# Statistical integration of diverse omics data

Bouhaddani, S. el

# 5

# Statistical integration of heterogeneous data with PO2PLS

## 5.1 Abstract

We propose Probabilistic O2PLS (PO2PLS), which is a reformulation of O2PLS and an extension of Probabilistic PLS, to describe the relation between all predictors and response variables by joint and specific latent variables. The set of predictors and responses are potentially high dimensional, where the number of variables exceed the sample size, and highly correlated, both within and between the sets of variables. In the PO2PLS model, the joint latent variables explain correlations across the two sets, while the specific variables explain covariance structure within each set. These latent variables are typically of much smaller dimensions than the original sets of variables.

The PO2PLS model is identifiable, and the parameters are estimated with maximum likelihood. A memory-efficient EM algorithm is used to simultaneously estimate joint and specific components under the identifiability constraints. Furthermore, the observed Fisher information is derived analytically and asymptotic standard errors for the parameters are obtained.

We investigate the performance of PO2PLS in terms of estimation accuracy and prediction quality, and compared it to several alternatives. Also, we apply PO2PLS in two different cohorts: in the first cohort, we consider transcriptomics and metabolomics data, in the second cohort, genetic and glycomic data are considered. We compare the results and top variables in the joint parts with previous literature and genomic databases. Results show that PO2PLS overall outperforms the alternative methods in the simulation study in terms of accuracy and prediction performance. In the data analyses, PO2PLS yields better interpretation of the joint latent variables and lower prediction error in a independent test cohort.

## 5.2 Introduction

The multivariate linear model is widely used to describe the relationship between a vector of response variables $y \in \mathbb{R}^q$ and a vector of predictors $x \in \mathbb{R}^p$. Although the model is general, we focus here on epidemiological applications in life sciences, where nowadays multiple 'omics' data are available, reflecting variation at several biological levels for the same set of subjects [19]. The underlying aim of these applications is to describe the relation between $x$ and $y$, where $x$ and $y$ are often high dimensional and correlated.

Studies that try to understand how $x$ and $y$ are related, have mainly been focusing on detecting and interpreting pair-wise relationships between, say, $x_i$ and $y_j$ (e.g. see [33]). However, within studies where many subjects are collected, a multivariate model for the relationship between $x$ and $y$ is feasible and can yield better interpretation of the true underlying mechanisms (also denoted by data integration, see [32]). These mechanisms can be described by unobservable random variables in a latent space of smaller dimensions than the original space of the observable variables (as is the case with, e.g., genetic or metabolic pathways [43]). In addition, these datasets are typically heterogeneous, where the sets of measurements on $x$ and $y$ may differ substantially in dimensions, scale, measurement platform and distribution [39].

In this paper, we consider datasets from two studies that contain data on several omics levels. These omics data are heterogeneous as they differ considerably,

in particular in dimensionality and measurement platform. In the DILGOM study, gene expression ($p \approx 10^4$) and metabolite abundances ($q = 137$) were measured on $N = 512$ participants [17, 18]. The aim was to identify molecular pathways underlying lipid metabolism and gene expression, in particular in relation to inflammation. Here, we address the same aim using multivariate In the second study, the Croatian Korcula study, genetic variants ($p \approx 10^5$) and IgG glycosylation ($q = 20$) are available on $N = 885$ subjects [21, 40]. The aim here was to identify genetic regions underlying changes in the glycosylation of immunoglobulins (Ig, i.e. antibodies).

We consider multivariate latent variable models that perform dimension reduction and take into account correlations within and between the datasets, also referred to as "omics data integration". Several methods have been proposed [32] to describe the relationship between $x$ and $y$. The majority of these methods are based on linear dimension reduction techniques [29] that construct 'joint parts' as an estimate of the common part of $x$ and $y$. Typically, these joint parts are composed of Joint Principal Components (JPCs): projections of the data that maximize a measure of relationship, such as covariance or correlation. However, to take into account the heterogeneity between $x$ and $y$ and improve model interpretation, specific parts should be included [38, 39]. Also, to facilitate statistical inference, probabilistic models with identifiable parameters are needed [8]. Furthermore, within a probabilistic framework, extensions can be made to accommodate complex study designs (e.g. multilevel and family data).

Among the methods that only include joint parts are Partial Least Squares (PLS) [41, 1], Canonical Correlation Analysis (CCA) [16] and Envelope Regression [4]. Methods that also incorporate specific parts are Two-way Orthogonal PLS (O2PLS) [38] and JIVE [25], where the former is more flexible than the latter in that it does not assume orthogonality between the joint and specific components.

Regarding probabilistic models for data integration, Probabilistic PLS (PPLS) has been developed [8] for homogeneous data, where no specific parts need to be estimated. In the machine learning community, Bayesian CCA (BCCA) was proposed as a prediction model to detect dependencies between multiple datasets, while correcting for data-specific parts [20]. Although the model is probabilistic, it is not identifiable, which is a drawback when parameter interpretation and inference is sought. Recently, a probabilistic method named Supervised Integrated Factor Analysis (SIFA) was proposed for estimating joint and specific components in multiple datasets [24]. However, SIFA assumes homogeneity of the JPCs, which is not realistic for heterogeneous omics data (see paragraph 5.3.1 and [9]).

We propose Probabilistic O2PLS (PO2PLS), a flexible probabilistic model that estimates JPCs and specific components in two heterogeneous datasets. Here, the JPCs are not assumed to be homogeneous nor required to be orthogonal to the specific components. The model parameters are identifiable, and estimated with maximum likelihood. A memory-efficient ECM algorithm [30] is used, where in each step a constrained optimization is solved to obtain identifiable estimates, even with high dimensional data.

The remainder of the paper is organized as follows. In Section 5.3, the PO2PLS model is developed and identifiability of the parameters is shown. Furthermore, maximum likelihood estimates are derived. In Section 5.4, the performance of PO2PLS is studied in a range of simulation scenarios. We focus on interpretability, but also evaluate prediction performance. In Section 5.5, PO2PLS is applied as illustration to

perform statistical integration of the datasets in the DILGOM and Korcula studies.

## 5.3 PO2PLS: model and estimation

### 5.3.1 The model

Let $x$ and $y$ be two random row-vectors of size $p$ and $q$, respectively. Note that $p$ and $q$ do not have to be equal. In the PO2PLS model, both $x$ and $y$ are expressed in terms of a joint part, a specific part, and a noise part. The joint parts involve random vectors $t$ and $u$ of size $r$, where $u$ is dependent on $t$. The specific parts involve independent random vectors $t_\perp$ and $u_\perp$ of size $r_x$ and $r_y$, respectively. The noise random vectors are denoted by $e$ ($p$-dimensional), $f$ ($q$-dimensional) and $h$ ($r$-dimensional). Here, $h$ represents heterogeneity in the joint parts. More precisely, the PO2PLS model for $x$ and $y$ is described by

$$\begin{aligned} x &= tW^{\mathrm{T}} + t_\perp W_\perp^{\mathrm{T}} + e \\ y &= uC^{\mathrm{T}} + u_\perp C_\perp^{\mathrm{T}} + f \\ u &= tB + h \end{aligned} \tag{5.1}$$

The parameter matrices $W$ ($p \times r$) and $C$ ($q \times r$) are called joint loadings. The matrices $W_\perp$ ($p \times r_x$) and $C_\perp$ ($q \times r_y$) are referred to as data-specific loadings.

The random vectors $e$ and $f$ are independent multivariate normally distributed random vectors, with zero mean and covariance matrices $\sigma_e^2 I_p$ and $\sigma_f^2 I_q$, respectively. Furthermore, $t$, $t_\perp$, $u_\perp$ and $h$ are zero mean multivariate normals, with diagonal covariance matrices $\Sigma_t$, $\Sigma_{t_\perp}$, $\Sigma_{u_\perp}$ and $\Sigma_h$, respectively. The covariance matrix of $u$ follows from (5.1): $\Sigma_u = B^{\mathrm{T}}\Sigma_t B + \Sigma_h$. Here, $B$ is a diagonal $r \times r$ matrix.

All parameters are collected in $\theta := [W, W_\perp, C, C_\perp, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \Sigma_h, \sigma_e^2, \sigma_f^2]$. It parametrizes the distribution of $(x, y) \sim \mathcal{N}(0, \Sigma_\theta)$ (the explicit expression for $\Sigma_\theta$ is given in the supplementary material).

The model for the relation between $u$ and $t$ is taken asymmetrically, as often a certain hierarchy is assumed for $x$ and $y$ [5]. For instance, it is reasonable to assume that genetic variability induces glycomic variation, therefore we model $u$ in terms of $t$.

**PO2PLS as a general data integration framework.** PO2PLS models the relationship between $x$ and $y$ through $t$ and $u$ as described in (5.1). The correlation between $u$ and $t$ is determined by $\Sigma_t$, $B$ and $\Sigma_h$. If the JPCs are assumed to be homogeneous, i.e. $u = t$, the PO2PLS model reduces to the SIFA and BCCA models. In this case, $B = I$ and $\Sigma_h = 0$, so $u$ and $t$ have the same scale and a correlation of one. Especially for heterogeneous omics data, the two sets of JPCs typically represent different biological mechanisms (e.g. genetic versus metabolic pathways). Therefore, they are biologically not perfectly correlated or on the same scale. Also, it has been shown that assuming homogeneity of JPCs can negatively affect estimation performance [9]. If additionally, the columns of the concatenated components ($WW_\perp$) and ($CC_\perp$) are orthogonal, the JIVE model is recovered. In this case, combinations of features involved in the joint and specific parts have to be orthogonal, which is a strong restriction and not likely to hold for omics data. The Probabilistic PLS model

is obtained by setting $\Sigma_{t_\perp}$ and $\Sigma_{u_\perp}$ to zero in (5.1). Therefore, PO2PLS can be seen as a framework in which several other data integration methods are contained.

Methods like CCA and Envelope Regression (ER) have similar models as PPLS and PO2PLS. However, the number of noise variance parameters to estimate is of order $O(p+q)$, whereas PO2PLS introduces one $\sigma_e^2$ and $\sigma_f^2$ for $x$ and $y$, respectively. When $p$ or $q$ is larger than the sample size (i.e. a high dimensional setting), CCA and ER estimators cannot be obtained due to singularity issues. Therefore, such models cannot be used for omics data integration.

In Table 5.1 an overview is shown with several methods and their features.

### 5.3.2 Identifiability of PO2PLS

Linear latent variable models are typically unidentifiable due to rotation indeterminacy of the loading components. For example, given a rotation matrix $R$ such that $RR^{\mathrm{T}} = I$, the models $x = tW^{\mathrm{T}}$ and $x = tRR^{\mathrm{T}}W^{\mathrm{T}}$ yield the same distribution for $x$. In PCA, the loading matrices are restricted to be semi-orthogonal, i.e. $W^{\mathrm{T}}W = I$, whereas in Factor analysis, the latent variables are standard normally distributed. However, these assumptions separately do not solve the rotation indeterminacy. In PO2PLS, identifiability can be obtained using assumptions similar to the two just mentioned, namely semi-orthogonal loading matrices and diagonal covariance matrices for the latent variables. Note that this generally leads to a constrained optimization over Stiefel manifolds, as we will see in subsection 5.3.3.

The assumptions in PO2PLS are firstly, $W^{\mathrm{T}}W = C^{\mathrm{T}}C = I_r$, $W_\perp^{\mathrm{T}}W_\perp = I_{r_x}$ and $C_\perp^{\mathrm{T}}C_\perp = I_{r_y}$. Additionally, $[WW_\perp]$ and $[CC_\perp]$ must not have linearly dependent columns. Note that the columns of $W_\perp$ and $C_\perp$ do not have to be orthogonal to the columns of $W$ and $C$, respectively. Second, the diagonal elements of $B$ are restricted to be non-negative. This does not restrict the PO2PLS model, as $t_k b_k$ is equal to $-t_k b_k$ in distribution, for $k = 1, \ldots, r$. Finally, the sequence $(\sigma_{t_k}^2 b_k)_{k=1}^r$ is assumed to be strictly decreasing in $k$. Regarding the number of components, we assume that $0 < r + r_x < p$ and $0 < r + r_y < q$, where $r$ is positive and both $r_x$ and $r_y$ are non-negative.

Given these assumptions, the loading matrices are identified up to sign. The other

Table 5.1: **An overview of several data integration methods and their features.** An 'X' indicates presence of a feature. The abbreviations 'High dim.', 'Probab' and 'Het. JPCs' stand for 'High dimensional estimation', 'Probabilistic' and 'Heterogeneous Joint Principal Components', respectively.

| Features | PLS | PPLS | CCA | BCCA | ER | O2PLS | JIVE | SIFA | PO2PLS |
|---|---|---|---|---|---|---|---|---|---|
| Joint | X | X | X | X | X | X | X | X | X |
| Specific | | | | X | X | X | X | X | X |
| High dim. | X | X | | X | | X | X | X | X |
| Probab. | | X | | X | X | | | X | X |
| Het. JPCs | X | X | | | X | X | | | X |

parameters in $\theta$ are uniquely identified. The following Theorem makes this precise.

**Theorem 5.3.1.** *Let $r$, $r_x$, $r_y$ and $\theta$ satisfy the above assumptions. Let $\Sigma_{\theta_1}$ and $\Sigma_{\theta_2}$ be the covariance matrices corresponding to PO2PLS parameters $\theta_1$ and $\theta_2$, and suppose $\Sigma_{\theta_1} = \Sigma_{\theta_2}$. Then $W_1 = W_2\Delta_W$, $C_1 = C_2\Delta_W$, $W_{\perp 1} = W_{\perp 2}\Delta_{W_\perp}$, $C_{\perp 1} = C_{\perp 2}\Delta_{C_\perp}$ for diagonal orthogonal matrices $\Delta_W, \Delta_{W_\perp}$ and $\Delta_{C_\perp}$, and all other parameters in $\theta_1$ and $\theta_2$ are equal.*

The proof is given in the supplementary material.

### 5.3.3 Maximum Likelihood Estimation of the parameters

We propose the maximum likelihood method to estimate $\theta$. Contrary to O2PLS, the estimation is simultaneous over both joint and specific parts. The log of the likelihood associated with the PO2PLS model (5.1) is given by

$$L(\theta|x,y) = -\frac{1}{2}\left\{(p+q)\log(2\pi) + \log|\Sigma_\theta| + (x,y)\Sigma_\theta^{-1}(x,y)^{\mathrm{T}}\right\}. \tag{5.2}$$

Note that $L$ is a complicated and highly non-linear function of $\theta$, and its computation requires computing and storing covariance matrices of size $(p+q)^2$. If the latent variables $t$, $u$, $t_\perp$ and $u_\perp$ would be observable, maximizing the log-likelihood becomes analytically tractable and computationally feasible, even for large $p$ and $q$. Therefore, we propose an EM algorithm to obtain maximum likelihood estimates for $\theta$.

Denote the complete data vector by $(x, y, t, u, t_\perp, u_\perp)$. For a current estimate $\theta'$, the EM algorithm considers the objective function

$$Q(\theta|x,y,\theta') := \mathbb{E}_{\theta'}\left[\log f(x,y,t,u,t_\perp,u_\perp|\theta)|x,y\right]. \tag{5.3}$$

Here, the complete data likelihood can be written (with abuse of notation) as

$$f(x,y,t,u,t_\perp,u_\perp|\theta) = \underbrace{f(x|t,t_\perp)}_{W,W_\perp,\sigma_e^2}\underbrace{f(y|u,u_\perp)}_{C,C_\perp,\sigma_f^2}\underbrace{f(u|t)}_{B,\Sigma_h}\underbrace{f(t)}_{\Sigma_t}\underbrace{f(t_\perp)}_{\Sigma_{t_\perp}}\underbrace{f(u_\perp)}_{\Sigma_{u_\perp}}. \tag{5.4}$$

These factors depend on distinct sets of parameters, yielding separate optimization problems.

The Expectation step involves a conditional expectation of the complete data likelihood. Since $f$ in (5.3) is a multivariate normal density, this expectation can be written in terms of the first and second conditional moments of the latent variables $t$, $u$, $t_\perp$ and $u_\perp$ given $x$ and $y$. Focusing on the first factor in (5.4), the conditional expectation of $\log f(x|t,t_\perp)$ is given by

$$-\frac{1}{2}\left\{Np\log(2\pi) + Np\log\sigma_e^2 + \sigma_e^{-2}\mathrm{tr}\,\mathbb{E}_{\theta'}\left[||x - tW^{\mathrm{T}} - t_\perp W_\perp^{\mathrm{T}}||_F^2\,|x,y\right]\right\}. \tag{5.5}$$

This expectation involves first and second conditional moments of the vector $(t, t_\perp)$ given $\theta'$, $x$ and $y$. These terms can be explicitly calculated and are given in the supplementary material.

In the Maximization step, the function in (5.5) is optimized over all semi-orthogonal matrices $W$ and $W_\perp$ (i.e. over the $V_{p,r}$ and $V_{p,r_x}$ Stiefel manifolds). To enforce semi-orthogonality, we introduce Lagrange multipliers $\Lambda_W$ and $\Lambda_{W_\perp}$. Maximizing (5.5)

over semi-orthogonal $W$ and $W_\perp$ is then equivalent to minimizing the following objective function

$$\mathbb{E}_{\theta'}\left[||x - tW^{\mathrm{T}} - t_\perp W_\perp^{\mathrm{T}}||_F^2 \,|x, y\right] + \Lambda_W\left(W^{\mathrm{T}}W - I_r\right) + \Lambda_{W_\perp}\left(W_\perp^{\mathrm{T}}W_\perp - I_{r_x}\right). \quad (5.6)$$

Note that the objective function involves both $W$ and $W_\perp$ and cannot be decoupled. Instead of numerical optimization, we consider a variant of EM that performs sequential optimization [30]. First, (5.6) is minimized over $W$, keeping $W_\perp$ constant. Then we minimize over $W_\perp$, keeping $W$ equal to its minimizer. Under standard conditions, this algorithm monotonically approaches a (local) maximum of the observed likelihood $L$ [30].

The above derivation is conditional on the dimensions of the latent spaces. Typically, the number of components $r$, $r_x$ and $r_y$ are unknown a priori. Strategies that can be used to select the number of PO2PLS components include cross-validation [11] and eigenvalue plots [28].

The expectation and maximization step for the other parts in (5.4) are calculated analogously (see the supplementary material). In this calculation, the following operator is used to obtain semi-orthogonal loading matrices.

**Definition 5.3.2.** *Let $A$ be a $p \times a$ full rank matrix with singular value decomposition $A = UDV^{\mathrm{T}}$. Let $R = VD$. Then we define the operator orth : $\mathbb{R}^{p\times a} \to \mathbb{R}^{p\times a}$ as*

$$\operatorname{orth}(A) = A(R^{\mathrm{T}})^{-1}. \quad (5.7)$$

Using this operator, the EM parameter updates are made explicit in Theorem 5.7.1, Appendix 5.7.

**Standard errors.** Maximum likelihood theory entails that, under regularity conditions, the estimator $\hat{\theta}$ has asymptotic distribution $\mathcal{N}(\theta, \Sigma_\theta)$, as the sample size goes to infinity. This also holds in factor analysis models [35]. By calculating the square root of $\Sigma_\theta$, standard errors for $\hat{\theta}$ are obtained. A well-known approach for estimating $\Sigma_\theta$ is the inverse observed Fisher information matrix. In an EM algorithm, this matrix is given by [26]:

$$\mathbb{E}\left[B(\hat{\theta})|X, Y\right] - \mathbb{E}\left[S(\hat{\theta})S(\hat{\theta})^{\mathrm{T}}|X, Y\right]. \quad (5.8)$$

Here, $S(\hat{\theta}) = \nabla L(\hat{\theta})$ and $B(\theta) = -\nabla^2 L(\hat{\theta})$ are the gradient and negative of the second derivative of the log likelihood $L$, respectively, evaluated in $\hat{\theta}$. The Fisher information matrix for PO2PLS is derived in the supplementary material.

## 5.4 Simulation study

We conduct a simulation study to evaluate the performance of PO2PLS estimates in several scenarios. We focus on interpretability of the estimators, but also consider predictive performance. Furthermore, PO2PLS is compared to PLS, O2PLS, PPLS and SIFA.

In the simulation scenarios, combinations of small and large sample sizes ($N = 100, 1000$), low and high dimensional $x$ and $y$ ($p = 2000, 10000; q = 25, 125$) are

considered. We additionally include two scenarios for the proportion of noise relative to the total variation: in the 'small noise proportion', we take 40% noise in both $x$ and $y$. In the 'large noise proportion' these values are 95% and 5% for $x$ and $y$, respectively. The impact of heterogeneity of joint parts is considered by increasing the joint residual variance $\Sigma_h$ from 0% to 80% of the total joint variance $\Sigma_u$. These scenarios are based on the two data analyses in Section 5.5.

The simulated data are generated from the PO2PLS model (5.1), with normally distributed latent variables and $r = r_x = r_y = 5$. In the homogeneous joint parts scenarios, we set $B = I$ and $\Sigma_h = 0$ to comply with the SIFA assumptions described in paragraph 5.3.1. The parameter values are drawn at random and kept fixed during simulation. The PO2PLS restrictions, described in Subsection 5.3.2, are then applied to these parameters.

Interpretability of each component is derived from the subset of variables that have the highest absolute loading value, since, in applications, the top features are followed-up for further investigation. For each joint component, we evaluate interpretability by sorting the estimated loading values and then calculating the proportion of true top 25% features among the estimated top 25% (i.e True Positives Rate, TPR). We then average these proportions across components to obtain an aggregated proportion across joint components. Regarding predictive performance, we calculate the RMSEP, defined as the square root of the average value of $||y - \hat{y}||_F^2$. Here, the RMSEP is calculated in both training and test data; the test data consist of $N = 10^4$ independent samples generated from the same model as the training data.

The EM algorithm is stopped if the log-likelihood increment is below $10^{-6}$ or $10^3$ steps are taken. For each setting in the normal distribution scenario, 1000 replicates are generated. We additionally match the sign and order of the estimated and true components.

Furthermore, we investigate the impact of rank misspecification and violation of the normality assumption. We fit the model using one component too many in the joint and specific parts, and we apply the algorithm to data generated from non-normally distributed latent variables. We consider a $t_2$, a Poisson $P_1$ and a binomial $B_{2,0.25}$ distribution, reflecting characteristics typically observed in omics data, such as heavy tailed, skewed and discrete measurements.

### 5.4.1 Results

The current implementation of SIFA (found on `GitHub:reagan0323/SIFA`) was unable to produce estimates in the scenarios where $p = 10000$. Therefore, we discuss only the lower dimensional setting and refer to the supplementary material regarding results in the high dimensional scenarios.

In Figure 5.1, boxplots of the TPR, difference in TPR with respect to PO2PLS, and RMSEP, respectively, are shown across several settings. In the RMSEP plot, the black line represents the prediction error when using the true parameter values on the test data.

Firstly, the TPR of PO2PLS and SIFA was generally above 25%, even for very noisy and heterogeneous data. PPLS underperformed in most scenarios compared to the other methods. O2PLS and PLS behaved very similar. Furthermore, when considering paired differences of the TPR between each method and PO2PLS, it can

be seen that PO2PLS generally has higher TPR within the simulation runs. This difference is more notable under large noise proportion and heterogeneous joint parts settings. Note that the difference of PPLS and PO2PLS is left out for better visual comparison of the other methods.

Secondly, regarding the prediction error in the heterogeneous joint parts scenarios, PO2PLS performs better than O2PLS, PPLS and SIFA, and has more realistic training error than PLS. Here, SIFA has the highest prediction error. Furthermore, PLS and O2PLS overfit to the data in noisy, small sample size scenarios.

The results for the other settings (rank misspecification, non-normal variables, and high dimensional data) are shown in the supplement.

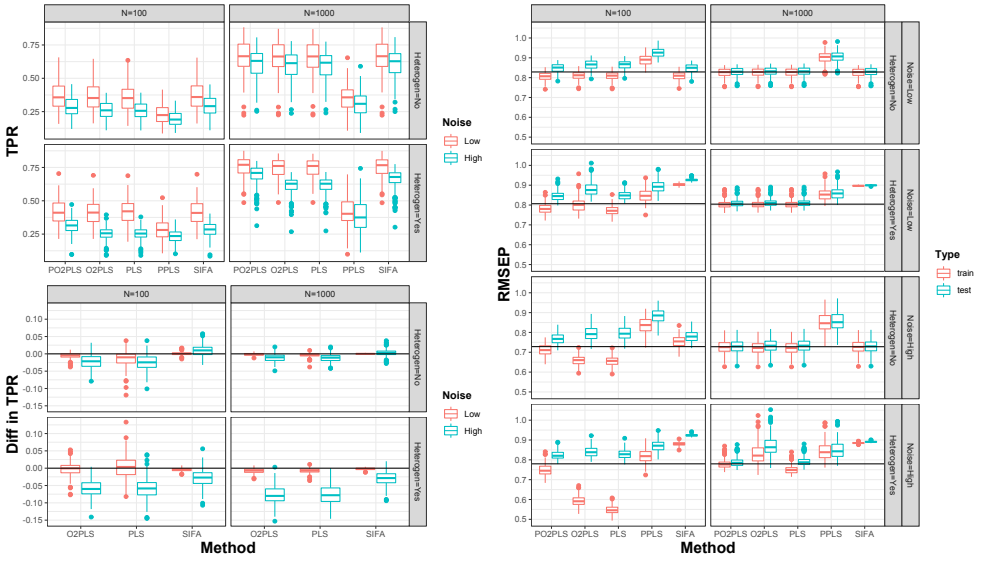


Figure 5.1: **True Positives Rate and Root MSE of Prediction.** *Upper left figure*: proportion of true top 25% among estimated top 25% (TPR) for each method, stratified by simulation scenario. *Lower left figure*: Difference in TPR of several methods and PO2PLS. Lower values are in favor of PO2PLS. In both plots, the left red boxplots correspond with the low noise scenario, the right blue boxplots with high noise. *Right figure*: Root mean squared error of prediction stratified by method and scenario. The black line represent the test error when using the true parameter values. The left red boxplots correspond with the training error, the right blue boxplots with the test error.

## 5.5 Analysis of heterogeneous omics data

As application of the PO2PLS model, we consider two analyses of heterogeneous omics data. Firstly, using PO2PLS, joint and specific variation in gene expression and metabolite levels is estimated. Secondly, PO2PLS is applied to estimate genetic contributions to glycomic variation. The results are compared with results obtained

with O2PLS.

SIFA was also applied to data from the two cohorts. However, the current implementation could not cope with the dimensionality of these data, due to unrealistic time and memory requirements: in the first analysis, one iteration took 30 minutes where 500 iterations is the default; in the second analysis, an out of memory error was issued.

**Transcriptomic–metabolomic analysis.** The link between lipid metabolite and gene expression levels has been investigated in the literature, especially in the context of coronary artery disease and atherosclerosis [2, 18]. Here, it was found that very-low-density- and high-density-lipoprotein (VLDL and HDL) metabolite concentrations were associated with expression levels of genes involved in inflammation and allergy [17]. The same data were analyzed with O2PLS, and was consistent with earlier reports [7].

In this data application, our aim is to elucidate joint and specific factors regarding lipid metabolites and gene expression. To this end, we apply PO2PLS to obtain joint and specific components from the transcriptomic and metabolomic measurements.

After pre-processing and filtering, data on $p = 7385$ expression probes and $q = 134$ metabolite concentrations were available for $N = 512$ participants from the DILGOM cohort. These data are denoted by $X$ and $Y$, respectively. For the analysis, 2 joint, 1 expression-specific, and 10 metabolite-specific components were retained (see Supplementary Material).

In Figure 5.2, the two metabolomic joint components (explaining 56% of total variance) are shown. Note that clusters of VLDL- LDL- and HDL-type metabolites are observed. The first joint component mostly represented VLDL metabolites, while in the second component VLDL and HDL were represented. The top 500 probes of the two transcriptomic joint components (explaining 28% of total variance) mapped to genes mostly involved in membrane localization and immune response, respectively.

For comparison, we applied O2PLS to these data. The metabolic components are similar to those reported in Figure 5.2. However, the gene annotation clusters are different, see Table 5.2. In particular, the first and second components of PO2PLS are better separated in terms of top gene annotation clusters. They also reflect current biological literature; the role of VLDL in membrane transport is well known [10], and the relation between VLDL, HDL and immunology genes has been investigated previously [17].

**Genetic–glycomic analysis.** Glycosylation is one of the most common post translational modifications that enriches the functionality of proteins in many biological processes, such as cell signaling, immune response and apoptosis [40]. Several studies have linked the composition of glycans to the risk and status of several diseases [13, 23, 37].

Glycan synthesis does not have a genetic template, rather, many glycosyltransferases and DNA binding proteins are involved in glycosylation [14]. Genetic regulation of glycosylation has been investigated, where SNP-glycan pairs were separately considered[21, 31, 40]. However, individual glycans abundances are highly correlated, and can be affected by several genes [36]. Therefore, a multivariate approach is more appropriate.

Figure 5.2: **Metabolomic PO2PLS joint components.** The first component is plotted on the x-axis against the second component. The colors and shapes represent the biological grouping of the metabolites: very-low-, low-, intermediate-, high-density-lipoproteins, fatty acids and others

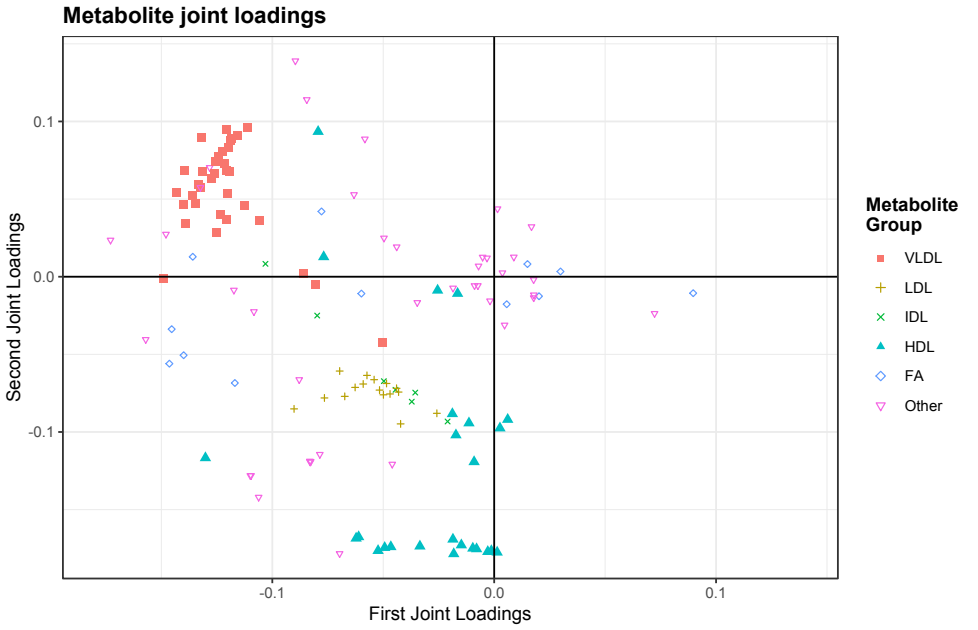


Table 5.2: **Annotation of transcriptomic and metabolomic joint components.** The metabolomic components approximately represent VLDL and HDL metabolite levels. The corresponding top 550 transcript probes are enriched for the given annotations.

| Component | Metabolites | Genes |
|---|---|---|
| 1 | VLDL | protein targeting to membrane; protein localization |
| 2 | VLDL vs HDL | immune/defense response; response to stimulus |

(a) Gene annotation clusters for Probabilistic O2PLS

| Component | Metabolites | Genes |
|---|---|---|
| 1 | VLDL | immune response; protein targeting to membrane |
| 2 | VLDL vs HDL | immune/defense response; response to stress |

(b) Gene annotation clusters for O2PLS

In this data application, our aim is to integrate genetic and glycomic data and understand genetic contributions to variation in glycan abundances. To this end, we apply PO2PLS to estimate the loading parameters and variance components.

We have data on 333858 genotyped Single Nucleotide Polymorphisms (SNPs) and 20 IgG1 glycan abundances, measured with nano-LCMS, for $N = 885$ participants in the Korcula cohort [22]. Both data sets contain highly correlated measurements (median Pearson correlation coefficient of the glycan data: 0.77, IQR: 0.19), and are heterogeneous, since they differ in scale, distribution and measurement error. Recently, these data were also analyzed with O2PLS [9].

Firstly, the SNPs were summarized on gene level, yielding a Genetic PCs (GPCs) dataset. Then, the GPCs and glycomic datasets were pre-processed, resulting in datasets $X$ ($p = 37819$) and $Y$ ($q = 20$), respectively. Furthermore, based on scree plots of the eigenvalues of the data matrices, 5 joint, 5 genetic-specific, and no glycan-specific components were retained [9].

Regarding the five IgG1 glycan joint components, they accounted for 95% of the total modeled IgG1 glycan variation. The amount of IgG1 variation that can be predicted with the Genetic PCs was 17%.

The loading values of each IgG1 glycan variable are depicted in Figure 5.3. They represent different aspects of glycans and their molecular structure, namely the 'average' glycan and presence of fucose, galactose and GlcNAc (last two components), respectively [9].

The five joint components in the Genetic PCs data set accounted for 2.3% of the total modeled variation. For the specific parts, this percentage was 2.6%. The top 500 genes in each Genetic PCs joint component were clustered using GSEA. The relevant clusters are shown in Table 5.3.

The top 500 genes in the Genetic PCs joint component seem to be involved in inflammatory pathways, signaling, transferases, and localized to the Golgi apparatus and Endoplasmic Reticulum.

PO2PLS was also applied to genetic and glycomic data in an independent study consisting of 714 participants from the Croatian Vis cohort. These results were consistent with above findings, indicating that the obtained joint components are not specific to one population.

The PO2PLS and O2PLS models trained in Korcula were then evaluated in Vis, in terms of prediction error of $Y$ given $X$. Here, the ratio of training and test error was 5/23 for O2PLS and 20/21 for PO2PLS, this is conform the simulation study that PO2PLS is less prone to overfitting.

Figure 5.3: **Glycomic PO2PLS joint components.** The components are separately plotted. The colors and shapes represent the biological grouping of the glycans. In the last row and column, a graphical representation of the structure of a particular glycan is shown.

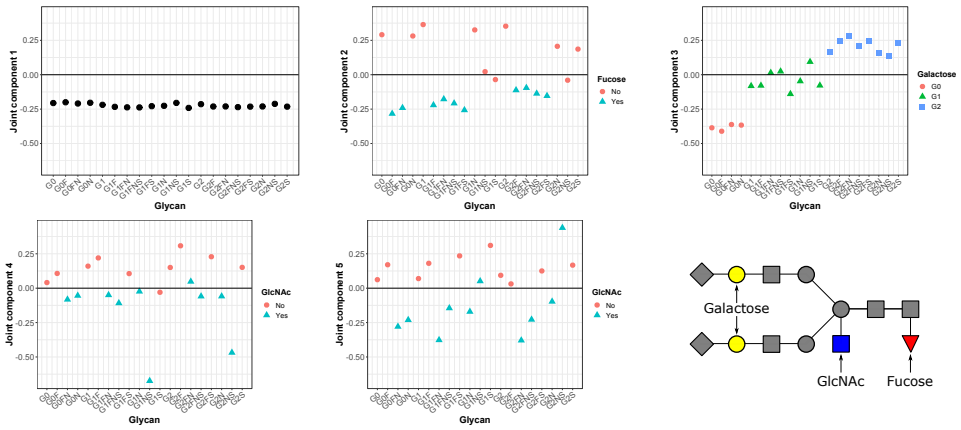


Table 5.3: **Annotation of genetic and glycan PO2PLS joint components.** The glycan components represent presence or absence of key molecules (fucose, galactose and GlcNAc). The corresponding top 500 genes are enriched for the given annotations.

| Component | Glycans | Genes |
|---|---|---|
| 1 | average | Inflammatory pathways; B-cell gene regulation through transcription factors *SP1*, *LEF1* and *ELK1* |
| 2 | fucose | Signalling; stimulus processing and transferases; localized to the Golgi apparatus |
| 3 | galactose | Signalling; stress and immune response; localized to the Golgi apparatus and Endoplasmic Reticulum |
| 4 and 5 | GlcNAc | Transferases |

## 5.6 Discussion

We propose Probabilistic Two-way Orthogonal Partial Least Squares (PO2PLS) to model the relation between two sets of variables, as well as characteristics specific to measurements of each set. The model parameters are shown to be identifiable, and an EM algorithm to compute the constrained maximum likelihood estimator is derived that is able to handle high dimensional data.

A simulation study is presented to evaluate the performance of PO2PLS, and compare it to alternative methods. Except for O2PLS, these alternatives do not model specific parts (PLS, PPLS), or joint heterogeneity (SIFA). Furthermore, SIFA could not be applied to high dimensional data due to computational constraints. These features are reflected in the performance of the corresponding methods. For example, SIFA underperformed in presence of joint heterogeneity. In small sample size and large noise level settings, PLS and O2PLS suffered from overfitting and had lower interpretability. This is contrary to the belief that PLS and O2PLS, as distribution free methods, are more suited in small sample size scenarios [42]. In our simulation study, PO2PLS yields better interpretation and prediction performance.

In both data analyses, the resulting top features could to be biologically linked according to the associated annotations. Moreover, the genetic-glycomic results were replicated in a second independent cohort. Also O2PLS was applied to these data; the biological interpretation was less obvious. Moreover, when comparing the training error of the first cohort with the test error in the second genetic-glycomic cohort, O2PLS overfitted to the first cohort. As concluded from the simulation study, this has a negative effect on interpretation of the top loadings.

When multiple cohorts are available, such as in Section 5.5, a meta-analysis of parameter estimates can be performed for more robust interpretation. Several methods have been proposed in the linear regression and factor analysis framework [6, 15, 3]. Both approaches use the Hessian matrix of the whole parameter vector, which is unfeasible to calculate for high dimensional data. PO2PLS can be extended by adding cohort-common and cohort-specific parameters to the model. Maximum likelihood estimation would then yield an 'optimal shared joint space' that incorporates information from each cohort.

In many epidemiological applications, a univariate outcome $z$ is available. Recent interest lies in using the relationship between molecular data $x$ and $y$ to model $z$ (e.g. [12]). Here, penalized regression approaches can be used with only one set of predictors $x$ or $y$. A more holistic approach is to use the information about $x$ and $y$, summarized by the low-dimensional component space, to explain variation in $z$. Based on the Probabilistic O2PLS framework, a model of $z$ given the joint and specific parts can be added. For example, by adding $z = (t, u, t_\perp, u_\perp)\beta_z + \epsilon_z$ to (5.1), the latent components will explain joint and specific variation in $x$ and $y$ and be associated to $z$.

Within epidemiological studies, two questions often arise regarding omics data integration: what is the effect size of the relationship between two omics data and which features are associated with this relationship. Previous data integration methods focused on estimating effect sizes and applied resampling methods to quantify the uncertainty of these estimates. Especially with high dimensional data, these methods are computationally demanding or difficult to carry out due to a non-standard

study design. As sample sizes are often high, the probabilistic PO2PLS framework provides an alternative approach based on the asymptotic Fisher information matrix (see Section 5.3) to assess significance of the estimated effect size and feature loadings. Alternatively, likelihood ratio tests can be used to compare two nested PO2PLS models. Since, in general, asymptotic ML theory is established for $N$ going to infinity, inference based on PO2PLS would be more reliable in large epidemiological cohorts. Also, the asymptotic behavior of PO2PLS when increasing dimensionality as well as sample size is yet to be investigated.

## 5.7 Appendices for Chapter 5

### Appendix A. An ECM algorithm for PO2PLS

**Theorem 5.7.1.** *Let $X$ and $Y$ be data matrices with $N$ i.i.d. PO2PLS replicates of $(x, y)$ across the rows. Let $r$, $r_x$ and $r_y$ be fixed, satisfying $\max(r + r_x, r + r_y) < N$. The loading matrix $W$ is estimated with the following iterative scheme in $k$, given known starting values for $k = 0$. Here, $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot|X, Y, \theta^k]$.*

$$
\begin{aligned}
W^{k+1} &= \operatorname{orth}\left(X^{\mathrm{T}}\,\mathbb{E}_k\left[T\right] - W_{\perp}^{k}\mathbb{E}_k\left[T_{\perp}^{\mathrm{T}}T\right]\right)\\
W_{\perp}^{k+1} &= \operatorname{orth}\left(X^{\mathrm{T}}\,\mathbb{E}_k\left[T_{\perp}\right] - W^{k+1}\mathbb{E}_k\left[T^{\mathrm{T}}T_{\perp}\right]\right)\\
C^{k+1} &= \operatorname{orth}\left(Y^{\mathrm{T}}\,\mathbb{E}_k\left[U\right] - C_{\perp}^{k}\mathbb{E}_k\left[U_{\perp}^{\mathrm{T}}U\right]\right)\\
C_{\perp}^{k+1} &= \operatorname{orth}\left(Y^{\mathrm{T}}\,\mathbb{E}_k\left[U_{\perp}\right] - C^{k+1}\mathbb{E}_k\left[U^{\mathrm{T}}U_{\perp}\right]\right)\\
B^{k+1} &= \mathbb{E}\left[U^{\mathrm{T}}T\right]\left(\mathbb{E}\left[T^{\mathrm{T}}T\right]\right)^{-1} \circ I_r\\
\Sigma_t^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[T^{\mathrm{T}}T\right] \circ I_r\\
\Sigma_{t_\perp}^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[T_{\perp}^{\mathrm{T}}T_{\perp}\right] \circ I_{r_x}\\
\Sigma_{u_\perp}^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[U_{\perp}^{\mathrm{T}}U_{\perp}\right] \circ I_{r_y}\\
\Sigma_h^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[H^{\mathrm{T}}H\right] \circ I_r\\
(\sigma_e^2)^{k+1} &= \frac{1}{Np}\operatorname{tr}\left(\mathbb{E}_k\left[E^{\mathrm{T}}E\right]\right)\\
(\sigma_f^2)^{k+1} &= \frac{1}{Nq}\operatorname{tr}\left(\mathbb{E}_k\left[F^{\mathrm{T}}F\right]\right)
\end{aligned}
\tag{5.9}
$$

The proof is given in the supplementary material.

## 5.8 Supplementary material for Chapter 5

### Variances and covariances

First, we derive the covariance matrix of $(x, y)$. The PO2PLS model for $x$ and $y$ is

$$\begin{aligned} x =& tW^{\mathrm{T}} + t_{\perp}W_{\perp}^{\mathrm{T}} + e \\ y =& uC^{\mathrm{T}} + u_{\perp}C_{\perp}^{\mathrm{T}} + f \\ u =& tB + h \end{aligned} \tag{5.10}$$

The covariance matrices of $x$ and $y$ are given by

$$\begin{aligned} \mathrm{Var}(x) &= \mathrm{Var}\left(tW^{\mathrm{T}} + t_{\perp}W_{\perp}^{\mathrm{T}} + e\right) = W\mathrm{Var}(t)W^{\mathrm{T}} + Wo\mathrm{Var}(t_{\perp})Wo^{\mathrm{T}} + \mathrm{Var}(e) \\ &= W\Sigma_t W^{\mathrm{T}} + W_{\perp}\Sigma_{t_{\perp}}W_{\perp}^{\mathrm{T}} + \sigma_e^2 I_p \\ \mathrm{Var}(y) &= \mathrm{Var}\left(uC^{\mathrm{T}} + u_{\perp}C_{\perp}^{\mathrm{T}} + f\right) = C\mathrm{Var}(u)C^{\mathrm{T}} + Co\mathrm{Var}(u_{\perp})Co^{\mathrm{T}} + \mathrm{Var}(f) \\ &= C(B^2\Sigma_t + \Sigma_h)C^{\mathrm{T}} + C_{\perp}\Sigma_{u_{\perp}}C_{\perp}^{\mathrm{T}} + \sigma_f^2 I_q \\ \mathrm{Cov}(x, y) &= \mathrm{Cov}\left(tW^{\mathrm{T}} + t_{\perp}W_{\perp}^{\mathrm{T}} + e, uC^{\mathrm{T}} + u_{\perp}C_{\perp}^{\mathrm{T}} + f\right) = W\mathrm{Cov}(t, u)C^{\mathrm{T}} \\ &= W\mathrm{Cov}(t, tB)C^{\mathrm{T}} = WB\Sigma_t C^{\mathrm{T}} \end{aligned} \tag{5.11}$$

The covariances between the observed and latent variables are given by

$$\begin{aligned} \mathrm{Cov}(x, t) &= \mathrm{Cov}\left(tW^{\mathrm{T}} + t_{\perp}W_{\perp}^{\mathrm{T}} + e, t\right) = W\mathrm{Var}(t) = W\Sigma_t \\ \mathrm{Cov}(x, t_{\perp}) &= \mathrm{Cov}\left(tW^{\mathrm{T}} + t_{\perp}W_{\perp}^{\mathrm{T}} + e, t_{\perp}\right) = W_{\perp}\mathrm{Var}(t_{\perp}) = W_{\perp}\Sigma_{t_{\perp}} \\ \mathrm{Cov}(x, u) &= \mathrm{Cov}\left(tW^{\mathrm{T}} + t_{\perp}W_{\perp}^{\mathrm{T}} + e, tB + h\right) = W\mathrm{Var}(t)B = W\Sigma_t B \\ \mathrm{Cov}(y, t) &= \mathrm{Cov}\left(uC^{\mathrm{T}} + u_{\perp}C_{\perp}^{\mathrm{T}} + f, t\right) = C\mathrm{Cov}(tB + h, t) = C\Sigma_t B \\ \mathrm{Cov}(y, u) &= \mathrm{Cov}\left(uC^{\mathrm{T}} + u_{\perp}C_{\perp}^{\mathrm{T}} + f, u\right) = C\mathrm{Cov}(tB + h, tB + h) = C\left(\Sigma_t B^2 + \Sigma_h\right) \\ \mathrm{Cov}(y, u_{\perp}) &= \mathrm{Cov}\left(uC^{\mathrm{T}} + u_{\perp}C_{\perp}^{\mathrm{T}} + f, u_{\perp}\right) = C_{\perp}\mathrm{Var}(u_{\perp}) = C_{\perp}\Sigma_{u_{\perp}} \end{aligned} \tag{5.12}$$

See e.g. [34] for more details.

Since $(x, y)$ is a linear transformation of $(t, u, t_{\perp}, u_{\perp}, e, f, h)$, its joint distribution is multivariate zero mean normal, and is parametrized by the covariance matrix:

$$\Sigma := \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} := \begin{bmatrix} W\Sigma_t W^{\mathrm{T}} + W_{\perp}\Sigma_{t_{\perp}}W_{\perp}^{\mathrm{T}} + \sigma_e^2 I_p & W\Sigma_t BC^{\mathrm{T}} \\ CB\Sigma_t W^{\mathrm{T}} & C\Sigma_u C^{\mathrm{T}} + C_{\perp}\Sigma_{u_{\perp}}C_{\perp}^{\mathrm{T}} + \sigma_f^2 I_q \end{bmatrix} \tag{5.13}$$

### Identifiability of PO2PLS

Identifiability entails that if the distribution of $(x, y)$ is given, there is only one corresponding set of parameters yielding this distribution. Since $(x, y)$ follows a zero mean normal distribution, identifiability can be stated as

$$\Sigma = \tilde{\Sigma} \quad \Longleftrightarrow \quad \theta = \tilde{\theta}, \tag{5.14}$$

where $\Sigma, \tilde{\Sigma}$ are two covariance matrices following the PO2PLS decomposition.

In the proof, we assume that $B$, $\Sigma_t$, $\Sigma_h$, $\Sigma_{t_\perp}$ and $\Sigma_{u_\perp}$ are diagonal, where each element is non-negative. Moreover, we assume that the diagonal elements of $\Sigma_t B$ are positive and in strictly decreasing order. Furthermore, we constrain $W$, $W_\perp$, $C$ and $C_\perp$ to have orthonormal columns. Additionally, the matrices $[W, W_\perp]$ and $[C, C_\perp]$ have full rank. Finally, we assume $0 < r + r_x < p$ and $0 < r + r_y < q$, where $p$ and $q$ are the dimensions of $x$ and $y$, respectively.

Following along the lines of proof in [8], we show that the off diagonal block matrix is identified.

**Theorem 5.8.1.** *Suppose* $W\Sigma_t BC^{\mathrm{T}} = \tilde{W}\tilde{\Sigma}_t\tilde{B}\tilde{C}^{\mathrm{T}}$. *Then* $W = \tilde{W}$, $C = \tilde{C}$ *and* $\Sigma_t B = \tilde{\Sigma}_t\tilde{B}$, *up to sign.*

*Proof.* Since $\Sigma_t B$ and $\tilde{\Sigma}_t\tilde{B}$ are diagonal with positive decreasing elements, both sides of the equation represent a singular value decomposition. This decomposition is unique up to sign [8]. Therefore, $W$, $C$ and $\Sigma_t B$ are identified up to sign. □

Identifiability of the specific parts is shown by projecting the covariance matrix onto a subspace orthogonal to the specific loadings. The following Theorem makes this precise.

**Theorem 5.8.2.** *Suppose*

$$W\Sigma_t W^{\mathrm{T}} + W_\perp\Sigma_{t_\perp}W_\perp^{\mathrm{T}} + \sigma_e^2 I_p = W\tilde{\Sigma}_t W^{\mathrm{T}} + \tilde{W_\perp}\tilde{\Sigma}_{t_\perp}\tilde{W_\perp}^{\mathrm{T}} + \tilde{\sigma}_e^2 I_p \tag{5.15}$$

*Then* $W_\perp = \tilde{W_\perp}$, $\Sigma_t = \tilde{\Sigma}_t$, $\Sigma_{t_\perp} = \tilde{\Sigma}_{t_\perp}$ *and* $\sigma_e^2 = \tilde{\sigma}_e^2$.

*Proof.* Let $VDV^{\mathrm{T}}$ and $\tilde{V}\tilde{D}\tilde{V}^{\mathrm{T}}$ be the eigenvalue decompositions of the left and right hand side of (5.15), respectively. Since $p > r + r_x$, the latter $p - r - r_x$ eigenvalues in $D$ and $\tilde{D}$ equal $\sigma_e^2$ and $\tilde{\sigma}_e^2$, respectively. Therefore, $\sigma_e^2 = \tilde{\sigma}_e^2$. The remaining part of Equation (5.15) is given by

$$W\Sigma_t W^{\mathrm{T}} + W_\perp\Sigma_{t_\perp}W_\perp^{\mathrm{T}} = W\tilde{\Sigma}_t W^{\mathrm{T}} + \tilde{W_\perp}\tilde{\Sigma}_{t_\perp}\tilde{W_\perp}^{\mathrm{T}}$$

Since the columns of both $W\Sigma_t W^{\mathrm{T}}$ and $W\tilde{\Sigma}_t W^{\mathrm{T}}$ span the same space, and the columns of $W$ are not linearly dependent of $W_\perp$ and $\tilde{W_\perp}$, we have that $\Sigma_t = \tilde{\Sigma}_t$. Both sides of the remainder of Equation (5.15) can be recognized as a spectral decomposition; such decompositions are unique up to sign, so we have that $W_\perp = \tilde{W_\perp}$ up to sign and $\Sigma_{t_\perp} = \tilde{\Sigma}_{t_\perp}$. □

Following the same reasoning as the proof above, we have identifiability of $C_\perp$ up to sign, $\Sigma_{u_\perp}$, $\sigma_f^2$ and $\Sigma_t B^2 + \Sigma_h$. As $\Sigma_t$ and $\Sigma_t B$ are identifiable, also $B$ and $\Sigma_h$ are identifiable.

## An EM algorithm for PO2PLS

Let $X$ and $Y$ be data matrices consisting of $N$ i.i.d. draws from $(x, y)$ along the rows. For empirical identifiability of the components, we assume $\max(r + r_x, r + r_y) < N$.

Let the complete data vector be $(x, y, t, u, t_\perp, u_\perp)$. For each current estimate $\theta'$, the EM algorithm considers the objective function

$$Q(\theta|x, y, \theta') := \mathbb{E}_{\theta'} \left[\log f(x, y, t, u, t_\perp, u_\perp|\theta)|x, y\right]. \tag{5.16}$$

Here, the complete data likelihood can be written (with abuse of notation) as

$$f(x, y, t, u, t_\perp, u_\perp|\theta) = \underbrace{f(x|t, t_\perp)}_{W,W_\perp,\sigma_e^2} \underbrace{f(y|u, u_\perp)}_{C,C_\perp,\sigma_f^2} \underbrace{f(u|t)}_{B,\Sigma_h} \underbrace{f(t)}_{\Sigma_t} \underbrace{f(t_\perp)}_{\Sigma_{t_\perp}} \underbrace{f(u_\perp)}_{\Sigma_{u_\perp}}. \tag{5.17}$$

These factors depend on distinct sets of parameters, yielding optimization problems over separate set of parameters. As $f$ in (5.16) is a multivariate normal density, $Q$ involves conditional expectations of the first and second moments of the latent variables $t$, $u$, $t_\perp$ and $u_\perp$ given $x$ and $y$. Focussing on the first factor in (5.17), the conditional expectation of $f(x|t, t_\perp)$ is given by

$$-\frac{1}{2}\left\{Np\log(2\pi) + Np\log\sigma_e^2 + \sigma_e^{-2}\mathrm{tr}\,\mathbb{E}\left[||x - tW^{\mathrm{T}} - t_\perp W_\perp^{\mathrm{T}}||_F^2\,|x, y\right]\right\}. \tag{5.18}$$

The EM algorithm first calculates the conditional expectation of the complete data log-likelihood (5.16), given the observed data $(x, y)$ . Note that, as $f$ in (5.16) is a multivariate normal density, this objective function involves conditional expectations of the first and second moments of the latent variables $t$, $u$, $t_\perp$ and $u_\perp$ given $x$ and $y$. For example, the $f(x|t, t_\perp)$ part in (5.17), given by (5.18) involves $\mathbb{E}\left[(TT_\perp)|X, Y\right]$ and $\mathbb{E}\left[(TT_\perp)^{\mathrm{T}}(TT_\perp)|X, Y\right]$. The following Lemma is used for calculating such conditional expectations.

**Lemma 5.8.1.** *Let $z \sim \mathcal{N}(0, \Sigma_z)$ be multivariate normal. If $(x|z) \sim \mathcal{N}\left(z\Gamma^{\mathrm{T}}, \Sigma_\epsilon\right)$ then*

$$x \sim \mathcal{N}\left(0, \Gamma\Sigma_z\Gamma^{\mathrm{T}} + \Sigma_\epsilon\right) \tag{5.19}$$

*and*

$$(z|x) \sim \mathcal{N}\left(x\Sigma_\epsilon^{-1}\Gamma\tilde{\Sigma}_z, \tilde{\Sigma}_z\right), \tag{5.20}$$

*with $\tilde{\Sigma}_z = \left\{\Sigma_z^{-1} + \Gamma^{\mathrm{T}}\Sigma_\epsilon^{-1}\Gamma\right\}^{-1}$.*

This Lemma can be applied by noting that Equation (5.10) can be rewritten as

$$(x, y) = (t, u, t_\perp, u_\perp)\begin{bmatrix} W & 0 & W_\perp & 0 \\ 0 & C & 0 & C_\perp \end{bmatrix}^{\mathrm{T}} + (e, f), \tag{5.21}$$

with

$$\mathrm{Var}((t, u, t_\perp, u_\perp)) = \begin{bmatrix} \Sigma_t & \Sigma_t B & 0 & 0 \\ \Sigma_t B & \Sigma_u & 0 & 0 \\ 0 & 0 & \Sigma_{t_\perp} & 0 \\ 0 & 0 & 0 & \Sigma_{u_\perp} \end{bmatrix}. \tag{5.22}$$

We take $z = (t, u, t_\perp, u_\perp)$ and $\Gamma$ the loading matrix as in (5.21). Then, Lemma 5.8.1 yields conditional expectations and variances of $z$ given $x$ and $y$.

Taking the derivative with respect to $W$, while fixing $W_\perp$, and setting it to zero yields

$$\begin{aligned}\hat{W} &= \left\{\left(X - t_\perp W_\perp^{\mathrm{T}}\right)^{\mathrm{T}} T\right\}\left\{T^{\mathrm{T}}T + \Lambda_W\right\}^{-1} \\ &= \operatorname{orth}\left\{\left(X - t_\perp W_\perp^{\mathrm{T}}\right)^{\mathrm{T}} T\right\}.\end{aligned} \tag{5.23}$$

Here, $\operatorname{orth}(A) = UV^{\mathrm{T}}$ with $U$ and $V$ the singular vectors of $A$. The last step is proven in [8]. The same argument can be applied to $W_\perp$, holding $W$ fixed at $\hat{W}$:

$$\begin{aligned}\hat{W}_\perp &= \left\{\left(X - T\hat{W}^{\mathrm{T}}\right)^{\mathrm{T}} t_\perp\right\}\left\{t_\perp^{\mathrm{T}} t_\perp + \Lambda_{W_\perp}\right\}^{-1} \\ &= \operatorname{orth}\left\{\left(X - T\hat{W}^{\mathrm{T}}\right)^{\mathrm{T}} t_\perp\right\}.\end{aligned} \tag{5.24}$$

In the same way, maximizers in the M-step are obtained for $W_\perp$, $C$ and $C_\perp$.

Regarding the variance parameters, consider the part of the log-likelihood given in (5.18). Taking the derivative with respect to $\sigma_e^2$ yields the well-known maximum likelihood estimator for the residual variance in a linear model:

$$\sigma_e^{2,next} = (Np)^{-1} E^{\mathrm{T}} E. \tag{5.25}$$

Similarly, the other variance parameter updates are calculated. The update for the inner regression matrix $B$ is given by the usual maximum likelihood estimator for the regression coefficient,

$$B^{next} = U^{\mathrm{T}} T (T^{\mathrm{T}} T)^{-1} \circ I_r. \tag{5.26}$$

Here, the off-diagonals are set to zero, since $B$ must be a diagonal matrix.

Taking into account that the latent variables are unobserved, we take the expected value of the respective log-likelihoods. Now the EM updates at step $k$ can be written as follows, starting with an initial guess at $k = 0$.

$$\begin{aligned}W^{k+1} &= \operatorname{orth}\left(X^{\mathrm{T}}\,\mathbb{E}_k\left[T\right] - W_\perp^k \mathbb{E}_k\left[T_\perp^{\mathrm{T}} T\right]\right) \\ W_\perp^{k+1} &= \operatorname{orth}\left(X^{\mathrm{T}}\,\mathbb{E}_k\left[T_\perp\right] - W^{k+1} \mathbb{E}_k\left[T^{\mathrm{T}} T_\perp\right]\right) \\ C^{k+1} &= \operatorname{orth}\left(Y^{\mathrm{T}}\,\mathbb{E}_k\left[U\right] - C_\perp^k \mathbb{E}_k\left[U_\perp^{\mathrm{T}} U\right]\right) \\ C_\perp^{k+1} &= \operatorname{orth}\left(Y^{\mathrm{T}}\,\mathbb{E}_k\left[U_\perp\right] - C^{k+1} \mathbb{E}_k\left[U^{\mathrm{T}} U_\perp\right]\right) \\ B^{k+1} &= \mathbb{E}\left[U^{\mathrm{T}} T\right]\left(\mathbb{E}\left[T^{\mathrm{T}} T\right]\right)^{-1} \circ I_r \\ \Sigma_t^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[T^{\mathrm{T}} T\right] \circ I_r \\ \Sigma_{t_\perp}^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[T_\perp^{\mathrm{T}} T_\perp\right] \circ I_{r_x} \\ \Sigma_{u_\perp}^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[U_\perp^{\mathrm{T}} U_\perp\right] \circ I_{r_y} \\ \Sigma_h^{k+1} &= \frac{1}{N}\mathbb{E}_k\left[H^{\mathrm{T}} H\right] \circ I_r \\ (\sigma_e^2)^{k+1} &= \frac{1}{Np}\operatorname{tr}\left(\mathbb{E}_k\left[E^{\mathrm{T}} E\right]\right) \\ (\sigma_f^2)^{k+1} &= \frac{1}{Nq}\operatorname{tr}\left(\mathbb{E}_k\left[F^{\mathrm{T}} F\right]\right)\end{aligned} \tag{5.27}$$

Reasonable initial guesses can be obtained by fitting an O2PLS model and extracting the estimates.

## Standard errors for PO2PLS

Standard errors for parameter values are obtained from the Fisher information matrix. This matrix can be obtained by calculating the conditional expectations of the first and second derivative of the complete data likelihood $L_{comp}$ [26]:

$$I_{obs} = \mathbb{E}\left[\Delta L_{comp}\right] - \mathbb{E}\left[\left(\nabla L_{comp}\right)\left(\nabla L_{comp}\right)^{\mathrm{T}}\right]. \tag{5.28}$$

First, we calculate these derivatives. Then, we calculate the conditional expectations of these expressions. Finally, we obtain the Fisher information matrix.

Consider the complete data likelihood (5.17). Define $\Gamma$ and $z$ as in (5.21) and Lemma 5.8.1. Furthermore, define the concatenated data matrix $D = (X, Y)$. The complete data likelihood with respect to $\Gamma$ is, up to a constant, proportional to

$$\lambda(\Gamma) := -\frac{1}{2}\sum_{i=1}^{N}\left(d_i - z_i\Gamma^{\mathrm{T}}\right)\Sigma_{(e,f)}^{-1}\left(d_i - z_i\Gamma^{\mathrm{T}}\right)^{\mathrm{T}}. \tag{5.29}$$

Here, $\Sigma_{(e,f)} := \mathrm{Var}((e,f))$ and is a diagonal matrix. We can write $\lambda$ as

$$\begin{aligned} -\frac{1}{2}\lambda(\mathrm{vec}(\Gamma)) = &\sum_{i=1}^{N} d_i\Sigma_{(e,f)}^{-1}d_i^{\mathrm{T}} \\ &- 2\sum_{i=1}^{N}\left(z_i \otimes d_i\Sigma_{(e,f)}^{-1}\right)\mathrm{vec}(\Gamma) \\ &+ \sum_{i=1}^{N}\mathrm{vec}(\Gamma)^{\mathrm{T}}\left(z_i \otimes \Sigma_{(e,f)}^{-\frac{1}{2}}\right)^{\mathrm{T}}\left(z_i \otimes \Sigma_{(e,f)}^{-\frac{1}{2}}\right)\mathrm{vec}(\Gamma) \end{aligned} \tag{5.30}$$

In this calculation, the identity $\mathrm{vec}(ABC) = (C^{\mathrm{T}} \otimes A)\mathrm{vec}(B)$ [27] was used, where $\otimes$ is the Kronecker product and vec is the vectorization operator.

Differentiating $\lambda$ with respect to $\mathrm{vec}(\Gamma)$ yields

$$\begin{aligned} S(\mathrm{vec}(\Gamma)) &:= \sum_{i=1}^{N}\left(z_i^{\mathrm{T}} \otimes \Sigma_{(e,f)}^{-1}d_i^{\mathrm{T}}\right) - \sum_{i=1}^{N}\left(z_i \otimes \Sigma_{(e,f)}^{-\frac{1}{2}}\right)^{\mathrm{T}}\left(z_i \otimes \Sigma_{(e,f)}^{-\frac{1}{2}}\right)\mathrm{vec}(\Gamma) \\ &= \sum_{i=1}^{N}\left(z_i^{\mathrm{T}} \otimes \Sigma_{(e,f)}^{-1}d_i^{\mathrm{T}}\right) - \sum_{i=1}^{N}\left(z_i^{\mathrm{T}}z_i \otimes \Sigma_{(e,f)}^{-1}\right)\mathrm{vec}(\Gamma) \end{aligned} \tag{5.31}$$

The negative of the second derivative of $\lambda$ with respect to $\mathrm{vec}(\Gamma)$ is

$$B(\mathrm{vec}(\Gamma)) := \sum_{i=1}^{N}\left(z_i^{\mathrm{T}}z_i \otimes \Sigma_{(e,f)}^{-1}\right) \tag{5.32}$$

Using these two expressions, we can calculate the Fisher information matrix for the parameter vector vec(Γ):

$$\begin{aligned}
I_{obs} &:= \mathbb{E}\left[B(\mathrm{vec}(\Gamma))|X,Y\right] - \mathbb{E}\left[\left(S(\mathrm{vec}(\Gamma))S(\mathrm{vec}(\Gamma))^{\mathrm{T}}\right)|X,Y\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\left(z_i^{\mathrm{T}} z_i \otimes \Sigma_{(e,f)}^{-1}\right)|d_i\right] \\
&- \mathbb{E}\left[\sum_{i=1}^{N}\left(z_i^{\mathrm{T}} z_i \otimes d_i \Sigma_{(e,f)}^{-2} d_i^{\mathrm{T}}\right)|d_i\right] \\
&+ \mathbb{E}\left[\sum_{i=1}^{N}\left(z_i^{\mathrm{T}} z_i \otimes \Sigma_{(e,f)}^{-1}\right)\mathrm{vec}(\Gamma)\left(z_i \otimes d_i \Sigma_{(e,f)}^{-1}\right)|d_i\right] \\
&+ \mathbb{E}\left[\sum_{i=1}^{N}\left(z_i^{\mathrm{T}} \otimes \Sigma_{(e,f)}^{-1} d_i^{\mathrm{T}}\right)\mathrm{vec}(\Gamma)^{\mathrm{T}}\left(z_i^{\mathrm{T}} z_i \otimes \Sigma_{(e,f)}^{-1}\right)|d_i\right] \\
&- \mathbb{E}\left[\sum_{i=1}^{N}\left(z_i^{\mathrm{T}} z_i \otimes \Sigma_{(e,f)}^{-1}\right)\mathrm{vec}(\Gamma)\mathrm{vec}(\Gamma)^{\mathrm{T}}\left(z_i^{\mathrm{T}} z_i \otimes \Sigma_{(e,f)}^{-1}\right)|d_i\right]
\end{aligned} \tag{5.33}$$

The standard errors of the loading elements are given by the diagonal elements of $-I_{obs}^{-1}$.

The information matrix for each column $k$ in Γ is then given by:

$$\begin{aligned}
I_{obs} &= \sum_{i=1}^{N}\mathbb{E}\left[z_{ik}^2|X,Y\right]\Sigma_{(e,f)}^{-1} \\
&- \sum_{i=1}^{N}\Sigma_{(e,f)}^{-1} d_i^{\mathrm{T}}\mathbb{E}\left[z_{ik}^2|X,Y\right] d_i \Sigma_{(e,f)}^{-1} \\
&+ 2\sum_{i=1}^{N}\Sigma_{(e,f)}^{-1} d_i^{\mathrm{T}}\mathbb{E}\left[z_{ik}^3|X,Y\right]\Gamma_k^{\mathrm{T}}\Sigma_{(e,f)}^{-1} \\
&- \sum_{i=1}^{N}\Sigma_{(e,f)}^{-1}\Gamma_k\mathbb{E}\left[(z_{ik}^4|X,Y\right]\Gamma_k^{\mathrm{T}}\Sigma_{(e,f)}^{-1}
\end{aligned} \tag{5.34}$$

The first and second conditional moments of $z_{ik}$ are given above. From these moments, the third and fourth moments can be obtained:

$$\begin{aligned}
\mathbb{E}\left[z_{ik}^3|X,Y\right] &= \mathbb{E}\left[z_{ik}|X,Y\right]^3 + 3\mathbb{E}\left[z_{ik}|X,Y\right]\mathrm{Var}\left(z_{ik}|X,Y\right) \\
\mathbb{E}\left[z_{ik}^4|X,Y\right] &= \mathbb{E}\left[z_{ik}|X,Y\right]^4 + 6\mathbb{E}\left[z_{ik}|X,Y\right]^2\mathrm{Var}\left(z_{ik}|X,Y\right) + 3\mathrm{Var}\left(z_{ik}|X,Y\right)^2.
\end{aligned} \tag{5.35}$$

## Simulation study

The results from the high dimensional scenario (without the SIFA method) is shown in Figure 5.4. The layout is the same as in Figure 5.1 in the main text, but excluding the column with the SIFA results. The conclusions are similar to those in the low-dimensional scenario.

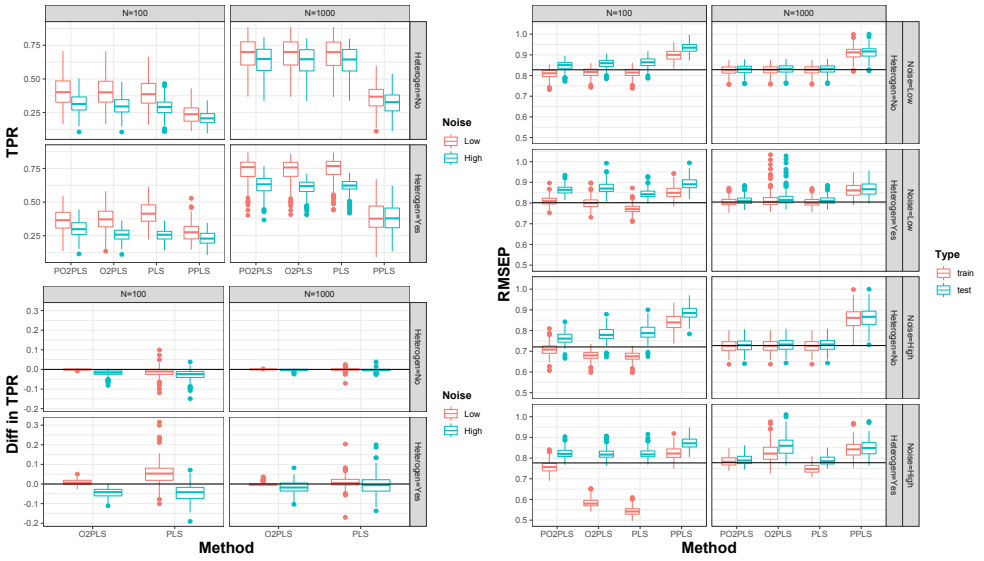


Figure 5.4: **True Positives Rate and Root MSE of Prediction (high dimensionality).** *Upper left figure*: proportion of true top 25% among estimated top 25% (TPR) for each method, stratified by simulation scenario. *Lower left figure*: Difference in TPR of several methods and PO2PLS. Lower values are in favor of PO2PLS. In both plots, the left red boxplots correspond with the low noise scenario, the right blue boxplots with high noise. *Right figure*: Root mean squared error of prediction stratified by method and scenario. The black line represent the test error when using the true parameter values. The left red boxplots correspond with the training error, the right blue boxplots with the test error.

Additional simulations are conducted to evaluate the performance of PO2PLS estimates in several scenarios. Robustness against violation of the normality assumption and rank misspecification is assessed. Finally, we investigate the impact of heterogeneity in the scale of the joint and specific parts on the PO2PLS estimates. The performance is compared to the performance of O2PLS and SIFA. Also, as often researchers are interested in the top features, we evaluate the accuracy of the top 25% estimated loading values.

In the first simulation scenario, combinations of large and small sample sizes ($N = 50, 500$), low and high dimensional $X$ ($p = 20, 200$) and small and large noise proportion ($10\%, 50\%$) were considered. Secondly, to evaluate robustness of PO2PLS, we considered a normal distribution, a $t$ distribution with two degrees of freedom, a Poisson distribution with rate one and a binomial distribution with two trials and success probability 0.25 for the latent variables. Note that these distributions reflect characteristics typically observed in omics data, such as skewed, heavy tailed and discrete variables. Thirdly, we investigated the impact of rank misspecification, by setting the number of components to be estimated to $(r, r_x, r_y) = (4, 3, 2)$. Finally, the impact of heterogeneity between joint and specific parts was evaluated by considering imbalanced joint and specific variance levels: $\text{tr}\Sigma_u \approx 10\text{tr}\Sigma_t$ and $\text{tr}\Sigma_{t_\perp} \approx 10\text{tr}\Sigma_t$, respectively.

The simulated data were generated from the PO2PLS model (5.10), with $r = 3$, $r_x = 2$ and $r_y = 1$. We took $B = I$ and $\Sigma_H = 0$ to comply with the SIFA assumptions. The other parameter values were drawn at random from a normal distribution for the loading values, and a uniform distribution on $[1, 3]$ for the variance parameters. The PO2PLS identifiability restrictions were applied to these parameters.

Estimation performance was measured by calculating the inner product of each estimated loading column with the corresponding true loading column. The accuracy of the top 25% loadings was measured by calculating the proportion of true top 25% among the estimated top. To avoid inflation of errors, we corrected the sign and order of the estimated components to match those of the true loading matrices.

The EM algorithm was considered converged when the log-likelihood increment was below $10^{-6}$ or when $10^4$ steps were taken. For each setting in the normal distribution scenario, 1000 replicates were generated. For the three non-normal distribution scenarios, we considered 200 replicates.

In the first scenario, the inner products of the joint and specific PO2PLS loadings were overall larger than 0.75, except for the specific loadings in the high noise and small sample size scenarios. An increase in performance was observed when the sample size was larger and when less noise was present. The performance was slightly better in the high dimensional scenario. Note that, in general, the first joint component was better estimated than the second and third. Furthermore, the $X$-specific components $W_\perp$ were better estimated than the $Y$-specific component $C_\perp$. Figures are shown in the supplement.

The proportion in the top 25% was higher in the scenarios with large sample size and low noise level. For the joint part, this proportion was also higher in the high dimensional scenarios, while the top 25% specific loadings were better recovered in the low dimensional scenarios.

Compared to O2PLS and SIFA, the PO2PLS estimates for the joint loadings performed similar. For the specific parts, the O2PLS performance tended to be lower.

Results for the scenarios with non-normally distributed variables were very similar (see supplementary material).

When ranks were chosen too high, the performance of the specific loadings was lower than when the rank was correctly specified, see Figure 5.7. For large sample size scenarios, SIFA did not perform well. Furthermore, for the high noise level and small sample size, the $Y$-specific loading estimates had larger error.

In Figure 5.9, results for the heterogeneity scenarios are shown. SIFA tended to underperform in scenarios where $p = 200$ and specific variation was much larger than joint variation. In presence of heterogeneity in joint parts, SIFA underperformed for large sample size, low noise level and high dimensions.

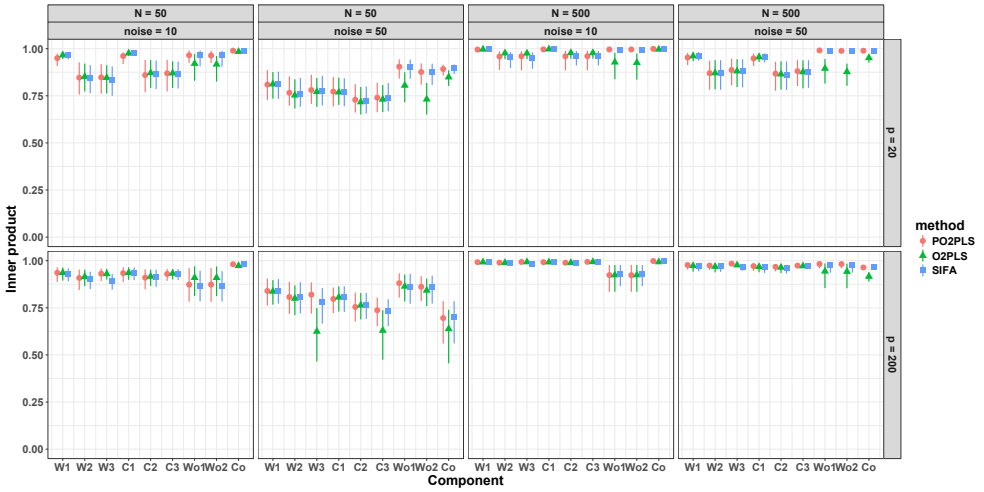


Figure 5.5: **Inner product of loadings; normally distributed variables.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

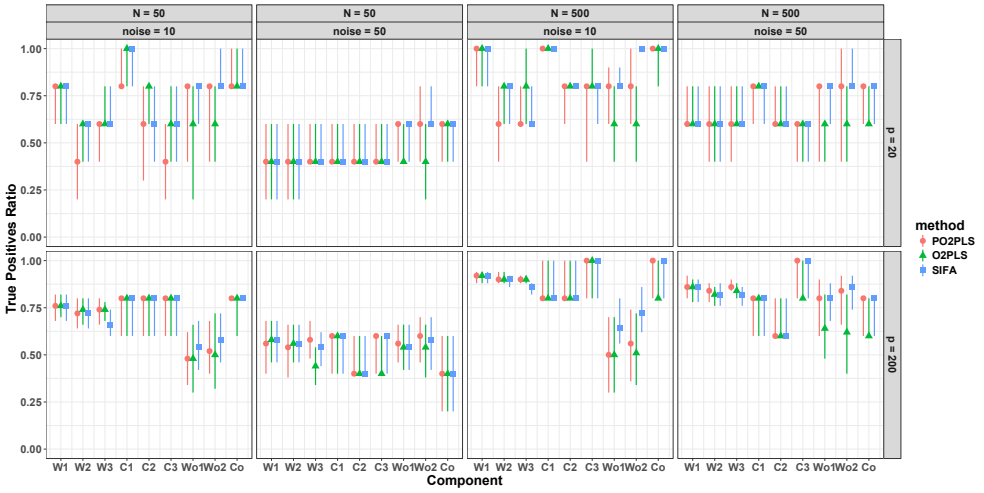


Figure 5.6: **Proportion of true positives in the top 25% of the estimated loadings; normally distributed variables.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

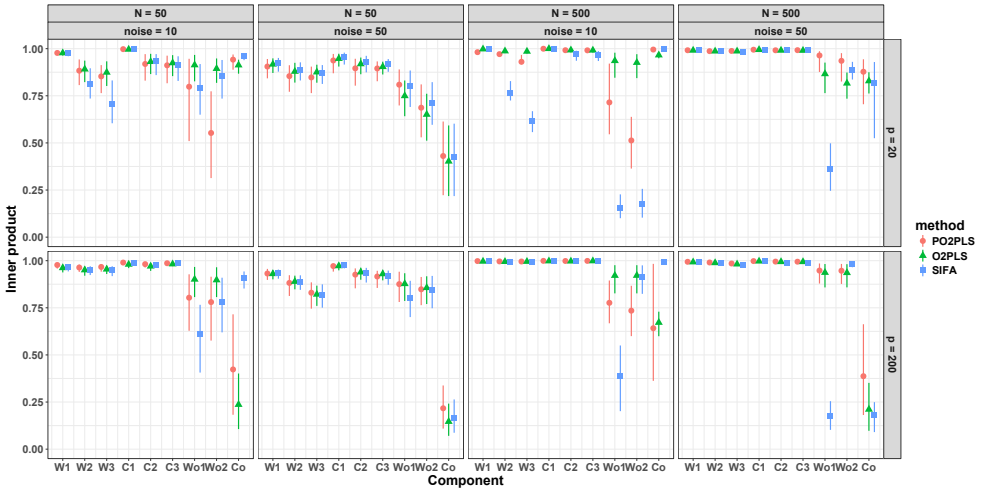


Figure 5.7: **Inner products of loadings; rank misspecification.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

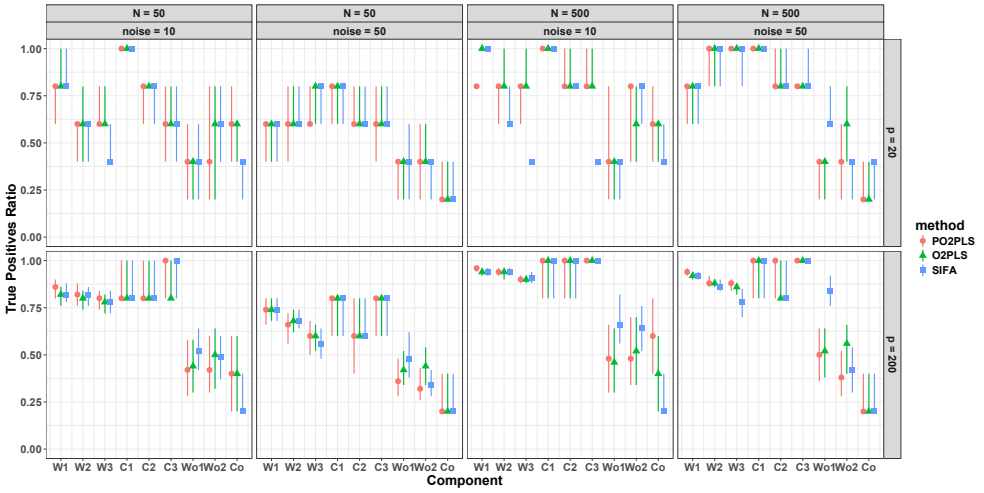


Figure 5.8: **Proportion of true positives in the top 25% of the estimated loadings; normally distributed variables.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

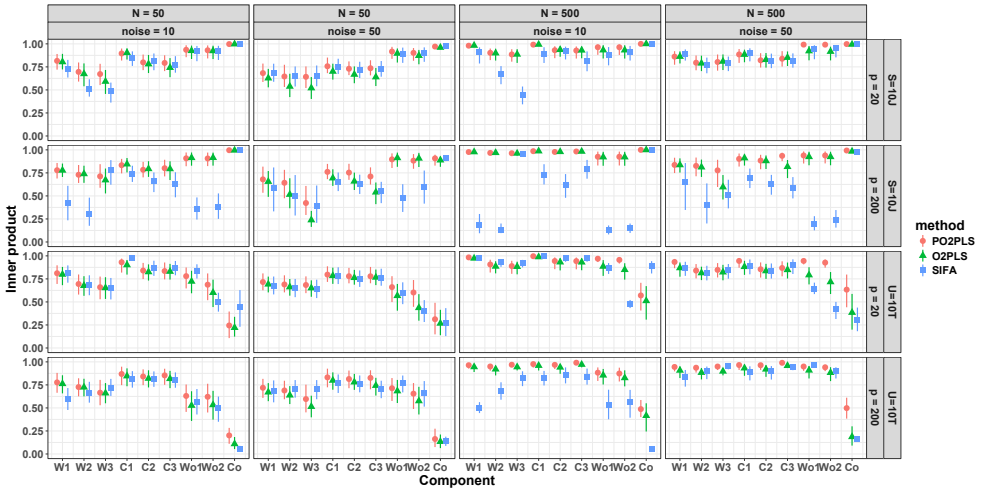


Figure 5.9: **Inner products of loadings; heterogeneity.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

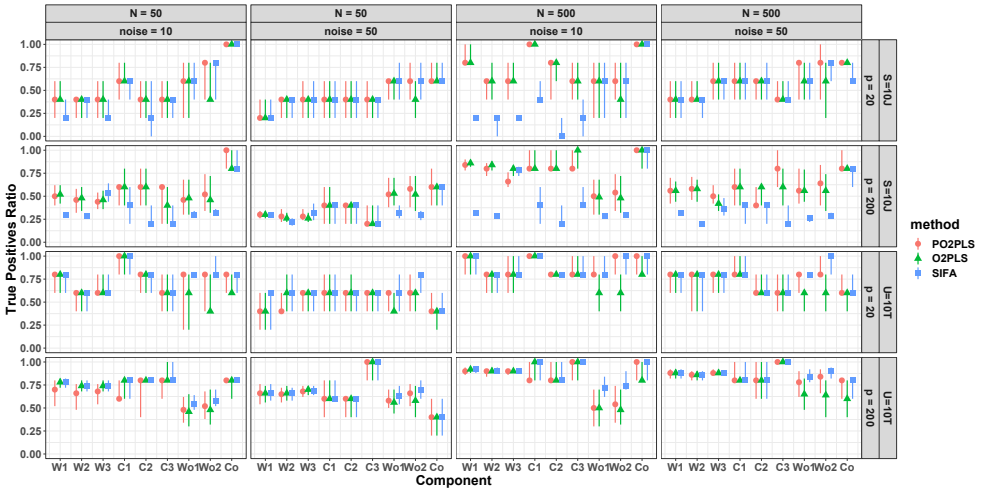


Figure 5.10: **Proportion of true positives in the top 25% of the estimated loadings; normally distributed variables.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

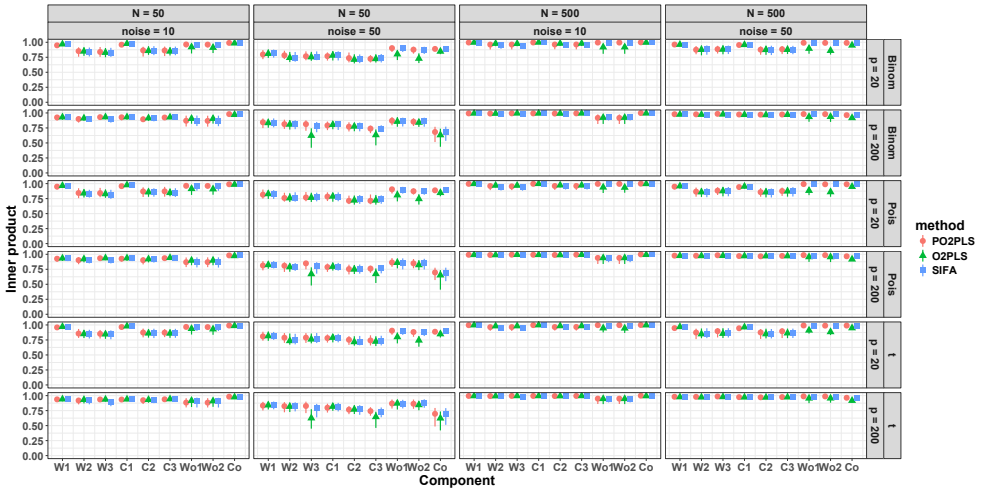


Figure 5.11: **Inner product of loadings; normally distributed variables.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

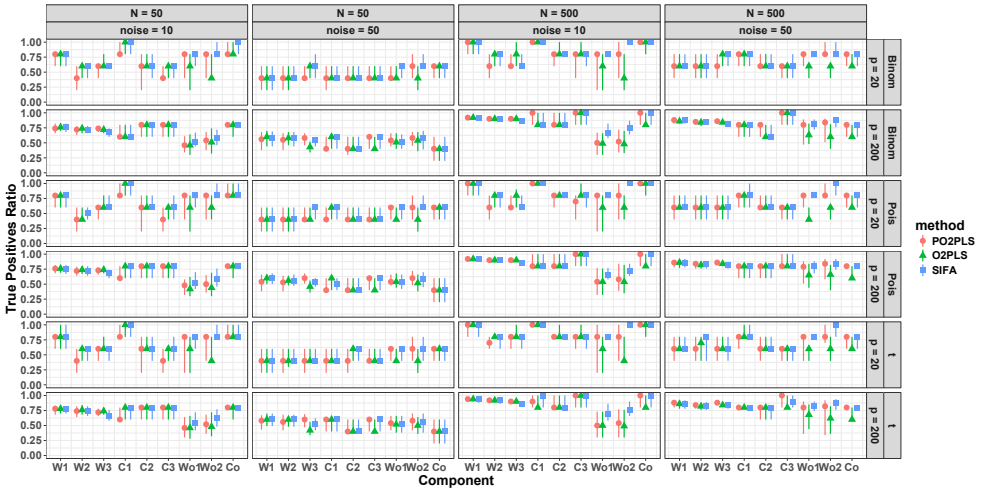


Figure 5.12: **Proportion of true positives in the top 25% of the estimated loadings; normally distributed variables.** Results for the low and high dimensional scenarios ($p = 20, 200$) are shown along the rows. Results for the low and high noise proportion scenarios ($10\%, 50\%$), nested within the small and large sample size scenarios ($N = 50, 500$), are shown along the columns.

# Bibliography

[1] 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.

[2] A. Chawla, Y. Barak, L. Nagy, D. Liao, P. Tontonoz, and R. M. Evans. PPAR-$\gamma$ dependent and independent effects on macrophage-gene expression in lipid metabolism and inflammation. *Nat. Med.*, 7(1):48–52, jan 2001.

[3] M. W.-L. Cheung. metaSEM: an R package for meta-analysis using structural equation modeling. *Front. Psychol.*, 5(January):1521, 2015.

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

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

[6] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Control. Clin. Trials*, 7(3):177–188, 1986.

[7] S. el Bouhaddani, J. Houwing-Duistermaat, P. Salo, M. Perola, G. Jongbloed, and H.-W. Uh. Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, 17(S2):S11, dec 2016.

[8] S. el Bouhaddani, H.-W. Uh, C. Hayward, G. Jongbloed, and J. Houwing-Duistermaat. Probabilistic partial least squares model: Identifiability, estimation and application. *J. Multivar. Anal.*, 167:331–346, sep 2018.

[9] S. el Bouhaddani, H. W. Uh, G. Jongbloed, C. Hayward, L. Klarić, S. M. Kiełbasa, and J. Houwing-Duistermaat. Integrating omics datasets with the OmicsPLS package. *BMC Bioinformatics*, 19(1), 2018.

[10] K. R. Feingold and C. Grunfeld. *Introduction to Lipids and Lipoproteins*. MDText.com, Inc., feb 2000.

[11] S. Geisser. Predictive Inference. *Philos. Sci.*, 24:180, 1993.

[12] S. M. Gross and R. Tibshirani. Collaborative regression. *Biostatistics*, 16(2):326–338, 2015.

[13] I. Gudelj and G. Lauc. Protein N-Glycosylation in Cardiovascular Diseases and Related Risk Factors, 2018.

[14] I. Gudelj, G. Lauc, and M. Pezer. Immunoglobulin G glycosylation in aging and diseases. *Cell. Immunol.*, 333(July):65–79, nov 2018.

[15] Q. He, H. H. Zhang, C. L. Avery, and D. Y. Lin. Sparse meta-analysis with high-dimensional data. *Biostatistics*, 17(2):205–220, 2016.

[16] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321, dec 1936.

[17] 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.

[18] 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.

[19] A. R. Joyce and B. Ø. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.*, 7(3):198–210, 2006.

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

[21] 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.

[22] G. Lauc, J. E. Huffman, M. Pučić, L. Zgaga, B. Adamczyk, A. Mužinić, M. Novokmet, O. Polašek, O. Gornik, J. Krištić, T. Keser, V. Vitart, B. Scheijen, H.-W. Uh, M. Molokhia, A. L. Patrick, P. McKeigue, I. Kolčić, I. K. Lukić, O. Swann, F. N. van Leeuwen, L. R. Ruhaak, J. J. Houwing-Duistermaat, P. E. Slagboom, M. Beekman, A. J. M. de Craen, A. M. Deelder, Q. Zeng, W. Wang, N. D. Hastie, U. Gyllensten, J. F. Wilson, M. Wuhrer, A. F. Wright, P. M. Rudd, C. Hayward, Y. Aulchenko, H. Campbell, and I. Rudan. Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. *PLoS Genet.*, 9(1):e1003225, jan 2013.

[23] G. Lauc, M. Pezer, I. Rudan, and H. Campbell. Mechanisms of disease: The human N-glycome. *Biochim. Biophys. Acta - Gen. Subj.*, 1860(8):1574–1582, aug 2016.

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

[25] 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.

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

[27] J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics - Third Edition.* Wiley series in probability and mathematical statistics., 2007.

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

[29] 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.

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

[31] C. Menni, T. Keser, M. Mangino, J. T. Bell, I. Erte, I. Akmačić, F. Vučković, M. P. Baković, O. Gornik, M. I. McCarthy, V. Zoldoš, T. D. Spector, G. Lauc, and A. M. Valdes. Glycosylation of immunoglobulin G: Role of genetic and epigenetic influences. *PLoS One*, 8(12):6–13, 2013.

[32] 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.

[33] 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.

[34] 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.

[35] A. Shapiro. Asymptotic theory of overparameterized structural models. *J. Am. Stat. Assoc.*, 81(393):142–149, 1986.

[36] S. A. Springer and P. Gagneux. Glycan evolution in response to collaboration, conflict, and constraint. *J. Biol. Chem.*, 288(10):6904–6911, 2013.

[37] E. Theodoratou, H. Campbell, N. T. Ventham, D. Kolarich, M. Pučić-Baković, V. Zoldoš, D. Fernandes, I. K. Pemberton, I. Rudan, N. A. Kennedy, M. Wuhrer, E. Nimmo, V. Annese, D. P. B. McGovern, J. Satsangi, and G. Lauc. The role of glycosylation in IBD. *Nat. Rev. Gastroenterol. Hepatol.*, 2014.

[38] 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.

[39] 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.

[40] A. Wahl, E. van den Akker, L. Klaric, J. Štambuk, E. Benedetti, R. Plomp, G. Razdorov, I. Trbojević-Akmačić, J. Deelen, D. van Heemst, P. Eline Slagboom, F. Vučković, H. Grallert, J. Krumsiek, K. Strauch, A. Peters, T. Meitinger,

C. Hayward, M. Wuhrer, M. Beekman, G. Lauc, and C. Gieger. Genome-wide association study on immunoglobulin G glycosylation patterns. *Front. Immunol.*, 9(FEB):1–14, 2018.

[41] H. Wold. Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivar. Anal. III (Proc. Third Internat. Symp. Wright State Univ., Dayton, Ohio, 1972)*, pages 383–407. Academic Press, New York, 1973.

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

[43] H. Zhu and L. Li. Biological pathway selection through nonlinear dimension reduction. *Biostatistics*, 12(3):429–444, 2011.