STATISTICAL MODELING OF AN OUTCOME VARIABLE WITH INTEGRATED OMICS DATA

Zhujie Gu

STATISTICAL MODELING OF AN OUTCOME VARIABLE WITH INTEGRATED OMICS DATA

ZHUJIE GU

STATISTICAL MODELING OF AN OUTCOME VARIABLE WITH INTEGRATED OMICS DATA

Zhujie GU

Author: Zhujie Gu
Layout: Zhujie Gu
Cover design: Publiss | www.publiss.nl
Print: Ridderprint | www.ridderprint.nl

ISBN: 978-94-6483-123-8

STATISTICAL MODELING OF AN OUTCOME VARIABLE WITH INTEGRATED OMICS DATA

STATISTISCHE MODELLERING VAN EEN UITKOMSTVARIABELE MET GEÏNTEGREERDE OMICS-DATA
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 22 mei 2023 des ochtends te 9.00 uur

door

Zhujie GU

geboren op 21 januari 1990
te Huzhou, China

Promotoren:
Prof. dr. M.C.J.M. Sturkenboom
Prof. dr. J.J. Houwing-Duistermaat

Copromotoren:
Dr. H.W. Uh
Dr. S. el Bouhaddani

Beoordelingscommissie:
Prof. dr. D.L. Oberski
Dr. W.N. van Wieringen
Prof. dr. J.H. Veldink
Prof. dr. ir. R.C.H. Vermeulen
Prof. dr. M. Zucknick

This thesis was accomplished with financial support from the European Union's Horizon 2020 research and innovation programme IMforFUTURE [grant number 721815], and the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking BigData@Heart [grant number 116074].

CONTENTS

1

General introduction

1

1.1. STATISTICAL BACKGROUND

Multi-omics efforts have taken center stage in biomedical research and can potentially develop new insights into human diseases [11]. The central statistical aim of these efforts is modeling the outcome disease in relationship with the correlation structure within and across omic layers. For a single omic layer, a well known traditional statistical method is generalized linear model (GLM) [15]. In this model, the relationship between an outcome $z \in \mathbb{R}$ and a set of omic features $x \in \mathbb{R}^p$ is given by

$$\eta[\mathbb{E}(z|x)] = \beta_0 + \beta^\top x, \tag{1.1}$$

where η is the link function connecting the conditional mean of the outcome variable z given x to a linear predictor; β_0 is the intercept, and β is the p-dimensional regression coefficient vector describing how the features in x relate to the outcome. The GLM model can be estimated by maximizing the log-likelihood. When x is high-dimensional and highly correlated, which is often the case for omics data, the estimate of β and its interpretation will be unreliable [14]. To address this problem, regularized regression approaches like ridge [10], lasso [4], or elastic net [26] (which is the combination of the former two) can be applied. Alternatively, latent variable method such as partial least squares univariate regression [22, 9] (or PLS1, to be clearly distinguished from the PLS later in the chapter) can be used.

The statistical challenge here is how to model the outcome utilizing multi-omic datasets. The most common application is to apply the GLM model in (1.1) to each omics dataset. To obtain an overall result, multiple single-omic analyses are typically followed by taking the union or the intersection of the associated genes or gene products across the omics layers. Such an approach lacks an overview of the structure within and across omics related to the outcome. Alternatively, one can stack multiple omics datasets into one and apply GLM on the stacked dataset. Multiple penalties might be applied to regularize the regression coefficients for each set of omic features separately. This type of models focuses on the relationship between an outcome variable and the omics datasets and does not model the joint distribution of the omics datasets thus lacks insight into the multi-omics structure. Moreover, they do not explicitly take into account the correlation structure between the omics datasets. Approaches that model an outcome with the correlation structure both within and between omics datasets are needed. In this thesis, such methods are proposed and studied.

To model the multivariate covariance structure within and across two datasets without an outcome, latent variable approaches such as partial least squares (PLS) [20, 2] have been developed. Latent variable approaches map both datasets from the original high-dimensional spaces to low-dimensional joint latent spaces which retain the relationship (we refer such methods to integrative methods in this thesis). Let $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ be two distinct sets of omic features of dimension p and q, respectively. The PLS model is given by

$$\begin{aligned} x &= tW^\top + e, \\ y &= uC^\top + f, \\ u &= tB + h, \end{aligned} \tag{1.2}$$

where the K-dimensional ($K << p$ or q) joint latent variables t (for x) and u (for y) capture the relationship between x and y, and their inner relationship is modelled by the $K \times K$ diagonal matrix B. $W(p \times K)$ and $C(q \times K)$ are the loading matrices indicating relative importance of each omic feature in x and y. The vectors e, f, h of dimension p, q, K respectively are the residuals.

These PLS-type of methods tailored for omics datasets (which will be introduced in the next section) do not model the outcome z. How to incorporate the regression model for the outcome z in (1.1) into the integrative model for two omics datasets x and y, or more broadly, how to jointly model the outcome z with two omics datasets taking into account the correlation structure within and across omic layers is not addressed in current literature. This thesis will build upon integrative methods towards a solution to this problem.

1.2. INTEGRATIVE METHODS FOR TWO OMICS - O2PLS

The analyses in this thesis involve several omic layers, including genomics, epigenomics, transcriptomics, glycomics, and metabolomics. A prominent characteristic present among these data is the complex dependence structure both within and between the datasets, as illustrated in Figure 1.1. The PLS model (1.2) addresses the correlation within dataset by projecting the dependent original omic features to independent latent variables, and models the dependence structure between datasets by associating the latent variables for each omics dataset.

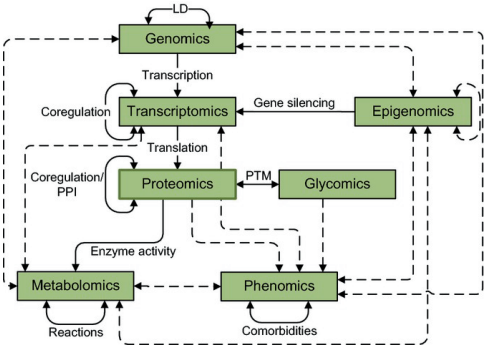


Figure 1.1: **Dependence structure within and between several types of omics data.** Each rectangle represents an omic layer, where arrows depict possible relationships between the omics. Dependence within an omic layer also exists. Figure taken from [25].

However, omics data are also heterogeneous in several aspects, with respect to the source of variation, dimensionality, and measurement platform. When integrating two heterogeneous omics datasets, the joint latent components t and u in the PLS model in (1.2) also contain (strong) omic-specific variation that is not involved in the joint variation of the two omics. Ignoring this omic-specific variation may lead to an erroneous representation of the true relationship. To account for the heterogeneity, the PLS model was extended to two-way orthogonal PLS (O2PLS) [18, 6], which includes data-

specific latent variables $t_\perp$ and $u_\perp$ that are independent of the joint and residual parts. For example, x is genomics and y is glycomics. The underlying model of O2PLS for their relationship is depicted in Figure 1.2. Here, t and u are joint latent variables modeling the relationship between genomics and glycomics, and $t_\perp$ and $u_\perp$ are specific latent variables that capture the variance of genomics not related to glycomics and variance of glycomics not related to genomics, respectively.

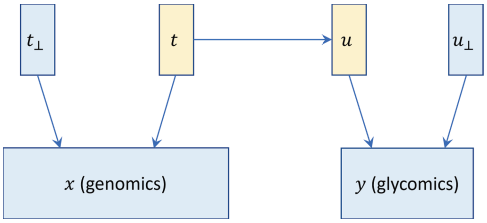


Figure 1.2: **O2PLS model for relationship between two omics datasets x and y.**

The equations for the underlying model of O2PLS are given by

$$\begin{aligned} x &= tW^\top + t_\perp W_\perp^\top + e, \\ y &= uC^\top + u_\perp C_\perp^\top + f, \\ u &= tB + h. \end{aligned} \tag{1.3}$$

where $W_\perp$ and $C_\perp$ are data-specific loading matrices.

Model (1.3) can be fitted by using algorithmic approaches [18, 6] or by using maximum likelihood estimation (MLE) methods [7, 8]. In the first approach, three sequential steps are applied. First, the data-specific parts are ignored and model (1.3) reduces to model (1.2) without data-specific parts. The parameters W and C are estimated per column by iteratively projecting (x, y) onto (t, u) via the current estimates for the respective columns in W and C, and vice versa [17, 21]. Next, with the initial estimate for the joint parts from the first step, model (1.3) reduces to PCA models for x and y where $W_\perp$ and $C_\perp$ are the loadings for the principal components. Lastly, these data-specific components are subtracted and model (1.3) again reduces to model (1.2) as in the first step. Likelihood-based approaches specify a multivariate normal distribution for the latent variables and the residual terms, and optimizes all the parameters of model (1.3) simultaneously. Directly maximizing the likelihood function is analytically and computationally not feasible. An EM algorithm [3] can be used as proposed in the probabilistic O2PLS (PO2PLS) [8]. In the expectation (E) steps, it calculates the first and second conditional moments of the latent variables, given the observed data and the current estimator of parameters (starting with an initial guess for the parameters). In the maximization (M) steps, the estimator is updated as the maximizer of the complete likelihood, where the expectations are used as predictions for the latent quantities. Under some assumptions, the sequence of estimates converge to a local optimum of the likelihood [3, 23]. Compared to the algorithmic approaches, the likelihood-based approach has the advantage of producing standard errors and hence facilitate statistical inference. For detailed descriptions and discussions of the integrative models, see [1].

1.3. MODELING THE OUTCOME

The model in (1.3) captures the joint variation between omics datasets in the joint latent variables t and u. These joint latent variables can be used to unveil the relationship between the outcome and the integrated omics (or the joint parts of omics). An exploratory analysis such as data visualization can be usually performed prior to any modeling of the outcome. The dimension reduction embedded in model (1.3) provides a convenient way to visualize the high-dimensional omics data and explore their relationship with the outcome. For statistical modeling, a two-stage approach can be considered where the outcome z is modeled in the second stage using linear models with the latent variables estimated from the first stage. Compared to two separate models, a one-step approach that models the joint distribution of the outcome and the omics is expected to yield better results. In this thesis, the main interest is modeling the outcome with the joint parts of omics. However, the data-specific latent variables can be of interest as well to uncover the mechanism underlying the outcome that involves only a specific omic layer. This will be discussed in the last chapter.

1.3.1. VISUALIZATION OF THE RELATIONSHIP

One of the main advantages of dimension reduction is that it enables visualization of the high-dimensional data. The number of joint latent variables K for the O2PLS model is small, usually taken between 1 and 5, based on cross-validation or scree plots of eigenvalues of $x^\top y$. The values of the joint latent variables (also called the joint component scores) for an individual can be estimated as $\hat{t} = x\hat{W}$, where $\hat{W}$ is the estimated joint loading matrix of x in model (1.3). These scores can be plotted against each other in a scatter plot (or joint score plot) to visually identify clusters of closely related individuals. Usually the scores of the first two joint components ($\hat{t}_{(1)}, \hat{t}_{(2)}$) are used as they explain the most of the covariance between omics. These scores are obtained in an unsupervised manner, i.e., the information of the outcome is not used. One can add the information of the outcome to a joint score plot by coloring the individuals based on the outcome status (for a binary outcome such as disease status) to see if the clusters overlap with the outcome. Figure 1.3 gives an example of such visualization from a study of hypertrophic cardiomyopathy (HCM) in Chapter 2, where regulomics and transcriptomics are integrated. The x- and y- axes are the values of the first and second joint latent variables (i.e., ($\hat{t}_{(1)}, \hat{t}_{(2)}$) and ($\hat{u}_{(1)}, \hat{u}_{(2)}$)). Each healthy control is marked in red and each HCM patient is colored in blue. The 95% confidence regions of each group is also added. From this plot, it is clear that the joint parts of both omics are associated with the outcome disease with a clear separation of the two groups.

1.3.2. TWO-STAGE MODELING OF THE OUTCOME

Two-stage approaches are considered where in the first stage, an integrative model is fit and the latent variables are estimated. In the second stage, the estimated latent variables from the first stage and the outcome z are modeled using various linear models, leading to different applications and interpretations.

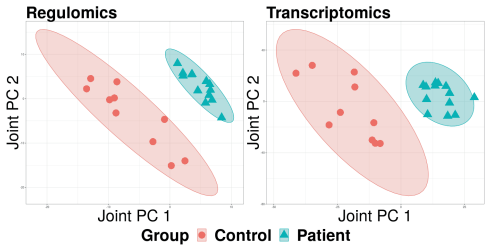


Figure 1.3: **Joint score plots of regulomics (left) and transcriptomics (right).** The x- and y- axes are $(\hat{t}_{(1)}, \hat{t}_{(2)})$ and $(\hat{u}_{(1)}, \hat{u}_{(2)})$ for regulomics and transcriptomics, respectively. HCM patients and controls were plotted in different colors. Ellipses are the 95% confidence regions of each group.

LATENT VARIABLES AS COVARIATES

We first consider modeling an outcome z using the GLM model in (1.1), with the estimated joint latent variables for omics from an O2PLS model as covariates. Specifically, model (1.3) is followed by

$$\eta[\mathbb{E}(z|t,u)] = \beta_0 + ta^\top + ub^\top, \tag{1.4}$$

where β_0 is the intercept, a and b are K-dimensional regression coefficients for t and u, respectively. Here, t and u can be highly correlated as they are designed to capture the covariance structure of x and y. This high correlation can cause the estimation of the regression coefficients to be unstable. We propose two solutions to address the problem.

Use t to represent both omics One solution is to include only t in model (1.4),

$$\eta[\mathbb{E}(z|t)] = \beta_0 + ta^\top. \tag{1.5}$$

The rationale is that t represents both omics well and u does not add much value (and vice versa) to modeling the outcome if t and u are highly correlated. This leads to a novel application of the two-stage approach in studies that involve genomics and heritable omics to construct scores from genomics representing the omics for modeling an outcome.

Let x be the genomics of an individual. The value of the k-th joint latent variable is estimated as

$$\hat{t}_{(k)} = x\hat{w}_{(k)} = \sum_{j=1}^{p} x_j \hat{w}_{(k)j}, \tag{1.6}$$

where $\hat{w}_{(k)}$ is the k-th estimated joint loading vector of x (i.e., the k-th column of $\hat{W}$). This is a weighted sum of all the alleles and is closely related to polygenic score (PGS) [5] which summarizes the estimated effect of many genetic variants on an individual's phenotype. A PGS is typically calculated as a weighted sum of m trait-associated alleles as

$$\text{PGS} = \sum_{j=1}^{m} G_j \hat{\beta}_j, \tag{1.7}$$

where G_j is the allele count for the j-th SNP, and $\hat{\beta}_j$ is the effect size of the j-th SNP estimated by a relevant genome-wide association study (GWAS). Both $\hat{t}_{(k)}$ in (1.6) and the PGS constructed for omic features in (1.7) are representation of the variation in omics that is related to genomics (namely the heritable part). Overall, they are stable over lifespan as the genomics do not change. Therefore can be used to model the outcome in (1.5) independent of time, and without the omics being measured (as long as the weights $\hat{W}$ and $\hat{\beta}$ are available from previous studies). The joint scores constructed from integrative methods have advantages over PGS as the covariance structure of both omics are modeled simultaneously, while the GWAS which PGS is based on typically examines pairwise associations. Details of the methodologies and their comparison and applications will be formulated in Chapter 3.

Re-parametrization Instead of discarding u in (1.4), another solution to address the correlation between t and u is to substitute $u = tB + h$ from the last equation in (1.3) and re-parametrize (1.4) as

$$\eta(\mathbb{E}[z|t,h]) = \beta_0 + ta^\top + (tB+h)b^\top = \beta_0 + t\tilde{a}^\top + h\tilde{b}^\top. \tag{1.8}$$

Here, h is the part in u that is independent of t. With this equivalent parametrization, instability due to high correlation in the linear predictor is reduced. The formulation in (1.8) will be revisited in Chapter 5.

LATENT VARIABLES AS PSEUDO-OUTCOMES

The latent variables can also be regarded as pseudo-outcomes in a linear model in the second stage. Separate models can be fitted for each k in $\{1,\dots,K\}$,

$$t_{(k)} = \beta_0 + z\beta + \epsilon, \tag{1.9}$$

where ϵ is the residual. Analogously, another K regression models are needed for $u_{(k)}$. The interpretation of regressions as (1.9) focuses on the effect of z on the omics. Such a formulation is natural to use on an outcome disease that is not effected by omics, for example, down syndrome (DS). In Chapter 4, an extended regression model of this type with covariates and random effects will be used to model the effect of DS on the joint part of methylation and glycomics.

Model (1.9) has an advantage that it models the estimation error from the first stage with the residual term ϵ. This estimation error is not modeled in (1.4), which often results in biased parameter estimates [16] and inaccurate estimates for the standard errors. This leads to the development of a joint modeling framework in the next section.

1.3.3. A ONE-STEP MODEL FOR THE JOINT DISTRIBUTION OF OUTCOME AND OMICS

From a statistical point of view, a simultaneous approach, rather than two consecutive steps, is expected to yield overall more accurate estimates for the parameters in each step. A model for the joint distribution of an outcome variable z and two omics datasets x and y can be formulated by combining the O2PLS model in (1.3) and the GLM model

in (1.1), via shared latent variables,

$$\begin{aligned} x &= tW^\top + t_\perp W_\perp^\top + e, \\ y &= uC^\top + u_\perp C_\perp^\top + f, \\ u &= tB + h, \\ \eta(\mathbb{E}[z]) &= \beta_0 + ta^\top + hb^\top. \end{aligned} \tag{1.10}$$

With assumptions on the joint distribution of the latent variables and the residual terms, this model can be estimated by maximizing the likelihood. The most commonly encountered outcome z in biomedical researches is normally or Bernoulli distributed, corresponding to an identity or logit link function η in (1.10). For a normally distributed z, the joint distribution of (x, y, z) is multivariate normal. The likelihood has the same form as that of the O2PLS model in (1.3), and a similar EM algorithm described in Section 1.2 can be used. For a binary outcome, the likelihood does not have an explicit form. While an EM algorithm can still be used, numerical integration is required to get approximate conditional expectations in the E steps, and nested iteration is needed to find a maximizer in the M steps. The computation for a binary outcome is therefore intense. In Chapter 5, this model will be formulated and studied.

1.3.4. FEATURE SELECTION

So far, we model an outcome via latent variables of omics, which are linear combinations of all the observed omic features. It is often of interest to prioritize features for further investigation aiming for biomarker development or drug targets. For integrative models (1.2), (1.3), and (1.10), the loadings in the matrices W and C indicate the relative importance of each observed omic feature for a corresponding latent variable. One can select the omic features with an absolute loading value above a certain threshold for further study. The selection of threshold here is arbitrary. Alternatively, one can add penalty functions to the likelihood function or estimating equation to obtain sparse loadings. A widely used penalty for this purpose is the least absolute shrinkage and selection operator (lasso) [4], which pushes small coefficients to exact zeros. Another useful penalty in the biomedical context is the group-wise L_2 penalty [24], which results in group-wise sparsity (i.e., features belonging to the same group will always be selected altogether). The groups are based on known biological knowledge, providing a way to augment the model with extra information and potentially improve statistical power [19, 24, 13]. For the PLS model in (1.2), lasso and group-wise L_2 penalty have been applied to get sparse loading vectors [26, 12, 13]. Sparse version of the O2PLS model in (1.3) will be developed in Chapter 2.

1.4. GENERAL OUTLINE OF THE THESIS

The remainder of this thesis is organized following the development of methodology for the outcome, from exploratory analysis of the outcome to modeling the outcome using two-stage approaches, and finally to joint modeling of the outcome with omics.

In Chapter 2, a sparse extension of the O2PLS method - group sparse O2PLS (GO2-PLS) - is developed. The method utilizes known group information among the features

to select relevant groups of features, by imposing group-wise penalties in the joint subspaces of omics. The method is implemented on methylation and glycomics from a population study, and regulomics and transcriptomics from a small case-control study of hypertrophic cardiomyopathy (HCM). In the latter, the relationship between the outcome HCM and the omics is explored visually using plots of the estimated joint scores, and the subsets of omic features selected are interpreted.

In Chapter 3, a novel way of modeling the outcome using genetic scores for omics is proposed. The genetic scores are constructed using various integrative methods and a newly proposed omic-PGS method. As mentioned in Section (1.3.2), the genetic scores for an omics can be computed without the omics data being measured. Only a fitted integrative model or GWAS summary statistics from previous studies is needed. We construct genetic scores for glycomics and metabolomics in two cohorts, and use the genetic scores to model BMI and type 2 diabetes.

In Chapter 4, we propose a two-stage approach to model down syndrome (DS) with methylation and glycomics. First, joint components representing methylation and glycomics are constructed using probabilistic O2PLS (PO2PLS). Each of these joint components is then used as pseudo-outcome and modeled via a linear mixed model with DS, age, sex as covariates and family as random effect. For the components that are significantly associated with DS, we identify the most important CpG sites and glycans and give interpretations.

In Chapter 5, a holistic model for joint modeling of an outcome and two omics, namely, GLM-PO2PLS is developed. The model identifiability is derived and EM algorithms to obtain maximum likelihood estimators of the parameters for the model with a normally or Bernoulli distributed outcome are developed. Test statistics are proposed to infer the association between the outcome and the omics, and their asymptotic distributions are derived. The method is applied on the same DS dataset as used in Chapter 4. The results are compared with previous studies as well as those in Chapter 4.

Finally in Chapter 6, a computationally more efficient two-stage EM algorithm is developed for the GLM-PO2PLS model with a binary outcome. Its relationship with the two-stage model in Chapter 4 is discussed. An extension of GLM-PO2PLS to allow omic-specific latent variables in the linear predictor of the outcome is proposed. The chapter concludes with other directions to incorporate various sources of information related to omics into outcome modeling.

1

BIBLIOGRAPHY

[1] el Bouhaddani, S. (2020). *Statistical integration of diverse omics data*. PhD thesis.

[2] de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263.

[3] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[4] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[5] Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3).

[6] el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., and Uh, H. W. (2016). Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, 17(2):S11.

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

[8] el Bouhaddani, S., Uh, H.-W., Jongbloed, G., and Houwing-Duistermaat, J. (2022). Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

[9] Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistics - Simulation and Computation*, 17(2):581–607.

[10] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

[11] Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing.

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

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

[14] Mackinnon, M. J. and Puterman, M. L. (1989). Collinearity in Generalized Linear Models. *Communications in Statistics - Theory and Methods*, 18(9):3463–3472.

[15] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Springer US, Boston, MA.

[16] Stefanski, L. A. and Carroll, R. J. (2007). Covariate Measurement Error in Logistic Regression.

[17] Tenenhaus, M. (2004). PLS Regression and PLS Path Modeling for Multiple Table Analysis. In *COMPSTAT 2004 — Proceedings in Computational Statistics*, pages 489–499. Physica, Heidelberg.

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

[19] Tyekucheva, S., Marchionni, L., Karchin, R., and Parmigiani, G. (2011). Integrating diverse genomic data using gene sets. *Genome Biology*, 12(10):R105.

[20] WOLD, H. (1973). Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. In *Multivariate Analysis–III*, pages 383–407.

[21] Wold, H. (1985). Partial Least Squares. *Encyclopedia of Statistical Sciences*, 6:581–591.

[22] Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. pages 286–293. Springer, Berlin, Heidelberg.

[23] Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103.

[24] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(1):49–67.

[25] Zierer, J., Menni, C., Kastenmüller, G., and Spector, T. D. (2015). Integration of 'omics' data in aging research: From biomarkers to systems biology. *Aging Cell*, 14(6):933–944.

[26] Zou, H. and Hastie, T. (2005). Erratum: Regularization and variable selection via the elastic net (Journal of the Royal Statistical Society. Series B: Statistical Methodology (2005) 67 (301-320)).

2

STATISTICAL INTEGRATION OF TWO OMICS DATASETS USING GO2PLS

Zhujie Gu, Said el Bouhaddani, Jiayi Pei, Jeanine Houwing-Duistermaat, Hae Won Uh. Statistical integration of two omics datasets using GO2PLS. *BMC Bioinformatics.* 2021; 22(1). DOI: 10.1186/s12859-021-03958-3

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. Since omics data are heterogeneous, the joint components in PLS will contain variation specific to each dataset. To account for this, Two-way Orthogonal Partial Least Squares (O2PLS) captures the heterogeneity by introducing 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 it might be of interest to identify a small subset relevant to the research question. To obtain sparsity, we extend O2PLS to Group Sparse O2PLS (GO2PLS) that utilizes biological information on group structures among variables and performs group selection in the joint subspace.

Results: The simulation study showed that introducing sparsity improved the feature selection performance. Furthermore, incorporating group structures increased robustness of the feature selection procedure. GO2PLS performed optimally in terms of accuracy of joint score estimation, joint loading estimation, and feature selection. We applied GO2PLS to datasets from two studies: TwinsUK (a population study) and CVON-DOSIS (a small case-control study). In the first, we incorporated biological information on the group structures of the methylation CpG sites when integrating the methylation dataset with the IgG glycomics data. The targeted genes of the selected methylation groups turned out to be relevant to the immune system, in which the IgG glycans play important roles. In the second, we selected regulatory regions and transcripts that explained the covariance between regulomics and transcriptomics data. The corresponding genes of the selected features appeared to be relevant to heart muscle disease.

Conclusions: GO2PLS integrates two omics datasets to help understand the underlying system that involves both omics levels. It incorporates external group information and performs group selection, resulting in a small subset of features that best explain the relationship between two omics datasets for better interpretability.

2.1. BACKGROUND

With the advancements in high throughput technology, multiple omics data are commonly available on the same subjects. To identify a set of relevant related features across the omics levels, these datasets need to be integrated and analyzed jointly. For statistical integration of omics data, there are several challenges to overcome: complex correlation structure within and between omics data, high-dimensionality ($p \gg n$, or "large p, small n"), heterogeneity between different omics datasets, and selection of relevant features in each dataset. To deal with the first two challenges, Partial Least Squares (PLS) has been proposed [5, 30]. Dimension reduction is achieved by decomposing two datasets X and Y into joint and residual subspaces. The joint (low-dimensional) subspace of one dataset represents the best approximation of X or Y based on maximizing the covariance of the two. However, by integrating two heterogeneous omics datasets, the PLS joint components also contain (strong) omic-specific variation. This heterogeneity can be caused by differences (e.g. between methylation and glycomics) in size, distribution, and measurement platform. Ignoring these omic-specific characteristics (variation specific to each of the data) in the model may lead to a biased representation of the underlying system. Two-way orthogonal partial least squares (O2PLS) [24, 9] 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. Hence, O2PLS accounts for the heterogeneity of two omics datasets. However, the resulting low-dimensional latent components spanning the joint subspaces are linear combinations of all the observed variables. Therefore, to select a small subset of relevant features for better interpretation, one can impose sparsity on the loadings of the principal components. A straightforward approach is to ignore all loadings smaller than some threshold value, effectively treating them as zero, which can be misleading [13].

Several sparse methods based on PLS have been proposed. Chun and Keleş proposed sparse PLS (SPLS) [8] which fits PLS on a reduced X space, consisting of preselected X-variables using a penalized regression. Sparse PLS (sPLS) by Lê Cao et al. [15] 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. Often it is of interest to select a group of features instead of individual features, e.g. features within a gene or a pathway. By so doing, one can improve power by identifying aggregate effects of the selected features [25, 31, 16]. Liquet et al. extended sPLS to group PLS (gPLS) [16], 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 to extend O2PLS to incorporate sparsity, called Group Sparse O2PLS (GO2PLS). GO2PLS obtains sparse solutions by pushing a large number of small non-zero weights (or loading values) to zeros, instead of employing hard thresholding using arbitrary cut-off values. Therefore, GO2PLS constructs joint low-dimensional latent components representing the underlying systems involving both omics levels while taking into account the heterogeneity of different omics data, incorporates external biological information such as known group structure, and performs variable selection by imposing group-wise penalties on the loading vectors in the joint subspaces.

For illustration, we apply GO2PLS to datasets from two studies. Firstly, TwinsUK is a population based study [19, 17], where methylation (482K CpG sites) and 22 immunoglobulin G (IgG) glycans were measured. A previous research [27] suggested the presence of an indirect influence of methylation on IgG glycosylation that may in part capture environmental exposures. We integrate the two omics datasets, aiming to identify genes of CpG sites affecting IgG glycosylation. In the CVON-DOSIS case-control study [CVO], regulomics (histone modification) and transcriptomics data were measured on 13 hypertrophic cardiomyopathy (HCM) patients and 10 controls. Histone modification can have an impact on gene expression. Therefore we integrate the two omics datasets and identify a small set of regulatory regions and transcripts explaining this relationship. Moreover, the extreme imbalance in a high-dimensional setting (33K ChIP-seq and 15K RNA-seq vs 23 subjects) poses computational challenges. The resulting selected features are further studied using gene set enrichment analysis [21]. Several possible scenarios containing these characteristics are designed and investigated in an extensive simulation study.

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 methylation and glycomics in the TwinsUK study and regulomics and transcriptomics in the CVON-DOSIS study. We conclude with a discussion and possible directions to further extend the method.

2.2. METHODS

2.2.1. DATA DESCRIPTION

TWINSUK DATASETS

Whole blood methylation (using Infinium HumanMethylation450 BeadChip) and IgG glycomics (Ultra Performance Liquid Chromatography) data were measured on 405 independent 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 [7] were removed. We kept only the CpG sites on CpG islands or surrounding areas (shelves and shores) that mapped to genetic regions. Age, sex, batch effect, and cell counts were corrected for using multiple regression. The glycomics dataset contains 22 glycan peaks. These peaks were normalized using median quotient (MQ) normalization [26], log-transformed, and adjusted for batch effect, age, and sex as well. The remaining 126299 CpG sites were then divided into 16892 groups based on their target genes (biological information from the UCSC database [14, UCS]). No group information was available for the glycomics data.

CVON-DOSIS DATASETS

In the CVON-DOSIS study, regulomics and transcriptomics datasets were measured on the samples taken from the heart tissues of 13 HCM patients and 10 healthy controls. HCM is a heart muscle disease that makes it harder for the heart to pump blood, leading to heart failure. The regulomics data were measured using ChIP-seq, providing counts

of histone modification H3K27ac in 33642 regulatory regions. The transcriptomics data contain counts of 15882 transcripts, measured by RNA-seq. The raw counts of regulomics data were normalized with reads per kilobase million (RPKM) to adjust for sequencing depth. Transcriptomics data were normalized with counts per million (CPM) with effective library size (estimated using the TMM method in EdgeR R package [18]). Further, both normalized data were log-transformed.

2.2.2. 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:

$$X = TW^\top + T_\perp P_\perp^\top + E,$$
$$Y = UC^\top + U_\perp Q_\perp^\top + F.$$

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

$$U = TB_T + H,$$
$$T = UB_U + \tilde{H}.$$

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:

$$\tilde{X} = (I_N - T_\perp (T_\perp^\top T_\perp)^{-1} T_\perp^\top) X,$$
$$\tilde{Y} = (I_N - U_\perp (U_\perp^\top U_\perp)^{-1} U_\perp^\top) Y,$$

where $T_\perp$ $U_\perp$ are estimates for the orthogonal subspaces, and I_N is identity matrix of size N. For more details see [24]. 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 $\tilde{Y}^\top \tilde{X}$ [9]. 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 $k = 1, \ldots, K$:

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

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 [29] algorithm, which starts with random initialization of the X-space score vector t and repeats a sequence of the following steps until convergence:

$$1)\ c_k = \frac{\tilde{Y}_k^\top t}{t^\top t}, \quad 2)\ \|c_k\|_2 \to 1, \quad 3)\ u = \tilde{Y}_k c_k,$$
$$4)\ w_k = \frac{\tilde{X}_k^\top u}{u^\top u}, \quad 5)\ \|w_k\|_2 \to 1, \quad 6)\ t = \tilde{X}_k w_k.$$

In step 1 and 4, Y_k and X_k are projected onto the X-space score vector t and the Y-space score u to get the loading vectors c_k and w_k. The loading vectors are then unitized (step 2 and 5) and used to calculated the new scores u and t. Convergence of the algorithm is guaranteed. A detailed description and proof of optimality of the O2PLS algorithm can be found in [24, 9].

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 criterion for all three parameters. As in [9], 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 $Y \to X$.

2.2.3. 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:

$$\max_{\|c_k\|_2=1, \|w_k\|_2=1} c_k^\top \tilde{Y}_k^\top \tilde{X}_k w_k + \lambda_c \|c_k\|_1 + \lambda_w \|w_k\|_1, \tag{2.1}$$

where λ_c, λ_w are penalization parameters that regulate the sparsity level. The optimization problem (2.1) can be solved [28] by iterating over the k-th pair of joint loadings,

$$c_k = \frac{S(\tilde{Y}_k^\top t, \lambda_c)}{\left\| S(\tilde{Y}_k^\top t, \lambda_c) \right\|_2}, \ w_k = \frac{S(\tilde{X}_k^\top u, \lambda_w)}{\left\| S(\tilde{X}_k^\top u, \lambda_w) \right\|_2}, \tag{2.2}$$

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:

$$\begin{aligned}\min_{c_k^{(m)}, w_k^{(j)}} \Bigg\{ & - \sum_{j=1}^{J} \sum_{m=1}^{M} {c_k^{(m)}}^\top {\tilde{Y}_k^{(m)}}^\top \tilde{X}_k^{(j)} w_k^{(j)} \\ & + \lambda_c \sum_{m=1}^{M} \sqrt{q_m} \left\| c_k^{(m)} \right\|_2 + \lambda_w \sum_{j=1}^{J} \sqrt{p_j} \left\| w_k^{(j)} \right\|_2 \\ & + \phi_c \left(\sum_{m=1}^{M} \left\| c_k^{(m)} \right\|_2^2 - 1 \right) + \phi_w \left(\sum_{j=1}^{J} \left\| w_k^{(j)} \right\|_2^2 - 1 \right) \Bigg\},\end{aligned} \tag{2.3}$$

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 Section 2.6.1). The solution takes the form:

$$\begin{aligned} c_k^{(m)} &= \frac{\left(\left\| {\tilde{Y}_k^{(m)}}^\top t \right\|_2 - \sqrt{q_m}\lambda_c \right)_+}{2\phi_c \left\| {\tilde{Y}_k^{(m)}}^\top t \right\|_2} \quad {\tilde{Y}_k^{(m)}}^\top t, \\ w_k^{(j)} &= \frac{\left(\left\| {\tilde{X}_k^{(j)}}^\top u \right\|_2 - \sqrt{p_j}\lambda_w \right)_+}{2\phi_w \left\| {\tilde{X}_k^{(j)}}^\top u \right\|_2} \quad {\tilde{X}_k^{(j)}}^\top u. \end{aligned} \tag{2.4}$$

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., (2.3) reduces to (2.1)). 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 (denote 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 [22], the "one-standard-error-rule" [12] can be applied to obtain a more stable CV result. The GO2PLS algorithm is described below:

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 (2.4) (or Formula (2.2) 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) $T = [t_1, \ldots, t_K]$; $U = [u_1, \ldots, u_K]$;
$W = [w_1, \ldots, w_K]$; $C = [c_1, \ldots, c_K]$.

2.3. SIMULATION STUDY

We evaluate the performance of GO2PLS in two scenarios. First, we investigate the ability to select the relevant groups under various scenarios, focusing on the joint subspace, where the group selection takes place. Second, we compare the performance of GO2PLS and SO2PLS with other methods: O2PLS, PLS, sPLS, and gPLS. We investigate joint score estimation, joint loading estimation, and feature selection performances.

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. 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 and U from Equation 2.2.2 are both standard normally distributed and have correlation 1). We perform 1000 simulation runs and record the number of the runs GO2PLS selected relevant groups; we compute the proportion of each truly relevant group (with non-zero loadings) being selected across the simulation runs (number of times being selected divided by 1000). The group importance measurement $\left\| X^{(j)\top} U \right\|_2 / \sqrt{p_j}$, that determines whether a group is selected or not is recorded for the first 5 groups (with loading value 1) to investigate the stability of the selection procedure.

In the second scenario, we vary the sample size N from 30 to 600, and 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 $T, T_\perp, U, U_\perp$ from Equation 2.2.2 are generated from normal distributions with zero mean. The relationship between the joint scores is represented by $U = T + H$, where H accounts for 20% of the variation in U. The noise matrices E, F are generated from normal distributions with zero mean and variance such that the variance of the noise matrix accounts for a proportion α ($0 < \alpha < 1$) of the variance of the data matrix (i.e., $\alpha = \text{Var}(E)/\text{Var}(X) = \text{Var}(F)/\text{Var}(Y)$). The 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 cosine of 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.

An overview of scenario settings is presented in Table 2.1, 2.2. To make a clearer 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).

2.3.1. RESULTS OF SIMULATION STUDY

Table 2.1: **Settings of Scenario 1** to study the performance of selecting relevant groups

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

The selection proportion is the number of times a relevant group being selected divided by the number of simulation runs. The $\left\|X^{(j)\top}U\right\|_2/\sqrt{p_j}$ is a measurement of group importance. It provides more information on the stability of the group selection procedure. We simulate groups with varying sizes to investigate the influence of group size on the group selection performance of GO2PLS.

Table 2.2: **Settings of Scenario 2** to compare the performances regarding joint score estimation, joint loading estimation, and feature selection

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	$[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$

$R^2_{\widehat{T}T}$ and $R^2_{\widehat{T}\widehat{U}}$ quantify the joint score estimation performance; TPR measures the feature selection performance; $W^\top\widehat{W}$ quantifies the joint loading estimation performance. The dimensions and number of relevant features are set based on the CVON-DOSIS study. Sample size N, the relative strength of orthogonal signal ($\sigma^2_{T_\perp}/\sigma^2_T$), and noise level α are varied.

SCENARIO 1

Figure 2.1 shows the selection proportion for each relevant group under each noise level. Compared to smaller groups, the proportion 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. Figure 2.2 shows the density of the group importance measurement $\left\|X^{(j)\top}U\right\|_2/\sqrt{p_j}$ 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 exceeds the threshold, the total area on the right side of the threshold under each density curve equals the selection proportion for the corresponding group. The measurement for larger relevant group shows higher precision at all noise levels. The threshold increases along with the noise.

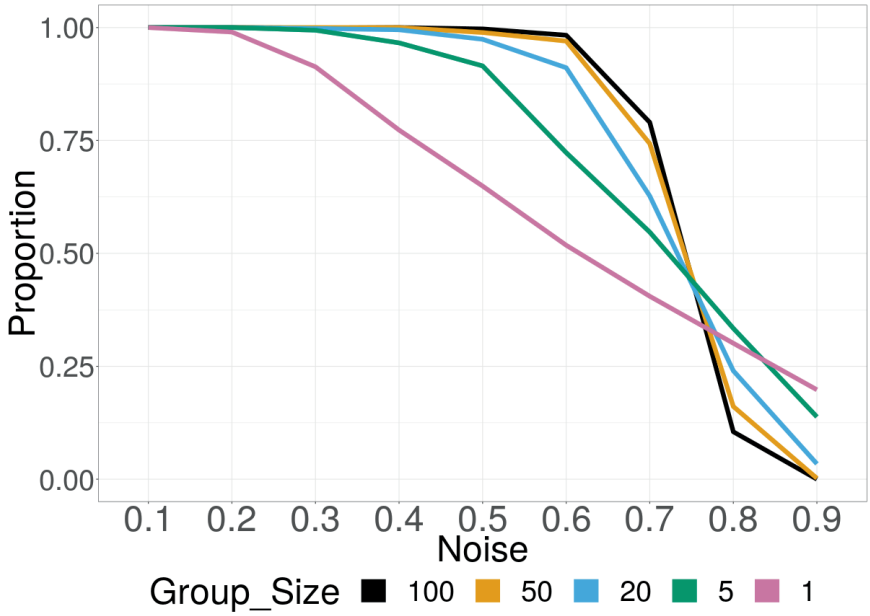


Figure 2.1: **Simulation Scenario 1: Selection proportion of relevant groups with different sizes under varying noise.** The proportion for larger groups is higher at low to moderate ($\alpha < 0.7$) noise levels, and shows robustness against increasing noise.

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). The top row of Figure 2.3 shows the performance measured by $R^2_{\hat{T}T}$ & $R^2_{\hat{T}\hat{U}}$ 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, $R^2_{\hat{T}T}$ of the various methods 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, $R^2_{\hat{T}\hat{U}}$ 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.

The bottom row of Figure 5.3 shows the score estimation performance under fixed relative orthogonal signal strength of 1, $\alpha = 0.1$, 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

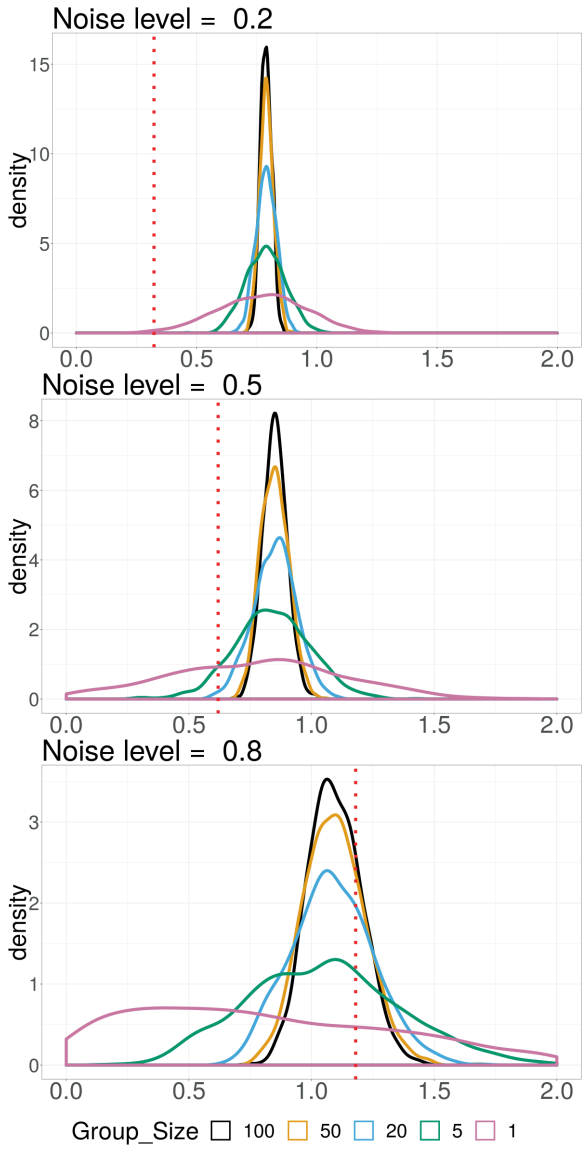


Figure 2.2: **Simulation Scenario 1: Density plot of estimated group importance measurement $\left\|X^{(j)\top}U\right\|_2/\sqrt{p_j}$ for each group size under 3 different noise levels.** The vertical dotted red line is the average threshold. When the measurement of a group is larger than the threshold, the group is selected. The total area on the right side of the threshold under each density curve equals to the selection proportion for the corresponding group. The less the density curve spreads out, the more stable is the estimate.

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

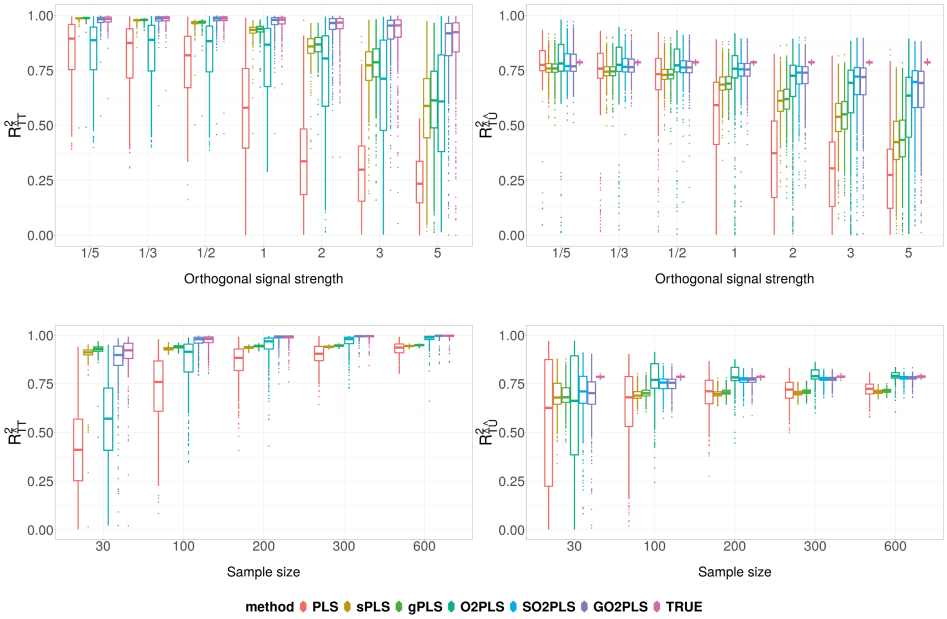


Figure 2.3: **Simulation Scenario 2: comparison of joint score estimation performance,** under varying relative orthogonal signal strength (top row), and varying sample size (bottom row). On the Y-axis, $R^2_{\hat{T}T}$ (left) and $R^2_{\hat{T}\hat{U}}$ (right) are the coefficient of determination of regressing T on $\hat{T}$, and $\hat{U}$ on $\hat{T}$, respectively, quantifying the joint score estimation performances. Boxes show the results of 500 repetition.

GO2PLS, SO2PLS, and O2PLS, respectively. In Figure 2.4, the top row shows the TPR and $W^\top \hat{W}$ under $N = 30$ and varying noise levels α from low to high. At all noise levels, GO2-PLS 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 the bottom row, 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.

2.4. APPLICATION TO DATA

We demonstrate SO2PLS and GO2PLS on datasets from two distinct studies. In the TwinsUK study, our aim is to integrate methylation and glycomics data and identify important groups of CpG sites underlying glycosylation. In the CVON-DOSIS study, we integrate regulomics and transcriptomics data and select a subset of genes and regions that drive their relationship.

2.4.1. TWINSUK STUDY

We performed GO2PLS on the data with 1 joint, no methylation-orthogonal, and 3 glycomics-orthogonal components based on 5-fold cross-validation. We set the sparsity parameters to select the top 100 groups in the methylation and kept all the 22 glycan variables.

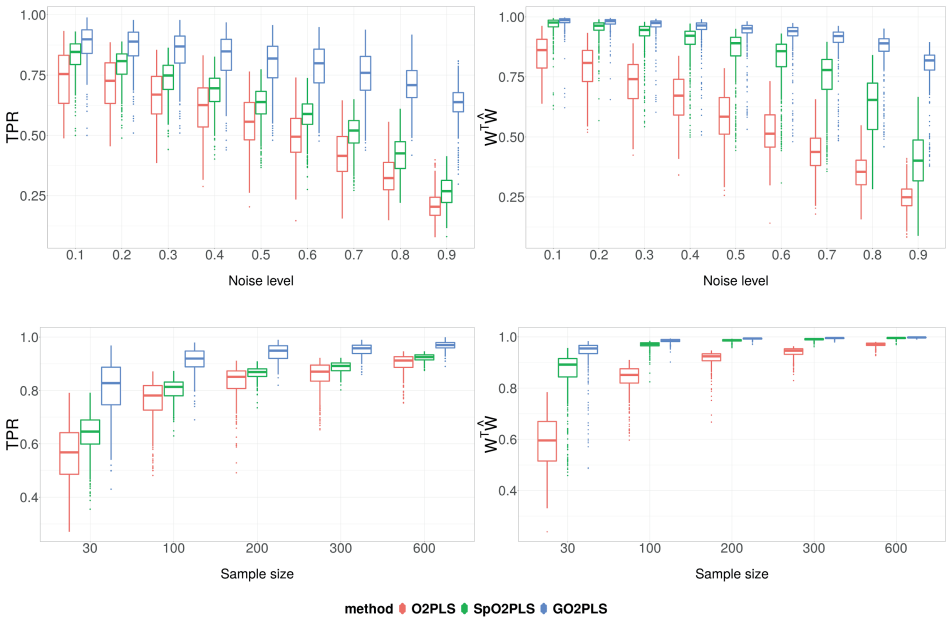


Figure 2.4: **Simulation Scenario 2: comparison of feature selection and joint loading estimation performance,** under varying noise level (top row), and varying sample size (bottom row). On the Y-axis are the True Positive Rate (left) and $W^\top \hat{W}$ (right), which is the cosine of the angle between the estimated loading vector $\hat{W}$ and the true one W. Boxes show the results of 500 repetition.

The selected CpG groups from GO2PLS were mapped to their targeted genes for interpretation.

We performed gene set enrichment analyses on the selected genes using the ToppGene Suite [6]. The results appeared to be related to immune response. We listed the most significant molecular function, biological process, and pathway in Table 2.3.

2.4.2. CVON-DOSIS STUDY

We applied SO2PLS on the regulomics and transcriptomics datasets, with 2 joint and 1 orthogonal components for each omics dataset. In each pair of the joint components, 1000 regulomics and 500 transcriptomics variables were selected. We then further identified 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). These genes are of interest since they are likely to be related to epigenetic regulation of gene expression. Genes corresponding to the selected transcripts were also identified. These gene sets identified from each joint component of the two omics data were investigated separately using gene set enrichment analysis. The top results were listed in Table 2.4. The GO analysis of the selected genes and regions showed terms related to HCM that were also found previously [11]. Due to the presence of the case-control status in both omics levels, we expect the joint components related to the disease. Plotting the joint scores of the two datasets showed a separation between HCM cases and

Table 2.3: **TwinsUK study: top results of gene set enrichment analysis**

GO2PLS	Name	pValue	FDR B&H
GO: Molecular Function	peptide antigen binding	1.42E-06	5.45E-04
GO: Biological Process	homophilic cell adhesion via plasma membrane adhesion molecules	1.82E-10	4.20E-07
	cell-cell adhesion via plasma-membrane adhesion molecules	5.46E-10	6.28E-07
	cell-cell adhesion	3.43E-07	2.63E-04
	interferon-gamma-mediated signaling pathway	1.01E-05	5.83E-03
Pathway (Source: KEGG)	Viral myocarditis	8.00E-08	9.60E-06
	Staphylococcus aureus infection	1.32E-06	7.92E-05
	Allograft rejection	3.77E-06	1.51E-04
	Graft-versus-host disease	5.54E-06	1.66E-04
	Type I diabetes mellitus	7.05E-06	1.69E-04
	Autoimmune thyroid disease	2.00E-05	3.66E-04
	Rheumatoid arthritis	2.14E-05	3.66E-04

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 [3, 20].

controls (Figure 2.5). For a comparison of score plots of PCA, PLS, O2PLS, and SO2PLS, please see Section 2.6.2.

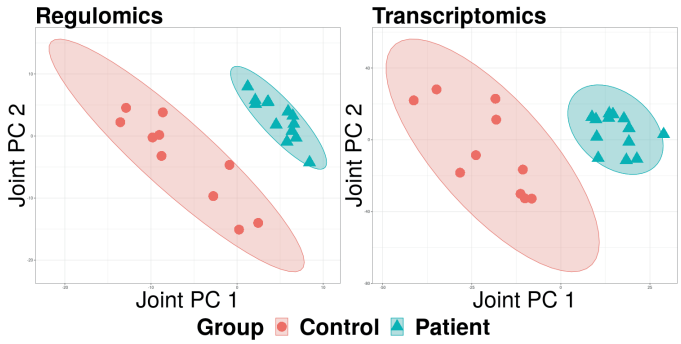


Figure 2.5: **CVON-DOSIS study: SO2PLS joint score plots** of regulomics (left) and transcriptomics (right). HCM patients and controls were plotted in different colors. Ellipses are the 95% confidence regions of each group.

Table 2.4: **CVON-DOSIS study: Gene set enrichment analysis results**

Joint component 1 - Regulomics	Name	pValue	FDR B&H
GO: Biological Process	muscle structure development	3.42E-08	2.09E-04
	muscle tissue development	1.43E-07	4.37E-04
	actin cytoskeleton organization	3.35E-07	6.70E-04
	cytoskeleton organization	4.40E-07	6.70E-04
	regulation of cellular response to stress	8.14E-07	9.93E-04
	striated muscle tissue development	1.20E-06	1.19E-03
	actin filament-based process	1.36E-06	1.19E-03
	organ growth	4.31E-06	3.28E-03
	heart development	5.85E-06	3.96E-03
GO: Cellular Component	contractile fiber	2.19E-07	1.17E-04
	myofibril	3.33E-07	1.17E-04
	I band	1.54E-06	3.60E-04
	Z disc	2.42E-06	4.25E-04
	sarcomere	4.82E-06	6.77E-04

Joint component 1 - Transcriptomics	Name	pValue	FDR B&H
GO: Biological Process	blood circulation	1.88E-08	4.19E-05
	circulatory system process	2.64E-08	4.19E-05
	regulation of system process	2.76E-08	4.19E-05
	ion transport	3.28E-08	4.19E-05
	positive regulation of developmental process	2.58E-07	2.63E-04
	neurogenesis	3.66E-07	2.70E-04
	heart contraction	3.71E-07	2.70E-04
Disease (Source: DisGeNET Curated)	Myocardial Failure	6.57E-09	1.20E-06
	Congestive heart failure	6.57E-09	1.20E-06
	Heart failure	6.57E-09	1.20E-06
	Left-Sided Heart Failure	6.57E-09	1.20E-06
	Heart Failure, Right-Sided	6.57E-09	1.20E-06

Joint component 2 - Regulomics	Name	pValue	FDR B&H
GO: Molecular Function	RNA binding	1.91E-19	2.63E-16
	unfolded protein binding	4.03E-09	2.17E-06
	catalytic activity, acting on DNA	4.74E-09	2.17E-06
	catalytic activity, acting on a tRNA	2.20E-08	7.57E-06

GO: Biological Process	cellular protein-containing complex assembly	2.67E-24	1.74E-20
	RNA processing	1.03E-15	3.36E-12
	ribonucleoprotein complex biogenesis	2.49E-15	5.41E-12
	amide biosynthetic process	9.83E-14	1.60E-10
	translational elongation	2.98E-13	3.24E-10
	translation	2.98E-13	3.24E-10
Pathway (BioSystems REACTOME)	Gene Expression	4.96E-14	6.31E-11

Joint component 2 - Transcriptomics	Name	pValue	FDR B&H
GO: Molecular Function	receptor antagonist activity	9.72E-09	7.44E-06
	receptor inhibitor activity	7.44E-08	2.85E-05
	signaling receptor activity	5.11E-05	1.06E-02
GO: Biological Process	negative regulation of execution phase of apoptosis	7.40E-10	2.75E-06
	vascular endothelial growth factor production	1.24E-09	2.75E-06
	regulation of vascular endothelial growth factor production	1.83E-08	2.02E-05
	cell-cell adhesion via plasma-membrane adhesion molecules	7.23E-08	6.40E-05
	positive regulation of cytokine biosynthetic process	8.79E-08	6.49E-05

Results from the gene set enrichment analysis using ToppGene on the selected genes and regions. In the upper two tables, the first joint regulomics and transcriptomics component is shown, respectively. The lower two tables are about the second joint components.

2.5. DISCUSSION AND CONCLUSION

Statistical integration of two omics datasets is becoming increasingly popular to gain insight into underlying biological systems. O2PLS is a method that integrates two heterogeneous datasets and takes into account omic-specific variation. The resulting joint and specific components are linear combinations of all variables, making interpretation difficult. To introduce sparsity and identify relevant groups, GO2PLS incorporates biological information on group structures to perform group selection in the joint subspace. Depending on the group size, such an approach may also lead to a higher selection probability of relevant features. We performed an extensive simulation study and showed that O2PLS-based methods generally outperformed PLS-based methods regarding joint score estimation when orthogonal variation was present in the data. Since PLS does not take into account orthogonal parts, the joint components also include part of the orthogonal variation. Further, when the sample size was small or the noise level

was high, penalized methods appeared to be much less prone to overfitting than non-penalized methods. This suggests that results based on GO2PLS are likely to be generalizable when applied to new datasets. Concerning feature selection, adding external group information led to higher TPR, and larger groups of relevant features had a higher proportion of being detected under a moderate noise level. We then applied GO2PLS to the TwinsUK study, where we selected 100 target genes comprising of CpG sites that are most related to IgG glycosylation. The results of the enrichment analysis on the selected genes showed GO-terms involving the immune system in which the IgG glycans play important roles. In the CVON-DOSIS study, we integrated regulomics and transcriptomics and identified 1000 regulatory regions and 500 transcripts, and mapped them to genes. Further analysis of the selected gene sets showed enrichment for terms related to heart muscle diseases. Moreover, the implementation of GO2PLS is computationally fast and memory efficient. It relies on an algorithm based on NIPALS that does not store large matrices of size $p \times q$ when performing the group-penalized optimization. A regular laptop (8G RAM, quad-core 2.6 GHz) was able to run GO2PLS on omics data from both case studies.

The group information should be chosen together with domain experts based on the research question and biological knowledge. For example, in our TwinsUK data application, we aimed to identify the genes comprising of CpG sites, rather than the individual CpG sites. Therefore, we grouped CpG sites in the same genetic region. Furthermore, the biological knowledge that close-by CpG sites tend to function together supported the choice of grouping. Different grouping information leads to a changed definition of groups, consequently the selected groups will have a different interpretation. An extra analysis in the TwinsUK study was performed using another grouping strategy. We grouped 55531 CpG sites that map to the promoter region (0-1500 bases upstream of the transcriptional start site (TSS)) of a gene to 14491 groups based on their targeted genes. We applied GO2PLS and selected 100 groups. Note that the size of these groups was smaller, and many CpG sites in gene bodies are excluded. Enrichment analysis did not result in significant results, supposedly due to weaker aggregated group effects. When the research goal is to identify individual features (e.g., in our CVON-DOSIS data application), or group information is not available, SO2PLS can be used.

In the CVON-DOSIS study, Plotting the first two joint components showed two distinct classes corresponding to the case-control status. This might be expected since the analysis was conditional on case-control status, yielding a correlation between the two omics datasets. This phenomenon is well known in regression analysis of secondary phenotypes [23], but not well studied in PLS type of methods. This is a topic of future research. Often omics data are collected to study their relationship with an outcome variable or to predict an outcome variable. To this end, our approach has to be extended to incorporate the outcome variable. Such an approach might also lead to a more sparse solution since the selected features have to be correlated among the three datasets. Further extensions of GO2PLS are to incorporate more than two omics datasets to represent the actual biological system even better.

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 [4], and PPLS [10] was proposed to provide a probabilistic frame-

work for PLS. It has been shown that the probabilistic counterpart has a lower bias in estimation and is robust to non-normally distributed variables [10]. More importantly, the probabilistic model will allow statistical inference, making it possible to interpret the relevance and importance of features in the population, and facilitating follow-up studies. These extensions of GO2PLS will be suited for various studies with more complicated designs.

To conclude, GO2PLS estimates joint latent components that represent underlying systems by integrating two omics data while taking into account the heterogeneity between different omics levels. It incorporates external information on group structures to perform group selection, leading to better interpretation.

2.6. SUPPLEMENTARY MATERIALS FOR CHAPTER 2

2.6.1. SOLVING THE OPTIMIZATION PROBLEM OF GO2PLS

The optimization problem of GO2PLS is block multi-convex. It can be solved by optimizing one block at a time, holding the others fixed [12]. The optimization problem for $w_k^{(j)}$ is:

$$\min_{w_k^{(j)}} \left\{ -c_k^\top \tilde{Y}_k^\top \tilde{X}_k^{(j)} w_k^{(j)} + \lambda_w \sqrt{p_j} \left\| w_k^{(j)} \right\|_2 + \phi_w \left(\left\| w_k^{(j)} \right\|_2^2 - 1 \right) \right\}.$$

We differentiate with regard to $w_k^{(j)}$ and set the derivative to 0, and solve for $w_k^{(j)}$:

$$-\tilde{X}_k^{(j)\top} u + \sqrt{p_j} \lambda_w s_j + 2\phi_w w_k^{(j)} = 0, \tag{2.5}$$

where s_j is the subdifferential of $\left\| w_k^{(j)} \right\|_2$ that has the following form:

$$s_j = \begin{cases} \frac{w_k^{(j)}}{\left\| w_k^{(j)} \right\|_2}, & \text{if } w_k^{(j)} \neq \mathbf{0} \\ \text{Any vector with } \left\| \widehat{s_j} \right\|_2 \in [-1,1], & \text{Otherwise.} \end{cases}$$

Rearranging equation (2.5), we have

$$w_k^{(j)} = \frac{\tilde{X}_k^{(j)\top} u - \sqrt{p_j}\lambda_w s_j}{2\phi_w} = \begin{cases} \mathbf{0}, & \left\| \tilde{X}_k^{(j)\top} u \right\|_2 \in [-\sqrt{p_j}\lambda_w, \sqrt{p_j}\lambda_w] \\ (\tilde{X}_k^{(j)\top} u - \sqrt{p_j}\lambda_w \frac{w_k^{(j)}}{\left\| w_k^{(j)} \right\|_2})/2\phi_w, & \text{Otherwise.} \end{cases}$$

If $w_k^{(j)} \neq \mathbf{0}$, we have

$$w_k^{(j)} = \frac{\left\| w_k^{(j)} \right\|_2}{2\phi_w \left\| w_k^{(j)} \right\|_2 + \sqrt{p_j}\lambda_w} \tilde{X}_k^{(j)\top} u. \tag{2.6}$$

By taking L_2 norm at both sides of (2.6), we can solve for $\left\| w_k^{(j)} \right\|_2$. Substituting it back to equation (2.6), we get

$$w_k^{(j)} = \frac{\left\| \tilde{X}_k^{(j)\top} u \right\|_2 - \sqrt{p_j}\lambda_w}{2\phi_w \left\| \tilde{X}_k^{(j)\top} u \right\|_2} \tilde{X}_k^{(j)\top} u.$$

Combining with the case when $w_k^{(j)} = \mathbf{0}$, we have the general form

$$w_k^{(j)} = \frac{\left(\left\| \tilde{X}_k^{(j)\top} u \right\|_2 - \sqrt{p_j}\lambda_w \right)_+}{2\phi_w \left\| \tilde{X}_k^{(j)\top} u \right\|_2} \tilde{X}_k^{(j)\top} u.$$

Similarly, the solution for $c_k^{(m)}$ can be obtained.

2.6.2. ADDITIONAL ANALYSIS OF CVON-DOSIS STUDY

Due to the case-control study design, we expected that the first few principal components of both datasets that explain most of the variance in the data are related to the disease status. We performed PCA on each dataset separately and plotted the scores of each data (Figure 2.6a) with different colors for the patients and controls. Though the 95% confidence regions of the two groups overlap, the scores of both data separated the groups quite well. We further investigated if the joint components that explain the covariance between the datasets were also related to the disease. We integrated the two omics data using PLS (2 joint components, Figure 2.6b), O2PLS (2 joint and 1 orthogonal components for each omics dataset, Figure 2.6c), and SO2PLS (Figure 2.6d). The group separation by the first 2 joint components appeared to be clearer comparing to PCA, especially when the data-specific variation was taken into account (i.e., in O2PLS and SO2PLS). More research is needed to quantify the performance of group separation and compare across methods.

The methods we applied are all unsupervised. It is interesting to incorporate the disease status in the model. It is our future work to develop supervised integrative approaches. For more discussions on future directions, please refer to the Discussion section in the article.

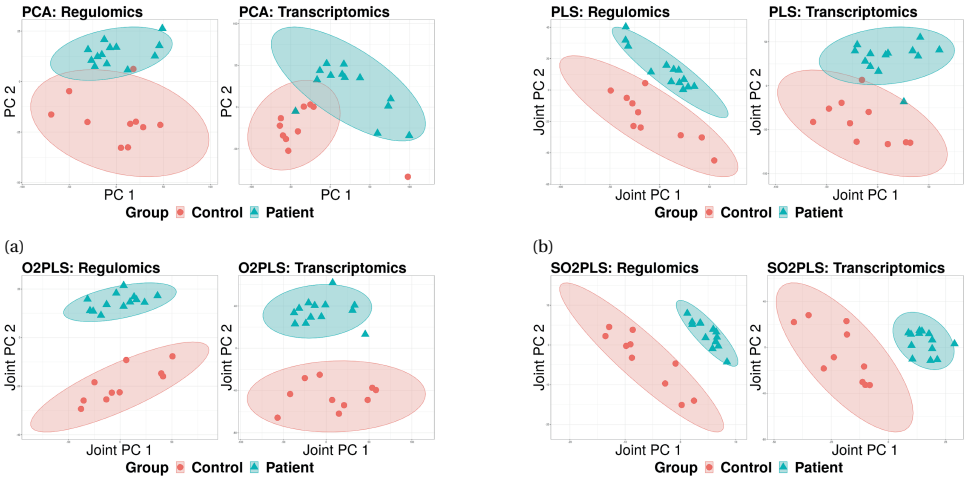


Figure 2.6: **Score plots of PCA (a), PLS (b), O2PLS (c), and SO2PLS (d); In each subplot, regulomics is on the left, and transcriptomics on the right.** PCA is non-integrative, hence performed on regulomics and transcriptomics separately; PLS, O2PLS, and SO2PLS were applied on the two omics datasets jointly. HCM patients and controls were plotted in different colors. Ellipses are the 95% confidence regions of each group.

BIBLIOGRAPHY

[CVO] CVON-DOSIS – Cardiovascular Research Consortium.

[UCS] UCSC Genome Browser Home.

[3] Benjamini, Y. and Hochberg, Y. (1995). 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.

[4] Bishop, C. M. and Tipping, M. E. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B*, 61(iii):611–622.

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

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

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

[8] Chun, H. and Keleş, S. (2010). 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.

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

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

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

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

[13] Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

[14] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006.

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

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

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

[18] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Technical report.

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

[20] Storey, J. D. (2002). A direct approach to false discovery rates. Technical Report 3.

[21] 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., and Mesirov, J. P. (2005). 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.

[22] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

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

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

[25] Tyekucheva, S., Marchionni, L., Karchin, R., and Parmigiani, G. (2011). Integrating diverse genomic data using gene sets. *Genome Biology*, 12(10):R105.

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

[27] 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., and Gieger, C. (2018). IgG glycosylation and DNA methylation are interconnected with smoking. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1862(3):637–648.

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

[29] WOLD, H. (1973). Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. In *Multivariate Analysis–III*, pages 383–407.

[30] Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J. (1984). 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.

[31] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(1):49–67.

3

OMIC-SCORES: A GENOMIC REPRESENTATION OF OMICS FOR OUTCOME MODELING

3.1. INTRODUCTION

Advances in high-throughput genomic technologies have resulted in a large amount of genomic data available in many studies. Many disease-associated genetic variants have been identified. However, these genetic variants do not directly cause diseases at the molecular level. They may affect other intermediate omic levels (such as glycomics, proteinomics, and metabolomics) that in turn induce molecular and physiological changes [43]. To better understand the underlying genetic architecture of diseases, the variation of these intermediate omic levels that is related to the genomics (namely the heritable part) needs to be incorporated into studies. The heritable part of omics is stable over lifespan, and can be predicted using genomic data. Moreover, to use this information in a study of an outcome disease, it is tempting to have a low-dimensional representation (such as a few scores) of the high-dimensional and correlated omics. In this paper, we propose methods to construct stable low-dimensional scores for omics data using genomics (which we call 'omic-score'), and implement the methods on glycomics and metabolomics, and incorporate the constructed omic-scores in studies of body mass index (BMI) and type 2 diabetes (T2D).

A widely used approach to estimate an individual's genetic liability to a trait is polygenic score (PGS), which is calculated as a weighted sum of trait-associated alleles. To compute a PGS, two elements are required: a method to estimate the weights for all the alleles, and a relevant GWAS study on which the weight estimates are based. Many methods for constructing a PGS were proposed, such as the most classic clumping and thresholding (C+T) and its extension stacked clumping and thresholding (SCT) [33], the shrinkage methods lassosum [22], SBLUP [35], the Bayesian methods MegaPRS [53], LDpred2 [32], DBSLMM [50] etc. Details can be found in the recent review papers [30, 28, 6]. GWASs for omic levels have been conducted as well in the past decade (e.g., for glycomics [40], proteinomics [42], and metabolomics [41]), hence it is possible to construct omic-score for each omic feature using PGS methods. The PGSs for each omic feature cannot be directly entered as covariates in a regression model for the disease outcome. Rather, a dimension reduction step is needed. For example, the first few principal components of all PGSs can be used as a low-dimensional representation of the genetics underlying the omics. These principal components are then taken as the covariates instead of the PGSs. We denote this two-stage approach 'omic-PGS', which performs standard PGS approach on each omic feature followed by a PCA step. Note that, even if PCA is a multivariate method, the PGSs are constructed based on univariate associations, ignoring the correlation between the original omic features.

This leads to the question whether PGS methods are optimal for (high dimensional) omics data. First of all, omic features often have a correlation structure that cannot be neglected [38]. Shen et al. [37] incorporated the correlation among omic features into the conventional GWAS and proposed a multi-phenotype method. However, the summary statistics produced can not be used for computing a PGS as the p-values and effect sizes are based on different combinations of phenotypes for each allele. Second, for high-dimensional omics data such as methylation, typically only a small subset of methylation CpG sites were studied in GWAS, and hence not possible to construct omic-scores for all the CpG sites.

As alternative approach, integrative methods can be used for dimension reduction

and to jointly analyse the genomic data X and the omic data Y. They take into account the correlation structure in both genomics and omics and can be used to construct a few scores for the whole omics dataset. Partial least squares (PLS) [3, 48] is a widely used integrative method that decomposes two datasets X and Y into joint and residual subspaces. The low-dimensional joint subspace of one dataset represents the best approximation of X or Y based on maximizing the covariance of the two. When applied to integrate genomics and omics data, the joint genomic components that correlate with the omics data can be a candidate of omic-score. However, genomic data contain rich information on many biological processes. When integrated with a specific omic dataset, the genomic variation uncorrelated with this omic dataset needs to be separated from the joint subspace of PLS so that the joint subspace captures the true relationship between the genomics and the omics dataset. Two-way orthogonal partial least squares (O2PLS) [45, 8] was developed to correct for this data-specific variation, hence is better suited for the task of estimating omic-score from genomics. We also consider various extensions of O2PLS, namely, sparse O2PLS (SO2PLS) [12], which imposes penalization on the combination weights so that the omic-scores are constructed from a small subset of relevant alleles, and probabilistic O2PLS (PO2PLS) [10], which is a likelihood-based method.

The rest of the paper is organized as follows. In Section 3.2.1, we first introduce the datasets and the study cohorts, and describe the preprocessing steps. We then propose the omic-PGS approach and integrative methods for constructing omic-scores in Section 3.2.2, followed by simulation designs to evaluate and compare these methods in Section 3.2.3. The data application is described in Section 3.2.4. In Section 3.3, we present the results of the simulation studies and the data analyses. We conclude with a discussion.

3.2. MATERIALS AND METHODS

3.2.1. DATASETS

The omic-scores will be based on SNP data and on gene-based data (summarized SNP data per gene). The genomic data (SNP or summarized SNP) will be denoted by X. As omics data sets we will consider glycomics and metabolomics, which we will refer to as Y. We will use body mass index (BMI) and type 2 diabetes (T2D) as outcome variables, referred to as z. The omic-scores for these omics datasets Y will be included as covariates in models of the outcome z to understand the associations between these omic-based covariates and the outcome or to predict the outcome. We will conduct data analysis in two cohorts, namely the Orkney complex disease study (ORCADES) and the TwinsUK [39, 25]. In the ORCADES cohort, both omics (glycomics, metabolomics) and outcomes (BMI, T2D) are available, along with the imputed genomics and demographic variables age and sex. In the TwinsUK cohort, only imputed genomics, BMI and demographic variables age and sex are available. The genomic dataset in TwinsUK is also used in simulation studies.

THE ORCADES COHORT

The ORCADES cohort consists of 1885 inhabitants from Orkney. Family structure is present in the cohort. To remove the family structure in the data, we estimated the kin-

ship coefficient between individuals from the genomic data using KING-robust [23] and excluded one individual from each pair with a kinship coefficient greater than 0.25 (i.e., the first-degree relatives such as parents, full siblings and children are removed). The individual excluded from a pair was determined in such a way that the number of participants was maximized (using PLINK 2 [34]). After kinship-based pruning our study comprised 1490 individuals. All these 1490 samples had missing genotype rate < 0.1.

We filtered out SNPs that had minor allele frequency (MAF) below 0.05. The 5314423 SNPs after the MAF filter were used to construct PGSs for both omic features and the outcome. For the integrative methods, we further pruned the SNPs with a r^2 threshold of 0.5 and removed SNPs that were not close to a gene (within +/- 50k base pair). The remaining 261644 SNPs were used for the SNP-based analyses (and referred to as the SNP data). For the gene-based analyses, we aggregated the SNPs around the same gene using PCA, and the first few genomic principal components (GPCs) that explained at least 80% of variance in each gene were taken. This resulted in a dataset with 95185 GPCs (referred to as the GPC data). The GPCs were approximately normally distributed.

The IgG glycomic data were measured using Ultra Performance Liquid Chromatography (UPLC) (details have been described in [17]. The data contain 23 glycan peaks. These peaks were normalized using log-transform of total area (logTA) [46], adjusted for batch effects using empirical Bayes method correction (R package 'sva' [18]), and corrected for age and sex using multiple regression.

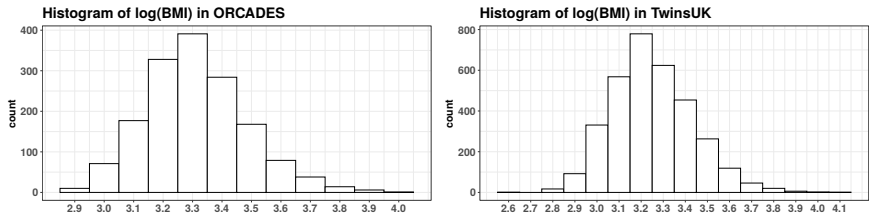
The metabolomic data were measured using the high throughput NMR metabolomics assay (Nightingale Health Ltd., Helsinki, Finland), consisting of 225 metabolite measures in molar concentration units. In this paper, we restrict our analysis to the 108 metabolites which overlap with the ones in the GWAS study of metabolomics conducted in [15]. We first excluded metabolites with a missing rate greater than 5% (0 metabolites excluded), and then replaced the zeros in the data with half of the lowest observed level for this metabolite, following the quality control steps in [13]. A Box-Cox transformation [4] with parameter 1/4 was then performed to reduce skewness [8]. Metabolite measures were set to missing based on a z-score cut-off of 6 [21]. The missing data were imputed once by chained equations (mice) [47]. Lastly, the imputed dataset was corrected for age and sex using multiple regression and the residuals were used.

The distribution of the observed log transformed BMI values is shown in left panel of Figure 3.1a. The type 2 diabetes status was self-reported. There are in total 53 cases among the 1490 individuals (prevalence 3.7%).

THE TWINSUK COHORT

To remove family structure, a twin from each pair was randomly chosen and discarded, resulting in 3465 independent samples. For data analysis, we further removed individuals who did not have a BMI measurement or had a BMI larger than 100, resulting in 3323 samples. To be able to compare the results obtained using the TwinsUK with the ones using the ORCADES cohort, we used the same set of SNPs, i.e. the set as described above obtained after pruning and quality control in ORCADES. The GPCs were calculated by using the PCA loadings obtained in ORCADES. The distribution of the log-transformed BMI measurements is shown in Figure 3.1b.

The simulation study is based on the genomic data from TwinsUK, hence not the same data as used for the data application. We used the 3465 independent samples and



(a) ORCADES (b) TwinsUK

Figure 3.1: **Histograms of log(BMI) in both cohort.**

retained the SNPs that satisfied all of the following quality control (QC) criteria: with no multi-character allele codes, genotype rate greater than 0.9, minor allele frequency greater than 0.05, and information content metric greater than 0.9. We then pruned the SNPs with a r^2 threshold of 0.5 and removed SNPs that were not close to a gene (within +/- 50k base pair), resulting in 140277 SNPs. The SNP dataset was further aggregated into 61747 GPCs in the same way described in Section 3.2.1.

GWAS SUMMARY STATISTICS

For constructing outcome-based and omic-based PGSs, we performed a discovery step where we retrieved summary statistics (estimated effect sizes for each SNP and p-values) from four external meta-analyses. The summary statistics of BMI were obtained from a meta-analysis for up to 339224 individuals from 125 studies published by Locke et al. [19]. The summary statistics of T2D were from a meta-analysis with 16 million genetic variants in 62892 T2D cases and 596424 controls of European ancestry conducted by Xue et al. [49]. The summary statistics used for glycomics were from a GWA study conducted by Klaric et al. [16]) on four cohorts of European descent with a combined sample size of 8090, where the associations of 77 ultraperformance liquid chromatography (UPLC) IgG N-glycan traits (including original glycan peaks and derived ones) with HapMap2 (release 22) imputed genomic data were studied. Note that both ORCADES and TwinsUK were included as discovery cohort in this study. The summary statistics for metabolomics were from a meta-analysis conducted by Kettunen et al. [15] using 14 cohorts from Europe, totaling up to 24,925 individuals. The study includes 123 circulating metabolic traits quantified by nuclear magnetic resonance (NMR).

3.2.2. STATISTICAL METHODS

In this subsection, we propose two approaches for constructing omic-scores. We first introduce omic-PGS, which performs standard PGS approach for each omic feature with an additional PCA step to reduce dimensionality. The integrative methods are then described, which perform joint dimension reduction on X and Y simultaneously. We will compare the proposed approaches in simulation studies and data applications.

OMIC-SCORES USING OMIC-PGS

The omic-PGS approach requires genomic data in three independent cohorts, namely, discovery, tuning, and target. The discovery cohort should have the omics data Y mea-

sured to obtain the GWAS summary statistics for each omic feature. Note that the GWAS summary statistics of the features in Y are also often available from previous studies. The tuning dataset contains the genomics and the outcome, which is used to tune parameters for omic-PGS. The target dataset contains genomics based on which the omic-scores will be constructed, and the outcome variable which will be modeled by the omic-scores. An illustration of the omic-PGS approach is given in Figure 3.2.

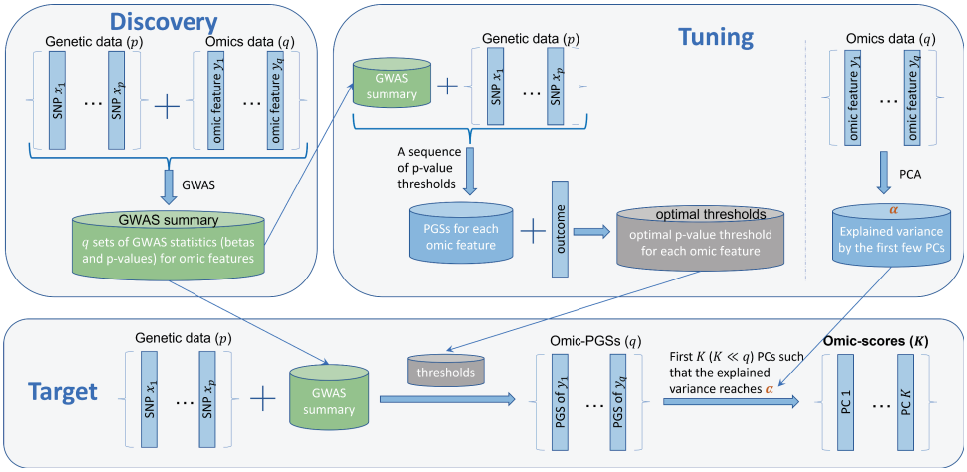


Figure 3.2: **Illustration of constructing omic-scores for omics using omic-PGS.**

For each individual, a polygenic score (PGS) of each omic feature (called omic-PGS) is calculated using the clumping and thresholding (C+T) method available in the software PRSice-2 [7]. The PGS of a omic feature for individual i is obtained as follows

$$\text{PGS}_i = \sum_{j=1}^{m} G_{ij}\hat{\beta}_j. \tag{3.1}$$

Here m is the total number of SNPs satisfying the p-value treshold in the discoveray cohort, G_{ij} is the allele count for the j-th SNP in individual i, and $\hat{\beta}_j$ is the effect sizes estimated by a relevant GWAS in a discovery cohort. For imputed genotypes, the expected values (real numbers between 0 and 2 known as the "dosage") are used for G_{ij} [51].

The p-value thresholds are selected by optimizing an objective function based on the PGS in a tuning cohort. Since the omic-PGSs will be used as covariates in a model for the outcome, we use the correlation between the omic-PGS and the outcome variable as objective function. When omics data are available, the corelation between the omics measurements and the omic-PGSs measures the ability of the omic-PGS to represent the omics data, which can be an alternative objective function for p-value thresholds selection.

We reduce the number of omic-PGSs by taking the first few principal components. These PCs capture the largest variances in the omic-PGSs and are independent to each other. Therefore they are good omic-scores and can be used when modeling the outcome. Note that the variances of the constructed omic-PGSs are not comparable, due

to different p-value thresholds and effect sizes of alleles used when constructing these omic- PGSs. To avoid a few omic-PGSs with relatively large variances dominating in the PCA, we standardize the variances of omic-PGSs before applying PCA.

To determine the number of omic-scores (or PCs of the omic-PGSs) to represent the omic-PGSs, we first determine the number of PCs in the original omics data by using a scree plot of eigenvalues and compute the proportion of explained variance α in the omics data by these PCs. This is again performed in the tuning cohort.

In the target cohort, we use the summary statistics from the discovery cohort and the p-value threshold from the tuning cohort to construct omic-PGS for each omic feature using equation (3.1). We then take K omic-scores in the omic-PGSs such that the explained variance in omic-PGSs is approximately equal to α, which was obtained in the tuning cohort.

OMIC-SCORES USING INTEGRATIVE METHODS

In this subsection, we will elaborate on our proposed integrative approaches to obtain omics-scores. In the same way as the omic-PGS approach, the constructed omic-scores will be used as covariates in a regression model for the outcome. We will consider O2PLS-based methods that also model the data-specific variation.

O2PLS Let X and Y be the genomic and the omic dataset of size $N \times p$ and $N \times q$, respectively. The O2PLS model decomposes the space into joint (T and U of size r), specific ($T_\perp$ and $U_\perp$ of size r_x resp. r_y) and residual (E and F of size p resp. q) subspaces. The relationship between X and Y is captured through the inner relation between T and U. The O2PLS model is written as

$$\begin{aligned} X &= TW^\top + T_\perp W_\perp^\top + E, \\ Y &= UC^\top + U_\perp C_\perp^\top + F, \\ U &= TB + H, \end{aligned}$$

where W ($p \times r$) and C ($q \times r$) are the loading matrices for the joint spaces of X and Y respectively and $W_\perp$ ($p \times r_x$) and $C_\perp$ ($q \times r_y$) are the loading matrices for the specific parts of X and Y respectively. The loading values indicate relative importance of each genomic and omic variable in forming the corresponding components. The $r \times r$ diagonal matrix B models the relationship between the joint components T and U. The O2PLS model can be estimated using the R package 'OmicsPLS' [9], and the details of the algorithm can be found in [45, 9].

The genomic variation related to the omics is captured in the genomic components T, which are linear combinations of the genetic variants. These linear combinations can be used as omic-scores for the omics data. For future samples X_{new} where only genomic data is available, the genomic joint components T_{new} can be computed using the genomic loadings W and $W_\perp$ from a fitted model as follows,

$$T_{new} = (X_{new} - X_{\perp new})W, \tag{3.2}$$

where $X_{\perp new}$ is the variation in X_{new} uncorrected to Y, which is estimated by projecting the new data X_{new} onto the X-specific subspace (details in [9]).

The genomic components in O2PLS are linear combinations of all genetic variants. Since we expect that only a limited number of SNPs are relevant, we also use the sparse version SO2PLS [12] which imposes an L_1 penalty on the joint loadings so that a large number of small non-zero weights are pushed to zero. The genomic components T_{new} are constructed as in (3.2), where W is sparse for SO2PLS.

PO2PLS PO2PLS [10] is a probabilistic extension of O2PLS. It assumes that a p-dimensional random vector x (containing genetic variants) and a q-dimensional random vector y (omic features) drawn from the population follow a multivariate normal distribution. Because of this assumption, we only fit PO2PLS on the GPCs that are approximately normally distributed and do not consider the discrete SNP data. The parameters of PO2PLS are estimated simultaneously by maximum likelihood. It appears that this approach is sensitive to model-misspecification (i.e., the number of joint and omic-specific components) [10], specifically using a too small number of components yields biased results. After estimation of the model, the genomic joint components for the omics can be computed using formula (3.2).

3.2.3. SIMULATION SETTINGS

To evaluate the performance of omic-PGS and the integrative methods (O2PLS and PO2PLS), we performed two simulation studies. In the first study, we simulated a glycomic dataset and an outcome using both the GPC and the SNP data from the TwinsUK study described in section 3.2.1. We aimed to investigate how well the outcome is modeled by the omic-scores constructed by each method in various scenarios. We computed the correlation between the omic-scores and the outcome variable. We also investigated whether these omic-scores have additional value to PGS. In the second study, we considered adding omic-scores of a second omics dataset (metabolomics) to the model. Both omics datasets are known to be associated with the outcomes [44, 14, 31, 29]. The main focus of this second simulation study was the accuracy of the method in identifying the relevant genomic variables.

SIMULATION I

We generated in total 23 glycans using 5 genomic components representing the 5 function groups see [36]. Let X ($N \times p$) be the GPC or SNP data, and Y ($N \times q$) be the simulated glycomic data. The relationship is given by

$$\begin{aligned} T &= XW, \\ Y &= TC^\top + F. \end{aligned} \tag{3.3}$$

Here, T is a $N \times 5$ matrix, containing 5 components, representing the 5 function groups, simulated as linear combinations of the columns of X. The $p \times 5$ weight matrix W for the 5 linear combinations was obtained by hard-thresholding the PCA loading vectors of X. The threshold was set to the 90th, 99th, and 99.9th percentile of the absolute loading values, representing causal proportions of 0.1, 0.01, and 0.001, respectively. To investigate the impact of the variance of T with respect to X, the hard-thresholding was performed on the first-fifth, sixth-tenth, and eleventh-fifteenth PC loadings. The $q \times 5$ matrix C

contains ones and zeros according to the biological grouping of the glycans, with each column representing one of five groups. The element $C_{j,k}$ is one if glycan j belongs to group k, and zero otherwise. The zero-mean normally distributed residual F accounts for $1-h_y^2$ of the total variance in Y. The heritability h_y^2 was set to 0.5 and 0.2. Note that most of the glycans have a heritability greater than 0.5 [24]. The outcome z was simulated as

$$z = TC^\top l + g.$$

Here, the q-dimensional column vector l was taken as the first principal component loading vector of the matrix $TC^\top$. It maps the genomic components underlying Y to the outcome z. The zero-mean normally distributed residual g accounts for 10% of the total variation in z.

We randomly split the data into 2000 discovery, 465 tuning, and 1000 target samples. We fitted O2PLS and PO2PLS models in the discovery set. the number of components for O2PLS and PO2PLS was specified as 5, 5, 0 for joint, genomic-specific and glycomics-specific, respectively. Because of the sensitivity of PO2PLS to the number of components, we considered a second PO2PLS model with 5, 10, 0 components (referred to as PO2PLS-2). The joint genomic components in the target set were then calculated as (3.2) (Note that the integrative methods do not use the tuning set). For omic-PGS, we performed GWAS for each variable in Y using the R package 'GenABEL' [1] in the discovery set, tuned the p-value thresholds in the tuning set, and computed omic-PGSs in the target set, following the steps in Figure 3.2. To be compared with integrative methods, we took 5 PCs (same as the number of joint components) of the correlated 23 omic-PGSs as omic-scores of Y. In the target set, the multiple genomic components of the integrative methods or PCs of omic-PGS were combined into a single score using the coefficients from the regression of z on the corresponding omic-scores in the discovery set. The prediction performance of the score was then measured by $R^2 = 1 - SSE/TSS$, where $SSE = \sum(z_i - \hat{z}_i)^2$, $TSS = \sum(z_i - \bar{z}_i)^2$. The whole process was repeated 20 times and the results for each simulation run were recorded.

SIMULATION II

The metabolomic data Y' ($N \times q'$) was generated using the same equations in (3.3). Specifically, we simulated 10 apolipoproteins using 4 components $T' = (t'_1, \ldots, t'_4)$ (eleventh-fourteenth PCs of the GPC data) representing 4 lipoprotein groups in [11], and a loading matrix C' mapping each apolipoprotein to groups. The causal proportion was set to 0.001, and the heritability was set to 0.7, based on the heritability of apolipoproteins [2]. The glycomic data Y was from one of the scenarios in the first simulation (GPC-based, causal proportion 0.001, sixth to tenth PCs, $h_y^2 = 0.5$). The outcome z' was then simulated from the genomic components underlying both Y and Y' as

$$z' = TC^\top l + T'C'^\top l' + g',$$

where l' is the loading vector of the first PC of $T'C'^\top$, and the residual g' accounts for 10% of total variance in z'.

We repeatedly split the data into discovery, tuning and target sets in the same way as the first simulation. All the methods were applied on the genomic data X and the two omics data Y and Y', resulting in two sets of omic-scores, namely one for the glycomics

dataset Y, and one for the metabolomics dataset Y'. For both omics, we calculated the true positive rate (TPR) of the top 40 genes. Here, the cut-off point of 40 is the number of genes in the method with the least number of genes selected to make the TPRs comparable across methods. To evaluate the joint prediction performance, we combined the two sets of omic-scores.

3.2.4. DATA APPLICATION

We applied omic-PGS and the integrative methods to the glycomics and metabolomic datasets available in the ORCADES cohort. We estimated the contribution of the omic-scores to the variation of BMI and of type 2 diabetes (T2D) measured by the adjusted-R^2 and Nagelkerke's R2 [27], respectively. For T2D, we also computed the area under the curve (AUC). Additionally, we studied whether the results with regard to the correlation between the omic-scores of the omics and BMI could be replicated in the TwinsUK cohort.

modeling BMI We first performed univariate regression of BMI on each glycan in ORCADES and selected the 10 glycans that had a nominal p-value of less than 0.05 for further analysis. Analogously, 87 metabolites were selected for the metabolomic dataset.

We randomly split the ORCADES cohort into discovery (N = 1000) and target (N = 490) sets. Note that the GWAS summary statistics for omic-PGS were obtained from external sources as described in Section 3.2.1, and therefore the discovery set here was used as tuning set.

In the discovery set, we determined the threshold of p-values for the omic-PGS using the GWAS summary statistics and we applied the integrative methods. Specifically, for each omic-PGS of an omic feature, we selected the p-value threshold, in such a way that the correlation between the omic-PGS and BMI was maximized. The correlation between each omic-PGS and its corresponding omic variable was also computed. The omic-PGSs were standardized and summarized to a few PCs. The number of PCs was chosen such that the proportion of explained variance in omic-PGSs was close to the proportion explained by the first 5 PCs (based on scree plot of eigenvalues) in the omics data. O2PLS and SO2PLS were applied on both the SNP data and the GPC data (referred to by adding a suffix -SNP and -GPC, respectively), and PO2PLS was only fitted on the GPC dataset. The number of joint and specific components in the integrative methods was chosen by visually identifying an elbow in the scree-plots. In the target set, omic-scores were constructed using the p-value thresholds selected (for omic-PGS) and loadings from the fitted models (for integrative methods) in the discovery set.

In both discovery and target sets, the proportion of explained variance of BMI was measured using adjusted-$R^2 = 1 - (1 - R^2)\frac{n-1}{n-K-1}$, where K is the number of omic-scores used (i.e., the number of PCs of omic-PGSs or the number of joint components). We calculated and compared the adjusted-R^2 using only PGS (BMI ~ PGS), only omic-scores of one method (BMI ~ omic-scores), and PGS and omic-scores combined (BMI ~ PGS + omic-scores).

To investigate the performance of the methods across cohorts, we constructed omic-scores in the TwinsUK cohort using the weights obtained in the ORCADES data and used these omic-scores to model BMI in the TwinsUK.

Predicting type 2 diabetes Similar to the BMI study, we performed uni-variate logistic regression of type 2 diabetes (T2D) on each glycan and metabolite and selected the 7 glycans and 78 metabolites that were significant in the ORCADES cohort.

We split the samples into discovery and target sets and applied methods in the discovery set. In the target set, we combined the omic-scores into a single genomic score using the regression coefficients from the logistic model in the discovery set. We tested whether the score was significantly associated with T2D by fitting a logistic regression of T2D on the score. We further measured the prediction performance by Nagelkerke's R2 and area under the curve (AUC).

3.3. RESULTS

3.3.1. RESULTS OF SIMULATION

SIMULATION I

Omics and outcome generated from GPC data For all proposed methods, the adjusted R^2 in the target set are shown in Figure 3.3. Overall, O2PLS outperformed the other methods. Concerning omic-PGS, it had a small advantage over the traditional PGS in most of the scenarios. As the causal proportion of X became smaller (columns left to right), the R^2 of O2PLS methods tended to decrease, while the performance of PGS methods increased. Comparing across the rows, when Y was generated from the higher numbered PCs of X (i.e., the genomic variation related to omics explains less variance in X), the R^2 of all the methods dropped. The PO2PLS model (with 5 X-specific components) appeared to decline most. PO2PLS with more X-specific components (PO2PLS-2) performed more robust compared to the model with less X-specific components. Comparing the left and right panels, the heritability of Y had little influence on the results.

The degree of overfitting of each method, measured by the ratio of R^2 in the target set to the R^2 in the discovery dataset is shown in Figure 3.4. The ratios for PO2PLS-2 were above 0.9 in all scenarios, while those for PO2PLS became unstable when Y was generated from the higher numbered PCs of X. The ratios for O2PLS were above 0.9 in most scenarios, and were slightly lower than PO2PLS-2. The ratios for omic-PGS were mostly between 0.6 and 0.75, which were lower compared to the integrative methods. The ratios for PGS were around 0.15 lower compared to those for omic-PGS. Recall that omic-PGS and PGS performed similarly in the target set regarding R^2 (see Figure 3.3). Omic-PGS which is based on the GWASs performed on the omic features showed less overfitting than PGS.

Omics and outcome generated from SNP data The adjusted R^2 in the target set and the ratio of R^2 in target to discovery are shown in Figure 3.5. The left panel shows that O2PLS outperformed the other methods in most of the scenarios regarding the R^2. PO2PLS-2 which was fitted on the GPC data performed poorly. Concerning the influence of causal proportion (columns) and covariance (rows) on the methods, we observed similar results as for the GPC data. In the right panel, O2PLS showed overfitting when

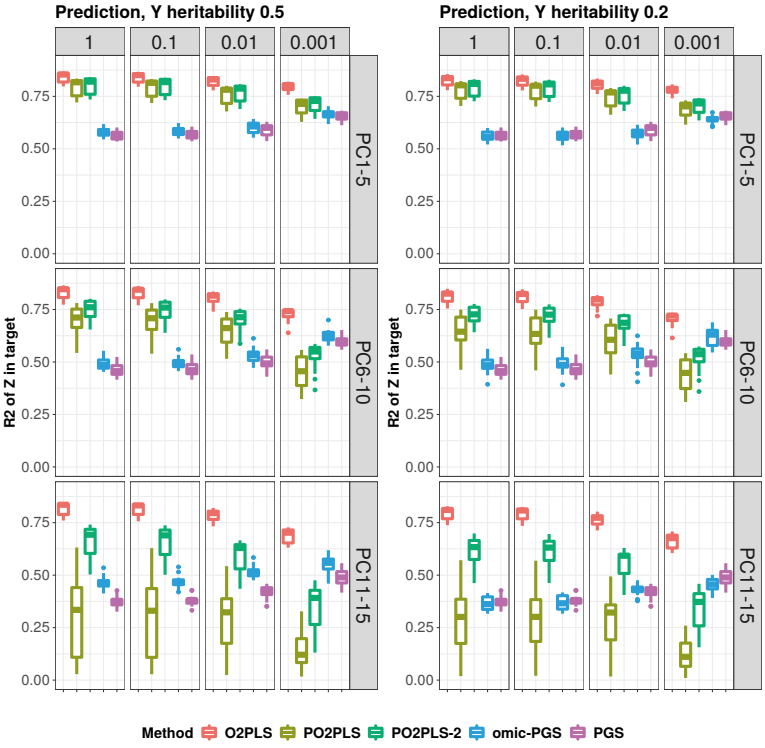


Figure 3.3: **Results of outcome prediction measured by adjusted R^2 in the target set.** The left and right panel show the results when the heritability of Y is 0.5, and 0.2, respectively. Rows of each panel indicates which PCs of the X dataset were used to generate Y. Columns show the causal proportion of GPCs. PO2PLS-2 is the PO2PLS model with 10 genomic-specific PCs. Boxes show the results on 20 randomly split datasets.

the causal proportion was small and the omics was generated from the higher numbered PCs. The ratios for PO2PLS-2 were unstable (with large boxes). The PGS methods showed most overfitting in all the scenarios.

SIMULATION II

The performance concerning prediction and feature selection for the scenario of two omics datasets Y and Y' is shown in Figure 3.6. The left panel shows the prediction performance evaluated on the combined joint components for two omics datasets. The performances were similar to the scenario of one dataset shown in Figure 3.3 (with causal proportion 0.001, PC 6-10). The middle and right panels show the TPR of the top 40 genes selected in the X-joint components for Y and Y', respectively. The TPR of O2PLS reached 1, suggesting the top genes in O2PLS were all relevant to the corresponding omics. For PO2PLS-2, the TPR for Y is higher than for Y' which has a smaller covariance with X. Although omic-PGS and PGS had similar prediction performance, omic-PGS had higher TPR for both omics.

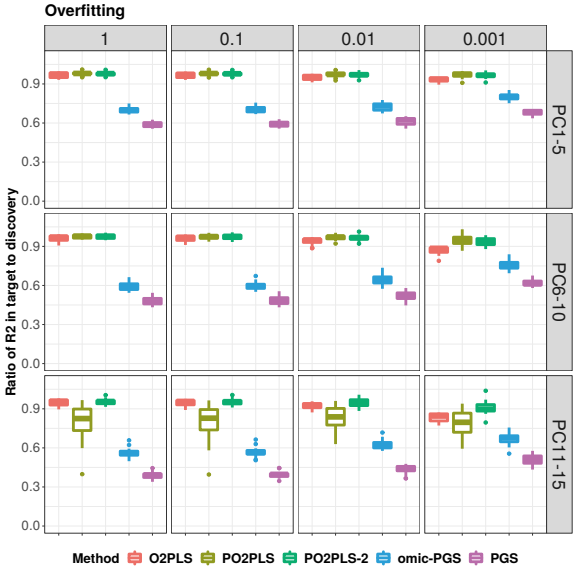


Figure 3.4: **Results of overfitting measured by the ratio of R^2 in target to the R^2 in discovery set.** The h_y^2 was set to be 0.5. A higher value close to 1 indicates less overfitting. Rows of each panel indicates which PCs of the X dataset were used to generate Y. Columns show the causal proportion of GPCs. PO2PLS-2 is the PO2PLS model with 10 genomic-specific PCs. Boxes show the results on 20 randomly split datasets.

3.3.2. RESULTS OF DATA APPLICATION

MODELING BMI

The correlation between each glycomic-PGS and its corresponding glycomic measurement is shown in Figure 3.7. From the plot, it can be concluded that the glycomic dataset was well represented by the glycomic-PGSs in both discovery and target sets, with most correlations above 0.7. The correlations of the metabolomic-PGSs with the corresponding metabolites are shown in Figure 3.8. The metabolomic data appeared to be less represented by its omic-scores, with most of the correlations below 0.2.

For the omic-PGS method, we took the first 6 PCs of the glycomic-PGSs and the first 17 PCs of the metabolomic-PGSs so that the explained variances in the omic-PGSs were close to the proportion of omics data explained by the first 5 PCs (79% in glycomics and 42% in metabolomics). For the integrative methods, the number of joint, genomic-specific, and omic-specific components were set to 5, 2, 0 for glycomics, and 7, 10, 0 for metabolomics, based on scree plots. The number of SNPs or GPCs retained in the SO2PLS were set to 1% (i.e., 2616 SNPs, or 952 GPCs with non-zero joint loadings).

The explained variance of BMI by the various omic-scores, along with the p-values of the F-test corresponding to the null hypothesis of no effect of the omic-scores are shown in Table 3.1. From the second and third columns of the table, the glycomic data in the discovery set explained 0.9% variance of BMI while the data in the target set explained 3.88% in the random split. The explained variances by the omic-PGS for glycomics were similar to those by the glycomic data. This was expected as the glycomic-PGSs repre-

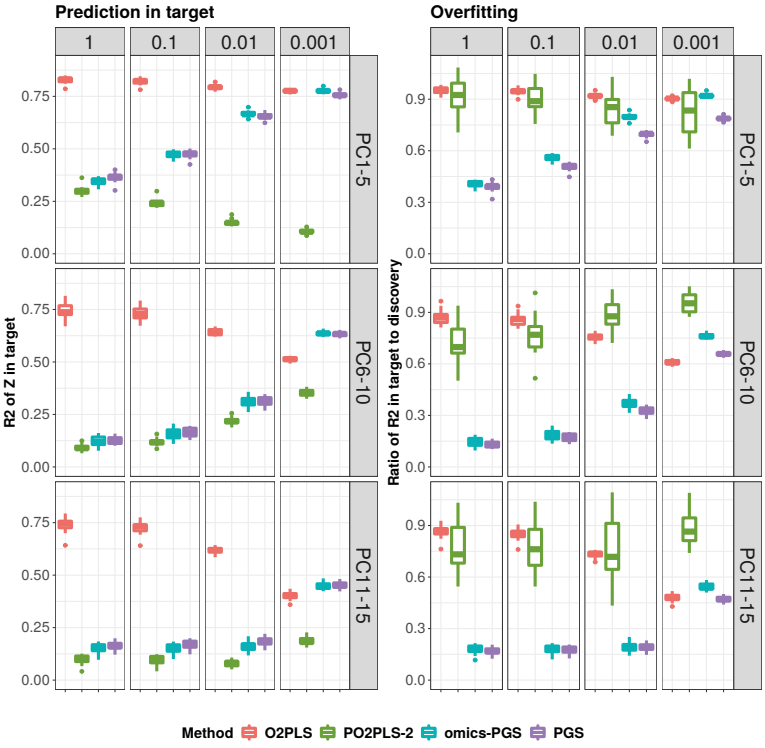


Figure 3.5: **Results of outcome prediction and overfitting, where the Y dataset was generated from the SNP data.** Rows of each panel indicates which PCs of the SNP dataset were used to generate Y. Columns show the causal proportion of SNPs. PO2PLS-2 is the PO2PLS model with 10 genomic-specific PCs. O2PLS was applied to the SNP data while PO2PLS-2 was fitted on the GPC data. Boxes show the results on 20 randomly split datasets.

sented the glycomic data well (see Figure 3.7). The corresponding F-test for the association between BMI and the omic-scores was significant in both discovery (0.031) and target (0.001) sets. Among the integrative methods, SO2PLS-SNP performed the best, but it explained less variance of BMI compared to omic-PGS. The fourth and fifth columns of Table 3.1 show the results for metabolomics. The metabolomic data explained more variance in BMI (17.91% in discovery and 20.66% in target) compared to the glycomic data. For omic-PGS, the omic-scores explained less BMI than the omics data, and the association with BMI in the target set was not significant. Among the integrative methods, O2PLS-GPC, O2PLS-SNP, and SO2PLS-SNP explained a larger proportion of the variance of BMI than the omics data in the discovery set, but did not explain the variance of BMI in the target set, suggesting overfitting in the discovery set. SO2PLS-GPC performed the best in the target set, explaining 2.26% of the variance of BMI, and its joint genomic components were significantly associated with BMI (p-value 0.012). The components of PO2PLS appeared not to be associated with BMI.

The omic-scores obtained with omic-PGS and SO2PLS-GPC and which were signif-

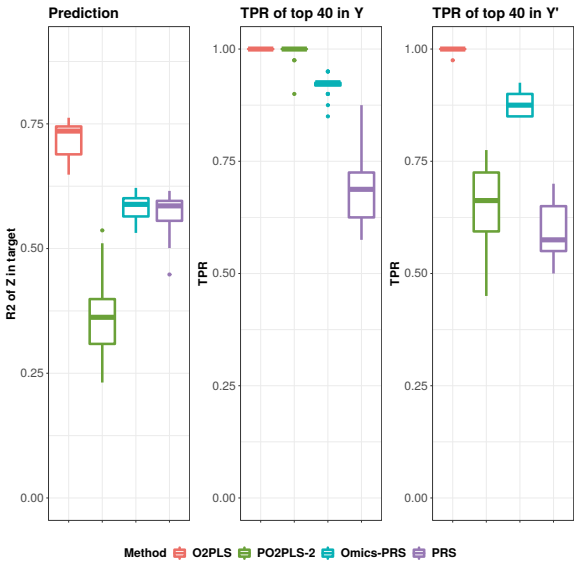
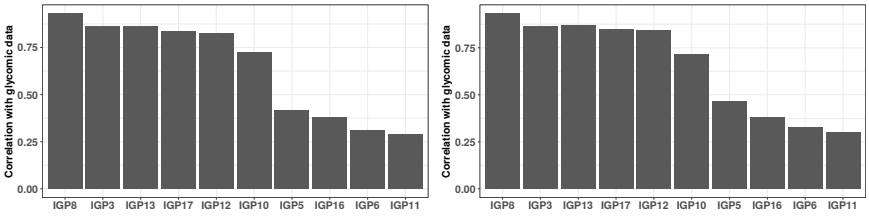


Figure 3.6: **Results of outcome prediction with two omics datasets and corresponding gene selection accuracy.** The prediction performance was evaluated on the combined joint components for two omics. The TPR was calculated using the top genes selected in the corresponding joint components for each omics dataset. The cut-off point of 40 was selected based on the number of genes in the method with the least genes selected. PO2PLS-2 is the PO2PLS model with 10 genomic-specific PCs. Boxes show the results on 20 randomly split datasets.



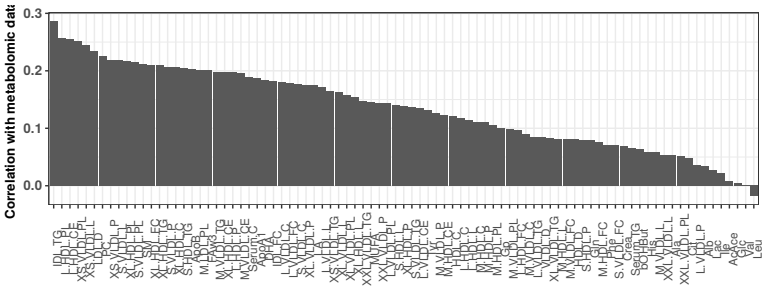
(a) Correlation in discovery data

(b) Correlation in target data

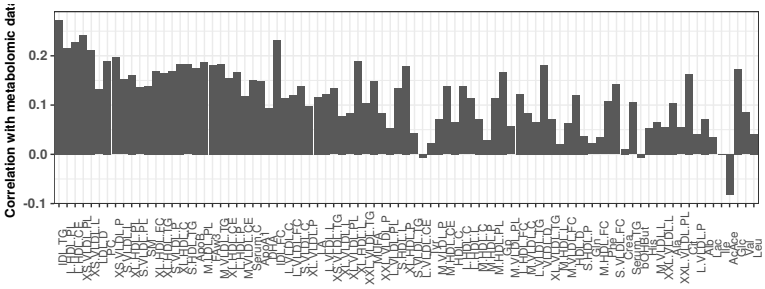
Figure 3.7: **Correlation of the constructed glycomic-PGSs with the corresponding measurement of glycans associated with BMI.** The glycans were sorted based on the correlations in the discovery set.

icantly associated with BMI were combined with the (BMI-based) PGS in one model. The explained BMI by the PGS was 13.1% in the discovery set and 13.74% in the target set. After combining with the omic-scores from omic-PGS, the explained variance increased to 15.65% in the discovery set and 17.04% in the target set. This improvement was mainly driven by the omic-scores for glycomics. Combining the PGS with the omic-scores from SO2PLS-GPC, the adjusted-R^2 increased to 27.2% in the discovery set, and moved slightly up to 13.98% in the target set.

We then investigated the genes selected in the SO2PLS-GPC for metabolomics, where the omic-scores were significantly associated with BMI. Among the 7 genomic compo-



(a) Correlation in discovery data



(b) Correlation in target data

Figure 3.8: **Correlation of the constructed metabolomic-PGSs with the corresponding measurement of metabolites associated with BMI.** The metabolites were sorted based on the correlations in the discovery set.

Table 3.1: **Explained variance of BMI by omic-scores in ORCADES cohort**

	Glycomics		Metabolomics	
	discovery	target	discovery	target
Omics data	0.90 (0.016)	3.88 (2.0e-04)	17.91 (< 2.2e-16)	20.66 (< 2.2e-16)
omic-PGS	**0.79 (0.031)**	**3.23 (0.001)**	2.54 (5.7e-4)	0.93 (-)
O2PLS-GPC	0.21 (-)	0 (-)	19.75 (<2.2e-16)	0 (-)
O2PLS-SNP	0.25 (-)	0 (-)	20.14 (<2.2e-16)	0.06 (-)
SO2PLS-GPC	0.31 (-)	0 (-)	**17.37 (<2.2e-16)**	**2.26 (0.012)**
SO2PLS-SNP	0.51 (-)	0.86 (-)	20.17 (<2.2e-16)	0.14 (-)
PO2PLS	0 (-)	0.18 (-)	0 (-)	0.74 (-)

Note: the numbers in the table are percentages of BMI explained (adjusted-$R^2 \times 100$), negative adjusted-R^2s are recorded as 0. P-values less than 0.05 are shown in brackets.

nents, the second and the fifth components were the most significant. Therefore, from each of these two components, we took the 952 GPCs with non-zero joint loadings and performed gene ontology (GO) enrichment analysis using the ToppGene Suite [5]. In the second component, the significant molecular function is cytoskeletal protein binding, which is associated with obesity and is involved in adipocyte lipid storage and metabolism [26]. The top 3 metabolites were apolipoprotein A-I, phosphatidylcholine, and total

cholesterol in HDL. Their role in the pathways underlying obesity needs further investigation. In the fifth component, the most significant term regarding molecular function is transmembrane receptor protein tyrosine phosphatase activity, and the motabolite with the largest joint loading value is tyrosine, showing a strong relevance.

We observed that SO2PLS showed less overfitting and performed better than O2PLS. Therefore, we performed a pre-screening on the genomic data before applying the integrative methods. We took the top 1000 SNPs with the smallest p-values for each glycan and metabolite based on the GWAS summary statistics in [16] and [15], respectively. This resulted in a union of 4124 SNPs (summarized to 1164 GPCs) for glycomics and a union of 11272 SNPs (1148 GPCs) for metabolomics. The number of joint, genomic-specific, and omic specific components were set to 4, 3, 0 for glycomics, and 6, 5, 0 for metabolomics based on scree plots.

The explained variance of BMI by the omic-scores using the pre-screened genomic data is shown in Table 3.2. For glycomics, the genomic components of O2PLS-GPC explained a little more variance (0.91%) in the target set compared to the unscreened case, but still small compared to the variance explained by omic-PGS (3.23%). For metabolomics, the components in O2PLS-GPC and O2PLS-SNP explained more variance in the target set (1.49% and 1.07%) compared to the same methods in the unscreened case, and were significantly associated with BMI. The explained variance was smaller than the best-performing SO2PLS-GPC in the unscreened case. The differences between the adjusted-R^2 of target and discovery sets reduced compared to those in Table 3.1, suggesting less overfitting. PO2PLS explained 0.67% in the target set, which was slightly less compared to the unscreened case, but the association was significant. Combining the omic-scores from O2PLS-GPC with (BMI-based) PGS, the adjusted-R^2 improved from 13.1% to 15.32% in the discovery set, and from 13.74% to 14.95% in the target set. Using the omic-scores from O2PLS-SNP, the combined variance explained increased similarly to 15.08% and 14.60% in discovery and target set, respectively. Combining with PO2PLS did not improve the performance of PGS.

Table 3.2: **Explained variance of BMI by omic-scores using pre-screened genomic data in ORCADES cohort**

	Glycomics		Metabolomics	
	discovery	target	discovery	target
O2PLS-GPC	0 (-)	0.91 (-)	**2.34 (4.8e-05)**	**1.49 (0.004)**
O2PLS-SNP	0.07 (-)	0 (-)	**1.64 (0.001)**	**1.07 (0.012)**
PO2PLS	0.31 (-)	0 (-)	0 (-)	**0.67 (0.039)**

Note: the numbers in the table are percentages of BMI explained (adjusted-$R^2 \times 100$), negative adjusted-R^2s are recorded as 0. P-values that are less than 0.05 are shown in brackets.

We conclude that in glycomics, omic-PGS outperformed integrative methods, while in metabolomics, SO2PLS-GPC on the unscreened data and O2PLS-GPC, O2PLS-SNP on the screened data performed the best. It is worth mentioning that although the O2PLS methods on the screened data explained less variance than SO2PLS-GPC (unscreened), the added value on the PGS was greater.

Lastly, we checked how well the linear combinations of GPCs and SNPs from the

ORCADES perform in the TwinsUK cohort. We used the p-value thresholds selected for omic-PGSs and the fitted loadings for the integrative methods in the ORCADES to calculate the omic-scores of glycomics and metabolomics in the TwinsUK. The first 6 PCs of glycomic-PGSs (accounting for 78% of the glycomic-PGSs) had an adjusted-R^2 (with BMI) of 0.94% (p-value 1.565e-06). The first 17 PCs of metabolomic-PGSs (48% of the metabolomic-PGSs) were not significantly associated with BMI. The joint genomic components of all the the integrative methods did not explain BMI and were not significant. Thus the results in ORCADES does not seem to replicate in the TwinsUK.

PREDICTION OF TYPE 2 DIABETES

The correlations of the constructed glycomic-PGSs and metabolomic-PGSs with the corresponding measurements of glycans and metabolites showed similar patterns as in the BMI study.

The number of PCs in the omic-PGS method was determined in the same way as described in the BMI study. The first 4 PCs of the glycomic-PGSs and the first 15 PCs of the metabolomic-PGSs were taken, which explained 70% and 46% of the corresponding omic-PGSs, respectively. The integrative methods were fitted on pre-screened SNP and/or GPC data. Based on scree plots, the number of joint, genomic-specific, and omic-specific components were set to 3, 4, 0 for glycomics, and 6, 5, 0 for metabolomics.

The Nagelkerke's pseudo R^2 and the p-value for the genomic score in the target set (linear combination of the omic-scores based on the regression coefficients from the discovery set) are shown in Table 3.3. For glycomics, the pseudo R^2 of the scores in the target set from omic-PGS was 11.75%, close to that of the glycomics data. The integrative methods had lower pseudo R^2 compared to omic-PGS, but the scores in the target set were all significant. PO2PLS performed the best among the integrative methods, with a pseudo R^2 of 3.56% (p-value 0.024). For metabolomics, omic-PGS had the highest pseudo R^2, but less than that of the metabolomic data, and the scores were not significant in the target set. The integrative methods showed overfitting for metabolomics, explaining little variance in the target set. Table 3.4 shows the AUCs of predicting T2D. The conclusions were similar.

Table 3.3: **Pseudo R^2 of T2D by omic-scores using pre-screened genomic data in ORCADES cohort**

	Glycomics		Metabolomics	
	discovery	target	discovery	target
Omics data	6.98	12.13 (1.9e-05)	23.90	15.29 (9.9e-07)
omic-PGS	**7.36**	**11.75 (5.5e-05)**	19.43	2.21 (-)
O2PLS-GPC	**2.72**	**3.01 (0.038)**	11.50	0.69 (-)
O2PLS-SNP	**2.31**	**2.92 (0.045)**	12.59	0.005 (-)
PO2PLS	**2.52**	**3.56 (0.024)**	3.51	0.49 (-)

Note: the numbers in the table are the pseudo R^2 of T2D (in percentage). P-values for the prediction score in the target set that were less than 0.05 are shown in brackets.

The performance of combining (T2D-based) PGS with the omic-scores from each method is shown in Table 3.5. The pseudo R^2 increased when combining the PGS with

Table 3.4: **AUC of T2D by omic-scores using pre-screened genomic data in ORCADES cohort**

	Glycomics		Metabolomics	
	discovery	target	discovery	target
Omics data	0.673	0.774	0.853	0.814
omic-PGS	0.704	0.749	0.821	0.619
O2PLS-GPC	0.624	0.632	0.757	0.563
O2PLS-SNP	0.627	0.630	0.765	0.517
PO2PLS	0.619	0.636	0.630	0.568

the omic-scores from all the methods, especially with the ones from omic-PGS, which almost doubled the R^2 from 10.46% to 20.78%. Regarding the AUC, the omic-PGS increased the AUC from 0.739 to 0.800. PO2PLS improved the AUC to 0.774, while the AUC with O2PLS (-GPC and -SNP) appeared to be on par with the AUC using only PGS. We conclude that omic-PGS performed the best for predicting T2D.

Table 3.5: **Pseudo R^2 and AUC of T2D combining PGS and omic-scores**

	Pseudo R^2		AUC	
	discovery	target	discovery	target
PGS	7.01	10.46	0.702	0.739
omic-PGS	33.01	20.78	0.878	0.800
O2PLS-GPC	19.10	12.99	0.794	0.744
O2PLS-SNP	20.26	12.20	0.785	0.734
PO2PLS	12.76	12.78	0.769	0.774

3.4. DISCUSSION

To better understand the genetic architecture and improve the prediction of outcomes, we proposed omic-PGS and integrative methods to incorporate inheritable information in omics into the linear model of the outcome, by constructing omic-scores of these omics. The advantages of using the omic-scores over the traditional outcome-based PGS (that does not utilize the omics data) regarding outcome prediction and feature selection were shown via simulation studies. The methods were applied to construct omic-scores of glycomics and metabolomics in two cohorts and appeared to increase the explained variance of BMI and improve the prediction performance of T2D.

The omic-PGS and the integrative methods have different advantages. The omic-PGS method utilizes the publicly available GWAS summary statistics for omic features to reconstruct omics datasets, thus the omics data at individual level are not needed. It can be used to incorporate information contained in any omic feature of which a GWAS is available to a study, as long as the genomics are available. Omic-PGS uses only a small subset of SNPs selected based on the p-values from the relevant GWAS, and it performs well when the causal proportion is small (see Figure 3.3). Note that many other PGS

methods can be implemented [28] in place of the classic C+T approach. We considered the C+T approach as it is well known and most often used in genomic studies. Moreover, it is one of the computationally faster algorithms. Comparing the performance of omic-PGS implementing other PGS methods is future work.

Comparing to omic-PGS, the integrative methods take into account the correlation structure in the genomics and omics simultaneously and construct a few omic-scores for the whole omics dataset. For the integrative methods, a large sample size is not needed, which is usually required in the GWA studies. O2PLS and PO2PLS are not sparse, they use all the genomic variables to construct the genomic components, therefore tend to perform better when the causal proportion is large. When the causal proportion is small, the constructed joint components contain a large proportion of noise by including a large amount of non-causal genomic variables. The sparse variant SO2PLS (or GO2PLS) can be used in this case. A detailed comparison of performance in sparse settings can be found in [12]. PO2PLS and O2PLS perform differently when the covariance between X and Y is small relative to the variance of X (see bottom rows of Figure 3.3). PO2PLS models the covariance of X and Y as well as their variances. As the covariance between X and Y becomes smaller, PO2PLS tends to weigh more towards the relatively larger X-specific variance in the joint subspace. Specifying more X-specific components in PO2PLS shifts part of the specific variance captured in the joint subspace into the X-specific subspace, which in turn improves the ability of the joint subspaces to capture the correlated variance in X and Y. This is also indicated by the improved performance of PO2PLS-2 in the simulation (PO2PLS with more X-specific components) in Figure 3.3. O2PLS only models covariance, it ignores the relative size of the data specific variance, therefore its performance is much less affected.

In data applications, the omic-scores modeled a relatively small proportion of the outcomes (BMI and T2D) compared to the traditional PGS. One reason is that the outcome variable is affected by many more omic levels other than the glycomics and metabolomics included in the model. For example, the BMI is also associated with transcriptomics [20], proteomics [52], etc. Therefore the omic-scores of one or two omic levels can only model a small part of the outcome. Another reason is the huge difference in the sample size used in the GWAS for the outcome and that available to our study. For example, for BMI, the contrast of sample size is 339K vs 1K. More research is need to investigate the performance of the omic-scores in modeling outcome under a large sample size. Nevertheless, we showed that omic-scores can provide omic-specific interpretations, and improved the performance of the traditional PGS when combined in the model for the outcome. Furthermore, by testing the associations between omic-scores based on various omics with an outcome, researchers can have a preliminary idea of the relevance of each omics to the outcome, without having to measure the omics. This helps to decide which omics to include in a study in a cost-efficient way.

To conclude, we proposed methods for constructing omic-scores based on genomics to better model an outcome disease and understand the genetic architecture underlying the disease. The omic-scores are stable over lifespan hence can be applied independent of age. They can be constructed using genomics and publicly available GWAS summary statistics, without omics data, therefore have potential to be implemented in a wide range of studies.

BIBLIOGRAPHY

[1] Aulchenko, Y. S., Ripke, S., Isaacs, A., and van Duijn, C. M. (2007). GenABEL: An R library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296.

[2] Beekman, M., Heijmans, B. T., Martin, N. G., Pedersen, N. L., Whitfield, J. B., DeFaire, U., Van Baal, G. C. M., Snieder, H., Vogler, G. P., Slagboom, P. E., and Boomsma, D. I. (2002). Heritabilities of apolipoprotein and lipid levels in three countries. *Twin Research*, 5(2):87–97.

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

[4] Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

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

[6] Choi, S. W., Mak, T. S. H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses.

[7] Choi, S. W. and O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, 8(7):1–6.

[8] el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., and Uh, H. W. (2016). Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, 17(2):S11.

[9] el Bouhaddani, S., Uh, H. W., Jongbloed, G., Hayward, C., Klarić, L., Kiełbasa, S. M., and Houwing-Duistermaat, J. (2018). Integrating omics datasets with the OmicsPLS package. *BMC Bioinformatics*, 19(1):371.

[10] el Bouhaddani, S., Uh, H.-W., Jongbloed, G., and Houwing-Duistermaat, J. (2022). Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

[11] Feingold, K. R. and Grunfeld, C. (2000). *Introduction to Lipids and Lipoproteins*. MDText.com, Inc.

[12] Gu, Z., el Bouhaddani, S., Pei, J., Houwing-Duistermaat, J., and Uh, H. W. (2021). Statistical integration of two omics datasets using GO2PLS. *BMC Bioinformatics*, 22(1).

[13] Hagenbeek, F. A., Pool, R., van Dongen, J., Draisma, H. H., Jan Hottenga, J., Willemsen, G., Abdellaoui, A., Fedko, I. O., den Braber, A., Visser, P. J., de Geus, E. J., Willems van Dijk, K., Verhoeven, A., Suchiman, H. E., Beekman, M., Slagboom, P. E., van Duijn, C. M., Barkey Wolf, J. J., Cats, D., Amin, N., Beulens, J. W., van der Bom, J. A., Bomer,

N., Demirkan, A., van Hilten, J. A., Meessen, J. M., Moed, M. H., Fu, J., Onderwater, G. L., Rutters, F., So-Osman, C., van der Flier, W. M., van der Heijden, A. A., van der Spek, A., Asselbergs, F. W., Boersma, E., Elders, P. M., Geleijnse, J. M., Ikram, M. A., Kloppenburg, M., Meulenbelt, I., Mooijaart, S. P., Nelissen, R. G., Netea, M. G., Penninx, B. W., Stehouwer, C. D., Teunissen, C. E., Terwindt, G. M., 't Hart, L. M., van den Maagdenberg, A. M., van der Harst, P., van der Horst, I. C., van der Kallen, C. J., van Greevenbroek, M. M., van Spil, W. E., Wijmenga, C., Zwinderman, A. H., Zhernikova, A., Jukema, J. W., Mei, H., Slofstra, M., Swertz, M., van den Akker, E. B., Deelen, J., Reinders, M. J., Harms, A. C., Hankemeier, T., Bartels, M., Nivard, M. G., and Boomsma, D. I. (2020). Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. *Nature Communications*, 11(1):1–11.

[14] Jin, Q. and Ma, R. C. W. (2021). Metabolomics in diabetes and diabetic complications: Insights from epidemiological studies.

[15] Kettunen, J., Demirkan, A., Würtz, P., Draisma, H. H., Haller, T., Rawal, R., Vaarhorst, A., Kangas, A. J., Lyytikäinen, L. P., Pirinen, M., Pool, R., Sarin, A. P., Soininen, P., Tukiainen, T., Wang, Q., Tiainen, M., Tynkkynen, T., Amin, N., Zeller, T., Beekman, M., Deelen, J., Van Dijk, K. W., Esko, T., Hottenga, J. J., Van Leeuwen, E. M., Lehtimäki, T., Mihailov, E., Rose, R. J., De Craen, A. J., Gieger, C., Kähönen, M., Perola, M., Blankenberg, S., Savolainen, M. J., Verhoeven, A., Viikari, J., Willemsen, G., Boomsma, D. I., Van Duijn, C. M., Eriksson, J., Jula, A., Järvelin, M. R., Kaprio, J., Metspalu, A., Raitakari, O., Salomaa, V., Eline Slagboom, P., Waldenberger, M., Ripatti, S., and Ala-Korpela, M. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications*, 7(1):1–9.

[16] Klarić, L., Tsepilov, Y. A., Stanton, C. M., Mangino, M., Sikka, T. T., Esko, T., Pakhomov, E., Salo, P., Deelen, J., McGurnaghan, S. J., Keser, T., Vučković, F., Ugrina, I., Krištić, J., Gudelj, I., Štambuk, J., Plomp, R., Pučić-Baković, M., Pavić, T., Vilaj, M., Trbojević-Akmačić, I., Drake, C., Dobrinić, P., Mlinarec, J., Jelušić, B., Richmond, A., Timofeeva, M., Grishchenko, A. K., Dmitrieva, J., Bermingham, M. L., Sharapov, S. Z., Farrington, S. M., Theodoratou, E., Uh, H. W., Beekman, M., Slagboom, E. P., Louis, E., Georges, M., Wuhrer, M., Colhoun, H. M., Dunlop, M. G., Perola, M., Fischer, K., Polasek, O., Campbell, H., Rudan, I., Wilson, J. F., Zoldoš, V., Vitart, V., Spector, T., Aulchenko, Y. S., Lauc, G., and Hayward, C. (2020). Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Science Advances*, 6(8).

[17] Krištić, J., Vučković, F., Menni, C., Klarić, L., Keser, T., Beceheli, I., Pučić-Baković, M., Novokmet, M., Mangino, M., Thaqi, K., Rudan, P., Novokmet, N., Šarac, J., Missoni, S., Kolčić, I., Polašek, O., Rudan, I., Campbell, H., Hayward, C., Aulchenko, Y., Valdes, A., Wilson, J. F., Gornik, O., Primorac, D., Zoldoš, V., Spector, T., and Lauc, G. (2014). Glycans are a novel biomarker of chronological and biological ages. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 69(7):779–789.

[18] Leek, J. T. (2014). Svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42(21):e161.

[19] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., Winkler, T. W., Wood, A. R., Workalemahu, T., Faul, J. D., Smith, J. A., Zhao, J. H., Zhao, W., Chen, J., Fehrmann, R., Hedman, K., Karjalainen, J., Schmidt, E. M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J. L., Bragg-Gresham, J. L., Buyske, S., Demirkan, A., Deng, G., Ehret, G., Feenstra, B., Feitosa, M., Fischer, K., Goel, A., Gong, J., Jackson, A. U., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Leach, I. M., Medina-Gomez, C., Medland, S. E., Nalls, M. A., Palmer, C. D., Pasko, D., Pechlivanis, S., Peters, M. J., Prokopenko, I., Shungin, D., Stančáková, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Isaacs, A., Albrecht, E., Ärnlöv, J., Arscott, G. M., Attwood, A. P., Bandinelli, S., Barrett, A., Bas, I. N., Bellis, C., Bennett, A. J., Berne, C., Blagieva, R., Blüher, M., Böhringer, S., Bonnycastle, L. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Caspersen, I. H., Chen, Y. I., Clarke, R., Daw, E. W., de Craen, A. J., Delgado, G., Dimitriou, M., Doney, A. S., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R. M., Garcia, M., Geller, F., Giedraitis, V., Gigante, B., Go, A. S., Golay, A., Goodall, A., Gordon, S. D., Gorski, M., Grabe, H. J., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., Groves, C. J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hengstenberg, C., Holmen, O., Hottenga, J. J., James, A. L., Jeff, J., Johansson, Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W., Laitinen, J., Lamina, C., Leander, K., Lee, N. R., Lichtner, P., Lind, L., Lindström, J., Lo, K. S., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P. K., Mahajan, A., McArdle, W. L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Mulas, A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Nagaraja, R., Nöthen, M. M., Nolte, I. M., Pilz, S., Rayner, N. W., Renstrom, F., Rettig, R., Ried, J. S., Ripke, S., Robertson, N., Rose, L. M., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F., Scott, W. R., Seufferlein, T., Shi, J., Smith, A. V., Smolonska, J., Stanton, A. V., Steinthorsdottir, V., Stirrups, K., Stringham, H. M., Sundström, J., Swertz, M. A., Swift, A. J., Syvänen, A. C., Tan, S. T., Tayo, B., Thorand, B., Thorleifsson, G., Tyrer, J., Uh, H. W., Vandenput, L., Verhulst, F. C., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Warren, H. R., Waterworth, D. M., Weedon, M. N., Wilkens, L., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., Brennan, E. P., Choi, M., Dastani, Z., Drong, A. W., Eriksson, P., Franco-Cereceda, A., Gådin, J. R., Gharavi, A. G., Goddard, M. E., Handsaker, R. E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J. Y., Liang, L., Lifton, R. P., Ma, B., McCarroll, S. A., McKnight, A. J., Min, J. L., Moffatt, M. F., Montgomery, G. W., Murabito, J. M., Nicholson, G., Nyholt, D. R., Okada, Y., Perry, J. R., Dorajoo, R., Reinmaa, E., Salem, R. M., Sandholm, N., Scott, R. A., Stolk, L., Takahashi, A., Van't Hooft, F. M., Vinkhuyzen, A. A., Westra, H. J., Zheng, W., Zondervan, K. T., Heath, A. C., Arveiler, D., Bakker, S. J., Beilby, J. P., Bergman, R. N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M., Cesana, G., Chakravarti, A., Chasman, D., Chines, P. S., Collins, F. S., Crawford, D., Cupples, L., Cusi, D., Danesh, J., de Faire, U., Den Ruijter, H. M., Dominiczak, A. F., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Felix,

S. B., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Franco, O. H., Gansevoort, R. T., Gejman, P. V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U. B., Hall, A. S., Harris, T. B., Hattersley, A. T., Hicks, A. A., Hindorff, L., Hingorani, A., Hofman, A., Homuth, G., Hovingh, G., Humphries, S. E., Hunt, S. C., Hyppönen, E., Illig, T., Jacobs, K. B., Jarvelin, M. R., Jöckel, K. H., Johansen, B., Jousilahti, P., Jukema, J., Jula, A., Kaprio, J., Kastelein, J. J., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Knekt, P., Kooner, J. S., Kooperberg, C., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lyssenko, V., Männistö, S., Marette, A., Matise, T., McKenzie, C. A., McKnight, B., Moll, F. L., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Madden, P. A., Pasterkamp, G., Peden, J. F., Peters, A., Postma, D. S., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O., Rankinen, T., Rao, D. C., Rice, T. K., Ridker, P., Rioux, J. D., Ritchie, M., Rudan, I., Salomaa, V., Samani, N., Saramies, J., Sarzynski, M. A., Schunkert, H., Schwarz, P. E., Sever, P., Shuldiner, A. R., Sinisalo, J., Stolk, R. P., Strauch, K., Tönjes, A., Trégouët, D. A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M. C., Völker, U., Waeber, G., Willemsen, G., Witteman, J. C., Zillikens, M. C., Adair, L. S., Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bornstein, S. R., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I., Dedoussis, G. V., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L., Haiman, C., Hamsten, A., Hui, J., Hunter, D. J., Hveem, K., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Metspalu, A., Moebus, S., Munroe, P., Njølstad, I., Oostra, B. A., Palmer, C. N., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Power, C., Quertermous, T., Rauramaa, R., Rivadeneira, F., Saaristo, T. E., Saleheen, D., Sattar, N., Schadt, E., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Thorsteinsdottir, U., Stumvoll, M., Tuomilehto, J., Uitterlinden, A. G., Uusitupa, M., van der Harst, P., Walker, M. C., Wallaschofski, H., Wareham, N., Watkins, H., Weir, D. R., Wichmann, H., Wilson, J. F., Zanen, P., Borecki, I., Deloukas, P., Fox, C. S., Heid, I. M., O'connell, J. R., Strachan, D. P., Stefansson, K., Van Duijn, C., Abecasis, G., Franke, L., Frayling, T. M., McCarthy, M. I., Visscher, P. M., Scherag, A., Willer, C. J., Boehnke, M., Mohlke, K. L., Lindgren, C. M., Beckmann, J. S., Barroso, I., North, K. E., Ingelsson, E., Hirschhorn, J. N., Loos, R. J., Speliotes, E. K., Thompson, J. R., Goldstein, B. A., König, I. R., Cazier, J. B., Grundberg, E., Havulinna, A. S., Ho, W. K., Hopewell, J. C., Eriksson, N., Lundmark, P., Lyytikäinen, L. P., Rafelt, S., Tikkanen, E., Van Zuydam, N., Voight, B. F., Ziegler, A., Altshuler, D., Balmforth, A. J., Braund, P. S., Burgdorf, C., Claudi-Boehm, S., Cox, D., Do, R., Doney, A. S., El Mokhtari, N., Fontanillas, P., Hager, J., Han, B. G., Hunt, S. E., Kang, H. M., Kessler, T., Knowles, J. W., Kolovou, G., Langford, C., Lokki, M. L., Lundmark, A., Meisinger, C., Melander, O., Maouche, S., Nikus, K., Rasheed, A., Rosinger, S., Rubin, D., Rumpf, M. P., Schäfer, A., Sivananthan, M., Song, C., Stewart, A. F., Thorgeirsson, G., van der Schoot, C. E., Wagner, P. J., Wells, G. A., Wild, P. S., Tsun-Po, Y., Basart, H., Brambilla, P., Cambien, F., Cupples, A. L., Dehghan, A., Diemert, P., Epstein, S. E., Evans, A., Ferrario, M. M., Gauguier, D., Hazen, S. L., Holm, H., Iribarren, C., Jang, Y., Kähönen, M., Kee, F., Kim, H. S., Klopp, N., Kuulasmaa, K., Laaksonen, R., Ouwehand, W. H., Parish, S., Park, J. E., Rader, D. J., Shah, S. H., Stark, K., Wallentin, L., Zimmermann, M. E., Nieminen, M. S., Sandhu, M. S., Pastinen, T., Zalloua, P. A., Siegbahn, A., Schreiber,

S., Ripatti, S., Blankenberg, S. S., O'donnell, C. J., Reilly, M. P., Collins, R., Roberts, R., Pattaro, C., Köttgen, A., Garnaas, M., Böger, C. A., Fuchsberger, C., Olden, M., Chen, M. H., Tin, A., Taliun, D., Li, M., Gao, X., Yang, Q., Hundertmark, C., Foster, M. C., O'seaghdha, C. M., Glazer, N. L., Liu, C. T., Struchalin, M., Li, G., Johnson, A. D., Gierman, H. J., Hwang, S. J., Atkinson, E. J., Lohman, K. K., Cornelis, M. C., Chouraki, V., Holliday, E. G., Sorice, R., Deshmukh, H., Ulivi, S., Chu, A. Y., Murgia, F., Imboden, M., Kollerits, B., Pistis, G., Launer, L., Aspelund, T., Eiriksdottir, G., Mitchell, B. D., Schmidt, H., Cavalieri, M., Rao, M., Hu, F. B., de Andrade, M., Turner, S. T., Ding, J., Andrews, J. S., Freedman, B. I., Döring, A., Kolcic, I., Zemunik, T., Boban, M., Minelli, C., Wheeler, H. E., Igl, W., Zaboli, G., Wild, S. H., Ellinghaus, D., Nöthlings, U., Jacobs, G., Biffar, R., Endlich, K., Ernst, F., Kroemer, H. K., Nauck, M., Stracke, S., Völzke, H., Aulchenko, Y., Polasek, O., Hastie, N., Vitart, V., Helmer, C., Wang, J. J., Ruggiero, D., Bergmann, S., Viikari, J., Nikopensius, T., Province, M. A., Ketkar, S., Colhoun, H. M., Doney, A., Robino, A., Giulianini, F., Krämer, B. K., Portas, L., Buckley, B. M., Adam, M., Thun, G. A., Paulweber, B., Haun, M., Sala, C., Metzger, M., Mitchell, P., Ciullo, M., Kim, S. K., Vollenweider, P., Palmer, C., Gasparini, P., Pirastu, M., Probst-Hensch, N. M., Kronenberg, F., Toniolo, D., Coresh, J., Schmidt, R., Siscovick, D., Kardia, S. L., Curhan, G., Franke, A., Parsa, A., Goessling, W., Kao, W. H., de Boer, I. H., Peralta, C. A., Akylbekova, E., Kramer, H., Arking, D. E., Franceschini, N., Egan, J., Hernandez, D. G., Townsend, R. R., Lumley, T., Psaty, B., Kestenbaum, B., Haritunians, T., Mooser, V., Florez, J. C., Meigs, J. B., Lu, X., Leak, T. S., Aasarød, K., Skorpen, F., Baumert, J., Devuyst, O., Mychaleckyj, J. C., Kedenko, L., Coassin, S., Hallan, S., Navis, G., Shlipak, M. G., Bull, S. B., Paterson, A. D., Rotter, J. I., Dreisbach, A. W., Anderson, C. A., Guo, Q., Henders, A., Lambert, A., Lee, S. H., Kraft, P., Kennedy, S. H., Macgregor, S., Missmer, S. A., Painter, J. N., Roseman, F., Treloar, S. A., Wallace, L., Forsblom, C., Isakova, T., McKay, G. J., Williams, W. W., Sadlier, D. M., Mäkinen, V. P., Swan, E. J., Boright, A. P., Ahlqvist, E., Keller, B. J., Huang, H., Ahola, A., Fagerholm, E., Gordin, D., Harjutsalo, V., He, B., Heikkilä, O., Hietala, K., Kytö, J., Lahermo, P., Lehto, M., Österholm, A. M., Parkkonen, M., Pitkäniemi, J., Rosengård-Bärlund, M., Saraheimo, M., Sarti, C., Söderlund, J., Soro-Paavonen, A., Syreeni, A., Thorn, L. M., Tikkanen, H., Tolonen, N., Tryggvason, K., Wadén, J., Gill, G. V., Prior, S., Guiducci, C., Mirel, D. B., Taylor, A., Hosseini, M., Parving, H. H., Rossing, P., Tarnow, L., Ladenvall, C., Alhenc-Gelas, F., Lefebvre, P., Rigalleau, V., Roussel, R., Tregouet, D. A., Maestroni, A., Maestroni, S., Falhammar, H., Gu, T., Möllsten, A., Cimponeriu, D., Mihai, I., Mota, M., Mota, E., Serafinceanu, C., Stavarachi, M., Hanson, R. L., Nelson, R. G., Kretzler, M., Panduru, N. M., Gu, H. F., Brismar, K., Zerbini, G., Hadjadj, S., Marre, M., Lajer, M., Waggott, D., Savage, D. A., Bain, S. C., Martin, F., Godson, C., Groop, P. H., Maxwell, A. P., Sengupta, S., Peloso, G. M., Ganna, A., Mora, S., Chang, H. Y., Den Hertog, H. M., Donnelly, L. A., Freitag, D. F., Gurdasani, D., Heikkilä, K., Johnson, T., Kaakinen, M., Kettunen, J., Li, X., Montasser, M. E., Petersen, A. K., Saxena, R., Service, S. K., Sidore, C., Surakka, I., Teslovich, T. M., Van den Herik, E. G., Volcik, K. A., Wu, Y., Asiki, G., Been, L. F., Burnett, M. S., Doring, A., Elliott, P., Eyjolfsson, G. I., Goodarzi, M. O., Gravito, M. L., Hartikainen, A. L., Hung, Y. J., Jones, M. R., Kaleebu, P., Khaw, K. T., Kim, E., Komulainen, P., Lehtimaki, T., Lin, S. Y., Lindstrom, J., Muller, G., Narisu, N., Nieminen, T. V., Nsubuga, R. N., Olafsson, I., Palotie, A., Papamarkou, T., Pomilla, C., Pouta, A., Ruokonen, A.,

Seeley, J., Silander, K., Tiret, L., van Pelt, L., Wainwright, N., Wijmenga, C., Young, E. H., Bennett, F., Boomsma, D. I., Burnier, M., Chen, Y. D., Feranil, A. B., Ferrieres, J., Freimer, N. B., Hsiung, C. A., Kesäniemi, A., Koudstaal, P. J., Krauss, R. M., Kyvik, K. O., Meneton, P., Moilanen, L., Sanghera, D. K., Sheu, W. H., Whitfield, J. B., Wolffenbuttel, B. H., Ordovas, J. M., Rich, S. S., Johnson, A., Johnson, L., Larson, M. G., Levy, D., Newton-Cheh, C., O'reilly, P. F., Palmas, W., Rice, K. M., Smith, A., Snider, H., Tobin, M., Verwoert, G., Verwoert, G. C., Pihur, V., Heath, S., Sõber, S., Arora, P., Zhang, F., Lucas, G., Milaneschi, Y., Parker, A. N., Fava, C., Fox, E. R., Go, M. J., Sjögren, M., Vinay, D., Alexander, M., Tabara, Y., Shaw-Hawkins, S., Whincup, P. H., Shi, G., Seielstad, M., Sim, X., Nguyen, K. D., Matullo, G., Gaunt, T. R., Onland-Moret, N. C., Cooper, M. N., Platou, C. G., Org, E., Hardy, R., Dahgam, S., Palmen, J., Kuznetsova, T., Uiterwaal, C. S., Adeyemo, A., Ludwig, B., Tomaszewski, M., Tzoulaki, I., Palmer, N. D., Chang, Y. P., Steinle, N. I., Grobbee, D. E., Morrison, A. C., Najjar, S., Hadley, D., Brown, M. J., Connell, J. M., Day, I. N., Lawlor, D. A., Lawrence, R. W., Ongen, H., Li, Y., Young, J. H., Bis, J. C., Bolton, J. A., Chaturvedi, N., Islam, M., Jafar, T. H., Kulkarni, S. R., Howard, P., Guarrera, S., Ricceri, F., Emilsson, V., Plump, A. S., Weder, A. B., Sun, Y. V., Scott, L. J., Peltonen, L., Vartiainen, E., Brand, S. M., Staessen, J. A., Wang, T. J., Burton, P. R., Artigas, M. S., Dong, Y., Wang, X., Zhu, H., Rudock, M. E., Heckbert, S. R., Smith, N. L., Wiggins, K. L., Doumatey, A., Shriner, D., Veldre, G., Viigimaa, M., Kinra, S., Prabhakaran, D., Tripathy, V., Langefeld, C. D., Rosengren, A., Thelle, D. S., Corsi, A. M., Singleton, A., Hilton, G., Salako, T., Iwai, N., Kita, Y., Ogihara, T., Ohkubo, T., Okamura, T., Ueshima, H., Umemura, S., Eyheramendy, S., Meitinger, T., Cho, Y. S., Kim, H. L., Scott, J., Sehmi, J. S., Hedblad, B., Nilsson, P., Smith, G. D., Raffel, L. J., Yao, J., Schwartz, S. M., Ikram, M., W, L., Mosley, T. H., Seshadri, S., Shrine, N. R., Wain, L. V., Zitting, P., Cooper, J. A., van Gilst, W. H., Janipalli, C. S., Mani, K., Yajnik, C. S., Mattace-Raso, F. U., Lakatta, E. G., Orru, M., Scuteri, A., Ala-Korpela, M., Kangas, A. J., Soininen, P., Tukiainen, T., Würtz, P., Ong, R. T., Dörr, M., Galan, P., Hercberg, S., Lathrop, M., Zelenika, D., Zhai, G., Meschia, J. F., Sharma, P., Terzic, J., Kumar, M., Denniff, M., Zukowska-Szczechowska, E., Wagenknecht, L. E., Fowkes, F., Charchar, F. J., Guo, X., Rotimi, C., Bots, M. L., Brand, E., Talmud, P. J., Nyberg, F., Laan, M., Palmer, L. J., van der Schouw, Y. T., Casas, J. P., Vineis, P., Ganesh, S. K., Wong, T. Y., Tai, E. S., Morris, R. W., Marmot, M. G., Miki, T., Chandak, G. R., Zhu, X., Elosua, R., Soranzo, N., Sijbrands, E. J., Uda, M., Vasan, R. S., Alizadeh, B. Z., de Boer, R. A., Boezen, H. M., Hillege, H. L., van der Klauw, M. M., Ormel, J., Rosmalen, J. G., Slaets, J. P., Lagou, V., Welch, R. P., Wheeler, E., Rehnberg, E., Rasmussen-Torvik, L. J., Lecoeur, C., Johnson, P. C., Sennblad, B., Salo, P., Timpson, N. J., Evans, D. M., St Pourcain, B., Bielak, L. F., Horikoshi, M., Navarro, P., Raychaudhuri, S., Chen, H., Rybin, D., Willems, S. M., Song, K., An, P., Marullo, L., Jansen, H., Pankow, J. S., Edkins, S., Varga, T. V., Oksa, H., Antonella, M., Kong, A., Herder, C., Antti, J., Small, K., Miljkovic, I., Atalay, M., Kiess, W., Smit, J. H., Campbell, S., Fowkes, G. R., Rathmann, W., Maerz, W., Watanabe, R. M., de Geus, E. J., Penninx, B. W., Toenjes, A., Peyser, P. A., Körner, A., Dupuis, J., Cucca, F., Balkau, B., Bouatia-Naji, N., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., Engert, J. C., Morgan, T., Spertus, J. A., Stoll, M., Girelli, D., McKeown, P. P., Patterson, C. C., Merlini, P. A., Berzuini, C., Bernardinelli, L., Peyvandi, F., Tubaro, M., Celli, P., Fetiveau, R., Marziliano, N., Casari, G., Galli, M., Ribichini, F., Rossi, M.,

Bernardi, F., Zonzin, P., Piazza, A., Yee, J., Friedlander, Y., Marrugat, J., Subirana, I., Sala, J., Ramos, R., Williams, G., Nathan, D. M., Macrae, C. A., Berglund, G., Asselta, R., Duga, S., Spreafico, M., Daly, M. J., Nemesh, J., Korn, J. M., Surti, A., Gianniny, L., Parkin, M., Burtt, N., Gabriel, S. B., Wright, B. J., Ball, S. G., Schunkert, I., Linsel-Nitschke, P., Lieb, W., Fischer, M., Grosshennig, A., Preuss, M., Scholz, M., Chen, Z., Wilensky, R., Matthai, W., Qasim, A., Hakonarson, H. H., Devaney, J., Pichard, A. D., Kent, K. M., Satler, L., Lindsay, J. M., Waksman, R., Knouff, C. W., Scheffold, T., Berger, K., Huge, A., Martinelli, N., Olivieri, O., Corrocher, R., Hólm, H., Xie, C., Ahmadi, K. R., Ainali, C., Bataille, V., Bell, J. T., Buil, A., Dermitzakis, E. T., Dimas, A. S., Durbin, R., Glass, D., Hassanali, N., Ingle, C., Knowles, D., Krestyaninova, M., Lowe, C. E., Meduri, E., Di Meglio, P., Montgomery, S. B., Nestle, F. O., Nica, A. C., Nisbet, J., O'rahilly, S., Parts, L., Potter, S., Sekowska, M., Shin, S. Y., Surdulescu, G., Travers, M. E., Tsaprouni, L., Tsoka, S., Wilk, A., Yang, T. P., Higashio, J., Williams, R., Nato, A., Ambite, J. L., Deelman, E., Manolio, T., Heiss, G., Taylor, K., Avery, C., Graff, M., Lin, D., Quibrera, M., Cochran, B., Kao, L., Umans, J., Cole, S., Maccluer, J., Person, S., Gross, M., Fornage, M., Durda, P., Jenny, N., Patsy, B., Arnold, A., Buzkova, P., Haines, J., Murdock, D., Glenn, K., Brown-Gentry, K., Thornton-Wells, T., Dumitrescu, L., Bush, W. S., Mitchell, S. L., Goodloe, R., Wilson, S., Boston, J., Malinowski, J., Restrepo, N., Oetjens, M., Fowke, J., Spencer, K., Pendergrass, S., Park, L., Tiirikainen, M., Kolonel, L., Cheng, I., Wang, H., Shohet, R., Stram, D., Henderson, B., Monroe, K., Anderson, G., Carlson, C., Prentice, R., Lacroix, A., Wu, C., Carty, C., Rosse, S., Young, A., Kocarnik, J., Lin, Y., Jackson, R., Duggan, D., Kuller, L., He, C., Sulem, P., Barbalic, M., Broer, L., Byrne, E. M., Gudbjartsson, D. F., McArdle, P. F., Porcu, E., van Wingerden, S., Zhuang, W. V., Lauc, L. B., Broekmans, F. J., Burri, A., Chen, C., Corre, T., Coviello, A. D., D'adamo, P., Davies, G., Deary, I. J., Ebrahim, S., Fauser, B. C., Ferreli, L., Folsom, A. R., Hankinson, S. E., Hass, M., Janssens, A. C., Karasik, D., Keyzer, J., Kiel, D. P., Lahti, J., Lai, S., Laisk, T., Laven, J. S., Liu, J., Lopez, L. M., Louwers, Y. V., Marongiu, M., Klaric, I. M., Masciullo, C., Melzer, D., Newman, A. B., Paré, G., Peeters, P. H., Pop, V. J., Räikkönen, K., Salumets, A., Stacey, S. N., Starr, J. M., Stathopoulou, M. G., Styrkarsdottir, U., Tenesa, A., Tryggvadottir, L., Tsui, K., van Dam, R. M., van Gils, C. H., van Nierop, P., Vink, J. M., Voorhuis, M., Widen, E., Wijnands-Van Gent, C. J., Yerges-Armstrong, L. M., Zgaga, L., Zygmunt, M., Buring, J. E., Crisponi, L., Demerath, E. W., Streeten, E. A., Murray, A., Visser, J. A., Lunetta, K. L., Elks, C. E., Cousminer, D. L., Koller, D. L., Lin, P., Smith, E. N., Warrington, N. M., Alavere, H., Berenson, G. S., Blackburn, H., Busonero, F., Chen, W., Couper, D., Easton, D. F., Eriksson, J., Foroud, T., Kilpeläinen, T. O., Li, S., Murray, S. S., Ness, A. R., Northstone, K., Peacock, M., Pennell, C. E., Pharoah, P., Rafnar, T., Rice, J. P., Ring, S. M., Schork, N. J., Segrè, A. V., Sovio, U., Srinivasan, S. R., Tammesoo, M. L., van Meurs, J. B., Young, L., Bierut, L. J., and Econs, M. J. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197.

[20] Ludwig-Słomczyńska, A. H., Seweryn, M. T., Kapusta, P., Pitera, E., Mantaj, U., Cyganek, K., Gutaj, P., Dobrucka, Ł., Wender-Ożegowska, E., Małecki, M. T., and Wołkow, P. P. (2021). The transcriptome-wide association search for genes and genetic variants which associate with BMI and gestational weight gain in women with type 1 diabetes. *Molecular Medicine*, 27(1):1–16.

[21] Macdonald-Dunlop, E., Taba, N., Klarić, L., Frkatović, A., Walker, R., Hayward, C., Esko, T., Haley, C., Fischer, K., Wilson, J. F., and Joshi, P. K. (2022). A catalogue of omics biological ageing clocks reveals substantial commonality and associations with disease risk. *Aging*, 14(2):623–659.

[22] Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469–480.

[23] Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.

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

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

[26] Moreno-Castellanos, N., Rodríguez, A., Rabanal-Ruiz, Y., Fernández-Vega, A., López-Miranda, J., Vázquez-Martínez, R., Frühbeck, G., and Malagón, M. M. (2017). The cytoskeletal protein septin 11 is associated with human obesity and is involved in adipocyte lipid storage and metabolism. *Diabetologia*, 60(2):324–335.

[27] Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination.

[28] Ni, G., Zeng, J., Revez, J. A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D. R., Coleman, J. R., Smoller, J. W., Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., Pers, T. H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S. A., Begemann, M., Belliveau, R. A., Bene, J., Bergen, S. E., Bevilacqua, E., Bigdeli, T. B., Black, D. W., Bruggeman, R., Buccola, N. G., Buckner, R. L., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R. M., Carr, V. J., Carrera, N., Catts, S. V., Chambert, K. D., Chan, R. C., Chen, R. Y., Chen, E. Y., Cheng, W., Cheung, E. F., Chong, S. A., Cloninger, C. R., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J. J., Davidson, M., Davis, K. L., Degenhardt, F., Del Favero, J., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A. H., Farrell, M. S., Frank, J., Franke, L., Freedman, R., Freimer, N. B., Friedl, M., Friedman, J. I., Fromer, M., Genovese, G., Georgieva, L., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J. I., Golimbet, V., Gopal, S., Gratten, J., de Haan, L., Hammer, C., Hamshere, M. L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A. M., Henskens, F. A., Herms, S., Hirschhorn, J. N., Hoffmann, P., Hofman, A., Hollegaard, M. V., Hougaard, D. M., Ikeda, M., Joa, I., Julià, A., Kahn, R. S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M. C., Kennedy, J. L., Khrunin, A., Kim, Y., Klovins,

J., Knowles, J. A., Konte, B., Kucinskas, V., Kucinskiene, Z. A., Kuzelova-Ptackova, H., Kähler, A. K., Laurent, C., Lee, J., Lee, S. H., Legge, S. E., Lerer, B., Li, M., Li, T., Liang, K. Y., Lieberman, J., Limborska, S., Loughland, C. M., Lubinski, J., Lönnqvist, J., Macek, M., Magnusson, P. K., Maher, B. S., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R. W., McDonald, C., McIntosh, A. M., Meier, S., Meijer, C. J., Melegh, B., Melle, I., Mesholam-Gately, R. I., Metspalu, A., Michie, P. T., Milani, L., Milanova, V., Mokrab, Y., Morris, D. W., Mors, O., Murphy, K. C., Murray, R. M., Myin-Germeys, I., Müller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D. A., Nestadt, G., Nicodemus, K. K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F. A., Oh, S. Y., Olincy, A., Olsen, L., Van Os, J., International Consortium, P. E., Pantelis, C., Papadimitriou, G. N., Papiol, S., Parkhomenko, E., Pato, M. T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D. O., Pietiläinen, O., Pimm, J., Pocklington, A. J., Powell, J., Price, A., Pulver, A. E., Purcell, S. M., Quested, D., Rasmussen, H. B., Reichenberg, A., Reimers, M. A., Richards, A. L., Roffman, J. L., Roussos, P., Ruderfer, D. M., Salomaa, V., Sanders, A. R., Schall, U., Schubert, C. R., Schulze, T. G., Schwab, S. G., Scolnick, E. M., Scott, R. J., Seidman, L. J., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J. M., Sim, K., Slominsky, P., So, H. C., Spencer, C. C., Stahl, E. A., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R. E., Strengman, E., Strohmaier, J., Stroup, T. S., Subramaniam, M., Suvisaari, J., Svrakic, D. M., Szatkiewicz, J. P., Söderman, E., Thirumalai, S., Toncheva, D., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B. T., Weiser, M., Wildenauer, D. B., Williams, N. M., Williams, S., Witt, S. H., Wolen, A. R., Wong, E. H., Wormley, B. K., Xi, H. S., Zai, C. C., Zheng, X., Zimprich, F., Wray, N. R., Stefansson, K., Visscher, P. M., Case-Control Consortium, W. T., Adolfsson, R., Andreassen, O. A., Blackwood, D. H., Bramon, E., Buxbaum, J. D., Børglum, A. D., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P. V., Gill, M., Gurling, H., Hultman, C. M., Iwata, N., Jablensky, A. V., Jönsson, E. G., Kendler, K. S., Kirov, G., Knight, J., Lencz, T., Levinson, D. F., Li, Q. S., Liu, J., Malhotra, A. K., McCarroll, S. A., McQuillin, A., Moran, J. L., Mortensen, P. B., Mowry, B. J., Nöthen, M. M., Ophoff, R. A., Owen, M. J., Palotie, A., Pato, C. N., Petryshen, T. L., Posthuma, D., Rietschel, M., Riley, B. P., Rujescu, D., Sham, P. C., Sklar, P., St Clair, D., Weinberger, D. R., Wendland, J. R., Werge, T., Daly, M. J., Sullivan, P. F., O'Donovan, M. C., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Air, T. M., Andlauer, T. F., Bacanu, S. A., Bækvad-Hansen, M., Beekman, A. T., Binder, E. B., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., Christensen, J. H., Clarke, T. K., Colodro-Conde, L., Couvy-Duchesne, B., Crawford, G. E., Davies, G., Deary, I. J., Derks, E. M., Direk, N., Dolan, C. V., Dunn, E. C., Eley, T. C., Hassan Kiadeh, F. F., Finucane, H. K., Foo, J. C., Forstner, A. J., Gaspar, H. A., Goes, F. S., Gordon, S. D., Grove, J., Hall, L. S., Hansen, C. S., Hansen, T. F., Hickie, I. B., Homuth, G., Horn, C., Hottenga, J. J., Howard, D. M., Ising, M., Jansen, R., Jones, I., Jones, L. A., Jorgenson, E., Kohane, I. S., Kraft, J., Kretzschmar, W. W., Kutalik, Z., Li, Y., Lind, P. A., MacIntyre, D. J., MacKinnon, D. F., Maier, R. M., Marchini, J., Mbarek, H., McGrath, P., McGuffin, P., Medland, S. E., Mehta, D., Middeldorp, C. M., Mihailov, E., Milaneschi, Y., Mondimore, F. M., Montgomery, G. W., Mostafavi, S., Mullins, N., Nauck, M., Ng, B., Nivard, M. G., O'Reilly, P. F., Oskarsson, H., Painter, J. N., Pedersen, C. B., Pedersen, M. G., Peterson, R. E., Peyrot, W. J., Pistis, G., Quiroz, J. A., Qvist, P., Rice, J. P., Rivera,

M., Mirza, S. S., Schoevers, R., Schulte, E. C., Shen, L., Shyn, S. I., Sinnamon, G. C., Smit, J. H., Smith, D. J., Streit, F., Tansey, K. E., Teismann, H., Teumer, A., Thompson, W., Thomson, P. A., Thorgeirsson, T. E., Traylor, M., Treutlein, J., Trubetskoy, V., Uitterlinden, A. G., Umbricht, D., Van der Auwera, S., van Hemert, A. M., Viktorin, A., Wang, Y., Weinsheimer, S. M., Wellmann, J., Willemsen, G., Wu, Y., Xi, H. S., Yang, J., Zhang, F., Arolt, V., Baune, B. T., Berger, K., Boomsma, D. I., Dannlowski, U., de Geus, E. J., DePaulo, J. R., Domschke, K., Grabe, H. J., Hamilton, S. P., Hayward, C., Heath, A. C., Kloiber, S., Lewis, G., Lucae, S., Madden, P. A., Magnusson, P. K., Martin, N. G., Nordentoft, M., Paciga, S. A., and Pedersen, N. L. (2021). A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biological psychiatry*, 90(9):611.

[29] Nikolac Perkovic, M., Pucic Bakovic, M., Kristic, J., Novokmet, M., Huffman, J. E., Vitart, V., Hayward, C., Rudan, I., Wilson, J. F., Campbell, H., Polasek, O., Lauc, G., and Pivac, N. (2014). The association between galactosylation of immunoglobulin G and body mass index. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 48:20–25.

[30] Pain, O., Glanville, K. P., Hagenaars, S. P., Selzam, S., Furtjes, A. E., Gaspar, H. A., Coleman, J. R., Rimfeld, K., Breen, G., Plomin, R., Folkersen, L., and Lewis, C. M. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genetics*, 17(5 May):e1009021.

[31] Payab, M., Tayanloo-Beik, A., Falahzadeh, K., Mousavi, M., Salehi, S., Djalalinia, S., Ebrahimpur, M., Rezaei, N., Rezaei-Tavirani, M., Larijani, B., Arjmand, B., and Gilany, K. (2022). Metabolomics prospect of obesity and metabolic syndrome; a systematic review.

[32] Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2020). LDpred2: Better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431.

[33] Privé, F., Vilhjálmsson, B. J., Aschard, H., and Blum, M. G. (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. *American Journal of Human Genetics*, 105(6):1213–1221.

[34] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

[35] Robinson, M. R., Kleinman, A., Graff, M., Vinkhuyzen, A. A., Couper, D., Miller, M. B., Peyrot, W. J., Abdellaoui, A., Zietsch, B. P., Nolte, I. M., Van Vliet-Ostaptchouk, J. V., Snieder, H., Medland, S. E., Martin, N. G., Magnusson, P. K., Iacono, W. G., McGue, M., North, K. E., Yang, J., and Visscher, P. M. (2017). Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, 1(1).

[36] Shadrina, A. S., Zlobin, A. S., Zaytseva, O. O., Klarić, L., Sharapov, S. Z., D Pakhomov, E., Perola, M., Esko, T., Hayward, C., Wilson, J. F., Lauc, G., Aulchenko, Y. S.,

and Tsepilov, Y. A. (2021). Multivariate genome-wide analysis of immunoglobulin G N-glycosylation identifies new loci pleiotropic with immune function. *Human Molecular Genetics*, 30(13):1259–1270.

[37] Shen, X., Klarić, L., Sharapov, S., Mangino, M., Ning, Z., Wu, D., Trbojević-Akmačić, I., Pučić-Baković, M., Rudan, I., Polašek, O., Hayward, C., Spector, T. D., Wilson, J. F., Lauc, G., and Aulchenko, Y. S. (2017). Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. *Nature Communications*, 8(1):1–10.

[38] Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies.

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

[40] Suhre, K. (2015a). A Table of all published GWAS with glycomics – Human Metabolic Individuality.

[41] Suhre, K. (2015b). A Table of all published GWAS with metabolomics – Human Metabolic Individuality.

[42] Suhre, K. (2017). A Table of all published GWAS with proteomics – Human Metabolic Individuality.

[43] Sun, Y. V. and Hu, Y. J. (2016). Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. *Advances in genetics*, 93:147.

[44] Testa, R., Vanhooren, V., Bonfigli, A. R., Boemi, M., Olivieri, F., Ceriello, A., Genovese, S., Spazzafumo, L., Borelli, V., Bacalini, M. G., Salvioli, S., Garagnani, P., Dewaele, S., Libert, C., and Franceschi, C. (2015). N-Glycomic changes in serum proteins in type 2 diabetes mellitus correlate with complications and with metabolic syndrome parameters. *PLoS ONE*, 10(3):e0119983.

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

[46] Uh, H. W., Klaric, L., Ugrina, I., Lauc, G., Smilde, A. K., and Houwing-Duistermaat, J. J. (2020). Choosing proper normalization is essential for discovery of sparse glycan biomarkers. *Molecular Omics*, 16(3):231–242.

[47] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

[48] Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J. (1984). 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.

[49] Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., Yengo, L., Lloyd-Jones, L. R., Sidorenko, J., Wu, Y., Agbessi, M., Ahsan, H., Alves, I., Andiappan, A., Awadalla, P., Battle, A., Beutner, F., Bonder, M. J. J., Boomsma, D., Christiansen, M., Claringbould, A., Deelen, P., Esko, T., Favé, M. J., Franke, L., Frayling, T., Gharib, S., Gibson, G., Hemani, G., Jansen, R., Kähönen, M., Kalnapenkis, A., Kasela, S., Kettunen, J., Kim, Y., Kirsten, H., Kovacs, P., Krohn, K., Kronberg-Guzman, J., Kukushkina, V., Kutalik, Z., Lee, B., Lehtimäki, T., Loeffler, M., Marigorta, U. M., Metspalu, A., Milani, L., Müller-Nurasyid, M., Nauck, M., Nivard, M., Penninx, B., Perola, M., Pervjakova, N., Pierce, B., Powell, J., Prokisch, H., Psaty, B., Raitakari, O., Ring, S., Ripatti, S., Rotzschke, O., Ruëger, S., Saha, A., Scholz, M., Schramm, K., Seppälä, I., Stumvoll, M., Sullivan, P., Teumer, A., Thiery, J., Tong, L., Tönjes, A., van Dongen, J., van Meurs, J., Verlouw, J., Völker, U., Võsa, U., Yaghootkar, H., Zeng, B., McRae, A. F., Visscher, P. M., Zeng, J., and Yang, J. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications*, 9(1):1–14.

[50] Yang, S. and Zhou, X. (2020). Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. *American Journal of Human Genetics*, 106(5):679–693.

[51] Yun, L., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation.

[52] Zaghlool, S. B., Sharma, S., Molnar, M., Matías-García, P. R., Elhadad, M. A., Waldenberger, M., Peters, A., Rathmann, W., Graumann, J., Gieger, C., Grallert, H., and Suhre, K. (2021). Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nature Communications 2021 12:1*, 12(1):1–13.

[53] Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications*, 12(1):1–9.

4

INVESTIGATING THE IMPACT OF DOWN SYNDROME ON METHYLATION AND GLYCOMICS WITH TWO-STAGE PO2PLS

Zhujie Gu, Said el Bouhaddani, Jeanine Houwing-Duistermaat, Hae Won Uh. Investigating the impact of Down syndrome on methylation and glycomics with two-stage PO2PLS. *Theoretical Biology Forum*. 2021; 114(1-2). DOI: 10.19272/202111401004

ABSTRACT

Down syndrome (DS) is a condition that leads to precocious and accelerated aging in affected subjects. Several alterations in DS cases have been reported at a molecular level,particularly in methylation and glycosylation. Investigating the relation between methylation, glycomics and DS can lead to new insights underlying the atypical aging. We consider a data integration approach, where we investigate how DS affects the parts of glycomics and methylation which are correlated, and which CpG sites and glycans are relevant. Our motivating datasets consist of methylation and glycomics data, measured on 29 DS patients and their unaffected siblings and mothers. The family-based case-control design needs to be taken into account when studying the relationship between methylation, glycomics and DS.

We propose a two-stage approach to first integrate methylation and glycomics data, and then link the joint information to Down syndrome. For the data integration step, we consider probabilistic two-way orthogonal partial least squares (PO2PLS). PO2PLS models two omics datasets in terms of low-dimensional joint and omic-specific latent components, and takes into account heterogeneity across the omics data. The relationship between the omics data can be statistically tested. The joint components represent the joint information in methylation and glycomics. In the second stage,we apply a linear mixed model to the relationship between DS and the joint methylation and glycomics components. For the components that are significantly associated with DS, we identify the most important CpG sites and glycans.

A simulation study is conducted to evaluate the performance of our approach. The results showed that the effects of DS on the omics data can be detected in a large sample size, and the accuracy of the feature selection was high in both small and large sample sizes. Our approach is applied to the DS datasets, a significant effect of DS on the joint components is found. The identified CpG sites and glycans appeared to be related to DS. Our proposed method that jointly analyzes multiple omics data with an outcome variable may provide new insight into the molecular implications of DS at different omics levels.

4.1. INTRODUCTION

Down syndrome (DS) is a common (14.47 per 10,000 live births [23]) but complex genetic condition, caused by a total or partial trisomy of the chromosome 21 (HSA21). Subjects with DS exhibit a number of characteristics associated with older age in the general population. This has led to the classical description of DS as a progeroid syndrome and as a model of precocious and accelerated aging [8, 24, 37, 12]. On the molecular level, alterations by DS were reported across multiple omics, including DNA methylation [2], transcriptomics [33], and glycomics [5]. It is worth noting that these alternations in omics also correspond to precocious and accelerated biological aging in DS subjects, based on the aging biomarkers developed on different omics levels [11, 12], e.g., the methylation-based epigenetic clocks [15, 16], and the glycomic-based GlycoAge [5]. However, previous studies on DS were conducted on a single omics level, overlooking the interactions between omics. For example, studies of glycomics alone ignore the possibility that changes in glycomics represent by-products of some other age-driven processes [19], say, DNA methylation. Therefore, a unified view on the molecular implications of DS requires a data integration approach that utilizes information across omics datasets.

Our motivating data come from a family-based case-control study consists of 29 families. Each family is composed of a DS subject, an unaffected sibling and the mother. Such design is often used to control for genetic and environmental influences in a study. Particularly, both methylation and glycomics can be influenced by genetics and also respond to environmental changes, therefore controlling the two factors allows us to focus on the molecular consequences of DS. The methylation dataset contains 450981 CpG sites and the glycomics data consist of 10 glycans. We aim to integrate the methylation and glycomics data, investigate the impact of DS on both omics, and identify the target CpG sites and glycans jointly.

A data integration approach poses several statistical and computational challenges: i) the high-dimensionality of the methylation data (i.e., 450K CpG sites vs 85 samples); ii) the complex correlation structure within and between omics; iii) the heterogeneity between omics (resulting from different source of variation, measurement platform, and biological representation); iv) and assessment of statistical evidence for a relation between omics data and between the DS. To deal with the first two issues, we consider partial least squares (PLS) [6, 35]. Dimension reduction is achieved by decomposing two omics datasets X and Y into low-dimensional joint and residual parts. The joint part of one dataset represents the best approximation of X or Y based on maximizing the covariance of the two. However, by integrating two heterogeneous omics datasets, the PLS joint parts also contain (strong) omic-specific variation. Two-way orthogonal partial least squares (O2PLS) [30, 9] was proposed to capture the omic-specific variation using a specific part for each omics dataset, making the joint parts better estimates for the true relationship between X and Y. O2PLS is algorithmic, thus cannot provide statistical evidence for the relationship between omics data. To allow for statistical inference, probabilistic two-way partial least squares (PO2PLS) [10] was recently developed, which models the two omics data in a probabilistic framework. Since PO2PLS performs statistical inference only on the relationship between the two omics datasets, the DS is not included in the model. To assess the relationship between DS and the two omics

jointly, we propose a two-stage PO2PLS approach.

In the first stage, PO2PLS is used to construct latent variables that reduce the dimension of omics data. Also, we test the association of the joint components from methylation and glycomics data. In the second stage, the joint latent variables identified from the first stage are considered pseudo outcomes influenced by having DS. A linear mixed model is used to include age, DS status, interaction of age and DS, sex for fixed effects, and a random intercept for each family in the model.

The rest of paper is organized as follows. We first describe the data in detail. In the methods section, an overview of the two-stage PO2PLS approach is presented, followed by the formulation of each step. Via a simulation study based on the DS dataset, the performance of the method is investigated. We then apply the method to the DS datasets and give interpretation of the results. We conclude with a discussion and possible directions to further extend the method.

4.2. METHODS

We propose a two-stage integration approach to investigate the effect of Down syndrome on methylation and glycomics jointly. In the first stage, PO2PLS decomposes the two omics datasets into joint and omic-specific parts, where the joint parts represent the relationship between methylation and glycomics. In the second stage, we model the relationship between Down syndrome and the joint parts. Additionally, we study how the joint parts reflect the effect of age and the interaction between age and DS.

An overview of the approach is illustrated in Figure 4.1.

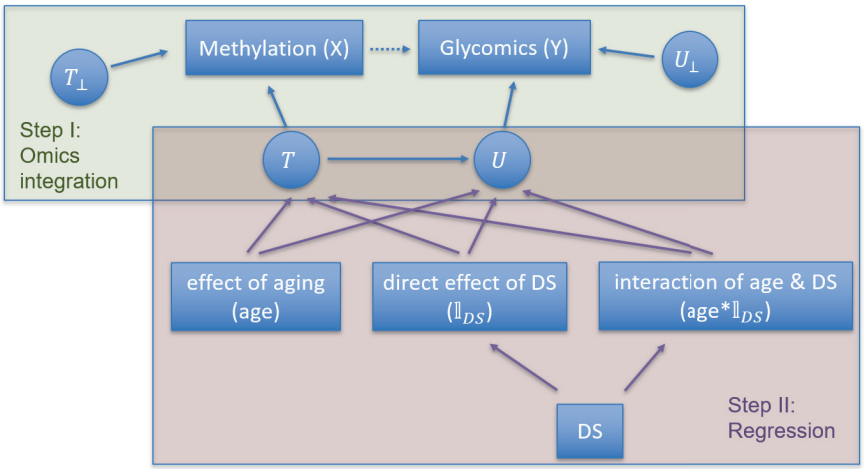


Figure 4.1: **An illustration of the two-stage PO2PLS approach.** In the first stage, low-dimensional joint components (T and U) are constructed to capture the relationship between the omics, while correcting for the omic-specific variation using specific components ($T_\perp$ and $U_\perp$). In the second, the components T and U are modeled with age, DS status, and their interaction.

4.2.1. STAGE I: JOINT MODELING OF METHYLATION AND GLYCOMICS

Let x and y be random vectors of dimensions p and q, respectively. In our context, x represents methylation and y represents glycomics. In the PO2PLS model [10], x and y are decomposed in joint, specific and residual parts. The joint parts consist of latent variables t and u of size K. The specific parts are given by latent variables $t_\perp$ of size K_x and $u_\perp$ of size K_y. Further, the residual random vectors are denoted by e, f, and h. In the PO2PLS model, h represents heterogeneity between t and u. A graphical depiction of the model is given in the green part of Figure 4.1. The PO2PLS model is written as

$$\begin{aligned} x &= tW^\top + t_\perp W_\perp^\top + e, \\ y &= uC^\top + u_\perp C_\perp^\top + f, \\ u &= tB + h. \end{aligned} \tag{4.1}$$

The loading parameters $W\,(p \times K)$ and $C\,(q \times K)$ are matrices that contain weights for each variable in x and y, respectively, per component. These weights indicate which variables are important for the joint latent variables and can be used to identify the most relevant variables. Similarly, the matrices $W_\perp\,(p \times K_x)$ and $C_\perp\,(q \times K_y)$ contain weights for the omic-specific latent variables. The $K \times K$ diagonal matrix B links the joint components t and u and hence represents the relationship between x and y.

The latent variables t, $t_\perp$, $u_\perp$, and the residuals h, e and f are assumed to be normally distributed with zero mean and respective covariance matrices Σ_t, $\Sigma_{t_\perp}$, $\Sigma_{u_\perp}$, Σ_h, $\sigma_e^2 I_p$, and $\sigma_f^2 I_q$. This yields a joint distribution of $(x, y) \sim \mathcal{N}(0, \Sigma_\theta)$, with covariance matrix Σ_θ depending on the parameters $\theta = \{W, C, W_\perp, C_\perp, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \sigma_e^2, \sigma_f^2, \Sigma_h\}$.

To estimate the parameters in θ, we use maximum likelihood. The log-likelihood function is written by

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

A direct optimization is complex and computationally infeasible in high dimensional settings (i.e., when p or q is large). In the PO2PLS model, an efficient Expectation–Maximization (EM) algorithm [7] is proposed. For a full description of the algorithm, we refer to [10].

STATISTICAL INFERENCE OF THE RELATIONSHIP BETWEEN OMICS

Within the PO2PLS framework, statistical evidence for the relationship between methylation and glycomics is assessed. We test the null hypothesis of no relationship between each pair of joint latent variables. We consider the following hypothesis for each component k between 1 and K, $H_0 : B_k = 0$ against $H_1 : B_k \neq 0$. Let X and Y be methylation and glycomics data matrices consisting of N draws from x and y. To test the hypothesis of no relationship between methylation and glycomics, the test statistic $T_B = \hat{B}_k / \widehat{SE}_{\hat{B}_k}$ is calculated for each component. Using an estimate for $SE_{\hat{B}_k}$ based on X and Y, and the asymptotic distribution of T_{B_k} (see [10]), a p-value is calculated per component. The components that showed significant association (at the nominal level of 0.05) are then used as outcome in a linear mixed model with Down syndrome and age. Since the components are not observed, they are reconstructed using a linear projection of X and Y

on W and C, respectively, as follows,

$$T = (X - X\hat{W}_\perp \hat{W}_\perp^\top)\hat{W},$$
$$U = (Y - Y\hat{C}_\perp \hat{C}_\perp^\top)\hat{C}.$$

All the parameter estimation and the statistical testing in the first stage were performed in R using 'PO2PLS' package (https://github.com/selbouhaddani/PO2PLS).

4.2.2. STAGE II: MODELING THE EFFECT OF DOWN SYNDROME AND AGING ON METHYLATION AND GLYCOMICS JOINTLY

In the second stage of our proposed framework, we take the significant joint components from the first stage as a representation of the joint information in methylation and glycomics. We investigate the direct effect of Down syndrome on the joint components, as well as whether a part of the joint variation in methylation and glycomics can be be explained by age and the interaction between age and DS.

We propose a linear mixed model with the joint components as outcome. The age, DS status and their interaction are taken as fixed effects. We further include sex as a fixed effect. The family structure is considered by including a random intercept for each family in the model. The full model is formulated as

$$T = \beta_0 + \beta_1 \text{Age} + \beta_2 \mathbb{1}_{DS} + \beta_3 \text{Age} * \mathbb{1}_{DS} + \beta_4 \text{Sex} + \text{Family} + \epsilon \tag{4.2}$$

where $\mathbb{1}_{DS}$ is one for Down syndrome cases and zero for controls. In the model, β_1, β_2, β_3, β_4 are the effect sizes of age, DS status, interaction between age and DS, and sex, respectively. The parameter β_0 is the intercept, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the individual-level error. These parameters are estimated using the restricted maximum likelihood (REML) [25] approach with R package 'lme4' [3]. Since the scale of the joint scores and effect sizes is not directly interpretable, we focus on the directions of the effects and the significance levels. For the components that have a significant association with DS or its interaction with age, we calculate the top loading values and their methylation sites and glycans and compare them to results from previous studies.

4.3. SIMULATION

A simulation study is conducted to evaluate the performance of our two-stage integration approach. We focus on detecting the effect of age, DS, and their interaction on the multi-omics joint parts, and the accuracy of identifying relevant features. We consider two scenarios with different sample sizes. In the small sample size scenario, we take $N = 85$, which is the same as in our data analysis. In the large sample size scenario, we choose $N = 900$, with 300 simulated Down syndrome patients, siblings and mothers. The dimensions of X and Y are $p = 400000$ and $q = 10$, similar as in our data analysis. Further, we take 1 joint component, and 1 specific component for each dataset

Rather than simulating T independently, we introduce a dependence on DS and age. We consider the fitted values from the linear model,

$$\hat{T} = \beta_0 + \beta_1 \text{Age} + \beta_2 \mathbb{1}_{DS} + \beta_3 \text{Age} * \mathbb{1}_{DS}.$$

Here, $\hat{T}$ represents the part of T that is dependent on age and DS. We do not include a random residual term, but regard $\hat{T}$ as fixed across the simulation runs. Namely, random variation is introduced in the observed X and Y via h, e, and f (4.1), which leads to uncertainty in predicting T from X and Y. In the small sample size scenario, we use the observed age and DS status from the data analysis. In the large sample size scenario, age is generated once per subject group (DS, sibling, mother). The ages of the 300 subjects in each group are generated from a normal distributions with mean and variance estimated for the corresponding group in our dataset. Across the simulation runs, the age and DS status are kept fixed. The parameters are $\beta_1 = 1$, $\beta_2 = 20$, and $\beta_3 = 0.5$ for the coefficients of age, Down syndrome, and their interaction. The intercept β_0 is chosen such that the resulting $\hat{T}$ has mean zero.

Table 4.1: **Simulation parameters**

Parameters	Description	Value/Distribution	Notes
N	Sample size	85; 900	
p	Dimension of X	400000	
q	Dimension of Y	10	
t	X joint component	Equation (3)	14.4% of variance in X
$t_\perp$	X specific component	$N(0, 180)$	8.7% of variance in X
e	Noise in X	$N(\mathbf{0}, 0.004 * \boldsymbol{I}_p)$	76.9% of variance in X
u	Y joint component	$tB + h$	66.9% of variance in Y
$u_\perp$	Y specific component	$N(0, 0.25)$	7.9% of variance in Y
f	Noise in Y	$N(\mathbf{0}, 0.08 * \boldsymbol{I}_q)$	25.2% of variance in Y
h	Noise in u	$N(0, 2)$	94.3% of variance in u
B	Coefficient $t \rightarrow u$	0.02	

Table 4.1 shows the parameter choices of the other components. They are chosen such that the proportions of explained variance in simulated data by each part match the proportions estimated from the real dataset. The loading values W are generated once from the $\mathcal{N}(1, 0.5^2)$ distribution. We then retain 1000 non-zero values, representing 1000 relevant CpG sites, and set all other values to zero. The loading values for the other components are generated once from a standard normal distribution. All loading vectors are scaled to have unit norm, which is a constraint in the PO2PLS model.

After choosing all values for age, DS, and the parameters, the data matrices X and Y are simulated 100 times based on the PO2PLS model (4.1), with $\hat{T}$ as a function of age and DS.

In both the small and large sample size scenarios, we fit two PO2PLS models to the simulated data. In the first fit, the correct number of components ($K = 1$) is used, while in the other fit, too many joint components ($K = 3$) are chosen. In the second stage, the first X-joint component is regressed on age, DS status, and their interaction term and the estimated coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ with their p-values are calculated. We visual-

ize the distributions of these estimates and calculate the proportion of each coefficient being significant (p-value < 0.05) across 100 simulation runs. In each run where at least one of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ is significant, we compare the top 1000 variables with the highest estimated absolute loading values to the true 1000 selected variables. The true positive rate (TPR = TP/(TP+FN), where TP = True Positive, FN = False Negative) is then calculated and averaged.

4.3.1. RESULTS

Results for the small sample size scenario are shown in Figure 4.2. The distribution of the regression coefficients and their p-values are plotted in the case where one joint component ($K = 1$) was used for PO2PLS. The estimates of the age effect $\hat{\beta}_1$, DS effect $\hat{\beta}_2$, and effect of interaction between age and DS status $\hat{\beta}_3$ were all symmetrically distributed. The means appeared to be biased towards zero. In terms of p-value, about 52% of $\hat{\beta}_1$ were significant (p-value < 0.05). The proportions of significant $\hat{\beta}_2$ and $\hat{\beta}_3$ were both below 0.05. The TPR of the estimated top 1000 variables was 0.93. In the case where too many joint components ($K = 3$) were specified in PO2PLS (Figure 4.3), the distributions of the estimated coefficients remained largely unchanged. The proportion of significant $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ slightly increased to 0.58, 0.05, and 0.05. The TPR decreased to 0.86.

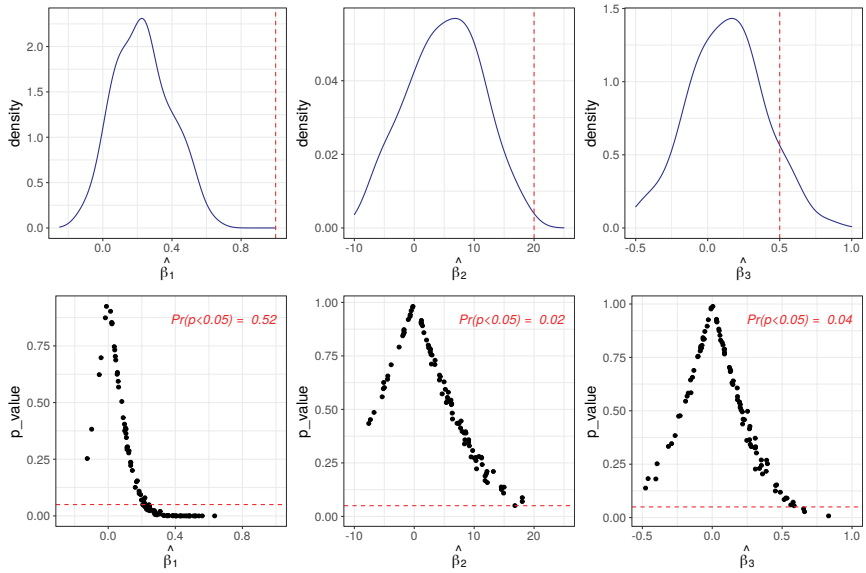


Figure 4.2: **Scenario N=85, 1 joint component: distribution of $\hat{\beta}$ (top) and their p-values (bottom).** In the columns are estimates of the age affect $\hat{\beta}_1$, DS effect $\hat{\beta}_2$, and effect of interaction between age and DS $\hat{\beta}_3$, respectively. The red vertical line in the left plot is the true value. The red horizontal line in the right plot is the threshold of p-value: 0.05.

Results for the large sample size scenarios are shown in Figure 4.4. The means of the coefficients when fitting $K = 1$ joint component were closer to the true values, compared to the $N = 85$ scenario. However, the distributions were more left-skewed. The propor-

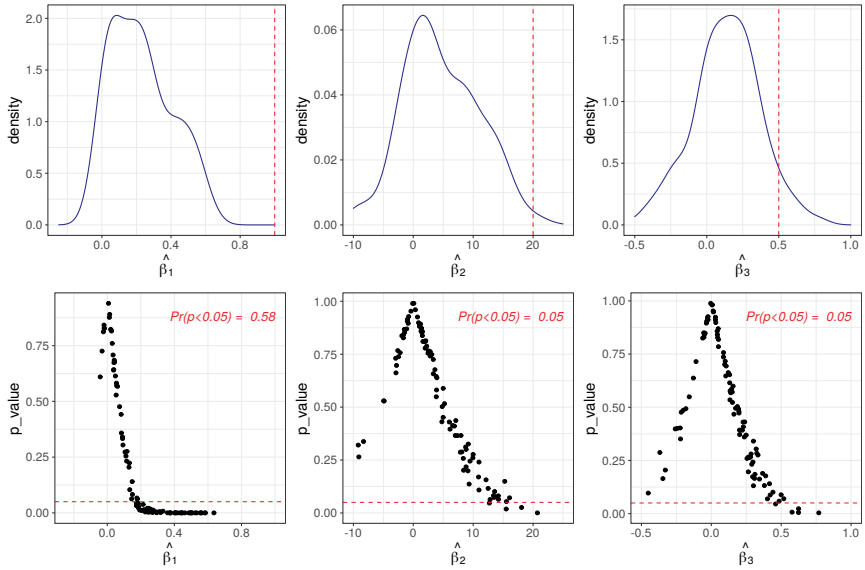


Figure 4.3: **Scenario N=85, 3 joint component: distribution of $\hat{\beta}$ (top) and their p-values (bottom).** The red vertical line in the left plot is the true value. The red horizontal line in the right plot is the threshold of p-value: 0.05.

tion of significant coefficients was higher: 0.98 for $\hat{\beta}_1$, and 0.82 for $\hat{\beta}_2$ and $\hat{\beta}_3$. The TPR also improved to 0.99. In the case where we choose $K = 3$ joint components, as shown in Figure 4.5, the performance of coefficient estimation improved compared to the normal sample size case. The means were closer to the true values and the standard deviation decreased. The proportion of significant $\hat{\beta}_2$ and $\hat{\beta}_3$ improved to 0.94 and 0.97. The TPR was around 0.98.

4.4. APPLICATION TO DOWN SYNDROME DATA

We apply the two-stage PO2PLS approach to the Down syndrome dataset. We first integrate the methylation and glycomics data. Then the effect of DS on the joint parts of both omics is investigated. Finally, for the joint components that are significantly associated with DS, we identify the target CpG sites and glycans.

4.4.1. DATA DESCRIPTION

Whole blood methylation (using Infinium HumanMethylation450 BeadChip) [2] and plasma N-glycans (using DSA-FACE) [5] data were measured on 29 trios composed by a down syndrome subject (DS), one non-affected sibling (SB), and the mother (MA). Due to missing data in 2 siblings, the sample size N is 85. The age of the DS group ranges from 10 to 43, with a median of 24. Ages of the siblings are roughly matched with the DS subjects, ranging from 14 to 52. The mothers are aged between 41 and 83, with a median of 57. The DS group has 18 males and 11 females, while the SB group contains

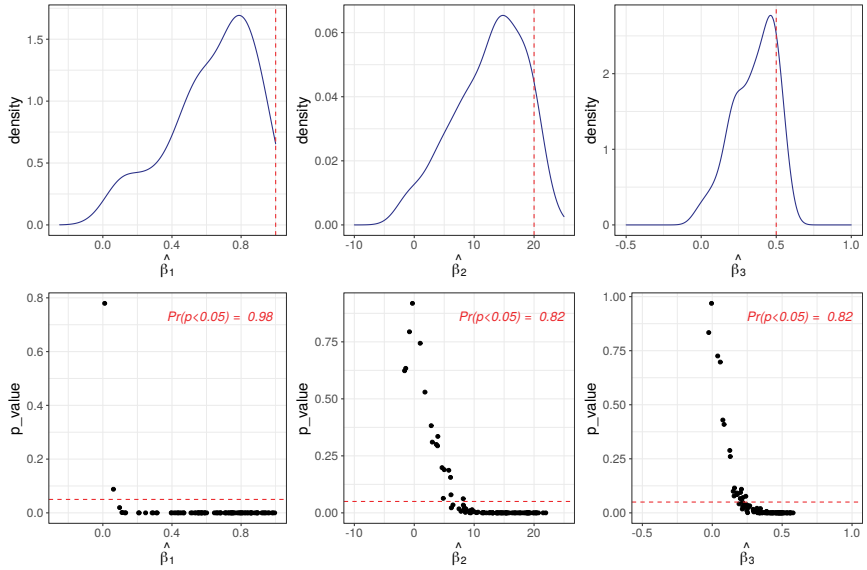


Figure 4.4: **Scenario N=900, 1 joint component: distribution of $\hat{\beta}$ (top) and their p-values (bottom).** The red vertical line in the left plot is the true value. The red horizontal line in the right plot is the threshold of p-value: 0.05.

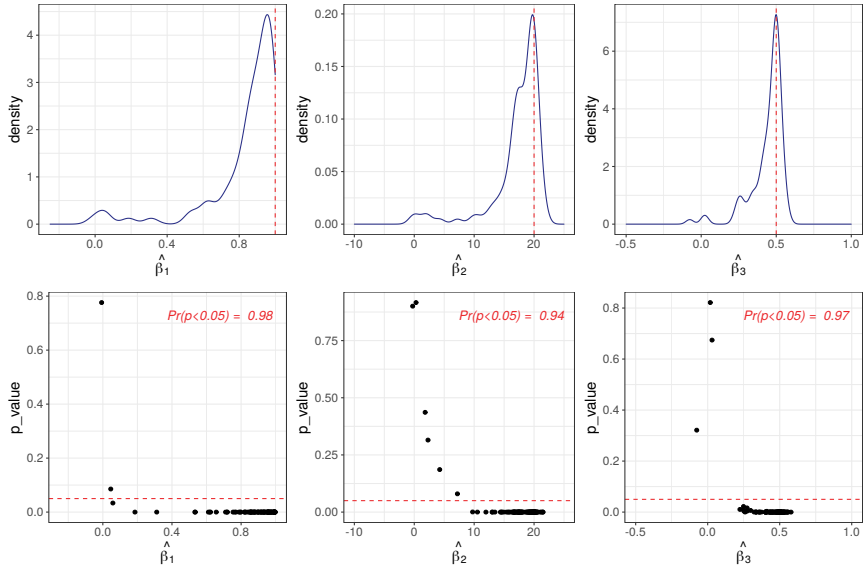


Figure 4.5: **Scenario N=900, 3 joint component: distribution of $\hat{\beta}$ (top) and their p-values (bottom).** The red vertical line in the left plot is the true value. The red horizontal line in the right plot is the threshold of p-value: 0.05.

7 males and 20 females. The methylation dataset consists of beta values (ratio of intensities between methylated and unmethylated alleles) at 450981 CpG sites after quality control following steps described in [2]. White blood cell counts are estimated from the methylation data and corrected for using R package 'Meffil' [21]. The glycomics dataset contains 10 glycan peaks, which are logTA normalized [31].

4.4.2. RESULTS

In the first stage, PO2PLS was applied to methylation (X) and glycomics (Y) with 3 joint and 1 specific component for each omics dataset. The null hypothesis of no relationship between each pair of methylation and glycomics joint components was tested. The p-values were 0.0006, 0.14, and 0.02 for the three joint components, respectively. With a threshold of 0.05 for the p-values, the first (T_1 for methylation and U_1 for glycomics) and third pair (T_3 and U_3) of joint components were significantly associated.

In Figure 4.6, we visualized the relationship between the first joint components and the three groups. It appeared that the DS group had a higher average of joint scores in both omics, comparing to the SB group. Note that the SB group is roughly of the same age as the DS group. The MA group, which is by design older, had a higher average score than the SB group. The average score in the DS group was higher than in the MA group, although the DS group was chronologically younger.

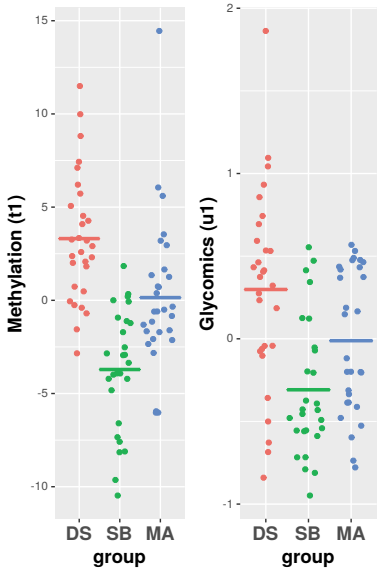


Figure 4.6: **Dot plot of the first joint components**. On the X axis are the three groups in different colors. On the Y axis are the scores of each individual. The mean score of each group is shown as a short horizontal line.

A linear mixed model (4.2) was then used to model the joint components in terms of age, DS status and their interaction. Only the components that had a significant relationship were considered, namely T_1, U_1, T_3 and U_3. We included age, DS status, their interaction, and sex as fixed effects, and further included a random intercept per family

in the model.

Table 4.2 shows the coefficients of interest and their p-values. In the first X-joint component, T_1, both age and DS had a significant positive effect, but their interaction term was not significant. Similarly, in U_1, age had a significant positive effect size, but the p-value of DS was slightly above 0.05. The DS effect size was also positive. The interaction term was not significant. In both T_3 and U_3, the effect of age was positive and significant.

Table 4.2: **Estimated effects of age, DS status, and their interaction**

Component	Age			DS			Age*DS		
	$\widehat{\beta_1}$	SE	p-value	$\widehat{\beta_2}$	SE	p-value	$\widehat{\beta_3}$	SE	p-value
T_1	0.088	0.027	0.002	7.240	2.142	0.001	-0.001	0.076	0.993
U_1	0.010	0.003	0.003	0.532	0.267	0.051	-0.001	0.009	0.908
T_3	0.051	0.030	0.090	2.135	2.364	0.370	0.003	0.084	0.974
U_3	0.012	0.002	<0.001	0.323	0.175	0.069	-0.005	0.006	0.459

Since the first pair of joint components were associated with DS, we further identify the relevant CpG sites and glycans corresponding to T_1 and U_1. In the first methylation joint component, the 1000 CpG sites with largest loading values were mapped to their respective target genes, yielding 496 genes. Next, GO enrichment analysis [1] was performed on this gene set using the GSEA software [22, 29]. The top significant GO terms were listed in Table 4.3. From these terms, the cellular component of neuron projection and synapse, biological process of neurogenesis, and neuron differentiation were shown to relate to DS [14, 17, 27, 26], while biological adhesion and cell-cell signaling are biological functions of plasma glycans [32, 18]. We further compared the set of 496 with genes obtained from a single point differential expression analysis in DS from a previous study [2]. The previously identified genes categorized into four main functions were also selected in our geneset: haematopoiesis (RUNX1, DLL1, EBF4), morphogenesis and development (HOXA2, HOXA4, HOXA5, HHIP, NCAM1), neuronal development (NAV1, EBF4, PRDM8, NCAM1, GABBR1), and regulation of chromatin structure (KDM2B, TET1).

For the first glycomics joint component, the loading values of each glycan is shown in Table 4.4. The glycans H3N4F1 and H3N5F1 had the largest absolute loading values. According to a previous study [5] on plasma glycans and DS, the first glycan was found to be the top discriminators of DS subjects and siblings. The second glycan was one of the main age discriminants.

4.5. DISCUSSION

We proposed a two-stage data integration approach to model the methylation and glycomics jointly and investigate the impact of Down syndrome on methylation and glycomics. In the first stage, we applied the PO2PLS model to integrate the two omics data. PO2PLS estimates low-dimensional joint and omic-specific latent components and test the relationship between the two datasets. The significantly associated joint components are then taken as surrogates for the joint information in methylation and glycomics. In the second stage, these joint components were modeled as a function of DS

Table 4.3: **Results of GO enrichment analysis**.

Gene Set Name	p-value	FDR q-value
GOCC_NEURON_PROJECTION	7.48E-22	7.61E-18
GOBP_BIOLOGICAL_ADHESION	1.47E-20	7.47E-17
GOBP_NEUROGENESIS	3.08E-19	1.05E-15
GOBP_NEURON_DIFFERENTIATION	2.22E-17	5.66E-14
GOMF_ADENYL_NUCLEOTIDE_BINDING	3.85E-17	7.84E-14
GOBP_CELL_CELL_SIGNALING	1.11E-16	1.88E-13
GOCC_SYNAPSE	5.40E-16	7.85E-13

The "p-value" column shows the p-value of each annotation derived by random sampling of the whole genome; the "FDR q-value" column provides the false discovery rate (FDR) analog of the p-value after correcting for multiple hypothesis testing [4, 28]. Complete list can be found in Additional file

Table 4.4: **Loading values of the glycans in the first joint component**.

Peak	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Structure	H3N4F1	H3N5F1	H4N4F1		H5N4	H5N4F1	H5N5F1	H6N5	H6N5F1	H7N6
Value	**-0.646**	**-0.473**	-0.316	-0.209	0.143	0.057	0.065	0.316	-0.086	0.284

status and aging using a linear mixed model, with sex as an additional covariate. Familial relationships are taken into account by specifying a random intercept per family.

We conducted a simulation study to evaluate the performance of the two-stage approach to detect effects of aging, DS status, and their interaction on the joint methylation-glycomics parts. In the large sample size scenarios, with 900 subjects, the proportions of significant coefficients were high for all the three effects, and the TPR of identified features was close to 1. In the small sample size scenarios, the effect of DS and its interaction with age were more difficult to detect. The proportions of significant coefficients for these two effects were both low. A reason for the reduction of the performance is the large amount of heterogeneity in the relation between the joint components of methylation and glycomics. The variance of h, representing this heterogeneity, was set to be 94.3% of the total variance of u (see Table 4.1). This number was estimated from the dataset. The underlying reason for a large heterogeneity between omics data might be that the biological link between methylation and glycomics involves many intermediate biological layers, e.g., transcription and protein expression [36, 34]. Combined with a small sample size, there is limited information to accurately estimate the true underlying joint components. The high noise level in the joint parts also implies a weak correlation between the joint components. Consequently, some joint variation tends to be estimated as omic-specific, especially if the number of joint components K is small. It might explain why the model with fewer joint components ($K = 1$) underperformed in detecting the effect of DS and its interaction with age on the joint components.

We applied the two-stage integration approach to the DS dataset. From the dot plot (Figure 4.6) of the first joint components, we observed difference in scores between groups. Since the effect of age was significant, these components might be seen as a

representation of biological age, similar to the biological aging models in [15, 20]. A significant and positive main effect of DS means that the biological age is higher in the DS group, suggesting existence of precocious aging, which could be considered as an intrinsic characteristic of DS [13]. The interaction effect of DS and age can be interpreted as accelerated aging, implying a faster rate of aging in the DS subjects compared to controls. In our analysis, the interaction was not significant, hence no evidence for accelerated aging in DS. However, as implied in the simulation results, with the small sample size, we had limited information to detect the interaction.

Our proposed two-stage approach involves, in the first step, a model for the two omics datasets that is independent of the Down syndrome status. In the second step, the estimated components are used to infer the effect of DS on the omics data. This leads to suboptimal efficiency when estimating the relation between Down syndrome and the omics datasets. Also, the additional estimation uncertainty from the first stage is ignored in the second stage. To model the DS status jointly with the two omics data and infer its role, a holistic framework is desired. The joint latent variables will not only represent the variation of the omics data, but also include information about the subject groups. It is suited for studies aiming at identifying mechanisms underlying the omics data that are particularly related to the subject groups. For complex study designs, such as the family-based case control study, it is also necessary to consider the family structure in the data integration approach. Including these functionalities in our methods will be the directions of our future work.

BIBLIOGRAPHY

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

[2] Bacalini, M. G., Gentilini, D., Boattini, A., Giampieri, E., Pirazzini, C., Giuliani, C., Fontanesi, E., Scurti, M., Remondini, D., Capri, M., Cocchi, G., Ghezzo, A., Rio, A. D., Luiselli, D., Vitale, G., Mari, D., Castellani, G., Fraga, M., Di Blasio, A. M., Salvioli, S., Franceschi, C., and Garagnani, P. (2015). Identification of a DNA methylation signature in blood cells from persons with down syndrome. *Aging*, 7(2):82–96.

[3] Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

[4] Benjamini, Y. and Hochberg, Y. (1995). 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.

[5] Borelli, V., Vanhooren, V., Lonardi, E., Reiding, K. R., Capri, M., Libert, C., Garagnani, P., Salvioli, S., Franceschi, C., and Wuhrer, M. (2015). Plasma N-Glycome Signature of Down Syndrome. *Journal of Proteome Research*, 14(10):4232–4245.

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

[7] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[8] Dyer, C. A. and Sinclair, A. J. (1998). The premature ageing syndromes: Insights into the ageing process.

[9] el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., and Uh, H. W. (2016). Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, 17(2):S11.

[10] Said el Bouhaddani, Hae-Won Uh, Geurt Jongbloed, and Jeanine Houwing-Duistermaat. Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, aug 2022.

[11] Franceschi, C., Garagnani, P., Gensous, N., Bacalini, M. G., Conte, M., and Salvioli, S. (2019). Accelerated bio-cognitive aging in Down syndrome: State of the art and possible deceleration strategies.

[12] Gensous, N., Bacalini, M. G., Franceschi, C., and Garagnani, P. (2020). Down syndrome, accelerated aging and immunosenescence.

[13] Gensous, N., Franceschi, C., Salvioli, S., Garagnani, P., and Bacalini, M. G. (2019). *Down syndrome, ageing and epigenetics*, volume 91.

[14] Haas, M. A., Bell, D., Slender, A., Lana-Elola, E., Watson-Scales, S., Fisher, E. M., Tybulewicz, V. L., and Guillemot, F. (2013). 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(10):78561.

[15] Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):115.

[16] Horvath, S., Garagnani, P., Bacalini, M. G., Pirazzini, C., Salvioli, S., Gentilini, D., Di Blasio, A. M., Giuliani, C., Tung, S., Vinters, H. V., and Franceschi, C. (2015). Accelerated epigenetic aging in Down syndrome. *Aging Cell*, 14(3):491–495.

[17] Huo, H. Q., Qu, Z. Y., Yuan, F., Ma, L., Yao, L., Xu, M., Hu, Y., Ji, J., Bhattacharyya, A., Zhang, S. C., and Liu, Y. (2018). Modeling Down Syndrome with Patient iPSCs Reveals Cellular and Migration Deficits of GABAergic Neurons. *Stem Cell Reports*, 10(4):1251–1266.

[18] Krautter, F. and Iqbal, A. J. (2021). Glycans and Glycan-Binding Proteins as Regulators and Potential Targets in Leukocyte Recruitment.

[19] Krištić, J., Vučković, F., Menni, C., Klarić, L., Keser, T., Beceheli, I., Pučić-Baković, M., Novokmet, M., Mangino, M., Thaqi, K., Rudan, P., Novokmet, N., Šarac, J., Missoni, S., Kolčić, I., Polašek, O., Rudan, I., Campbell, H., Hayward, C., Aulchenko, Y., Valdes, A., Wilson, J. F., Gornik, O., Primorac, D., Zoldoš, V., Spector, T., and Lauc, G. (2014a). Glycans are a novel biomarker of chronological and biological ages. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 69(7):779–789.

[20] Krištić, J., Vučković, F., Menni, C., Klarić, L., Keser, T., Beceheli, I., Pučić-Baković, M., Novokmet, M., Mangino, M., Thaqi, K., Rudan, P., Novokmet, N., Šarac, J., Missoni, S., Kolčić, I., Polašek, O., Rudan, I., Campbell, H., Hayward, C., Aulchenko, Y., Valdes, A., Wilson, J. F., Gornik, O., Primorac, D., Zoldoš, V., Spector, T., and Lauc, G. (2014b). Glycans are a novel biomarker of chronological and biological ages. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 69(7):779–789.

[21] Min, J. L., Hemani, G., Smith, G. D., Relton, C., and Suderman, M. (2018). Meffil: Efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, 34(23):3983–3989.

[22] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273.

[23] Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., Anderson, P., Mason, C. A., Collins, J. S., Kirby, R. S., and Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004-2006. *Birth Defects Research Part A - Clinical and Molecular Teratology*, 88(12):1008–1016.

[24] Patterson, D. and Cabelof, D. C. (2012). Down syndrome as a model of DNA polymerase beta haploinsufficiency and accelerated aging. *Mechanisms of Ageing and Development*, 133(4):133–137.

[25] Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58(3):545.

[26] Sobol, M., Klar, J., Laan, L., Shahsavani, M., Schuster, J., Annerén, G., Konzer, A., Mi, J., Bergquist, J., Nordlund, J., Hoeber, J., Huss, M., Falk, A., and Dahl, N. (2019). Transcriptome and Proteome Profiling of Neural Induced Pluripotent Stem Cells from Individuals with Down Syndrome Disclose Dynamic Dysregulations of Key Pathways and Cellular Functions. *Molecular Neurobiology*, 56(10):7113–7127.

[27] Stagni, F., Giacomini, A., Emili, M., Guidi, S., and Bartesaghi, R. (2018). Neurogenesis impairment: An early developmental defect in Down syndrome.

[28] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(3):479–498.

[29] 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., and Mesirov, J. P. (2005). 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.

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

[31] Uh, H. W., Klaric, L., Ugrina, I., Lauc, G., Smilde, A. K., and Houwing-Duistermaat, J. J. (2020). Choosing proper normalization is essential for discovery of sparse glycan biomarkers. *Molecular Omics*, 16(3):231–242.

[32] Varki, A. (2017). Biological roles of glycans. *Glycobiology*, 27(1):3–49.

[33] Vilardell, M., Rasche, A., Thormann, A., Maschke-Dutz, E., Pérez-Jurado, L. A., Lehrach, H., and Herwig, R. (2011). Meta-analysis of heterogeneous Down Syndrome data reveals consistent genome-wide dosage effects related to neurological processes. *BMC Genomics*, 12(1):1–16.

[34] 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., and Gieger, C. (2018). IgG glycosylation and DNA methylation are interconnected with smoking. *Biochimica et Biophysica Acta - General Subjects*, 1862(3):637–648.

[35] Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J. (1984). 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.

[36] Zierer, J., Menni, C., Kastenmüller, G., and Spector, T. D. (2015). Integration of 'omics' data in aging research: From biomarkers to systems biology. *Aging Cell*, 14(6):933–944.

[37] Zigman, W. B. (2013). Atypical aging in down syndrome. *Developmental Disabilities Research Reviews*, 18(1):51–67.

5

JOINT MODELING OF AN OUTCOME VARIABLE AND INTEGRATED OMIC DATASETS USING GLM-PO2PLS

Zhujie Gu, Said el Bouhaddani, Jeanine Houwing-Duistermaat, Hae Won Uh. This work has been submitted and reviewed.

ABSTRACT

In many studies of human diseases, multiple omic datasets are measured. Typically, these omic datasets are studied one by one with the disease, thus the relationship between omics are 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. In this article, we extend dimension reduction methods which model the joint part of omics to a novel 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 model is illustrated by a Down syndrome study where Down syndrome and two omics – methylation and glycomics – are jointly modeled.

5.1. 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 GWAS conducted to identify the genetic risk variants associated with more than 3000 traits and human diseases [55, 51]. Other examples include studies of methylation data to pinpoint differentially methylated regions of DNA as indicators of many diseases [41, 34], and studies of glycomic data to gain insight into the role of post-translational modification of protein in disease pathways [48, 47]. Though these studies on a single omic dataset provided biological insights of diseases from various aspects, they ignored the correlations among the omic levels. Analyzing multiple linked omic datasets jointly can bring further insights into the biological system underlying diseases. In this paper, we propose a new model for two omic datasets and an outcome variable, where the relationship of the omic datasets with the outcome is modeled via the joint parts of the omic 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 [35], 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 [15, 16, 4, 8] and glycomics [15, 6, 9]. 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 [54]. We will use our new model to jointly analyze DS and both omics, aiming to investigate whether the molecules involved in the relationship between methylation and glycomics are related to DS.

Generalized linear models (GLM) are flexible models which link linear predictors with an outcome variable via functions such as identity and logit [33]. However, omic datasets are often high-dimensional ($p >> N$) and the features are highly correlated, which leads to (near) collinearity, hence GLM cannot directly be employed. A solution is to use its penalized variants such as ridge regression [20] and elastic net [60], which handle the dimensionality and correlation better. To incorporate information from more than one dataset, stacked sets of omic features can be modeled. However, such approaches do not model the relationship between omics, and hence cannot provide insight on the joint part. Furthermore, when the omic datasets are heterogeneous, regressing on a stacked dataset can lead to inferior performance than using only one of the available omic sources [39, 49].

To model the joint part of two omic datasets, several dimension reduction methods have been developed, which map both datasets from the original high-dimensional spaces to low-dimensional spaces including a joint space which retains the relationship. Among these, PLS [56, 10] simultaneously decomposes two datasets x and y into joint and residual subspaces. The joint (low-dimensional) subspace of one dataset represents the best approximation of x or y. The joint subspaces of PLS may contain omic-specific variation due to the presence of heterogeneity between the omics. To sepa-

rate this omic-specific variation from the joint subspaces, two-way orthogonal partial least squares (O2PLS) [50, 12] was proposed which decomposes each dataset into joint, data-specific, and residual subspaces. The data-specific subspaces in x and y capture variation unrelated to each other, and improve the estimates of the joint subspaces for the true relationship between x and y. A drawback of PLS and O2PLS is that they are algorithmic, and hence do not model the whole distribution. To enable statistical inference, likelihood-based probabilistic approaches such as supervised integrated factor analysis (SIFA) [26] and probabilistic PLS (PPLS) [13] were proposed. Recently, a general framework probabilistic O2PLS (PO2PLS) [14] was proposed which contains SIFA and PPLS as specific cases. In PO2PLS, the relationship between two omics is inferred by using a Wald-type test-statistic to test the hypothesis that the joint latent variables of the two omics are related. However, PO2PLS is unsupervised, i.e., the outcome variable is not used in the joint dimension reduction. In this paper, we will propose a new model GLM-PO2PLS which extends PO2PLS by including an outcome variable in the model next to the omic datasets. The relationship between two omic 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. 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 omic 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 5.2, 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 5.3, 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 5.4, we apply GLM-PO2PLS to the motivating DS datasets. We conclude with a discussion.

5.2. METHODS

The GLM-PO2PLS was developed based on PO2PLS model which has been described in detail elsewhere [14]. 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.

5.2.1. THE GLM-PO2PLS MODEL

GLM-PO2PLS jointly models an outcome variable z with two omic 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

$$\begin{gathered} x = tW^\top + t_\perp W_\perp^\top + e, \quad y = uC^\top + u_\perp C_\perp^\top + f, \quad u = tB + h, \\ \eta(\mathbb{E}[z]) = \beta_0 + ta^\top + ub^\top, \end{gathered} \tag{5.1}$$

with β_0 the intercept, a and b both row-vectors of size K and η the link function which links the outcome z to the linear predictor $\beta_0 + ta^\top + ub^\top$. Note that the equations in the first row of (5.1) are identical to 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 (5.1) can equivalently be written in terms of t and the h (the part in u independent of t), i.e.

$$\eta(\mathbb{E}[z]) = \beta_0 + ta^\top + (tB + h)b^\top = \beta_0 + t\tilde{a}^\top + h\tilde{b}^\top,$$

where $\tilde{a} = a + Bb^\top$ and $\tilde{b} = b$. With this equivalent parametrization, instability due to near colinearity in the linear predictor of z is reduced.

In the remainder of the paper, we use the rightmost form in (1.8) and omit the tildes on a and b.

5.2.2. 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, $\eta(v) = v$. We assume that the outcome is centered and since t and h have zero-mean, the intercept β_0 can be omitted. We assume that the residual g ($g = z - ta^\top - hb^\top$) 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

$$\Sigma_\theta = \begin{bmatrix} W\Sigma_t W^\top + W_\perp \Sigma_{t_\perp} W_\perp^\top + \sigma_e^2 I_p & W\Sigma_t BC^\top & W\Sigma_t a^\top \\ CB\Sigma_t W^\top & C\Sigma_u C^\top + C_\perp \Sigma_{u_\perp} C_\perp^\top + \sigma_f^2 I_q & C(\Sigma_h b^\top + B\Sigma_t a^\top) \\ a\Sigma_t W^\top & (a\Sigma_t B + b\Sigma_h)C^\top & a\Sigma_t a^\top + b\Sigma_h b^\top + \sigma_g^2 \end{bmatrix}, \tag{5.2}$$

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 [14]. 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, $[WW_\perp]$ 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 1. *Let (x, y, z) follow 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 [14] 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 (5.2). 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. □

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 the (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,

$$\begin{aligned} Q(\theta|\theta') =& \mathbb{E}[\log f(x,y,z,t,u,t_\perp,u_\perp)] = \mathbb{E}[\log f(x,y,z|t,u,t_\perp,u_\perp)] + \mathbb{E}[\log f(t,u,t_\perp,u_\perp)] \\ =& \underbrace{\mathbb{E}[\log f(z|t,u)]}_{Q_{\{a,b,\sigma_g^2\}}} + \underbrace{\mathbb{E}[\log f(x|t,t_\perp)]}_{Q_{\{W,W_\perp,\sigma_e^2\}}} + \underbrace{\mathbb{E}[\log f(y|u,u_\perp)]}_{Q_{\{C,C_\perp,\sigma_f^2\}}} + \underbrace{\mathbb{E}[\log f(u|t)]}_{Q_{\{B,\Sigma_h\}}} \\ &+ \underbrace{\mathbb{E}[\log f(t)]}_{Q_{\Sigma_t}} + \underbrace{\mathbb{E}[\log f(t_\perp)]}_{Q_{\Sigma_{t_\perp}}} + \underbrace{\mathbb{E}[\log f(u_\perp)]}_{Q_{\Sigma_{u_\perp}}}. \end{aligned} \tag{5.3}$$

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 [14] for the expectation and maximization regarding these terms.

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

$$Q_{\{a,b,\sigma_g^2\}} = -\frac{1}{2}\left\{\log(2\pi\sigma_g^2) + \frac{1}{\sigma_g^2}\operatorname{tr}\mathbb{E}\left[(z - ta^\top - (u - tB)b^\top)^\top(z - ta^\top - (u - tB)b^\top)\right]\right\}. \tag{5.4}$$

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 [14]. The conditional moments involved in (5.4) 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 (5.3) 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 (5.4). 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.

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 the joint parts t and u,

and given by the equation $\eta(\mathbb{E}[z]) = ta^\top + hb^\top$ in (5.1). 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,

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

For each component-wise test, we consider the null hypothesis of no relationship between z and the k-th pair of joint components,

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

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

$$T_{full} = \hat{\alpha}\Pi_{\hat{\alpha}}^{-1}\hat{\alpha}^\top, \tag{5.5}$$

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

$$T_{comp.wise} = \hat{\alpha_k}\Pi_{\hat{\alpha_k}}^{-1}\hat{\alpha_k}^\top. \tag{5.6}$$

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 (Shapiro 1986) applied to the GLM-PO2PLS model.

Theorem 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 $2K$ 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 [28],

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

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 of the 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 $\hat{\alpha}$ and $\hat{\theta}/\hat{\alpha}$ are asymptotically independent 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.

5.2.3. THE GLM-PO2PLS MODEL WITH A BINARY OUTCOME

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

$$x = tW^\top + t_\perp W_\perp^\top + e, \quad y = uC^\top + u_\perp C_\perp^\top + f, \quad u = tB + h,$$
$$\text{logit}(p(z)) = \beta_0 + ta^\top + hb^\top.$$

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 $\mathcal{N}(\beta_0, a\Sigma_t a^\top + b\Sigma_h b^\top)$. The joint distribution $(x, y, \text{logit}(p(z)))$ is multivariate normal with mean vector $(0_{p+q}, \beta_0)$ and covariance matrix Σ_θ in (5.2) excluding σ_g^2. The collection of parameters in the GLM-PO2PLS model with a binary outcome is $\theta = \{W, C, W_\perp, C_\perp, \beta_0, a, b, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \sigma_e^2, \sigma_f^2, \Sigma_h\}$.

Identifiability of GLM-PO2PLS with a binary outcome Theorem 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 $(0_{p+q}, \beta_0)$ 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 $(0_{p+q}, \beta_0) = (0_{p+q}, \tilde{\beta}_0)$ and $\Sigma_{\theta/g^2} = \Sigma_{\tilde{\theta}/g^2}$. It is clear that $\beta_0 = \tilde{\beta}_0$ from the equivalence of the mean vectors. The identifiablity of a and b can be shown from the equivalence of covariance matrices analogously as in the proof of Theorem 1. This shows the identifiability of all the parameters in GLM-PO2PLS with a binary outcome.

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 $v = (t, u)$ and $\xi = (t_\perp, u_\perp)$,

$$\ell(\theta; x, y, z) = \log \int_{(v,\xi)} f(x,y,z|v,\xi,\theta) f(v,\xi|\theta) d(v,\xi). \tag{5.7}$$

To estimate (5.7), numerical integration is needed. Note that given v, the binary outcome z is independent of x, y and ξ, thus the conditional density $f(x,y,z|v,\xi)$ in (5.7) can be factorized as $f(x,y,z|v,\xi) = p(z|v) f(x,y|v,\xi)$. The factorization enables to integrate out the specific random vector ξ, hence reducing the dimension of the integral to

$2K$,

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

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

$$p(z|v) = \begin{cases}\left(1+\exp\{-(\beta_0 + ta^\top + (u-tB)b^\top)\}\right)^{-1} & z=1,\\ \left(1+\exp\{\beta_0 + ta^\top + (u-tB)b^\top\}\right)^{-1} & z=0.\end{cases} \tag{5.8}$$

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

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

where the covariance matrices involved are:

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

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

$$Q(\theta|\theta') = \mathbb{E}[\log f(x,y,z,v|\theta)|x,y,z,\theta']. \tag{5.9}$$

Expectation step based on numerical integration Analogously to (5.3), the conditional expectation in (5.9) can be decomposed to factors that depend on distinct sets of parameters,

$$\begin{aligned}Q(\theta|\theta') &= \mathbb{E}[\log f(x,y,z,v)] = \mathbb{E}[\log f(x,y,z|v)] + \mathbb{E}[\log f(v)]\\ &= \underbrace{\mathbb{E}[\log p(z|v)]}_{Q_{\{\beta_0,a,b\}}} + \underbrace{\mathbb{E}[\log f(x|t)]}_{Q_{\{W,W_\perp,\sigma_e^2,\Sigma_{t_\perp}\}}} + \underbrace{\mathbb{E}[\log f(y|u)]}_{Q_{\{C,C_\perp,\sigma_f^2,\Sigma_{u_\perp}\}}} + \underbrace{\mathbb{E}[\log f(u|t)]}_{Q_{\{B,\Sigma_h\}}} + \underbrace{\mathbb{E}[\log f(t)]}_{Q_{\Sigma_t}}.\end{aligned} \tag{5.10}$$

Here, the first conditional expectation $Q_{\{\beta_0,a,b\}}$ has no closed form,

$$Q_{\{\beta_0,a,b\}} = \int [\log p(z|v)]f(v|x,y,z,\theta')dv = \frac{1}{f(x,y,z)}\int [\log p(z|v)]p(z|v)f(x,y|v)f(v)dv.$$

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

$$\int \varphi(v)p(z|v)f(x,y|v)f(v)dv \approx \sum_{m_1=1}^{M}\ldots\sum_{m_{2K}=1}^{M}\varphi(v=v_m)p(z|v=v_m)f(x,y|v=v_m)w_{m_1}\ldots w_{m_{2K}} \tag{5.11}$$

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 (5.11) more suitable based on the distribution of ν [27].

The other terms in (5.10) 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

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

which can be numerically calculated with (5.11).

Maximization step based on gradient descent Maximizing $Q_{\{\beta_0,a,b\}}$ requires iterations as its derivative with respect to $\beta = (\beta_0, a, b)$ 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

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

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 (5.10) can be found in the supplementary materials.

5.3. 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 (5.1), with the equations for the continuous and binary outcomes being $z_c = ta^\top + hb^\top + g$, and $z_b \sim$ Bernoulli$((1 + \exp\{-(\beta_0 + ta^\top + hb^\top)\})^{-1})$.

5.3.1. 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 (5.1). 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 simplicity of the model and computational efficiency. For each setting, 500 replications are generated. The settings are summarized in Table 5.1.

Table 5.1: **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%

The metrics used to assess the performance are listed in Table 5.2. 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. 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 mea-

sured by true positives rate (TPR) calculated as the proportion of true top 25% features among the 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).

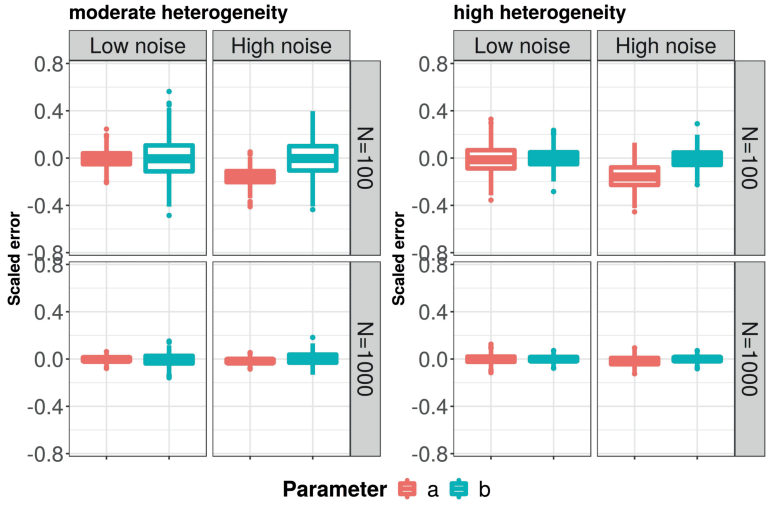
Table 5.2: **Metric of simulation**

Category	Metric	Calculation	Competing methods
Coefficient estimation	Scaled error	$(\hat{a}-a)/a$, $(\hat{b}-b)/b$	
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

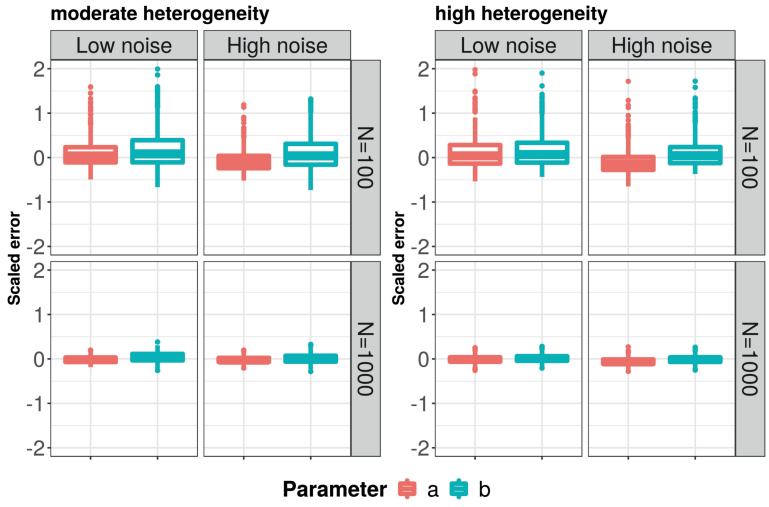
5.3.2. RESULTS OF SIMULATION STUDY

In Figure 5.1, results of the coefficient estimation in high-dimensional settings are depicted. Figure 5.1a 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 5.1b. 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 5.2 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 5.2a. 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



(a) **Continuous outcome z_c**

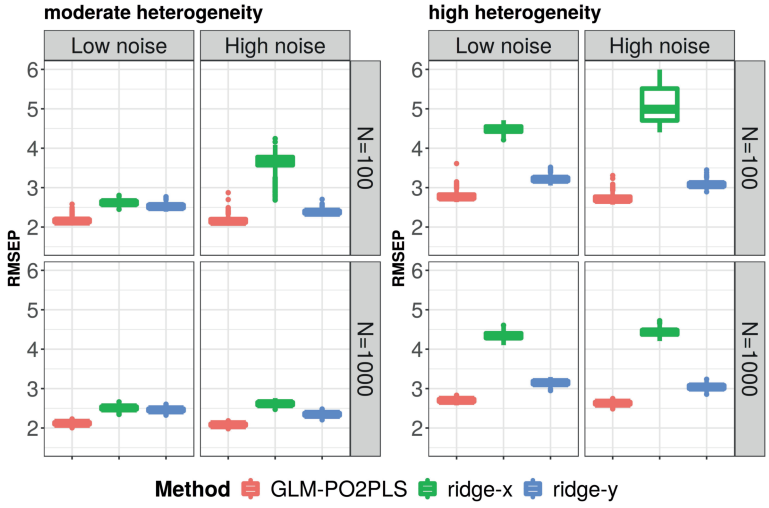


(b) **Binary outcome z_b**

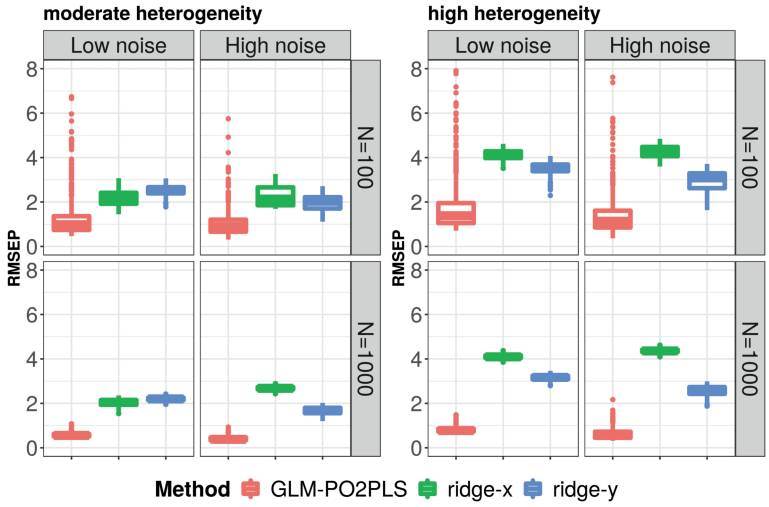
Figure 5.1: **Performance of coefficient estimation for continuous (a) and binary (b) outcome.** The y-axis shows the scaled estimation error as defined in Table 5.2. In the moderate and high heterogeneity settings, h account for 40% and 80% of total variance in $u = tB + h$, respectively. Boxes show the results of 500 repetitions.

of GLM-PO2PLS was less affected. For the binary outcome z_b , GLM-PO2PLS still outperformed ridge regression as shown in in Figure 5.2b. 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.

Lastly, we briefly present the key results for loading estimation and feature selection. Overall, the loading estimates were accurate for both continuous and binary outcomes,



(a) **Continuous outcome z_c**



(b) **Binary outcome z_b**

Figure 5.2: **Performance of outcome prediction for continuous (a) and binary (b) outcome.** y-axis shows the RMSEP as defined in Table 5.2. Boxes show the results of 500 repetitions.

with most inner products between the estimated and the true loadings approaching 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 5.1a. 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. The details are given in the supplementary material.

5.4. 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 [22], and both methylation and glycomics are associated with biological age [21, 25], we expect the DS patients to be more similar to their mothers than siblings.

5.4.1. 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.

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' [31]. Age and sex were corrected for using multiple regression. 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 [52] and corrected for age and sex.

We will fit a GLM-PO2PLS continuous model and a GLM-PO2PLS binary model to these data. 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 [54] that suggested the presence of an indirect influence of methylation on glycosylation.

5.4.2. RESULTS OF DS DATA ANALYSIS

For the GLM-PO2PLS continuous model, we used 3 joint and 1 specific component for each omic 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 (5.1). 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 5.3 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 5.3 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}.

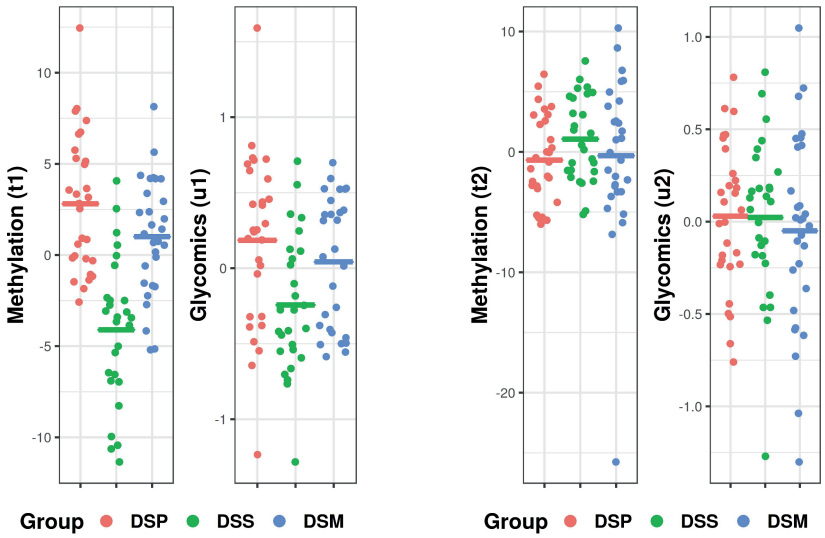


Figure 5.3: **Joint scores of the first (left) and second (right) pair of joint components.** On the y-axis are the scores of each individual colored by different groups. The mean score of each group is shown as a horizontal line.

Table 5.3: **Results of testing for no relationship between DS and joint components**

	T_{full} in (5.5)	$T_{comp.wise}$ in (5.6)		
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

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 [32, 46]. The top three significant GO terms were listed in Table 5.4. Among these terms, the cell-cell signaling is a biological function of plasma glycans [53, 24]. The cellular component of neuron projection and biological process of neurogenesis were shown to relate to DS [18, 23, 43, 42]. We further searched the mapped geneset in the DisGeNET database [36] 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 [36]. 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 GDA score regarding Down syndrome (i.e., with the

most evidence of association with DS) in the DisGeNET database, and found the gene RCAN1 which relates to epigenetics was among our top genes mapped from methylation. It has been revealed that RCAN1 plays a critical upstream role in epigenetic regulation of adult neurogenesis [7], hence important in the pathogenesis of Down syndrome [58].

For the first glycomics joint component, the loading values of each glycan can be found in the supplementary. 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 omic dataset. We chose for 1 joint component based on the test results in the continuous model shown in Table 5.3. 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 5.4. The top two terms were related to membrane organelle, more specifically, Golgi apparatus, which is required for accurate glycosylation [59]. 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.

Table 5.4: **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, 45]. Complete list can be found in supplementary material.

Although for the continuous and binary models the sets of top CpG sites appeared to be relevant to glycosylation and DS, there was little overlap between the top CpG sites. This might be explained by the different number of joint components specified for the two models. Therefore, we performed an additional analysis using a filtered dataset, which was obtained by subtracting the second and third joint components, and the specific components of the GLM-PO2PLS continuous model from the omic data. The estimated parameters of the fitted GLM-PO2PLS binary model were very similar to the ones in the continuous model. More specifically, the inner product of the joint loading vectors from the two models reached 0.99. The correlation between the corresponding joint components of the two models was also high (at 0.99). Note that an inner product of 1 means the loading vectors are the same. Regarding the interpretation of top CpG

sites and glycans, we refer to the results in the continuous model, as the loadings were very similar, so as the top features identified.

5.5. 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 omic 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 in various scenarios, and 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 omic datasets, on top of their relationship with the outcome.

The methylation dataset analyzed with GLM-PO2PLS was also analyzed by Bacalini et al [4] using single point approaches. They identified four categories. 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). Therefore, the association between DS and HOXA6 appeared to be weak in the data.

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 $\mathcal{O}(M^{2K})$, 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 [37, 38]. Another strategy is to decompose the $2K$-dimensional integration to K 2-dimensional integrations. This will reduce the computational complexity to $\mathcal{O}(K \times M^2)$.

To calculate the p-values for the tests in (5.5) and (5.6), 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 [40]. 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. While providing a holistic overview, for the binary outcome variable, the approach is computationally intensive. In our data analysis, it appeared that modeling the binary outcome as continuous provided similar information. Alternatively, a two-stage approach might be used. We recently proposed a two-stage PO2PLS approach [17], 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 out-

come. 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 [44]. 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 $p(z|v)$ in (5.8) 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. Another direction is to generalize the model to incorporate more than two omic 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 [29]. 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 omic datasets and as a base for further extensions.

5.6. SUPPLEMENTARY MATERIALS FOR CHAPTER 5

This section is structured into two parts: methods, and simulation results. Each part contains additional materials for the respective section in the main article.

In section 5.6.1 and 5.6.2, we give mathematical details for the EM algorithms for GLM-PO2PLS continuous and binary model, respectively. We then prove the asymptotic normality of the estimator, and give equations for the observed Fisher information matrix needed in calculating the test statistics in section 5.6.3. In section 5.6.4, we show simulation results omitted in the main article.

5.6.1. AN EM ALGORITHM FOR GLM-PO2PLS WITH A NORMALLY DISTRIBUTED OUTCOME

Let X, Y and Z be data matrices consisting of N observations of (x, y, z). For empirical identifiability of the components, we assume $\max(K + K_x, K + K_y) < N$.

In the E step, the expectation of the complete data log likelihood for one subject can

be decomposed to factors that depend on distinct sets of parameters as follows,

$$\begin{aligned}Q(\theta|\theta') =& \mathbb{E}[\log f(x,y,z,t,u,t_\perp,u_\perp)] = \mathbb{E}[\log f(x,y,z|t,u,t_\perp,u_\perp)] + \mathbb{E}[\log f(t,u,t_\perp,u_\perp)] \\ =& \underbrace{\mathbb{E}[\log f(z|t,u)]}_{Q_{\{a,b,\sigma_g^2\}}} + \underbrace{\mathbb{E}[\log f(x|t,t_\perp)]}_{Q_{\{W,W_\perp,\sigma_e^2\}}} + \underbrace{\mathbb{E}[\log f(y|u,u_\perp)]}_{Q_{\{C,C_\perp,\sigma_f^2\}}} + \underbrace{\mathbb{E}[\log f(u|t)]}_{Q_{\{B,\Sigma_h\}}} \\ &+ \underbrace{\mathbb{E}[\log f(t)]}_{Q_{\Sigma_t}} + \underbrace{\mathbb{E}[\log f(t_\perp)]}_{Q_{\Sigma_{t_\perp}}} + \underbrace{\mathbb{E}[\log f(t_\perp)]}_{Q_{\Sigma_{u_\perp}}}.\end{aligned} \tag{5.12}$$

In this equation, the conditioning on x, y, z and θ' is dropped, to simplify notation. Given the observed data for N subjects, the factorized conditional expectations in (5.12) are calculated as follows. Let $(T,U,T_\perp,U_\perp)$ be the collection of row vectors $(t,u,t_\perp,u_\perp)$ for N subjects.

$$\begin{aligned}Q_{\{a,b,\sigma_g^2\}} &= -\frac{1}{2}\Big\{N\log(2\pi\sigma_g^2) + \frac{1}{\sigma_g^2}\operatorname{tr}\mathbb{E}\left[(Z-Ta^\top-(U-TB)b^\top)^\top(Z-Ta^\top-(U-TB)b^\top)\right]\Big\}, \\ Q_{\{W,W_\perp,\sigma_e^2\}} &= -\frac{1}{2}\Big\{Np\log(2\pi\sigma_e^2) + \frac{1}{\sigma_e^2}\operatorname{tr}\mathbb{E}\left[(X-TW^\top-T_\perp W_\perp^\top)^\top(X-TW^\top-T_\perp W_\perp^\top)\right]\Big\}, \\ Q_{\{C,C_\perp,\sigma_f^2\}} &= -\frac{1}{2}\Big\{Nq\log(2\pi\sigma_f^2) + \frac{1}{\sigma_f^2}\operatorname{tr}\mathbb{E}\left[(Y-UC^\top-U_\perp C_\perp^\top)^\top(Y-UC^\top-U_\perp C_\perp^\top)\right]\Big\}, \\ Q_{\{B,\Sigma_h\}} &= -\frac{1}{2}\Big\{NK\log(2\pi) + N\log|\Sigma_h| + \operatorname{tr}\mathbb{E}\left[(U-TB)^\top(U-TB)\Sigma_h^{-1}\right]\Big\}, \\ Q_{\Sigma_t} &= -\frac{1}{2}\Big\{NK\log(2\pi) + N\log|\Sigma_t| + \operatorname{tr}\mathbb{E}\left[T^\top T\Sigma_t^{-1}\right]\Big\}, \\ Q_{\Sigma_{t_\perp}} &= -\frac{1}{2}\Big\{NK_x\log(2\pi) + N\log|\Sigma_{t_\perp}| + \operatorname{tr}\mathbb{E}\left[T_\perp^\top T_\perp\Sigma_{t_\perp}^{-1}\right]\Big\}, \\ Q_{\Sigma_{u_\perp}} &= -\frac{1}{2}\Big\{NK_y\log(2\pi) + N\log|\Sigma_{u_\perp}| + \operatorname{tr}\mathbb{E}\left[U_\perp^\top U_\perp\Sigma_{u_\perp}^{-1}\right]\Big\}.\end{aligned} \tag{5.13}$$

Here, the conditional expectations involve the first and second conditional moments of the vector $(t,u,t_\perp,u_\perp)$ given x,y,z and θ' for each subject. Since the complete data vector for one subject $(x,y,z,t,u,t_\perp,u_\perp)$ follows a multivariate normal distribution with zero mean and known covariance matrix, the conditional distribution $(t,u,t_\perp,u_\perp|x,y,z)$ for each subject can be calculated explicitly following lemma 3 in [14] as follows:

$$(t,u,t_\perp,u_\perp|x,y,z) \sim \mathcal{N}\left((x,y,z)\Sigma_e\Gamma\tilde{\Sigma}_m,\tilde{\Sigma}_m\right),$$

where

$$\tilde{\Sigma}_m = \{\Sigma_m^{-1} + \Gamma^\top \Sigma_\epsilon^{-1} \Gamma\}^{-1},$$
$$\Sigma_\epsilon = \begin{bmatrix} I_p\sigma_e^2 & 0 & 0 \\ 0 & I_q\sigma_f^2 & 0 \\ 0 & 0 & \sigma_g^2 \end{bmatrix},$$
$$\Gamma = \begin{bmatrix} W & 0 & W_\perp & 0 \\ 0 & C & 0 & C_\perp \\ a - bB & b & 0 & 0 \end{bmatrix},$$
$$\Sigma_m = \begin{bmatrix} \Sigma_t & \Sigma_t B & 0 & 0 \\ \Sigma_t B & \Sigma_u & 0 & 0 \\ 0 & 0 & \Sigma_{t_\perp} & 0 \\ 0 & 0 & 0 & \Sigma_{u_\perp} \end{bmatrix}.$$

In the M step, we set the partial derivatives of the conditional expectations in (5.13) to zero and get an update of each parameter.

Regarding the first term $Q_{\{a,b,\sigma_g^2\}}$, taking partial derivatives with respect to $\alpha = (a, b)$ yields

$$\hat{\alpha} = Z^\top \mathbb{E}[(T, H)]\, \mathbb{E}[(T, H)^\top (T, H)]^{-1}.$$

This resembles the usual maximum likelihood estimator for the regression coefficient in a linear regression model where Z is regressed on $\mathbb{E}[(T, H)]$. Taking partial derivatives with respect to σ_g^2 yields the well-known maximum likelihood estimator for the residual variance

$$\hat{\sigma_g^2} = \frac{1}{N} \operatorname{tr}\mathbb{E}[(Z - (T, H)(a, b)^\top)^\top (Z - (T, H)(a, b)^\top)] = \frac{1}{N} \operatorname{tr}\mathbb{E}[G^\top G].$$

Regarding $Q_{\{W,W_\perp,\sigma_e^2\}}$ which involves optimization over semi-orthogonal loading matrices W and $W_\perp$, Lagrange multipliers Λ_W and $\Lambda_{W_\perp}$ are introduced. The objective function to minimize is then

$$\operatorname{tr}\mathbb{E}\left[(X - TW^\top - T_\perp W_\perp^\top)^\top (X - TW^\top - T_\perp W_\perp^\top)\right] + \Lambda_W (W^\top W - I_K) + \Lambda_{W_\perp} (W_\perp^\top W_\perp - I_{K_x}). \tag{5.14}$$

Note that the objective function involves both W and $W_\perp$ and cannot be decoupled. We adopt here the same strategy used in [14] that performs sequential optimization [30]. First, (5.14) is minimized over W, keeping $W_\perp$ constant,

$$\hat{W} = (X^\top \mathbb{E}[T] - W_\perp \mathbb{E}[T_\perp^\top T])(E[T^\top T] + \Lambda_W)^{-1} = \operatorname{orth}(X^\top \mathbb{E}[T] - W_\perp \mathbb{E}[T_\perp^\top T]), \tag{5.15}$$

where $\operatorname{orth}(A) = JV^\top$ with J and V the singular vectors of A. The last step is proven in [13]. Next, (5.14) is optimized over $W_\perp$, keeping W equal to its minimizer,

$$\hat{W_\perp} = (X^\top \mathbb{E}[T_\perp] - \hat{W}\, \mathbb{E}[T^\top T_\perp])(E[T_\perp^\top T_\perp] + \Lambda_{W_\perp})^{-1} = \operatorname{orth}(X^\top \mathbb{E}[T_\perp] - \hat{W}\, \mathbb{E}[T^\top T_\perp]).$$

In the same way, estimates for semi-orthogonal loading matrices C and $C_\perp$ can be obtained.

5

For the matrices that are restricted to be diagonal, for example, the inner regression matrix B, we set the off-diagonals to zero using its Hadamard product with an identity matrix as follows,

$$\hat{B} = \mathbb{E}[U^\top T](\mathbb{E}[T^\top T])^{-1} \circ I_K.$$

Now the EM updates at step j can be written as follows, starting with an initial guess for $j = 0$. Denote $\mathbb{E}_j[\cdot] = \mathbb{E}[\cdot|X, Y, Z, \theta^j]$.

$$\begin{aligned}
(a,b)^{j+1} &= Z^\top \mathbb{E}_j[(T,H)]\mathbb{E}_j[(T,H)^\top(T,H)]^{-1} \\
W^{j+1} &= \text{orth}(X^\top \mathbb{E}_j[T] - W_\perp^j \mathbb{E}_k[T_\perp^\top T]) \\
W_\perp^{j+1} &= \text{orth}(X^\top \mathbb{E}_j[T_\perp] - W^{j+1}\mathbb{E}_j[T^\top T_\perp]) \\
C^{j+1} &= \text{orth}(Y^\top \mathbb{E}_j[U] - C_\perp^j \mathbb{E}_k[U_\perp^\top U]) \\
C_\perp^{j+1} &= \text{orth}(Y^\top \mathbb{E}_j[U_\perp] - C^{j+1}\mathbb{E}_j[U^\top U_\perp]) \\
B^{j+1} &= \mathbb{E}_j[U^\top T](\mathbb{E}_j[T^\top T])^{-1} \circ I_K \\
\Sigma_t^{j+1} &= \frac{1}{N}\mathbb{E}_j[T^\top T] \circ I_K \\
\Sigma_{t_\perp}^{j+1} &= \frac{1}{N}\mathbb{E}_j[T_\perp^\top T_\perp] \circ I_{K_x} \\
\Sigma_{u_\perp}^{j+1} &= \frac{1}{N}\mathbb{E}_j[U_\perp^\top U_\perp] \circ I_{K_y} \\
\Sigma_h^{j+1} &= \frac{1}{N}\mathbb{E}_j[H^\top H] \circ I_K \\
(\sigma_e^2)^{j+1} &= \frac{1}{Np}\,\text{tr}(\mathbb{E}_j[E^\top E]) \\
(\sigma_f^2)^{j+1} &= \frac{1}{Nq}\,\text{tr}(\mathbb{E}_j[F^\top F]) \\
(\sigma_g^2)^{j+1} &= \frac{1}{N}\,\text{tr}(\mathbb{E}_j[G^\top G])
\end{aligned} \tag{5.16}$$

5.6.2. AN EM ALGORITHM FOR GLM-PO2PLS MODEL WITH A BERNOULLI DISTRIBUTED OUTCOME

The associated log-likelihood for the GLM-PO2PLS model with a Bernoulli distributed outcome involves an integral with respect to $(\nu, \xi) = ((t,u),(t_\perp,u_\perp))$ of dimension $(2K + K_x + K_y)$. By integrating out ξ, the dimension of integral can be reduced to $2K$ as follows

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

The conditional probability mass/density functions involved are given by

$$
\begin{aligned}
p(z|v) &= \begin{cases} \Big(1+\exp\{-(\beta_0 + ta^\top + (u-tB)b^\top)\}\Big)^{-1}, & z=1 \\ \Big(1+\exp\{\beta_0 + ta^\top + (u-tB)b^\top\}\Big)^{-1}, & z=0, \end{cases} \\
f(x|v) &= (2\pi)^{-\frac{p}{2}} |\Sigma_{x|t}|^{-\frac{1}{2}} exp\Big(-\frac{1}{2}(x-tW^\top)\Sigma_{x|t}^{-1}(x-tW^\top)^\top\Big), \\
f(y|v) &= (2\pi)^{-\frac{q}{2}} |\Sigma_{y|u}|^{-\frac{1}{2}} exp\Big(-\frac{1}{2}(y-uC^\top)\Sigma_{y|u}^{-1}(y-uC^\top)^\top\Big), \\
f(v) &= (2\pi)^{-K} |\Sigma_v|^{-\frac{1}{2}} exp\Big(-\frac{1}{2} v\Sigma_v^{-1} v^\top\Big),
\end{aligned}
\tag{5.17}
$$

where

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

Note here that the determinant and the inverse of a $p \times p$ matrix $\Sigma_{x|t}$ (and a $q \times q$ matrix $\Sigma_{y|u}$) are required in (5.17). Calculating them directly is not feasible, even for a moderate p (or q). Following matrix determinant lemma [19] and Woodbury matrix identity [57], we perform the following transformation, such that only calculation of the determinant and inverse of a $K_x \times K_x$ (or $K_y \times K_y$) matrix is required. Here, the transformation utilizes the semi-orthogonality constraint of $W_\perp^\top W_\perp = I_{K_x}$.

$$
\begin{aligned}
\log|\Sigma_{x|t}| &= \log\Big| I_{K_x} + \frac{1}{\sigma_e^2}\Sigma_{t_\perp} \Big| + p\log\sigma_e^2, \\
\Sigma_{x|t}^{-1} &= \frac{1}{\sigma_e^2}\Big(I_p - \frac{1}{\sigma_e^2} W_\perp (\Sigma_{t_\perp}^{-1} + \frac{1}{\sigma_e^2} I_{K_x})^{-1} W_\perp^\top \Big),
\end{aligned}
$$

The determinant and inverse of $\Sigma_{y|u}$ are calculated analogously.

In the EM algorithm, we consider the partial complete data vector (x, y, z, v). For each current estimate θ', the algorithm optimizes for each subject the objective function

$$
Q(\theta|\theta') = \mathbb{E}[\log f(x, y, z, v|\theta)|x, y, z, \theta'].
\tag{5.18}
$$

Similar to (5.12), the conditional expectation in (5.18) can be decomposed to factors that depend on distinct sets of parameters,

$$
\begin{aligned}
Q(\theta|\theta') &= \mathbb{E}[\log f(x, y, z, v)] = \mathbb{E}[\log f(x, y, z|v)] + \mathbb{E}[\log f(v)] \\
&= \underbrace{\mathbb{E}[\log p(z|v)]}_{Q_{\{\beta_0,a,b\}}} + \underbrace{\mathbb{E}[\log f(x|t)]}_{Q_{\{W,W_\perp,\sigma_e^2,\Sigma_{t_\perp}\}}} + \underbrace{\mathbb{E}[\log f(y|u)]}_{Q_{\{C,C_\perp,\sigma_f^2,\Sigma_{u_\perp}\}}} + \underbrace{\mathbb{E}[\log f(u|t)]}_{Q_{\{B,\Sigma_h\}}} + \underbrace{\mathbb{E}[\log f(t)]}_{Q_{\Sigma_t}}.
\end{aligned}
\tag{5.19}
$$

The first term $Q_{\{\beta_0,a,b\}}$ needs to be estimated with numerical integration for each observed instance (X_i, Y_i, Z_i) with respect to $v_i = (T_i, U_i)$ as follows,

$$
\begin{aligned}
Q_{\{\beta_0,a,b\}} &= \sum_{i=1}^N \int [\log p(Z_i|v_i)] f(v_i|X_i, Y_i, Z_i, \theta') dv_i \\
&= \sum_{i=1}^N \frac{\int [\log p(Z_i|v_i)] p(Z_i|v) f(X_i|v_i) f(Y_i|v_i) f(v_i) dv_i}{\int p(Z_i|v_i) f(X_i|v_i) f(Y_i|v_i) f(v_i) dv_i}.
\end{aligned}
\tag{5.20}
$$

The numerator and denominator for each subject in (5.20) can be approximated separately by

$$\int \varphi(v)p(z|v)f(x,y|v)f(v)dv \approx \sum_{m_1=1}^{M} \dots \sum_{m_{2K}=1}^{M} \varphi(v=v_m)p(z|v=v_m)f(x,y|v=v_m)w_{m_1}\dots w_{m_{2K}}, \tag{5.21}$$

with the function φ being $\log p(z|v)$ and 1, respectively.

Given the observed data for N subjects, the other terms in (5.19) are given by

$$\begin{aligned}
Q_{\{W,W_\perp,\sigma_e^2,\Sigma_{t_\perp}\}} &= -\frac{1}{2}\Big(Np\log(2\pi) + N\log|\Sigma_{x|t}| + \mathrm{tr}\mathbb{E}\left[(X-TW^\top)^\top\Sigma_{x|t}^{-1}(X-TW^\top)\right]\Big),\\
Q_{\{C,C_\perp,\sigma_f^2,\Sigma_{u_\perp}\}} &= -\frac{1}{2}\Big(Nq\log(2\pi) + N\log|\Sigma_{y|u}| + \mathrm{tr}\mathbb{E}\left[(Y-UC^\top)^\top\Sigma_{y|u}^{-1}(Y-UC^\top)\right]\Big),\\
Q_{\{B,\Sigma_h\}} &= -\frac{1}{2}\Big\{NK\log(2\pi) + N\log|\Sigma_h| + \mathrm{tr}\mathbb{E}\left[(U-TB)^\top(U-TB)\Sigma_h^{-1}\right]\Big\},\\
Q_{\Sigma_t} &= -\frac{1}{2}\Big\{NK\log(2\pi) + N\log|\Sigma_t| + \mathrm{tr}\mathbb{E}\left[T^\top T\Sigma_t^{-1}\right]\Big\}.
\end{aligned} \tag{5.22}$$

The conditional expectations in (5.22) involve calculation of the first and second conditional moments of $v_i = (T_i, U_i)$ for each subject i. The conditional moments are given by

$$\begin{aligned}
\mathbb{E}[v_i|X_i,Y_i,Z_i,\theta'] &= \frac{\int v_i p(Z_i|v_i)f(X_i|v_i)f(Y_i|v_i)f(v_i)dv_i}{\int p(Z_i|v_i)f(X_i|v_i)f(Y_i|v_i)f(v_i)dv_i}\\
\mathbb{E}[v_i^\top v_i|X_i,Y_i,Z_i,\theta'] &= \frac{\int v_i^\top v_i p(Z_i|v_i)f(X_i|v_i)f(Y_i|v_i)f(v_i)dv_i}{\int p(Z_i|v_i)f(X_i|v_i)f(Y_i|v_i)f(v_i)dv_i}
\end{aligned}$$

Here, the integrals are numerically calculated with (5.21).

In the M step, maximizing $Q_{\{\beta_0,a,b\}}$ requires iteration. We propose a one-step gradient descent strategy to find an update of $\beta = (\beta_0, a, b)$ along the direction of the gradient given by

$$\nabla Q_\beta = \sum_{i=1}^{N}\left\{\frac{\int[\frac{\partial}{\partial\beta}\log p(Z_i|v_i)]p(Z_i|v)f(X_i|v_i)f(Y_i|v_i)f(v_i)dv_i}{\int p(Z_i|v_i)f(X_i|v_i)f(Y_i|v_i)f(v_i)dv_i}\right\}^\top.$$

A step size that guarantees the increase of Q_β is searched using the backtracking rule [2].

To estimate the semi-orthogonal joint loading matrix W, we relax the orthogonality constraint temporarily, and obtain an intermediate estimator $\hat{W}_*$ by setting the partial derivative of $Q_{\{W,W_\perp,\sigma_e^2,\Sigma_{t_\perp}\}}$ with respect to W to zero,

$$\hat{W}_* = X^\top\mathbb{E}[T]\mathbb{E}[T^\top T]^{-1}.$$

To impose the orthogonality constraint, the "orth" operator in (5.15) is used,

$$\hat{W} = \mathrm{orth}(\hat{W}_*).$$

This strategy was also used in [26] for estimation of orthogonal loading matrices.

The parameters $W_\perp, \sigma_e^2, \Sigma_{t_\perp}$ are contained in $\Sigma_{x|t}$. Therefore, we take derivative of $Q_{\{W,W_\perp,\sigma_e^2,\Sigma_{t_\perp}\}}$ with respect to $\Sigma_{x|t}$ as follows,

$$\frac{\partial Q_{\{W,W_\perp,\sigma_e^2,\Sigma_{t_\perp}\}}}{\partial \Sigma_{x|t}} = \Sigma_{x|t}^{-1} - \frac{1}{N}\Sigma_{x|t}^{-1}\mathbb{E}[(X-TW^\top)^\top(X-TW^\top)]\Sigma_{x|t}^{-1}. \tag{5.23}$$

Since $\Sigma_{x|t} = W_\perp \Sigma_{t_\perp} W_\perp^\top + \sigma_e^2 I_p$ is a full-rank matrix, by setting (5.23) to zero, we get the following relationship,

$$\frac{1}{N}\mathbb{E}[(X-TW^\top)^\top(X-TW^\top)] = \Sigma_{x|t} = W_\perp \Sigma_{t_\perp} W_\perp^\top + \sigma_e^2 I_p.$$

Taking trace of both sides, σ_e^2 can be estimated as

$$\hat{\sigma_e^2} = \frac{1}{p}(\frac{1}{N}\mathrm{tr}\mathbb{E}[(X-TW^\top)^\top(X-TW^\top)] - \mathrm{tr}[\Sigma_{t_\perp}]).$$

Note that $W_\perp$ and $\Sigma_{t_\perp}$ can be obtained by eigendecomposition of the real symmetric matrices $(\frac{1}{N}\mathbb{E}[(X-TW^\top)^\top(X-TW^\top)] - \hat{\sigma_e^2} I_p)$. Here, power iteration is used to avoid processing a $p \times p$ matrix. Parameters in $Q_{\{C,C_\perp,\sigma_f^2,\Sigma_{u_\perp}\}}$ are estimated analogously.

Using the same notation as in (5.16), the EM algorithm updates parameters in step j as follows:

$$\begin{aligned}
(\beta_0, a, b)^{j+1} &= (\beta_0, a, b)^j + s^{j+1}\nabla Q_\beta^{j+1} \\
W^{j+1} &= \mathrm{orth}(X^\top \mathbb{E}_j[T]\mathbb{E}_k[T^\top T]^{-1}) \\
C^{j+1} &= \mathrm{orth}(Y^\top \mathbb{E}_j[U]\mathbb{E}_k[U^\top U]^{-1}) \\
B^{j+1} &= \mathbb{E}_j[U^\top T](\mathbb{E}_j[T^\top T])^{-1} \circ I_K \\
\Sigma_t^{j+1} &= \frac{1}{N}\mathbb{E}_j[T^\top T] \circ I_K \\
\Sigma_h^{j+1} &= \frac{1}{N}\mathbb{E}_j[H^\top H] \circ I_K \\
(\sigma_e^2)^{j+1} &= \frac{1}{p}(\frac{1}{N}\mathrm{tr}\mathbb{E}_j[(X-TW^\top)^\top(X-TW^\top)] - \mathrm{tr}[\Sigma_{t_\perp}^j]) \\
(\sigma_f^2)^{j+1} &= \frac{1}{q}(\frac{1}{N}\mathrm{tr}\mathbb{E}_j[(Y-UC^\top)^\top(Y-UC^\top)] - \mathrm{tr}[\Sigma_{u_\perp}^j]) \\
W_\perp^{j+1} &= \text{eigen vectors of } (\frac{1}{N}\mathbb{E}_j[(X-TW^{j+1^\top})^\top(X-TW^{j+1^\top})] - (\sigma_e^2)^{j+1} I_p) \\
C_\perp^{j+1} &= \text{eigen vectors of } (\frac{1}{N}\mathbb{E}_j[(Y-UC^{j+1^\top})^\top(Y-UC^{j+1^\top})] - (\sigma_f^2)^{j+1} I_q) \\
\Sigma_{t_\perp}^{j+1} &= \mathrm{diag}[\text{eigen values of } (\frac{1}{N}\mathbb{E}_j[(X-TW^{j+1^\top})^\top(X-TW^{j+1^\top})] - (\sigma_e^2)^{j+1} I_p)] \\
\Sigma_{u_\perp}^{j+1} &= \mathrm{diag}[\text{eigen values of } (\frac{1}{N}\mathbb{E}_j[(Y-UC^{j+1^\top})^\top(Y-UC^{j+1^\top})] - (\sigma_f^2)^{j+1} I_q)]
\end{aligned}$$

5.6.3. THE ASYMPTOTIC DISTRIBUTION

Recall the GLM-PO2PLS model with a normally distributed outcome,

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

The parameters are collected in $\theta = \{W, C, W_\perp, C_\perp, a, b, B, \Sigma_t, \Sigma_{t_\perp}, \Sigma_{u_\perp}, \sigma_e^2, \sigma_f^2, \Sigma_h, \sigma_g^2\}$, and the associated log-likelihood is given by

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

Here, the covariance matrix Σ_θ is

$$\Sigma_\theta = \begin{bmatrix} W\Sigma_t W^\top + W_\perp \Sigma_{t_\perp} W_\perp^\top + \sigma_e^2 I_p & W\Sigma_t BC^\top & W\Sigma_t a^\top \\ CB\Sigma_t W^\top & C\Sigma_u C^\top + C_\perp \Sigma_{u_\perp} C_\perp^\top + \sigma_f^2 I_q & C(\Sigma_h b^\top + B\Sigma_t a^\top) \\ a\Sigma_t W^\top & (a\Sigma_t B + b\Sigma_h)C^\top & a\Sigma_t a^\top + b\Sigma_h b^\top + \sigma_g^2 \end{bmatrix}. \tag{5.24}$$

We show here that 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 applies to the GLM-PO2PLS continuous model. It has been shown in [14] that the proposition applies to the PO2PLS model. Similarly, we define τ as the mapping from a θ' to $\Sigma_{\theta'}$, given in (5.24), and the discrepancy function F as $F(S; \Sigma_{\theta'}) = N\log|\Sigma_{\theta'}| + \operatorname{tr} S\Sigma_{\theta'}^{-1} - N\log|S| + \operatorname{tr} SS^{-1}$, where S is the maximum likelihood estimator of the covariance matrix of (x, y, z). The function F can be recognized as the discrepancy of two log-likelihoods evaluated at S and $\Sigma_{\theta'}$, respectively, with a minimizer $\Sigma_{\hat{\theta}}$. The mapping function τ is analytic and quadratic in θ'. It follows from the definition of F and the regularity of the normal log-likelihood ℓ that F is non-negative, zero only if $S = \Sigma_{\theta'}$, and positive everywhere else. Also, ℓ, thus F, is twice continuously differentiable and since GLM-PO2PLS is identifable, F has a positive definite Hessian at θ'. Then, Proposition 4.2 in Shapiro (1986) states that the elements of $\Sigma_{\hat{\theta}}$ are asymptotically normally distributed. Theorem 2 in the main article follows,

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

Covariance matrix of the coefficients Let $\alpha = (a, b)$ and $\alpha_k = (a_k, b_k)$. The two test statistics $T_{full} = \hat{\alpha}\Pi_{\hat{\alpha}}^{-1}\hat{\alpha}^\top$, and $T_{comp.wise} = \hat{\alpha_k}\Pi_{\hat{\alpha_k}}^{-1}\hat{\alpha_k}^\top$ involve calculation of the covariance matrix of the coefficients $\Pi_{\hat{\alpha}}$.

Let $\mathcal{I}(\hat{\theta})$ be the observed Fisher information matrix. 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 $\hat{\alpha}$ and $\hat{\theta}/\hat{\alpha}$ are asymptotically independent 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 $2K \times 2K$ observed Fisher

information matrix $\mathscr{I}(\hat{\alpha})$ is given by

$$\begin{aligned}\mathscr{I}(\hat{\alpha}) &= \sum_{i=1}^{N} \mathbb{E}[B_i(\hat{\alpha})|\psi_i] - \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[S_i(\hat{\alpha})S_j(\hat{\alpha})^\top|\psi_i;\psi_j] \\ &= \sum_{i=1}^{N} \mathbb{E}[B_i(\hat{\alpha})|\psi_i] - \sum_{i=1}^{N} \mathbb{E}[S_i(\hat{\alpha})S_i(\hat{\alpha})^\top|\psi_i] - \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} \mathbb{E}[S_i(\hat{\alpha})|\psi_i]\mathbb{E}[S_j(\hat{\alpha})|\psi_j]^\top\end{aligned} \tag{5.25}$$

Here, $S_i(\hat{\alpha}) = \frac{1}{\sigma_g^2}\left((T_i,H_i)^\top Z_i - (T_i,H_i)^\top(T_i,H_i)\hat{\alpha}^\top\right)$, and $B_i(\hat{\alpha}) = \frac{1}{\sigma_g^2}(T_i,H_i)^\top(T_i,H_i)$. Note that (5.25) involves conditional expectations of cubic and quadratic terms of (T_i,H_i). It can be re-formulated in terms of the first and second conditional moments $\mu_i = \mathbb{E}[(T_i,H_i)]$, and $V_i = \mathbb{E}[(T_i,H_i)^\top(T_i,H_i)]$, which are readily available from the E step of EM algorithm,

$$\begin{aligned}\mathscr{I}(\hat{\alpha}) = \frac{1}{(\hat{\sigma_g^2})^2}\sum_{i=1}^{N}\Big\{&\hat{\sigma_g^2}V_i - Z_i^2V_i + Z_i\Big(\mu_i^\top\hat{\alpha}V_i + (\mu_i^\top\hat{\alpha}V_i)^\top + \hat{\alpha}\mu_i^\top(V_i - 2\mu_i^\top\mu_i)\Big) \\ &+ Z_i\Big(\mu_i^\top\hat{\alpha}V_i + (\mu_i^\top\hat{\alpha}V_i)^\top + \hat{\alpha}\mu_i^\top(V_i - 2\mu_i^\top\mu_i)\Big)^\top \\ &- 2V_i\hat{\alpha}^\top\hat{\alpha}V_i - \mu_i\hat{\alpha}^\top\hat{\alpha}\mu_i^\top(V_i - 2\mu_i^\top\mu_i) - \mathrm{tr}[\hat{\alpha}^\top\hat{\alpha}(V_i - \mu_i^\top\mu_i)]V_i \\ &- \sum_{j=1,j\neq i}^{N}(Z_i\mu_i^\top - V_i\hat{\alpha}^\top)(Z_j\mu_j^\top - V_j\hat{\alpha}^\top)^\top\Big\}.\end{aligned}$$

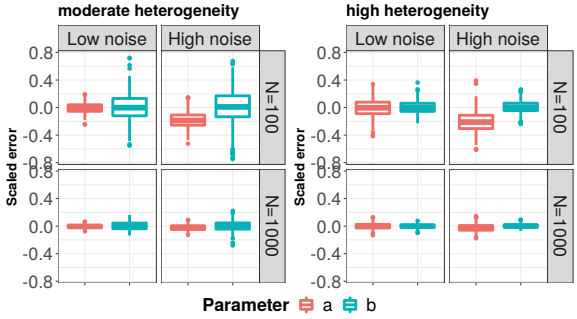
5.6.4. ADDITIONAL SIMULATION RESULTS

We evaluated the performance of GLM-PO2PLS for both normally and Bernoulli distributed outcome under different combinations of sample size, dimensionality, heterogeneity level, and noise level. In the main article, we only described the results of coefficient estimation and outcome prediction in high-dimensional settings. Here, we show the results of coefficient estimation and outcome prediction in low-dimensional settings, and the results of loading estimation and feature selection in both high and low dimensions.

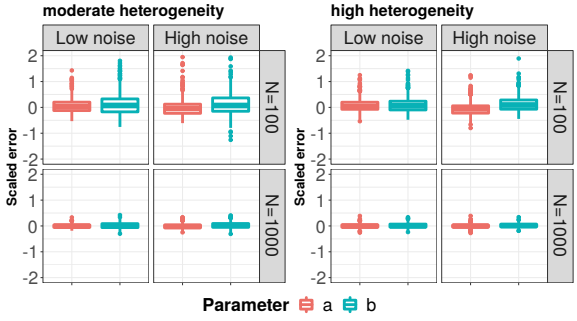
Figure 5.4 and Figure 5.5 depict the results of coefficient estimation and outcome prediction in low-dimensional settings, respectively. Overall, the performance regarding both metrics were very similar in low-dimensional and high-dimensional settings. The conclusions in high dimensional settings in the main article also hold for low dimensional settings.

Figure 5.6 shows the inner products of the estimated loading vectors and the corresponding true loadings. Overall, the loading estimation was accurate except for the x-loadings in the scenarios of small sample size, high noise level. Be reminded that in these scenarios where the x-loadings were not well estimated, the coefficient a for the x-joint components in the linear predictor of z tended to be underestimated.

Figure 5.7 shows the results regarding performance of feature selection. From Figure 5.7a for continuous outcome in high-dimensional settings, in the scenarios with small sample size and moderate heterogeneity, the median TPR of GLM-PO2PLS was 0.90 under low noise level and decreased to 0.62 when the dataset was more noisy. The



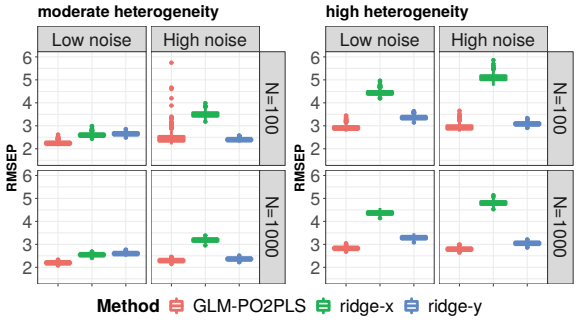
(a) Continuous z



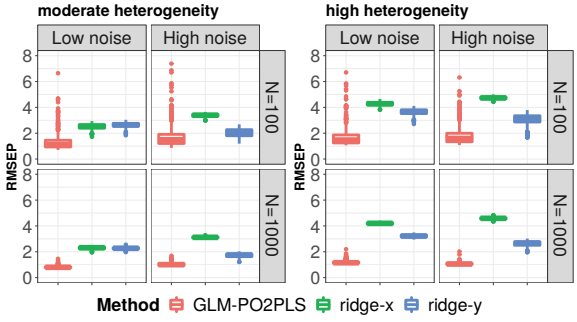
(b) Binary z

Figure 5.4: **Performance of coefficient estimation in low-dimensional settings.** The y-axis shows the scaled estimation error. Boxes show the results of 500 repetitions.

TPR of ridge-x stayed around 0.25 regardless of the noise level. When the sample size was large, the median TPR of GLM-PO2PLS increased to 0.96 under low noise level and to 0.86 under high noise level. The median TPR of ridge-x increased to 0.50 in both noise levels. The amount of heterogeneity did not have much impact on the TPRs. Comparing Figure 5.7b for binary outcome to Figure 5.7a, GLM-PO2PLS performed similarly well, while the performance of ridge-x improved. In low dimensions (Figure 5.7c and Figure 5.7d), the TPRs of GLM-PO2PLS decreased slightly, compared to the TPRs in high-dimensional settings. On the contrary, the TPR of ridge-x increased substantially. Note that GLM-PO2PLS still outperformed ridge-x in low dimensions.

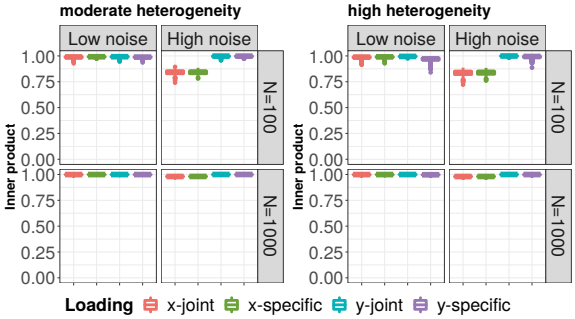


(a) Continuous z

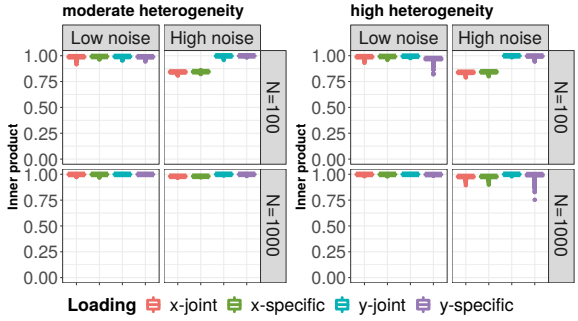


(b) Binary z

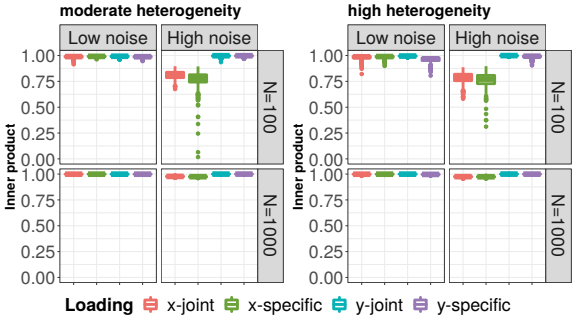
Figure 5.5: **Performance of outcome prediction in low-dimensional settings.** y-axis shows the root mean square error of prediction (RMSEP). Boxes show the results of 500 repetitions.



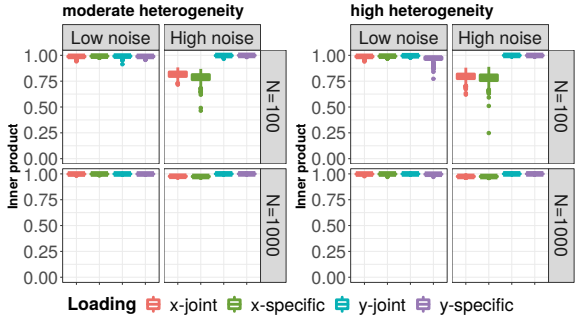
(a) Continuous z (high-dimensional)



(b) Binary z (high-dimensional)

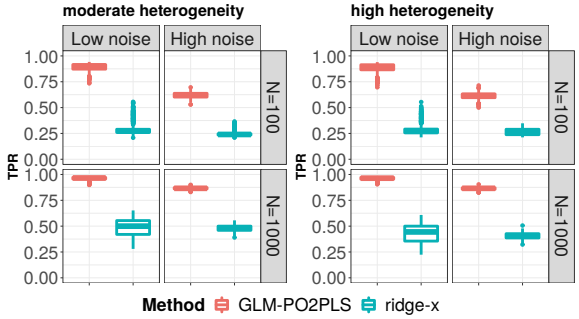


(c) Continuous z (low-dimensional)

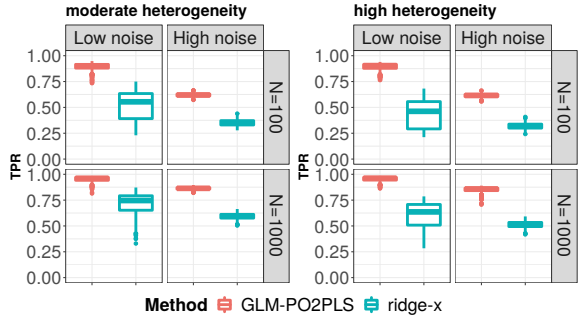


(d) Binary z (low-dimensional)

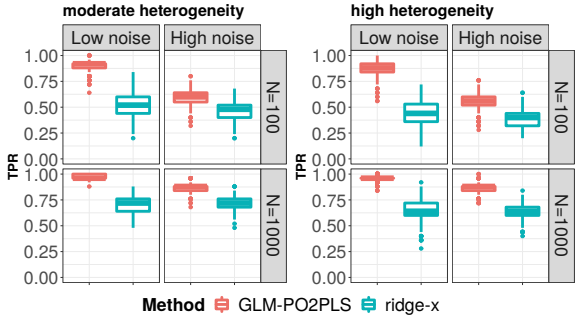
Figure 5.6: **Performance of loading estimation.** y-axis shows the inner product of the estimated loading vectors and the corresponding true loadings. Inner product of 1 suggests the loading vector is accurately estimated. Boxes show the results of 500 repetitions.



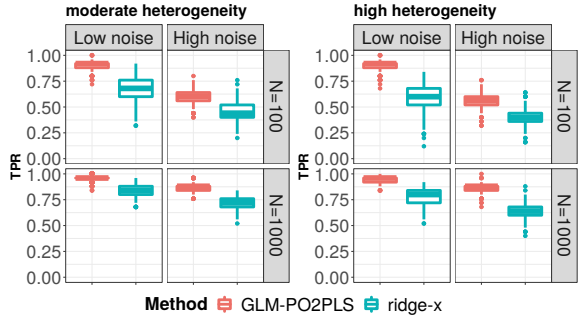
(a) Continuous z (high-dimensional)



(b) Binary z (high-dimensional)



(c) Continuous z (low-dimensional)



(d) Binary z (low-dimensional)

Figure 5.7: **Performance of feature selection.** y-axis shows the true positive rate calculated on 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). Boxes show the results of 500 repetitions.

BIBLIOGRAPHY

[1] Abramowitz, M. and Irene Stegun (1972). Numerical interpolation, differentiation, and integration. *Handbook of mathematical functions*, pages 877–925.

[2] Armijo, L. (1966). Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3.

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

[4] Bacalini, M. G., Gentilini, D., Boattini, A., Giampieri, E., Pirazzini, C., Giuliani, C., Fontanesi, E., Scurti, M., Remondini, D., Capri, M., Cocchi, G., Ghezzo, A., Rio, A. D., Luiselli, D., Vitale, G., Mari, D., Castellani, G., Fraga, M., Di Blasio, A. M., Salvioli, S., Franceschi, C., and Garagnani, P. (2015). Identification of a DNA methylation signature in blood cells from persons with down syndrome. *Aging*, 7(2):82–96.

[5] Benjamini, Y. and Hochberg, Y. (1995). 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.

[6] Borelli, V., Vanhooren, V., Lonardi, E., Reiding, K. R., Capri, M., Libert, C., Garagnani, P., Salvioli, S., Franceschi, C., and Wuhrer, M. (2015). Plasma N-Glycome Signature of Down Syndrome. *Journal of Proteome Research*, 14(10):4232–4245.

[7] Choi, C., Kim, T., Chang, K. T., and Min, K.-T. (2019). DSCR1-mediated TET1 splicing regulates miR-124 expression to control adult hippocampal neurogenesis. *The EMBO Journal*, 38(14):e101293.

[8] Ciccarone, F., Valentini, E., Malavolta, M., Zampieri, M., Bacalini, M. G., Calabrese, R., Guastafierro, T., Reale, A., Franceschi, C., Capri, M., Breusing, N., Grune, T., Morenoe Villanueva, M., Bürkle, A., and Caiafa, P. (2018). DNA Hydroxymethylation Levels Are Altered in Blood Cells from Down Syndrome Persons Enrolled in the MARK-AGE Project. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 73(6):737–744.

[9] Cindric, A., Vuckovic, F., Borelli, V., Juric, J., Deris, H., Murray, A., Alic, I., Groet, J., Petrovic, D., and Hamburg, S. (2021). Accelerated biological aging in people with Down syndrome with full and segmental trisomy 21 begins in childhood as revealed by immunoglobulin G glycosylation. *Research Square*, pages 1–29.

[10] de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263.

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[12] el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., and Uh, H. W. (2016). Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, 17(2):S11.

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

[14] Said el Bouhaddani, Hae-Won Uh, Geurt Jongbloed, and Jeanine Houwing-Duistermaat. Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, aug 2022.

[15] Franceschi, C., Garagnani, P., Gensous, N., Bacalini, M. G., Conte, M., and Salvioli, S. (2019). Accelerated bio-cognitive aging in Down syndrome: State of the art and possible deceleration strategies.

[16] Gensous, N., Bacalini, M. G., Franceschi, C., and Garagnani, P. (2020). Down syndrome, accelerated aging and immunosenescence.

[17] Gu, Z., El Bouhaddani, S., Houwing-Duistermaat, J., and Uh, H.-w. (2021). Investigating the impact of Down syndrome on Methylation and Glycomics with two-stage PO2PLS. *Theoretical Biology Forum*, (114(1-2)):29–44.

[18] Haas, M. A., Bell, D., Slender, A., Lana-Elola, E., Watson-Scales, S., Fisher, E. M., Tybulewicz, V. L., and Guillemot, F. (2013). 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(10):78561.

[19] Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer New York.

[20] Hoerl, A. E. and Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 42(1):80.

[21] Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):115.

[22] Horvath, S., Garagnani, P., Bacalini, M. G., Pirazzini, C., Salvioli, S., Gentilini, D., Di Blasio, A. M., Giuliani, C., Tung, S., Vinters, H. V., and Franceschi, C. (2015). Accelerated epigenetic aging in Down syndrome. *Aging Cell*, 14(3):491–495.

[23] Huo, H. Q., Qu, Z. Y., Yuan, F., Ma, L., Yao, L., Xu, M., Hu, Y., Ji, J., Bhattacharyya, A., Zhang, S. C., and Liu, Y. (2018). Modeling Down Syndrome with Patient iPSCs Reveals Cellular and Migration Deficits of GABAergic Neurons. *Stem Cell Reports*, 10(4):1251–1266.

[24] Krautter, F. and Iqbal, A. J. (2021). Glycans and Glycan-Binding Proteins as Regulators and Potential Targets in Leukocyte Recruitment.

[25] Krištić, J., Vučković, F., Menni, C., Klarić, L., Keser, T., Beceheli, I., Pučić-Baković, M., Novokmet, M., Mangino, M., Thaqi, K., Rudan, P., Novokmet, N., Šarac, J., Missoni, S., Kolčić, I., Polašek, O., Rudan, I., Campbell, H., Hayward, C., Aulchenko, Y., Valdes, A., Wilson, J. F., Gornik, O., Primorac, D., Zoldoš, V., Spector, T., and Lauc, G. (2014). Glycans are a novel biomarker of chronological and biological ages. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 69(7):779–789.

[26] Li, G. and Jung, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, 73(4):1433–1442.

[27] Liu, Q. and Pierce, D. A. (1994). A Note on Gauss-Hermite Quadrature. *Biometrika*, 81(3):624.

[28] Louis, T. A. (1982). Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

[29] Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641.

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

[31] Min, J. L., Hemani, G., Smith, G. D., Relton, C., and Suderman, M. (2018). Meffil: Efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, 34(23):3983–3989.

[32] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273.

[33] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370.

[34] Nishiyama, A. and Nakanishi, M. (2021). Navigating the DNA methylation landscape of cancer.

[35] Patterson, D. (2007). Genetic mechanisms involved in the phenotype of down syndrome.

[36] Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855.

[37] Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable Estimation of Generalized Linear Mixed Models using Adaptive Quadrature. *The Stata Journal: Promoting communications on statistics and Stata*, 2(1):1–21.

[38] Rockwood, N. J. (2021). Efficient Likelihood Estimation of Generalized Structural Equation Models with a Mix of Normal and Nonnormal Responses. *Psychometrika*, 86(2):642–667.

[39] Rodríguez-Girondo, M., Salo, P., Burzykowski, T., Perola, M., Houwing-Duistermaat, J., and Mertens, B. (2018). Sequential double cross-validation for assessment of added predictive ability in high-dimensional omic applications. *Annals of Applied Statistics*, 12(3):1655–1678.

[40] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, 81(393):142–149.

[41] Sheikhpour, M., Maleki, M., Ebrahimi Vargoorani, M., and Amiri, V. (2021). A review of epigenetic changes in asthma: methylation and acetylation.

[42] Sobol, M., Klar, J., Laan, L., Shahsavani, M., Schuster, J., Annerén, G., Konzer, A., Mi, J., Bergquist, J., Nordlund, J., Hoeber, J., Huss, M., Falk, A., and Dahl, N. (2019). Transcriptome and Proteome Profiling of Neural Induced Pluripotent Stem Cells from Individuals with Down Syndrome Disclose Dynamic Dysregulations of Key Pathways and Cellular Functions. *Molecular Neurobiology*, 56(10):7113–7127.

[43] Stagni, F., Giacomini, A., Emili, M., Guidi, S., and Bartesaghi, R. (2018). Neurogenesis impairment: An early developmental defect in Down syndrome.

[44] Stefanski, L. A. and Carroll, R. J. (1985). Covariate Measurement Error in Logistic Regression. 13(4):1335–1351.

[45] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(3):479–498.

[46] 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., and Mesirov, J. P. (2005). 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.

[47] Sugár, S., Tóth, G., Bugyi, F., Vékey, K., Karászi, K., Drahos, L., and Turiák, L. (2021). Alterations in protein expression and site-specific N-glycosylation of prostate cancer tissues. *Scientific Reports*, 11(1):1–12.

[48] Tabang, D. N., Ford, M., and Li, L. (2021). Recent Advances in Mass Spectrometry-Based Glycomic and Glycoproteomic Studies of Pancreatic Diseases.

[49] Tissier, R., Mar Rodr´ıguez-Girondo, and Houwing-Duistermaat, J. J. (2018). Integration of several omic sources in prediction models using network-based approaches. *Biometrical Journal*.

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

[51] Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers 2021 1:1*, 1(1):1–21.

[52] Uh, H. W., Klaric, L., Ugrina, I., Lauc, G., Smilde, A. K., and Houwing-Duistermaat, J. J. (2020). Choosing proper normalization is essential for discovery of sparse glycan biomarkers. *Molecular Omics*, 16(3):231–242.

[53] Varki, A. (2017). Biological roles of glycans. *Glycobiology*, 27(1):3–49.

[54] 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., and Gieger, C. (2018). IgG glycosylation and DNA methylation are interconnected with smoking. *Biochimica et Biophysica Acta - General Subjects*, 1862(3):637–648.

[55] Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348.

[56] WOLD, H. (1973). Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. In *Multivariate Analysis–III*, pages 383–407.

[57] Woodbury, M. (1950). *Inverting modified matrices*, volume 42. Princeton Univ.

[58] Yun, Y., Zhang, Y., Zhang, C., Huang, L., Tan, S., Wang, P., Vilariño-Gúell, C., Song, W., and Sun, X. (2021). Regulator of calcineurin 1 is a novel RNA-binding protein to regulate neuronal apoptosis. *Molecular Psychiatry*, 26(4):1361–1375.

[59] Zhang, X. and Wang, Y. (2016). Glycosylation Quality Control by the Golgi Structure.

[60] Zou, H. and Hastie, T. (2005). Erratum: Regularization and variable selection via the elastic net (Journal of the Royal Statistical Society. Series B: Statistical Methodology (2005) 67 (301-320)).

6

Further extensions and outlook on GLM-PO2PLS

The central aim of this thesis is to model an outcome disease with the correlation structure within and across omic layers. Towards this aim, we developed GLM-PO2PLS in the previous chapter. In this chapter, we give further thoughts on the limitations of the method and propose possible extensions and future directions. Specifically, we first propose a computationally feasible algorithm to overcome the computational issue for the GLM-PO2PLS binary model with more than 1 joint component and apply it on the Down syndrome (DS) dataset. Then we briefly describe an extended GLM-PO2PLS model with both joint and data-specific latent variables in the linear predictor for the outcome. It is followed by a further discussion on the importance and challenges of deriving the asymptotic normality for the GLM-PO2PLS binary model. The chapter is concluded with future directions of work.

6.1. GLM-PO2PLS BINARY MODEL WITH MULTIPLE JOINT COMPONENTS

In the previous chapter, the number of joint components in the GLM-PO2PLS binary model is limited to 1 due to the computational complexity of the EM algorithm. In the data analysis of Down syndrome, a two-stage ad-hoc approach was implemented, where we first fitted a GLM-PO2PLS continuous model with 3 joint components and 1 specific component for both methylation and glycomics, and then we subtracted the second and third joint components, and the specific components of the GLM-PO2PLS continuous model from the omics data, and fitted a single-component GLM-PO2PLS binary model on the filtered omics datasets. For easier reference, we call this approach the 'two-stage filtering' approach in rest of the chapter. Let $\tilde{x}$ and $\tilde{y}$ be the filtered omics datasets, the single-component GLM-PO2PLS binary model in the second stage is given by

$$\begin{aligned} \tilde{x} &= t_{(1)} w_{(1)}^\top + e, \\ \tilde{y} &= u_{(1)} c_{(1)}^\top + f, \\ u_{(1)} &= t_{(1)} B_{(1)} + h_{(1)}, \\ \text{logit}(p(z)) &= \beta_0 + t_{(1)} a_{(1)} + h_{(1)} b_{(1)}, \end{aligned} \tag{6.1}$$

where the latent variables and parameters with a subscript '$_{(1)}$' are the corresponding subset regarding the first pair of joint latent variables.

The rationale of utilizing the estimates from a continuous model is that the latent variables in the two models represent similar underlying biological mechanisms. The subset of parameters associated with the omics x and y are expected to be similar between the continuous and binary models. The main difference between the two models lies in the estimation of the coefficients $\beta = (\beta_0, a, b)$, which should be therefore updated by fitting a binary model.

In model (6.1), the coefficients β is only partly updated for $a_{(1)}$ and $b_{(1)}$, and the effects of the other joint latent variables on the outcome z are not taken into account. We generalize this ad-hoc approach to a two-stage EM algorithm that updates the whole vector of coefficients β in the second stage.

6.1.1. A TWO-STAGE EM ALGORITHM FOR THE GLM-PO2PLS BINARY MODEL

The algorithm involves fitting a fast GLM-PO2PLS continuous model in the first stage, and updating the estimates of the parameters corresponding to each joint component $k = 1, \ldots, K$ sequentially, using a one-component GLM-PO2PLS binary model in the second stage. The whole coefficient vector $\beta = (\beta_0, a, b)$ is always updated simultaneously rather than sequentially.

Let the estimate of the GLM-PO2PLS continuous model parameters be $\hat{\theta_0} = \{\hat{W}, \hat{C}, \hat{W_\perp}, \hat{C_\perp}, \hat{B}, \hat{\Sigma_t}, \hat{\Sigma_h}, \hat{\Sigma_{t_\perp}}, \hat{\Sigma_{u_\perp}}, \hat{\sigma_e^2}, \hat{\sigma_f^2}, \hat{a}, \hat{b}, \hat{\sigma_g^2}\}$. For each k, we fix the following subset of $\hat{\theta_0}$ corresponding to the joint components other than k (referred to by adding a subscript '$_{(|k)}$' to the parameter)

$$\hat{\theta}_{(|k)} = \{\hat{W}_{(|k)}, \hat{C}_{(|k)}, \hat{W_\perp}, \hat{C_\perp}, \hat{B}_{(|k)}, \hat{\Sigma}_{t(|k)}, \hat{\Sigma}_{h(|k)}, \hat{\Sigma_{t_\perp}}, \hat{\Sigma_{u_\perp}}, \hat{\sigma_e^2}, \hat{\sigma_f^2}\},$$

and update the rest of the parameters corresponding to the k-th joint components (referred to by adding a subscript '$_{(k)}$')

$$\theta_{(k)} = \{w_{(k)}, c_{(k)}, B_{(k)}, \sigma^2_{t(k)}, \sigma^2_{h(k)}, \beta_0, a, b\} \tag{6.2}$$

using a binary model given by

$$\begin{aligned} x &= t_{(k)} w_{(k)}^\top + \hat{t}_{(|k)} \hat{W}_{(|k)}^\top + \hat{t_\perp} \hat{W_\perp}^\top + e, \\ y &= u_{(k)} c_{(k)}^\top + \hat{u}_{(|k)} \hat{C}_{(|k)}^\top + \hat{u_\perp} \hat{C_\perp}^\top + f, \\ u_{(k)} &= t_{(k)} B_{(k)} + h_{(k)}, \\ \text{logit}(p(z)) &= \beta_0 + (t_{(k)}, \hat{t}_{(|k)}) a^\top + (h_{(k)}, \hat{h}_{(|k)}) b^\top, \end{aligned} \tag{6.3}$$

where $t_{(k)}$ and $u_{(k)}$ are the k-th joint latent variables; the other joint latent variables $\hat{t}_{(|k)}$ and $\hat{u}_{(|k)}$ are predicted from the continuous model as $\hat{v}_{(|k)} = (\hat{t}_{(|k)}, \hat{u}_{(|k)}) = \mathbb{E}[v_{(|k)}|x, y, z, \hat{\theta_0}]$; $\hat{h}_{(|k)})$ is derived as $\hat{u}_{(|k)} - \hat{t}_{(|k)} \hat{B}_{(|k)}$; the specific latent variables $\hat{t_\perp}$ and $\hat{u_\perp}$ are predicted from the continuous model as $(\hat{t_\perp}, \hat{u_\perp}) = \mathbb{E}[(t_\perp, u_\perp)|x, y, z, \hat{\theta_0}]$.

The binary model (6.3) is equivalent to

$$\begin{aligned} \tilde{x} &= t_{(k)} w_{(k)}^\top + e, \\ \tilde{y} &= u_{(k)} c_{(k)}^\top + f, \\ u_{(k)} &= t_{(k)} B_{(k)} + h_{(k)}, \\ \text{logit}(p(z)) &= \beta_0 + (t_{(k)}, \hat{t_{(|k)}}) a^\top + (h_{(k)}, \hat{h_{(|k)}}) b^\top. \end{aligned} \tag{6.4}$$

where $\tilde{x} = x - \hat{t}_{(|k)} \hat{W}_{(|k)}^\top - \hat{t_\perp} \hat{W_\perp}^\top$ is the filtered x matrix, and similarly, $\tilde{y}$ is the filtered y matrix.

the log-likelihood of the model (6.4) for the k-th component is given by

$$\begin{aligned} \ell_{(k)}(\theta_{(k)}; \tilde{x}, \tilde{y}, z, \hat{v_{(|k)}}) &= \log \int_{v_{(k)}} f(\tilde{x}, \tilde{y}, z | v_{(k)}, \hat{v_{(|k)}}, \theta_{(k)}) dv_{(k)} \\ &= \log \int_{v_{(k)}} p(z | v_{(k)}, \hat{v_{(|k)}}, \theta_{(k)}) f(\tilde{x} | v_{(k)}, \theta_{(k)}) f(\tilde{y} | v_{(k)}, \theta_{(k)}) f(v_{(k)}) dv_{(k)}. \end{aligned}$$

where

$$p(z|\nu_{(k)},\hat{\nu_{(|k)}},\theta_{(k)}) = \begin{cases} \left(1+\exp\{-(\beta_0+(t_{(k)},\hat{t_{(|k)}})a^\top+(h_{(k)},\hat{h_{(|k)}})b^\top)\}\right)^{-1}, & z=1 \\ \left(1+\exp\{\beta_0+(t_{(k)},\hat{t_{(|k)}})a^\top+(h_{(k)},\hat{h_{(|k)}})b^\top\}\right)^{-1}, & z=0, \end{cases}$$

$$f(\tilde{x}|\nu_{(k)},\theta_{(k)}) = (2\pi\sigma^2_{e(k)})^{-\frac{p}{2}} exp\Big(-\frac{1}{2\sigma^2_{e(k)}}(\tilde{x}-t_{(k)}w^\top_{(k)})(\tilde{x}-t_{(k)}w^\top_{(k)})^\top\Big),$$

$$f(\tilde{y}|\nu_{(k)},\theta_{(k)}) = (2\pi\sigma^2_{f_{(k)}})^{-\frac{q}{2}} exp\Big(-\frac{1}{2\sigma^2_{f_{(k)}}}(\tilde{y}-u_{(k)}c^\top_{(k)})(\tilde{y}-u_{(k)}c^\top_{(k)})^\top\Big),$$

$$f(\nu_{(k)}) = (2\pi\sigma^2_{t(k)})^{-\frac{1}{2}} exp\Big(-\frac{1}{2\sigma^2_{t(k)}}\nu_{(k)}\nu^\top_{(k)}\Big).$$

Estimating the binary model in the second stage An EM algorithm similar to that proposed for the binary model in Chapter 5 can be used to estimate the parameters $\theta_{(k)}$ in model (6.4). Denote the complete data vector by $(\tilde{x},\tilde{y},z,\nu_{(k)})$. For each current estimate $\theta'_{(k)}$, it considers the objective function

$$Q(\theta_{(k)}|\theta'_{(k)}) = \mathbb{E}[\log f(\tilde{x},\tilde{y},z,\nu_{(k)}|\hat{\nu}_{(|k)},\theta_{(k)})|\tilde{x},\tilde{y},z,\hat{\nu}_{(|k)},\theta'_{(k)}],$$

which can be decomposed to factors that depend on distinct subsets of $\theta_{(k)}$,

$$\begin{aligned} Q(\theta_{(k)}|\theta'_{(k)}) =& \mathbb{E}[\log f(\tilde{x},\tilde{y},z,\nu_{(k)}|\hat{\nu}_{(|k)}] = \mathbb{E}[\log f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu}_{(|k)})] + \mathbb{E}[\log f(\nu_{(k)}|\hat{\nu}_{(|k)})] \\ =& \underbrace{\mathbb{E}[\log p(z|\nu_{(k)},\hat{\nu}_{(|k)})]}_{Q_{\{\beta_0,a,b\}}} + \underbrace{\mathbb{E}[\log f(\tilde{x}|t_{(k)})]}_{Q_{\{w_{(k)},\sigma^2_{t(k)}\}}} + \underbrace{\mathbb{E}[\log f(\tilde{y}|u_{(k)})]}_{Q_{\{c_{(k)}\}}} \\ &+ \underbrace{\mathbb{E}[\log f(u_{(k)}|t_{(k)})]}_{Q_{\{B_{(k)},\sigma^2_{h(k)}\}}} + \underbrace{\mathbb{E}[\log f(t_{(k)})]}_{Q_{\{\sigma^2_{t(k)}\}}}. \end{aligned} \tag{6.5}$$

Here, the first conditional expectation $Q_{\{\beta_0,a,b\}}$ is given by

$$Q_{\{\beta_0,a,b\}} = \frac{\int_{\nu_{(k)}}[\log p(z|\nu_{(k)},\hat{\nu}_{(|k)})]f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu_{(|k)}},\theta_{(k)})d\nu_{(k)}}{\int_{\nu_{(k)}} f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu_{(|k)}},\theta_{(k)})d\nu_{(k)}},$$

and the other terms are explicit functions with respect to

$$\mathbb{E}[\nu_{(k)}|\tilde{x},\tilde{y},z,\hat{\nu}_{(|k)}] = \frac{\int_{\nu_{(k)}}\nu_{(k)}f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu_{(|k)}},\theta_{(k)})d\nu_{(k)}}{\int_{\nu_{(k)}} f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu_{(|k)}},\theta_{(k)})d\nu_{(k)}},$$

$$\mathbb{E}[\nu^\top_{(k)}\nu_{(k)}|\tilde{x},\tilde{y},z,\hat{\nu}_{(|k)}] = \frac{\int_{\nu_{(k)}}\nu^\top_{(k)}\nu_{(k)}f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu_{(|k)}},\theta_{(k)})d\nu_{(k)}}{\int_{\nu_{(k)}} f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu_{(|k)}},\theta_{(k)})d\nu_{(k)}}.$$

These integrals can be approximated using Gauss–Hermite quadrature as follows

$$\int \varphi(\nu_{(k)})f(\tilde{x},\tilde{y},z|\nu_{(k)},\hat{\nu_{(|k)}})d\nu_{(k)} \approx \sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\varphi(\nu_{(k)}=\nu_m)f(\tilde{x},\tilde{y},z|\nu_{(k)}=\nu_m,\hat{\nu_{(|k)}})w_{m_1}w_{m_2} \tag{6.6}$$

where φ is any function of $\nu_{(k)}$ and M is the number of nodes per dimension. Details of the nodes vector ν_m and weights vector $w_m = (w_{m_1}, w_{m_r})$ can be found in Chapter 5. Note that the most time-consuming component of the EM algorithm for GLM-PO2PLS binary model is evaluating the joint density $f(\tilde{x}, \tilde{y}, z|\nu_{(k)} = \nu_m, \nu_{\hat{(|k)}})$ in (6.6) at each node (or ν_m value). In the EM algorithm in Chapter 5, the joint density is evaluated at M^{2K} nodes in each EM iteration. By fixing $\nu_{(|k)}$ at $\nu_{\hat{(|k)}}$, the joint density needs to be evaluated at M^2 nodes in each EM iteration. Therefore the computational complexity of the algorithm does not grow exponentially with the number of joint component K, allowing $K > 1$.

The conditional expectation $Q_{\{\beta_0,a,b\}}$ can be maximized using the one-step gradient descent strategy in Chapter 5, and the other terms in (6.5) can be maximized by setting the first derivatives to zero.

One can keep updating $\theta_{(k)}$ for $k = 1, \ldots, K$ until convergence. We expect the loading estimates in the second stage will stay close to the initial estimate of the continuous model from the first stage, and the major update will be for the coefficients (a, b). Since the coefficient vector (a, b) is estimated jointly for any k, a iteration of k is not necessary.

6.1.2. RELATIONSHIP WITH TWO-STAGE PO2PLS MODEL

The two-stage estimation procedure for the GLM-PO2PLS binary model shares similarity with the two-stage PO2PLS model in Chapter 4 in the sense that they both estimate the loadings and predict the latent variables in the first stage and estimate the regression coefficients in the second stage. However, there are essential differences between the two. Firstly, two-stage PO2PLS contains two separate models. the one in the first stage models the omics datasets x and y, and the one in the second stage models the outcome z. The two-stage EM algorithm is an approach to obtain an estimate for the parameters of a one-stage model that jointly models (x, y, z). Secondly, the two-stage PO2PLS constructs the latent variables in an unsupervised manner, i.e., the latent variables only explain the variance and covariance of x and y. The latent variables in the first stage of the two-stage EM algorithm are predicted in a supervised manner from a GLM-PO2PLS model, thus they do use the information in outcome variable z. Lastly, two-stage PO2PLS uses the predicted latent variables as pseudo outcomes and z as covariates in several separate regression models in the second stage, and thus the interpretation is different from a logistic regression model with z as outcome. Alternatively, the predicted latent variables can be used as covariates in a logistic regression in the second stage. This ignores the errors in the covariates (predicted latent variables) and can cause biased inference results and attenuated predicted probabilities in the logistic regression [7]. In the two-stage EM algorithm, the uncertainty is partly taken into account by the joint modeling of (x, y, z) through one pair of joint latent variables in the second stage. We expect the results to be less biased.

6.1.3. APPLICATION ON DS DATASET AND COMPARISON WITH OTHER METHODS

We implemented the two-stage EM algorithm on the Down syndrome dataset described in Chapter 5 and compared the results with GLM-PO2PLS continuous model, the two-stage filtering approach, and two-stage PO2PLS. The focus is on the coefficients (a, b).

We compared the estimates as well as their p-values, assuming asymptotic normality holds for GLM-PO2PLS binary model.

We first fitted a GLM-PO2PLS continuous model with 3 joint, 1 methylation-specific, and 1 glycomic-specific components. For the two-stage EM algorithm, we set $k = 1$ and updated the parameters $\theta_{(1)}$ defined in (6.2) with model (6.4). The two-stage filtering model (6.1) was fitted on the filtered omics data (removing all the other components estimated in the continuous model apart from the first joint). For two-stage PO2PLS, a PO2PLS model with the same number of joint and specific components was first fitted, the estimated latent variables based on the fitted PO2PLS model were then used as covariates in a logistic regression. The standard error of $(\hat{a}, \hat{b})$ in GLM-PO2PLS models were estimated using the inverse of observed information matrix and p-values for the full test $H_0 : a = b = 0$ and component-wise test $H_0 : a_k = b_k = 0$ were computed using the corresponding chi-square test statistics (details in Chapter 5). The p-value of chi-square test for the logistic regression in the two-stage PO2PLS was performed using function 'anova' [1] in R. The results are presented in Table 6.1.

The top rows of Table 6.1 shows the estimates of each pair of (a_k, b_k). The GLM-PO2PLS continuous model uses an identity link, thus the estimates are not comparable to the other methods. The estimates of (a_1, b_1) were similar in the two-stage EM and the filtering approach. For the two-stage PO2PLS, the estimate of b_1 appeared to be smaller. As mentioned in the previous section, two-stage PO2PLS is unsupervised, hence the association between the joint latent variables and the outcome can be weaker than GLM-PO2PLS. The bottom rows shows the p-values of the corresponding test. The continuous model had the smallest p-value for the full test, and this significance was driven mainly by the first pair of joint components. The two-stage EM showed similar results, but less significant than the continuous model. The filtering approach had only the first pair in the model and yielded a p-value slightly less than the two-stage EM. The full test in the two-stage PO2PLS was more significant than that in the two-stage EM. Note that the two-stage PO2PLS does not taken into account the uncertainty in predicting the latent variables, hence can under-estimate the true variance.

6.2. LINKING DATA-SPECIFIC PART TO THE OUTCOME

An assumption of GLM-PO2PLS model is that the effect of the two omics datasets on the outcome is solely through the joint parts of x and y. This is quite restrictive, as there could be direct effects of x and y on z that are not joint. Furthermore, as discussed in Chapter 5, linking the omic-specific parts to the outcome can provide insights into the biological system underlying the outcome that is unique to a particular omic level, and it can lead to improved outcome prediction performance in some studies. Therefore, we propose an extended model which links the outcome to both the joint and data-specific parts.

Inheriting the notations from the GLM-PO2PLS model, the extended model is given

by

$$\begin{aligned} x &= tW^\top + t_\perp W_\perp^\top + e, \\ y &= uC^\top + u_\perp C_\perp^\top + f, \\ u &= tB + h, \\ \eta(\mathbb{E}[z]) &= \beta_0 + ta^\top + hb^\top + t_\perp a_\perp^\top + u_\perp b_\perp^\top, \end{aligned} \tag{6.7}$$

where the K_x-dimensional vector $a_\perp$ and the K_y-dimensional vector $b_\perp$ are the coefficients for the x- and y- specific latent variables $t_\perp$ and $u_\perp$, respectively. It can be shown using the Theorem 2.1 in Chapter 5 that the extended model (6.7) is identifiable under the same conditions.

From the methodological point of view, the extended model might also be more stable. Imagine an x-specific variation (specific with respect to only the other omics y, not z) that is associated with the outcome z. Such a scenario is not uncommon in practice, for example, a genomic effect (of x) on a cardiovascular disease (outcome z) that has nothing to do with glycomics (y). In this case, the first three equations of GLM-PO2PLS (or the PO2PLS part) would 'categorize' this variation to the x-specific subspace $t_\perp$. However, the last equation $\eta(\mathbb{E}[z]) = \beta_0 + ta^\top + hb^\top$ tends to capture this variation in the x-joint subspace t. Thus, where this variation ends up in the model becomes unpredictable. The variation can also be split into both subspaces, making interpretation difficult. In the extended model (6.7), such a variation will clearly be captured by $t_\perp$.

For a normally distributed outcome z, the data vector (x, y, z) follows a multivariate normal distribution $\mathcal{N}(0, \Sigma_\theta)$, with a covariance matrix given by

$$\Sigma_\theta = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_z^2 \end{bmatrix},$$

where σ_x^2, σ_y^2 and σ_{xy} are the same as in the GLM-PO2PLS model and

$$\begin{aligned} \sigma_{xz} &= a\Sigma_t W^\top + a_\perp \Sigma_{t_\perp} W_\perp^\top, \\ \sigma_{yz} &= (a\Sigma_t B + b\Sigma_h) C^\top + b_\perp \Sigma_{u_\perp} C_\perp^\top, \\ \sigma_z^2 &= a\Sigma_t a^\top + b\Sigma_h b^\top + a_\perp \Sigma_{t_\perp} a_\perp^\top + b_\perp \Sigma_{u_\perp} b_\perp^\top + \sigma_g^2. \end{aligned}$$

An efficient EM algorithm similar to the one proposed for GLM-PO2PLS continuous model in Chapter 5 can be used to estimate the model.

For a binary outcome z, the log-likelihood of the observed data involves a integral of dimension $2K + K_x + K_y$,

$$\ell(\theta; x, y, z) = \log \int_{(v,\xi)} f(x, y, z | v, \xi, \theta) f(v, \xi | \theta) d(v, \xi), \tag{6.8}$$

where $v = (t, u)$ is a row vector of joint latent variables and $\xi = (t_\perp, u_\perp)$ is a row vector of specific latent variables. In GLM-PO2PLS binary model, the joint distribution of (x, y, z) conditional on v has an explicit form. Therefore the dimension off the integral

is reduced to $2K$ by integrating out ξ. This is not possible in the extended model (6.7). The EM algorithm proposed for GLM-PO2PLS binary model in Chapter 5 will not work due to the computational complexity (of $\mathcal{O}(M^{2K+K_x+K_y})$). The two-stage EM algorithm proposed in this Chapter might be used.

6.3. ASYMPTOTIC PROPERTIES OF GLM-PO2PLS BINARY MODEL

GLM-PO2PLS and its extensions were developed towards the aim of modeling the relationship of the outcome variable with omics jointly. To formally establish the association, statistical testing needs to be performed, which relies on the asymptotic properties. For example, in Section 6.1.3, the p-values of the statistical testing and hence the conclusions depend on the assumption of the asymptotic distribution of the estimator. An alternative way is to use resampling methods, like permutation or bootstrapping. However, in high-dimensional settings, such methods can be computationally cumbersome. Therefore, the asymptotic properties need to be studied.

For GLM-PO2PLS continuous model, we showed that the estimator for the parameters is asymptotically normally distributed and the test statistic proposed asymptotically follows a chi-square distribution. Extending the derivation of the asymptotic properties for the GLM-PO2PLS continuous model to the binary model is challenging. Firstly, in the continuous model, the observed data follows a zero-mean multivariate normal distribution which is uniquely defined by the covariance matrix

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

The mapping (denote τ) from the parameter vector θ to the moment structure is thus explicit. For a binary z which follows a Bernoulli distribution, the moment structure of the data and the mapping function τ are not explicit. Therefore it is difficult to show that τ satisfies certain regularity conditions [6]. Secondly, the likelihood function of the continuous model is an explicit function of the moment structure of data

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

while the likelihood function of the binary model is an integral as (6.8), and again, the moment structure of the data is unclear. It is not possible to parameterize the likelihood in terms of the moments as for a normally distributed outcome.

A second approach we have explored to prove the asymptotic normality is extending the proof of asymptotic properties of a logistic regression model with measurement error, where the independent variables contain error and the true values are unobserved. However, in a logistic regression model with measurement error, the observed data can be formulated in terms of the unobserved true values and vice versa, and the asymptotic properties are studied by investigating the estimator obtained by regressing the outcome on the observed variables [7]. In the GLM-PO2PLS model, the relationship between the observed data and the unobserved latent variables is more complicated.

Recall the model

$$x = tW^\top + t_\perp W_\perp^\top + e, \quad y = uC^\top + u_\perp C_\perp^\top + f, \quad u = tB + h,$$
$$\eta(\mathbb{E}[z]) = \beta_0 + ta^\top + ub^\top.$$

It is not possible to formulate the unobserved latent variables t or u in terms of the observed omics data x and y, because the joint latent variables are shared by more than one equation, and each omics dataset contains an additional data-specific subspace.

The derivation of asymptotic properties for the binary model is future work.

6.4. FUTURE DIRECTIONS

In this chapter, we proposed a computationally feasible two-stage EM algorithm for estimating the GLM-PO2PLS binary model with more than one joint components. The algorithm was implemented on the DS dataset and the results were interpreted. A major limitation of the GLM-PO2PLS model is assuming the effect of omics on the outcome goes solely through the joint parts. This limitation was addressed by allowing the data-specific parts in the linear predictor of the outcome in an extended model. Apart from the data-specific parts, one might also want to add other covariates into the linear predictor for the outcome. Actually, model (6.4) in the second stage of the two-stage EM algorithm can be regarded as a binary model with $2K-2$ covariates. Thus the corresponding algorithm can be used to estimate a GLM-PO2PLS binary model with covariates (one joint component). Estimating a GLM-PO2PLS continuous model with covariates is expected to be much easier and computationally efficient.

Often the number of observed individuals or samples in multiple omics datasets studies is relatively small and novel datasets are augmented by current biological knowledge to increase efficiency in parameter estimation. To include this type of information, penalty functions that push the estimators in the direction of the prior can be added to the likelihood function or estimating equation. For example, van de Wiel et al. [8] developed an approach that utilizes information brought by links between genes and an outcome (co-data). Li and Li [4] proposed an approach that directly incorporates a given network into a penalised regression model. For joint analysis of two omics datasets, we have developed GO2PLS in Chapter 2 which incorporates group structures in the variable selection process. More available information sources (such as databases and networks) can be added to the model to further aid parameter estimation. Research on multi-omics methods with multiple penalty functions to include these multi-source information are needed.

In many studies (such as the TwinsUK), longitudinal omics data are available [9]. The interest is often to model the effect of the omics on a survival outcome such as onset of a disease using the history of the omics. Joint models are the golden standard for this type of datasets [5]. However, the joint model is computationally intense and cannot be applied directly on high-dimensional longitudinal omics. It needs to be combined with method extracting time-dependent latent components representing the longitudinal omics datasets. A functional form for the omics data over time can be used. For one omics dataset, unsupervised method such as functional PCA [3] and supervised method such as functional PLS [2] exist. For multiple longitudinal omics, novel methods that generalizes these single-omic methods and incorporate the outcome jointly are needed.

Table 6.1: **Coefficient estimates and p-values in DS data analysis**

Method		GLM-PO2PLS continuous	two-stage EM	two-stage filtering	two-stage PO2PLS
Parameter estimation	$(\hat{a}_1, \hat{b}_1)$	(-0.044, -0.122)	(-0.286, -0.942)	(-0.277, -0.940)	(-0.282, -0.794)
	$(\hat{a}_2, \hat{b}_2)$	(0.015, -0.113)	(0.076, -0.507)	-	(0.088, -0.482)
	$(\hat{a}_3, \hat{b}_3)$	(-0.011, -0.204)	(-0.066, -1.024)	-	(-0.068, -1.107)
p-value of test	$\hat{a} = \hat{b} = 0$	6.32×10^{-5}	7.46×10^{-3}	3.07×10^{-4}	3.13×10^{-4}
	$\hat{a}_1 = \hat{b}_1 = 0$	1.35×10^{-5}	3.56×10^{-4}	3.07×10^{-4}	-
	$\hat{a}_2 = \hat{b}_2 = 0$	0.15	0.36	-	-
	$\hat{a}_3 = \hat{b}_3 = 0$	0.20	0.30	-	-

BIBLIOGRAPHY

[1] Chambers, J. M. and Hastie, T. (1992). Statistical models in S. page 608.

[2] Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *https://doi.org/10.1214/11-AOS958*, 40(1):322–352.

[3] James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

[4] Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182.

[5] Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

[6] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, 81(393):142–149.

[7] Stefanski, L. A. and Carroll, R. J. (1985). Covariate Measurement Error in Logistic Regression. 13(4):1335–1351.

[8] van de Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N., and Wilting, S. M. (2016). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine*, 35(3):368–381.

[9] Verdi, S., Abbasian, G., Bowyer, R. C., Lachance, G., Yarand, D., Christofidou, P., Mangino, M., Menni, C., Bell, J. T., Falchi, M., Small, K. S., Williams, F. M., Hammond, C. J., Hart, D. J., Spector, T. D., and Steves, C. J. (2019). TwinsUK: The UK Adult Twin Registry Update. *Twin Research and Human Genetics*, 22(6):523–529.

SUMMARY

For many human disease studies with multiple omics datasets available, a central aim is to get a multi-angle view of the disease by modeling the disease using multiple omics datasets jointly. The statistical challenges for this aim include 1) high dimensionality of the omics datasets; 2) complex correlation structure within each omic layer; 3) presence of both correlation and heterogeneity between different omic layers; and 4) different distributions of the outcomes. Most of the currently available methods either model a disease outcome with a single omic layer, overlooking the correlation between omics, or model the relationship between omics without considering the outcome. This thesis develops holistic statistical methods for jointly analyzing an outcome and two omics datasets.

The first chapter gives an introduction to the regression models (generalized linear model and its extensions) for modeling an outcome with a single omics dataset and latent variable models (partial least squares and its extensions) for integrating two omics datasets without an outcome. The aim of this thesis is to develop a statistical method that models an outcome and two omics jointly. We therefore build upon the integrative latent variable methods and propose approaches to incorporate an outcome variable. We start with an example of visually exploring the relationship between an outcome and integrated omics which are represented by the low-dimensional latent components constructed using the integrative methods. We then describe how two-stage methods can be used to model an outcome with integrated omics and statistically infer the relationships. Two-stage design can lead to biased inference, we therefore describe how to obtain unbiased results using a one-stage approach that models the joint distribution of an outcome and two omics datasets. The chapter finishes with an outline of the thesis.

In Chapter 2, we develop a sparse integrative method called group sparse two-way orthogonal partial least squares (GO2PLS), which is an extension of the latent variable method O2PLS. The method utilizes known group information among the features to select relevant groups of features, and use these relevant features to construct joint latent components. Simulation studies show that the accuracy of the constructed components is robust against high noise levels. The method is illustrated on methylation and glycomics from a population study, and regulomics and transcriptomics from a small case-control study of hypertrophic cardiomyopathy (HCM). Enrichment analysis in the population study shows involvement of the selected methylation CpG sites in the immune system where glycans play important roles. In the HCM study, the scatter plots of the estimated joint scores show separation of the HCM patients from the healthy controls. And the subset of omic features selected appears highly relevant to the outcome disease.

In Chapter 3 and Chapter 4, two-stage approaches are proposed to first integrate two omics data in the first stage, and then model an outcome variable in the second stage using the joint latent components constructed in the first stage. In Chapter 3, a novel

way of modeling an outcome using genetic scores for omics is proposed. The genetic scores are constructed from the genomics data, and contain the heritable information of an omics layer. Various ways of constructing genetic scores are explored using integrative methods and polygenic score (PGS) methods. One of the advantages of such genetic scores is that they can be computed without future observations of omics once the model is fitted. Simulation studies show that the genetic scores have predictive value for the outcome variable. We construct genetic scores for glycomics and metabolomics in two cohorts, and use the genetic scores to model BMI and type 2 diabetes (T2D). It appears that the explained variance of BMI is increased and the prediction performance of T2D is improved.

In Chapter 4, we propose a two-stage probabilistic O2PLS approach to model Down syndrome (DS) with methylation and glycomics in a family-based case-control study. First, joint components representing methylation and glycomics are constructed using probabilistic O2PLS (PO2PLS) and the association between the two omics is statistically infered. Each of these joint components is then used as pseudo-outcome and modeled via a linear mixed model with DS, age, sex as covariates and family as random effect. The first pair of joint components is significantly associated with DS, and its score plots show that the DS patients are more similar to the mothers than to the siblings possibly due to accelerated and prematured aging in the DS patients. Further, we identify the most important CpG sites and glycans in constructing the joint components. The CpG sites are related to both DS the functionality of glycans, and the selected glycans are shown to be discriminators of DS in a previous study.

The two-stage models are computationally fast, but the uncertainty in the estimates of the first stage is not taken into account in the second stage, which might lead to biased inference. In Chapter 5, a one-stage model for joint modeling of an outcome and two omics, namely, GLM-PO2PLS is developed. The model identifiability is derived and expectation–maximization (EM) algorithms to obtain maximum likelihood estimators of the parameters for the model with a normally or Bernoulli distributed outcome are developed. Test statistics are proposed to infer the association between the outcome and the omics, and their asymptotic distributions are derived. In the simulation study, we compare the outcome prediction performance of GLM-PO2PLS with ridge regression, which models the outcome with each omics dataset separately, hence does not model the correlation between omics. Results show the advantage of joint modeling over two separate models. The method is then applied on the same DS dataset as used in Chapter 4. The results support the conclusions in Chapter 4. An important gene related to DS is identified which is missed in Chapter 4.

There are several limitations of GLM-PO2PLS, among which is the heavy computational burden of the EM algorithm for a binary outcome, due to the numerical integration required in each iteration. In Chapter 6, a computationally more efficient two-stage EM algorithm is developed for the GLM-PO2PLS model with a binary outcome. It first estimates a GLM-PO2PLS continuous model regarding the binary outcome as a normally distributed variable, and then fits several one-component GLM-PO2PLS binary models sequentially to update each pair of joint latent variables, treating the other pairs as known. The algorithm makes the computation of a GLM-PO2PLS binary model with multiple components feasible. The chapter conclude with relevant extensions and fu-

ture directions of GLM-PO2PLS, including descriptions of the extended GLM-PO2PLS model allowing omic-specific latent variables in the linear predictor of the outcome.

To conclude, in this thesis we propose multiple ways to model an outcome variable with integrated omics data. The methods are validated via extensive simulations and demonstrated on various datasets. Depending on the study design and research question, each type of method has its advantages and value. Together with free and open-source software available online, the work in the thesis provides useful tools to biomedical research and helps further development of biostatistical methodology in the field of omics research.

SAMENVATTING

Voor veel studies naar ziektes waarbij meerdere omics-datasets beschikbaar zijn, is het doel om een 'multi-angle' beeld van de ziekte te verkrijgen. Dit kan d.m.v. een statistisch model die de relatie tussen de ziekte en alle gemeten omics-datasets modelleert. De statistische uitdagingen hierbij zijn 1) de hoge dimensionaliteit van de omics-datasets; 2) de complexe correlatiestructuur binnen elke omics-dataset; 3) de aanwezigheid van zowel correlatie als heterogeniteit tussen verschillende omics-datasets; en 4) de verschillende verdelingen van de uitkomsten. De meeste huidige methoden modelleren ofwel een ziekte-uitkomst met één omics-dataset, waarbij dus de correlatie tussen omics over het hoofd wordt gezien, ofwel de relatie tussen omics-datasets zonder rekening te houden met de uitkomst.

Dit proefschrift ontwikkelt statistische methoden voor het gezamenlijk analyseren van een uitkomst en twee omics-datasets. Het eerste hoofdstuk geeft een introductie tot de regressiemodellen ('generalized linear model' en zijn uitbreidingen) voor het modelleren van een uitkomst met één omics-dataset en latent variabele modellen ('partial least squares' en zijn uitbreidingen) voor het integreren van twee omics-datasets zonder een uitkomst. Het doel van dit proefschrift is het ontwikkelen van een statistische methode die een uitkomst en twee omics gezamenlijk modelleert. We bouwen daarom voort op de integratieve latent variabele methoden en stellen verschillende methoden voor om een uitkomstvariabele in het model op te nemen. We beginnen met een voorbeeld van een visueel onderzoek naar de relatie tussen een uitkomst en geïntegreerde omics die worden vertegenwoordigd door de laag-dimensionale latente componenten die zijn geconstrueerd met behulp van de integratieve methoden. Vervolgens beschrijven we hoe twee-staps-methoden kunnen worden gebruikt om een uitkomst met geïntegreerde omics te modelleren en statistisch inferentie te doen over de relaties. Een twee-staps methode kan leiden tot vertekende inferentie, daarom beschrijven we hoe we zuivere ('*unbiased*') resultaten kunnen verkrijgen met behulp van een één-staps benadering die de gezamenlijke verdeling van een uitkomst en twee omics-datasets modelleert. Het hoofdstuk eindigt met een overzicht van het proefschrift.

In Hoofdstuk 2 ontwikkelen we een integratieve methode genaamd group sparse two-way orthogonal partial least squares (GO2PLS), die een uitbreiding is van de latente variabele methode O2PLS. De methode maakt gebruik van vooraf bekende groepsinformatie over de variabelen om relevante groepen van variabelen te selecteren en deze relevante variabelen te gebruiken om gezamenlijke latente componenten te construeren. Simulatiestudies tonen aan dat de nauwkeurigheid van de geconstrueerde componenten robuust is tegen hoge niveaus van ruis. De methode wordt geïllustreerd met een voorbeel van methylation en glycomics datasets gemeten in een populatiestudie, en regulomics en transcriptomics gemeten in een kleine case-control studie naar hypertrofische cardiomyopathie (HCM). Gen-verrijkings-analyse ('*gene enrichment analysis*') in de populatiestudie toont betrokkenheid van de geselecteerde CpG-sites in het

immuunsysteem waarin glycanen een belangrijke rol spelen. In de HCM studie laten de spreidingsplots van de geschatte gezamenlijke scores een structuur zien waarbij de HCM patiënten van de gezonde controles onderscheiden kunnen worden. Daarnaast blijkt een aantal van de geselecteerde omics variabelen relevant te zijn voor de ziekte.

In Hoofdstuk 3 en Hoofdstuk 4 worden twee-staps methoden voorgesteld om eerst twee omics data te integreren in de eerste stap en vervolgens de relatie te modelleren tussen de uitkomstvariabele en de gezamenlijke latente componenten uit de eerste stap. In Hoofdstuk 3 wordt een nieuwe manier voorgesteld om een uitkomstvariabele te modelleren met behulp van genetische scores voor omics. De genetische scores worden geconstrueerd uit de genetische data en bevatten de erfelijke informatie van een omics laag. Verschillende manieren om genetische scores te construeren worden onderzocht, namelijk integratieve methoden en polygene score (PGS) methoden. Een van de voordelen van dergelijke genetische scores is dat ze kunnen worden berekend zonder toekomstige observaties van omics zodra het gefitte model beschikbaar is. Simulatiestudies tonen aan dat de genetische scores voorspellende waarde hebben voor de uitkomstvariabele. We construeren genetische scores voor glycomics en metabolomics variabelen beschikbaar in twee cohorten, en gebruiken deze om genetische scores om body mass index (BMI) en type 2 diabetes (T2D) te modelleren. Het lijkt erop dat in de verkregen modellen de verklaarde variantie van BMI is verhoogd en dat T2D beter wordt voorspeld.

In Hoofdstuk 4 stellen we een twee-staps probabilistische O2PLS-aanpak voor om het Downsyndroom (DS) te modelleren met behulp van methylation en glycomics datasets in een op families gebaseerde case-controlstudie. Eerst worden de gezamenlijke componenten die methylation en glycomics vertegenwoordigen, geconstrueerd door probabilistische O2PLS (PO2PLS) toe te passen en wordt de associatie tussen de twee omics statistisch getoetst. Elk van deze gezamenlijke componenten wordt vervolgens gebruikt als pseudo-outcome en gemodelleerd via een lineair mixed model met DS, leeftijd en geslacht als covariaten. Correlatie binnen families wordt gemodelleerd met een random effect. Het eerste paar gezamenlijke componenten is statistisch significant geassocieerd met DS en de scoreplots tonen aan dat de mensen met DS meer overeenkomsten vertonen met hun moeders dan met hun broers en zussen. Mogelijk is dit het gevolg van versnelde en vroegtijdige veroudering bij mensen met DS. Verder identificeren we de belangrijkste CpG-sites en glycanen bij het construeren van de gezamenlijke componenten. De CpG-sites hebben betrekking op zowel DS als op de functionaliteit van glycanen, en de geselecteerde glycanen blijken discriminatoren te zijn van DS in een eerdere studie.

De twee-staps methoden zijn rekenkundig snel, maar de onzekerheid in de schattingen van de eerste stap wordt niet meegenomen in de tweede stap, wat kan leiden tot een vertekende inferentie. In Hoofdstuk 5 wordt een een-staps methode ontwikkeld voor het gezamenlijk modelleren van een uitkomst en twee omics, namelijk GLM-PO2PLS. Identificeerbaarheid van het model wordt afgeleid en expectation-maximization (EM) algoritmes worden ontwikkeld om maximum likelihood schatters van de parameters te verkrijgen voor het model met een normaal of Bernoulli verdeelde uitkomst. Teststatistieken worden voorgesteld om de associatie tussen de uitkomst en de omics te toetsen en hun asymptotische verdelingen worden afgeleid. In de simulatiestudie vergelijken we de prestaties van GLM-PO2PLS met die van ridge-regressie met betrekking tot het

voorspellen van de uitkomst. Ridge modelleert de uitkomst afzonderlijk met elk omics dataset, waardoor de correlatie tussen omics niet wordt meegenomen. De resultaten laten het voordeel zien van het gezamenlijk modelleren ten opzichte van twee afzonderlijke modellen. De methode wordt vervolgens toegepast op dezelfde DS-dataset als in Hoofdstuk 4. De resultaten ondersteunen de conclusies uit Hoofdstuk 4. Daarnaast wordt een belangrijk gen gerelateerd aan DS geïdentificeerd dat in Hoofdstuk 4 gemist is.

GLM-PO2PLS heeft verschillende tekortkomingen, waaronder de zware computationele last van het EM-algoritme voor een binair uitkomst, als gevolg van de numerieke integratie die bij elke iteratie van het algoritme gemaakt moet worden. In Hoofdstuk 6 wordt een meer computationeel efficiënt twee-staps-EM-algoritme ontwikkeld voor het GLM-PO2PLS-model met een binair uitkomst. Hierbij wordt eerst een GLM-PO2PLS-continu model geschat waarbij de binaire uitkomst wordt beschouwd als een normaal verdeelde variabele en vervolgens worden verschillende GLM-PO2PLS-binaire modellen sequentieel toegepast om elk paar van gezamenlijke latente variabelen te updaten, waarbij de andere paren als bekend worden beschouwd. Het algoritme maakt de berekening van een GLM-PO2PLS-binaire model met meerdere componenten computationeel haalbaar. Het hoofdstuk sluit af met relevante uitbreidingen en toekomstige richtingen van GLM-PO2PLS, waaronder beschrijvingen van het uitgebreide GLM-PO2PLS-model dat omic-specifieke latente variabelen mogelijk maakt in de lineaire predictor van de uitkomst.

Om samen te vatten, in dit proefschrift worden meerdere manieren om een uitkomstvariabele te modelleren met geïntegreerde omics-data voorgesteld. De methoden worden gevalideerd via uitgebreide simulaties en geïllustreerd met verschillende datasets. Afhankelijk van de onderzoeksopzet en de onderzoeksvraag heeft elk type methode zijn eigen voordelen. Samen met gratis en open-source software die online beschikbaar is, biedt het werk in dit proefschrift nuttige tools voor biomedisch onderzoek en helpt het bij de verdere ontwikkeling van biostatistische methodologie in het veld van omics-onderzoek.

LIST OF PUBLICATIONS

- S. Gill, A. Karwath, H.W. Uh, V.R. Cardoso, **Z. Gu**, …, D. Kotecha. (2023) Artificial intelligence to enhance clinical value across the spectrum of cardiovascular healthcare. European Heart Journal, 44(9).

- **Z. Gu**, S. el Bouhaddani, J.J. Houwing-Duistermaat, H.W. Uh. (2021). Investigating the impact of Down syndrome on methylation and glycomics with two-stage PO2PLS. Theoretical Biology Forum, 114(1-2).

- **Z. Gu**, S. el Bouhaddani, J. Pei, J.J. Houwing-Duistermaat, H.W. Uh. (2021). Statistical integration of two omics datasets using GO2PLS. BMC Bioinformatics, 22(1).

- J. Pei, M. Schuldt, E. Nagyova, **Z. Gu**, …, M. Harakalova. (2021). Multi-omics integration identifies key upstream regulators of pathomechanisms in hypertrophic cardiomyopathy due to truncating MYBPC3 mutations. Clinical Epigenetics, 13(1).

ACKNOWLEDGEMENTS

I would like to thank my supervisors for all the scientific supervision and non-scientific supports. Prof. dr. Jeanine Houwing-Duistermaat, the programme IMforFUTURE you coordinated made me feel a strong sense of belonging and immense pride. Your dedication to statistics sets an example for what a statistician should strive to be. Your support throughout all the struggles boosted my confidence and made the completion of this work possible. Dr. Hae-Won Uh, thank you for entrusting me with this opportunity to embark on this journey and continuously fighting for my funding after IMforFUTURE. Your passion for challenges has had a profound influence on me, and our conversations have helped me find a balance between confidence and self-criticism. Dr. Said el Bouhaddani, without you, I am not sure how I would have completed this work. You were always there with enormous patience and willingness to explain every detail that I struggled to comprehend. All these details and many insightful discussions with you have become crucial building blocks of this thesis. Prof. dr. ir. René Eijkemans, thank you for being my mentor until the very end of your illustrious life. Your keen insight on scientific research, combined with your witty sense of humour, has left a lasting impact on my academic pursuits. And your encouragement and kindness will be greatly missed.

I would like to express my gratitude to the committee members for their interest in my work and their investment of time in reading and commenting on this thesis: prof. dr. D.L. Oberski, dr. W.N. van Wieringen, prof. dr. J.H. Veldink, prof. dr. ir. R.C.H. Vermeulen, prof. dr. M. Zucknick, and dr. J. de Ridder.

I would like to thank the senior fellows in the IMforFUTURE network for all the trainings during our annual meetings and supports during my secondments. I am especially grateful to prof. dr. Gastone Castellani for arranging my secondment in Bologna and providing me with invaluable guidance on network methods. Your support for my postdoc application was also invaluable. Additionally, I must say that I fell in love with the Italian accent during our meetings. I would also like to extend my gratitude to prof. dr. Paolo Garagnani for the clear explanation of the mechanism of methylation, and for providing me the Down syndrome dataset which was analysed in Chapter 4 & 5. Dr. Lucija Klarić, thank you for the valuable discussions we had during my secondment in Genos and afterwards. These discussions deepened my understanding of glycans and their associated methods, which was crucial in the development of my thesis.

Big hugs to my fellow friends in the IMforFUTURE programme. The great time we had together during our annual meetings, public engagement activities, secondments, and all the other occasions will always be memorized. Particularly, I would like to thank Azra for many inputs and discussions on glycans, and arranging the amazing month in Zagreb with Samira during my secondment in Genos. Arianna, thank you for helping me obtain the ORCADES data with detailed documentation, and the laughters in Zagreb. Iva, Bologna is amazing, and Omiš is more so. Anna, thank you for hosting me in

Bologna with Iva and coming to Utrecht for your secondment. I still wonder how many kilometres we ran in the chaotic Utrecht Marathon. Maarten, we had great time sharing an apartment in Bologna, and the pancake you taught me to make on the last day has become my signature dessert. Shafiq, you made Cambridge feel like home upon my first arrival. I am so glad that we ended up in the same department and neighbourhood after IMforFUTURE.

I am grateful for all the colleagues in the biostatistics department, Julius Center, UMC Utrecht for the friendly and welcoming working environment. Prof. dr. Miriam Sturkenboom, thank you for your support in facilitating my PhD registration and completion. Bert, I always know that I can rest assured that everything is in good hands with you. Marian, you are the first one buying me a beer in the Netherlands, and I am also glad that we remain in touch. I would also like to thank the collaborators at the UMCU: prof. dr. F. Asselbergs, dr. M. Harakalová, dr. M. Mokry, dr. J. Pei, and dr. R. Berbers.

I would like to express my heartfelt gratitude to the referees who provided invaluable support at the outset of my academic journey. Prof. Jianmin Li, I still remember how you supported my decision to switch from mechanical engineering to finance during my bachelor's studies. I am grateful that we have stayed in touch for over a decade and even had the opportunity to meet in Paris. Prof. Jin Zhu, thank you for your guidance on my bachelor's thesis, and I only wish that I had written it better. Mr. Qinghao Li, the last conversation we had in Shanghai has had a profound influence on my approach to life. Your advice to be persistent and patient has stayed with me, and I am greatly inspired by your remarkable journey in leading NewBanker to become a prominent leader in the industry.

I wish to express my thanks to prof. dr. Damian Clancy for providing valuable advices and information on PhD application during my master's study. The course 'advanced statistical methods' that you lectured was one of my favourite and built a solid foundation for my statistical work.

I am honoured to express my appreciation to dr. Jessica Barrett for acknowledging my work and providing me the opportunity to continue my journey in MRC, Cambridge. It is truly a privilege to do research in such a vibrant and supportive institute.

I am filled with tremendous love and gratefulness towards my family. Hong Zhu, my mother, and Jinxin Gu, my father, as your only son, my every decision has impacts on your life, and I wish there have been less negative ones. I hope this piece of achievement brings you confidence in the future, and most importantly, fills you with pride. Xuefen Li, my grandmother, I regret that I am absent in your 80s, but I am happily taking your crown as the most educated person in our family. Yongsheng Gu, my grandfather, thoughts of you when I am feeling down give me comfort and strength. I believe I am always blessed by you.

I would like to express my gratefulness and love to my girlfriend, Sisi Pu. I am truly blessed to have your selfless love and unconditional support for my career. Your remarkable virtues of honesty, humility, compassion, gratitude, and integrity are just a few of the many qualities that shine so brightly within you, and they continue to inspire me to become a better person each and every day. Your companionship has filled so many gaps in my life and brought immeasurable joy. I am so grateful for all that you do and

for the incredible person that you are. And when I say that you are extraordinarily wise, I mean it.

I am deeply grateful to my life-long friend, dr. Xiang Zheng. This entire journey would not have taken place without your call from Boston which gave me the strength to get up off my knees on the very darkest day. As you always say, my life is like a movie, and the scene where you walked into the movie two decades ago is one of the most impressive and unforgettable moments. Sitting on the last row of the classroom on the first day of high school, I glanced at the back door, there you were, late, with yellowish hair and in slippers. I hoped you would not sit beside me, but you did, and became the first classmate to talk to me - asking if you could copy my homework. Little did I know that you would become an integral part of my life, offering unwavering support throughout my academic journey, from master's degree to PhD, to postdoc, and beyond. Thank you for believing in my potential more than I do myself, and I am glad that I have lived up to it.

I am indebted to another life-long friend, dr. Yu Shen. You have been my role model since primary school. Our countless discussions on all kinds of topics have helped me overcome doubts and confusion in my life. I am grateful that you were the one to welcome me on my first trip to a western country. The solemn Princeton campus, the fresh noble price chocolate, what a luxury start of my academic journey. As I sit in my office in Cambridge today, I often reflect on the time six years ago when I sat in your office, applying for master programmes.

This work is not possible without the support, kindness, generosity, trust, understanding, and love from all of the following lovely people after the catastrophic collapse of Kuailu. Despite direct and indirect losses, they showed me hope, gave me faith in humanity, and brought me the courage and strength to move on.

I appreciate the good and tough times spent together with my former team members. Particularly, C. Han, J. Qin, L. Zhang, and T. Kai, I am thankful to have had you in my team and for having each other's back.

F. Feng and X. Zhu, even though you never explicitly mentioned it, I am well aware of the tremendous help you and your families provided me in the aftermath. Please know that I am and always will be deeply grateful for this.

Dr. H. Tan, in many ways, you showed me what a well-educated person is truly made of. Your encouragement and all the help you offered to me and my family will forever be remembered.

H. Sun, I cannot imagine the extra pressure this unfortunate event has placed on your shoulders, and I am heartily sorry that I could not be of more help. Your creativity, humour, insightfulness, modesty, and diligence are just a few of the many virtues that make you an indispensable leader and a highly respected person whom I deeply admire.

Q. Shen and X. Tan, few have the tremendous integrity that you possess to do what you insisted in difficult times. You may still remember teaching me to pronounce words properly almost thirty years ago, but what you may not know is that you have also given me invaluable lessons on how to live life decently over these years.

H. Zhang, thank you for the countless favours you've done for me over the past nine years, both professional and personal. I wish you, your lovely daughter, and your much-loved mother a very happy life ahead.

My sincere appreciation goes to all my relatives who have provided constant care and support all along the way, from my childhood to my school years, from my time in Shanghai to my years abroad. I would like to extend a special thank you to my granduncle and grandaunt S. Zhang & P. Xiang, as well as the families of W. Hu & J. Xue, J. Xue & M. Jin, J. Gu & J. Jiang, M. Shen, and J. Shi for the tremendous understanding and patience in the aftermath.

I owe a debt of gratitude to the family of Thomas Qin (Roger Qin, Anna Wu, and George Zhu), for all the support, generosity, hospitality, and understanding. Thomas, you were like an elder brother to me and welcomed me into your family. Every big and small thing you and your family did for me is etched in my memory. Although we have lost touch, the time I spent with you is a unique, happy, and meaningful part of my 20s that will always be deeply cherished. I hope that one day, we can rekindle our friendship. From the bottom of my heart, I wish good health and happiness to your whole family.

I extend my warmest thanks to my coach and my former business partner Y. Xu and his wife S. Sun. Day after day, you reminded me of the fire within me hidden behind thick, dark clouds. I will never forget the stormy days when you provided me with a safe haven in Bridge No.8, a place where I found peace and gathered strength to face the challenges ahead. I'd also like to thank Z. Zhou and Y. Su for their thoughtful gesture of introducing me to Y. Xu and for the Buddhist book they gifted me, which granted me a new perspective of the world.

Words cannot express how in debt of gratitude I am to Qianlin Shi, who stood with me against all odds in the darkest days. In those times that challenged me to the very core and determined who I am, your companionship, elegance, positivity, and resilience guided me to be a better person than I would have been without you. I sincerely hope that you always remember how extraordinary and deserving you are of all the beauty in this world, and I wish you a fantastic life that aligns with your exceptional qualities with all my heart.

CURRICULUM VITÆ

ZHUJIE GU was born on January 21st, 1990, in Nanxun, Huzhou, China. He began his undergraduate studies in Mechanical Engineering in 2008, but soon discovered a deeper interest in Finance. In 2012, he graduated with a Bachelor's degree in Economics and embarked on a career in the financial sector in Shanghai. From 2012 to 2016, he held various positions, including independent financial advisor, consultant team leader, consultant cluster manager, and CEO of a coaching business.

In 2017, he left China to pursue further studies in Scotland. The following year, he obtained his Master of Science (MSc) degree with distinction in Actuarial Science from Heriot-Watt University. In June 2018, he began his position as a Marie Skłodowska-Curie early stage researcher in the biostatistics department of Julius Center, University Medical Center Utrecht. Funded by the European Union's Horizon 2020 research and innovation programme IMforFUTURE, as well as the EU/EFPIA BigData@Heart grant, he worked on the development of statistical methods for integrative analysis of multiple omics with an outcome, which resulted in this thesis. The chapters of his thesis were presented at several international biometrical conferences.

In September 2022, he joined as a postdoctoral fellow at the biostatistics unit of Medical Research Council, University of Cambridge. His current research focuses on dynamic risk prediction using large-scale electronic health records data.