C060 Vignette

Martin Sill, Thomas Hielscher, Natalia Becker, Manuela Zucknick

1 Introduction

Data from published gene expression studies are often deposited in public data repositories, for example on the Gene Expression Omnibus (GEO) website by the NCBI (National Center for Biotechnology Information): `http://www.ncbi.nlm.nih.gov/geo`. We find the Metzeler *et al.* data under GEO accession number GSE12417.

Here should follow a detailed description of

1. the problem (what do we want to do)
2. the existing methods and R software (what exists in glmnet package and which functions are missing)
3. the data set

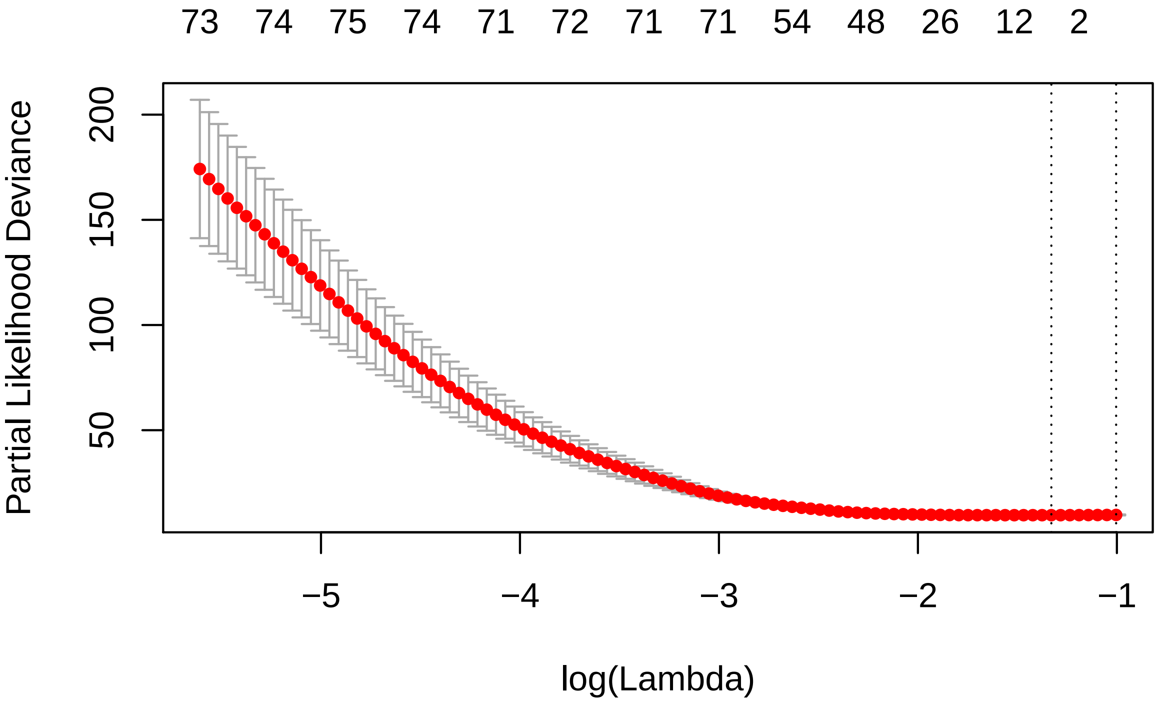

Figure 1: Cross-validated partial likelihood function, including upper and lower standard deviations, as a function of $\log \lambda$ for the AML data set.

2 Lasso penalised Cox PH regression model

We tune the lasso penalty parameter by 10-fold cross-validation using the cross-validated partial log-likelihood function as the loss function. The resulting penalty parameter value leads to a final lasso model with 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.

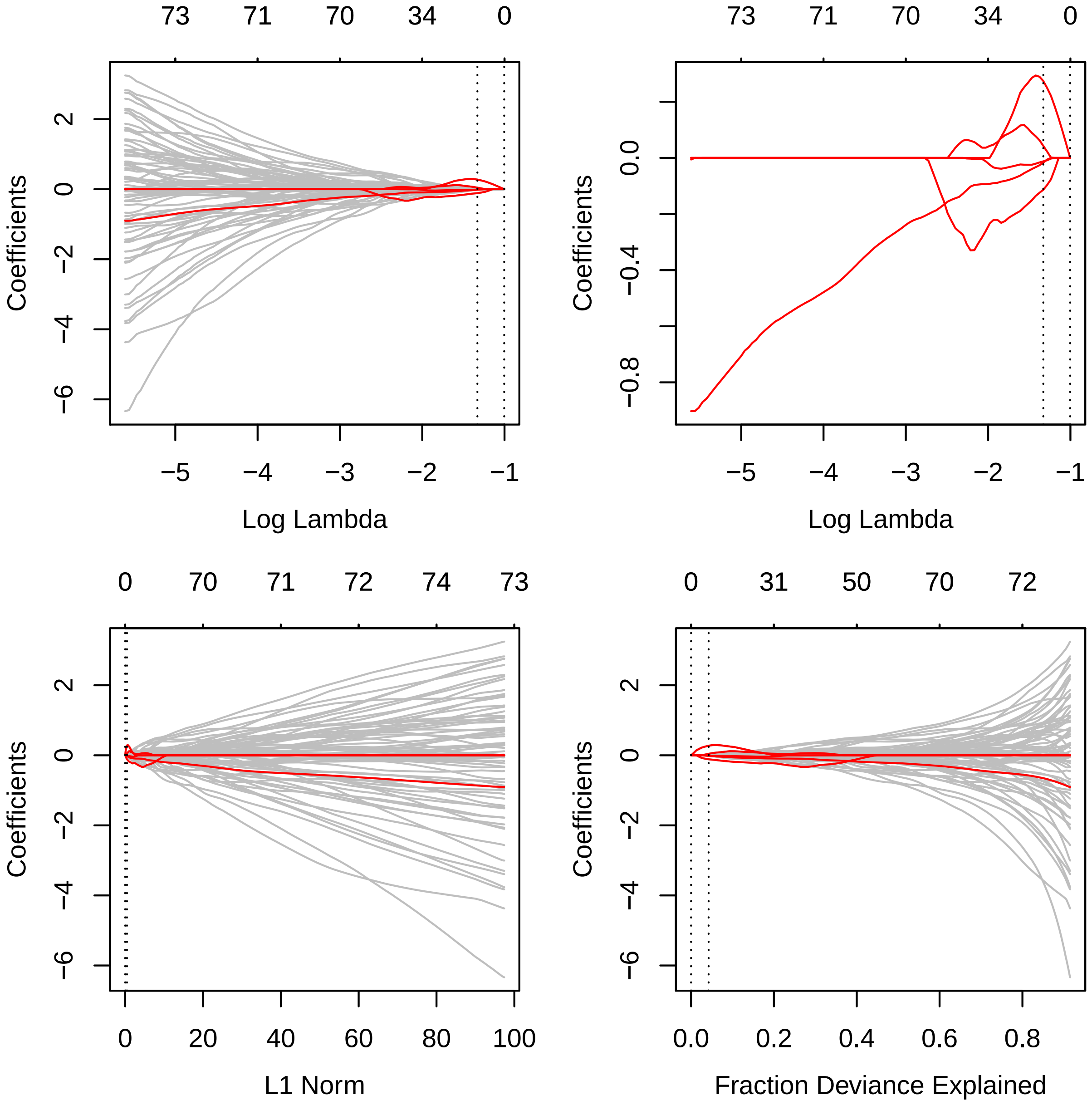

Figure 2: Coefficient paths for lasso penalised Cox PH regression model applied to the AML data set.

At this point we would like to assess the prediction performance of the lasso model. We can do this with bootstrapped prediction error curves and corresponding integrated Brier score values (see Thomas' functions adapted for peperr/pec).

Once we have seen that this model is not very satisfactory, we can attempt to improve the model in two ways. First, we can fit an elastic net model rather than lasso (and use Natalia's search algorithm for that). And second, we can assess the stability of the lasso (and elastic net) models by stability selection and identify the most stable features (using Martin's stabilityselection.R script).

2.1 Prediction error curves

3 Elastic net penalised Cox PH regression model

4 Stability selection

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

```
206932_at
     2823
```

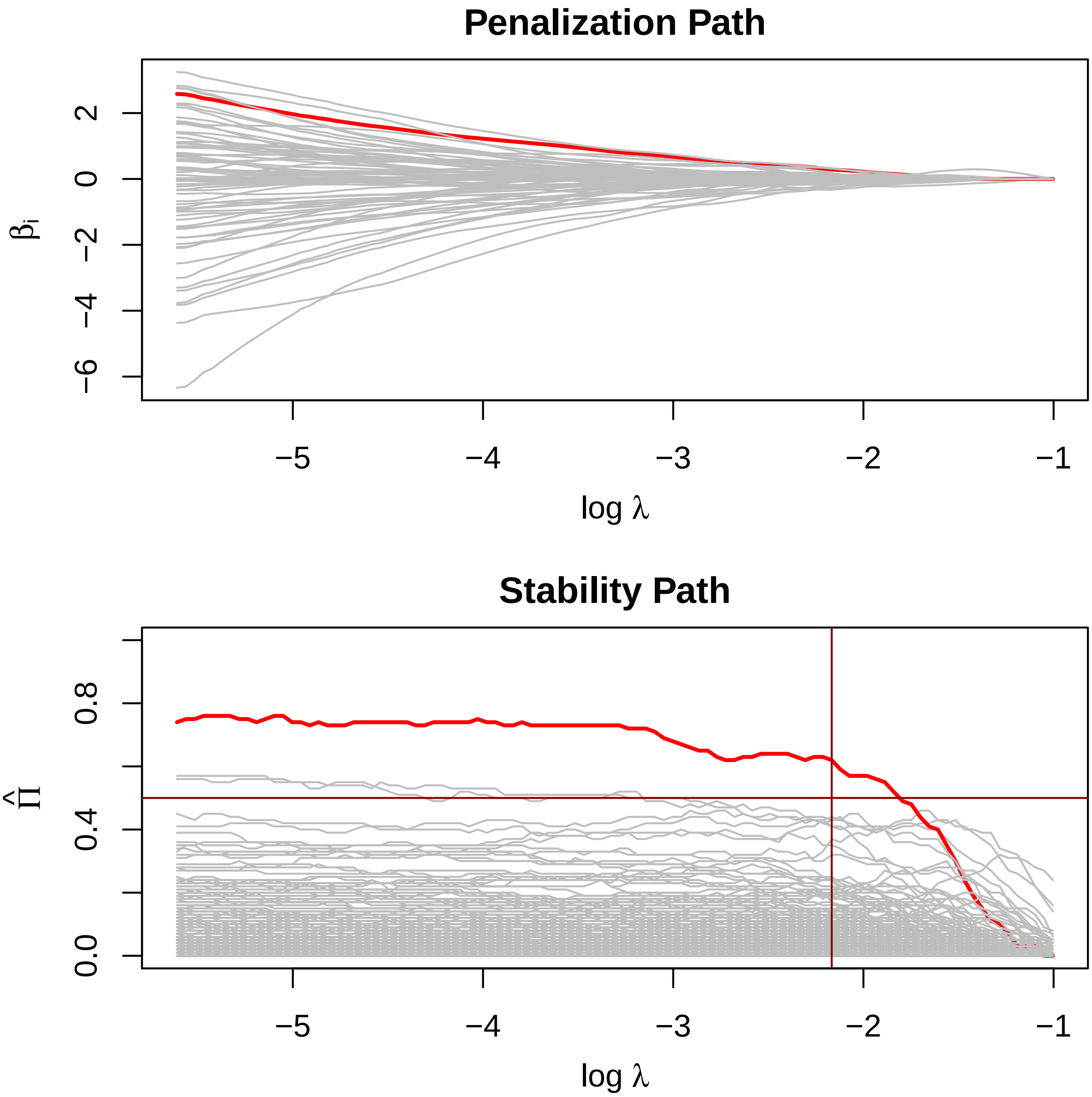

Figure 3: Coefficient and stability paths for lasso penalised Cox PH regression model applied to the AML data set.

5 Summary

6 Session Information

The version number of 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, stats, utils
- Other packages: Biobase 2.14.0, Matrix 1.0-5, cacheSweave 0.6-1, filehash 2.2-1, genefilter 1.36.0, glmnet 1.7.3, lattice 0.20-6, stashR 0.3-5
- Loaded via a namespace (and not attached): AnnotationDbi 1.16.19, DBI 0.2-5, IRanges 1.12.6, RSQLite 0.11.1, annotate 1.32.3, digest 0.5.2, grid 2.14.1, splines 2.14.1, survival 2.36-14, tools 2.14.1, xtable 1.7-0