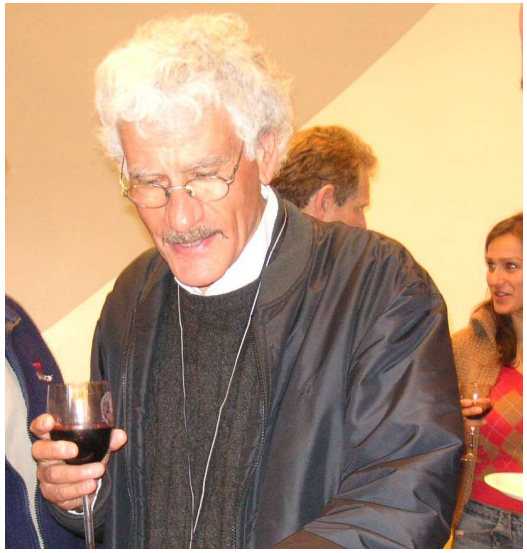
Learning with Sparsity Constraints

Trevor Hastie

Stanford University

recent joint work with Rahul Mazumder, Jerome Friedman and Rob Tibshirani




earlier work with Brad Efron, Iain Johnstone, Ji Zhu, Saharon Rosset, Hui Zou and Mee-Young Park

Linear Models in Data Mining

As datasets grow *wide*—i.e. many more predictors than samples—linear models have regained popularity.

Document classification: bag-of-words can leads to $p = 20K$ features and $N = 5K$ document samples.

Image deblurring, classification: $p = 65K$ pixels are features, $N = 100$ samples.

Genomics, microarray studies: $p = 40K$ genes are measured for each of $N = 100$ subjects.

Genome-wide association studies: $p = 500K$ SNPs measured for $N = 2000$ case-control subjects.

In all of these we use linear models — e.g. linear regression, logistic regression, Cox model. Since $p \gg N$, we have to regularize.

Forms of Regularization

We cannot fit linear models with $p > N$ without some constraints. Common approaches are

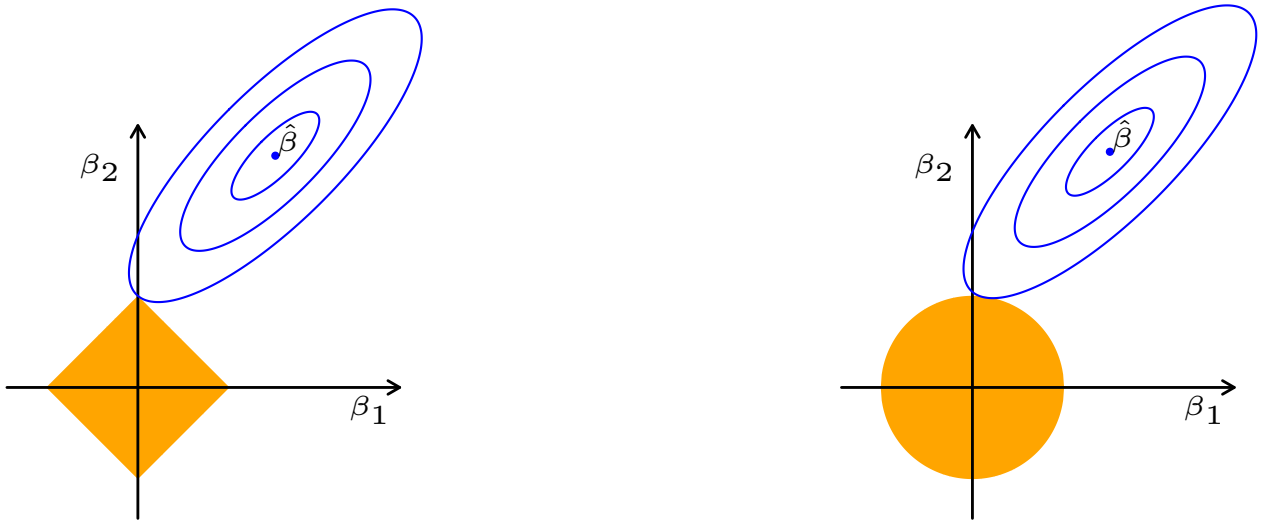
Forward stepwise adds variables one at a time and stops when overfitting is detected. This is a *greedy* algorithm, since the model with say 5 variables is not necessarily the best model of size 5.

Best-subset regression finds the subset of each size k that fits the model the best. Only feasible for small p around 35.

Ridge regression fits the model subject to constraint $\sum_{j=1}^{p} \beta_j^2 \leq t$. Shrinks coefficients toward zero, and hence controls variance. Allows linear models with arbitrary size p to be fit, although coefficients always in row-space of X.

Lasso regression fits the model subject to constraint $\sum_{j=1}^{p} |\beta_j| \leq t$.

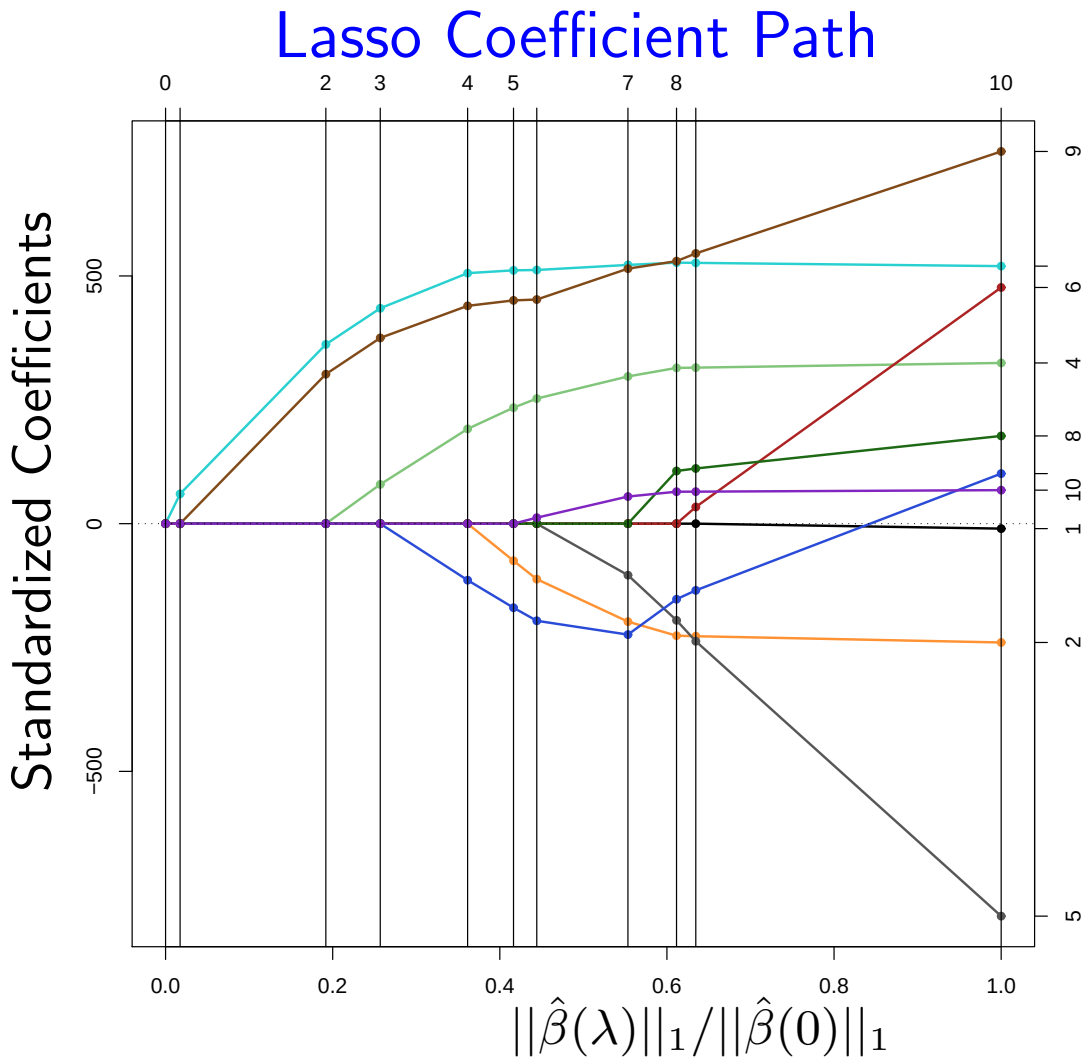
Lasso does variable selection and shrinkage, while ridge only shrinks.



Brief History of ℓ_1 Regularization

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

- Wavelet *Soft Thresholding* (Donoho and Johnstone 1994) in orthonormal setting.
- Tibshirani introduces *Lasso* for regression in 1995.
- Same idea used in *Basis Pursuit* (Chen, Donoho and Saunders 1996).
- Extended to many linear-model settings e.g. Survival models (Tibshirani, 1997), logistic regression, and so on.
- Gives rise to a new field *Compressed Sensing* (Donoho 2004, Candes and Tao 2005)—near exact recovery of sparse signals in very high dimensions. In many cases ℓ_1 a good surrogate for ℓ_0.



Lasso: $\hat{\beta}(\lambda) = \text{argmin}_\beta \sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda||\beta||_1$

History of Path Algorithms

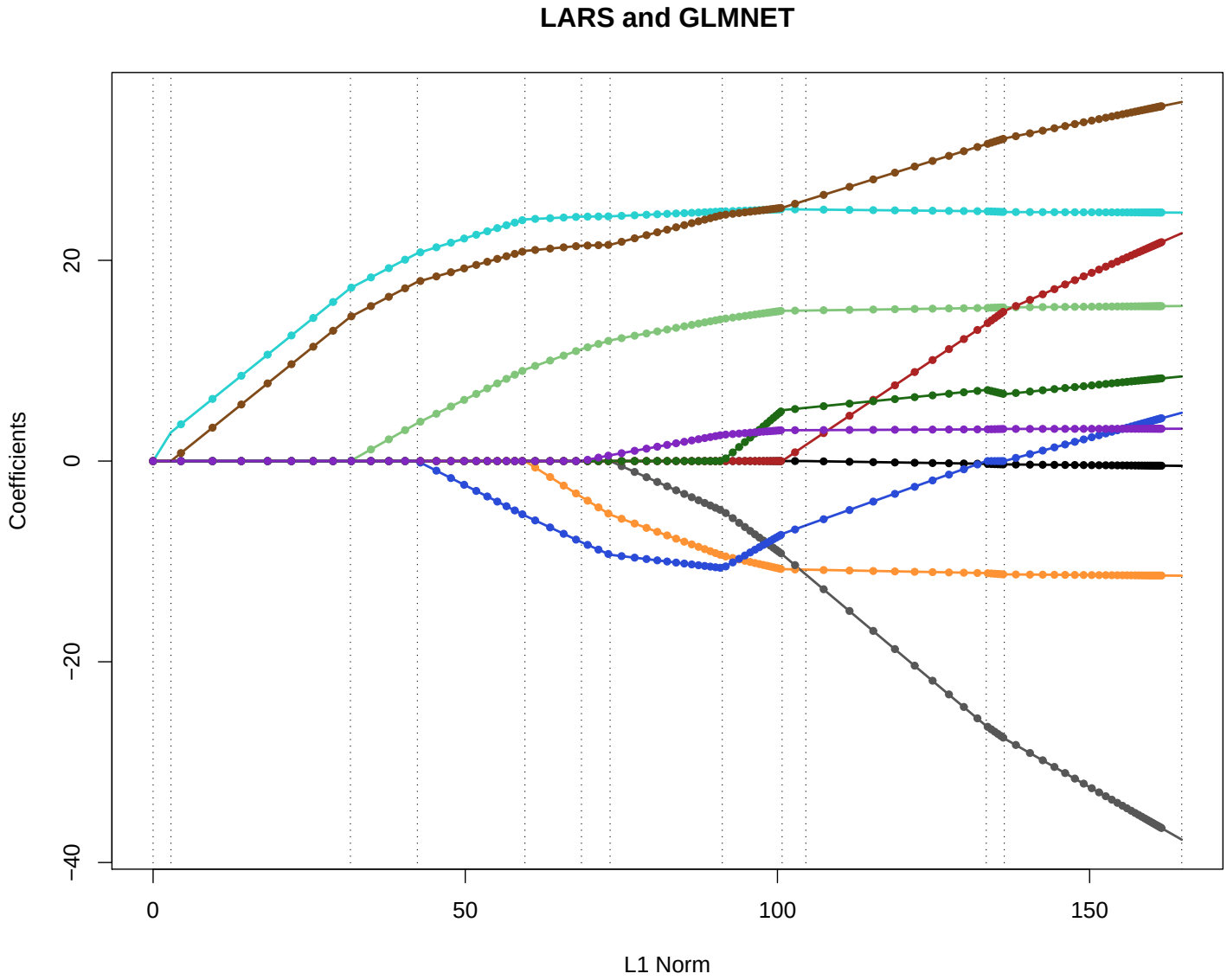
Efficient path algorithms for $\hat{\beta}(\lambda)$ allow for easy and exact cross-validation and model selection.

- In 2001 the LARS algorithm (Efron et al) provides a way to compute the entire lasso coefficient path efficiently at the cost of a full least-squares fit.
- 2001 – present: path algorithms pop up for a wide variety of related problems: Grouped lasso (Yuan & Lin 2006), support-vector machine (Hastie, Rosset, Tibshirani & Zhu 2004), elastic net (Zou & Hastie 2004), quantile regression (Li & Zhu, 2007), logistic regression and glms (Park & Hastie, 2007), Dantzig selector (James & Radchenko 2008), ...
- Many of these do not enjoy the piecewise-linearity of LARS, and seize up on very large problems.

Coordinate Descent

- Solve the lasso problem by coordinate descent: optimize each parameter separately, holding all the others fixed. Updates are trivial. Cycle around till coefficients stabilize.
- Do this on a grid of λ values, from λ_{max} down to λ_{min} (uniform on log scale), using warms starts.
- Can do this with a variety of loss functions and additive penalties.

Coordinate descent achieves dramatic speedups over all competitors, by factors of 10, 100 and more.


LARS and GLMNET
Coefficients
20
0
−20
−40
0
50
100
150
L1 Norm

Speed Trials on Large Datasets

Competitors:

`glmnet` Fortran based R package using coordinate descent. Covers GLMs and Cox model.

`l1logreg` Lasso-logistic regression package by Koh, Kim and Boyd, using state-of-art interior point methods for convex optimization.

`BBR/BMR` Bayesian binomial/multinomial regression package by Genkin, Lewis and Madigan. Also uses coordinate descent to compute posterior mode with Laplace prior—the lasso fit.

Logistic Regression — Real Datasets

Name	Type	N	p	**glmnet**	**l1logreg**	**BBR BMR**
			Dense			
Cancer	14 class	144	16,063	57	NA	2.1 hrs
Leukemia	2 class	72	3571	0.65	55.0	450
			Sparse			
Internet ad	2 class	2359	1430	5.0	20.9	34.7
Newsgroup	2 class	11,314	777,811	28	3.5 hrs	

Timings in seconds (unless stated otherwise). For Cancer, Leukemia and Internet-Ad, times are for ten-fold cross-validation over 100 λ values; for Newsgroup we performed a single run with 100 values of λ, with $\lambda_{min} = 0.05\lambda_{max}$.

A brief history of coordinate descent for the lasso

1997 Tibshirani's student Wenjiang Fu at U. Toronto develops the "shooting algorithm" for the lasso. Tibshirani doesn't fully appreciate it.

A brief history of coordinate descent for the lasso

1997 Tibshirani's student Wenjiang Fu at U. Toronto develops the "shooting algorithm" for the lasso. Tibshirani doesn't fully appreciate it.

2002 Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for the lasso. Hastie implements it, makes an error, and Hastie +Tibshirani conclude that the method doesn't work.

A brief history of coordinate descent for the lasso

1997 Tibshirani's student Wenjiang Fu at U. Toronto develops the "shooting algorithm" for the lasso. Tibshirani doesn't fully appreciate it.

2002 Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for the lasso. Hastie implements it, makes an error, and Hastie +Tibshirani conclude that the method doesn't work.

2006 Friedman is external examiner at PhD oral of Anita van der Kooij (Leiden) who uses coordinate descent for elastic net. Friedman, Hastie + Tibshirani revisit this problem. Others have too — Shevade and Keerthi (2003), Krishnapuram and Hartemink (2005), Genkin, Lewis and Madigan (2007), Wu and Lange (2008), Meier, van de Geer and Buehlmann (2008).

Coordinate descent for the lasso

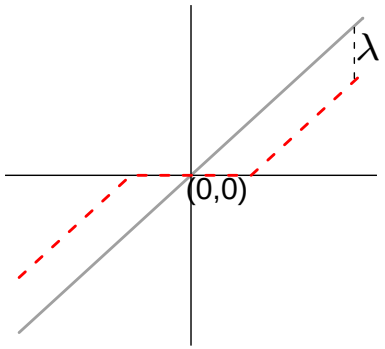
$$\min_\beta \frac{1}{2N}\sum_{i=1}^N (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^p |\beta_j|$$

Suppose the p predictors and response are standardized to have mean zero and variance 1. Initialize all the $\beta_j = 0$.

Cycle over $j = 1, 2, \ldots, p, 1, 2, \ldots$ till convergence:

- Compute the partial residuals $r_{ij} = y_i - \sum_{k\neq j} x_{ik}\beta_k$.
- Compute the simple least squares coefficient of these residuals on jth predictor: $\beta_j^* = \frac{1}{N}\sum_{i=1}^N x_{ij}r_{ij}$
- Update β_j by *soft-thresholding*:

$$\begin{aligned}\beta_j &\leftarrow S(\beta_j^*, \lambda)\\ &= \text{sign}(\beta_j^*)(|\beta_j^*| - \lambda)_+\end{aligned}$$



Elastic-net penalty family

Family of convex penalties proposed in Zou and Hastie (2005) for $p \gg N$ situations, where predictors are correlated in groups.

$$\min_\beta \frac{1}{2N}\sum_{i=1}^N (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p P_\alpha(\beta_j)$$

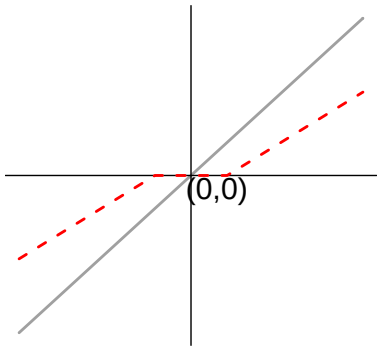
$$\text{with } P_\alpha(\beta_j) = \tfrac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|.$$

α creates a compromise between the *lasso* and *ridge*.

Coordinate update is now

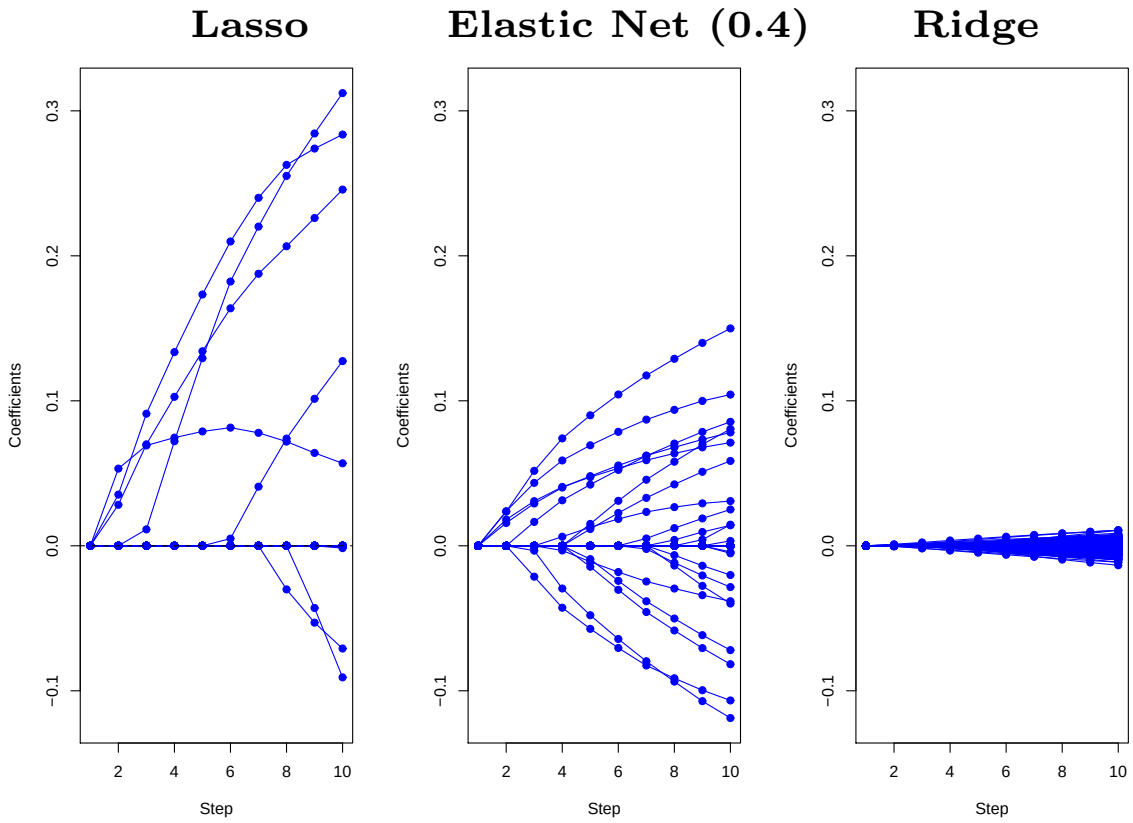
$$\beta_j \leftarrow \frac{S(\beta_j^*,\ \lambda\alpha)}{1+\lambda(1-\alpha)}$$

where $\beta_j^* = \frac{1}{N}\sum_{i=1}^N x_{ij} r_{ij}$ as before.



GLMNET: coordinate descent for elastic net family

Friedman, Hastie and Tibshirani 2008



Leukemia Data, Logistic regression, N=72, p=3571, first 10 steps shown.

GLMNET R package

New version of R GLMNET package includes Gaussian, Poisson, Binomial, Multinomial and Cox models.

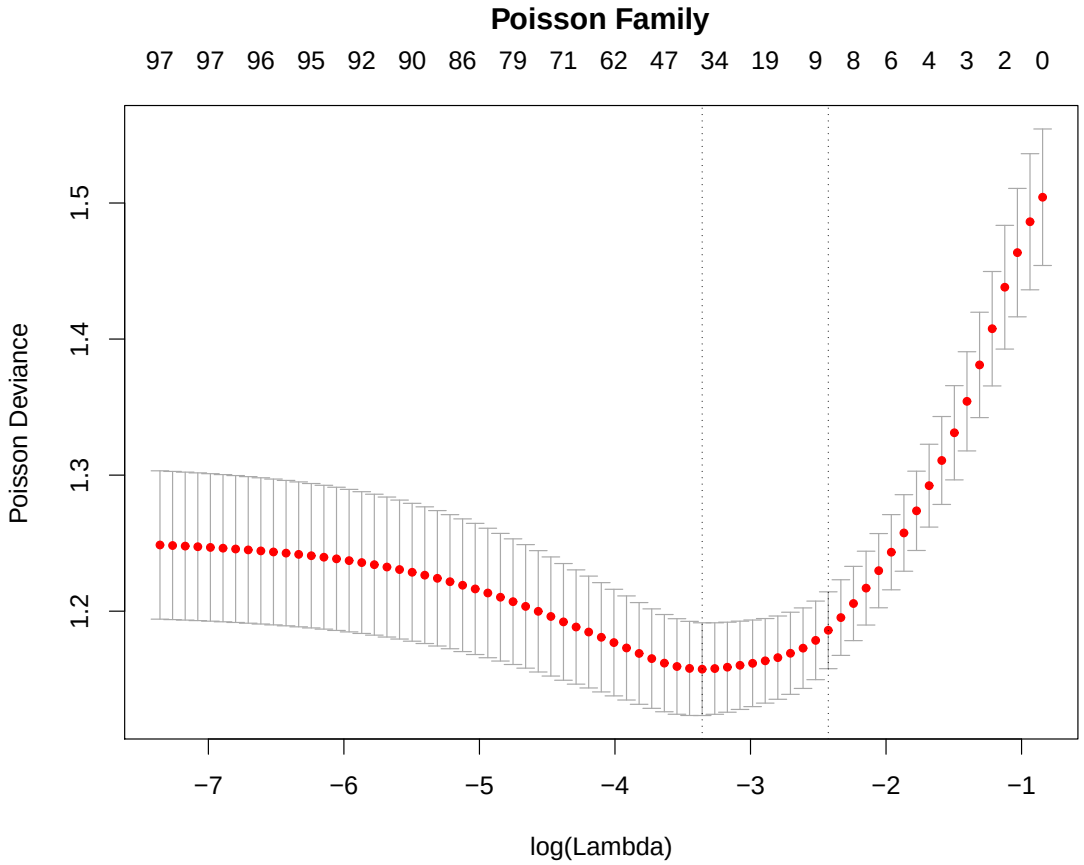
Computes entire regularization path efficiently for lasso — elastic net — ridge penalties.

Has built-in cross-validation functions for selecting tuning parameters.

Can handle very large datasets, and exploit sparsity in X matrix.

Dramatic speedups using *strong-rule* variable screening (Tibshirani et al 2010, Wu et al 2009).

Cross Validation to select λ



K-fold cross-validation is easy and fast. Here K=10, and the true model had 10 out of 100 nonzero coefficients.

Problem with Lasso

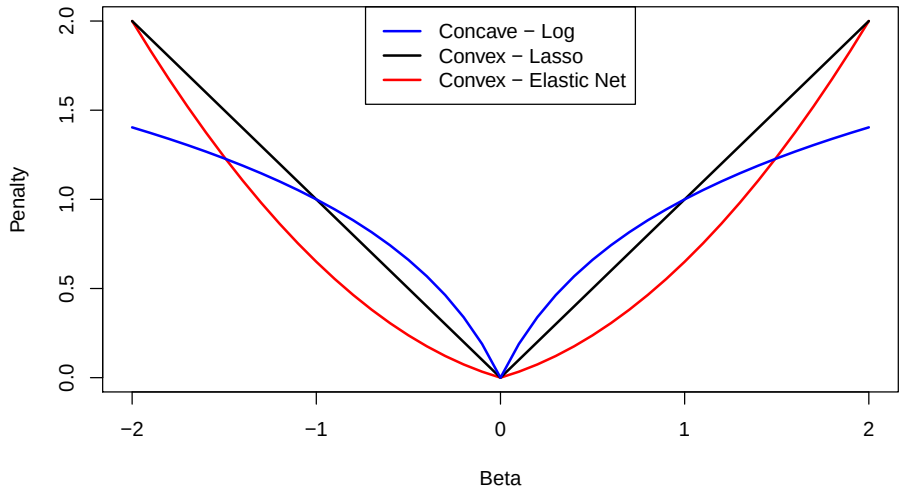
When p is large and the number of relevant variables is small:

- to screen out spurious variables, λ should be large, causing bias in retained variables.
- decreasing λ to reduce bias floods the model with spurious variables.

Many approaches to modify the lasso to address this problem. Here we focus on *concave penalties* (Mazumder et al, JASA 2011 *in press*).

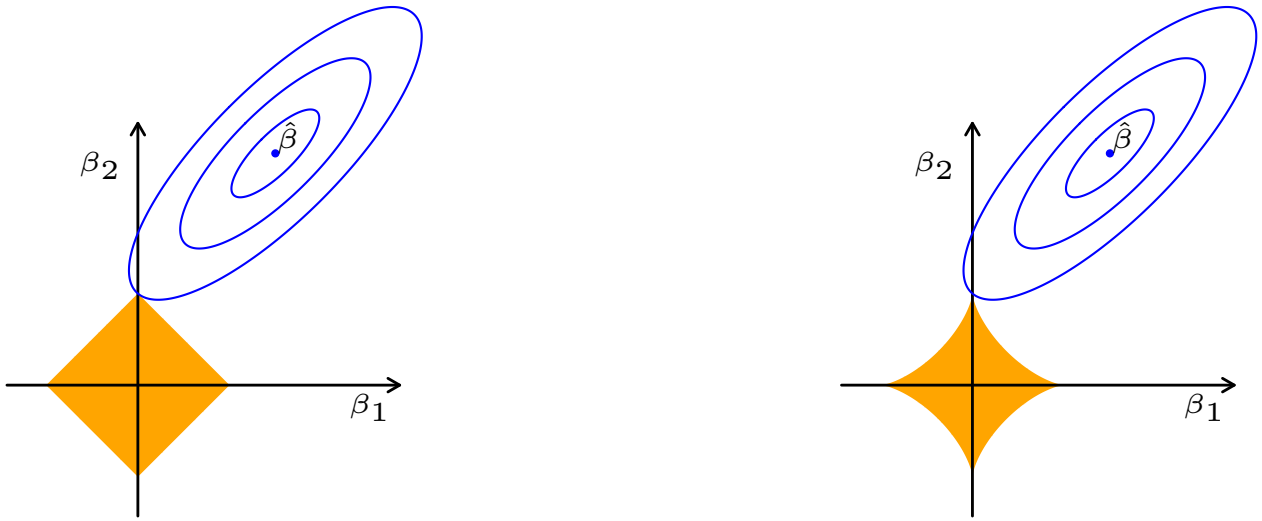
Concave penalties

- Penalize small coefficients more severely than lasso, leading to sparser models.
- Penalize larger coefficients less, and reduce their bias.
- Concavity causes multiple minima and computational issues.



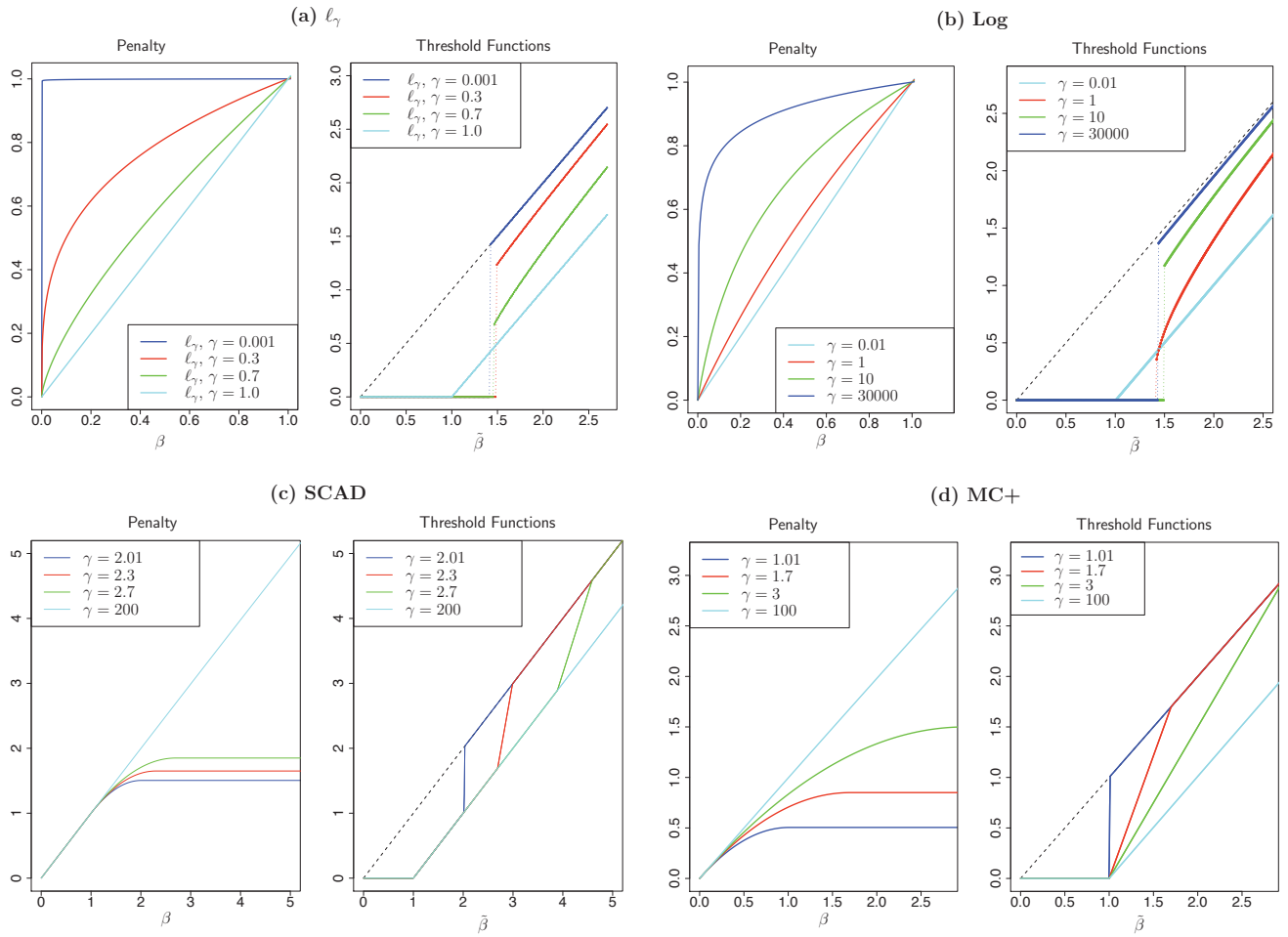
$$\min_\beta \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} P_\gamma(\beta_j)$$

Constraint region for concave penalty

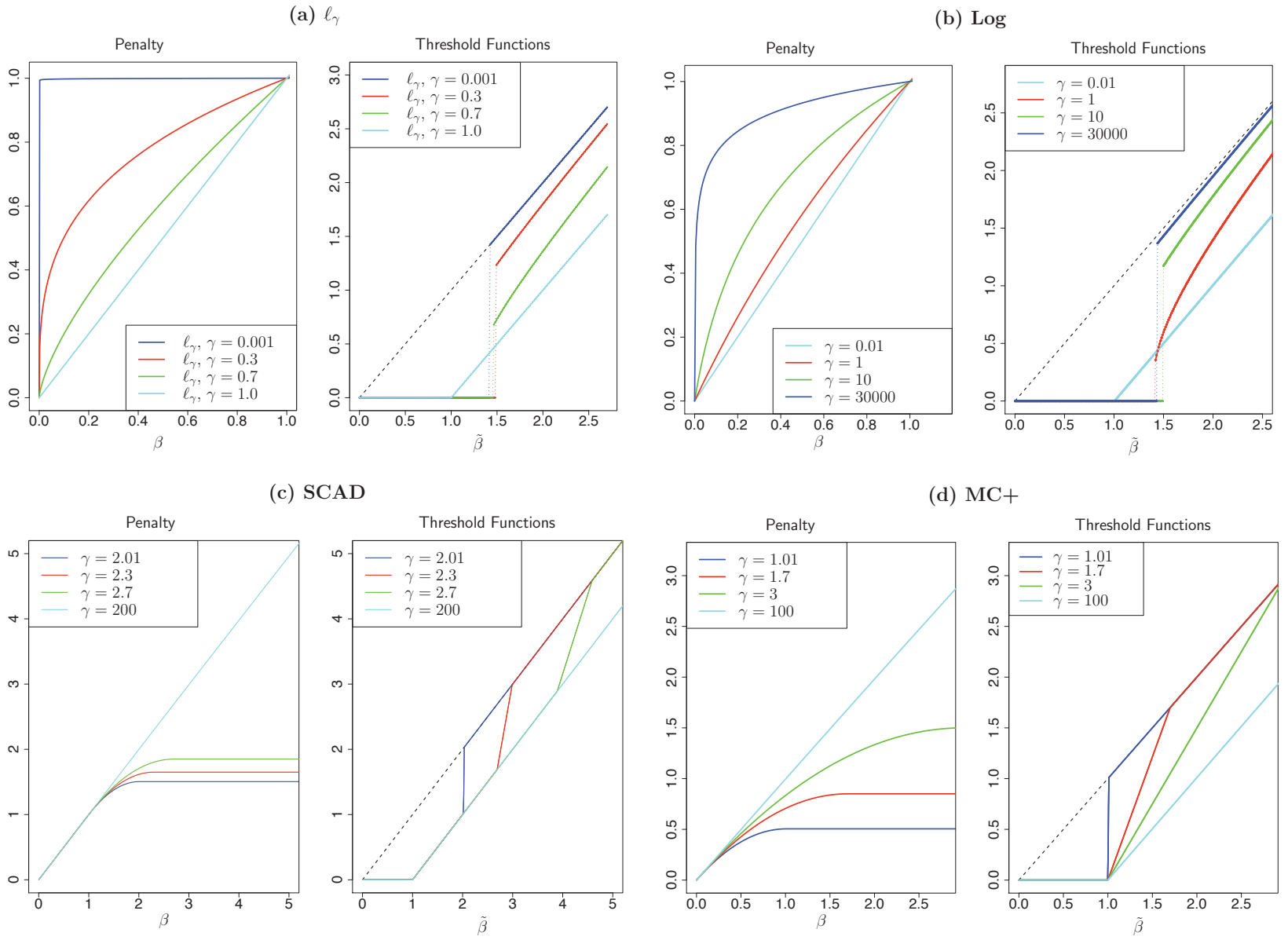


Shown are ℓ_1 (lasso) and ℓ_q penalty $\sum_{j=1}^{p} |\beta_j|^q \leq t$ with $q = 0.7$.
Note that ℓ_0 regularization corresponds to best-subset selection.

Penalty families and threshold functions



(a) Friedman and Frank (1993) (b) Friedman (2008) (c) Fan and Li (2001) (d) Zhang (2010), Zhang&Huang(2008). Since symmetric about zero, only positive side shown.


(a) ℓ_γ
Penalty
Threshold Functions
ℓ_γ, $\gamma = 0.001$
ℓ_γ, $\gamma = 0.3$
ℓ_γ, $\gamma = 0.7$
ℓ_γ, $\gamma = 1.0$
β
$\tilde{\beta}$
(b) Log
Penalty
Threshold Functions
$\gamma = 0.01$
$\gamma = 1$
$\gamma = 10$
$\gamma = 30000$
β
$\tilde{\beta}$
(c) SCAD
Penalty
Threshold Functions
$\gamma = 2.01$
$\gamma = 2.3$
$\gamma = 2.7$
$\gamma = 200$
β
$\tilde{\beta}$
(d) MC+
Penalty
Threshold Functions
$\gamma = 1.01$
$\gamma = 1.7$
$\gamma = 3$
$\gamma = 100$
β
$\tilde{\beta}$

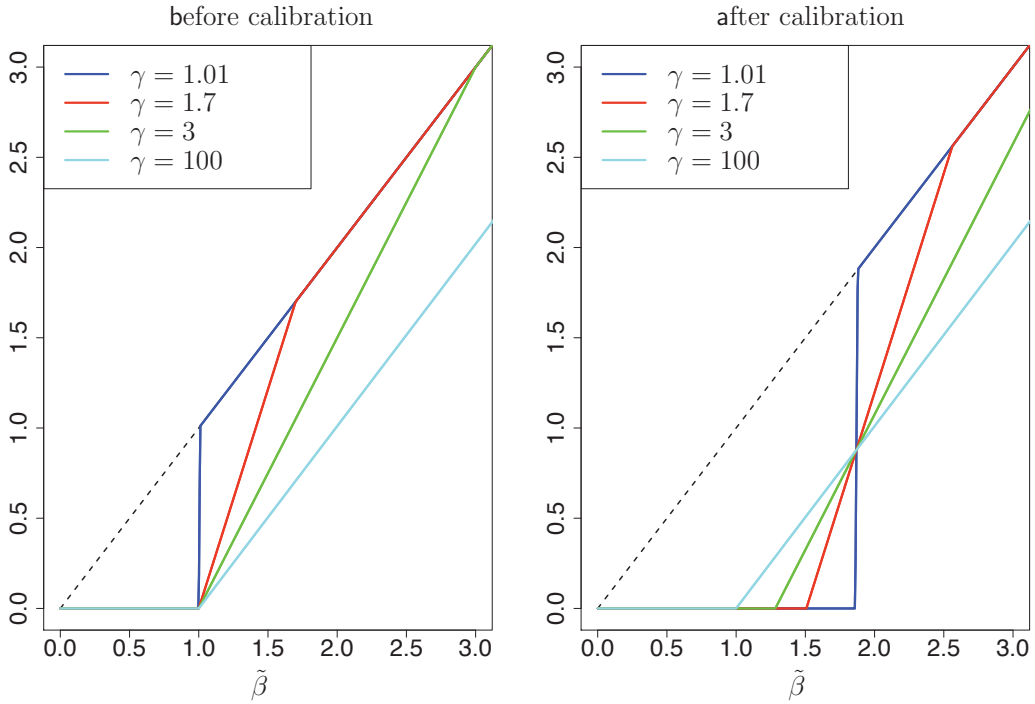
SPARSENET: coordinate descent for concave families

Mazumder, Friedman and Hastie (2009)

- Start with lasso family and obtain regularization path by coordinate descent.
- Move down family to slightly concave penalty, using lasso solutions as warm starts.
- Continue in this fashion till close to best subset penalty.

Results in regularization surface for concave penalty families.

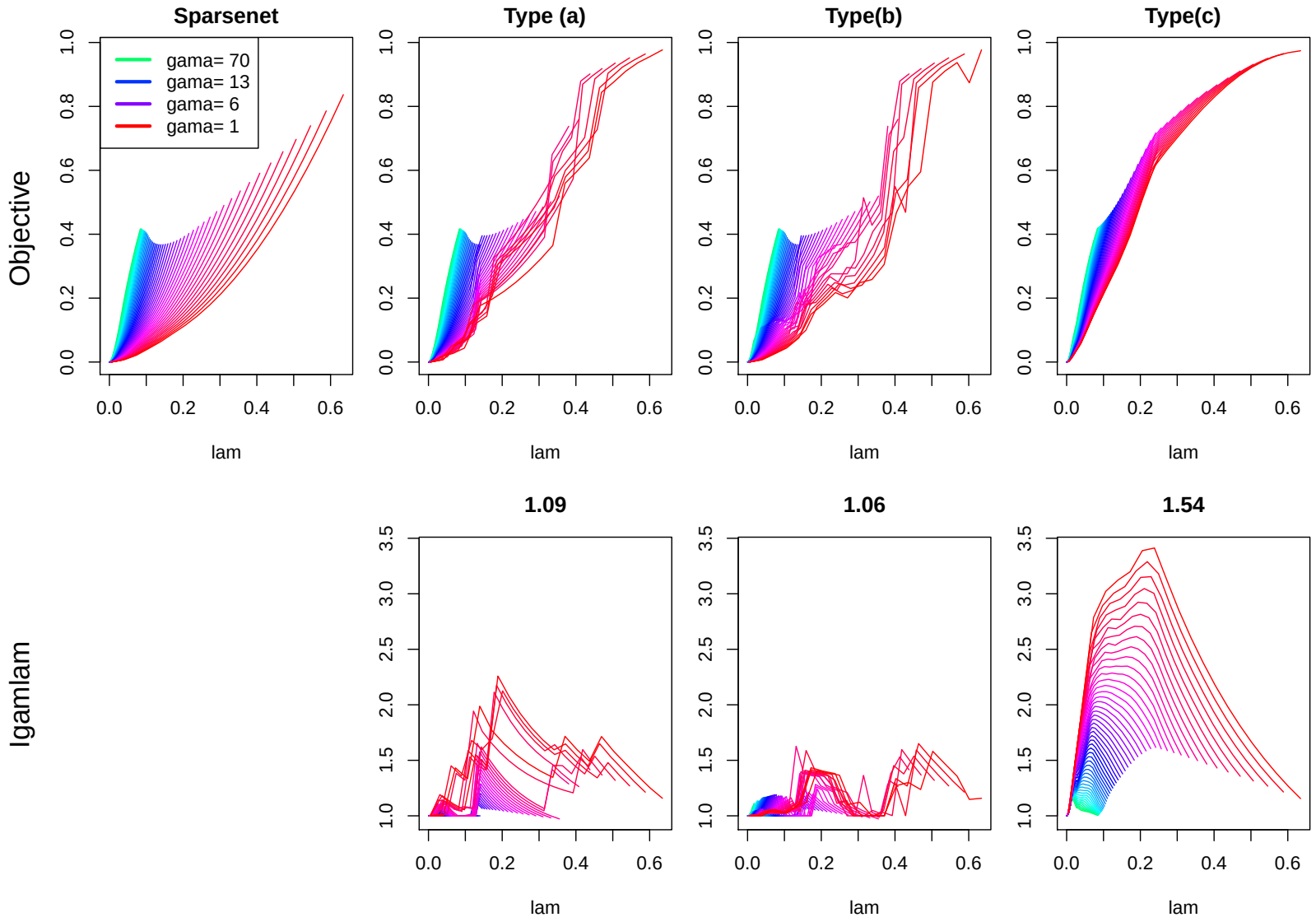
Related ideas in She (2008) using threshold functions, but not coordinate descent.



We prefer Zhang's MC+ penalty *calibrated for constant df* for fixed λ. Effective *df* computed as for lasso (Zou et al, 2007). As γ changes from *lasso* to *best subset*, shrinking threshold increases setting more coefficients to zero.

Properties

- Threshold functions are continuous in γ.
- Univariate optimization problems are convex over sufficient range of γ to cover lasso to subset regression, resulting in unique coordinate solutions.
- Monotonicity in shrinking threshold.
- Algorithm provably converges to (local) minimum of penalized objective.
- Empirically outperforms other algorithms for optimizing concave penalized problems.
- Inherits all theoretical properties of MC+ penalized solutions (Zhang, 2010 AoS).


Sparsenet
Type (a)
Type(b)
Type(c)
gama= 70
gama= 13
gama= 6
gama= 1
Objective
lam
1.09
1.06
1.54
Igamlam

Summary and Future Work

- ℓ_1 regularization (and variants) has become a powerful tool with the advent of wide data.
- Coordinate descent is fastest known algorithm for solving many of these problems—along a path of values for the tuning parameter.
- Currently developing fast algorithms using *group lasso* for screening interactions (gene-gene, gene-environment) in GLM models with $p \gg N$.
- Developing methods for estimating sequential FDR for path algorithms.