Priority-Lasso:
prediction
using
multi-omics
data

Roman
Hornung

Introduction

priority-Lasso
estimation
algorithm

Application

Conclusions

Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data

Simon Klau, Vindi Jurinovic, **Roman Hornung**, Tobias Herold, Anne-Laure Boulesteix

LMU Munich
Institute for Medical Information Processing, Biometry and Epidemiology

March, 26th, 2018

Last few years: more and more **data** available that feature **omics** measurements of **several types for the same patients** (**multi-omics data**),

e.g. gene expression **and** DNA methylation data **and** miRNA.

⇒ **new** possibility: **combine several types** of omics data **for prediction** modeling

⇒ **new methodological developments necessary** that take **block structures** in high-dimensional covariate data into account

Priority-Lasso:
prediction
using
multi-omics
data

Roman
Hornung

Introduction

priority-Lasso
estimation
algorithm

Application

Conclusions

Introduction

Different types of omics data ("**blocks**") are associated with **different costs** and are **differently well established**.

$\Rightarrow$ Medical **doctors prioritize some blocks** over others and **predictive information overlaps between blocks**.

$\Rightarrow$ **idea** of our new approach **priority-Lasso**:

Retrieve overlapping predictive information from blocks of high priority rather than from blocks of low priority.

$\Rightarrow$ **blocks of low priority less used** in the resulting prediction rules

Priority-Lasso:
prediction
using
multi-omics
data

Roman
Hornung

Introduction

priority-Lasso
estimation
algorithm

Application

Conclusions

Prediction:

Fit a **Lasso** model considering **only** the covariates in the **block with highest priority**:
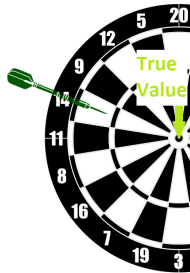
$$y_{i*} = \hat{\beta}_0 + x_{i*1}^{(\pi_1)}\hat{\beta}_1^{(\pi_1)} + \ldots + x_{i*p_{\pi_1}}^{(\pi_1)}\hat{\beta}_{p_{\pi_1}}^{(\pi_1)} + \epsilon_{i*}$$

by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p_{\pi_1}} x_{ij}^{(\pi_1)}\beta_j^{(\pi_1)}\right)^2 + \lambda^{(\pi_1)}\sum_{j=1}^{p_{\pi_1}}|\beta_j^{(\pi_1)}|.$$

Rationale: **Use all** predictive **information** contained in the block **with highest priority**.

Priority-Lasso:
prediction
using
multi-omics
data

Roman
Hornung

Introduction

priority-Lasso
estimation
algorithm

Application

Conclusions

Using the fitted values from **Step 1 as offset** fit a **Lasso** model considering the covariates in the **block with second highest priority**:

Prediction:

$$y_{i*} = \underbrace{\hat{\beta}_0 + \mathbf{x}_{i*}^{(\pi_1)T} \hat{\boldsymbol{\beta}}^{(\pi_1)}}_{\hat{\eta}_{1,i*}(\boldsymbol{\pi})} + \mathbf{x}_{i*}^{(\pi_2)T} \hat{\boldsymbol{\beta}}^{(\pi_2)} + \tilde{\epsilon}_{i*}$$

by minimizing

$$\sum_{i=1}^{n} \left(y_i - \hat{\eta}_{1,i}(\boldsymbol{\pi}) - \mathbf{x}_i^{(\pi_2)T} \boldsymbol{\beta}^{(\pi_2)} \right)^2 + \lambda^{(\pi_2)} \sum_{j=1}^{p_{\pi_2}} |\beta_j^{(\pi_2)}|.$$

Rationale: **Use all** that predictive **information** from the **block with second-highest priority** that is **not contained in** the **block with highest priority**.

Priority-Lasso: prediction using multi-omics data

Roman Hornung

Introduction

priority-Lasso estimation algorithm

Application

Conclusions

Using the fitted values from **Step 2 as offset** fit a **Lasso** model considering the covariates in the **block with third highest priority**:

Prediction:

$$y_{i*} = \underbrace{\hat{\beta}_0 + \boldsymbol{x}_{i*}^{(\pi_1)T}\hat{\boldsymbol{\beta}}^{(\pi_1)}}_{\hat{\eta}_{1,i*}(\boldsymbol{\pi})} + \boldsymbol{x}_{i*}^{(\pi_2)T}\hat{\boldsymbol{\beta}}^{(\pi_2)} + \boldsymbol{x}_{i*}^{(\pi_3)}\hat{\boldsymbol{\beta}}^{(\pi_3)} + \tilde{\tilde{\epsilon}}_{i*}$$

$$\hat{\eta}_{2,i*}(\boldsymbol{\pi})$$

by minimizing

$$\sum_{i=1}^{n}\left(y_i - \hat{\eta}_{2,i}(\boldsymbol{\pi}) - \boldsymbol{x}_i^{(\pi_3)T}\boldsymbol{\beta}^{(\pi_3)}\right)^2 + \lambda^{(\pi_3)}\sum_{j=1}^{p_{\pi_3}}|\beta_j^{(\pi_3)}|.$$

Rationale: **Use all** that predictive **information** from the **block with third-highest priority** that is **contained neither in** the block with **highest** nor in that with **second-highest priority**.

- For $m = 4, \ldots, M$ (number of blocks): Using the fitted values from **Step** $m - 1$ **as offset** fit **Lasso** model considering the covariates in the **block with** m**th highest priority**.

- desired **effect**: **Include variables** from each block **only if** they provide **additional information to** that of variables from **blocks with higher priority**;

  possibly **exclusion of whole blocks** ($\Rightarrow$ reduction of costs/efforts)

- **Problem**: Due to **overfitting**, **offsets overestimate information** contained in the blocks.

  $\Rightarrow$ **not all additional information** of blocks with low priority **exploited**

  $\Rightarrow$ sub-optimal prediction performance

  — **Solution: cross-validated offsets**

- Algorithm works **analoguously for other types of outcome** than metric outcomes.

- priority-Lasso **implemented for continuous, binary, and survival outcomes** in the R package prioritylasso, available on CRAN in version 0.2.1

- **two datasets** on **acute myeloid leukemia** (AML):

  1. **training dataset**: Affymetric data, 447 patients
  2. **test dataset**: RNAseq data, 250 patients

- **survival outcome**

- **4 blocks**: established score (ELN2017), clinical covariates, gene mutation data, gene expression data

- **comparison with** standard **Lasso**

- aspects of interest: included variables, prediction accuracy

- included variables:
  - **Priority-Lasso** selected **fewer variables than Lasso**, in particular fewer from blocks with low priority.
    $\Rightarrow$ sparser models ✓
  - As expected priority-Lasso with **cross-validated offsets** selected **more variables**, in particular more variables from blocks **with lower priority**.

- prediction accuracy:
  - **priority-Lasso comparable** to **Lasso**
  - version of priority-Lasso **with cross-validated offsets slightly better**

- **Priority-Lasso** is a prediction method for **multi-omics data** that allows practitioners to impose a **priority sequence among the blocks**.

- Variables from **blocks of lower priority are excluded if** they do **not carry additional information** to that contained in blocks of higher priority.

- This leads to **sparse models**.

- The **prediction performance** is **comparable to** that of standard **Lasso**.

Data:

Büchner, T., Krug, U., Gale, R. P., Heinecke, A., Sauerland, M., Haferlach, C., et al. (2016).
Age, not therapy intensity, determines outcomes of adults with acute myeloid leukemia.
Leukemia **30,** 1781–1784.

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., et al. (2016).
Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel.
Blood **129,** 424–447.

Kreuzer, K.-A., Spiekermann, K., Lindemann, H. W., Lengfelder, E., Graeven, U., Staib, P., et al. (2013).
High efficacy and significantly shortened neutropenia of dose-dense S-HAM as compared to standard double induction: first results of a prospective randomized trial (AML-CG 2008).
Blood **122,** 619.

Priority-Lasso:
prediction
using
multi-omics
data

Roman
Hornung

Introduction

priority-Lasso
estimation
algorithm

Application

Conclusions

priority-Lasso — Technical Report:

📄 Klau, S., Jurinovic, V., Hornung, R., Herold, T.,
Boulesteix, A.L. (2017).
Priority-Lasso: a simple hierarchical approach to the
prediction of clinical outcome using multi-omics data.
Technical Report No. **213,** *Department of Statistics, LMU.*

priority-Lasso — R package prioritylasso:

📄 Klau, S., Hornung, R. (2017).
prioritylasso: Analyzing Multiple Omics Data with an
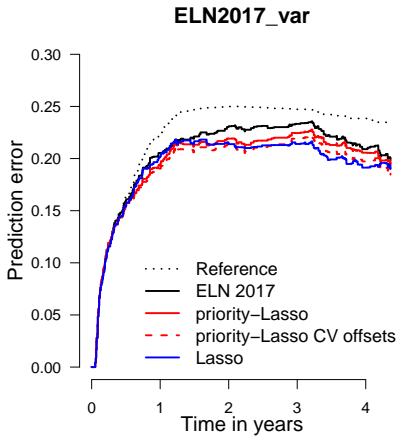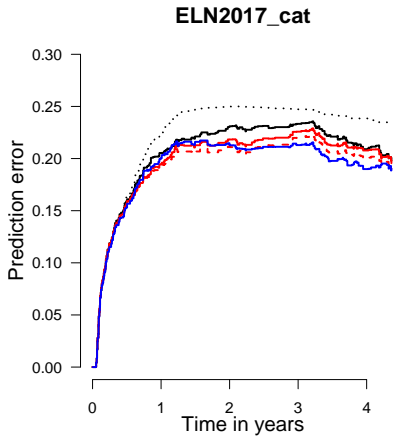Offset Approach.
R package version 0.2.1.
http://cran.r-project.org/web/packages/
prioritylasso/index.html

Priority-Lasso: prediction using multi-omics data

Roman Hornung