Models Can Lie

An Illustration from Heart Rate Variablity Data

Raju Rimal and Veronika Lindberg

2015/12/09

Norwegian University of Science and Technology
Norwegian University of Life Sciences

NTNU

Norges miljø- og biovitenskapelige universitet

Overview

1. Background

2. How data looks like

3. Classification with series stack

4. Classification with series averaged over Series repetition

5. Classification with series averaged over Person-Event Combination

6. Some Comparison

Background

Some Background

1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis

Some Background

1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis
2. CPPLS integrate CCA with PLS to select relevant variables for response

Some Background

1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis
2. CPPLS integrate CCA with PLS to select relevant variables for response
3. Cross-validation is performed over the observations on each a) frequency window b) series c) person-event combination

Some Background

1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis
2. CPPLS integrate CCA with PLS to select relevant variables for response
3. Cross-validation is performed over the observations on each a) frequency window b) series c) person-event combination
4. Three variation of dataset is used

Some Background

1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis
2. CPPLS integrate CCA with PLS to select relevant variables for response
3. Cross-validation is performed over the observations on each a) frequency window b) series c) person-event combination
4. Three variation of dataset is used
    - Transpose of each frequency windows stacked together

Some Background

1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis
2. CPPLS integrate CCA with PLS to select relevant variables for response
3. Cross-validation is performed over the observations on each a) frequency window b) series c) person-event combination
4. Three variation of dataset is used
    - Transpose of each frequency windows stacked together
    - The average frequencies over time for each Series

Some Background

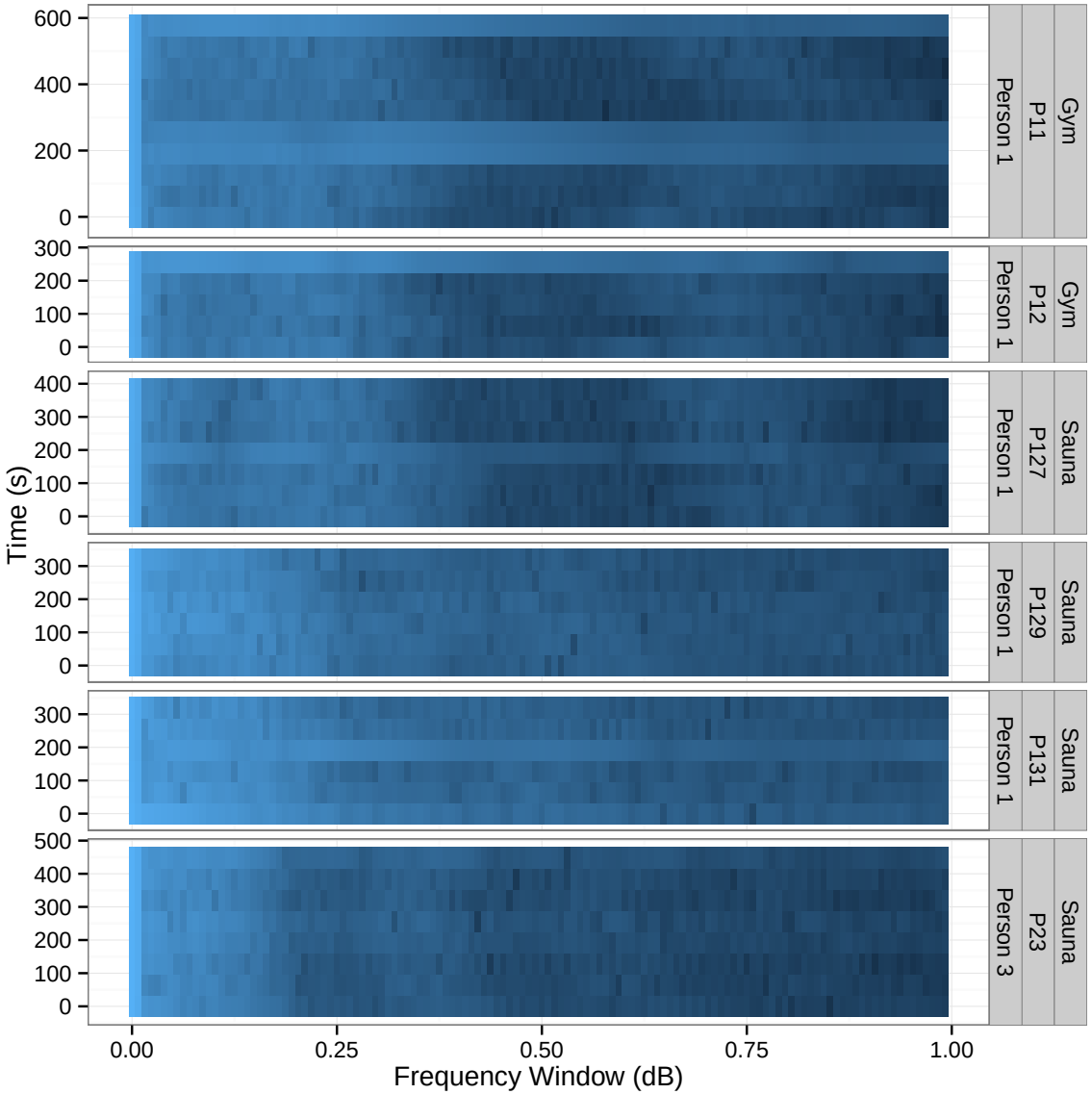
1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis
2. CPPLS integrate CCA with PLS to select relevant variables for response
3. Cross-validation is performed over the observations on each a) frequency window b) series c) person-event combination
4. Three variation of dataset is used
    - Transpose of each frequency windows stacked together
    - The average frequencies over time for each Series
    - The average frequencies over time for each person-event combination

Some Background

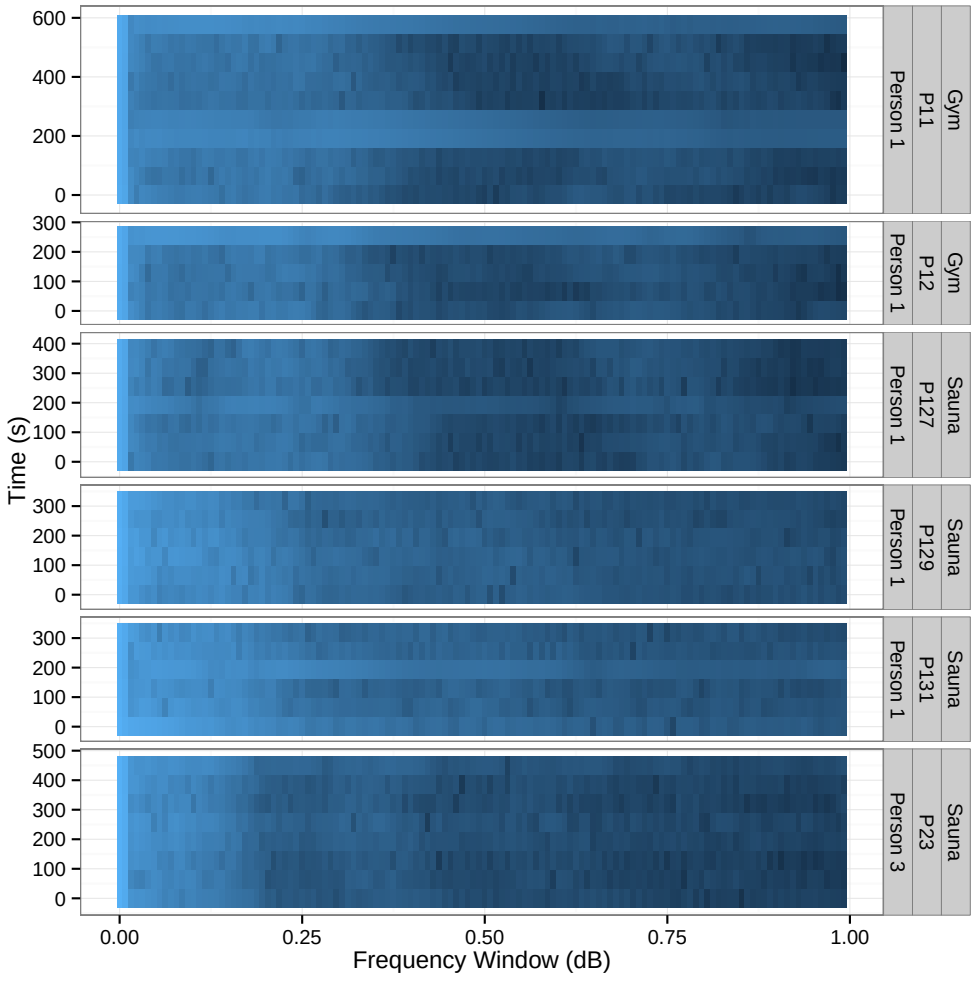
1. PCR, PLS and Canonically Powered PLS (CPPLS) is used in the analysis
2. CPPLS integrate CCA with PLS to select relevant variables for response
3. Cross-validation is performed over the observations on each a) frequency window b) series c) person-event combination
4. Three variation of dataset is used
    - Transpose of each frequency windows stacked together
    - The average frequencies over time for each Series
    - The average frequencies over time for each person-event combination
5. LDA model is used for discriminant analysis using the scores obtained form each of the latent variable model with cross-validation implemented

How data looks like

How data looks like

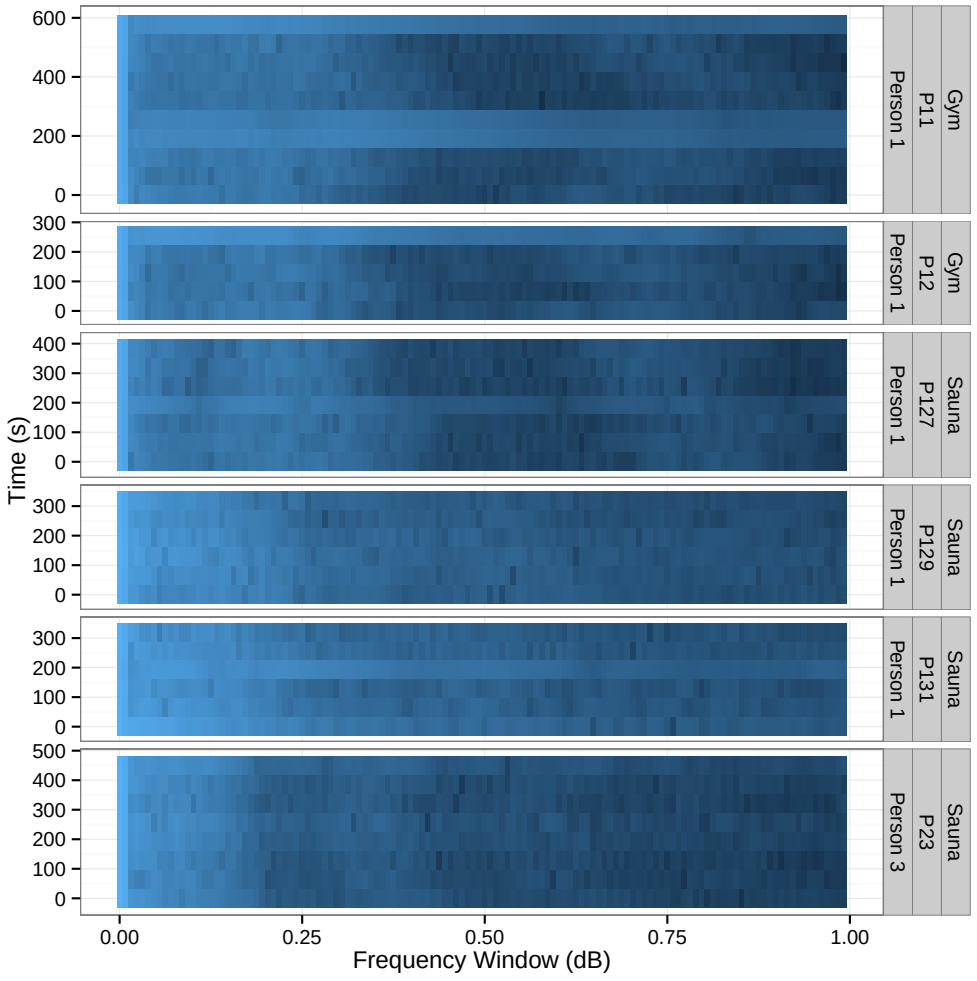


How data looks like



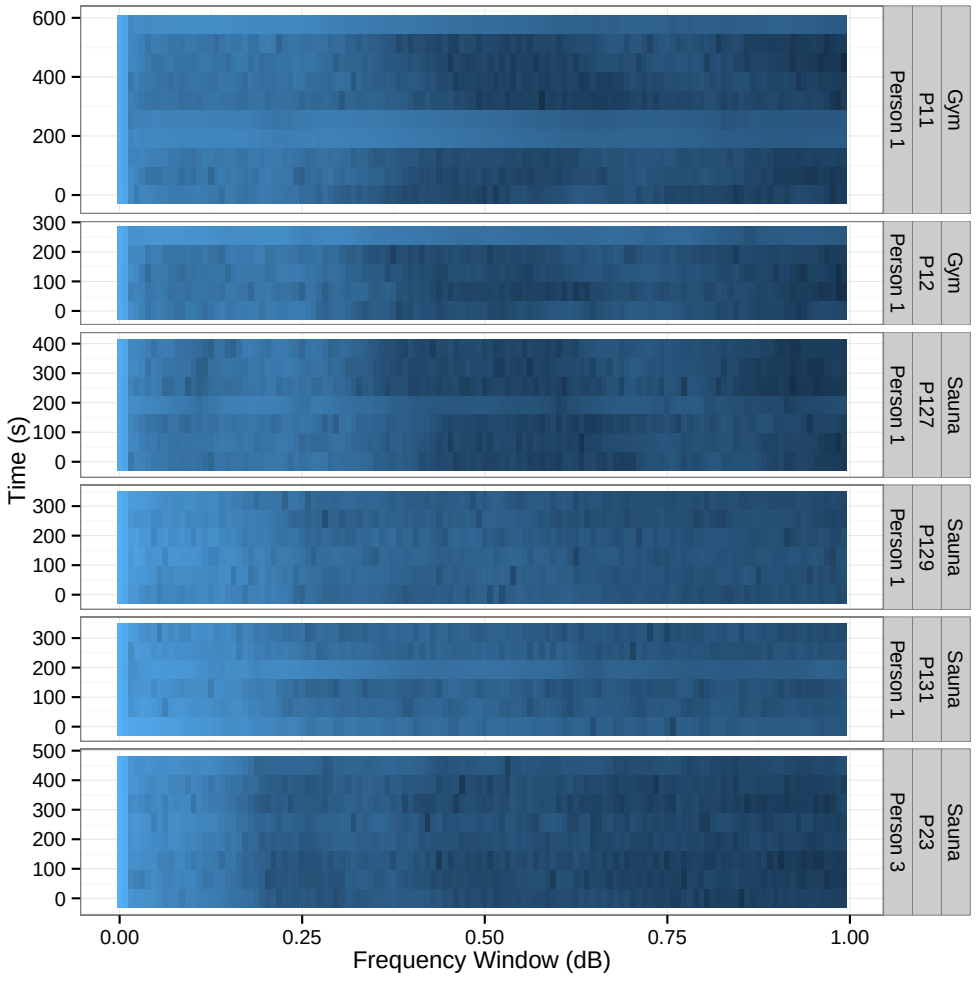
- Each block represent a series divided into several windows (rows), 128 columns each with 16 overlaps. The cell contains the frequency values obtained from fast fourier transform

How data looks like



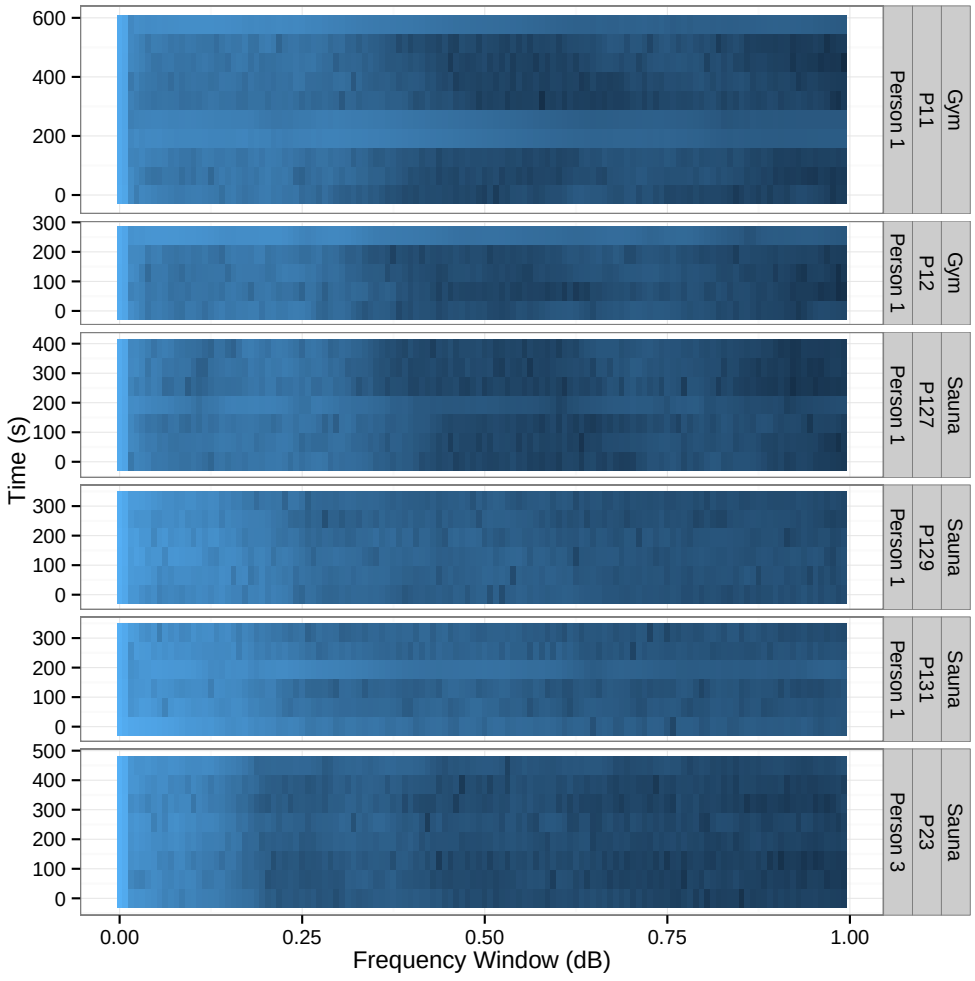
- Each person may have involved into multiple activities which may have replications

How data looks like



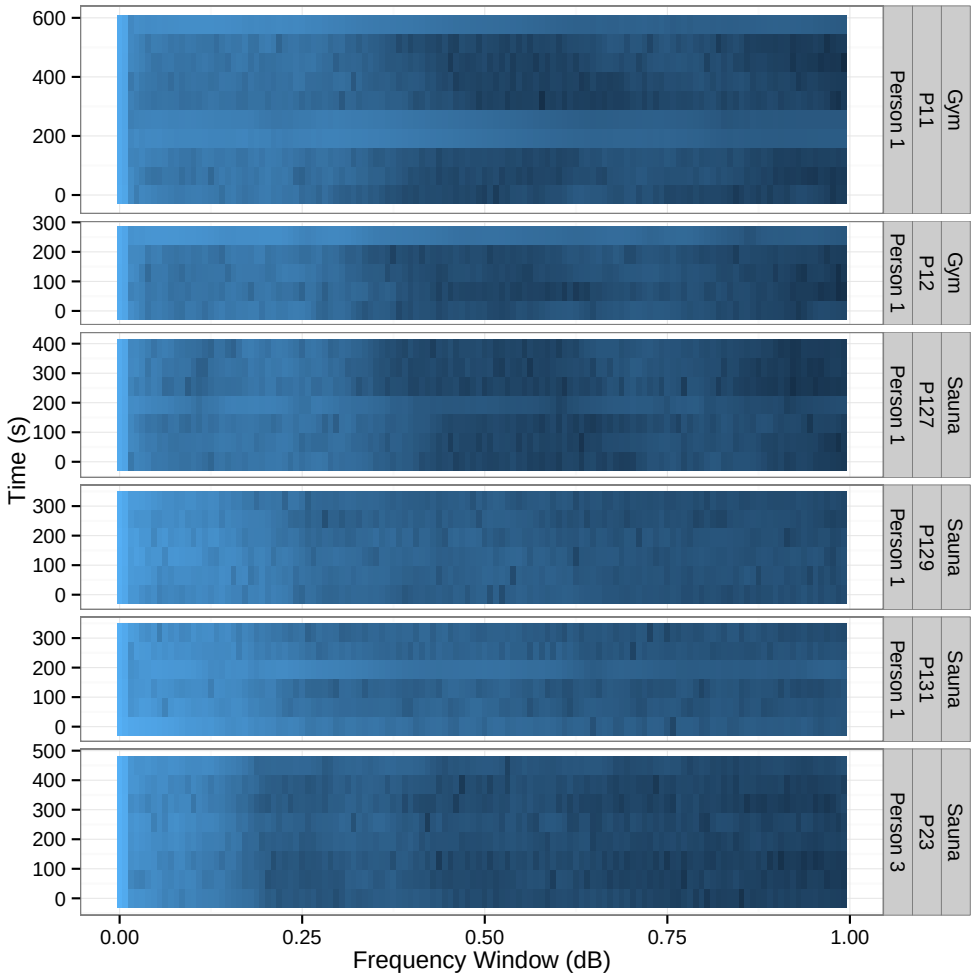
- Set 1: Each windows are stacked in a row to form a big matrix (may suffer from repeated measurement). This contains different parts of same series in various rows.

How data looks like



- Set 2: Each series are averaged over different time points. Each row corresponds to one series.

How data looks like

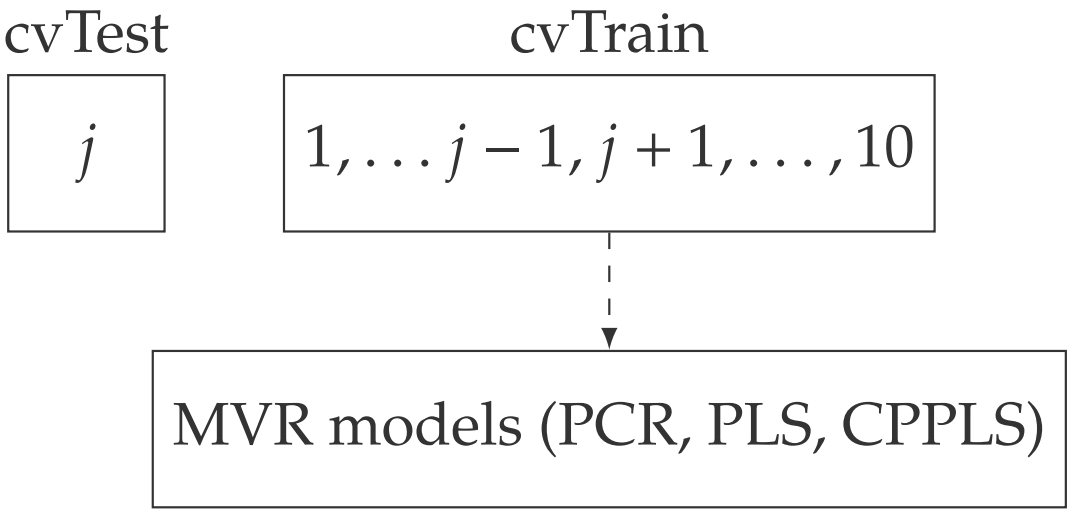


- Set 3: A person can have multiple series of same activity (replication), the third set is averaged over each person-event conbination. In this case each row corresponds to some specific event for some specific person

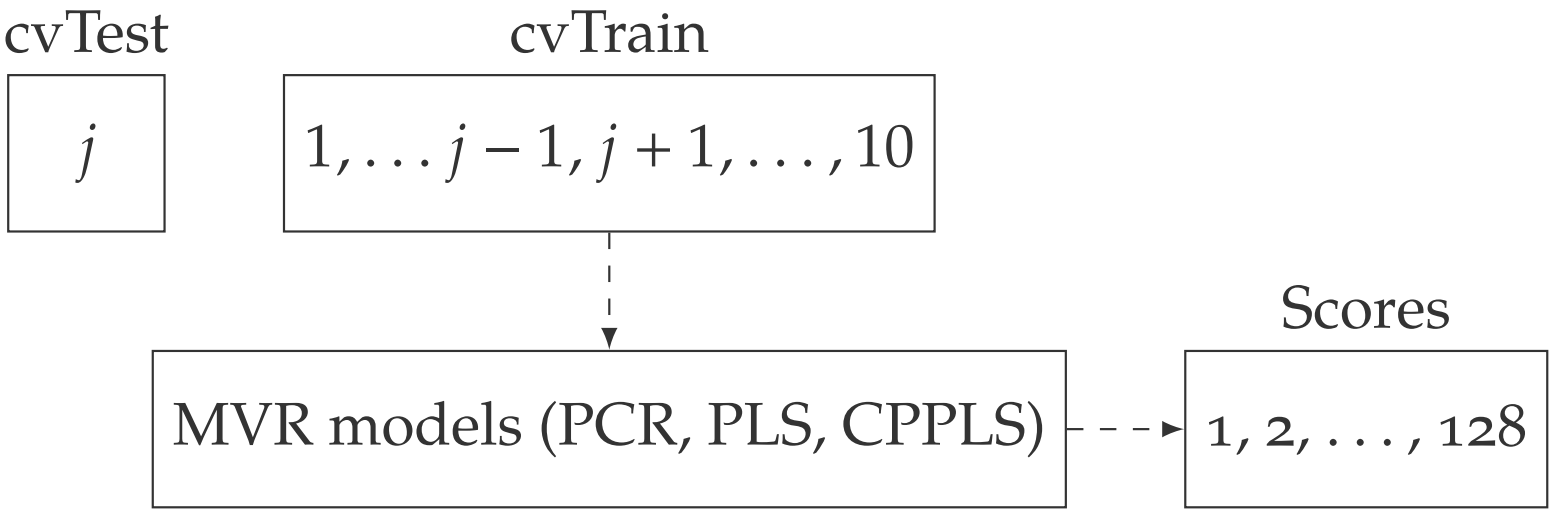
Cross-Validation

cvTest	cvTrain
j	$1, \ldots j-1, j+1, \ldots, 10$

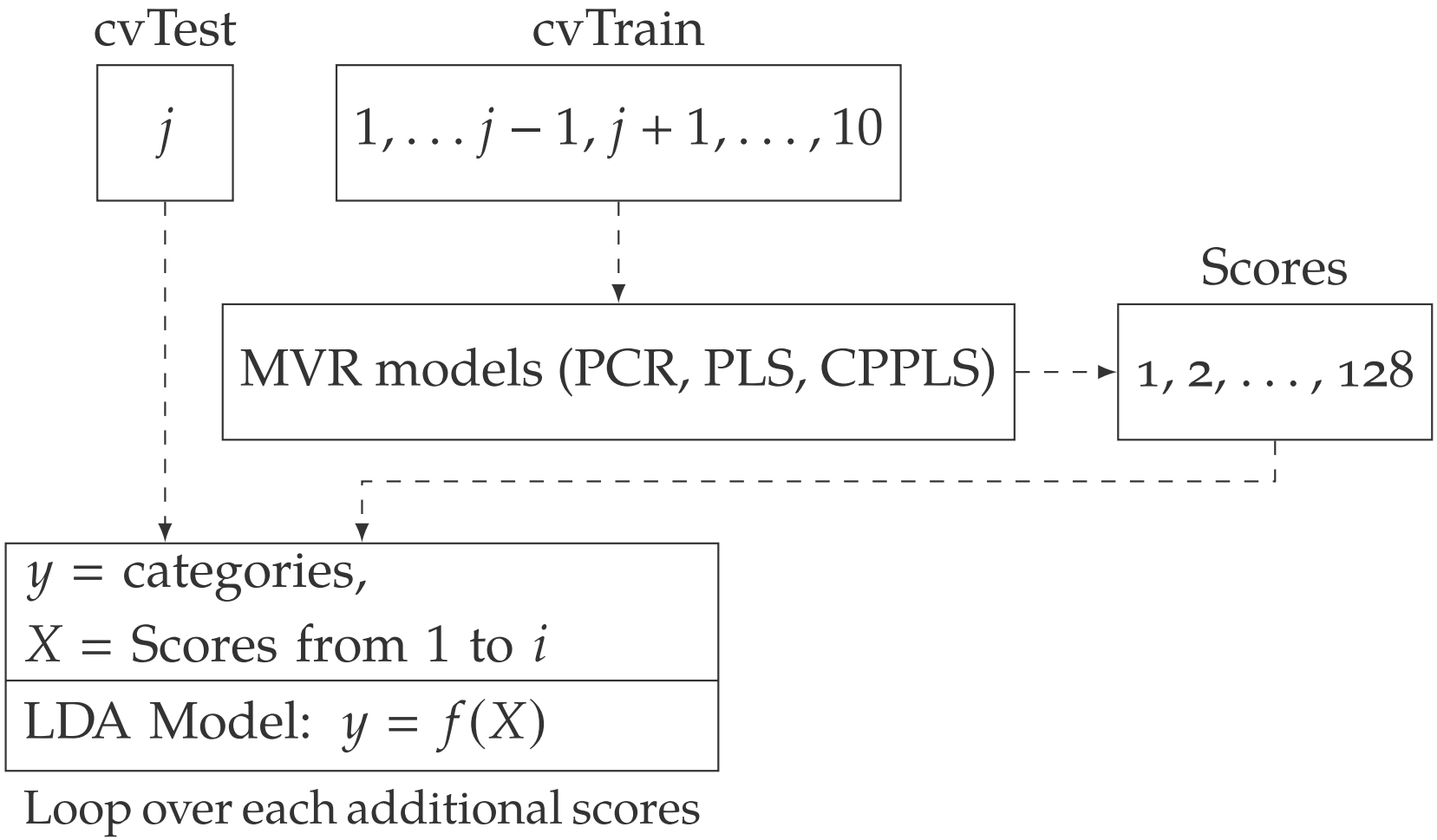
Cross-Validation



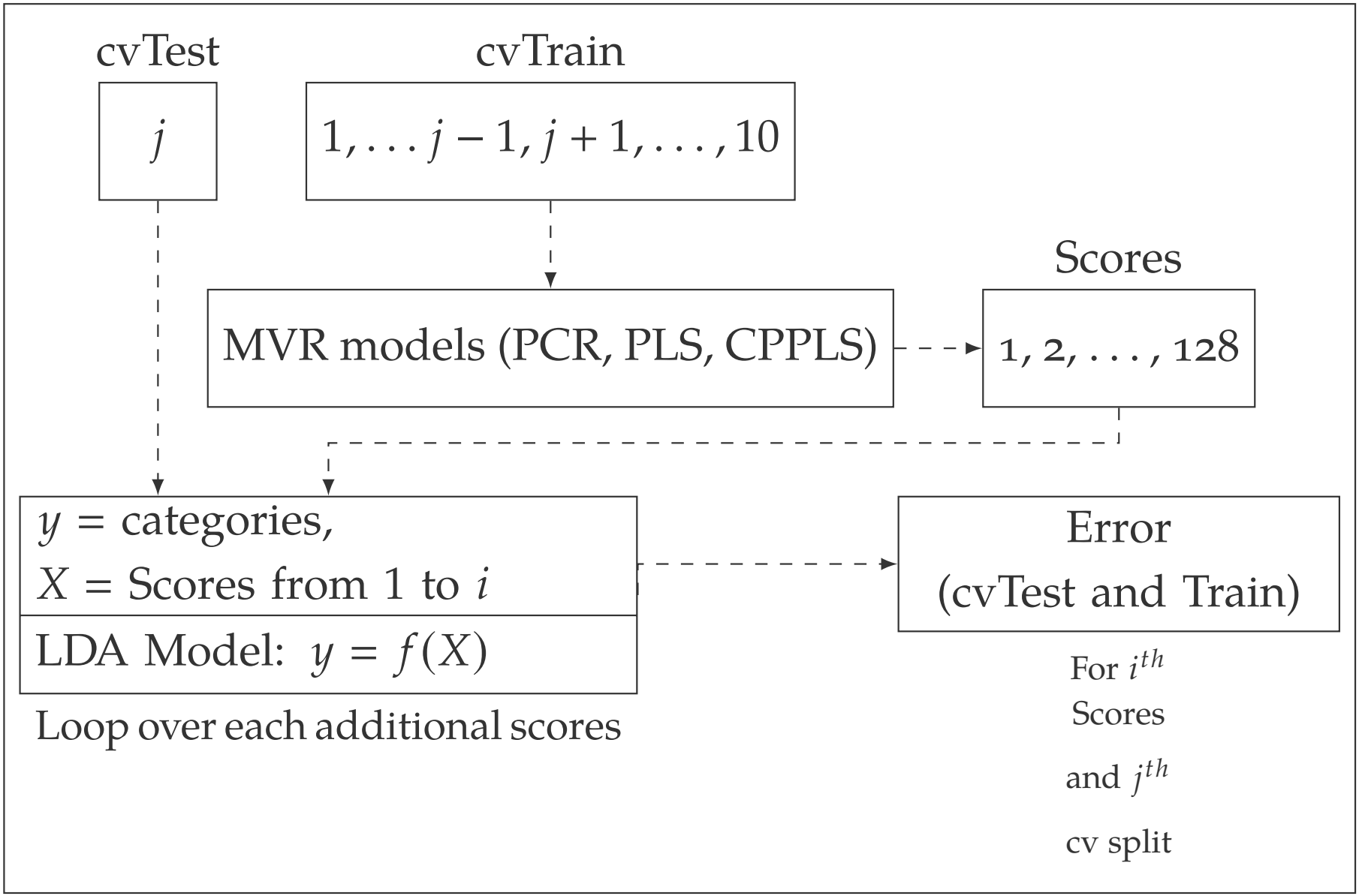
Cross-Validation



Cross-Validation



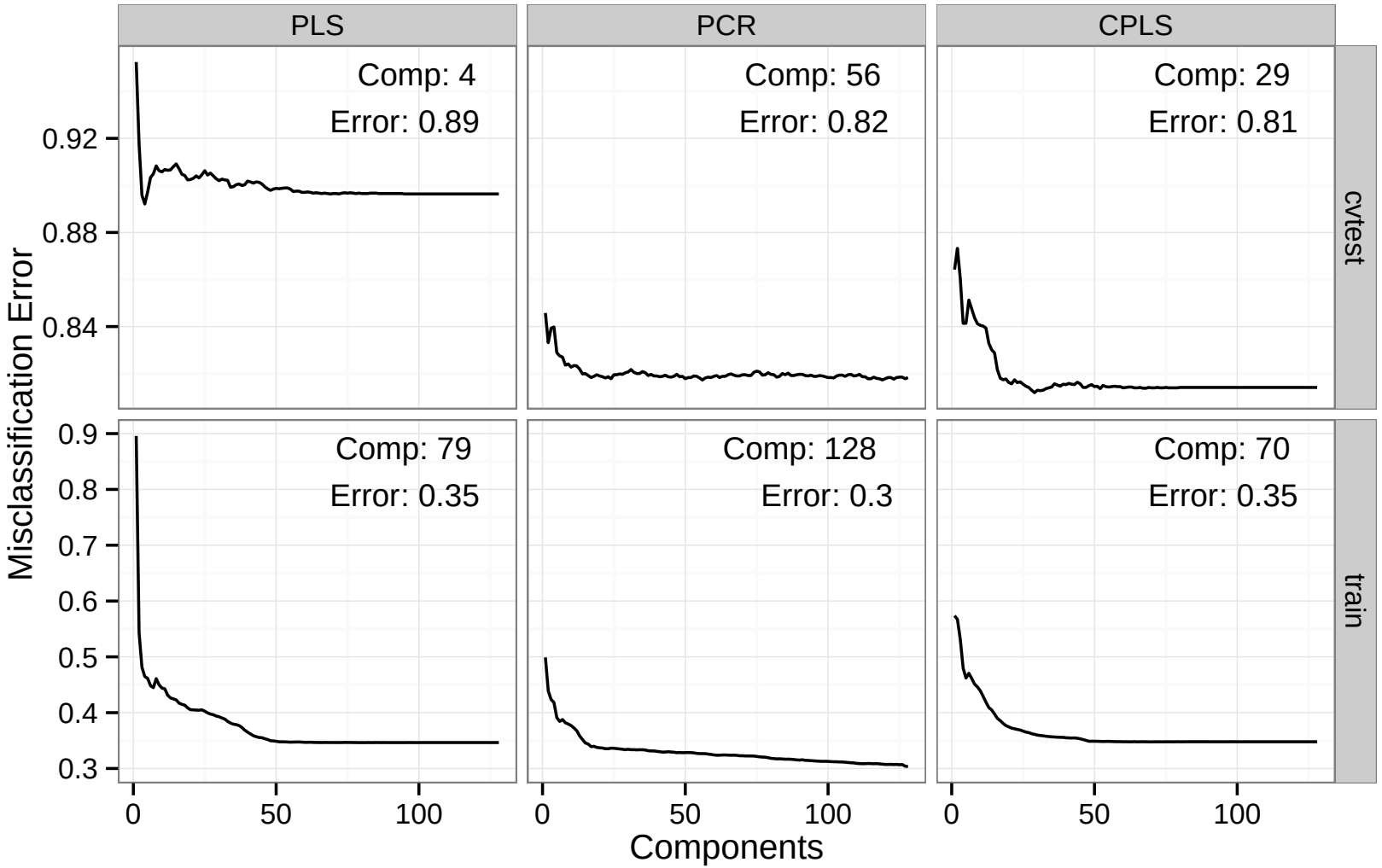
Cross-Validation



Cross-validation Loop

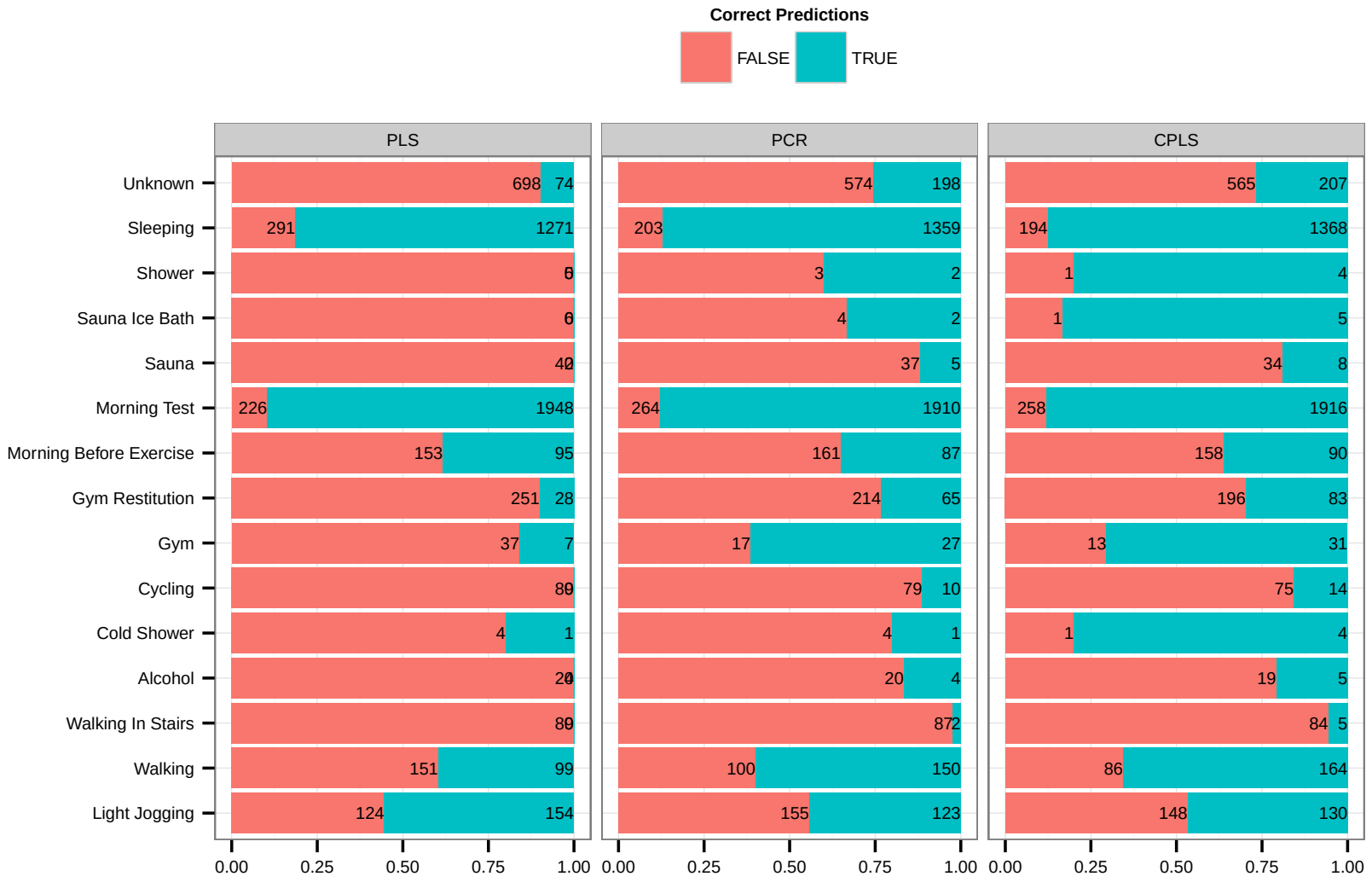
Classification with series stack

Training and Cross-validation Errors



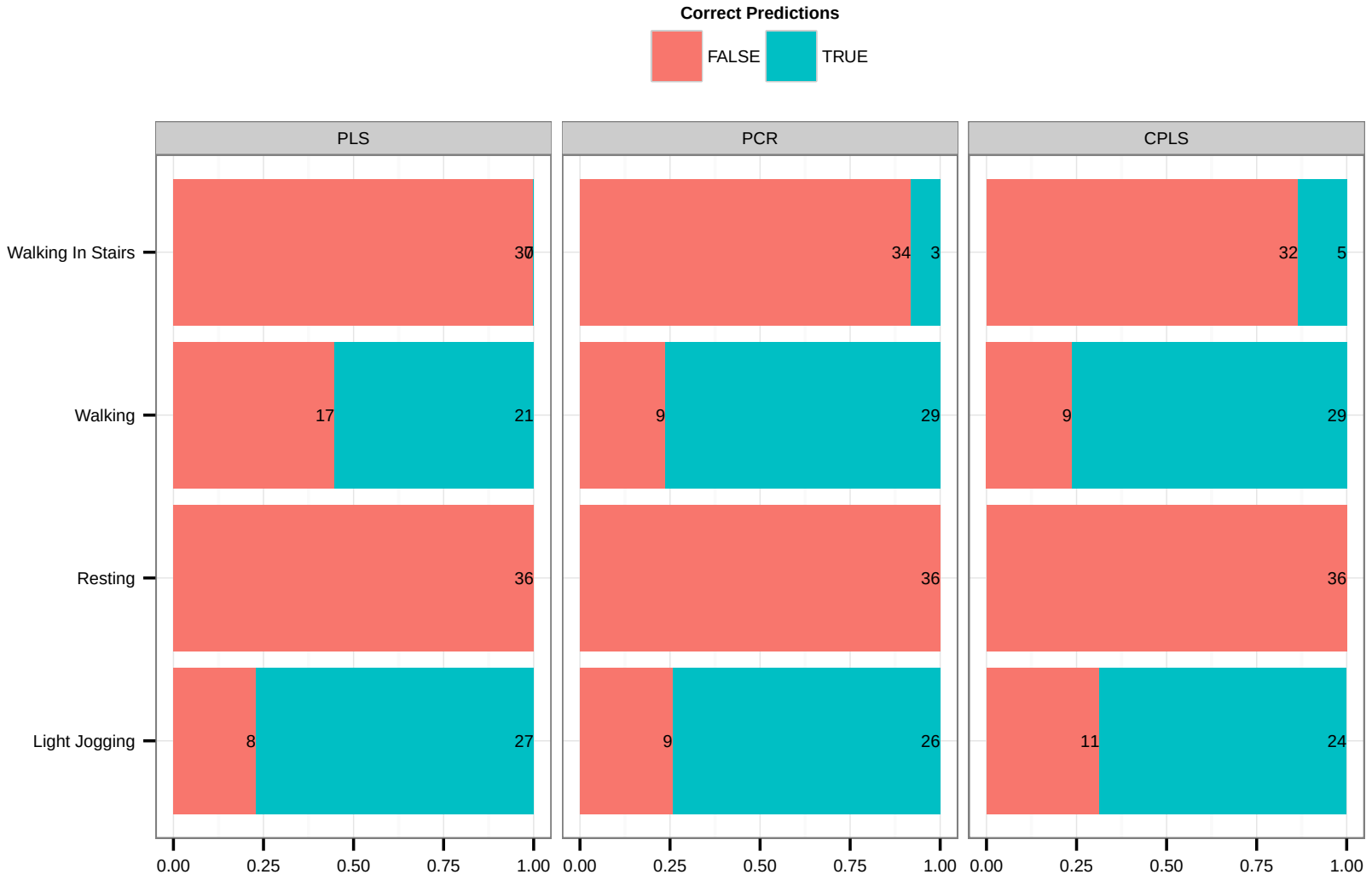
Misclassifications

Training Misclassifications



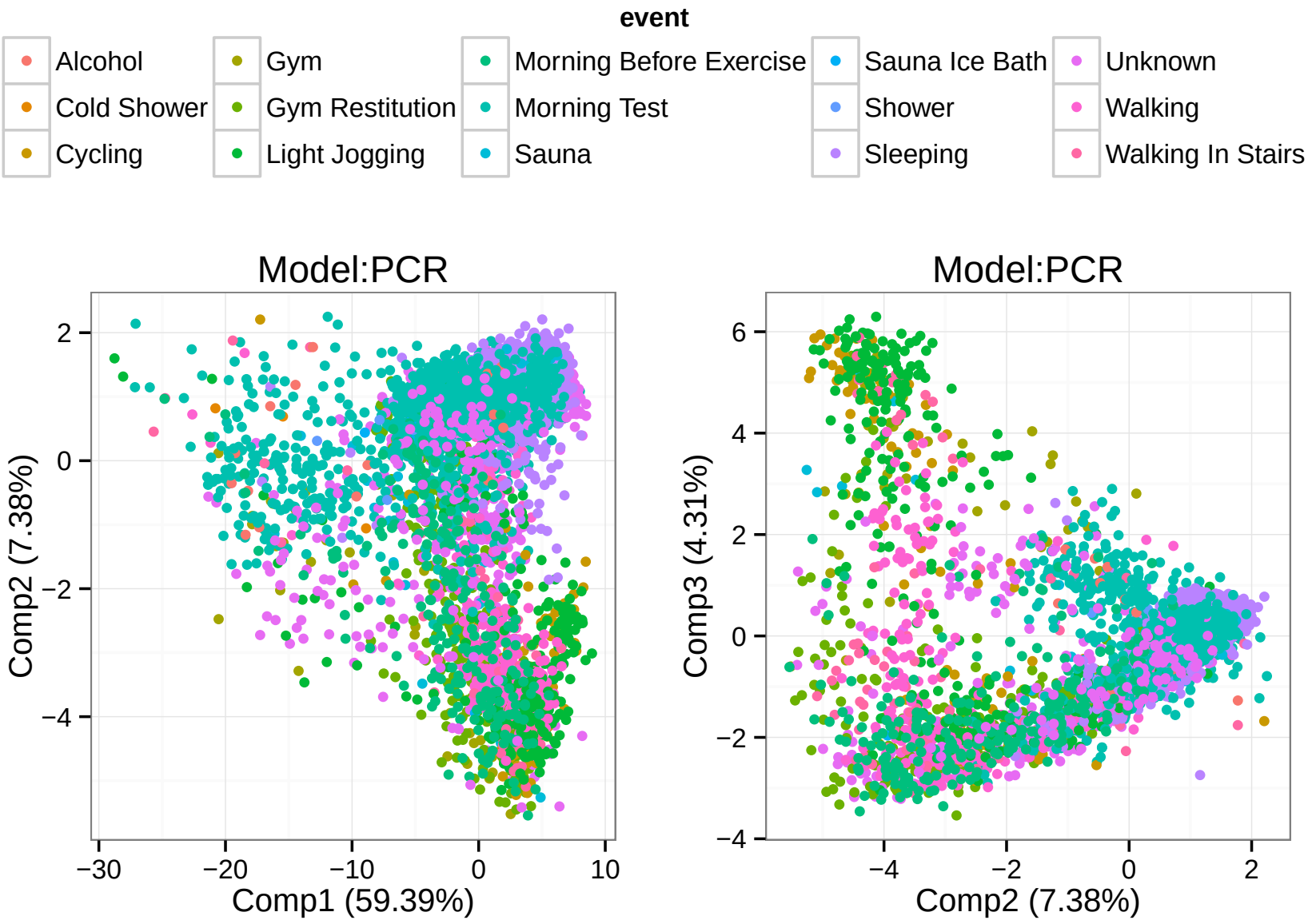
Misclassifications

Test Misclassification



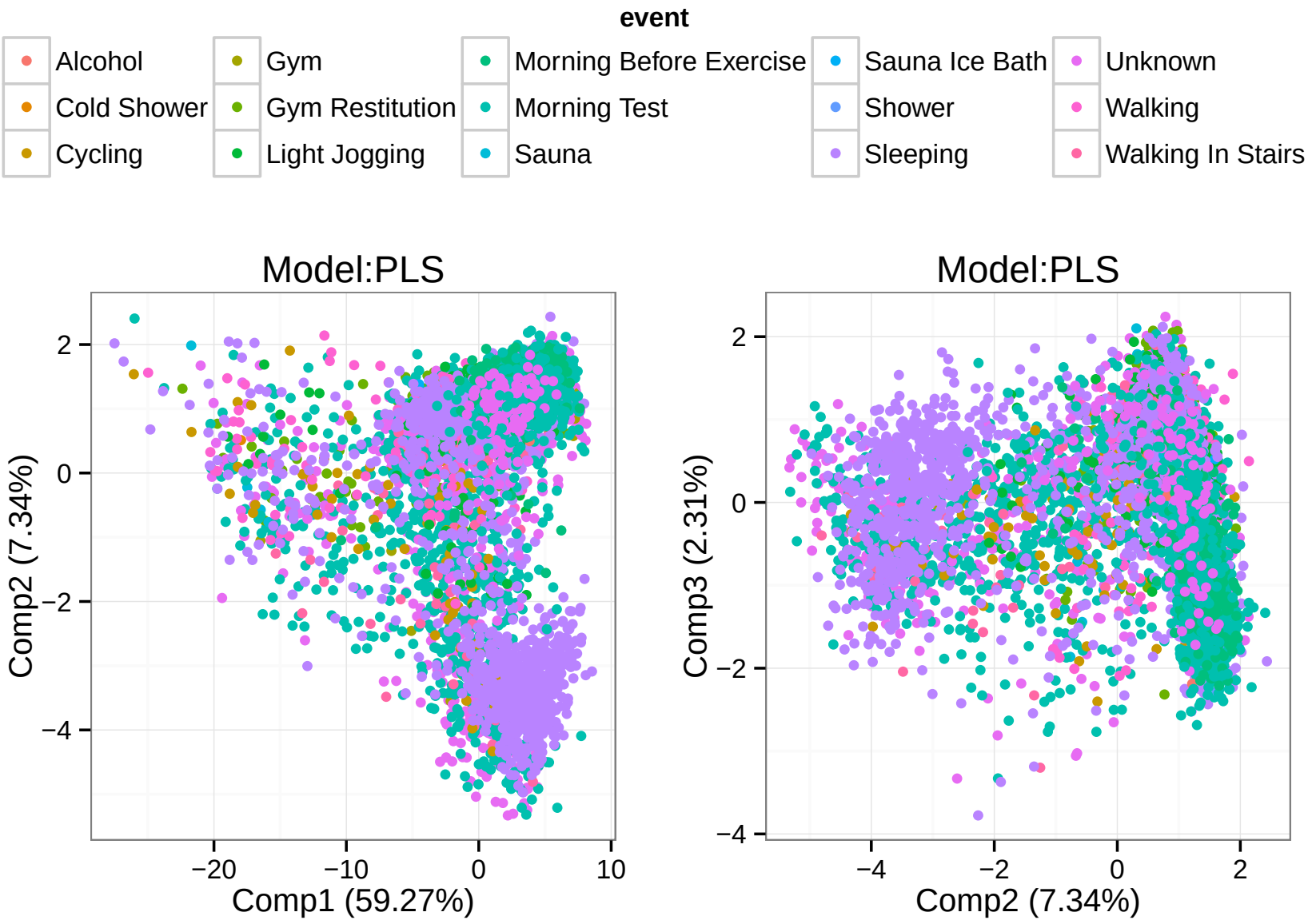
Plotting Scores

Scoreplot for PCR model



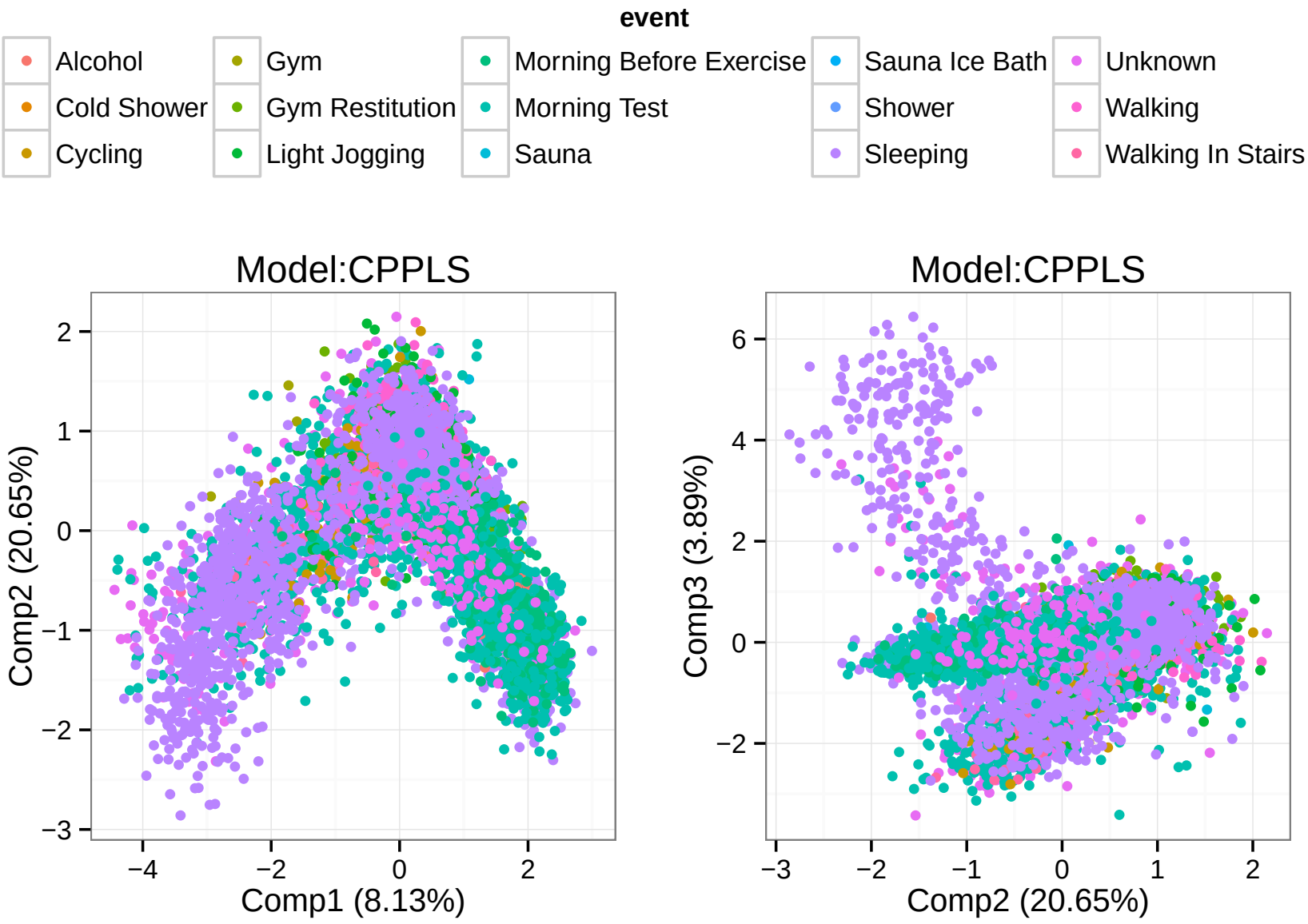
Plotting Scores

Scoreplot for PLS model



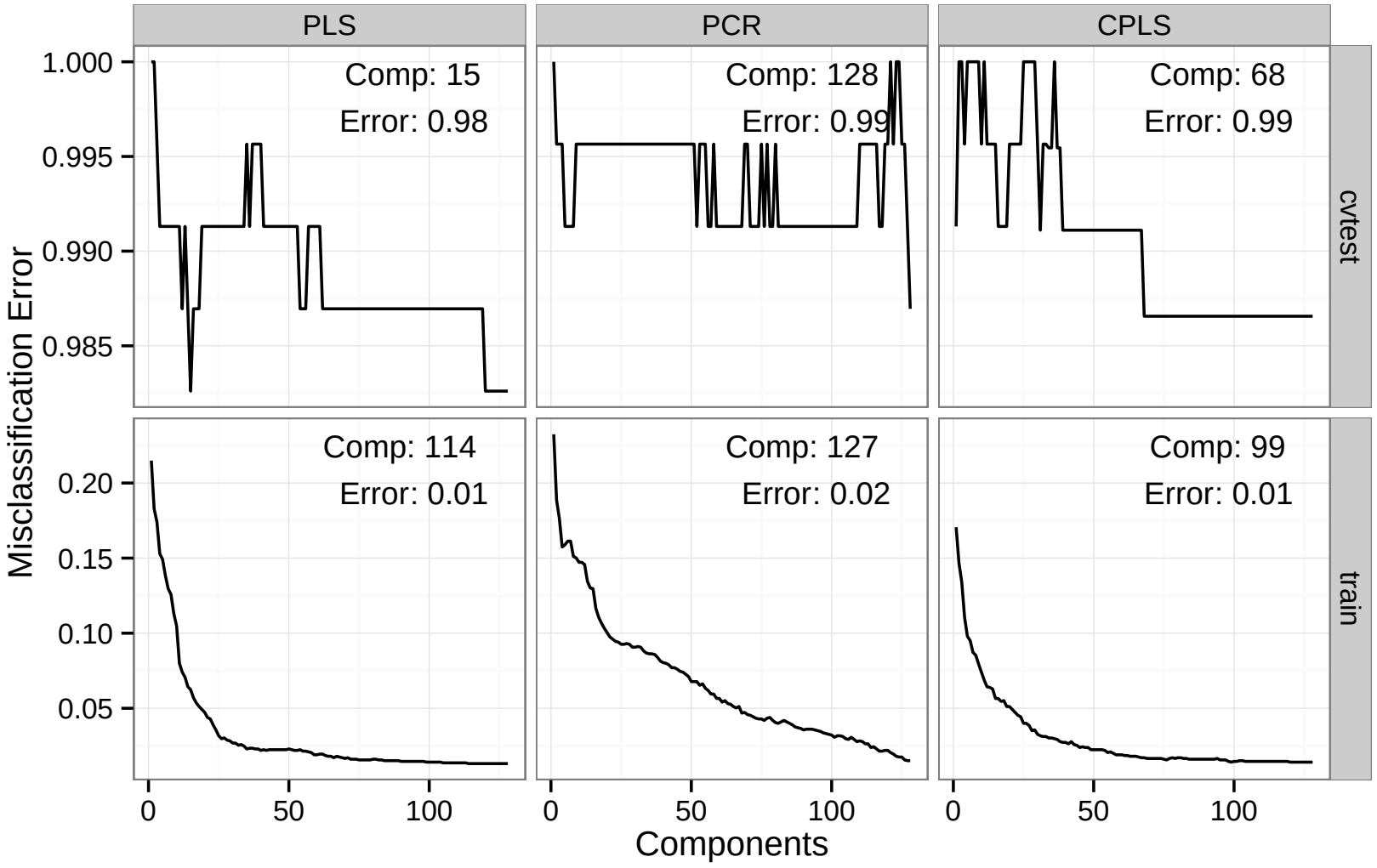
Plotting Scores

Scoreplot for CPPLS model



Classification with series averaged over Series repetition

Training and Cross-validation Errors

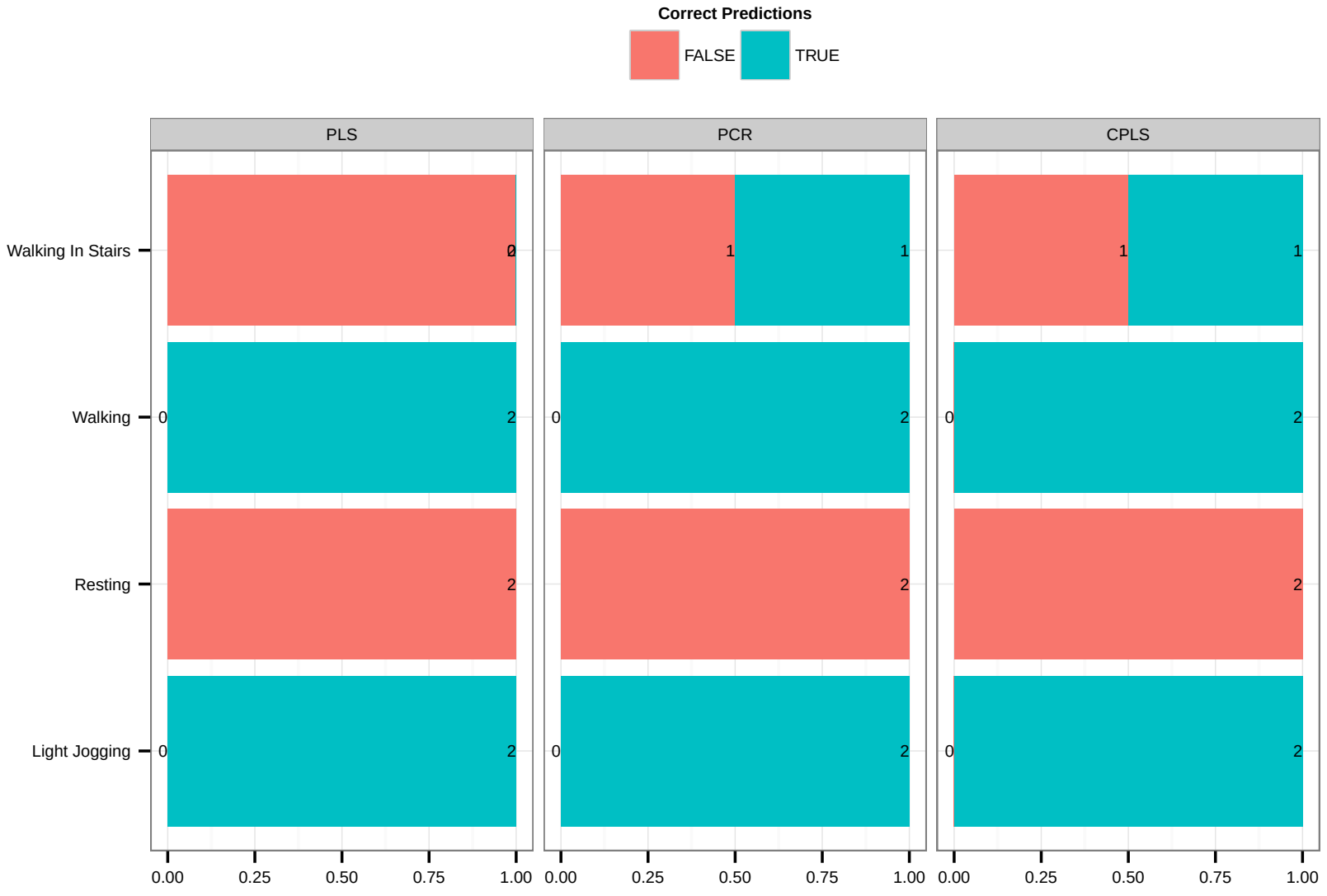


Misclassifications

Training Misclassifications

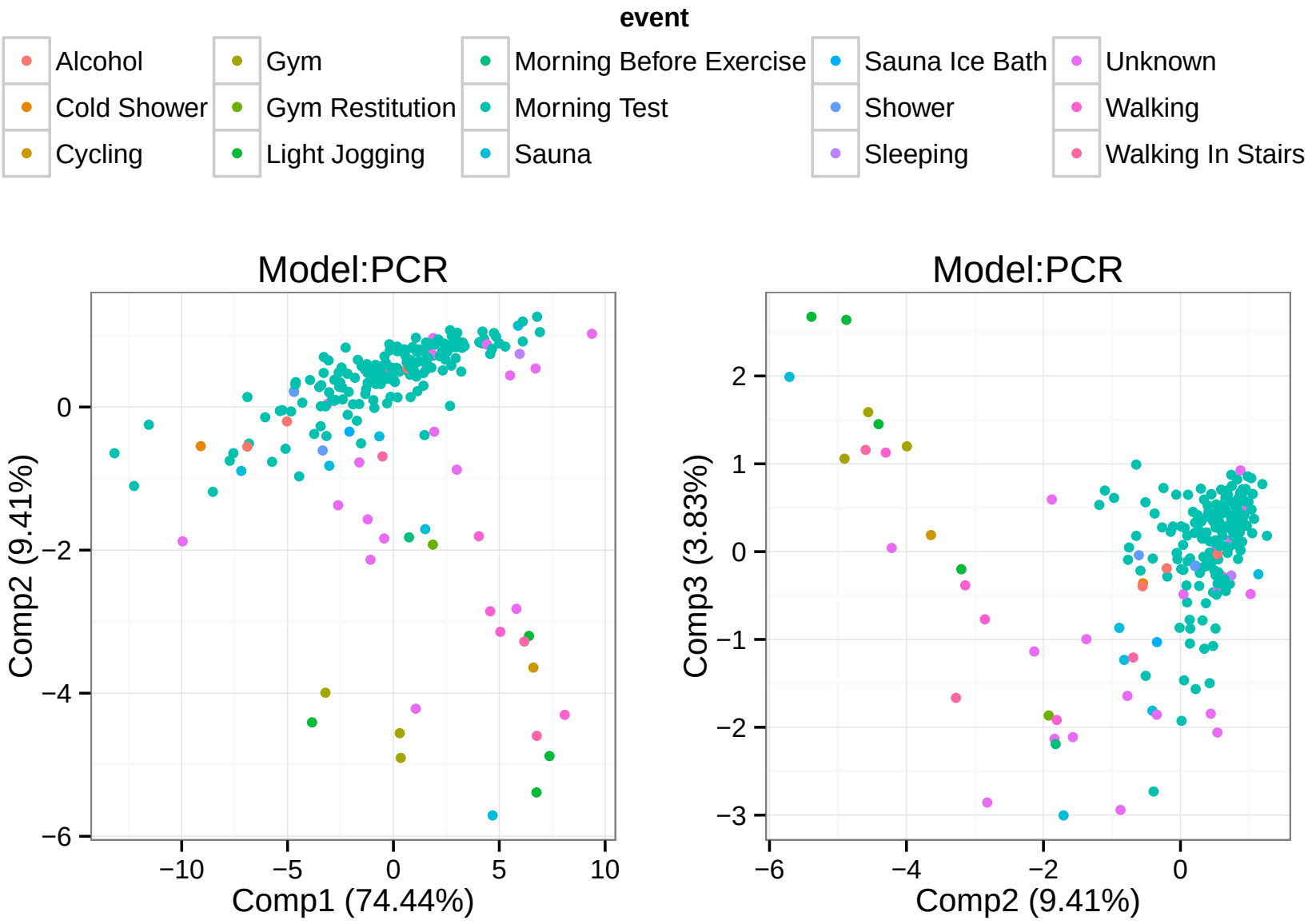
Misclassifications

Test Misclassification



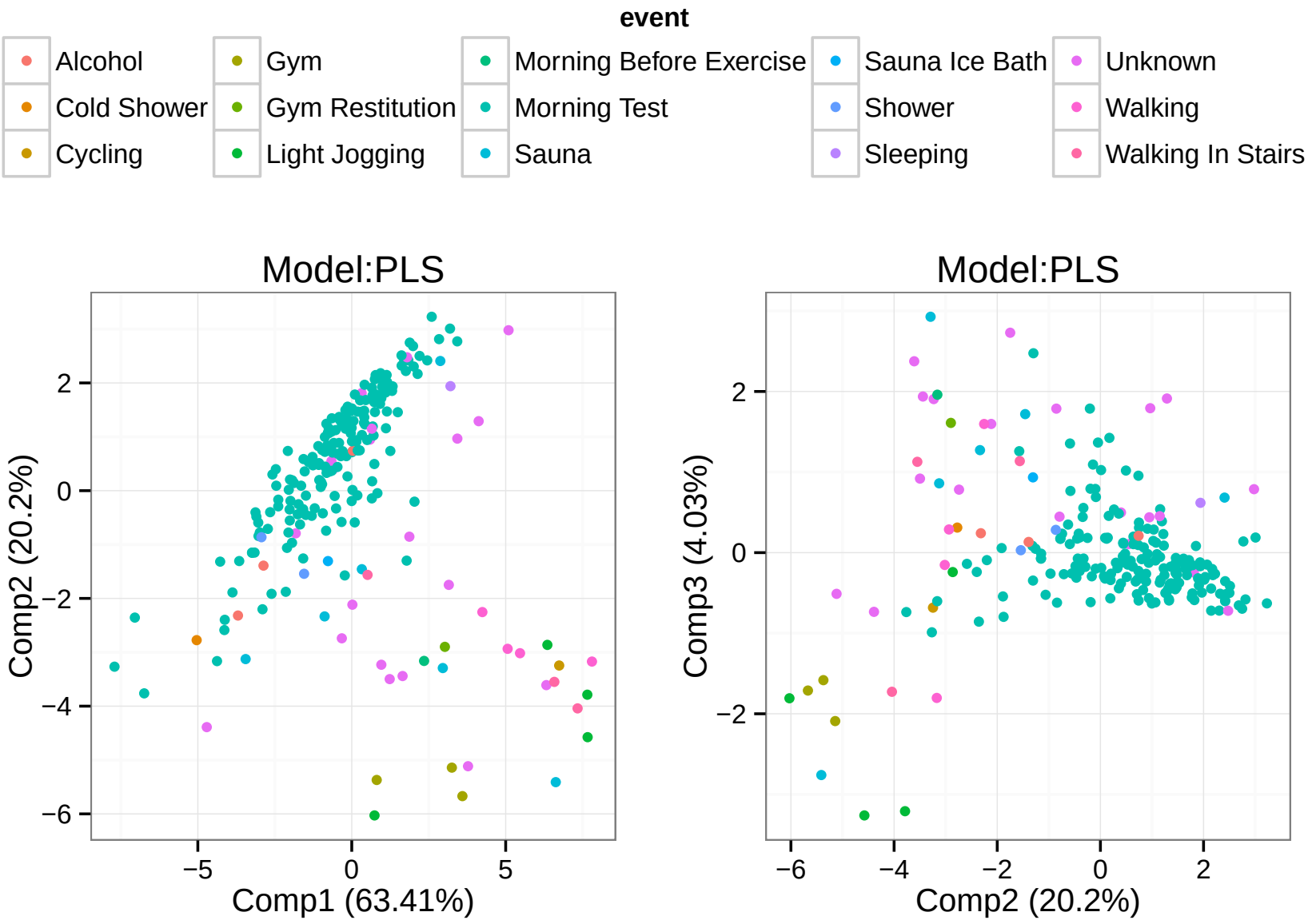
Plotting Scores

Scoreplot for PCR model



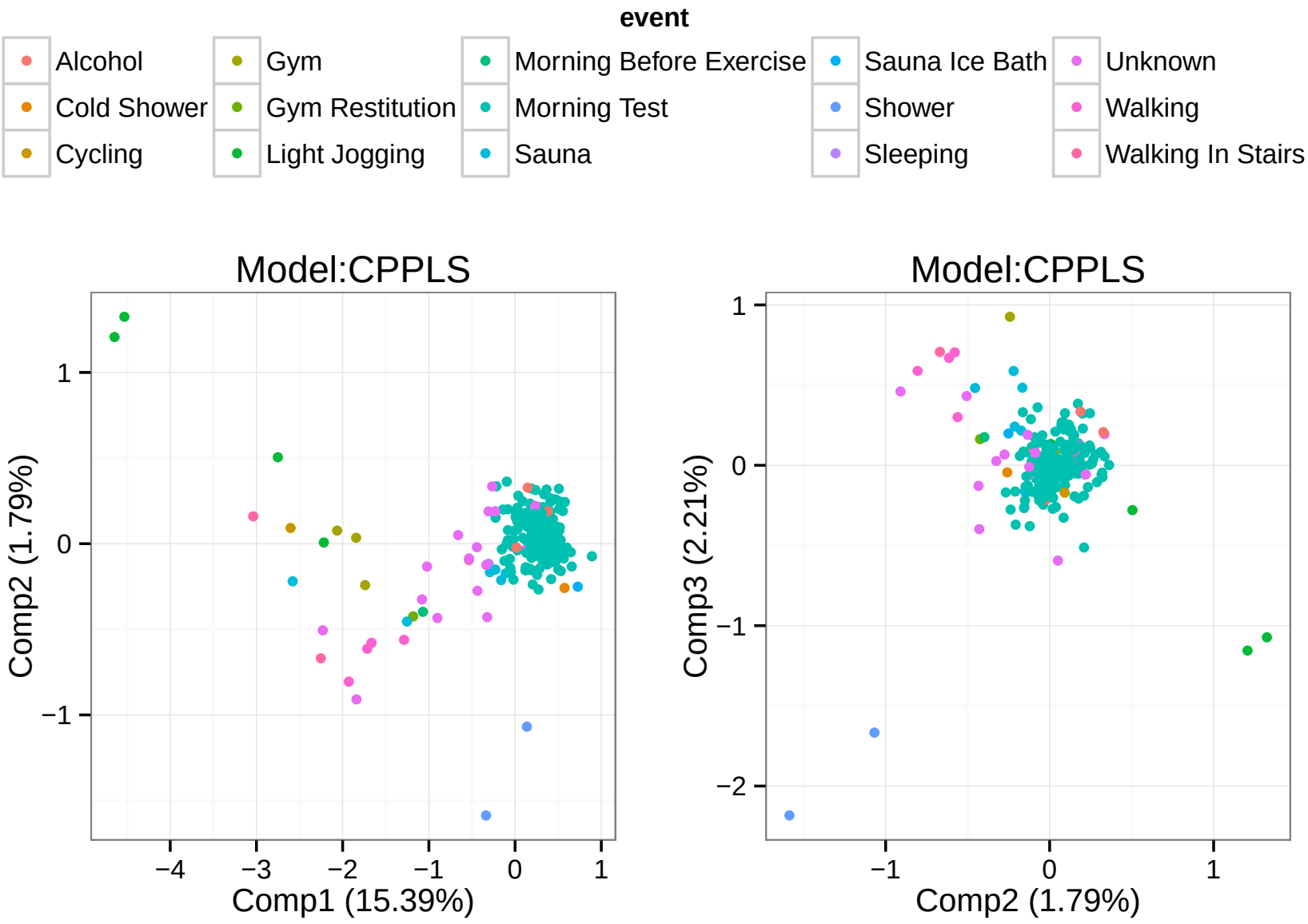
Plotting Scores

Scoreplot for PLS model



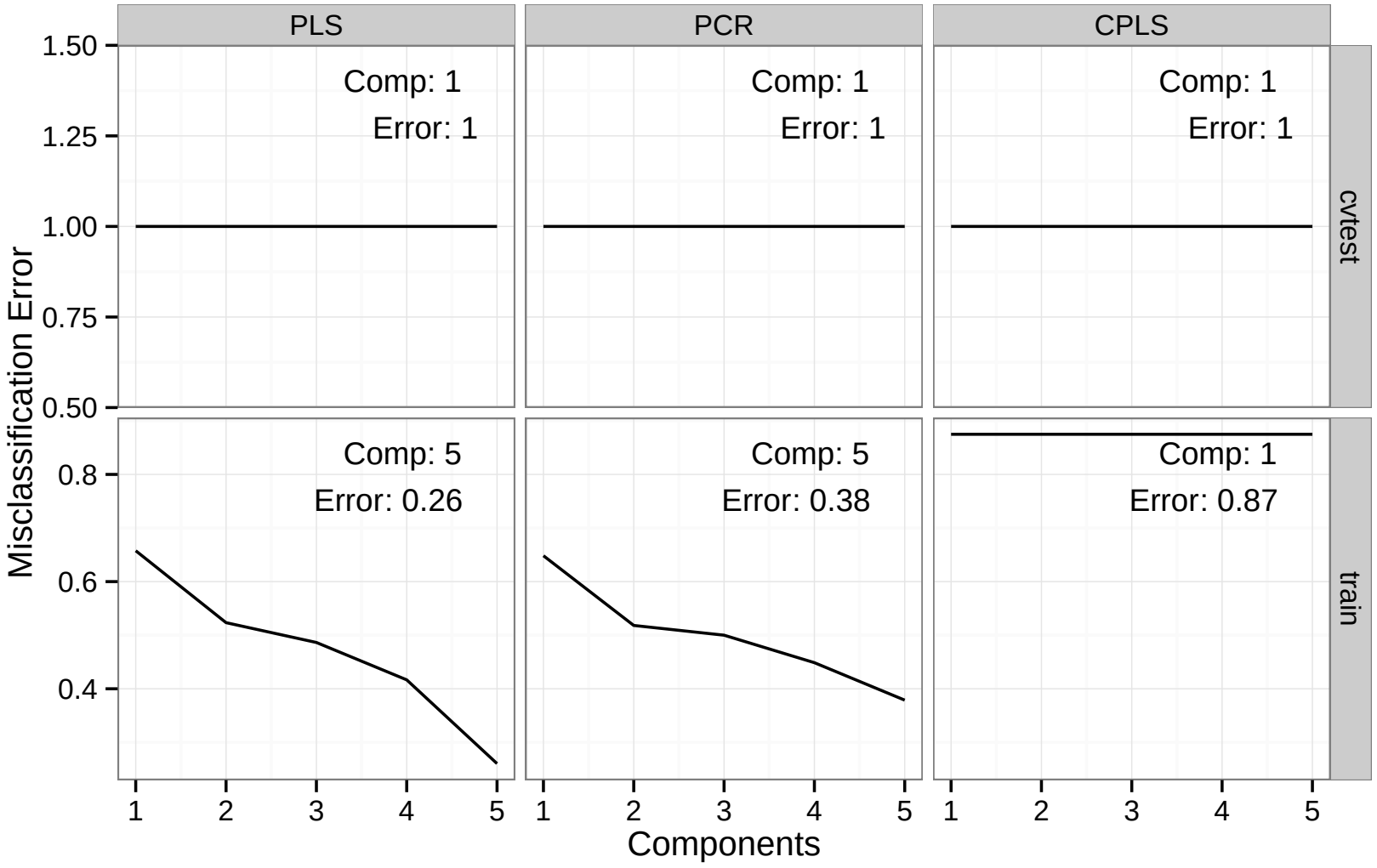
Plotting Scores

Scoreplot for CPPLS model



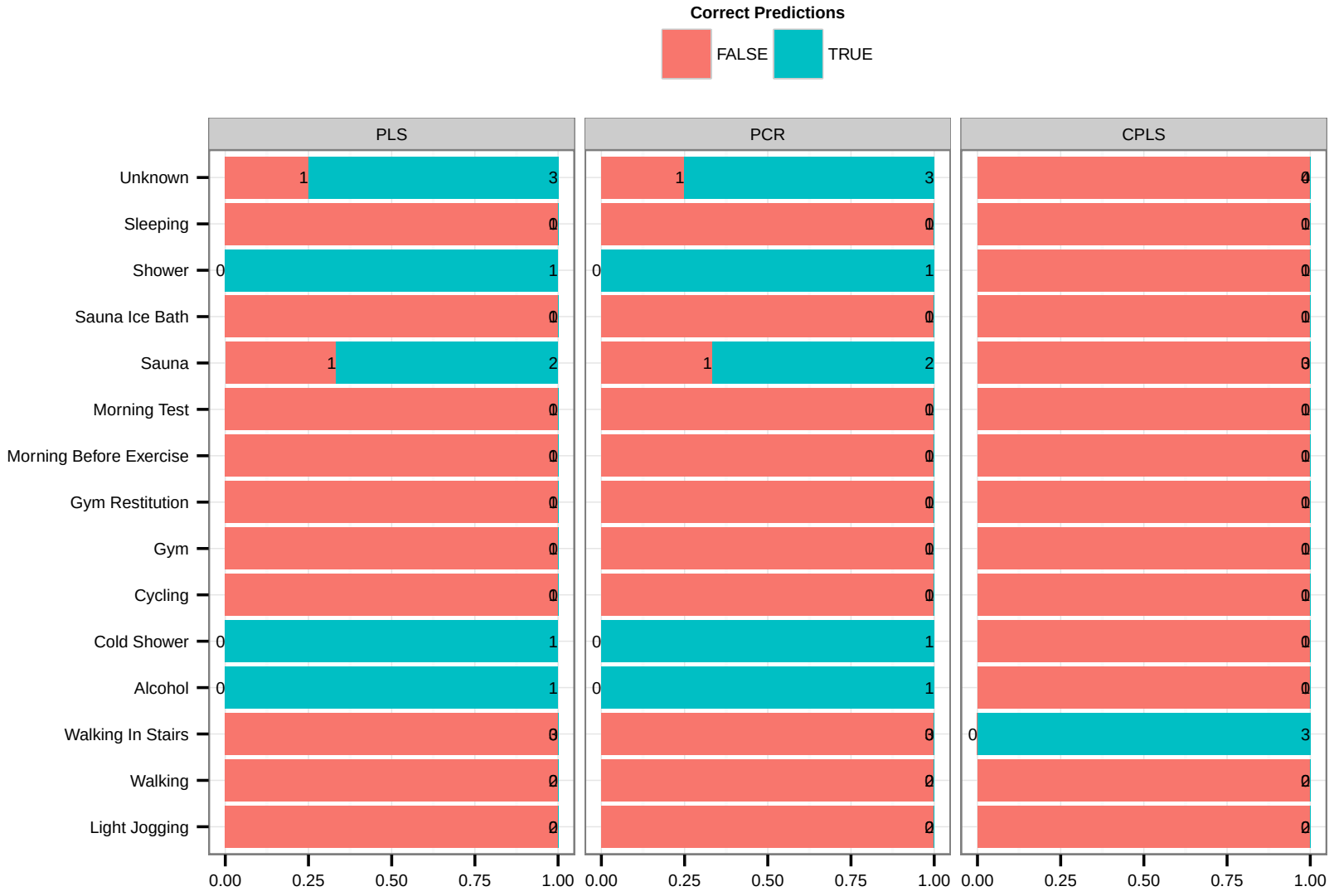
Classification with series averaged over Person-Event Combination

Training and Cross-validation Errors



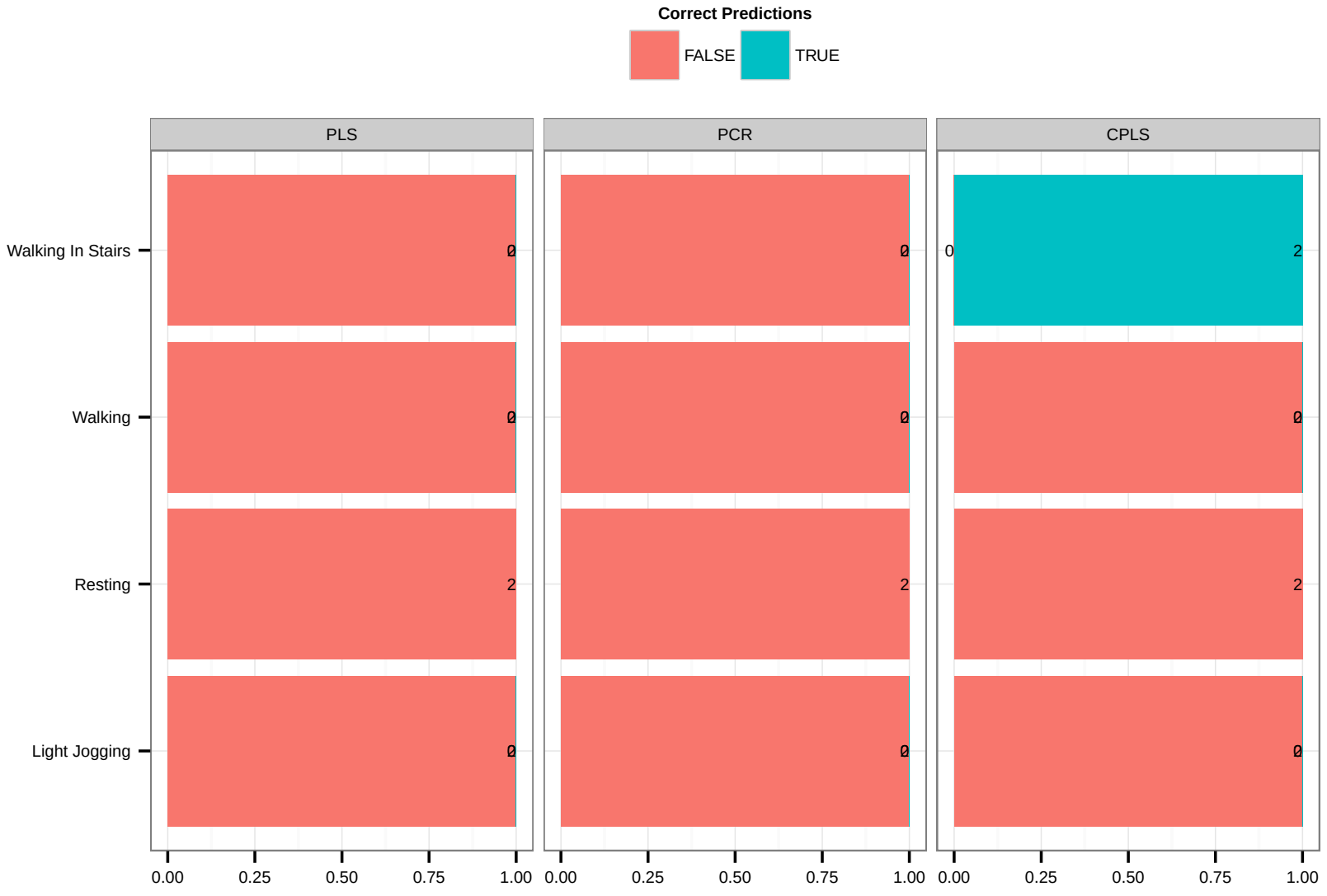
Misclassifications

Training Misclassifications



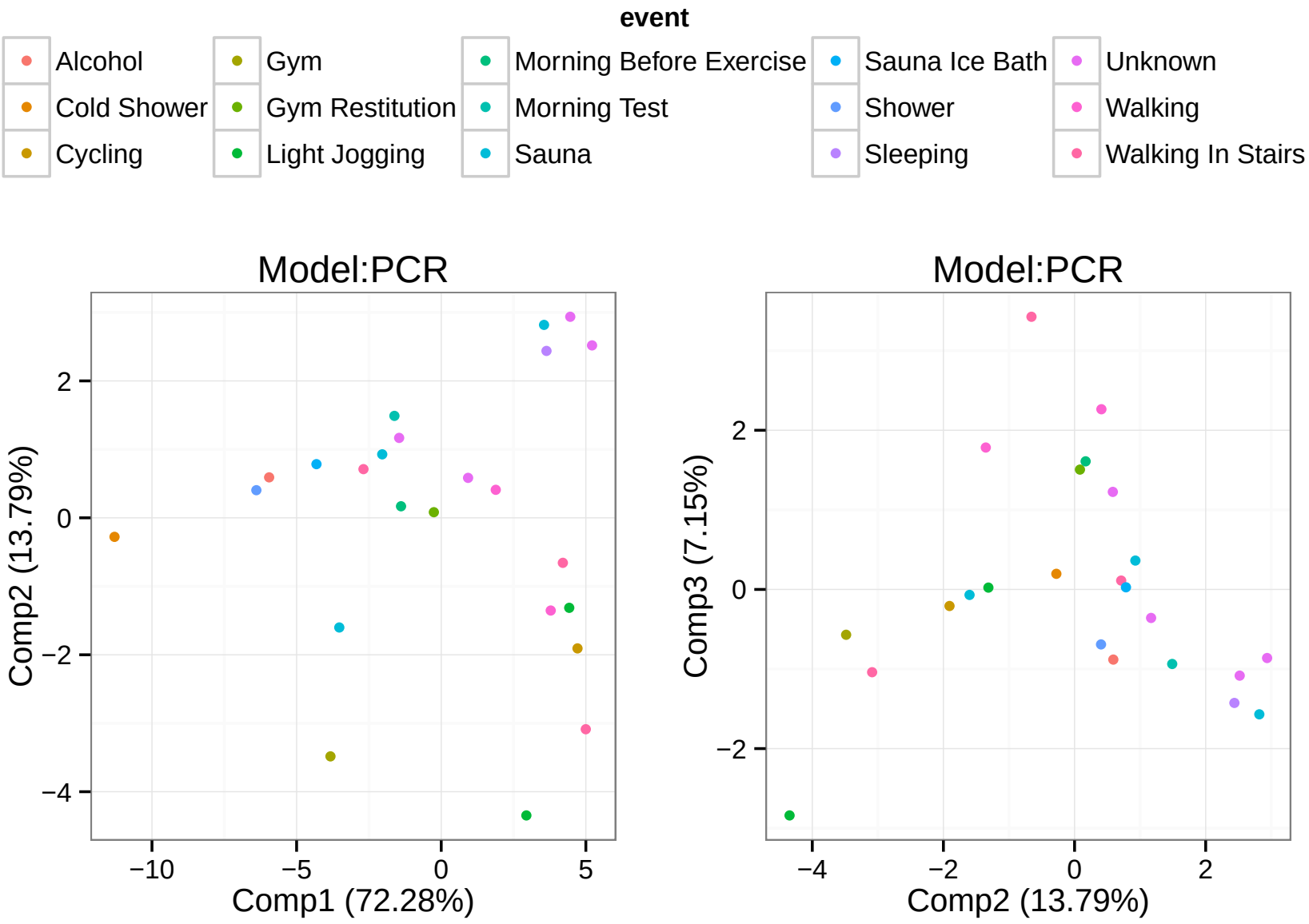
Misclassifications

Test Misclassification



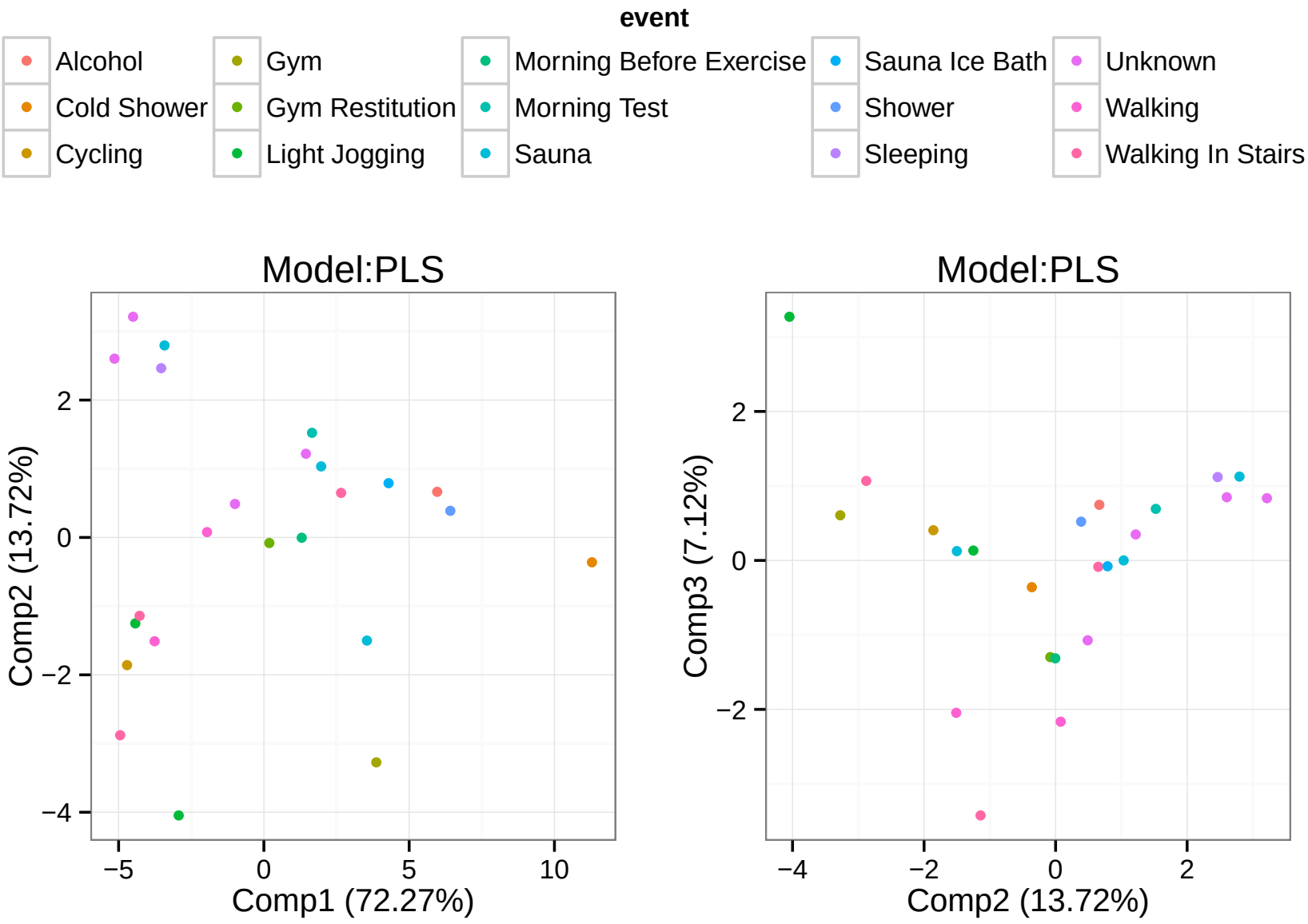
Plotting Scores

Scoreplot for PCR model



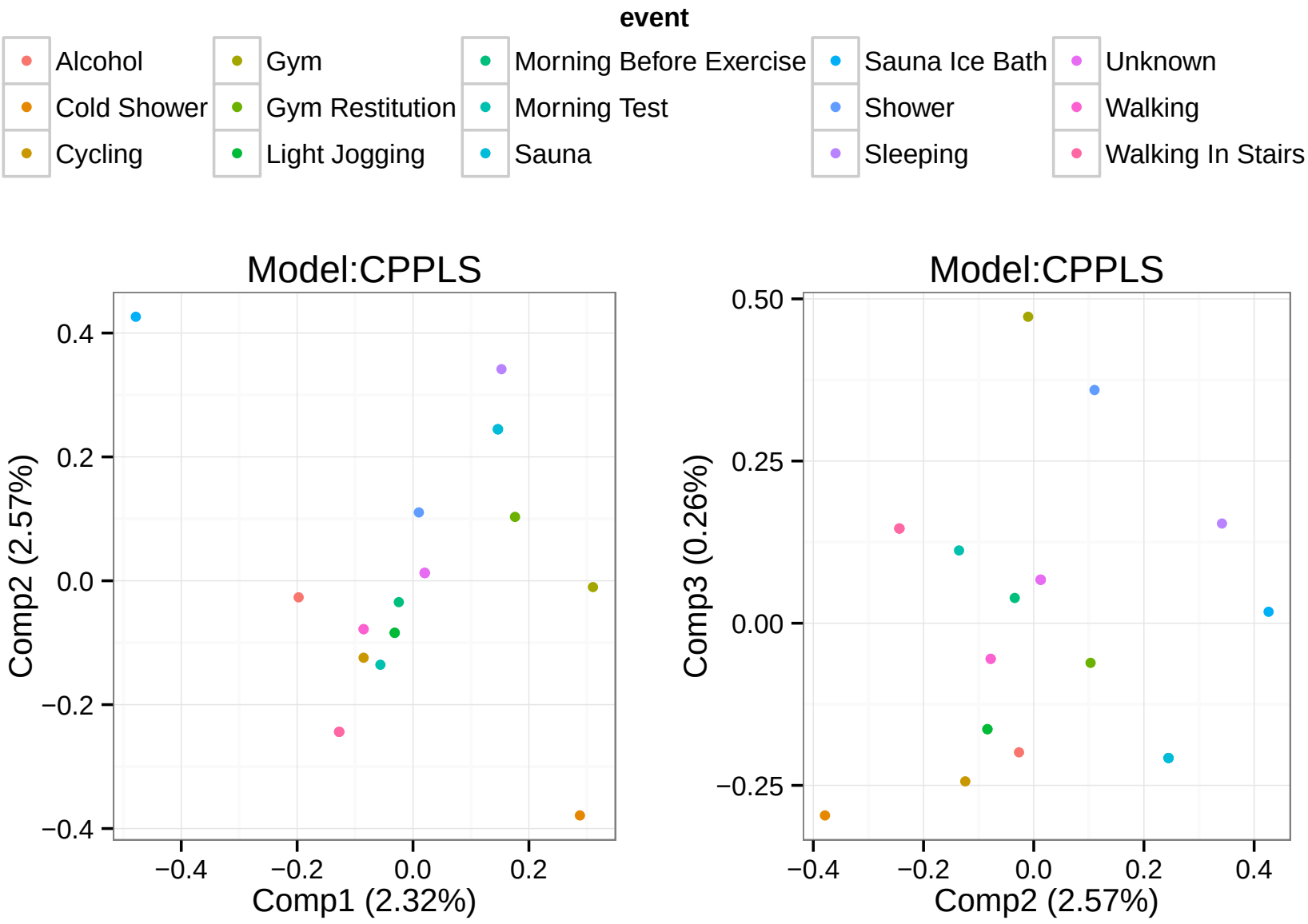
Plotting Scores

Scoreplot for PLS model



Plotting Scores

Scoreplot for CPPLS model



Some Comparison

Misclassification Errors

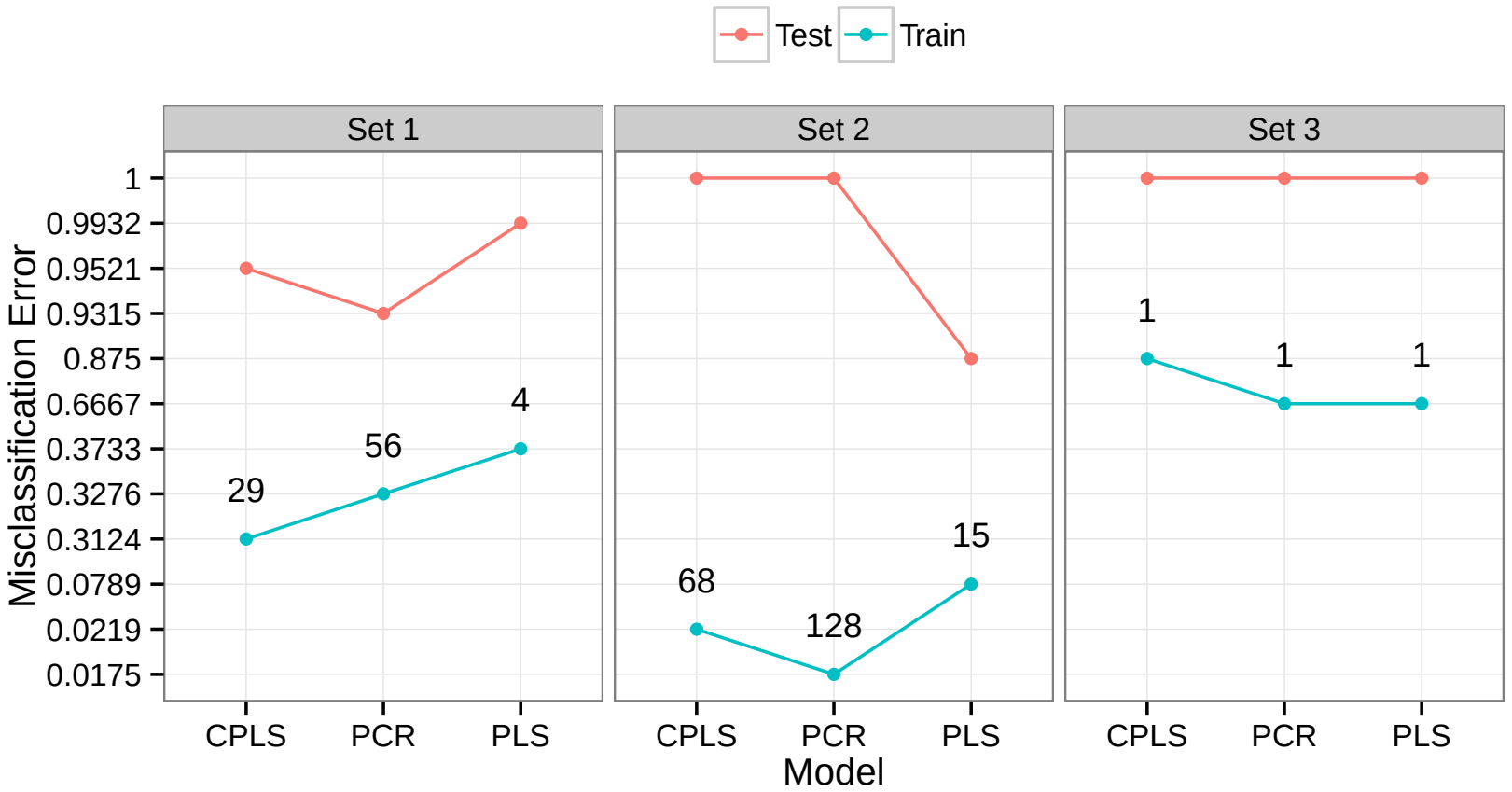


Figure: Training and Test Misclassification Error for all the three models. The LDA models were fitted with the scores obtained from three models with components (number above each points) needed to get minimum cross-validation error.

References

[1] M Dowle et al. *data.table: Extension of Data.frame*. R package version 1.9.6. 2015. URL: http://CRAN.R-project.org/package=data.table.

[2] Ulf G Indahl, Kristian Hovde Liland, and Tormod Næs. "Canonical partial least squares—a unified PLS approach to classification and regression problems". In: *Journal of Chemometrics* 23.9 (2009), pp. 495–504.

[3] Uwe Ligges, Tom Short, and Paul Kienzle. *signal: Signal Processing*. R package version 0.7-6. 2015. URL: http://CRAN.R-project.org/package=signal.

[4] Harald Martens and Magni Martens. *Multivariate analysis of quality: an introduction*. John Wiley & Sons, 2001.

[5] Harald Martens and Tormod Naes. *Multivariate calibration*. John Wiley & Sons, 1992.

[6] Hadley Wickham. "ggplot: An Implementation of the Grammar of Graphics". In: *R package version 0.4. 0* (2006).

[7] Hadley Wickham. "reshape2: Flexibly reshape data: a reboot of the reshape package". In: *R package version* 1.2 (2012).

[8] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.11. 2015. URL: http://CRAN.R-project.org/package=knitr.