technische universität
dortmund

Survival Models built from Gene Expression Data using Gene Groups as Covariates

Kai Kammers, Jörg Rahnenführer
Email: kammers@statistik.uni-dortmund.de

Contents

technische universität
dortmund

- Introduction
  - Combination of gene expression data and survival data
- Statistical Models and Methods
  - Cox Model
  - Penalized Regression Models
  - Cross-validation
  - Evaluation criteria and procedure
- Results
  - Penalized package in **R**
  - Application to leukemia dataset
- Outlook

Introduction

technische universität
dortmund

Goal **Prediction of survival times from gene expression data with high level of interpretability of estimated models**

Motivation

- Models with good prediction accuracy and parsimony property
- **Problem:** Number of genes by far larger than number of observations (individuals) ($p >> n$)
- Use procedures to select genes that are relevant to patient survival and to build a predictive model for future patients
- Classify future patients into clinically relevant high- and low-risk groups based on the gene expression profile and survival times of previous patients

Introduction

technische universität
dortmund

Prediction of survival from expression data

- Many single genes as covariates in survival models
- Dimension reduction through gene selection
- Evaluation of prediction error with suitable measures

Gene group testing

- Define gene groups through Gene Ontology (GO)
- GO groups: Gene expression values are summarized
  (mean, median, maybe other robust measures)
- Identify significant GO groups:
  Analyze and interpret these groups as well as single genes contained in the groups

Cox Model

technische universität dortmund

Cox proportional hazards model for hazard of cancer recurrence or death at time *t*

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 X_1 + ... + \beta_p X_p) = \lambda_0(t)\exp(\beta' X)$$

Estimation of the regression coefficients (in classical setting with $n > p$) by maximizing the log partial likelihood

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta' x_i - \log \left(\sum_{j \in R(t_i)} \exp(\beta' x_j) \right) \right]$$

Methods for Prediction

technische universität dortmund

Univariate selection

- Fit univariate Cox model for each gene/GO group
- Arrange genes/GO groups according to increasing p-values
- Fit multivariate Cox model using λ top ranked genes/GO groups

Penalized Regression

- Lasso Regression (L1 penalty)
  - Penalized log partial likelihood: $l(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|$
- Ridge Regression (L2 penalty)
  - Penalized log partial likelihood: $l(\beta) - \lambda \sum_{j=1}^{p} \beta_j^2$

For all methods, we choose λ via log partial likelihood cross-validation

Cross-validation

technische universität dortmund

Choose tuning parameter λ which maximizes the cross-validated log partial likelihood

$$CVPL(\lambda) = \sum_{k=1}^{K} \left[l(\hat{\beta}_{(-k)}(\lambda)) - l_{(-k)}(\hat{\beta}_{(-k)}(\lambda)) \right]$$

$l(\beta)$	log partial likelihood with all subjects
$l_{(-k)}(\beta)$	log partial likelihood when *k*th fold is left out, *k = 1,…,K*
$\hat{\beta}_{(-k)}(\lambda)$	Estimate of β obtained by a given prediction method when the *k*th fold is left out

Optimal value of λ is chosen to maximize the sum of the contributions of each fold to the log partial likelihood

Evaluation Criteria

technische universität dortmund

Log rank test

- Assign patients to subgroups based on their prognosis, e.g. into one with ‘good’ and one with ‘bad’ prognosis
- Patient *i* in the test set is assigned to the ‘bad’ group if its prognostic index is above the median of the prognostic indices
- Log rank test: use p-value as an evaluation criterion

Prognostic index

- Prognostic index as a single continuous covariate in a Cox model on the test data set
- Likelihood-ratio test: look at p-value to evaluate a method’s performance

Evaluating Procedure

technische universität dortmund

Algorithm (for a fixed prediction method)

- For each of S random splits into training and test data sets
  - Find the optimal tuning parameter $\hat{\lambda}_{train}$ by *K*-fold cross-validation using the training data set
  - Given $\hat{\lambda}_{train}$, estimate the vector of regression coefficients $\hat{\beta}_{train}$ on the whole training data set
  - Calculate the values of the two performance criteria on the test data set

Comparison of performance with boxplots

Dataset: DLBCL data from Rosenwald et al. (2002)

- 7399 gene expression measurements
- 240 patients with diffuse large-B-cell lymphoma (DLBCL)

`penalized` - Package

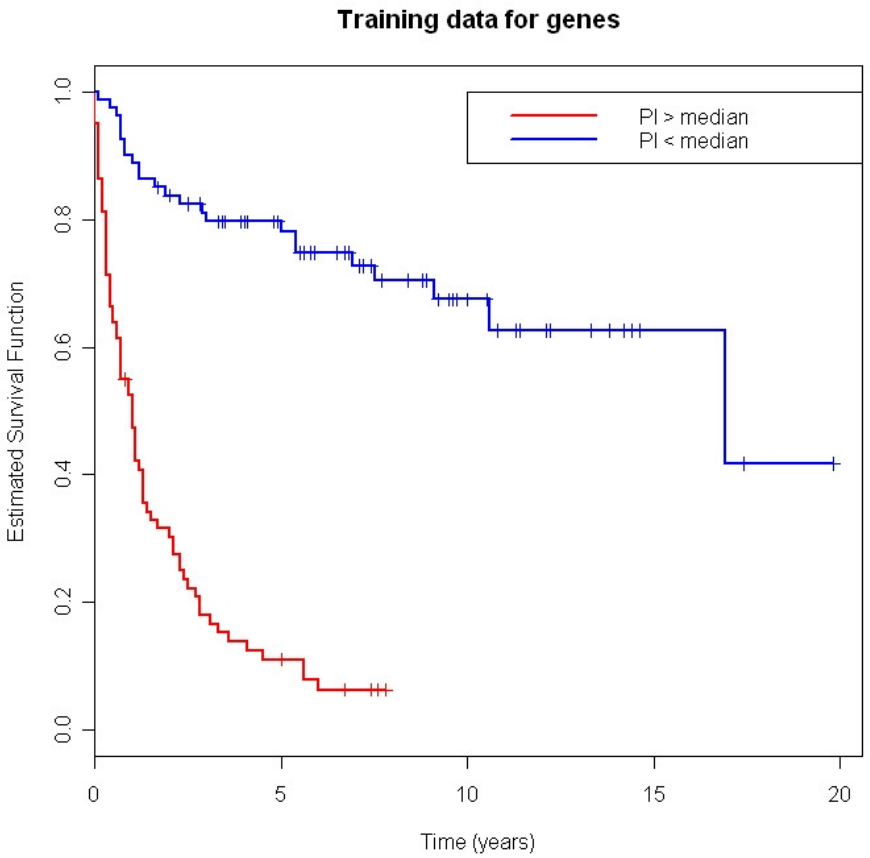
technische universität dortmund

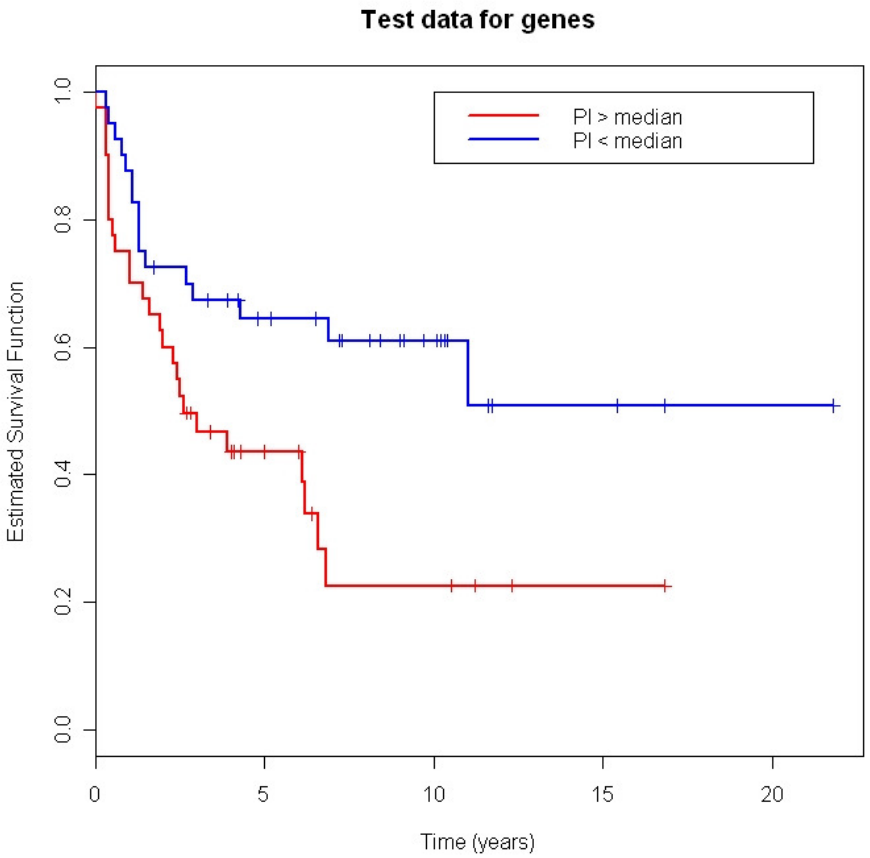
- **penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model**
- A package for fitting possibly high dimensional penalized regression models.
- Penalty structure can be any combination of an L1 penalty (lasso), an L2 penalty (ridge) and a positivity constraint on the regression coefficients.
- Supported regression models are linear, logistic and poisson regression and the **Cox Proportional Hazards model**.
- Cross-validation routines allow optimization of the tuning parameters.
- Version:0.9-21, 2008-04-25, Author: Jelle Goeman

Results

technische universität dortmund

Lasso Regression - one split - median cutoff



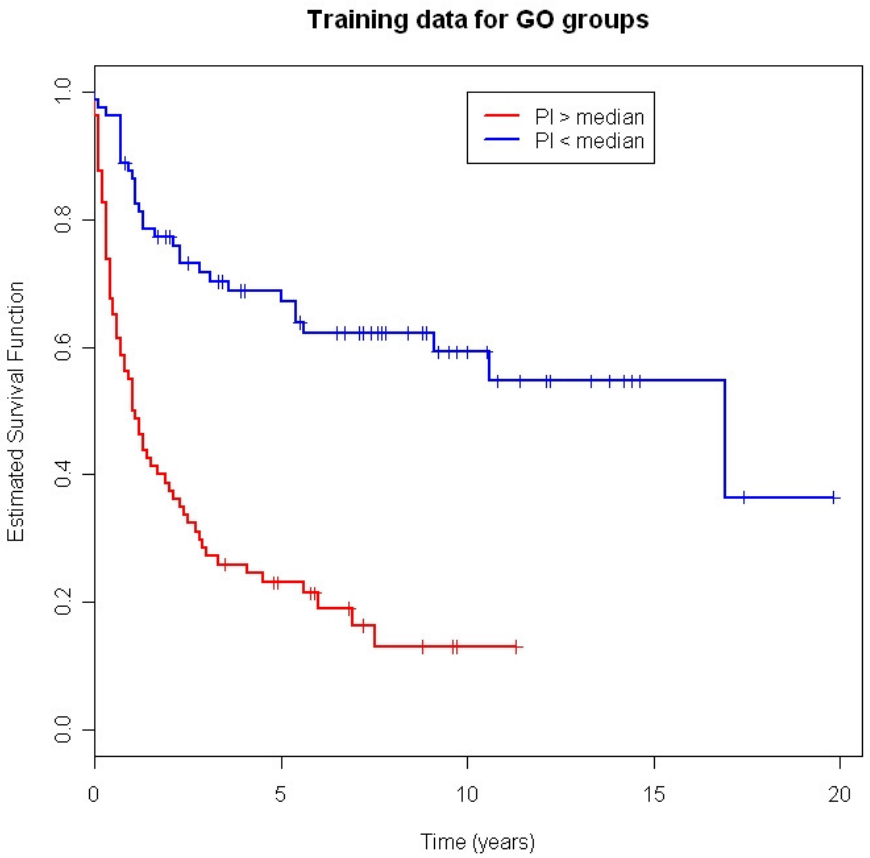


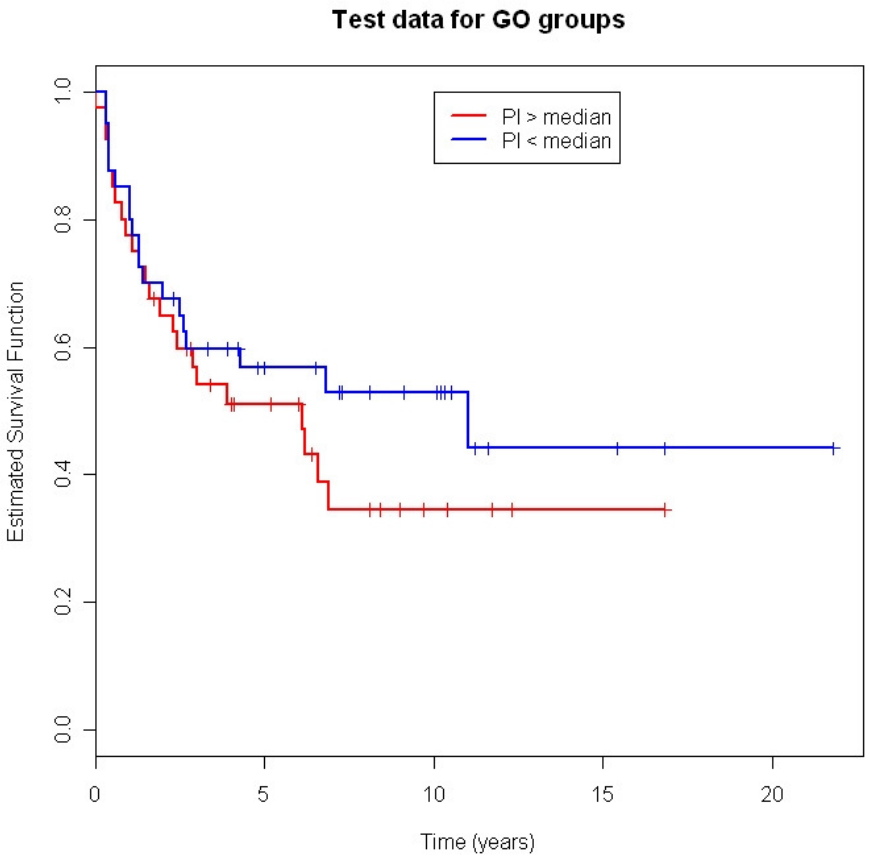
Log-rank test: $p < 10^{-10}$ $p = 0.01$

Results

technische universität dortmund

Lasso Regression - one split - median cutoff



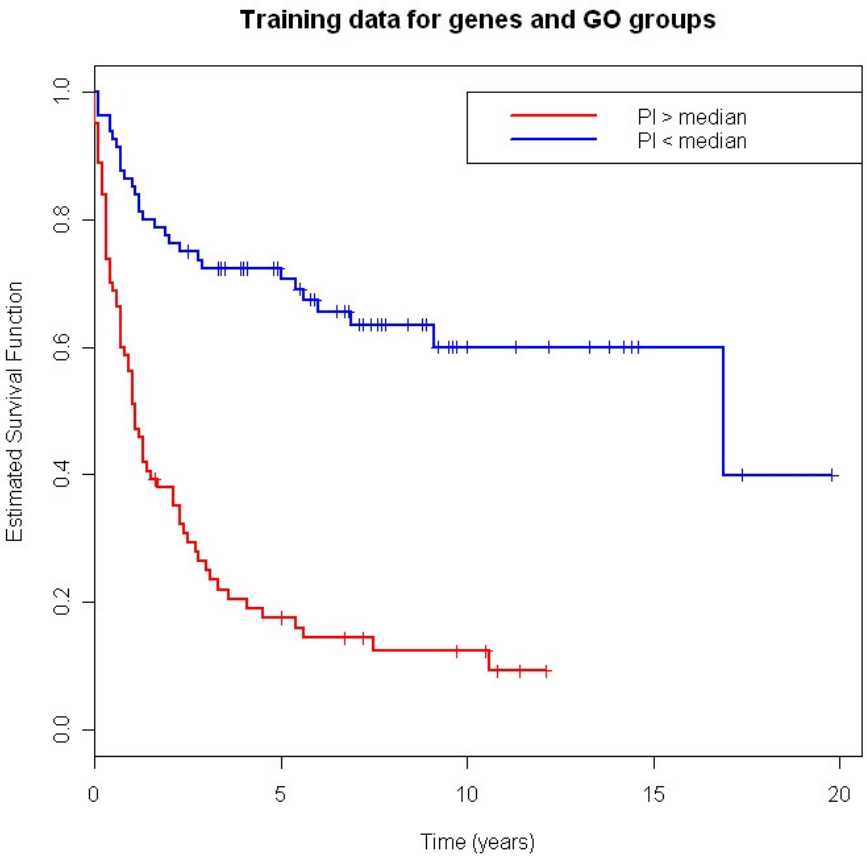


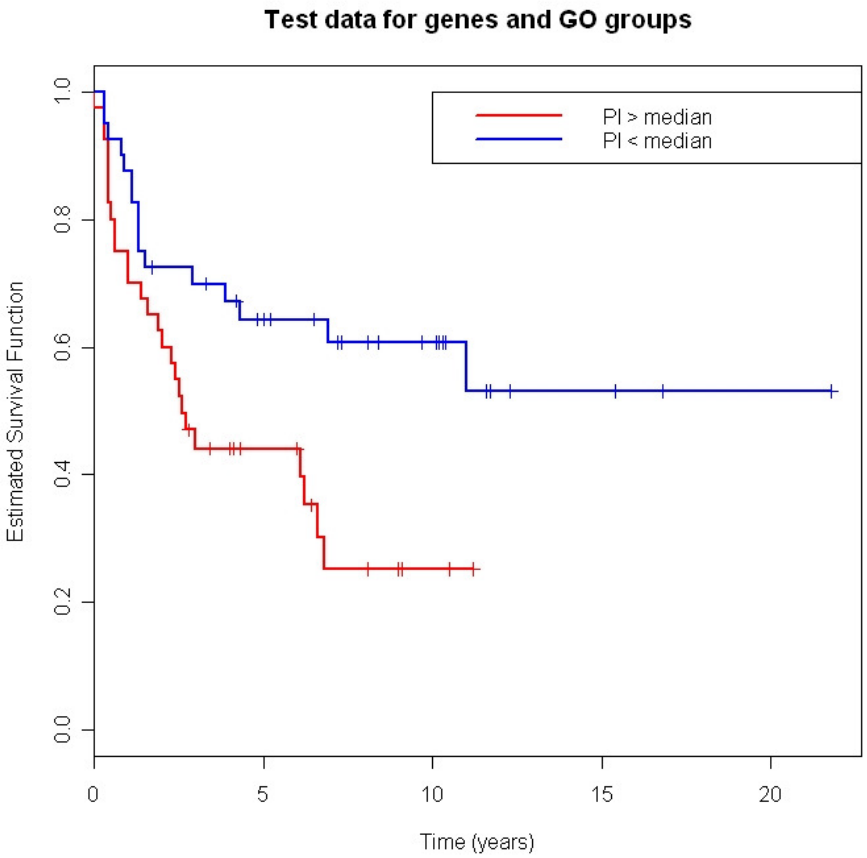
Log-rank test: $p < 10^{-10}$ $p = 0.329$

Results

technische universität dortmund

Lasso Regression - one split - median cutoff





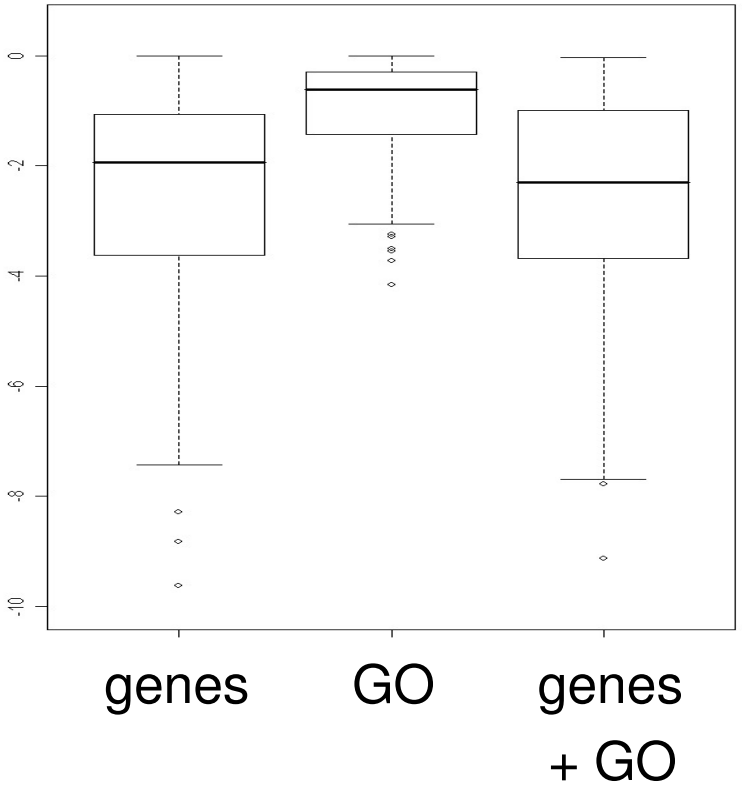
Log-rank test: $p < 10^{-10}$ $p = 0.001$

Results

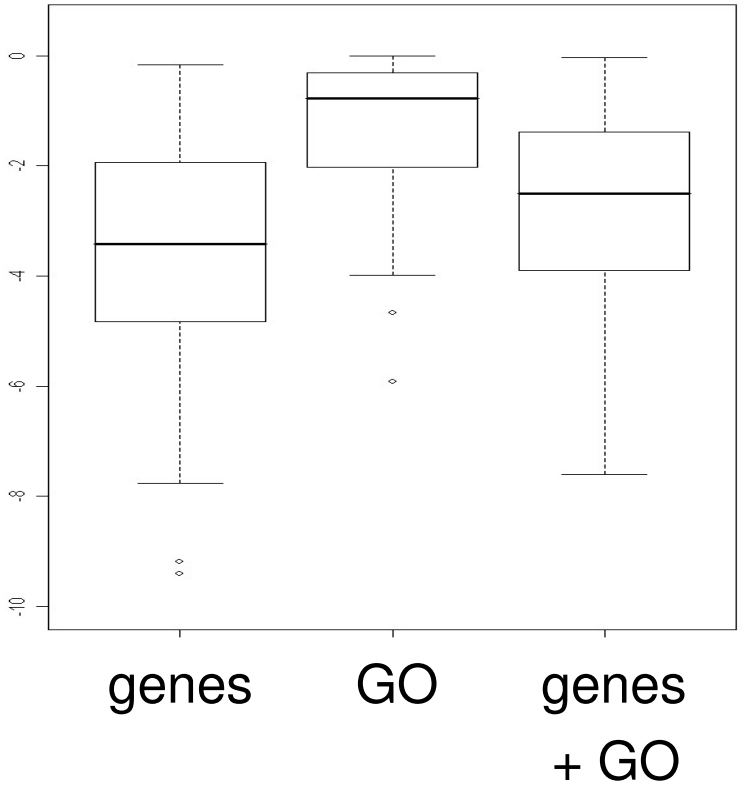
technische universität dortmund

Log-rank test - 100 random splits into training and test data

method: univariate



method: Lasso

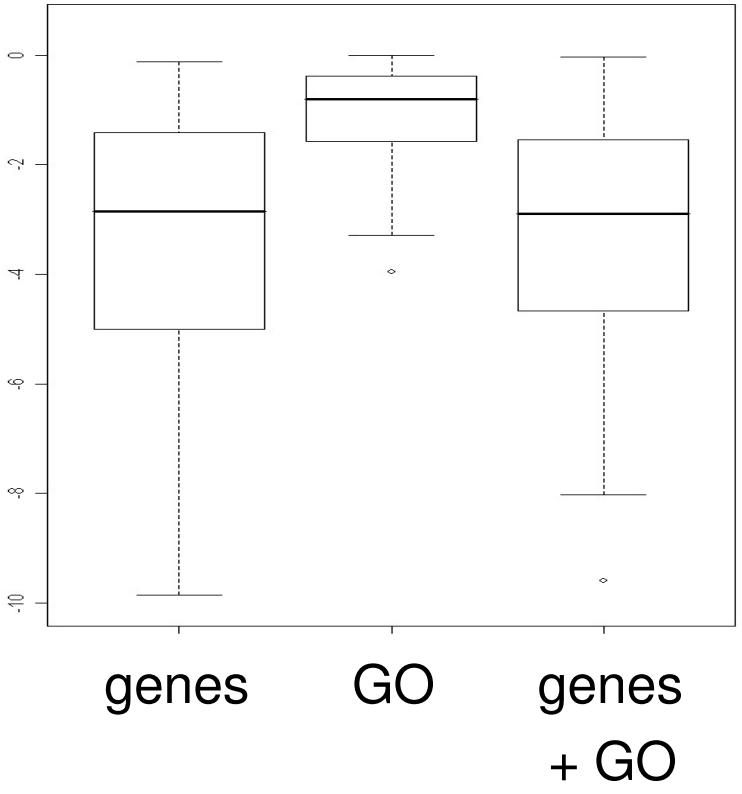


Results

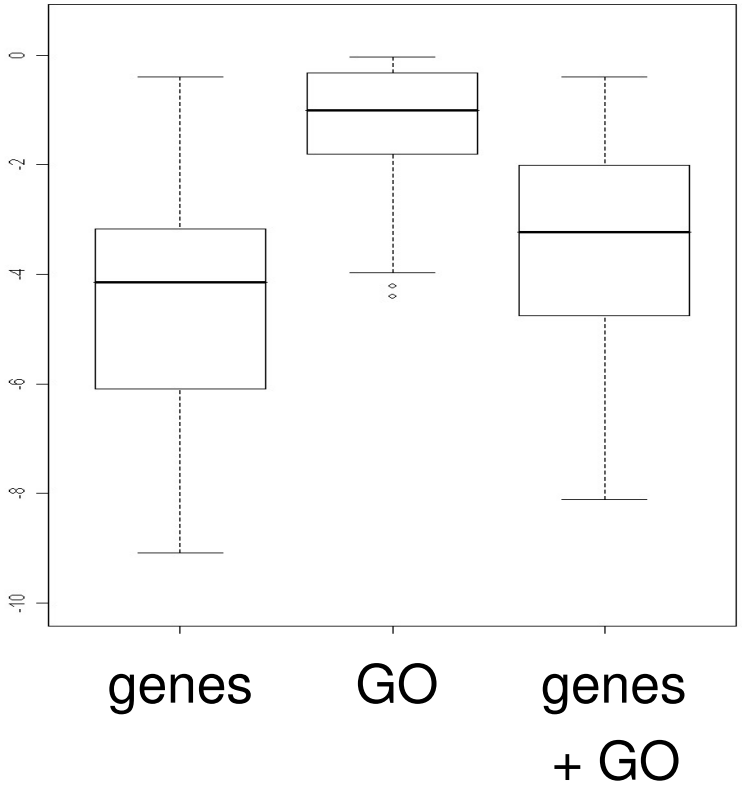
technische universität dortmund

Prognostic Index - 100 random splits into training and test data

method: univariate



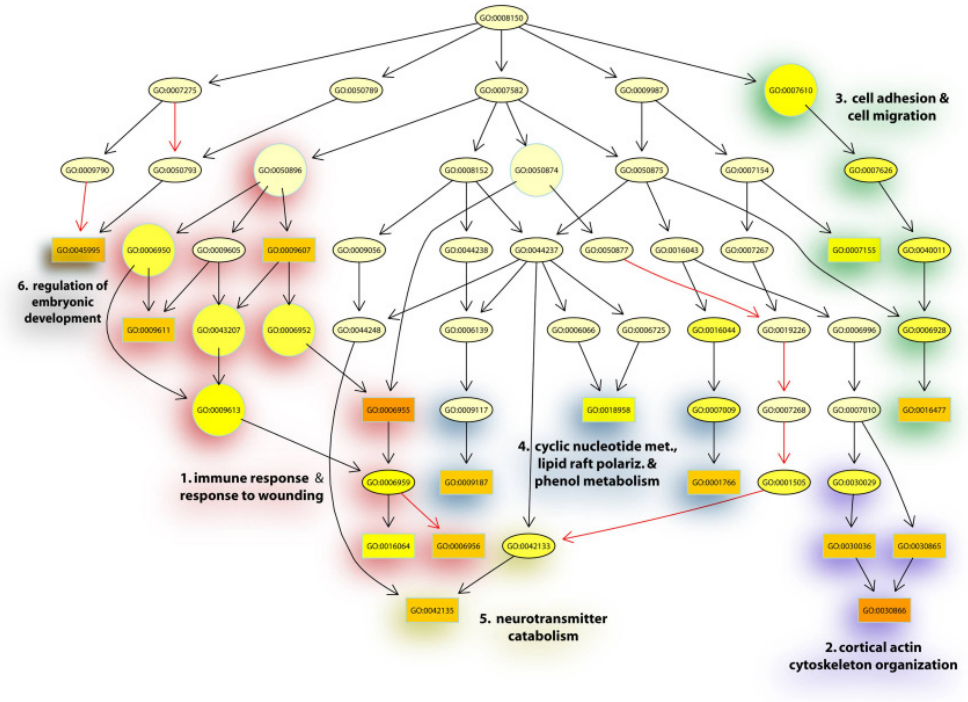
method: Lasso



Outlook

technische universität dortmund

- Additional methods for prediction/evaluation
- Robust measures to summarize gene expression values for one GO group
- Coping with high correlations in GO groups
- Integrate GO graph structure
  - Remove correlations between neighboring GO groups and construct survival models using only significant GO groups
  - Analyze single genes obtained from these GO groups



References

technische universität dortmund

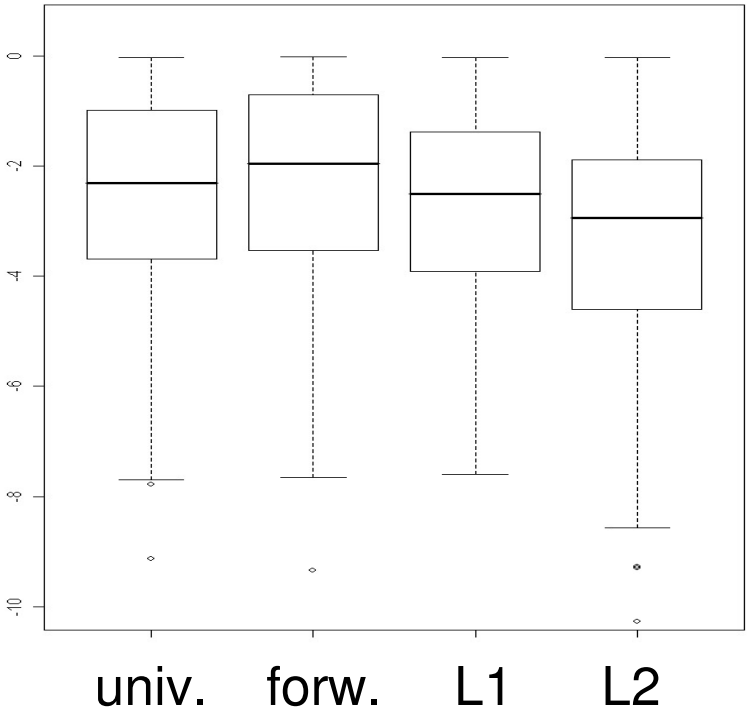
- H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi and O. C. Lingjaerde: **Predicting survival from microarray data - a comparative study,** *Bioinformatics* 23(16): 2080-2087, **2007**
- J. Gui and H. Li: **Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data,** *Bioinformatics* 21(13): 3001-3008, **2005**
- A. Gerds and M. Schumacher: **Efron-Type Measures of Prediction Error for Survival Analysis,** *Biometrics*, Jul **2007**
- GO Consortium: **The Gene Ontology (GO) database and informatics resource,** *Nucleic Acids Research* 32:D258–D261, 2004. Oxford University Press.
- A. Alexa, J. Rahnenführer, T. Lengauer: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure,** *Bioinformatics* 22(13): 1600-1607, **2006**
- W. A. Schulz, A. Alexa, V. Jung, C. Hader, M. J. Hoffmann, M. Yamanaka, S. Fritzsche, A. Wlazlinski, M. Müller, T. Lengauer, R. Engers, A. R. Florl, B. Wullich, J. Rahnenführer: **Factor interaction analysis for chromosome 8 and DNA methylation alterations highlights innate immune response suppression and cytoskeletal changes in prostate cancer,** *Molecular Cancer* 6:14, **2007**

Results

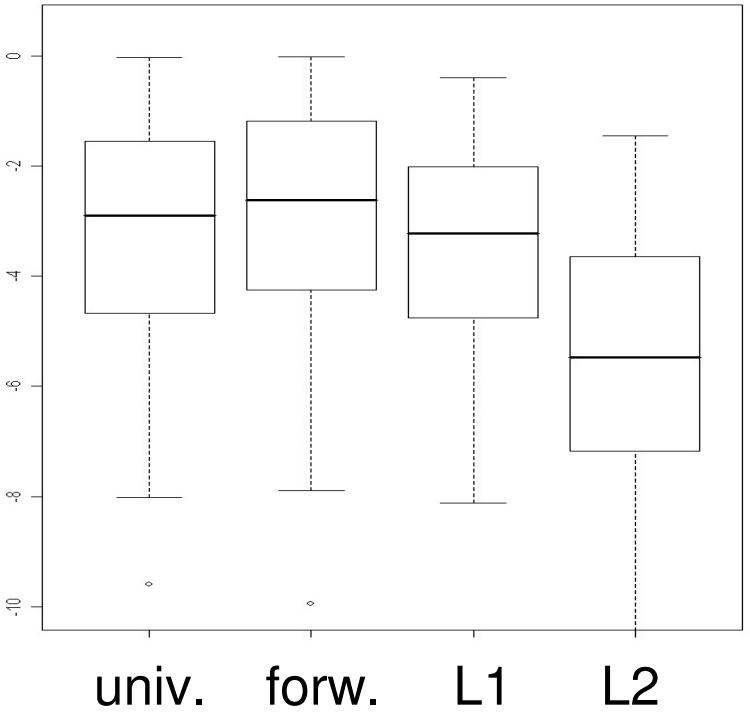
technische universität
dortmund

All methods - 100 random splits into training and test data

log-rank test



prognostic index



Results

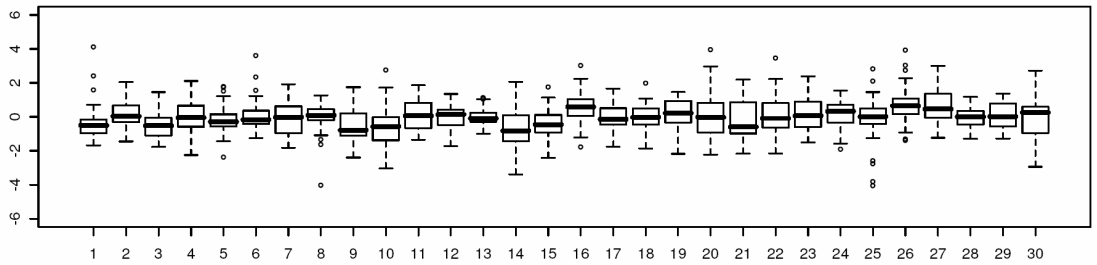
technische universität dortmund

10 most significant GO groups (univariate selection, one split)

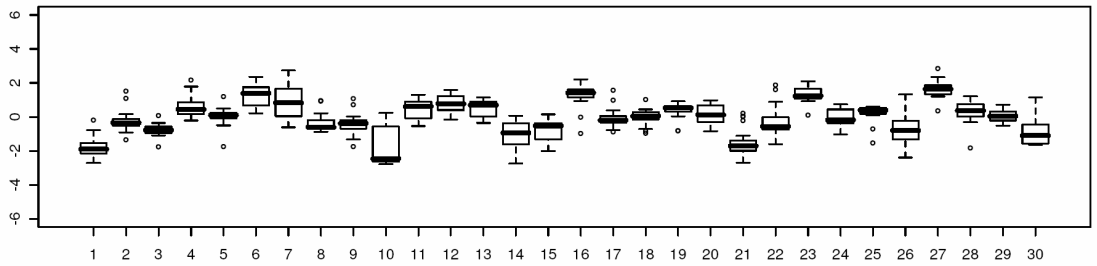
GO Group	P-value	P-value adjusted	#Genes	Function
01562	0.00049	0.47	3	Response to protozoan
40012	0.00053	0.47	2	Regulation of locomotion
40029	0.00142	0.54	4	Regulation of gene expression, epigenetic
30149	0.00149	0.54	3	Sphingolipid catabolic process
51310	0.00151	0.54	5	Metaphase plate congression
01816	0.00312	0.63	19	Cytokine production
50764	0.00359	0.63	6	Regulation of phagocytosis
21700	0.00363	0.63	11	Development maturation
30282	0.00366	0.63	1	Bone mineralization
02268	0.00370	0.63	7	Fillicular dendritic cell differentiation

Results

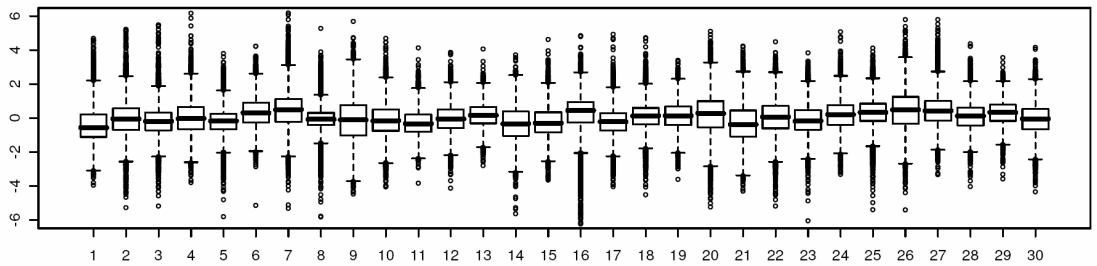
technische universität
dortmund



Non-significant GO group



Significant GO group



All genes