Least Absolute Shrinkage And Selection Operator (LASSO)

Peter von Rohr

22.03.2021

Fixed Linear Effect Model

- Back to

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

- All $\beta_0, \beta_1, \ldots, \beta_p$ into vector β of length $(p + 1)$

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

- Only random componente: ϵ with

$$E(\epsilon) = 0 \text{ and } var(\epsilon) = I * \sigma^2$$

Parameter Estimation

- Least Squares

$$\hat{\beta}_{LS} = argmin_{\beta}||y - X\beta||^2$$

- Normal Equations

$$(X^T X)\hat{\beta}_{LS} = X^T y$$

- Existence of $(X^T X)^{-1}$?

1. Yes: $\hat{\beta}_{LS} = (X^T X)^{-1} X^T y$
2. No: $b_0 = (X^T X)^{-} X^T y$

with $(X^T X)^{-}$ being a generalized inverse of $(X^T X)$

Generalized Inverse

- System of equations

$$Ax = y$$

with coefficient matrix A, vector of unknowns x and vector of right hand side y

- If A^{-1} exists, then unknowns $x = A^{-1}y$
- If A^{-1} does not exist, $x = A^{-}y$ is one solution with A^{-} being a generalized inverse
- Generalized inverse A^{-} defined by

$$AA^{-}A = A$$

Solutions

- Why is A^- a solution
  - if $AA^-A = A$, then $AA^-Ax = Ax$
  - when $Ax = y$, this gives $A(A^-y) = y$
  - hence $A^-y = x$ is a solution
- If A^- is a generalized inverse of A then $Ax = y$ has solutions

$$\tilde{x} = A^-y + (A^-A - I)z$$

for aribitrary z

- Proof

$$A\tilde{x} = AA^-y + A(A^-A - I)z = AA^-y + (AA^-A - AI)z = AA^-y = y$$

because $AA^-A = A$.

Results

- $b_0 = (X^T X)^- X^T y$ is a solution to $(X^T X) b_0 = X^T y$
- But b_0 is not unique, because for any $(X^T X)^-$

$$\tilde{b}_0 = (X^T X)^- X^T y + ((X^T X)^- (X^T X) - I)z$$

is also a solution

- b_0 cannot be an estimate for β

Estimable Functions

Idea: construct linear functions ($q^T\beta$) of the parameters β such that

- estimator can be found from b_0
- independent of choice of b_0

Such linear functions $q^T\beta$ must satisfy

$$q^T\beta = t^T E(y)$$

for any vector t, then $q^T\beta$ is **estimable**

- Determine q as

$$q^T = t^T X$$

Invariance to b_0

When $q^T\beta$ is estimable, then

- $q^T b_0$ is always the same, independent of choice of b_0
- Why?
- With $q^T = t^T X$

$$q^T b_0 = t^T X b_0 = t^T X(X^T X)^- X^T y$$

is independent of choice of b_0 because $X(X^T X)^- X^T$ is independent of choice of $(X^T X)^-$

Summary

Use of generalized inverse $(X^T X)^-$ of normal equations yields

- solutions b_0
- estimatble functions $q^T b_0$ which estimate $q^T \beta$
- independent of b_0

But for genomic data

- no possibility to determine important SNP loci
- need an alternative to least squares

Alternatives To Least Squares

Desirable properties

1. **Subset Selection**: determine important predictors
2. **Shrinkage**: limit parameter estimates to certain area
3. **Dimension Reduction**: Reduce p predictors to m linear combinations where $m < p$

LASSO

- ... stands for Least Absolute Shrinkage and Selection Operator
- ... combines subset selection (1) and shrinkage (2)
- shrinkage is achieved by introduction of penality term
- subset selection is due to the form of penalty term

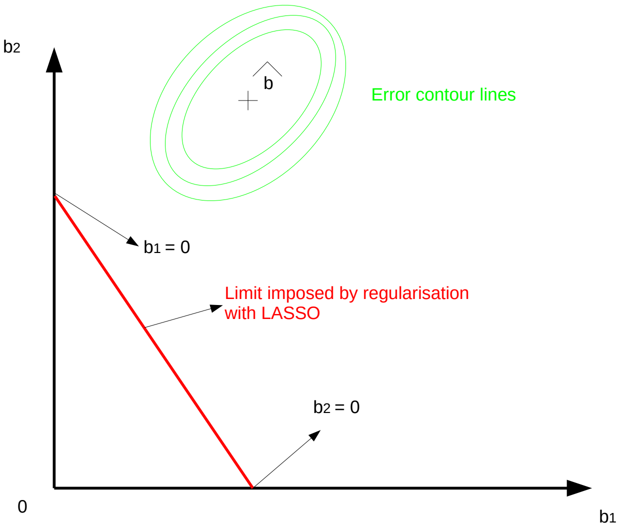
Shrinkage

- penalty term added to least squares criterion

$$\hat{\beta}_{LASSO} = argmin_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- large values of $|\beta_j|$ are penalized compared to small $|\beta_j|$

Subset Selection



Find λ

- λ is an additional parameter to be estimated from data
- use cross validation
  - split data randomly into training set $(80 - 90\%)$ and test set $(10 - 20\%)$
  - assume a certain λ value and do parameter estimation with training data
  - try to predict test data with estimated parameters
  - repeat this many times
  - take that λ with the best predictive performance