UMC Utrecht

Feature Selection

Zhujie Gu

Leiden Annual Meeting

Feb 19 2020

IM FOR FUTURE

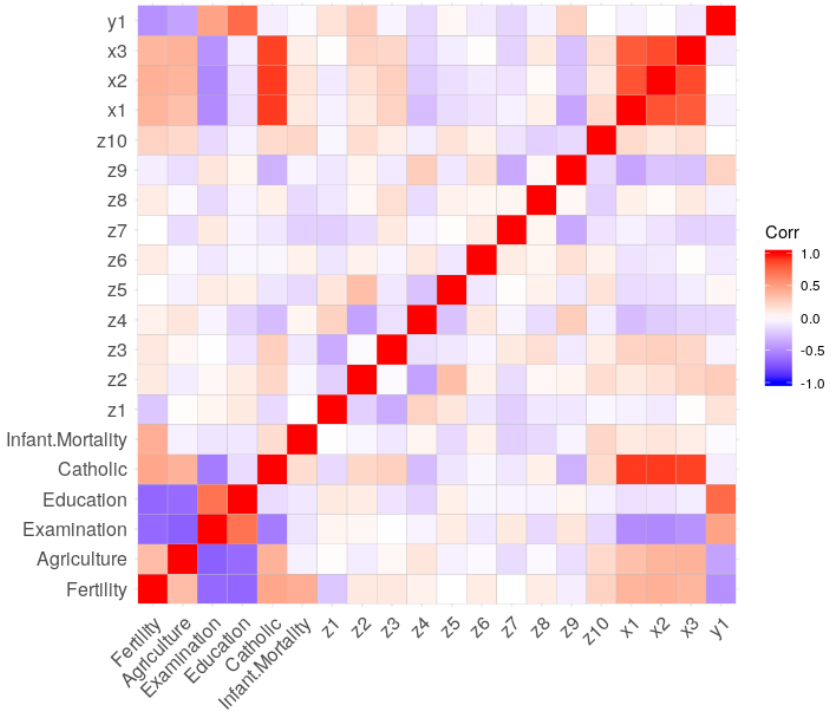
Universitair Medisch Centrum Utrecht

Recap - Purposes of feature selection

- Better prediction
  - Select combination of features which predicts well
  - Bias–variance tradeoff
- Better interpretation
  - Select relevant features
- Different purposes can lead to different desired selecting strategies:
  - Example: highly correlated features x1, x2, for prediction, select one of them, for interpretation, select both of them
  - Today we focus on better prediction

Recap - Multivariate vs Univariate regression

- Compare significant features of multivariate regression (Q4) and those of univariate regression (Q8)
  - What causes the difference?
  - Check correlation among variables
  - Variance Inflation Factor (VIF) values



Advanced methods for feature selection

- General form of penalized regression

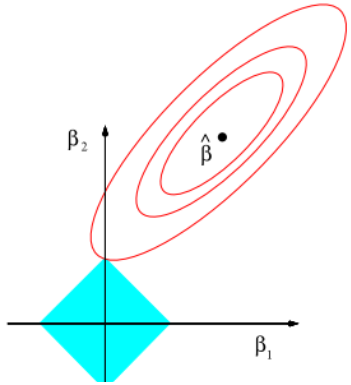
$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \mathbf{P}_\lambda(\beta)$$

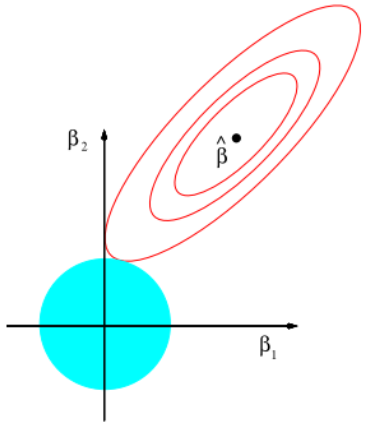
- **Lasso** (least absolute shrinkage and selection operator): shrink some coefficients to exact 0; does not handle correlated data well

$$\mathbf{P}_\lambda(\beta) = \lambda \|\beta\|_1$$

- **Ridge:** shrink coefficients but not to 0; handle correlated data

$$\mathbf{P}_\lambda(\beta) = \lambda \|\beta\|_2^2$$



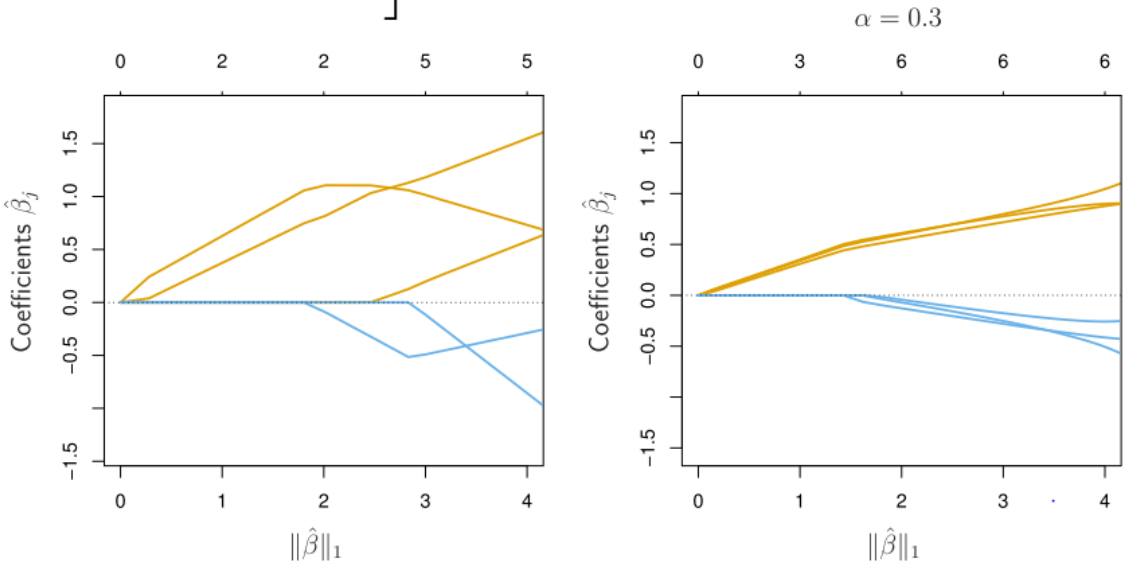


Advanced methods for feature selection

- **Elastic net:** a compromise between Lasso and Elastic net

$$\mathbf{P}_\lambda(\beta) = \lambda \left[\frac{1}{2}(1-\alpha) \|\beta\|_2^2 + \alpha \|\beta\|_1 \right]$$

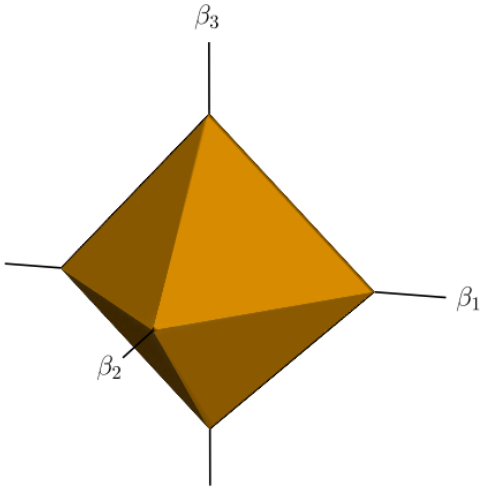
- Example of coefficient path of Lasso (left) and Elastic net (right) for correlated data (two correlated groups of variables)

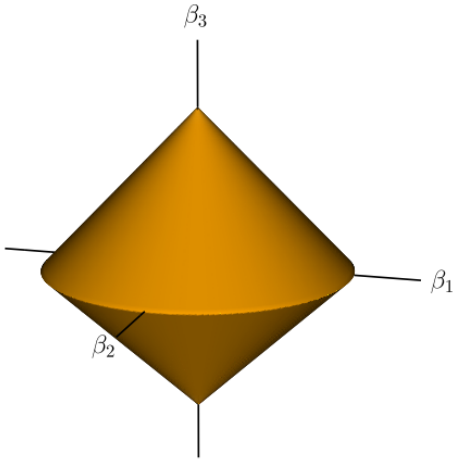


Advanced methods for feature selection

- **Group lasso:** result in group-wise sparsity by shrinking groups of coefficients to 0

$$\mathbf{P}_\lambda(\beta) = \lambda \sum_{j=1}^{J} \sqrt{p_j}\, \|\beta_j\|_2$$





Evaluating prediction performance

- Training and testing datasets
  - Why split the data
  - Be careful about case/control proportion for binary outcome
- Overfitting?
- What do you see in homework Q5?
  - Different results every run
  - Averaging over multiple runs

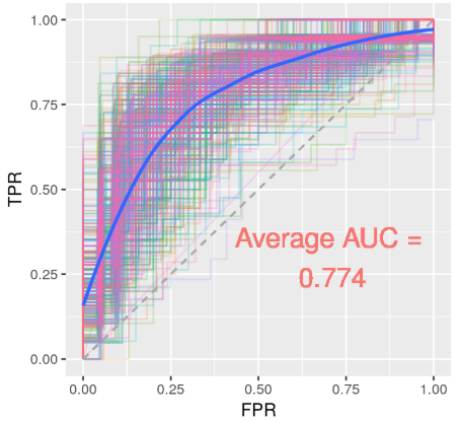
Example on a microbiome dataset

- Data description:
  - n=78
  - p=139 (p > n)
  - Outcome (37 case, 41 control)
- Aim: to compare the prediction performance of the bacteria selected by different methods
- Measure: averaged AUC
  - Binary outcome
  - Data split into training and testing (50-50)
  - Repeat 500 times to have stable results

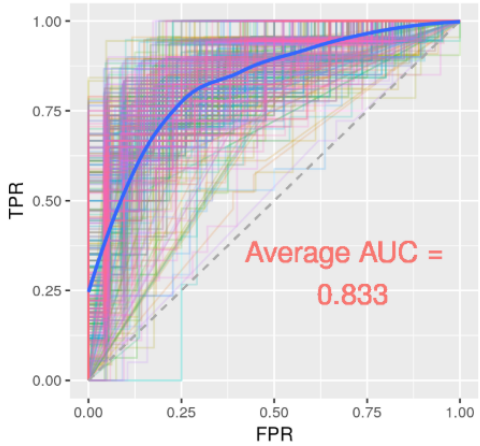
Results

- Here, we show prediction results from 3 methods
  - Single point analysis: 4 significant bacteria
  - Elastic net: top 5 selected bacteria
  - Random forest: top 5 selected bacteria

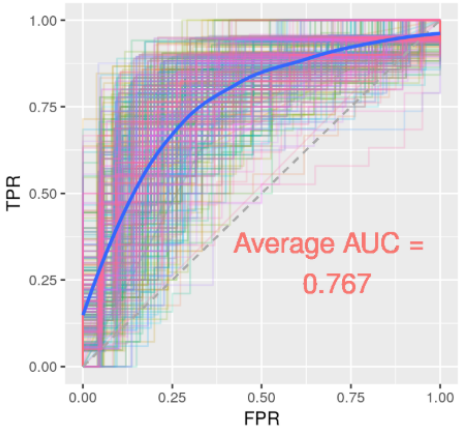
Single point



Elastic net

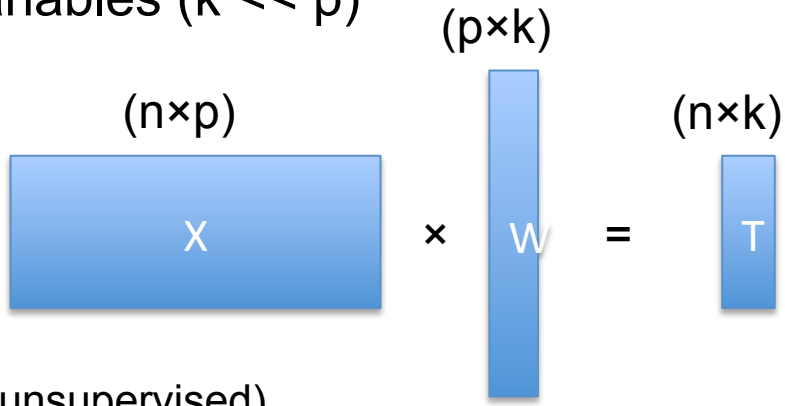


Random forest



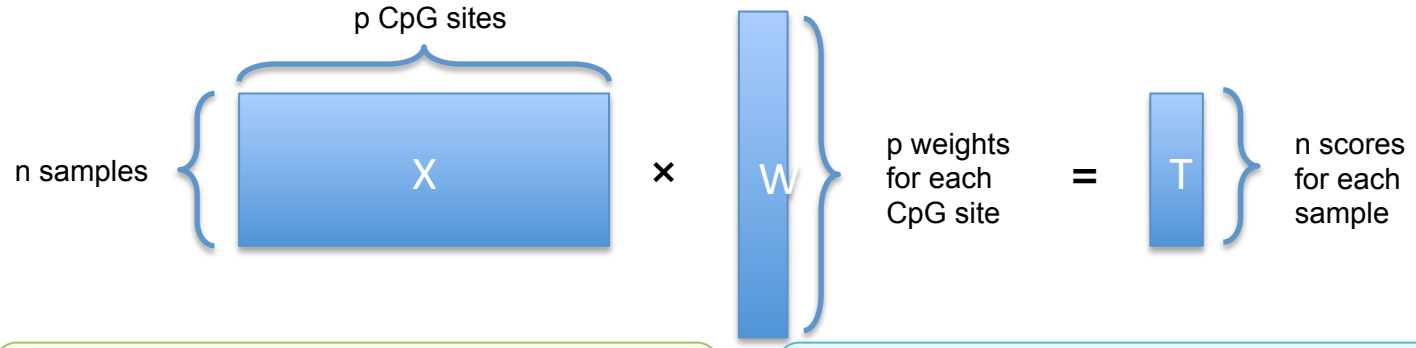
Dimension reduction

- reducing the number of random variables p by constructing k latent variables (k << p)



  - PCA (unsupervised)
  - PLS1 (supervised)
  - PLS2 (allow Y to be a matrix)

Dimension reduction + feature selection



Feature selection is performed by shrinking weights of irrelevant features (CpG) to 0. Features with non-zero weight are selected

Group selection is performed by shrinking weights of the whole irrelevant group (CpG on the same island) to 0. The features in the same group have either zero or non-zero weights altogether

Dimension reduction + feature selection

- Each latent variable **t** is linear combination of all the observed variables **x**
- Imposing L1-norm penalty on the weights **w** shrinks small weights to exact 0 (very similar to Lasso in regression)
- Reduce dimensionality and select relevant features in each latent variable at the same time
- Example methods
  - Sparse PCA
  - Sparse PLS

OMG!!

So many methods !
How do I choose?

Simulation – YOU KNOW THE TRUTH

- What's your purpose? (e.g. investigate if method A is better than other methods in prediction)
- What are the scenarios? (e.g. based on your data?)
- What's "better"? (e.g. Low prediction error, robustness…)
- What are the measures? (e.g. MSE …)
- Which methods to compare?
- Implementation
- Reproducible?

Acknowledgements

- Prof. Jeanine Houwing-Duistermaat
- Dr. Hae-Won Uh
- S. (Said) el Bouhaddani
- Roos-Marijn Berbers

Reference

- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: The lasso and generalizations. *Statistical Learning with Sparsity: The Lasso and Generalizations*, *84*(1), 1–337. https://doi.org/10.1201/b18401