Supplementary Materials for Joint Modeling of An Outcome Variable and Integrated Omics Datasets Using GLM-PO2PLS

Zhujie Gu\textsuperscript{1,2}, Hae-Won Uh\textsuperscript{1}, Jeanine Houwing-Duistermaat\textsuperscript{3,1,4}, and Said el Bouhaddani\textsuperscript{1}

\textsuperscript{1}Dept. of Data Science and Biostatistics, Julius Centre, UMC Utrecht, The Netherlands
\textsuperscript{2}Medical Research Council Biostatistics Unit, University of Cambridge, UK
\textsuperscript{3}Dept. of Statistics, University of Leeds, UK
\textsuperscript{4}Dept. of Mathematics, Radboud University, Nijmegen, The Netherlands

This document is supplement to the main article, “Joint Modeling of An Outcome Variable and Integrated Omics Datasets Using GLM-PO2PLS”. It is structured into two parts: Methods, and Simulation results. Each part contains additional materials for the respective section in the main article.

In section S 1, we first give mathematical details for the EM algorithms for both GLM-PO2PLS continuous and binary models. 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 S 2, we show simulation results omitted in the main article and the additional simulations.

S 1 Methods

S 1.1 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 $\max(r + K_x, r + K_y) < N$.

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{equation}
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)]
\end{equation}

\begin{align*}
&= \mathbb{E}[\log f(z|t, u)] + \mathbb{E}[\log f(x|t, t_\perp)] + \mathbb{E}[\log f(y|u, u_\perp)] \\
&+ \mathbb{E}[\log f(u|t)] + \mathbb{E}[\log f(t)] + \mathbb{E}[\log f(t_\perp)] + \mathbb{E}[\log f(u_\perp)].
\end{align*}

In this equation, the conditioning on $x$, $y$, $z$ and $\theta'$ is dropped, to simplify notation. Given the observed data for $N$ subjects, the factorized conditional expectations in (1) 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.

\[
Q_{(a,b,\sigma^2_e)} = -\frac{1}{2} \left\{ N \log(2\pi\sigma^2_e) + \frac{1}{\sigma^2_e} \text{tr} \mathbb{E} \left[ (Z - Ta^\top - (U - TB)b^\top)^\top(Z - Ta^\top - (U - TB)b^\top) \right] \right\},
\]

\[
Q_{(w,W_\perp,\sigma^2_e)} = -\frac{1}{2} \left\{ Np \log(2\pi\sigma^2_e) + \frac{1}{\sigma^2_e} \text{tr} \mathbb{E} \left[ (X - TW^\top - T_\perp W_\perp^\top)^\top(X - TW^\top - T_\perp W_\perp^\top) \right] \right\},
\]

\[
Q_{(c,C_\perp,\sigma^2_e)} = -\frac{1}{2} \left\{ Nq \log(2\pi\sigma^2_e) + \frac{1}{\sigma^2_e} \text{tr} \mathbb{E} \left[ (Y - UC^\top - U_\perp C_\perp^\top)^\top(Y - UC^\top - U_\perp C_\perp^\top) \right] \right\},
\]

\[
Q_{(b,\Sigma_h)} = -\frac{1}{2} \left\{ NK \log(2\pi) + N \log |\Sigma_h| + \text{tr} \mathbb{E} \left[ (U - TB)^\top(U - TB)\Sigma^{-1}_h \right] \right\},
\]

\[
Q_{\Sigma_t} = -\frac{1}{2} \left\{ NK \log(2\pi) + N \log |\Sigma_t| + \text{tr} \mathbb{E} \left[ T^\top T\Sigma^{-1}_t \right] \right\},
\]

\[
Q_{\Sigma_{u\perp}} = -\frac{1}{2} \left\{ NK \log(2\pi) + N \log |\Sigma_{u\perp}| + \text{tr} \mathbb{E} \left[ U_\perp^\top U_{u\perp} \Sigma^{-1}_{u\perp} \right] \right\},
\]

\[
(2)
\]

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 \(\theta'\) 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 [3] as follows:

\[
(t, u, t_\perp, u_\perp|x, y, z) \sim \mathcal{N}\left( (x, y, z)\Sigma_h \tilde{\Sigma}_m, \tilde{\Sigma}_m \right),
\]

where

\[
\tilde{\Sigma}_m = \left\{ \Sigma_m^{-1} + \Gamma^\top \Sigma^{-1}_e \Gamma \right\}^{-1},
\]

\[
\Sigma_e = \begin{bmatrix}
I_p \sigma^2_e & 0 & 0 \\
0 & I_q \sigma^2_e & 0 \\
0 & 0 & \sigma^2_g
\end{bmatrix},
\]

\[
\Gamma = \begin{bmatrix}
W & 0 & W_\perp & 0 \\
0 & C & 0 & C_\perp \\
-a - bb & b & 0 & 0
\end{bmatrix},
\]

\[
\Sigma_m = \begin{bmatrix}
\Sigma_t & \Sigma_t B & 0 & 0 \\
\Sigma_t B & \Sigma_u & 0 & 0 \\
0 & 0 & \Sigma_{u\perp} & 0 \\
0 & 0 & 0 & \Sigma_{u\perp}
\end{bmatrix}.
\]

In the M step, we set the partial derivatives of the conditional expectations in (2) to zero and get an update of each parameter.

Regarding the first term \(Q_{(a,b,\sigma^2_e)}\), 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 \(\sigma^2_g\) yields the well-known maximum likelihood estimator for the residual variance

\[
\hat{\sigma^2_g} = \frac{1}{N} \text{tr} \mathbb{E}[(Z - (T, H)(a, b)^\top)^\top(Z - (T, H)(a, b)^\top)] = \frac{1}{N} \text{tr} \mathbb{E}[G^\top G].
\]
Regarding $Q_{\{W,W_{\perp}\sigma^2\}}$ which involves optimization over semi-orthogonal loading matrices $W$ and $W_{\perp}$, Lagrange multipliers $\Lambda_W$ and $\Lambda_{W_{\perp}}$ are introduced. The objective function to minimize is then

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

(3)

Note that the objective function involves both $W$ and $W_{\perp}$ and cannot be decoupled. We adopt here the same strategy used in [3] that performs sequential optimization [7]. First, (3) is minimized over $W$, keeping $W_{\perp}$ constant,

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

where $\text{orth}(A) = JV^\top$ with $J$ and $V$ the singular vectors of $A$. The last step is proven in [2].

Next, (3) is optimized over $W_{\perp}$, keeping $W$ equal to its minimizer,

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

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

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 $k$ can be written as follows, starting with an initial guess for $k = 0$. Denote $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|X, Y, Z, \theta^k]$.

$$(a, b)^{k+1} = Z^\top \mathbb{E}_k[(T, H)] \mathbb{E}_k[(T, H)^\top (T, H)]^{-1}$$

$$(a, b)^{k+1} = \text{orth}(X^\top \mathbb{E}_k[T] - W_{\perp}^k \mathbb{E}_k[T_{\perp}^\top T])$$

$$(a, b)^{k+1} = \text{orth}(X^\top \mathbb{E}_k[T_{\perp}] - W_{\perp}^k \mathbb{E}_k[T_{\perp}^\top T])$$

$$C^{k+1} = \text{orth}(Y^\top \mathbb{E}_k[U] - C_{\perp}^k \mathbb{E}_k[U_{\perp}^\top U])$$

$$C_{\perp}^{k+1} = \text{orth}(Y^\top \mathbb{E}_k[U_{\perp}] - C_{\perp}^{k+1} \mathbb{E}_k[U_{\perp}^\top U_{\perp}])$$

$$B^{k+1} = \mathbb{E}_k[U^\top T] (\mathbb{E}_k[T_{\perp}^\top T])^{-1} \circ I_K$$

$$\Sigma_{t}^{k+1} = \frac{1}{N} \mathbb{E}_k[T^\top T] \circ I_K$$

$$\Sigma_{t_{\perp}}^{k+1} = \frac{1}{N} \mathbb{E}_k[T_{\perp}^\top T_{\perp}] \circ I_{K_{\perp}}$$

$$\Sigma_{u_{\perp}}^{k+1} = \frac{1}{N} \mathbb{E}_k[U_{\perp}^\top U_{\perp}] \circ I_{K_{\perp}}$$

$$\Sigma_{h}^{k+1} = \frac{1}{N} \mathbb{E}_k[H^\top H] \circ I_K$$

$$(\sigma_t^2)^{k+1} = \frac{1}{N_p} \text{tr} (\mathbb{E}_k[E^\top E])$$

$$(\sigma_f^2)^{k+1} = \frac{1}{N_q} \text{tr} (\mathbb{E}_k[F^\top F])$$

$$(\sigma_g^2)^{k+1} = \frac{1}{N} \text{tr} (\mathbb{E}_k[G^\top G])$$

### S 1.2 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 $\xi$, the dimension of integral can be reduced to $2K$ as follows
\[
\ell(\theta; x, y, z) = \log \int_{(\nu, \xi)} f(x, y, z|\nu, \xi, \theta) f(\nu, \xi|\theta) d(\nu, \xi) = \log \int_{\nu} p(\nu) f(x|\nu) f(y|\nu) f(\nu) d\nu.
\]
The conditional probability mass/density functions involved are given by
\[
p(\nu)|\nu = \begin{cases} 
\left(1 + \exp\{-a_0 + ta + (u - tB)b^T\}\right)^{-1}, & z = 1 \\
\left(1 + \exp\{a_0 + ta + (u - tB)b^T\}\right)^{-1}, & z = 0,
\end{cases}
\]
\[
f(x|\nu) = (2\pi)^{-\frac{d}{2}}|\Sigma_{x|\nu}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - tW^\top)^\top \Sigma_{x|\nu}^{-1}(x - tW^\top)\right),
\]
\[
f(y|\nu) = (2\pi)^{-\frac{d}{2}}|\Sigma_{y|\nu}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y - uC^\top)^\top \Sigma_{y|\nu}^{-1}(y - uC^\top)\right),
\]
\[
f(\nu) = (2\pi)^{-K}|\Sigma_\nu|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \nu^\top \Sigma_\nu^{-1} \nu\right),
\]
where
\[
\Sigma_{x|\nu} = W_\perp \Sigma_{t|\nu} W_\perp^\top + \sigma_e^2 I_p,
\]
\[\Sigma_{y|\nu} = C_\perp \Sigma_{u|\nu} C_\perp^\top + \sigma_f^2 I_q,
\]
\[\Sigma_\nu = \begin{bmatrix} \Sigma_t & \Sigma_t B \\ B \Sigma_t & \Sigma_u \end{bmatrix}.
\]
Note here that the determinant and the inverse of a $p \times p$ matrix $\Sigma_{x|\nu}$ (and a $q \times q$ matrix $\Sigma_{y|\nu}$) are required in (6). Calculating them directly is not feasible, even for a moderate $p$ (or $q$). Following matrix determinant lemma [4] and Woodbury matrix identity [6], 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 $W_\perp W_\perp^\top = I_{K_x}$.
\[
\log |\Sigma_{x|\nu}| = \log \left|I_{K_x} + \frac{1}{\sigma_e^2} \Sigma_{t|\nu}\right| + p \log \sigma_e^2,
\]
\[
\Sigma_{x|\nu}^{-1} = \frac{1}{\sigma_e^2} \left(I_p - \frac{1}{\sigma_e^2} W_\perp (\Sigma_{t|\nu}^{-1} + \frac{1}{\sigma_e^2} I_{K_x})^{-1} W_\perp^\top\right),
\]
The determinant and inverse of $\Sigma_{y|\nu}$ are calculated analogously.

In the EM algorithm, we consider the partial complete data vector $(x, y, z, \nu)$. For each current estimate $\theta'$, the algorithm optimizes for each subject the objective function
\[
Q(\theta|\theta') = \mathbb{E}[\log f(x, y, z, \nu|\theta)|x, y, z, \theta'].
\]
Similar to (1), the conditional expectation in (7) can be decomposed to factors that depend on distinct sets of parameters,
\[
Q(\theta|\theta') = \mathbb{E}[\log f(x, y, z, \nu)] = \mathbb{E}[\log f(x, y, z|\nu)] + \mathbb{E}[\log f(\nu)] = \mathbb{E}[\log p(\nu)] + \mathbb{E}[\log f(x|\nu)] + \mathbb{E}[\log f(y|\nu)] + \mathbb{E}[\log f(u|\nu)] + \mathbb{E}[\log f(t)]
\]
\[
Q_{(nu,a,b)} \quad Q_{(w,w_\perp,\sigma_e^2,u_\perp)} \quad Q_{(c,c_\perp,\sigma_f^2,u_\perp)} \quad Q_{\Sigma_t} \quad Q_{\Sigma_u}.
\]
The first term $Q_{(a_0, a, b)}$ needs to be estimated with numerical integration for each observed instance $(X_i, Y_i, Z_i)$ with respect to $\nu_i = (T_i, U_i)$ as follows,
\[
Q_{(a_0, a, b)} = \sum_{i=1}^{N} \int \log p(Z_i|\nu_i) f(\nu_i|X_i, Y_i, Z_i, \theta') d\nu_i = \sum_{i=1}^{N} \int \frac{\int \log p(Z_i|\nu_i) p(Z_i|\nu_i) f(X_i|\nu_i) f(Y_i|\nu_i) f(\nu_i)d\nu_i}{\int p(Z_i|\nu_i) f(X_i|\nu_i) f(Y_i|\nu_i) f(\nu_i)d\nu_i} d\nu_i.
\]
The conditional expectations in (11) involve calculation of the first and second conditional moments of \( \nu_i = (T_i, U_i) \) for each subject \( i \). The conditional moments are given by

\[
\mathbb{E}[\nu_i | X_i, Y_i, Z_i, \theta^*] = \frac{\int \nu_i p(Z_i | \nu_i) f(X_i | \nu_i) f(Y_i | \nu_i) f(\nu_i) d\nu_i}{\int p(Z_i | \nu_i) f(X_i | \nu_i) f(Y_i | \nu_i) f(\nu_i) d\nu_i},
\]

\[
\mathbb{E}[\nu_i^\top | X_i, Y_i, Z_i, \theta^*] = \frac{\int \nu_i^\top \nu_i p(Z_i | \nu_i) f(X_i | \nu_i) f(Y_i | \nu_i) f(\nu_i) d\nu_i}{\int p(Z_i | \nu_i) f(X_i | \nu_i) f(Y_i | \nu_i) f(\nu_i) d\nu_i}.
\]

Here, the integrals are numerically calculated with (10).

In the M step, maximizing \( Q_{\{a_0, a, b\}} \) requires iteration. We propose a one-step gradient descent strategy to find an update of \( \beta = (a_0, a, b) \) along the direction of the gradient given by

\[
\nabla Q_\beta = \sum_{i=1}^N \left( \frac{\partial}{\partial \beta} \log p(Z_i | \nu_i) p(Z_i | \nu) f(X_i | \nu_i) f(Y_i | \nu_i) f(\nu_i) d\nu_i \right)^\top.
\]

A step size that guarantees the increase of \( Q_\beta \) is searched using the backtracking rule [1].

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 \( Q_{\{W, W_+ \sigma^2, \Sigma_{\perp} \}} \) 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 (4) is used,

\[
\hat{W} = \text{orth}(\hat{W}).
\]

This strategy was also used in [5] for estimation of orthogonal loading matrices.

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

\[
\frac{\partial Q_{\{W, W_+ \sigma^2, \Sigma_{\perp} \}}}{\partial \Sigma_{x|t}} = \Sigma_{x|t}^{-1} - \frac{1}{N} \Sigma_{x|t}^{-1} \mathbb{E}[(X - TW^\top)^\top (X - TW^\top)] \Sigma_{x|t}^{-1}.
\]
Since $\Sigma_{xlt} = W_\perp \Sigma_{lt} W_\perp^T + \sigma_e^2 I_p$ is a full-rank matrix, by setting (12) to zero, we get the following relationship,
\[
\frac{1}{N} \mathbb{E}[(X - TW^T)^\top (X - TW^T)] = \Sigma_{xlt} = W_\perp \Sigma_{lt} W_\perp^T + \sigma_e^2 I_p.
\]
Taking trace of both sides, $\sigma_e^2$ can be estimated as
\[
\hat{\sigma}_e^2 = \frac{1}{p} \left( \frac{1}{N} \text{tr} \mathbb{E}[(X - TW^T)^\top (X - TW^T)] - \text{tr}[\Sigma_{lt}] \right).
\]
Note that $W_\perp$ and $\Sigma_{lt}$ can be obtained by eigendecomposition of the real symmetric matrices $(\frac{1}{N} \mathbb{E}[(X - TW^T)^\top (X - TW^T)] - \sigma_e^2 I_p)$. 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), the EM algorithm updates parameters in step $k$ as follows:
\[
(a_0, a, b)^{k+1} = (a_0, a, b)^k + s^{k+1} \nabla Q^{k+1}_\beta
\]
\[
W^{k+1} = \text{orth}(X^\top \mathbb{E}_k[T]E_k[T^\top T]^{-1})
\]
\[
C^{k+1} = \text{orth}(Y^\top \mathbb{E}_k[U]E_k[U^\top U]^{-1})
\]
\[
B^{k+1} = \mathbb{E}_k[U^\top T](E_k[T^\top T]^{-1} \circ I_K)
\]
\[
\Sigma_{lt}^{k+1} = \frac{1}{N} \mathbb{E}_k[T^\top T] \circ I_K
\]
\[
\Sigma_u^{k+1} = \frac{1}{N} \mathbb{E}_k[H^\top H] \circ I_K
\]
\[
(\sigma_e^2)^{k+1} = \frac{1}{p} \left( \frac{1}{N} \text{tr} \mathbb{E}_k[(X - TW^T)^\top (X - TW^T)] - \text{tr}[\Sigma_{lt}^{k+1}] \right)
\]
\[
(\sigma_f^2)^{k+1} = \frac{1}{q} \left( \frac{1}{N} \text{tr} \mathbb{E}_k[(Y - UC^\top)^\top (Y - UC^\top)] - \text{tr}[\Sigma_u^{k+1}] \right)
\]
\[
W_\perp^{k+1} = \text{eigen vectors of } \left( \frac{1}{N} \mathbb{E}_k[(X - TW^{k+1})^\top (X - TW^{k+1})] - (\sigma_e^2)^{k+1} I_p \right)
\]
\[
C_\perp^{k+1} = \text{eigen vectors of } \left( \frac{1}{N} \mathbb{E}_k[(Y - UC^{k+1})^\top (Y - UC^{k+1})] - (\sigma_f^2)^{k+1} I_q \right)
\]
\[
\Sigma_{lt}^{k+1} = \text{diag}[ \text{eigen values of } \left( \frac{1}{N} \mathbb{E}_k[(X - TW^{k+1})^\top (X - TW^{k+1})] - (\sigma_e^2)^{k+1} I_p \right)]
\]
\[
\Sigma_u^{k+1} = \text{diag}[ \text{eigen values of } \left( \frac{1}{N} \mathbb{E}_k[(Y - UC^{k+1})^\top (Y - UC^{k+1})] - (\sigma_f^2)^{k+1} I_q \right)]
\]

### S 1.3 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 $\theta = \{W, C, W_\perp, C_\perp, a, b, B, \Sigma, \Sigma_{lt}, \Sigma_u, \Sigma_e, \Sigma_f, \Sigma_h, \Sigma_g\}$, and the associated log-likelihood is given by
\[
\ell(\theta; x, y, z) = -\frac{1}{2}((p + q + 1) \log(2\pi) + \log |\Sigma_\theta| + (x, y, z)^\top \Sigma_\theta^{-1}(x, y, z)^\top).
\]
Here, the covariance matrix $\Sigma_\theta$ is
\[
\Sigma_\theta = \begin{bmatrix}
W\Sigma_t W^\top + W_\perp \Sigma_{lt} W_\perp^\top + \sigma_e^2 I_p & W\Sigma_tC^\top & W\Sigma_t a^\top \\
CW_\perp W^\top & C\Sigma_u C^\top + C_\perp \Sigma_{lt} C_\perp^\top + \sigma_f^2 I_q & C(\Sigma_h b^\top + B\Sigma_h a^\top) \\
a_\Sigma_i W^\top & (a\Sigma_i B + b\Sigma_h) C^\top & a_\Sigma_i a^\top + b_\Sigma_h b^\top + \sigma_g^2
\end{bmatrix}.
\]
We show here that under certain regularity conditions, consistency of the estimator $\hat{\theta}$ and its asymptotic distribution $\mathcal{N}(\theta, \Pi_0)$ follows from Shapiro’s Proposition 4.2 applies to the GLM-PO2PLS-Normal model. It has been shown in [3] that the proposition applies to the PO2PLS model. Similarly, we define $\tau$ as the mapping from a $\theta'$ to $\Sigma_{\theta'}$, given in (13), and the discrepancy function $F$ as $F(S; \Sigma_{\theta'}) = N \log |\Sigma_{\theta'}| + tr SS^{-1} - N \log |S| + tr SS^{-1}$, 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 $\tau$ is analytic and quadratic in $\theta'$. It follows from the definition of $F$ and the regularity of the normal log-likelihood $\ell$ that $F$ is non-negative, zero only if $S = \Sigma_{\theta'}$, and positive everywhere else. Also, $\ell$, thus $F$, is twice continuously differentiable and since GLM-PO2PLS is identifiable, $F$ has a positive definite Hessian at $\theta'$. 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,

$\mathcal{N}^{1/2}(\hat{\theta} - \theta) \longrightarrow \mathcal{N}(0, \Pi_0)$.

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

Let $\mathcal{I}(\hat{\theta})$ be the observed Fisher information matrix. To obtain $\Pi_\alpha$, 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 $\hat{\alpha}_k^2$ is non-random, $\mathcal{I}^{-1}(\hat{\alpha})$ can be calculated, and be used to approximate $\Pi_\alpha$. The $2K \times 2K$ observed Fisher information matrix $\mathcal{I}(\hat{\alpha})$ is given by

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

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

(14)

Here, $S_i(\hat{\alpha}) = \frac{1}{\sigma^2} ((T_i, H_i)^\top Z_i - (T_i, H_i)^\top (T_i, H_i)^\alpha^\top)$, and $B_i(\hat{\alpha}) = \frac{1}{\sigma^2} (T_i, H_i)^\top (T_i, H_i)$. Note that (14) 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,

$$
\mathcal{I}(\hat{\alpha}) = \frac{1}{(\sigma^2)^2} \sum_{i=1}^{N} \left\{ \sigma^2 \mathbb{E}V_i - Z_i^2 V_i + Z_i (\mu_i^\top \hat{\alpha} V_i + (\mu_i^\top \hat{\alpha} V_i)^\top + \hat{\alpha} \mu_i^\top (V_i - 2\mu_i^\top \mu_i)) 
$$

$$
+ Z_i (\mu_i^\top \hat{\alpha} V_i + (\mu_i^\top \hat{\alpha} V_i)^\top + \hat{\alpha} \mu_i^\top (V_i - 2\mu_i^\top \mu_i)) ^\top 
$$

$$
- 2 \mathbb{E} \hat{\alpha}^\top \hat{\alpha} V_i - \mu_i \hat{\alpha}^\top \hat{\alpha} \mu_i (V_i - 2\mu_i^\top \mu_i) - \text{tr}[\hat{\alpha}^\top \hat{\alpha} (V_i - \mu_i^\top \mu_i)] V_i 
$$

$$
- \sum_{j=1, j \neq i}^{N} (Z_i \mu_j^\top - V_i \hat{\alpha}^\top)(Z_j \mu_i^\top - V_j \hat{\alpha}^\top) ^\top \right\}
$$

**S 2 Simulation**

In this section, we first show the simulation results omitted in the main article. We then describe the additional simulations and show the corresponding results.
S 2.1 Simulation results of the main article

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, type I error and outcome prediction in low-dimensional settings, and the results of loading estimation and feature selection in both high and low dimensions.

Figure S1 depicts the results of coefficient estimation in low-dimensional settings. Overall, the performance was similar in low-dimensional and high-dimensional settings. In the scenarios with small sample size and high noise, the errors were larger than those in the high-dimensional settings.

Figure S2 shows the type I error in low-dimensional settings. Overall, they were higher than those in the high-dimensional settings. In scenarios with small sample size and high noise level, the type I errors were above 10%. The type I errors decreased below 6% when the sample size was increased to 1000, and moved closer to 5% when the sample size further increased to 10000.

Figure S3 depicts the results of outcome prediction in low-dimensional settings. The performance was similar to that in the high-dimensional settings. The conclusions in high dimensional

Figure S1: Performance of coefficient estimation in low-dimensional settings. The y-axis shows the scaled estimation error. Boxes show the results of 500 repetitions.

Figure S2: Type I error in low-dimensional settings. The error bars show the standard errors of the estimation. The dotted horizontal line is the significance level of 5%.

Figure S3: Performance of outcome prediction in low-dimensional settings. The conclusions in high dimensional
settings in the main article also hold for low dimensional settings.

Figure S3: **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.

Figure S4 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 α for the x-joint components in the linear predictor of z tended to be underestimated.

Figure S4: **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.

Figure S5 shows the results regarding performance of feature selection. From Figure S5a 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 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 S5b for binary outcome to Figure S5a, GLM-PO2PLS performed similarly well, while the performance of ridge-x improved. In low dimensions (Figure S5c and Figure S5d), 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.

Figure S5: 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.

### S 2.2 Additional simulations

We perform two additional simulations to evaluate the estimation accuracy of \( a \) and \( b \) with multiple components, and the power of the chi-square test using parameters estimated from the Down syndrome dataset. We consider here only high dimensional omics (\( p = 2000 \), \( q = 25 \)) with a continuous outcome.

#### S 2.2.1 Coefficient estimation with multiple components

To investigate the coefficient estimation performance with multiple components, we set the number of joint, x-specific, and y-specific components to 2. The coefficient vectors \( a = (a_1, a_2) \) and \( b = (b_1, b_2) \) are set to \((2, 2)\) and \((1, 1)\), respectively. The other parameters are kept the same.
as specified in Table 1 of the main manuscript. The errors $\hat{a}_k - a_k$ and $\hat{b}_k - b_k$ are standardized by $a_k$ and $b_k$ (where $k = 1, 2$) to exclude the influence of the parameter scale.

Figure S6 shows the scaled error of each coefficient. The errors of the first pair $(a_1, b_1)$ were similar to the errors of the second pair $(a_2, b_2)$. Compared to the results from a model with 1 pair of joint components, the estimation accuracy was similar.

Figure S6: Performance of coefficient estimation with multiple components. The y-axis shows the scaled estimation error. 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.

### S 2.2.2 Power in settings estimated from dataset

We investigate the empirical power under settings estimated from the Down syndrome data. The noise levels are 80% in $x$, 20% in $y$, and 75% in $z$. The heterogeneity level (proportion of $h$ in $u$) is 90%. The ratio of the explained variance by the x-specific part and x-joint part (i.e., $\sigma^2_{x_i}/\sigma^2_t$) is 2, and the ratio of y-specific part to y-joint ($\sigma^2_{u_i}/\sigma^2_u$) is 0.1. The coefficient $a$ and $b$ are set to 0.05 and 0.1, respectively. The matrix $B$ is $0.04 \times I$. We investigate two sample sizes (100 and 1000), and four additional noise levels in $z$ (60%, 90%, 95%, and 99%). 500 replications are simulated.

Figure S7 shows the empirical power under different noise levels in the outcome. Under a small sample size, the empirical power was 0.998 at the noise level of 75%, which was estimated from the Down syndrome data. The empirical power remained high at 0.862 with a higher noise level of 90%. With a large sample size of 1000, the empirical power was 1 with all the noise levels below 95% and dropped to 0.814 under an extreme noise level of 99%.
Figure S7: **Empirical power.** The noise proportions in the outcome $z$ is shown on the x-axis, among which 75% is estimated from the DS dataset.

**References**

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

[2] Said el Bouhaddani, Hae Won Uh, Caroline Hayward, Geurt Jongbloed, and Jeanine Houwing-Duistermaat. Probabilistic partial least squares model: Identifiability, estimation and application. *Journal of Multivariate Analysis*, 167:331–346, 2018.

[3] Said el Bouhaddani, Hae-Won Uh, Geurt Jongbloed, and Jeanine Houwing-Duistermaat. Statistical Integration of Heterogeneous Data with PO2PLS. mar 2021.

[4] David A. Harville. *Matrix Algebra From a Statistician’s Perspective*. Springer New York, 1997.

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

[6] A Woodbury Max. *Inverting modified matrices*. Princeton Univ., 1950.

[7] Xiao Li Meng and Donald B Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, jun 1993.