

Adaptive Lasso for correlated predictors

Keith Knight

Department of Statistics
University of Toronto

e-mail: keith@utstat.toronto.edu

This research was supported by NSERC of Canada.

OUTLINE

1. Introduction
2. The Lasso under collinearity
3. Projection pursuit with the Lasso
4. Example: Diabetes data

1. INTRODUCTION

- Assume a linear model for $\{(\mathbf{x}_i, Y_i) : i = 1, \dots, n\}$:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n) \end{aligned}$$

- Assume that the predictors are centred and scaled to have mean 0 and variance 1.
 - We can estimate β_0 by \bar{Y} — least squares estimator.
 - Thus we can assume that $\{Y_i\}$ are centred to have mean 0.
- In many applications, p can be much greater than n .
- In this talk, we will assume implicitly that $p < n$.

Shrinkage estimation

- **Bridge regression:** Minimize

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|^\gamma$$

for some $\gamma > 0$.

- Includes the Lasso (Tibshirani, 1996) and ridge regression as special cases with $\gamma = 1$ and 2 respectively.
 - For $\gamma \leq 1$, it's possible to obtain exact 0 parameter estimates.
 - Many other variations of the Lasso: elastic nets (Zou & Hastie, 2005), fused lasso (Tibshirani *et al.*, 2006) among others.
 - The Dantzig selector of Candès & Tao (2007) is similar in spirit to the Lasso.

- **Stagewise fitting:** Given $\hat{\beta}^{(k)}$, minimize

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i^T \hat{\beta}^{(k)} - \mathbf{x}_i^T \phi)^2$$

over ϕ with all but 1 (or a small number) of its elements equal to 0. Then define

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \epsilon \hat{\phi} \quad (0 < \epsilon \leq 1)$$

and repeat until “convergence”.

- This is a special case of **boosting** (Shapire, 1990).
- Also related to LARS (Efron *et al.*, 2004), which in turn is related to the Lasso.

2. THE LASSO UNDER COLLINEARITY

- For given λ , the Lasso estimator $\hat{\beta}(\lambda)$ can be defined in a number of equivalent ways:

1. $\hat{\beta}(\lambda)$ minimizes

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \quad \text{subject to } \sum_{j=1}^p |\beta_j| \leq t(\lambda);$$

2. $\hat{\beta}(\lambda)$ minimizes

$$\sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\beta})^2 \quad \text{subject to } \left| \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) x_{ij} \right| \leq \lambda$$

for $j = 1, \dots, p$.

- The advantage of the Lasso is that it produces exact 0 estimates while $\hat{\beta}(\lambda)$ is a smooth function of λ .
 - This is very useful when $p \gg n$ to produce “sparse” models.
- However, when the predictors $\{\mathbf{x}_i\}$ are highly correlated then $\hat{\beta}(\lambda)$ may contain too many zeroes.
- This is not necessarily undesirable but some important effects may be missed as a result.
 - How does one interpret a “sparse” model under high collinearity?

Question: Why does this happen?

Answer: Redundancy in the constraints

$$\left| \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) x_{ij} \right| \leq \lambda \quad \text{for } j = 1, \dots, p$$

due to collinearity; that is, we don't have p independent constraints.

- The Dantzig selector minimizes $\sum_j |\beta_j|$ subject to similar constraints on the correlations, and thus will tend to behave similarly.

- For LS estimation ($\lambda = 0$), we have

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \mathbf{x}_i^T \mathbf{a} = 0$$

for any \mathbf{a} .

- Similarly, we could try to consider estimates $\tilde{\boldsymbol{\beta}}$ such that

$$\left| \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}) \mathbf{x}_i^T \mathbf{a}_\ell \right| \leq \lambda$$

for some set of vectors (projections) $\{\mathbf{a}_\ell : \ell \in \mathcal{L}\}$.

- If the set \mathcal{L} is finite, we can incorporate predictors $\{\mathbf{a}_\ell^T \mathbf{x}\}$ into the Lasso.

Example: Principal components regression ($|\mathcal{L}| = p$) where $\mathbf{a}_1, \dots, \mathbf{a}_p$ are the eigenvectors of

$$C = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T.$$

However ...

- Projections obtain via PC are based solely on information in the design.
- Moreover, they need not be particularly easy to interpret.
- More generally, there's no problem in taking $|\mathcal{L}| \gg p$.

3. PROJECTION PURSUIT WITH THE LASSO

- For collinear predictors, it's often desirable to consider projections of the original predictors.
- Given predictors x_1, \dots, x_p and projections $\{\mathbf{a}_\ell : \ell \in \mathcal{L}\}$, we want to identify “interesting” (data-driven) projections $\mathbf{a}_{\ell_1}, \dots, \mathbf{a}_{\ell_p}$ and define new predictors $\mathbf{a}_{\ell_1}^T \mathbf{x}, \dots, \mathbf{a}_{\ell_p}^T \mathbf{x}$.
- We can take \mathcal{L} to be very large – but the projections we consider should be easily interpretable.
 - Coordinate projections (i.e. original predictors).
 - Sums and differences of two or more predictors.

Question: How do we do this?

Answer: Two possibilities:

- Use the Lasso on the projections.
 - But we need to worry about the choice of λ .
 - The “active” projections will depend on λ .
- Look at the Lasso solution as $\lambda \downarrow 0$.
 - This identifies a set of p projections.
 - These projections can be used in the Lasso.

Question: What happens to the Lasso solution as $\lambda \rightarrow 0$?

- Suppose that $\hat{\beta}(\lambda)$ minimizes

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

and that

$$C = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$$

is singular.

- Define

$$\mathcal{D} = \left\{ \phi : \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \phi)^2 = \min_{\beta} \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \beta)^2 \right\}.$$

Proposition: For the Lasso estimate $\beta(\lambda)$, we have

$$\lim_{\lambda \downarrow 0} \hat{\beta}(\lambda) = \operatorname{argmin} \left\{ \sum_{j=1}^p |\phi_j| : \phi \in \mathcal{D} \right\}.$$

“Proof”. Assume (for simplicity) that the minimum RSS is 0. Then $\hat{\beta}(\lambda)$ minimizes

$$Z_\lambda(\beta) = \frac{1}{\lambda} \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \beta)^2 + \sum_{j=1}^p |\beta_j|.$$

As $\lambda \downarrow 0$, the first term of Z_λ blows up for $\beta \notin \mathcal{D}$ and is exactly 0 for $\beta \in \mathcal{D}$. The conclusion follows using convexity of Z_λ .

Corollary: The Dantzig selector estimator has the same limit as $\lambda \downarrow 0$.

- In our problem, define $t_{i\ell}$ to be a scaled version of $\mathbf{a}_{\ell}^T \mathbf{x}_i$.
- The model now becomes

$$\begin{aligned} Y_i &= \sum_{\ell \in \mathcal{L}} \phi_{\ell} t_{i\ell} + \varepsilon_i \\ &= \mathbf{t}_i^T \boldsymbol{\phi} + \varepsilon_i \quad (i = 1, \dots, n) \end{aligned}$$

- We estimate $\boldsymbol{\phi}$ by minimizing

$$\sum_{\ell \in \mathcal{L}} |\phi_{\ell}| \quad \text{subject to} \quad \sum_{i=1}^n (Y_i - \mathbf{t}_i^T \boldsymbol{\phi}) \mathbf{t}_i = \mathbf{0}.$$

- This can be solved using linear programming methods.
 - Software for the Lasso tends to be unstable as $\lambda \downarrow 0$.

Asymptotics:

- Assume $p < r = |\mathcal{L}|$ are fixed and $n \rightarrow \infty$.
- Define matrices

$$C = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$$
$$D = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{t}_i \mathbf{t}_i^T$$

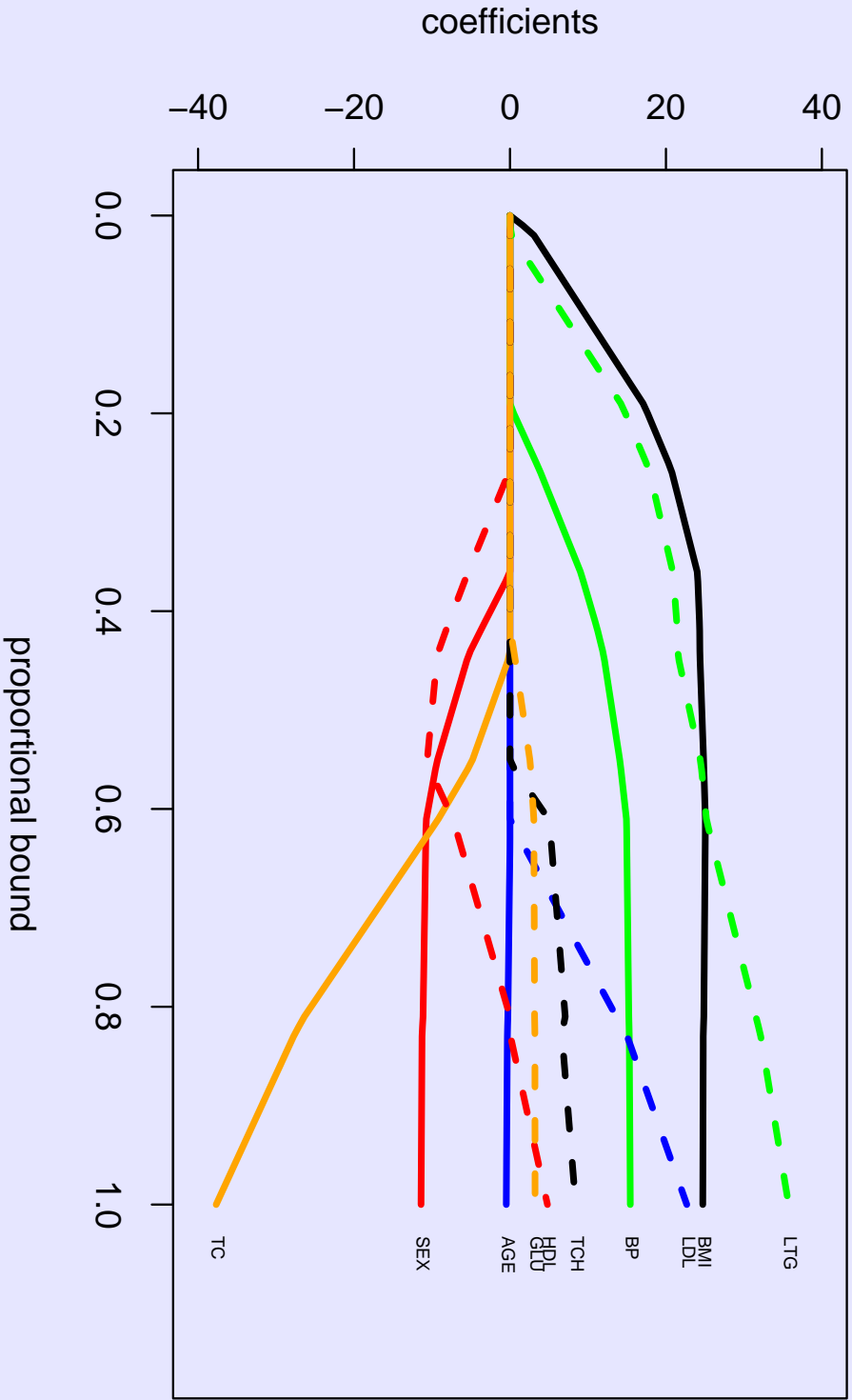
where C is non-singular and D singular with rank p .

- Then $\widehat{\phi}_n \xrightarrow{p}$ some ϕ_0 .
- We also have $\sqrt{n}(\widehat{\phi}_n - \phi_0) \xrightarrow{d} \mathbf{V}$ where the distribution of \mathbf{V} is concentrated on the orthogonal complement of the null space of D .

4. EXAMPLE

Diabetes data (Efron *et al.*, 2004)

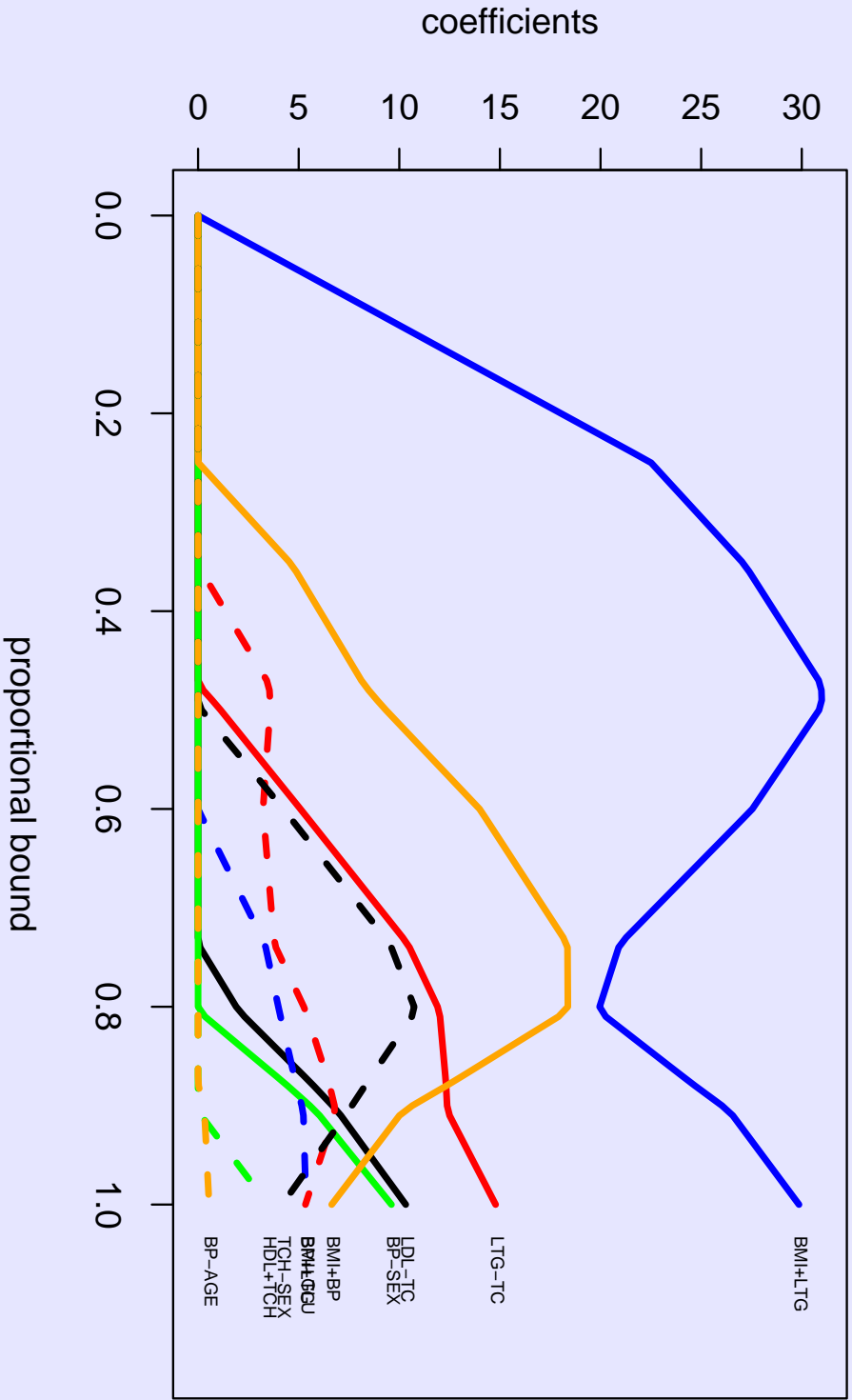
- Response: measure of disease progression.
- Predictors: age, sex, BMI, blood pressure, and 6 blood serum measurements (TC, LDL, HDL, TCH, LTG, GLU).
 - Some predictors are quite highly correlated.
- Analysis indicates that the most important variables are LTG, BMI, BP, TC, and sex.
- Look at coordinate-wise projections as well as pairwise sums and differences (100 projections in total).



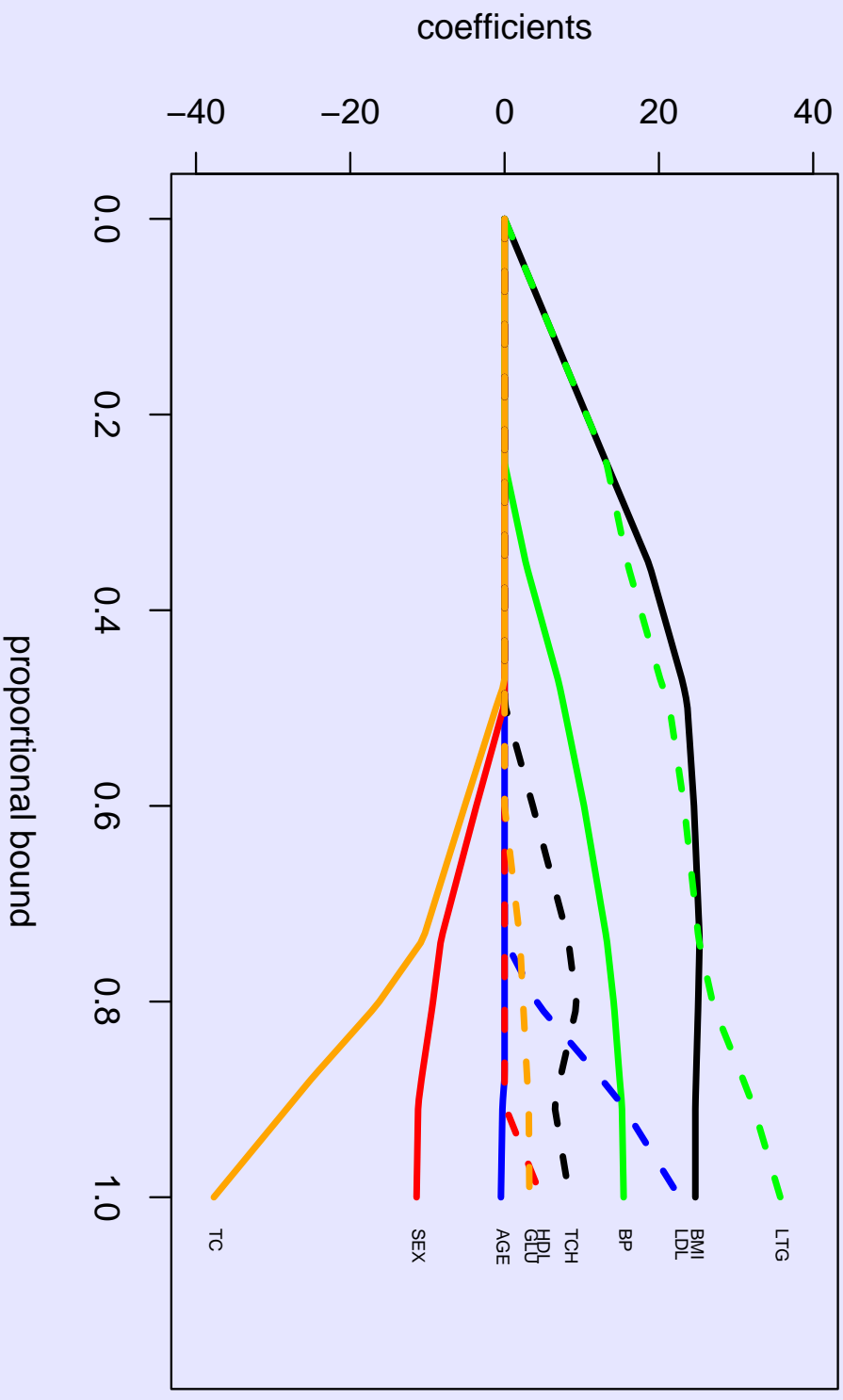
Lasso plot for original predictors.

Results: Estimated projections

Projections	Estimates
BMI + LTG	29.86
LTG – TC	14.79
LDL – TC	10.32
BP – SEX	9.61
BMI + BP	6.64
BMI + GLU	5.36
BP + LTG	5.33
TCH – SEX	4.18
HDL + TCH	3.48
BP – AGE	0.55



Lasso plot for the 10 identified projections.



Lasso trajectories for original predictors using the projections.