Adaptive lasso, MCP, and SCAD

Patrick Breheny

February 28

Introduction

- Although the lasso has many excellent properties, it is a biased estimator and this bias does necessarily not go away as $n \to \infty$
- For example, in the orthonormal case,

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

- Thus, the bias of the lasso estimate for a truly nonzero variable is about λ for large regression coefficients coefficients.

Adaptive lasso: Motivation

- Given that the bias of the estimate is determined by λ, one approach to reducing the bias of the lasso is to use the weighted penalty approach we saw last time: $\lambda_j = w_j \lambda$
- If one was able to choose the weights such that the variables with large coefficients had smaller weights, then we could reduce the estimation bias of the lasso while retaining its sparsity property
- Indeed, by more accurately estimating $\boldsymbol{\beta}$, one would even be able to improve on the variable selection accuracy of the lasso

Adaptive lasso: Motivation (cont'd)

- All of this may seem circular in the sense that if we already knew which regression coefficients were large and which were small, we wouldn't need to be carrying out a regression analysis in the first place
- However, it turns out that the choice of $\mathbf{w}$ does not need to be terribly precise in order to realize benefits from this approach
- In practice, one can obtain reasonable values for $\mathbf{w}$ from any consistent initial estimator of $\boldsymbol{\beta}$

Adaptive lasso

- Let $\widetilde{\beta}$ denote the initial estimate (from, say, OLS or the lasso)
- The *adaptive lasso* estimate $\widehat{\beta}$ is then defined as the argument minimizing the following objective function:

$$Q(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}, \mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_j w_j \left|\beta_j\right|,$$

  where $w_j = |\widetilde{\beta}_j|^{-1}$
- Note that this weighting scheme assigns smaller weights to larger regression coefficients, and that if the initial estimate $\widetilde{\beta}_j = 0$, we have $w_j = \infty$, so $\widehat{\beta}_j = 0$.

Two-stage vs. pathwise approaches

- In the above approach, known as a *two-stage approach*, a single initial estimate $\widetilde{\boldsymbol{\beta}}$ is made, which in turn produces a single set of weights $\mathbf{w}$, which are held constant across all values of λ
- An alternative approach, known as a *pathwise approach* is to let the weights change with λ:

$$w_j(\lambda) = w(\widetilde{\beta}_j(\lambda))$$

- Here, the initial estimate is typically a lasso estimator, so that λ has the same meaning for the initial estimator as it does for the re-weighted, or adaptive, estimator

Alternative weighting strategies

- There are many possibilities besides $w_j = |\widetilde{\beta}_j|^{-1}$ for choosing weights based on initial estimates
- Really, any nonincreasing function $w(\beta)$ would be a reasonable way to choose weights, and could be used in either a two-stage or adaptive approach, although the resulting estimators may be quite different
- For example, one might allow $w_j = |\widetilde{\beta}_j|^{-\gamma}$ or $w_j = 1\{|\widetilde{\beta}_j| \leq \tau\}$

Hybrid and relaxed lasso approaches

- A more extreme weighting scheme is

$$w_j = \begin{cases} 0 & \text{if } \widetilde{\beta}_j \neq 0, \\ \infty & \text{if } \widetilde{\beta}_j = 0 \end{cases}$$

- When applied in a two-stage fashion, this approach is known as the *lasso-OLS hybrid* estimator (i.e., we use the lasso for variable selection and OLS for estimation)
- When the approach is applied in a pathwise fashion, it is known as the *relaxed lasso*

Single-stage approaches to bias reduction

- The adaptive lasso consists of a two-stage approach involving an initial estimator to reduce bias for large regression coefficients
- An alternative single-stage approach is to use a penalty that tapers off as β becomes larger in absolute value
- Unlike the absolute value penalty employed by the lasso, a tapering penalty cannot be convex

Folded concave penalties

- Rather, the penalty function $P(\beta|\lambda)$ will be concave with respect to $|\beta|$
- Such functions are often referred to as *folded concave penalties*, to clarify that while $P(\cdot)$ itself is neither convex nor concave, it is concave on both the positive and negative halves of the real line, and also symmetric (or folded) due to its dependence on the absolute value

Objective function for folded concave penalties

- Consider the objective function

$$Q(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p(|\beta_j|\,; \lambda, \gamma),$$

  where $p(\theta|\lambda, \gamma)$ is a concave function defined on $[0, \infty)$
- Unlike the lasso, many concave penalties depend on λ in a non-multiplicative way, so $p(\theta|\lambda) \neq \lambda p(\theta)$
- Furthermore, they typically involve a tuning parameter γ that controls the concavity of the penalty (i.e., how rapidly the penalty tapers off)

SCAD

- A variety of nonconvex penalties have been proposed; one of the earliest and most influential was the smoothly clipped absolute deviations (SCAD) penalty:

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

  for $\gamma > 2$
- Note that SCAD coincides with the lasso until $|\beta| = \lambda$, then smoothly transitions to a quadratic function until $|\beta| = \gamma\lambda$, after which it remains constant for all $|\beta| > \gamma\lambda$

SCAD: Derivative

- It is typically more instructive to consider a penalty's derivative – i.e., the contribution made by the penalty to the penalized estimating equations (KKT conditions)
- The derivative of the SCAD penalty is

$$\dot{p}(\theta|\lambda,\gamma) = \begin{cases} \lambda & \text{if } \theta \leq \lambda, \\ \frac{\gamma\lambda-\theta}{\gamma-1} & \text{if } \lambda < \theta < \gamma\lambda, \\ 0 & \text{if } \theta \geq \gamma\lambda \end{cases}$$

- The SCAD penalty retains the penalization rate (and bias) of the lasso for small coefficients, but continuously relaxes the rate of penalization as the absolute value of the coefficient increases

MCP

The idea behind the minimax concave penalty (MCP) is very similar:

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

for $\gamma > 1$

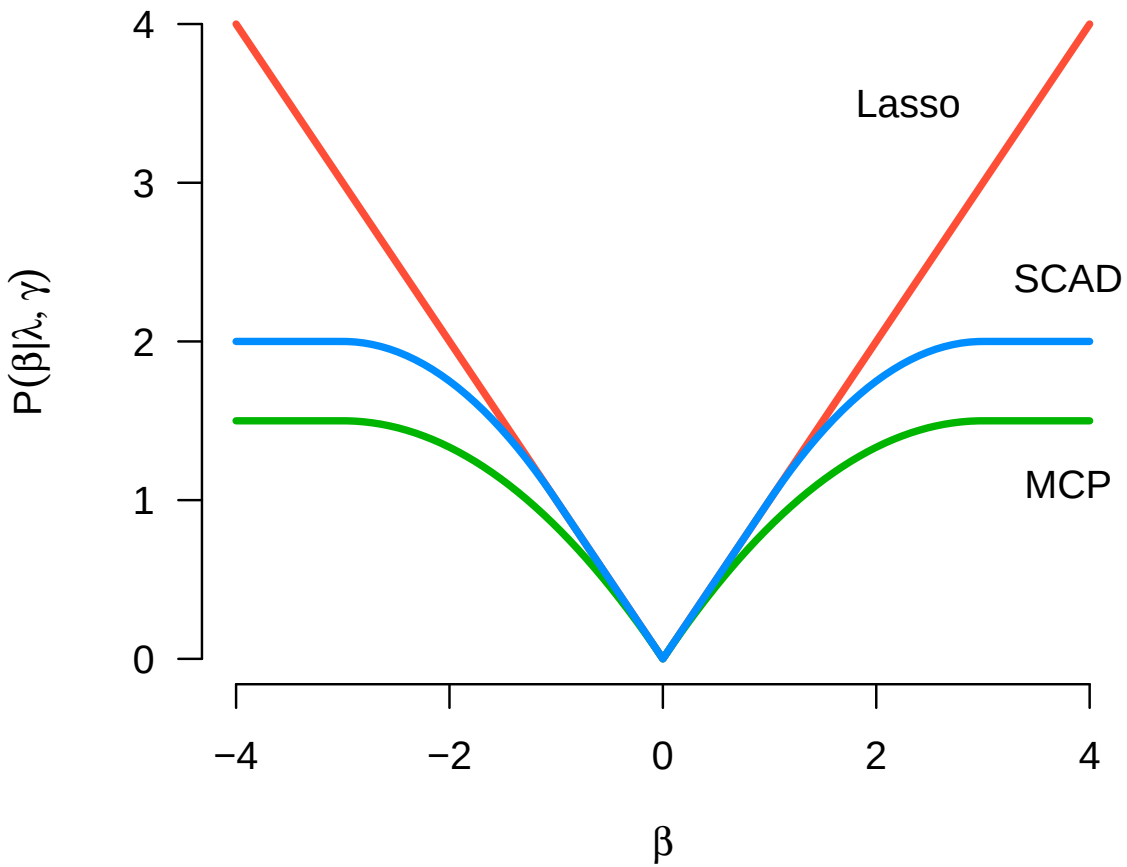
MCP: Derivative

- Its derivative is

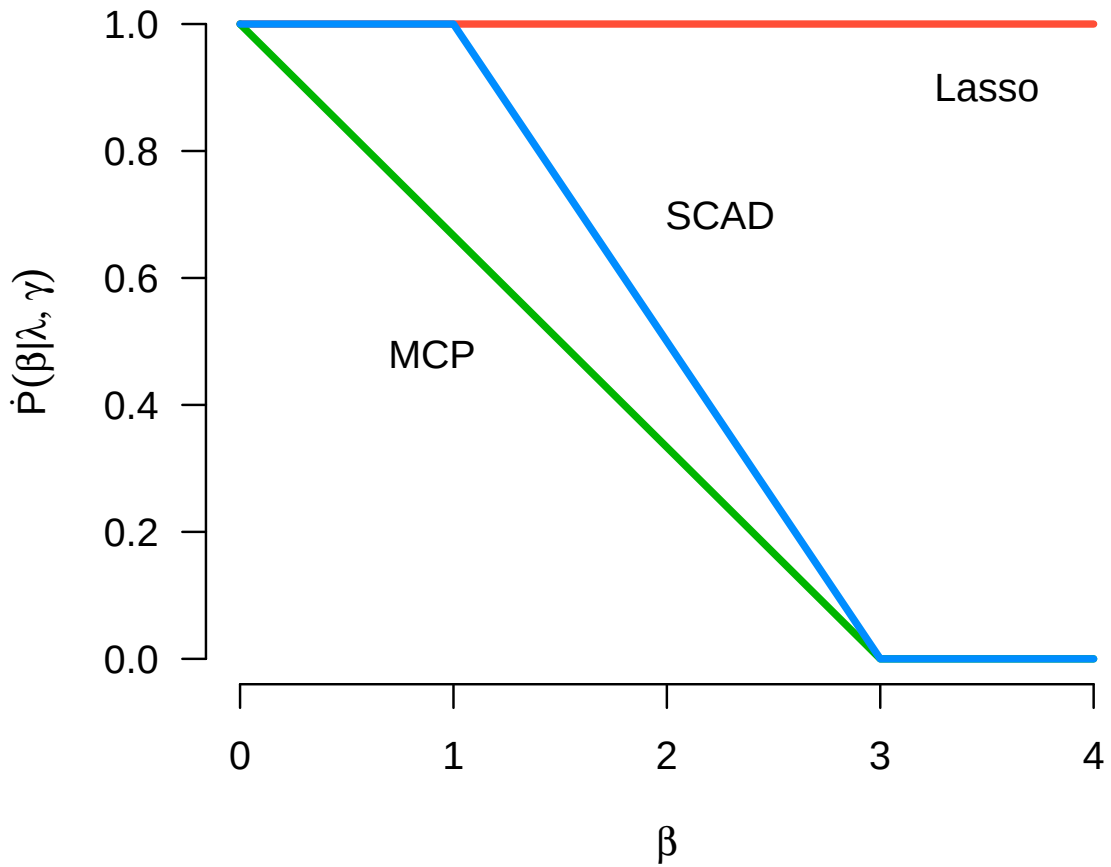
$$\dot{p}(\theta|\lambda,\gamma) = \begin{cases} \lambda - \frac{\theta}{\gamma}, & \text{if } \theta \leq \gamma\lambda, \\ 0, & \text{if } \theta > \gamma\lambda. \end{cases}$$

- As with SCAD, MCP starts out by applying the same rate of penalization as the lasso, then smoothly relaxes the rate down to zero as the absolute value of the coefficient increases
- In comparison to SCAD, however, the MCP relaxes the penalization rate immediately while with SCAD the rate remains flat for a while before decreasing

SCAD and MCP: Illustration



SCAD and MCP: Illustration (cont'd)



Remarks

- These plots illustrate the sense in which the MCP is *minimax concave*
- Out of all functions p continuously differentiable on $(0, \infty)$ that satisfy $\dot{p}(0 + |\lambda) = \lambda$ and $\dot{p}(\theta|\lambda) = 0$ for all $t \geq \gamma\lambda$, the MCP minimizes the maximum concavity

$$\kappa = \sup_{0<\theta_1<\theta_2} \frac{\dot{p}(\theta_1|\lambda) - \dot{p}(\theta_2|\lambda)}{\theta_2 - \theta_1}.$$

- As the figure shows, the derivatives of SCAD and MCP are equal at 0 and again at $\gamma\lambda$, but MCP has a concavity of $\kappa = 1/\gamma = 1/3$ over this region while SCAD has a maximum concavity of $\kappa = 1/(\gamma - 1) = 1/2$

MCP & firm thresholding

- As with the lasso, MCP and SCAD have closed-form solutions in the orthonormal case that provide insight into how the methods work
- For MCP, the univariate solution is known as the *firm thresholding operator*:

$$F(z|\lambda,\gamma) = \begin{cases} \frac{\gamma}{\gamma-1} S(z|\lambda) & \text{if } |z| \leq \gamma\lambda, \\ z & \text{if } |z| > \gamma\lambda \end{cases},$$

where $z = \mathbf{x}^\top \mathbf{y}/n$ denotes the unpenalized (OLS) solution

Remarks: Firm thresholding

- As $\gamma \to \infty$, the firm thresholding operator becomes equivalent to the soft thresholding operator: $F(z|\lambda, \gamma) \to S(z|\lambda)$
- As $\gamma \to 1$, it becomes equivalent to hard thresholding
- Thus, as γ changes, the solution bridges the gap between soft and hard thresholding; hence the name "firm thresholding"

SCAD thresholding

- The SCAD solution is similar, although somewhat more complicated
- The SCAD thresholding operator is

$$T_{\text{SCAD}}(z|\lambda,\gamma) = \begin{cases} S(z|\lambda), & \text{if } |z| \leq 2\lambda, \\ \frac{\gamma-1}{\gamma-2} S(z|\frac{\gamma\lambda}{\gamma-1}), & \text{if } 2\lambda < |z| \leq \gamma\lambda, \\ z, & \text{if } |z| > \gamma\lambda \end{cases}$$

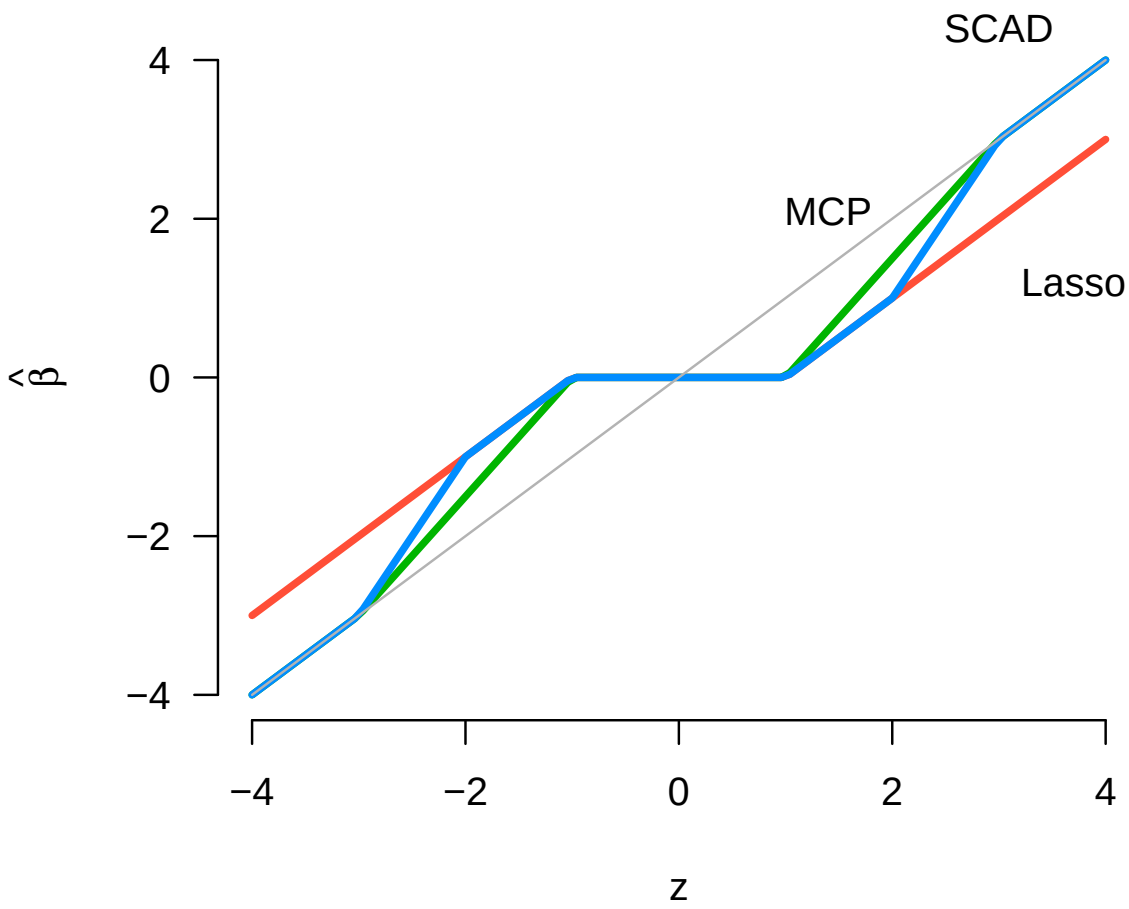
Remarks: SCAD thresholding

- As with MCP, $T_{\text{SCAD}}(\cdot|\lambda,\gamma) \to S(\cdot|\lambda)$ as $\gamma \to \infty$
- However, as $\gamma \to 2$, $T_{\text{SCAD}}(\cdot|\lambda,\gamma)$ does not converge to hard thresholding; instead, it converges to

$$\begin{cases} S(z;\lambda), & \text{if } |z| \le 2\lambda, \\ z, & \text{if } |z| > 2\lambda \end{cases}$$

- In other words, both T_{SCAD} and F converge to discontinuous functions as γ approaches its minimum value: for the firm thresholding operator F, the solution jumps from 0 to λ as z exceeds λ, while for the SCAD thresholding operator T_{SCAD}, the solution jumps from λ to 2λ as z exceeds 2λ

SCAD and MCP thresholding



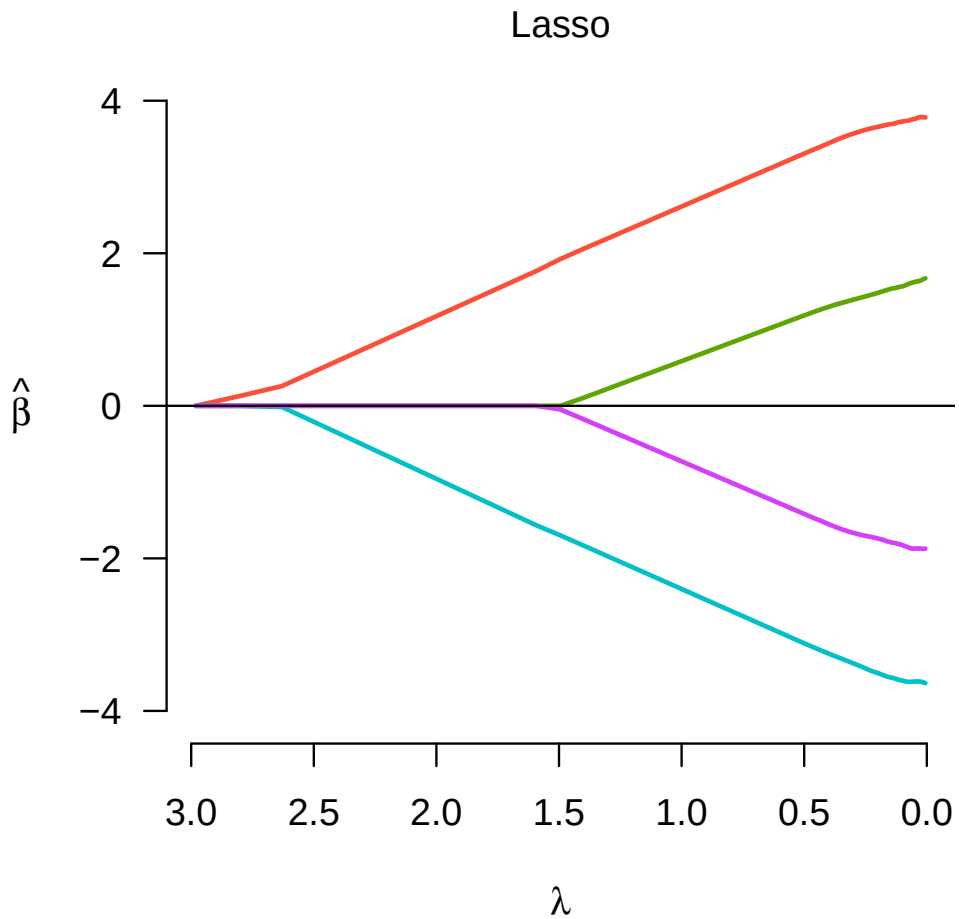
Solution paths

- To get a sense of how the MCP, SCAD, and adaptive lasso estimates compare to those of the regular lasso, we consider here the solution paths for the four penalties fit to the same data
- We generate data from the linear regression model

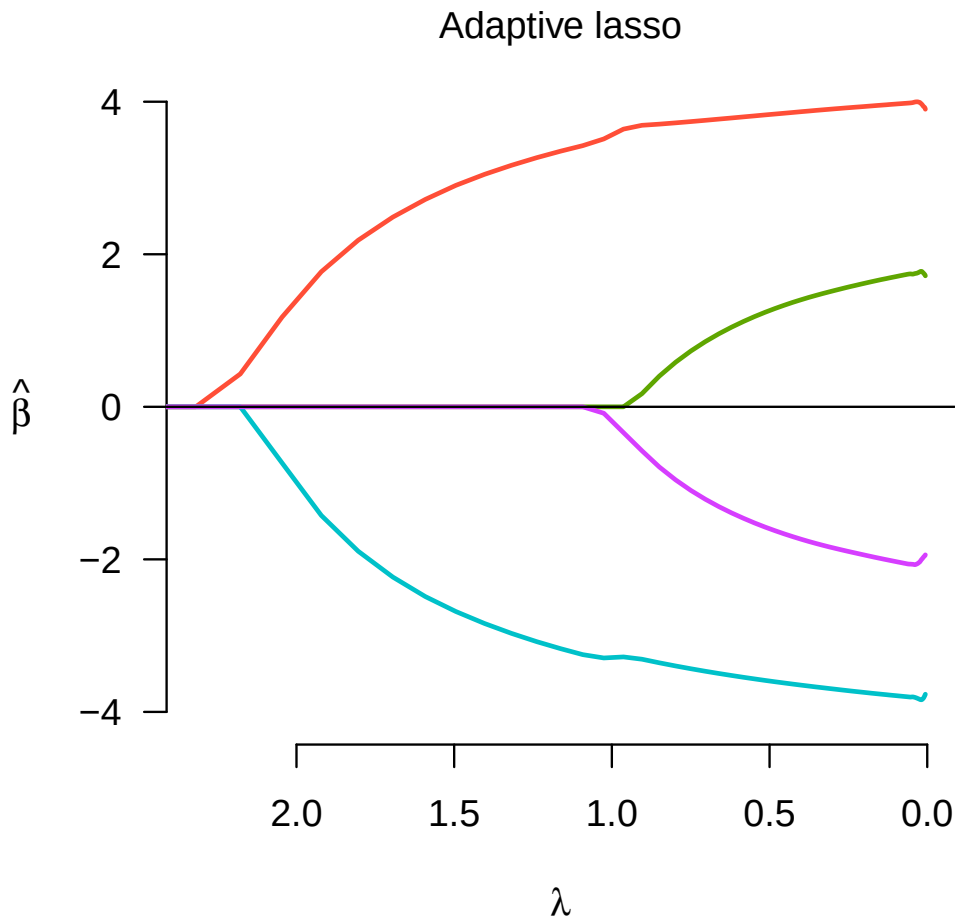
$$y_i = \sum_{j=1}^{1000} x_{ij}\beta_j + \epsilon_i, i = 1, \ldots, 200,$$

where $(\beta_1, \ldots, \beta_4) = (4, 2, -4, -2)$ and the remaining coefficients are zero

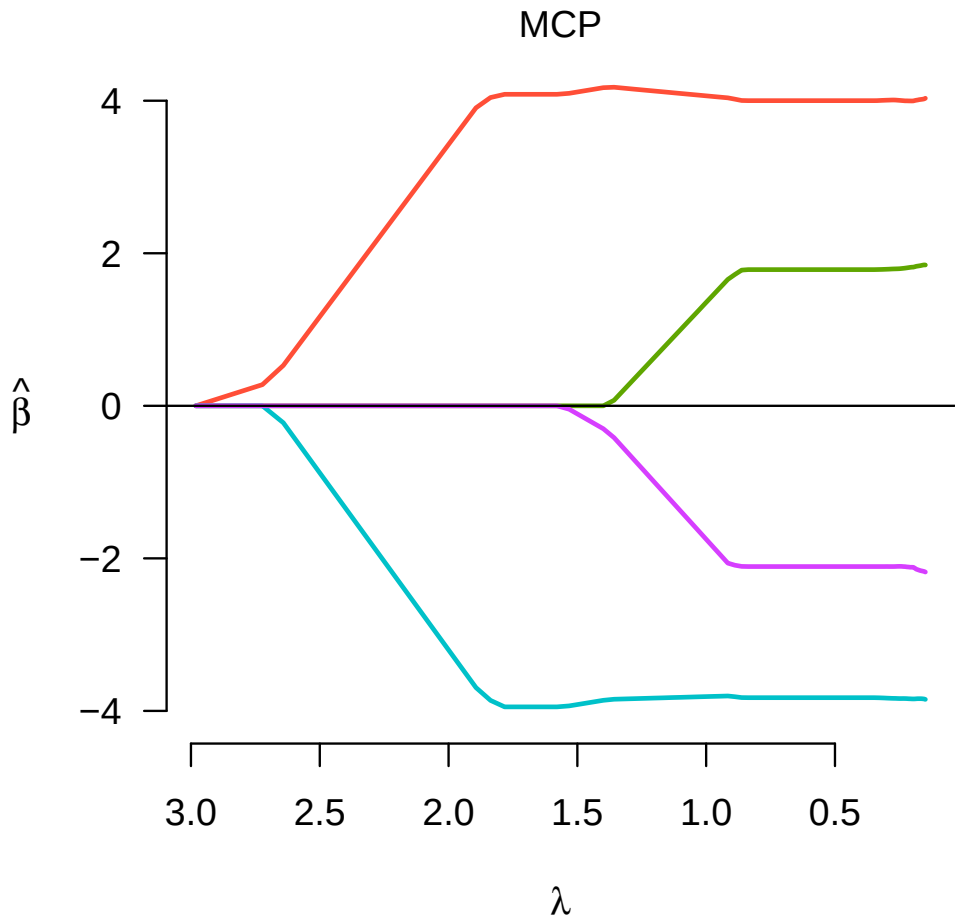
Solution path: Lasso



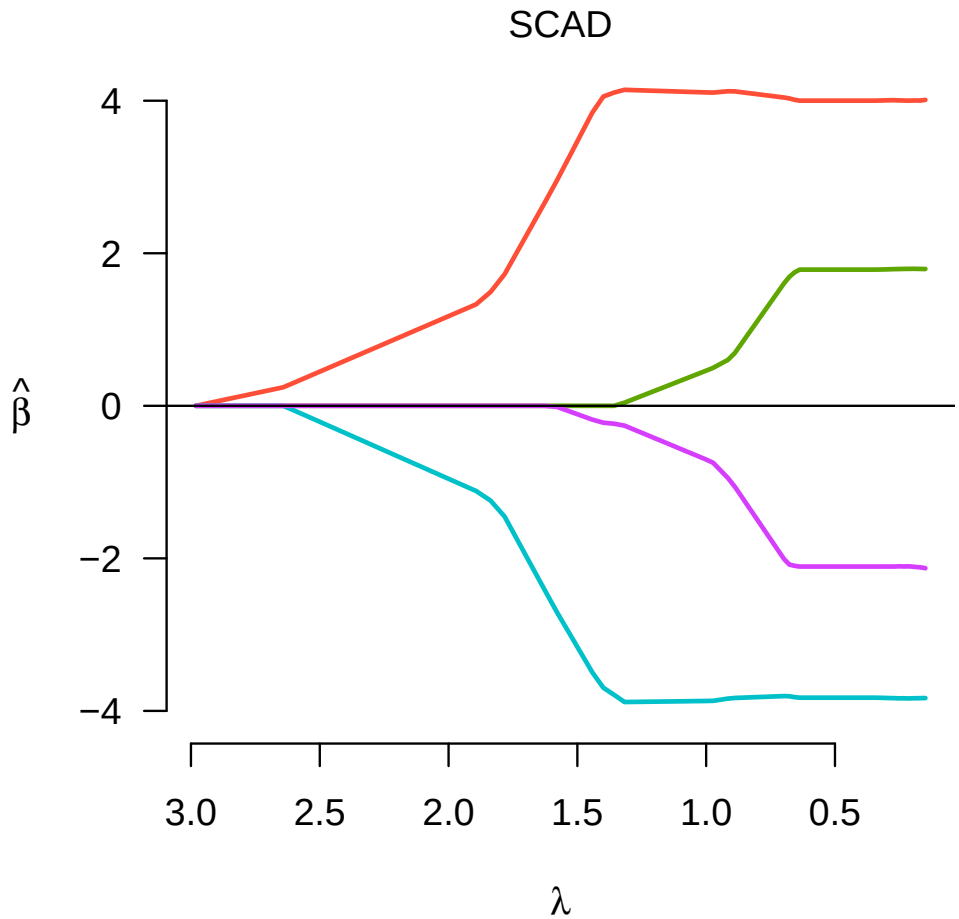
Solution path: Adaptive lasso (pathwise)



Solution path: MCP ($\gamma = 3$)



Solution path: SCAD ($\gamma = 4$)



Remarks

- The primary way in which adaptive lasso, SCAD, and MCP differ from the lasso is that they allow the estimated coefficients to reach large values more quickly than the lasso
- In other words, although the methods all shrink most of the coefficients towards zero, MCP, SCAD, and the adaptive lasso apply less shrinkage to the nonzero coefficients; this is what we refer to in the book as *bias reduction*

Remarks (cont'd)

- In this example, one can clearly see the piecewise components of MCP and SCAD
- In particular, it is worth noting that both MCP and SCAD possess an interval of λ values over which all the estimates are flat – over this region, the estimates are the same as those of ordinary least squares regression, but with only the four variables with nonzero effects included
- These estimates are referred to as the *oracle* estimates

The role of γ in SCAD and MCP

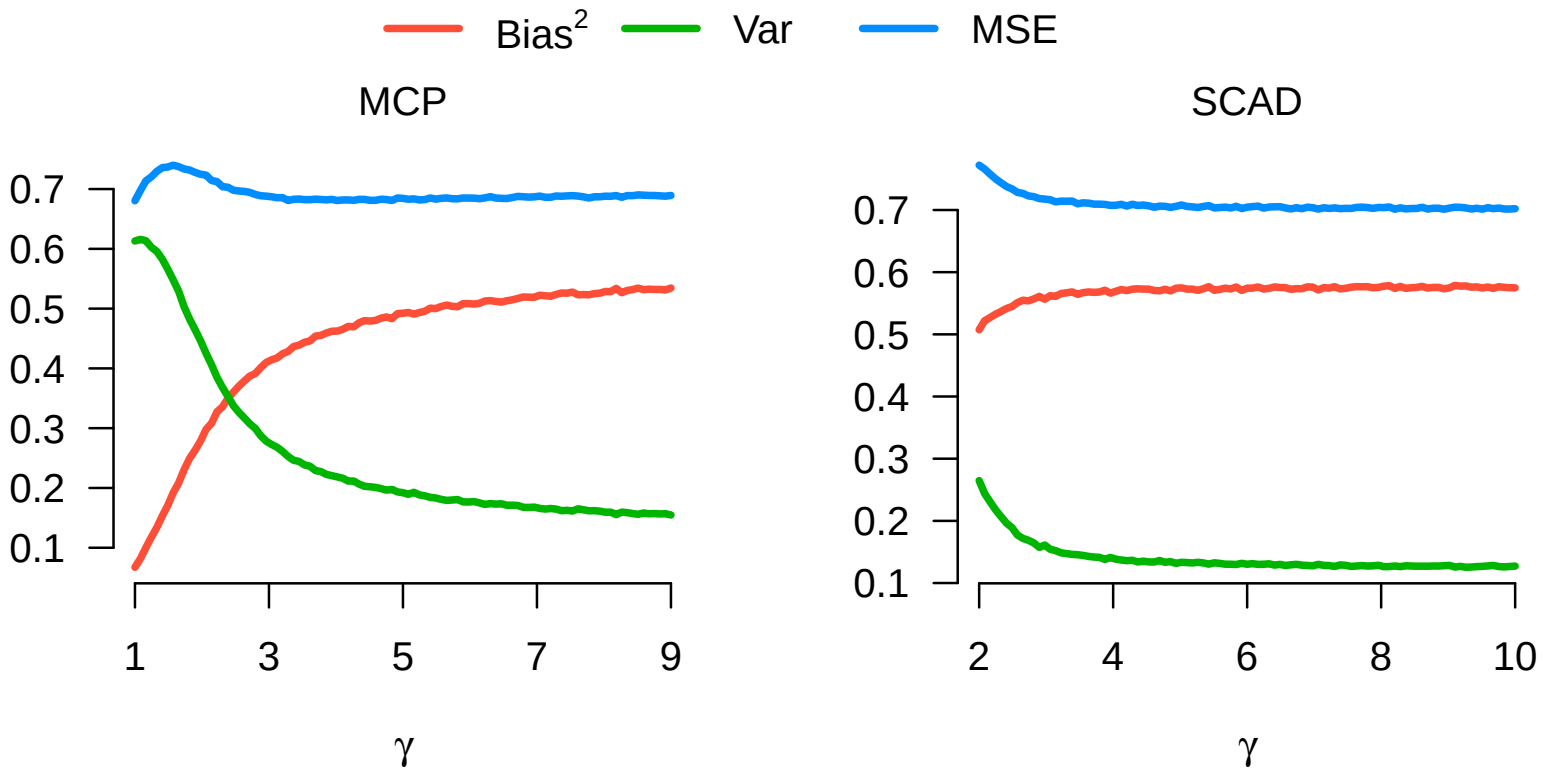
- As discussed previously, the tuning parameter γ for the SCAD and MCP estimates controls how fast the penalization rate goes to zero
- This, in turn, affects the bias of the estimates as well as the stability of the estimate in the sense that as the penalty becomes more concave, there is a greater chance for multiple local minima to exist
- As $\gamma \to \infty$, both the MCP and SCAD penalties converge to the ℓ_1 penalty
- As γ approaches its minimum value, bias is minimized, but both estimates become unstable

γ and the bias-variance tradeoff

- "Stability" here refers to the optimization sense that an objective function with a single, well-defined minimum is stable while optimization problems with multiple local minima tend are unstable
- However, the same remarks apply with respect to the statistical properties of the estimators, in the sense that a more highly variable estimator is less stable
- For SCAD and MCP, lower values of γ produce more highly variable, but less biased, estimates

Bias-variance tradeoff: Illustration

For $\sigma^2 = 6$, $\lambda = 1$, $n = 10$, and there is a single feature with $\beta = 1$:



Effect of γ on solution paths

Same data as the earlier path example (MCP paths shown)