Part 1: overview

- Principal Component Analysis (PCA)
- Partial Least Squares (PLS)
- Two-way Orthogonal PLS (O2PLS)
  - Population model
  - Multi-omics data integration
- Post-hoc analyses using external databases

Introduction

- We have seen PCA and PLS
- Looked at them from an algorithmic point of view
- What is the underlying population model?
    - What are the random variables, and what are the parameters?

Population model for PCA

$$x = tW^{\top} + e$$

- t are scores (the PCs)
- W are weights (or loadings), parameters
- e are residuals
- Parameters are estimated such that the variance of Xw is maximized

Population model for PLS

$$x = tW^\top + e$$
$$y = uC^\top + f$$
$$u = tB + h$$

- t and u are latent (hidden) scores (the joint PCs)
- w and c are weights (or loadings), parameters
- The relation between x and y is fully captured by t and u, via the third relation
- Parameters are estimated such that the covariance between Xw and Yc is maximized

Interpretation of the model

$$x = tW^\top + e$$
$$y = uC^\top + f$$
$$u = tB + h$$

- x and y are two datasets, say genetic and metabolomic data
- t and u are latent variables underlying these data
- These latent variables vary, and through W and C cause variation in the two datasets
- In the above context, t and u could be methylation/glycomic pathways
- Then, W and C would tell us which CpGs and glycans are involved

Omics-specific variation

- PCA models the variance of X (or Y)
- PLS models the covariance
- Suppose both sources are present, independently
  - E.g. some pathways connect CpGs and glycans
  - Other pathways are there for self-maintenance
- Capture both parts at the same time
- Need to extend the PLS model

O2PLS model and data-specific parts

$$
\begin{gathered}
x = tW^\top + t_s W_s^\top + e \\
y = uC^\top + u_s C_s^\top + f \\
u = tB + h
\end{gathered}
$$

- x and y are two datasets, say genetic and metabolomic data
- In addition to the PLS joint weights and scores, we have specific weights and scores
- The relation between x and y is still fully captured by t and u

Estimating the O2PLS components

- Three step estimation
  - First estimate several PLS components, they will contain both joint and specific parts
  - From that, estimate specific parts only and subtract
  - Finally, estimate again PLS on the “corrected” data
- Implemented and on CRAN: *OmicsPLS*
  - Obtain loadings with: *loadings*
  - Obtain scores with: *scores*

Number of components

- Until now, we did not mention how to find out the number of components needed
- For PCA, this number is the number of PCs
- For PLS, this is the number of joint PCs
- For O2PLS, this is
  - The number of joint PCs
  - And the number of X-specific components
  - And Y-specific components
- Standard way is to do cross-validation (not covered now)

O2PLS: summary

- For given datasets X and Y, we want to inspect their relation
- Want to capture data-specific parts as well
- O2PLS estimates joint and specific parts, consisting of weights and scores
- One can interpret or plot the weights and scores to understand which features/samples are most important

Exercises

- Load the Down Syndrome data and run PCA to obtain the first principal component of the glycans
  - *svd(glycomics, nu=0, nv=1)$v*
- Now run *o2m* using both methylation and glycomics data and obtain the first joint principal component
  - *fit <- o2m(methylation, glycomics, 4, 2, 6)*
  - *loadings(fit, "Yj", subset=1)*
- Compare the two loading vectors

Part 1: overview

- Principal Component Analysis (PCA)
- Partial Least Squares (PLS)
- Two-way Orthogonal PLS (O2PLS)
  - Population model
  - Multi-omics data integration
- Post-hoc analyses using external databases