Regularization

Big Data Lectures – Chapter 4

Lexin Li

Division of Biostatistics
University of California, Berkeley

Outline

- list of topics:
  - (partial) motivation: **really large p**
  - general framework
  - L_2 regularization
  - L_1 regularization
  - variants of L_1 regularization
  - additional remarks

General Framework

Introduction

- why regularization?
  - can effectively handle $n < p$, highly correlated predictors
  - can improve interpretation
  - can incorporate prior subject knowledge, e.g., structure, smoothness
  - can stabilize the estimates and improve the risk property

  — powerful, principled, simple

Introduction

- why regularization?
  - can effectively handle $n < p$, highly correlated predictors
  - can improve interpretation
  - can incorporate prior subject knowledge, e.g., structure, smoothness
  - can stabilize the estimates and improve the risk property

    — powerful, principled, simple

- a general regularization problem:

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda J(f) \right]$$

  - $L(y, f(x))$ is a point-wise loss function, $J(f)$ is a penalty functional, and $\mathcal{H}$ is a space of candidate functions

- examples:
  - ridge regression; lasso; adaptive lasso; group lasso; fused lasso; elastic net; SCAD; smoothing splines; manifold regularization; ...

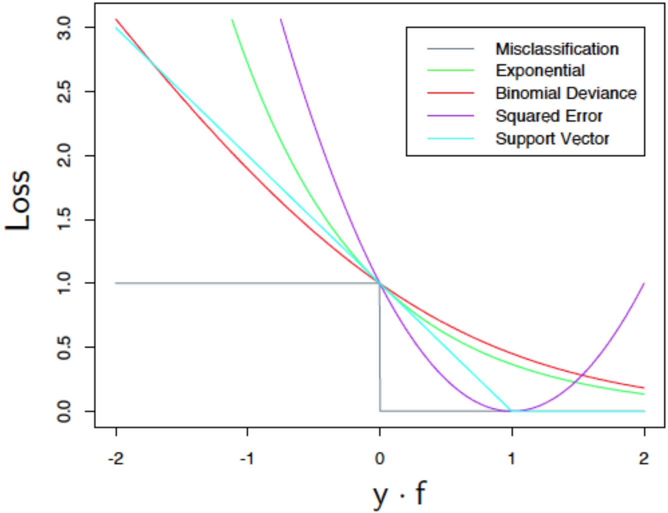
Loss function

- classification $y = \pm 1$:
  - misclassification loss: $I(\mathrm{sign}(f(\boldsymbol{x})) \neq y)$
  - exponential loss: $\exp(-yf)$
  - binomial deviance: $\log(1 + \exp(-2yf))$
  - squared error: $(y - f)^2 = (1 - yf)^2$
  - SVM hinge loss: $[1 - yf]_+$

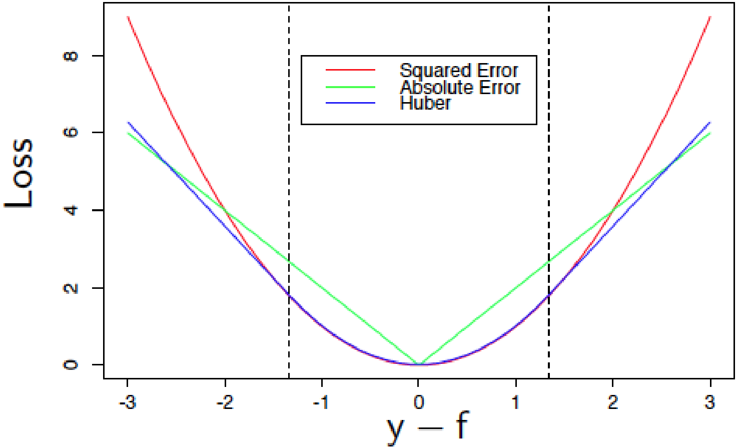
  - all are monotone decreasing functions of the **margin** $yf(\boldsymbol{x})$
  - the margin plays a role analogous to the **residual** $y - f(\boldsymbol{x})$ in regression
  - positive margin $y_i f(\boldsymbol{x}_i) > 0$ correctly classified; negative margin $y_i f(\boldsymbol{x}_i) < 0$ misclassified

- regression:
  - squared error: $(y - f)^2$
  - absolute error: $|y - f|$
  - Huber loss: $(y - f)^2$ for $|y - f| \leq \delta$ and $2\delta|y - f| - \delta^2$ o.w.

Loss function



Loss function



L_1 and L_2 Regularization

L_2 regularization

- **ridge regression**:

$$\min \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- equivalent form:

$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \le \tau$$

- matrix form: after centering all predictors and response

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

where $\|\boldsymbol{\beta}\|_2 = (\sum_{j=1}^{p} \beta_j^2)^{1/2}$, $\mathbf{y} \in \mathbb{R}^{n\times 1}, \mathbf{X} \in \mathbb{R}^{n\times p}$

L_2 regularization

- ridge (closed form) solution:

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda \boldsymbol{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

  - add a positive constant to the diagonal of $\mathbf{X}^\top\mathbf{X}$ before inversion
  - hence, can handle $n < p$ and/or highly correlated predictors

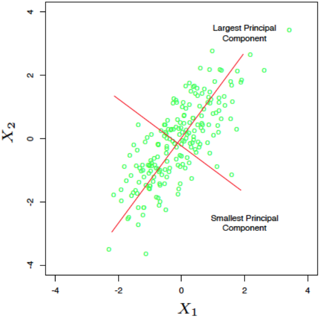
- let $\mathbf{X} = \boldsymbol{UDV}^\top$ singular value decomposition

$$\begin{aligned}
\mathbf{X}\hat{\beta}^{\text{ols}} &= \boldsymbol{UU}^\top\mathbf{y} = \sum_{j=1}^{p} \boldsymbol{u}_j\boldsymbol{u}_j^\top\mathbf{y} \\
\mathbf{X}\hat{\beta}^{\text{ridge}} &= \boldsymbol{UD}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{DU}^\top\mathbf{y} = \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda}\boldsymbol{u}_j^\top\mathbf{y}
\end{aligned}$$

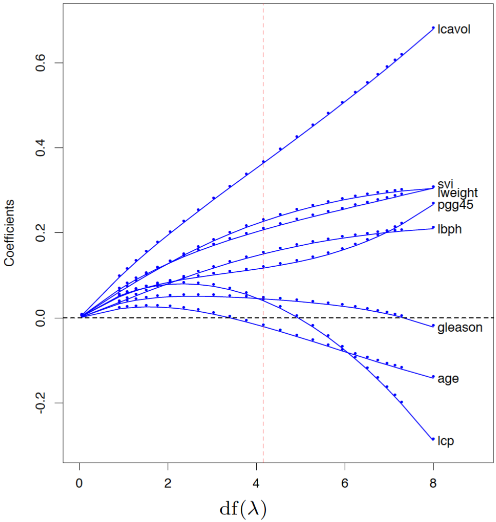
  - compute the coordinates of $\mathbf{y}$ wrt the orthonormal basis $\boldsymbol{U}$ then **shrink** these coordinates by the factors $d_j^2/(d_j^2 + \lambda)$

L_2 regularization

- ridge shrinkage:
  - a greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2
  - the small singular values d_j correspond to directions in the column space of $\boldsymbol{X}$ having small variance
- ridge regression shrinks those least important principal components directions the most
- the underlying assumption: the response tends to vary most in the directions of high variance of the predictors



L_2 regularization



L_1 regularization

- lasso:

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

- equivalent form:

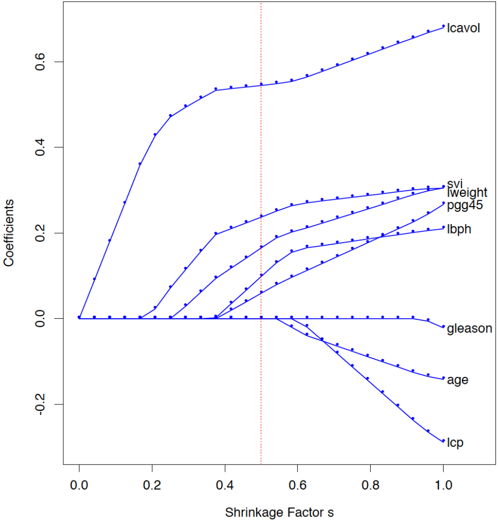
$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq \tau$$

- matrix form: after centering all predictors and response

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \, \|\boldsymbol{\beta}\|_1$$

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

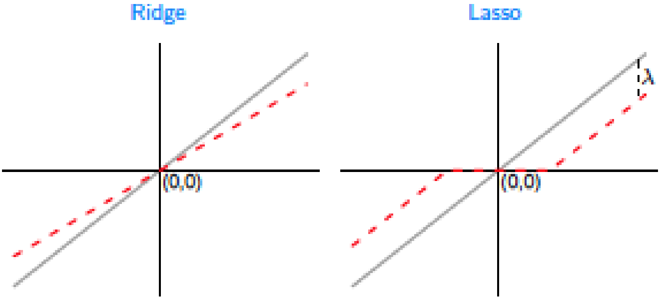
L_1 regularization



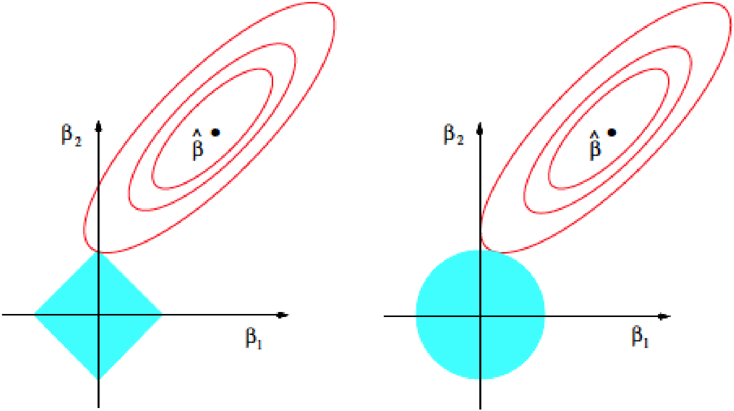
Comparison of L_1 and L_2 regularizations

- for orthonormal columns of $\mathbf{X}$:

$$\begin{array}{ll} \text{ridge} & \hat{\beta}_j/(1+\lambda) \\ \text{lasso} & \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+ \end{array}$$



Comparison of L_1 and L_2 regularizations



L_1 regularization

- **lasso shrinkage**:
  - lasso shrinks those coefficients whose magnitudes are less than a threshold λ, and shifts those by the amount of threshold whose magnitudes are greater
  - the underlying assumption: the response is only associated with a usually very small subset of the predictors — **sparsity principle**

- **piecewise linear solution path**: in general

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}(\lambda) &= \arg\min_{\beta} \left[R(\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta}) \right] \\
R(\boldsymbol{\beta}) &= \sum_{i=1}^{n} L(y_i, \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)
\end{aligned}
$$

then the solution path is piecewise linear if

  - R is quadratic or piecewise quadratic as a function of $\boldsymbol{\beta}$, and
  - J is piecewise linear in $\boldsymbol{\beta}$

L_1 regularization

- computation via least angle regression: entire solution path
  - a forward stepwise strategy
  - extremely efficient, requiring the same order of computation as that of a single least squares fit
- computation via coordinate descent: solution over a grid of λ values
  - rewrite the optimization problem to isolate β_j

$$\sum_{i=1}^{n} \left(y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k^{(t)}(\lambda) - x_{ij}\beta_j \right)^2 + \lambda \sum_{k \neq j} |\hat{\beta}_k^{(t)}(\lambda)| + \lambda |\beta_j|$$

  - the explicit soft-thresholding solution

$$\hat{\beta}_j^{(t+1)} = s\left(\sum_{i=1}^{n} x_{ij}[y_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k^{(t)}(\lambda)], \lambda \right)$$

    where $s(u, \lambda) = \mathrm{sign}(u)(|u| - \lambda)_+$
  - super fast since soft thresholding is fast and often yields 0

L_1 regularization – variants

- standardization:
  - without loss of generality, assume all predictors and response are centered, so no intercept term
  - predictors are often marginally standardized to have variance 1

L_1 regularization – variants

- standardization:
  - without loss of generality, assume all predictors and response are centered, so no intercept term
  - predictors are often marginally standardized to have variance 1

- **adaptive lasso**:

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1^w$$

  - $\|\boldsymbol{\beta}\|_1^w = \sum_{j=1}^p w_j |\beta_j| = \sum_{\mathbf{j=1}}^{\mathbf{p}} |\beta_{\mathbf{j}}| / |\beta_{\mathbf{j}}^{\text{ols}}|$
  - **intuition**: if $|\beta_j^{\text{ols}}|$ is large, give less penalization; if $|\beta_j^{\text{ols}}|$ is small, give more penalization
  - nice asymptotic properties / **oracle properties**: (a) can select all truly relevant predictors with probability going to 1; (b) can estimate the true linear model coefficients with same efficiency as if the subset were known a priori

L_1 regularization – variants

- **group lasso**:

$$\min \|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2$$

  - predictors are divided into G groups $\mathbf{X}_1, \ldots, \mathbf{X}_G$
  - $\|\boldsymbol{\beta}_g\|_2 = \sqrt{\boldsymbol{\beta}_g^\mathsf{T} \boldsymbol{\beta}_g}$ (Euclidean norm, not squared)
  - encourage the **entire group** of predictors to drop out; suitable for, e.g., dummy variables coding one discrete predictor

- **fussed lasso**:

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\Delta(\boldsymbol{\beta})\|_1$$

  - $\|\Delta(\boldsymbol{\beta})\|_1 = \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$
  - encourage sparsity of the coefficients and also sparsity of their **successive differences**; suitable for when features can be **ordered** in a meaningful way; e.g., a functional curve, mass spectroscopy

L_1 regularization – variants

- **elastic net**:

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$$

  - a combination of L_1 and L_2 regularization
  - encourage a **grouping** effect, where strongly correlated predictors tend to be in or out of the model together
  - particularly useful for $p >> n$

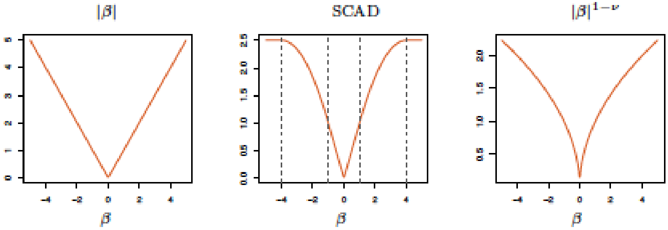
- **Dantzig selector**:

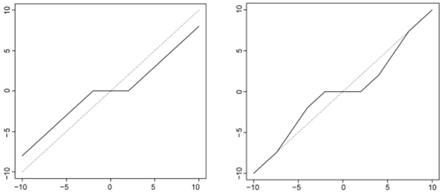
$$\min \|\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \quad \text{subject to } \|\boldsymbol{\beta}\|_1 \le \tau$$

  - $\|\cdot\|_\infty$ is the L_∞ norm, i.e., the maximum absolute value of the components of the vector
  - replace squared error loss in lasso by the maximum absolute value of its gradient
  - some interesting mathematics properties; useful for $p >> n$

L_1 regularization – variants

- SCAD:
  - replace the lasso penalty with the smoothly clipped absolute deviation penalty so that larger coefficients are shrunken less severely
  - nice oracle properties





More examples of regularization

- there are **many many many** regularization methods:
  different penalty functions for different purposes, for different models, e.g., mixed model, longitudinal model, survival model, ...
- **inference** in regularized estimation!
- what assumptions are imposed?
- any alternative methods?

Additional readings

- Hastie, H., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning*. Springer. Chapter 3