Received: 6 July 2021 | Accepted: 17 June 2022

DOI: 10.1111/rssc.12583

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series C Applied Statistics

Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS)

Said el Bouhaddani[1] | **Hae-Won Uh**[1] | **Geurt Jongbloed**[2] | **Jeanine Houwing-Duistermaat**[1,3,4]

[1]Department of Data Science and Biostatistics, UMC Utrecht, Utrecht, The Netherlands

[2]Delft Institute of Applied Mathematics, TU Delft, Delft, The Netherlands

[3]Department of Statistics, University of Leeds, Leeds, UK

[4]Department of Statistical Sciences, University of Bologna, Bologna, Italy

Correspondence
Said el Bouhaddani, Department of Data Science and Biostatistics, UMC Utrecht, Utrecht, The Netherlands.
Email: s.elbouhaddani@umcutrecht.nl

Abstract

The availability of multi-omics data has revolutionized the life sciences by creating avenues for integrated system-level approaches. Data integration links the information across datasets to better understand the underlying biological processes. However, high dimensionality, correlations and heterogeneity pose statistical and computational challenges. We propose a general framework, probabilistic two-way partial least squares (PO2PLS), that addresses these challenges. PO2PLS models the relationship between two datasets using joint and data-specific latent variables. For maximum likelihood estimation of the parameters, we propose a novel fast EM algorithm and show that the estimator is asymptotically normally distributed. A global test for the relationship between two datasets is proposed, specifically addressing the high dimensionality, and its asymptotic distribution is derived. Notably, several existing data integration methods are special cases of PO2PLS. Via extensive simulations, we show that PO2PLS performs better than alternatives in feature selection and prediction performance. In addition, the asymptotic distribution appears to hold when the sample size is sufficiently large. We illustrate PO2PLS with two examples

from commonly used study designs: a large population cohort and a small case–control study. Besides recovering known relationships, PO2PLS also identified novel findings. The methods are implemented in our R-package *PO2PLS*.

KEYWORDS

EM algorithm, global test, heterogeneity, identifiability, latent variable models, probabilistic O2PLS

1 | INTRODUCTION

Many methods have been developed to analyse multiple omics datasets measured on the same person. Under the hypothesis that these different omics datasets represent different stages of biological processes, new insights can be obtained when these datasets are jointly analysed (Richardson et al., 2016). Examples of datasets are: genome-wide DNA markers reflecting the genetic code, transcriptomics and epigenetics, providing information on expressed and silenced genes, proteomics measuring the abundance of proteins. Methods to link the information in these datasets have to address challenges such as high dimensionality of the data, high correlation between variables within and across datasets, and the presence of heterogeneity among datasets. The heterogeneity is caused by measuring different biological levels and using different technologies to measure them. In this paper, we propose a novel probabilistic approach (PO2PLS) to address these challenges and appropriately model the relationship between two datasets x and y.

The PO2PLS model takes its underlying ideas from the algorithmic O2PLS method (Trygg & Wold, 2003). O2PLS is a latent variable approach that estimates joint and data-specific components, which are linear combinations of the variables in x and y. It identifies these components sequentially rather than simultaneously, starting with the component that has the largest variance. The drawbacks of this method are: due to the sequential approach, there is a risk of overfitting, the components are not uniquely defined, and standard errors have to be obtained via bootstrapping, which is not feasible in high-dimensional data.

The key innovation of this paper is an alternative probabilistic approach. We assume a multivariate normal distribution for x and y. By posing mild constraints, all parameters of the model are identifiable. For parameter estimation, we use maximum likelihood. With likelihood-based data integration methods, a path from x to y can be explicitly specified, making the estimation of a relationship more efficient. For maximizing the likelihood, a computationally efficient EM algorithm is derived that specifically targets high-dimensional data. A global test statistic is formulated for the null hypothesis of no relationship between x and y. Estimating the standard errors is not straightforward, since the standard regularity conditions of maximum likelihood estimators being asymptotically normally distributed may not hold for overparameterized models (such as latent variable models) (Sun et al., 2015). We will derive the asymptotic distribution of our estimators as well as the distribution of our proposed test statistic; we apply the mathematical theory that establishes asymptotic properties of estimators of overparameterized models (Shapiro, 1983). Note that the size of the resulting asymptotic covariance matrix increases quadratically with the number of x and y variables; hence we propose an approximation that is very fast to compute.

Other latent variable approaches are available. Examples of algorithmic methods that only include joint parts are partial least squares (PLS) (Wold, 1973) and canonical correlation

analysis (CCA) (Hotelling, 1936). A second algorithmic method also incorporating data-specific parts is JIVE (Lock et al., 2013). JIVE is less flexible than O2PLS as it restricts the joint components of x and y to be exactly equal (details are given in Section 2.2). We have recently shown that when this assumption does not hold, convergence problems may arise, and the performance of the estimators might be poor (el Bouhaddani, Uh, Jongbloed, et al., 2018).

Examples of other likelihood approaches are envelope regression (Cook & Zhang, 2015) and probabilistic PLS (PPLS) (el Bouhaddani, Uh, Hayward, et al., 2018). Envelope regression fully models the covariance structure and is therefore not suited for high-dimensional data. PPLS uses a simpler covariance structure with fewer parameters and is applicable to high-dimensional datasets, but does not allow for data-specific components. In contrast to PPLS, SIFA (Li & Jung, 2017) models specific components. However, just as JIVE, SIFA assumes the joint components to be exactly equal and might not perform well if this condition does not hold. SIFA and PPLS can be viewed as specific cases of the PO2PLS model.

We will illustrate the method by analysing data from two studies with a different design. The first study is a population study in which DNA markers ($p \approx 10^5$) and glycomics ($q = 20$) data are available for $N = 885$ subjects (Wahl et al., 2018). This study has been part of genome-wide association studies (GWAS), testing associations between a marker and a glycan using single pair methods. These methods do not take into account the high correlations among the measurements. We will perform a global test for an association between genetic markers and glycan abundances interlinking all variables, and delineate which genes and glycans contribute most to this association. The second study is a case–control study. It consists of epigenetics and transcriptomics data ($p, q \approx 10^4$) for 23 subjects, of which 13 suffered from hypertrophic cardiomyopathy (HCM) and 10 are healthy controls. Differential expression analyses are usually performed to test the relationship between each pair of measurements. Instead, we globally test for an association between epigenetic activity and gene transcription.

The contributions of this paper are fourfold. First, we propose a novel model, PO2PLS, for the relationship between two omics datasets. Second, a computationally efficient EM algorithm is derived to estimate the parameters using maximum likelihood. Third, we formulate a global test for the null hypothesis of no relationship between x and y. Fourth, the added value of our methods is demonstrated by applying them to omics datasets from two different studies. Software code implementing the estimation procedure and global test is freely available as an R package on GitHub (https://github.com/selbouhaddani/PO2PLS) and will soon be released on CRAN.

The remainder of the article is organized as follows. In Section 2, the PO2PLS model is formulated, and identifiability of the parameters is proven. Furthermore, maximum likelihood estimates are derived, and a global test of the relationship between x and y is proposed. In Section 3, the performance of PO2PLS is studied over a range of simulation scenarios. We focus on feature selection, prediction performance, type I error and power of the statistical test. In Section 4, PO2PLS is applied to the case studies to test and describe the relation between two sets of omics variables. We conclude with a discussion.

2 | A PROBABILISTIC DATA INTEGRATION FRAMEWORK

2.1 | The PO2PLS model

Let x and y be two random row vectors of size p and q respectively. 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, with r usually a small number. 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, leading to differences between t and u. More precisely, the PO2PLS model for x and y is described by

$$x = tW^{\mathrm{T}} + t_\perp W_\perp^{\mathrm{T}} + e, \quad y = uC^{\mathrm{T}} + u_\perp C_\perp^{\mathrm{T}} + f, \quad u = tB + h. \tag{1}$$

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 parameter B is a diagonal $r \times r$ matrix.

The random vectors e and f are independent multivariate normally distributed, 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 normally distributed variables, with diagonal covariance matrices Σ_t, $\Sigma_{t_\perp}$, $\Sigma_{u_\perp}$ and Σ_h respectively. The covariance matrix of u follows from (1): $\Sigma_u = B^{\mathrm{T}}\Sigma_t B + \Sigma_h$.

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 parameterizes the distribution of $(x, y) \sim \mathcal{N}(0, \Sigma_\theta)$ (the explicit expression for Σ_θ is given in the supplementary materials).

Note that the model for the relationship between u and t (the inner relation) is taken asymmetrically, as often a certain hierarchy is assumed for x and y (Crick, 1970). For instance, it is reasonable to assume that genetic variability induces glycomic variation and not the other way around, so a model for u in terms of t better reflects the underlying biology.

2.2 | PO2PLS as a general data integration framework

PO2PLS models the relationship between x and y through t and u as described in (1). It can be seen as a generalization of other models. The first model is SIFA, which assumes that the joint principal components (JPCs) are exactly equal, that is, $u = t$. In this case, $B = I$ and $\Sigma_h = 0$, so u and t have the same scale and the identity as correlation matrix. For heterogeneous datasets, the two sets of JPCs represent different mechanisms (e.g. genetic vs. glycomic pathways). Therefore, they may not be perfectly correlated or on the same scale. Also, assuming homogeneity of datasets can negatively affect estimation performance (el Bouhaddani, Uh, Jongbloed, et al., 2018). The second model JIVE, assumes $u = t$ and orthogonality between columns in the concatenated loading matrices $(W, W_\perp)$ and $(C, C_\perp)$. In this case, combinations of variables involved in the joint and specific parts have to be orthogonal, which is a strong restriction. The third model is the probabilistic PLS model which is obtained by setting $\Sigma_{t_\perp}$ and $\Sigma_{u_\perp}$ to zero in (1).

In the envelope regression (ER) model, the number of noise variance parameters to estimate is of order $O(p + q)$, whereas PPLS and PO2PLS introduce one σ_e^2 and σ_f^2 for x and y respectively. When the covariance matrix of x or y is singular, the ER estimator is not well-defined.

The estimation approach we adopt in PO2PLS is maximum likelihood. Computationally, we propose an EM algorithm (details about estimation is found below in Section 2.4). Other probabilistic approaches, such as ER, directly optimize the likelihood over Grassmann manifolds (Cook & Zhang, 2015). However, calculation of the covariance of (x, y) is not feasible in high dimensions. Alternatively, joint and specific components can be sequentially estimated. For example, the O2PLS estimator (see el Bouhaddani et al., 2016; Trygg & Wold, 2003) is as follows: first, the covariance between xW and yC is optimized, then the covariance between xW and $x - xWW^{\mathrm{T}}$ is optimized to get estimates for the specific parts, and finally, after subtracting these parts, the covariance between x^*W and y^*C is optimized with the star indicating a deflation step. Our EM

TABLE 1 Features of several data integration methods. An 'X' indicates the presence of a feature. All methods estimate a joint part. The first row indicates methods that also estimate specific components $W_\perp$ and $C_\perp$. The second row indicates methods that are based on a probability distribution for x and y. For the next row, an X is placed if the method does not restrict the model to $u = t$. The last row indicates methods that can cope with data where p, q are larger than the sample size

Properties	PLS	PPLS	ER	O2PLS	JIVE	SIFA	PO2PLS
Specific			X	X	X	X	**X**
Probabilistic		X	X			X	**X**
$u = tB + h$	X	X	X	X			**X**
High-dimensional	X	X		X	X	X	**X**

implementation of PO2PLS appears to be competitive with fast algorithmic approaches in terms of memory usage and is reasonably fast in high-dimensional settings (see Section 3). In Table 1, an overview is shown with several methods and their features.

2.3 | Identifiability of the PO2PLS model

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 $x = (tR)(WR)^{\mathrm{T}}$ yield the same x while W and WR are not the same. Note that if $C_\perp v(t)$ is diagonal with distinct elements, $C_\perp v(tR)$ is not diagonal unless R is also diagonal. In PCA, the loading matrices are restricted to be semi-orthogonal, that is, $W^{\mathrm{T}}W = I$, whereas in factor analysis, the latent variables are standard normally distributed. In PO2PLS, identifiability can be obtained using similar assumptions, namely semi-orthogonal loading matrices and diagonal covariance matrices for the latent variables.

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 positive. 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 $(\sigma^2_{t_k} b_k)_{k=1}^r$ 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 and the other parameters in θ are uniquely identified. The following theorem makes this precise.

Theorem 1. *Let r, r_x, r_y and θ satisfy the above assumptions. Let Σ_{θ_1} and Σ_{θ_2} be covariance matrices corresponding to PO2PLS parameters θ_1 and θ_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 Δ_W, $\Delta_{W_\perp}$ and $\Delta_{C_\perp}$, and all other parameters in θ_1 and θ_2 are equal.*

Proof. The part of the proof pertaining to W and C follows from the spectral theorem and its uniqueness property. Indeed, consider the off-diagonal block of Σ_θ, given by $W\Sigma_t BC^{\mathrm{T}}$. This can be recognized as a spectral decomposition with unique eigenvalues and eigenvectors up to a sign. The part of the proof pertaining to $W_\perp$ and $C_\perp$ uses the linear independence between W and $W_\perp$: one can find vectors orthogonal to W but not to $W_\perp$ to eliminate W from the equation $\Sigma_{\theta_1} = \Sigma_{\theta_2}$. Note that the common assumption $W^{\mathrm{T}}W_\perp = 0$ is not necessary here. A complete proof is given in the supplementary materials.

2.4 | Maximum likelihood estimation of the parameters

To estimate θ, the following log-likelihood function under the PO2PLS model (1) is maximized

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

Note that L is a complicated and highly non-linear function of θ, and its computation requires computing and storing covariance matrices of size $(p+q)^2$. We implement a memory-efficient and analytically tractable EM algorithm (Dempster et al., 1977) to obtain maximum likelihood estimates for θ. Contrary to the sequential O2PLS algorithm, the estimation is simultaneous over both joint and specific parts.

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

$$Q(\theta|x,y,\theta') := \mathbb{E}_{\theta'}\left[\log f(x,y,t,u,t_\perp,u_\perp|\theta)|x,y\right]. \tag{3}$$

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

$$f(x,y,t,u,t_\perp,u_\perp|\theta) = f(x|t,t_\perp)f(y|u,u_\perp)f(u|t)f(t)f(t_\perp)f(u_\perp). \tag{4}$$

These factors depend on distinct sets of parameters. For example $f(x|t,t_\perp)$ depends only on W, $W_\perp$ and σ_e^2, yielding separate optimization problems.

The expectation step involves a conditional expectation of the complete data likelihood. Since f in (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 (4), the conditional expectation of $\log f(x|t,t_\perp)$ is given by

$$-\frac{1}{2}\{Np\log(2\pi) + Np\log\sigma_e^2 + \sigma_e^{-2}\mathrm{tr}\,\mathbb{E}_{\theta'}[\|x - tW^{\mathrm{T}} - t_\perp W_\perp^{\mathrm{T}}\|_F^2|x,y]\}. \tag{5}$$

This expectation involves first and second conditional moments of the vector $(t, t_\perp)$ given θ', x and y. We give explicit expressions for these terms in the supplementary materials.

In the maximization step, the function in (5) is optimized over all semi-orthogonal matrices W and $W_\perp$. By introducing Lagrange multipliers Λ_W and $\Lambda_{W_\perp}$, maximizing (5) over semi-orthogonal W and $W_\perp$ is then equivalent to minimizing the following objective function

$$\mathbb{E}_{\theta'}[\|x - tW^{\mathrm{T}} - t_\perp W_\perp^{\mathrm{T}}\|_F^2\,|x,y] + \Lambda_W(W^{\mathrm{T}}W - I_r) + \Lambda_{W_\perp}(W_\perp^{\mathrm{T}}W_\perp - I_{r_x}). \tag{6}$$

Note that the objective function involves both W and $W_\perp$ and cannot be decoupled. Instead of numerical optimization, we sequentially optimize over W and then $W_\perp$, which leads to an algorithm that converges to the maximum likelihood estimates. Under standard conditions, this algorithm monotonically approaches a (local) maximum of the observed likelihood L (Meng & Rubin, 1993).

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 (Geisser, 1993) and eigenvalue (scree) plots (Mardia et al., 1979).

The expectation and maximization step for the other parts in (4) are calculated analogously (see the supplementary material). In this calculation, the orthogonalization operator is used to obtain semi-orthogonal loading matrices, defined as follows.

Definition 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 $orth(A) = A(R^{\mathrm{T}})^{-1}$.*

Using this operator, the EM parameter updates are made explicit in Theorem 4 in the Appendix.

2.5 | Statistical inference: Formulation of a global test

One of the statistical challenges in data integration is to assess the statistical evidence for the relationship between x and y. In our model, this relationship is represented by the equation $u = tB + h$ in (1). In general, the matrix B is an $r \times r$ diagonal matrix with elements B_k, representing the relation between joint components t_k and u_k. Since these components are independent, we propose the null hypothesis of no relationship for each B_k,

$$H_0 : B_k = 0 \quad \text{against} \quad H_1 : B_k \neq 0. \tag{7}$$

To test this null hypothesis, we propose the following test statistic,

$$T_{B_k} = \hat{B}_k / \hat{SE}_{\hat{B}_k}. \tag{8}$$

We refer to (7) with (8) as the global test and denote any of the T_{B_k} by simply T_B. Global here means across all variables simultaneously, contrary to single variable association studies. To apply the global test statistic in practice, the asymptotic distribution of all parameters θ, including B, needs to be derived. Since our model is overparameterized, for example the matrix W contains r redundant parameters, standard maximum likelihood theory does not hold.

Consistency of the estimator $\hat{\theta}_N$ and its asymptotic distribution can still be derived, with N the number of independent and identically distributed samples drawn from (x, y) under the PO2PLS model. It is known that the sample covariance matrix $S_N = 1/N \sum_{i=1}^{N} (x_i, y_i)^{\mathrm{T}}(x_i, y_i)$ is consistent and asymptotically distributed. Moreover, S_N converges to the PO2PLS covariance matrix Σ_θ almost surely as the sample size goes to infinity. These results can be used to show that $\hat{\theta}_N$ is consistent and asymptotically normally distributed, as shown in the following theorem.

Theorem 3. *Let S_N be the sample covariance matrix of $(x_i, y_i)_{i=1}^{N}$. Let g be the mapping from any PO2PLS parameter set θ' to $\Sigma_{\theta'}$, and $F(\cdot,\cdot)$ be a discrepancy function from any $(S_N, \Sigma_{\theta'})$ to the set of positive real numbers. Under certain regularity conditions on g and F, we get that $\Sigma_{\hat{\theta}_N}$ is consistent and asymptotically normally distributed, that is,*

$$N^{1/2}(\Sigma_{\hat{\theta}_N} - \Sigma_\theta) \to \mathcal{N}(0, \Xi). \tag{9}$$

Moreover, the same holds for $\hat{\theta}_N$,

$$N^{1/2}(\hat{\theta}_N - \theta) \to \mathcal{N}(0, \Pi_\theta). \tag{10}$$

Proof. After choosing a suitable F, the proof of the first part in this theorem relies on showing that in the limit, $\Sigma_{\hat{\theta}_N}$ is 'close enough' to S_N. As a result, $\Sigma_{\hat{\theta}_N}$ is also consistent and asymptotically normally distributed. The second part can be proven by showing that the mapping g from θ to Σ_θ is analytic, implying that $\hat{\theta}_N$ is consistent and asymptotically normally distributed. Further details are given in the supplementary materials.

Given the asymptotic covariance matrix of $\hat{\theta}$, Π_θ, standard errors are obtained by calculating the square root of the diagonal elements of Π_θ. An estimate of Π_θ is obtained from the inverse observed Fisher information matrix given by Louis (1982),

$$\mathcal{I}_{\hat{\theta}} = \mathbb{E}[B(\hat{\theta})|X, Y] - \mathbb{E}\left[S(\hat{\theta})S(\hat{\theta})^{\mathrm{T}}|X, Y\right]. \tag{11}$$

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 derivation of the Fisher information matrix for the parameters of the PO2PLS model is given in the supplementary materials.

To obtain standard errors for $\hat{B}$, the submatrix of $\mathcal{I}_{\hat{\theta}}^{-1}$ with respect to B has to be calculated. However, this requires inverting a matrix of size $O((pqr)^2)$, which is computationally infeasible even for moderate p and q. Under the assumptions that $\hat{B}$ and $\hat{\theta}/\hat{B}$ are asymptotically independent and $\hat{\Sigma_h}$ is non-random, the observed Fisher information matrix $\mathcal{I}_{\hat{B}}$ and thus $SE_{\hat{B}}$ are given by

$$\mathcal{I}_{\hat{B}} = \hat{\Sigma}_h^{-1}\mathbb{E}[t^{\mathrm{T}}t|x, y] - \hat{\Sigma}_h^{-2}\mathbb{E}[(u - t\hat{B})^{\mathrm{T}}tt^{\mathrm{T}}(u - t\hat{B})|x, y]. \tag{12}$$

Details of the derivation of this formula are given in the supplementary materials. Note that the first part on the right-hand side is the Fisher information matrix based on the general linear model, had t and u been observed. Standard errors for $\hat{B}$ are given by the square root of the diagonal elements of $\mathcal{I}_{\hat{B}}^{-1}$. Thus, to test the global hypothesis (7), we apply our statistic T_B and calculate the corresponding p-values.

3 | SIMULATION STUDY

We conduct a simulation study to evaluate the performance of PO2PLS in terms of feature selection, prediction and performance of our global test. Four metrics are considered: true positive rates, root mean squared error of the prediction, type I error and power. We compare PO2PLS to existing approaches PLS, O2PLS, PPLS and SIFA, covering algorithmic and probabilistic methods with and without specific parts (see Table 1). We investigate robustness against model assumptions. Finally, we assess computational efficiency.

For performance in feature selection and prediction ability, we consider combinations of small and large sample sizes ($N = 100, 1000$) and low- and high-dimensional data ($p = 2000, 10{,}000$; $q = 25, 125$). We also include two proportions of noise relative to the total variation: in the 'small noise proportion', we set the variance of e and f to be 40% of the variance of x and y. In the 'large noise proportion', these values are 95% and 5% for x and y respectively. We set $B = I$ and $\Sigma_h = 0$ to comply with the SIFA assumptions, see Section 2.2. The impact of heterogeneity of joint parts is considered by increasing the joint residual variance Σ_h from 0% to 80% of the total joint variance Σ_u. Finally, we set r, r_x and r_y to five components. These scenarios are commonly encountered in data analysis.

To assess the feature selection performance, we calculate the proportion of true top 25% features among the estimated top 25% (i.e. true positive rate, TPR). We then average these

proportions across components to obtain an aggregated measure. Predictive performance is measured by calculating the RMSEP, defined as the square root of $\mathbb{E}\|y - \hat{y}\|^2$ with $\hat{y}$ predicted from x. 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.

To evaluate the performance of the PO2PLS global test T_B described in Section 2.5, we first estimate the type I error for increasing sample size of 50, 500, 5000 and 10,000. The dimension of x is set to 20. The number of simulation replicates here is 50,000. Next, we consider increasing dimensionality, namely $p = 20, 200, 2000$. The sample size is set to 500, and we replicate 2000 times. The type I error is calculated as the proportion of rejecting the null hypothesis $B = 0$ at a 5% level when simulating under this hypothesis. Next, we estimate the power of T_B in (8). We compare four procedures, namely using the normal distribution for T_B with calculated standard errors using our approximation of its covariance matrix, with standard errors obtained from parametric and from non-parametric bootstrapping, and with using the empirical distribution of T_B via permutations. The proportion of false and true rejections are reported and compared for increasing B. We consider a sample size of 50 and 500, and dimensionality p of 20 and 200. The number of bootstrap and permutation iterations is 250 and 500, respectively, and we repeat 500 times. In all three simulations, the dimension of y is kept to 5, the noise proportion is 50%, and we set $r = 2$, $r_x = 1$ and $r_y = 0$.

Three additional simulation studies are carried out to study the robustness of PO2PLS against model deviations. PO2PLS is applied to high-dimensional simulated datasets from a selected case–control study design, mimicking the second data analysis in Section 4. We compare the error of predicting the outcome using the PO2PLS joint components with aforementioned alternatives. Then, we assess the impact of rank misspecification when fitting PO2PLS, by estimating too few components, and the impact of non-normality of the latent variables, using four commonly encountered distributions. Finally, we study the computational efficiency of the PO2PLS implementation, measured by the CPU time and memory demand of the EM algorithm. Details of these simulations and results are given in the supplementary materials.

3.1 | Simulation results

We first present the accuracy and prediction performance in the low-dimensional setting, see Figure 1. Boxplots of the accuracy and prediction error are shown across the scenarios. Differences in accuracy with respect to PO2PLS are also shown. In terms of feature selection, PO2PLS performed good compared to the other methods. When considering the TPR difference between each method and PO2PLS per simulation run, PO2PLS generally had the highest TPR. This difference tends to increase with larger noise proportions and more heterogeneous joint parts settings. The differences between PO2PLS and PPLS are not shown for better visual comparison. Regarding the prediction error, PO2PLS generally performed better than the other methods. SIFA had the highest prediction error when heterogeneity between the joint parts was present. Furthermore, PLS and O2PLS seemed to overfit in noisy, small sample size scenarios: the training error was lower than the test error compared to the other methods. In the high-dimensional settings, similar results were obtained. Details can be found in the supplementary materials. For the high-dimensional settings, the implementation of SIFA gave 'out-of-memory' errors. Hence, we could not include SIFA in these comparisons.

Results for the global inference are shown in Figure 2. The type I error of the PO2PLS test was around 5% for increasing sample size and dimensionality. Based on the proportion of rejections

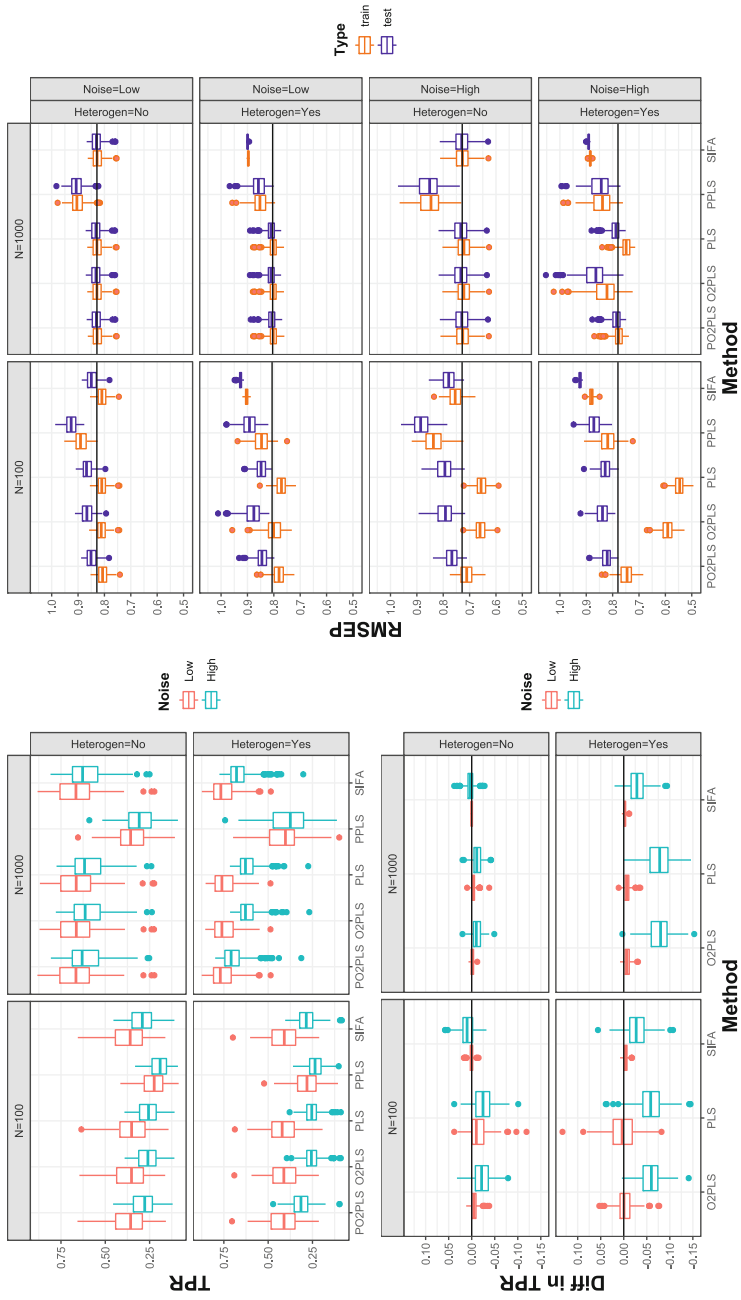

FIGURE 1 Simulation study: feature selection and prediction error. *Upper left figure:* proportion of true top 25% among estimated top 25% (TPR) for each method, stratified by simulation scenario. The left and right boxplots represent low and high noise, respectively. *Lower left figure:* Difference in TPR of several methods and PO2PLS. Lower values are in favour of PO2PLS. *Right figure:* Root mean squared error of prediction stratified by method and scenario. The left and right boxplots represent the training and test error respectively. The black line represents the median test error when using true parameter values. [Colour figure can be viewed at wileyonlinelibrary.com]

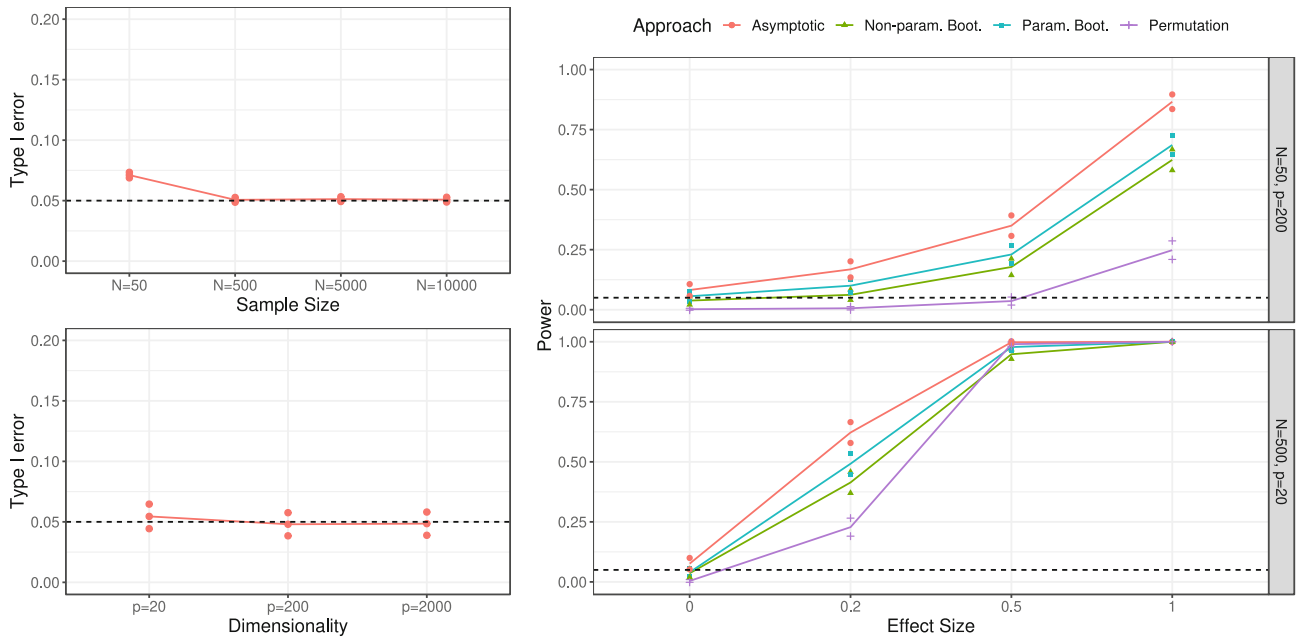

FIGURE 2 Simulation study: global inference. *Left panel*: Type I error of the global test based on asymptotic PO2PLS statistic, for increasing sample size and $p = 20$ (upper plot), and increasing dimensionality and $N = 500$ (lower plot). These are based on 50,000 and 2000 replicates respectively. For all plots, 95% confidence intervals are added based on the binomial distribution, indicated by the points above and below the lines. The dashed horizontal lines represent the 5% rejection level. *Right panel*: Rejection proportions of the global test performed by the four approaches (asymptotic, non-parametric, parametric bootstrapping and permutations), for increasing effect size. The sample size was $N = 50$ with $p = 200$ (upper plot) and $N = 500$ with $p = 20$ (lower plot). These are based on 500 replicates. [Colour figure can be viewed at wileyonlinelibrary.com]

under the null hypothesis (7), the PO2PLS test had type I error around 5% for all but the smallest sample size; in that case, the type I error was about 7%. It also had more power under the alternative than the other approaches, with the permutation test being severely underpowered in small sample size.

We briefly present the key results of the additional simulations. In the selected case–control simulation study, PO2PLS had highest TPR, and suffered less from overfitting than PLS and O2PLS. When estimating one component less than the true number of joint and specific components, PO2PLS performed similarly to the algorithmic methods (PLS and O2PLS). Furthermore, PO2PLS was robust against non-normal distributions. Finally, the increase in CPU time and memory usage for increasing data dimensionality was similar across the methods. The full findings are given in the supplementary materials.

4 | APPLICATIONS TO OMICS DATASETS

We illustrate the PO2PLS model with datasets from two different studies. First, PO2PLS is applied to test and estimate genetic contributions to glycomic variation in a population-based cohort. Second, PO2PLS is used to infer a relation between DNA regulation and gene expression using data from a case–control study. Here we also investigate whether the joint components reveal the case–control status and whether the top features overlap with findings in cardiovascular diseases. For comparison, we also applied O2PLS to these datasets.

4.1 | Data integration in a population cohort

Glycosylation is one of the most common post-translational modifications that enrich the functionality of proteins in many biological processes, such as cell signalling, immune response and apoptosis (Wahl et al., 2018). Previously, GWAS were performed between pairs of single nucleotide polymorphisms (SNPs) and glycans to investigate genetic regulation of glycosylation (Lauc et al., 2010; Wahl et al., 2018). However, glycans abundances are highly correlated and associated with multiple genes. For example, the glycan G0 was found associated with multiple genes, including *FUT8*, and this gene was itself associated with multiple glycans (Klarić et al., 2020). Therefore, a multivariate approach might provide new insights. We first confirm that genetics play a significant role in regulating of glycans. Then, we investigate whether the joint glycan components represent biological structures. Finally, we compare our top genes with genes identified in GWAS.

Genetic and glycomic data were measured, yielding 333,858 genotyped SNPs and 20 IgG1 glycan abundances for $N = 885$ participants in the Croatian Korcula cohort (Lauc et al., 2010). The SNPs were aggregated on the gene level by combining SNPs around the same gene with PCA, yielding a genetic PCs (GPCs) dataset. Then, the GPCs and glycomics datasets were pre-processed, resulting in datasets X ($p = 37{,}819$) and Y ($q = 20$) respectively. Based on scree plots of the eigenvalues of X^TX, X^TY and Y^TY, five joint, five genetic-specific and no glycan-specific components were retained.

A global test for the association between genetics and glycans was performed using PO2PLS. The T_B statistic for each component was between four (for the first component) and three (for the last component). With corresponding p-values of 10^{-5} and 10^{-3}, there is statistical evidence of a relationship between genetics and glycans.

The loading values of each glycan variable for the five joint components are depicted in Figure 3. Each joint glycan component appears to represent different aspects of glycans and their molecular structure. While the first component represents the 'average' glycan (first component), the second component represents presence of fucose, the third component represents the presence of galactose, and the last two components represent GlcNAc (el Bouhaddani, Uh, Jongbloed, et al., 2018). The top gene in the second joint genetic component is *FUT8* which has been linked to fucosylation (Lauc et al., 2010). Note that the second glycan component reflects 'presence of fucose'. The same article reports more genes linked to glycosylation that we did not find, but their GWAS results are based on imputed genetic data from multiple cohorts. With our joint approach, several other top genes were found, for example, *DNAJC10* and *AKAP9*, that have links to synthesis and degradation of glycoproteins or (more generally) with inflammation and immune responses.

A second independent study of 714 participants from the Croatian Vis cohort is available. To replicate our findings in the Korcula cohort, we apply PO2PLS to this cohort and compare the components underlying the genetics and glycomics data. The results from the second study, shown in the supplementary material, are consistent with the above findings, indicating that the obtained components are not specific to one study. Finally, we compared the prediction error of Y given X of the models estimated with PO2PLS and O2PLS in Korcula, evaluated using the data from Vis. The ratio of training (Korcula) and test (Vis) error appeared to be 5/23 for O2PLS and 20/21 for PO2PLS. This is conform the simulation study that O2PLS is prone to overfitting.

4.2 | Data integration in a case–control study

HCM is a rare heart muscle disease negatively affecting blood circulation and leading to heart failure. Several studies have shown that several molecular factors, such as epigenetics and gene transcription, play an important role in HCM (Hemerich et al., 2019). We investigate whether epigenetic variation affects transcription and test this relationship using PO2PLS. Since the samples consist of HCM cases and controls, an obvious question is whether one of the joint components represents this segregation of cases and controls.

Data on epigenetics (DNA regulation) and transcriptomics (gene expression) are available, obtained from the heart tissue of 13 HCM patients and 10 controls. Epigenetic data were measured using ChIP-seq, yielding regulation levels of 33,642 regions after pre-processing. Transcriptomics data were measured using RNA-seq, yielding 15,882 expression levels after pre-processing (TMM normalization, followed by log transformation). Statistical challenges are the small sample size of 23 and the large number of features (around 45,000).

PO2PLS is applied to the epigenetics (X) and transcriptomics (Y) data, using two joint components and one specific component for both datasets. These numbers are determined using scree plots. The T_B test statistic for the first component was 9.12, and 2.35 for the second component. The p-values were smaller than 0.001 for the first component and 0.018 for the second, so the two component were statistically significant.

To investigate whether the top genes in the joint components are involved in cardiovascular outcomes, we clustered the 500 genes with highest loading values in the first joint PC using DisGeNET (a database of gene–disease associations (Sabater-Molina et al., 2018)). The top 10 most significant clusters appear to represent a broad spectrum of cardiovascular diseases (Table 2).

In Figure 4, PO2PLS scores are plotted for the first two joint components, and each dot is coloured according to its case–control status. The plots indicate that the first joint component

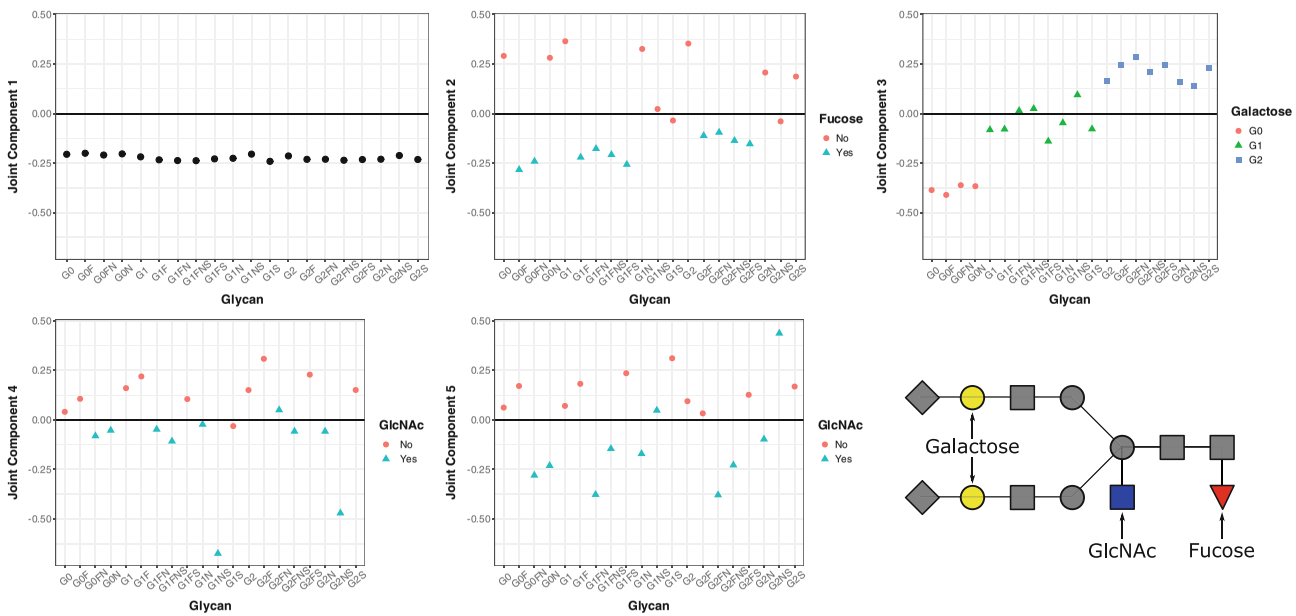

FIGURE 3 Glycomic PO2PLS joint components. The five JPCs are plotted one by one, from left to right, from top to bottom. The dots represent the loading values of each glycan, indicating their importance in the genetic–glycomic relationship. The colours and shapes represent the biological grouping of the glycans. In the last row and column, a graphical representation of the structure of a particular glycan is shown. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Annotation of genes in the first transcriptomics joint PC in the HCM analysis. Using PO2PLS, the top 500 genes were clustered using DisGeNET (a database of gene–disease associations). These top 500 genes are primary drivers of the association with epigenetics across HCM cases and controls. Here, *p*-values are calculated with a Fisher exact test and corrected for multiple testing. The 10 most significant clusters are shown

Clusters	Disease name	*p*-value (FDR B&H)
Disease cluster 1	Hypertensive disease	2.53e−7
Disease cluster 2	Arteriosclerosis	2.57e−6
Disease cluster 3	Atherosclerosis	2.57e−6
Disease cluster 4	Coronary heart disease	7.42e−6
Disease cluster 5	Arthritis	1.19e−5
Disease cluster 6	Aortic valve stenosis	2.16e−5
Disease cluster 7	Coronary artery disease	2.29e−5
Disease cluster 8	Cardiovascular diseases	2.29e−5
Disease cluster 9	Gestational diabetes	2.67e−5
Disease cluster 10	Heart failure	5.27e−5

picked up the case–control segregation. Additionally, the O2PLS scores are plotted, showing a similar pattern as PO2PLS.

5 | DISCUSSION

We propose probabilistic two-way orthogonal partial least squares (PO2PLS) to model the relationship between two sets of variables x and y in the presence of data-specific characteristics. Our method is suited for heterogeneous, high-dimensional, correlated datasets commonly available in the life sciences. For estimation, we proposed a memory-efficient EM algorithm. For testing, we derived a global test statistic and its approximate distribution under the null hypothesis of no relationship between x and y.

Via an extensive simulation study, we showed that PO2PLS often performed better than PPLS, SIFA, O2PLS and PLS. In terms of feature selection and prediction, it performed better than PLS, PPLS and SIFA when heterogeneity exists between the datasets. These results were expected since, contrary to the other methods, PO2PLS incorporates the heterogeneity in the model and, therefore, better estimates the joint components. PO2PLS performed better than O2PLS and PLS in terms of prediction when the datasets are small. For noisy and small datasets, PO2PLS also had a better true positive rate than O2PLS and PLS. PO2PLS had a smaller risk of overfitting, probably because it models all the available information in the data. This reduction in overfitting was also confirmed in studying the relationship between genetic data and glycans, for which we had a replication cohort. The common belief is that PLS and O2PLS, as distribution-free methods, are more suited for small sample size scenarios than probabilistic methods (Wold, 1985). Contrary to this belief, in these scenarios, PO2PLS yielded a better true positive rate and prediction performance. Simulations showed that PO2PLS is robust against model deviations such as using too small a number of components and non-normality of the data. Note that the prediction performance of all methods can be improved by cross-validation, although less effective in small sample size scenarios and computationally intensive for the likelihood-based methods.

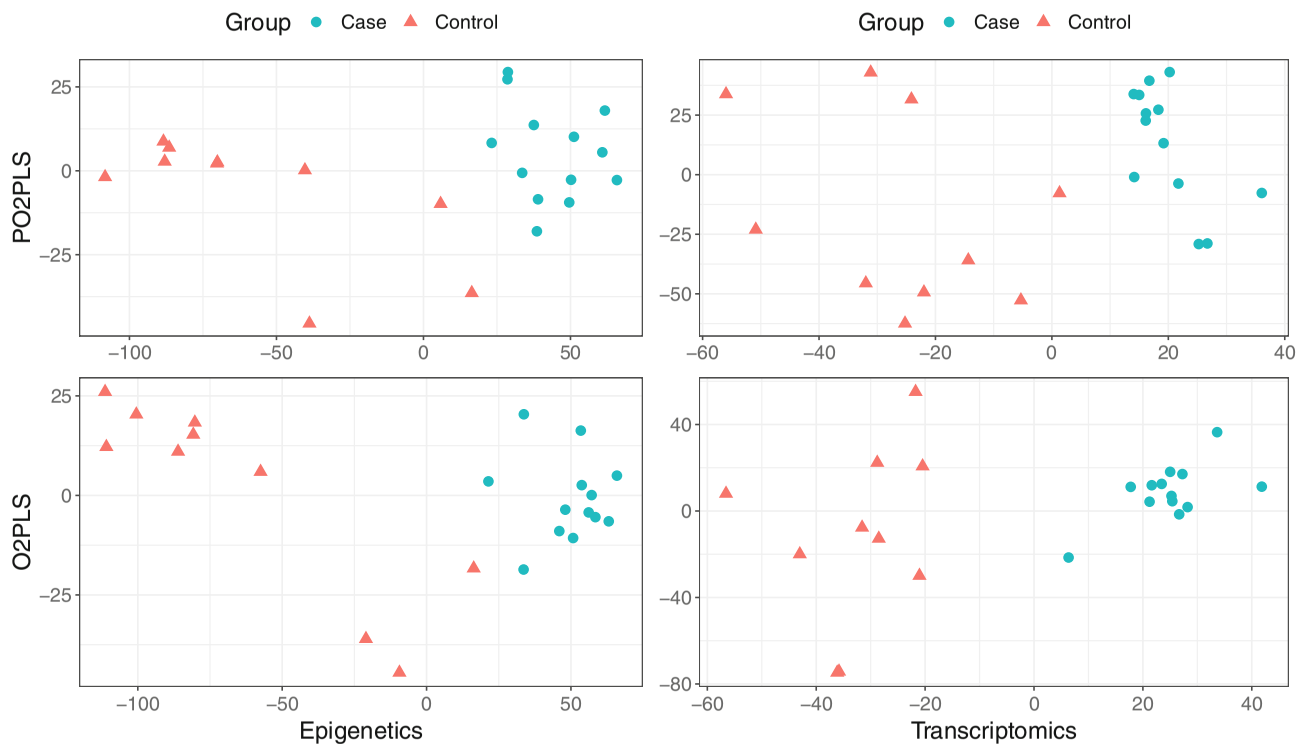

FIGURE 4 Joint principal component scores across HCM cases and controls. The two joint component scores are plotted against each other, where the first JPC is on the *x*-axis. The *upper plots* show transcriptomics resp. epigenetics scores from the PO2PLS fit. The *lower plots* show O2PLS scores. Each dot represents either an HCM patient (blue circle) or a control (red triangle). [Colour figure can be viewed at wileyonlinelibrary.com]

We also showed with simulations that our proposed global test statistic for testing the null hypothesis of no relationship is asymptotically normally distributed and performs well in terms of type I error and power. In algorithmic latent variable approaches, testing for a relationship is carried out by empirically estimating the distribution of the test statistic. Since this is time-consuming, evidence for the relevance of the top features (e.g. genes, proteins, glycans) is instead obtained by relating the findings to historical ones (Domingo-Fernández et al., 2019). For example, it is tested whether specific molecular pathways or interaction networks are over-represented in the top feature ranking. Such an approach has the advantage that prior domain knowledge is incorporated. A drawback is a focus on existing findings and a bias against novel discoveries. Moreover, it is often unclear how much evidence exists for pathways and networks in these databases. Each database uses its own scoring mechanisms, often not based on a formal scoring method. Furthermore, there might be a lack of information in the context for new diseases or measurement techniques, and using the information on related diseases or datasets may result in incorrect conclusions about relations (Mubeen et al., 2019). Our proposed testing procedure quantifies the evidence in a data-driven way; hence it may generate new hypotheses.

PO2PLS was applied to omics data from two case studies. The first one is a typical epidemiological population cohort, designed to identify new molecular drivers and build omics predictors for common diseases. These studies are also well suited to study relationships between multiple omics datasets. We applied PO2PLS to genetic and glycomics datasets. The relationship between genetics and glycomics was statistically significant, which confirmed the known high heritability of glycans and the multiple hits of GWAS (Zaytseva et al., 2020). We did not replicate all GWAS findings since we restricted ourselves to genotyped SNPs in a gene's neighbourhood. On the other hand, modelling the joint distribution of glycans and genes also led to new findings. We replicated the estimated components with relevant features in a second cohort study. The second case study was a small case–control study. To identify molecular markers for rare diseases, omics datasets are measured in cases and controls. Typically, these datasets have a small sample size, either because of the limited number of available cases (rare disease) or costs. We applied PO2PLS to epigenetic and transcriptomic data in HCM cases and controls. The relationship between the two sets was statistically significant. Clustering of the top genes using DisGeNET showed that the top genes are in gene clusters associated with several cardiovascular diseases. Moreover, when plotting the first two joint components against each other, a structure representing case–control status was evident. This might be expected since all analyses are conditional on the outcome status, and the outcome is a collider for features of the datasets that affect the outcome variable (Balliu et al., 2015; Tissier et al., 2017).

There are many future directions along the lines of the current work. First, it is of interest to include the outcome variable in the model. In omics research, several penalized regression models have been proposed to identify sets of variables related across the different datasets x or y which predict an outcome variable z (Vinga, 2020). These approaches do not model the within and across correlations and are hard to interpret when correlations between x and y are present (Tissier, 2018). Extending the probabilistic O2PLS framework to model the joint distribution of (x, y, z), addresses these challenges. This joint distribution can be specified conditional on latent joint and specific variables. In such a framework, the relation between x and y is modelled, and their association with the outcome z is simultaneously incorporated and estimated. Second, it might be of interest to model more than two omics datasets. Here, the interest may lie in inferring relations between three (or more sets) of variables, say x_1, x_2 and x_3 with joint latent variables t_1, t_2 and t_3 respectively. Derivations of estimation and inference for PO2PLS can be extended analogously. A complication is that the direction of the relationship between the sets of variables needs to

be specified via inner relationship equations, which might be unknown. The majority of current data integration approaches for more than two datasets avoid this issue by specifying the same set of latent variables t for all sets of variables (Meng et al., 2016), similar to SIFA (Section 2.2). Such an approach does not perform well for heterogeneous datasets, as shown in our simulations (Section 3). Another approach proposes optimizing a sum of objective functions for each pair of datasets (Löfstedt & Trygg, 2011) while accounting for heterogeneity in the joint parts. Third, to improve feature selection, a penalty term can be added to the likelihood function to incorporate prior belief about which variables are more important or belong together. For the algorithmic O2PLS, such an extension was recently proposed. For the probabilistic PO2PLS approach, prior belief can be included by adding a penalty function to the likelihood.

ACKNOWLEDGEMENTS

This work was supported by EU Horizon 2020 under Grant 721815 (IMforFUTURE); ERA-Net E-rare-3 JTC 2018 (MSA-Omics), EU IMI under Grant 116074 (BigData@Heart) and EU FP7-Health under Grant 305280 (MIMOmics). The authors acknowledge M. Harakalova, and M. Mokry, UMC Utrecht Dept. of Cardiology, for providing data from the CVON-DOSIS HCM study; and C. Hayward and L. Klarić, University of Edinburgh MRC Institute of Genetics & Molecular Medicine, for providing data from the CROATIA Korcula and Vis cohorts.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available upon request from third parties. The CROATIA Korcula and Vis data can be obtained upon request directed towards Prof. C. Hayward, IGMM, University of Edinburgh (caroline.hayward@igmm.ed.ac.uk). For the CVON-DOSIS HCM data, a request can be sent to Dr. M. Harakalova (m.harakalova@umcutrecht.nl). Restrictions apply to the availability of these data, which were used under license for this study.

SUPPORTING MATERIALS

Proofs, simulations and details for PO2PLS (pdf). This document contains additional materials for the methods, simulation and data analysis sections. First, details and proofs of theoretical variances and covariances, identifiability, maximum likelihood estimation and asymptotic results are derived. Then, additional results of the simulation study are shown. Finally, the results of the extra data analysis are shown.

ORCID

Said el Bouhaddani https://orcid.org/0000-0002-2279-4337

REFERENCES

Balliu, B., Tsonaka, R., Boehringer, S. & Houwing-Duistermaat, J. (2015) A retrospective likelihood approach for efficient integration of multiple omics factors in case-control association studies. *Genetic Epidemiology*, 39, 156–165.

el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G. & Uh, H.-W. (2016) Evaluation of O2PLS in omics data integration. *BMC Bioinformatics*, 17, S11.

el Bouhaddani, S., Uh, H.-W., Hayward, C., Jongbloed, G. & Houwing-Duistermaat, J. (2018) Probabilistic partial least squares model: identifiability, estimation and application. *Journal of Multivariate Analysis*, 167, 331–346.

el Bouhaddani, S., Uh, H.-W., Jongbloed, G., Hayward, C., Klarić, L., Kie lbasa, S.M., et al. (2018) Integrating omics datasets with the omicsPLS package. *BMC Bioinformatics*, 19, 371.

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

Crick, F. (1970) Central dogma of molecular biology. *Nature*, 227, 561–563. Available from: https://doi.org/10.1038/227561a0

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the {EM} algorithm. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 39, 1–38.

Domingo-Fernández, D., Hoyt, C.T., Bobis-Álvarez, C., Marín-Llaó, J. & Hofmann-Apitius, M. (2019) ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Systems Biology and Applications Impact Factor 2019*, 5, 43.

Geisser, S. (1993) Predictive inference. *Philosophy Science*, 24, 180.

Hemerich, D., Pei, J., Harakalova, M., van Setten, J., Boymans, S., Boukens, B.J. et al. (2019) Integrative functional annotation of 52 genetic loci influencing myocardial mass identifies candidate regulatory variants and target genes. *Circulation Genomic and Precision Medicine*, 12, 76–83.

Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, 28, 321.

Klarić, L., Tsepilov, Y.A., Stanton, C.M., Mangino, M., Sikka, T.T., Esko, T. et al. (2020) Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Science Advances*, 6, eaax0301.

Lauc, G., Essafi, A., Huffman, J.E., Hayward, C., Knežević, A., Kattla, J.J. et al. (2010) Genomics meets glycomics-the first gwas study of human N-glycome identifies HNF1A as a master regulator of plasma protein fucosylation. *PLoS Genetics*, 6, 1–14.

Li, G. & Jung, S. (2017) Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, 73, 1433–1442.

Lock, E.F., Hoadley, K.A., Marron, J.S. & Nobel, A.B. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7, 523–542.

Löfstedt, T. & Trygg, J. (2011) OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics*, 25, 441–455.

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 44, 226–233.

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979) *Multivariate analysis*. Cambridge, MA: Academic Press.

Meng, X.-L. & Rubin, D.B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80, 267–278.

Meng, C., Zeleznik, O.A., Thallinger, G.G., Kuster, B., Gholami, A.M. & Culhane, A.C. (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17, bbv108.

Mubeen, S., Hoyt, C.T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H. & Domingo- Fernández, D. (2019) The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Frontiers in Genetics*, 10, 1203.

Richardson, S., Tseng, G.C. & Sun, W. (2016) Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, 3, 181–209.

Sabater-Molina, M., Pérez-Sánchez, I., Hernández del Rincón, J. & Gimeno, J. (2018) Genetics of hypertrophic cardiomyopathy: a review of current state. *Clinical Genetics*, 93, 3–14.

Shapiro, A. (1983) Asymptotic distribution theory in the analysis of covariance structures (a unified approach). *South African Statistical Journal*, 17, 33–81.

Sun, Q., Zhu, H., Liu, Y. & Ibrahim, J.G. (2015) SPReM: sparse projection regression model for high-dimensional linear regression. *Journal of the American Statistical Association*, 110, 289–302.

Tissier, R. (2018) Statistical methods for the analysis of complex omics data. Ph.D. thesis, Leiden University, Leiden.

Tissier, R., Tsonaka, R., Mooijaart, S.P., Slagboom, E. & Houwing-Duistermaat, J.J. (2017) Secondary phenotype analysis in ascertained family designs: application to the Leiden longevity study. *Statistics in Medicine*, 36(14), 2288–2301.

Trygg, J. & Wold, S. (2003) O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*, 17, 53–64.

Vinga, S. (2020) Structured sparsity regularization for analyzing high-dimensional omics data. *Briefings in Bioinformatics*, 2020, 1–11.

Wahl, A., van den Akker, E., Klaric, L., Štambuk, J., Benedetti, E., Plomp, R. et al. (2018) Genome-wide association study on immunoglobulin G glycosylation patterns. *Frontiers in Immunology*, 9, 1–14.

Wold, H. (1973) Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate Analysis III (Proceedings of the 3rd Symphony Wright State University, Dayton, Ohio, 1972)*, New York: Academic Press, pp. 383–407.

Wold, H. (1985) Partial least squares. *Encyclopedia of Statistical Sciences*, 6, 581–591.

Zaytseva, O.O., Freidin, M.B., Keser, T., Štambuk, J., Ugrina, I., Šimurina, M. et al. (2020) Heritability of human plasma N-glycome. *Journal of Proteome Research*, 19, 85–91.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: el Bouhaddani, S., Uh, H.-W., Jongbloed, G. & Houwing-Duistermaat, J. (2022) Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(5), 1451–1470. Available from: https://doi.org/10.1111/rssc.12583

APPENDIX: AN EM ALGORITHM FOR PO2PLS

Theorem 4. *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,* $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot|X, Y, \theta^k]$.

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

The proof is given in the supplementary materials.