Regularization and shrinkage estimation

•••

Evangelina López de Maturana & Oscar González-Recio

Recap


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Recap

Large p small n problem

Need regularization methods

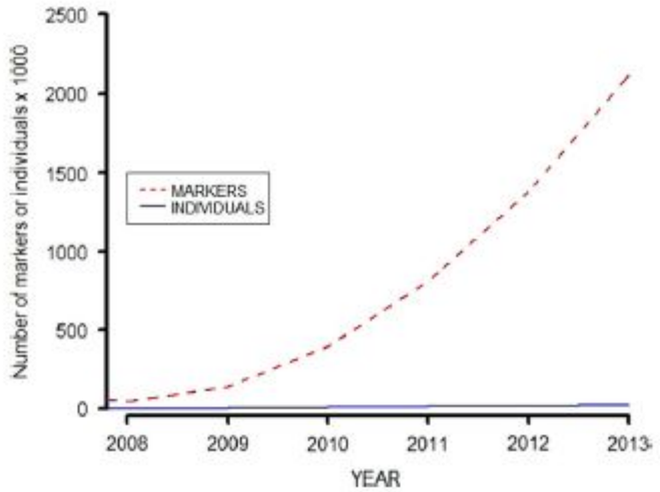


Fig. 1. Relationship between the number of marker effects to be estimated in regression models and the number of genotyped individuals with phenotype in livestock populations.

⚠️ Run out of degrees of freedom in the model

Recap

- many tests → many false positives
    - e.g. 2000 (independent) tests, α=0.05 → How many expected false positives?

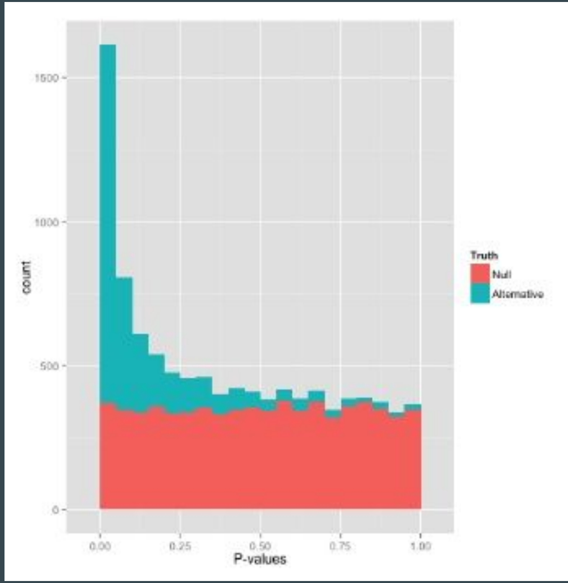
GUESS

Recap

- many tests ⟶ many false positives
  - e.g. 2000 (independent) tests, α=0.05 ⟶ How many expected false positives?

100 false positives by chance alone

- multiple testing problem
- many SNPs, many statistical tests, many p-values
  - Some *p*-values are lower than the α significance level just by chance

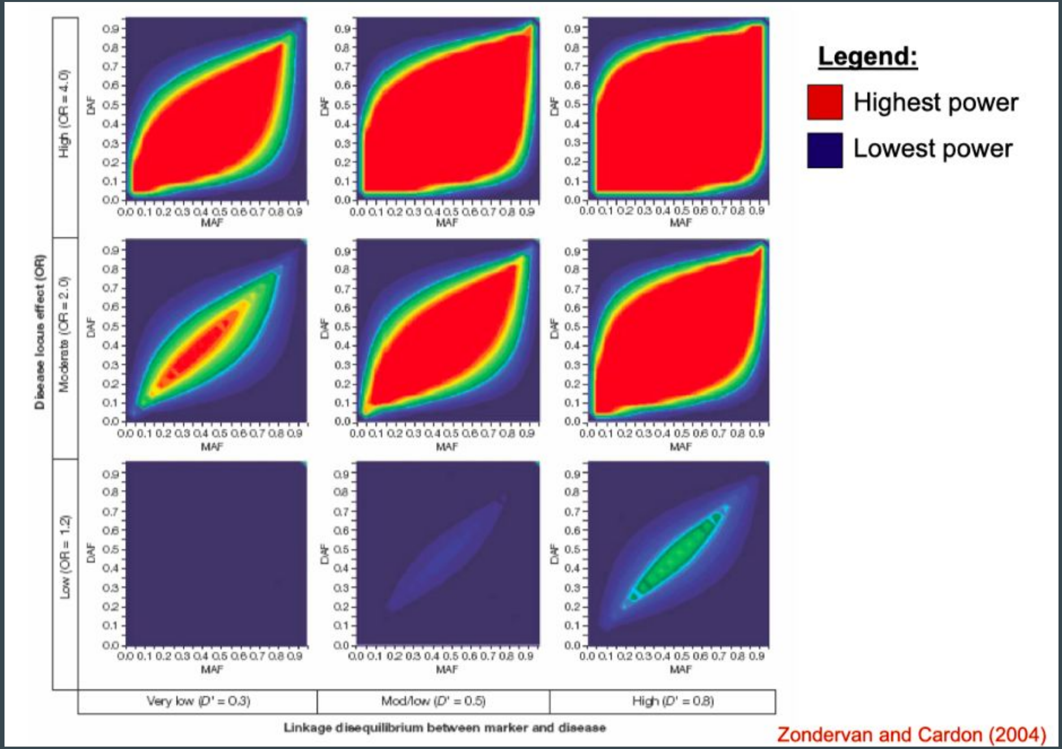


Main difficulties to capture genetic marker signal

- Low effect size (*impairs statistical power*)
- LD between markers (*markers share QTL information*)
- Epistasis and interaction (*models assume linearity*)
- Many markers with (probably) null effect (*sparsity*)

Main difficulties to capture genetic marker signal

- Sample size
- Magnitude of effect
- MAF marker
- MAF qtl
- Range of LD
- Likelihood of the model
- Experimental design



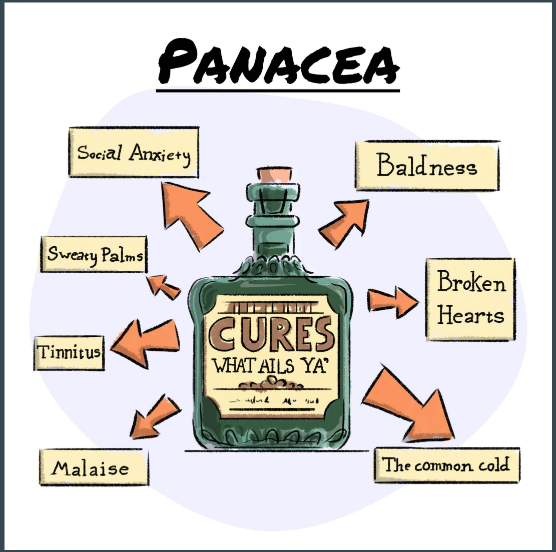
Prediction not that much interest on real effect of markers

How to cope with the problem

- Increase the sample size (e.g. Bio Banks, large reference populations)
- Reduce the number of tests
  - Based on LD
  - Choose relevant regions (functional analysis)
- Decrease the significance threshold
  - Bonferroni correction
  - False discovery rate
  - q-values
- Go Bayesian...

How to cope with the problem

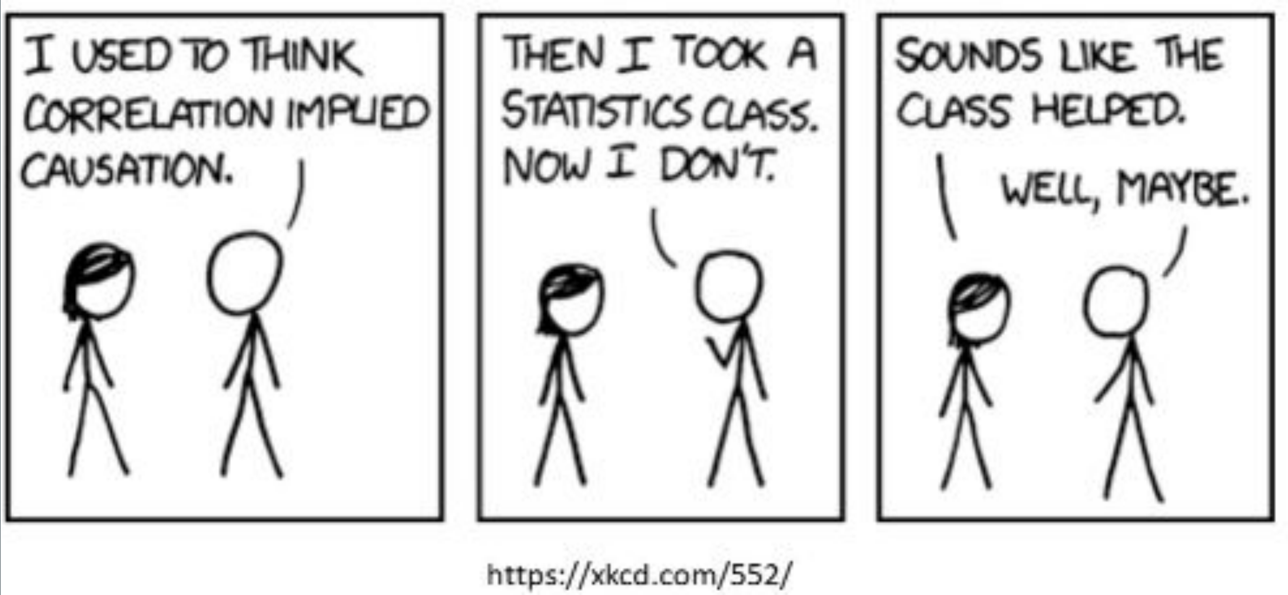
- Increase the sample size (e.g. Bio Banks, lare reference populations)
- Reduce the number of tests
  - Based on LD
  - Choose relevant regions (functional analysis)
- Decrease the significance threshold
  - Bonferroni correction
  - False discovery rate
  - q-values
- Go Bayesian...



Find a causal mutation

A needle in a haystack

Remember



Prediction aims at good correlation or classification performance!

Luckily:
Predict genomic values is different from finding causal mutations

Inference vs Prediction

Inference

- Determine the effect of a covariate on the response
- Determine the causal relationship between a covariate and the response
- More difficult (in general)

Prediction

- Educated guess of the outcome
- Expected behaviour in the future
- Based on proxies/markers

Inference vs Prediction

GWAS goal

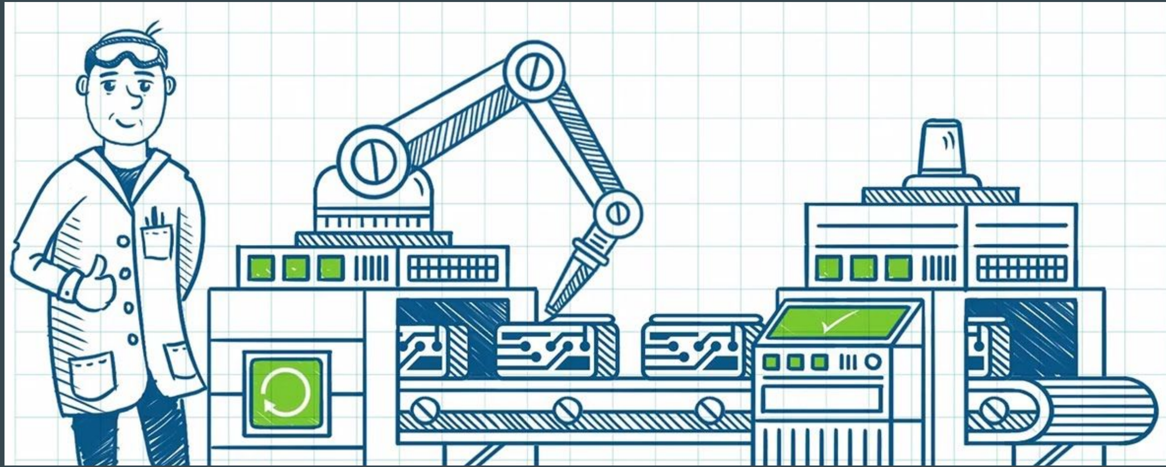
- Detect genomic markers/regions associated to phenotypes (traits) of interest
- Find biological pathways of interest
- Interaction between treatments/drugs and genes

GWP goal

- Calculate the genomic risk score /predisposition of a disease or trait
- Calculate the genomic merit of individuals
- Predict future performance
- Sire/dam selection in animal breeding
- Don’t need to know the real marker effect

Prediction

Aim: construct a predictive (statistical) machine that provides an accurate guess of a yet to be observed phenotype



Large dimensionality problem

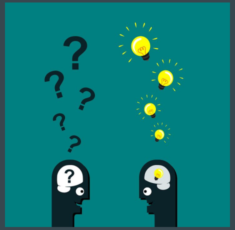
$$y = X\beta + \epsilon,$$

- Minimize a loss function (usually minimize MSE) $\hat{\beta} = \arg\min_{\beta} \sum_{i} (y_i - \beta_1 - \beta_2 x_i)^2,$
- p>>>>n
  - Run out of degrees of freedom (*cannot estimate so many unknowns*).
- Need a more restrictive penalty function $\hat{\beta}_{\mathrm{R}} = \arg\min_{\beta}\{\|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2\},$
- Or set a number of unknowns to zero (feature selection e.g. LASSO or elastic net)

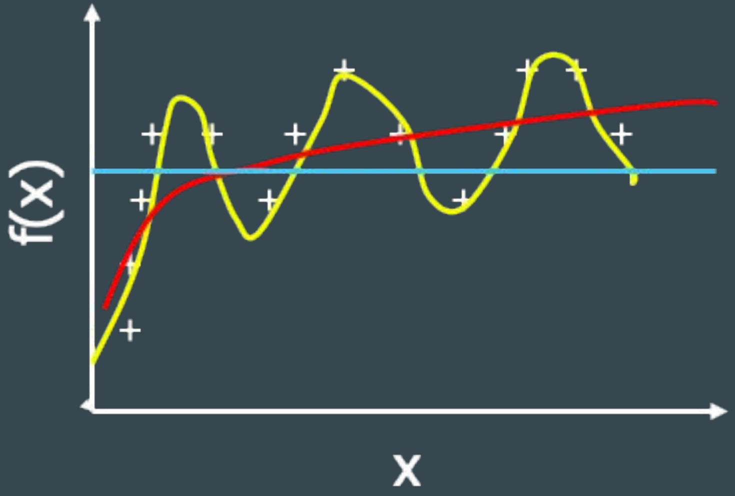
$$\|\beta\|_2^2 \le c_2(\lambda_2) \qquad \|\beta\|_0 \le c_0(\lambda_0) \qquad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

Need to modelize the data

Caution with overfitting with so many variables (SNPs)

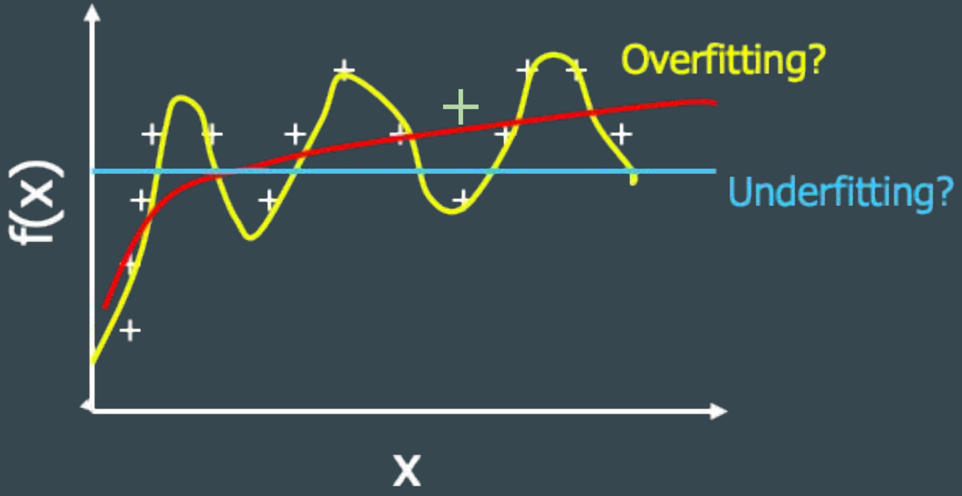


Which model does fit data the best?



Need to modelize the data

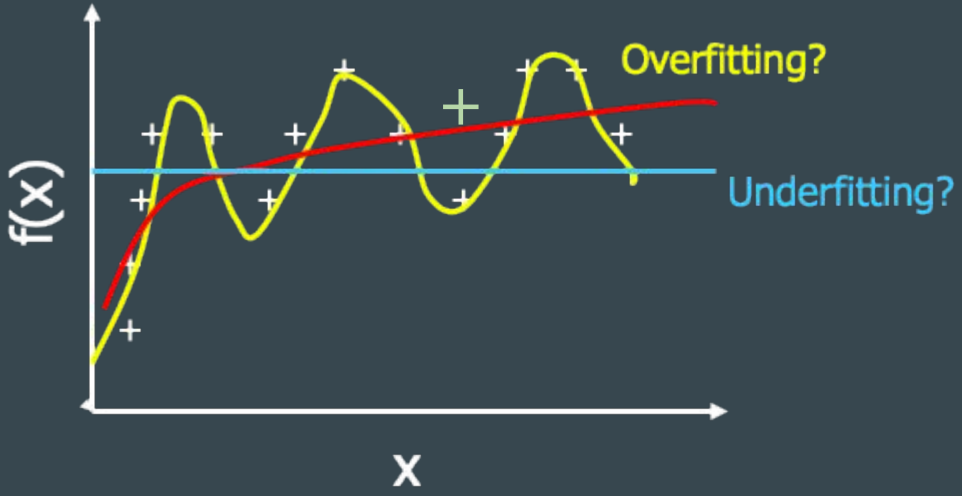
Caution with overfitting with so many variables (SNPs)



Need to modelize the data

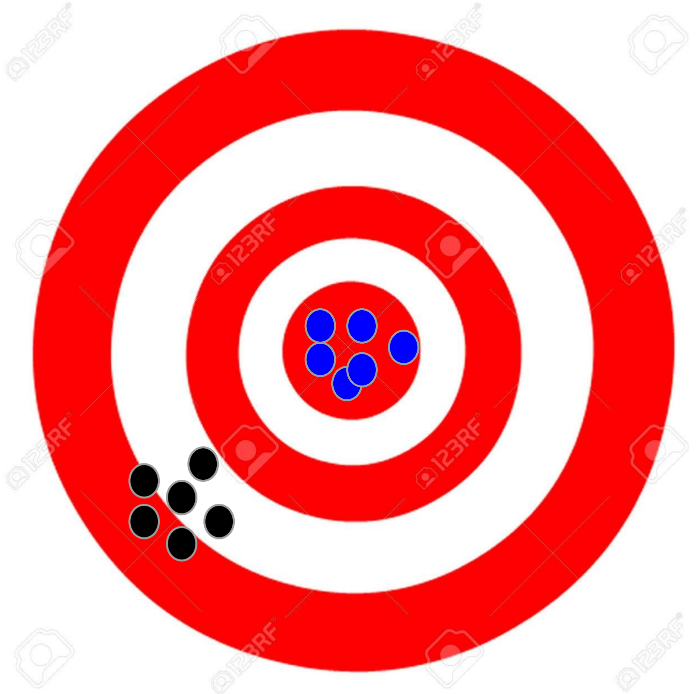
Caution with overfitting with so many variables (SNPs)

Important to control the variance-bias trade-off



Bias

a systematic distortion of a statistical result due to a factor not allowed for in its derivation

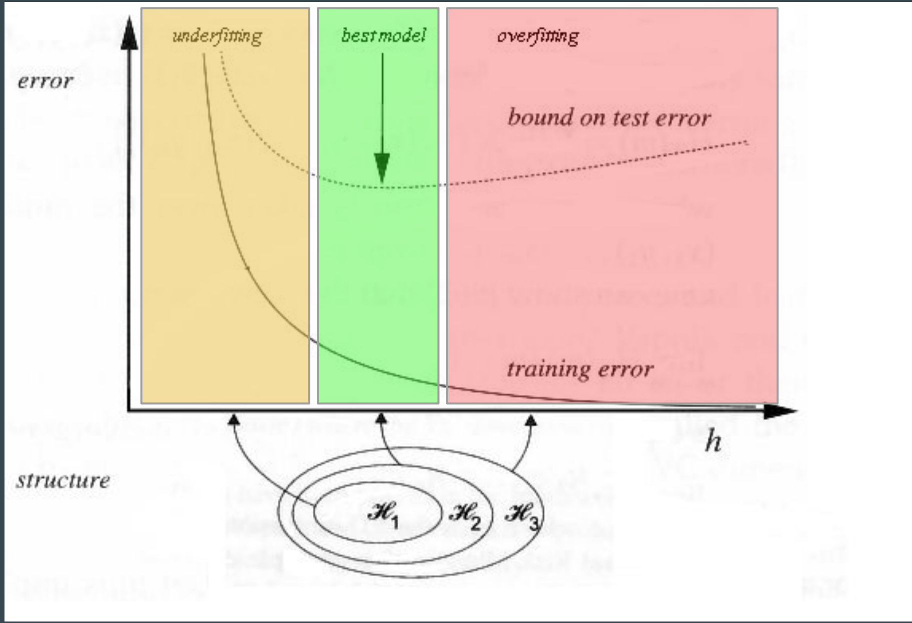


Variance

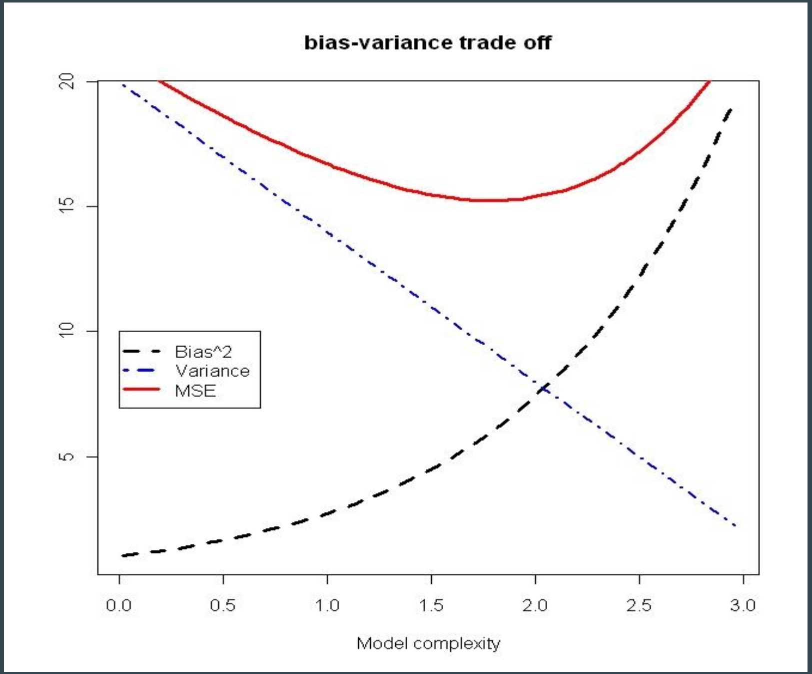
It measures how far a set of numbers is spread out from their average value

Bias - Variance trade off

Increase model complexity might overfit the data, but impair predictive ability



Bias - Variance trade off



Parameters that control the variance-bias trade off

- Variances in the mixed models
- Lambda parameter in the LASSO
- Hyperparameters in Machine Learning methods

How to control large dimensionality

- Applied more restrictive loss functions

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} (y_i - \beta_1 - \beta_2 x_i)^2,$$

$$\hat{\beta}_{\mathrm{R}} = \arg\min_{\beta}\{\|y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2\},$$

$$J[\mathbf{g}(\mathbf{x})|\lambda] = \frac{1}{2}[\mathbf{y} - \mathbf{W}\boldsymbol{\theta} - \mathbf{g}(\mathbf{x})]'\mathbf{R}^{-1}[\mathbf{y} - \mathbf{W}\boldsymbol{\theta} - \mathbf{g}(\mathbf{x})] + \frac{\lambda}{2}\|\mathbf{g}(\mathbf{x})\|_{\mathcal{H}}^2,$$

- Lasso, Bayesian LASSO, RKHS

How to control large dimensionality

- Apply simple models repeatedly over the data or residuals from previous iteration
  - Neural Networks, LASSO, Boosting, Random Forest
- Resample data during inference to increase variability of results
  - Random Forest, Bagging
- Variable selection
  - Random Forest, Neural Networks
- Use cross validation to train predictive ability

How to control large dimensionality

- Apply simple models repeatedly over the data or residuals from previous iteration
  - Neural Networks, LASSO, Boosting, Random Forest
- Resample data during inference to increase variability of results
  - Random Forest, Bagging
- Variable selection
  - Random Forest, Neural Networks
- Use cross validation to train predictive ability

More than one option

Considerations

- If a model have very high goodness of fit, it will not generalize the data (low predictive ability)
- Check predictive ability tuning hyperparameters and using internal and external cross validation.
-