Regularized regression

Readings for today

- Chapter 6: Linear model selection and regularization. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer

Topics

1. Ridge regression
2. LASSO
3. Elastic net

Ridge regression

Dimensionality

<u>Dimensionality of a model:</u> n x p

As n →p, dimensionality increases & model variance increases

$$\begin{pmatrix} x_{1,1} \\ \cdots \\ x_{n,1} \end{pmatrix} \rightarrow \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,15} \\ \cdots & & & \\ x_{n,1} & x_{n,2} & \cdots & x_{n,15} \end{pmatrix} \rightarrow \uparrow \text{model flexibility}$$

How do you select (or remove) variables in the fitting process itself?

(James et al. 2013)

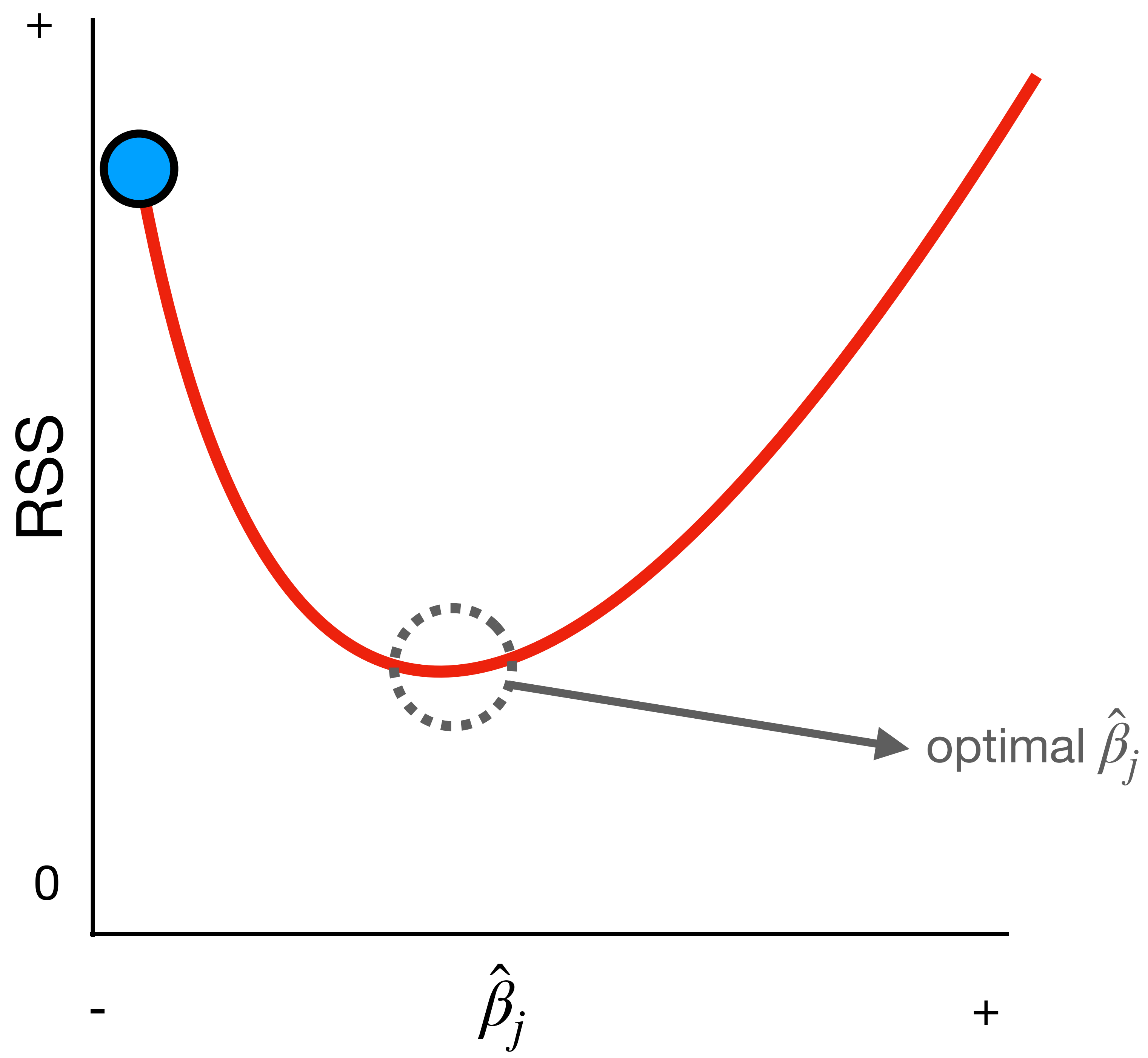
OLS Objective

Residual Sums of Squares (RSS):

$$
\begin{aligned}
RSS &= e_1^2 + \ldots + e_n^2 \\
&= (y_1 - \sum_{j=1}^{p} \hat{\beta}_j x_{1,j})^2, \ldots, \sum_{j=1}^{p} \hat{\beta}_j x_{n,j})^2 \\
&= \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})^2
\end{aligned}
$$

Objective: $\min(\log(L)) = min(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2)$

$$= min(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})^2)$$



(James et al. 2013)

Ridge regression

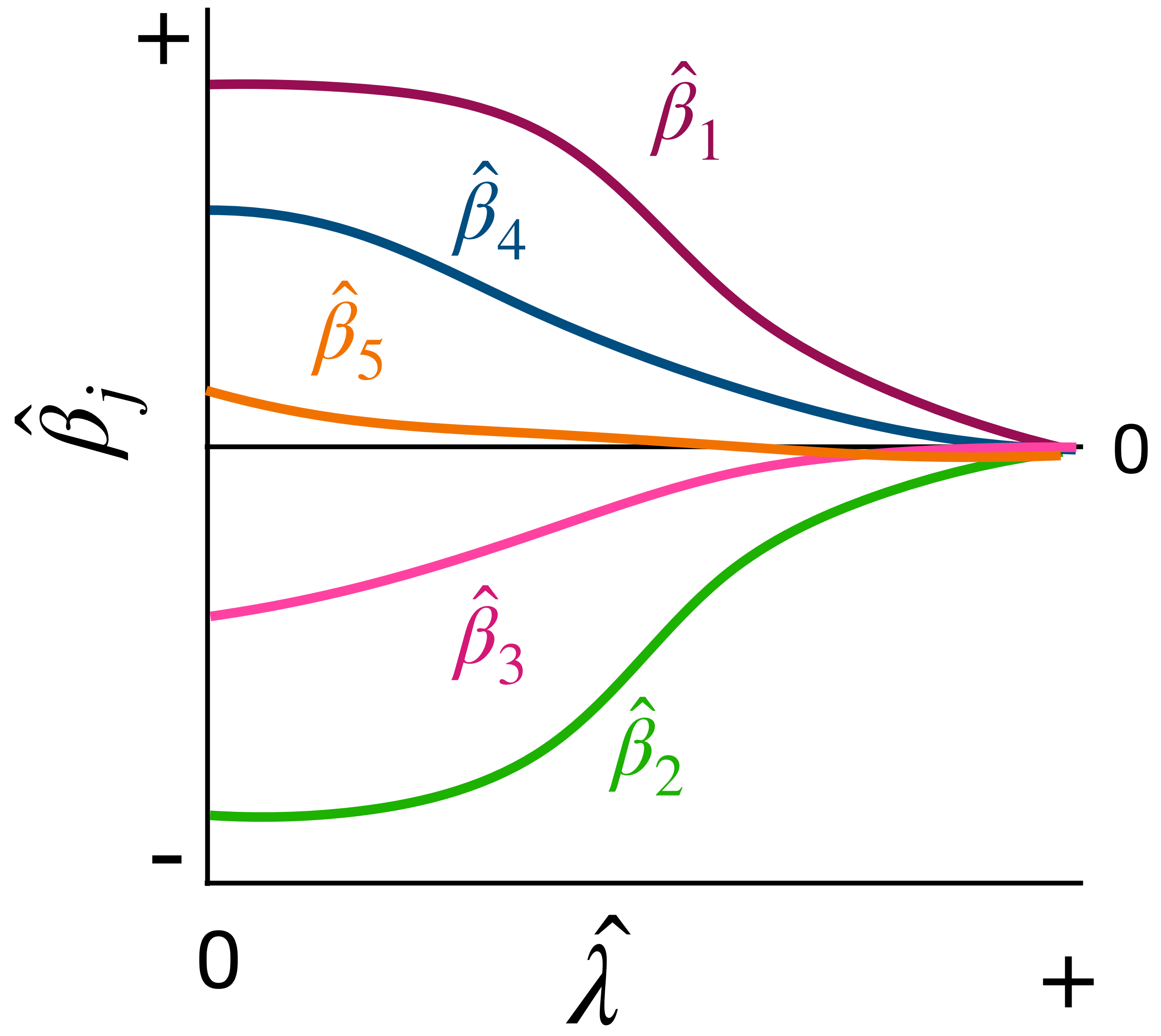
Goal: Penalize "weak" effects so they have minimal impact on your model.

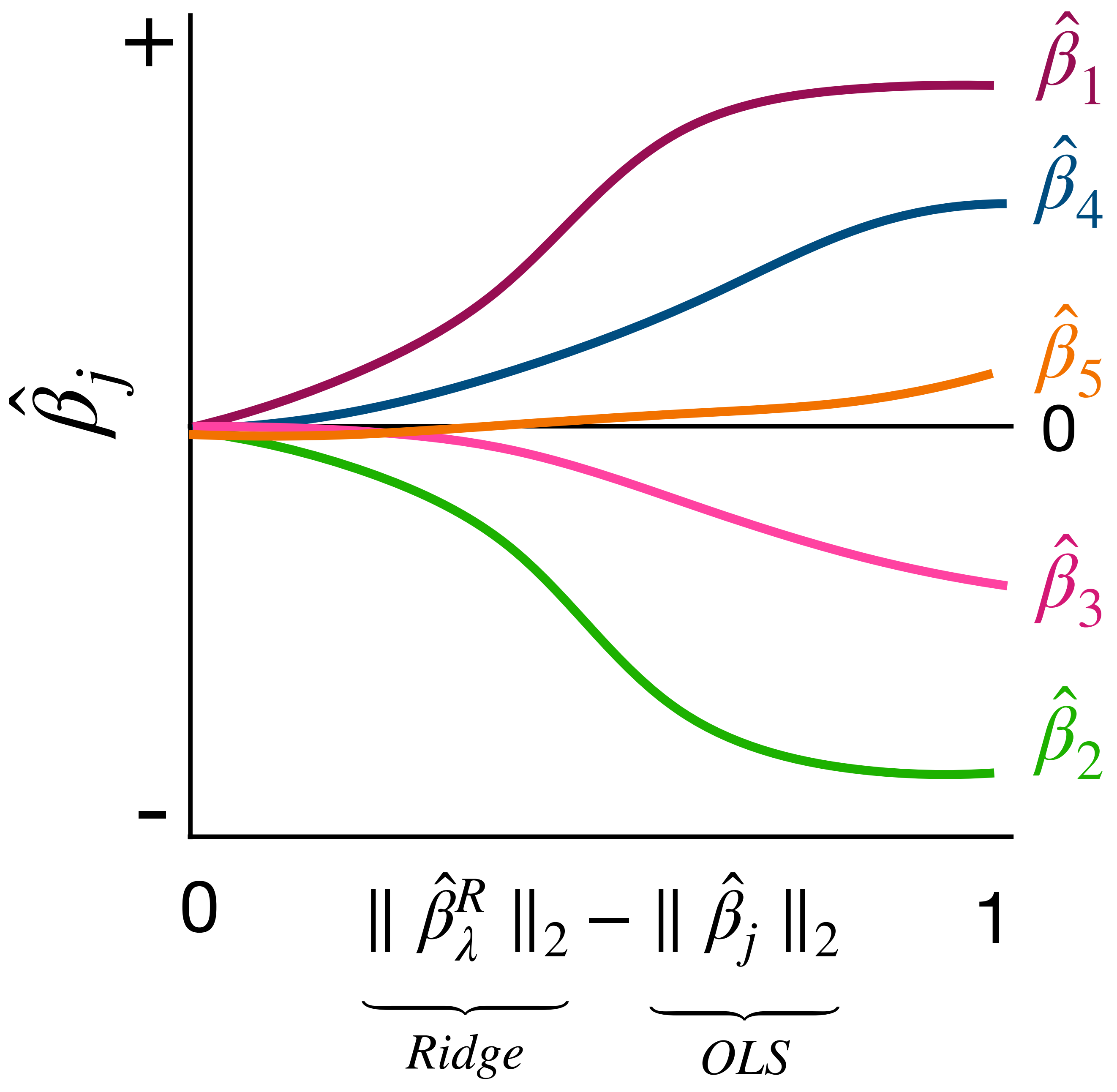
Objective: $$\min(\log(L)) = min(\underbrace{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\hat{\beta}_j x_{i,j})^2}_{\text{minimized residual squared error}} + \underbrace{\hat{\lambda}\sum_{j=1}^{p}\hat{\beta}_j^2}_{\text{shrinkage penalty}})$$

$\hat{\lambda}$ = sparsity parameter

(James et al. 2013)

Effects on coefficients





- Ridge will *always* return all p predictors.
- $\uparrow \lambda$ results in weaker $\hat{\beta}_j's$.
- Ridge manages the bias-variance tradeoff by reducing the influence of weak predictor variables.

Does *not* select features!

(James et al. 2013)

Finding λ

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \\ y_{m+1} \\ \vdots \\ y_n \end{pmatrix} \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & & \vdots \\ x_{m,1} & \cdots & x_{m,p} \\ x_{m+1,1} & \cdots & x_{m+1,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & & \vdots \\ x_{m,1} & \cdots & x_{m,p} \end{pmatrix}$$

Test set

Training set

$$\begin{pmatrix} y_{m+1} \\ \vdots \\ y_k \\ y_{k+1} \\ \vdots \\ y_n \end{pmatrix} \begin{pmatrix} x_{m+1,1} & \cdots & x_{m+1,p} \\ \vdots & & \vdots \\ x_{k,1} & \cdots & x_{k,p} \\ x_{k+1,1} & \cdots & x_{k+1,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

Tune $\hat{\lambda}$

$$\begin{pmatrix} y_{m+1} \\ \vdots \\ y_k \end{pmatrix} \begin{pmatrix} x_{m+1,1} & \cdots & x_{m+1,p} \\ \vdots & & \vdots \\ x_{k,1} & \cdots & x_{k,p} \end{pmatrix}$$

Fit $\hat{\beta}_j's$

$$\begin{pmatrix} y_{k+1} \\ \vdots \\ y_n \end{pmatrix} \begin{pmatrix} x_{k+1,1} & \cdots & x_{k+1,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

$\hat{\lambda}$ is a free parameter that needs to be fit

(James et al. 2013)

Example: high variance model

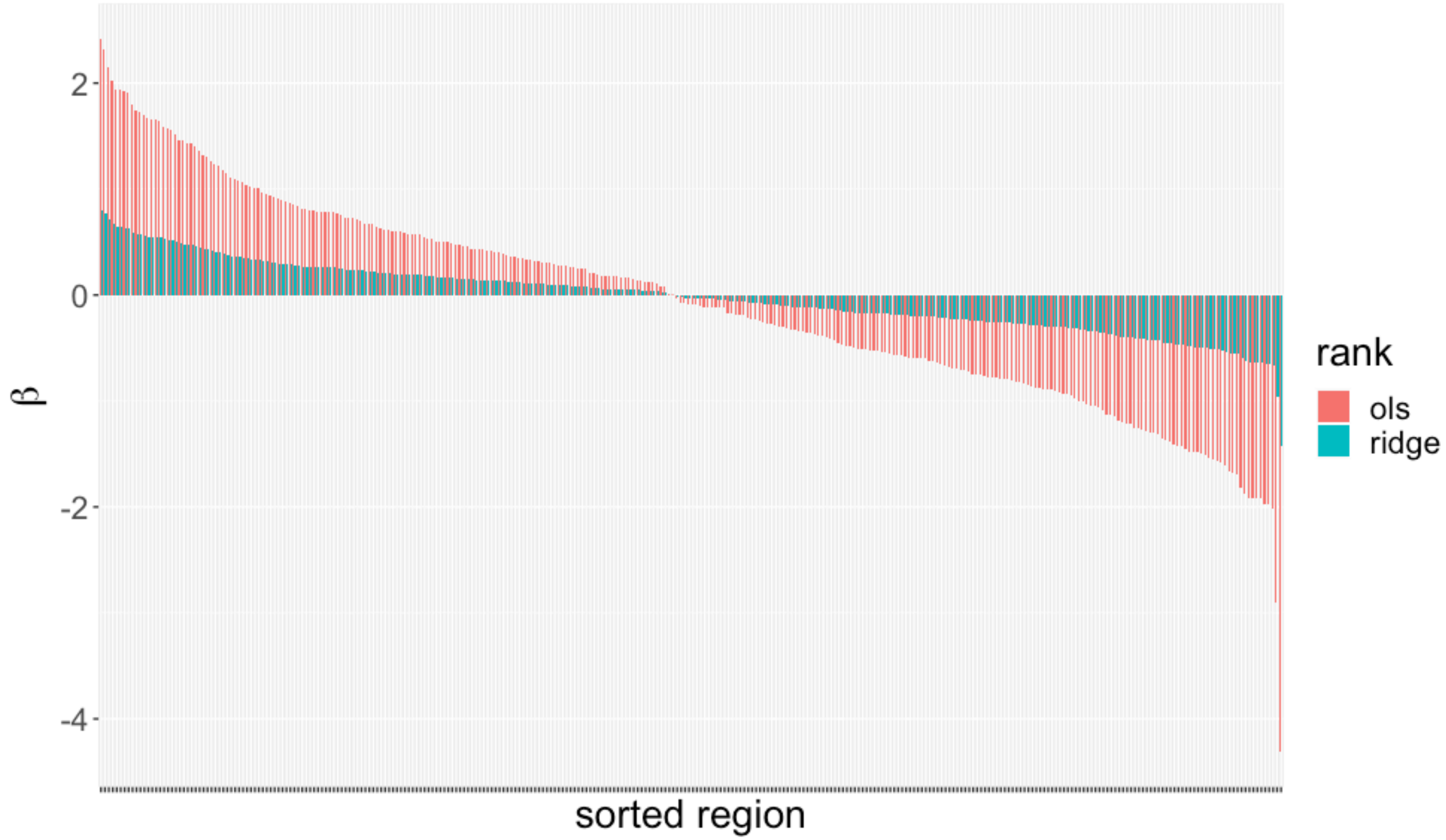
<u>Q:</u> What brain areas predict reaction times across trials?

$$R = 300$$
$$n = 500$$

$$Y_{rt,t} = \hat{\beta}_0 + \sum_{j=1}^{R} \hat{\beta}_j X_{t,j}$$

$Y_{rt,t}$: reaction time on trial t

$X_{j,t}$: fMRI response of region j on trial t



(James et al. 2013)

The pros and cons of Ridge regression

Advantages:

- Manages the bias-variance tradeoff well when $\frac{p}{n}$ is very high.
- Best solution for high variance models.

Disadvantages:

- Does *not* actually do feature selection (simply pushes $\hat{\beta}_j's$ close to zero).
- Is not scale invariant
  - OLS: $cX \rightarrow c\hat{\beta}$
  - Ridge: $cX \neq c\hat{\beta}$

(James et al. 2013)

LASSO

Least Absolute Shrinkage & Selection Operator (LASSO)

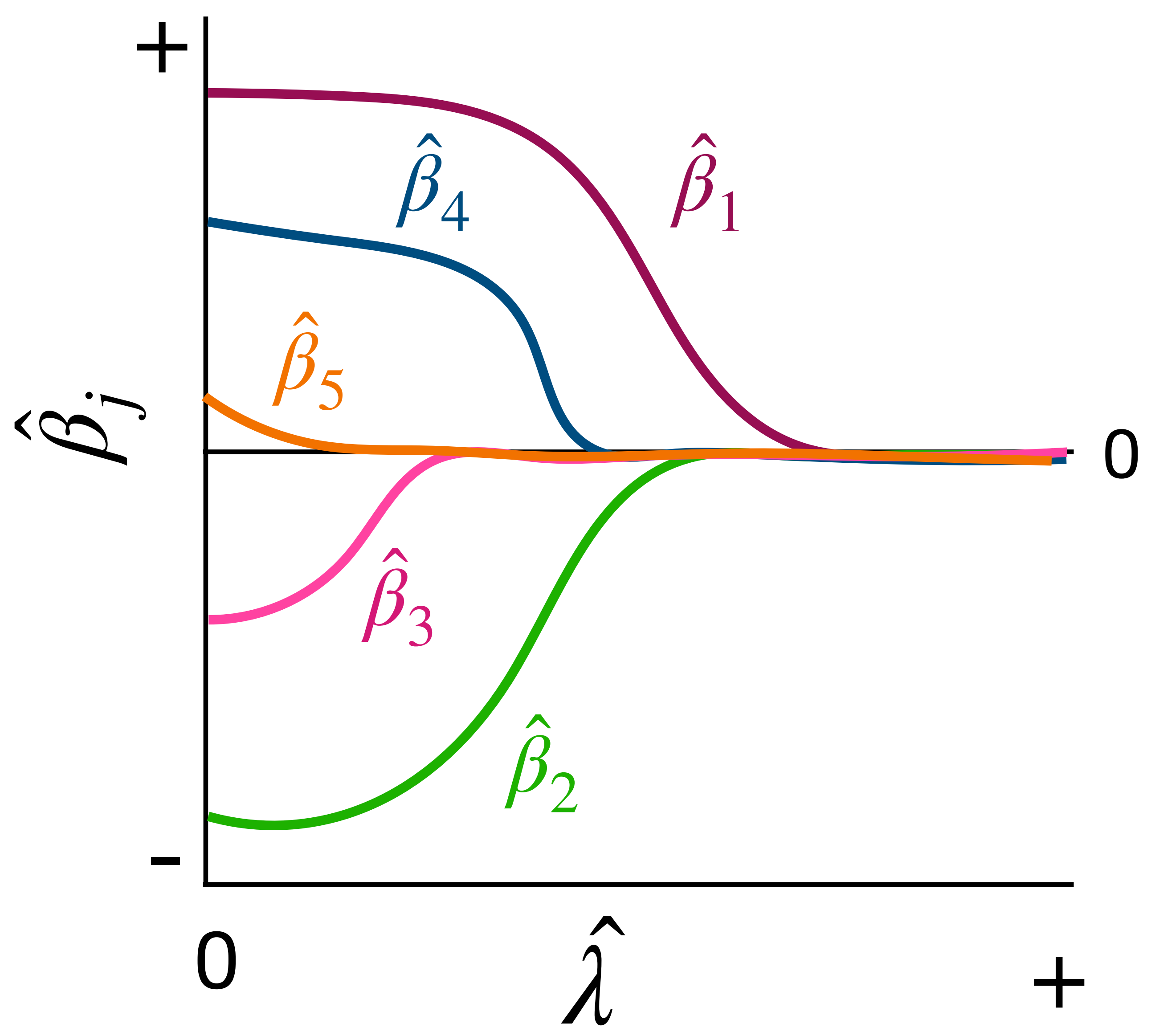
Goal: Remove "weak" effects altogether

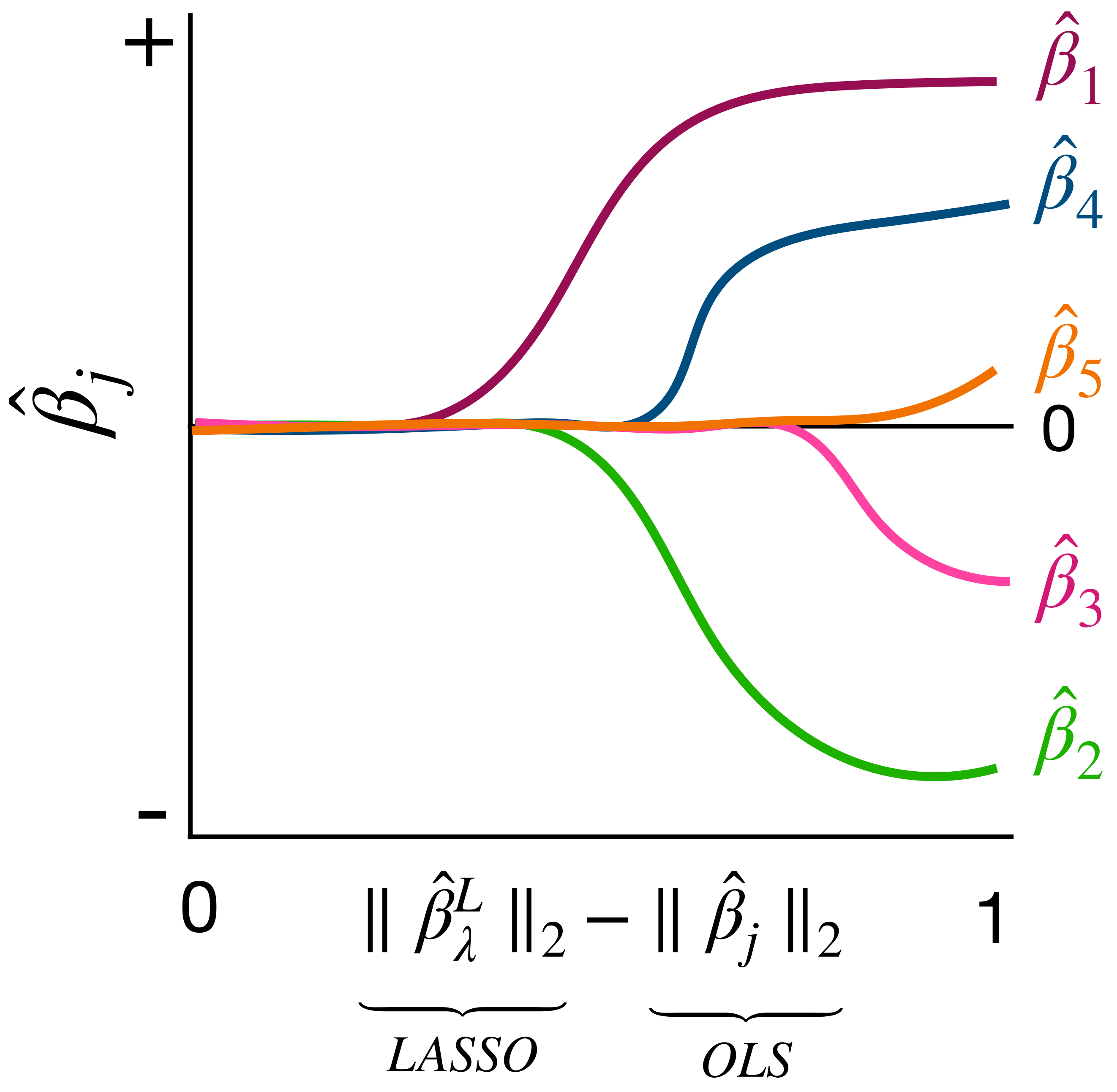
Objective: $$\min(\log(L)) = min(\underbrace{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\hat{\beta}_j x_{i,j})^2}_{\text{minimized residual squared error}} + \underbrace{\hat{\lambda}\sum_{j=1}^{p}|\hat{\beta}_j|}_{\text{shrinkage penalty}})$$

$\hat{\beta}_j's$ can now shrink to 0.

(James et al. 2013)

Effects on coefficients





- LASSO returns $< p$ predictors with $\lambda > 0$.
- $\uparrow \lambda$ results in more $\hat{\beta}_j's = 0$.
- LASSO does feature selection in the model fit process by setting coefficients to zero.

(James et al. 2013)

Symmetry: Ridge vs. LASSO

Assume:

$$X = I = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \quad p = n$$

OLS:

$$\min(\log(L)) = min(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})^2)$$

$$\hat{\beta}_i = y_i$$

Ridge:

$$\min(\log(L)) = min(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})^2 + \hat{\lambda} \sum_{j=1}^{p} \hat{\beta}_j^2)$$

$$\hat{\beta}_i^R = \frac{y_i}{1 + \hat{\lambda}}$$

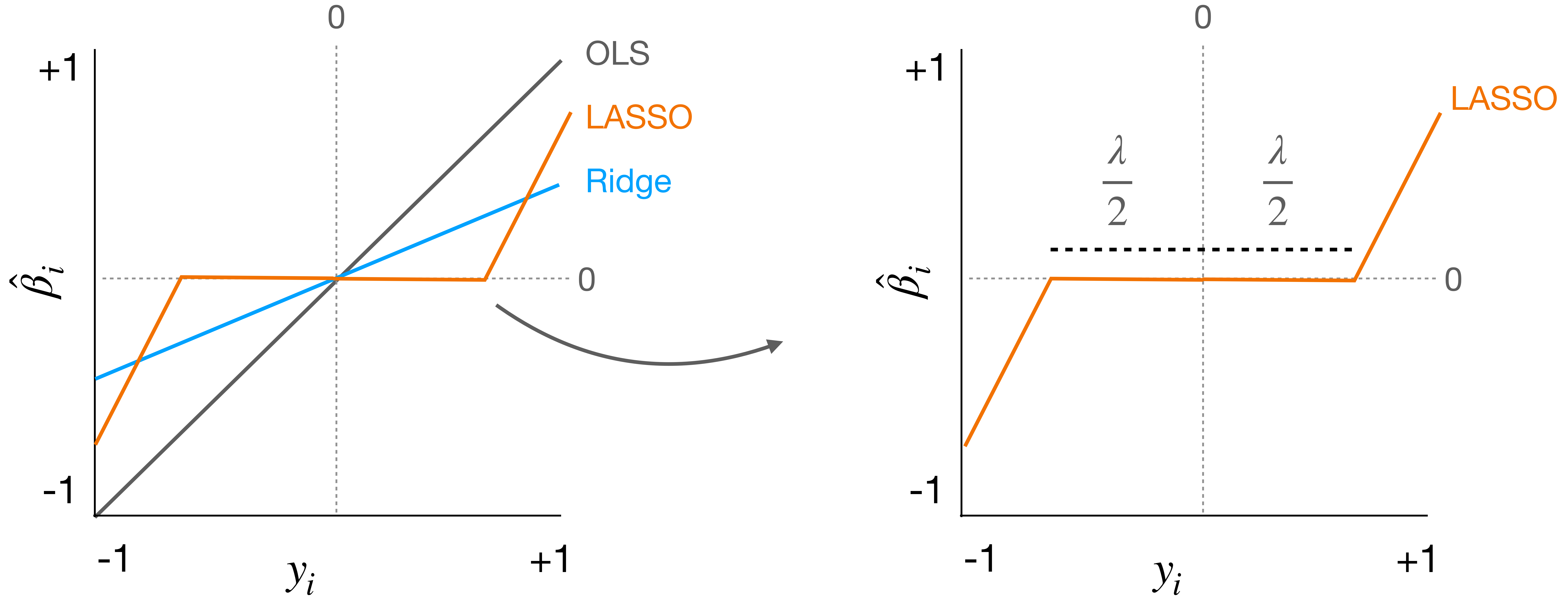
LASSO:

$$\min(\log(L)) = min(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})^2 + \hat{\lambda} \sum_{j=1}^{p} |\hat{\beta}_j|)$$

$$\hat{\beta}_i^L = \begin{cases} y_i - \frac{\hat{\lambda}}{2}, & \text{if } y_i > \frac{\hat{\lambda}}{2} \\ 0, & \text{if } |y_i| \leq \frac{\hat{\lambda}}{2} \\ y_i + \frac{\hat{\lambda}}{2}, & \text{if } y_i < \frac{\hat{\lambda}}{2} \end{cases}$$

(James et al. 2013)

Symmetry: Ridge vs. LASSO



(James et al. 2013)

Example: high variance model

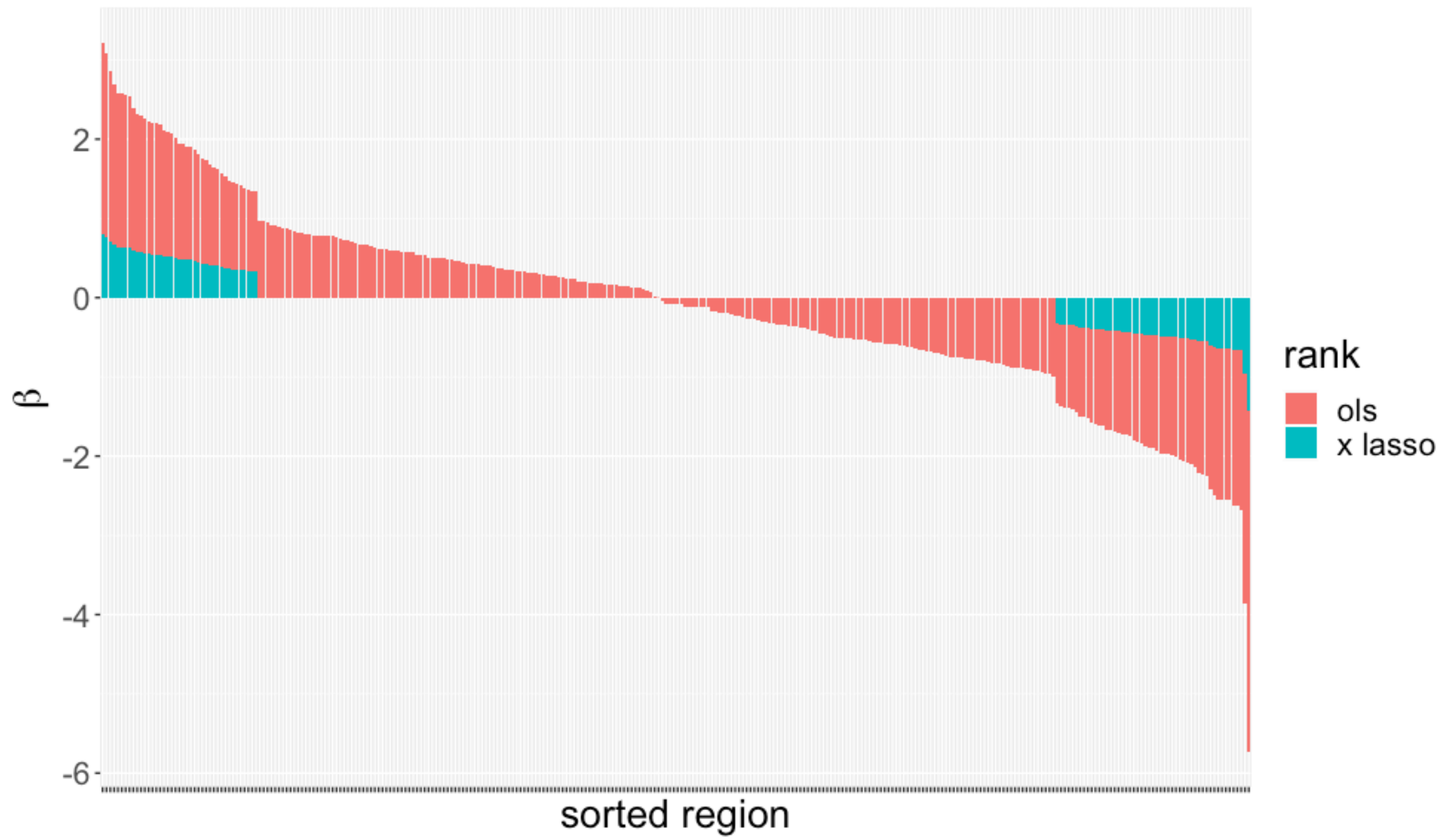
Q: What brain areas predict reaction times across trials?

$$Y_{rt,t} = \hat{\beta}_0 + \sum_{j=1}^{R} \hat{\beta}_j X_{t,j}$$

$$R = 300$$
$$n = 500$$

$Y_{rt,t}$: reaction time on trial t

$X_{j,t}$: fMRI response of region j on trial t



(James et al. 2013)

Best subset as LASSO

LASSO can implement the best subset algorithm with a change in the sparsity penalty.

<u>Best subset penalty:</u>

$$\min(\log(L)) = min(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\hat{\beta}_j x_{i,j})^2 + \hat{\lambda}\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0))$$

$$I(\hat{\beta}_j \neq 0) = \begin{cases} 1, & \text{if } \hat{\beta}_j \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

(James et al. 2013)

Ridge vs. LASSO vs. Subset

1. LASSO picks the “best” of any correlated/clustered set of predictor variables, X.

2. Ridge picks the strongest subset of correlated/clustered predictor variables, X.

3. LASSO is more conservative than Ridge or best subset selection.

(James et al. 2013)

Elastic net

Elastic net

Goal: Little bit of ridge and little bit of LASSO

Objective: $$\min(\log(L)) = min(\underbrace{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\hat{\beta}_j x_{i,j})^2}_{\text{minimized residual squared error}} + \underbrace{\hat{\alpha}\hat{\lambda}\sum_{j=1}^{p}\hat{\beta}_j^2}_{\text{partial ridge penalty}} + \underbrace{(1-\hat{\alpha})\hat{\lambda}\sum_{j=1}^{p}|\hat{\beta}_j|}_{\text{partial LASSO penalty}})$$

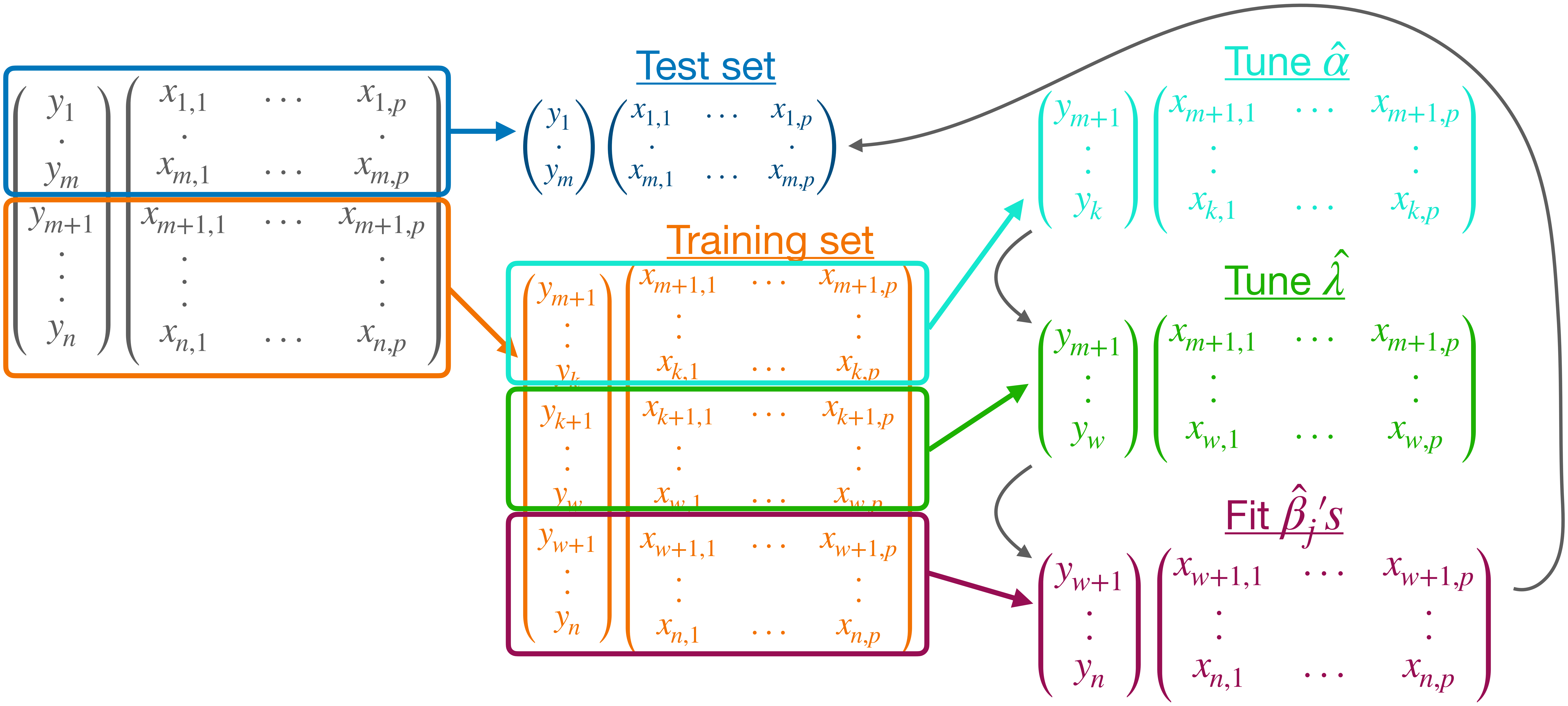
Ridge: $\alpha = 1$

α = mixing parameter → LASSO: $\alpha = 0$

Elastic Net: $0 < \alpha < 1$

(James et al. 2013)

Finding α



(James et al. 2013)

Take home message

- Regularized regression models are a fast, computationally efficient way of managing model complexity by implementing feature selection (or sparsity) in the model fitting process itself.