Omics data integration with the OmicsPLS R-package

Said el Bouhaddani

August 2017

Learning objectives

- Decomposition of variation
- Covariation/correlation
- Loadings and scores
- Fit and interpret OmicsPLS output

O2PLS Method

- Trygg & Wold, 2003
- Decomposition:

$$X = TW^\top + T_\perp W_\perp^\top + E$$

$$Y = UC^\top + U_\perp C_\perp^\top + F$$

- Joint part: $U = TB + H$
- Find W and C such that T and U have high covariance.
- W and C corrected for independent latent variation specific for X and Y.
- n joint components, nx X-specific components, ny Y-specific components

OmicsPLS R package: Overview

- Input data `X` and `Y`, rows correspond to **the same** subjects
- Number of components `n`, `nx` and `ny`.
  - Main fitting function `o2m(X, Y, n, nx, ny)`
  - *Simultaneous* estimation of all components per part
  - Stripped version is also present: `stripped = TRUE`
  - Automatic switching to high dimensional mode with `p_thresh = 3000`
  - Output: list of class `o2m`
- `print/plot/summary/predict/loadings`: see `help("___.o2m", "OmicsPLS")`
- For a complete overview: `?OmicsPLS`

Data to be analyzed

- DILGOM population study
- Gene expression ($p = 6272$)
- Metabolites ($q = 137$)
- $N = 191$ participants

Data analysis

- Install from cran:
- `install.packages("OmicsPLS")`
- `library(OmicsPLS)`

OmicsPLS workflow I: choosing number of components

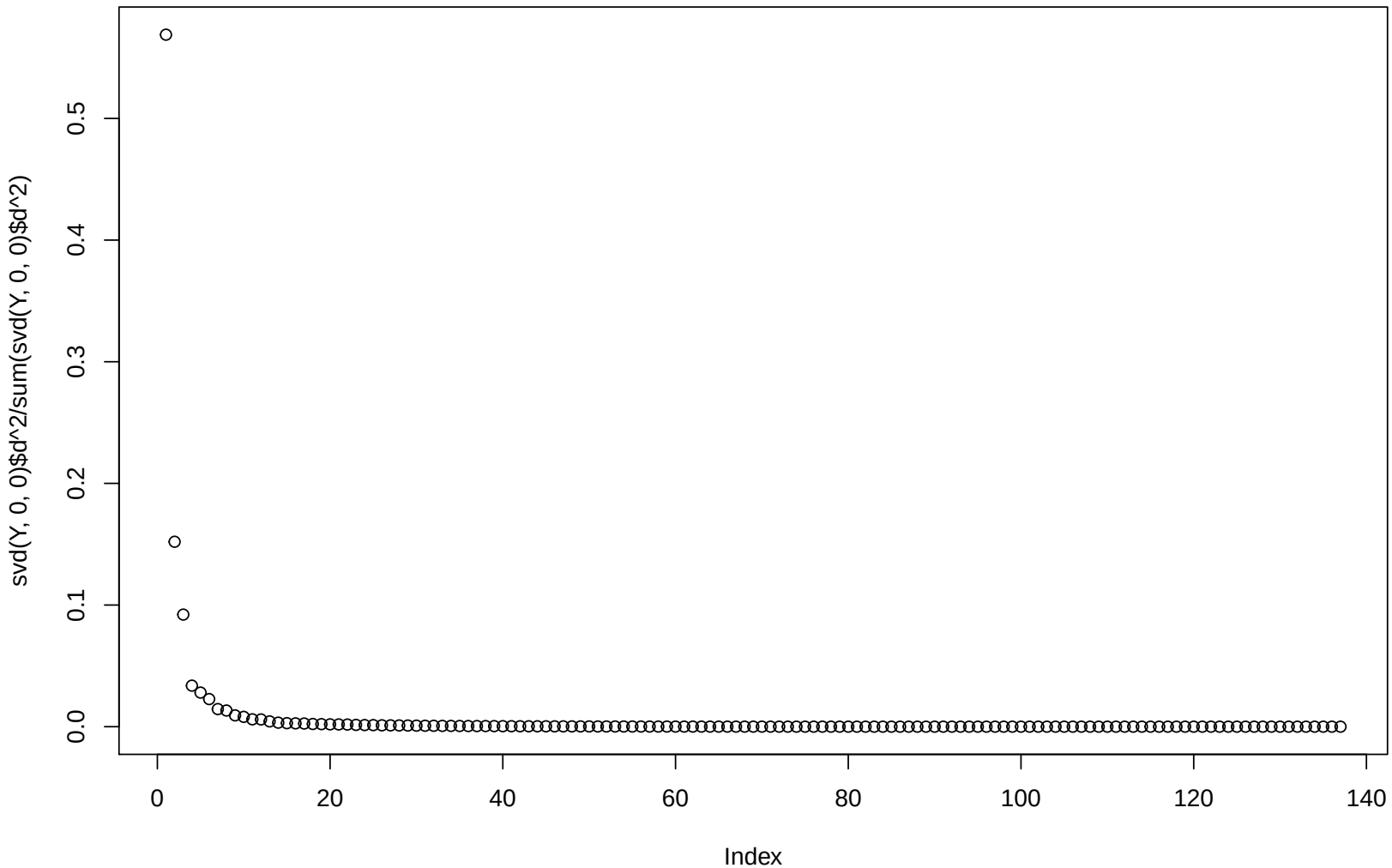
- Recall: `o2m(X, Y, n, nx, ny)`
- First need to know how many components needed
- Two popular approaches
  - Cross-validation
  - Eigenvalue plot
- First approach useful for prediction
- Second approach useful for exploration

Eigenvalue plots

- Consider metabolite data `Y`
- Eigenvalues are given by `svd(Y, 0, 0)$d^2`
- Often useful to plot relative contribution of each eigenvalue

Eigenvalue plots

```
plot(svd(Y, 0, 0)$d^2 / sum(svd(Y, 0, 0)$d^2))
```



Cross-validation

- Idea:
  - Choose which numbers of components to consider
  - Set aside part of data for testing
  - Fit O2PLS on rest of data
  - Calculate prediction error on test data
  - Repeat for other numbers of components
  - Choose the best ones

Cross-validation

```
crossval_o2m(X, Y, 1:3, 0, 0,
 nr_folds = 2,
 nr_cores = parallel::detectCores())
```

```
> **********
> Elapsed time: 11.22 sec
> *******
> Minimal 2-CV error is at ax=0 ay=0 a=2
> *******
> Minimum is 0.610606
> **********
```

OmicsPLS workflow II: fitting

- Main function (again) `o2m(X, Y, n, nx, ny)`
- See help file `?o2m`
- Stripped version `stripped = TRUE`
- High dimensional version `p_thresh` and `q_thresh`

Example

- Fit O2PLS: X = RNA, Y = Metabolites
- Low dimensional mode, since $q < 3000$

```
fit <- o2m(X, Y, n = 1, nx = 8, ny = 1)
```

Inspecting the results

```
fit
```

```
> O2PLS fit
> with 1 joint components
> and  8 orthogonal components in X
> and  1 orthogonal components in Y
> Elapsed time: 3.13 sec
```

Inspecting the results

```
fit
```

```
> O2PLS fit
> with 1 joint components
> and  8 orthogonal components in X
> and  1 orthogonal components in Y
> Elapsed time: 3.13 sec
```

Some timings (i5 laptop) with `stripped = TRUE`

Problem size	Timing Low D	Timing High D
1000 vars	3 sec	15 sec
5000 vars	300 sec	130 sec

OmicsPLS workflow III: Summarizing

```
summary(fit)
```

```
*** Summary of the O2PLS fit ***
-  Call: o2m(X = X, Y = Y, n = 1, nx = 8, ny = 1)

-  Modeled variation
-- Total variation:
in X: 116016.8
in Y: 2516.821

-- Joint, Orthogonal and Noise as proportions:

            data X data Y
Joint        0.013  0.522
Orthogonal   0.490  0.069
Noise        0.497  0.410
[TRUNCATED...]
```

Questions

```
-- Joint, Orthogonal and Noise as proportions:

           data X data Y
Joint       0.013  0.522
```

- How much of X is explained by Joint?
- How much of X is explained by Y?

Questions

```
-- Joint, Orthogonal and Noise as proportions:

           data X data Y
Joint       0.013  0.522
```

- How much of X is explained by Joint?
- How much of X is explained by Y?

```
-- Predictable variation in Y-joint part by X-joint part:
Variation in Yhat relative to U: 0.671
-- Predictable variation in X-joint part by Y-joint part:
Variation in Xhat relative to T: 0.671
```

OmicsPLS workflow IV: Plotting

- Plot the loadings: `plot(fit, loading_name, i, j, use_ggplot2, label, ...)`
- Returns ggplot2 object
- In the `...` you can use all kinds of ggplot2 commands (col, alpha, size)
- You can add layers to the plot, or customize theme

Some fancy stuff

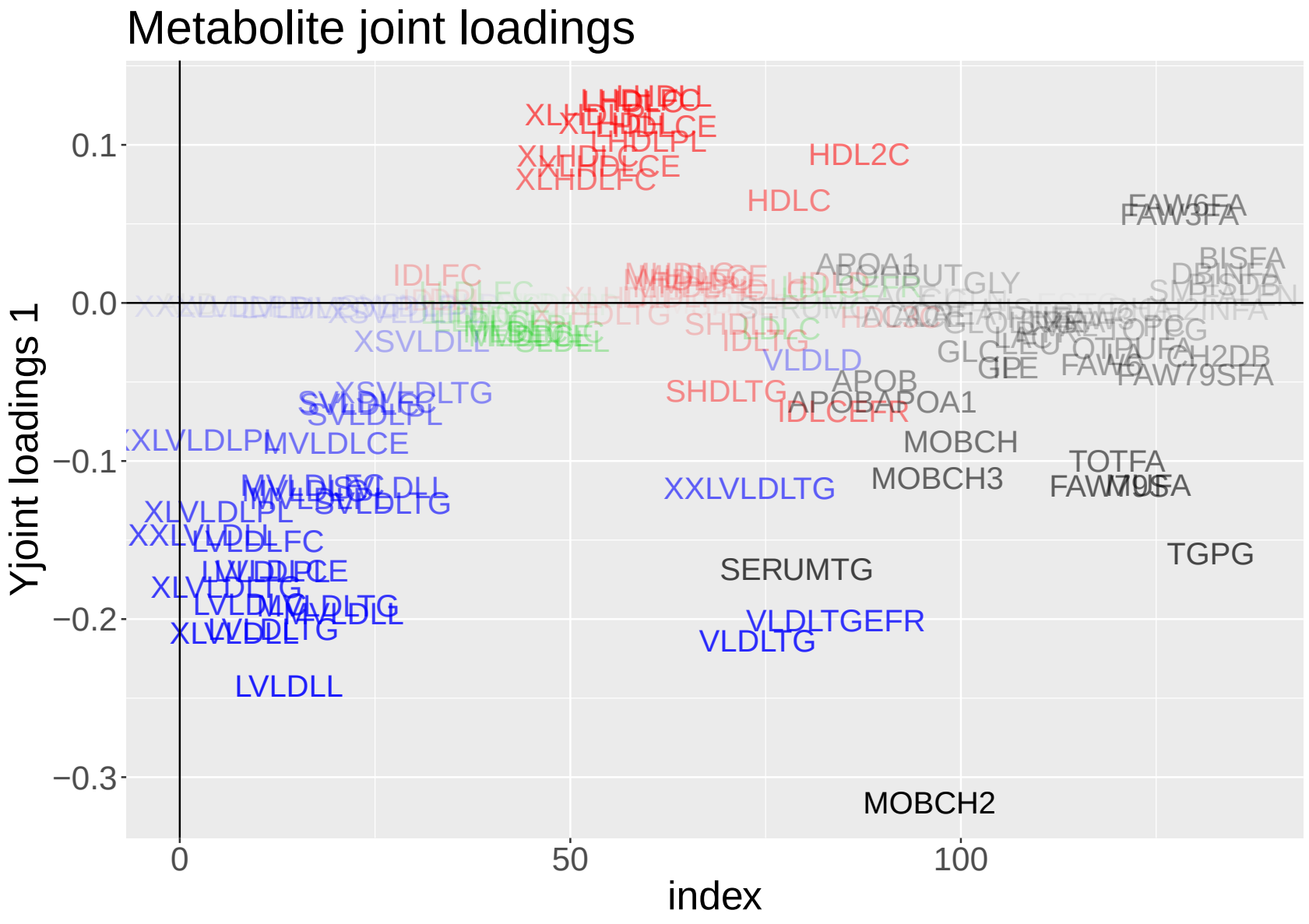
- Example:

```
alp <- loadings(fit, "Yjoint", 1) %>% abs %>% sqrt
```

- `cols` contains labeling {VLDL, LDL, HDL, other}

```
plot(fit, "Yj", i=1, label = "colnames",
     size = 6, alpha = alp/max(alp), col = cols) +
  theme(text = element_text(size = 22)) +
  ggtitle("Metabolite joint loadings")
```

Previous code results in:



Summary

- OmicsPLS package for omics data analysis
- Install via CRAN: `install.packages("OmicsPLS")`
- Overview of Package: `?OmicsPLS`
- Main function: `o2m`, see also `?o2m`

Remarks

- Acknowledgments
  - Jeanine Houwing-Duistermaat, LUMC, Leeds
  - Hae-Won Uh, LUMC
  - Geurt Jongbloed, TU Delft
  - Szymon Kielbasa, LUMC
- Please cite if you use it:

```
citation("OmicsPLS")
```