6.2 Lasso

$Y = 1.2 + 2.4X_1 + 0.0 X_2 - 1.7 X_3 + \varepsilon$

One drawback to ridge regression is that the estimated $\hat{\beta}_j$'s always take on nonzero values, even if they are very small.

The lasso (least absolute shrinkage & selection operator) is an alternative that minimizes

$$(y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_1$$

$$= (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso shrinks some estimated β_j's to exactly zero, so serves as a variable selection method, & generates sparse models

6.3 Another view of ridge/lasso

An equivalent to write the ridge problem is

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$$

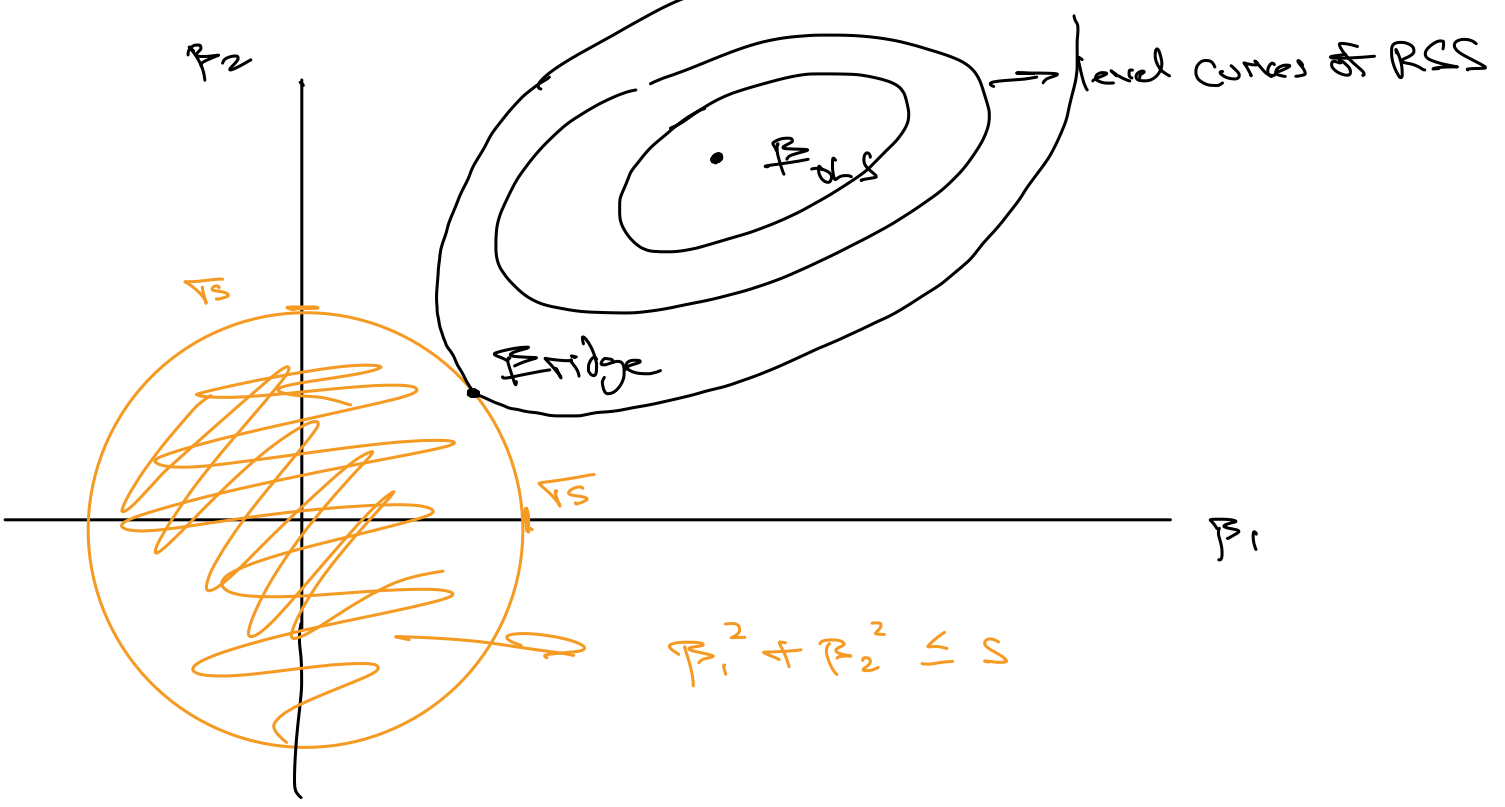
subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$ where s is a regularization parameter

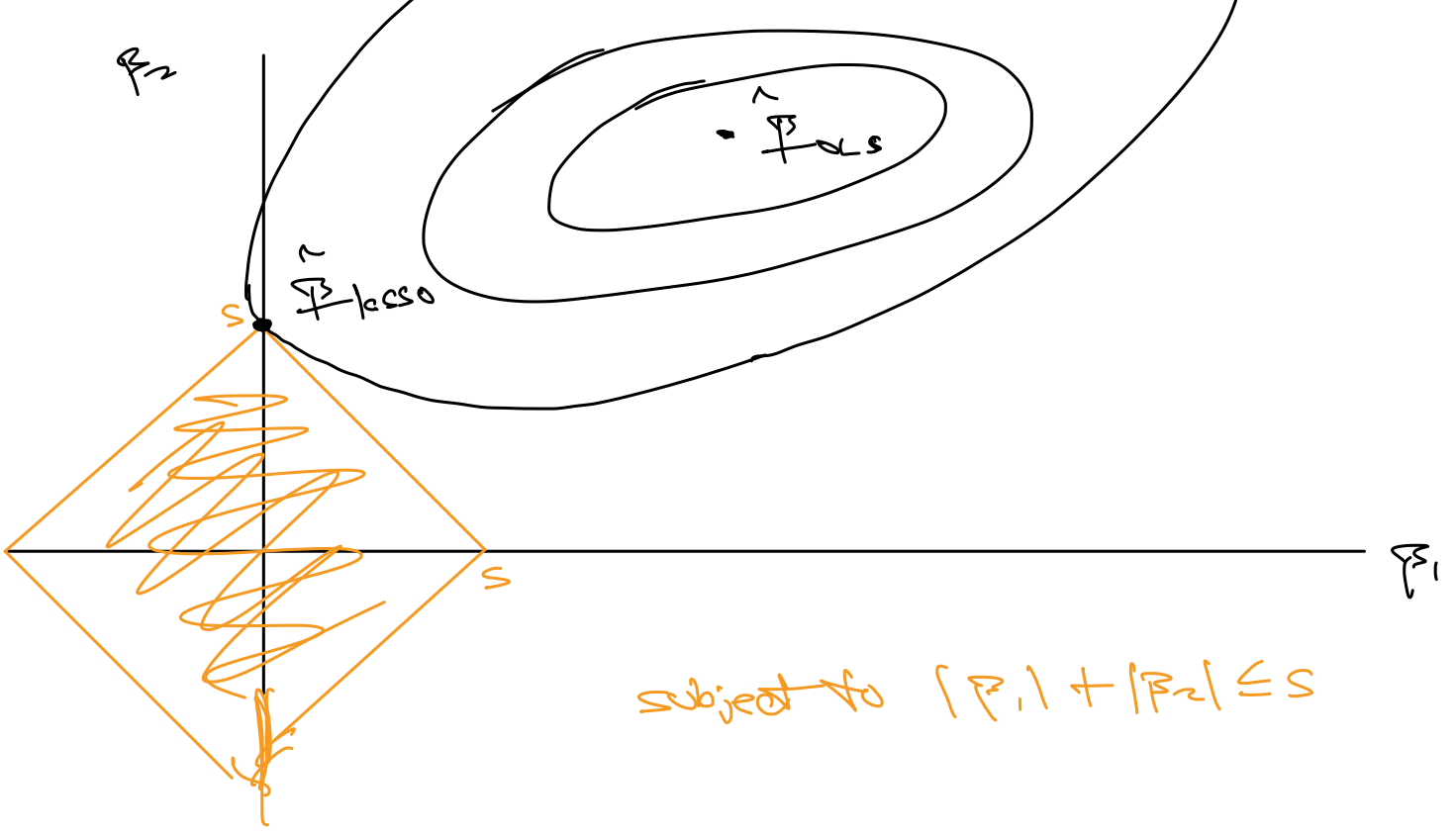
as $s \to \infty$ get the OLS estimates

as $s \to 0$ the model becomes the null model $Y = \beta_0 + \varepsilon$

For lasso, $\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$

subject to $\sum_{j=1}^{p} |\beta_j| \leq s$


β_2
level curves of RSS
$\hat{\beta}_{OLS}$
$\sqrt{s}$
Ridge
$\sqrt{s}$
β_1
$\beta_1^2 + \beta_2^2 \leq s$


β_2
$\hat{\beta}_{ols}$
$\tilde{\beta}_{lasso}$
s
s
β_1
subject to $|\beta_1| + |\beta_2| \leq s$

6.4 A Bayesian interpretation

In Bayesian statistics, we represent beliefs (uncertainty) knowledge about a parameter β using a prior distribution

then, given data, update beliefs (uncertainty) confidence/estimates in a posterior distribution.

Think of $Y \sim U(0, \beta)$ or $Exp(\beta)$ or $N(\beta, 1)$ or $N(0, \beta)$

We get data from a data likelihood, for $Y_1, \ldots, Y_n$

$$f(y \mid \beta) = \text{data likelihood}$$

we seek

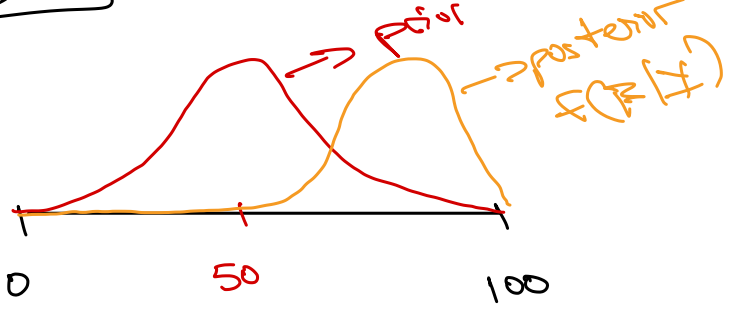
$$f(\beta \mid y) = \frac{f(y \mid \beta)\,\pi(\beta)}{f(y)}$$

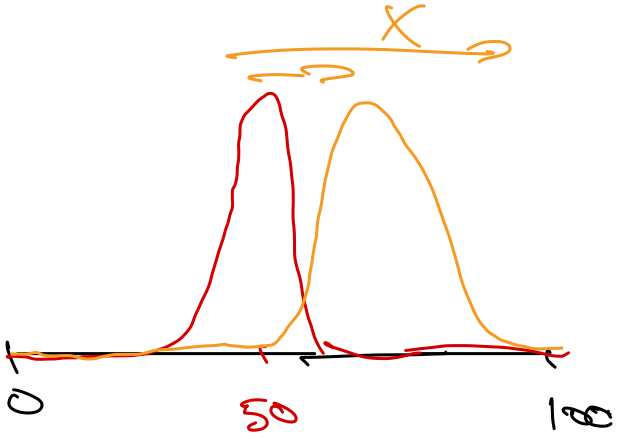
where $\pi(\beta)$ is a prior distribution, $f(\beta|y)$ is posterior distribution.

$\pi(\beta)$ represents our beliefs about β before seeing any data.

Ex

β = average score on exam





get data ⇒ everybody > 90%

same data