

Fast Sparse Regression and Classification

Jerome H. Friedman

Stanford University

PREDICTION (Regression/Classification)

y = outcome/response variable

$\mathbf{x} = \{x_1, \dots, x_n\}$ predictors

Goal: $\hat{y} = F(\mathbf{x})$

Want good $F(\mathbf{x})$

LINEAR MODEL

$$F(\mathbf{x}; \mathbf{a}) = a_0 + \sum_{j=1}^n a_j x_j$$

a_0 = intercept

$\{a_j\}_1^n$ = coefficients

ACCURACY

Cost for error: $L(y, F)$

$$L(y, F) = |y - F|, (y - F)^2, \quad y \in R$$

$y \in \{-1, 1\}$:

$$L(y, F) = \log(1 + e^{-yF}) \quad \text{logistic reg.}$$

$$L(y, F) = (1 - yF)_+ \quad \text{SVM}$$

Many many more

Any log-likelihood $\Rightarrow L(y, F)$

PREDICTION RISK

$$R(\mathbf{a}) = E_{\mathbf{x},y}L(y, F(\mathbf{x}; \mathbf{a}))$$

Optimal solution: $\mathbf{a}^* = \arg \min_{\mathbf{a}} R(\mathbf{a})$

$p(\mathbf{x}, y)$ unknown $\Rightarrow \mathbf{a}^*$ unknown

STATISTICAL (MACHINE) LEARNING

Training data: $\{y_i, \mathbf{x}_i\}_1^N \sim p(\mathbf{x}, y)$

$$\hat{R}(\mathbf{a}) = \frac{1}{N} \sum_{i=1}^N L \left(y_i, a_0 + \sum_{j=1}^n a_j x_{ij} \right)$$

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a})$$

If not $N \gg n$, not very good!

$$R(\hat{\mathbf{a}}) \gg R(\mathbf{a}^*) \quad (\text{high variance})$$

REGULARIZATION (biased learning)

$$\hat{\mathbf{a}}(t) = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a}) \quad \text{s.t.} \quad P(\mathbf{a}) \leq t$$

$P(\mathbf{a}) \geq 0$ constraining function

$t \geq P(\hat{\mathbf{a}})$: no constraint \Rightarrow no bias / max. variance

$t = 0$: max. constraint \Rightarrow max. bias / min. variance

$0 < t < P(\hat{\mathbf{a}}) \Rightarrow$ bias–variance trade–off

EQUIVALENT FORMULATION

$$\hat{\mathbf{a}}(\lambda) = \arg \min_{\mathbf{a}} [\hat{R}(\mathbf{a}) + \lambda \cdot P(\mathbf{a})]$$

Here $P(\mathbf{a}) =$ “penalty”

$$\infty \leq \lambda \leq 0 \sim 0 \leq t \leq P(\hat{\mathbf{a}})$$

$\hat{\mathbf{a}}(\lambda) \sim$ 1-dim. path of solutions $\in S^{n+1}$

$S^{n+1} =$ parameter space

MODEL SELECTION (λ)

$$\lambda^* = \arg \min_{0 \leq \lambda \leq \infty} R(\hat{\mathbf{a}}(\lambda))$$

Model selection criterion $\tilde{R}(\mathbf{a})$:

surrogate for $R(\mathbf{a})$ computed from $\{y_i, \mathbf{x}_i\}_1^N$

$$\hat{\lambda} = \arg \min_{0 \leq \lambda \leq \infty} \tilde{R}(\hat{\mathbf{a}}(\lambda))$$

$\hat{\mathbf{a}}(\hat{\lambda}) =$ selected model

$\tilde{R}(\mathbf{a})$ depends on $L(y, F)$ & $P(\mathbf{a})$

Cross-validation: any $L(y, F)$ & $P(\mathbf{a})$

PENALTY SELECTION

\mathbf{a}^* = point in S^{n+1}

Choose penalty that induces paths that come close to \mathbf{a}^*

$$\min_{\lambda} |R(\hat{\mathbf{a}}(\lambda)) - R(\mathbf{a}^*)| = \text{small}$$

Depends on \mathbf{a}^*

Choose $P(\mathbf{a})$ based on knowledge of \mathbf{a}^*

SPARSITY

Fraction of non influential variables

$$\text{sparsity}(\mathbf{a}) = S(\mathbf{a}) = \#(|a_j| = \text{small})/n$$

$$\text{Assumption: } \hat{\mathbf{a}} \simeq \mathbf{a}^* \Rightarrow S(\hat{\mathbf{a}}) \simeq S(\mathbf{a}^*)$$

Choose $P(\mathbf{a})$ s.t. $S(\hat{\mathbf{a}}(\lambda^*)) \simeq S(\mathbf{a}^*)$

Don't know $S(\mathbf{a}^*)$?

Family of penalties $P_\gamma(\mathbf{a})$: γ regulates $S(\hat{\mathbf{a}})$

bridging sparse \rightarrow dense solutions

Model selection to jointly estimate (γ, λ)

("bridge regression": Frank & Friedman 1993)

POWER FAMILY

$$P_\gamma(\mathbf{a}) = \sum_{j=1}^n |a_j|^\gamma$$

With $L(y, F) = (y - F)^2$:

$\gamma = 2$: ridge-regression (dense)

$\gamma = 1$: lasso (moderately sparse)

$\gamma = 0$: (all) subsets selection (sparsest)

$0 \leq \gamma \leq 2$ bridges subset \rightarrow ridge

Note: $\gamma \geq 1 \Rightarrow$ convex, $\gamma < 1 \Rightarrow$ non convex

Elastic Net (Zou & Hastie 2005)

$$P_{\beta}(\mathbf{a}) = \sum_{j=1}^n (\beta - 1) a_j^2 / 2 + (2 - \beta) |a_j|$$

$1 \leq \beta \leq 2$ (bridges lasso \rightarrow ridge):

convex: moderately sparse \rightarrow dense

Generalized Elastic Net (non convex)

$$P_{\beta}(\mathbf{a}) = \sum_{j=1}^n \log((1 - \beta) |a_j| + \beta)$$

$0 \leq \beta < 1$ (bridges subset selection \rightarrow lasso):

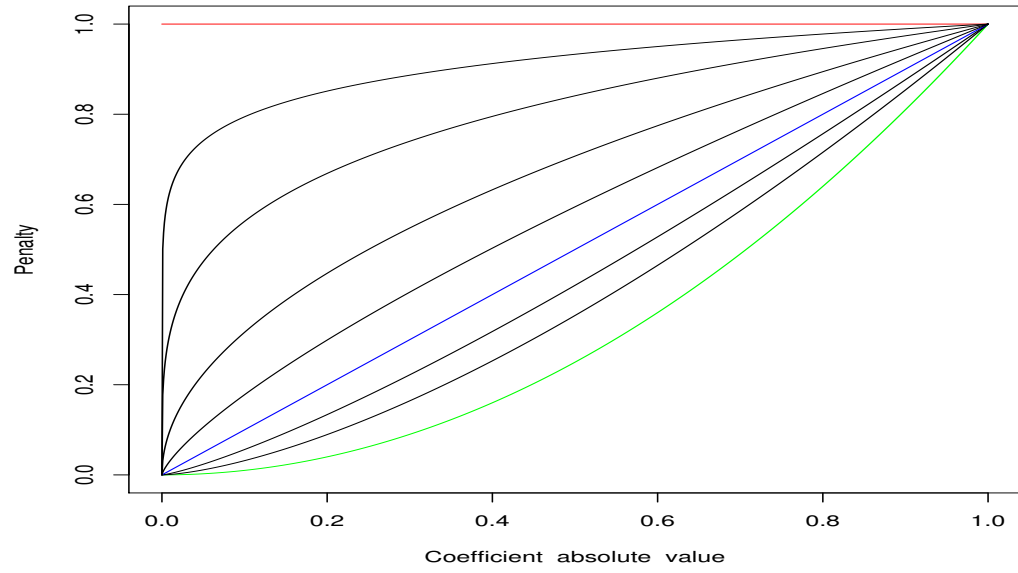
very sparse \rightarrow moderately sparse

non + convex: $0 \leq \beta \leq 2$ bridges very sparse \rightarrow dense

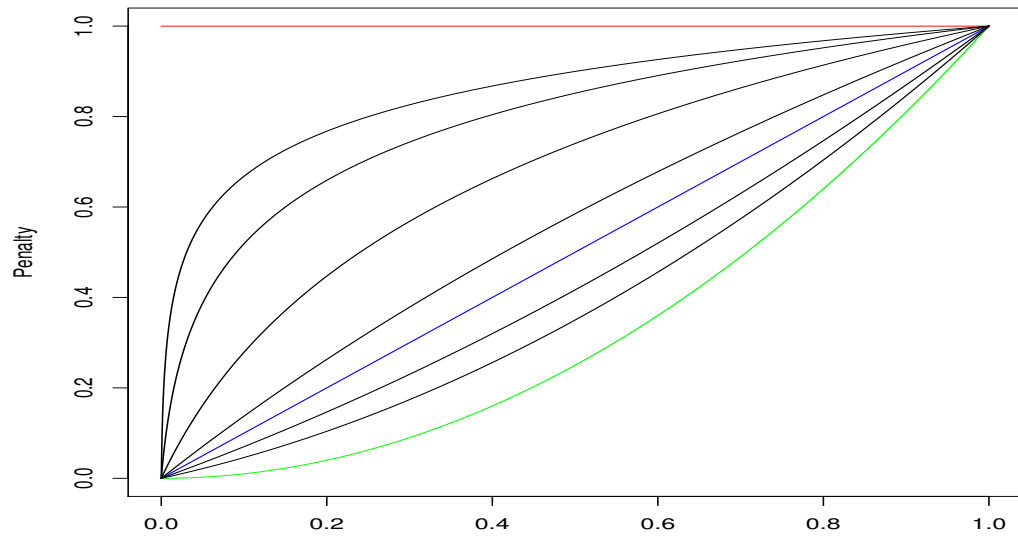
Better statistical & computational properties

Method works for both + many more

Power family



Generalized elastic net



BRIDGE REGRESSION

(1) Repeatedly solve:

$$\hat{\mathbf{a}}_{\beta}(\lambda) = \arg \min_{\mathbf{a}} [\hat{R}(\mathbf{a}) + \lambda \cdot P_{\beta}(\mathbf{a})]$$

$$0 \leq \beta \leq 2, \quad 0 \leq \lambda \leq \infty$$

(2) $(\hat{\beta}, \hat{\lambda}) \leftarrow$ model selection criterion

(3) $\hat{\mathbf{a}}_{\hat{\beta}}(\hat{\lambda}) =$ solution

Big challenge: fast enough algorithm for (1)

Especially for $P_{\beta}(\mathbf{a}) =$ non convex

DIRECT PATH SEEKING

Goal: directly produce path \simeq given $P(\mathbf{a})$

without repeatedly optimizing

$\nu \geq 0$: path length; $\Delta\nu > 0$: increment

$\mathbf{d}(\nu)$ = direction in parameter space

Algorithm

Initialize: $\nu = 0$; $\hat{\mathbf{a}}(0) = 0$

Loop {

$$\hat{\mathbf{a}}(\nu + \Delta\nu) = \hat{\mathbf{a}}(\nu) + \mathbf{d}(\nu) \cdot \Delta\nu$$

$$\nu \leftarrow \nu + \Delta\nu$$

}

Until ($\hat{R}(\hat{\mathbf{a}}(\nu)) = \min$; $\nu = \nu_{\max}$)

Numerical optimization: $\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a})$

Choose $\mathbf{d}(\nu)$ & $\Delta\nu$ such that

$$\hat{\mathbf{a}}(\nu) : (0 \leq \nu \leq \nu_{\max})$$

$$\simeq \hat{\mathbf{a}}(\lambda) = \arg \min_{\mathbf{a}} [\hat{R}(\mathbf{a}) + \lambda \cdot P(\mathbf{a})]$$

$$(\infty \geq \lambda \geq 0)$$

$(\mathbf{d}(\nu), \Delta\nu)$ depend on $L(y, F)$ & $P(\mathbf{a})$

Multiple numerical optimizations replaced by

single numerical optimization

EXAMPLES

$$L(y, F) = (y - F)^2:$$

PLS \simeq ridge-regression ($\beta = 2$)

LAR \simeq lasso ($\beta = 1$)

Stepwise \simeq all-subsets ($\beta = 0$)

Extend to more $L(y, F)$ and/or $P(\mathbf{a})$

Bridge regression: $P_\beta(\mathbf{a})$ ($0 \leq \beta \leq 2$)

Generalized Path Seeking (GPS)

Fast algorithm for:

(1) any convex $L(y, F)$

(2) any $P(\mathbf{a})$ s.t. $\frac{\partial P(\mathbf{a})}{\partial |a_j|} \geq 0$

i.e. $P(\mathbf{a})$ monotone $\uparrow |a_j|$

Definitions

$\nu \geq 0$: path length

$\Delta\nu > 0$: small increment

$$g_j(\nu) = - \left[\frac{\partial \hat{R}(\mathbf{a})}{\partial a_j} \right]_{\mathbf{a}=\hat{\mathbf{a}}(\nu)} \quad (\text{loss} + \text{data})$$

$$p_j(\nu) = \left[\frac{\partial P(\mathbf{a})}{\partial |a_j|} \right]_{\mathbf{a}=\hat{\mathbf{a}}(\nu)} \quad (\text{penalty})$$

$$\lambda_j(\nu) = g_j(\nu) / p_j(\nu) \quad (\text{combination})$$

- 1 Initialize: $\nu = 0; \{\hat{a}_j(0) = 0\}_1^n$
- 2 Loop {
- 3 Compute $\{\lambda_j(\nu)\}_1^n$
- 4 $S = \{j \mid \lambda_j(\nu) \cdot \hat{a}_j(\nu) < 0\}$
- 5 if ($S = \text{empty}$) $j^* = \arg \max_j |\lambda_j(\nu)|$
- 6 else $j^* = \arg \max_{j \in S} |\lambda_j(\nu)|$
- 7 $\hat{a}_{j^*}(\nu + \Delta\nu) = \hat{a}_{j^*}(\nu) + \Delta\nu \cdot \text{sign}(\lambda_{j^*}(\nu))$
- 8 $\{\hat{a}_j(\nu + \Delta\nu) = \hat{a}_j(\nu)\}_{j \neq j^*}$
- 9 $\nu \leftarrow \nu + \Delta\nu$
- 10 } Until $\lambda(\nu) = 0$ ($\nu = \nu_{\max}$)

THEOREM

$\hat{\mathbf{a}}(\lambda) = \text{exact path } (\infty \geq \lambda \geq 0)$

$\hat{\mathbf{a}}(\nu) = \text{GPS path } (0 \leq \nu \leq \nu_{\max})$

If for all $\lambda > \lambda_0$

all $\{\hat{a}_j(\lambda)\}_1^n$ are continuous and monotone

Then for all $\lambda > \lambda_0$

$\hat{\mathbf{a}}(\nu) = \hat{\mathbf{a}}(\lambda)$, as $\Delta\nu \rightarrow 0$

i.e. GPS produces exact path

Otherwise: $\hat{a}_j(\nu) \simeq \hat{a}_j(\lambda)$

When $\hat{a}_j(\lambda)$ becomes non monotone:

$\hat{a}_j(\nu)$ tends to slightly delay becoming non monotone

(smoother more stable paths)

When $\hat{a}_j(\lambda)$ discontinuous ($\gamma < 1, \beta < 1/2$):

$\hat{a}_j(\nu) =$ continuous

\sim interpolates between $\hat{a}_j(\lambda)$ discontinuities

(smoother more stable paths)

Logistic Regression / Classification

South African heart transplant data

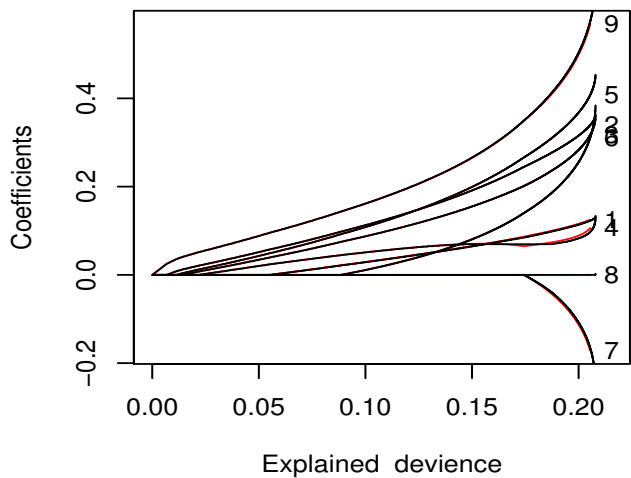
$y \in \{1, -1\} = \{\text{success, failure}\}$

$$L(y, F) = \log(1 + e^{-yF})$$

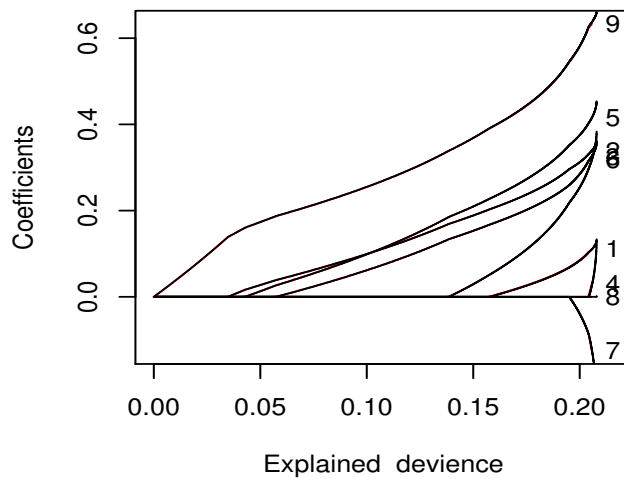
$$n = 9, \quad N = 462$$

red = exact (convex), black = GPS

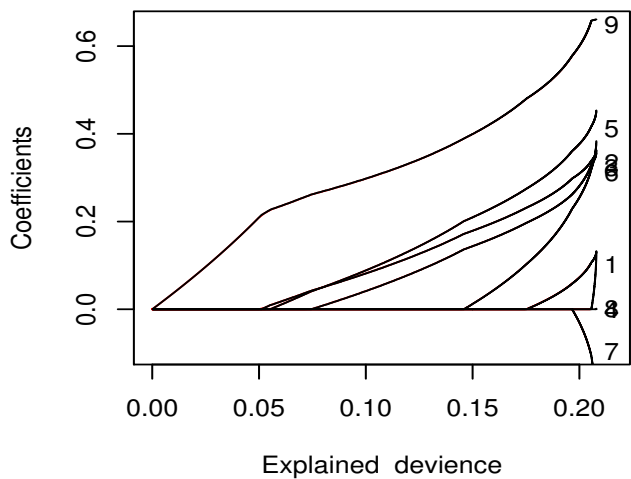
Elastic Net 1.9



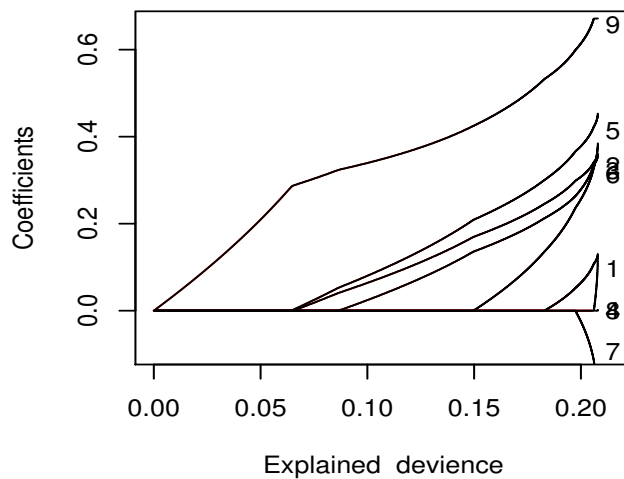
Elastic Net 1.5



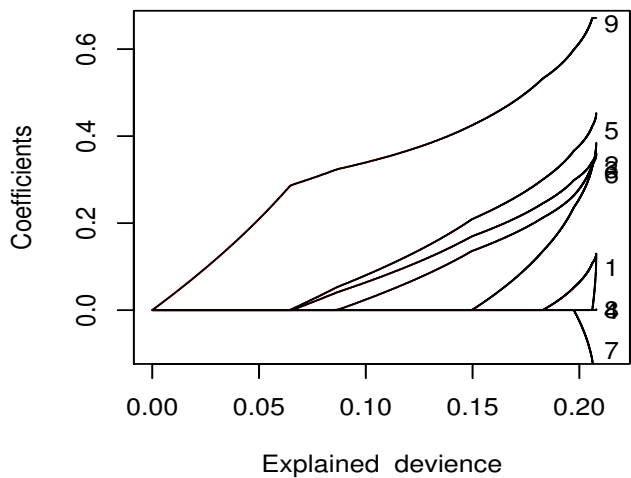
Elastic Net 1.25



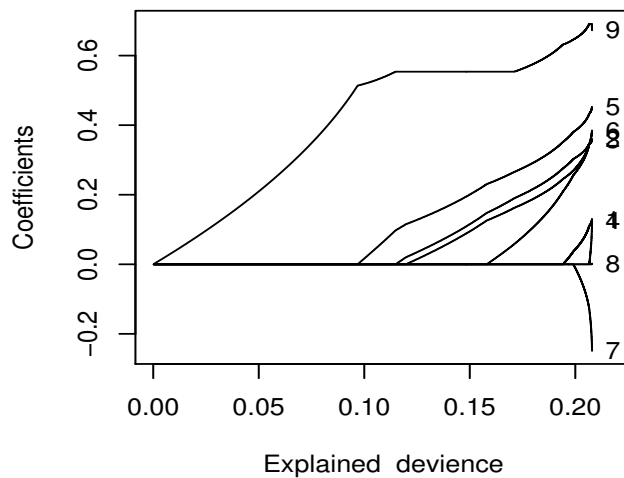
Lasso



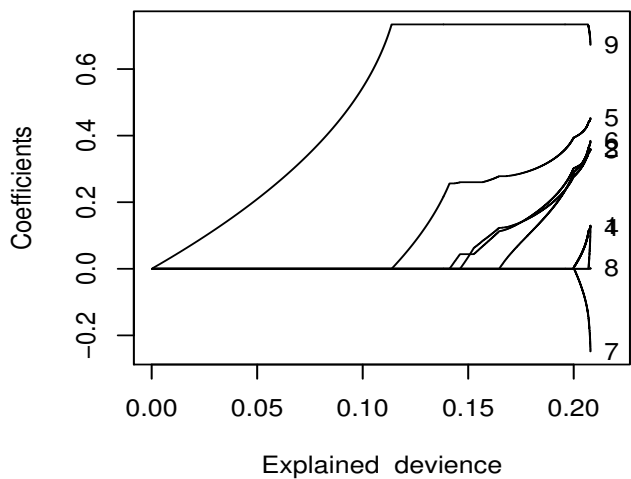
Lasso



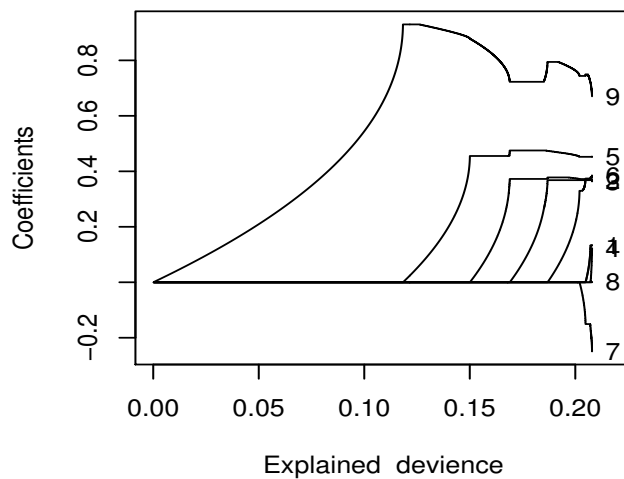
Elastic Net 0.5



Elastic Net 0.25



Elastic Net 0.0



Regression: under-determined example

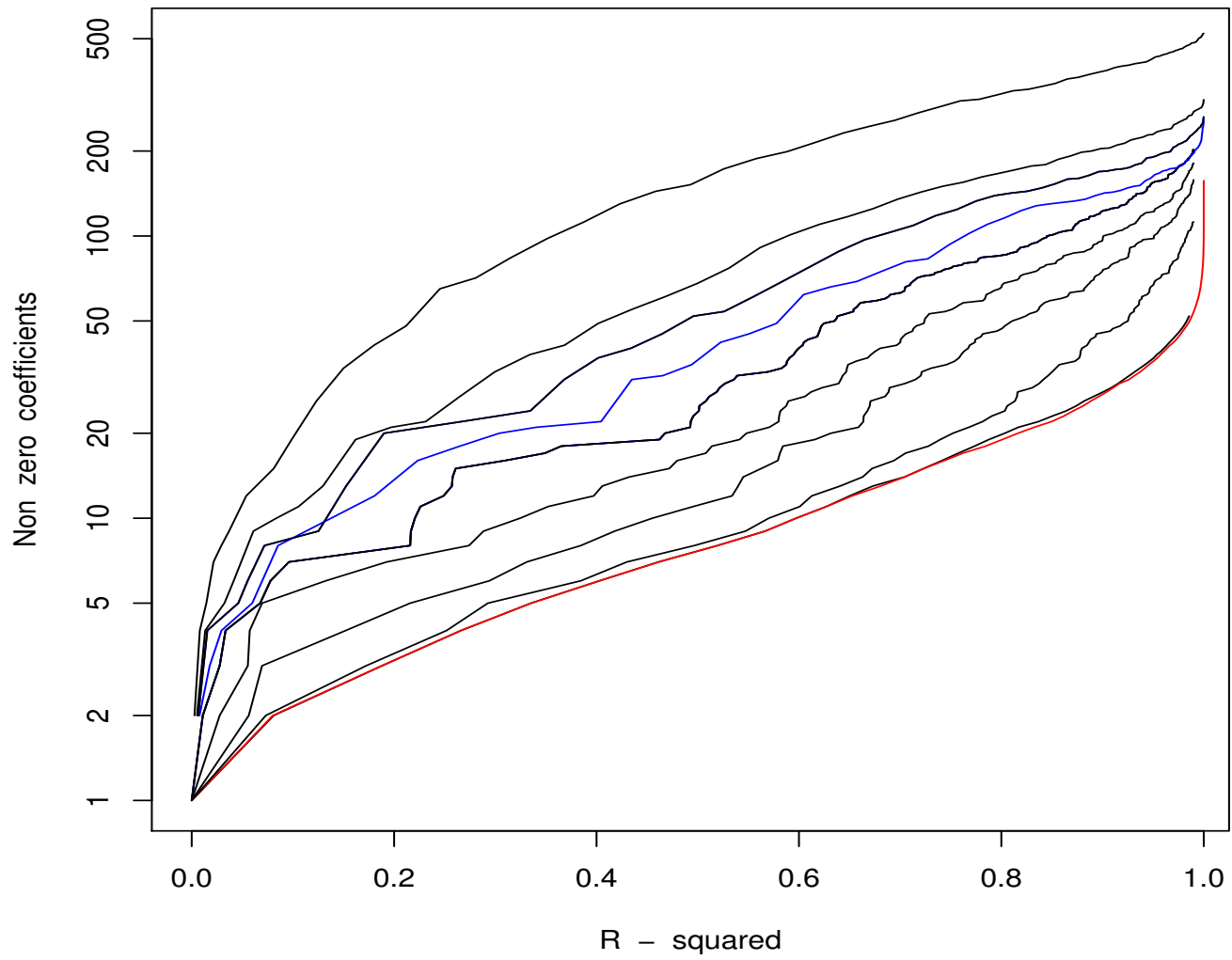
$$n = 10000, \quad N = 200$$

$$\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{C}); \quad C_{jj} = 1, \quad C_{jk} = 0.4$$

$$y_i = \sum_{j=1}^n a_j^* x_{ij} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2); \quad \sigma \sim 3/1 \text{ signal/noise}$$

$$|a_j^*| = [31 - j]_+, \quad \text{sign}(a_{j+1}^*) = -\text{sign}(a_j^*)$$



$\beta \in \{1.9, 1.7, 1.5, 1.0 \text{ (lasso, blue)}, 0.5, 0.3, 0.2, 0.1,$

THEREFORE

$P_\beta(\mathbf{a}) =$ generalized elastic net

$\beta \downarrow \Rightarrow S(\hat{\mathbf{a}}) \uparrow$ monotonically

at all path points

Penalty Selection (β)

Regression: under-determined example

$$n = 10000, \quad N = 200$$

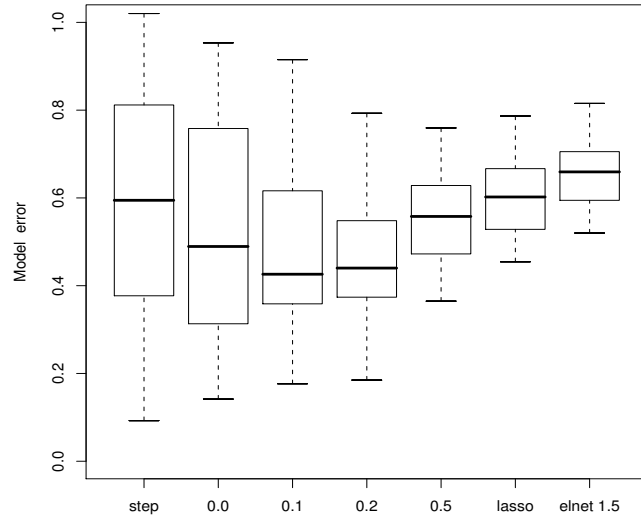
50 data sets $\sim p(\mathbf{x}, y)$

Dist. of pred. risk at optimal path point \mathbf{a}^*

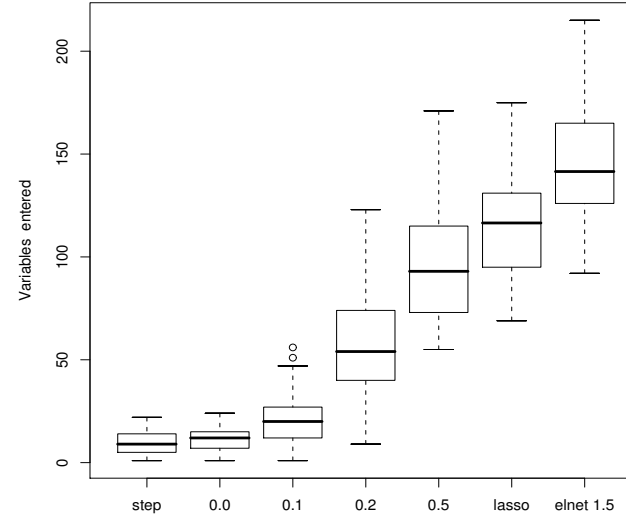
Methods: GPS $\beta \in \{0.0, 0.1, 0.2, 0.5\}$

forward stepwise, lasso, elastic net (1.5)

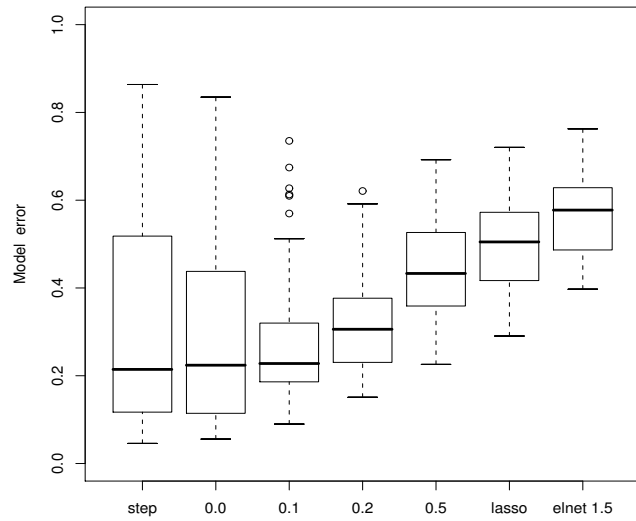
Corr = 0.4



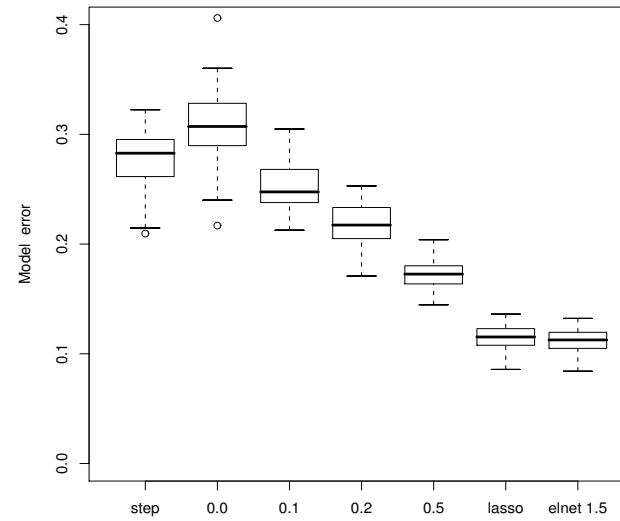
Corr = 0.4



Corr = 0.0



Corr = 0.4, a > 0



CONCLUSIONS

(1) $n \gtrsim N \Rightarrow$ regularization essential

(2) best penalty depends on

sparsity $\{a_j^*\}$, $\{sign(a_j^*)\}$ and \mathbf{x} – distribution

(2) need bridge regression to chose

best penalty β & path point λ

(3) results same for logistic regression

Post-processing Selectors

$$(1) \quad \tilde{\mathbf{a}}(\lambda) = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a}) + \lambda P(\mathbf{a})$$

$$P(\mathbf{a}) = \text{convex (lasso)}$$

$$(2) \quad A(\lambda) = \{j\}_{\tilde{a}_j(\lambda) \neq 0} \quad (\text{active variables at } \lambda)$$

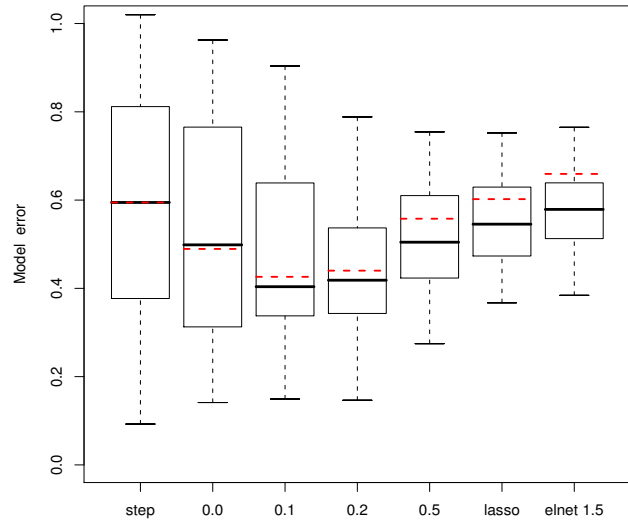
$$(3) \quad \hat{\mathbf{a}}(\lambda) = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a}) \quad \text{s.t.} \quad \{a_j = 0\}_{j \notin A(\lambda)}$$

Intuition (sparse problems):

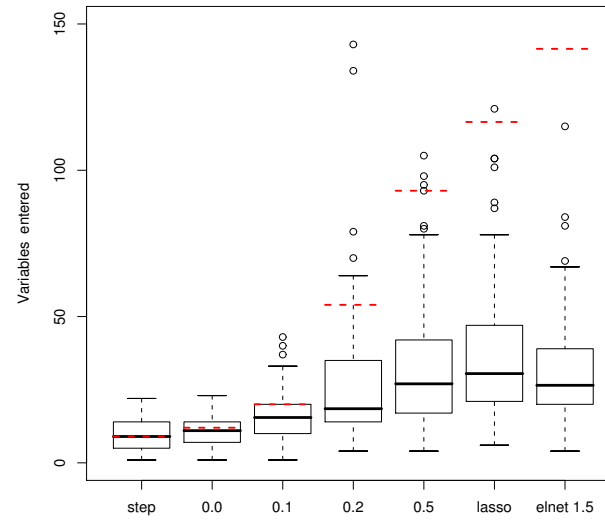
$\tilde{\mathbf{a}}(\lambda) \simeq$ selects correct variables

but over shrinks their values

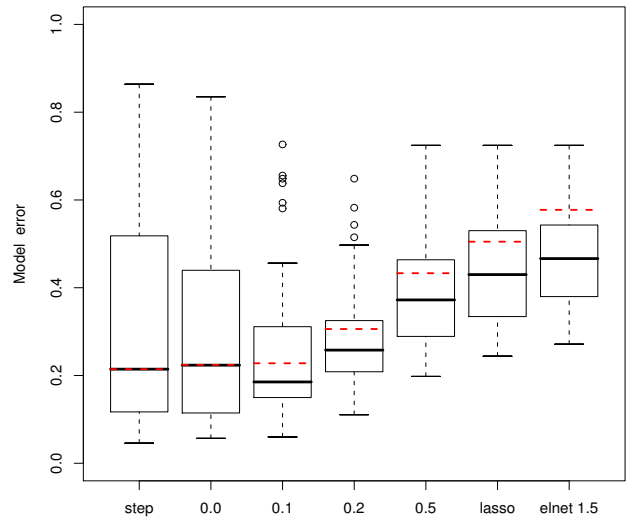
Corr = 0.4, Post



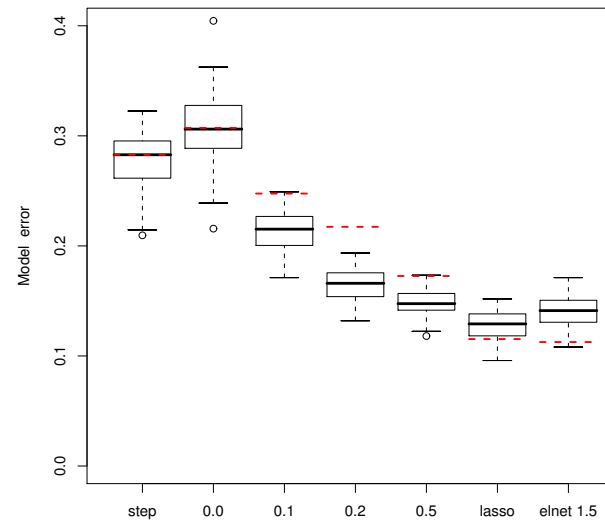
Corr = 0.4, Post



Corr = 0, Post



Corr = 0.4, a > 0, Post



CONCLUSIONS

(1) when sparse non convex $P(\mathbf{a})$ is best:

better variable *selection* & shrinkage

(2) best direct methods \rightarrow best variable selectors

(3) results same for logistic regression

Generalized Path Seeking

For same $L(y, F)$ & $P(\mathbf{a})$:

paths close to exact solutions

same sparseness properties

Can be applied with:

non convex $P(\mathbf{a}) \Rightarrow$ sparser than lasso

Used as variable selector \rightarrow further improvement

Multinomial regression

TALK

<http://www-stat.stanford.edu/~jhf/talks/lyon1.pdf>

PAPER

<http://www-stat.stanford.edu/~jhf/ftp/GPSpaper.pdf>

R – INTERFACE

<http://www-stat.stanford.edu/~jhf/R-GPS.html>