Univariate and Multivariate Feature Selection in R

Nikolay Oskolkov, MRG Group Leader, LIOS, Riga, Latvia

R course, 06.02.2026

@NikolayOskolkov

@oskolkov.bsky.social

Personal homepage:
https://nikolay-oskolkov.com

Image generated by ChatGPT

Session content

Topics we'll cover in this session:

1) Feature prioritization as a way to reduce data dimensionality

2) Univariate feature selection: differential gene expression analysis

3) Multivariate feature selection: LASSO, PLS and LDA

4) Overfitting and underfitting, cross-validation and hyperparameter tuning

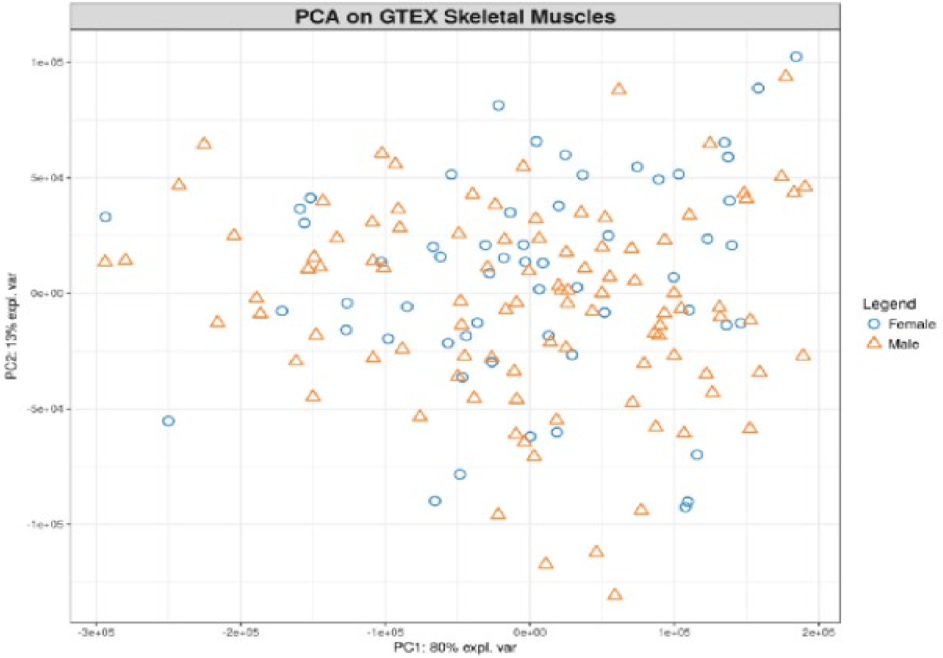
Univariate Feature Selection

```r
X <- read.table("GTEX_SkeletalMuscles_157Samples_1000Genes.txt",
                header=TRUE, row.names=1, check.names=FALSE, sep="\t")
X <- X[,colMeans(X) >= 1]
Y <- read.table("GTEX_SkeletalMuscles_157Samples_Gender.txt",
                header=TRUE, sep="\t")$GENDER
library("mixOmics")
pca.gtex <- pca(X, ncomp=10)
plot(pca.gtex)
plotIndiv(pca.gtex, group = Y, ind.names = FALSE, legend = TRUE,
          title = 'PCA on GTEX Skeletal Muscles')
```

ReadGTEX.R hosted with ❤ by GitHub view raw

```r
rho <- vector()
p <- vector()
a <- seq(from=0, to=dim(X)[2], by=100)
for(i in 1:dim(X)[2])
{
  corr_output <- cor.test(X[,i], as.numeric(Y), method="spearman")
  rho <- append(rho,as.numeric(corr_output$estimate))
  p <- append(p,as.numeric(corr_output$p.value))
  if(isTRUE(i%in%a)==TRUE){print(paste("FINISHED ",i," FEATURES",sep=""))}
}
output <- data.frame(GENE=colnames(X), SPEARMAN_RHO=rho, PVALUE=p)
output$FDR <- p.adjust(output$PVALUE, method="fdr")
output <- output[order(output$FDR, output$PVALUE, -output$SPEARMAN_RHO), ]
head(output,10)
```

UnivarFeatureSelect.R hosted with ❤ by GitHub view raw




```
##                           GENE SPEARMAN_RHO       PVALUE          FDR
## 256   ENSG00000184368.11_MAP7D2   -0.5730196 4.425151e-15 2.416132e-12
## 324     ENSG00000110013.8_SIAE     0.3403994 1.288217e-05 3.516833e-03
## 297   ENSG00000128487.12_SPECC1   -0.3003621 1.323259e-04 2.408332e-02
## 218    ENSG00000162512.11_SDC3     0.2945390 1.807649e-04 2.467441e-02
## 38    ENSG00000129007.10_CALML4    0.2879754 2.549127e-04 2.783647e-02
## 107  ENSG00000233429.5_HOTAIRM1   -0.2768054 4.489930e-04 4.085836e-02
## 278   ENSG00000185442.8_FAM174B   -0.2376098 2.731100e-03 2.130258e-01
## 421   ENSG00000234585.2_CCT6P3    -0.2322268 3.426233e-03 2.338404e-01
## 371    ENSG00000113312.6_TTC1      0.2284351 4.007655e-03 2.431310e-01
## 269 ENSG00000226329.2_AC005682.6  -0.2226587 5.064766e-03 2.523944e-01
```

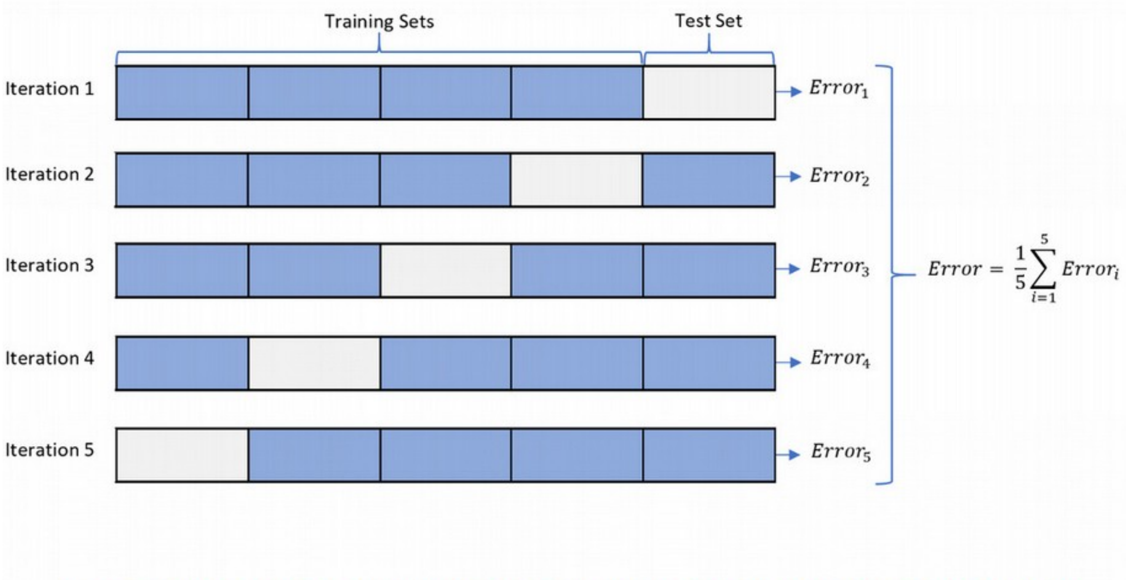
Generally acknowledged that univariate feature selection has poor predictive capacity compared to multivariate feature selection

Multivariate Feature Selection: LASSO

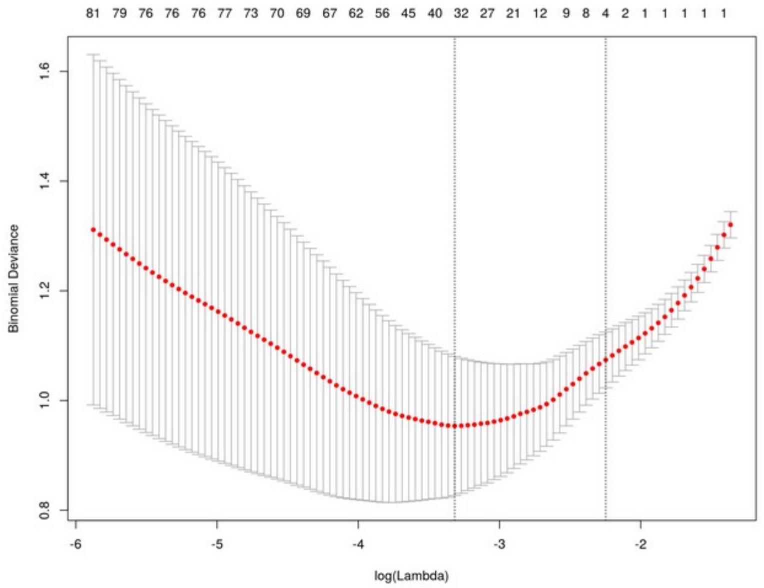
$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$\text{OLS} = (Y - \beta_1 X_1 - \beta_2 X_2)^2$$

$$\text{Penalized OLS} = (Y - \beta_1 X_1 - \beta_2 X_2)^2 + \lambda(|\beta_1| + |\beta_2|)$$



Cross-validation is a standard way to tune model hyperparameters such as λ in LASSO



Regularizations are Priors in Bayesian stats

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon; \quad Y \sim N(\beta_1 X_1 + \beta_2 X_2, \sigma^2) \equiv \mathrm{L}(\mathrm{Y} \mid \beta_1, \beta_2)$$

- **Maximum Likelihood** principle: maximize probability to observe data given parameters:

$$\mathrm{L}(\mathrm{Y} \mid \beta_1, \beta_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(\mathrm{Y} - \beta_1 \mathrm{X}_1 - \beta_2 \mathrm{X}_2)^2}{2\sigma^2}}$$

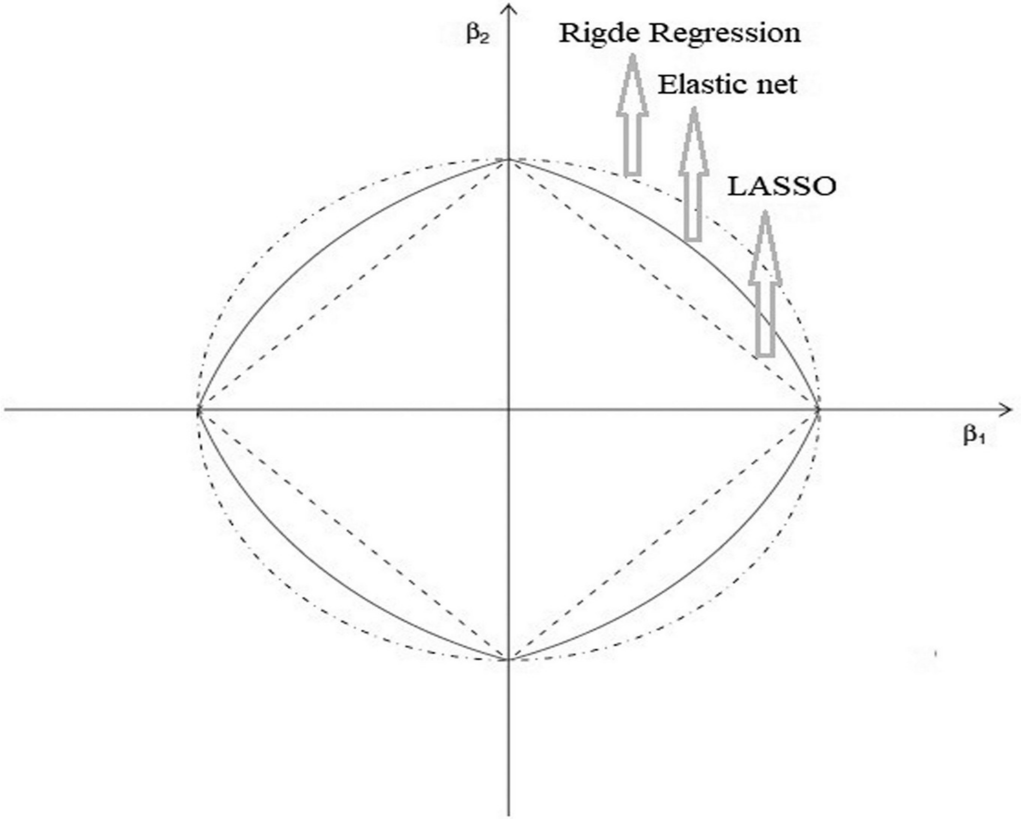
- **Bayes theorem**: maximize posterior probability of observing parameters given data:

$$\mathrm{Posterior}(\mathrm{params} \mid \mathrm{data}) = \frac{\mathrm{L}(\mathrm{data} \mid \mathrm{params}) * \mathrm{Prior}(\mathrm{params})}{\int \mathrm{L}(\mathrm{data} \mid \mathrm{params}) * \mathrm{Prior}(\mathrm{params})\, \mathrm{d}(\mathrm{params})}$$

$$\mathrm{Posterior}(\beta_1, \beta_2 \mid \mathrm{Y}) \sim \mathrm{L}(\mathrm{Y} \mid \beta_1, \beta_2) * \mathrm{Prior}(\beta_1, \beta_2) \sim \exp^{-\frac{(\mathrm{Y} - \beta_1 \mathrm{X}_1 - \beta_2 \mathrm{X}_2)^2}{2\sigma^2}} * \exp^{-\lambda(|\beta_1| + |\beta_2|)}$$

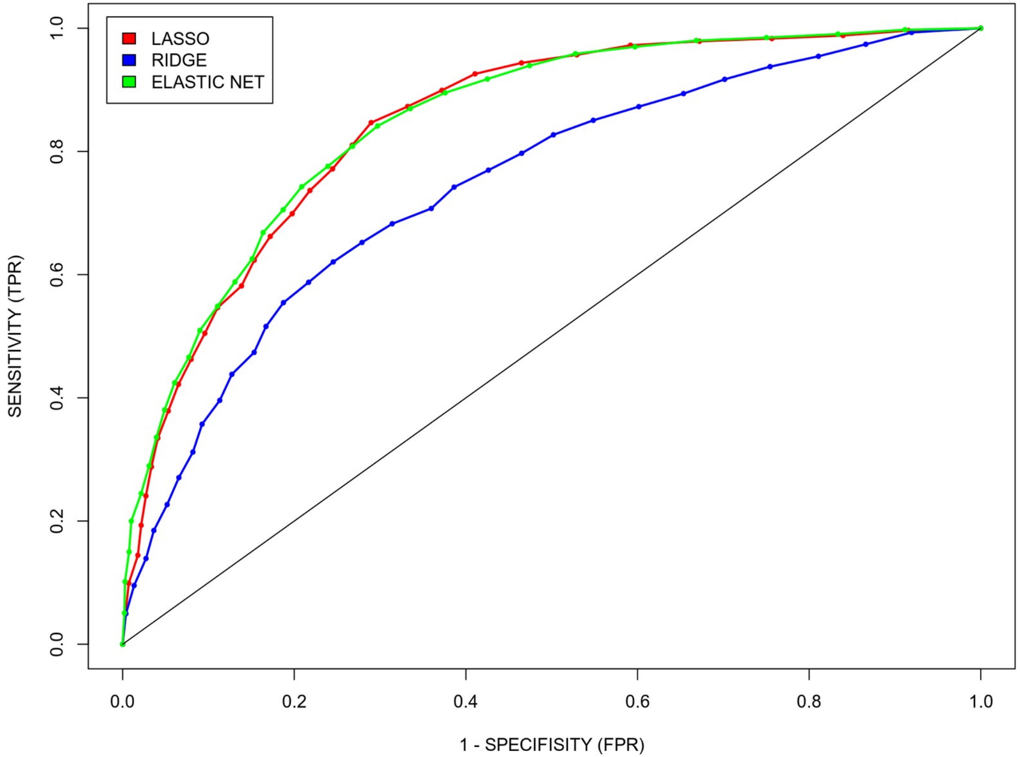
$$-\log\left[\mathrm{Posterior}(\beta_1, \beta_2 \mid \mathrm{Y})\right] \sim (\mathrm{Y} - \beta_1 \mathrm{X}_1 - \beta_2 \mathrm{X}_2)^2 + \lambda(|\beta_1| + |\beta_2|)$$

Lasso vs. Ridge vs. Elastic Net



$$\text{Lasso} : |\beta_1| + |\beta_2| \leq \lambda$$

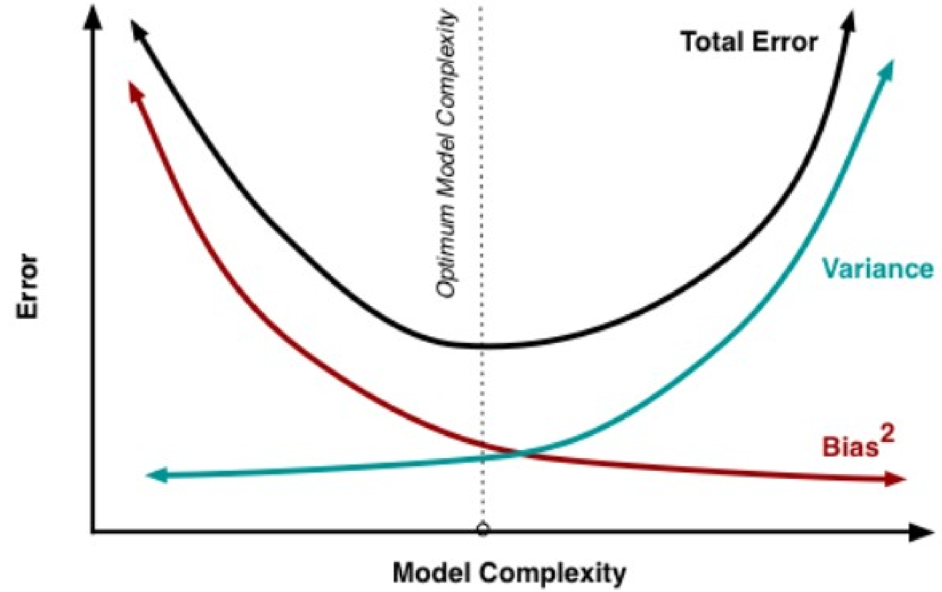
$$\text{Ridge} : \beta_1^2 + \beta_2^2 \leq \lambda$$

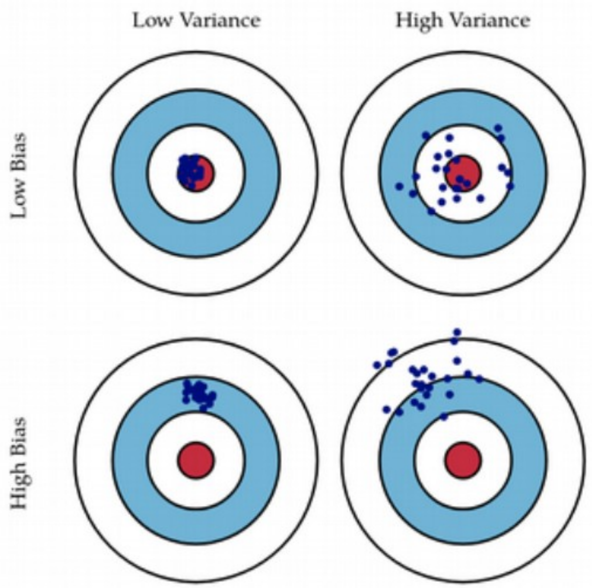


Lasso is more conservative

Ridge is more permissive

Penalized regression interpretation

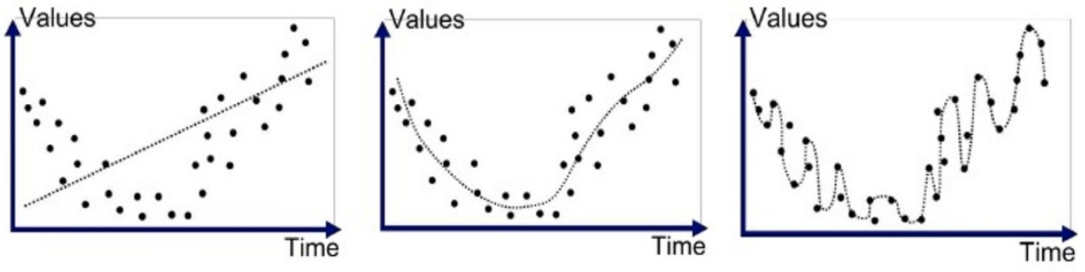




$$Y = f(X) \Longrightarrow \text{Reality}$$

$$Y = \hat{f}(X) + \text{Error} \Longrightarrow \text{Model}$$

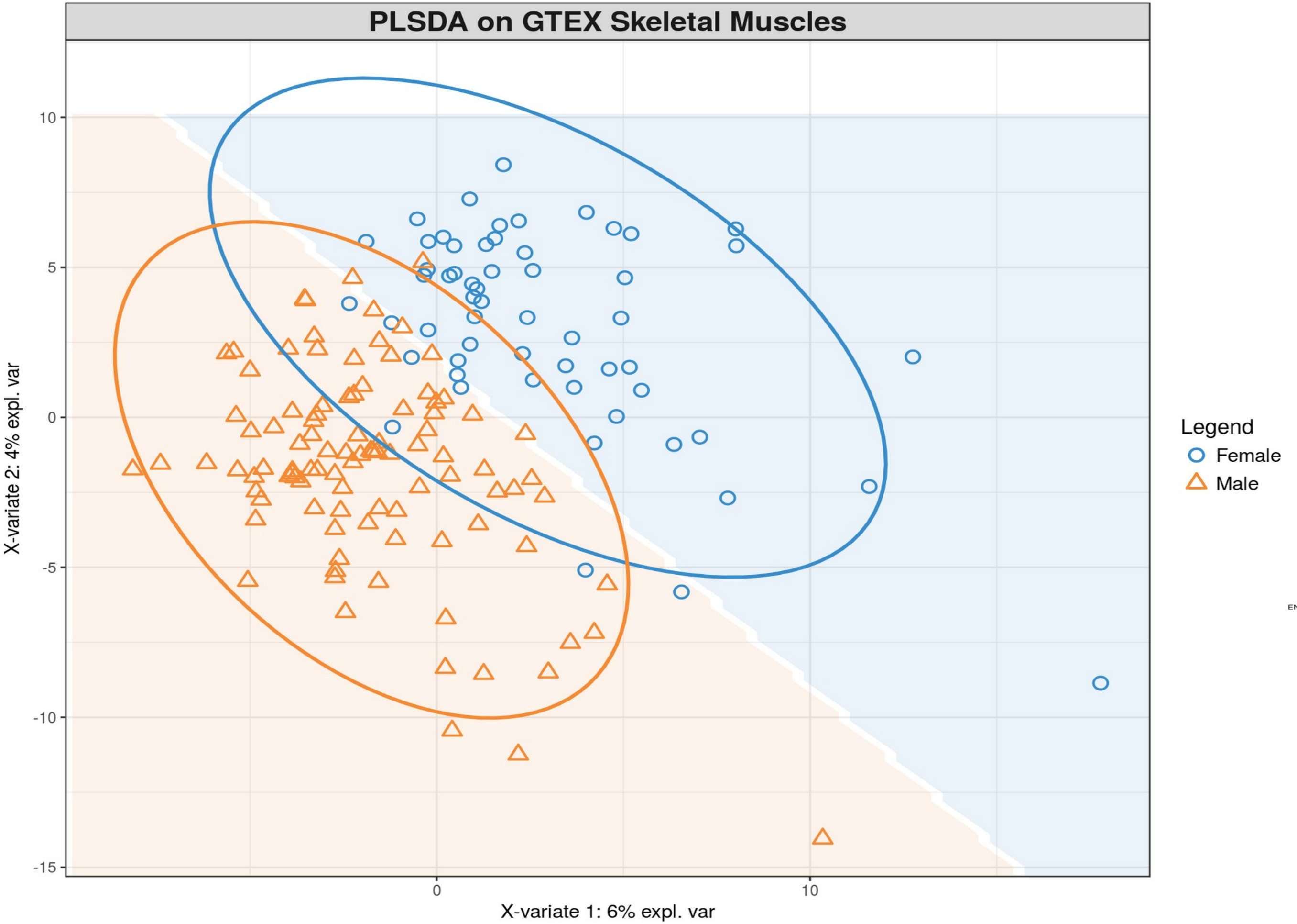
$$\text{Error}^2 = (Y - \hat{f}(X))^2 = \text{Bias}^2 + \text{Variance}$$

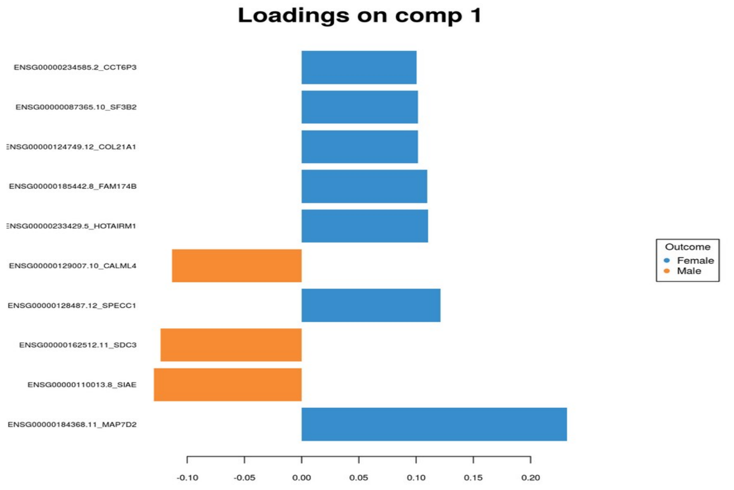


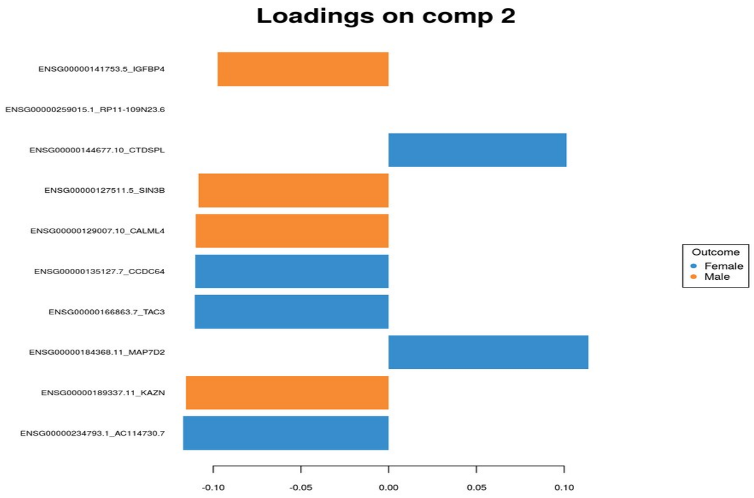
Underfitted Good Fit/Robust Overfitted

LASSO – high bias, low variance

Multivariate Feature Selection: PLS

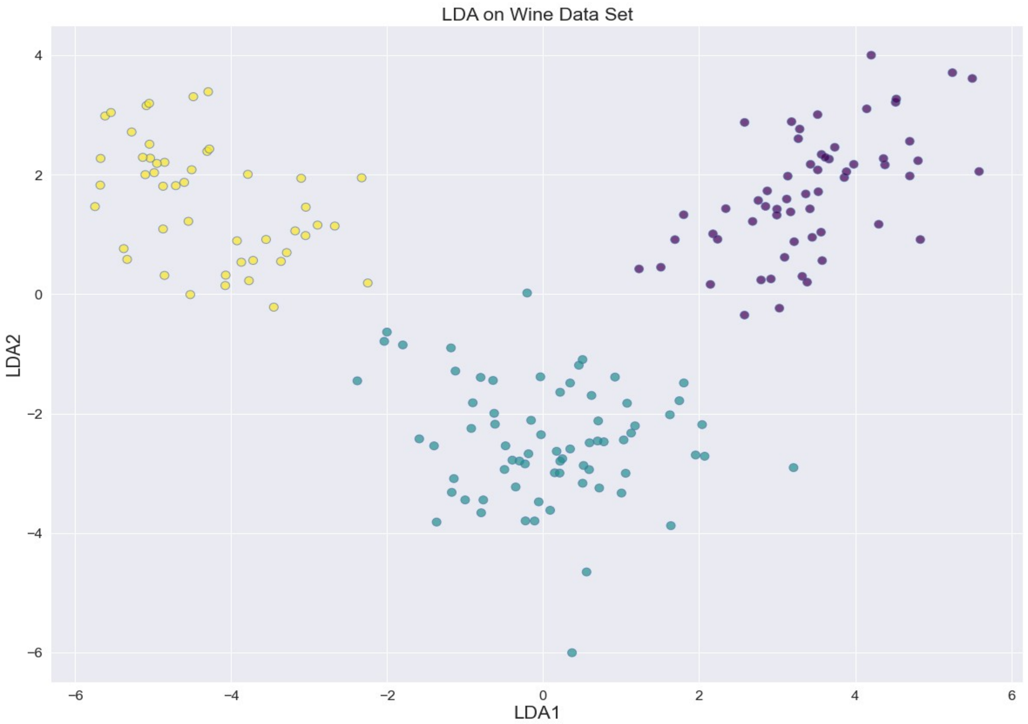






Select features that separate two groups of samples the most

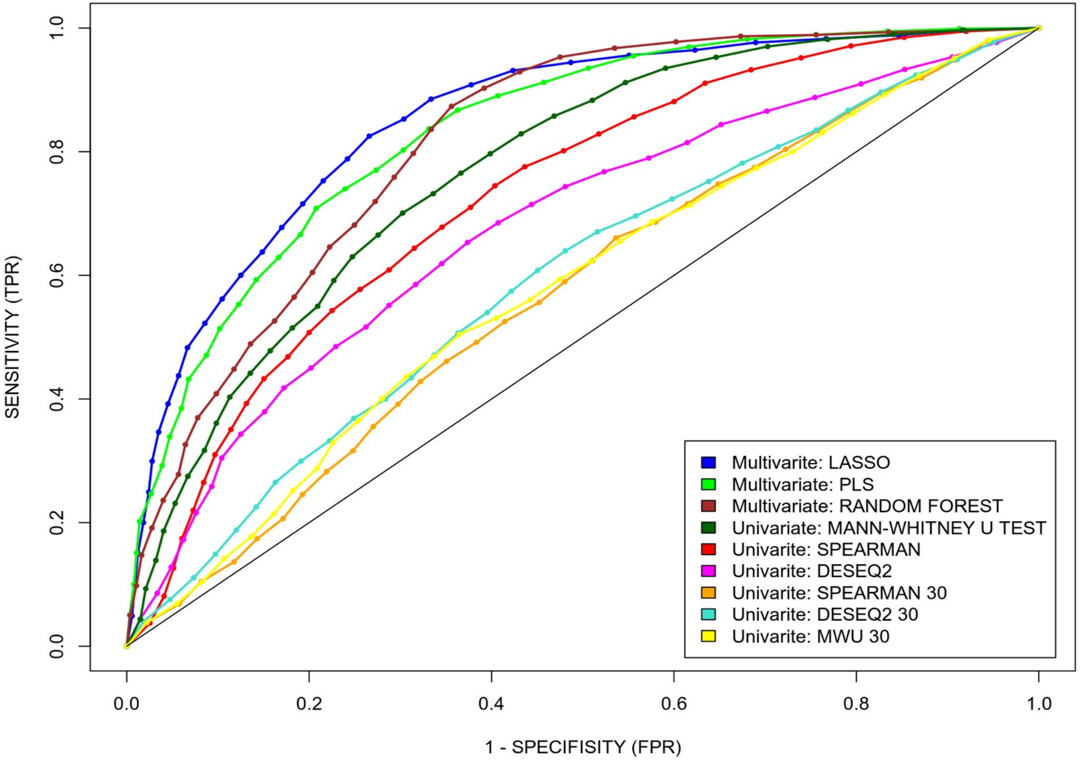
Multivariate Feature Selection: Linear Discriminant Analysis (LDA)



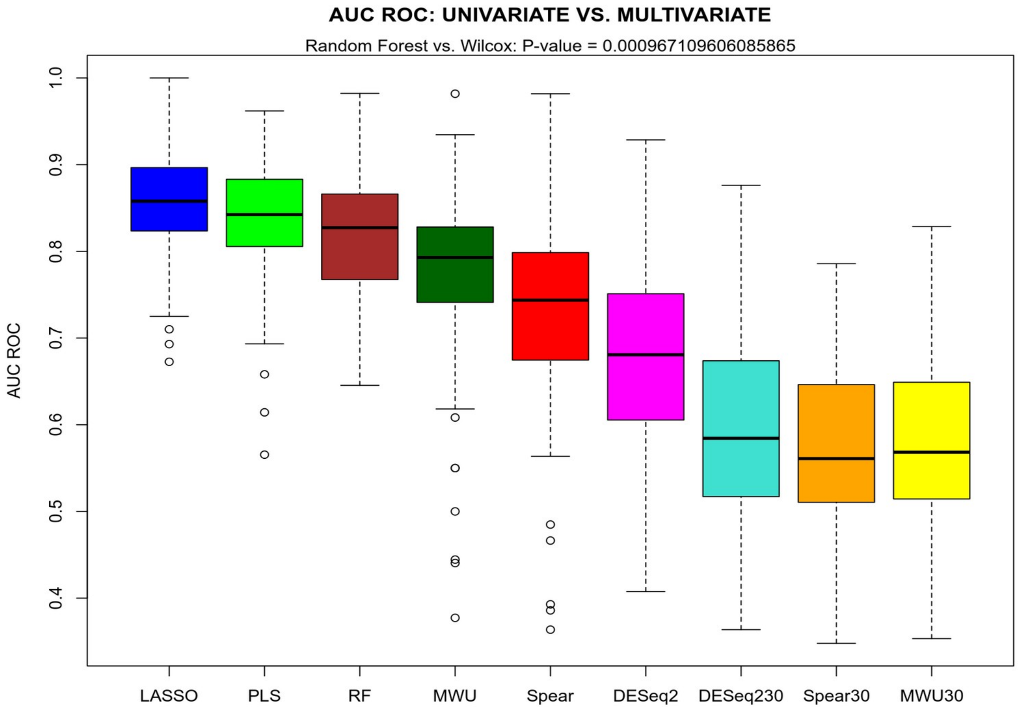
Minimize variance within clusters and maximize variance between clusters

Similar to what ANOVA is doing, therefore LINEAR Discriminant Analysis (LDA)

Univariate vs. Multivariate Prediction



Multivariate methods (LASSO, PLS, RanFor) have significantly higher AUC ROC than univariate methods (Spear, MWU, DESeq2) on skeletal muscle gene expression data



If you find a dataset where univariate feature selection has higher predictive capacity than multivariate one, please let me know

Session summary

Take home messages of the session:

1) Univariate feature selection tests feature by feature for association with the phenotype of interest

2) Multivariate feature selection tests all available features simultaneously

3) Multivariate feature selection has generally higher predictive power than univariate feature selection

4) LASSO can be viewed as a bridge between Frequentist stats, Bayesian stats and Machine Learning

Acknowledgments: LIOS + TARGETWISE

Latvian Institute of Organic Synthesis

2027
National Development Plan

Funded by
the European Union