

# Outline: (Linear) Models in Large Feature Spaces

## Traditional Feature Selection

(F. Chiaromonte)

Introduction to Statistical Learning  
Chapter 6 Section 1, Lab 1

## Traditional (“hard”) feature selection: (most) **NON-CONVEX** constrained optimization

L0 norm counts non-0 coefficients, but size of each non-0 coefficient is unconstrained.

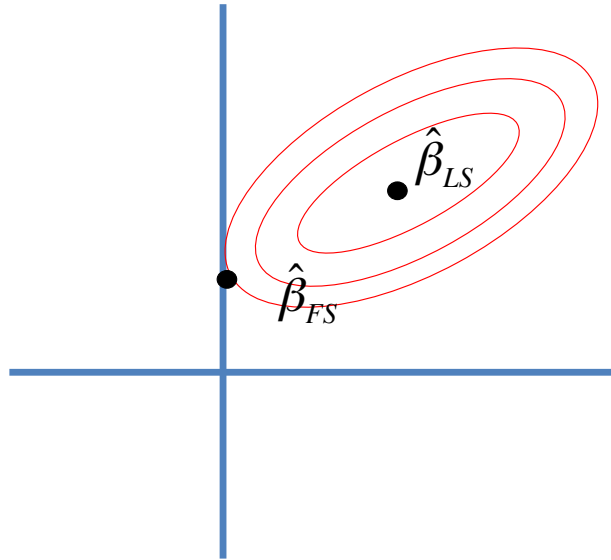
Much harder, previously not computationally viable; now *Mixed Integer Optimization*.

### Some (non-traditional) references:

- Bertsimas D., King A., Mazumder R. (2016) Best subset selection via a modern optimization lens. AOS 44(2) 813-852.
- Kenney A., Chiaromonte F., Felici G. (2020) MIP-BOOST: Efficient and Effective  $L_0$  Feature Selection for Linear Regression. JCGS.

### Important remarks on the potential of MIP:

- Further integer constraints to capture structure, e.g., feature groups.
- Incorporate election of units/observations to robustify the fit; Insolia L., Kenney S., Chiaromonte F., Felici G. (2021) Simultaneous feature selection and outlier detection with optimality guarantees. Biometrics.



$$\| \underline{Y} - \underline{X}\beta \|^2 = \min_{\beta \in \mathbb{R}^p}$$

$$\sum_{j=1}^p \text{Ind}(\beta_j \neq 0) \leq c \quad \boxed{\text{size constraint}}$$

SIZE CONSTRAINT TUNED BY CROSS VALIDATION  
... the traditional **Best Subset Selection** problem

Back to traditional statistical approaches:

- **Feature (Subset) Selection**

- Best Subset Selection
- Step-wise (sequential) Selection

Followed by LS fit of the (smaller) model comprising the selected features.

- **Dimension Reduction**

- Principal Components (unsupervised reduction)
- Sufficient Dimension Reduction (supervised reduction)

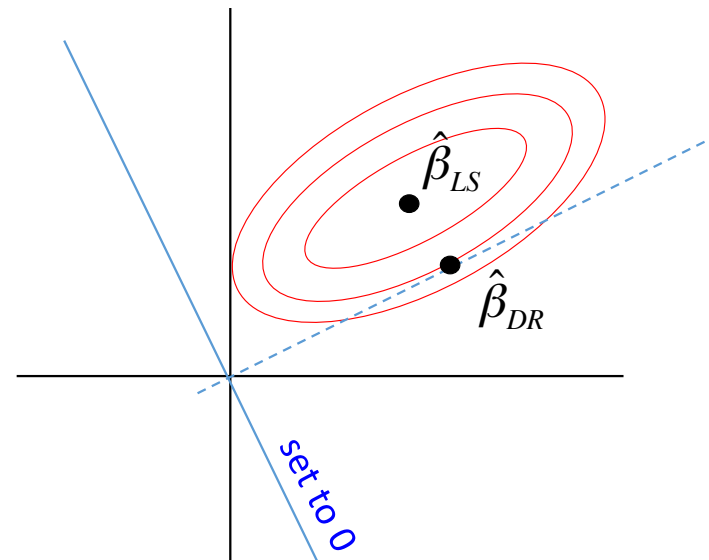
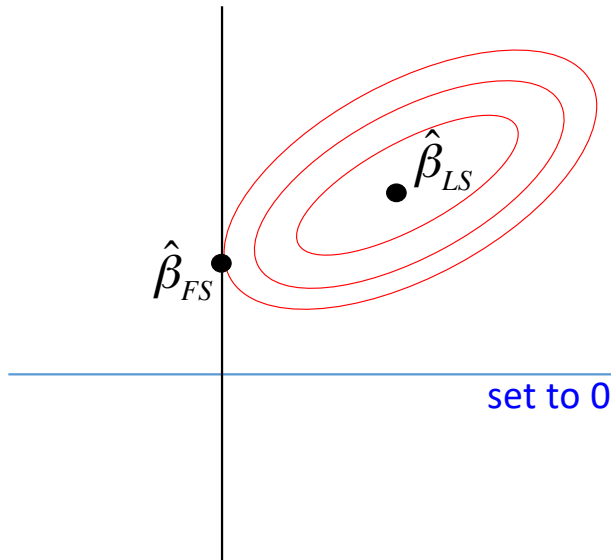
Followed by LS fit of the (smaller) model comprising the selected linear combinations

Also these can be thought of in the framework of constrained LS: *Linear constraints* to force  $\beta$  in a coordinate space, or a generic linear subspace, of  $\mathbb{R}^p$ . ***BUT WE NEED AN ADDITIONAL “INGREDIENT”!***

=====

Dimension Reduction: ISLR also describes **Partial Least Squares**. But does not describe Sufficient Dimension Reduction techniques.

Cartoons with  $p=2$  and  $c=1$  ( $d=1$ )



$$\| \underline{Y} - \underline{X}\beta \|^2 = \min_{\beta \in \mathbb{R}^p}$$

$$H_{FS}\beta = 0_{(p-c)}$$

linear constraints

$$\| \underline{Y} - \underline{X}\beta \|^2 = \min_{\beta \in \mathbb{R}^p}$$

$$V_{DR}\beta = 0_{(p-d)}$$

THE (maximal) **COUNT** OF FEATURES  $c$  OR THE **DIMENSION**  $d$  IS FIXED... SO IS THE **(O.N.) BASIS** OF THE FEATURE SPACE FROM WHICH WE FORM THE  $(p-c)$  or  $(p-d)$  CONSTRAINTS.

*Feature Selection: How do we produce a  $p \times p$  matrix  $H$  expressing an O.N. basis in the feature space?*

(we then focus on its last  $(p-c)$  rows to create the linear constraints with  $H_{FS}$ )

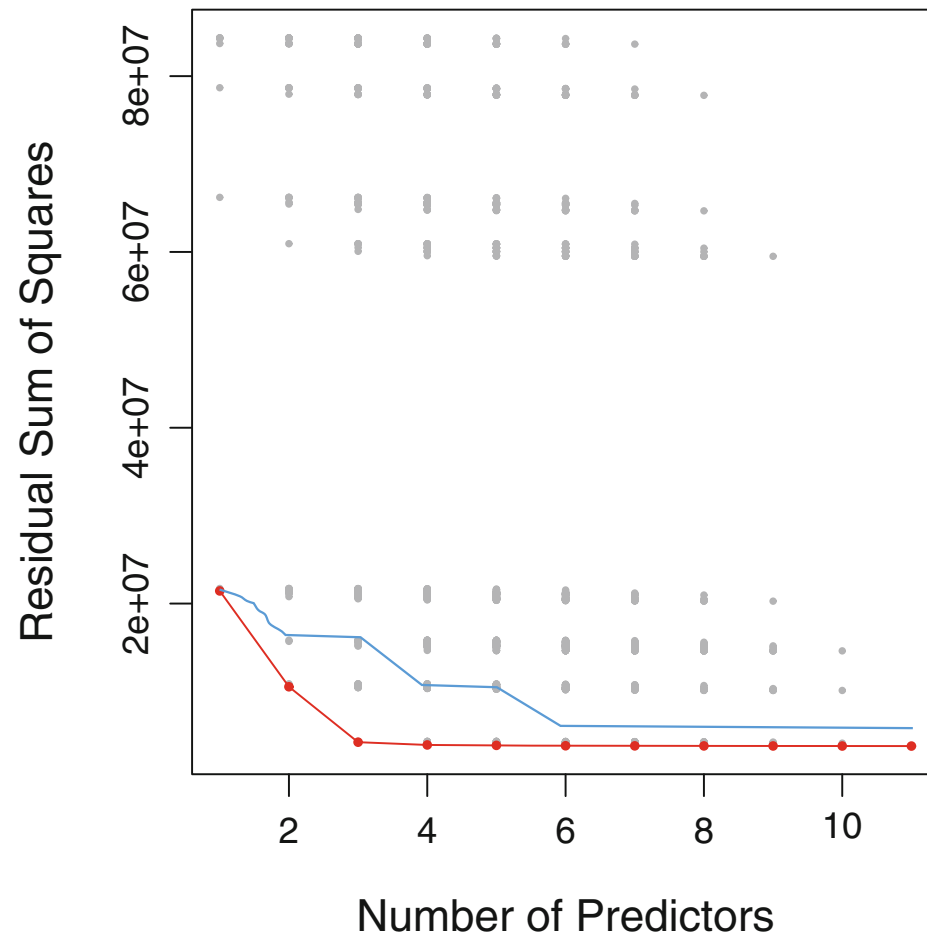
Feature (subset) Selection:

Make a sequence of nested subsets adding one feature at a time, i.e., make an ordering of the features.

This corresponds to a permutation of the elements of the Canonical O.N. basis  $\{e_1, \dots, e_p\}$ , and provides

**$H$  (a permutation matrix)**

Rigorous for feature selection implemented through **stepwise (sequential) methods**, not for **best subset methods** which find the best subset of each size – in principle such subsets may not be nested (but the intuition still works).



The RSS of  $2^p$  possible models (subsets of features).

For each size  $c$ , there are  $\binom{p}{c}$  possible models.

On the lower boundary (red) are the best models of each size; best subsets. May or may not be a nested sequence.

The blue path represents a nested sequence of “good” models obtained stepwise. May depart from the lower boundary.

**Algorithm 6.1** *Best subset selection*

1. Let  $\mathcal{M}_0$  denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For  $k = 1, 2, \dots, p$ :
  - (a) Fit all  $\binom{p}{k}$  models that contain exactly  $k$  predictors.
  - (b) Pick the best among these  $\binom{p}{k}$  models, and call it  $\mathcal{M}_k$ . Here *best* is defined as having the smallest RSS, or equivalently largest  $R^2$ .
3. Select a single best model from among  $\mathcal{M}_0, \dots, \mathcal{M}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC, or adjusted  $R^2$ .

**Best Subset** approach requires fitting ALL possible models (“brute” enumeration).

**Algorithm 6.2** *Forward stepwise selection*

1. Let  $\mathcal{M}_0$  denote the *null model*, which contains no predictors.
2. For  $k = 0, \dots, p - 1$ :
  - (a) Consider all  $p - k$  models that augment the predictors in  $\mathcal{M}_k$  with one additional predictor.
  - (b) Choose the *best* among these  $p - k$  models, and call it  $\mathcal{M}_{k+1}$ . Here *best* is defined as having smallest RSS or highest  $R^2$ .
3. Select a single best model from among  $\mathcal{M}_0, \dots, \mathcal{M}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC, or adjusted  $R^2$ .

**Stepwise** (sequential) approaches do not; they explore fewer models. Computationally leaner.

Hybrid stepwise approaches mix forward and backward progression.

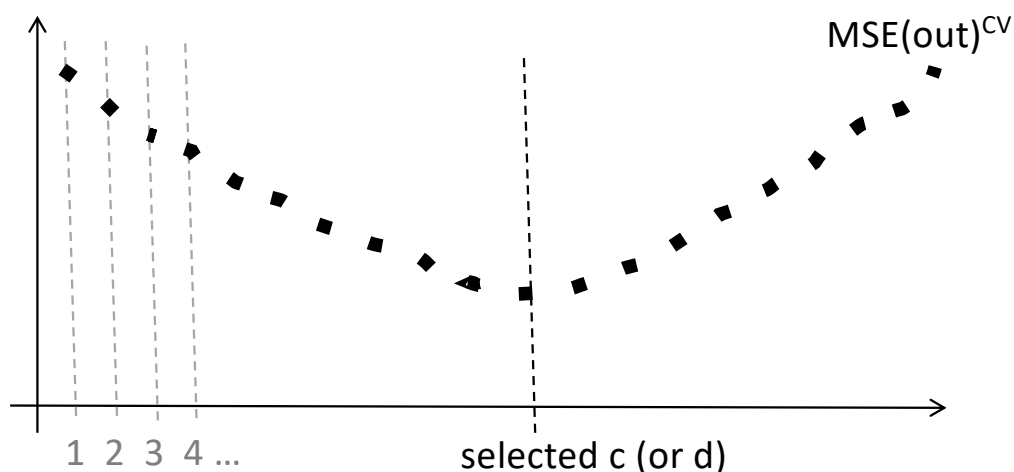
**Algorithm 6.3** *Backward stepwise selection*

1. Let  $\mathcal{M}_p$  denote the *full model*, which contains all  $p$  predictors.
2. For  $k = p, p - 1, \dots, 1$ :
  - (a) Consider all  $k$  models that contain all but one of the predictors in  $\mathcal{M}_k$ , for a total of  $k - 1$  predictors.
  - (b) Choose the *best* among these  $k$  models, and call it  $\mathcal{M}_{k-1}$ . Here *best* is defined as having smallest RSS or highest  $R^2$ .
3. Select a single best model from among  $\mathcal{M}_0, \dots, \mathcal{M}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC, or adjusted  $R^2$ .

(Step 3 in each algorithm; next...)

*How do we select the (maximal) count  $c$ , or dimension  $d$ , along an ordered sequence of features or linear combinations?*

Of course, once  $H$  (or  $V$ ) are fixed we could perform the selection estimating the out-of-sample MSE associated to each value of  $c$  or  $d$ , e.g., by cross-validation (same as for the penalty parameter in Ridge or LASSO).



Historically though other (non-computation based) approaches and criteria have been used.



## CRITERIA FOR FEATURE (SUBSET) SELECTION

The number of features  $c$  ( $= d$  if formulae below) is selected optimizing a criterion that approximates an estimate of the out-of-sample performance – but indirectly, not directly as in cross-validation; no intensive computation required. For instance:

Adjusted  $R^2 = 1 - \frac{\text{RSS}/(n - d - 1)}{\text{TSS}/(n - 1)}$ . **Adjusted R-squared** provides the simplest solution, each sum of squares is divided by its degrees of freedom.

$C_p = \frac{1}{n} (\text{RSS} + 2d\hat{\sigma}^2)$  **Mallow's  $C_p$**  formula in the book, proportional to original formula. If  $\hat{\sigma}^2$  is an unbiased estimate of the error variance,  $C_p$  is an unbiased estimate of the out-of-sample MSE (can take the  $s^2$  estimate from the full model, least likely to carry bias).

Strike a balance between RSS (error on the training data) and  $d$  (size, complexity of the model)

**Akaike and Bayes Information Criteria** can be used for model selection whenever models are fitted by Maximum Likelihood.

For linear models with additive Gaussian errors their expressions become:

$$AIC = \frac{1}{n\hat{\sigma}^2} (RSS + 2d\hat{\sigma}^2) \quad \text{Behaves exactly like } C_p \text{ here!}$$

$$BIC = \frac{1}{n} (RSS + \log(n)d\hat{\sigma}^2) \quad \text{The weight of the adjustment uses } \log(n), \text{ depending on the sample size, instead of 2. For } n > 7 \text{ BIC has a steeper penalty than AIC } (C_p); \text{ it favors smaller models.}$$

General equations for Information Criteria:

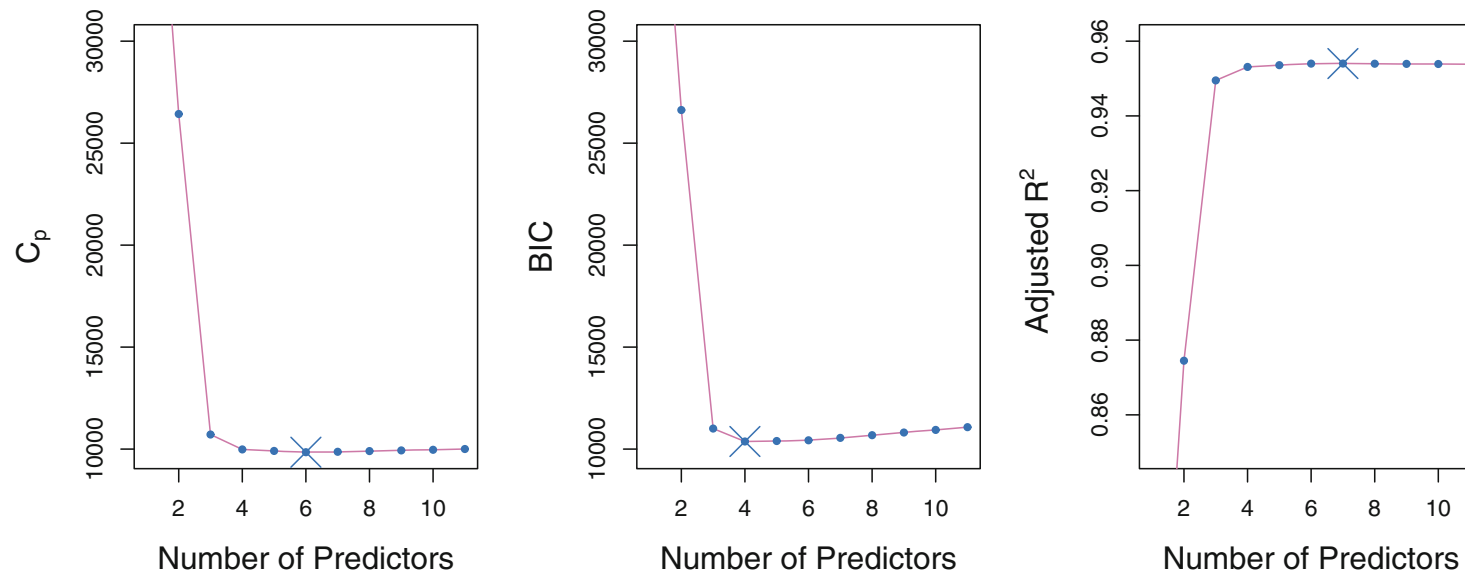
$$AIC = -2 \log L(\hat{\vartheta}_{ML}) + 2d$$

$$BIC = -2 \log L(\hat{\vartheta}_{ML}) + \log(n)d$$

$$L(\hat{\vartheta}_{ML}) = \text{optimal likelihood value on MLE of parameter vector}$$

$$d = \# \text{ of free parameters being estimated}$$

## HOW DO THE TRADITIONAL CRITERIA LOOK IN PRACTICE?



This behavior is rather common; even though the criteria are NOT monotone and ought to have a minimum/maximum, in many applications they do not clearly discriminate unambiguously the number of features (the model) to be used – rather flat ranges.

Would you pick 3 instead of the “technical” optima, which btw differ across criteria?