A Robust Hybrid of Ridge and Lasso

Art B. Owen

Department of Statistics

Stanford University

Penalized regression

Minimize

$$\sum_{i=1}^{n}(y_i - \mu - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \text{(Ridge)}$$

or

$$\sum_{i=1}^{n}(y_i - \mu - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \text{(Lasso)}$$

Where

Response	$y_i \in \mathbb{R}$
Predictor	$x_i \in \mathbb{R}^p$
Coefficient	$\beta \in \mathbb{R}^p$
Intercept	$\mu \in \mathbb{R}$
Residual	$\varepsilon_i \equiv y_i - \mu - x_i'\beta$

Pros and cons

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\beta_j \neq 0}$$

Ridge: usually $\|\hat{\beta}\|_0 = n$ (not sparse)

Lasso: $\|\hat{\beta}\|_0 < n$ (might be too sparse)

Folklore: Ridge may be better for prediction

Hybrid

We seek a penalty that behaves like lasso at small β_j like ridge at large β_j

Can be done by $\|\beta\|_1 + \gamma\|\beta\|_2^2$

Hastie, Zou Elastic Net

Or by $\|\beta\|_1 + \gamma\|\beta\|_2$ or $\|\beta\|_1 + \gamma\|\beta\|_\infty$

Zhao, Rocha, Yu Composite Absolute Penalties

New hybrid

Min $\sum_{i=1}^{n} L(y_i - \mu - x_i'\beta) + \sum_{j=1}^{p} P(\beta_j)$

Choose $P(z)$ that is

like $|z|$ for small z

like z^2 for large z

Recall Huber's loss function

$$\mathcal{H}(z) = \begin{cases} z^2 & |z| \le 1 \\ 2|z| - 1 & |z| \ge 1 \end{cases}$$

Treats small errors like L_2 (efficiency)

Treats large errors like L_1 (robustness)

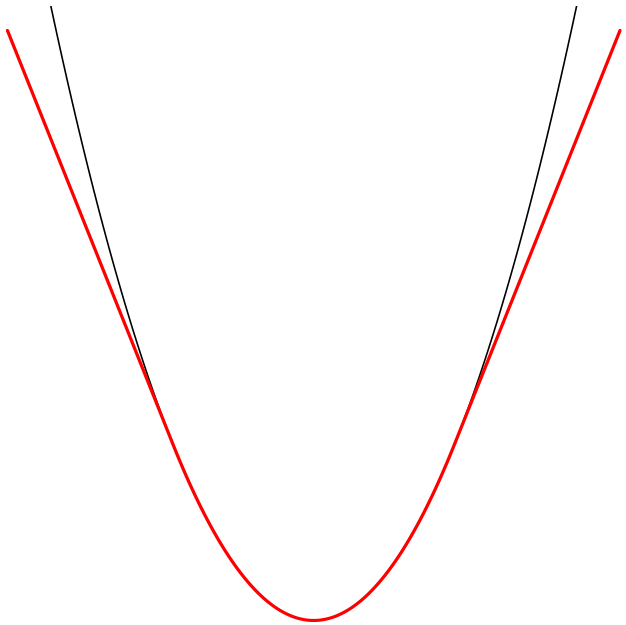
We want the opposite behavior

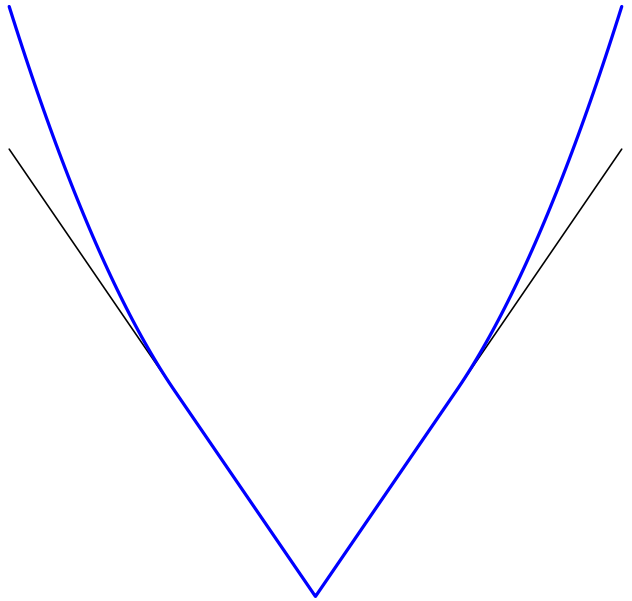
Recipe

1. Carpentry: find $P(z) \sim |z|$ for small z and $\sim z^2$ for large
2. Find dividing line between small and large (concomitant scale)
3. Avoid parameter search: make criterion convex in scale and β
4. Express result as quadratic program

Huber and Berhu

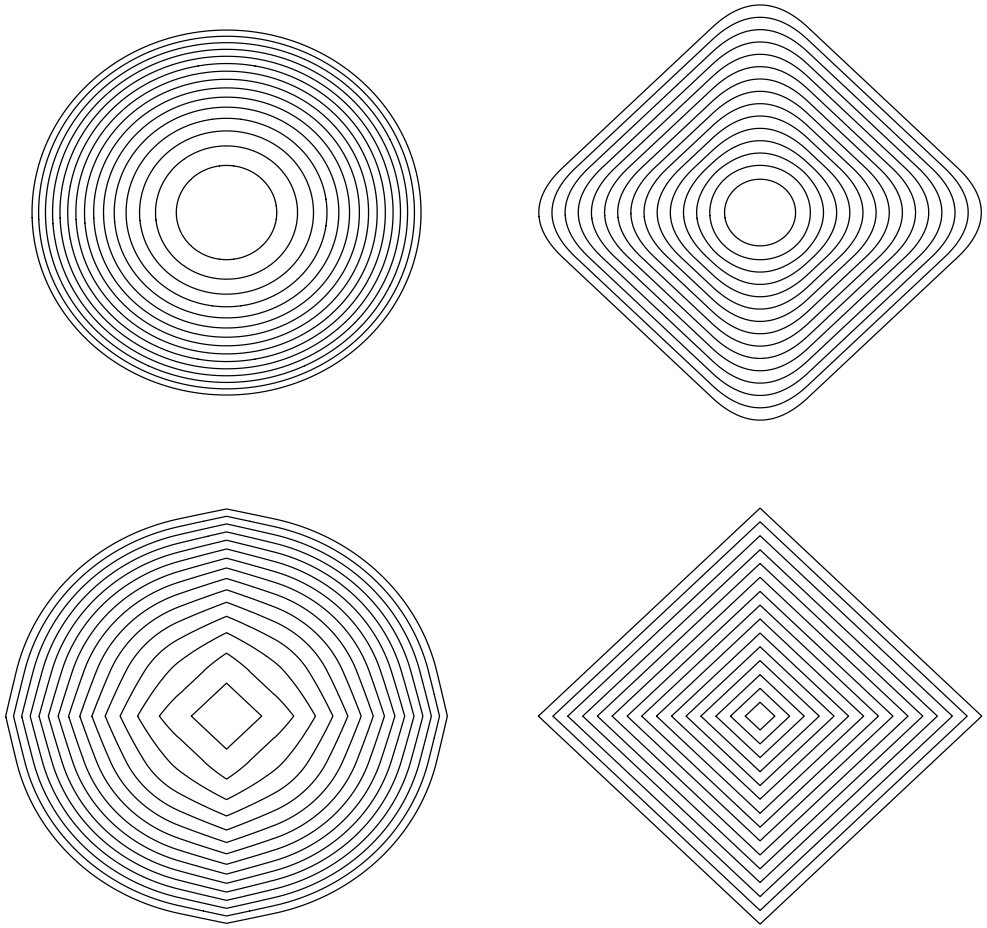
$$\mathcal{H}(z) = \begin{cases} z^2 & |z| \le 1 \\ 2|z| - 1 & |z| \ge 1 \end{cases} \qquad \mathcal{B}(z) = \begin{cases} |z| & |z| \le 1 \\ \frac{z^2+1}{2} & |z| \ge 1 \end{cases}$$





L_2 Huber

Berhu L_1



Concomitant scale

Parametrized Huber

$$\mathcal{H}_M(z) = M^2\mathcal{H}\Big(\frac{z}{M}\Big) = \begin{cases} z^2 & |z| \le M \\ 2M|z| - M^2 & |z| \ge M \end{cases}$$

e.g.: $M = 1.35$ for 95% efficiency at $N(0,1)$

Scaled Huber

$$\sum_{i=1}^{n} \mathcal{H}_M\Big(\frac{y_i - \mu - x_i'\beta}{\sigma}\Big) + \lambda \sum_{j=1}^{p} P(\beta_j)$$

Robust σ crucial

Also need scale for penalty on β_j

Estimation of σ

Common method: alternate between estimating σ and β

Better:

$$n\sigma + \sum_{i=1}^{n} \mathcal{H}_M\left(\frac{y_i - \mu - x_i'\beta}{\sigma}\right)\sigma$$

is jointly convex in μ, β, σ

optimal μ β unaffected by outer scaling and shift

Theorem

$$\rho(z) \qquad \text{convex on } z \in \mathcal{I} \subseteq \mathbb{R}$$

$$\Longrightarrow \rho\left(\frac{z}{\sigma}\right) \times \sigma \quad \text{convex on } (z, \sigma) \in \mathcal{I} \times (0, \infty)$$

Proved in Huber (1981)

Perspective

Huber's theorem known as the "perspective transformation"

by Boyd & Vandenberghe (2004)

Huber uses $\sigma + \rho(\cdot/\sigma)\sigma$ instead of $\rho(\cdot/\sigma)\sigma$

For $\rho(z) = z^2$ we get

z^2/σ convex on $(z, \sigma) \in \mathbb{R} \times (0, \infty)$

We'll see this

"quad_over_lin"

function later

Scale continued

$\sigma + \mathcal{H}(\varepsilon/\sigma)\sigma$ Convex 1st term keeps $\hat{\sigma}$ from being too large

For least squares

$$f(\mu, \sigma) = \sum_{i=1}^{n} \sigma + \left(\frac{y_i - \mu}{\sigma}\right)^2 \sigma = n\sigma + \sum_{i=1}^{n} (y_i - \mu)^2/\sigma$$

Any $\sigma > 0$ minimizing μ is $\bar{y}$

$$\frac{\partial}{\partial \sigma} f(\mu, \sigma) = n - \sum_{i=1}^{n} (y_i - \mu)^2/\sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2$$

Gets the $N(\mu, \sigma^2)$ MLEs from a convex criterion (unlike the log likelihood)

Extends to regression

Concomitant scale for L_1

$$n\sigma + \sum_{i=1} \frac{|y_i - \mu - x_i'\beta|}{\sigma} \times \sigma$$

$$= \quad n\sigma + \sum_{i=1} |y_i - \mu - x_i'\beta|$$

Degenerate

Always get $\hat{\sigma} \downarrow 0$

Huber's concomitant scale fails for L_1 loss

Also for Huber's loss (L_1 outside) when $M \leq 1$

So take $M > 1$

Combined criterion

$$n\sigma + \sum_{i=1}^{n} \mathcal{H}_M\left(\frac{\varepsilon_i}{\sigma}\right)\sigma + \lambda\left[p\tau + \sum_{j=1}^{p} \mathcal{B}_M\left(\frac{\beta_j}{\tau}\right)\tau\right]$$

Jointly convex in $(\mu, \beta, \sigma, \tau)$ for any λ

Fix M (eg at 1.35 for both functions)

Trace over $0 \le \lambda \le \infty$

Scaling Berhu

$$\mathcal{B}_M(z) = M\mathcal{B}\left(\frac{z}{M}\right) = \begin{cases} |z| & |z| \le M \\ \frac{z^2+M^2}{2M} & |z| \ge M \end{cases}$$

What happens for ridge regression?

Replace $\mathcal{H}_H$ and $\mathcal{H}_B$ by L_2:

$$n\sigma + \sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma}\right)^2\sigma + \lambda\left[p\tau + \sum_{j=1}^{p}\left(\frac{\beta_j}{\tau}\right)^2\tau\right]$$

Recall $n\,\hat{\sigma}^2 = \|\varepsilon\|_2^2$ also $p\,\hat{\tau}^2 = \|\beta\|_2^2$

End up with

$$2\sqrt{n}\,\|\varepsilon\|_2 + \lambda 2\sqrt{p}\,\|\beta\|_2 \quad \propto \quad \|\varepsilon\|_2 + \sqrt{\frac{p}{n}}\,\lambda\,\|\beta\|_2$$

instead of $\|\varepsilon\|_2^2 + \lambda\|\beta\|_2^2$

σ and τ fall out

ridge trace does not change (relabel λ)

Implementation in cvx

Matlab functions of Grant, Boyd, Ye (2006)

Disciplined convex programming

Ridge/lasso regression

$$p \in \{1, 2, \infty\}$$

```
cvx_begin
    variables mu beta(p)
    minimize norm(y−mu−x∗beta,2) + lambda ∗ norm(beta,p)
cvx_end
```

cvx features

calles SeDuMi solver

very fast to prototype

very slow to execute (no warm start)

handles growing list of convex optimization problems

symbolic smarts to verify convexity

Huber as quadratic program

cvx has Huber function but no concomitant scale

Quadratic program for $\mathcal{H}_M(z)$ Grant, Boyd, Ye

$$\begin{array}{ll} \text{minimize} & w^2 + 2Mv \\ \text{subject to} & |z| \le v + w \\ & w \le M \\ & v \ge 0, \end{array}$$

$\sigma + \mathcal{H}_M(z/\sigma)\sigma$ translates to

$$\begin{array}{ll} \text{minimize} & \sigma + (w^2 + 2Mv)\sigma \\ \text{subject to} & |z|/\sigma \le v + w \\ & w \le M \\ & v \ge 0 \\ & \sigma \ge 0 \end{array}$$

Simplification

After $w \rightarrow w/\sigma$ and $v \rightarrow v/\sigma$

$$
\begin{aligned}
\text{minimize} \quad & \sigma + w^2/\sigma + 2Mv \\
\text{subject to} \quad & |z| \leq v + w \\
& w \leq M\sigma \\
& v \geq 0 \\
& \sigma \geq 0
\end{aligned}
$$

$\mathcal{B}(z)$ as quadratic program

$$\begin{aligned} \text{minimize} \quad & v + w^2/(2M) + w \\ \text{subject to} \quad & |z| \leq v + w \\ & v \leq M \\ & w \geq 0 \end{aligned}$$

With scale

$$\begin{aligned} \text{minimize} \quad & \tau + v + w^2/(2M\tau) + w \\ \text{subject to} \quad & |z| \leq v + w \\ & v \leq M\tau \\ & w \geq 0 \\ & \tau \geq 0 \end{aligned}$$

Summed Huber loss with scale

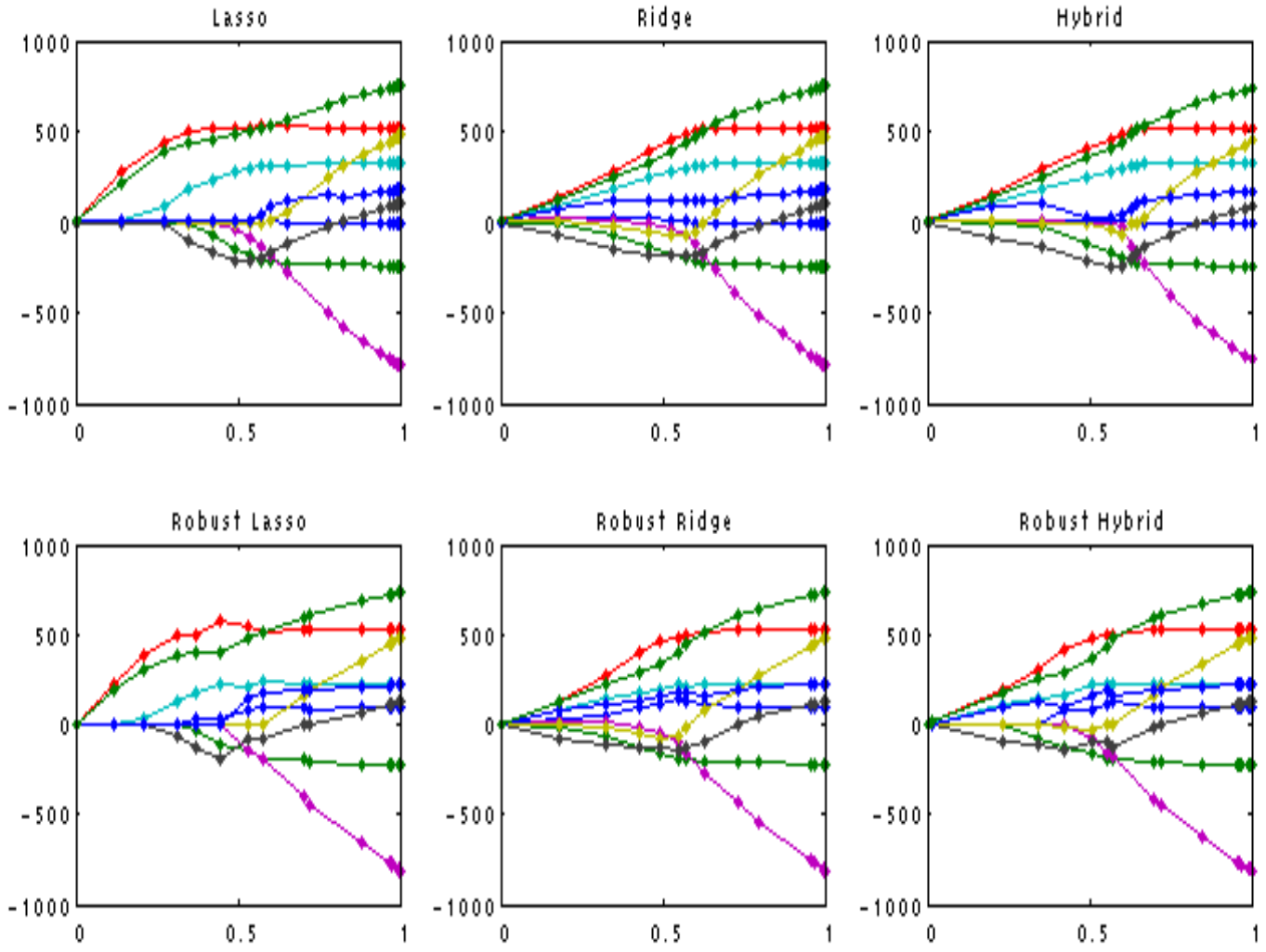
$$
\begin{aligned}
\text{minimize} \quad & n\sigma + \textstyle\sum_{i=1}^{n} w_i^2/\sigma + 2Mv_i \\
\text{subject to} \quad & |\varepsilon_i| \le v + w \\
& w \le M\sigma \\
& v \ge 0 \\
& \sigma \ge 0
\end{aligned}
$$

Similar for Berhu penalty

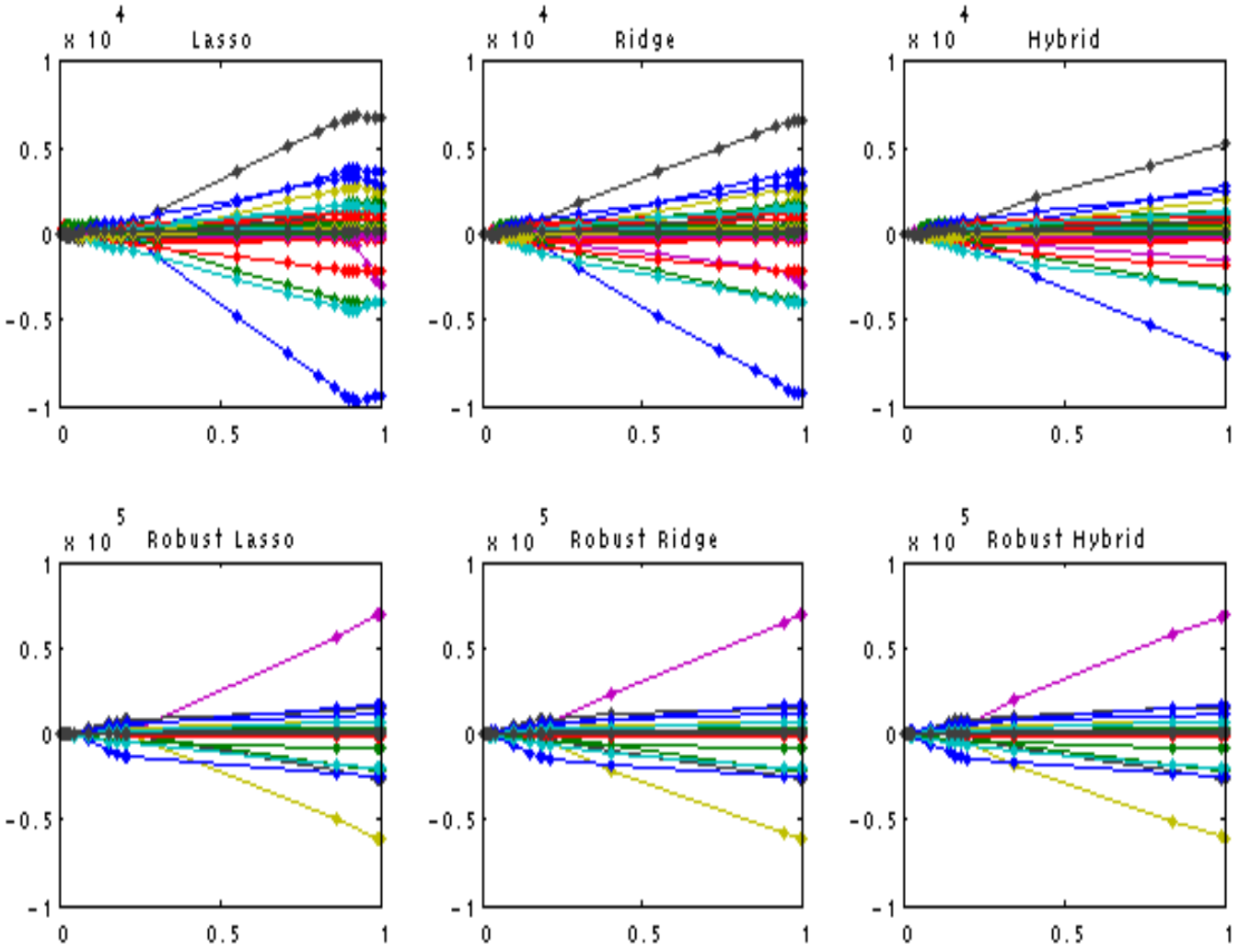
Cvx code for hybrid

```
cvx_begin
    variables mu res(n) w(n) v(n) coef(p) sig ww(p) vv(p) tau;
    minimize quad_over_lin(res,sig) + 2*ML*sum(v) + n*sig +...
        B*( p*tau + sum( abs(coef)) + quad_over_lin( ww, tau)./(2.*MP))
    subject to
        res = yy - mu - xx*coef
        w <= ML*sig
        sig >= 0
        abs(res) <= v+w
        v >= 0
        abs(coef) <= vv+ww
        vv <= MP * tau
        ww >= 0
cvx_end
```

Diabetes example



Full quadratic model



Conclusions

An L_1 vs L_2 tradeoff for both penalty and loss

Both can be handled similarly

Concomitant scale fits in

Hybrid performs like ridge on subset of params

Next steps

3) Compare prediction and coefficient estimation

2) Pick λ automatically

1) Make it faster

Acknowledgments

Michael Grant

Organizers: Joe Verducci, Xiaotong Shen, John Lafferty

NSF DMS-0306612

Sponsors: NSF, AMS, IMS, SIAM