On weak and strong oracle properties

Yongdai Kim

Seoul National University, Korea

Introduction

- Variable selection is important for high dimensional models.
- Traditional approaches such as stepwise selections are
  - computationally intensive
  - hard to draw sampling properties
  - unstable
- Alternative approach is sparse, which means some coefficients are exactly zero, penalized approaches including
  - bridge regression (Frank and Friedman, 1993)
  - Lasso (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1997)
  - SCAD (Smoothly Clipped Absolute Deviation, Fan and Li 2001)

Introduction

Sparse penalized approaches

- Data: $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ where $y_i \in R$ and $\mathbf{x}_i \in R^p$.
- General form of sparse penalized estimators

$$\hat{\beta} = \text{argmin}_{\beta} C_n(\beta),$$

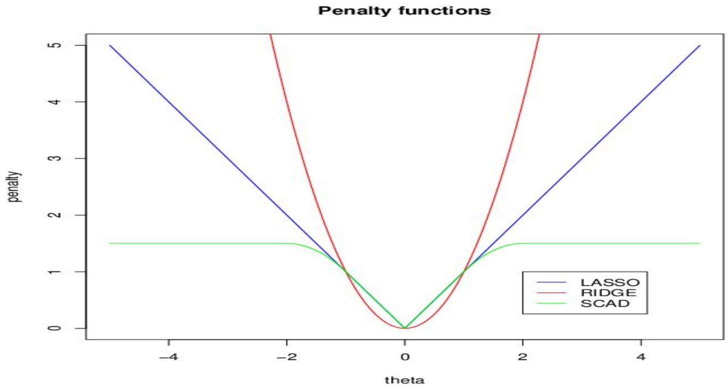
  where

$$C_n(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \beta)^2 / 2n + \lambda \sum_{j=1}^{p} J_\lambda(|\beta_j|)$$

  for some penalty function J.
- Various penalty functions
  - Bridge: $J(\beta) = \beta^q, q > 0$
  - Lasso: $J(\beta) = \beta$
  - SCAD:

$$\begin{aligned} J_\lambda(\beta) &= \lambda \beta I(0 \leq \beta < \lambda) \\ &+ \left(\frac{a\lambda(\beta - \lambda) - (\beta^2 - \lambda^2)/2}{(a-1)} + \lambda^2 \right) I(\lambda \leq \beta \leq a\lambda) \\ &+ \left(\frac{(a-1)\lambda^2}{2} + \lambda^2 \right) I(\beta \geq a\lambda). \end{aligned}$$

Introduction



Introduction

- The penalties are nondifferentiable at 0, which is necessary for sparsity.
- The Lasso is convex while the bridge and SCAD penalties are nonconvex. Nonconvexity is necessary for unbiasedness of estimated coefficients.

Introduction

Theme of the talk

- The theme of the talk is about the **oracle property** of noncovex penalized estimators.
- The oracle property means that the penalized estimator is asymptotically equivalent to the oracle estimator that is the ideal estimator obtained only with signal variables without penalization.
- Many noncovex penalties such as the bridge and SCAD penalties possess the oracle property.
- In practice, however, only a local minimum (of the penalized sum of squared residuals) is given, and it is extremely difficult (almost impossible) to check if a given local minimum is (asymptotically) the oracle estimator.
- In this sense, the oracle property of a nonconvex penalty is practically meaningful only when reasonable local minima are asymptotically equivalent to the oracle estimator.

Introduction

Theme of the talk

- The objectives of the talk are
  - to demonstrate that there are reasonable but bad local minima for some nonconvex penalties that have the oracle property;
  - to give necessary conditions to ensure the uniqueness of local minima;
  - to show that certain nonconvex penalties have the unique local minimum.

Three modes of the oracle property

Notations

- Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be n many response-covariates pairs where $y_i \in R$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) \in R^p$.
- Let $\mathbf{y} = (y_1, \ldots, y_n)'$ and $X^j = (x_{1j}, \ldots, x_{nj})'$
- For $\pi \subset \{1, \ldots, p\}$, let $\mathbf{X}_\pi = (X^j, j \in \pi)$ and $\beta_\pi = (\beta_j, j \in \pi)$.
- Let β^* be the true regression coefficient and let $\mathscr{A} = \{j : \beta_j^* \neq 0\}$.
- Let $\hat{\beta}^o$ be the oracle estimator defined as

$$\hat{\beta}^o = \mathrm{argmin}_{\beta, \beta_j = 0, j \in \mathscr{A}^c} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \beta)^2.$$

Three modes of the oracle property

Definition of the standard oracle property

- $\hat{\beta}$ is said to possess the oracle property if there exists a sequence of λ_n such that with $\lambda = \lambda_n$

$$\Pr(\hat{\beta} = \hat{\beta}^o) \to 1.$$

- Kim et al. (2008) for SCAD and Huang et al. (2008) for bridge when $p < n$.
- A slightly weaker definition is
  - (*) $\Pr(\hat{\mathscr{A}} = \mathscr{A}^*) \to 1$, where $\hat{\mathscr{A}} = \{k : \hat{\beta}_j \neq 0\}$,
  - (**)

$$\sqrt{n}(\hat{\beta}_{\mathscr{A}^*} - \beta^*_{\mathscr{A}^*}) \overset{d}{\approx} \sqrt{n}(\hat{\beta}^o_{\mathscr{A}^*} - \beta^*_{\mathscr{A}^*}).$$

Three modes of the oracle property

Definition of weak oracle property

- Let $\mathscr{L}(\lambda)$ be the set of all local minima of the penalized sum of squared residuals.
- The penalty is said to have the weakly oracle property if there exists a sequence of λ_n such that

$$\Pr(\hat{\beta}^o \in \mathscr{L}(\lambda_n)) \to 1.$$

- Fan and Li (2001), Fan and Peng (2004), Kim et al. (2008).
- For the SCAD penalty, the weak oracle property holds for $p > n$.
- A slightly weaker version is that there exists a local minimum satisfying (*) and (**).

Three modes of the oracle property

Definition of the strong oracle property

- Let $\mathscr{L}(\lambda)$ be the set of all local minima of the penalized sum of squared residuals.
- The panalty is said to have the strongly oracle property if there exists a sequence of λ_n such that

$$\Pr(\mathscr{L}(\lambda_n) = \{\hat{\beta}^o\}) \to 1.$$

- That is, the oracle estimator is the unique local minimum.
- A slightly weaker version is that there exists a **unique** local minimum satisfying (*) and (**).

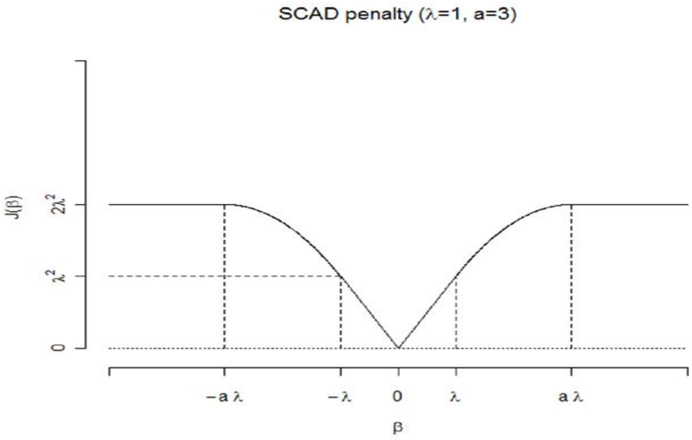
Sufficient conditions for the weak oracle property

Class of penalty functions

- Let $\nabla(\beta) = dJ_\lambda(\beta)/d\beta$.
- Let $\phi = \nabla(0+)$
- There exist positive constants γ and η such that $\nabla(\beta) \leq \gamma$ for $\beta > \eta$.

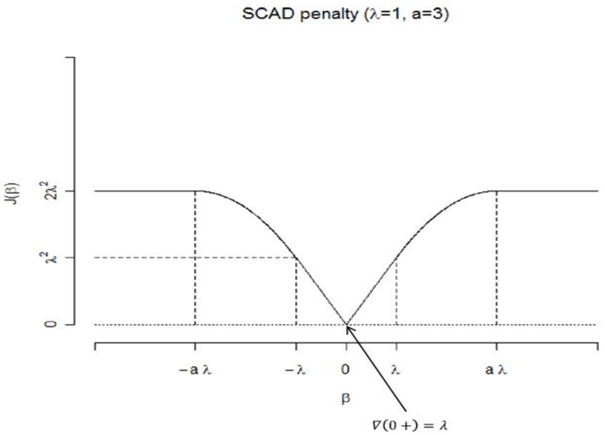
Sufficient conditions for the weak oracle property

Example: SCAD



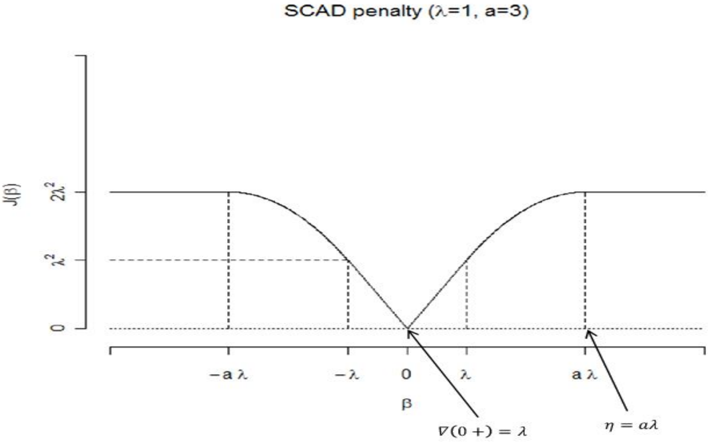
Sufficient conditions for the weak oracle property

Example: SCAD



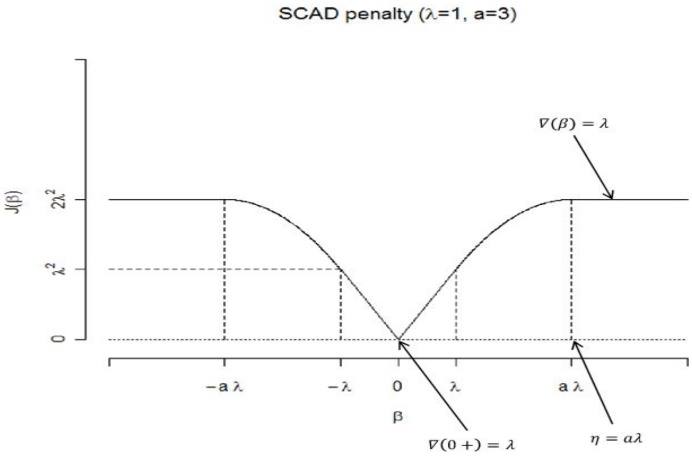
Sufficient conditions for the weak oracle property

Example: SCAD



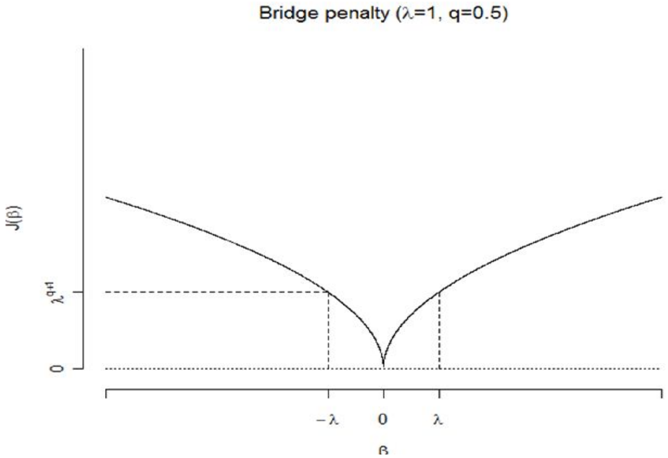
Sufficient conditions for the weak oracle property

Example: SCAD



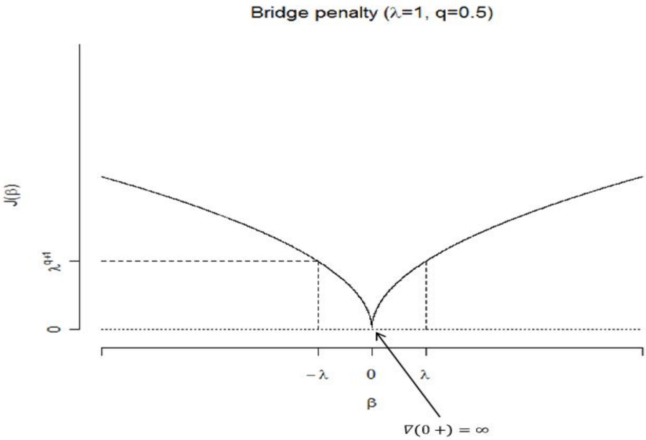
Sufficient conditions for the weak oracle property

Example: Bridge



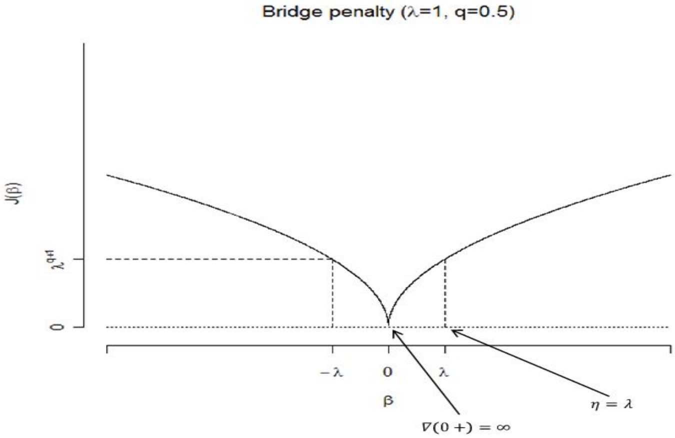
Sufficient conditions for the weak oracle property

Example: Bridge



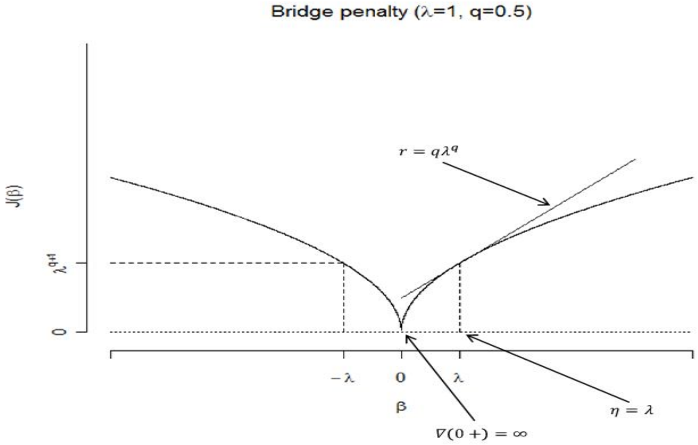
Sufficient conditions for the weak oracle property

Example: Bridge



Sufficient conditions for the weak oracle property

Example: Bridge



Sufficient conditions for the weak oracle property

Necessary conditions of a local minimum

(1) For $j \in \hat{\mathscr{A}}$

$$X^{j'}(\mathbf{y} - \mathbf{X}\hat{\beta})/n = \mathrm{sign}(\hat{\beta}_j)\nabla(|\hat{\beta}_j|),$$

(2) For $j \in \hat{\mathscr{A}}^c$,

$$\left| X^{j'}(\mathbf{y} - \mathbf{X}\hat{\beta})/n \right| \leq \phi.$$

Sufficient conditions for the weak oracle property

Sufficient conditions

- Suppose the necessary conditions hold.
- Let $\mathbf{H}$ be the $\#(\hat{\mathscr{A}}) \times \#(\hat{\mathscr{A}})$ matrix whose entries are $\partial^2 C_n(\beta)/\partial\beta_k\partial\beta_l$ for $k, l \in \hat{\mathscr{A}}$.
- If $\mathbf{H}$ is a positive definite, then $\hat{\beta}$ is a local minimum.

(*) The positive definiteness of $\mathbf{H}$ holds when the smallest eigenvalue of $\mathbf{X}'_{\hat{\mathscr{A}}}\mathbf{X}_{\hat{\mathscr{A}}}$ is larger than the negative sum of the second derivatives of $J_\lambda(\beta)$ at $\hat{\beta}_k, k \in \hat{\mathscr{A}}$. Hence, for most penalties, it holds when the smallest eigenvalue of $\mathbf{X}'_{\hat{\mathscr{A}}}\mathbf{X}_{\hat{\mathscr{A}}}$ is sufficiently large.

Sufficient conditions for the weak oracle property

Necessary conditions for the weak oracle property

1. For $j \in \mathscr{A}$, $|\hat{\beta}_j^o| > \eta$
2. $\gamma = o(1/\sqrt{n})$.
3. $\phi > \sqrt{2\sigma^2 \log p/n}$
4. $\log p/n \to 0$ and $\sqrt{n}\lambda \to \infty$.

Sufficient conditions for the weak oracle property

Remark

- Conditions 1 and 2 are needed for $\hat{\beta}_{\mathscr{A}}$ is asymptotically equivalent to $\hat{\beta}^{o}_{\mathscr{A}}$.
- Conditions 3 and 4 are need for $\hat{\mathscr{A}} = \mathscr{A}$ asymptotically.
- It is not difficult to see that the SCAD and bridge satisfy the conditions as long as the true signal coefficients are sufficiently large, and so they have the weak oracle property.

An example of bad local minima

Weak oracle property for the bridge penalty.

- Recall $J_\lambda(\beta) = \lambda\beta^q, q \in (0,1)$.
- Let $\lambda^q = o(1/\sqrt{n})$.
- Then, the bridge has the weak oracle property provided

$$\min_{j\in\mathscr{A}} |\beta_j^*| > n^{-1/2q}. \tag{1}$$

- This is because $\gamma = q\lambda^q = o(1/\sqrt{n})$ and $\phi = \infty$.

(*) Huang et al (2008) proved that the standard oracle property holds for $p < n$.

(*) The condition (1) is much weaker than the standard condition $\min_{j\in\mathscr{A}} |\beta_j^*| > \sqrt{\log p/n}$. That is, the bridge estimator can detect small signals better.

An example of bad local minima

Bad local minima

- Let $\mathscr{B}$ be a subset of $\{1, \ldots, p\}$.
- Let $\hat{\beta}^{\mathscr{B}}$ be the bridge estimator with covariates only in $\mathscr{B}$.
- Then, it is a local minimum of the bridge penalized sum of squared residuals with all covariates.

An example of bad local minima

Proof

- Let $\hat{\mathscr{A}} = \{j : \hat{\beta}_j^{\mathscr{B}} \neq 0\}$.
- By the necessary condition of local minima, we have
  - For $j \in \hat{\mathscr{A}}$
    $$X^{j'}(\mathbf{y} - \mathbf{X}\hat{\beta})/n = \mathrm{sign}(\hat{\beta}_j)\nabla(|\hat{\beta}_j|),$$
  - For $j \in \mathscr{B} - \hat{\mathscr{A}}$,
    $$\left|X^{j'}(\mathbf{y} - \mathbf{X}\hat{\beta})/n\right| \leq \phi.$$
- Since $\phi = \infty$, we have
  - For $j \in \hat{\mathscr{A}}$
    $$X^{j'}(\mathbf{y} - \mathbf{X}\hat{\beta})/n = \mathrm{sign}(\hat{\beta}_j)\nabla(|\hat{\beta}_j|),$$
  - For $j \in \hat{\mathscr{A}}^c$,
    $$\left|X^{j'}(\mathbf{y} - \mathbf{X}\hat{\beta})/n\right| \leq \phi.$$

An example of bad local minima

Example of an algorithm producing a bad local minimum

- Consider the following augmented penalized sum of squared residuals

$$S(\beta,\theta)=\frac{1}{2n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i'\beta)^2+\sum_{j=1}^{p}\theta_j^{1-1/q}|\beta_j|+\lambda\sum_{j=1}^{p}\theta_j.$$

- It is easy to see that

$$\min{}_{\theta:\theta_j\geq 0}S(\beta,\theta)=\frac{1}{2n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i'\beta)^2+\lambda\sum_{j=1}^{p}|\beta_j|^q.$$

- Algorithm
  - Initialize $\hat{\beta}$
  - Iterate until convergence
    - $$\hat{\theta}=\operatorname{argmin}_{\theta:\theta_j\geq 0}S(\hat{\beta},\theta)$$
    - $$\hat{\beta}=\operatorname{argmin}_{\beta}S(\beta,\hat{\theta}).$$

An example of bad local minima

- When $\hat{\beta}_j = 0$ initially, then $\hat{\theta}_j = 0$ and so $\hat{\beta}_j = 0$ forever.
- That is, the solution obtained by the algorithm strongly depends on the initial solution.
- If a signal variable is dropped in the initial solution, it will be dropped in the final solution.
- We may start with an initial solution with all coefficients being nonzero.
- Then, the final solution depends on the sizes of coefficients, and it is not obvious where the algorithm converges.

Necessary condition for the strong oracle property

Necessary conditions for the strong oracle property

- $\phi = o(\min_{j\in\mathscr{A}} |\beta_j^*|)$ to avoid under selection.
- $\eta < \min_{j\in\mathscr{A}} |\beta_j^*|$ and $\gamma = o(1/\sqrt{n})$ for $\hat{\beta}_{\mathscr{A}}$ to be asymptotically equivalent to $\hat{\beta}^o_{\mathscr{A}}$.

Necessary condition for the strong oracle property

Examples of penalties satisfying the necessary conditions

- SCAD
- MCP (minimmax concave penalty) of Zhang (2010)
- Truncated l_1 regression of Shen, Zhu and Pan (2010)
- Steamless l_0 penalty of Dicker, Hauang and Lin (201?)

Necessary condition for the strong oracle property

Theorem for the strong oracle property

- If
  - The smallest eigenvalue of $\mathbf{X}'\mathbf{X}/n$ is large (i.e. $p < n$);
  - $\min_{j\in\mathscr{A}} |\beta_j^*| > O(1/\sqrt{n})$;
  - the penalty satisfies the necessary conditions of the strong oracle property,
- then, the penalized estimator has the strong oracle property.
- Proof) Kim and Kwon (2011, To appear in Biometrika)

Necessary condition for the strong oracle property

Remark for the strong oracle property on high dimension

- For high-dimensional models (i.e. $p > n$), it would be too much to expect the strong oracle property.
- However, we can expect the uniqueness of local minima whose sparsity is bounded.
- In fact, we can show this kind of the restrict strong oracle property under the sparse Riesz condition (a condition on $\mathbf{X}$).
- See, Kim and Kwon (2011) for details.

Conclusion

- Bad local minima is a real problem for nonconvex penalties.
- Too sparse penalties (i.e. $\nabla(0+)$ is too large) may suffer from bad local minima problems.
- The weak oracle property is not enough.
- The strong oracle property is important for a given noncovex penalty to be practically useful.
- Care should be given to develop a new nonconvex penalty.