l_1 Regularization: Efficient and Effective

Saharon Rosset (IBM Research), Ji Zhu (Michigan)

Collaborators: Trevor Hastie, Rob Tibshirani (Stanford)

Outline

My talk:

- Introduction: Regularized optimization and the regularized path
- The Lasso and Least Angle Regression
- Relationship of l_1 regularization and boosting
- Sparseness propert(ies) of l_1 regularization
- Efficient l_1 regularization through piecewise linear solution paths

Next talk (Ji Zhu):

Designing efficient algorithms for Support Vector Machines using path-following methods

Regularized optimization

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_i C(y_i, \mathbf{x}_i\beta) + \lambda J(\beta)$$

- C is a convex loss, describing "goodness of fit" of our model to training data
  - Regression: $C(y, f) = C(y - f)$ function of residual
  - Classification: $C(y, f) = C(yf)$ function of margin
- $J(\beta)$ is a model complexity penalty.
  Typically $J(\beta) = \|\beta\|_q^q$ i.e. penalize l_q norm of model, $q \geq 1$.
- $\lambda \geq 0$ is a regularization parameter
  - As $\lambda \to 0$, we approach non-regularized model
  - As $\lambda \to \infty$, we get that $\hat{\beta}(\lambda) \to 0$

Examples

- Regularized linear regression:

$$\hat{\beta}(\lambda) = \min_{\beta} \sum_i (y_i - \mathbf{x}_i \beta)^2 + \sum_j \|\beta_j\|_q^q$$

  Squared error loss: $C(y, f) = (y - f)^2$

  - Ridge regression uses l_2 penalty $J(\beta) = \|\beta\|_2^2$
  - The Lasso (Tibshirani 96) uses l_1 penalty $J(\beta) = \|\beta\|_1$

- Support Vector Machines:

  Hinge loss: $C(y, f) = (1 - yf)_+$

  - Standard (2-norm) SVM uses l_2 penalty $\|\beta\|_2^2$
  - 1-norm SVM uses l_1 penalty $\|\beta\|_1$

The components of a regularized optimization problem

Loss: describes “goodness of fit” to training data

- Classic statistical view: corresponds to likelihood
- Should also consider robustness and computation

Penalty: limits model search, prevents overfitting

- Bayesian interpretation: prior on model space
- Should also consider sparseness and computation

Regularization parameter balances loss and penalty

The regularized solution path

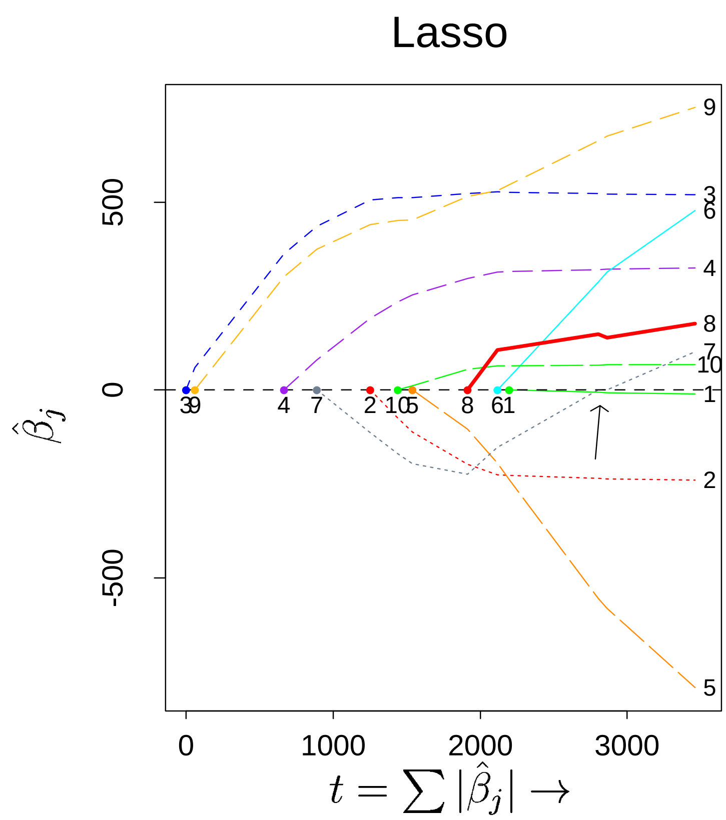
Fixing the loss, penalty and data, and varying the regularization parameter we get the "path of solutions"

$$\{\hat{\beta}(\lambda)\ ,\ 0 \leq \lambda < \infty\}$$

This is a 1-dim curve through $\mathbb{R}^p$.

- Interesting statistically, as the set of solutions to problems of interest (Bayesian interpretation: changing prior variance)
- Often interesting computationally, as it has properties which allow efficient "tracking" of this path

Example: Lasso solution path in $\mathbb{R}^{10}$



(from Efron et al. (2004). Least Angle Regression. Annals of Statistics)

Least Angle Regression

Efron et al (2004), Annals of Statistics

Consider the Lasso:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_i (y_i - \mathbf{x}_i\beta)^2 + \lambda \sum_j |\beta_j|$$

and its relation to two other regularization approaches:

- Stagewise regression: add variables one by one, fit residual (as opposed to stepwise, where we re-do the fit)
- Least Angle Regression: new, geometrically motivated approach with efficient algorithm

Least Angle Regression: Main Results

1. Efficient “path following” algorithm for lasso.
   - Use geometry of the curve $\{\hat{\beta}(\lambda)\ ,\ 0 \leq \lambda < \infty\}$ to track it
2. Close relationship between stagewise regression and lasso
   - By analogy, has implications for analysis of Boosting

We re-interpret these results and generalize them:

- To other regularized optimization problems
- To methods used in classification and machine learning

l_1 regularization: efficient and effective

Highlights:

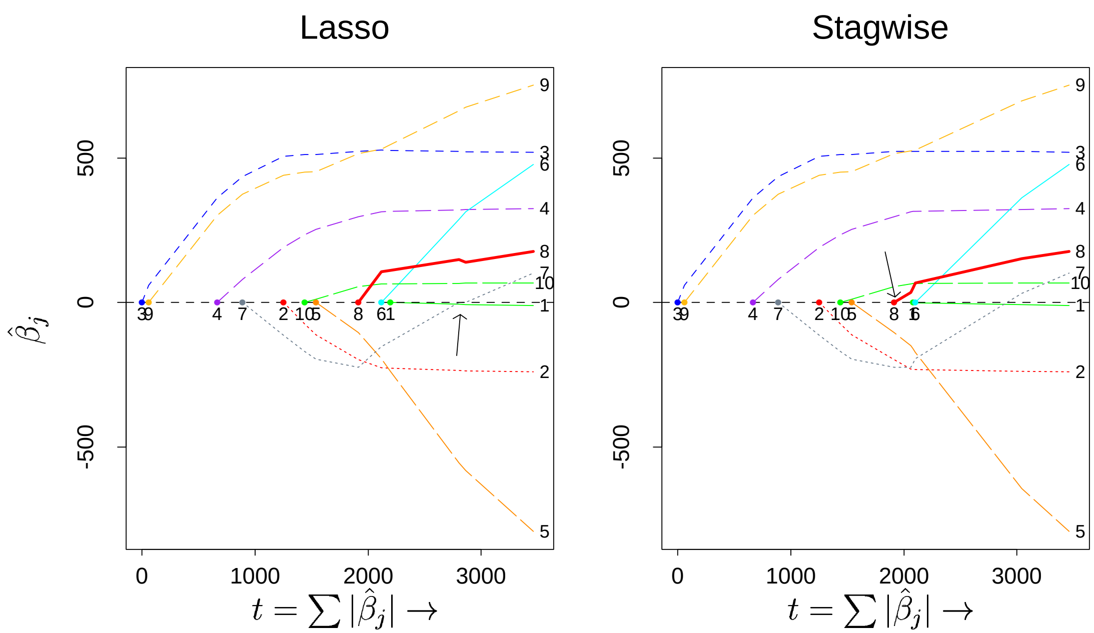
- Boosting as approximate l_1 regularization
  - Allows approximate l_1 regularization in high (even infinite) dimensional spaces
- The sparseness propert(ies) of l_1 regularization
- The piecewise linearity property of l_1 penalized solution paths
  - Design new, efficient algorithms for popular methods
  - Define new, robust regularized methods which we can solve efficiently

Boosting as approximate l_1 regularized path

Boosting and l_1 regularization

Hastie et al. (2001) argue and LARS makes more formal:

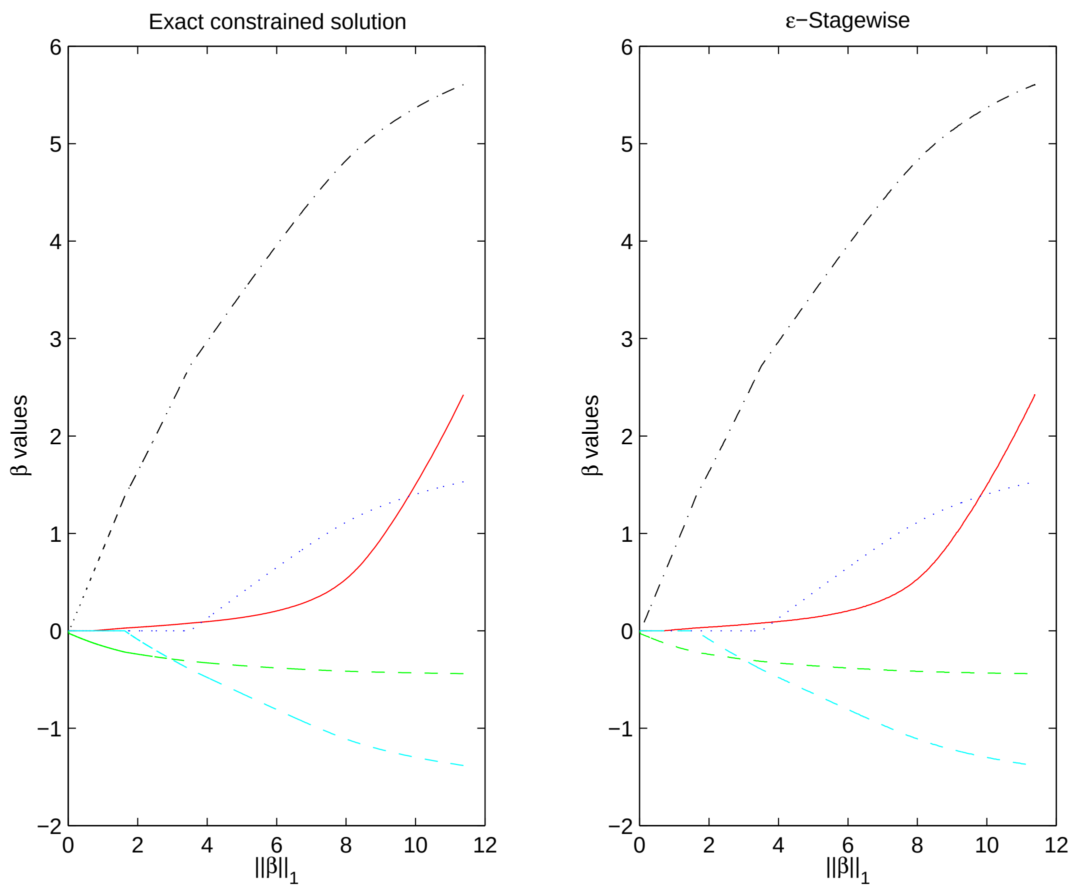
Boosting *(AKA forward stage-wise)* with squared error loss is very similar to lasso



Does this extend beyond sq. loss?

Yes, this is property of the l_1 penalty, not the loss.

l_1-penalized logistic regression example:



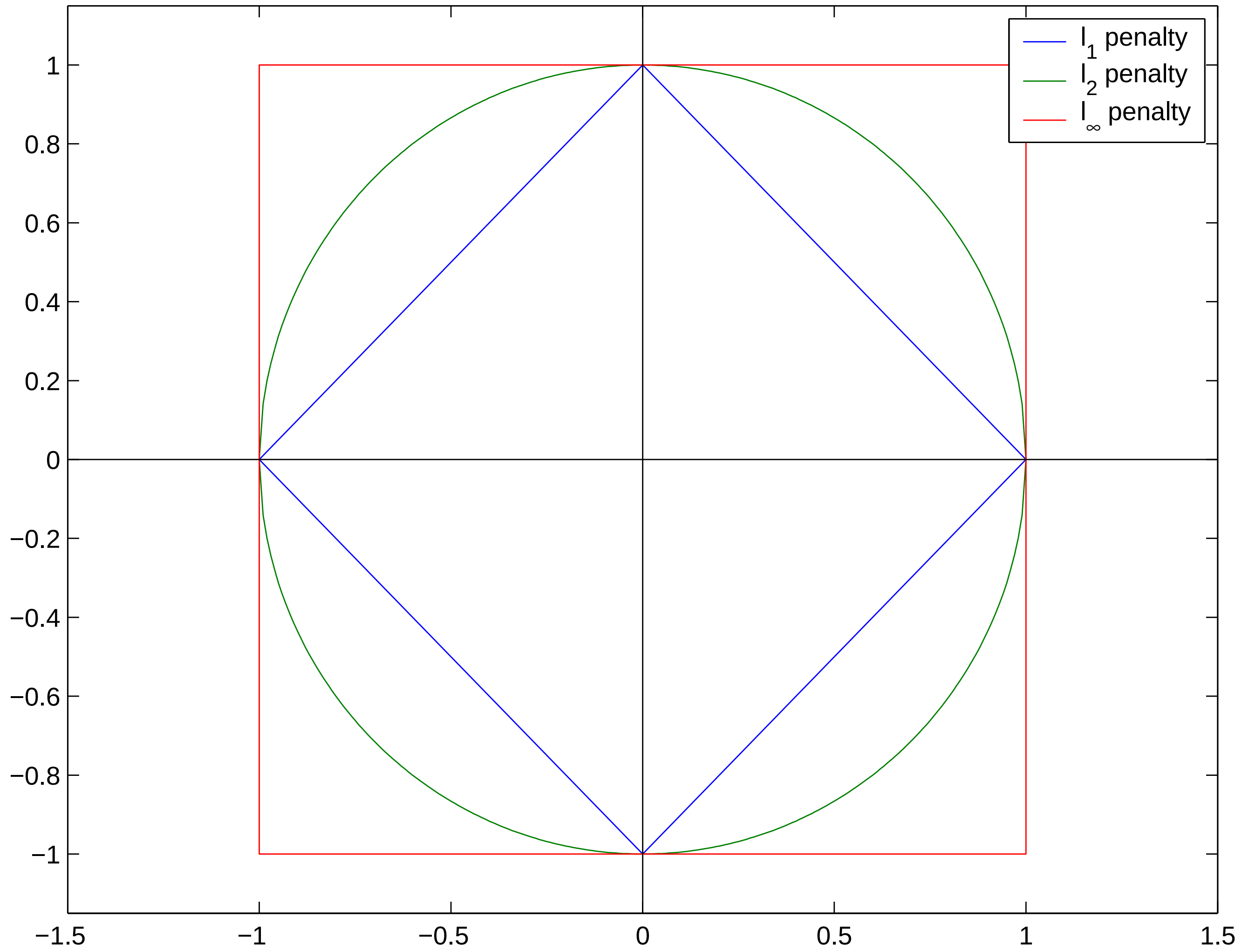
Conclusion: Boosting and l_1 reg.

"Boosting can be described as a coordinate-descent search, approximately following the path of l_1-constrained optimal solutions to its loss criterion, and converging, in the separable case, to a "margin maximizer" in the l_1 sense."

Rosset, Zhu & Hastie (2004). *Boosting as a Regularized Path to a Maximum Margin Classifier.* Journal of Machine Learning Research

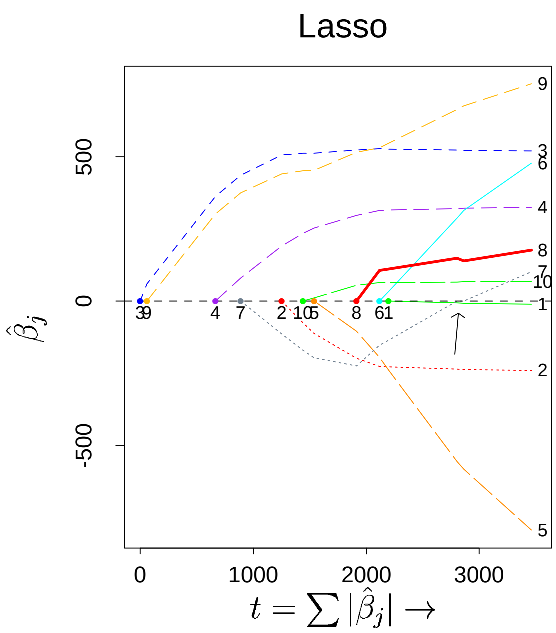
Sparseness propert(ies) of l_1 regularized path

l_1, l_2 and l_∞ penalties in $\mathbb{R}^2$



Sparseness of l_1 penalty: $n > p$

Shape of l_1 penalty implies sparseness. For large values of λ only few non-zero coefficients.



Sparseness: $p > n$

For any convex loss, assuming only “non-redundancy”:

Theorem (Rosset et al. 2004)

Any l_1 regularized solution has at most n non-zero components

Corollary

The limiting interpolating (or margin maximizing) solution also has at most n non-zero components

Some implications of sparseness

- Variable selection (obviously)
- l_1-regularized problems are "easier" than, say, l_2-regularized ones
  - Can give good solutions in $p >> n$ situations
    See:

    Friedman, Hastie, Rosset, Tibshirani, Zhu (2004). *Discussion of three boosting papers*. Annals of Statistics

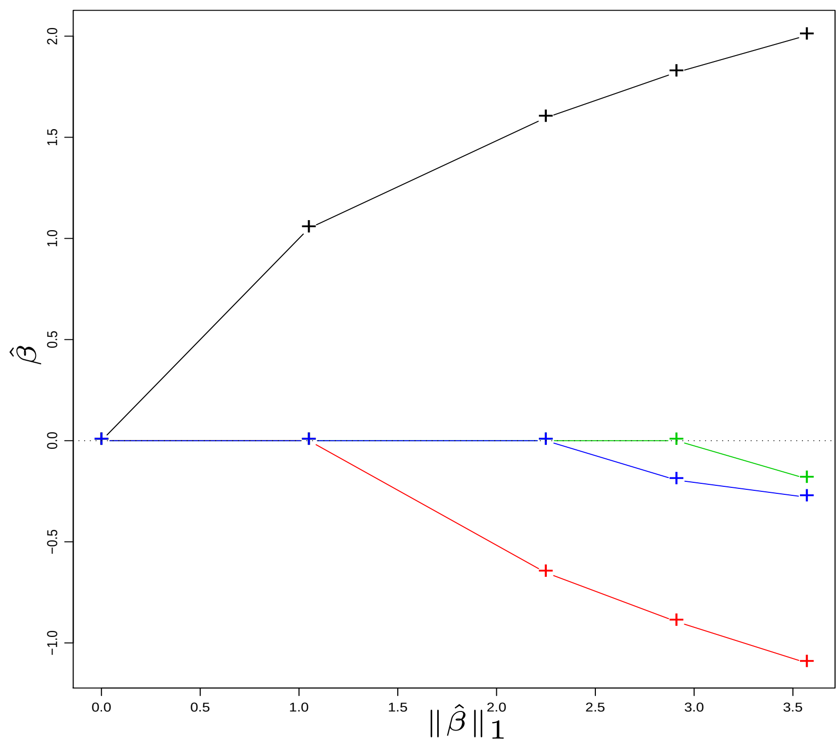
    Ng (2004). *Feature selection, l_1 vs l_2 regularization and rotational invariance*. ICML-04

Piecewise linear regularized solution paths

The piecewise linear property

We want $\{\hat{\beta}(\lambda)\ ,\ 0 \leq \lambda < \infty\}$ to be piecewise linear in $\mathbb{R}^p$ as function of λ.

For lasso established by Osborne et al. (2001) and LARS paper



Our key questions:

- What is the fundamental property of (loss, penalty) pairs which yields piecewise linearity?
- Are there efficient algorithms to generate these regularized paths?
- Are there statistically interesting members in these families?

Rosset & Zhu (2004). Piecewise linear regularized solutions paths.

What makes paths piecewise linear?

Some algebra gives us the following Lemma:

A sufficient condition for piecewise linearity is that:

- *The loss C is piecewise quadratic*
- *The penalty J is piecewise linear*

Practically, this condition is also necessary

Building blocks for PWL regularized optimization problems

Piecewise quadratic loss:

- Squared error loss: regression: $(y - r)^2$, classification: $(1 - yr)^2$
- Huber's loss (robust):

$$C(y, \mathbf{x}\beta) = \begin{cases} (y - \mathbf{x}\beta)^2 & \text{if } |y - \mathbf{x}\beta| \leq m \\ m^2 + 2m(|y - \mathbf{x}\beta| - m) & \text{otherwise} \end{cases}$$

- Piecewise linear loss: regression: $|y - r|$, classification: $(1 - yr)_+$

Piecewise linear penalty:

- l_1 penalty: $J(\beta) = \sum_j |\beta_j|$ (gives sparse solutions)
- l_∞ penalty: $J(\beta) = max_j |\beta_j|$ (statistical motivation?)

Some interesting examples of PWL

(with efficient algorithms)

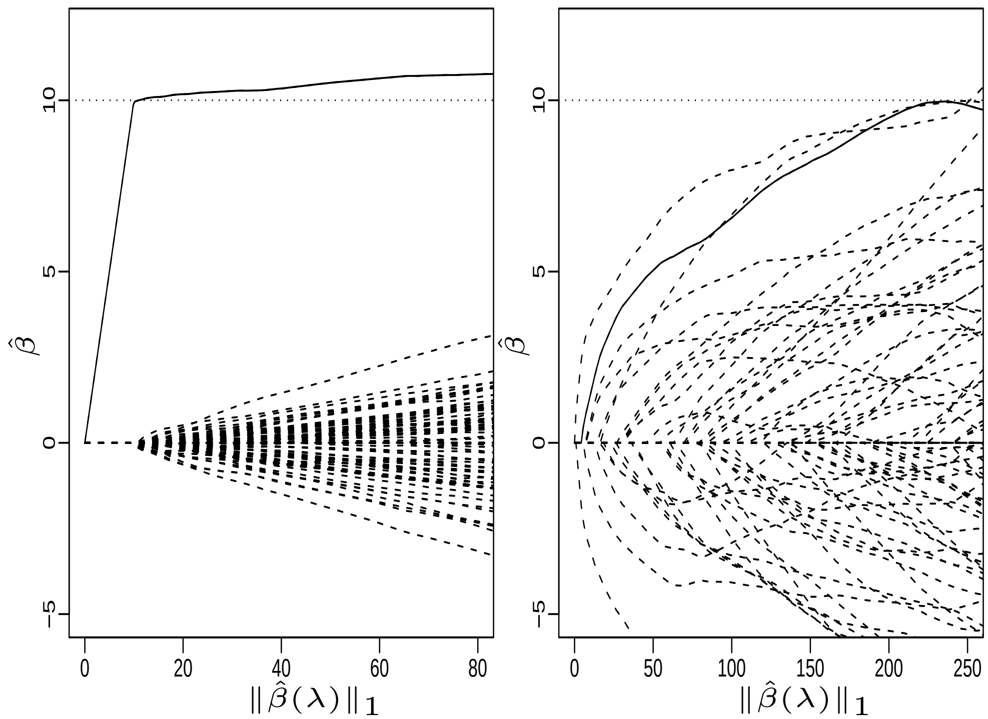
Robustifying the lasso

- $n = 100$, $p = 80$.
- All x_{ij} are i.i.d $N(0,1)$ and the true model is:

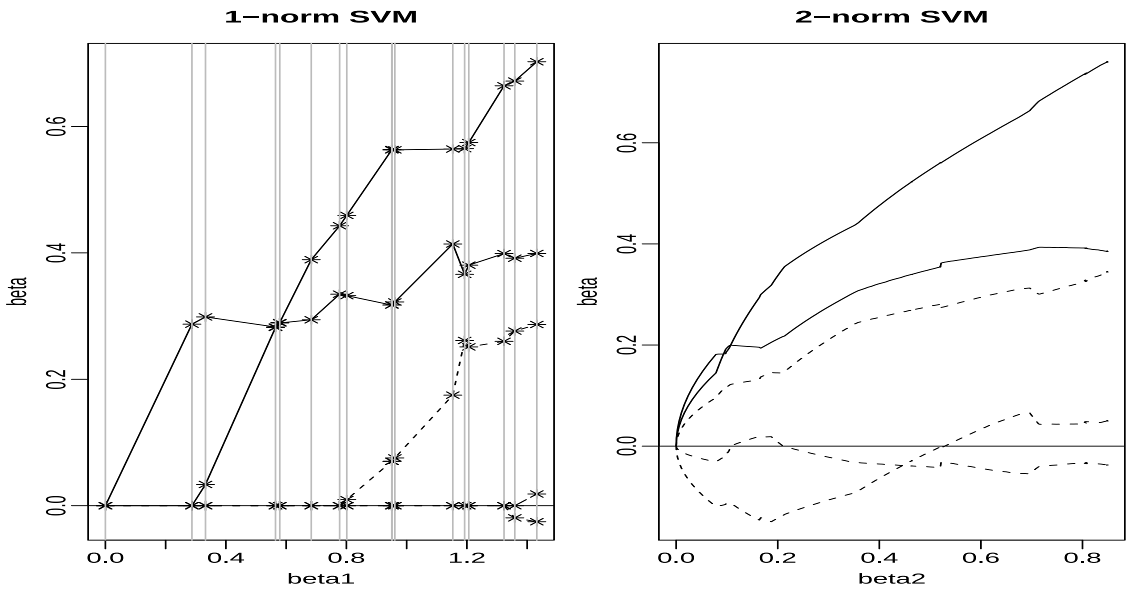
$$\begin{aligned} y_i &= 10 \cdot x_{i1} + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} 0.9 \cdot N(0,1) + 0.1 \cdot N(0,100) \end{aligned}$$

- Sparsity implies l_1 penalty is appropriate
- Compare l_1-regularized paths using Huber's loss and squared error loss

The Huberized lasso (left) and the lasso (right)



Classification: 1-norm and 2-norm Support Vector Machines



Zhu, Rosset, Hastie & Tibshirani. (2003) *1-norm SVM*, NIPS-03

Hastie, Rosset, Tibshirani & Zhu. (2004) *The entire regularization path of SVM*. Journal of Machine Learning Research.

Multiple penalty problem: Protein Mass Spectroscopy

(Tibshirani, Saunders, Rosset, Zhu & Knight, JRSSB, to appear)

- Predictors are “experssion levels” along a spectrum of masses for proteins.
- Want to constrain model while keeping coefficients “smooth”.
- Solution: l_1 penalty on coefficients, l_1 penalty on successive differences:

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_{\beta} \sum_i (y_i - \mathbf{x}_i \beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_j |\beta_j - \beta_{j-1}|$$

- Solution path is piecewise affine in (λ_1, λ_2)

Summary

Implicit or explicit l_1 regularization is prevalent in practical methods:

- Parametric regularization: lasso, 1-norm SVM
- Basis expansions: Wavelet thresholding, basis pursuit
- Implicit: boosting

Has favorable statistical and computational properties:

- Sparseness
- With appropriate loss, allows PWL solution paths

We use PWL property to:

- Design new, efficient algorithms for popular methods, like SVM
- Define new, robust regularized methods we can solve efficiently