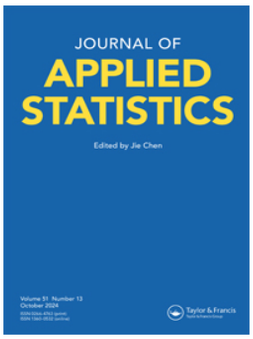


Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cjas20

Joint modeling of an outcome variable and integrated omics datasets using GLM-PO2PLS

Zhujie Gu, Hae-Won Uh, Jeanine Houwing-Duistermaat & Said el Bouhaddani

To cite this article: Zhujie Gu, Hae-Won Uh, Jeanine Houwing-Duistermaat & Said el Bouhaddani (2024) Joint modeling of an outcome variable and integrated omics datasets using GLM-PO2PLS, Journal of Applied Statistics, 51:13, 2627-2651, DOI: 10.1080/02664763.2024.2313458

To link to this article: https://doi.org/10.1080/02664763.2024.2313458

JOURNAL OF APPLIED STATISTICS
2024, VOL. 51, NO. 13, 2627–2651
https://doi.org/10.1080/02664763.2024.2313458

Taylor & Francis
Taylor & Francis Group

Joint modeling of an outcome variable and integrated omics datasets using GLM-PO2PLS

Zhujie Gu [a,b], Hae-Won Uh[a], Jeanine Houwing-Duistermaat [a,c,d] and Said el Bouhaddani[a]

[a]Department of Data Science and Biostatistics, Julius Centre, UMC Utrecht, Utrecht, The Netherlands; [b]Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, UK; [c]Department of Statistics, University of Leeds, Leeds, UK; [d]Department of Mathematics, Radboud University, Nijmegen, The Netherlands

ABSTRACT

In many studies of human diseases, multiple omics datasets are measured. Typically, these omics datasets are studied one by one with the disease, thus the relationship between omics is overlooked. Modeling the joint part of multiple omics and its association to the outcome disease will provide insights into the complex molecular base of the disease. Several dimension reduction methods which jointly model multiple omics and two-stage approaches that model the omics and outcome in separate steps are available. Holistic one-stage models for both omics and outcome are lacking. In this article, we propose a novel one-stage method that jointly models an outcome variable with omics. We establish the model identifiability and develop EM algorithms to obtain maximum likelihood estimators of the parameters for normally and Bernoulli distributed outcomes. Test statistics are proposed to infer the association between the outcome and omics, and their asymptotic distributions are derived. Extensive simulation studies are conducted to evaluate the proposed model. The method is illustrated by modeling Down syndrome as outcome and methylation and glycomics as omics datasets. Here we show that our model provides more insight by jointly considering methylation and glycomics.

ARTICLE HISTORY
Received 18 July 2022
Accepted 23 January 2024

KEYWORDS
Dimension reduction; PLS methods; multiple omics; generalized linear models; data integration

1. Introduction

The biological mechanisms underlying human diseases are often complex. Diverse omics datasets represent various aspects of these mechanisms. Recent advances in high-throughput technologies have made it affordable to measure these omic levels for many studies. Typically, these datasets are studied one-by-one. A good example is the analysis of genomic data in more than 5700 Genome-Wide Association Studies (GWAS) conducted to identify the genetic risk variants associated with more than 3000 traits and human diseases [58,62]. Other examples include studies of methylation data to pinpoint differentially

CONTACT Zhujie Gu zhujie.gu@mrc-bsu.cam.ac.uk MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02664763.2024.2313458.

methylated regions of DNA as indicators of many diseases [39,47], and studies of glycomic data to gain insight into the role of post-translational modification of proteins in disease pathways [53,54]. Although these studies provided biological insights of diseases on a single omic level, they ignored correlations among the omic levels. Analyzing multiple linked omics datasets jointly can bring further insights into the biological system underlying diseases. In this paper, we propose a new model for two omics datasets and an outcome variable, where the relationship of the omics datasets with the outcome is modeled via the joint parts of the omics datasets.

Our motivating dataset comes from a family-based case-control study of Down syndrome (DS). DS is the most frequent genomic aneuploidy with an incidence of approximately 1 in 700 live-newborn [41], caused by the trisomy of all or part of chromosome 21 (trisomy 21). Studies at the molecular level of DS have reported several alterations in methylation [4,9,17,19] and glycomics [6,10,17]. These alternations are mainly discovered by testing the mean difference of a single CpG site or glycan between the DS subjects and healthy controls. Furthermore, these studies were conducted on each omic level separately, overlooking the influence of methylation on glycosylation [61]. We aim to jointly model DS in terms of both omics, and investigate whether the molecules involved in the relationship between methylation and glycomics are related to DS.

One way to model multiple omics data and an outcome is to employ penalized linear regression such as ridge [24] and lasso [55]. Applying these methods to stacked omics data is found to lead to inferior performance compared to using only one of the omics datasets [45,56]. In contrast, IPF-LASSO [7] uses an omic-specific penalty. However, these approaches do not take into consideration the correlation structure between omics.

On the other hand, several methods are available that model the correlation between omics, but not the outcome. They decompose the omics data into joint and residual parts [35]. When fitted to multi-omics data that are heterogeneous in dimensionality, scale, and measurement platform, it has been shown that modeling omic-specific part simultaneously with the joint part improves estimation and interpretability of the joint components [13,15,18,31,33,57]. See [15] for an overview of these methods. To model an outcome variable, a second step is needed. JIVE-prediction [28] based on JIVE [33] and two-stage PO2PLS [21] based on PO2PLS [15] are examples of such two-stage approaches. However, two-stage approaches do not use the information in the outcome to guide the dimension reduction, hence provide less insight [23]. Furthermore, the error from the first stage is not taken into account in the second stage, leading to incorrect inference of the association between the outcome and the omics [50].

In the recent literature, several one-stage methods have been proposed where the outcome is directly modeled with multi-omics data. JACA [64] combines CCA with linear discriminant analysis (LDA) to find the linear combinations of features that lead to best classification of the outcome. Cooperative learning [12] was developed based on collaborative regression [20], searching for one linear combination from each dataset that jointly minimizes a loss function regarding the outcome, while keeping the prediction from each dataset close. Neither of the two models omic-specific variations, hampering the interpretation of the estimated joint components for heterogeneous omics data.

In this paper, we propose a new model GLM-PO2PLS, which extends PO2PLS by including an outcome variable in the model next to the omics datasets. The relationship between two omics data is modeled by joint and omic-specific latent variables to deal

with possible heterogeneity. The joint latent variables are linked to an outcome variable by a generalized linear model. Unlike the two-stage approach [21], the model is estimated simultaneously. We develop EM algorithms to obtain maximum likelihood estimators of the parameters for normally and Bernoulli distributed outcomes. The relationship between the outcome variable and the omics and that between two omics datasets can be inferred. The code is available on GitHub (github.com/zhujiegu/GLM-PO2PLS).

The rest of the paper is organized as follows. In Section 2, the PO2PLS model is recapped, and the GLM-PO2PLS model is formulated. The EM algorithms to estimate its parameters are proposed. Also, two chi-square tests of the relationship between outcome and both omics are proposed. In Section 3, the performance of GLM-PO2PLS is studied for a range of simulation scenarios where the focus is on parameter estimation and outcome prediction performance. In Section 4, we apply GLM-PO2PLS to the motivating DS datasets. We conclude with a discussion.

2. Methods

The GLM-PO2PLS was developed based on PO2PLS model which has been described in detail elsewhere [16]. Briefly, let x and y be two random row-vectors of dimensions p and q, respectively. In PO2PLS, x and y are decomposed into joint (t and u of size K), specific ($t_\perp$ and $u_\perp$ of size K_x resp. K_y) and residual (e and f of size p resp. q) parts. Heterogeneity between the joint parts is represented by an additional random vector h. The PO2PLS model is written as

$$x = tW^\top + t_\perp W_\perp^\top + e, \quad y = uC^\top + u_\perp C_\perp^\top + f, \quad u = tB + h,$$

where $W\,(p \times K)$ and $C\,(q \times K)$ are the loading matrices for the joint spaces of x and y respectively and $W_\perp\,(p \times K_x)$ and $C_\perp\,(q \times K_y)$ are the loading matrices for the specific parts of x and y respectively. The $K \times K$ diagonal matrix B models the relationship between the joint components t and u. With regard to the random vectors, we assume that t, $t_\perp$, $u_\perp$, h are zero mean multivariate normally distributed, with diagonal covariance matrices Σ_t, $\Sigma_{t_\perp}$, $\Sigma_{u_\perp}$, Σ_h, respectively. Since $u = tB + h$, the covariance matrix of u is $\Sigma_u = B^\top \Sigma_t B + \Sigma_h$. The residual random vectors e, and f are independent normally distributed, with zero mean and respective diagonal covariance matrices, $\sigma_e^2 I_p$, and $\sigma_f^2 I_q$, where I_p and I_q are identity matrices of size p and q. The number of components K, K_x, and K_y can be determined based on the scree plots of the eigenvalues of $x^\top y$, $x^\top x$ and $y^\top y$, or based on a cross-validation (CV) procedure [13]. Note that the PO2PLS model is asymmetrical, where y is modeled in terms of x. In most cases, it is reasonable to assume that the effect goes from one omics to another (e.g. from methylation to glycomics in the Down syndrome data), therefore an asymmetrical model better reflects the underlying biology.

2.1. The GLM-PO2PLS model

GLM-PO2PLS jointly models an outcome variable z with two omics datasets x and y, where it is assumed that the effect of x and of y on z is solely through the joint parts of x and y.

Using the same notations as in the PO2PLS model, the GLM-PO2PLS model is given by

$$x = tW^\top + t_\perp W_\perp^\top + e, \quad y = uC^\top + u_\perp C_\perp^\top + f, \quad u = tB + h,$$
$$\eta(\mathbb{E}[z]) = a_0 + ta^\top + ub^\top, \tag{1}$$

with a_0 the intercept, a and b both row-vectors of size r and η the link function which links the outcome z to the linear predictor $a_0 + ta^\top + ub^\top$. The equations for x and y follow the PO2PLS model. Since the joint latent variables t and u are linked to x, y, and z, GLM-PO2PLS jointly models the outcome and two omics.

Now, u is a linear function of t, namely $u = tB + h$. Hence, the model for z in (1) can equivalently be written in terms of t and the h (the part in u independent of t), i.e.

$$\eta(\mathbb{E}[z]) = a_0 + ta^\top + (tB + h)b^\top = a_0 + t\tilde{a}^\top + h\tilde{b}^\top, \tag{2}$$

where $\tilde{a} = a + Bb^\top$ and $\tilde{b} = b$. With this equivalent parametrization, instability due to near collinearity in the linear predictor of z is reduced. The coefficient $\tilde{a}$ models the total effect of x and $\tilde{b}$ models the direct effect of y on z.

In the remainder of the paper, we use the rightmost form in (2) and omit the tildes on a and b.

2.2. The GLM-PO2PLS model with a normally distributed outcome

In this subsection, we first consider a continuous outcome z. The details for a binary z is then given in the next subsection. As link function, we use the identity, $\eta(\nu) = \nu$. We assume that the outcome is centered and since t and h have zero-mean, the intercept a_0 can be omitted. We assume that the residual $g = z - ta^\top - hb^\top$ is normally distributed, $g \sim \mathcal{N}(0, \sigma_g^2)$. Since (x, y, z) is linearly dependent on $(t, u, t_\perp, u_\perp, e, f, h, g)$, it follows a multivariate normal distribution $\mathcal{N}(0, \Sigma_\theta)$, with a covariance matrix given by

$$\Sigma_\theta = \begin{bmatrix} W\Sigma_t W^\top + W_\perp \Sigma_{t_\perp} W_\perp^\top + \sigma_e^2 I_p & W\Sigma_t BC^\top & W\Sigma_t a^\top \\ CB\Sigma_t W^\top & C\Sigma_u C^\top + C_\perp \Sigma_{u_\perp} C_\perp^\top + \sigma_f^2 I_q & C(\Sigma_h b^\top + B\Sigma_t a^\top) \\ a\Sigma_t W^\top & (a\Sigma_t B + b\Sigma_h)C^\top & a\Sigma_t a^\top + b\Sigma_h b^\top + \sigma_g^2 \end{bmatrix}, \tag{3}$$

where $\theta = \{W, C, W_\perp, C_\perp, a, b, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \sigma_e^2, \sigma_f^2, \Sigma_h, \sigma_g^2\}$ is the collection of GLM-PO2PLS model parameters.

Identifiability of GLM-PO2PLS. Latent variable models are typically unidentifiable due to rotation indeterminacy of the loading components. In PO2PLS, identifiability up to sign has been shown under mild conditions [16]. Namely, the loading matrices are semi-orthogonal, i.e. $W^\top W = C^\top C = I_K$, $W_\perp^\top W_\perp = I_{K_x}$, and $C_\perp^\top C_\perp = I_{K_y}$. Additionally, matrices $[WW_\perp]$ and $[CC_\perp]$ do not have linearly dependent columns. Furthermore, the covariance matrices for the latent variables Σ_t, Σ_u, $\Sigma_{t_\perp}$, $\Sigma_{u_\perp}$ are diagonal. Finally, the diagonal elements of B are positive and the diagonal elements of $\Sigma_t B$ are strictly decreasing.

We show that these conditions also guarantee the identifiability (up to sign) of the GLM-PO2PLS model.

Theorem 2.1: *Let (x, y, z) follow the model assumptions of the GLM-PO2PLS model where z is normally distributed. Additionally, let the parameters satisfy the PO2PLS conditions as described above. It follows that the GLM-PO2PLS model parameters are identifiable up to a sign.*

Proof: Let $f(x, y, z|\theta) = f(x, y, z|\tilde{\theta})$ be identical joint distributions under two sets of parameters θ and $\tilde{\theta}$. Then we necessarily have $f(x, y|\theta) = f(x, y|\tilde{\theta})$. Since $(x, y|\theta)$ follows a zero mean multivariate normal distribution, its distribution is uniquely defined by the covariance matrix $\Sigma_{x,y|\theta}$. Thus $\Sigma_{x,y|\theta} = \Sigma_{x,y|\tilde{\theta}}$ follows. It has been proven in [16] that if $\Sigma_{x,y|\theta} = \Sigma_{x,y|\tilde{\theta}}$ holds, then the parameters involved (i.e. $\{W, C, W_\perp, C_\perp, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \sigma_e^2, \sigma_f^2, \Sigma_h\}$) are identified, up to sign.

For a normally distributed z, the random vector (x, y, z) follows a zero mean multivariate normal distribution, and its distribution is uniquely defined by the covariance matrix Σ_θ in (3). It follows from $f(x, y, z|\theta) = f(x, y, z|\tilde{\theta})$ that $\Sigma_\theta = \Sigma_{\tilde{\theta}}$. Now let $a\Sigma_t W^\top = \tilde{a}\tilde{\Sigma}_t \tilde{W}^\top$. Since $\Sigma_t W^\top = \tilde{\Sigma}_t \tilde{W}^\top$ and is of full rank, we have $a = \tilde{a}$. Similarly, we have $b = \tilde{b}$, and $\sigma_g^2 = \tilde{\sigma_g^2}$ from $b\Sigma_h C^\top = \tilde{b}\tilde{\Sigma}_h \tilde{C}^\top$ and $a\Sigma_t a^\top + b\Sigma_h b^\top + \sigma_g^2 = \tilde{a}\tilde{\Sigma}_t \tilde{a}^\top + \tilde{b}\tilde{\Sigma}_h \tilde{b}^\top + \tilde{\sigma_g^2}$, respectively. This shows identifiability of all the parameters in θ. ■

2.2.1. Maximum likelihood estimation

Since the GLM-PO2PLS model is a latent variable model and the likelihood factorizes in terms which can be maximized separately, we propose an EM algorithm [11] to obtain maximum likelihood estimates of the model parameters.

Suppose we observe (x, y, z) for N subjects. Since we assume a multivariate normal distribution of $(x, y, z) \sim \mathcal{N}(0, \Sigma_\theta)$, the log-likelihood for one subject is given by

$$\ell(\theta; x, y, z) = -\frac{1}{2}\{(p + q + 1)\log(2\pi) + \log|\Sigma_\theta| + (x, y, z)\Sigma_\theta^{-1}(x, y, z)^\top\}.$$

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

$$Q(\theta|\theta') = \mathbb{E}[\log f(x, y, z, t, u, t_\perp, u_\perp|\theta)|x, y, z, \theta'].$$

Expectation step. The conditional expectation of the complete data log likelihood can be decomposed into different terms,

$$\begin{aligned} Q(\theta|\theta') &= \mathbb{E}[\log f(x, y, z, t, u, t_\perp, u_\perp)] = \mathbb{E}[\log f(x, y, z|t, u, t_\perp, u_\perp)] \\ &\quad + \mathbb{E}[\log f(t, u, t_\perp, u_\perp)] \\ &= \underbrace{\mathbb{E}[\log f(z|t, u)]}_{Q_{\{a,b,\sigma_g^2\}}} + \underbrace{\mathbb{E}[\log f(x|t, t_\perp)]}_{Q_{\{W,W_\perp,\sigma_e^2\}}} + \underbrace{\mathbb{E}[\log f(y|u, u_\perp)]}_{Q_{\{C,C_\perp,\sigma_f^2\}}} \\ &\quad + \underbrace{\mathbb{E}[\log f(u|t)]}_{Q_{\{B,\Sigma_h\}}} + \underbrace{\mathbb{E}[\log f(t)]}_{Q_{\Sigma_t}} + \underbrace{\mathbb{E}[\log f(t_\perp)]}_{Q_{\Sigma_{t_\perp}}} + \underbrace{\mathbb{E}[\log f(u_\perp)]}_{Q_{\Sigma_{u_\perp}}}. \end{aligned} \tag{4}$$

In this equation, the conditioning on x, y, z and θ' is dropped, to simplify notation. The individual conditional expectations depend on distinct sets of parameters, yielding separate optimization tasks. Compared to PO2PLS, the extra parameters in GLM-PO2PLS $\{a, b, \sigma_g^2\}$ are included in the first term $Q_{\{a,b,\sigma_g^2\}}$. Therefore, we focus on the optimization of $Q_{\{a,b,\sigma_g^2\}}$ with respect to $\{a, b, \sigma_g^2\}$. The rest of the terms are identical to the factorized densities in the original PO2PLS EM algorithm, we refer to the PO2PLS paper [16] for the expectation and maximization regarding these terms.

In the expectation step, $Q_{\{a,b,\sigma_g^2\}}$ is calculated as

$$Q_{\{a,b,\sigma_g^2\}} = -\frac{1}{2}\left\{\log(2\pi\sigma_g^2) + \frac{1}{\sigma_g^2}\,\mathrm{tr}\,\mathbb{E}\left[(z - ta^\top - (u - tB)b^\top)^\top(z - ta^\top - (u - tB)b^\top)\right]\right\}. \quad (5)$$

Here, the first and second conditional moments of the vector (t, u) given x, y, z and θ' are involved. Since $(x, y, z, t, u, t_\perp, u_\perp)$ follows a multivariate normal distribution with zero mean and known covariance matrix, the conditional density $f(t, u, t_\perp, u_\perp|x, y, z)$ can be calculated following Lemma 3 in [16]. The conditional moments involved in (5) can then be obtained from the mean and the covariance matrix of $(t, u, t_\perp, u_\perp|x, y, z)$ (see the Supplementary material for details).

Maximization step. In the maximization (M) step, each conditional expectation in (4) can be optimized separately. Here, we restrict to the description of the term involving the outcome, namely, maximize the $Q_{\{a,b,\sigma_g^2\}}$ as given in Equation (5). Note that the coefficient vector (a, b) can be separately optimized from the residual parameter σ_g^2, as in the standard linear regressions. We first calculate the derivative with respect to (a, b) and set it to 0, yielding

$$\frac{\partial Q_{\{a,b,\sigma_g^2\}}}{\partial(a, b)} = 0 \Rightarrow (\hat{a}, \hat{b}) = z^\top\,\mathbb{E}[(t, h)]\,\mathbb{E}[(t, h)^\top(t, h)]^{-1}.$$

where the conditional moments are calculated in the E step. The maximization with respect to the parameter σ_g^2 can then be performed similarly. Details are given in the supplementary material.

2.2.2. Statistical inference

The GLM-PO2PLS method allows for statistical inference on the relationship between the omic data and the outcome. This relationship is captured by t and h, and given by the equation $\eta(\mathbb{E}[z]) = ta^\top + hb^\top$ in (1). Here, we propose two tests, one full test for the relationship between z and all the joint components together, and one component-wise test for the relationship between z and each pair of joint components.

For the full test, we consider the null hypothesis,

$$H_0 : a = b = \mathbf{0} \quad \text{against } H_1 : a \neq \mathbf{0} \text{ or } b \neq \mathbf{0}.$$

For each component-wise test, we consider the null hypothesis of no relationship between z and the k-th ($k = 1, \ldots, K$) pair of joint components,

$$H_0 : a_k = b_k = 0 \quad \text{against } H_1 : a_k \neq 0 \text{ or } b_k \neq 0.$$

where a_k and b_k are the coefficients for t_k and h_k, respectively.

Let $\alpha = (a, b)$ and $\alpha_k = (a_k, b_k)$. The full test statistic is given by

$$T_{full} = \hat{\alpha}\Pi_{\hat{\alpha}}^{-1}\hat{\alpha}^\top, \tag{6}$$

where $\Pi_{\hat{\alpha}}^{-1}$ is the inverse of the covariance matrix of $\hat{\alpha}$. And the pair-wise test statistic is given by:

$$T_{comp.wise} = \hat{\alpha_k}\Pi_{\hat{\alpha_k}}^{-1}\hat{\alpha_k}^\top. \tag{7}$$

To calculate the (asymptotic) distribution of these test statistics, the asymptotic distribution of all parameters θ needs to be derived.

Asymptotic distribution. Under certain regularity conditions, consistency of the estimator θ and its asymptotic distribution $\mathcal{N}(\theta, \Pi_\theta)$ follows from Shapiro's Proposition 4.2 [46] applied to the GLM-PO2PLS model.

Theorem 2.2: *Let $\hat{\theta}$ be the maximum likelihood estimator for θ from the GLM-PO2PLS model. When the sample size N approaches infinity, the distribution of $\hat{\theta}$ converges to a normal distribution, i.e.*

$$N^{1/2}(\hat{\theta} - \theta) \longrightarrow \mathcal{N}(0, \Pi_\theta)$$

Details and proofs are given in the supplement.

In particular, $\hat{\alpha} = (\hat{a}, \hat{b})$ is asymptotically normally distributed. Therefore, the test statistics T_{full} and $T_{comp.wise}$ follow a chi-square distribution with $2r$ resp. 2 degrees of freedom. An estimate of Π_θ is obtained from the inverse observed Fisher information matrix. Let ψ_i be an instance of observed data (x, y, z) and ζ_i be the latent variables involved. In an EM algorithm, this matrix is given by [34],

$$\mathcal{I}(\hat{\theta}) = \sum_{i=1}^{N} \mathbb{E}[B_i(\hat{\theta})|\psi_i] - \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[S_i(\hat{\theta})S_j(\hat{\theta})^\top|\psi_i; \psi_j]$$

where $S_i(\hat{\theta}) = \nabla l(\hat{\theta}; \psi_i, \zeta_i)$ and $B_i(\hat{\theta}) = -\nabla^2 l(\hat{\theta}; \psi_i, \zeta_i)$ are the gradient and negative second derivative of the log complete likelihood of instance i, respectively, evaluated at $\hat{\theta}$.

To obtain $\Pi_{\hat{\alpha}}$, the submatrix of $\mathcal{I}^{-1}(\hat{\theta})$ corresponding to $\hat{\alpha}$ (denote $\mathcal{I}^{-1}(\hat{\alpha})$) has to be calculated. However, inverting $\mathcal{I}(\hat{\theta})$ is computationally infeasible, even for moderate dimensions. Under additional assumptions that $\hat{\alpha}$ is asymptotically independent from the rest of the parameters and $\hat{\sigma}_g^2$ is non-random, $\mathcal{I}^{-1}(\hat{\alpha})$ can be calculated, and be used to approximate $\Pi_{\hat{\alpha}}$. The details are given in supplementary materials.

2.3. The GLM-PO2PLS model with a binary outcome

For a binary outcome, we use a Bernoulli distribution for z and the logit link function $\eta(\nu) = \text{logit}(\nu) = \log[\nu(1-\nu)^{-1}]$. The model is then given by

$$x = tW^\top + t_\perp W_\perp^\top + e, \quad y = uC^\top + u_\perp C_\perp^\top + f, \quad u = tB + h,$$
$$\text{logit}(p(z)) = a_0 + ta^\top + hb^\top.$$

Here, $p(z) = \Pr(z = 1|t, h)$ is the conditional probability of the random variable z being 1, given t and h. Note that the probability $p(z)$ is logit-normally distributed, therefore the linear predictor $\text{logit}(p(z))$ follows a normal distribution $\mathcal{N}(a_0, a\Sigma_t a^\top + b\Sigma_h b^\top)$. The joint distribution $(x, y, \text{logit}(p(z)))$ is multivariate normal with mean vector $(0_{p+q}, a_0)$ and covariance matrix Σ_θ in (3) excluding σ_g^2. The collection of parameters in the GLM-PO2PLS model with a binary outcome is $\theta = \{W, C, W_\perp, C_\perp, a_0, a, b, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \sigma_e^2, \sigma_f^2, \Sigma_h\}$.

Identifiability of GLM-PO2PLS with a binary outcome. Theorem 2.1 also appears to hold for a binary z that follows a Bernoulli distribution, under the same conditions. The proof is similar. Specifically, let $f(x, y, z|\theta) = f(x, y, z|\tilde{\theta})$ be identical joint distributions under two sets of parameters θ and $\tilde{\theta}$. Then $f(x, y|\theta) = f(x, y|\tilde{\theta})$, thus $\Sigma_{x,y|\theta} = \Sigma_{x,y|\tilde{\theta}}$ holds regardless of the distribution of z. The conclusion follows that the parameters involved in PO2PLS model (i.e. $\{W, C, W_\perp, C_\perp, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \sigma_e^2, \sigma_f^2, \Sigma_h\}$) are identified up to sign. Now consider $(x, y, \text{logit}(p(z)))$ which is multivariate normally distributed with mean vector $(0_{p+q}, a_0)$ and covariance matrix Σ_θ excluding σ_g^2 (denote Σ_{θ/g^2}). Since the mapping $f(x, y, z|\theta) \mapsto f(x, y, \text{logit}(p(z))|\theta)$ is one-to-one, it follows that $f(x, y, \text{logit}(p(z))|\theta) = f(x, y, \text{logit}(p(z))|\tilde{\theta})$. Necessarily, the means and covariance matrices of two identical multivariate normal distributions are equivalent, thus $(0_{p+q}, a_0) = (0_{p+q}, \tilde{a}_0)$ and $\Sigma_{\theta/g^2} = \Sigma_{\tilde{\theta}/g^2}$. It is clear that $a_0 = \tilde{a}_0$ from the equivalence of the mean vectors. The identifiability of a and b can be shown from the equivalence of covariance matrices analogously as in the proof of Theorem 2.1. This shows the identifiability of all the parameters in GLM-PO2PLS with a binary outcome.

2.3.1. EM algorithm for a binary outcome

For a Bernoulli distributed outcome, the log-likelihood of the observed data involves an integral of dimension $2K + K_x + K_y$. Let $\nu = (t, u)$ and $\xi = (t_\perp, u_\perp)$,

$$\ell(\theta; x, y, z) = \log \int_{(\nu,\xi)} f(x, y, z|\nu, \xi, \theta) f(\nu, \xi|\theta)\, \mathrm{d}(\nu, \xi). \tag{8}$$

To estimate (8), numerical integration is needed. Note that given ν, the binary outcome z is independent of x, y and ξ, thus the conditional density $f(x, y, z|\nu, \xi)$ in (8) can be factorized as $f(x, y, z|\nu, \xi) = p(z|\nu) f(x, y|\nu, \xi)$. The factorization enables to integrate out the specific random vector ξ, hence reducing the dimension of the integral to $2K$,

$$\begin{aligned}
\ell(\theta; x, y, z) &= \log \int_{(\nu,\xi)} p(z|\nu) f(x, y|\nu, \xi) f(\nu, \xi)\, \mathrm{d}(\nu, \xi) \\
&= \log \int_\nu p(z|\nu) \left[\int_\xi f(x, y|\nu, \xi) f(\xi|\nu)\, \mathrm{d}\xi \right] f(\nu)\, \mathrm{d}\nu \\
&= \log \int_\nu p(z|\nu) f(x, y|\nu) f(\nu)\, \mathrm{d}\nu \\
&= \log \int_\nu p(z|\nu) f(x|\nu) f(y|\nu) f(\nu)\, \mathrm{d}\nu.
\end{aligned}$$

Here, the probability mass function $p(z|\nu)$ is given by

$$p(z|\nu) = \begin{cases} \left(1 + \exp\{-(a_0 + ta^\top + (u - tB)b^\top)\}\right)^{-1} & z = 1, \\ \left(1 + \exp\{a_0 + ta^\top + (u - tB)b^\top\}\right)^{-1} & z = 0. \end{cases} \tag{9}$$

The probability density functions $f(x|\nu)$, $f(y|\nu)$, and $f(\nu)$ follow from the following multivariate normal distributions,

$$x|\nu \sim \mathcal{N}(tW^\top, \Sigma_{x|t}), \quad y|\nu \sim \mathcal{N}(uC^\top, \Sigma_{y|u}), \quad \nu \sim \mathcal{N}(0, \Sigma_\nu)$$

where the covariance matrices involved are:

$$\Sigma_{x|t} = W_\perp \Sigma_{t_\perp} W_\perp^\top + \sigma_e^2 I_p, \quad \Sigma_{y|u} = C_\perp \Sigma_{u_\perp} C_\perp^\top + \sigma_f^2 I_q, \quad \Sigma_\nu = \begin{bmatrix} \Sigma_t & \Sigma_t B \\ B\Sigma_t & \Sigma_u \end{bmatrix}.$$

Denote the partial complete data vector by (x, y, z, ν). For each current estimate θ', the EM algorithm for a binary outcome considers the objective function

$$Q(\theta|\theta') = \mathbb{E}[\log f(x, y, z, \nu|\theta)|x, y, z, \theta']. \tag{10}$$

Expectation step based on numerical integration. Analogously to (4), the conditional expectation in (10) can be decomposed to factors that depend on distinct sets of parameters,

$$\begin{aligned} Q(\theta|\theta') &= \mathbb{E}[\log f(x, y, z, \nu)] = \mathbb{E}[\log f(x, y, z|\nu)] + \mathbb{E}[\log f(\nu)] \\ &= \underbrace{\mathbb{E}[\log p(z|\nu)]}_{Q_{\{a_0,a,b\}}} + \underbrace{\mathbb{E}[\log f(x|t)]}_{Q_{\{W,W_\perp,\sigma_e^2,\Sigma_{t_\perp}\}}} + \underbrace{\mathbb{E}[\log f(y|u)]}_{Q_{\{C,C_\perp,\sigma_f^2,\Sigma_{u_\perp}\}}} + \underbrace{\mathbb{E}[\log f(u|t)]}_{Q_{\{B,\Sigma_h\}}} + \underbrace{\mathbb{E}[\log f(t)]}_{Q_{\Sigma_t}}. \end{aligned} \tag{11}$$

Here, the first conditional expectation $Q_{\{a_0,a,b\}}$ has no closed form,

$$\begin{aligned} Q_{\{a_0,a,b\}} &= \int [\log p(z|\nu)] f(\nu|x, y, z, \theta') \, \mathrm{d}\nu \\ &= \frac{1}{f(x, y, z)} \int [\log p(z|\nu)] p(z|\nu) f(x, y|\nu) f(\nu) \, \mathrm{d}\nu. \end{aligned}$$

To obtain an approximation of the multivariate integral, Gauss–Hermite quadrature can be used. For an integral of form $\int \varphi(\nu) p(z|\nu) f(x, y|\nu) f(\nu) \, \mathrm{d}\nu$, where φ is any function, we approximate it with

$$\begin{aligned} &\int \varphi(\nu) p(z|\nu) f(x, y|\nu) f(\nu) \, \mathrm{d}\nu \\ &\quad \approx \sum_{m_1=1}^{M} \cdots \sum_{m_{2K}=1}^{M} \varphi(\nu = \nu_m) p(z|\nu = \nu_m) f(x, y|\nu = \nu_m) w_{m_1} \ldots w_{m_{2K}} \end{aligned} \tag{12}$$

with nodes vector $\nu_m = (\nu_{m_1}, \ldots \nu_{m_K}) = \sqrt{2}(\Sigma_\nu^{1/2})^\top \nu_m^*$ and weights vector $w_m = (w_{m_1}, \ldots w_{m_K}) = w_m^*/\sqrt{\pi}$. Here, M is the number of sampling nodes, $\Sigma_\nu^{1/2}$ is the Cholesky

decomposition of Σ_ν, and ν_m^* and w_m^* are nodes and weights of a M-point standard Gauss–Hermite quadrature rule, which can be found on page 924 in [1]. The transformation from the standard quadrature nodes ν_m^* to ν_m is to make the sampling range of the integrand in (12) more suitable based on the distribution of ν [32].

The other terms in (11) have explicit expressions in terms of the first and second conditional moments of the vector ν given x, y, z and θ' (see for details in the Supplementary materials). Note that the conditional moments of ν are in forms of integrals as follows

$$\mathbb{E}[\nu|x,y,z,\theta'] = \int \nu f(\nu|x,y,z)\,\mathrm{d}\nu = \frac{1}{f(x,y,z)}\int \nu p(z|\nu)f(x,y|\nu)f(\nu)\,\mathrm{d}\nu,$$
$$\mathbb{E}[\nu^\top\nu|x,y,z,\theta'] = \int \nu^\top\nu f(\nu|x,y,z)\,\mathrm{d}\nu = \frac{1}{f(x,y,z)}\int \nu^\top\nu p(z|\nu)f(x,y|\nu)f(\nu)\,\mathrm{d}\nu,$$

which can be numerically calculated with (12).

Maximization step based on gradient descent. Maximizing $Q_{\{a_0,a,b\}}$ requires iterations as its derivative with respect to $\beta = (a_0, a, b)$ has no analytical solutions. To find an update of β in each EM iteration, we propose a one-step gradient descent strategy. The gradient of Q_β is given by

$$\begin{aligned}\nabla Q_\beta &= \left[\frac{\partial Q_\beta}{\partial\beta}\right]^\top = \left[\frac{1}{f(x,y,z)} * \frac{\partial}{\partial\beta}\int [\log p(z|\nu)]p(z|\nu)f(x,y|\nu)f(\nu)\,\mathrm{d}\nu\right]^\top \\ &= \left[\frac{1}{f(x,y,z)}\int \frac{\partial \log p(z|\nu)}{\partial\beta}p(z|\nu)f(x,y|\nu)f(\nu)\,\mathrm{d}\nu\right]^\top\end{aligned}$$

To guarantee the increase of Q_β in each EM iteration, we search for a step size along the direction of the gradient using the backtracking rule (also known as the Armijo rule) [2]. It is performed by starting with an initial step size of $s = 1$ for movement along the gradient, and iteratively shrinking the step size ($s \leftarrow 0.8 * s$) until an increase of Q_β exceeds the expected increase based on the local gradient. More precisely, we keep shrinking the step size until the following ascent condition is met:

$$Q_{(\beta+s\nabla Q_\beta)} \geq Q_\beta + 0.5 * s\nabla Q_\beta \nabla Q_\beta^\top.$$

The maximization of the other conditional expectation terms in (11) can be found in the supplementary materials.

3. Simulation

We conduct a simulation study to evaluate the performance of GLM-PO2PLS. Both continuous outcome z_c and binary outcome z_b are investigated. The datasets are simulated following the GLM-PO2PLS model in (1), with the equations for the continuous and binary outcomes being $z_c = ta^\top + hb^\top + g$, and $z_b \sim \text{Bernoulli}((1 + \exp\{-(a_0 + ta^\top + hb^\top)\})^{-1})$.

3.1. Simulation settings

We consider combinations of small and large sample sizes ($N = 100, 1000$) with low and high dimensionalities ($p = 100, 2000; q = 10, 25$). The latent variables t, $t_\perp$, $u_\perp$ are simulated from standard normal distribution, and $u = tB + h$ following Equation (1). Here, B is the identity matrix and the joint residual h in u that is independent of t determines the level of heterogeneity in the joint parts. To investigate the impact of heterogeneity levels, we vary the variance of h to account for 40% and 80% of the total variance in u. The residual terms e, f are generated from zero-mean normal distributions. In the low noise level scenario, we set the noise proportion in x and y to both 40%. In the high noise level scenario, we investigate the performance of GLM-PO2PLS when integrating a very noisy large dataset and a less noisy small dataset, by increasing the noise in x to 95% and decreasing the noise in y to 5%. The noise term g for the continuous outcome is generated from a zero-mean normal distribution, accounting for 20% of variation in z_c. All the loading matrices are generated from standard normal distribution and then semi-orthogonalized. The coefficients a and b are set to 2 and 1, respectively. The number of joint and specific components is set to 1. For each setting, 500 replications are generated. The settings are summarized in Table 1.

The metrics used to assess the performance are listed in Table 2. We first study the estimation accuracy of the coefficients a and b. The errors $(\hat{a} - a)$ and $(\hat{b} - b)$ are standardized by a and b to exclude the influence of the parameter scale. For the continuous outcome, we evaluate the type I error and power of the chi-square test in (6). For type I error, we generate the outcome from a standard normal distribution independent of (x, y) and calculate the proportion of false positives (at significance level of 5%) in 10,000 repetitions. For power, we compute the empirical power as the proportion of true positives. The performance of outcome prediction is assessed by root mean square error of prediction (RMSEP), defined as $(\mathbb{E}[(\hat{z_c} - z_c)^2])^{\frac{1}{2}}$ for continuous outcome z_c, and $(\mathbb{E}[(\text{logit}(p(z_b)) - \text{logit}(\hat{p}(z_b)))^2])^{\frac{1}{2}}$ for binary outcome z_b. We compare the performance of GLM-PO2PLS with ridge regression fitted separately on x (denote ridge-x) and on y (denote ridge-y). The shrinkage hyper-parameter in ridge regressions is searched using a 10-fold cross-validation for each fit. The prediction performance is evaluated on an independent test dataset of size 1000. The accuracy of loading estimation is measured by the inner product between the estimated and the true loading vectors. The performance of feature selection is measured by true positives rate (TPR) calculated as the proportion of true top 25% features among the

Table 1. Summary of simulation settings.

Notations	Description	Setting/Distribution
N	Sample size	Small: 100 Large: 1000
$p; q$	Dimension of x, y	Low: 100,10 High: 2000,25
h	Heterogeneity between joint latent variables t and u	Normal Moderate: 40% of variance in u High: 80% of variance in u
e, f	Noise in x, y	Normal Low: 40%, 40% High: 95%, 5%

Table 2. Metrics of simulation.

Category	Metric	Calculation	Competing methods
Coefficient estimation	Scaled error	$(\hat{a}-a)/a, (\hat{b}-b)/b$	
Statistical inference	Type I error Empirical power	Proportion of false positives under null hypothesis Proportion of true positives	
Outcome prediction	RMSEP	$(\mathbb{E}[(\hat{z_c}-z_c)^2])^{\frac{1}{2}}$, $(\mathbb{E}[(\text{logit}(p(z_b))-\text{logit}(\hat{p}(z_b)))^2])^{\frac{1}{2}}$	ridge-x, ridge-y
Loading estimation	Inner product	$W^\top\hat{W}, W_\perp^\top\hat{W}_\perp, C^\top\hat{C}, C_\perp^\top\hat{C}_\perp$	
Feature selection	TPR of top 25%	TP/(TP+FN)	ridge-x

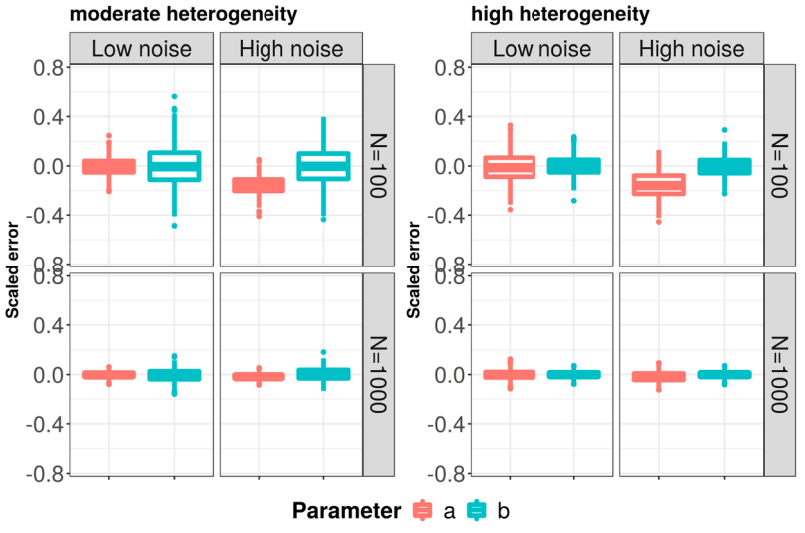
estimated top 25% in x (i.e. the top 25% of features in x with the largest absolute loading values in GLM-PO2PLS, or with the largest absolute regression coefficients in ridge regression). Additionally, we evaluate the estimation accuracy of a and b with 2 joint and 2 specific components, and the power under more noisy scenarios in supplementary materials.

3.2. Results of simulation study

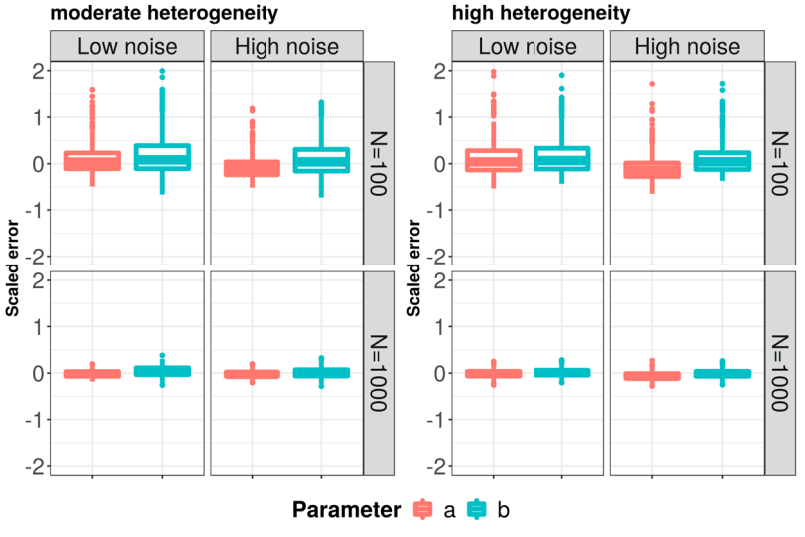
In Figure 1, results of the coefficient estimation in high-dimensional settings are depicted. Figure 1(a) shows that for the continuous outcome, overall, the scaled errors of both $\hat{a}$ and $\hat{b}$ were small. When the sample size was small and the noise was high, the scaled error $(\hat{a}-a)/a$ was mostly negative, suggesting that a was underestimated. For a large sample size, the estimators appeared to be unbiased. When the heterogeneity between the joint components was increased (from the left panel to the right), the joint residual h had larger variance relative to t and explained a larger proportion of z. Consequently, the estimation of the coefficient b (for h) became more stable, while the estimation of a (for t) became less stable. The results for a binary outcome are shown in Figure 1(b). Under a small sample size, the parameter estimation was less stable than the continuous case (note that the scale of y-axis in subplot (a) and (b) are different). The long upper whiskers suggested that the coefficients were overestimated in a some simulation runs. For a large sample size, the scaled errors for all coefficients were close to 0 and stable. Overall, the results for low dimensions were similar, except that the estimation of b was less stable in low dimensions compared to that in high dimensions. Details are given in the supplementary material.

Figure 2 shows the type I error in high-dimensional settings. With a small sample size of 100, the type I error was around 7–7.5%, slightly higher than 5%. When the sample size increased to 1000 and 10,000, the type I error decreased to 5%. The empirical power was 1 for all the scenarios.

Figure 3 shows the results regarding outcome prediction in high-dimensional settings. For the continuous outcome, GLM-PO2PLS outperformed both ridge-x and ridge-y as shown in Figure 3(a). The small boxes suggest that the prediction was very similar in each repetition, hence stable. Ridge-y performed similarly as GLM-PO2PLS, while ridge-x under-performed. When the noise in x was increased, the performance of ridge-x deteriorated, especially when the sample size was small. The larger noise proportion in x barely affected the performance of GLM-PO2PLS. Increasing the heterogeneity made the RMSEP of ridge-x higher, as x explained less variation in z, while the performance of GLM-PO2PLS was less affected. For the binary outcome z_b, GLM-PO2PLS still outperformed



(a) Continuous outcome z_c.



(b) Binary outcome z_b.

Figure 1. Performance of coefficient estimation for continuous (a) and binary (b) outcome. The y-axis shows the scaled estimation error as defined in Table 2. In the moderate and high heterogeneity settings, h account for 40 and 80% of total variance in $u = tB + h$, respectively. Boxes show the results of 500 repetitions. (a) Continuous outcome z_c and (b) Binary outcome z_b.

ridge regression as shown in Figure 3(b). When the sample size increased, the prediction of GLM-PO2PLS was less skewed and more stable. The conclusions also hold in low dimensions, details are given in the supplementary material.

The results for loading estimation and feature selection are given in the supplementary material. Overall, the loading estimates were accurate for both continuous and binary outcomes, with most inner products between the estimated and the true loadings approaching

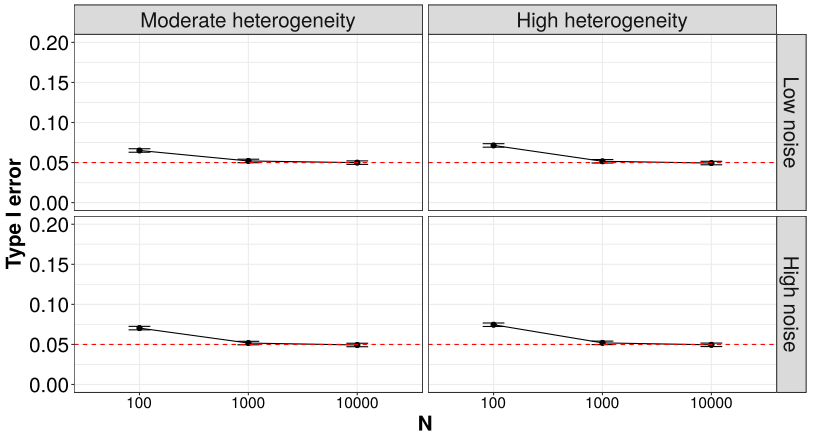


Figure 2. Type I error in high-dimensional settings. The error bars show the standard errors of the estimation. The dotted horizontal line is the significance level of 5%.

the optimum. When the sample size was small and the noise level was high, the accuracy of loading estimation for x dropped. This was the same setting in which $\hat{a}$ was biased as is shown in Figure 1(a). Regarding feature selection, the lowest median TPR of GLM-PO2PLS was 0.62 in the scenario with a small sample size, large noise proportion, and high heterogeneity. In the other scenarios, the median TPR was above 0.85.

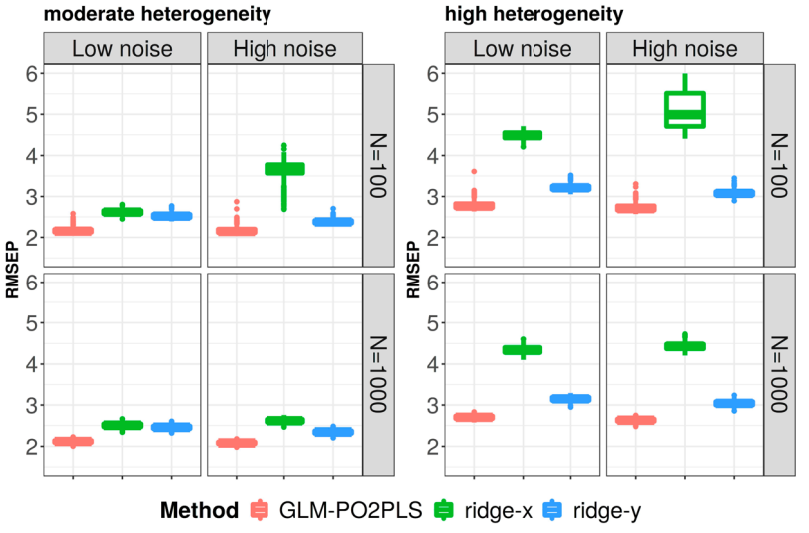
Lastly, we briefly present the key results of the additional simulations. Regarding parameter estimation performance when increasing the number of joint and specific components, we had similar results compared with those with one joint and one specific component (Supplementary Figure S6). Regarding power, when simulating under the settings based on the DS data, the empirical power was 0.998 (75% of noise in the outcome). When increasing the noise in the outcome to 90%, the empirical power remained high at 0.862.

4. Application to down syndrome study

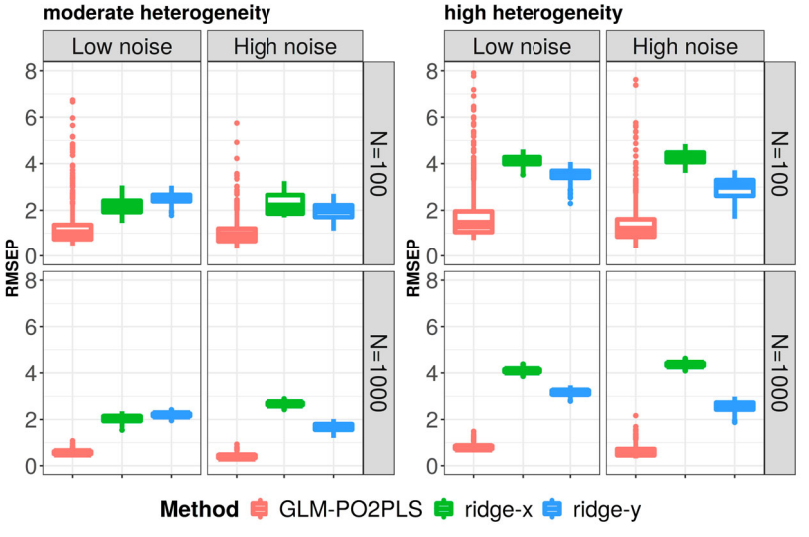
We apply the GLM-PO2PLS model to the Down syndrome dataset, aiming to investigate whether the relationship between methylation and glycomics is associated to DS, and select the relevant molecules involved in the relationship. Since Down syndrome is often considered as a model for aging [26], and both methylation and glycomics are associated with biological age [25,30], we expect the DS patients to be more similar to their mothers than siblings.

4.1. Data description

The Down syndrome study includes 29 families. Each family consists of one Down syndrome patient (DSP), one non-affected sibling (DSS), and their mother (DSM). The family-based design is used to control for genetic and environmental influences. Two DSS are missing. Thus, the total sample size N is equal to 85. The ages of the DSPs range from 10 to 43, with a median of 24 years. The ages of the siblings are roughly matched with the DS patients, ranging from 14 to 52 years. The mothers have ages between 41 and 83, with a median of 57 years.



(a) Continuous outcome z_c.



(b) Binary outcome z_b

Figure 3. Performance of outcome prediction for continuous (a) and binary (b) outcome. y-axis shows the RMSEP as defined in Table 2. Boxes show the results of 500 repetitions.(a) Continuous outcome z_c and (b) Binary outcome z_b.

For each individual, the whole blood methylation was measured using Infinium HumanMethylation450 BeadChip (Infinium 450k). After quality control following steps described in [4], 450981 CpG sites were retained. Beta value was derived at each CpG site as the ratio of intensities between methylated and unmethylated alleles. White blood cell counts were estimated from the beta values and corrected for using R package 'Meffil' [37]. For each CpG measurement, we performed a processing step by taking as measurements the residuals of a linear regression with the CpG measurement as outcome and sex and age

as covariates. The glycomic dataset consists of 10 plasma N-glycans measured using DNA sequencer-assisted fluorophore-assisted carbohydrate electrophoresis (DSA-FACE) [6]. These glycans were logTA normalized [59] and corrected for age and sex in the same way as the CpG sites.

Our implementation of GLM-PO2PLS binary model is limited to one joint component due to computational complexity. We first apply the computationally more efficient continuous GLM-PO2PLS model and identify potentially significant joint components. We then apply binary GLM-PO2PLS with one joint component and interpret the results. We set methylation as x, glycomics as y, and the DS status as z. The direction from methylation to glycomics (x to y) was chosen based on previous research [61] that suggested the presence of an indirect influence of methylation on glycosylation.

4.2. Results of DS data analysis

For the GLM-PO2PLS continuous model, we used 3 joint and 1 specific component for each omics dataset based on the scree plots of the eigenvalues of $x^\top y$, $x^\top x$ and $y^\top y$.

We first present the results regarding the relationship between methylation and glycomics, which is represented by the first three equations of the GLM-PO2PLS in (1). The p -value for each pair of methylation and glycomics joint components was 0.0007, 0.03, and 0.20, respectively. Using a threshold of 0.05 for statistical significance, the first (t_1 for methylation and u_1 for glycomics) and second pair (t_2 and u_2) of joint components were significantly associated. Figure 4 shows the scores of the first two pairs of joint components. For both t_1 and u_1, the DSPs were closer to the DSMs, than the DSS group, which was in line with our expectation. No noticeable patterns were observed in the second pair of joint components.

Table 3 shows the results regarding the relationship between the DS status and the omics. The significant test statistic T_{full} suggests that DS was associated with the two omics. Component-wise, only the first pair was significant, with a p-value of 6.32×10^{-5}.

Since t_1 and u_1 were significantly associated with DS, we investigated the CpG sites and glycans in the first component pair. In the first methylation joint component, the 1000 CpG sites with the largest loading values were mapped to their respective target genes, yielding 493 genes. Next, gene ontology (GO) enrichment analysis [3] was performed on this gene set using the GSEA software [38,52]. The top three significant GO terms were listed in Table 4. Among these terms, the cell-cell signaling is a biological function of plasma glycans [29,60]. The cellular component of neuron projection and biological process of neurogenesis were shown to relate to DS [22,27,48,49]. We further searched the mapped geneset in the DisGeNET database [42] for human diseases. The significant diseases found were chronic myeloid leukemia (q-value 0.0004), common acute lymphoblastic leukemia (q-value 0.045), and glioblastoma multiforme (q-value 0.045). Research has shown that children with Down syndrome have an increased risk for developing acute lymphoblastic leukemia [42]. For chronic myeloid leukemia and glioblastoma multiforme, we did not find evidence linking them with DS. We then checked the 7 genes with the highest gene-disease association score (which is a quantification of the association between a gene and a disease taking into account the number and type of sources and the number of publications supporting the association) regarding Down syndrome in the DisGeNET database, and found the gene RCAN1 which relates to epigenetics was among our top genes mapped

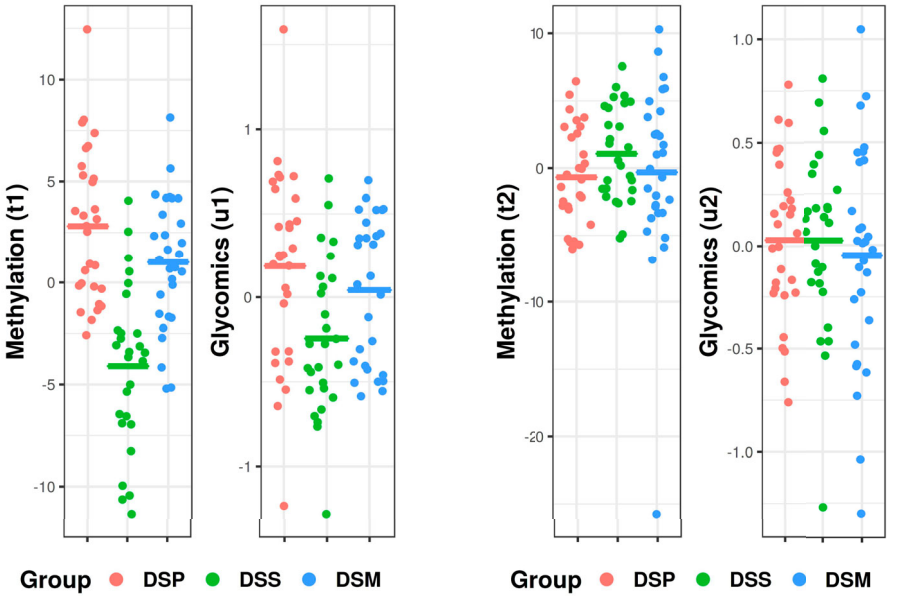


Figure 4. Joint scores of the first (left) and second (right) pair of joint components. On the y-axis are the scores of each individual colored by different groups. The mean score of each group is shown as a horizontal line.

Table 3. Results of testing for no relationship between DS and joint components.

	T_{full} in (6)	$T_{comp.wise}$ in (7)		
H_0	$a = b = 0$	$a_1 = b_1 = 0$	$a_2 = b_2 = 0$	$a_3 = b_3 = 0$
p-value	6.32×10^{-5}	1.35×10^{-5}	0.15	0.20

Table 4. Top 3 GO terms of the mapped genesets in GLM-PO2PLS continuous and binary models.

Gene Set Name (continuous model)	p-value	FDR q-value
GOBP CELL CELL SIGNALING	1.61e−13	2e−9
GOCC NEURON PROJECTION	2.96e−13	2e−9
GOBP NEUROGENESIS	4.08e−13	2e−9
Gene Set Name (binary model)	**p-value**	**FDR q-value**
GOCC ORGANELLE SUBCOMPARTMENT	4.22e−12	4.3e−8
GOCC GOLGI APPARATUS	8.27e−11	4.1e−7
GOCC VESICLE MEMBRANE	1.21e−10	4.1e−7

The p-value of each annotation was derived by random sampling of the whole genome; the FDR q-value provides the false discovery rate (FDR) analog of the p-value after correcting for multiple hypothesis testing [5,51].

from methylation. It has been revealed that RCAN1 plays a critical upstream role in epigenetic regulation of adult neurogenesis [8], hence important in the pathogenesis of Down syndrome [63].

For the first glycomics joint component, the glycan H3N4F1 had the largest absolute loading value. According to the result of a previous study [6] on plasma glycans and DS, H3N4F1 was the top discriminators of DS subjects and siblings.

Next we fitted a GLM-PO2PLS binary model with 1 joint and 1 specific component for each omics dataset. We chose for 1 joint component based on the test results in the continuous model shown in Table 3. The relationship between the two omics was significant with a p -value of 0.022. The top 1000 CpG sites were identified and mapped to genes. The most significant GO terms of the geneset are shown in Table 4. The top two terms were related to membrane organelle, more specifically, Golgi apparatus, which is required for accurate glycosylation [65]. Terms related to DS, such as neurogenesis (q-value 2.33e−6), neuron differentiation (1.11e−5), and synapse (1.33e−5) were also significant. Regarding glycomics, the glycan with the largest absolute loading value was H3N4F1, which was also identified in the GLM-PO2PLS continuous model.

5. Discussion

Motivated by the studies on the relationship among Down syndrome, methylation and glycomics, we developed a new statistical model GLM-PO2PLS, which simultaneously models the relationships among an outcome variable and two heterogeneous omics datasets. We studied in detail the models for normally and Benoulli distributed outcome variables. The identifiability of the model was established and EM algorithms were developed. For testing, we proposed two chi-square test statistics T_{full} and $T_{comp.wise}$ and derived their asymptotic distributions.

Via a simulation study, we have shown that the model parameters were well estimated, and the test statistic performed well in various scenarios. The outcome prediction performance of GLM-PO2PLS was robust against high noise and heterogeneity between omics. GLM-PO2PLS predicted the outcome better than ridge regressions, because it considers all the information in the data jointly, while ridge used each dataset separately. Another advantage of GLM-PO2PLS over ridge regression is that it can provide insights into the relationship between two omics datasets, on top of their relationship with the outcome.

Recently, a similar one-stage method supervised JIVE (sJIVE) [40] was published. It extends JIVE to model a continuous outcome as a linear combination of the joint and omic-specific components of multiple omics. The method estimates its parameters by optimizing a loss function containing both the omics and the continuous outcome, and thus finds a compromise between modeling the omics and predicting the outcome. The joint components of each omics data are assumed to be identical in sJIVE, following the same assumption in JIVE. A previous simulation study in [14] showed that such an assumption might be restrictive and lead to inferior performance when heterogeneity between omics is present (e.g. in our DS dataset). To perform statistical inference, resampling techniques such as bootstrapping are required since sJIVE is not likelihood-based. As sJIVE models a continuous outcome, the performance of applying sJIVE on non-normally distributed outcome is unclear. Systematically comparing GLM-PO2PLS to sJIVE is future work.

The methylation and glycomics dataset were also analyzed by Bacalini *et al*. [4] and Borelli *et al*. [6] using single point approaches for association with Down syndrome, respectively. Since for both omics datasets association with aging have been found and Down syndrome is an aging model, it makes sense to analyze them jointly in this study. Concerning methylation, Bacalini *et al*. identified four categories of genes. Most of the genes in these categories were also in our obtained geneset: haematopoiesis (RUNX1, DLL1, EBF4, PRDM16), morphogenesis and development (HOXA2, HOXA4, HHIP,

NCAM1), neuronal development (NAV1, EBF4, PRDM8, NCAM1), and regulation of chromatin structure (PRDM8, KDM2B). In total four genes mentioned in [4] were not in our gene list, namely, HOXA5, TET1, GABBR1, and HOXA6. It appeared that three out of these four genes rank just below our cut-off point of 1000, namely the CpG site with largest loading value in HOXA5, TET1, and GABBR1 ranked 1059, 1142, and 1535 out of 450K respectively. Concerning the fourth gene HOXA6, we performed univariate logistic regressions of the Down syndrome outcome on each of the 20 CpG sites located in the genetic region, and only identified one significant CpG site (p -value of 0.018). In comparison, the other selected genes from the HOXA family members have more significant CpG sites (such as HOXA2 with 22, HOXA4 with 16, the borderline HOXA5 with 13), and smaller p -values for the most associated CpG sites (HOXA2 0.0003, HOXA5 0.003). Furthermore, there is little evidence linking HOXA6 to the functions of glycans. Therefore, our proposed GLM-PO2PLS seems to better identify the CpG sites relevant to both DS and glycomics.

It is worth mentioning that we expect differences between our approach and the single-omic studies. The single-omic approaches did not consider the presence of correlation between CpG sites and glycomics when modeling the association of CpG sites and Down syndrome. Therefore, some methylation-specific genes that are unrelated with glycomics can rank lower in the joint components in our analysis. Furthermore, in GLM-PO2PLS, we focus on the joint part and the omic-specific parts are not linked to the outcome variable, and hence the top genes mapped from the methylation-specific components are not necessarily associated with the outcome DS. In this regard, an extension of our model which also considers the omic-specific parts in the linear predictor for the outcome variable can provide further insights into the disease from omic-specific aspects.

We have shown evidence for association between the mapped gene set and Down syndrome. Nonetheless, The dedicated 'Down syndrome' set in the DisGeNET database was not enriched in our gene set. One reason could be that very few studies have been conducted on DS with methylation data. Furthermore, common diseases and cancers are usually more frequently studied, resulting in possible publication bias in the database. We searched the genes identified by both our study and [4] (namely, RUNX1, DLL1, EBF4, HOXA2, HOXA4, HHIP, NCAM1, NAV1, PRDM8, KDM2B) in the DisGeNET database, and found none of these genes has their highest association score (i.e. amount of evidence) with DS. For example, the RUNX1 gene had the highest association score of 0.8 with acute myeloid leukemia, and a score of only 0.1 with DS.

When estimating a GLM-PO2PLS binary model, we rely on numerical integration. The computational complexity of the numerical estimation is $\mathcal{O}(M^{2r})$, with M nodes per dimension. In practice this means that the binary model can only include 1 joint component. A computationally feasible solution is to include only one pair of the joint components in the linear predictor for the binary outcome. Such a model might be suited for our Down syndrome analysis where only one pair of joint components was associated to the outcome. However, the assumption that only one pair of joint components is related to the outcome might not apply to other studies. Therefore, a more efficient numerical integration strategy is needed. One strategy is to use adaptive quadrature. Although for a fixed number of nodes M, the adaptive quadrature is computationally more complex than its non-adaptive counterpart we used, the adaptive variant needs a smaller M to reach an equally precise approximation, thus can be more efficient [43,44]. Another strategy is

to decompose the $2r$-dimensional integration to r 2-dimensional integrations. This will reduce the computational complexity to $\mathcal{O}(r \times M^2)$.

To calculate the p -values for the tests in (6) and (7), we derived the asymptotic normality of the estimator for the parameters of GLM-PO2PLS with a normally distributed outcome. Asymptotic normality was proved by showing that the mapping (denote τ) from the parameter vector θ to the moment structure as well as the discrepancy function with respect to the moments satisfy certain regularity conditions [46]. For the GLM-PO2PLS model with a binary outcome, there is not an explicit mapping function τ, and it is difficult to parameterize the likelihood in terms of the moments. Therefore, while the p -values for the binary model can be calculated assuming the asymptotic normality holds, it is unclear whether they are correct. The derivation of asymptotic normality for the binary model is future work.

In this paper our aim was to use one model for all the data, and model the relationships between the omics simultaneously with their relationship with an outcome to obtain a holistic overview. The method we proposed to estimate the model provides unbiased and efficient estimators of the model. However, for a binary outcome, the approach is computationally intensive. In specific situations, one may prefer two-stage approaches which is computationally faster. We recently proposed a two-stage PO2PLS approach [21], where we first constructed a few joint latent components that represent the two omics, then linked these latent components to the outcome variable using a linear regression model. In the implementation of two-stage PO2PLS to the DS dataset, the latent variables from the first stage were used as outcomes in several separate regression models in the second stage, thus the interpretation was different from a logistic regression model with DS as outcome. Alternatively, the latent variables can also be used as covariates in the second stage. However, the latent variables contain errors from the dimension reduction process. Ignoring these errors in the covariates can cause attenuated predicted probabilities in the logistic regression [50]. Therefore, to correctly model the outcome, the two-stage approach needs to be augmented with a measurement error model for the latent variables. Here more research is needed.

Several extensions of GLM-PO2PLS might be relevant. For an outcome variable from other members of the exponential family (e.g. Poisson, gamma, etc.), the corresponding EM algorithm can be obtained by modifying the EM algorithm for the binary outcome by replacing $p(z|\nu)$ in (9) with the corresponding conditional probability mass/density function. Regarding the relationship between omics and outcome, the omic-specific latent variables are not included in the linear predictor for the outcome variable in GLM-PO2PLS. As discussed above, linking the omic-specific parts to the outcome might provide further insights. Furthermore, since the omic-specific latent variable might also be predictive of the outcome, a model where all the latent variables are linked to the outcome can lead to improved outcome prediction performance in some studies. Extending GLM-PO2PLS to such a model will increase the computational complexity to $\mathcal{O}(M^{2r+K_x+K_y})$ for non-normal outcomes. In GLM-PO2PLS, we assume that the individuals are independent of each other and family structure is not taken into account. If this assumption is violated, the standard error of the coefficients (a, b) might be underestimated. An extension accounting for family structure in the model will be better suited for studies with non-independent participants. Another direction is to generalize the model to incorporate more than two omics datasets jointly with an outcome. Such an extension would require to specify the directions of the

relationships among more than two sets of variables. A workaround might be to model a common set of latent variables for all sets of variables [36]. For some studies, the directions of the relationships are clear (e.g. among genetics, methylation, and glycomics), and specifying the direction in the model and allowing the joint latent variables for each set of variables to differ can improve model performance. However, the computation will also be intensive for a binary outcome.

To conclude, GLM-PO2PLS is a promising method to model an outcome with two omics datasets and as a base for further extensions.

Disclosure statement

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors were supported by the following financial support for the research, authorship, and/or publication of this article: Zhujie Gu was supported by the European Union's Horizon 2020 research and innovation programme IMforFUTURE [grant number 721815]; the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking BigData@Heart [grant number 116074]; and Medical Research Council [programme number MC-UU-00002/5]. Said el Bouhaddani was supported by ERA-Net E-Rare JTC 2018 (MSA-omics) [40-44000-98-2006/ 90030376507].

Data availability statement

The DNA methylation data used in this study are available at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE52588.

ORCID

Zhujie Gu http://orcid.org/0000-0001-7675-8000
Jeanine Houwing-Duistermaat http://orcid.org/0000-0002-4505-7137

References

[1] M. Abramowitz and Irene A. Stegun, *Numerical interpolation, differentiation, and integration*, in *Handbook of Mathematical Functions*, M. Abramowitz and I.A. Stegun, eds., Dover Publications, 1972, pp. 877–925. Available at https://books.google.com/books/about/Handbook_of_Mathematical_Functions.html?id=V3ZQAAAAMAAJ.

[2] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pac. J. Math. 16 (1966), pp. 1–3. Available at https://projecteuclid.org/journals/pacific-journal-of-mathematics/volum.

[3] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene ontology: Tool for the unification of biology*, 2000.

[4] M.G. Bacalini, D. Gentilini, A. Boattini, E. Giampieri, C. Pirazzini, C. Giuliani, E. Fontanesi, M. Scurti, D. Remondini, M. Capri, G. Cocchi, A. Ghezzo, A.D. Rio, D. Luiselli, G. Vitale, D. Mari, G. Castellani, M. Fraga, A.M. Di Blasio, S. Salvioli, C. Franceschi, and P. Garagnani, *Identification of a DNA methylation signature in blood cells from persons with down syndrome*, Aging 7 (2015), pp. 82–96.

[5] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. R. Stat. Soc. Ser. B (Methodol.) 57 (1995), pp. 289–300.

[6] V. Borelli, V. Vanhooren, E. Lonardi, K.R. Reiding, M. Capri, C. Libert, P. Garagnani, S. Salvioli, C. Franceschi, and M. Wuhrer, *Plasma N-Glycome signature of down syndrome*, J. Proteome. Res. 14 (2015), pp. 4232–4245.
[7] A.L. Boulesteix, R. De Bin, X. Jiang, and M. Fuchs, *IPF-LASSO: integrative l1-penalized regression with penalty factors for prediction based on multi-Omics data*, Comput. Math. Methods Med. 2017 (2017), pp.1–14.
[8] C. Choi, T. Kim, K.T. Chang, and K.T. Min, *DSCR1-mediated TET1 splicing regulates miR-124 expression to control adult hippocampal neurogenesis*, EMBO. J. 38 (2019), pp. e101293. doi:10.15252/embj.2018101293
[9] F. Ciccarone, E. Valentini, M. Malavolta, M. Zampieri, M.G. Bacalini, R. Calabrese, T. Guastafierro, A. Reale, C. Franceschi, M. Capri, N. Breusing, T. Grune, M. Morenoe Villanueva, A. Bürkle, and P. Caiafa, *DNA hydroxymethylation levels are altered in blood cells from down syndrome persons enrolled in the MARK-AGE project*, J. Gerontol. Ser. A. Bio. Sci. Med. Sci. 73 (2018), pp. 737–744. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5946825/.
[10] A. Cindric, F. Vuckovic, V. Borelli, J. Juric, H. Deris, A. Murray, I. Alic, J. Groet, D. Petrovic, and S. Hamburg, *Accelerated biological aging in people with down syndrome with full and segmental trisomy 21 begins in childhood as revealed by immunoglobulin G glycosylation*, Res. Sq. (2021), pp. 1–29.
[11] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm* , J. R. Stat. Soc. Ser. B. (Methodol). 39 (1977), pp. 1–22.
[12] D.Y. Ding, S. Li, B. Narasimhan, and R. Tibshirani, *Cooperative learning for multiview analysis*, Proc. Natl. Acad. Sci. USA 119 (2022), pp. e2202113119. doi:10.1073/pnas.2202113119
[13] S. el Bouhaddani, J. Houwing-Duistermaat, P. Salo, M. Perola, G. Jongbloed, and H.W. Uh, *Evaluation of O2PLS in omics data integration*, BMC. Bioinform. 17 (2016), pp. S11. doi:10.1186/s12859-015-0854-z
[14] S. el Bouhaddani, H.W. Uh, G. Jongbloed, C. Hayward, L. Klarić, S.M. Kiełbasa, and J. Houwing-Duistermaat, *Integrating omics datasets with the OmicsPLS package*, BMC. Bioinform. 19 (2018), pp. 371.
[15] S. el Bouhaddani, H.W. Uh, G. Jongbloed, and J. Houwing-Duistermaat, *Statistical integration of heterogeneous omics data: probabilistic two-way partial least squares (PO2PLS)*, J. R. Stat. Soc. Ser. C (Appl. Stat.) 71 (2022), pp. 1451–1470. doi:10.1111/rssc.12583
[16] S. el Bouhaddani, H.W. Uh, G. Jongbloed, and J. Houwing-Duistermaat, *Statistical integration of heterogeneous data with PO2PLS*, 2021.
[17] C. Franceschi, P. Garagnani, N. Gensous, M.G. Bacalini, M. Conte, and S. Salvioli, *Accelerated bio-cognitive aging in Down syndrome: state of the art and possible deceleration strategies*, 2019. doi:10.1111/acel.12903
[18] I. Gaynanova and G. Li, *Structural learning and integrative decomposition of multi-view data*, Biometrics 75 (2019), pp. 1121–1132. doi:10.1111/biom.13108
[19] N. Gensous, M.G. Bacalini, C. Franceschi, and P. Garagnani, *Down syndrome, accelerated aging and immunosenescence*, 2020. doi:10.1007/s00281-020-00804-1
[20] S.M. Gross and R. Tibshirani, *Collaborative regression*, Biostatistics 16 (2015), pp. 326–338. doi:10.1093/biostatistics/kxu047
[21] Z. Gu, S. El Bouhaddani, J. Houwing-Duistermaat, and H.W. Uh, *Investigating the impact of down syndrome on methylation and glycomics with two-stage PO2PLS*, Theor. Biol. Forum. 114 (2021), pp. 29–44. Available at http://digital.casalini.it/5213807.
[22] M.A. Haas, D. Bell, A. Slender, E. Lana-Elola, S. Watson-Scales, E.M. Fisher, V.L. Tybulewicz, and F. Guillemot, *Alterations to dendritic spine morphology, but not dendrite patterning, of cortical projection neurons in Tc1 and Ts1Rhr mouse models of Down syndrome*, PLoS. ONE. 8 (2013), pp. e78561. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3813676/.
[23] I.S. Helland, *On the structure of partial least squares regression*, Commun. Stat. Simul. Comput. 17 (1988), pp. 581–607.
[24] A.E. Hoerl and R.W. Kennard, *Ridge regression: biased estimation for nonorthogonal problems*, Technometrics 42 (2000), pp. 80–86.

[25] S. Horvath, *DNA methylation age of human tissues and cell types*, Genome. Biol. 14 (2013), pp. R115. Available at http://genomebiology.com//14/10/R115.
[26] S. Horvath, P. Garagnani, M.G. Bacalini, C. Pirazzini, S. Salvioli, D. Gentilini, A.M. Di Blasio, C. Giuliani, S. Tung, H.V. Vinters, and C. Franceschi, *Accelerated epigenetic aging in down syndrome*, Aging. Cell. 14 (2015), pp. 491–495.
[27] H.Q. Huo, Z.Y. Qu, F. Yuan, L. Ma, L. Yao, M. Xu, Y. Hu, J. Ji, A. Bhattacharyya, S.C. Zhang, and Y. Liu, *Modeling down syndrome with patient iPSCs reveals cellular and migration deficits of GABAergic neurons*, Stem. Cell. Reports. 10 (2018), pp. 1251–1266. doi:10.1016/j.stemcr.2018.02.001
[28] A. Kaplan and E.F. Lock, *Prediction with dimension reduction of multiple molecular data sources for patient survival*, Cancer Inform. 16 (2017), pp. 117693511771851. doi:10.1177/1176935117718517
[29] F. Krautter and A.J. Iqbal, *Glycans and glycan-binding proteins as regulators and potential targets in leukocyte recruitment*, 2021.
[30] J. Krištić, F. Vučković, C. Menni, L. Klarić, T. Keser, I. Beceheli, M. Pučić-Baković, M. Novokmet, M. Mangino, K. Thaqi, P. Rudan, N. Novokmet, J. Šarac, S. Missoni, I. Kolčić, O. Polašek, I. Rudan, H. Campbell, C. Hayward, Y. Aulchenko, A. Valdes, J.F. Wilson, O. Gornik, D. Primorac, V. Zoldoš, T. Spector, and G. Lauc, *Glycans are a novel biomarker of chronological and biological ages*, J. Gerontol. Ser. A. Bio. Sci. Med. Sci. 69 (2014), pp. 779–789. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4049143/.
[31] G. Li and S. Jung, *Incorporating covariates into integrated factor analysis of multi-view data*, Biometrics 73 (2017), pp. 1433–1442.
[32] Q. Liu and D.A. Pierce, *A note on gauss-Hermite quadrature*, Biometrika 81 (1994), pp. 624.
[33] E.F. Lock, K.A. Hoadley, J.S. Marron, and A.B. Nobel, *Joint and individual variation explained (JIVE) for integrated analysis of multiple data types*, Ann. Appl. Stat. 7 (2013), pp. 523–542. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3671601/ https://genome.unc.edu/jive/.
[34] T.A. Louis, *Finding the observed information matrix when using the EM algorithm*, J. R. Stat. Soc. Ser. B. (Methodol). 44 (1982), pp. 226–233.
[35] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
[36] C. Meng, O.A. Zeleznik, G.G. Thallinger, B. Kuster, A.M. Gholami, and A.C. Culhane, *Dimension reduction techniques for the integrative analysis of multi-omics data*, Brief. Bioinform. 17 (2016), pp. 628–641. Available at https://pubmed.ncbi.nlm.nih.gov/26969681/.
[37] J.L. Min, G. Hemani, G.D. Smith, C. Relton, and M. Suderman, *Meffil: efficient normalization and analysis of very large DNA methylation datasets*, Bioinformatics 34 (2018), pp. 3983–3989. Available at https://pubmed.ncbi.nlm.nih.gov/29931280/.
[38] V.K. Mootha, C.M. Lindgren, K.F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M.J. Daly, N. Patterson, J.P. Mesirov, T.R. Golub, P. Tamayo, B. Spiegelman, E.S. Lander, J.N. Hirschhorn, D. Altshuler, and L.C. Groop, *PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes*, Nat. Genet. 34 (2003), pp. 267–273. Available at https://www.nature.com/articles/ng1180.
[39] A. Nishiyama and M. Nakanishi, *Navigating the DNA methylation landscape of cancer*, 2021.
[40] E.F. Palzer, C.H. Wendt, R.P. Bowler, C.P. Hersh, S.E. Safo, and E.F. Lock, *sJIVE: supervised joint and individual variation explained*, Comput. Stat. Data Anal. 175 (2022), pp.107547.
[41] D. Patterson, *Genetic mechanisms involved in the phenotype of down syndrome*, 2007.
[42] J. Piñero, J.M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L.I. Furlong, *The DisGeNET knowledge platform for disease genomics: 2019 update*, Nucleic. Acids. Res. 48 (2020), pp. D845–D855. Available at https://academic.oup.com/nar/article/48/D1/D845/5611674.
[43] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, *Reliable estimation of generalized linear mixed models using adaptive quadrature*, Stata J. Promoting Commun. Stat. Stata. 2 (2002), pp. 1–21.

[44] N.J. Rockwood, *Efficient likelihood estimation of generalized structural equation models with a mix of normal and nonnormal responses*, Psychometrika 86 (2021), pp. 642–667. doi:10.1007/s11336-021-09770-5
[45] M. Rodríguez-Girondo, P. Salo, T. Burzykowski, M. Perola, J. Houwing-Duistermaat, and B. Mertens, *Sequential double cross-validation for assessment of added predictive ability in high-dimensional omic applications*, Ann. Appl. Stat. 12 (2018), pp. 1655–1678. Available at https://projecteuclid.org/journals/annals-of-applied-statistics/volume.
[46] A. Shapiro, *Asymptotic theory of overparameterized structural models*, J. Am. Stat. Assoc. 81 (1986), pp. 142–149.
[47] M. Sheikhpour, M. Maleki, M. Ebrahimi Vargoorani, and V. Amiri, *A review of epigenetic changes in asthma: methylation and acetylation*, 2021. doi:10.1186/s13148-021-01049-x
[48] M. Sobol, J. Klar, L. Laan, M. Shahsavani, J. Schuster, G. Annerén, A. Konzer, J. Mi, J. Bergquist, J. Nordlund, J. Hoeber, M. Huss, A. Falk, and N. Dahl, *Transcriptome and proteome profiling of neural induced pluripotent stem cells from individuals with down syndrome disclose dynamic dysregulations of key pathways and cellular functions*, Mol. Neurobiol. 56 (2019), pp. 7113–7127. doi:10.1007/s12035-019-1585-3
[49] F. Stagni, A. Giacomini, M. Emili, S. Guidi, and R. Bartesaghi, *Neurogenesis impairment: an early developmental defect in Down syndrome*, 2018.
[50] L.A. Stefanski and R.J. Carroll, *Covariate measurement error in logistic regression*, Ann. Stat. 13 (1985), pp. 1335–1351.
[51] J.D. Storey, *A direct approach to false discovery rates*, J. R. Stat. Soc. Ser. B: Stat. Methodol. 64 (2002), pp. 479–498.
[52] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov, *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*, Proc. Natl. Acad. Sci. USA 102 (2005), pp. 15545–15550. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1239896/.
[53] S. Sugár, G. Tóth, F. Bugyi, K. Vékey, K. Karászi, L. Drahos, and L. Turiák, *Alterations in protein expression and site-specific N-glycosylation of prostate cancer tissues*, Sci. Rep. 11 (2021), pp. 1–12. Available at https://www.nature.com/articles/s41598-021-95417-5.
[54] D.N. Tabang, M. Ford, and L. Li, *Recent advances in mass spectrometry-based glycomic and glycoproteomic studies of pancreatic diseases*, 2021.
[55] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B. (Methodol.) 58 (1996), pp. 267–288.
[56] R. Tissier, J. Houwing-Duistermaat, and M. Rodríguez-Girondo, *Improving stability of prediction models based on correlated omics data by using network approaches*, PLOS ONE 13 (2018), pp. e0192853. doi:10.1371/journal.pone.0192853
[57] J. Trygg and S. Wold, *O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter*, J. Chemom. 17 (Jan. 2003), pp. 53–64. doi:10.1002/cem.775
[58] E. Uffelmann, Q.Q. Huang, N.S. Munung, J. de Vries, Y. Okada, A.R. Martin, H.C. Martin, T. Lappalainen, and D. Posthuma, *Genome-wide association studies*, Nat. Rev. Methods Primers 2021 (2021), pp. 1–21. Available at https://www.nature.com/articles/s43586-021-00056-9.
[59] H.W. Uh, L. Klaric, I. Ugrina, G. Lauc, A.K. Smilde, and J.J. Houwing-Duistermaat, *Choosing proper normalization is essential for discovery of sparse glycan biomarkers*, Mol. Omics 16 (2020), pp. 231–242.
[60] A. Varki, *Biological roles of glycans*, Glycobiology 27 (2017), pp. 3–49. Available at https://pubmed.ncbi.nlm.nih.gov/27558841/.
[61] A. Wahl, S. Kasela, E. Carnero-Montoro, M. van Iterson, J. Štambuk, S. Sharma, E. van den Akker, L. Klaric, E. Benedetti, G. Razdorov, I. Trbojević-Akmačić, F. Vučković, I. Ugrina, M. Beekman, J. Deelen, D. van Heemst, B.T. Heijmans, B.I.O.S. Consortium, M. Wuhrer, R. Plomp, T. Keser, M. Šimurina, T. Pavić, I. Gudelj, J. Krištić, H. Grallert, S. Kunze, A. Peters, J.T. Bell, T.D. Spector, L. Milani, P.E. Slagboom, G. Lauc, and C. Gieger, *IgG glycosylation and DNA methylation are interconnected with smoking*, Biochim. Biophys.

Acta – Gen. Sub. 1862 (2018), pp. 637–648. Available at https://www.sciencedirect.com/science/article/pii/S0304416517303410?dgcid=raven_sd_recommender_email.
[62] K. Watanabe, S. Stringer, O. Frei, M. Umićević Mirkov, C. de Leeuw, T.J. Polderman, S. van der Sluis, O.A. Andreassen, B.M. Neale, and D. Posthuma, *A global overview of pleiotropy and genetic architecture in complex traits*, Nat. Genet. 51 (2019), pp. 1339–1348.
[63] Y. Yun, Y. Zhang, C. Zhang, L. Huang, S. Tan, P. Wang, C. Vilariño-Gúell, W. Song, and X. Sun, *Regulator of calcineurin 1 is a novel RNA-binding protein to regulate neuronal apoptosis*, Mol. Psychiatry. 26 (2021), pp. 1361–1375. Available at https://pubmed.ncbi.nlm.nih.gov/31451750/.
[64] Y. Zhang and I. Gaynanova, *Joint association and classification analysis of multi-view data*, Biometrics 78 (2021), pp. 1614–1625.
[65] X. Zhang and Y. Wang, *Glycosylation quality control by the golgi structure*, 2016.