

# C060 Vignette

Martin Sill, Thomas Hielscher, Natalia Becker, Manuela Zucknick

---

## Abstract

*Keywords:*

---

## 1. Introduction

Penalized regression models provide a statistically appealing method to build prediction models from high-dimensional data sources. Since the introduction of the lasso for linear regression models (Tibshirani, 1996), the methodology has been extended to generalized linear regression models, time-to-event endpoints (Tibshirani, 1997) etc. In addition to the well-known  $L_1$ - (lasso) and  $L_2$ -norms (ridge) penalty functions, various other penalties have been proposed in recent years to select features and/or estimate their effects. In particular, we will use the elastic net penalty function (Zou and Hastie, 2005), which is a linear combination of the  $L_1$ - and  $L_2$ -norms.

With ever increasing data, the properties of the algorithm to actually fit the model have become almost as important as the statistical model itself. In 2010, Friedman, Hastie and Tibshirani proposed a coordinate descent algorithm (Friedman et al., 2010) for generalized linear regression models, which has since then been extended to penalized Cox PH regression models (Simon et al., 2011). Due to its efficiency this algorithm is considered one of the state-of-the-art approaches to estimate penalized regression models with lasso, ridge or elastic net penalty terms.

This algorithm has also been implemented in R in the `glmnet` package. The package provides functions to tune and fit regression models, plot the results, and make predictions. However, in practical applications, where often an independent validation data set is lacking, some additional features and routines are desirable as part of a complete data analysis. We have assembled some functions that enhance the existing functionality of the `glmnet` package or allow to use it within the framework of other existing R packages. These functions have been useful in our daily work here at

the German Cancer Research Center where prognostic modelling of patient survival data is of particular interest. Therefore, for illustration purposes we focus on penalized Cox PH regression models in this article. But most of the functions are applicable to all types of regression models implemented in the `glmnet` package.

We provide R functions to perform stability selection (Meinshausen and Bühlmann, 2010) in a computationally efficient way using `glmnet` which allows to select the most stable features at a given error level. We also provide an approach to select the optimal parameter combination  $(\alpha, \lambda)$  for elastic net penalties using an interval-search algorithm (Froehlich and Zell, 2005) which is often faster and more accurate than a standard grid search (Ref for a comparison study?). Another very useful addition for real-life applications of `glmnet` for building prognostic models is the provision of wrapper functions to allow the computation of resampling-based prediction error curves with the framework of the R package `peperr` (Porzelius et al., 2009). The `peperr` package makes it computationally feasible to assess the predictive accuracy of a penalized Cox PH regression model via resampling methods even for very large-scale applications by employing parallel computing. We also provide the possibility to speed up stability selection by parallel computing using the functionalities of the R base package `parallel`.

## 2. Data application

Throughout this article we use the gene expression data set on cytogenetically normal acute myeloid leukemia (CN-AML) by Metzeler et al. (2008) and corresponding clinical data in order to illustrate a typical application of penalized Cox PH regression models with the aim of developing a prognostic model for patient survival while at the same time identifying the most influential gene expression features. To simulate the typical situation that only one data set is available for model training and evaluation, we only use the data set that was used as validation data in the original publication by Metzeler et al. (2008). The data can be accessed from the Gene Expression Omnibus (GEO) data repository (<http://www.ncbi.nlm.nih.gov/geo>) by the National Center for Biotechnology Information (NCBI). We find the Metzeler *et al.* data set under GEO accession number GSE12417.

The data set contains gene expression data for 79 patient samples measured with Affymetrix HG-U133 Plus 2.0 microarrays. The median survival time of these 79 patients was 17.6 months with 40% censoring.

### 3. Methods and algorithms

*[Manuela: What we need here is not just the description of the statistical methods, but also of the specific algorithms that are implemented in the R package.]*

#### 3.1. Penalized generalized linear models and Cox models

An efficient implementation for fitting generalized linear models and Cox proportional hazards models with regularization by the lasso or elastic net penalty terms is provided by the R package `glmnet` (Friedman et al., 2010; Simon et al., 2011). This implementation uses a coordinate descent algorithm for fitting the models for a specified penalty parameter value  $\lambda$ . The computation of an entire regularization path across a range of values  $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$  with `glmnet` is very fast, because previously computed solutions for  $\{\lambda_1, \dots, \lambda_{j-1}\}$  are used as a 'hot' starting value for the computation of  $\lambda_j$ .

Models are fitted by maximizing the penalized log-likelihood function for generalized linear models and the penalized partial log-likelihood for Cox models. Cross-validation can be performed to decide which model (i.e. which penalty parameter values) to choose by using the cross-validated (partial) log-likelihood as the loss function. *[Manuela: Is the approach by Verweij and van Houwelingen (1993) used for Cox models?]* *[Manuela: The remainder of the text in this section 3.1 is from the Benner et al. (2010) paper and needs to be rewritten and adapted to this manuscript.]*

The penalized (partial) log-likelihood is given by

$$l_n(\beta) - \sum_{j=1}^p p_{\lambda^*}(|\beta_j|) \quad (1)$$

where  $l_n(\beta)$  denotes the (partial) log-likelihood given  $n$  observations. The dimension of the parameter vector  $\beta$  is  $p$  and  $p_{\lambda}(|\cdot|)$  is the penalty function with tuning parameter  $\lambda^*$ . Correspondingly, the objective function employed in `glmnet` is (with  $\lambda = \lambda^*/n$ ):

$$-l_n(\beta)/n + \sum_{j=1}^p p_{\lambda}(|\beta_j|). \quad (2)$$

### 3.2. $L_2$ -penalized Cox regression

Penalized maximum likelihood estimation in Cox regression with the ridge penalty

$$p_\lambda(|\beta|) = \lambda\beta^2 \quad (3)$$

was introduced by Verweij and van Houwelingen (1994). Ridge penalized regression does not perform model selection and results in downwardly biased parameter estimates. On the other hand, it has been found to produce models with good prediction performance in high-dimensional genomic applications (e.g. Bøvelstad et al., 2007).

### 3.3. $L_1$ -penalized Cox regression

Tibshirani (1997) proposed to use an  $L_1$ -penalized Cox model with

$$p_\lambda(|\beta|) = \lambda|\beta| \quad (4)$$

and described a technique, called the lasso for "least absolute shrinkage and selection operator", for parameter estimation. The  $L_1$ -penalty has the advantage over the  $L_2$ -penalty of shrinking some of the coefficients to zero, i.e. it has an in-built variable selection procedure.

### 3.4. The elastic net

Zou and Hastie (2005) introduced the elastic net, which employs a combination of the  $L_1$ - and  $L_2$ -penalty

$$p_{\lambda_1, \lambda_2}(|\beta|) = \lambda_1|\beta| + \lambda_2\beta^2. \quad (5)$$

Zou and Hastie (2005) rescale the initial solutions from the optimization of the doubly-penalized log-likelihood function by the factor  $1 + \lambda_2$ , in order to reduce the effect of the double shrinkage. The elastic net performs variable selection and parameter estimation equivalently to the lasso. But the additional  $L_2$ -penalty term distributes the weight to more variables, such that the elastic net tends to select more variables than the lasso. This is especially the case in situations with high correlation, where the lasso would select only one variable of a set of highly correlated variables.

Throughout this manuscript we use an alternative parametrization of the elastic net penalty function equivalently to the formulation used in the `glmnet` package:

$$p_{\alpha, \lambda}(|\beta|) = \lambda \times ((1 - \alpha)\frac{1}{2}|\beta|^2 + \alpha|\beta|). \quad (6)$$

- 3.4.1. The interval-search algorithm to select the optimal elastic net parameter combination*
- 3.5. Stability selection*
- 3.6. Prediction error curves for survival models*

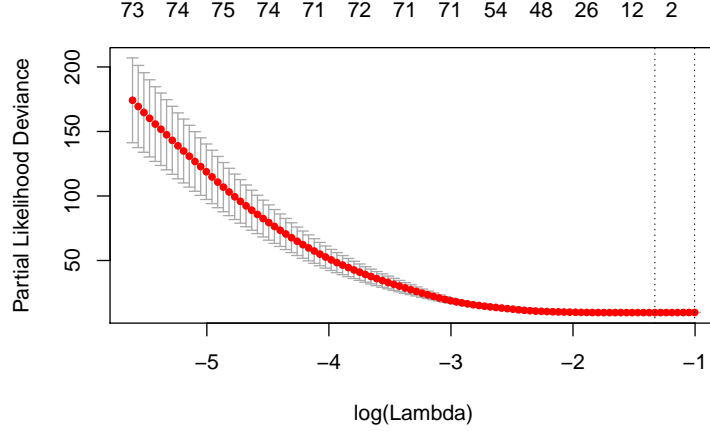


Figure 1: Cross-validated partial log-likelihood function, including upper and lower standard deviations, as a function of  $\log \lambda$  for the AML data set.

## 4. Application and demonstration of software

### 4.1. Starting off: Lasso-penalised Cox model

We can apply the `glmnet` function to fit a lasso-penalized Cox model to the CN-AML data set. The function call with default penalty parameter settings will fit the lasso model for 100  $\lambda$  data derived values:

```
> fit <- glmnet(y=y, x=t(exprs(eset)), family="cox")
```

We tune the lasso penalty parameter by 10-fold cross-validation using the `cv.glmnet` function. The loss function, i.e. the cross-validated partial log-likelihood, is shown in Figure 1 including upper and lower standard deviations as a function of  $\log \lambda$  for the AML data set. The penalty parameter value minimizing the loss function is  $\lambda = 0.265$  and corresponds to a final lasso model with the following 5 selected features:

```
203640_at 204419_x_at 222462_s_at 226169_at 233371_at
-0.11339033 -0.01664530 0.27420521 0.04300559 -0.01216429
```

The selected features are highlighted as red lines in the coefficient paths shown in Figure 2. The coefficient path shows the development of the regression coefficient estimates with increasing regularization.

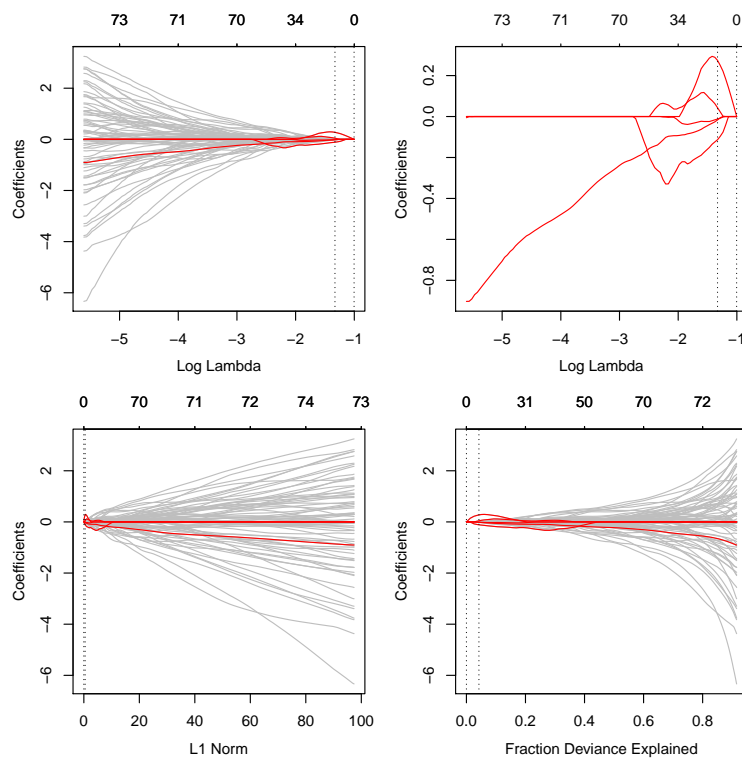


Figure 2: Coefficient paths for lasso penalised Cox PH regression model applied to the AML data set.

#### 4.2. Resampling based prediction errors

Once the final prognostic model is selected, we need to assess its prediction accuracy for future patients, frequently also in comparison with established clinico-pathological prognostic markers. In many applications no independent validation data set is available. The same data set need to be used to develop and assess the prognostic model. This is even more problematic for high-dimensional data, where the risk of overfitting is much more present. Resampling-based methods can be used to unbiasedly estimate the predictive accuracy of the prognostic model in this situation. This is also called internal validation or pre-validation.

The R package **peperr** (Porzelius et al., 2009) provides a modular framework for survival and binary endpoints, i.e. prognostic and classification models. Wrapper functions for new or customized prediction model algorithms can be defined and passed to the generic call function **peperr**. In case of prognostic models, algorithm specific wrapper functions for model fitting, tuning and prediction are required. Wrapper functions for selected machine learning approaches are already implemented.

Prediction accuracy is per default assessed with prediction error curves based on the time-dependent Brier score (Graf et al., 1999). But it is also possible to define and use customized accuracy measures.

We defined additional wrapper functions for the **glmnet** algorithm for fitting (**fit.glmnet**) and tuning (**complexity.glmnet**) the model, and predicting survival probabilities (**predictProb.glmnet**) based on the fitted model and the estimated baseline hazard from the training data.

We estimate the  $L_1$ -penalized Cox PH regression model for overall survival starting with the 10.000 most varying probe sets using **glmnet**. The .632+ bootstrap estimator is calculated based on subsampling (Binder and Schumacher, 2008) using only 20 bootstrap samples for illustration.

```
> peperr_obj <- peperr(response=Surv(eset$os, eset$os_status), x=t(exprs(eset)),
+                       fit.fun=fit.glmnet, args.fit=list(standardize=F, family="cox"),
+                       complexity=complexity.glmnet, args.complexity=list(standardize
+                       trace=F, RNG="fixed", seed=0815,
+                       indices=resample.indices(n=dim(eset)[2], sample.n = 20, method
```

Individual bootstrap results can be visualized with the **plot.peperr** function from the **peperr** package showing the selected complexity parameters, out-of-bag prediction error curves as well as the prediction error integrated over time, and the predictive partial log-likelihood (PLL) values.



```
> plot.peperr.curves(peperr_obj, at.risk=T)
```

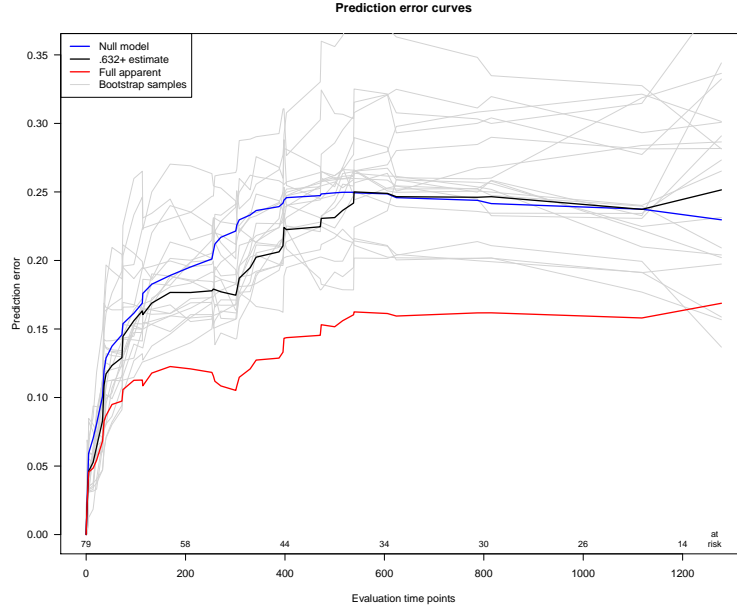


Figure 3: Prediction error curves

In order to calculate the predictive PLL values again an algorithm specific wrapper (here `PLL.coxnet`) needs to be defined.

In addition, we provide a slightly modified version of the prediction error curves plot function from the `peperr` package which allows to display the number still at risk (`plot.peperr.curves`) as shown in figure 3.

The `peperr` package is designed for high-dimensional covariates data and allows for various types of parallel computations. Here, we re-run the calculations on 3 CPUs in parallel using a socket cluster on a Windows OS.

```
> peperr_obj_parallel <- peperr(response=Surv(eset$os, eset$os_status), x=t(exprs(eset
+                                     fit.fun=fit.glmnet,args.fit=list(standardize=F, famil
+                                     complexity=complexity.glmnet, args.complexity=list(st
+                                     trace=F, RNG="fixed",seed=0815, cpus=3, parallel=T, c
+                                     load.list=list(functions=c("basesurv")),
+                                     indices=resample.indices(n=dim(eset)[2], sample.n = 2
```

Additional arguments can be passed directly to the `glmnet` call by specifying additional arguments for the fitting or tuning procedure. Here, we in-

clude patient's age as mandatory model variable into the prognostic model, i.e. age is not subject to penalization.

```
> peperr_obj_parallel_mandatory <- peperr(response=Surv(eset$os, eset$os_status), x=
+                                     fit.fun=fit.glmnet, args.fit=list(standard
+                                     penalty.
+                                     complexity=complexity.glmnet, args.comple
+                                     trace=F, RNG="fixed", seed=0815, cpus=3, p
+                                     load.list=list(functions=c("basesurv")),
+                                     indices=resample.indices(n=dim(eset)[2],
```

For classification models, the same wrapper functions for fitting and tuning the model are called. Model performance measures shipped with the **peperr** packages are misclassification rate and Brier score.

We extended functionality of the Brier score (**aggregation.brier**) and misclassification rate (**aggregation.misclass**) calculation for the **glmnet** algorithm, and defined AUC under the ROC curve (**aggregation.auc**) as additional performance measure. For binary responses, the **peperr** package does not quite provide the same modular flexibility as for time-to-event endpoints. The predicted class probability is calculated within the performance/aggregation function by calling the algorithm specific predict function. Whenever a new algorithm is incorporated the aggregation function has to be modified and overwritten accordingly.

For illustration purpose only we use survival status as binary response variable and build the  $L_1$ -penalized logistic regression model with **glmnet**. All three prediction accuracy measures are calculated. Again, computations are run in parallel. Results of 20 bootstrap runs are displayed in plain boxplots (figure 4).

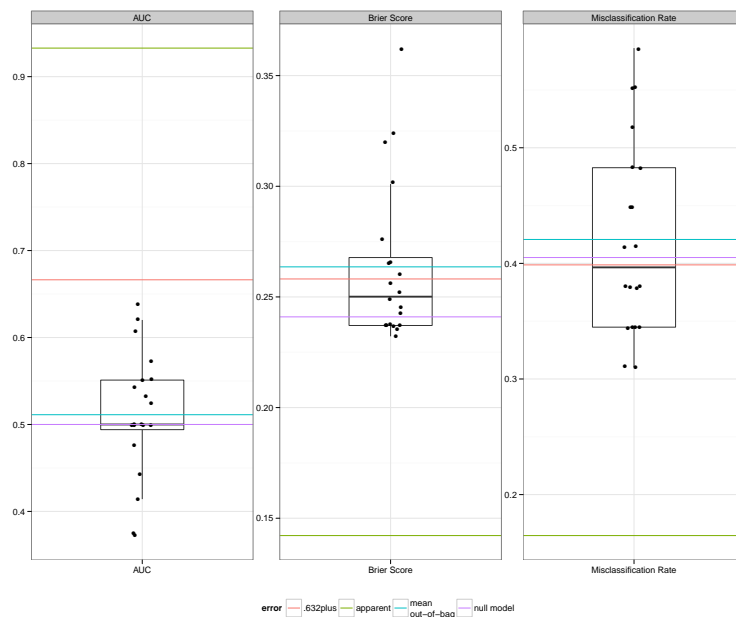


Figure 4: Different bootstrap performance measures for binary classification

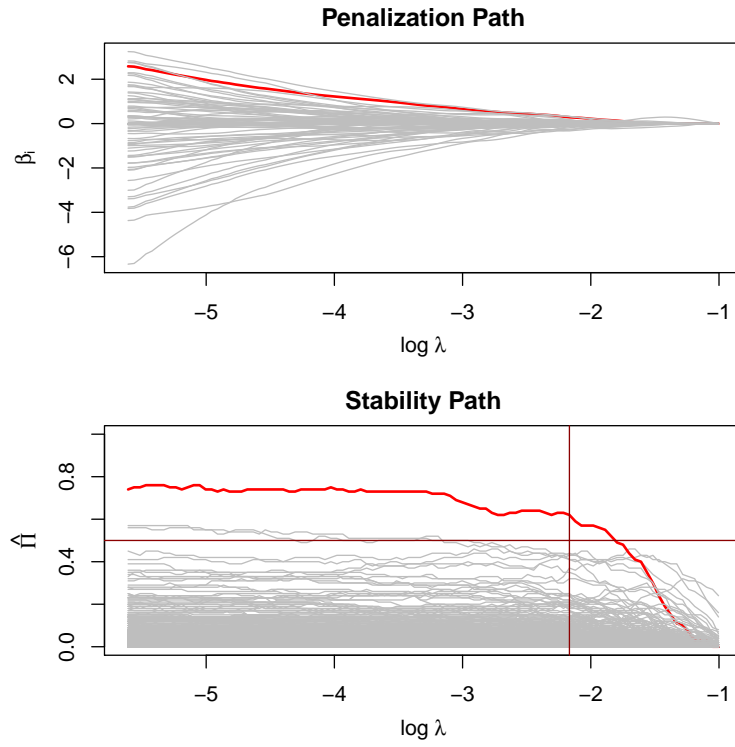


Figure 5: Coefficient and stability paths for lasso penalised Cox PH regression model applied to the AML data set.

#### 4.3. Stability selection

Stable features (with  $\hat{\pi} > 0.5$  at  $\lambda = 0.115$ ) are:

206932\_at  
2823

Table 1: Summary of visited points in the parameter space

|    | alpha   | lambda    | deviance | n.features |
|----|---------|-----------|----------|------------|
| 1  | 0.67605 | 0.5421144 | 9.71152  | 0          |
| 2  | 0.17194 | 0.5174744 | 9.66298  | 1          |
| 3  | 0.81895 | 0.4939544 | 9.70872  | 1          |
| 4  | 0.31188 | 0.4715034 | 9.69917  | 2          |
| 5  | 0.61671 | 0.4500729 | 9.70060  | 2          |
| 6  | 0.47035 | 0.4296164 | 9.70494  | 4          |
| 7  | 0.51860 | 0.4100896 | 9.70339  | 6          |
| 8  | 0.63074 | 0.3914504 | 9.70020  | 9          |
| 9  | 0.05830 | 0.3736584 | 9.64177  | 12         |
| 10 | 0.10655 | 0.3566751 | 9.65779  | 14         |

#### 4.4. Elastic net penalised Cox PH regression model

##### 4.4.1. Tune both $\lambda$ and $\alpha$ by interval search

The task is to find such a setting of tuning parameters  $(\alpha, \lambda)$ , for which the 10-fold cross validation error of the model is minimal. Instead of using a fixed grid, an interval search approach is applied. REFF

The parameter space of tuning parameters is defined as follows:

```

lower upper
alpha    0    1

```

The second tuning parameter  $\lambda$  will be found for each given  $\alpha$  via regularization path. Thus, the two dimensional parameter space has the form  $(0, 1) \times \mathbb{R}$ . The seed is 1234. For each given  $\alpha$  an optimal lambda is defined as largest value of lambda such that error is within 1 standard error of the minimum. (parameter type.min = 'lambda.1se').

The visited points in the parameter space  $(\alpha, \lambda)$  and resulting cross-validated deviances are summarised in table 1.

The minimal mean cross-validated deviance over the folds of 9.633 is reached for optimal parameter pair  $(\alpha, \lambda) = (0.013, 28.235)$ .

The 'visited' points with corresponding deviance and number of selected features in each model are presented in figure 1.

```
> summary(cofn)
```

```

      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
-0.0941500 -0.0071050 -0.0001162 -0.0002743  0.0064470  0.0913500

```

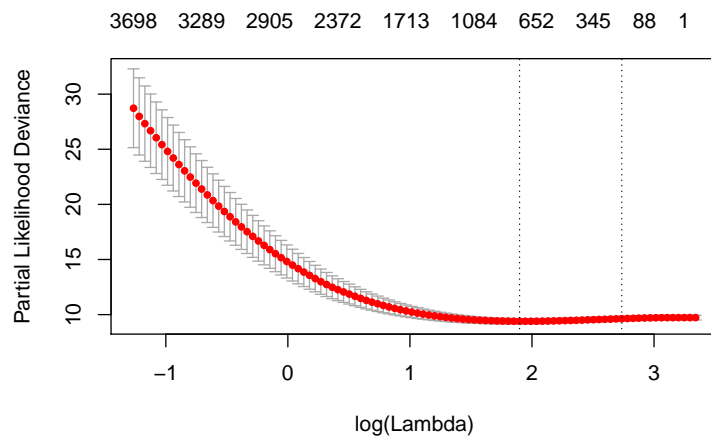


Figure 6: deviance as a function of  $\log(\alpha)$  for the AML data set.

```
> head(sort(cofn))
```

```
      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
-0.09415166 -0.09328023 -0.09295101 -0.09193752 -0.09130928 -0.09004991
```

```
> tail(sort(cofn))
```

```
      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
0.08676626 0.08727578 0.08815779 0.08842483 0.09008347 0.09134615
```

A feature selected by the stability algorithm is in the set of selected features

```
> '206932_at' %in% names.cof
```

```
[1] TRUE
```

**TODO : plots**

## 5. Conclusions

## 6. Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 2.14.1 (2011-12-22), x86\_64-apple-darwin9.8.0
- Locale: C/en\_US.UTF-8/C/C/C/C
- Base packages: base, datasets, grDevices, graphics, methods, parallel, splines, stats, tools, utils
- Other packages: Biobase 2.14.0, GEOquery 2.21.9, MASS 7.3-18, MCRestimate 2.10.0, Matrix 1.0-5, cacheSweave 0.6-1, class 7.3-3, cluster 1.14.2, codetools 0.2-8, corpcor 1.6.2, e1071 1.6, filehash 2.2-1, genefilter 1.36.0, ggplot2 0.9.0, glmnet 1.7.3, golubEsets 1.4.10, lattice 0.20-6, lhs 0.7, limma 3.10.3, mlegp 3.1.2, pamr 1.54, penalizedSVM 1.1, peperr 1.1-6, snow 0.3-9, snowfall 1.84, stashR 0.3-5, statmod 1.4.14, survival 2.36-14, tgp 2.4-7, xtable 1.7-0
- Loaded via a namespace (and not attached): AnnotationDbi 1.16.19, DBI 0.2-5, IRanges 1.12.6, RColorBrewer 1.0-5, RCurl 1.91-1, RSQLite 0.11.1, XML 3.9-4, annotate 1.32.3, colorspace 1.1-1, dichromat 1.2-4, digest 0.5.2, grid 2.14.1, memoise 0.1, munsell 0.3, plyr 1.7.1, proto 0.3-9.2, randomForest 4.6-6, reshape2 1.2.1, scales 0.2.0, stringr 0.6

## References

- Binder, H., Schumacher, M., 2008. Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology* 7.
- Bøvelstad, H.M.M., Nygård, S., Størvold, H.L.L., Aldrin, M., Borgan, O., Frigessi, A., Lingjærde, O.C.C., 2007. Predicting survival from microarray data - a comparative study. *Bioinformatics* 23, 2080–2087.
- Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33, 1–22.
- Froehlich, H., Zell, A., 2005. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization, in: *Proceedings of the International Joint Conference of Neural Networks*, pp. 1431–1438.
- Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M., 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18, 2529–2545.
- Meinshausen, N., Bühlmann, P., 2010. Stability selection. *Journal of the Royal Statistical Society - Series B: Statistical Methodology* 72, 417–473.
- Metzeler, K., Hummel, M., Bloomfield, C., Spiekermann, K., Braess, J., Sauerland, M.C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S., Maharry, K., Paschka, P., Larson, R., Berdel, W., Buchner, T., Wormann, B., Mansmann, U., Hiddemann, W., Bohlander, S., Buske, C., 2008. An 86 probe set gene expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* 112(10), 4193–4201.
- Porzelius, C., Binder, H., Schumacher, M., 2009. Parallelized prediction error estimation for evaluation of high-dimensional models. *Bioinformatics* 25, 827–829.
- Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2011. Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software* 39, 1–13.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological* 58, 267–288.



- Tibshirani, R., 1997. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 16, 385–395.
- Verweij, P.J.M., van Houwelingen, H.C., 1993. Cross-validation in survival analysis. *Statistics in Medicine* 12, 2305–2314.
- Verweij, P.J.M., van Houwelingen, H.C., 1994. Penalized likelihood in Cox regression. *Statistics in Medicine* 13, 2427–2436.
- Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67(2), 301–320.