Least angle regression

Antti Ajanki

antti.ajanki@hut.fi

Linear problems

- $\boldsymbol{x}_i$ are predictor variables and y_i is the response
- Find β_j which minimize squared error

$$\sum_i (y_i - \hat{\boldsymbol{\mu}}_i)^2$$

- where $\hat{\boldsymbol{\mu}}_i = \sum_j \beta_j x_{ij}$
- Ordinary least squares: $\boldsymbol{\beta} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- Let’s assume that $\boldsymbol{y}$ and all $\boldsymbol{x}_i$ are centered to zero and $\boldsymbol{x}_i$ have unit length

Forward selection

- Forward selection
  - Start with empty model
  - Select variable most correlated with output
  - Linear regression from the variable to y
  - Project other predictors orthogonally to the selected variable
  - Repeat
- Usually overly greedy
- Eliminates variables correlated with selected ones

LARS compared to other algorithms

- Least Angle Regression (LARS) ”less greedy” than ordinary least squares
- Two quite different algorithms, Lasso and Stagewise, give similar results
- LARS tries to explain this
- Significantly faster than Lasso and Stagewise

Lasso

- Lasso is a constrained version of OLS

$$\min \textstyle\sum_i (y_i - \hat{\boldsymbol{\mu}}_i)^2$$
$$\text{subject to } \textstyle\sum_j |\beta_j| \leq t$$

- Can be solved with quadratic optimization or with iterative techniques
- Parsimonious: $\beta_j \neq 0$ only for some j
- Increasing t selects more variables

Stagewise regression

- Forward stagewise linear regression
  - Choose $\boldsymbol{x}_j$ with highest current correlation $c_j = \boldsymbol{x}_j^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$
  - Take a small step $0 < \epsilon < |c_j|$ in the direction of selected $\boldsymbol{x}_j$
  - $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} + \epsilon \cdot \mathrm{sign}(c_j) \cdot \boldsymbol{x}_j$
  - Repeat
- ”Large” step size of $|c_j|$ would lead to ordinary least squares

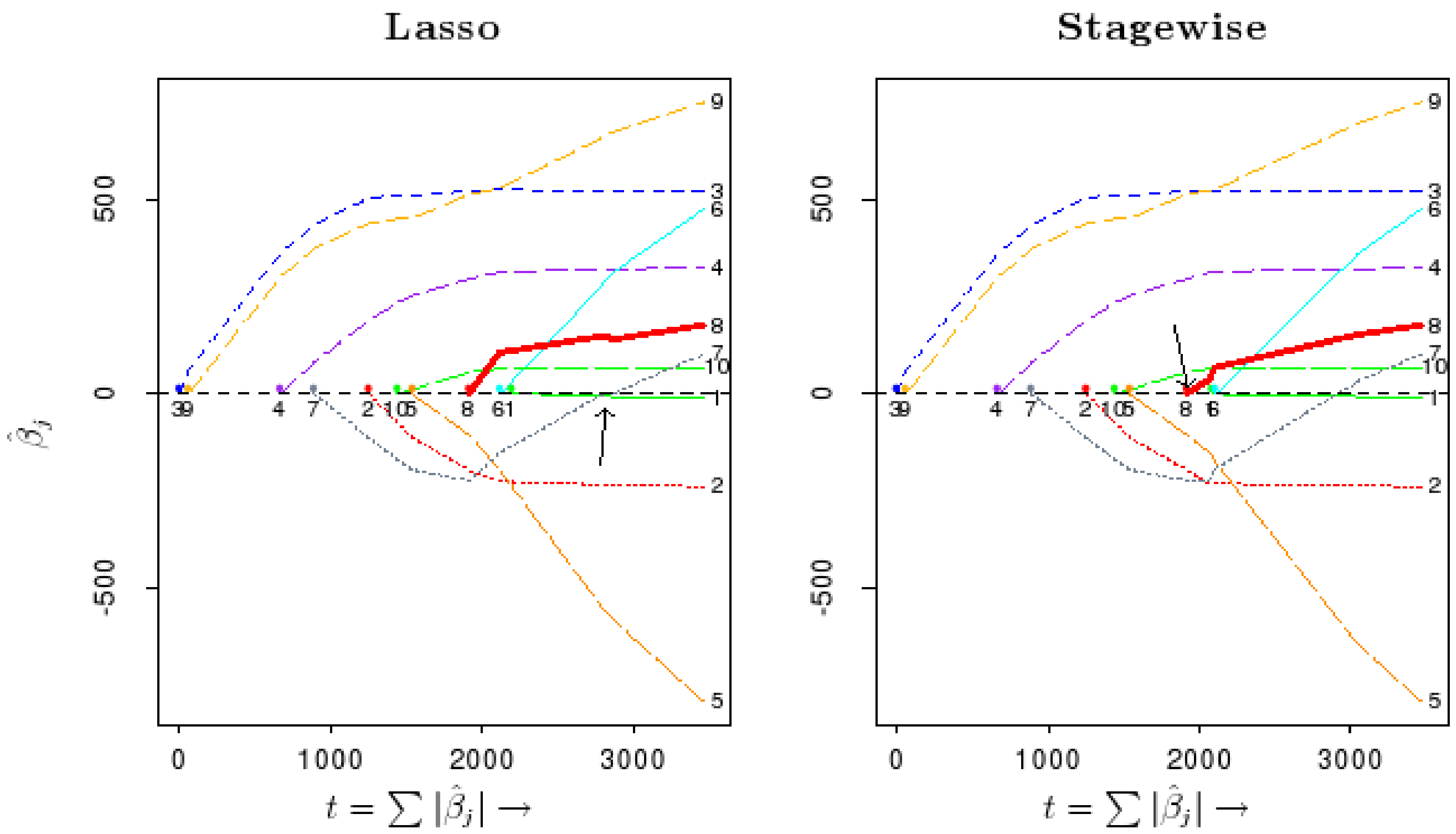
Non-linear extension

- Stagewise idea can be easily expanded non-linearly
- Boosting
  - Fit regression tree to residuals
  - Finds the most correlated tree
  - Take a small step in the direction of fitted values
  - Repeat

Diabetes data

- Main example in the paper
- $n = 442$ patients
- 10 variables: age, body mass index, blood pressure, serum measurements, $\ldots$
- Response variable is ”a quantitative measure of disease progression one year later”
- Variable selection problem: which of the variables affect the disease

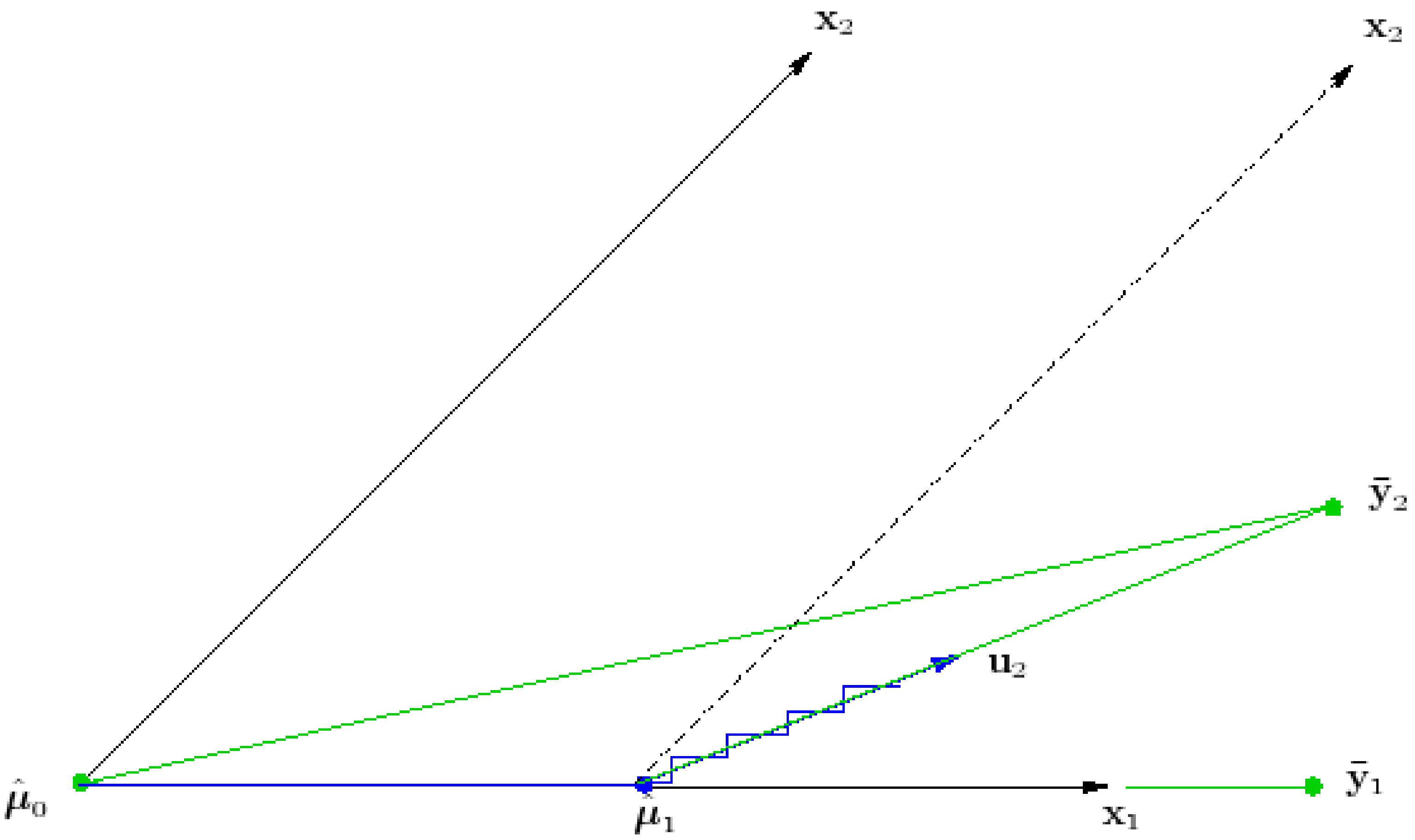
Comparsion of Lasso and Stagewise



LARS

- Least Angle Regression
  - Start with empty set
  - Select $\boldsymbol{x}_j$ that is most correlated with residuals $\boldsymbol{y} - \hat{\boldsymbol{\mu}}$
  - Proceed in the direction of $\boldsymbol{x}_j$ until another variable $\boldsymbol{x}_k$ is equally correlated with residuals
  - Choose equiangular direction between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$
  - Proceed until third variable enters the active set, etc
- Step is always shorter than in OLS

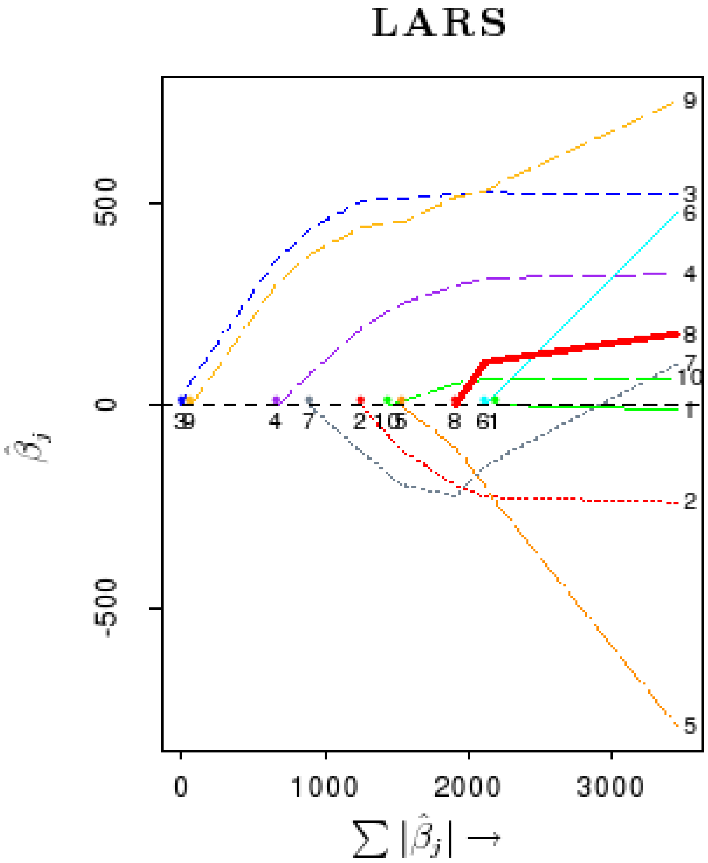
Geometrical presentation



Computing LARS

- Every step a new variable enters the active set $\Rightarrow$ no more steps than variables
- New dicrection can be solved with linear algebra
- Step length by iterating over all variables not in active set
- By cleverly updating estimates from previous iteration, the computation cost will be comparable to OLS

LARS results



Lasso modification

- LARS can be modified to give Lasso solution
- In the Lasso algorithm signs of the β_j and c_j must agree
- Take only as long LARS step as possible without changing the sign
- This works, if only one variable at a time enters the active set
- Unlike LARS, variables can be removed from active set

Stagewise modification

- The Stagewise step is positive combination of active set variables
- LARS has no sign restrictions on the direction vector
- Project LARS vector to positive convex cone
- This leads to Stagewise solution (assuming Stagewise step size $\rightarrow 0$)

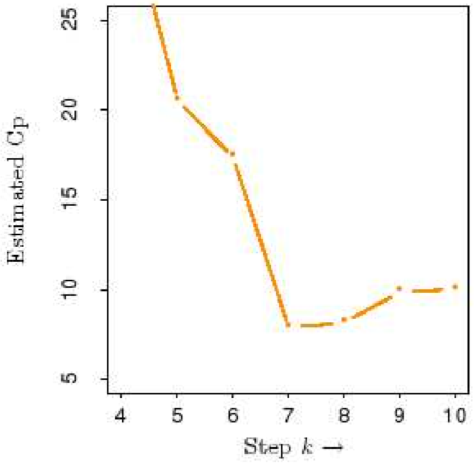
Other modifications

- *LARS/OLS hybrid*
  - Select the model with LARS, but the parameter values with OLS
- *Main effects first*
  - Run LARS until most important varibles are in the active set
  - Restart with predictor variables replaced by interaction terms between already selected variables

Stopping criteria

- When to stop?
- $C_p(\hat{\boldsymbol{\mu}})$ is an unbiased estimator of $E\left(\frac{||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}||^2}{\sigma^2}\right)$
- Simple formula for C_p in the kth LARS step:

$$C_p(\hat{\boldsymbol{\mu}}_k) = ||\boldsymbol{y} - \hat{\boldsymbol{\mu}}_k||^2/\bar{\sigma}^2 - n + 2k$$



Summary

- Lasso and Stagewise can be seen as modifications of LARS
- This explains similar results
- LARS is more efficient to compute
- Other modifications