Shrinkage methods

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

- examples: ridge and lasso regressions, elastic net...

Goals:
- improve prediction performance with respect to standard regression
- choose most important features, reducing total number of features (feature selection)

Shrinkage methods were invented as alternative to standard feature selection
- ↳ keep or discard each variable
- ↳ discrete ⇒ highly variable methods (depending on train data)

↳ continuous process:
penalize coeff's smoothly (damping) ⇒ less variance

Ridge and Lasso regressions penalize large coefficients β_is

Why large coeff's are bad?
- they impose high variability (when values of X_i's change slightly)
- they may appear as a result of correlation between variables and produce false impression of importance of some β_is

(example: X_k and X_m are dependent, then large positive coeff's β_k cancels large negative coeff's β_m (think $X_k = -X_m$), so both can grow without penalty in LSS-loss, but maybe should not be that large)

Ridge regression

- directly penalizes <u>size</u> of coefficients

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

$\lambda > 0$

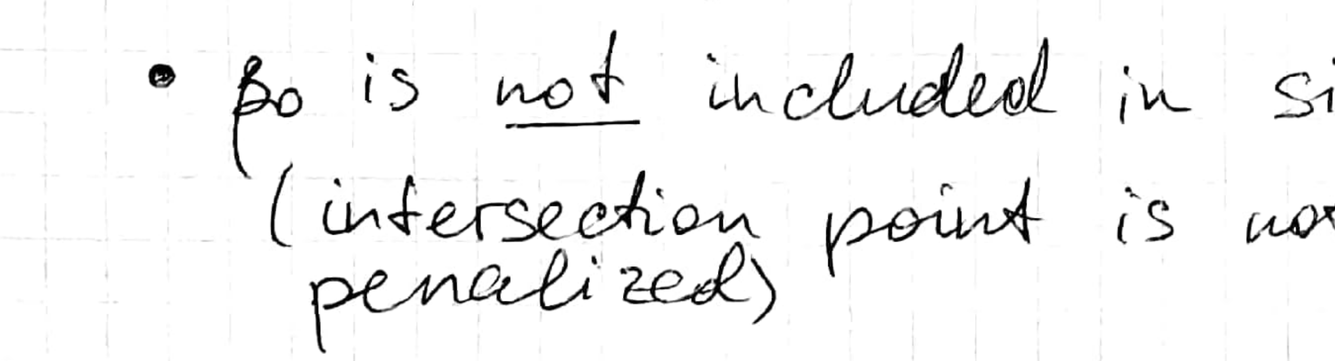
or

$$\hat{\beta}_{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$

explicit size constraint

- t (or λ) is chosen to minimize prediction error on cross validation set



- ridge solutions are <u>not</u> equivalent under scaling of coeff's $\Rightarrow$ <u>normalization</u> is a common preprocessing
- and centering $x_{ij} \mapsto x_{ij} - \text{mean}(x_{ij})$
- β_0 is <u>not</u> included in size penalty $\sum_{j=1}^{p} \beta_j^2$ (intersection point is not penalized)

- Ridge solutions in the matrix form

$$RSS(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \longrightarrow \min$$

$$\hat{\beta}_{Ridge} = (X^T X + \lambda \cdot Id)^{-1} X^T y \quad (1)$$

sometimes used as definition of ridge regression

- λ - positive constant on diagonal in (1) - makes matrix invertible

- Using SVD-decomposition $X = UDV^T$ we can write

→ $X\hat{\beta} = UU^T y$ (for standard regression solution $\hat{\beta}$ $U^T y$ - coord of y in basis U)

→ $X\hat{\beta}_{ridge} = X(X^TX + \lambda Id)^{-1}X^Ty = \dots$

$= UD(D^2 + \lambda \cdot Id)^{-1} D U^T y =$

$= \sum_{j=1}^{p} u_j \cdot \underbrace{\frac{d_j^2}{d_j^2 + \lambda}} \cdot u_j^T \cdot y$ u_j - columns of U

(also • finds $u_j^T y$ - coord of y in basis U ≤ 1
• then shrinks them by factor $\frac{d_j^2}{d_j^2+\lambda}$, i.e.

(!) shrinks less important principal components more

- $df(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda} = \text{trace}\left(X(X^TX+\lambda I)^{-1}X^T\right)$

effective degrees of freedom

$\lambda = 0$ $df = p$ degrees of freedom

$\lambda \to \infty$ $df \to 0$

(idea: p dimensions, features, Ridge regression does not cast any of them to zero but effectively gives less freedom for coefficient choice when $\lambda \to$, for $\lambda = \infty$ all coeffs are zero)

LASSO - Least Absolute Shrinkage and Selection Operator regression

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

Lagrangian form

or

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

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

- l_1-penalty instead of l_2 $\Rightarrow$

  no closed form expression for the solution

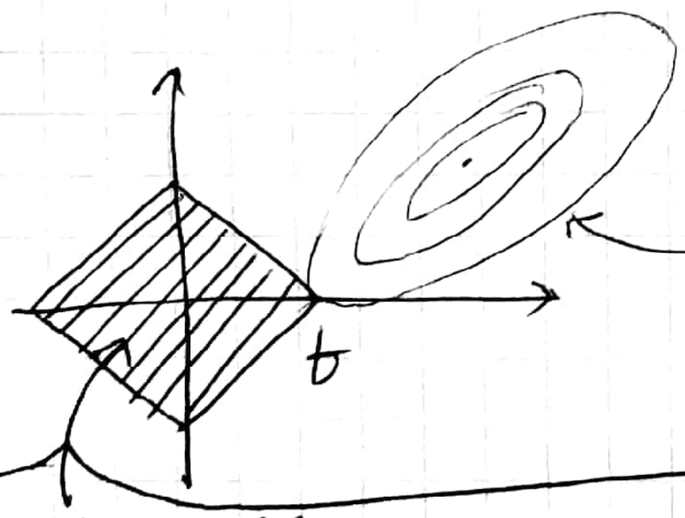
  however efficient algorithms are available to compute "path" solutions (for all t values) at the same cost as for ridge

- if $t \geq t_0 := \sum_{j=1}^{p} |\hat{\beta}_j|$ — coeff's of standard regression.

  $\Rightarrow \hat{\beta}_j^{lasso} = \hat{\beta}_j$

  elif $t = \varepsilon \cdot t_0$

  $\Rightarrow$ coeff's are shrunk ε times on average, but many of them tend to be exactly zero



t. unit ball for l_1-norm points inside satisfy $\sum_{j=1}^{2} |\beta_j| \leq t$

label curves for sum of least squares

optimal β — smallest level curve that touches unit ball

This tend to happen on a sharp angle

$\Rightarrow$ SPARSITY

$\Rightarrow$ CONTINUOUS FEATURE SELECTION

Comparison of Ridge and Lasso

0. Ridge has closed form solution, lasso does not
1. Lasso does sparse selection, ridge does not
2. For highly-correlated variables - ridge will shrink 2 coeff's towards each other, lasso picks one and sends other to zero (most likely)
3. Ridge penalizes largest β's more, lasso is uniform

In orthogonal case (orthogonal data vectors)

- Ridge does proportional shrinkage
- Lasso translates all coeff's down by λ, truncating at zero ("soft thresholding")
- We can also do "hard thresholding" - put all coeff's that are $\leq M$ to zero ("more variance")

Other options

- Bridge regression

$$\hat{\beta}_b = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \cdot \sum_{j=1}^{p} |\beta_j|^q \right\}$$

$\lambda, q \geq 0$

- q-norm regularization (generalization of ridge: $q=2$ and lasso: $q=1$)
- can think about $|\beta_j|^q$ as a prior distribution of β_j (i.e. assumption about "how coeff's should look like in general")
- $q<1$ - non-convex constraint (hard for optimization)
  - very sharp angles - more mass on coordinate directions
- ESL book: "one might consider estimating q from data, ... not worth the effort for the extra variance incurred"

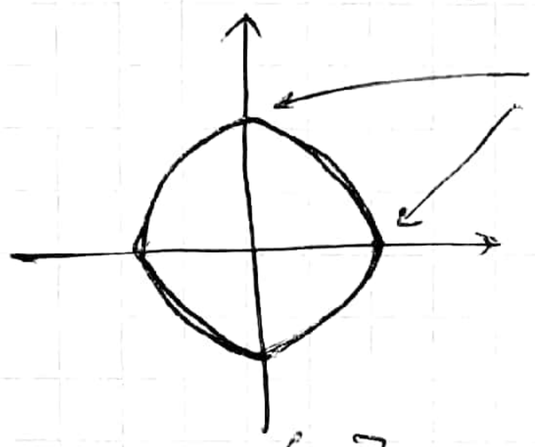
- choose $q \in (1, 2)$ suggests compromise between Bridge and lasso, but it loses the ability of lasso to make sparsity - zero coeff's (due to sharp angles)

- Elastic net (better compromise)

  - penalty $\lambda \cdot \sum_{j=1}^{p} (\alpha \cdot \beta_j^2 + (1-\alpha) \cdot |\beta_j|)$

  - selects coordinates like lasso + shrinks them like ridge

  - unit ball



  - relatively new (Zue & Hastie 2005)

- LAR - Least Angle Regression (Efron, 2004) - see ESL book pp 73-79

- NNG - non-negative Garotte (Breiman, 1995)