Statistical Integration of Multiple Omics Datasets Using GO2PLS

Zhujie Gu[1], Said el Bouhaddani[1], Jiayi Pei[2], Jeanine Houwing-Duistermaat[1, 3], and Hae-Won Uh[1]

[1]Dept. of Biostatistics and Research Support, Julius Centre, UMC Utrecht, The Netherlands
[2]Dept. of Cardiology, UMC Utrecht, The Netherlands
[3]Dept. of Statistics, University of Leeds, United Kindom

Abstract

Background: Nowadays, multiple omics data are measured on the same samples in the belief that these different omics datasets represent various aspects of the underlying biological systems. Integrating these omics datasets will facilitate the understanding of the systems. For this purpose, various methods have been proposed, such as Partial Least Squares (PLS), decomposing two datasets into joint and residual subspaces. Omics data are heterogeneous, and the joint subspaces estimated in PLS contain orthogonal variations unrelated to one another. Alternatively, Two-way Orthogonal Partial Least Squares (O2PLS) captures the heterogeneity by introducing the orthogonal subspaces and better estimates the joint subspaces. However, the latent components spanning the joint subspaces in O2PLS are linear combinations of all variables, while the interests of domain experts might be in a small subset. To obtain sparsity, we extend O2PLS to Group Sparse O2PLS (GO2PLS) that performs feature selection. Furthermore, features in the data often have group structures, and incorporating this information might improve the reliability of the selection procedure.

Results: The simulation study showed that introducing sparsity improved the performance concerning feature selection. Furthermore, incorporating group structures increased the precision and power of the feature selection procedure. GO2PLS performed optimally in terms of accuracy of joint score estimation, joint loading estimation, and feature selection. For illustration, we applied GO2PLS to datasets from two studies: CVON-DOSIS (small case-control study) and TwinsUK (population study). In the first, by integrating regulomics and transcriptomics data, joint components using GO2PLS discriminated cardiomyopathy patients and controls better than PCA and PLS. Genes selected based on the regulatory regions and transcripts appeared to be relevant to heart muscle disease. In the second, we incorporated the information on the group structures of the methylation CpG sites when integrating the methylation dataset with the glycomics data. The selected methylation groups turned out to be relevant to the immune system, in which glycans play an important role.

Conclusions: GO2PLS is an efficient approach for integrating two heterogeneous omics datasets to gain insights. It performs feature selection in both datasets, enhanced by incorporating known group structures, thereby resulting in a small subset of features for further investigation.

Keywords— Integration of Omics data, Dimension reduction, Feature selection, Group structure, O2PLS

1 Background

With the advancements in high throughput technology, multiple omics data are commonly available on the same subjects. Traditional statistical methods such as principal component analysis (PCA) can only handle one omics dataset at a time, overlooking the correlation between datasets. To better understand the underlying biological systems that involve multiple omics levels and identify potential biomarkers from each omics level, these datasets need to be analyzed jointly. Developing statistical methods for this purpose poses several challenges [1], among which we consider two major ones. The first is to account for the heterogeneity between different omics datasets. Each of the omics datasets (e.g., methylation and glycomics) has different sources of variation, distribution, dimensionality, etc. Ignoring these omic-specific characteristics in the model may lead to a biased representation of the underlying system [2]. The second is to introduce sparsity in the results and identify a small subset of relevant features. Omics datasets are usually high dimensional ($p \gg N$). Non-sparse methods produce results where all the variables get non-zero weights. To select a small subset of variables for further investigation, feature selection is needed to obtain sparse results by pushing a large number of small non-zero weights to zeros. In this work, we propose a method that deals with the heterogeneity and yields sparse results at the same time.

For omics data integration, Partial Least Squares (PLS) has been proposed and reviewed [3, 4]. PLS decomposes two datasets X and Y into joint and residual subspaces. The joint subspace of one dataset represents the best approximation of this dataset based on maximizing the covariance of the two datasets. However, it also contains variation that is unrelated to the other dataset. Therefore, it does not represent the true relationship between X and Y. Two-way orthogonal partial least squares (O2PLS) [5, 2] was proposed to decompose each dataset into joint, orthogonal, and residual subspaces. The orthogonal subspaces in X and Y capture variation unrelated to each other, making the joint subspaces better estimates for the true relation between X and Y.

Although O2PLS can deal with heterogeneity across omics datasets, the latent components spanning the joint subspaces are linear combinations of all the observed variables. To perform feature selection, several sparse methods based on PLS have been proposed. Chun and Keleş proposed sparse PLS (SPLS) [6] which fits PLS on a reduced X space. Sparse PLS (sPLS) by Lê Cao et al. [7] imposes L_1 penalty on the singular value decomposition (SVD) of the covariance matrix of X and Y, resulting in sparse loading vectors for both datasets. When group structures across the features are present (e.g., genes within the same pathway or CpG sites in the same genetic region), the aggregated effects of the features in the same group can be detected more easily [8]. Therefore, considering groups of features can make the biomarker identification procedure more effective than considering individual features [9, 10]. Liquet et al. extended sPLS to group PLS (gPLS) [10], imposing group-wise L_2 penalties on the loadings of the pre-defined feature groups. It

results in group-wise sparsity (i.e., features belonging to the same group will always be selected altogether).

In this work, we propose Group Sparse O2PLS (GO2PLS) to construct joint latent components representing the underlying systems involving both omics levels while taking into account the heterogeneity of different omics data, and to perform feature selection, enhanced by incorporating known group information of features.

Our motivating datasets come from two studies. In the CVON-DOSIS study, we aim to identify a small set of non-genetic determinants underlying hypertrophic cardiomyopathy (HCM). HCM is a heart muscle disease that makes it harder for the heart to pump blood, leading to heart failure. Regulomics (histone modification) and transcriptomics data are available, measured on 13 HCM cases and 10 controls. Histone modification is crucial for explaining the variation of the transcriptomics data. Nonetheless, both omics data contain variations that are not relevant to each other. For example, genetic mutation rather than epigenetic regulation partially explains variations of the gene expression. These omic-specific variations need to be taken into account to construct latent components that represent the non-genetic mechanisms when integrating the two omics datasets. To identify a small set of features for interpretation, we consider feature selection.

In the TwinsUK study [11, 12], methylation and immunoglobulin G (IgG) glycomics datasets were measured on 405 independent samples. IgG is an antibody important for immune response, whose functional diversity is mainly achieved by glycosylation[13]. Methylation as the most common epigenetic modification has an important role in the glycosylation pathways[14, 15]. The aim is to integrate the two omics datasets for studying the relationship between epigenetic modification and IgG glycosylation in general. The IgG glycomics data contain a source of variation unrelated to methylation, because other omics levels other than epigenomics, such as genomics, also has an impact on IgG glycosylation [16]. These glycomic-specific variations coming from other sources need to be taken into account when integrating the two omics datasets. Moreover, in the methylation data, group structures are present as neighboring CpG sites tend to be highly correlated [17]. A method capable of utilizing the grouping information to identify the groups of CpG sites that are relevant for IgG glycosylation is desired.

This paper is organized as follows. In the methods section, an overview of O2PLS is presented, followed by the formulation of GO2PLS. Via a simulation study, we explore the properties of GO2PLS and compare its performance to other competitive methods. We then apply GO2PLS to integrate regulomics and transcriptomics in the CVON-DOSIS study, and methylation and glycomics in the TwinsUK study. We conclude with a discussion and possible directions to further extend the method.

2 Methods

Two-way Orthogonal Partial Least Squares (O2PLS)

let X and Y be two data matrices with the number of rows equal to the sample size N and the number of columns equal to the dimensionality p and q, respectively. Let the number of joint, X-orthogonal (unrelated to Y) and Y-orthogonal components be K, K_x and K_y, respectively, where K, K_x and K_y are typically much smaller than p and q. The O2PLS model decomposes X and Y as follows:

$$\begin{aligned} X &= TW^\top + T_\perp P_\perp^\top + E, \\ Y &= UC^\top + U_\perp Q_\perp^\top + F. \end{aligned} \tag{1}$$

The relation between X and Y is captured through the inner relation between T and U,

$$\begin{aligned} U &= TB_T + H, \\ T &= UB_U + \tilde{H}. \end{aligned} \tag{2}$$

In this model, the scores are: $T\,(N \times K)$, $U\,(N \times K)$, $T_\perp\,(N \times K_x)$, $U_\perp\,(N \times K_y)$. They represent projections of the observed data X and Y to lower-dimensional subspaces. The loadings, $W\,(p \times K)$, $C\,(q \times K)$, $P_\perp\,(p \times K_x)$, $Q_\perp\,(q \times K_y)$, indicate relative importance of each X and Y variable in forming the corresponding scores. Further, $E\,(N \times p)$, $F\,(N \times q)$, $H\,(N \times K)$, $\tilde{H}\,(N \times K)$, represent the residual matrices.

In O2PLS, estimates of the joint subspaces are obtained by first filtering out the orthogonal variation. The filtered data matrices $\tilde{X}$ and $\tilde{Y}$ are constructed as follows:

$$\begin{aligned} \tilde{X} &= (I_N - T_\perp (T_\perp^\top T_\perp)^{-1} T_\perp^\top) X, \\ \tilde{Y} &= (I_N - U_\perp (U_\perp^\top U_\perp)^{-1} U_\perp^\top) Y, \end{aligned} \tag{3}$$

where $T_\perp\, U_\perp$ are estimates for the orthogonal subspaces, and I_N is identity matrix of size N. For more details see [5]. The joint parts maximize the covariance between the joint scores $T = \tilde{X}W$ and $U = \tilde{Y}C$. Here, W and C consist of loading vectors $(w_1, \ldots, w_K)$ and $(c_1, \ldots, c_K)$, which can be found as the right and left singular vectors of the covariance matrix $\tilde{Y}^\top \tilde{X}$ [2]. Calculating and storing $\tilde{Y}^\top \tilde{X}$ of dimension $q \times p$ can be cumbersome for high dimensional omics data. Therefore we consider the following optimization problem sequentially for components $k = 1, \ldots, K$:

$$\max_{\|c_k\|_2 = 1, \|w_k\|_2 = 1} c_k^\top \tilde{Y}_k^\top \tilde{X}_k w_k, \tag{4}$$

where parameters w_k, c_k are the loading vectors of the k-th joint components and $\tilde{X}_k$, $\tilde{Y}_k$ are the filtered data matrices after $k-1$ times of deflation. This can be solved efficiently using NIPALS [18] algorithm. A detailed description and proof of optimality of the O2PLS algorithm can be found in [5, 2].

While standard cross-validation (CV) over a 3-dimensional grid is often used to determine the optimal number of components K, K_x, and K_y, the procedure is not optimal for O2PLS, since there is not a single optimization criteria for all three parameters. As in [2], we use an alternative CV procedure that first performs a 2-dimensional grid search of K_x and K_y, with a fixed K, to optimize prediction performance of $T \to U$ and $U \to T$. Then a sequential search of optimal K is conducted to minimize the sum of mean squared errors (MSE) of prediction concerning $X \to Y$ and $Y \to X$.

Group Sparse O2PLS (GO2PLS)

GO2PLS extends O2PLS by introducing a penalty in the NIPALS optimization on the filtered data $\tilde{X}$ and $\tilde{Y}$. This penalty encourages sparse, or group-sparse solutions for the joint loading matrices W and C, leading to a subset of the original features corresponding to non-zero loading values being selected in each joint component.

Briefly, we introduce an L_1 penalty on each pair of joint loading vectors. The optimization problem for the k-th pair of joint loadings c_k, w_k is:

$$\max_{\|c_k\|_2 = 1, \|w_k\|_2 = 1} c_k^\top \tilde{Y}_k^\top \tilde{X}_k w_k + \lambda_c \|c_k\|_1 + \lambda_w \|w_k\|_1 , \tag{5}$$

where λ_c, λ_w are penalization parameters that regulate the sparsity level. The optimization problem (5) can be solved [19] by iterating over the k-th pair of joint loadings,

$$c_k = \frac{S(\tilde{Y}_k^\top t, \lambda_c)}{\left\| S(\tilde{Y}_k^\top t, \lambda_c) \right\|_2}, \; w_k = \frac{S(\tilde{X}_k^\top u, \lambda_w)}{\left\| S(\tilde{X}_k^\top u, \lambda_w) \right\|_2}, \tag{6}$$

where $t = \tilde{X}_k w_k$ and $u = \tilde{Y}_k c_k$. Here, $S(\cdot)$ is the soft thresholding operator: $S(a, \text{const}) = sgn(a)(|a| - \text{const})_+$ (const ≥ 0 is a non-negative constant, $(x)_+$ equals to x if $x > 0$ and equals to 0 if $x \leq 0$).

To perform group selection, we impose group-wise L_2 penalty on the joint loading vectors. Let $\tilde{X}$ and $\tilde{Y}$ be partitioned into J ($J \leq p$) and M ($M \leq q$) groups, respectively. The submatrices $\tilde{X}^{(j)}$ and $\tilde{Y}^{(m)}$ ($j = 1, \ldots, J$; $m = 1, \ldots, M$) contain the j-th and m-th group of variables, with corresponding loading vectors $w^{(j)}$ (of size p_j) and $c^{(m)}$ (of size q_m). The optimization problem for the k-th pair of loading vectors $c_k = ({c_k^{(1)}}^\top, \ldots, {c_k^{(M)}}^\top)^\top$ and $w_k = ({w_k^{(1)}}^\top, \ldots, {w_k^{(J)}}^\top)^\top$ can be written as follows:

$$\begin{aligned} \min_{c_k^{(m)}, w_k^{(j)}} \Bigg\{ & -\sum_{j=1}^{J} \sum_{m=1}^{M} {c_k^{(m)}}^\top {\tilde{Y}_k^{(m)}}^\top \tilde{X}_k^{(j)} w_k^{(j)} \\ & + \lambda_c \sum_{m=1}^{M} \sqrt{q_m} \left\| c_k^{(m)} \right\|_2 + \lambda_w \sum_{j=1}^{J} \sqrt{p_j} \left\| w_k^{(j)} \right\|_2 \\ & + \phi_c \left(\sum_{m=1}^{M} \left\| c_k^{(m)} \right\|_2^2 - 1 \right) + \phi_w \left(\sum_{j=1}^{J} \left\| w_k^{(j)} \right\|_2^2 - 1 \right) \Bigg\}, \end{aligned} \tag{7}$$

where the last two terms are reformulations of the unit norm constraints on c_k and w_k, with ϕ_c and ϕ_w being the Lagrangian multipliers. The effective penalization parameters on each group (λ_c, λ_w) are adjusted by the square root of the group size to correct for the fact that larger groups are more likely to be selected. This optimization problem can be solved using block coordinate descent (for details, see Additional file 1). The solution takes the form:

$$\begin{aligned} c_k^{(m)} &= \frac{\left(\left\| {\tilde{Y}_k^{(m)}}^\top t \right\|_2 - \sqrt{q_m} \lambda_c \right)_+}{2\phi_c \left\| {\tilde{Y}_k^{(m)}}^\top t \right\|_2} \; {\tilde{Y}_k^{(m)}}^\top t, \\ w_k^{(j)} &= \frac{\left(\left\| {\tilde{X}_k^{(j)}}^\top u \right\|_2 - \sqrt{p_j} \lambda_w \right)_+}{2\phi_w \left\| {\tilde{X}_k^{(j)}}^\top u \right\|_2} \; {\tilde{X}_k^{(j)}}^\top u. \end{aligned} \tag{8}$$

The $\tilde{X}$-variables within the j-th group will have non-zero weights if $\left\| {\tilde{X}_k^{(j)}}^\top u \right\|_2$ (i.e., the contribution of the whole group to the covariance) is larger than the size-adjusted penalization parameter $\sqrt{p_j}\lambda_w$. In the same way, the $\tilde{Y}$-variables within the m-th group will be assigned non-zero loading values if $\left\| {\tilde{Y}_k^{(m)}}^\top t \right\|_2 > \sqrt{q_m}\lambda_c$.

Note that when all the groups have size 1, the summation of group-wise L_2 penalties is equivalent to an L_1 penalty on the unpartitioned loading vector and individual features will be selected (i.e., (7) reduces to (5)). In this specific case, to avoid confusion, we call the method Sparse O2PLS (SO2PLS). When the penalization parameters $\lambda_w = \lambda_c = 0$, GO2PLS becomes to O2PLS. If the number of orthogonal components $K_x = K_y = 0$, GO2PLS, SO2PLS, O2PLS are equivalent to gPLS, sPLS, and PLS, respectively.

The k-th pair of joint loadings are orthogonalized with respect to the previous $k-1$ loading vectors. Let π be an index set for selected variables in w_k. The orthogonalization is achieved by first projecting $w_k^{(\pi)}$ onto $span\{w_1^{(\pi)}, \ldots, w_{k-1}^{(\pi)}\}$, and then subtracting this projection from $w_k^{(\pi)}$. When the previous $k-1$ components do not select any variable in π, $span\{w_1^{(\pi)}, \ldots, w_{k-1}^{(\pi)}\}$ is actually a zero subspace and no orthogonalization is needed.

To determine the optimal sparsity level, it is more convenient and intuitive to focus on the number of selected $\tilde{X}$, $\tilde{Y}$ groups (donote h_x, h_y, respectively). If prior biological knowledge does not already specify certain h_x and h_y, cross-validation can be used to search for combinations of h_x and h_y that maximize the covariance between each pair of estimated joint components $\mathrm{Cov}(\hat{t}, \hat{u})$. Similar to LASSO [20], the "one-standard-error-rule" [21] can be applied to obtain a more stable CV result. The GO2PLS algorithm is described below:

3 Simulation Study

We evaluate the performance of GO2PLS in two scenarios. First, we investigate the power to detect relevant groups with different group sizes, focusing on the joint subspace, where the feature selection takes place. Second, we compare the performance of GO2PLS with other methods: SO2PLS, O2PLS, PLS, sPLS, and gPLS. We investigate joint score estimation, joint loading estimation, and feature selection performances. An overview of scenario settings is presented in Table 1, 2.

In the first scenario, we set the number of variables in X and Y to be $p = 5000$ and $q = 20$, respectively. There are 10 groups of variables in X with non-zero loading values. The first 5 groups have group sizes of 100, 50, 20, 5, and 1, respectively, in which all the variables have loading values equal to 1. The remaining 5 groups are of size 10, with loading values of variables equal to 5. Note that large loading values are assigned to the latter 5 groups to make the detection of the first 5 groups more difficult, thus making the effect of group size on the detection power easier to observe. The remaining variables have zero loading values and are divided into groups of size 10. All the Y-variables have the same loading values and are not grouped. The sample size N is set to 30. We simulate both data matrices with 1 joint component ($T = U \in \mathcal{N}(0,1)$). The statistic, $\frac{\left\| X^{(j)\top} U \right\|_2}{\sqrt{(p_j)}}$, that determines whether a group is selected or not is recorded for the first 5 groups (with loading value 1) over 1000 repetition under each noise level, and the power of detecting each of these groups is calculated.

In the second scenario, we set $p = 20000$ and $q = 10000$, mimicking the dimensionality of the CVON-DOSIS datasets. Both X- and Y- variables are evenly divided into 1000 groups. For each joint component, we select 50 relevant groups and assign non-zero loadings to the variables contained in them. Within each group, variables have the same loading values: 1 for the first group, 2 for the second,..., and 50 for the last relevant group. We set the number of joint components $K = 2$ and the number of orthogonal components $K_x = K_y = 1$. The scores $T, T_\perp, U, U_\perp$ are generated from normal distributions with zero

Algorithm: GO2PLS

1 Get $\tilde{X}$ and $\tilde{Y}$ by removing orthogonal variation from X and Y:

(I) Apply NIPALS on X and Y, get an initial estimate of score matrices T, U and loading matrices W, C;

(II) $E = X - TW^\top$; $F = Y - UC^\top$;

(III) $W_\perp = K_x$ left singular vectors of SVD $(E^\top T)$;
$C_\perp = K_y$ left singular vectors of SVD $(F^\top U)$;
$T_\perp = XW_\perp$; $U_\perp = YC_\perp$;

(IV) $\tilde{X} = (I - T_\perp(T_\perp^\top T_\perp)^{-1}T_\perp^\top)X$;
$\tilde{Y} = (I - U_\perp(U_\perp^\top U_\perp)^{-1}U_\perp^\top)Y$.

2 Calculate joint loadings and joint scores sequentially:

(I) Let $\tilde{X}_1 = \tilde{X}; \tilde{Y}_1 = \tilde{Y}$;

(II) For $k = 1, 2, \ldots, K$:

(a) Iterate between c_k and w_k until convergence, following Formula (8) (or Formula (6) for SO2PLS);

(b) Orthogonalization of c_k, w_k with regard to the previous $k-1$ loading vectors;

(c) $t_k = \tilde{X}_k w_k$; $u_k = \tilde{Y}_k c_k$;

(d) $p_k = \tilde{X}_k^\top t_k/(t_k^\top t_k)$; $q_k = \tilde{Y}_k^\top u_k/(u_k^\top u_k)$;

(e) $\tilde{X}_{k+1} = \tilde{X}_k - t_k p_k^\top$; $\tilde{Y}_{k+1} = \tilde{Y}_k - u_k q_k^\top$;

(III) $T = [t_1, \ldots, t_K]$; $U = [u_1, \ldots, u_K]$;
$W = [w_1, \ldots, w_K]$; $C = [c_1, \ldots, c_K]$.

mean. The relationship between the joint scores is represented by $U = T + H$, where H accounts for 20% of the variation in U. Sample size N, ratio of the variance of the orthogonal components to the variance of the joint components ($\sigma^2_{T_\perp}/\sigma^2_T$), and noise level α are varied. For evaluating the accuracy of the joint score estimation, we computed $R^2_{\widehat{T}T} = 1 - \sum(\widehat{T} - T)^2/\sum T^2$ and $R^2_{\widehat{T}\widehat{U}} = 1 - \sum(\widehat{U} - \widehat{T})^2/\sum \widehat{U}^2$, which quantify how well the true parameter T and the estimated Y-joint component $\widehat{U}$ can be explained by the estimated X-joint component $\widehat{T}$. The performance of feature selection and the accuracy of estimated loadings are evaluated by true positive rate (TPR = TP/(TP+FN), where TP = True Positive, FN = False Negative) and $W^\top \widehat{W}$, which represents the angle between the estimated loading vector and the true one. The performances of all methods are evaluated on an independent test dataset of size 1000. For each setting, 500 replications are generated.

To make a clear comparison of the behavior across all the methods, we use the optimum values for the tuning parameters (number of components and number of relevant variables

or groups).

Table 1: Settings to study the property of GO2PLS

Measure	$\frac{\left\|X^{(j)\top}U\right\|_2}{\sqrt{(p_j)}}$; Selection frequency
$p; q$	$5000; 20$
relevant group sizes	$100; 50; 20; 5; 1$
N	30
noise level α	$[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$

Table 2: Settings to compare the performances

Methods	GO2PLS; SO2PLS; O2PLS; gPLS; sPLS; PLS
Measure	$R^2_{\widehat{T}T}$, $R^2_{\widehat{T}\widehat{U}}$, TPR, $W^\top\widehat{W}$
$p; q$	$20,000; 10,000$
relevant $p; q$	$1000; 500$
N	$[30, 100, 200, 300, 600]$
$\sigma^2_{t_\perp}/\sigma^2_t$	$[1/5, 1/3, 1/2, 1, 2, 3, 5]$
noise level α	$[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$

Results of simulation study

Scenario 1

Fig 1 shows the distribution of the statistic $\frac{\left\|X^{(j)\top}U\right\|_2}{\sqrt{(p_j)}}$ for the first 5 relevant groups with different group sizes under 3 different noise levels. The vertical dotted lines indicate the average threshold given the correct number of relevant groups. Since a group will be selected if $\frac{\left\|X^{(j)\top}U\right\|_2}{\sqrt{(p_j)}}$ exceeds the threshold, the total area on the right side of the threshold under each density curve is proportional to the detection power for the corresponding group. The statistic for larger relevant group shows higher precision at all noise levels. The threshold increases along with the noise. Fig 2 shows the detection power for each relevant group under each noise level. Compared to smaller groups, the detection power for larger groups is higher at low to moderate ($\alpha < 0.7$) noise levels, and shows robustness against increasing noise. When the noise level is very high ($\alpha > 0.8$), the method loses power to detect relevant group of any size, particularly, of larger size.

Scenario 2

The performance of the joint score estimation is compared focusing on the difference between methods with orthogonal parts (GO2PLS, SO2PLS, O2PLS) and their counterparts without the "O2" filtering (gPLS, sPLS, PLS). Fig 3 shows the performance measured by $R^2_{\widehat{T}T}$ & $R^2_{\widehat{T}\widehat{U}}$ under $N = 30$, $\alpha = 0.1$ and varying relative orthogonal signal strength from one fifth to five times of the joint signal. In the left panel, $R^2_{\widehat{T}T}$ of the various methods

Figure 1: Density plot of estimated $\frac{\left\|X^{(j)\top}U\right\|_2}{\sqrt{(p_j)}}$ for each group size under 3 different noise levels. The vertical dotted line indicating the average threshold.

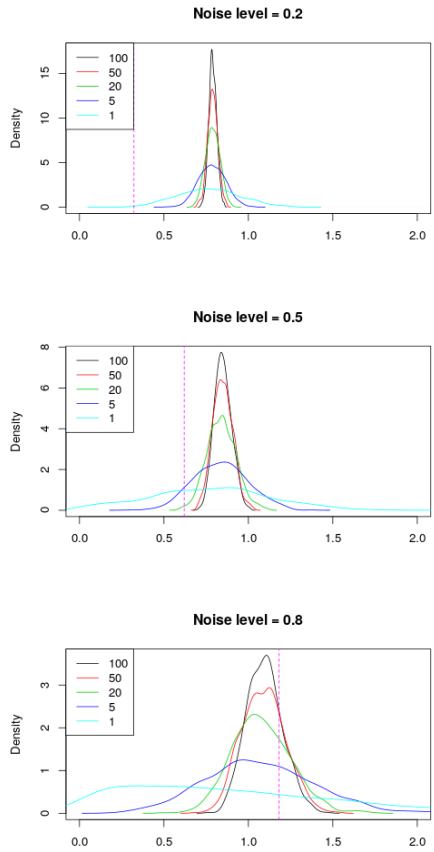


Figure 2: Detection power of relevant groups with different sizes

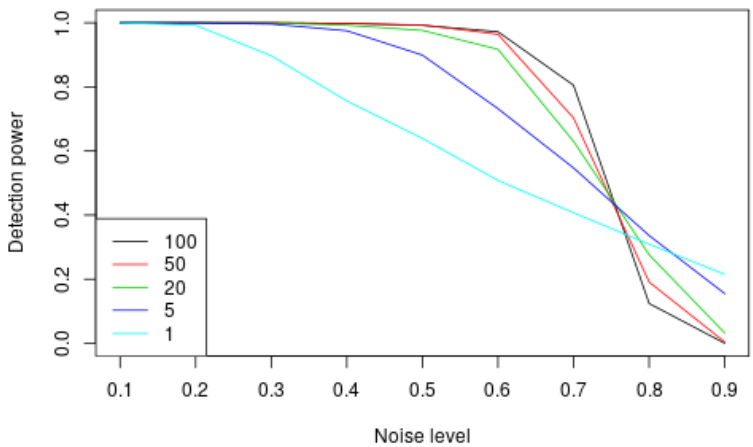


Figure 3: **Joint score estimation performance**. X-axis is the relative orthogonal signal strength; Y-axis is $R^2_{\widehat{T}T}$ (left), $R^2_{\widehat{T}\widehat{U}}$ (right)

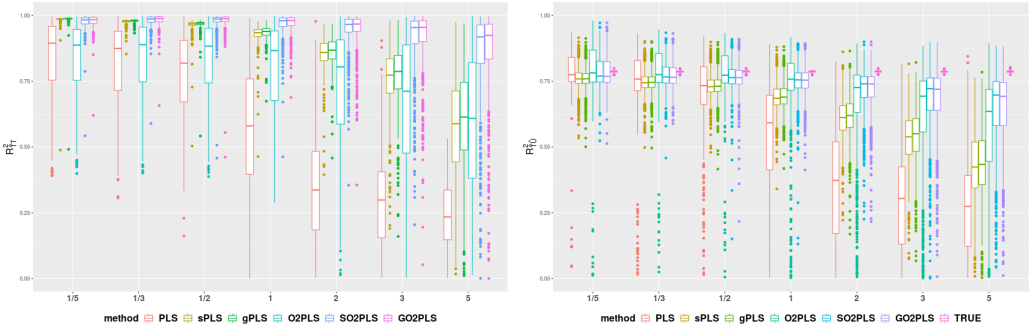
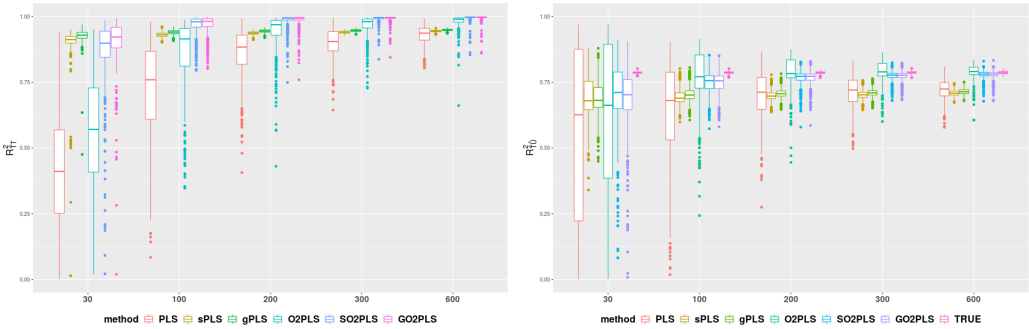


Figure 4: **Joint score estimation performance**. X-axis is sample size N; Y-axis is $R^2_{\widehat{T}T}$ (left), $R^2_{\widehat{T}\widehat{U}}$ (right)



is depicted, representing how well the joint component $\hat{T}$ captured the true underlying T. Overall, penalized methods performed better than non-penalized ones, especially when the orthogonal variation is relatively small. PLS performed poorly compared to O2PLS, when the orthogonal variation exceeds the joint variation. As the orthogonal variation further increases, performances of sPLS and gPLS deteriorated, while SO2PLS and GO2PLS were less affected. In the right panel, $R^2_{\widehat{T}\widehat{U}}$ is presented, an estimate of the true parameters R^2_{TU}, capturing correlation of T and U. Across different settings, O2PLS-based methods performed better, especially when the orthogonal variation is large.

Fig 4 shows the score estimation performance under fixed relative orthogonal signal strength of 1 and varying sample size N from 30 to 600. Penalized methods performed better compared to non-penalized methods in general, when the sample size is small. Regardless of the sample size, O2PLS-based methods outperformed PLS-based methods.

Lastly, we present the results of GO2PLS, SO2PLS, and O2PLS with regard to feature selection and estimation of joint loadings. Results of PLS-based methods are not included since the performances of gPLS, sPLS, and PLS in this regard are very similar to GO2PLS, SO2PLS, and O2PLS, respectively. Fig 5 shows the TPR and $W^\top \hat{W}$ under $N = 30$ and varying noise levels α from low to high. At all noise levels, GO2PLS had higher TPR than SO2PLS and O2PLS, and performed robustly against increasing noise. Regarding $W^\top \hat{W}$, GO2PLS outperformed the other two as well. In Fig 6, when increasing sample size at a fixed noise level of 0.5, the variance appeared to decrease and the performances of all the methods converged. Overall, GO2PLS outperformed others.

Figure 5: **Feature selection and joint loading estimation performance**. X-axis is noise level; Y-axis is TPR (left) and $W^\top \hat{W}$ (right)

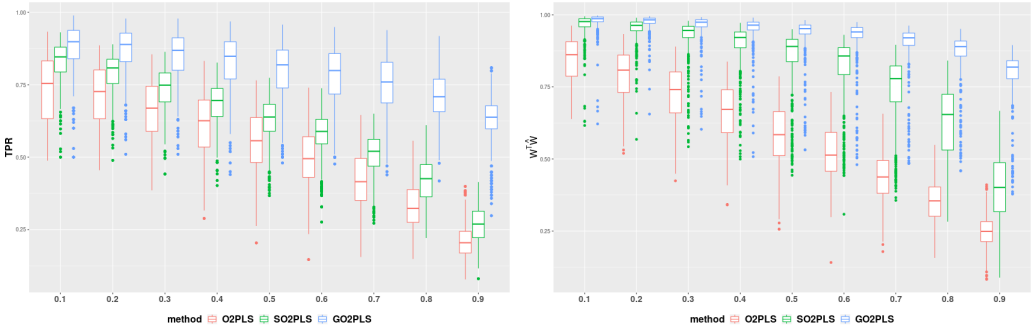
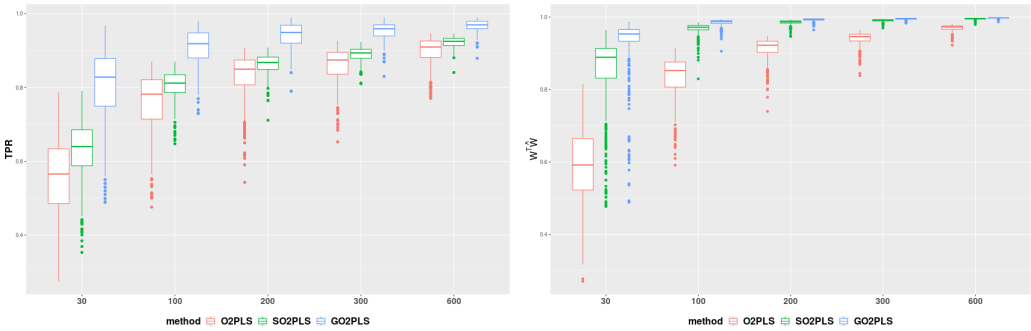


Figure 6: **Feature selection and joint loading estimation performance**. X-axis is sample size N; Y-axis is TPR (left) and $W^\top \hat{W}$ (right)



4 Application to data

We demonstrate SO2PLS and GO2PLS on datasets from two distinct studies. We first apply SO2PLS to the CVON-DOSIS study to integrate regulomics and transcriptomics data and select the features that contribute most to the joint components. The integration helps us gain insight into the underlying systems that involve both omics levels, and the feature selection makes it easier to perform further investigation. Next, we apply GO2PLS to the methylation and IgG glycomics data from TwinsUK study, which have noticeable differences with the CVON-DOSIS data (e.g., different study designs, omics levels, dimensionality, etc.). We utilize the group structures in the methylation data to help the selection of CpG sites related to IgG glycosylation. The selection at a group level also makes interpretation more convenient and intuitive.

CVON-DOSIS study

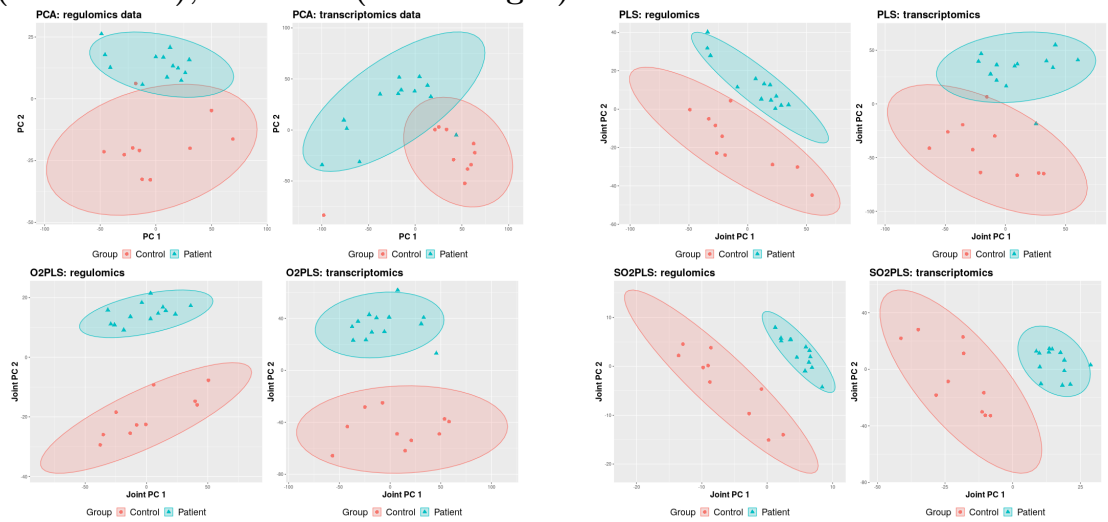
Regulomics and transcriptomics datasets were measured on the samples taken from the heart tissue of 13 HCM patients and 10 healthy controls. The regulomics data were measured using ChIP-seq, providing counts of histone modification H3K27ac in 33642 regulatory regions. The transcriptomics data were measured using RNA-seq, containing counts of 15882 transcripts. The raw counts of regulomics data were normalized using reads per kilobase million (RPKM) to adjust for sequencing depth. Transcriptomics data were normalized using counts per million (CPM) with effective library size (estimated using the TMM method in EdgeR R package [22]). Further, both normalized data were

log-transformed.

Fig 7 shows the score plots of PCA, PLS, O2PLS, and SO2PLS. Performing PCA on each dataset separately did not distinguish the two groups well (i.e., overlapping confidence regions and misclassified samples). Integrating the two datasets with PLS showed improved separation of the two groups, although confidence regions of the groups were close to each other. When applying O2PLS with 2 joint and 1 orthogonal components for each omics dataset, group separation became clearer. SO2PLS (1000 regulomics and 500 transcriptomics variables were selected in each pair of the joint components) showed a similar separation of patients and controls, with smaller confidence regions.

Fig 8 shows the loading values of O2PLS and SO2PLS. For O2PLS, all the loading values are non-zero. It is difficult to select although an ellipse could be drawn based on a certain confidence level and the features outside of the confidence region could be selected. For SO2PLS, only the selected features have non-zero loading values, making the plot easier to read and interpret.

Figure 7: **Labeled score plots of PCA (topleft), PLS (topright), O2PLS (bottomleft), SO2PLS (bottomright)**



The genes corresponding to the promoter regions where the selected 1000 histone modification locates (using ± 10K window from the transcription start site of the gene) were identified for further investigation. These genes are of interest since they are likely to be regulated. A gene set enrichment analysis [23] was performed on these genes using the ToppGene Suite [24], and the top results were listed in Table 3, where the "pValue" column shows the p-value of each annotation derived by random sampling of the whole genome; the "FDR B&H" column provides the false discovery rate (FDR) analog of the p-value after correcting for multiple hypothesis testing [25, 26]. The 500 transcripts selected were also interpreted by performing an enrichment analysis on the corresponding genes (results in Table 4). The genes identified from both regulomics and transcriptomics data appeared to be highly related to heart muscle development and heart failure, which is in line with previous studies in HCM [27].

TwinsUK study

Whole blood methylation (using Infinium HumanMethylation450 BeadChip) and IgG glycomics (Ultra Performance Liquid Chromatography) data were measured on 405 indepen-

Figure 8: **Loading plots of O2PLS (top) and SO2PLS (bottom)**

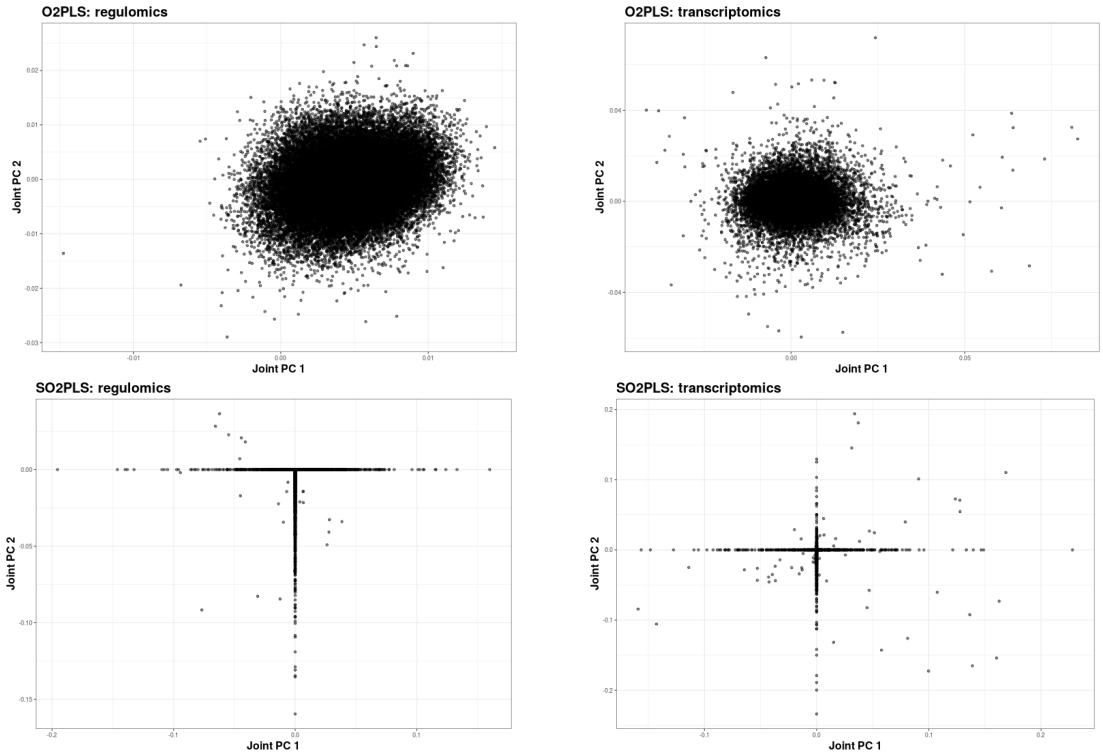


Table 3: Gene set enrichment analysis results of regulomics data in CVON-DOSIS study

	Name	pValue	FDR B&H
GO: Biological Process	actin cytoskeleton organization	5.265E-8	2.625E-4
	muscle structure development	1.125E-7	2.625E-4
	muscle tissue development	2.623E-7	3.869E-4
	striated muscle tissue development	1.932E-6	2.029E-3
	regulation of cellular response to stress	2.282E-6	2.054E-3
	striated muscle cell differentiation	7.865E-6	6.194E-3
	heart development	1.290E-5	8.128E-3
	regulation of muscle tissue development	4.679E-5	2.268E-2
GO: Cellular Component	contractile fiber	4.953E-8	2.512E-5
	myofibril	7.046E-8	2.512E-5
	I band	8.307E-7	1.974E-4
	Z disc	1.221E-6	2.176E-4
	sarcomere	3.564E-6	5.082E-4

dent individuals, among which 392 are females and 13 are males. The age ranges from 18 to 81, with a median of 58. The methylation dataset consists of beta values (ratio of intensities between methylated and unmethylated alleles) at 482563 CpG sites. CpG sites with missing values, on allosomes, or labeled cross-active [28] were removed. We kept only the CpG sites on CpG islands or surrounding areas (shelves and shores) that mapped to genetic regions. The remaining 126299 CpG sites were then divided into 16892 groups

Table 4: Gene set enrichment analysis results of transcriptomics data in CVON-DOSIS study

	Name	pValue	FDR B&H
GO: Biological Process	blood circulation	2.849E-8	5.556E-5
	ion transport	2.872E-8	5.556E-5
	circulatory system process	7.434E-8	1.386E-4
	heart contraction	4.901E-7	4.229E-4
	regulation of heart contraction	1.584E-6	6.219E-4
	regulation of blood circulation	2.343E-6	6.664E-4
Disease (Source: DisGeNET Curated)	Myocardial Ischemia	5.016E-9	1.006E-5
	Heart failure	9.128E-9	1.220E-5
	Heart Decompensation	2.454E-8	1.968E-5
	Myocardial Failure	5.069E-8	2.258E-5
	Hypertensive disease	3.176E-7	1.158E-4

such that CpG sites in the same group map to the same gene. Age, sex, batch effect, and cell counts were corrected using multiple regression. The glycomics dataset contains 22 glycan peaks. These peaks were normalized using median quotient (MQ) normalization [29], log-transformed, and adjusted for batch effect, age, and sex as well. No group information was used for the glycomics data.

We performed GO2PLS on the data with 1 joint, no methylation-orthogonal, and 3 glycomics-orthogonal components based on 5-fold cross-validation. Since IgG glycosylation was reported to be highly complex and to involve many genes [15], we set the sparsity parameters to select the top 100 groups in the methylation and kept all the 22 glycan variables. We also applied SO2PLS selecting 100 individual CpG sites instead of groups of CpG sites for comparison. The selected CpG groups from GO2PLS and individual CpG sites from SO2PLS were all mapped to genes for interpretation.

Among the selected genes from SO2PLS, about 65% were also selected by GO2PLS. We performed gene set enrichment analyses on the selected genes from GO2PLS and SO2PLS. Results from both methods appeared to be related to immune response. In particular, the results from GO2PLS seemed more significant than the results from SO2PLS. We listed the most significant molecular function, biological process, and pathway in Table 5 and Table 6.

A further analysis excluding CpG sites that do not map to the promoter region (0-1500 bases upstream of the TSS) of a gene was performed. There were 55531 CpG sites and 14491 groups remained. Similarly, we applied GO2PLS and selected 100 CpG groups. Among the selected, 24 groups overlapped with the previous results of GO2PLS. Gene set enrichment analysis did not found significant results. The reason might be that the CpG sites remote from annotated TSSs also play a role in the immune system. Excluding these CpG sites resulted in a loss of information and made the groups smaller, which lowered the power of the method, as demonstrated in the simulation study. Moreover, the gene sets for comparison in the enrichment analysis are based on transcriptomics and proteomics databases. The findings based on methylation data might be incomplete.

Table 5: SO2PLS: top results of gene set enrichment analysis in TwinsUK study

	Name	pValue	FDR B&H
GO: Molecular Function	peptide antigen binding	2.033E-5	7.971E-3
GO: Biological Process	homophilic cell adhesion via plasma membrane adhesion molecules	9.293E-8	2.162E-4
Pathway (Source: KEGG)	Viral myocarditis Allograft rejection Graft-versus-host disease	1.058E-5 3.625E-5 4.918E-5	1.760E-3 4.522E-3 4.908E-3

Table 6: GO2PLS: top results of gene set enrichment analysis in TwinsUK study

	Name	pValue	FDR B&H
GO: Molecular Function	peptide antigen binding	4.470E-7	1.618E-4
GO: Biological Process	interferon-gamma-mediated signaling pathway	2.885E-6	6.335E-3
Pathway (Source: KEGG)	Viral myocarditis Staphylococcus aureus infection Allograft rejection Graft-versus-host disease Translocation of ZAP-70 to Immunological synapse Rheumatoid arthritis Autoimmune thyroid disease	1.674E-8 3.508E-7 1.246E-6 1.839E-6 4.067E-6 5.886E-6 6.720E-6	9.089E-6 9.524E-5 2.255E-4 2.496E-4 3.681E-4 3.705E-4 3.705E-4

5 Discussion and conclusion

Statistical integration of two omics datasets is becoming increasingly popular to gain insight into underlying biological systems. It is essential to take omic-specific variation into account when integrating heterogeneous omics data. Among latent variable models, O2PLS separates the omic-specific variation from joint subspaces by introducing orthogonal subspaces. We extended O2PLS to GO2PLS to perform group selection in the joint parts. Compared to the joint components of O2PLS, which are linear combinations of all the observed variables, the joint components of GO2PLS are constructed by only a small subset of relevant features. GO2PLS incorporates known group structures in the feature selection procedure to increase detection power. Furthermore, it is often more convenient to interpret the selection at a group level than at an individual level.

The simulation study showed that O2PLS-based methods generally outperformed PLS-based methods regarding joint score estimation when orthogonal variation was present in the data. Due to the absence of the orthogonal subspaces, PLS-based methods captured part of the orthogonal variation in the joint subspaces, making the estimation biased. When the sample size was small or the noise level was high, non-penalized methods suffered from overfitting, while penalized methods appeared to be much less prone to that. Therefore, a sparse solution increased the generalizability of the results. Concerning feature selection, GO2PLS had higher TPR than SO2PLS. This result was in line with that shown in the simulation study; the larger groups of relevant features had a higher proba-

bility of being detected under a reasonable noise level.

We applied SO2PLS and GO2PLS to the CVON-DOSIS and the TwinsUK studies. In the first study, due to the heterogeneity present in the individual dataset, SO2PLS better separated the cases and controls than O2PLS or PCA. We visualized the loadings of both SO2PLS and O2PLS. The loading plots of SO2PLS summarized the data better than the ones of O2PLS since each joint component only contained a small number of non-zero values. Gene set enrichment analyses showed that the genes selected by SO2PLS appeared to be highly related to heart muscle disease. In the second study, we utilized the group structures in the methylation dataset and selected 100 groups of CpG sites. The results of GO2PLS and SO2PLS were similar, but the enrichment analysis of selected genes by GO2PLS showed more significance than the results of SO2PLS.

Group information has a large impact on the feature selection procedure. A different way of grouping features will likely lead to a different selection of groups and features. We have shown that GO2PLS has more power to detect large relevant groups. However, it is based on the assumption that all the features in the group are relevant. If a large group contains both relevant and irrelevant features, the results based on this group information become unreliable. At the group level, irrelevant features possibly make the group-size-adjusted contribution to the covariance smaller. Consequently, the group might be less likely to be detected, lowering the power of the method. At the individual feature level, the irrelevant features in a selected group are regarded as selected as well, resulting in a higher false positive rate. Therefore, group information should be chosen with care.

Typically, omics data are collected, aiming to identify biomarkers for an outcome variable. To this end, a supervised or semi-supervised extension of GO2PLS needs to be developed. The (semi-)supervised approach might also lead to a more sparse solution since the selected features have to be correlated between the omics datasets and the outcome variable. We have shown the advantage of jointly analyzing two omics datasets over non-integrative analysis. Further extending GO2PLS to incorporate more than two omics levels might better represent the actual biological system. Finally, it is possible to extend the GO2PLS algorithm to a probabilistic model. Extending latent variable methods to probabilistic models is not new. PCA was extended to Probabilistic PCA in [30], and PPLS [31] was proposed to provide a probabilistic framework for PLS. It has been shown that the probabilistic counterpart has a lower bias in estimation and is robust to non-normally distributed variables [31]. More importantly, the probabilistic model will allow statistical inference, making it possible to interpret the relevance and importance of features in the population. These extensions of GO2PLS will be suited for various studies with more complicated designs.

To conclude, GO2PLS is efficient in estimating latent components that represent underlying systems. It is capable of utilizing known group structures to enhance the feature selection procedure, providing a small number of potential biomarkers for further investigation.

6 Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The R scripts and functions for GO2PLS are publicly available in the OmicsPLS R package `https://cran.r-project.org/package=OmicsPLS` and can be installed in R via `install.packages("OmicsPLS")`.

Because of the sensitive nature of the data collected for the CVON-DOSIS study, requests to access the dataset from qualified researchers trained in human subject confidentiality protocols may be sent to the corresponding authors.

Individual level methylation and glycomics data from TwinsUK are not permitted to be shared or deposited due to the original consent given at the time of data collection. However, access to methylation and glycomics data can be applied for through the TwinsUK data access committee. For information on access and how to apply, visit http://www.twinsuk.ac.uk/data-access/submission-procedure/.

Competing interests

The authors declare that they have no competing interests.

Funding

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme IMforFUTURE under H2020-MSCA-ITN grant agreement number 721815.

Authors' contributions

ZG performed the mathematical work, simulation study, data anaylysis, and wrote the manuscript. JH and H-WU provided the underlying idea of the methods. SB supported the programming. All authors contributed to the discussion of the methods, simulation and data analysis. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank M. Harakalova, and M. Mokry from the Dept. of Cardiology, UMC Utrecht for providing the CVON-DOSIS data and discussion on the analysis of the CVON-DOSIS datasets. We thank M. Michels and J. van der Velden for providing the HCM tissues, the biobank of UMC Utrecht, the biobank of the Washington University School of Medicine, and the Sydney Heart Bank for providing non-failing donor tissue.

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking BigData@Heart grant (116074).

TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation (CDRF), Zoe Global Ltd and the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

References

[1] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., Tegnér, J.: Data integration in the era of omics: current and future challenges. BioMed Central (2014). doi:10.1186/1752-0509-8-S2-I1. http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-8-S2-I1

[2] el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., Uh, H.W.: Evaluation of O2PLS in Omics data integration. BMC Bioinformatics **17**(2), 1–20 (2016). doi:10.1186/s12859-015-0854-z

[3] Boulesteix, A.L., Strimmer, K.: Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics **8**(1), 32–44 (2007). doi:10.1093/bib/bbl016

[4] Wold, S., Ruhe, A., Wold, H., Dunn, III, W.J.: The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. SIAM Journal on Scientific and Statistical Computing **5**(3), 735–743 (1984). doi:10.1137/0905052. arXiv:1308.0863v1

[5] Trygg, J., Wold, S.: O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. Journal of Chemometrics **17**(1), 53–64 (2003). doi:10.1002/cem.775

[6] Chun, H., Keleş, S.: Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Series B: Statistical Methodology **72**(1), 3–25 (2010). doi:10.1111/j.1467-9868.2009.00723.x

[7] Lê Cao, K.A., Rossouw, D., Robert-Granié, C., Besse, P.: A sparse PLS for variable selection when integrating omics data. Statistical Applications in Genetics and Molecular Biology **7**(1) (2008). doi:10.2202/1544-6115.1390

[8] Tyekucheva, S., Marchionni, L., Karchin, R., Parmigiani, G.: Integrating diverse genomic data using gene sets. Genome Biology **12**(10), 105 (2011). doi:10.1186/gb-2011-12-10-r105

[9] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B: Statistical Methodology **68**(1), 49–67 (2006). doi:10.1111/j.1467-9868.2005.00532.x

[10] Liquet, B., De Micheaux, P.L., Hejblum, B.P., Thiébaut, R.: Group and sparse group partial least square approaches applied in genomics context. Bioinformatics **32**(1), 35–42 (2016). doi:10.1093/bioinformatics/btv535

[11] Spector, T.D., Williams, F.M.K.: The UK Adult Twin Registry (TwinsUK). Twin Research and Human Genetics **9**(6), 899–906 (2006). doi:10.1375/twin.9.6.899

[12] Moayyeri, A., Hammond, C.J., Hart, D.J., Spector, T.D.: The UK adult twin registry (twinsUK resource). Twin Research and Human Genetics **16**(1), 144–149 (2013). doi:10.1017/thg.2012.89

[13] Wahl, A., Kasela, S., Carnero-Montoro, E., van Iterson, M., Štambuk, J., Sharma, S., van den Akker, E., Klaric, L., Benedetti, E., Razdorov, G., Trbojević-Akmačić, I., Vučković, F., Ugrina, I., Beekman, M., Deelen, J., van Heemst, D., Heijmans, B.T., B.I.O.S. Consortium, Wuhrer, M., Plomp, R., Keser, T., Šimurina, M., Pavić, T., Gudelj, I., Krištić, J., Grallert, H., Kunze, S., Peters, A., Bell, J.T., Spector, T.D., Milani, L., Slagboom, P.E., Lauc, G., Gieger, C.: IgG glycosylation and DNA methylation are interconnected with smoking. Biochimica et Biophysica Acta (BBA) - General Subjects **1862**(3), 637–648 (2018). doi:10.1016/J.BBAGEN.2017.10.012

[14] Zoldoš, V., Grgurević, S., Lauc, G.: Epigenetic regulation of protein glycosylation. BioMolecular Concepts **1**(3-4), 253–261 (2010). doi:10.1515/bmc.2010.027

[15] Lauc, G., Vojta, A., Zoldoš, V.: Epigenetic regulation of glycosylation is the quantum mechanics of biology. Biochimica et Biophysica Acta (BBA) - General Subjects **1840**(1), 65–70 (2014). doi:10.1016/J.BBAGEN.2013.08.017

[16] Wahl, A., van den Akker, E., Klaric, L., Štambuk, J., Benedetti, E., Plomp, R., Razdorov, G., Trbojević-Akmačić, I., Deelen, J., van Heemst, D., Eline Slagboom, P., Vučković, F., Grallert, H., Krumsiek, J., Strauch, K., Peters, A., Meitinger, T., Hayward, C., Wuhrer, M., Beekman, M., Lauc, G., Gieger, C.: Genome-wide association study on immunoglobulin G glycosylation patterns. Frontiers in Immunology **9**(FEB) (2018). doi:10.3389/fimmu.2018.00277

[17] Sofer, T., Schifano, E.D., Hoppin, J.A., Hou, L., Baccarelli, A.A.: A-clustering: A novel method for the detection of co-regulated methylation regions, and regions associated with exposure. Bioinformatics **29**(22), 2884–2891 (2013). doi:10.1093/bioinformatics/btt498

[18] WOLD, H.: Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. Multivariate Analysis–III, 383–407 (1973). doi:10.1016/B978-0-12-426653-7.50032-6

[19] Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3), 515–534 (2009). doi:10.1093/biostatistics/kxp008

[20] Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996). doi:10.1111/j.2517-6161.1996.tb02080.x

[21] Hastie, T., Tibshirani, R., Wainwright, M.: Statistical learning with sparsity: The lasso and generalizations. Statistical Learning with Sparsity: The Lasso and Generalizations **84**(1), 1–337 (2015). doi:10.1201/b18401

[22] Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Technical report (2010). `http://genomebiology.com/2010/11/3/R25`

[23] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America **102**(43), 15545–15550 (2005). doi:10.1073/pnas.0506580102

[24] Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G.: ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research **37**(SUPPL. 2) (2009). doi:10.1093/nar/gkp427

[25] Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (1995). doi:10.1111/j.2517-6161.1995.tb02031.x

[26] Storey, J.D.: A direct approach to false discovery rates. Technical Report 3 (2002). doi:10.1111/1467-9868.00346

[27] Gao, J., Collyer, J., Wang, M., Sun, F., Xu, F.: Genetic dissection of hypertrophic cardiomyopathy with myocardial rna-seq. International Journal of Molecular Sciences **21**(9) (2020). doi:10.3390/ijms21093040

[28] Chen, Y.-a.A., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., Weksberg, R.: Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics **8**(2), 203–209 (2013). doi:10.4161/epi.23470

[29] Uh, H.-W., Klarić, L., Ugrina, I., Lauc, G., Smilde, A.K., Houwing-Duistermaat, J.J.: Choosing proper normalization is essential for discovery of sparse glycan biomarkers. Molecular Omics (2020). doi:10.1039/c9mo00174c

[30] Bishop, C.M., Tipping, M.E.: Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society. Series B **61**(iii), 611–622 (1999). doi:10.1111/1467-9868.00196

[31] el Bouhaddani, S., Uh, H.W., Hayward, C., Jongbloed, G., Houwing-Duistermaat, J.: Probabilistic partial least squares model: Identifiability, estimation and application. Journal of Multivariate Analysis **167**, 331–346 (2018). doi:10.1016/j.jmva.2018.05.009. 1706.03597

7 Additional Files

Additional file 1

This pdf contains the details of solving the optimization problem (7).