Foundations of Machine Learning
CentraleSupélec — Fall 2017

7. Regularized linear regression

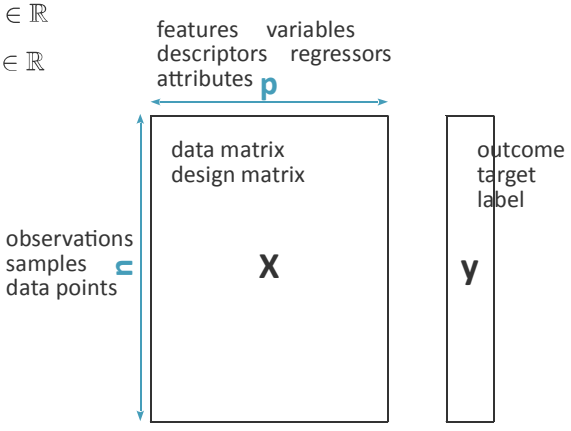
Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
chloe-agathe.azencott@mines-paristech.fr

Learning objectives

- Understand **regularization** as a means to control model complexity.
- Define **Lasso**, **ridge regression**, **elastic net.**
- Understand the role of the **l1 and l2 norms** in regularization
- Interpret **solution paths** for Lasso and ridge regression.

Regression setting

$x_j^i \in \mathbb{R}$
$y^i \in \mathbb{R}$

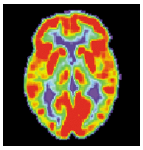


Large p, small n

E.g.

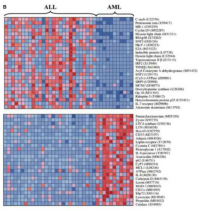
- **neuroimaging**

  thousands of brain regions / pixels / voxels
  much fewer patients



- **genetics and genomics**

  thousands of genes, millions of SNPs...
  usually, at best thousands of patients



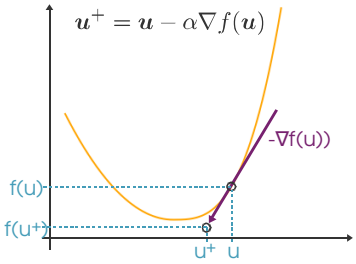
When X^TX not inversible

$$(X^\top X)\hat{\beta} = X^\top y$$

- Pseudo-inverse
- Linear system of p equations:

Numerical methods

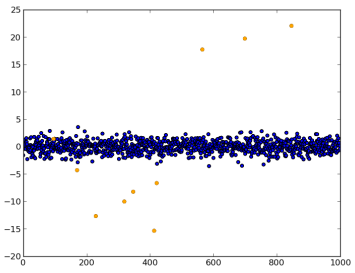
- **Gaussian elimination**
- **LU decomposition**
- **Gradient descent**



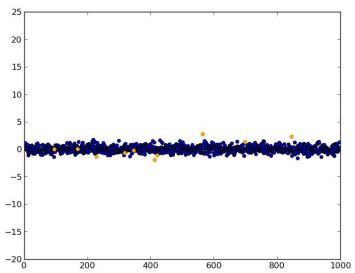
Linear regression when p >> n

Simulated data: p=1000, n=100, 10 causal features

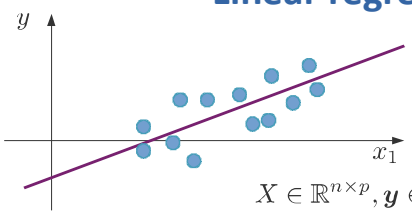
True coefficients



Predicted coefficients



Linear regression



$$\boldsymbol{x} \in \mathbb{R}^p$$

$$f(\boldsymbol{x}|\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j x_j + \beta_0$$

$$X \in \mathbb{R}^{n \times p}, \boldsymbol{y} \in \mathbb{R}^n$$

Least-squares fit (equivalent to MLE under the assumption of Gaussian noise):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) = (X^\top X)^{-1} X^\top \boldsymbol{y}$$

Properties of the least-squares fit estimate

- **Unbiased** $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- **Explicit form** $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$
- **Computational time**

When p is large

- **p > n:**

- **Multicollinearity:**
- Large p **reduces interpretability** of the model

Would prefer a small subset of features with strong effects (= large coefficients).

Regularization

- Minimize

  Prediction error + **λ** penalty on model complexity

- **Biased estimator** when λ ≠ 0.
- Trade bias for a smaller variance.
- λ can be set by cross-validation.

- Simpler model ≈ fewer parameters

  → **shrinkage:** drive the coefficients of the parameters towards 0.

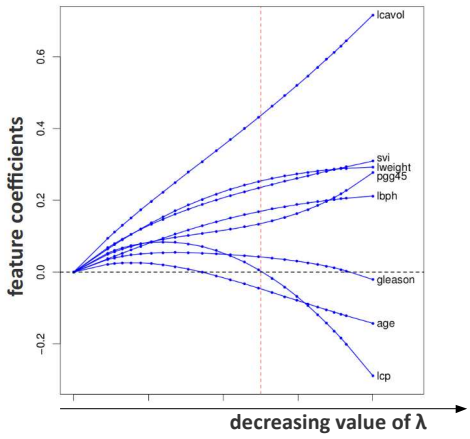
Ridge regression

- **Sum-of-squares penalty**

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_2^2$$

- **Ridge regression estimator:**

Ridge regression solution path



Standardization

- **Multiply x_j by a constant:**
  - For **standard linear regression**
  - For **ridge regression**

Ridge regression

- **Grouped selection:**
  - correlated variables get similar weights
  - identical variables get identical weights
- Ridge regression shrinks coefficients towards 0 but does not result in a **sparse model**.
- **Sparsity:**
  - many coefficients get a weight of 0
  - they can be eliminated from the model.

Lasso

- **L1 penalty**

$$||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_p|$$

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} ||y - X\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$$

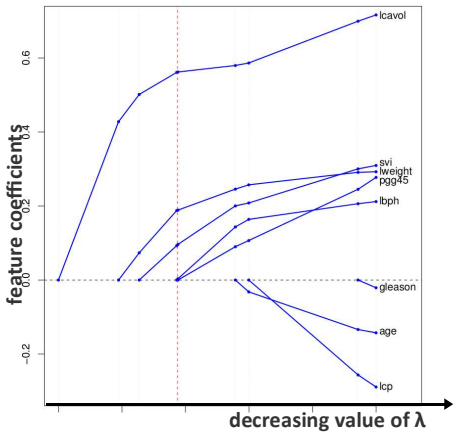
- aka **basis pursuit** (signal processing)
- no closed-form solution
- Equivalent to

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} ||y - X\boldsymbol{\beta}||_2^2 \quad \text{s.t. } ||\boldsymbol{\beta}||_1 \leq t$$

for a unique one-to-one match between t and λ.

Optimization problem:

Lasso solution path



Forward stepwise regression

- Build model **sequentially**, adding one variable at a time
  - Start with the intercept
  - At each step, add the variable that **most improves the fit**
  - **Stop when** $||\beta||_1 \leq t$
- Greedy solution

Least Angle Regression

At each step, add "only as much of a variable as needed"

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.

4a. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

5. Continue in this way until all p predictors have been entered.

Approaches to dimensionality reduction

- **Feature selection**

  Choose m < p features, ignore the remaining (p-m)

  - **Filtering** approaches

    Apply a statistical measure to assign a score to each feature (correlation, χ^2-test).

  - **Wrapper** approaches

    Search problem: Find the best set of features for a given predictive model.

  - **Embedded** approaches

    Simultaneously fit a model and learn which features should be included.

All these feature selection approaches are **supervised.**

Elastic Net

- **Combine lasso** and **ridge regression**

$$\hat{\boldsymbol{\beta}}_{\text{enet}} = \arg\min_{\boldsymbol{\beta}} ||y - X\boldsymbol{\beta}||_2^2 + \lambda \left(\alpha||\boldsymbol{\beta}||_2^2 + (1-\alpha)||\boldsymbol{\beta}||_1\right)$$

  - **Select variables** like the lasso.
  - **Shrinks together coefficients of correlated variables** like the ridge regression.

Lq-norm regularization

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||Y - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_q^q \qquad ||\boldsymbol{\beta}||_q = \left(\sum_{j=1}^{p} |\beta_j|^q\right)^{1/q}$$

Equivalently:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} ||Y - X\boldsymbol{\beta}||_2^2 \text{ s. t. } ||\boldsymbol{\beta}||_q^q \leq s$$

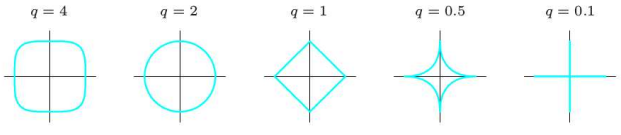
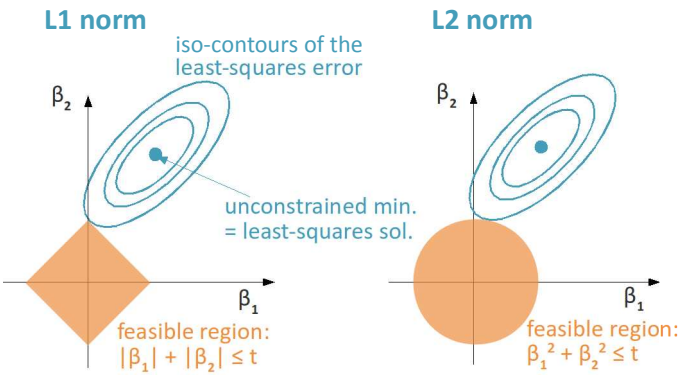


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

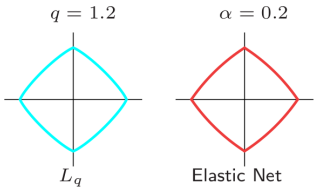
Lasso vs. ridge



Elastic net

- **Elastic penalty**

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda \left(\alpha||\beta||_2^2 + (1-\alpha)||\beta||_1\right)$$



Structured regularization

Group lasso

Use K predefined groups of variables that are known to "work" together and expected to be either all active or all inactive together.

E.g.: genes belonging to the same biological pathway.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||y - \sum_{k=1}^{K} X_k \boldsymbol{\beta}_k||_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k}\, ||\boldsymbol{\beta}_k||_2$$

Features belonging to group k

Size of group k

Other examples of structured penalties

- **Overlapping groups**

  Jacob et al. (2009). Group lasso with overlap and graph lasso. *ICML.*

- **Graphs**

  Li & Li (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. App. Stats.*

- **Trees**

  Zhao et al. (2006). Grouped and hierarchical model selection through composite absolute penalties. *Ann. Stat.*

- **Multiple related tasks**

  Obozinski et al. (2006). Multitask feature selection. *Technical Report, UC Berkeley.*

Minimize SSE + λ x regularizer

- **Ridge regression**
  - gives similar weights to similar variables
  - not sparse
- **Lasso**
  - randomly picks one of several correlated variables
  - sparse
- **Elastic net**
  - selects variables like the lasso
  - shrinks together the coefficients of correlated variables.
- **Many other regularizers** are possible

References

- *A Course in Machine Learning.* `http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf`
  - **Regularization:** Chap 7.2–7.3
- *The Elements of Statistical Learning.* `http://web.stanford.edu/~hastie/ElemStatLearn/`
  - **Regularization:** Chap 10.12
  - **Ridge regression:** Chap 3.4.1
  - **Lasso:** Chap 3.4.2
  - **LAR:** Chap 3.4.4
  - **Elastic Net:** Chap 4.2