MAS202M Final Project

Kári Hlynsson

University of Iceland

April 2024

Overview

- Part A: Data cleaning / clustering (30%)
  - Removal of duplicated samples
  - Treatment of missing values (`NA`s)
  - Hierarchical clustering, K-means and PCA
- Part B: Prediction / inference (70%)
  - `EHBP1`: LASSO and Random Forest
  - `temp`: SVMs and KNN

Part A

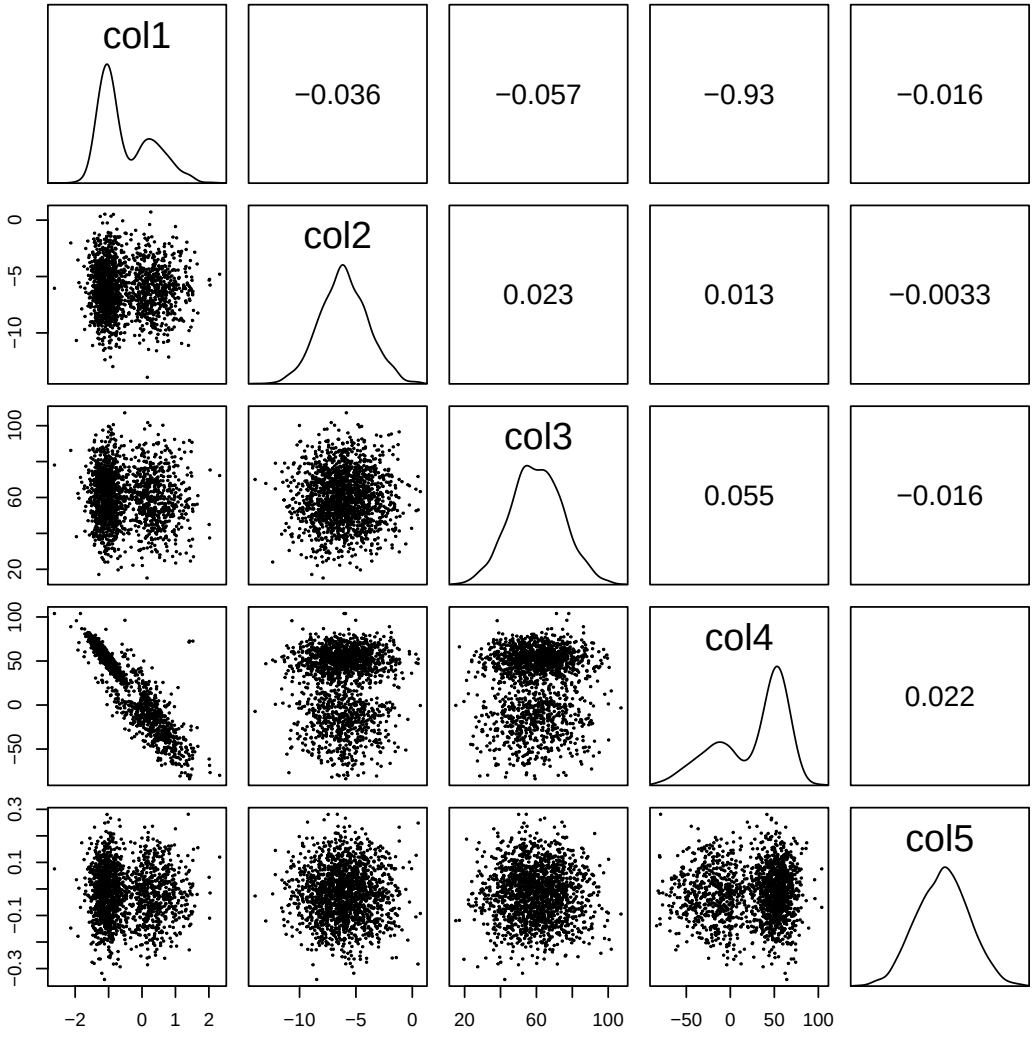
Data cleaning

- Raw data is 1772×8, reduced to 1672×5
- The first column, `index`, was removed
- Covariates `col3` and `col6` were identical except at points where the other was `NA`. The missing observations in `col3` were derived from `col6`; the latter was removed
- 97 observations were found to be duplicates and were removed from the data set
- 3 observations were duplicates but contained one `NA` field, and were not initially detected. These were also removed
- Following removal of duplicated, the `id` column held no relevant information and was subsequently removed

Data cleaning

- The raw data contained 20 missing values
- Means of respective covariates was used to impute these values

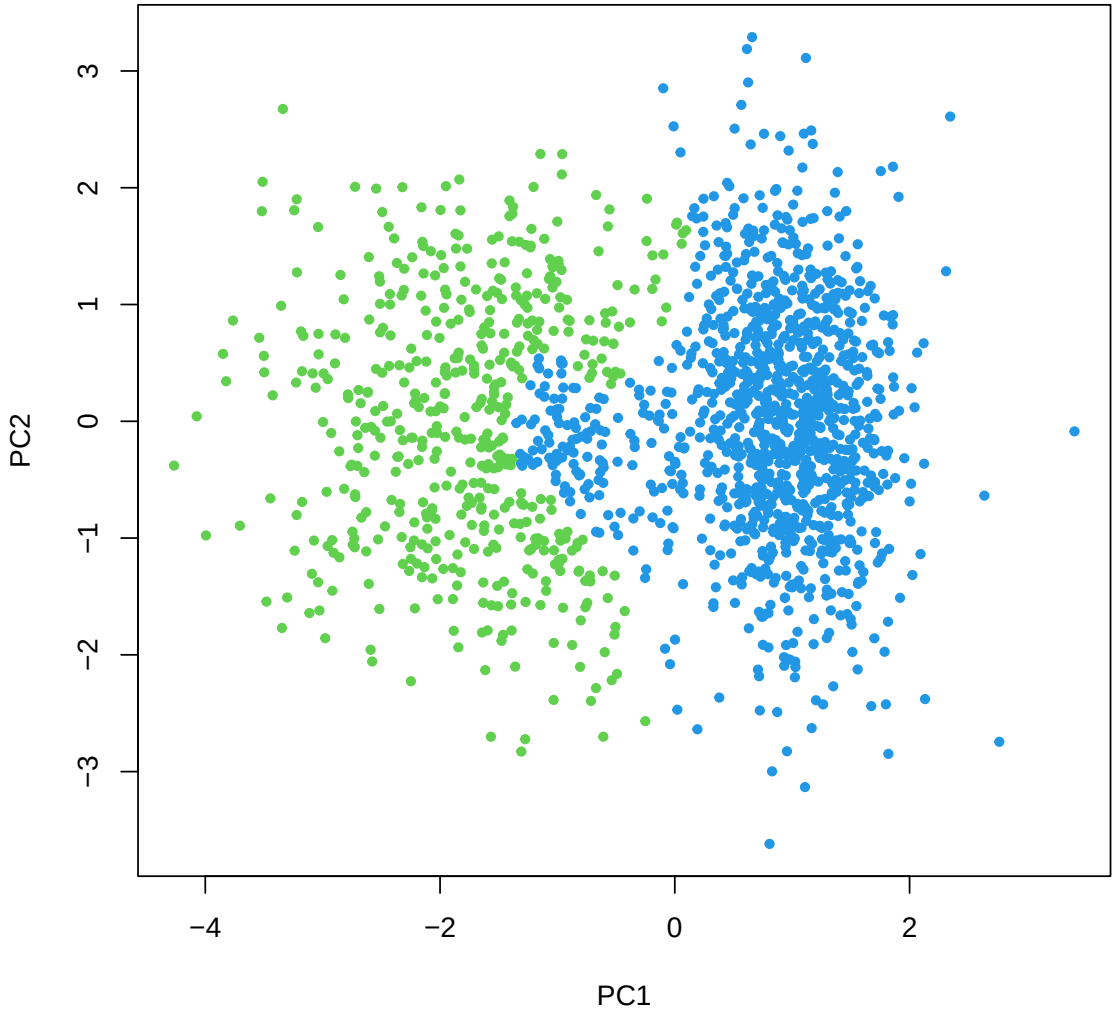
Clustering


col1
−0.036
−0.057
−0.93
−0.016
col2
0.023
0.013
−0.0033
col3
0.055
−0.016
col4
0.022
col5
0
−5
−10
100
60
20
100
50
0
−50
0.3
0.1
−0.1
−0.3
−2
0
1
2
−10
−5
0
20
60
100
−50
0
50
100

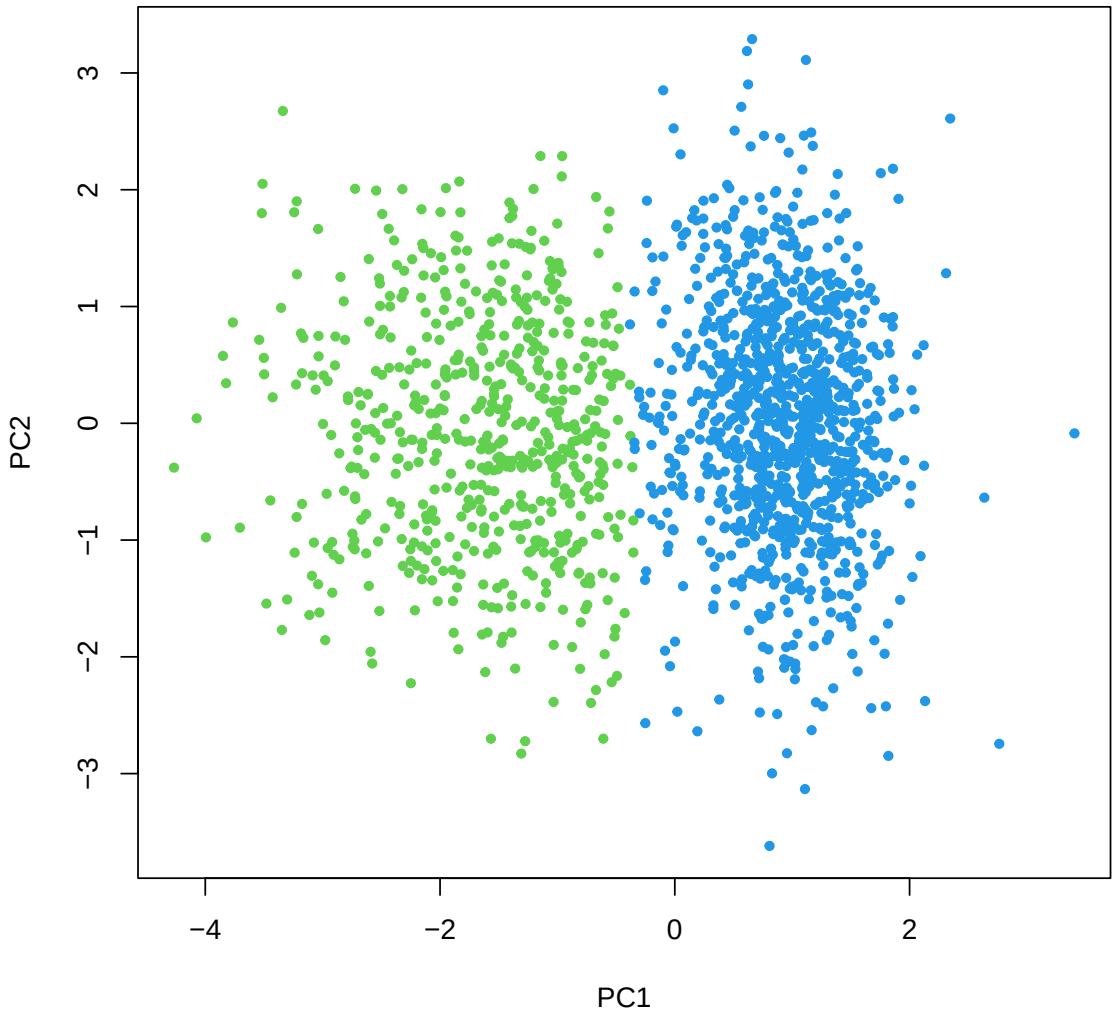
Clustering

- To cluster the data, hierarchical clustering and K-means were considered
  - Each of the two methods was run on *raw* and *scaled* data, as well as on the first two *principal components* of the data
  - Clustering accuracy was mainly assessed through visual confirmation
- Optimal clustering results were obtained for K-means clustering on PCA data

Clustering – Hierarchical



Clustering – K-means



Part B

Quantitative variable

- We will consider `EHPB1`
- LASSO and random forest regression

LASSO

- High dimensional setting, $n \ll p$, motivates use of methods that *reduce number of covariates* used in modelling
- LASSO was chosen for this task in the linear setting. Other potential candidates include ridge, PCR, PLS, AIC step procedures, etc.
- Assumes a *linear relationship* between response and covariates which may be overly stringent
- Optimise over cost hyperparameter, λ:

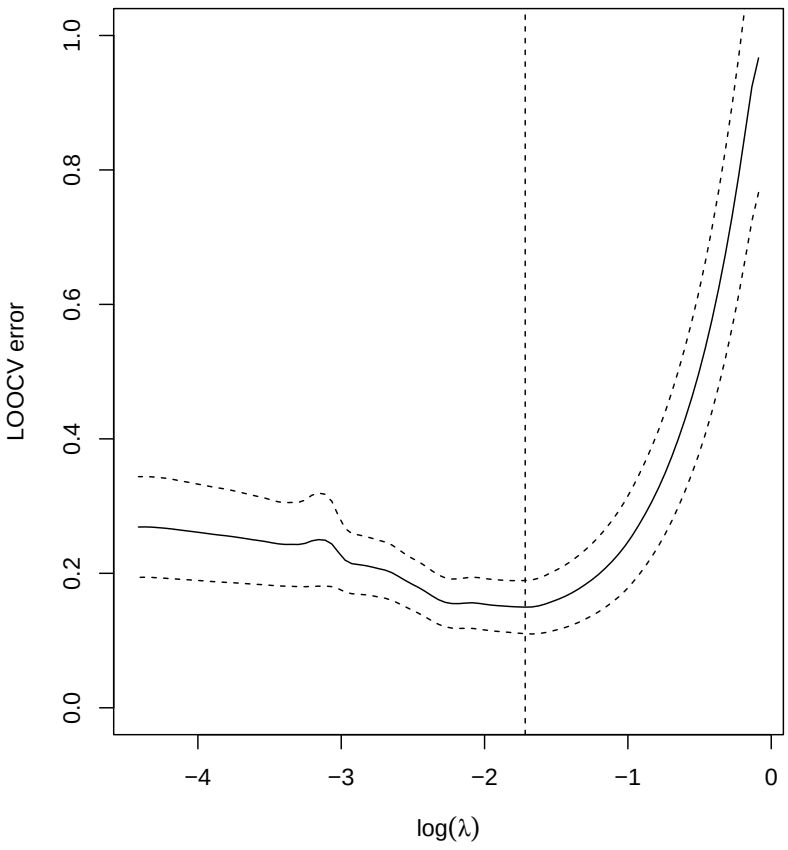
$$\beta_{\text{LASSO}} = \underset{\beta}{\text{argmin}} \left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \left| \sum_{j=1}^{p} \beta_j \right| \right]$$

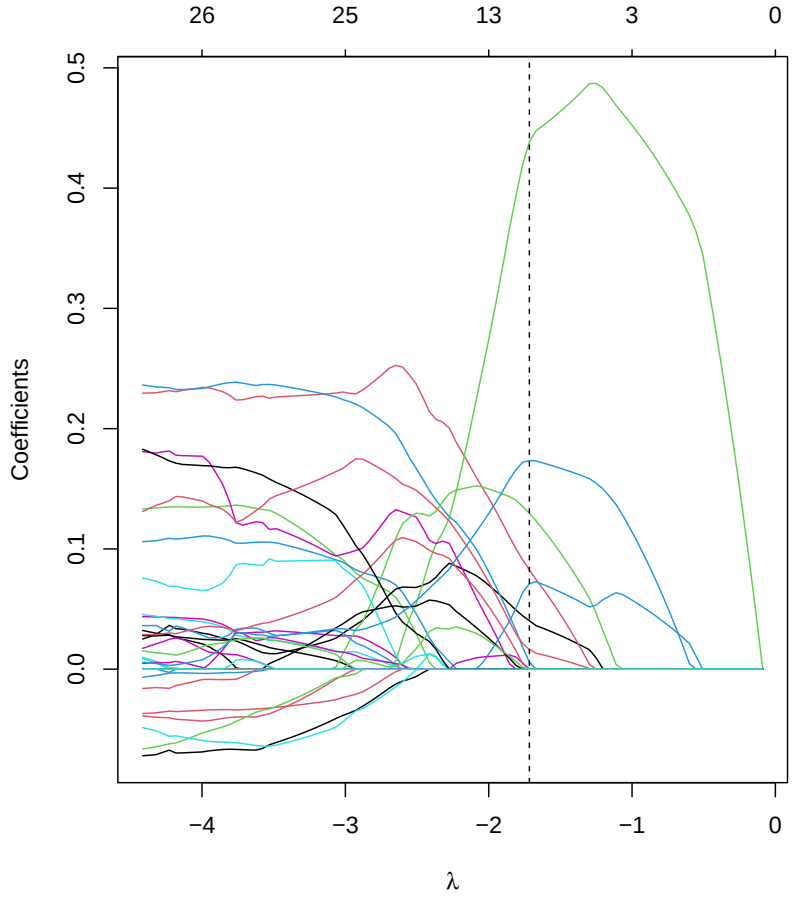
LASSO

- Optimal value of λ determined using 75% train-test split using LOOCV. This gave $\lambda \approx 0.1797$
- Test set MSE $= 0.1080$
- 8 predictors were non-zero, specifying the model

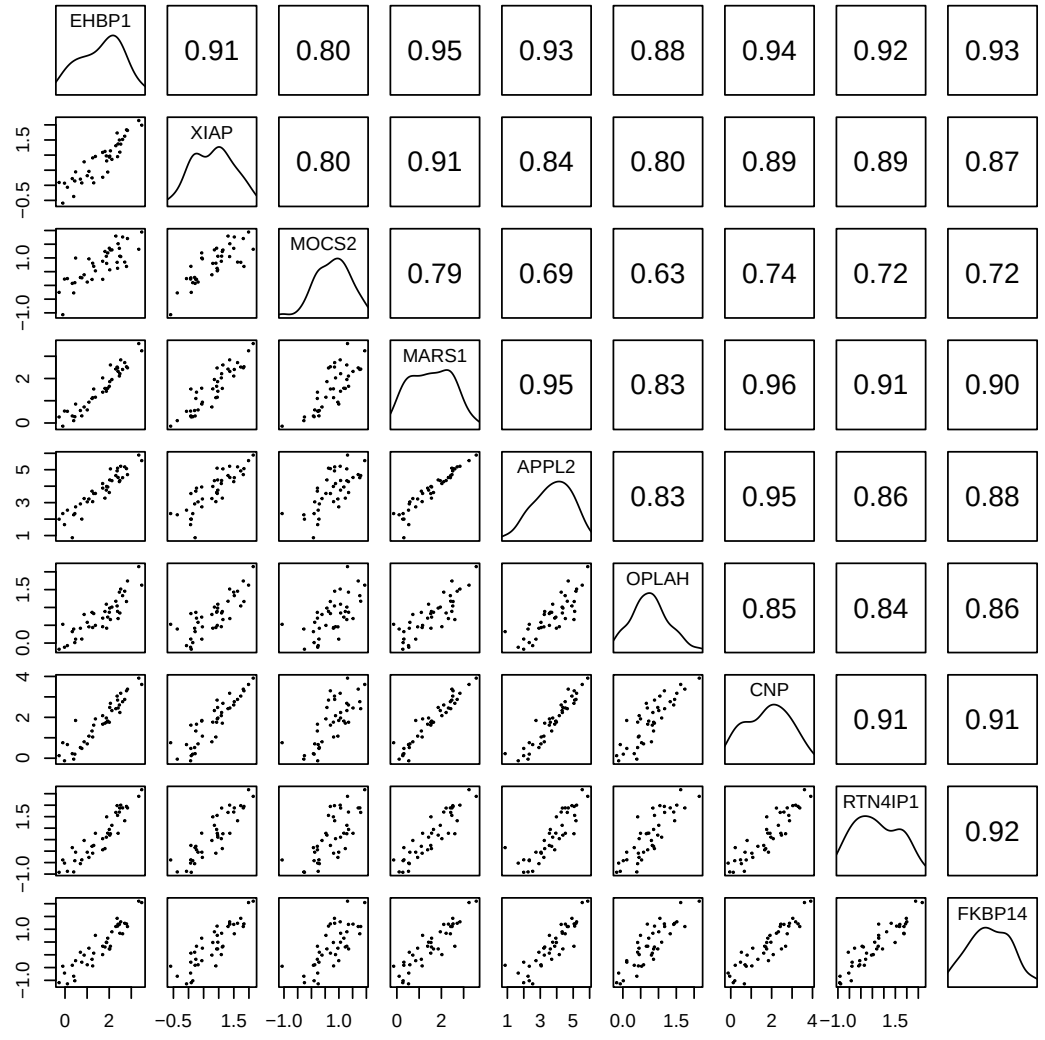
$$\begin{aligned}\texttt{EHBP1} = 0.5504 &+ 0.0403 \times \texttt{XIAP} + 0.0826 \times \texttt{MOCS2} \\ &+ 0.4383 \times \texttt{MARS1} + 0.006 \times \texttt{APPL2} \\ &+ 0.1296 \times \texttt{OPLAH} + 0.071 \times \texttt{CNP} \\ &+ 0.0164 \times \texttt{RTN4IP1} + 0.1734 \times \texttt{FKBP14}\end{aligned}$$

LASSO


LOOCV error
0.0
0.2
0.4
0.6
0.8
1.0
−4
−3
−2
−1
0
log(λ)


26
25
13
3
0
Coefficients
0.0
0.1
0.2
0.3
0.4
0.5
−4
−3
−2
−1
0
λ

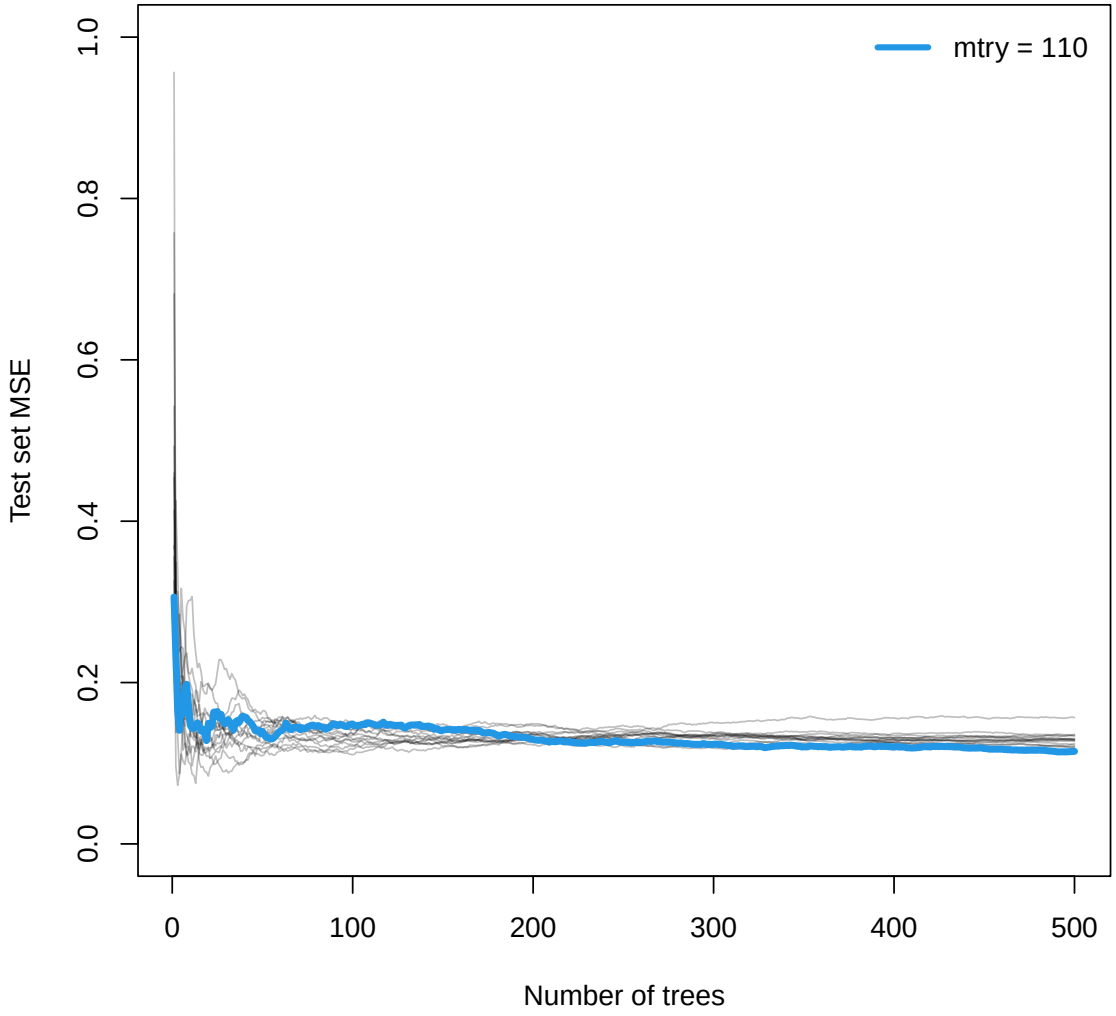
LASSO


EHBP1
0.91
0.80
0.95
0.93
0.88
0.94
0.92
0.93
XIAP
0.80
0.91
0.84
0.80
0.89
0.89
0.87
MOCS2
0.79
0.69
0.63
0.74
0.72
0.72
MARS1
0.95
0.83
0.96
0.91
0.90
APPL2
0.83
0.95
0.86
0.88
OPLAH
0.85
0.84
0.86
CNP
0.91
0.91
RTN4IP1
0.92
FKBP14
−0.5 1.5
−1.0 1.0
0 2
1 3 5
0.0 1.5
4
−1.0 1.5

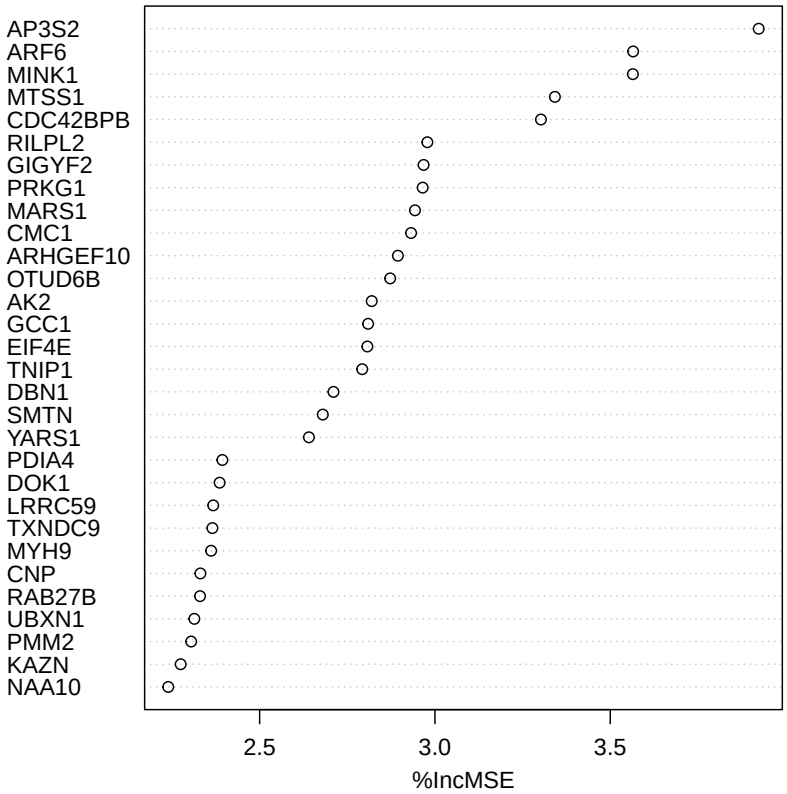
Random forest

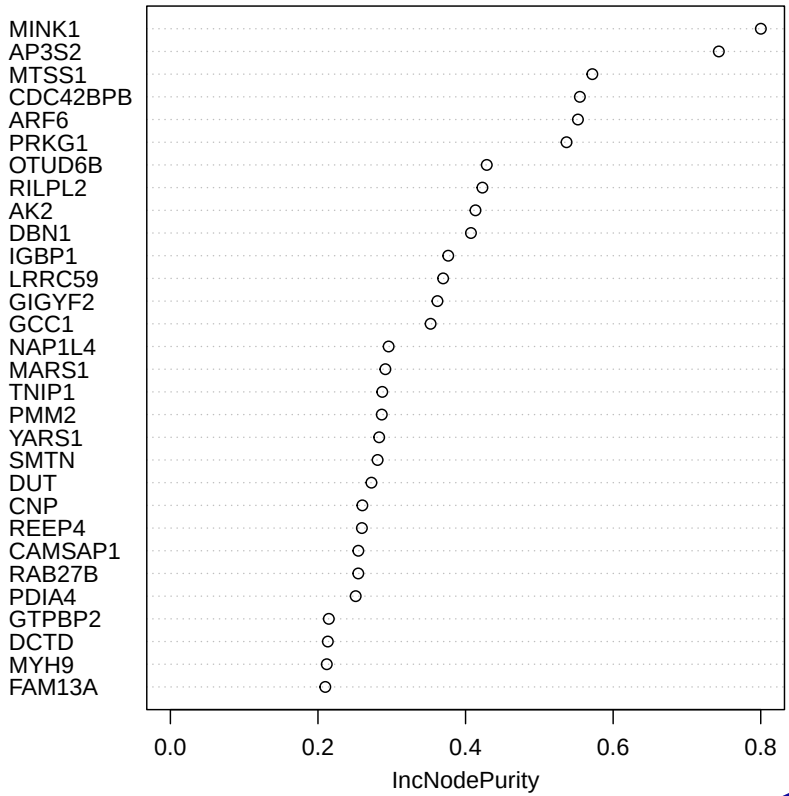
- Random forest regressor was tuned for optimal `mtry` with `ntree` = 500, yielding `mtry` = 110
- Works well in the *high-dimensional setting* $n \ll p$ due to restricting decision trees to a *subset of predictors*
- More *candidate causal predictors* uncovered – favourable in ”breadth-first” scanning, e.g. for biomarkers
- Test set MSE of 0.1147, marginally unfavourable to LASSO.

Random forest



Random forest



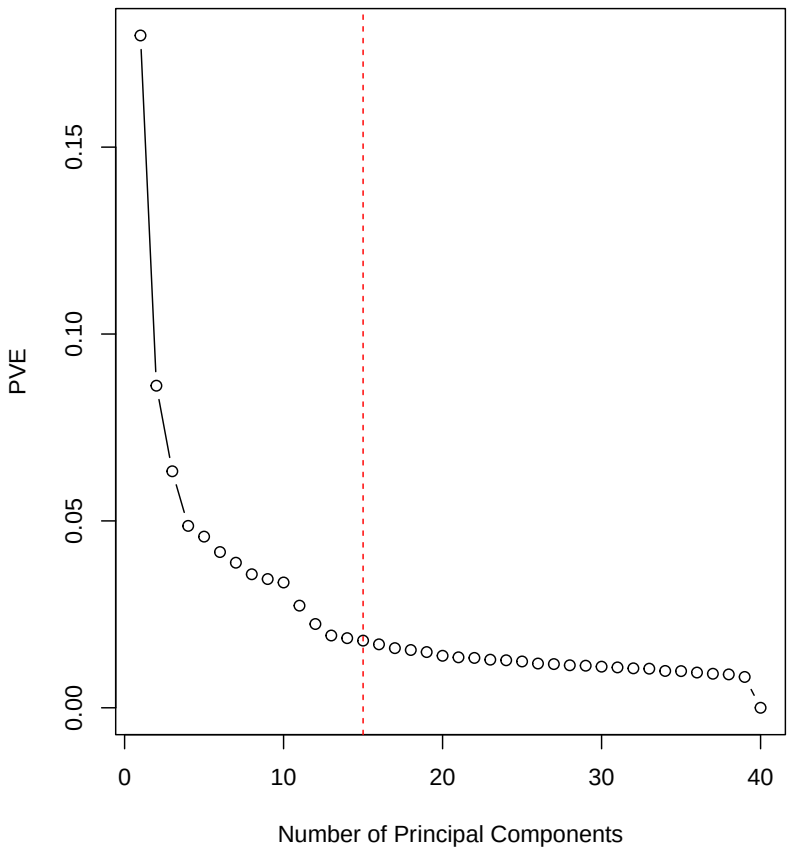


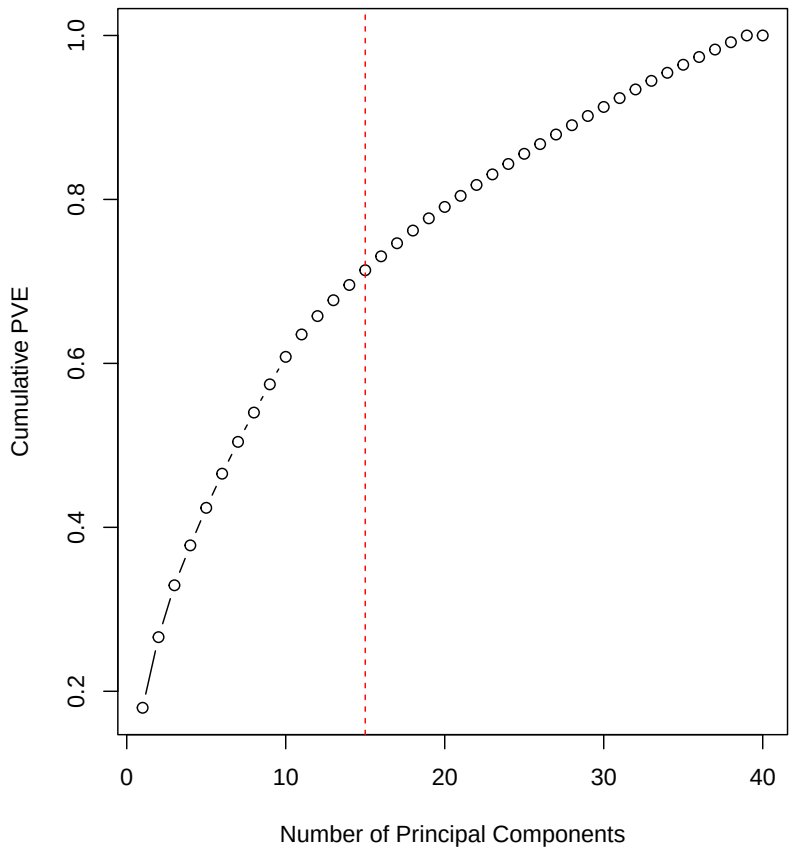
Random forest

Categorical variable

- We will study the `temp` variable
- SVM and KNN
- Fitted on data transformed using PCA to reduce computational complexity (15 components, 71.4% var.)

Categorical variable





SVM

- Optimised over three kernels; linear (SVC), polynomial and radial and respective parameters
  - Linear: `cost`
  - Polynomial: `cost`, `degree`
  - Radial: `cost`, `gamma`
- Linear (`cost` = 0.0785) and radial kernels gave (`cost` = 54.5559, `gamma` = 0.0034) best performance with 70% accuracy
- Polynomial kernel (`cost` = 6.1585, `degree` = 1) lagged behind with 65% accuracy

KNN

- Strong out-of-the-box classifier, *nonparametric*
- Tends to suffer in high-dimensional settings
- Optimised for the number of nearest observations to inspect when classifying, K
- $K = 9$ gave the optimal accuracy of 75%

KNN