Statistical Modeling of Omics Data Using Two-Stage-PO2PLS

He Li[1], Zhujie Gu[1,3], Said el Bouhaddani[2], Jeanine Houwing-Duistermaat[1,4]

[1]Dept. of Mathematics, Radboud University, Nijmegen, NL
[2]Dept. of Data Science and Biostatistics, Julius Centre, UMC Utrecht, NL
[3]Medical Research Council Biostatistics Unit, University of Cambridge, UK
[4]Dept. of Statistics, University of Leeds, UK

CNC23, August 2023

Radboud Universiteit

Background

- Consider the following datasets:
  - X: $N \times p$, high dimensional, easy to measure.
  - Y: $N \times q$, not easy to measure.
  - Z: Outcome, single variable.
- Question: Model relationship between Z and Y (through X).
- Two-stage approaches.
  - Stage 1: Obtain the summary of Y in terms of X.
  - Stage 2: Fit a model for Z with the summary.
- Challenges: For stage 1, several methods exist, but not clear which performs better.

Our Aim

- Comparison of three methods via simulations:
  - Univariate.
  - Algorithmic (O2PLS). } Multivariate
  - Likelihood (PO2PLS). }
- Two simulation models.

Univariate Method

(1) Stage 1: Lasso regression for each column j of Y.

$$\hat{\beta}_j^{Lasso} = \arg\min_{\beta_j} \sum_{i=1}^{N} \|Y_{ij} - \sum_{k=1}^{p} X_{ik}\beta_{kj}\|^2 + \lambda \sum_{k=1}^{p} \|\beta_{kj}\|,$$

$$\hat{Y}_j = X\hat{\beta}_j^{Lasso}.$$

- $\hat{Y}_j$ goes to stage 2.

(2) Stage 2: Ridge regression between Z and $\hat{Y}$.

$$\hat{\gamma}^{Ridge} = \arg\min_{\gamma} \sum_{i=1}^{N} \|Z_i - \sum_{k=1}^{q} \hat{Y}_{ik}\gamma_k\|^2 + \lambda \sum_{k=1}^{q} \|\gamma_k\|^2,$$

$$\hat{Z} = \hat{Y}\hat{\gamma}^{Ridge}.$$

O2PLS

(1) Stage 1: Model for random vectors x and y.

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

- x, y are decomposed into joint (t, u, dim=r), specific ($t_\perp$, dim=r_x; $u_\perp$, dim=r_y), residual (e, f) parts.
- W ($W_\perp$), C ($C_\perp$) are the loading matrices for the joint (specific) parts of x, y respectively.
- B is diagonal.
- t is the summary for y based on x, $\hat{T}$ goes to stage 2.

(2) Stage 2: Model for Z.

$$\hat{Z} = a_0 + \hat{T}a^\top.$$

PO2PLS

(1) Stage 1: Model for x and y.

- Assume the latent variables are normally distributed.
- Estimate the parameters by maximum likelihood.

(2) Stage 2: Model for Z.

$$\hat{Z} = a_0 + \hat{T}a^\top.$$

Generate Data from PO2PLS

- Generate data as follows:

$$x = tW^\top + t_\perp W_\perp^\top + e$$
$$y = uC^\top + u_\perp C_\perp^\top + f$$
$$u = tB + h$$
$$z = a_0 + ta^\top + g$$

Step 1: Generate t, $t_\perp$, $u_\perp$.

Step 2: Generate e, h, g by given the proportion.

Step 3: Generate x, u, z.

Step 4: Generate y by given the proportion of f.

Generate Data from PCA

- Generate y and z as follows:

$$y = tC^\top + f$$
$$z = tC^\top \ell + g$$

where t is the score of x, and fixed (once generated), x is a fixed vector.

Step 1: Generate f, g by given the proportion.

Step 2: Generate y, z.

Simulation Procedure

- Two simulation models.
- Split each dataset into training and testing sets.
- Predict z^{test} based on the three methods.
- Evaluate the performance by

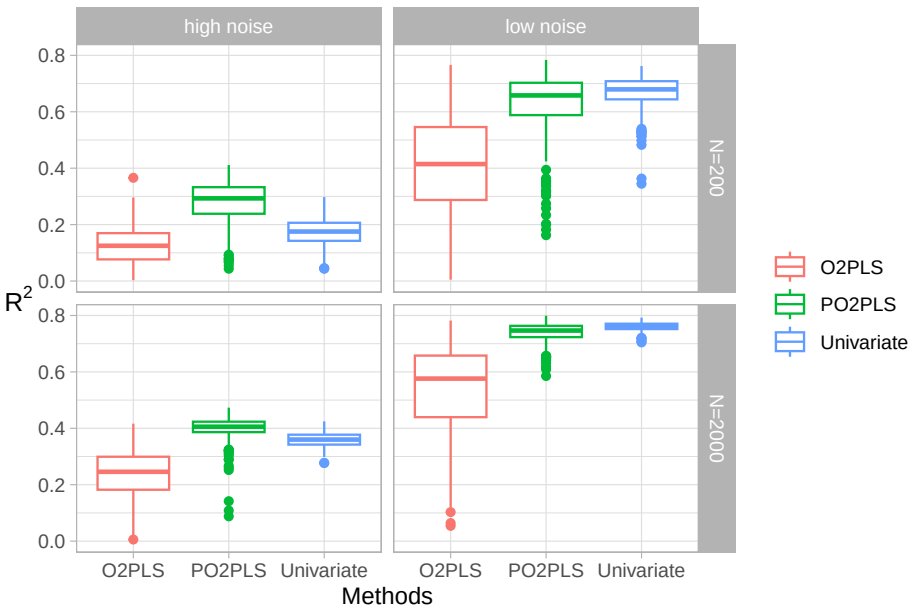
$$R^2 = 1 - \frac{\sum(z_i^{test} - \hat{z}^{test})^2}{\sum(z_i^{test} - \bar{z}^{test})^2}.$$

where $\bar{z}^{test}$ is the average value of testing set.

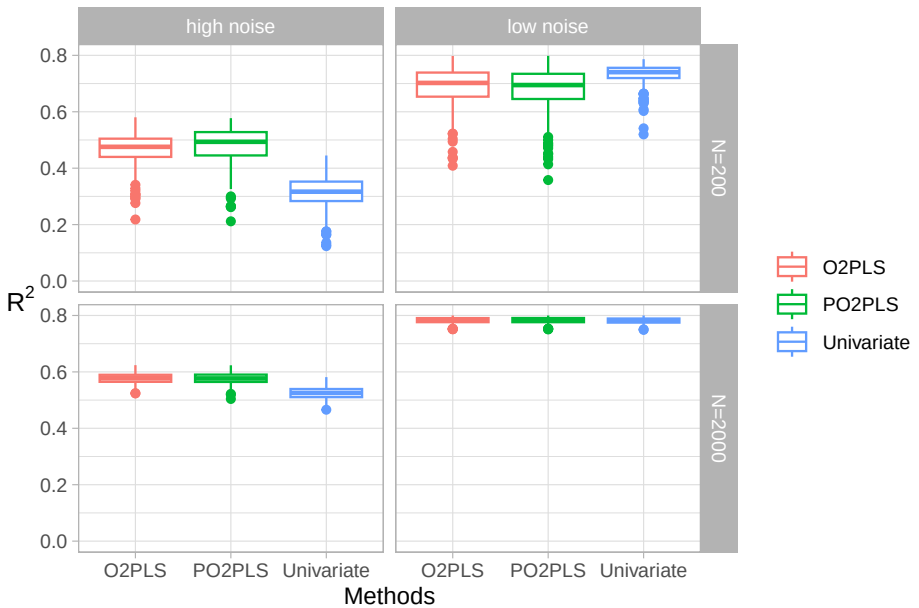
Parameters Settings

- Components of joint parts:
  - PO2PLS-based: $r = r_x = r_y = 5$,
  - PCA-based: $r = 5, r_x = 1, r_y = 0$.
- Sample size:
  - training sets: N=200, 2000,
  - testing sets: 1000.
- Dimension:
  - low dimension: $p = 100, q = 10$,
  - high dimension: $p = 2000, q = 25$.
- Proportion of residuals in x and y:
  - low noise: $(\alpha_x, \alpha_y) = (0.4, 0.4)$,
  - high noise: $(\alpha_x, \alpha_y) = (0.95, 0.05)$.
- Proportion of residuals in u:
  - small heterogeneity: $\alpha_{tu} = 0.4$,
  - large heterogeneity: $\alpha_{tu} = 0.8$.

PO2PLS-Based Simulation: Results



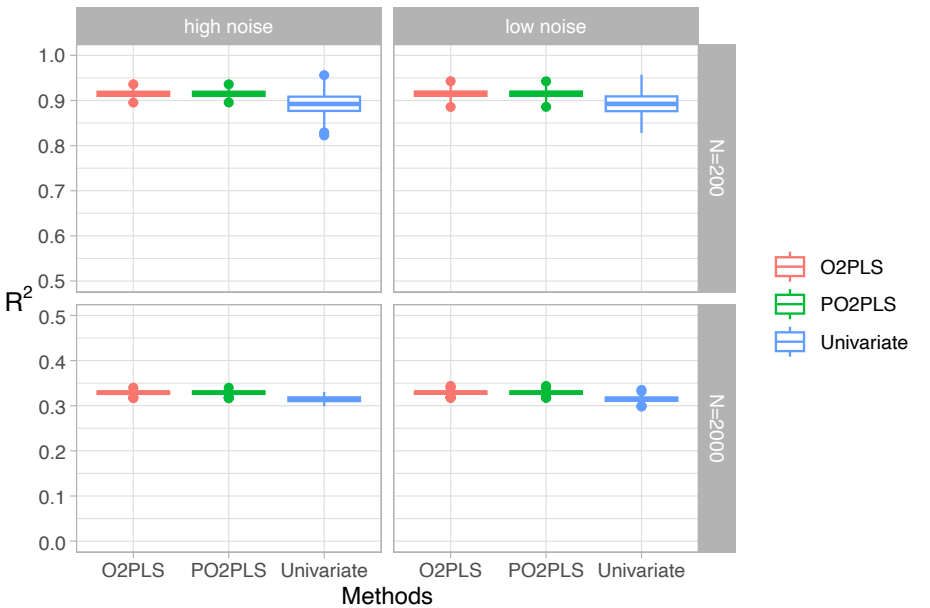
(a) Low dimension, $\alpha_{tu} = 0.8$



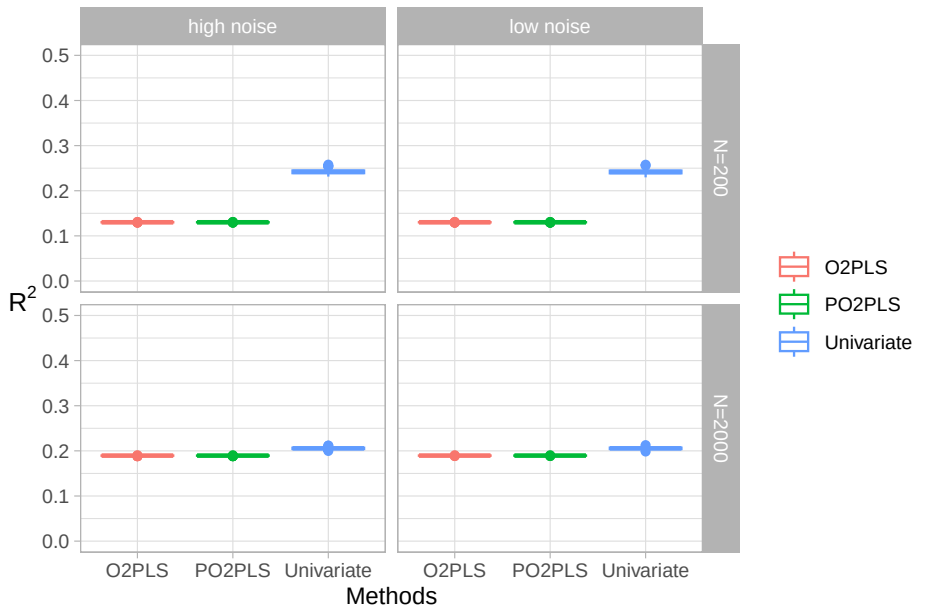
(b) High dimension, $\alpha_{tu} = 0.8$

$\alpha_{tu} = 0.4$ performs similar with $\alpha_{tu} = 0.8$.

PCA-Based Simulation: Results



(c) Low dimension, $\alpha_{tu} = 0.8$



(d) High dimension, $\alpha_{tu} = 0.8$

$\alpha_{tu} = 0.4$ performs similar with $\alpha_{tu} = 0.8$.

Simulation: Conclusion

- All methods perform well under heterogeneity.
- PO2PLS performs better in PO2PLS-based simulation, especially under low dimension: $(N > p)$.
- Univariate method performs better in PCA-based simulation, especially for high dimension $(p > N)$.

ORCADES Analysis

Dataset from Orkney complex disease study (ORCADES)

- X: 1523 $\times$ 2402, selected SNPs from GWAS.
- Y: 1523 $\times$ 87, metabolomics.
- Z: 1523 $\times$ 1, BMI.
- 1000 individuals for training, 523 for testing.

Table 1: Explained variance of BMI

	training	testing
2-stage-Univariate	3.71	2.31
2-stage-O2PLS	1.64	1.07
2-stage-PO2PLS	0	0.67

Conclusions

- Simulations show that PO2PLS performs well when x is random.
- Univariate method performs better in data analysis, which seems to agree with the PCA-based simulation, namely under the high dimensional condition.
- Future work: Integration of two omics datasets in longitudinal studies.

Reference


Cook RD (2022).
A slice of multivariate dimension reduction.
Journal of Multivariate Analysis, 188: 104812.


El Bouhaddani S, Uh HW 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, 71(5), 1451–1470.


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


Trygg J, Wold S (2002).
Orthogonal projections to latent structures (O-PLS).
Journal of Chemometrics, 16(3), 119–128.

Thank you!

He Li

Email: he.li@math.ru.nl