Article:
El Bouhaddani, S, Uh, H-W, Hayward, C et al. (2 more authors) (2018) Probabilistic partial least squares model: Identifiability, estimation and application. Journal of Multivariate Analysis, 167. pp. 331-346. ISSN 0047-259X

https://doi.org/10.1016/j.jmva.2018.05.009

Probabilistic partial least squares model: Identifiability, estimation and application

Said el Bouhaddani[a,*], Hae-Won Uh[b,a], Caroline Hayward[e], Geurt Jongbloed[c], Jeanine Houwing-Duistermaat[d,a]

[a] *Department of Medical statistics and bioinformatics, Leiden University Medical Center, The Netherlands*
[b] *Department of Biostatistics and Research Support UMC Utrecht, div. Julius Centrum, University Medical Center Utrecht, The Netherlands*
[c] *Department of Applied Mathematics, Delft University of Technology, The Netherlands*
[d] *Department of Statistics, University of Leeds, United Kingdom*
[e] *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Scotland*

Abstract

With a rapid increase in volume and complexity of data sets, there is a need for methods that can extract useful information, for example the relationship between two data sets measured for the same persons. The Partial Least Squares (PLS) method can be used for this dimension reduction task. Within life sciences, results across studies are compared and combined. Therefore, parameters need to be identifiable, which is not the case for PLS. In addition, PLS is an algorithm, while epidemiological study designs are often outcome-dependent and methods to analyze such data require a probabilistic formulation. Moreover, a probabilistic model provides a statistical framework for inference. To address these issues, we develop Probabilistic PLS (PPLS). We derive maximum likelihood estimators that satisfy the identifiability conditions by using an EM algorithm with a constrained optimization in the M step. We show that the PPLS parameters are identifiable up to sign. A simulation study is conducted to study the performance of PPLS compared to existing methods. The PPLS estimates performed well in various scenarios, even in high dimensions. Most notably, the estimates seem to be robust against departures from normality. To illustrate our method, we applied it to IgG glycan data from two cohorts. Our PPLS model provided insight as well as interpretable results across the two cohorts.

Keywords: Dimension reduction, EM algorithm, Identifiability, Inference, Probabilistic partial least squares

1. Introduction

With the exponentially growing volume of data sets, multivariate methods for reducing dimensionality are an important research area in statistics. For combining two data sets, Partial Least Squares (PLS) regression [28] is a popular dimension reduction method [1]. PLS decomposes variation in each data set in a joint part and a residual part. The joint part is a linear projection of one data set on the other that best explains the covariance between the two data sets. These projections are obtained by iterative algorithms, such as NIPALS [28]. Partial Least Squares is popular in chemometrics [3]. In this field, the focus is on development of algorithms with good prediction performance, while the underlying model is less important. For applications in life sciences, interpretation of parameter estimates is necessary to gain understanding of the underlying molecular mechanisms.

For interpretation, a model needs to be identifiable. A model is said to be unidentifiable if the model corresponds to more than one set of parameter values. For PLS, rotation of the parameters does not change the model [26]. Hence, PLS does not provide an identifiable model. By constraining the parameter space, identifiability can be obtained. This involves solving a challenging optimization problem, since PLS requires estimating a structured covariance matrix [19].

For many problems in life sciences the study design needs to be accounted for, and algorithmic approaches such as PLS cannot be applied. Hence, a probabilistic formulation is necessary. Since likelihood method provides asymptotic standard errors of parameter estimates, computer-intensive resampling procedures can be avoided.

*Corresponding author
Email addresses: S.el_Bouhaddani@lumc.nl (Said el Bouhaddani), H.W.Uh@umcutrecht.nl (Hae-Won Uh), Caroline.Hayward@igmm.ed.ac.uk (Caroline Hayward), G.Jongbloed@tudelft.nl (Geurt Jongbloed), J.Duistermaat@leeds.ac.uk (Jeanine Houwing-Duistermaat)

Also for other dimension reduction techniques, probabilistic methods have been developed. In 1999, Tipping and Bishop [23] developed the Probabilistic Principal Component Analysis (PPCA), in order to deal with missing data and dependent samples. In 2005, Bach and Jordan [2] developed Probabilistic Canonical correlation analysis (PCCA). However, for both PPCA and PCCA the model parameters are not identifiable, since rotation of the parameters does not change the model [2, 23]. In addition, in 2015, simultaneous envelopes models have been developed [4] for 'low-dimensional' settings. Further, Probabilistic PLS Regression and Probabilistic PLS have been proposed [14, 30]. For all these approaches, the model parameters are not identifiable.

In this paper we propose the Probabilistic Partial Least Squares (PPLS) model and show that the model parameters are identifiable up to a sign. We propose to maximize the PPLS likelihood with an EM algorithm that decouples the likelihood into several factors involving distinct sets of parameters. In the M step, a constrained optimization problem is solved by using a matrix of Lagrange multipliers.

The rest of the paper is organized as follows: In Section 2 we develop the PPLS model and establish identifiability of the model parameters. We develop an efficient algorithm for estimating the PPLS parameters. In Section 3 we study the performance of the PPLS estimators via simulations. In Section 4 we illustrate the PPLS model with two data matrices from two cohorts. We finish with a discussion.

2. Model and estimation

2.1. The PPLS model

Let x and y be two random row-vectors of dimension p and q, respectively. The Probabilistic Partial Least Squares (PPLS) model describes the two random vectors in terms of a joint part and a noise part. The joint part consists of correlated latent vectors, denoted by t and u, while the noise part consists of isotropic normal random vectors referred to as e, f and h. The dimension of t and u is denoted by r. The PPLS model describing the relationship between x, y and the joint and noise parts is

$$x = tW^\top + e, \quad y = uC^\top + f, \quad u = tB + h. \tag{1}$$

Specifically, $e = (e_1, \ldots, e_p)$, $f = (f_1, \ldots, f_q)$ and $h = (h_1, \ldots, h_r)$ are independent with zero mean and referred to as noise variables. The distributions of e, f and h are multivariate normal with positive definite covariance matrix proportional to the identity matrix,

$$e \sim \mathcal{N}(0, \sigma_e^2 I_p), \quad f \sim \mathcal{N}(0, \sigma_f^2 I_q), \quad h \sim \mathcal{N}(0, \sigma_h^2 I_r).$$

The latent vector $t = (t_1, \ldots, t_r)$ is an r-dimensional multivariate normal vector with with zero mean and diagonal positive definite covariance matrix $\Sigma_t = \mathrm{diag}(\sigma_{t_1}^2, \ldots, \sigma_{t_r}^2)$, so

$$t \sim \mathcal{N}(0, \Sigma_t). \tag{2}$$

The matrix $B = \mathrm{diag}(b_1, \ldots, b_r)$ is a diagonal matrix of size r, containing regression coefficients of u on t. Finally W $(p \times r)$ and C $(q \times r)$ are parameter matrices, referred to as loadings. The PPLS model for the random p-dimensional row-vector x and random q-dimensional row-vector y is given in Eq. (1). Let θ be the parameters of the PPLS model, i.e.,

$$\theta = (W, C, B, \Sigma_t, \sigma_e, \sigma_f, \sigma_h). \tag{3}$$

The PPLS model and its parameters are formulated conditional on the value of the dimension of the latent space r.

The PPLS model (1) assumes a multivariate normal distribution for the observable random vectors x and y. The covariance between x and y is modeled by the regression of the latent vector u on t. The distribution of (x, y) is $\mathcal{N}(0, \Sigma)$ with density given, for $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, by

$$f(x, y) = (2\pi)^{-(p+q)/2} |\Sigma|^{-\frac{1}{2}} \mathrm{e}^{(x,y)\Sigma^{-1}(x,y)^\top},$$

and covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{x,y} \\ \Sigma_{y,x} & \Sigma_y \end{pmatrix} = \begin{pmatrix} W\Sigma_t W^\top + \sigma_e^2 I_p & W\Sigma_t BC^\top \\ CB\Sigma_t W^\top & C(B^2\Sigma_t + \sigma_h^2 I_r)C^\top + \sigma_f^2 I_q \end{pmatrix}. \tag{4}$$

This follows from the normality property and from computing the variances and covariances of the random vectors; see Appendix A for the details.

2.2. Identifiability of probabilistic PLS

To establish identifiability of the PPLS model, some assumptions about its parameters have to be made. First, we assume that $0 < r < \min(p, q)$. Second, we assume that the diagonal elements of B are positive, $b_k > 0$ for $k \in \{1, \ldots, r\}$. This will not restrict the model, since $t_k b_k$ is equal to $-t_k b_k$ in distribution. To identify the order of the loading vectors, the elements of $(\sigma_{t_k}^2 b_k)_{k=1}^r$ are assumed to be strictly decreasing with k. Finally, we assume that the loading matrices W and C are orthogonal, i.e., $W^\top W = C^\top C = I_r$. Together with the diagonality of Σ_t in (2), it implies identifiability of all parameters up to sign. This is shown in the following theorem.

Theorem 1. *Let r be fixed such that $0 < r < \min(p, q)$. Let $(x, y)_1$ and $(x, y)_2$ be generated by the PPLS model* (1) *having covariance matrix Σ_1 and Σ_2 with underlying parameters θ_1 and θ_2 as defined in* (3)*, respectively. Then $\Sigma_1 = \Sigma_2$ implies that $W_1 = W_2\Delta$, $C_1 = C_2\Delta$ for some diagonal matrix Δ with on the diagonal elements $\delta_i \in \{-1, 1\}$, for $i \in \{1, \ldots, r\}$, and all other parameters in θ_1 and θ_2 are equal.*

The formal proof is given in Appendix B. Identifiability up to sign can be represented by a diagonal orthogonal matrix, namely a diagonal matrix with diagonal elements in $\{-1, 1\}$. For example, taking the model for x in (1), we may substitute W by WR_S and t by tR_S, where R_S is a diagonal orthogonal matrix, and get

$$x = tR_S R_S^\top W^\top + e = \sum_{j=1}^{r} t_j (R_S)_{jj}^2 w_j^\top + e.$$

Since $(R_S)_{jj}^2 = 1$ and the distribution of t_j and $-t_j$ is the same, the right-hand side reduces to the original model for x in (1). Note that the PPLS model is not invariant under general rotation matrices. Take a general rotation matrix R, then we still get

$$x = tW^\top + e = tRR^\top W^\top + e,$$

since $RR^\top = I_r$. Inspecting the covariance of TR we see that $\mathrm{cov}(TR) = R^\top \Sigma_t R$, which is not diagonal if R is not diagonal, and violates the PPLS model assumption on Σ_t in Eq. (2).

2.3. Estimating the parameters

Unlike the iterative PLS methods, we propose a simultaneous approach for estimating the parameters, while taking the constraints in the PPLS model into account. Given the number of PPLS components, r, the log likelihood of an independent and identically distributed (iid) sample $(X, Y) = \{(X_1, Y_1)^\top, \ldots, (X_N, Y_N)^\top\}^\top$ of size N from (x, y) is

$$L(\theta) = -\frac{N(p+q)}{2} - \frac{N}{2} \ln |\Sigma| - \frac{N}{2} \mathrm{tr}(S\Sigma^{-1}) \tag{5}$$

with $S = N^{-1} \sum_{i=1}^{N} (X_i, Y_i)^\top (X_i, Y_i)$ and Σ as in Eq. (4). To ensure empirical identifiability, we assume that $r < N$. Note that the data dimensionality p and q may be larger than N. For estimation of θ, maximum likelihood is used.

The log likelihood (5) depends in a non-linear way on the theoretical covariance matrix Σ, which contains the loadings and variances. Optimizing this function directly is a non-trivial task, especially in high dimensions (i.e. when p and q are large). However, the PPLS model allows for a more simple (but iterative) optimization approach. Indeed, the maximum likelihood estimates for θ are a least squares type solution if the latent variables t and u are observed, as the model for x and y in (1) involves known t and u. In contrast, knowing θ allows for reconstruction of t and u by computing their conditional means given x and y. Alternating these two scenarios is actually an Expectation-Maximization (EM) [5] algorithm, with observed data (x, y) and missing data (t, u).

The EM algorithm. The joint distribution of the complete data (x, y, t, u) can (with abuse of notation) be decomposed as

$$f(x, y, t, u) = f(x|t) f(y|u) f(u|t) f(t). \tag{6}$$

This follows from

$$f(x, y, t, u) = f(x, y|t, u) f(t, u) = f(x|t, u) f(y|t, u) f(t, u).$$

The second equation is implied by the fact that x and y are independent given t and u. The first two factors in the right-hand side can be rewritten as $f(x|t, u) = f(x|t)$ and $f(y|t, u) = f(y|u)$, since x and u are independent given t,

and y and t are independent given u. The last factor can be rewritten as $f(u|t)f(t)$, yielding Eq. (6). The logarithm of the first three factors in the product in (6) can be written as

$$\ln f(X|T) = -\frac{Np}{2\pi\sigma_e^2} - \frac{1}{2\sigma_e^2}\sum_{i=1}^{N}\|X_i - T_i W^\top\|^2,$$
$$\ln f(Y|U) = -\frac{Nq}{2\pi\sigma_f^2} - \frac{1}{2\sigma_f^2}\sum_{i=1}^{N}\|Y_i - U_i C^\top\|^2,$$
$$\ln f(U|T) = -\frac{Nr}{2\pi\sigma_h^2} - \frac{1}{2\sigma_h^2}\sum_{i=1}^{N}\|U_i - T_i B\|^2.$$

Denote by $L_{\text{Comp}} = \ln f(X, Y, T, U)$ the complete data log-likelihood, and define

$$Q(\theta) = \mathrm{E}\{L_{\text{Comp}}(\theta)|X, Y, \theta'\},$$

where the expectation is taken conditional on the observed X and Y, and θ' is a fixed current estimate of the parameters. By optimizing Q over all allowed θ, we get a non-negative increase in the *observed* log-likelihood L. Moreover, by iterating this process of taking the expectation and maximizing over θ, the estimates in general converge to a stationary point or, in particular, a (possibly local) maximum of L [5, 29]. The expectation step calculates the conditional expectation of the missing data given the observed data given by $Q(\theta)$, which may in general involve intractable integration. However, for the exponential family, in particular the multivariate normal family, the complete likelihood depends on the complete data only via the sufficient statistics (called $t(\mathbf{x})$ in [5]), which are given in terms of the first and second moments of the complete data for the multivariate normal distribution. Computing $Q(\theta)$ implies computing the expected first and second moment of the latent variables: $\mathrm{E}\,(T|X, Y, \theta)$, $\mathrm{E}\,(T^\top T|X, Y, \theta)$, $\mathrm{E}\,(U|X, Y, \theta)$, $\mathrm{E}\,(U^\top U|X, X, \theta)$ and $\mathrm{E}\,(U^\top T|X, Y, \theta)$; see Appendix C for details. Moreover, the decomposition in (6) allows for optimization of $\mathrm{E}\{\ln f(X|T)\}$, $\mathrm{E}\{\ln f(Y|U)\}$ and $\mathrm{E}\{\ln f(U|T)\}$ separately, while only considering parameters involved in each factor. Maximizing Q over θ yields parameter estimates for the next iteration in the EM algorithm. This leads us to the following theorem.

Theorem 2. *Let X and Y be an observed data sample of size N, generated according to the PPLS model* (1)*. Let r be fixed such that $0 < r < \min(N, p, q)$. The parameters in θ can be estimated with an EM algorithm, yielding the following iterative scheme in k with given starting values for $k = 0$:*

$$W^{k+1} = X^\top \mathrm{E}(T|X, Y, \theta^k)(L_W^\top)^{-1}; \quad C^{k+1} = Y^\top \mathrm{E}(U|X, Y, \theta^k)(L_C^\top)^{-1};$$
$$B^{k+1} = \mathrm{E}(U^\top T|X, Y, \theta^k)\{\mathrm{E}(T^\top T|X, Y, \theta^k)\}^{-1} \circ I_r; \; \Sigma_t^{k+1} = \frac{1}{N}\mathrm{E}(T^\top T|X, Y, \theta^k) \circ I_r; \; (\sigma_h^2)^{k+1} = \frac{1}{Nr}\mathrm{trE}(H^\top H|X, Y, \theta^k);$$
$$(\sigma_e^2)^{k+1} = \frac{1}{Np}\mathrm{trE}(E^\top E|X, Y, \theta^k); \quad (\sigma_f^2)^{k+1} = \frac{1}{Nq}\mathrm{trE}(F^\top F|X, Y, \theta^k);$$

where L_W and L_C are such that

$$L_W L_W^\top = \mathrm{E}(T^\top|X, Y, \theta^k)\, X\, X^\top\, \mathrm{E}(T|X, Y, \theta^k), \quad L_C L_C^\top = \mathrm{E}(U^\top|X, Y, \theta^k)\, Y\, Y^\top\, \mathrm{E}(U|X, Y, \theta^k).$$

The proof for Theorem 2 and the expressions for the conditional expectations are given in Appendix C. Note the dependency of W^{k+1} and C^{k+1} on the matrices L_W and L_C. These matrices ensure orthogonality of W^{k+1} and C^{k+1} in each iteration:

$$(W^{k+1})^\top W^{k+1} = L_W^{-1}\,\tilde{T}^\top\, X\, X^\top\, \tilde{T}(L_W^\top)^{-1} = L_W^{-1}\, L_W\, L_W^\top\, (L_W^\top)^{-1} = I_r,$$

where $\tilde{T} = \mathrm{E}(T|X, Y, \theta^k)$. The exact forms of L_W and L_C are not unique. Two choices are the eigenvectors of $\mathrm{E}(T^\top|X, Y, \theta^k)\, X\, X^\top\, \mathrm{E}(T|X, Y, \theta^k)$ and the lower triangular matrix of $X^\top\, \mathrm{E}(T|X, Y, \theta^k)$ in the Cholesky decomposition. Note that these two orthogonalization matrices are straightforward to calculate with standard linear algebra tools. Since the PPLS model is identifiable, all choices for L_W and L_C will lead to the same optimum as the iteration number k tends to infinity.

Standard errors for PPLS. Asymptotic standard errors for maximum likelihood estimators are found by inverting the observed Fisher information matrix. Following the reasoning of [16], the observed information may be given by

$$\mathrm{E}\{B(\hat{\theta})|X, Y\} - \mathrm{E}\{S(\hat{\theta})S(\hat{\theta})^\top|X, Y\}.$$

Here $S(\hat{\theta}) = \nabla\lambda(\hat{\theta})$ and $B(\theta) = -\nabla^2\lambda(\hat{\theta})$ are the gradient and negative of the second derivative of the log likelihood $\lambda(\theta)$ respectively evaluated in the MLE $\hat{\theta}$. The explicit form of the asymptotic covariance matrix of w_k is given in Appendix D. The square root of the diagonal elements are the asymptotic standard errors for the corresponding loading estimates.

Finding the number of components r. Available approaches to determine the number of PPLS components r are minimizing a cross-validated loss function [9], visually inspecting eigenvalues of a covariance matrix [17], and selecting the number of components needed to achieve a certain proportion of variance explained by the components. In this paper we apply the last approach.

The PLS and PPLS algorithms are available as R packages at `github.com/selbouhaddani` under repository OmicsPLS and PPLS, respectively.

3. Simulation study

To evaluate the performance of the PPLS estimates, a simulation study was conducted. The aim was (1) to investigate the performance of PPLS for various scenarios, (2) to evaluate robustness of the PPLS estimates against departures from the normality assumption, (3) to compare the performance of the loading estimates with other probabilistic approaches, and (4) to compare the asymptotic PPLS standard errors with the bootstrap standard errors.

The simulated data were generated according to the PPLS model (1). The number of components was chosen to be 3, both in the data generation and in the estimation. We considered combinations of small and large sample size ($N \in \{50, 500\}$), low and high dimensionality ($p \in \{20, 1000\}$), and small and large proportion of noise (denoted by $\alpha_n \in \{0.1, 0.5\}$). The robustness of PPLS was evaluated by considering four different continuous and discrete distributions for the latent variables t, u, e, f and h; we chose the normal distribution, the t distribution with two degrees of freedom, the Poisson distribution with rate 1, and the Binomial distribution with two trials and success probability 0.25. These distributions cover a wide range of characteristics typically observed in omics data: heavy tailed, skewed and/or discrete. The latent variables were scaled to have mean zero and variances as specified below. All scenarios are summarized in Table 1.

The true loading values per component were generated from the normal density function with parameters μ and σ, denoted by $N(x; \mu, \sigma^2)$, as follows

$$w_{j,k} = N\{j; (1/2 + 1/10j)p, 1/10p\}, \quad c_{j,k} = N\{j; (3/5 + 1/10j)q, 1/10q\}.$$

The second columns in W and C were orthonormalized with respect to the first columns, and the third columns were orthonormalized with respect to the first two columns; we used a Gram–Schmidt procedure for both operations. The elements of the diagonal matrix B were set to $b_k = \mathrm{e}^{\ln(1.5)-3(k-1)/10} = (1.5, 1.11, 0.82)$, for Σ_t we chose $\sigma_{t_k} = \mathrm{e}^{-(k-1)/10} = (1, 0.90, 0.82)$.

For comparing the parameter estimates with the true values θ, we computed the bias and the variance of the estimates. To deal with the identifiability up to sign, we multiplied each estimated loading vector by -1 if the inner product of the estimated loading vector and the true loading vector was negative. Moreover, we swapped columns in W and C to maintain the same ordering as the ordering in the true loadings. This was done to avoid inflation of the bias or variance due to a wrong sign or ordering of the individual components.

PPLS estimates were compared to PLS estimates (with orthogonal loadings, see [20] for an overview) for all scenarios above. For comparing PPLS with PPCA and PCCA, we constructed a 'null model', i.e., $B = 0$, as well as $B \neq 0$. We used the same scenarios as above, but we only considered the normal distribution.

Regarding standard errors for PPLS loadings, we compared asymptotic standard errors (as in Section 2) and bootstrap standard errors [27]. One set of two data matrices X and Y was simulated from a PPLS model with $p = q = 20$ normally distributed variables. Based on these data, asymptotic and bootstrap standard errors were calculated. The number of bootstrap replicates was 1000. Furthermore, simulation-based standard errors for the loadings (based on standard deviations over 1000 data matrices drawn from the PPLS model used to generate the

Table 1: Overview of the simulation scenarios. The noise level is defined as the proportion of variation in the noise matrices E, F and H relative to the total variation in X, Y and U respectively.

Sample size	$N = (50,\ 500)$
Dimensionality	$p = q = (20,\ 1000)$
Noise level	$\alpha_n = (0.1,\ 0.5)$
Distribution of t, u, e, f and h	$\{\mathcal{N}(0,1),\ t_2,\ \mathcal{P}(1),\ \mathcal{BIN}(2, 0.3)\}$

original data) were included as reference. Low and high noise levels ($\alpha_n = 0.1$ resp. $\alpha_n = 0.5$), and small, large and 'extra large' sample sizes ($N = 50$, $N = 500$ and $N = 5000$, respectively) were considered. In the 'extra large' sample size scenario, no simulation-based reference was calculated. The PPLS estimation algorithm was considered to be converged when either the log-likelihood increment was below 10^{-6}, or 10^4 EM steps were made. For each scenario, 1000 replicates are used.

3.1. Results

Results for the loadings. The biases and variances of the estimated first component W_1 for the low dimensional case for normally distributed latent variables are graphically depicted in Figure 1. A black dot represents the average estimated PPLS loading value across 1000 simulations, whereas the width of the black dashed vertical line equals two times the standard deviation across 1000 simulations. The red star and red dashed vertical line represent the average loading value and twice the standard deviation for the PLS estimates. The true loading values are represented by a step function with steps at each index $j \in \{1, \ldots, p\}$. Results for other components and scenarios are included in the Online Supplement.

Comparing the estimates for the *first* loading component W_1, a better performance of PPLS compared to PLS was observed in terms of bias. In all scenarios the bias of the PPLS estimators were about the same as or less than the bias of the PLS estimators. Both estimators showed larger bias towards zero for higher absolute loading values. The biases decreased with a larger sample size and lower noise level. The biases of both estimators were very similar across different distributions. In the scenario where there is 50% noise and few (50) samples the variance of the PPLS estimators tended to be slightly larger than the variances of the PLS estimators when the true loading values were larger. This was observed across all distributions. The variances of the PPLS estimates were about the same or lower than the PPLS estimates in all other scenarios, where either the noise level was less or more samples were available. For both PPLS and PLS estimators the variances tended to increase with higher loading values. The variances decreased with larger sample size and lower noise level. The variances of bots estimators were very similar across different distributions. For the loading component C_1 and their PLS and PPLS estimators the same conclusions were obtained.

For the *second* loading component W_2 (shown in the Online Supplement), the biases of the PPLS loading estimates were as good as, and often better than the PLS loading estimates, especially at lower values. In the scenarios of 50% noise and a small sample size ($N = 50$) the bias was slightly larger for PPLS estimators compared to PLS estimates when the loading values were larger. Both estimators showed larger bias towards zero for higher loading values. The biases decreased with a larger sample size and lower noise level. For all distributions, the biases of both estimators were very similar. The variances of the PPLS estimators were as good as or lower than the PLS estimators, except in the scenario in which both the noise level was high (50%) and the sample size was small (50). In this scenario the variances of the PPLS estimators were still lower if the true loading values were close to zero, but higher for higher loading values. For both PPLS and PLS estimators the variances tended to increase with higher loading values. The variances decreased with larger sample size and lower noise level. The variances of both estimators were very similar across different distributions. For the loading component C_2 and their PLS and PPLS estimators the same conclusions were obtained.

For the *third* loading components W_3 and C_3 (shown in the Online Supplement), the same observations were made as for the first loading components W_1 and C_1, both for the biases as for the variances.

For the high and extra high-dimensional case, the same results were obtained for the loadings W and C. See the Online Supplement for more details.

With regard to the comparison of PPLS with PPCA and PCCA, PPLS performed better than PCCA and similar to PPCA in most scenarios. Details are given in the Online Supplement.

Table 2: Proportion of correct order of loadings W and C across 1000 simulation replicates. These were obtained for different values of the dimensionality (high = 1000 variables, low = 20 variables), sample size (large = 500 subjects, small = 50 subjects) and noise level (high = 50% noise, low = 10% noise).

Dimensionality	Sample Size	Noise Level	Correct Ordering Proportion
low	large	low	1.000
		high	0.989
	small	low	0.932
		high	0.435
high	large	low	0.990
		high	0.985
	small	low	0.940
		high	0.665

Results for the variance parameters. The performance of the estimators of the variance parameters B, σ_t, σ_e, σ_f and σ_h were also evaluated, the results are shown in Figure 2. We did not compare with PLS as these model parameters are not present in the PLS framework. For sake of comparison, we calculated the relative biases and variances of the estimates with respect to the true corresponding parameter value. The biases of the PPLS estimators for all variance parameters were very small for large sample size ($N = 500$), regardless of the noise. For small sample size ($N = 50$), the first two diagonal elements of B and Σ_t were overestimated, while the last component was underestimated. The noise parameters σ_e and σ_f were underestimated in these scenarios, while the estimator for σ_h was unbiased, except in combination with a low noise level (10%). The relative biases decreased slightly with lower noise level, except for the earlier mentioned σ_h, and decreased more with larger sample size. The relative variances of the estimators of B, Σ_t and σ_h were larger than the variances of the estimators of σ_e and σ_f. For B, there was a slight increase in variance across the three components. The variances decreased slightly with lower noise and more with larger sample size. The variances slightly decreased in the scenario of high dimensionality and high noise level. The same observations were made across the different distributions.

Ordering of the loadings. We compared the ordering of the true loadings W and C with the ordering of the estimated loadings. This provides a proportion across the 1000 simulation replicates in which the ordering matched. In Table 2, the proportion of correct orderings of W for the scenario with normally distributed latent variables is shown for different scenarios. It can be seen that the proportion of correct orderings tends to be lower with smaller sample size and with higher noise level. Moreover, if the sample size is small, the proportion of correct orderings is much lower with higher noise. A higher dimensionality has a slightly negative impact on the correct ordering proportion when the sample size is larger, but a positive impact in the small sample size scenario. Especially, when also the noise level is high, this can be considerable. The same observations were made for the other distributions. Exactly the same proportions were observed for the loadings C.

Comparison of PPLS standard errors. The results for low noise level are shown in Figure 3. In all scenarios, the asymptotic standard errors were smaller than the bootstrap standard errors for nearly all loading elements. In particular, for high loading values the difference between asymptotic and bootstrap standard errors tended to be large. This difference decreased with larger sample size: In the 'extra large' sample size, the bootstrap and asymptotic standard errors had similar magnitude. Similar observations were made for other distributions. For details, see the Online Supplement.

4. Data analysis

To illustrate the Probabilistic Partial Least Squares model, we apply it to IgG glycan datasets. Glycans, in particular IgG glycans, play an important role in the innate immune system, as well as in cell signaling. IgG2 glycans are less abundant than IgG1 glycans and more difficult to measure. Therefore, by using the relationships between IgG1 and IgG2 glycans, the characteristics of IgG2 can be better estimated. Hence, we will use IgG1 as X matrix, and IgG2 as Y matrix.

In total, 40 IgG glycans were measured, of which $p = 20$ are of subclass IgG1 and $q = 20$ are of subclass IgG2. These data were measured in two cohorts (CROATIA_Korcula with 951 participants and CROATIA_Vis with 796

participants) [12]. The data matrices containing IgG1 and IgG2 glycan variables are denoted by X_m and Y_m, with $m \in \{1, 2\}$, where $m = 1$ corresponds to CROATIA_Korcula and $m = 2$ corresponds to CROATIA_Vis. We apply PPLS to IgG1 and IgG2 glycans in each cohort *separately* and compare results.

In Figure 4, heatmaps of the correlations within and between the IgG1 and IgG2 glycans are shown, from which it is clear that there are many highly positive correlations between and within IgG1 and IgG2 in each data set. The RV coefficient [18], which generalizes the *squared* Pearson correlation coefficient to two matrices, was about 0.60 and 0.45 for CROATIA_Korcula and CROATIA_Vis cohorts respectively.

To determine the number of latent variables to use, we considered the total amount of variance explained by the latent space relative to the total amount of variation in the data: $\|T_m\|/\|X_m\|$ and $\|U_m\|/\|Y_m\|$ for $m \in \{1, 2\}$. By using four components, at least 89% of the total variation in each of the matrices X_1, X_2, Y_1 and Y_2 was accounted for.

For both cohorts, we fitted the PPLS models using $r = 4$ latent components. The amount of overlap in each cohort, estimated by $\mathrm{tr}\hat{\Sigma}_{x,y}/\mathrm{tr}\hat{\Sigma}_y$ given in (4), was 58% and 46% for CROATIA_Korcula and CROATIA_Vis cohorts, respectively. The PPLS loadings are inspected to identify which IgG glycans contribute most to this overlap. The estimated IgG1 loadings $w_{j,k}$, $j \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, 4\}$, for both cohorts and both subclasses are depicted in Figure 5. The first joint component is proportional to the average glycan, as all glycans get about the same loading value. The second joint component involves especially G0 and G2 glycan subtypes, in which they are negatively correlated. Inspection of the loading values for the third component shows contibutions of fucosylated and non-fucosylated glycan subtypes. In the fourth component a pattern of positive and negative loading values is visible regarding the presence and absence of bisecting GlcNAc, respectively. The large loading value for G1NS is remarkable. The same conclusions hold for IgG2, as the estimated loading values were very similar. It is interesting to note that the observed patterns within components potentially reflect enzymatic synthesis where monosaccharides are added to glycan substrates [22]. Additionally, similar patterns are seen reflecting the inflammatory characteristics of glycans in aging and several different diseases [13]. Finally, the observed loading patterns were strikingly similar for both cohorts.

5. Discussion

We proposed PPLS to model the covariance matrix of two data sets. Maximum likelihood estimators for the model parameters were derived by solving a constrained optimization problem, and the parameter loadings were shown to be identifiable up to sign. This property ensures that PPLS estimates are comparable across several studies.

Our simulation study showed that the PPLS estimators had good performance and lower bias compared to PLS. Most notably, the performance of PPLS was robust to misspecification of the distribution of the variables. A smaller sample size and high noise level had a negative effect on the accuracy of the estimates, but large loading values were still non-zero. Also, compared to Probabilistic CCA estimates, the PPLS estimates were less biased and more efficient. For high-dimensional data, PCCA estimates have larger bias and higher variance. This is likely to be caused by the unstable inverse sample covariance matrix calculated when using PCCA. Moreover, if the number of variables is larger than the sample size, PCCA estimates cannot be obtained. Therefore, especially in omics data analysis, PPLS provides more robust findings.

As an illustration of the PPLS model, we analyzed IgG glycomics data from two cohorts. The high correlations in the data (Figure 4) and the use of multiple cohorts illustrate the applicability of PPLS to facilitate combination of results derived from different experimental settings. We found that the estimated loading values were almost identical across the two cohorts (Figure 5).

When multiple cohorts are available, a meta-analysis on the parameter estimates can be performed. In ordinary regression models, this has been addressed for both low-dimensional [6] and high-dimensional [10] design matrices. When there is no access to all data, PPLS estimates can be combined by using standard meta-analysis approaches [6]. Such an approach requires that the PPLS parameter estimates are identifiable and asymptotically normally distributed. For the PLS framework, several approaches to combine estimates across cohorts were developed when there is access to all data. A group-PLS approach was considered [15] to incorporate several groups of subjects in the model. The authors showed that under certain assumptions this approach provided better predictions than a model without group effects. However their model is not identifiable and requires $N > p$. Another method is based on weighted least squares to combine data from different studies with potentially different covariates [11]. An alternative method, when access to data is possible, is to estimate joint parts between the data sets and the studies simultaneously. This yields a joint space with variables that have high loading values in most

studies. For example, in [25], a non-probabilistic approach is pursued in a least squares estimation method using PCA. Performing data integration across studies, while taking into account uncertainties within each study, is one of our topics for future research, and will lead to more powerful analysis of the IgG glycans across cohorts.

To assess the statistical significance of loadings, the probabilistic framework provides alternative approaches to jackknifing and bootstrapping [27]. The observed Fisher information matrix can be used to estimate standard errors for individual loading parameters. For small sample sizes, bootstrap approaches appears to better reflect the uncertainty of the parameters. For large enough sample sizes, the asymptotic standard errors are close to the simulation-based standard errors. Typically, in epidemiological studies, the sample size is large enough to use asymptotic standard errors.

In this paper we ignored the fact that different biological 'omics' measurements have different error structures. An extension of Partial Least Squares was proposed to correct for systematic variation (variation induced by latent variables uncorrelated to the other data set) in the data sets, named Two-Way Orthogonal PLS (O2PLS) [8, 24]. Such an extension can be pursued for PPLS by adding for both X and Y in (1) a set of independent latent variables multiplied by their loading parameters. We are currently working on exploring the possibilities of a Probabilistic O2PLS for heterogeneous data sets.

Acknowledgments

The authors would like to thank the Editor-in-Chief, the Associate Editor and the referees for their valuable comments and suggestions. This work has been supported by the European Unions Seventh Framework Programme (FP7-Health-F5-2012) under grant agreement number 305280 (MIMOmics). The CROATIA_Vis and CROATIA_Korcula studies were funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006-018947), FP7 contract BBMRI-LPC (grant No. 313010), Croatian Science Foundation (grant 8875) and the Republic of Croatia Ministry of Science, Education and Sports (216-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools, Institute for Anthropological Research in Zagreb and the Croatian Institute for Public Health. Glycome analysis was supported by the European Commission HighGlycan (contract No. 278535), MIMOmics (contract No. 305280), HTP-GlycoMet (contract No. 324400), IntegraLife (contract No. 315997). The IgG glycan data are available upon request.

Appendix A Variances and covariances

The covariance matrix of (x, y) is given in (4). First note that $\mathrm{var}(u) = \mathrm{var}(tB + h) = B^2\Sigma_t + \sigma_h^2 I_r$, then compute

$$\begin{aligned}
\mathrm{var}(x) &= \mathrm{var}(tW^\top + e) = W\mathrm{var}(t)W^\top + \mathrm{var}(e) = W\Sigma_t W^\top + \sigma_e^2 I_p,\\
\mathrm{var}(y) &= \mathrm{var}(UC^\top + f) = C\mathrm{var}(u)C^\top + \mathrm{var}(f) = C(B^2\Sigma_t + \sigma_h^2 I_r)C^\top + \sigma_f^2 I_q,\\
\mathrm{cov}(x, y) &= \mathrm{cov}(tW^\top + e, uC^\top + f) = W\mathrm{cov}(t, u)C^\top = W\mathrm{cov}(t, tB)C^\top = WB\Sigma_t C^\top.
\end{aligned}$$

The covariances between the observed and latent variables are as follows

$$\begin{aligned}
\mathrm{cov}(x, t) &= \mathrm{cov}(tW^\top + e, t) = W\mathrm{var}(t) = W\Sigma_t,\\
\mathrm{cov}(x, u) &= \mathrm{cov}(tW^\top + e, tB + h) = W\mathrm{var}(t)B = W\Sigma_t B,\\
\mathrm{cov}(y, t) &= \mathrm{cov}(uC^\top + f, t) = C\mathrm{cov}(tB + h, t) = C\Sigma_t B,\\
\mathrm{cov}(y, u) &= \mathrm{cov}(uC^\top + f, u) = C\mathrm{cov}(tB + h, tB + h) = C(\Sigma_t B^2 + \sigma_h^2 I_r).
\end{aligned}$$

See, e.g., [21] for more details.

Appendix B Identifiability of PPLS

For establishing identifiability of the PPLS model, we need to prove 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 is equivalent to

$$\Sigma = \tilde{\Sigma} \quad \Leftrightarrow \quad \theta = \tilde{\theta},$$

where $\Sigma, \tilde{\Sigma}$ is defined, through $\theta, \tilde{\theta}$, in (4). The following lemma will be very useful in establishing identifiability.

Lemma 1. (Singular Value Decomposition). *Let $W, \tilde{W}$ be $p \times r$ and $C, \tilde{C}$ be $q \times r$, all orthogonal matrices. Let $D, \tilde{D}$ be $r \times r$ diagonal with r distinct positive elements on the diagonal. Then $WDC^\top = \tilde{W}\tilde{D}\tilde{C}^\top$ (B.1) implies $W = \tilde{W}\Delta$, $C = \tilde{C}\Delta$ for some diagonal matrix Δ of size $r \times r$ with on the diagonal elements $\delta_i \in \{-1, 1\}$ and $D = \tilde{D}$.*

Proof. Let $A_1 = WDC^\top$ and $A_2 = \tilde{W}\tilde{D}\tilde{C}^\top$. Consider $A_i A_i^\top$ and $A_i^\top A_i$, $i \in \{1, 2\}$. The assertion (B.1) then implies the following.

$$A_1 A_1^\top = WD^2W^\top = \tilde{W}\tilde{D}^2\tilde{W}^\top = A_2 A_2^\top; \quad A_1^\top A_1 = CD^2C^\top = \tilde{C}\tilde{D}^2\tilde{C}^\top = A_2^\top A_2.$$

Note that both $WD^2W^\top$ and $\tilde{W}\tilde{D}^2\tilde{W}^\top$ are eigenvalue decompositions, as D^2 and $\tilde{D}^2$ are diagonal and $W, \tilde{W}$ and $C, \tilde{C}$ are orthogonal. The spectral theorem for matrices [7] then implies that whenever the elements in $D^2, \tilde{D}^2$ are distinct, the corresponding columns in $W, \tilde{W}$ and $C, \tilde{C}$ are equal up to multiplication with the same sign. We thus get $W = \tilde{W}\Delta$, $C = \tilde{C}\Delta$ and $D = \tilde{D}$. □

Using this lemma, we show identifiability of the off-diagonal block part of the covariance matrix as given in Eq. (4).

Lemma 2. *If for matrices $W, \tilde{W}$, $C, \tilde{C}$ and diagonal $B, \tilde{B}$ and $\Sigma_t, \tilde{\Sigma}_t$, given as in the PPLS model, $W\Sigma_t BC^\top = \tilde{W}\tilde{\Sigma}_t\tilde{B}\tilde{C}^\top$, then $W = \tilde{W}\Delta$, $C = \tilde{C}\Delta$ and $\Sigma_t B = \tilde{\Sigma}_t\tilde{B}$.*

Proof. Applying Lemma 1 with $D = \Sigma_t B$ and $\tilde{D} = \tilde{\Sigma}_t\tilde{B}$ gives the desired result, since $\Sigma_t B$ and $\tilde{\Sigma}_t\tilde{B}$ are diagonal matrices with distinct ordered elements. □

Given $\Sigma_{x,y}$ we can identify W and C up to sign and the product $\Sigma_t B$. We now show that in particular also the individual parameters Σ_t and B are identified from the upper diagonal block matrix Σ_x.

Lemma 3. *If for matrix W, diagonal matrices Σ_t and $\tilde{\Sigma}_t$ and positive numbers $\sigma_e^2, \tilde{\sigma}_e^2$, given as in the PPLS model, $W\Sigma_t W^\top + \sigma_e^2 I_p = W\tilde{\Sigma}_t W^\top + \tilde{\sigma}_e^2 I_p$ (B.2), then $\sigma_e = \tilde{\sigma}_e$ and $\Sigma_t = \tilde{\Sigma}_t$.*

Proof. Suppose (B.2) holds. Since $r < p$ and $p > 1$, one can find a unit vector $w_\perp$ such that $W^\top w_\perp = 0$. Multiplying with such vector yields $\sigma_e^2 w_\perp = \tilde{\sigma}_e^2 w_\perp$. Multiplying again with $w_\perp^\top$ yields $\sigma_e^2 = \tilde{\sigma}_e^2$. It follows that we can identify σ_e^2. We can now reduce (B.2) to $W\Sigma_t W^\top = W\tilde{\Sigma}_t W^\top$. Pre-multiplying with $W^\top$ and post-multiplying with W on both sides yields $\Sigma_t = \tilde{\Sigma}_t$. □

We have seen in Theorem 2 that we can identify $\Sigma_t B$. Since we identified Σ_t we get identifiability of B. The remaining parameters σ_h^2 and σ_f^2 are now shown to be identified using the lower block diagonal Σ_y.

Lemma 4. *If for matrices C, B, Σ_t, $\sigma_f^2, \tilde{\sigma}_f^2$ and $\sigma_h^2, \tilde{\sigma}_h^2$, given as in the PPLS model, the assertion $\Sigma_y = \tilde{\Sigma}_y$ holds, i.e.,*

$$C(B^2\Sigma_t + \sigma_h^2 I_r)C^\top + \sigma_f^2 I_q = C(B^2\Sigma_t + \tilde{\sigma}_h^2 I_r)C^\top + \tilde{\sigma}_f^2 I_q,$$

then $\sigma_f^2 = \tilde{\sigma}_f^2$ and $\sigma_h^2 = \tilde{\sigma}_h^2$.

Proof. In Theorem 3 take W equal to C, σ_e^2 equal to σ_f^2, $\tilde{\sigma}_e^2$ equal to $\tilde{\sigma}_f^2$, and the diagonal covariance matrices Σ_t and $\tilde{\Sigma}_t$ equal to $\Sigma_t B^2 + \sigma_h^2 I_p$ and $\Sigma_t B^2 + \tilde{\sigma}_h^2 I_p$. We find that we can identify $\Sigma_t B^2 + \sigma_h^2$ and σ_f^2. Since we already identified Σ_t and B, we have also identifiability of σ_h^2. □

We conclude with the proof of Theorem 1.

Proof. Suppose $\Sigma = \tilde{\Sigma}$. This is true if and only if

$$\Sigma_{x,y} = \tilde{\Sigma}_{x,y}, \quad \Sigma_x = \tilde{\Sigma}_x, \quad \Sigma_y = \tilde{\Sigma}_y. \tag{B.1}$$

Applying Lemma 2 to the first equation, we identify W and C up to sign. Considering Lemma 3 together with Lemma 2, the second equation implies identifiability of Σ_t, B and σ_e. The three Lemmas 2, 3 and 4 together with the last equation imply identifiability of σ_h and σ_f.

□

Appendix C An Expectation-Maximization algorithm for PPLS

To obtain parameter estimates in the PPLS model, maximum likelihood is used. The EM algorithm is an iterative procedure for maximizing the observed log-likelihood (5) and consists of an Expectation step and a Maximization step. The following Lemma is convenient to make the expectation step explicit.

Lemma 5. *Let the pair* (z, x) *be jointly multivariate normal row vectors with zero mean and covariance matrix*

$$\begin{pmatrix} \Sigma_z & \Sigma_{z,x} \\ \Sigma_{x,z} & \Sigma_x \end{pmatrix}.$$

Then $z|x$ *is normally distributed with conditional mean* $\mathrm{E}(z|x) = x\Sigma_x^{-1}\Sigma_{x,z}$*, and conditional covariance matrix* $\mathrm{var}(z|x) = \Sigma_z - \Sigma_{z,x}\Sigma_x^{-1}\Sigma_{x,z}$*. Secondly, if* $z = (t, u)$*,* $\mathrm{cov}(t, x) = \Sigma_{t,x}$ *and* $\mathrm{cov}(x, u) = \Sigma_{x,u}$*, then the conditional covariance between* t *and* u *is* $\mathrm{cov}(t, u|x) = \mathrm{cov}(t, u) - \Sigma_{t,x}\Sigma_x^{-1}\Sigma_{x,u}$*.*

Proof. The proof for the first part of the Lemma is found in [21]. The second part follows from the off diagonal blocks of $\mathrm{var}(z|x)$. □

Expectation. The conditional first moments can be obtained by applying Lemma 5 while substituting t or u for z and (x, y) for x.

$$\mu_t = \mathrm{E}(t|x, y, \theta) = (x, y)\,\Sigma^{-1}\,\mathrm{cov}\{(x, y), t\}, \quad \mu_u = \mathrm{E}(u|x, y, \theta) = (x, y)\,\Sigma^{-1}\,\mathrm{cov}\{(x, y), u\}.$$

The same substitution can be made for the conditional second moments. Using $\mathrm{E}(a^\top b|z) = \mathrm{cov}(a, b|z) + \mathrm{E}(a|z)^\top \mathrm{E}(b|z)$, we get

$$\begin{aligned} C_{TT} &= \mathrm{E}(t^\top t|x, y, \theta) = I_r - \mathrm{cov}\{t, (x, y)\}\,\Sigma^{-1}\,\mathrm{cov}\{(x, y), t\} + \mathrm{cov}\{t, (x, y)\}\,\Sigma^{-1} S \Sigma^{-1}\,\mathrm{cov}\{(x, y), t\}, \\ C_{UU} &= \mathrm{E}(u^\top u|x, y, \theta) = I_r - \mathrm{cov}\{u, (x, y)\}\,\Sigma^{-1}\,\mathrm{cov}\{(x, y), u\} + \mathrm{cov}\{u, (x, y)\}\,\Sigma^{-1} S \Sigma^{-1}\,\mathrm{cov}\{(x, y), u\}, \end{aligned}$$

where S is the biased sample covariance matrix of (x, y). The conditional cross term equals

$$C_{UT} = \mathrm{E}(u^\top t|x, y, \theta) = \Sigma_t B - \mathrm{cov}\{u, (x, y)\}\,\Sigma^{-1}\,\mathrm{cov}\{(x, y), t\} + \mathrm{cov}\{u, (x, y)\}\,\Sigma^{-1} S \Sigma^{-1}\,\mathrm{cov}\{(x, y), t\}$$

The covariances are given by

$$\mathrm{cov}\{(x, y), t\} = \begin{pmatrix} W\Sigma_t \\ C\Sigma_t B \end{pmatrix}, \quad \mathrm{cov}\{(x, y), u\} = \begin{pmatrix} W\Sigma_t B \\ C(\Sigma_t B + \sigma_h^2 I_r) \end{pmatrix}.$$

Although the the conditional expectations involve random variables and parameters, in the maximization step the calculated quantities are considered fixed and known.

Maximization. The maximization step involves maximizing the complete-data likelihood (6), we have seen that it can be decomposed in distinct factors. This allows optimization of the expected complete data likelihood to be split into four sub-maximizations, given by the individual factors and their respective parameters in the following annotated product:

$$\underbrace{f(x|t)}_{W,\sigma_e}\,\underbrace{f(y|u)}_{C,\sigma_f}\,\underbrace{f(u|t)}_{B,\sigma_h}\,\underbrace{f(t)}_{\Sigma_t}$$

Moreover, it will become apparent that each parameter within each component can be decoupled, yielding a separate maximization per component per parameter. We focus on the part of $f(x|t)$ that depends on W, which is given by

$$\begin{aligned} \mathrm{E}\{\ln f(X|T)|X, Y\} &= -\mathrm{E}(\|X - TW^\top\|^2|X, Y) + \text{const.} \\ &= \mathrm{tr}(-X^\top X + 2X^\top \mu_t W^\top - WC_{TT}W^\top) + \text{const.} \end{aligned}$$

To take into account the constraints on W, namely $W^\top W = I_r$, we introduce a matrix of Lagrange multipliers Λ. We get as objective function

$$\mathrm{tr}(-X^\top X + 2X^\top \mu_t W^\top - WC_{TT}W^\top) - \mathrm{tr}\{(W^\top W - I_r)\Lambda\}.$$

Differentiating with respect to W yields $2X^\top\mu_t - 2WC_{TT} - 2W\Lambda = 2W\,(C_{TT} + \Lambda) - 2X^\top\mu_t$. One may choose Λ so that $C_{TT} + \Lambda$ is invertible. In a maximum W then satisfies $W = X^\top\mu_t\,(C_{TT} + \Lambda)^{-1}$. We want to find a Λ such that the constraint holds, i.e.,

$$I_r = W^\top W = \{(C_{TT} + \Lambda)^{-1}\}^\top\mu_t^\top XX^\top\mu_t\,(C_{TT} + \Lambda)^{-1}\,, \quad \mu_t^\top XX^\top\mu_t = (C_{TT} + \Lambda)^\top\,(C_{TT} + \Lambda)$$

The last identity can be recognized as a Cholesky or Eigenvalue decomposition.

$$\mu_t^\top XX^\top\mu_t = (C_{TT} + \Lambda)^\top\,(C_{TT} + \Lambda) = L_tL_t^\top$$

with L_t the lower triangular matrix of a Cholesky decomposition of $\mu_t^\top XX^\top\mu_t$. Note that L_t exists, since $\mu_t^\top XX^\top\mu_t$ is always positive semi-definite. Choosing $\Lambda = L_t^\top - C_{TT}$, we get as update $W = X^\top\mu_t(L_t^\top)^{-1}$. Following the same reasoning, we obtain for the $f(Y|U)$ part $C = Y^\top\mu_u(L_u^\top)^{-1}$, where L_u is the lower triangular matrix from the Cholesky decomposition of $\mu_u^\top YY^\top\mu_u$.

The parameter B involves maximizing $\ln f(U|T)$, which is given by

$$-\|U - TB\|^2 = -\mathrm{tr}(U^\top U - 2U^\top TB + BT^\top TB) + \mathrm{const.}$$

Taking the conditional expectation with respect to (x, y) yields $-\mathrm{tr}\,\mathrm{E}(U^\top U - 2U^\top TB + BT^\top TB|X, Y)$. Differentiating with respect to B and equating to the zero matrix yields

$$B\mathrm{E}(T^\top T|X, Y) = \mathrm{E}(U^\top T|X, Y) \quad B = \mathrm{E}(U^\top T|X, Y)\{\mathrm{E}(T^\top T|X, Y)\}^{-1}$$

To incorporate the constraint that B should be diagonal, we set the diagonal elements to zero, yielding

$$B = \mathrm{E}(U^\top T|X, Y)\{\mathrm{E}(T^\top T|X, Y)\}^{-1} \circ I_r,$$

with $\circ$ the element-wise (Hadamard) product operator.

For the covariance matrix of Σ_t, we consider $\ln f(T)$ which is given by

$$2\ln f(T) = \mathrm{const.} - N\ln|\Sigma_T| - \mathrm{tr}(T^\top T\Sigma_T^{-1}) = \mathrm{const.} + N\ln|\Sigma_T^{-1}| - \mathrm{tr}(T^\top T\Sigma_T^{-1}).$$

After taking the conditional expectation of the last expression, it can be differentiated with respect to Σ_t^{-1}, which yields

$$2\frac{\partial}{\partial\,\Sigma_T^{-1}}\ln f(T) = N\Sigma_T - \mathrm{E}(T^\top T|x, y) = 0, \quad \Sigma_T = N^{-1}\mathrm{E}(T^\top T|x, y) \circ I_r$$

The last Hadamart product ensures Σ_t is diagonal.

To maximize over σ_e^2, we consider $\ln f(X|T)$ and note that $E = X - TW^\top$. Then $\ln f(X|T)$ is given by

$$2\ln f(X|T) = \mathrm{const.} - Np\ln|\sigma_e^2| - \sigma_e^{-2}\mathrm{tr}(E^\top E) = \mathrm{const.} + Np\ln\sigma_e^{-2} - \sigma_e^{-2}\mathrm{tr}(E^\top E)$$

After taking the conditional expectation of the last expression, we differentiate it with respect to σ_e^{-2}, yielding

$$2\frac{\partial}{\partial\,\sigma_e^{-1}}\ln f(X|T) = Np\sigma_e^2 - \mathrm{E}(E^\top E|X, Y) = 0, \quad \sigma_e^2 = (Np)^{-1}\mathrm{E}(E^\top E|X, Y)$$

The same derivation can be applied to $\ln f(y|u)$ and $\ln f(u|t)$ to find

$$\sigma_f^2 = (Nq)^{-1}\mathrm{E}(F^\top F|X, Y), \quad \sigma_h^2 = (Nr)^{-1}\mathrm{E}(H^\top H|X, Y)$$

Appendix D Asymptotic standard errors for PPLS loadings

Using notation as in [16] we define

$$\lambda(W_k) = -\frac{1}{2\sigma_e^2}\mathrm{tr}(X^\top X - 2X^\top t_kw_k^\top + w_kt_k^\top t_kw_k^\top)$$

to be the part of the log likelihood depending on w_k. We calculate the following first and second derivatives.

$$S(w_k) = \nabla\lambda = \sigma_e^{-2}(X^\top t_k - w_k t_k^\top t_k), \quad B(w_k) = -\nabla^2\lambda = \sigma_e^{-2}(t_k^\top t_k)I_p.$$

We obtain

$$\begin{aligned}
\sigma_e^4 S(w_k)S(w_k)^\top &= X^\top t_k t_k^\top X - 2X^\top t_k t_k^\top t_k w_k^\top + w_k t_k^\top t_k t_k^\top t_k w_k^\top, \\
\sigma_e^4 \mathrm{E}\{S(w_k)S(w_k)^\top|X,Y\} &= X^\top \mathrm{E}(t_k t_k^\top|X,Y)X - 2X^\top \mathrm{E}(t_k t_k^\top t_k|X,Y)w_k^\top + w_k \mathrm{E}(t_k^\top t_k t_k^\top t_k|X,Y)w_k^\top \\
&= \sigma_k^2 X^\top X - 2X^\top(\mu_k\|\mu_k\|_2^2 + 3\mu_k\sigma_k^2)w_k^\top + w_k(\|\mu_k\|_2^4 + 6\|\mu_k\|_2^2\sigma_k^2 + 3\sigma_k^4)w_k^\top.
\end{aligned}$$

Here $\mu_k = \mathrm{E}(t_k|X,Y)$ and $\sigma_k = \mathrm{E}(t_k^\top t_k|X,Y)$. For explicit expressions of these expectations, see Appendix C. For the second derivative we get $\mathrm{E}\{B(w_k)|X,Y\} = \sigma_k^2 I_p/\sigma_e^2$. The observed Fisher information is now

$$I_{\text{obs}} = \mathrm{E}\{B(w_k)|X,Y\} - \mathrm{E}\{S(w_k)S(w_k)^\top|X,Y\},$$

and the asymptotic covariance matrix of w_k is $-I_{\text{obs}}^{-1}$. The square root of the diagonal elements are the standard errors of the corresponding loading elements.

References

[1] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS Regression), Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 97–106.

[2] F.R. Bach, M. Jordan, A Probabilistic Interpretation of Canonical Correlation Analysis, Dept. Stat. Univ. California, Berkeley, CA, Tech. Rep 688 (2005) 1–11.

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

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

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1977) 1–38.

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

[7] M. Eaton, Multivariate Statistics: A Vector Space Approach, Wiley, New York, 1983.

[8] S. el Bouhaddani, J. Houwing-Duistermaat, P. Salo, M. Perola, G. Jongbloed, H.W. Uh, Evaluation of O2PLS in Omics data integration, BMC Bioinformatics 17.

[9] S. Geisser, Predictive Inference, Taylor & Francis, London, 1993.

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

[11] X. Huang, W. Pan, X. Han, Y. Chen, L.W. Miller, J. Hall, Borrowing information from relevant microarray studies for sample classification using weighted partial least squares, Comput. Biol. Chem. 29 (2005) 204–211.

[12] 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, I. Rudan, Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers, PLoS Genet. 9 (2013) e1003225.

[13] G. Lauc, M. Pezer, I. Rudan, H. Campbell, Mechanisms of disease: The human N-glycome, Biochim. Biophys. Acta 1860 (2016) 1574–82.

[14] S. Li, J.O. Nyagilo, D.P. Dave, W. Wang, B. Zhang, J. Gao, Probabilistic partial least squares regression for quantitative analysis of Raman spectra, Int. J. Data Min. Bioinform. 11 (2015) 223–243.

[15] Y. Li, P. Uden, D. von Rosen, A two-step PLS inspired method for linear prediction with group effect, Sankhyā Indian J. Stat. 75 (2013) 96–117.

[16] T.A. Louis, Finding the observed information matrix when using the EM algorithm, J. R. Stat. Soc. Ser. B 44 (1982) 226–233.

[17] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate analysis, Academic Press, London, 1979.

[18] P. Robert, Y. Escoufier, A unifying tool for linear multivariate statistical methods: The rv-coefficient, J. R. Stat. Soc. Ser. C 25 (1976) 257–265.

[19] B. Roś, F. Bijma, J.C. de Munck, M.C.M. de Gunst, Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure, J. Multivariate Anal. 143 (2016) 345–361.

[20] R. Rosipal, N. Krämer, Overview and Recent Advances in Partial Least Squares, in: Proc. 2005 Int. Conf. Subspace, Latent Struct. Featur. Sel., Vol. 3940 of SLSFS'05, Springer, New York, 2006, pp. 34–51.

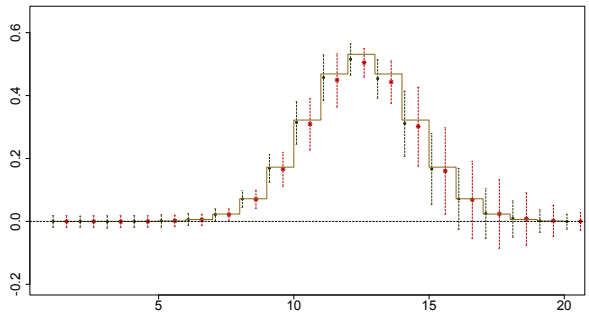
[21] G.A.F. Seber, A.J. Lee, Linear Regression Analysis, 2nd ed., Wiley, Hoboken, NJ, 2003.

[22] N. Taniguchi, K. Honke, M. Fukuda, H. Narimatsu, Y. Yamaguchi, T. Angata (Eds.), Handbook of Glycosyltransferases and Related Genes, Springer Japan, Tokyo, 2014.

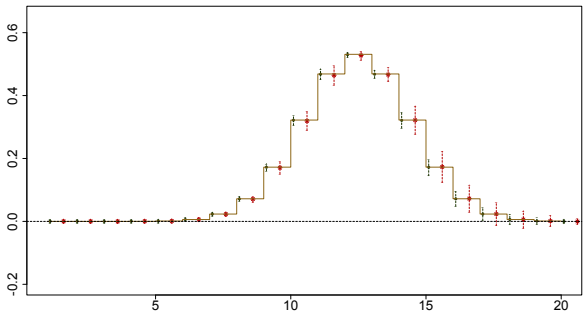
[23] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc. Ser. B 61 (1999) 611–622.

[24] J. Trygg, S. Wold, O2-PLS, a two-block (XY) latent variable regression (LVR) method with an integral OSC filter, J. Chemom. 17 (2003) 53–64.

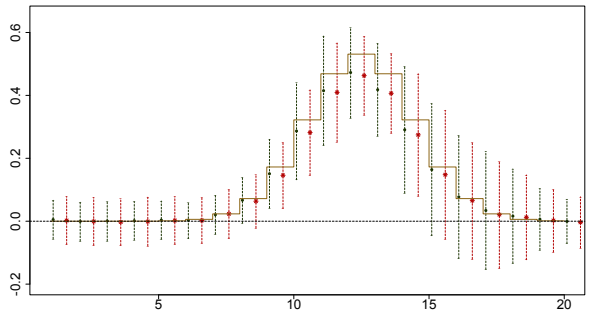
[25] K. Van Deun, A.K. Smilde, M.J. van der Werf, H.A.L. Kiers, I. Van Mechelen, A structured overview of simultaneous component based data integration, BMC Bioinformatics 10 (2009) 246.
[26] H. Wang, Q. Liu, Y. Tu, Interpretation of partial least-squares regression models with VARIMAX rotation, Comput. Stat. Data Anal. 48 (2005) 207–219.
[27] R. Wehrens, W.E. van der Linden, Bootstrapping principal component regression models, J. Chemom. 11 (1997) 157–171.
[28] 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), Academic Press, New York, 1973, pp. 383–407.
[29] C.F.J. Wu, On the convergence properties of the EM algorithm, Ann. Stat. 11 (1983) 95–103.
[30] J. Zheng, Z. Song, Z. Ge, Probabilistic learning of partial least squares regression model: Theory and industrial applications, Chemom. Intell. Lab. Syst. 158 (2016) 80–90.



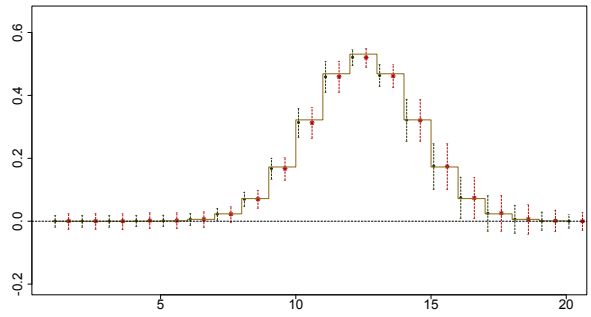
(a) Low noise (10%); small sample size ($N = 50$)



(b) Low noise (10%); large sample size ($N = 500$)

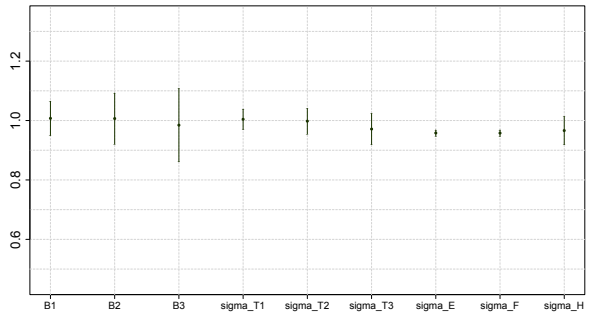


(c) High noise (50%); small sample size ($N = 50$)

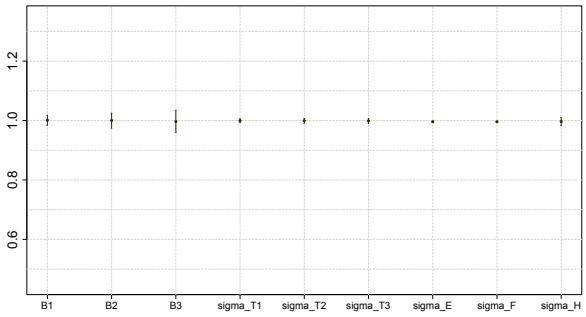


(d) High noise (50%); large sample size ($N = 500$)

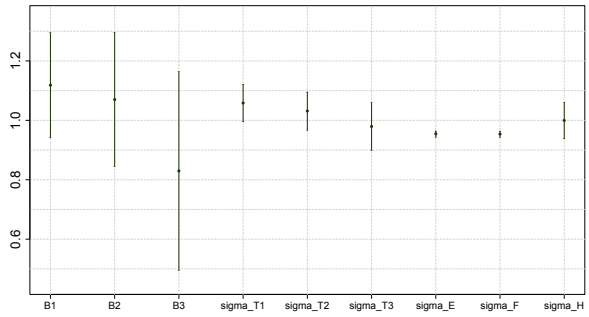
Figure 1: True and estimated loadings W_1 over 1000 simulation replications. The black dots and dashed vertical lines (on the left of each pair) represent PPLS estimates, the red stars and dashed vertical lines (on the right of each pair) represent PLS estimates. The dots and stars are the average loading values across 1000 simulation replications; the width of the dashed lines are twice the standard deviations. The results are for normally distributed latent variables (t, e, f and h) and low dimensionality ($p = q = 20$ variables).



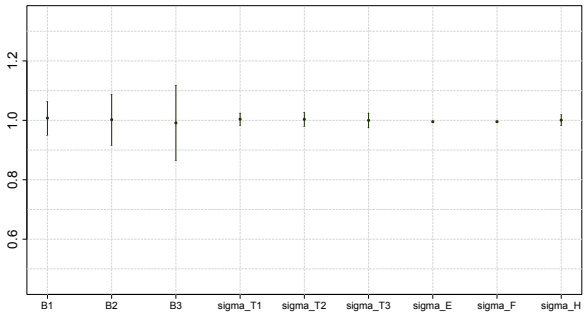
(a) Low noise (10%); small sample size ($N = 50$)



(b) Low noise (10%); large sample size ($N = 500$)



(c) High noise (50%); small sample size ($N = 50$)



(d) High noise (50%); large sample size ($N = 500$)

Figure 2: True and estimated variance parameters B, Σ_t, σ_e, σ_f and σ_h over 1000 simulation replications. The dots are the average values across 1000 simulation replications; the width of the dashed lines are twice the standard deviations. The results are for normally distributed latent variables (t, e, f and h) and low dimensionality ($p = q = 20$ variables).

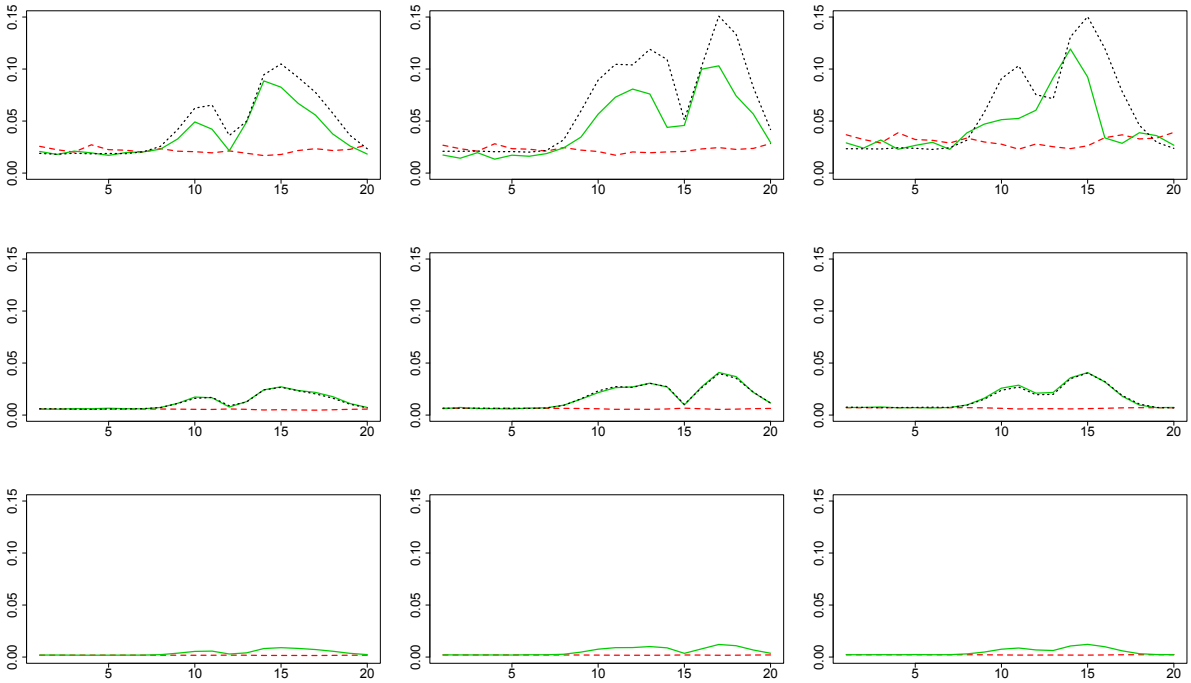
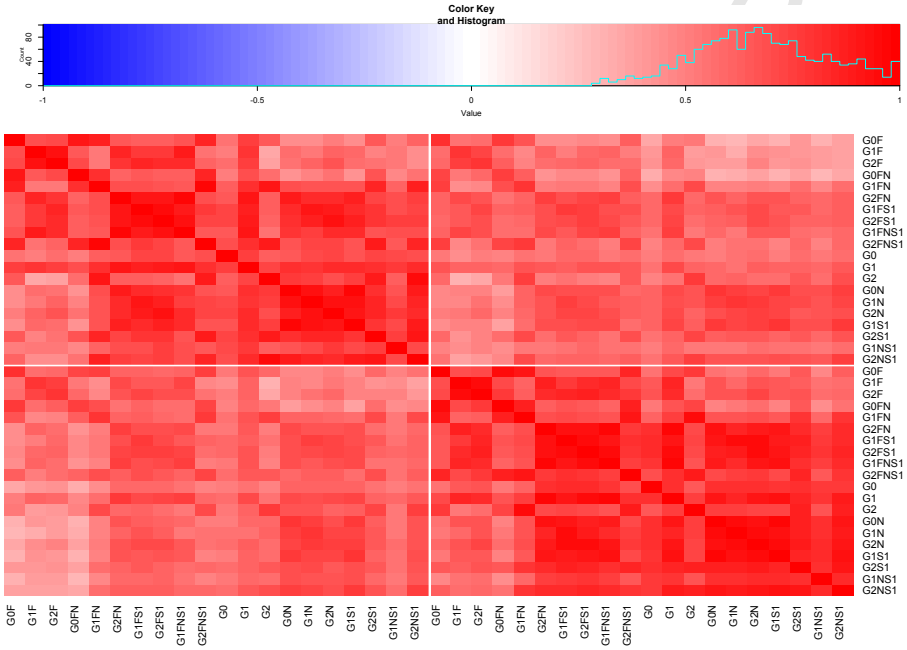
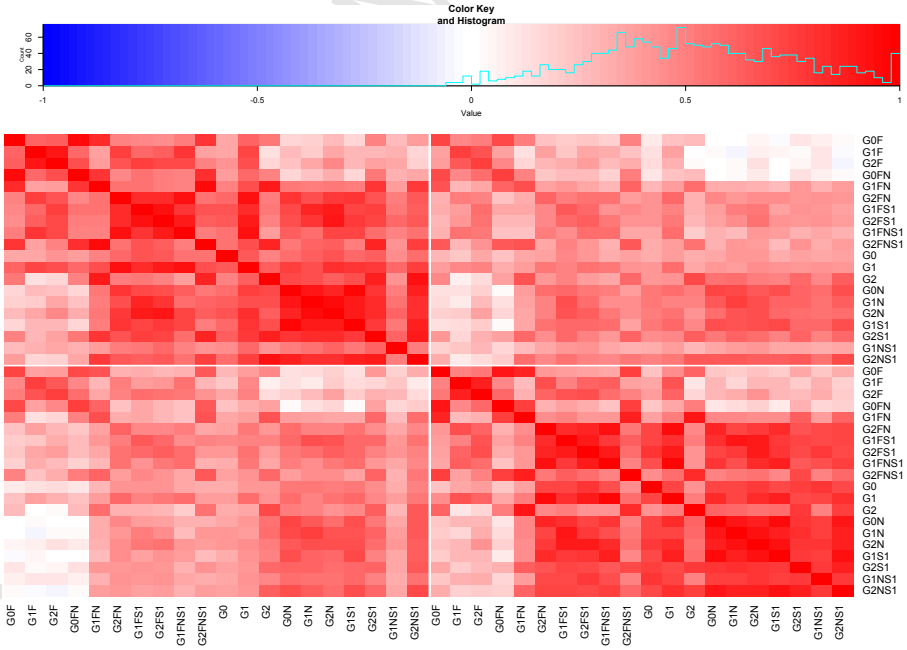


Figure 3: Standard errors of the W loading elements per component. Bootstrap standard errors (solid green line), asymptotic standard errors (dashed red line) and simulation-based standard errors (dotted black line) are plotted for the loading estimates in each component. Plots for the three sample sizes (small $N = 50$, large $N = 500$, 'extra large' $N = 5000$) are shown along the rows. The three loading components (W_1, W_3 and W_3) are plotted column wise. The last row does not include simulation-based standard errors. The data are generated from a normal distribution with $p = q = 20$ variables and low noise level ($\alpha_n = 0.1$).

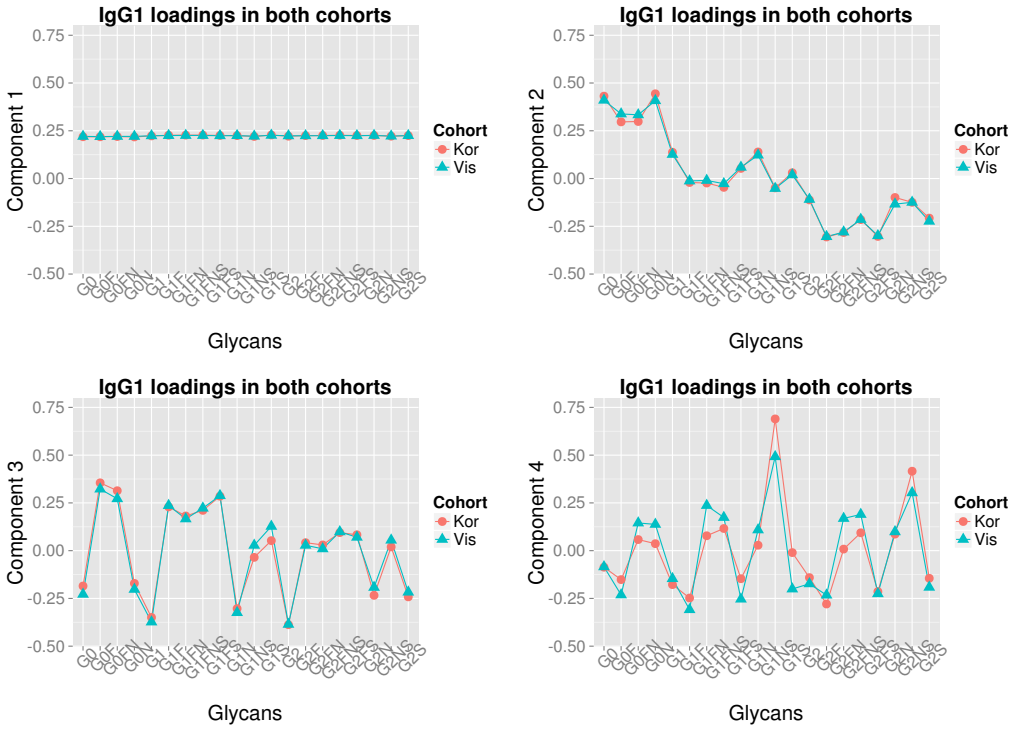


(a) The first (CROATIA_Korcula) cohort

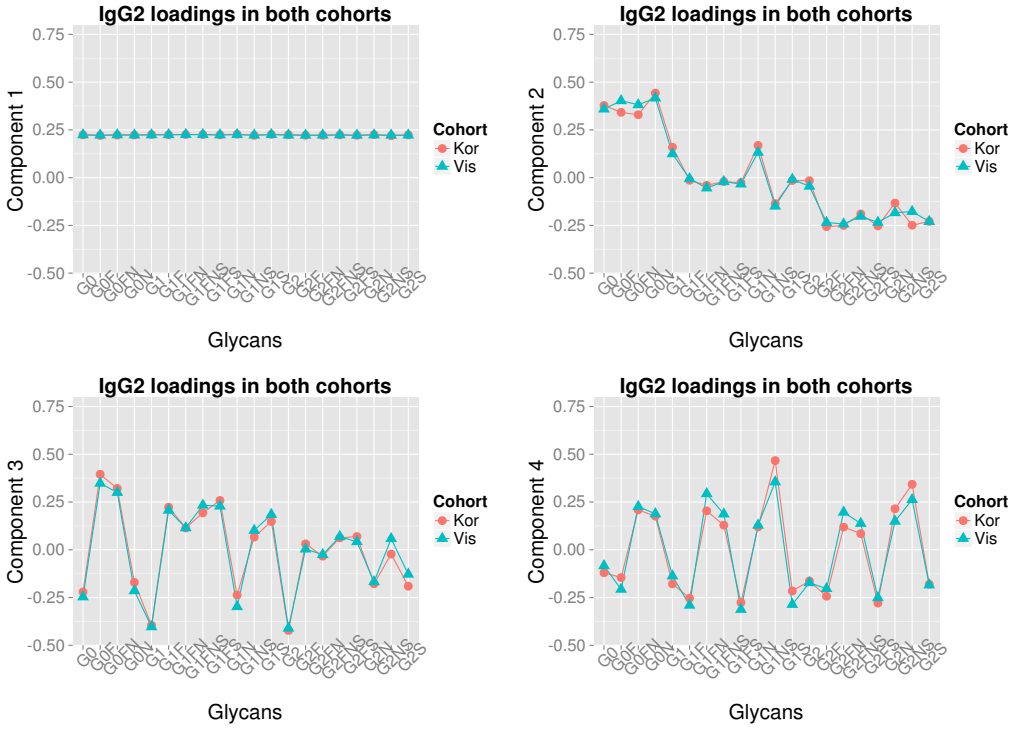


(b) The second (CROATIA_Vis) cohort

Figure 4: Heatmaps of the correlations between IgG1 and IgG2 glycans. In left the correlations of the CROATIA_Korcula cohort is shown. In right the CROATIA_Vis cohort is shown. The upper-left and lower-right block are the within subclass correlations, the off-diagonal block contains the correlations between IgG1 and IgG2 glycans. In both cohorts the glycans exhibit high positive correlations, especially between glycans within the IgG1 and IgG2 subclasses.



(a) The IgG1 glycan loadings (W) for both cohorts.



(b) The IgG2 glycan loadings (C) for both cohorts.

Figure 5: Loadings per component for both cohorts. In the top four plots loading values of IgG1 glycans (W) are plotted per glycan. The red dots connected by red lines are for the CROATIA_Korcula loadings. The four loading vectors are plotted left-to-right and top-to-bottom. The blue triangles and lines are for the CROATIA_Vis cohorts. In the bottom four plots the IgG2 glycan loading values are plotted in the same order and style.