GHENT UNIVERSITY

Feature Discovery in Small-Sized Experiments in Early Drug Development

Steven Wallaert – Promotor: Prof. Dr. Ir. Olivier Thas

Context

- Pre-clinical pharmacological research
- Biomarker discovery
- (Multi-) Omics

Problem

- Increasing popularity machine learning

- High hopes for better, more, easier discoveries

However

- High dimensionality: up to 10.000 and more features

- Extremely small sample sizes (10 to 50)

And

- Little is known about the performance of methods in these extreme situations

Therefore

- Need for a neutral comparison study

Research questions

- How do ‘traditional’ hypothesis tests compare to ‘modern’ statistical methods in these situations?

- Which methods are better suited for different scenarios?

- What are limitations and weaknesses of the different methods?

Proposal of guidelines: “How (not) to”

Included methods

- Student’s t-test with FDR control

- Student’s t-test with empirical Bayes based selection bias correction using Tweedie’s formula

- Logistic regression with L1 regularization

- Random forests based RFE

- Support vector machine based RFE

Simulation study

- Variety of scenarios
  - Data generating mechanism
  - Sample size
  - Number of features
  - Number of predictive features
  - Degree of discriminativeness

- Estimands
  - Number of true/false detections
  - Chance of true/false detection
  - Discriminative ability:
    - AUC
      - Bias
      - Variance

Challenges

- Selection of methods
- Computational demands
- Integration of results and visualization

Example analysis