Regularized regression

$y = X\beta + \varepsilon$, X $n \times p$ design matrix

→ if X's are correlated ⇒ high variance for $\hat{\beta}$'s

⇒ correlated $\hat{\beta}$'s

⇒ Difficult to interpret model

→ Source of problem? $\hat{\beta} = (X'X)^{-1}X'y$

the inverse numerically unstable when p large cmp n and/or X's are correlated

What to do? → Regularize the fit

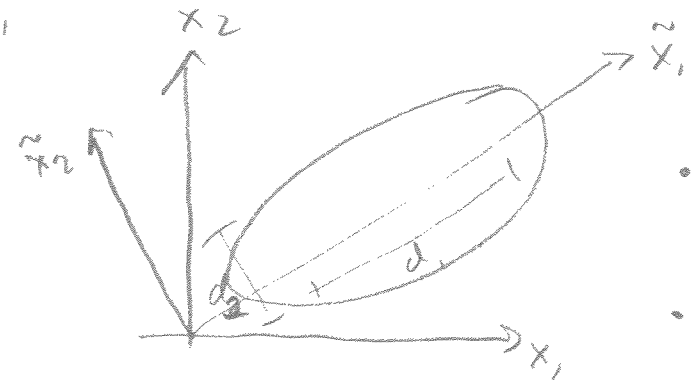
Principal Component Regression

$X = UDV'$ Singular value decomposition $U'U = I$, $V'V = I$

$X'X = VD^2V'$

/

eigen(X)



- $\tilde{X}_1$ and $\tilde{X}_2$ are uncorrelated
- $V(\tilde{X}_2) = d_2^2 < V(\tilde{X}_1) = d_1^2$

- X_1 and X_2 are correlated
- direction V_1 (1st column in V) i.e. maximum X-spread in this direction
- $V(\tilde{X}_1) = d_1^2$

So PC = new x-variables such that

$V(\tilde{x}_1) > V(\tilde{x}_2) >$

If we rotate X-variables $\Rightarrow$ $\tilde{X}$-variables

The new variables are uncorrelated $\Rightarrow$ new $\hat{\tilde{\beta}}$'s are uncorrelate

$$y = \tilde{X}\tilde{\beta} + \varepsilon \Rightarrow \hat{\tilde{\beta}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y \quad , \tilde{X} = XV$$
$$\tilde{\beta} = V'\beta$$

$$\Downarrow$$

$$V(\hat{\tilde{\beta}}) = \sigma^2 (\tilde{X}'\tilde{X})^{-1}$$

$$= \sigma^2 (V'X'XV)^{-1} = \sigma^2 (V'VD^2V'V)^{-1}$$

$$= \sigma^2 D^{-2}$$

↑ diagonal, so all $\hat{\tilde{\beta}}$'s are uncorrelated and

$V(\hat{\tilde{\beta}}_1) < V(\hat{\tilde{\beta}}_2) <$

$\sigma^2 d_1^{-2}$ $\sigma^2 d_2^{-2}$ ...

In new coordinate system, the 1st variable $\tilde{x}_1$ has the largest spread and so its coefficient estimate $\hat{\tilde{\beta}}_1$ is the most precise.

To regularize - just use the first k components with large spread.

Pros of PC regression — simple to use

— keep only $\tilde{x}$-variables for which we can estimate $\tilde{\beta}$ w high precision

(ON — each new $\tilde{x}$-variable = linear combo of all x-variables ⇒ difficult to interpret.

Regularized regression

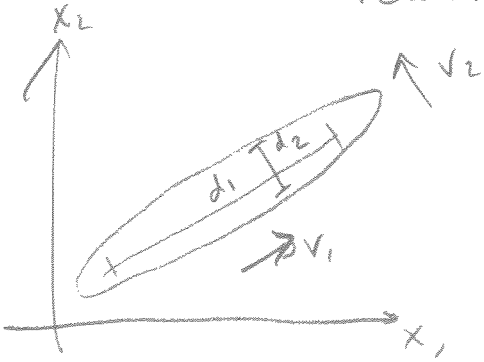
Note $V(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 V D^{-2} V' = \sigma^2 \sum_{k=1}^{p} d_k^{-2} v_k v_k'$

spectral decomposition

Total $MSE(\hat{\beta}) = \text{Trace}\{V(\hat{\beta})\} = \sigma^2 \text{Tr}\{(X'X)^{-1}\} = \sigma^2 \sum_{k=1}^{p} d_k^{-2}$

large if <u>any</u> d_k is small.

When does that happen? Whenever there are x's that are highly correlated so information in X is redundant.



Here d_2 is small

So LS estimates

- $E(\hat{\beta}) = \beta$ — do unbiased
- $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ — can be large if any x_i are correlated

Can we reduce variance by allowing for some bias?

Source of high variance = numerical instability of $X'X$ inverse

→ stabilize it → $\hat{\beta}_R = (X'X + rI)^{-1} X'y$

add r to the diagonal of $X'X$ before taking inverse

→ artificially suppress correlation between x's

What are $\hat{\beta}_R$ properties?

$\hat{\beta}_R = (X'X + rI)^{-1} X'y \rightarrow E(\hat{\beta}_R) = (X'X + rI)^{-1} X'X\beta$

$= (X'X + rI)^{-1} (X'X + rI - rI)\beta$

$= \beta - r(X'X + rI)^{-1}\beta$

bias

controlled by r
if $r = 0$, no bias

depends on β
large $\beta \rightarrow$ large bias

$$V(\hat{\beta}_R) = \sigma^2 (X'X + rI)^{-1} X'X (X'X + rI)^{-1}$$

$$MSE(\hat{\beta}_R) = Var + bias^2$$

$$= (X'X + rI)^{-1} \left(\sigma^2 (X'X) + r^2 \beta\beta' \right) (X'X + rI)^{-1}$$

$$\text{Total MSE} = \text{Trace}\{MSE(\hat{\beta}_R)\}$$

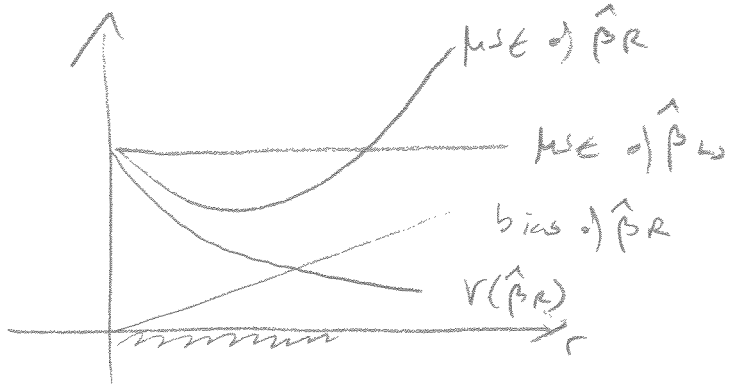
$$= \text{Trace}\left\{ V (D^2 + rI)^{-1} V' \left[\sigma^2 V D^2 V' + r^2 \beta\beta' \right] V (D^2 + rI)^{-1} V' \right\}$$

$$= \text{Trace}\left\{ V (D^2 + rI)^{-1} \left[\sigma^2 D^2 + r^2 \tilde{\beta}\tilde{\beta}' \right] (D^2 + rI)^{-1} V' \right\}$$

diagonal matrix

$$= \sum_{k=1}^{p} \frac{\sigma^2 d_k^2 + r^2 \tilde{\beta}_k^2}{(d_k^2 + r)^2} \quad < \sum_{k=1}^{p} \sigma^2 d_k^{-2} \quad \text{for some } r$$

↑ MSE of $\hat{\beta}_R$ ↑ MSE of $\hat{\beta}_{LS}$



r-values for which $\hat{\beta}_R$ better than $\hat{\beta}_{LS}$

Use CV to find r in practice. — best prediction error.

Another way of looking at this:

Shrinkage estimators

What does adding r to diagonal of $X'X$ do?

If X's are uncorrelated s.t. $X'X = D^2$

$$\Rightarrow (X'X + rI) = (D^2 + rI) = \begin{pmatrix} d_1^2 + r & & & \\ & d_2^2 + r & & \\ & & \ddots & \\ & & & d_p^2 + r \end{pmatrix}$$

$(X'X + rI)^{-1} X'y$ = each $\hat{\beta}_j$ shrunk by a factor r

General; $\hat{\beta}_S = \dfrac{\hat{\beta}_{LS}}{1+c}$ $\quad c > 0$

$\rightarrow$ bias $\dfrac{c\beta}{1+c}$

$\rightarrow$ variance $\dfrac{V(\hat{\beta}_{LS})}{(1+c)^2} < V(\hat{\beta}_{LS})$

If X's are not uncorrelated

$$X'X + rI = V D^2 V' + rI = V (D^2 + rI) V'$$

affects mostly the PC directions where d_k^2 is small.

Idea behind shrinkage is that by not allowing $\hat{\beta}$'s to get too big we cover ourselves from risk of poor estimates having a huge impact on prediction - just make everything smaller = bias↑ some and variance ↓ a lot.

We can arrive at this solution by doing

Penalized regression

LS: $\min_{\beta} (y-X\beta)'(y-X\beta) = \|y-X\beta\|^2 = \sum_{i=1}^{n} (y_i - x_i'\beta)^2$

Penalized LS: $\min_{\beta} (y-X\beta)'(y-X\beta)$

subject to $\|\beta\|^2 \leq \tau$

same thing as $\beta'\beta \leq \tau$

$\sum_{j=1}^{p} \beta_j^2 \leq \tau$

So, we want to fit the data to a model using LS but we only accept solutions where the total magnitude of β^2's is less than τ - either all small - or one big = τ rest 0 and anything in between.

How to solve constrained optimization problems?

$$\min_{\beta} (y-X\beta)'(y-X\beta)$$

$\Rightarrow$ Lagrange multipliers

subject to $\beta'\beta \leq \tau$

$$\min_{\beta} (y-X\beta)'(y-X\beta) + \lambda \beta'\beta$$

solve this for different values of λ and find the smallest λ so that solution satisfies the constraint $\beta'\beta \leq \tau$

If $\lambda = 0 \Rightarrow$ LS solution $\Rightarrow$ if $\hat{\beta}_{LS}'\hat{\beta}_{LS} \leq \tau$ done!

If not increase $\lambda \Rightarrow$ if $\hat{\beta}'\hat{\beta} = \tau$ done!

If not increase λ

Solution given λ:

$$\min_{\beta} \; (y - X\beta)'(y - X\beta) + \lambda \beta'\beta$$

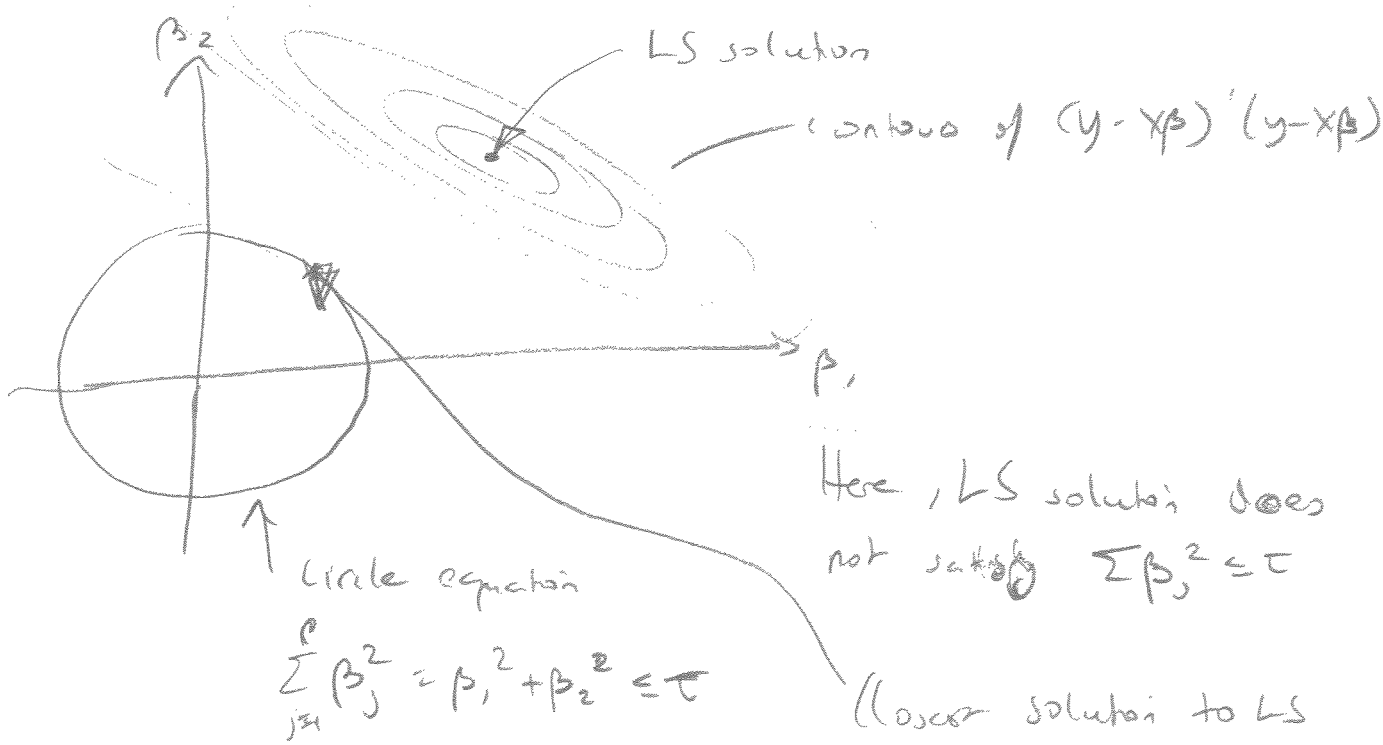
$$\frac{\partial}{\partial \beta} \; \text{[curved line]} \; = 0$$

$$\Rightarrow \quad -2X'(y - X\beta) + 2\lambda\beta = 0$$

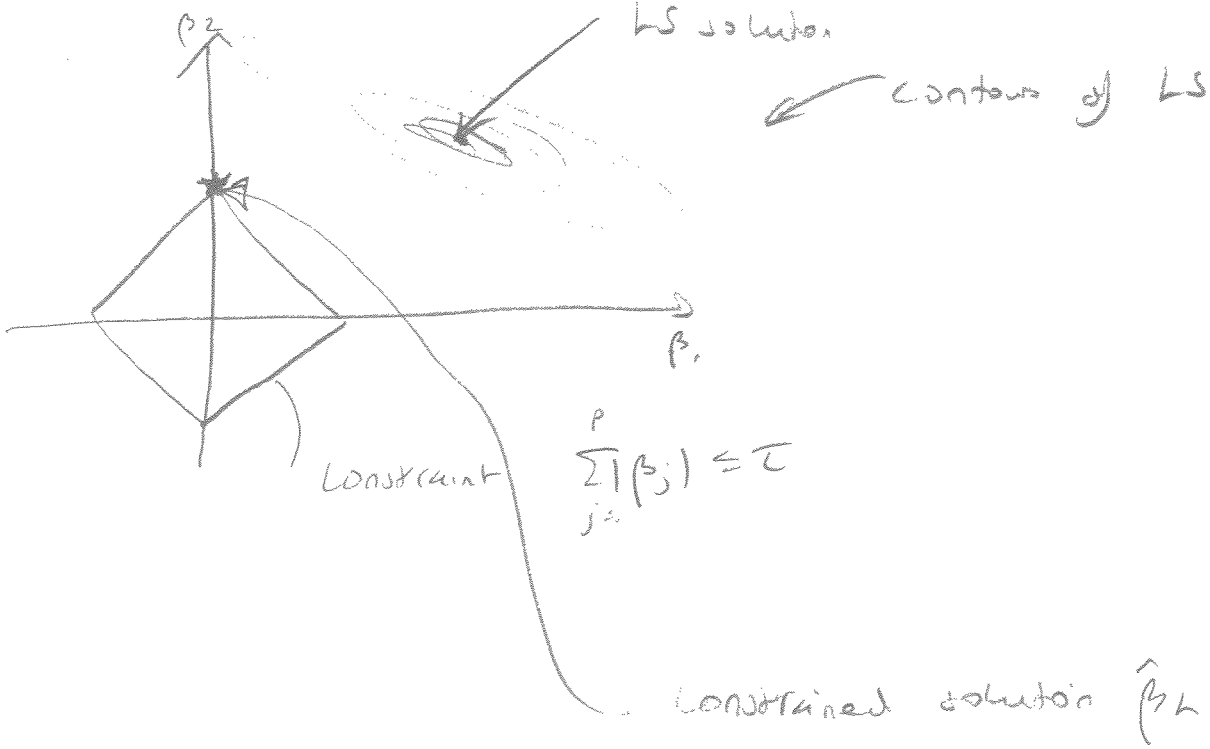
$$\Rightarrow \quad (X'X + \lambda I)\beta = X'y \quad \Rightarrow \quad \hat{\beta} = (X'X + \lambda I)^{-1} X'y$$

$$= \hat{\beta}_R \; \checkmark$$

Another way of looking at it:



Recently — explore the use of other constraint shapes

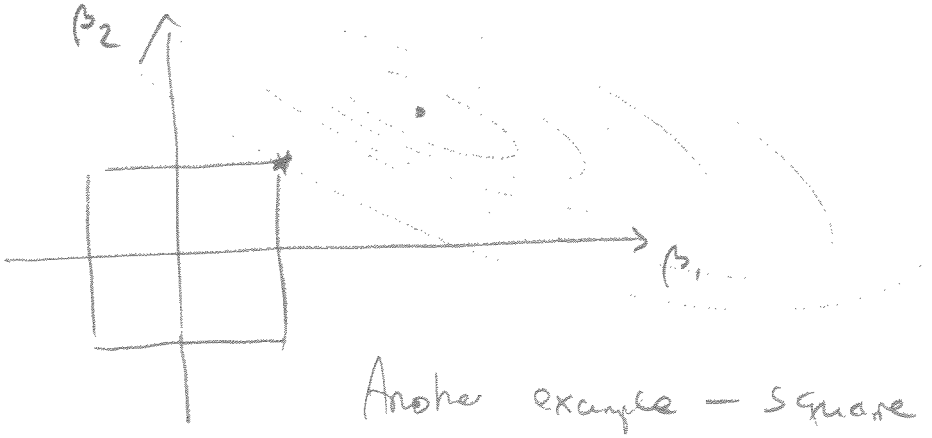


Using this diamond-shape constraint is called LASSO

It does automatic model selection.

Since constraint region has sharp angles on the axis of β-space, we encourage solutions where some $\hat{\beta} = 0$

Above example $\hat{\beta}_1 = 0$ and $\hat{\beta}_2 \neq 0$.



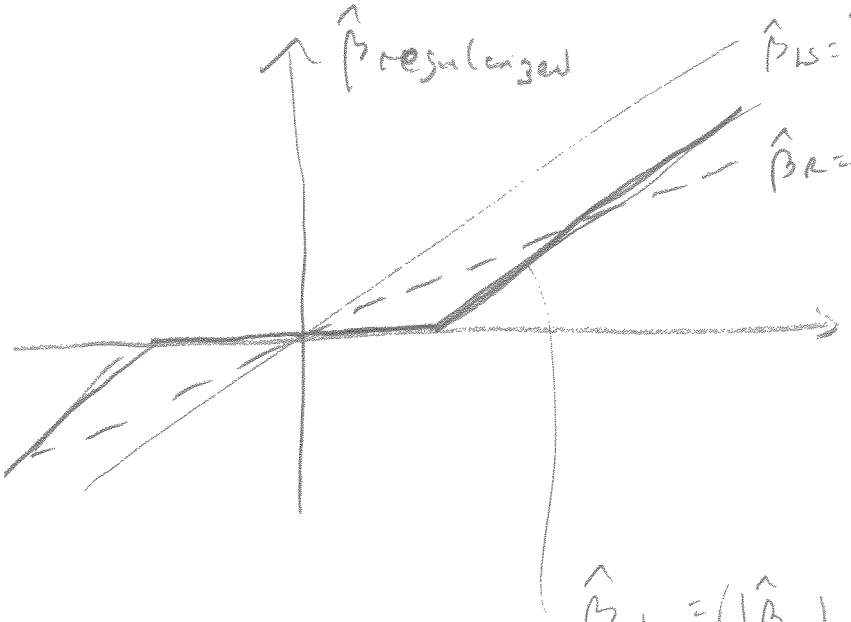
Another example — square constraint

$$\max(|\beta_1|, |\beta_2|) \leq \tau$$

$\Rightarrow$ encourages $\hat{\beta}_1 = \hat{\beta}_2$

Recently - lots of work in statistics to come up with automated model selection using interesting constraint regions like these.

Another look at the regularized estimates



$\hat{\beta}_{LS} = \hat{\beta}_{LS}$ → what we get when we use LS and no regularization

$\hat{\beta}_R = \frac{\hat{\beta}_{LS}}{1+c}$

notice how bias increases with β and how $\hat{\beta}_R$ is never 0 unless $\hat{\beta}_{LS} = 0$ exactly (never happens)

$$\hat{\beta}_L = (|\hat{\beta}_{LS}| - \delta)_+ \operatorname{sign}(\hat{\beta}_{LS})$$

$$= \begin{cases} \hat{\beta}_{LS} - \delta & \text{if } \hat{\beta}_{LS} > \delta \\ 0 & \text{if } |\hat{\beta}_{LS}| \leq \delta \\ \hat{\beta}_{LS} + \delta & \text{if } \hat{\beta}_{LS} < -\delta \end{cases}$$

: soft thresholding of $\hat{\beta}_{LS}$

- notice bias is constant