Likelihood Adaptive Modified Penalty and its properties

Yang Feng

Columbia University

International Workshop on Perspectives on High-dimensional Data Analysis

June 9, 2011

Outline

Outline

1. Introduction
2. Likelihood Adaptive Modified Penalty (LAMP)
3. Theoretical Properties
   - Asymptotic stability
   - Oracle Properties
4. Coordinate Descent Algorithm
5. Simulation
6. Summary and Future Work

History of Penalty Functions

- L_0 penalty: AIC (Akaike, 1973), C_p (Mallows, 1973), BIC (Schwartz, 1978), RIC (Foster and George, 1994), Extended BIC (Chen and Chen, 2008), ...
- L_q penalty (bridge penalty): $0 < q \leq 2$, Frank and Friedman (1993)
- Non-negative garrote: Breiman (1995)
- L_1 penalty (LASSO): Tibshirani (1996), Osborne et. al (2000), Efron et. al (2004)
- SCAD penalty: Fan and Li (2001)
- Elastic Net: Zou and Hastie (2005)
- Adaptive LASSO: Zou (2006)
- Minimax concave penalty (MCP): Zhang (2010)
- ......
- ......

Sparsity and Stability

- Convex penalties: very stable, efficient algorithms.
- Non-convex penalties: gain sparsity, but sacrifice some stability.
- SCAD and MCP: choose an appropriate γ.

Goal

Good balance between

Sparsity and Stability.

Outline

Negative Absolute Prior (NAP)

- $X_1, \cdots, X_n$, i.i.d, with pdf $f(x, \boldsymbol{\beta})$
- **Negative Absolute Prior (NAP)** for $\boldsymbol{\beta}$:
  $p(\boldsymbol{\beta}) \propto \prod_{i=1}^{k} 1/f(C_i, |\boldsymbol{\beta}|)$ for some integer $k > 0$ and k constants $C_i, i = 1, \cdots, k$.
- The prior: remove noisy sample information by selecting appropriate C_i's.
- $|\boldsymbol{\beta}|$: treats unknown sign of $\boldsymbol{\beta}$.
- A conjugate prior if $p(\boldsymbol{\beta}) \propto \prod_{i=1}^{k} f(C_i, \boldsymbol{\beta})$.

Example

$\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, i.i.d, with pdf $f(\alpha + \boldsymbol{X}^T\beta)$. Then $\beta_j \propto 1/f(c_1 + c_2|\beta_j|)$ for some constants c_1 and c_2 is a NAP for $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$.

Negative Absolute Prior (NAP)

- $X_1, \cdots, X_n$, i.i.d, with pdf $f(x, \boldsymbol{\beta})$
- **Negative Absolute Prior (NAP)** for $\boldsymbol{\beta}$: $p(\boldsymbol{\beta}) \propto \prod_{i=1}^{k} 1/f(C_i, |\boldsymbol{\beta}|)$ for some integer $k > 0$ and k constants $C_i, i = 1, \cdots, k$.
- The prior: remove noisy sample information by selecting appropriate C_i's.
- $|\boldsymbol{\beta}|$: treats unknown sign of $\boldsymbol{\beta}$.
- A conjugate prior if $p(\boldsymbol{\beta}) \propto \prod_{i=1}^{k} f(C_i, \boldsymbol{\beta})$.

Example

$\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, i.i.d, with pdf $f(\alpha + \boldsymbol{X}^T\beta)$. Then $\beta_j \propto 1/f(c_1 + c_2|\beta_j|)$ for some constants c_1 and c_2 is a NAP for $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$.

Negative Absolute Prior (NAP)

- $X_1, \cdots, X_n$, i.i.d, with pdf $f(x, \boldsymbol{\beta})$
- **Negative Absolute Prior (NAP)** for $\boldsymbol{\beta}$: $p(\boldsymbol{\beta}) \propto \prod_{i=1}^{k} 1/f(C_i, |\boldsymbol{\beta}|)$ for some integer $k > 0$ and k constants $C_i, i = 1, \cdots, k$.
- The prior: remove noisy sample information by selecting appropriate C_i's.
- $|\boldsymbol{\beta}|$: treats unknown sign of $\boldsymbol{\beta}$.
- A conjugate prior if $p(\boldsymbol{\beta}) \propto \prod_{i=1}^{k} f(C_i, \boldsymbol{\beta})$.

Example

$\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, i.i.d, with pdf $f(\alpha + \boldsymbol{X}^T \boldsymbol{\beta})$. Then $\beta_j \propto 1/f(c_1 + c_2|\beta_j|)$ for some constants c_1 and c_2 is a NAP for $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$.

Likelihood Adaptive Modified Penalty (LAMP)

- For generalized linear model $Y_i|\boldsymbol{X}_i;\boldsymbol{\theta} \sim^{i.i.d} \exp(T(y)\xi - g(\xi) + h(x)), i = 1, \cdots, n$, where $\xi = \boldsymbol{X}^T\boldsymbol{\theta}$, and $\boldsymbol{X} = (1, X_1, X_2, \cdots, X_p)^T$, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\tau)^\tau$.
- Log-likelihood:

$$l = \frac{1}{n}\sum_{i=1}^{n}[T(Y_i)\xi_i - g(\xi_i)],$$

- We propose to maximize:

$$l - \sum_{j=1}^{p} \frac{\lambda_n^2}{-g'(\alpha_1)\lambda_0}[\alpha_1 + \frac{\lambda_0}{\lambda_n}|\beta_j| - g(\alpha_1 + \frac{\lambda_0}{\lambda_n}|\beta_j|)].$$

- *Taking a NAP*: bayesian interpretation.

Likelihood Adaptive Modified Penalty (LAMP)

- For convenience, take away $\frac{\lambda_0}{\lambda_n}|\beta_j|$, define **Likelihood Adaptive Modified Penalty**:

$$p_{\lambda_n}(\beta_j) = \frac{\lambda_n^2}{g'(\alpha_1)\lambda_0}[g(\frac{\lambda_0}{\lambda_n}\beta_j + \alpha_1) - g(\alpha_1)], \qquad (1)$$

where α_1 and λ_0 are positive constants.

- For identifiability purpose: $p_{\lambda_n}(0) = 0$ and $p'_{\lambda_n}(0) = \lambda_n$.
- Note that:

$$p'_{\lambda_n}(\theta) = \lambda_n g'(\frac{\lambda_0}{\lambda_n}\theta + \alpha_1)/g'(\alpha_1),$$

$$p''_{\lambda_n}(\theta) = \lambda_0 g''(\frac{\lambda_0}{\lambda_n}\theta + \alpha_1)/g'(\alpha_1).$$

Likelihood Adaptive Modified Penalty (LAMP)

- For convenience, take away $\frac{\lambda_0}{\lambda_n}|\beta_j|$, define **Likelihood Adaptive Modified Penalty**:

$$p_{\lambda_n}(\beta_j) = \frac{\lambda_n^2}{g'(\alpha_1)\lambda_0}[g(\frac{\lambda_0}{\lambda_n}\beta_j + \alpha_1) - g(\alpha_1)], \qquad (1)$$

  where α_1 and λ_0 are positive constants.
- For identifiability purpose: $p_{\lambda_n}(0) = 0$ and $p'_{\lambda_n}(0) = \lambda_n$.
- Note that:

$$p'_{\lambda_n}(\theta) = \lambda_n g'(\frac{\lambda_0}{\lambda_n}\theta + \alpha_1)/g'(\alpha_1),$$

$$p''_{\lambda_n}(\theta) = \lambda_0 g''(\frac{\lambda_0}{\lambda_n}\theta + \alpha_1)/g'(\alpha_1).$$

Likelihood Adaptive Modified Penalty (LAMP)

- For convenience, take away $\frac{\lambda_0}{\lambda_n}|\beta_j|$, define **Likelihood Adaptive Modified Penalty**:

$$p_{\lambda_n}(\beta_j) = \frac{\lambda_n^2}{g'(\alpha_1)\lambda_0}[g(\frac{\lambda_0}{\lambda_n}\beta_j + \alpha_1) - g(\alpha_1)], \qquad (1)$$

  where α_1 and λ_0 are positive constants.
- For identifiability purpose: $p_{\lambda_n}(0) = 0$ and $p'_{\lambda_n}(0) = \lambda_n$.
- Note that:

$$p'_{\lambda_n}(\theta) = \lambda_n g'(\frac{\lambda_0}{\lambda_n}\theta + \alpha_1)/g'(\alpha_1),$$

$$p''_{\lambda_n}(\theta) = \lambda_0 g''(\frac{\lambda_0}{\lambda_n}\theta + \alpha_1)/g'(\alpha_1).$$

Quick Facts

- For linear regression, $g(\xi) = \xi^2/2$, LAMP reduces to Elastic Net (Zou and Hastie, 2005);
- Smooth $g(\cdot)$, smooth LAMP in $(0, \infty)$.
- SCAD and MCP $\notin \mathcal{C}^2(0, \infty)$.
- Intuition: smoothness in penalty $\rightarrow$ more stable optimization problem.

Quick Facts

- For linear regression, $g(\xi) = \xi^2/2$, LAMP reduces to Elastic Net (Zou and Hastie, 2005);
- Smooth $g(\cdot)$, smooth LAMP in $(0, \infty)$.
- SCAD and MCP $\notin \mathcal{C}^2(0, \infty)$.
- Intuition: smoothness in penalty $\rightarrow$ more stable optimization problem.

Other LAMP Examples

- Logistic regression: $g(\xi) = \log(1 + e^{\xi}) - \xi$, $p_{\lambda_n}(\theta) = \frac{\lambda_n^2(1+\rho)}{\lambda_0} \log[\frac{(1+\rho)e^{\lambda_0/\lambda_n\theta}}{1+\rho e^{\lambda_0/\lambda_n\theta}}]$, sigmoid penalty;
- Poisson regression: $g(\xi) = e^{-\xi}$, poisson penalty;
- Gamma regression: $g(\xi) = -k\log(\xi)$;
- Inverse gaussian regression: $g(\xi) = -l\sqrt{\xi}$;
- Probit model: $g(\xi) = -\log(\Phi(\xi))$.

Other LAMP Examples

- Logistic regression: $g(\xi) = \log(1 + e^{\xi}) - \xi$, $p_{\lambda_n}(\theta) = \frac{\lambda_n^2(1+\rho)}{\lambda_0} \log[\frac{(1+\rho)e^{\lambda_0/\lambda_n \theta}}{1+\rho e^{\lambda_0/\lambda_n \theta}}]$, sigmoid penalty;
- Poisson regression: $g(\xi) = e^{-\xi}$, poisson penalty;
- Gamma regression: $g(\xi) = -k \log(\xi)$;
- Inverse gaussian regression: $g(\xi) = -l\sqrt{\xi}$;
- Probit model: $g(\xi) = -\log(\Phi(\xi))$.

Comparison with other penalties

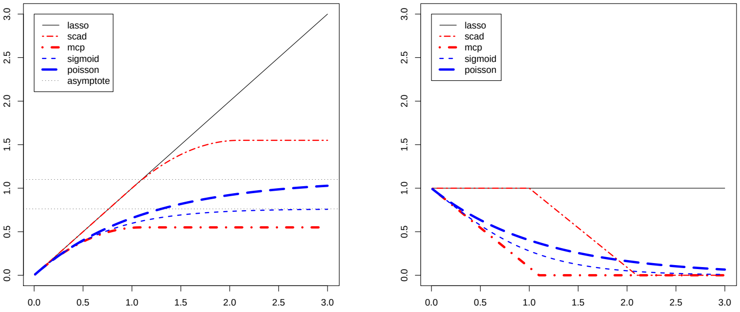


Figure: Penalty functions and their derivatives of the Lasso, SCAD, MCP, Sigmoid(for logistic), and Poisson. We choose $\lambda = 1$, $\gamma = 1.1$ for MCP, $\gamma = 2.1$ for SCAD, $\lambda_0 = 1/1.1$ for Sigmoid and Poisson penalty, and $\rho = 1$ for Sigmoid penalty. We choose parameters this way to keep the maximum concavity of these penalties the same.

Outline

Asymptotic stability

Minimize random function

$$M_n(\boldsymbol{x}, \theta) = \frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i, \theta) + r_n(\theta), \theta \in \Theta.$$

Denote by $\mathrm{arglmin} M_n(x, \theta)$

$$\{\theta^* \in \Theta | \text{for some } c > 0, M_n(\boldsymbol{x}, \theta^*) \le M_n(\boldsymbol{x}, \theta), \forall \|\theta - \theta^*\| \le c\}.$$

Two types of asymptotic stability

F_n denotes the domain for $\boldsymbol{x}$, $\mathrm{diam}_d(A) \triangleq \max_{\boldsymbol{x},\boldsymbol{y}\in A} \|\boldsymbol{x}-\boldsymbol{y}\|_d$ for measure d and set A.

We call $\mathrm{arglmin} M_n(\boldsymbol{x},\theta)$ satisfies **weak asymptotic stability** if $\forall \epsilon > 0, \forall \boldsymbol{x} \in F_n$ fixed,

$$\mathrm{P}(\overline{\lim_{n\to\infty}}\, \overline{\lim_{\delta\to 0}}\, \mathrm{diam}_d \bigcup_{\substack{\|\boldsymbol{\delta}\|_d<\delta \\ \boldsymbol{x}+\boldsymbol{\delta}\in F_n}} \{\mathrm{arglmin} M_n(\boldsymbol{x}+\boldsymbol{\delta},\theta)\} = 0) = 1; \quad (2)$$

Two types of asymptotic stability (Cont')

$\mathrm{arglmin} M_n(\boldsymbol{x}, \theta)$ satisfies **strong asymptotic stability** if *weak asymptotic stability* holds and $\forall \epsilon > 0, \forall \boldsymbol{x} \in F_n$ fixed,

$$\varliminf_{n \to \infty} \mathrm{P}(\varlimsup_{\delta \to 0} \mathrm{diam}_d \bigcup_{\substack{\|\boldsymbol{\delta}\|_d < \delta \\ \boldsymbol{x}+\boldsymbol{\delta} \in F_n}} \{\mathrm{arglmin} M_n(\boldsymbol{x}+\boldsymbol{\delta}, \theta)\} = 0) = 1. \quad (3)$$

Example of asymptotic stability

Let $\boldsymbol{x}_n^{(1)}$ and $\boldsymbol{x}_n^{(2)}$ be two local minimizers of $M_n(\boldsymbol{x}, \theta)$.

- Weak asymptotic stability:
  $P(\lim_{n\to\infty} \|\boldsymbol{x}_n^{(1)} - \boldsymbol{x}_n^{(2)}\|_d = 0) = 1$.
- Strong asymptotic stability:
  $P(\lim_{n\to\infty} \boldsymbol{x}_n^{(1)} = \boldsymbol{x}_n^{(2)}) = 1$.

Asymptotic stability

Lemma

Suppose $\forall \boldsymbol{x} \in F_n$. $m_n(\boldsymbol{x}, \theta) = \frac{1}{n}\sum_{i=1}^n m(\boldsymbol{x}_i, \theta)$ *is a smooth strictly convex function of* θ*.* $\forall$ *fixed* θ, $m_n(\boldsymbol{x}, \theta)$ *is continuous in* $\boldsymbol{x}$*.*

1. $r_n(\theta) > 0, r_n(\theta) \to 0$ *uniformly in* $\theta \in \Theta$ *as* $n \to \infty$*,*
2. *Each* $\boldsymbol{x} \in F_n$ *is a limit point.*
3. $m_n(\boldsymbol{x}, \theta)$ *is an open map:* $F_n \to [\min(m_n(\boldsymbol{x}, \theta)), +\infty)$ *with* θ *fixed,*
4. $\mathrm{E}M_n(\boldsymbol{x}, \theta)$ *is strictly convex at* $\theta_0 = \arg\min \mathrm{E}m_n(\boldsymbol{x}, \theta)$.

$\mathrm{arglmin} M_n(\boldsymbol{x}, \theta)$ *has weak asymptotic stability If (1)-(3) hold; strong asymptotic stability if (1)-(4) hold.*

Oracle Properties for LAMP

Theorem

We assume p *is fixed. Let* $\{\hat{\boldsymbol{\beta}}\}$ *be the set of local maximizers of* $\tilde{l}(\boldsymbol{\beta}) = \sum_{i=1}^{n}[T(y_i)\xi_i - g(\xi_i)] - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|)$*, where* $\xi_i = \tilde{X}_i^T\boldsymbol{\theta}$*. Under some regularity conditions,*

- *any* $\hat{\boldsymbol{\beta}}$ *has estimation consistency and oracle property (in the sense defined by Fan and Li, 2001),*
- $\{\hat{\boldsymbol{\beta}}\}$ *has strong asymptotic stability.*

Outline

Algorithm

We use the popular coordinate descent algorithm (Friedman et. al, 2010; Keerthi and Shevade, 2007) to calculate the solution path.

1. Start with $\lambda_{\max}$ where all the estimates are 0.
2. Do optimization on a grid $\tau\lambda_{\max} = \lambda_K < \lambda_{K-1} < \cdots < \lambda_1 = \lambda_{\max}$. When $\lambda = \lambda_{i+1}$, use the solution from $\lambda = \lambda_i$ as the initial value to speed up. Here $\tau = 0.01$, $K = 100$.
3. Inside each λ_i, we cycle through all the variables and update their coefficients until converge.

Algorithm

We use the popular coordinate descent algorithm (Friedman et. al, 2010; Keerthi and Shevade, 2007) to calculate the solution path.

1. Start with $\lambda_{\max}$ where all the estimates are 0.
2. Do optimization on a grid $\tau\lambda_{\max} = \lambda_K < \lambda_{K-1} < \cdots < \lambda_1 = \lambda_{\max}$. When $\lambda = \lambda_{i+1}$, use the solution from $\lambda = \lambda_i$ as the initial value to speed up. Here $\tau = 0.01$, $K = 100$.
3. Inside each λ_i, we cycle through all the variables and update their coefficients until converge.

Remarks on the algorithm

- In Step 3, when updating each variable, use Newton-Raphson algorithm.
- In Step 3, we calculate a *violation* measure for all the variables and update the variables according to the descending order of the violation (Keerthi and Shevade, 2007).

$$\mathrm{viol}_j = \begin{cases} |F_j|, & \text{if } j = 0; \\ \max\{0, -n\lambda_n - F_j, -n\lambda_n + F_j\}, & \text{if } \beta_j = 0, j > 0; \\ |F_j - n\,\mathrm{sgn}(\beta_j) p'_{\lambda_n}(|\beta_j|)|, & \text{if } \beta_j \neq 0, j > 0. \end{cases}$$

Outline

Logistic Regression

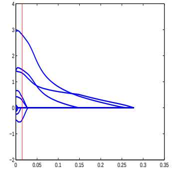
$n = 200$, logistic regression model with intercept $\alpha = -2$ and $\boldsymbol{\beta} = (3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0)^T$. $X_i \sim N(0, 1)$ and $Corr(X_i, X_j) = 0.5^{|i-j|}$. MRME stands for median of ratios of ME of a selected model to that of the ordinary least squares estimate under the oracle model. ME here stands for "model error" defined as:

$$\mathrm{ME}(\hat{\theta}) = E[E_{\hat{\theta}}(Y|\boldsymbol{X}) - E_{\theta_0}(Y|\boldsymbol{X})]^2.$$

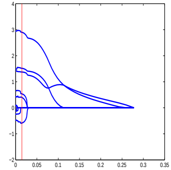
Logistic Regression (Cont')

Penalties	TP	FP	cf	of	uf	L1	L2	MRME
sig(.08)	2.67	0.56	0.37	0.30	0.33	2.28	1.22	0.51
sig(.09)	2.72	0.59	0.47	0.25	0.28	2.37	1.24	0.49
sig(.1)	2.69	0.55	0.42	0.27	0.31	2.43	1.26	0.50
lasso	2.99	3.25	0.04	0.95	0.01	2.54	1.18	0.50
alasso(3)	2.82	0.98	0.37	0.45	0.18	2.27	1.15	0.47
alasso(4)	2.69	0.52	0.44	0.25	0.31	1.92	1.05	0.41
alasso(5)	2.50	0.15	0.43	0.08	0.49	1.86	1.08	0.52
mcp(2.8)	2.79	0.66	0.37	0.42	0.21	2.05	1.13	0.48
mcp(2.9)	2.79	0.68	0.41	0.38	0.21	2.21	1.17	0.54
mcp(3)	2.81	0.87	0.36	0.45	0.19	2.33	1.21	0.50
scad(2.4)	2.80	1.01	0.25	0.55	0.20	2.25	1.20	0.54
scad(2.7)	2.87	1.15	0.30	0.57	0.13	2.43	1.27	0.55
scad(3)	2.89	1.24	0.24	0.65	0.11	2.20	1.17	0.50

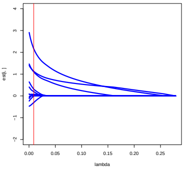
Solution Paths



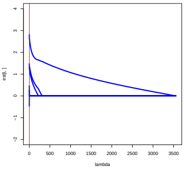
(a) Sigmoid (0.05)



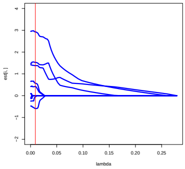
(b) Sigmoid (0.09)



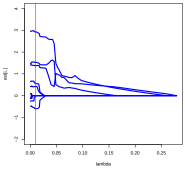
(c) Lasso



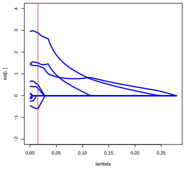
(d) ALasso (4)



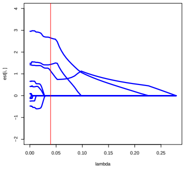
(e) SCAD (3)



(f) SCAD (2.7)



(g) MCP (2.9)



(h) MCP (2)

Figure: Solution Paths for all the penalties using different λ_0 or γ for selection.

Corresponding penalties

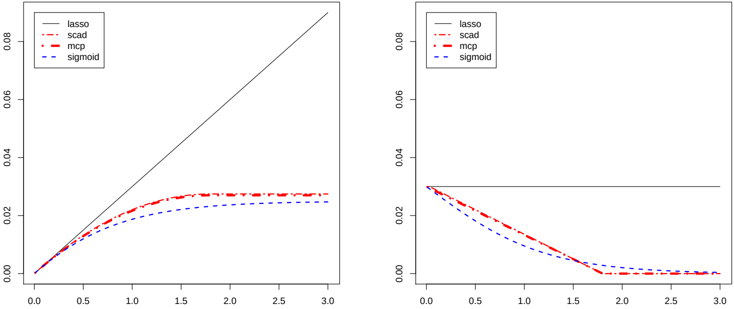


Figure: Left Panel: penalty functions. Right Panel: the first derivative of the penalties

Test the stability

Test stability of different penalties on logistic regression.

1. 10-fold cross-validation $\longrightarrow$ mean standard error of the estimates for the coefficients.
2. Perturb the sample with gaussian error and do the previous simulation.
3. Repeat each simulation for 100 times

Test the stability (Cont')

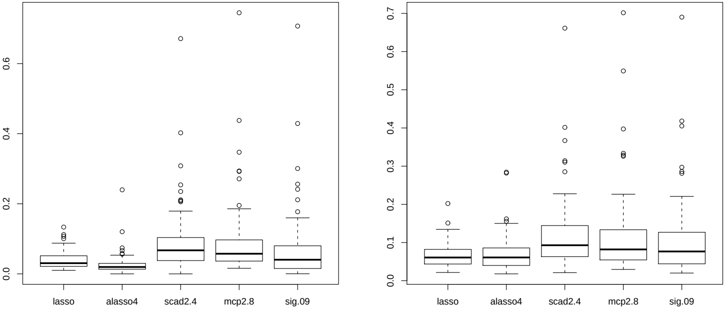


Figure: Boxplots for the mean standard error. Left: without random error. Right: with random error.

Stability and Sparsity Balance: Revisited

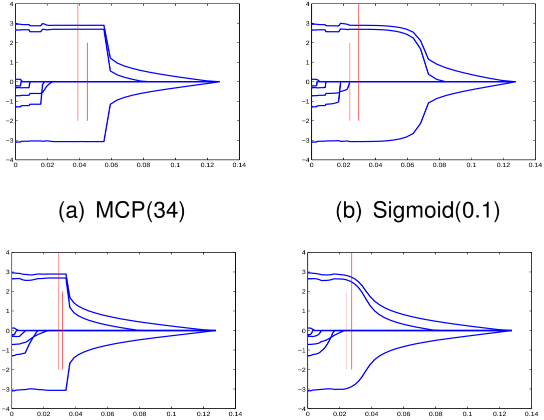
We use the same logistic regression model but with
$\alpha = -4, \boldsymbol{\beta} = (-2, 1.5, 0, 0, 2, 0, 0, 0)^T$.

Penalty	λ_0	False0	True0
Sigmoid	.04	0	4.69
Sigmoid	.05	.01	4.79
Sigmoid	.06	.01	4.85
Sigmoid	.07	.01	4.86
Sigmoid	.09	.01	4.89
Sigmoid	.10	0	4.89
Sigmoid	.12	0	4.89
Sigmoid	.14	0	4.89
Sigmoid	.16	0	4.89
Sigmoid	.18	0	4.89

Penalty	γ	False0	True0
MCP	1.1	0	4.89
MCP	4	0	4.89
MCP	12	0	4.89
MCP	27	0	4.89
MCP	34	.01	4.90
MCP	41	.01	4.88
MCP	48	0	4.83
MCP	55	0	4.77
MCP	62	0	4.71
MCP	69	0	4.65

Table: BIC: 100 simulations

Solution Path Comparison



(a) MCP(34)

(b) Sigmoid(0.1)

(c) MCP(69)

(d) Sigmoid(0.04)

Figure: λ_0 and γ is chosen to make the penalties have similar sparsity.

Poisson Regression

$Y|\boldsymbol{X} \sim \text{Poisson}(\boldsymbol{X}^T\boldsymbol{\beta} + \alpha)$, $\boldsymbol{X}$ follows the same distribution and correlation structure as before, and the true parameters $\boldsymbol{\beta} = (.6, .4, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)^T, \alpha = -1$. We perform 100 simulations with sample size $n = 100$.

Poisson Regression (Cont')

Penalties	TP	FP	cf	of	uf	L1	L2	MRME
poi(2.7)	2.82	0.47	0.53	0.29	0.18	0.61	0.34	0.21
poi(3)	2.80	0.52	0.56	0.24	0.20	0.65	0.36	0.22
poi(10)	2.80	0.61	0.51	0.29	0.20	0.66	0.36	0.23
lasso	3.00	4.05	0.04	0.96	0.00	1.01	0.43	0.37
alasso(4)	2.69	1.24	0.29	0.41	0.30	0.84	0.43	0.31
alasso(5)	2.52	0.64	0.31	0.27	0.42	0.80	0.44	0.28
alasso(6)	2.25	0.44	0.25	0.16	0.59	0.88	0.50	0.34
mcp(1.1)	2.81	0.84	0.46	0.35	0.19	0.74	0.38	0.30
mcp(1.4)	2.80	0.48	0.52	0.28	0.20	0.64	0.35	0.23
mcp(1.7)	2.80	0.83	0.47	0.33	0.20	0.73	0.38	0.28

Outline

Summary

- Motivation: balance between stability and sparsity
- Solution: Likelihood Adaptive Modified Penalty (LAMP)
- Conclusion: strong asymptotic stability and oracle properties.

Future Work

- Diverging p.
- Characterization of degrees of stability.

Acknowledgements

- Organizers of the conference.
- Collaborators: Tengfei Li, Wen Yu, Zhiliang Ying, and Hong Zhang.