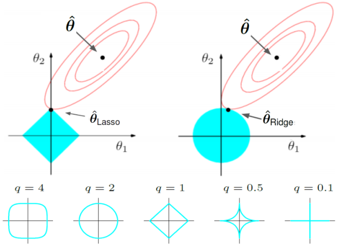
Introduction to Machine Learning

Regularization
Other Regularizers



Learning goals

- L1/L2 regularization induces bias
- Lq (quasi-)norm regularization
- L0 regularization
- SCAD and MCP

RIDGE AND LASSO ARE BIASED ESTIMATORS

Although ridge and lasso have many nice properties, they are biased estimators and the bias does not (necessarily) vanish as $n \to \infty$.

For example, in the orthonormal case ($\mathbf{X}^\top \mathbf{X} = \boldsymbol{I}$) the bias of the lasso is

$$\begin{cases} \mathbb{E}\left|\widehat{\theta}_j - \theta_j\right| = 0 & \text{if } \theta_j = 0 \\ \mathbb{E}\left|\widehat{\theta}_j - \theta_j\right| \approx \theta_j & \text{if } |\theta_j| \in [0, \lambda] \\ \mathbb{E}\left|\widehat{\theta}_j - \theta_j\right| \approx \lambda & \text{if } |\theta_j| > \lambda \end{cases}$$

To reduce the bias/shrinkage of regularized estimators various penalties were proposed, a few of which we briefly introduce now.

LQ REGULARIZATION ▸ Fu and Knight 2000

Besides *L*1/*L*2 we could use any *Lq* (quasi-)norm penalty $\lambda\|\boldsymbol{\theta}\|_q^q$

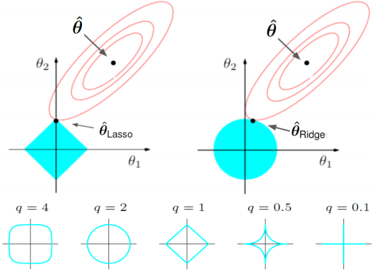
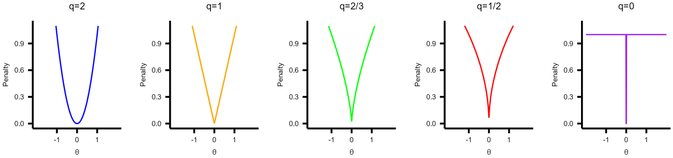


Figure: *Top:* loss contours and *L*1/*L*2 constraints. *Bottom:* Constraints for *Lq* norms $\sum_j |\theta_j|^q$.

- For $q < 1$ penalty becomes non-convex but for $q > 1$ no sparsity is achieved
- Non-convex *Lq* has nice properties like **oracle property** ▸ Zou and Hastie 2005: consistent (+ asy. unbiased) param estimation and var selection
- Downside: non-convexity makes optimization even harder than *L*1 (no unique global minimum but multiple local minima)

L0 REGULARIZATION

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 := \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \sum_j |\theta_j|^0.$$



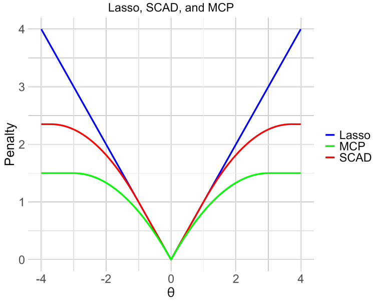
- L0 "norm" simply counts the nr of non-zero params
- Induces sparsity more aggressively than $L1$, but does not shrink
- AIC and BIC are special cases of $L0$
- $L0$-regularized risk is not continuous or convex
- NP-hard to optimize; for smaller n and p somewhat tractable, efficient approximations are still current research

SCAD ▸ Fan and Li 2001

Smoothly Clipped Absolute Deviations:
non-convex, $\gamma > 2$ controlls how fast penalty “tapers off”

$$\text{SCAD}(\theta \mid \lambda, \gamma) = \begin{cases} \lambda|\theta| & \text{if } |\theta| \leq \lambda \\ \frac{2\gamma\lambda|\theta| - \theta^2 - \lambda^2}{2(\gamma - 1)} & \text{if } \lambda < |\theta| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2} & \text{if } |\theta| \geq \gamma\lambda \end{cases}$$

- Lasso, quadratic, then const
- Smooth
- Contrary to lasso/ridge, SCAD continuously relaxes penalization rate as $|\theta|$ increases above λ

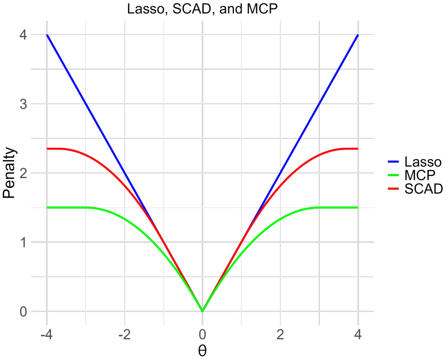


MCP ▸ Zhang 2010

Minimax Concave Penalty:
also non-convex; similar idea as SCAD with $\gamma > 1$

$$MCP(\theta|\lambda,\gamma) = \begin{cases} \lambda|\theta| - \frac{\theta^2}{2\gamma}, & \text{if } |\theta| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\theta| > \gamma\lambda \end{cases}$$

- As with SCAD, MCP starts by applying same penalization rate as lasso, then smoothly reduces rate to zero as $|\theta| \uparrow$
- Different from SCAD, MCP immediately starts relaxing the penalization rate, while for SCAD rate remains flat until $|\theta| > \lambda$
- Both SCAD and MCP possess oracle property: they can consistently select true model as $n \to \infty$ while lasso may fail

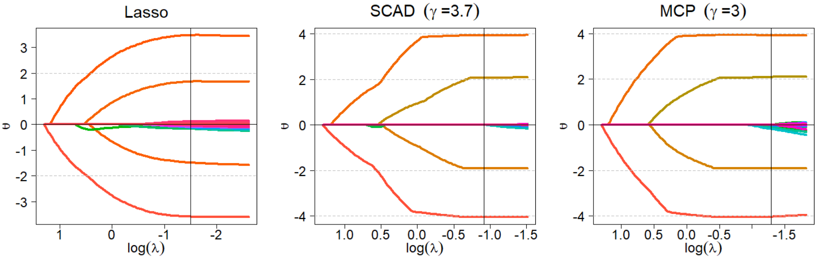


EXAMPLE: COMPARING REGULARIZERS

Let’s compare coeff paths for lasso, SCAD, and MCP.

We simulate $n = 100$ samples from the following DGP:

$$y = \mathbf{x}^\top \boldsymbol{\theta} + \varepsilon\,, \quad \boldsymbol{\theta} = (4, -4, -2, 2, 0, \ldots, 0)^\top \in \mathbb{R}^{1500}, \quad x_j, \varepsilon \sim \mathcal{N}(0, 1)$$



Vertical lines mark optimal λ from 10CV.

Conclusion: Lasso underestimates true coeffs while SCAD/MCP achieve unbiased estimation and better variable selection