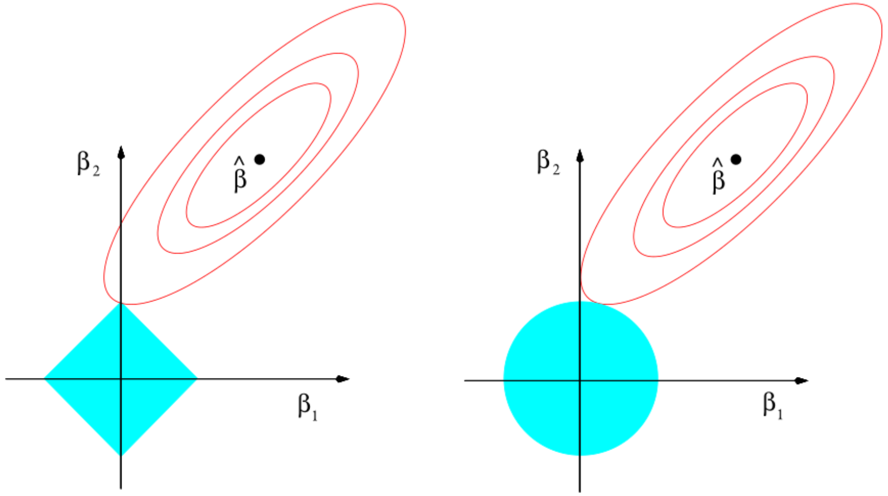
Distributed Lasso

SEBASTIEN { DUBOIS, LEVY }

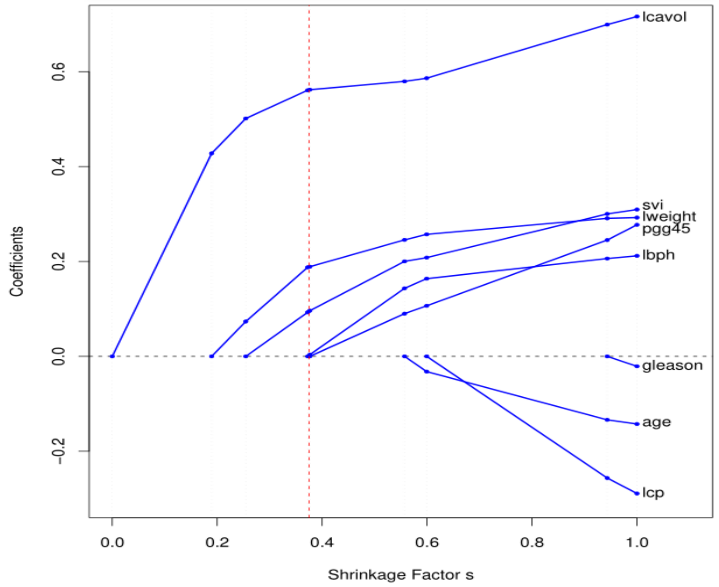
The Lasso

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Sparsity



Coefficient Path



Existing distributed methods

By SGD (Spark)

- Sparsity not guaranteed
- Problem when $p \gg n$
- Coefficient path ?

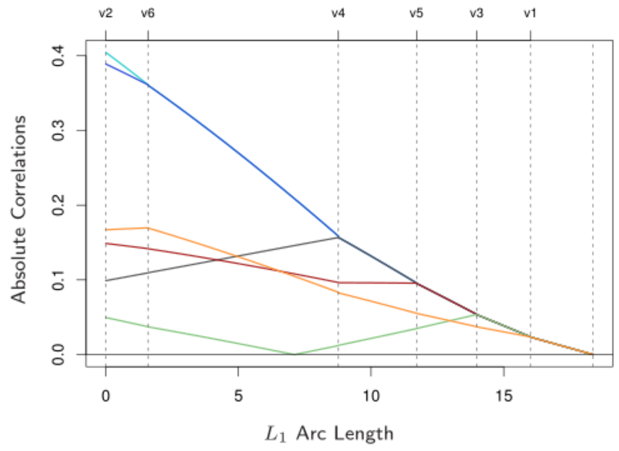
Shotgun (Distributed Coordinate Descent)

- Too much communication
- Convergence issues (need locks)
- $n \gg p$? (Data stored by column)

Least Angle Regression (LAR)

Algorithm

- Relax the penalty at each iteration
- Add features most correlated with the residual
- Increase coefficients to keep correlation tied



Correlations with residual

Computations

- Dot products between columns
- `k x k` matrix inversion
- `min(n,p)` iterations

$$\delta_k \leftarrow (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k$$

New coefficient direction

Distributed LAR

Tall and Skinny
$p \ll n$

- p^2 fits in memory,
  n does not

- Data stored by rows

- B machines, I iterations:
  - Communication:
    $$O(p^2 B)$$
  - Computations:
    $$O\left(pI\left(p + \log(B) + \frac{n}{B}\right)\right)$$

Small and Fat
$p \gg n$

- n^2 fits in memory,
  p does not

- Data stored by columns

- B machines, I iterations:
  - Communication:
    $$O(n^2 B)$$
  - Computations:
    $$O\left(nI\left(n + \log(B) + \frac{p}{B}\right)\right)$$

Almost Square
$p \sim n$

- p and n fit in memory,
  n^2 and p^2 do not

- If $\mathrm{I} < \sqrt{n}$ both methods work

- More iterations:
  - Inversion using distributed block matrix
  - SGD to solve linear system
  - Forward Stagewise