

INFORMATION THEORETICAL OPTIMALITY OF VARIABLE SELECTION WITH MINIMAX CONCAVE PENALTY

Cun-Hui Zhang

Department of Statistics and Biostatistics, Rutgers University

czhang@stat.rutgers.edu

Research partially supported by the NSF and NSA

**Beijing International Conference
on Machine Learning and Data Mining
June 16-19, 2008**

Thanks for the invitation!

THE PROBLEM

ℓ_1 SELECTOR

SIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

Data: response vector \mathbf{y} and covariate/feature vectors as columns of $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$; In linear regression

$$\mathbf{y} = \sum_{j=1}^p \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$$

Difficulties: small n , large p ; often $p \gg n$ in bioinformatics, fMRI, networks applications and more

Aims of variable selection:

- Parsimonious and more interpretable **models** (sparsity)
- **Estimation** with (oracle) efficiency
- **Prediction** of a response variable (w/o over fitting)
- **Applications** or **further experiments**

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

Methodologies:

- Subset selection is computationally impossible
- Penalized methods: **loss + continuous penalty**
- Optimization in restricted parameter space
- Threshold gradient descent: find a trajectory with selection features and converging to minimum loss

Problems:

- Still **computationally** nontrivial for large datasets
- Which **method**, **penalty**/norm, and at what **level**?
- Statistical properties: **selection consistency**, prediction or generalization error, estimation efficiency, **information limits**

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

Basis pursuit/LASSO: Chen-Donoho (94), Tibshirani (96)

$$\hat{\beta} = \arg \min_{\beta} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

Computation: LARS, Osborne et al (00), Efron et al (04)

Related methods: Boosting, Schapire (90), Freund-Schapire (96), Friedman-Hastie-Tibshirani (00); Dantzig selector, Candes-Tao (05), ...

Prediction and estimation: Greenshtein-Ritov (04), Bunea-Tsybakov-Wegkamp (06), Meinshausen-Yu (06), van de Geer (06), Zhang-Huang (06), ...

Extension to GLM and general loss: Gerkin-Lewis-Madigan (04), Zhao-Yu (04), van de Geer (06), ...

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

Selection/sign consistency: Meinshausen-Buhlmann (06), Tropp (06), Zhao-Yu (06), Wainwright (06) proved

$$P\left\{\text{sgn}(\hat{\beta}_j) = \text{sgn}(\beta_j) \quad \forall j\right\} \rightarrow 1,$$

under quite strong conditions:

$$\max_{i \notin A^\circ} \left\| \mathbf{x}'_i \mathbf{X}_{A^\circ} (\mathbf{X}'_{A^\circ} \mathbf{X}_{A^\circ})^{-1} \right\|_1 < \kappa < 1$$

with $A^\circ = \{j : \beta_j \neq 0\}$ and $\mathbf{X}_A = \{\mathbf{x}_j, j \in A\}$, and

$$\beta_* = \min_{j \in A^\circ} |\beta_j| \geq M\sigma \sqrt{|A^\circ|(\log p)/n}$$

under the standardization $\|\mathbf{x}_j\|_2^2 = n$.

Information limits: Wainwright (07) proved that for standard Gaussian \mathbf{X} , the LASSO attains the info limit

$$\text{selection consistency} \Leftrightarrow \beta_* \geq M\sigma\sqrt{(\log p)/n}$$

General designs: Zhang (2007) proved

$$\text{selection consistency} \Rightarrow \frac{n\beta_*^2}{\log p} \geq \frac{\sigma^2}{2}(1 + o(1))$$

provided $\|\mathbf{x}_j\|_2^2 = n \forall j$ or $E\|\mathbf{x}_j\|_2^2 = n \forall j$.

- Lower bounds derived using Birgé (83) & Yatracos (88)
- LASSO does not perform well in selection consistency for moderate d^o , due to its **bias**

$$\mathbf{X}_{\hat{A}}(\mathbf{y} - \mathbf{X}\hat{\beta})/n = \text{sgn}(\hat{\beta}_{\hat{A}})\lambda, \quad \text{as in KKT}$$

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

TABLE: LASSO, MC+(Zhang, 07) and SCAD(Fan-Li, 01)

$\%(\hat{A} = A^\circ)$ & mean(# steps) based on 100 replications

$n = 300, p = 200, \|\mathbf{x}_j\|^2 = n, E\mathbf{X}'\mathbf{X}/n \neq \mathbf{I}_p$

$\beta_*/\sigma = 1/2, \lambda/\hat{\sigma} = \{2^k(\log p)/n\}^{1/2}$ for rows $k = 1, 2, 3$

$ A^\circ = 10$			$ A^\circ = 20$			$ A^\circ = 40$		
lasso	mc+	scad	lasso	mc+	scad	lasso	mc+	scad
0.34	0.76	0.70	0.06	0.78	0.61	0.01	0.84	0.24
12	16	26	23	32	51	48	65	132
0.88	0.97	0.93	0.41	0.81	0.49	0.01	0.11	0.00
11	11	14	21	21	27	42	41	57
0.39	0.40	0.39	0.07	0.08	0.07	0.00	0.00	0.00
10	10	10	17	17	17	31	28	32

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

TABLE: Standard Gaussian \mathbf{X} , $\lambda/\sigma = \sqrt{2(\log p)/n}$ Nonzero β_j are independent $\pm(\beta_* + |N(0, 1)|)$

$n, p, A^\circ $	200, 1000, 8			800, 20000, 18	
β_*/σ	$4(\log n)/\sqrt{n}$			$5(\log n)/\sqrt{n}$	
method	lasso	mc+	scad	mc+	mc+(\hat{\sigma})
$\%(\hat{A} = A^\circ)$	0.08	0.86	0.87	0.91	0.91
mean(# steps)	11	17	25	37	37
med($\ \hat{\beta} - \beta\ ^2$)	1.52	0.18	0.18	0.09	0.09
β_*/σ	$(\log n)/\sqrt{n}$				
$\%(\hat{A} = A^\circ)$	0.07	0.54	0.30		
mean(# steps)	11	13	17		
med($\ \hat{\beta} - \beta\ ^2$)	1.34	0.34	0.76		

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

TABLE: IID Gaussian rows in $\mathbf{X} = (x_{ij})$, $E x_{ij} x_{ik} = (3/4)^{|j-k|}$ Nonzero β_j are independent $\pm(1 + |N(0, 1)|)$

(n, p, A°) penalty level λ	$(250, 10000, 30)$			
	$\sqrt{2(\log p)/n}$		$\sqrt{10(\log p)/n}$	
	mc+	lasso	mc+	lasso
$\% \{ \widehat{A} = A^\circ \}$	80	0	83	0
$\widehat{E}(\text{FN})$	0.02	1.46	0.03	4.38
$\widehat{E}(\text{FP})$	0.29	115	0.23	68
$\widehat{E}(\text{steps})$	108	172	108	102
$\widehat{E} \ \widehat{\beta} - \beta^\circ\ ^2$	0.35	23	0.36	38
$\widehat{E} \ \mathbf{X}\widehat{\beta} - \mathbf{X}\beta^\circ\ ^2$	56	789	58	3084

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

Controlling the false negative: Zhang-Huang (06)

$$P\left\{\text{sgn}(\hat{\beta}_j) = \text{sgn}(\beta_j) \forall j \in A^o, |\hat{A}| = O(|A^o|)\right\} \rightarrow 1,$$

under $\beta_* \geq M_1 \sigma \sqrt{|A^o|(\log p)/n}$, $M_2 |A^o| + 1 \leq d^*$ and the SRC: for all $|A| \leq m \leq d^*$ and $\|\mathbf{u}\|_2^2 = 1/n$

$$c_* \leq \|\mathbf{X}_A \mathbf{u}\|_2^2 \leq c^*$$

- Consistency via post LASSO selection: Meinshausen (06), Zou (06, adaptive LASSO)
- The gap with the factor $\sqrt{|A^o|}$ persists
- Needs to capture A^o before $|\hat{A}| = n$

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

TABLE: Sparse recovery: $\sigma = 0$, $\lambda = 0+$

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ with standard Gaussian $\{\mathbf{X}, \beta_j, j \in A^o\}$

From the top row: $\% \{\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}\}$, $\hat{E}[\text{FN} | \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}]$.

$\hat{E}[\#\text{steps} | \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}]$, $\hat{E}[\#\text{steps} | \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}]$

(n, p, m^o)					
$(100, 2000, 15)$		$(100, 2000, 28)$		$(200, 10000, 40)$	
mc+	ℓ_1	mc+	ℓ_1	mc+	ℓ_1
100	51	73	0	100	0
	2	19	13		18
32	65	87		102	
	144	513	153		311

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

Nearly unbiased selection: Fan-Li (01), Fan-Peng (04), ...

- Removing the bias of PLSE: Using penalty $\sum_j \rho(|\beta_j|; \lambda)$ satisfying $\rho'(0+; \lambda) = \lambda$, $\rho'(t; \lambda) = 0$, $t \geq \gamma\lambda$
- SCAD: $\rho'(t; \lambda) = \lambda I_{[0,1]} + (1 - (t - 1)/(\gamma - 1))_+ I_{(1,\infty)}$

Non-convex minimization algorithms: Fan-Li (01), Hunter-Li (05), Zou-Li (06); iterative approximation of a local (hopefully global) minimum of penalized loss

More: Stepwise regression/matching pursuit; Bridge penalty, Frank-Friedman (93); Adaptive LASSO, Zou (06, adaptive LASSO), Huang-Ma-Zhang (06); Twin boosting, Bühlmann (07); Correlation screening, Fan-Lv (07)

MC+: THE *minimax concave penalty* (MCP)

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

MCP: Zhang (07)

$$\begin{aligned}\rho(t; \lambda) &= \lambda \int_0^t (1 - x/(\gamma\lambda))_+ dx, \\ \rho'(t; \lambda) &= \lambda(1 - t/(\gamma\lambda))_+, t \geq 0.\end{aligned}$$

Among all continuous spline penalty functions satisfying $\rho'(0+; \lambda) = 0$ and $\rho'(t; \lambda) = 0$ on $[\gamma\lambda, \infty)$, the MCP

- minimizes the maximum concavity $\sup_{t>0} -\rho''(t; \lambda)$
- has the smallest number of knots
- has smaller bias $\rho'(t; \lambda)$ than the SCAD

These properties has consequences in computational complexity and selection consistency

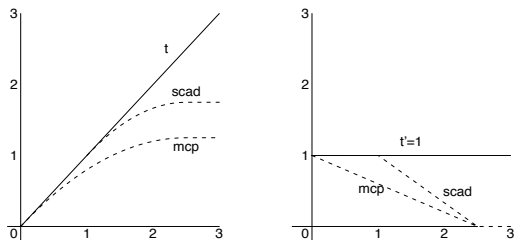


FIGURE: The ℓ_1 penalty $\rho(t; \lambda) = \lambda t$ along with the MCP and the SCAD penalty, $\lambda = 1$ and $t > 0$. Left: $\rho(t; 1)$; Right: $\rho'(t; 1)$

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

PLUS: local minimization of the penalized ℓ_2 loss

$$\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^p \lambda^2 \rho(|\beta_j|/\lambda)$$

for quadratic splines $\rho(t)$ satisfying $\rho'(0+) = 1$

- Goal: find the **sparsest local minimizer**
- Strategy: trace a continuous **path of critical points** beginning with 0 and ends with a perfect fit
- Estimating equation via subdifferentiation:

$$\begin{cases} \mathbf{x}'_j(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n = \text{sgn}(\beta_j)\lambda\rho'(|\beta_j|/\lambda), & \beta_j \neq 0 \\ |\mathbf{x}'_j(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n| \leq \lambda, & \beta_j = 0 \end{cases}$$

- Implementation: solve a finite sequence of linear problems as in LARS; same computational cost per step

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

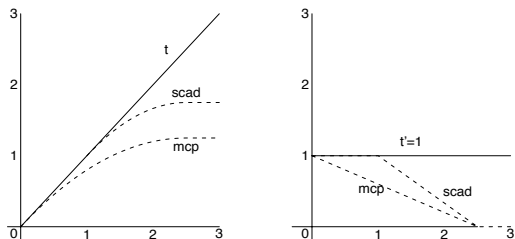


FIGURE: For computational efficiency, we need (i) small number of knots: 1,2,3 for ℓ_1 , MCP, SCAD; (ii) small maximum concavity: 0, $1/\gamma$, $1/(\gamma - 1)$; and (iii) small bias $\rho'(t)$

MC+: MCP & PLUS

$$\rho'(t; \lambda) = \lambda(1 - t/(\gamma\lambda))_+$$

Choice of tuning parameters λ and γ :

- We choose the **“universal” penalty/threshold level**

$$\lambda = \hat{\sigma} \sqrt{2(\log p)/n}$$

- We choose the **sparsest local minimizer** among all solutions for given λ
- We choose γ for certain **sparse convexity** of the penalized loss given $\#\{j : \beta_j \neq 0\}$, to provide some **uniqueness** and to control computational complexity ($\gamma = \infty$ for the LASSO)

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

TABLE: Choice of γ

$\%(\hat{A} = A^o)$ & mean(# steps) based on 100 replications
 $n = 300, p = 2000, |A^o| = 30, \|\mathbf{x}_j\|^2 = n, E\mathbf{X}'\mathbf{X}/n \neq \mathbf{I}_p$
 $\lambda/\sigma = \{2^{k-1}(\log p)/n\}^{1/2}$ for rows $k = 1, 2, 3$
 SCAD for $\gamma = 2.4^*$, LASSO for $\gamma = \infty$

	γ for $\beta_*/\sigma = 1/2$				γ for $\beta_*/\sigma = 3/8$					
	1.1	1.4	1.7	2.4*	∞	1.1	1.4	1.7	2.4*	∞
0.02	0.02	0.02	0.02	0.01	0.00	0.02	0.02	0.02	0.00	0.00
366	104	78.3	240	80.0	680	167	100	231	75.2	
0.93	0.93	0.93	0.01	0.00	0.26	0.14	0.05	0.00	0.00	
353	98.4	72.7	128	53.9	392	111	61.4	56.4	46.4	
0.53	0.16	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
342	80.8	42.5	40.2	36.1	152	35.3	19.9	23.8	23.8	

Good λ : $\lambda \approx \sigma \sqrt{2(\log p)/n}$ in row 2

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

The **sparse Riesz condition** (SRC): for all $|A| \leq d^*$
 $1/\gamma < c_* \leq c_{\min}(\mathbf{X}'_A \mathbf{X}_A/n) \leq c_{\max}(\mathbf{X}'_A \mathbf{X}_A/n) \leq c^*$

Theorem: Suppose $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$. Let $\lambda_* = \sigma \sqrt{2(\log p)/n}$
 be the **universal penalty level**. The **sign consistency**

$$P\left\{\text{sgn}(\hat{\beta}(\hat{\lambda})) = \text{sgn}(\beta)\right\} \rightarrow 0$$

holds for the MC+ at a random $\hat{\lambda}$ under the conditions:

- (i) The SRC with $M_2|A^o| + 1 \leq d^*$ ($p \gg n$ allowed)
- (ii) $\beta_* \geq (\gamma M_1 + \sqrt{\gamma})\lambda_*$ (\asymp **information limit**)
- (iii) $P(\lambda_* \leq \hat{\lambda} \leq M_1\lambda_*) \rightarrow 1$ (estimated λ allowed),
 where M_1 and M_2 are constants depend on $\{c_*, c^*, \gamma\}$ only.

- The **bias** of the LASSO interferes with selection consistency
- **MC+** = **MCP & PLUS** for the sparsest local minimizer of the penalized loss, unbiased selection & fast computation of solution path
- Good λ for variable/feature selection: **universal penalty** $\sigma\sqrt{2(\log p)/n}$
- **Selection consistency** proved under the SRC
- The **information limit** for selection consistency attained up to a constant factor
- More ...

THE PROBLEM

 ℓ_1 SELECTORSIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

THE PROBLEM

ℓ_1 SELECTOR

SIMULATION
RESULTS

MC+

SELECTION
CONSISTENCY

CONCLUSION

Thanks!