# INTERPLAY BETWEEN GUT MICROBIOME AND IMMUNE SYSTEM

M. Knigge

Asst. Prof. Dr. S. Sanna

Asst. Prof. Dr. Y. Li

umcg

Hanze
University of Applied Sciences
Groningen

# OVERVIEW

- PROJECT GOAL
- BACKGROUND
- WORKFLOW

# PROJECT GOAL

> *The aim is to find causality links between microbiome composition / function and immune system with a statistical machine learning approach.*

# BACKGROUND

- LINEAR REGRESSION
- MULTIPLE LINEAR REGRESSION
- ORDINARY LEAST SQUARES
- GAUSS MARKOV THEOROM
- SHRINKAGE ESTIMATORS
- RIDGE REGRESSION
- LASSO
- ELASTIC NET REGULARIZATION
- K-FOLD CROSS-VALIDATION
- ONE-SAMPLE MENDELIAN RANDOMIZATION

# LINEAR REGRESSION

- Science of fitting lines to patterns of data
- Used for predicting and forcasting
- Simpel regression model

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Dependent variable $Y$

Intercept $\beta_0$

Coefficient $\beta_1$

Independent variable $X_1$

Error term $\epsilon$

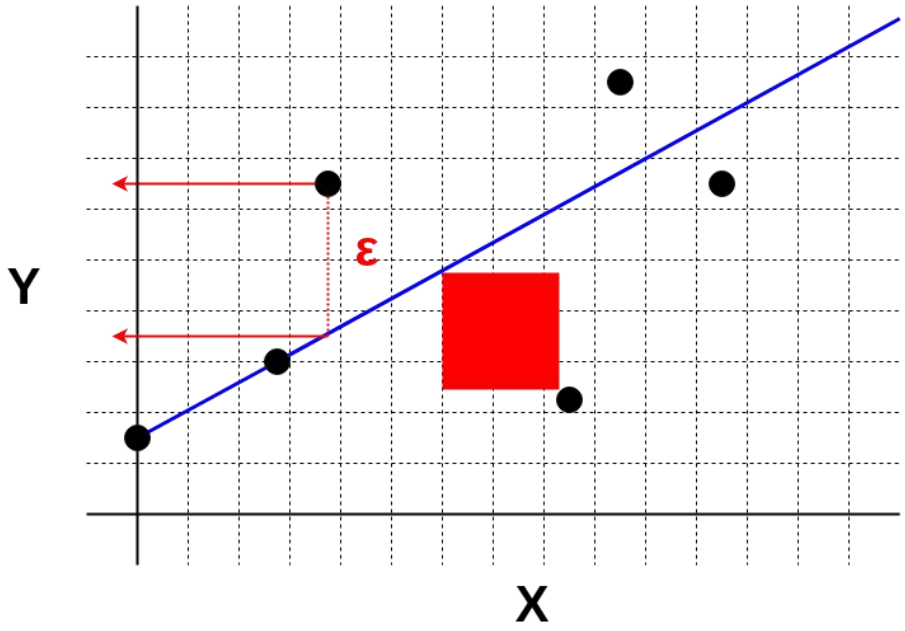
# MULTIPLE LINEAR REGRESSION

- Essentially the same as simple linear regression
- Multiple coefficients and independent variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \epsilon$$

# ORDINARY LEAST SQUARES

- Method for estimating uknown parameters in linear model
- Goal is to minimize difference in observed and predicted
- Sum of  vertical distance between observation and prediction
- Smaller difference, better model

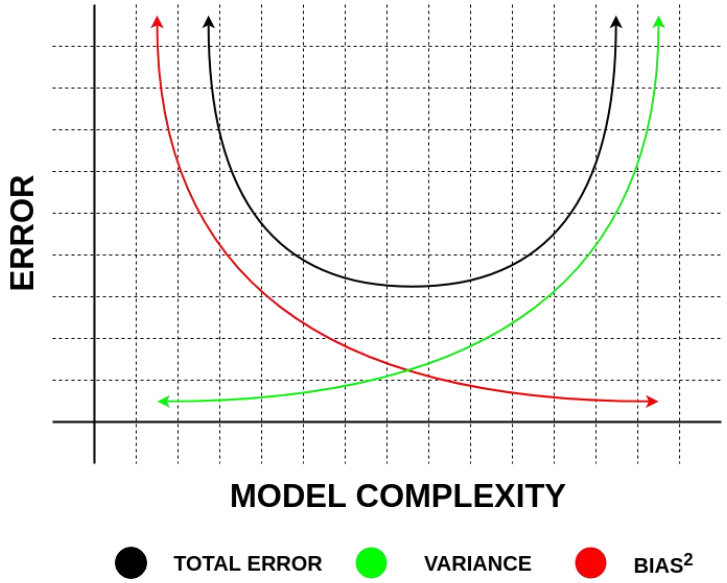


# ORDINARY LEAST SQUARES

- Sum of Squares representation of error in model
- Prediction error two components:
    * Error due to Bias
    * Error due to Variance
- Bias and Variance Tradeoff
    * Model complexity
    * Gauss Markov Theorem

# GAUSS MARKOV THEOREM

- OLS has the smallest variance
- Unbiased OLS has the smallest Mean Squared Error
- Can there be a biased estimator with a smaller mean squared error?



# SHRINKAGE ESTIMATORS

- Replace OLS estimates $\beta_k$

$$\beta'_k = \frac{1}{1+\lambda}\beta_k$$

- $\lambda == 0 = \beta_k$
- if $\lambda$ gets large, estimate approaches minimal value
- Shrinkage estimator $\lambda$

# SHRINKAGE ESTIMATORS

- Right choice of $\lambda$, better MSE
- Estimate not unbiased, but better variance

    * unbiased OLS  has smallest variance

- What is lost, is made up for with variance

# RIDGE REGRESSION

- Modeling technique to solve collinearity in OLS with $\lambda$
- OLS unbiased, but high variance, far from true value
- Adding a degree of bias to regression, Ridge reduces variance.
- Shrinking by penalty function on OLS
- $\ell_2$ shrinks coefficients of correlated values toward each other
- Drawback
    * $k$ identical predictors
    * ideal with many independent variables
    * depends on allele frequence and sample size
    * can not perform variable selection

# LASSO

- $\ell_1$ equivalent to minimizing RSS plus penalty on coefficients
- Can perform variable selection
- Drawback
    * $k$ identical variables
    * among clusters, selects one predictor and ignores the rest
    * Can not select more predictor variables than sample size
- $\ell_2$ pushes coefficients towards zero with proportional effect to coefficients
- $\ell_1$ exerts the same force on all non-zero coefficients

# ELASTIC NET REGULARIATION

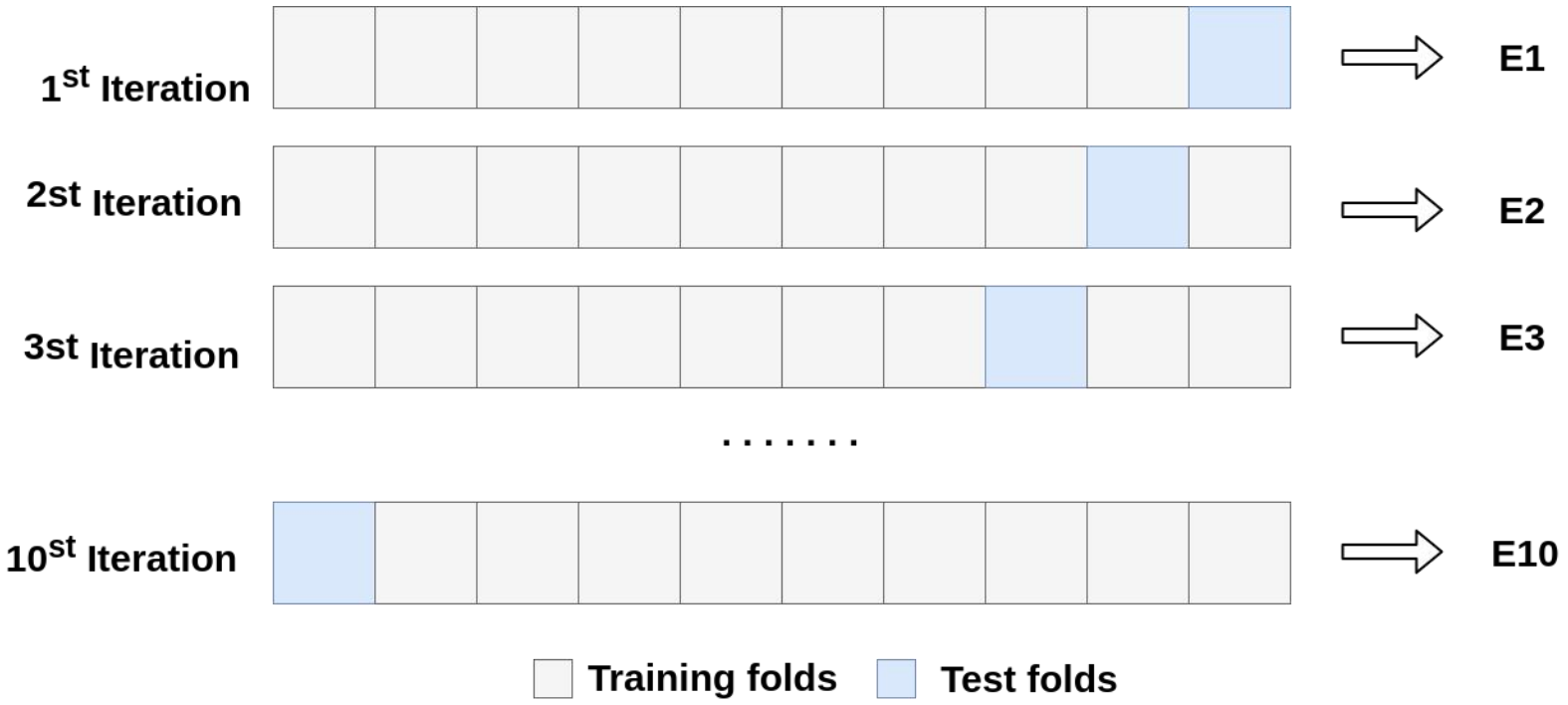
- Elastic net forms a hybrid of the $\ell_1, \ell_2$ penalties
- Elastic net penalty parameter $\alpha$ determines the weight that should be given to LASSO or Ridge
- close to 0 is Ridge
- close to 1 is LASSO

# K-FOLD CROSS-VALIDATION

- A K-fold partition of the sample space is created
- The original sample is randomly partitioned into K equal sized
- Of the K subsamples
    * a single subsample is retained as the test set
    * and the remaining as training data set
- The cross-validation process is then repeated K times

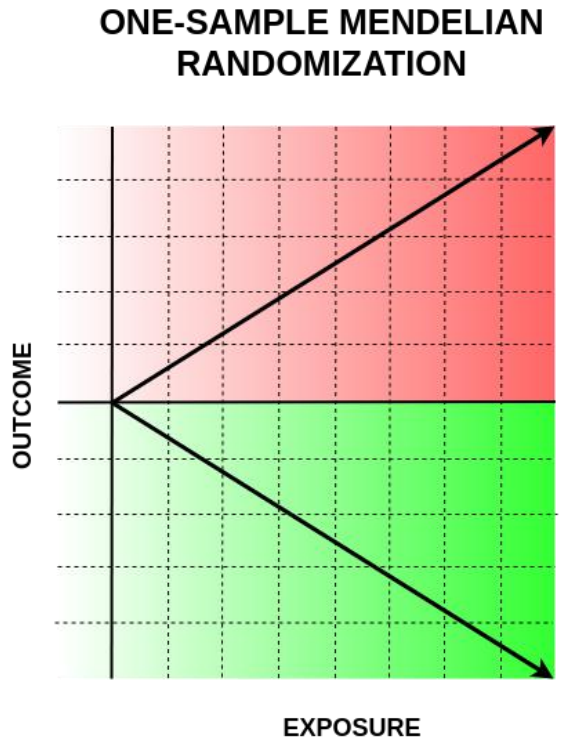
- This procedures takes place for different $\alpha$ levels with Elastic Net

# K-FOLD CROSS-VALIDATION

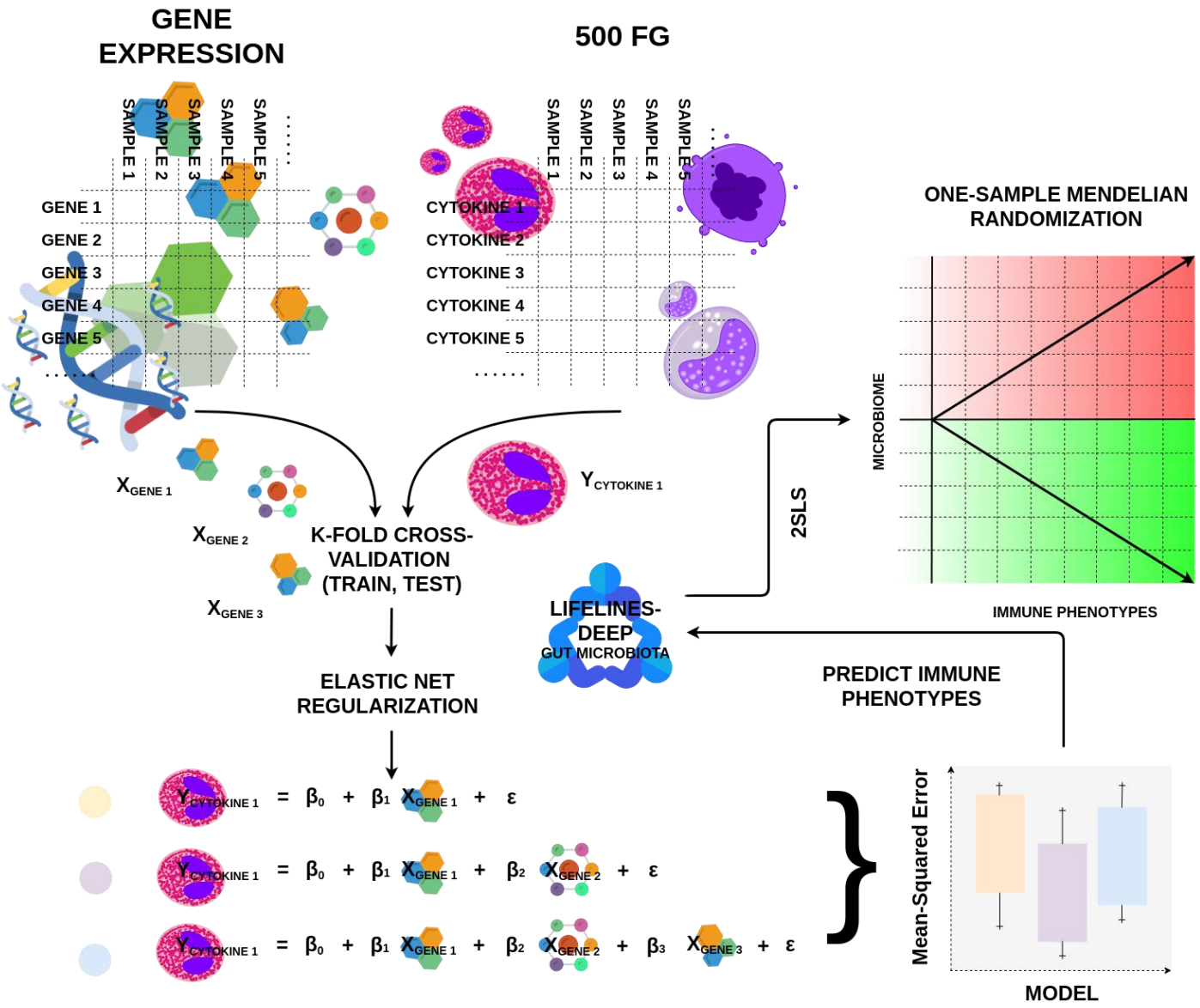


# ONE-SAMPLE MENDELIAN RANDOMIZATION

- Basic implementation of MR
- Two-Stage Least Squares regression



# WORKFLOW


GENE EXPRESSION
500 FG
SAMPLE 1
SAMPLE 2
SAMPLE 3
SAMPLE 4
SAMPLE 5
GENE 1
GENE 2
GENE 3
GENE 4
GENE 5
CYTOKINE 1
CYTOKINE 2
CYTOKINE 3
CYTOKINE 4
CYTOKINE 5
ONE-SAMPLE MENDELIAN RANDOMIZATION
MICROBIOME
IMMUNE PHENOTYPES
2SLS
$X_{GENE\ 1}$
$X_{GENE\ 2}$
$X_{GENE\ 3}$
$Y_{CYTOKINE\ 1}$
K-FOLD CROSS-VALIDATION (TRAIN, TEST)
LIFELINES-DEEP
GUT MICROBIOTA
ELASTIC NET REGULARIZATION
PREDICT IMMUNE PHENOTYPES
$Y_{CYTOKINE\ 1} = \beta_0 + \beta_1 X_{GENE\ 1} + \varepsilon$
$Y_{CYTOKINE\ 1} = \beta_0 + \beta_1 X_{GENE\ 1} + \beta_2 X_{GENE\ 2} + \varepsilon$
$Y_{CYTOKINE\ 1} = \beta_0 + \beta_1 X_{GENE\ 1} + \beta_2 X_{GENE\ 2} + \beta_3 X_{GENE\ 3} + \varepsilon$
Mean-Squared Error
MODEL