858 SNPs were genotyped. Also, glycomics measurements are available, consisting of 50 IgG glycan abundances. In this study, one of the aims is to find genetic contributions to changes in glycosylation. Univariate tests were performed for each pair of genetic variants and glycans separately. However, the glycan measurements are highly correlated and often contain substantial measurement error. In this thesis, the relation between both datasets is modeled and the genetic contributions to glycan variation is estimated.

## 1.2 Modeling the relationship

Many statistical approaches that relate $y$ to $x$ focus on detecting linear associations between features. Two research questions are formulated: how strong is the association between $x$ and $y$, and which variables are (most) involved? Traditional approaches are based on statistical measures of association between pairs of features $x_j$ and $y_{j'}$ in both datasets [29]. To this end, p-values are calculated for each test statistic for the null hypothesis of no association, and multiple testing corrections are applied to

control the probability or rate of falsely rejecting any null hypothesis among all tests. Note that these methods do not take into account correlations among the features, as the p-values are obtained from univariate analyses. Furthermore, when considering all pairs of two high dimensional omics data, the large number of tests lead to computational issues and loss of statistical power. We consider multivariate approaches that model the relation between $x$ and $y$ simultaneously.

An approach to model the relationship between two omics datasets, reflecting biology, involves unobserved 'pathways' that explain correlations within and between the measurements. These pathways are represented by latent variables, say $t \in \mathbb{R}^r$ and $u \in \mathbb{R}^r$, capturing the variation common to two datasets. Typically, $r$ is much smaller than $p$ and $q$. The latent variable approach firstly implies that given $t$ and $u$, $x$ and $y$ are conditionally independent. Secondly, since $r << p, q$, such methods also yield dimensionality reduction. In this thesis, these methods are called *(omics) data integration methods*. A latent variable approach to model the relationship between two datasets $X$ and $Y$ takes into account the correlation structure of the datasets, often has good statistical power compared to univariate testing methods, and, due to the dimensionality reduction step, is computationally attractive [24].

### 1.2.1 Unifying the data integration methods.

A statistical framework for data integration of $x$ and $y$ based on latent pathways is the Structural Equations Modeling framework (SEM). An SEM is a model for $x$ and $y$ in terms of $t$ and $u$, given by

$$\begin{aligned} x &= tW^{\mathrm{T}} + e, \\ y &= uC^{\mathrm{T}} + f, \\ u &= uA + tB + h. \end{aligned} \tag{1.3}$$

The first two equations are referred to as the outer model, while the last equation is called the inner model. The latent variables $t$ and $u$ represent the pathways underlying $x$ and $y$, respectively. The loading matrices $W$ and $C$ represent the association strength of each feature for the respective variable in $t$ and $u$. It is further assumed that $(I - A)$ is non-singular and that the error variables $e$, $f$ and $h$ are independent of each other and of $t$ and $u$. Furthermore, the regression matrix $B$ represents the relations between the latent variables $t$ and $u$. The matrix $A$ captures relationships among the $u$, and provides a flexible framework for explaining multi-level structure in $y$. For high dimensional omics data, such additional correlation structure is usually ignored, and the matrix $A = 0$ is taken. A graphical representation is given in Figure 1.2. It is worth mentioning that in Chapter 6, this additional structure will be exploited.

Many data integration methods for omics data have been proposed. To give an indication, *more than 20* methods were compared and their application to multiple omics data was discussed [21]. Several of these methods can be unified in the general structural equation modeling (SEM) framework, as they model $(x, y)$ in terms of $(t, u)$. The main difference between these models is in the parametrization of the residual error covariance matrices. For example, in the Canonical Correlation Analysis (CCA) [7] model, the covariance matrices of $e$ and $f$ are assumed to be diagonal with $p + q$

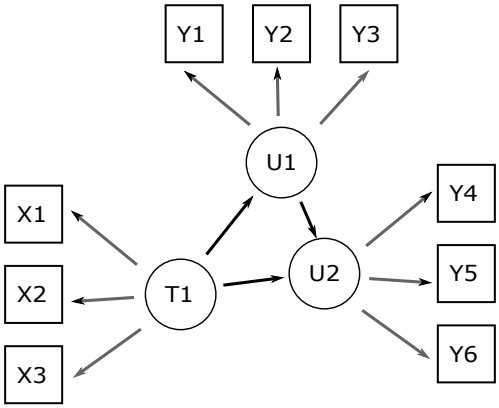


Figure 1.2: **An SEM model with two sets of measurements $X$ and $Y$ and three latent variables.** In the inner model, $U2$ depends on $T1$ as well as $U1$, while $U1$ depends on $T1$ only. The outer model relates the $X$ variables to $T1$, the first three $Y$ variables to $U1$, and the last three to $U2$.

error variance terms, i.e. $\Sigma_e = \text{diag}(\sigma^2_{e_j})$ and $\Sigma_f = \text{diag}(\sigma^2_{f_j})$. Partial Least Squares (PLS) [2] restricts the diagonal elements to be exactly equal, yielding two covariance matrices proportional to the identity matrix with two error terms, i.e. $\Sigma_e = \sigma^2_e I_p$ and $\Sigma_f = \sigma^2_f I_q$. The Envelope Regression (ER) [3] framework in a way compromises between the two by allowing $(p+q) - \text{rank}(t,u)$ degrees of freedom for the error covariance matrices; it restricts the space spanned by $e$ and $f$ to be orthogonal on the space spanned by $t$ and $u$, respectively. In the Redundancy Analysis (RA) method, a diagonal residual variance matrix is assumed for $x$, while one variance parameter is retained for $y$, yielding $p+1$ variance terms.

Note that in many proposed methods, the inner relation is reduced to an equality $u = t$. This has drawbacks when dealing with heterogeneous omics data. Firstly, $t$ and $u$ represent unobserved pathways from different biological layers. These pathways are biologically not perfectly correlated or on the same scale. Therefore, some interpretation is lost when assuming $u = t$, as the corresponding loading components do not reflect the true biological mechanisms. Secondly, in Chapter 3 and 5 it is shown that this assumption may have a negative effect on the performance of the method.

For data integration methods, identifiability of the parameters needs to be investigated. If a model is unidentifiable, several solutions can be constructed that cannot be distinguished in terms of model fit. This can be problematic when individual parameters are interpreted. Roughly stated, identifiability can be established with two assumptions: independence between and within the joint and residual latent variables, and orthogonal loading matrices. One of the most well-known identifiability issues in the SEM is the 'freedom of rotation', [39, 11]. Without restrictions on the model parameters in (1.3), one can take a rotation matrix $R$, such that $RR^{\mathrm{T}} = I$, and introduce it in, e.g., the equation $x = tRR^{\mathrm{T}}W^{\mathrm{T}} + e$. The model remains unchanged, with the joint latent variable as $tR$ and corresponding loadings $WR$. Even if the loadings are assumed to be orthogonal, this issue remains. However, note that $\text{Cov}(tR) = R^{\mathrm{T}}\Sigma_t R$ is not diagonal in general. If independence among the $t$ is assumed, then a non-trivial

rotation leads to joint latent variables that violate this assumption. Identifiability is studied in more detail in Chapter 4 and 5.

### 1.2.2 Estimating the data integration models

To estimate the SEM (1.3), two approaches have been considered in the literature [40]. The first approach is based on a sequence of least squares estimations, the second approach optimizes the likelihood given a probabilistic formulation of (1.3). In the first approach, the parameters $W$ and $C$ are estimated per column by iteratively projecting $(x, y)$ onto $(t, u)$ via current estimates for the respective columns in $W$ and $C$, and vice versa [41, 33]. An initial guess for $W$ and $C$ is required to start the algorithm. This approach is also known as the PLS path modeling algorithm and converges to a PLS, CCA or RA solution, depending on the exact form of the projection. For example, an update of the form $w \propto x^{\mathrm{T}}u$ yields PLS, while $w \propto (x^{\mathrm{T}}x)^{-1}x^{\mathrm{T}}u$ yields CCA (see [33] page 194). Note that in the CCA case, the inverse of $x^{\mathrm{T}}x$ is required. With high dimensional datasets, where $p$ or $q$ is large or larger than the sample size, this inverse is unstable or does not exist, therefore CCA is not suitable for omics data integration. Many data integration methods in this category rely on algorithmic descriptions in which it is difficult to incorporate the identifiability conditions without resorting to post-hoc modifications. As a result, most of these methods fail to produce unique solutions.

A probabilistic approach to estimate SEMs has also been considered for data where $N > p$. By specifying a distribution for the latent variables, a likelihood can be formulated and maximized to obtain estimates. Typically, the latent variables are assumed to be jointly normally distributed, with zero mean and unknown covariance matrix $\Sigma$. Given the model specified in (1.3) and a distributional form for the latent variables, the covariance matrix equals

$$\Sigma = \begin{bmatrix} W\Sigma_t W^{\mathrm{T}} + \Sigma_e & W\Sigma_t BC^{\mathrm{T}} \\ CB^{\mathrm{T}}\Sigma_t W^{\mathrm{T}} & C\left\{B^{\mathrm{T}}\Sigma_t B + \Sigma_h\right\}C^{\mathrm{T}} + \Sigma_f \end{bmatrix}. \tag{1.4}$$

The log-likelihood of the data $X$ and $Y$ then takes the following form:

$$L = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left((X,Y)^{\mathrm{T}}(X,Y)\Sigma^{-1}\right), \tag{1.5}$$

where $N$ is the sample size. For this model, the sample covariance matrix $S = N^{-1}(X,Y)^{\mathrm{T}}(X,Y)$ is a sufficient statistic. Maximum likelihood methods effectively optimize the similarity between the theoretical and sample covariance matrix, given the observed data. Note that when maximizing over all covariance matrices $\Sigma$, without assuming any latent structure, the solution is known to be $S$ [12]. However, when a structure as in (1.4) is assumed, directly optimizing $L$ becomes difficult and iterative procedures are needed.

Optimization methods such as Newton-Raphson (e.g. [25]) can be used. Note that these methods typically depend on the Hessian matrix which is, if $p$ or $q$ is large, computationally infeasible as $O((p+q)^2)$ numbers need to be stored. A memory-efficient approach to obtain maximum likelihood estimates is the EM algorithm [5] where $O(p+q)$ numbers are stored in each step (shown in Chapter 4 and 5). Under the normal distributions assumption, the EM algorithm first calculates the expected

value of the first and second moments of the latent variables of the model (1.3), given the observed data and an initial guess for the parameters. In the second step, maximizers of the complete likelihood are obtained, where the expectations are used as predictions for the latent quantities. This step involves a constrained optimization, to incorporate the identifiability restrictions on the parameters. The two steps are alternated, and under some assumptions, the sequence of estimates converge to a local optimum of the likelihood [5, 42].

Data integration within a probabilistic framework with identifiable parameters facilitates statistical inference. Under certain conditions, the maximum likelihood estimates asymptotically follow a normal distribution around the true parameter values [38, 1]. The covariance matrix of these estimates is given by the inverse of the Fisher information matrix. In an EM algorithm, this matrix can be estimated in the last step, by applying the missing information principle [19]. Based on this matrix, standard errors and, if asymptotic normality is assumed, p-values for the parameter estimates can be calculated.

## 1.3 Modeling the data-specific characteristics

In the SEM (1.3), the covariance matrix of $(x, y)$, given in (1.4), is decomposed in joint and residual parts. As discussed before, data-specific characteristics introduce additional covariance structure in the variance matrix of $x$ and $y$ (see [14], section 3.4). Fitting an SEM without taking into account data-specific variation can lead to misleading results regarding the joint parts, since the estimates need to account for the specific variation using the joint components. Especially if the residual covariance matrices have fewer degrees of freedom, this appears to be an issue. Furthermore, the estimated shared components can erroneously represent specific components if this is the most prominent type of variation in the data. Finally, modeling data-specific parts facilitates further interpretation of the unrelated parts in the data.

The SEM model can be extended to include data-specific components, by including latent variables $t_s$ and $u_s$, independent of the joint and residual parts:

$$\begin{aligned} x &= tW^{\mathrm{T}} + t_s W_s^{\mathrm{T}} + e, \\ y &= uC^{\mathrm{T}} + u_s C_s^{\mathrm{T}} + f. \end{aligned} \tag{1.6}$$

The parameters $W_s$ and $C_s$ are data-specific loading matrices of appropriate size. The theoretical covariance matrix of $(x, y)$ is then given by

$$\Sigma = \begin{bmatrix} W\Sigma_t W^{\mathrm{T}} + W_s \Sigma_{t_s} W_s^{\mathrm{T}} + \Sigma_e & W\Sigma_t BC^{\mathrm{T}} \\ CB^{\mathrm{T}}\Sigma_t W^{\mathrm{T}} & C\left\{B^{\mathrm{T}}\Sigma_t B + \Sigma_h\right\}C^{\mathrm{T}} + C_s \Sigma_{u_s} C_s^{\mathrm{T}} + \Sigma_f \end{bmatrix}. \tag{1.7}$$

Note that without the specific parts, the covariance structure represented by, e.g., $W_s\Sigma_{t_s}W_s^{\mathrm{T}}$ is absorbed by the joint covariance structure and $\Sigma_e$.

To estimate the extended model (1.6), the same two approaches can be used as in Section 1.2.2, namely least squares and maximum likelihood. In the first approach, two separate models are fitted. Using an initial guess for the joint parts, (1.6) reduces to PCA models for $x$ and $y$ where $W_s$ and $C_s$ are the PCs. Next, these components are subtracted and the model reduces to an SEM without data-specific parts as in (1.3).

This approach is used especially in chemometrics, where the Two-way Orthogonal PLS (O2PLS) algorithm was proposed [35]. The JIVE method [18] alternates between the two steps until convergence. The DISCO-SCA algorithm [30] has a somewhat different algorithm: it first estimates a regular SEM (1.3), and performs a post-hoc correction by rotating the solution to also represent specific components. Among these methods, O2PLS has the advantage that it does not assume that $u = t$.

The maximum likelihood approach involves optimizing all parameters of the model (1.6) simultaneously. Since directly calculating the score function is analytically and computationally not feasible, an EM algorithm can be used. In the EM algorithm, the joint and specific loading matrices should also be optimized simultaneously. However, two *conditional* maximization [22] steps can be performed, where the joint parts are optimized while keeping the specific parts fixed and vice verse. This approach also yields a sequence of estimates that converge to a local optimum of the likelihood [22]. This approach is used in SIFA [16] with the assumption that $u = t$. The methodology in Chapter 5 is based on this approach, where in the M step a constrained optimization problem is solved to satisfy the identifiability conditions.

**Free and open-source software.** Software development for data integration methods is an important way to stimulate scientific advances in this area. By providing free access to the source code, experiments can be replicated and the methodology can be validated. Potential solutions for drawbacks and further improvements can be incorporated in the source code, with better and robuster software being the end result. Moreover, new methodology can be compared to alternatives to verify added benefits of the new method. Finally, free access to the software and visualization tools is of direct benefit for users of the methods.

Especially in this field, methodology is sophisticated and difficult to implement from scratch. Furthermore, high dimensional data pose challenges for an implementation in terms of memory usage and computation speed. Several software packages exist for omics data integration on various computing platforms (see [17], Table 2). However, some of the data integration methods are only available via expensive software packages (e.g. O2PLS in SIMCA [36]), or on commercial computing platforms (e.g. DISCO-SCA and SIFA on Matlab). Other implementations cannot be applied to high dimensional data, where $p$ or $q$ are large, due to computational issues. Some tools fail to converge to a solution when applied to heterogeneous omics data. Therefore, continuous development of new methodology and software tools is needed to advance in the field of omics data integration.

## 1.4 General outline of the thesis

The remainder of this thesis is structured into two parts. The first part, consisting of Chapter 2 and 3, studies current data integration methods, evaluates a specific method in omics data from population cohorts, and implements it in a free and open access software package. The second part, covering Chapter 4 and 5, proposes a probabilistic data integration framework to model the relation on a population level. The estimators are obtained with maximum likelihood using an EM algorithm that can deal with high dimensional variables. A discussion is included at the end,

summarizing the methodology in this thesis and discussing, with a case study, future directions in omics data integration.

In Chapter 2, omics data integration in population cohorts is evaluated and discussed. Here, transcriptomics and metabolomics data from the DILGOM study are decomposed in joint, specific and residual parts using O2PLS. The methodology in Chapter 2 simultaneously estimates a relation as in (1.6) between transcripts and metabolites, using O2PLS. The top genes in the resulting joint transcriptomic components (represented by $t$ in (1.6)) as well as top metabolites in the joint metabolomic components (represented by $u$) are further investigated using pathway analyses and compared to the previous results.

In Chapter 3, a free and open-source software package that implements O2PLS is proposed. The aim of this Chapter is to facilitate the use of O2PLS for high dimensional omics data, by developing a memory-efficient algorithm and providing several visualization tools. The package, OmicsPLS, is evaluated in a simulation study in terms of accuracy and speed and compared to an implementation of JIVE [26]. OmicsPLS is also applied to genetic and glycomic data from the Korcula cohort to investigate how IgG1 glycans are related to genetic variants.

Chapter 4 presents a probabilistic method to integrate two homogeneous omics datasets from large population cohorts, based on the model in (1.3). The proposed model, Probabilistic PLS (PPLS), is inspired by PLS in the sense that it also uses isometric normal distributions for the residual latent variables. The PPLS model is fitted by maximizing the likelihood using an EM algorithm. Identifiability constraints are incorporated by solving a constrained optimization in the M step, and the asymptotic standard errors are derived. The PPLS model is applied to data from the Korcula and Vis cohorts separately, and the results are compared.

Chapter 5 extends the probabilistic framework of PPLS to accomodate heterogeneous omics data. The resulting method, Probabilistic O2PLS (PO2PLS), is similar to (1.6) with identifiable parameters. Maximum likelihood estimates are calculated with EM, as well as standard errors. In a simulation study, the PO2PLS method is compared to PLS, PPLS, O2PLS and SIFA in terms of interpretation and prediction performance. PO2PLS is then applied in two data integration analyses: transcriptomics and metabolomics from the DILGOM cohort (Chapter 2), and genetics and glycomics from the Korcula cohort (Chapter 3). The results are compared with O2PLS: in the first data analysis we compare pathway interpretation of the top genes, in the second analysis we validate the results in the Vis cohort.

Finally, in Chapter 6, we briefly review the methodology presented in this thesis. Further extensions are discussed regarding omics data integration in population cohorts, where the sampling design is not standard; it cannot be assumed that the observations on $x$ and $y$ are i.i.d. A case study is presented using data on genetics, methylation and triglycerides, measured on several time points where subjects are grouped in families. The chapter concludes with other future directions to extend the data integration methodology presented in this thesis.

# Bibliography

[1] T. W. Anderson and H. Rubin. Statistical Inference in Factor Analysis. *Proc. Third Berkeley Symp. Math. Stat. Probab.*, 5:111–150, 1956.

[2] A. L. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Br. Bioinform*, 8(1):32–44, 2007.

[3] R. D. Cook and X. Zhang. Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25, 2015.

[4] F. H. C. Crick. Central Dogma of Molecular Biology, 1970.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B*, 39(1):1–38, 1977.

[6] C. García, J. García, M. López Martín, and R. Salmerón. Collinearity: revisiting the variance inflation factor in ridge regression. *J. Appl. Stat.*, 42(3):648–661, 2015.

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Comput.*, 16(12):2639–2664, dec 2004.

[8] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, jan 2009.

[9] M. Inouye, J. Kettunen, P. Soininen, K. Silander, S. Ripatti, L. S. Kumpula, E. Hämäläinen, P. Jousilahti, A. J. Kangas, S. Männistö, M. J. Savolainen, A. Jula, J. Leiviskä, A. Palotie, V. Salomaa, M. Perola, M. Ala-Korpela, and L. Peltonen. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol. Syst. Biol.*, 6(441):441, dec 2010.

[10] M. Inouye, K. Silander, E. Hamalainen, V. Salomaa, K. Harald, P. Jousilahti, S. Männistö, J. G. Eriksson, J. Saarela, S. Ripatti, M. Perola, G.-J. B. van Ommen, M.-R. Taskinen, A. Palotie, E. T. Dermitzakis, and L. Peltonen. An Immune Response Network Associated with Blood Lipid Levels. *PLoS Genet.*, 6(9):e1001113, sep 2010.

[11] R. I. Jennrich and P. M. Bentler. Exploratory Bi-factor Analysis. *Psychometrika*, 76(4):537–549, 2011.

[12] K. G. Jöreskog, U. H. Olsson, and F. Y. Wallentin. *Multivariate Analysis with LISREL*. Springer Series in Statistics. Springer International Publishing, 2016.

[13] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, 44(D1):D457–D462, jan 2016.

[14] A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *J. Mach. Learn. Res.*, 14:965–1003, 2013.

[15] G. Lauc, A. Essafi, J. E. Huffman, C. Hayward, A. Knežević, J. J. Kattla, O. Polašek, O. Gornik, V. Vitart, J. L. Abrahams, M. Pučić, M. Novokmet, I. Redžić, S. Campbell, S. H. Wild, F. Borovečki, W. Wang, I. Kolčić, L. Zgaga, U. Gyllensten, J. F. Wilson, A. F. Wright, N. D. Hastie, H. Campbell, P. M. Rudd, and I. Rudan. Genomics meets glycomics-the first gwas study of human N-glycome identifies HNF1A as a master regulator of plasma protein fucosylation. *PLoS Genet.*, 6(12):1–14, 2010.

[16] G. Li and S. Jung. Incorporating Covariates into Integrated Factor Analysis of Multi-View Data. *Biometrics*, 73(4):1433–1442, dec 2017.

[17] Y. Li, F.-X. Wu, and A. Ngom. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.*, 19(2):325–340, dec 2018.

[18] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.*, 7(1):523–542, 2013.

[19] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *J. Roy. Stat. Soc. Ser. B*, 44:226–233, 1982.

[20] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, 1979.

[21] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.*, 17(October 2015):bbv108, 2016.

[22] X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[23] H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, 47(D1):D419–D426, jan 2019.

[24] L. H. Nguyen and S. Holmes. Ten quick tips for effective dimensionality reduction. *PLOS Comput. Biol.*, 15(6):e1006907, jun 2019.

[25] J. Nocedal and S. Wright. *Numerical optimization*. Springer, 2006.

[26] M. J. O'Connell and E. F. Lock. R.JIVE for exploration of multi-source molecular data. *Bioinformatics*, 32(June):btw324, 2016.

[27] T. Padayachee, T. Khamiakova, Z. Shkedy, M. Perola, P. Salo, and T. Burzykowski. The Detection of Metabolite-Mediated Gene Module Co-Expression Using Multivariate Linear Models. *PLoS One*, 11(2):e0150257, feb 2016.

[28] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.*, 16(2):85–97, 2015.

[29] E. Saccenti, H. C. J. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. W. B. Hendriks. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3):361–374, 2014.

[30] M. Schouteden, K. Van Deun, T. F. Wilderjans, and I. Van Mechelen. Performing DISCO-SCA to search for distinctive and common information in linked data. *Behav. Res. Methods*, 46(2):576–587, nov 2013.

[31] G. A. F. Seber and A. J. Lee. *Linear regression analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2003.

[32] S. D. Silvey. Multicollinearity and Imprecise Estimation. *J. R. Stat. Soc. Ser. B*, 31(3):539–552, sep 1969.

[33] M. Tenenhaus. Pls Regression and Pls Path Mod- Eling for Multiple Table Analysis. *Analysis*, 2004.

[34] R. Tissier, J. Houwing-Duistermaat, and M. Rodríguez-Girondo. Improving stability of prediction models based on correlated omics data by using network approaches. *PLoS One*, 13(2):e0192853, feb 2018.

[35] J. Trygg and S. Wold. O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.*, 17(1):53–64, 2003.

[36] UMetrics. SIMCA O2PLS software, 2017.

[37] F. M. van der Kloet, P. Sebastián-León, A. Conesa, A. K. Smilde, and J. A. Westerhuis. Separating common from distinctive variation. *BMC Bioinformatics*, 17(S5):S195, dec 2016.

[38] A. W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 1998.

[39] H. Wang, Q. Liu, and Y. Tu. Interpretation of partial least-squares regression models with VARIMAX rotation. *Comput. Stat. Data Anal.*, 48(1):207–219, jan 2005.

[40] J. C. Westland. *Structural Equation Models*, volume 22 of *Studies in Systems, Decision and Control*. Springer International Publishing, Cham, 2015.

[41] H. Wold. Partial least squares. *Encycl. Stat. Sci.*, 6(2):581–591, 1985.

[42] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, 1983.

[43] I. S. L. Zeng and T. Lumley. Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science). *Bioinform. Biol. Insights*, 12:117793221875929, jan 2018.

[44] J. Zierer, C. Menni, G. Kastenmüller, and T. D. Spector. Integration of 'omics' data in aging research: from biomarkers to systems biology. *Aging Cell*, 14(6):933–944, dec 2015.