Probabilistic PLS method for statistical integration of omics data (PO2PLS)

Said el Bouhaddani[1], Hae-Won Uh[1], Jeanine Houwing-Duistermaat[2,3]

[1] Dept. of Biostatistics and Research support, UMC Utrecht, The Netherlands

[2] Dept. of Statistics, University of Leeds, Leeds, The UK

[3] Dept. of Statistical Sciences, University of Bologna, Bologna, Italy

E-mail for correspondence: s.elbouhaddani@umcutrecht.nl

Abstract: Nowadays, data are collected on several biological levels, e.g., genomics, transcriptomics, epigenetics. We focus on the joint analysis or "integration" of these datasets. Challenges for data integration are high dimensionality, strong correlations, and heterogeneity across omics datasets (due to differing biological levels and measurement platforms). Several methods address parts of these challenges and are popular for data integration. However, they do not provide statistical evidence for a relation between the datasets.

We propose PO2PLS, a probabilistic latent variable framework for the relation between two datasets. The PO2PLS model includes joint and specific components that are linear combinations of the original variables. PO2PLS reduces dimensionality, captures correlations and addresses heterogeneity. For estimation, we develop a memory-efficient EM algorithm, and we show that the estimator is consistent and asymptotically normal, even for high dimensional data. We propose a "global" test for the relation and derive its asymptotic distribution.

We evaluate the estimation and testing performance of PO2PLS with simulations. We illustrate the PO2PLS inference framework with two motivating studies: a population cohort with genetics and glycomics data, and a case-control cohort on hypertrophic cardiomyopathy with epigenetics and transcriptomics data. This demonstrates the potential of PO2PLS as a statistical framework in data integration.

Key words: joint principal components; omics data; inference; high dimensionality; PLS

el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., & Uh, H.-W. (2016). Evaluation of O2PLS in Omics data integration. BMC Bioinformatics, 17(S2), S11.

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