Outline: Linear Models in Large Feature Spaces, Ridge and LASSO
(F. Chiaromonte)

Introduction to Statistical Learning

(Linear Models review: Chapter 3)

Ridge & LASSO: Chapter 6 Sections 2, Lab 6 parts 1,2,3

A QUICK REVIEW OF SOME CONCEPTS AND NOTATION CONCERNING LINEAR MODELS

A classical **linear model** expresses a continuous response Y through an additive function comprising

- an intercept (constant)
- p terms (linear in the slope parameters)
- a random error independent from the random mechanisms underlying the terms.

On a generic observation "i":

$$y_i = \beta_o + \beta' x_i + \varepsilon_i$$

Terms: X = $(X_1 \ldots X_p)$' features (predictors) or specified transformations and functions of the features (e.g. powers, products; observable).

Coefficient parameters: β_o intercept, $\beta = (\beta_1 \ldots \beta_p)$' slopes for each term.

Regression function: the conditional expected value $E(Y|X_1 \ldots X_p) = \beta_o + \beta'X$

Independent random error: added to the regression function and assumed to have mean $E(\varepsilon)=0$ and the same variance parameter $var(\varepsilon)=\sigma^2$ on all observations (homoschedasticity).

In order to develop **inferential procedures**, the random error is also often assumed to be Gaussian:

$$\varepsilon \sim N_1(0,\sigma^2)$$

(T-based confidence intervals and tests for the slope coefficients and the mean response, T-based prediction intervals, F-based ANOVA tests).

The observations are assumed to be iid from the joint distribution of ($Y, X_1 \ldots X_p$). If $X_1 \ldots X_p$ are viewed as fixed (conditioning; not random), the errors are assumed to be iid across observations. Thus one has, for the vector of errors associated to the n observations in the sample:

$$\underline{\varepsilon}_{(n)} \sim N_n(0,\sigma^2 I)$$

Model in **matrix notation**:

$$\underline{Y}_{(n)} = \underline{1}_{(n)}\beta_o + \underline{X}_{(n,p)}\beta + \underline{\varepsilon}_{(n)}$$

$$\text{for simplicity } \underline{Y}_{(n)} = \underline{X}_{(n,p+1)}\beta + \underline{\varepsilon}_{(n)}$$

$$\underline{X}_{(n,p+1)} = (\underline{1}_{(n)}\ \underline{X}_{(n,p)}) \quad \beta = \begin{pmatrix} \beta_o \\ \beta \end{pmatrix}$$

Estimating model parameters (fitting):

Because of linearity in the coefficient parameters, whatever the terms represent, fitting can be performed through **least squares** with an explicit, close form solution.

An estimate of the error variance is obtained dividing the minimized sum of squared deviations (Residual Sum of Squares; RSS) by the appropriate number of degrees of freedom.

$$\hat{\beta} = \text{argmin} \| \underline{Y} - \underline{X}\beta \|^2 = (\underline{X}' \underline{X})^{-1} \underline{X}' \underline{Y}$$

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \| \underline{Y} - \underline{X}\hat{\beta} \|^2 = \frac{1}{n-(p+1)} RSS$$

Implemented in most statistical software packages, including R. See https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html

---

* Intercept = “slope” of the constant term 1 (or assume response is centered, intercept = 0)

If the model is correct (i.e. not misspecified, though some of the coefficients may in fact be 0) the LS coefficient estimators, as well as the LS estimator of the error variance, are **unbiased** for the corresponding parameters.

$$E(\hat{\beta}) = \beta \qquad E(\hat{\sigma}^2) = \sigma^2$$

Moreover, again if the model is not misspecified, the LS coefficient estimators are **BLUE** i.e. the best (most accurate; minimum sampling variance) among unbiased estimators expressed as linear functions of the response observations – **Gauss-Markov Theorem**.

Finally, if one assumes Gaussian error, the LS coefficient estimators coincide with the **Maximum Likelihood** (ML) estimators – and the ML estimator for the error variance, which is biased, divides the minimized sum of squares by n instead of the degrees of freedom.

Why? The exponent of the Gaussian likelihood is inversely proportional to the sum of squared deviations:

$$L(\beta \mid \underline{Y}; \underline{X}) \propto \exp\left\{-\frac{1}{2\sigma^2} \| \underline{Y} - \underline{X}\beta \|^2\right\}$$

Residuals diagnostics and **remedial measures:**

Lots of techniques that, based on residuals (observable proxies for the unobservable errors) aim to detect departures from

$$\underline{\varepsilon}_{(n)} \sim N_n(0,\sigma^2 I)$$

i.e. Gaussian iid random errors with 0 mean and shared variance σ^2.

Lots of approaches to manipulate the data, e.g.:

- variance stabilizing transformations of the response to address heteroschedasticity
- identification and removal of outliers/influential observations (*case diagnostics*)

and the model specification, e.g.:

- add/remove terms to address mean patterns in the residuals

to come closer to meeting the assumptions.

Also important, **Multicollinearity**:

Linear dependencies among predictors increase the variance in effect estimates, may even make LS solution unstable.

Diagnose through

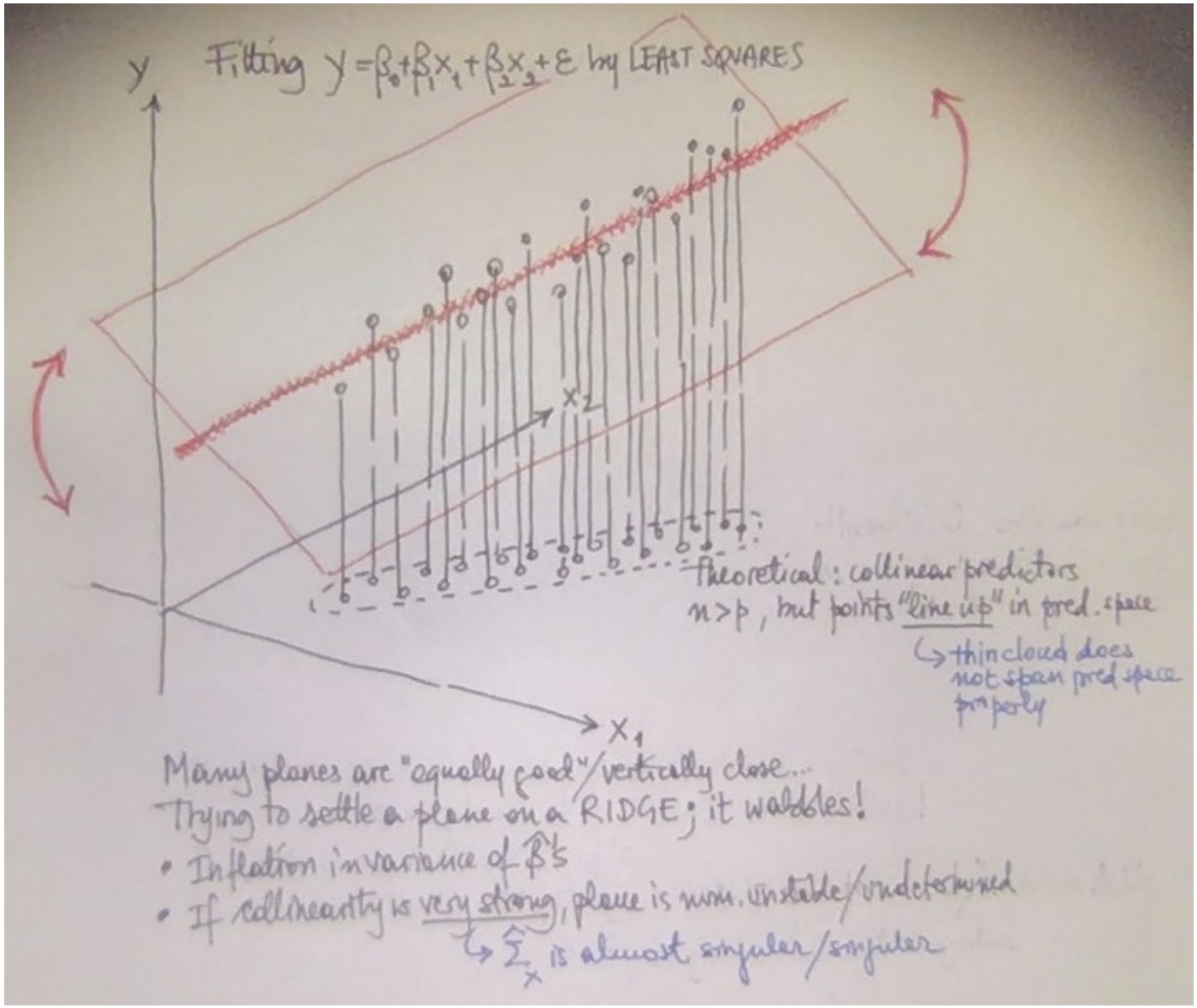
- Pair-wise correlations or scatter-plot matrices
- More complete, **partial R-squares** and **Variance Inflation Factors** (do not depend on the response)

$R^2_{X_j|\ other\ Xs}$ $\qquad\qquad$ $VIF_j = \dfrac{1}{1 - R^2_{X_j|\ other\ Xs}}$

Remedies

- stabilize the fit (Ridge)
- eliminate some predictors (LASSO, Best Subsets)
- reduce dimension creating “composite” predictors

… coming


Fitting $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ by LEAST SQUARES
y
x_1
x_2
theoretical: collinear predictors
$n > p$, but points "line up" in pred. space
↳ thin cloud does not span pred. space for poly
Many planes are "equally good"/vertically close...
Trying to settle a plane on a RIDGE; it wobbles!
• Inflation in variance of $\hat{\beta}$'s
• If collinearity is very strong, plane is num. unstable/undetermined
↳ $\hat{\Sigma}_x$ is almost singular/singular

Generalizations:

- **GENERALIZED LEAST SQUARES**,

Recover BLUE if observations present heteroschedastic and/or correlated errors

- **GENERALIZED LINEAR MODELS**.

Use link functions and different assumptions on the stochastic components underlying the data to create models for non-continuous responses, e.g.:

  - Counts (Poisson regression)
  - Binary labels (logit or Binomial regression; probit regression)
  - Multi-class labels (Multinomial regression)

All (including the standard Normal regression) implemented in the R package **GLM**

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html

We have reviewed a very powerful, rich and versatile framework for supervised data analysis.

Why do we need to go beyond it?

Even when all the assumptions comprised in the modeling are right, so in particular the LS coefficient estimators are unbiased, their accuracy (notwithstanding the Gauss-Markov Theorem) may deteriorate severely when the feature space is large relative to the (iid) data at our disposal...

When does this happen?

When the features (terms) have strong linear associations to each other (multicollinearity) and/or n is not very large relative to p (possibly n<p) the data cloud does not "span" the feature space properly; it is "thin", possibly it collapses in lower dimension.

As a consequence

- The LS coefficient estimators undergo variance inflation; if the data collapses (almost collapses) in lower dimension the estimates are non-uniquely determined (numerically unstable)
- Relatedly, we run into overfitting; the in-sample MSE may be very small, but the out-of-sample MSE is very large.

We will focus on a variety of techniques to overcome this problem.

Overarching idea: *constrain the LS as to reduce the estimators' variance and emiliorate overfitting*.

This *introduces a bias*, but the bias may be minor relative to the gain in variance – so that accuracy overall improves.

Additionally, constraining may result in a smaller, more *parsimonious and interpretable model* (some features are eliminated, or focus shifts to a small number of composite features – linear combinations).

Note: the same approach can be extended to the case of Generalized Linear Models (e.g., a logistic or multinomial regression for binary or multi-class classification, as a measure of accuracy one can considers in-sample and out-of-sample misclassification rates)

- **Shrinkage/Regularization**
  - Ridge Regression
  - LASSO Regression

  Penalized version of LS produces an alternative estimator.

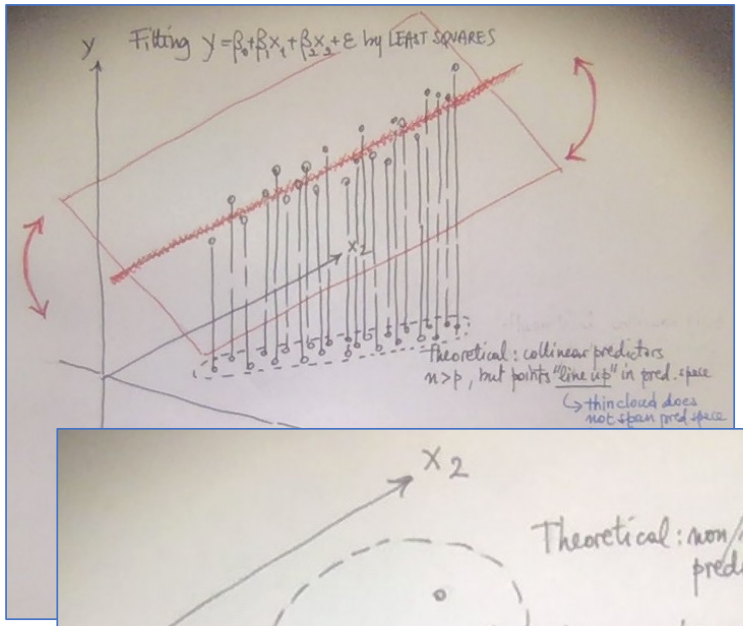
Ridge ('60-'70s): fitting procedure to *overcome multicollinearity* in regressions with highly interdependent features.

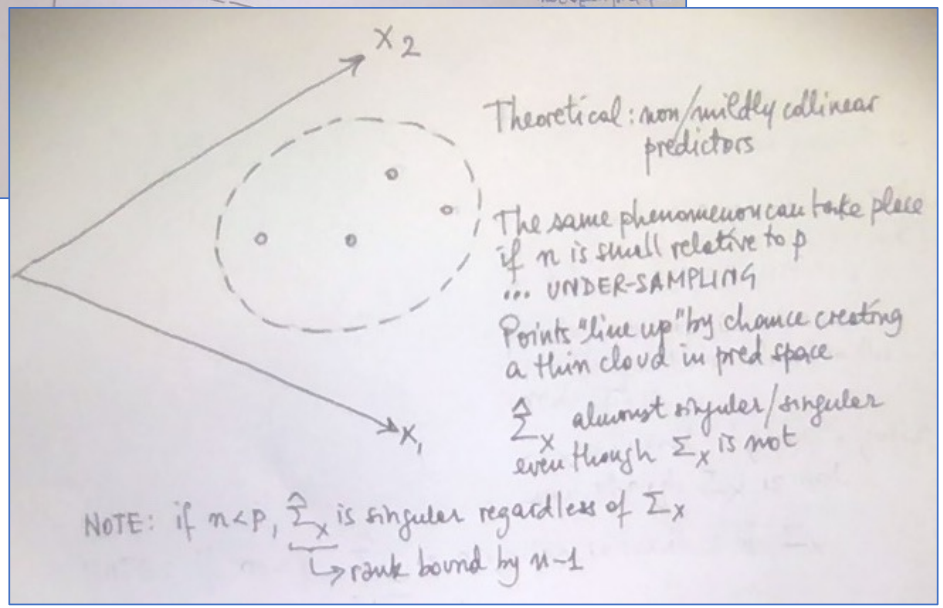
Loosely, in the LS, replaces $S_X=\underline{X}'\underline{X}$ (pxp sample covariance matrix of centered X's) with $S_X(\lambda)=\underline{X}'\underline{X} + \lambda I_p$, $\lambda>0$ which is always invertible – spherically "increasing" the features' spread in the data decreases the sampling variability of the estimator.

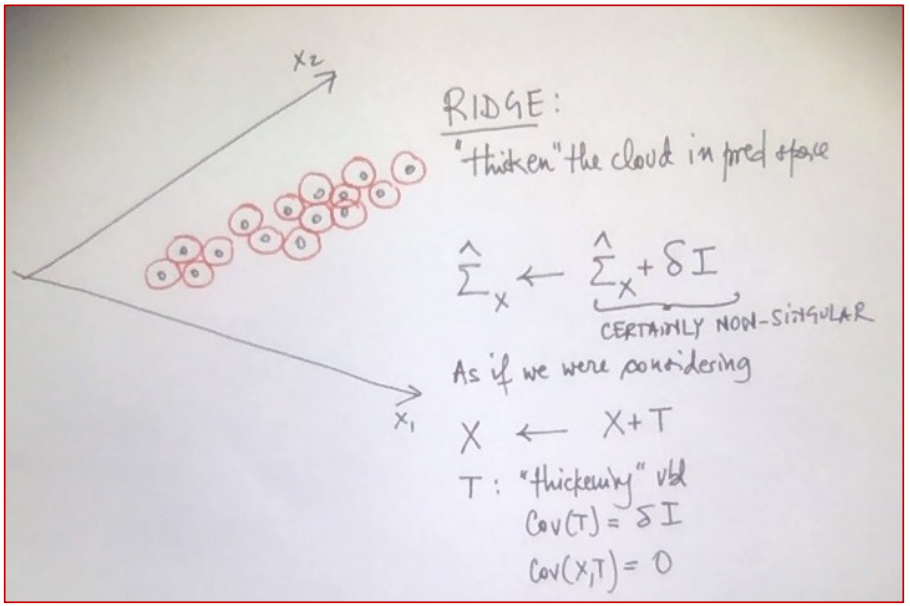
LASSO ('90s): fitting procedure to *produce sparse solutions* in regressions with a large number of features, only a subset of which is expected to matter.

Loosely, it performs "soft" features selection (more below), without specifying how many coefficients should be set to 0.

Both formulated as a constrained LS: *Size constraint* on β using different norms in R^p.


Fitting $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$ by LEAST SQUARES
y
x_2
Theoretical: collinear predictors
$n>p$, but points "line up" in pred. space
↳ thin cloud does not span pred space


X_2
X_1
Theoretical: non/mildly collinear predictors
The same phenomenon can take place if n is small relative to p
... UNDER-SAMPLING
Points "line up" by chance creating a thin cloud in pred space
$\hat{\Sigma}_X$ almost singular/singular even though Σ_X is not
NOTE: if $n<p$, $\hat{\Sigma}_X$ is singular regardless of Σ_X
↳ rank bound by $n-1$


x_2
x_1
RIDGE:
"thicken" the cloud in pred space
$\hat{\Sigma}_X \leftarrow \hat{\Sigma}_X + \delta I$
CERTAINLY NON-SINGULAR
As if we were considering
$X \leftarrow X+T$
T: "thickening" v.b.
$Cov(T) = \delta I$
$Cov(X,T) = 0$

Constrained Least Squares (using different norms)

Ridge

$$\hat{\beta}_{Ridge} = \text{argmin}\left\{ \| \underline{Y} - \underline{X}\beta \|^2 + \lambda \| \beta \|_{(2)} \right\}$$

$$\| \beta \|_{(2)} = \sum_{j=1}^{p} \beta_j^2 \quad \text{L2 Norm}$$

LASSO

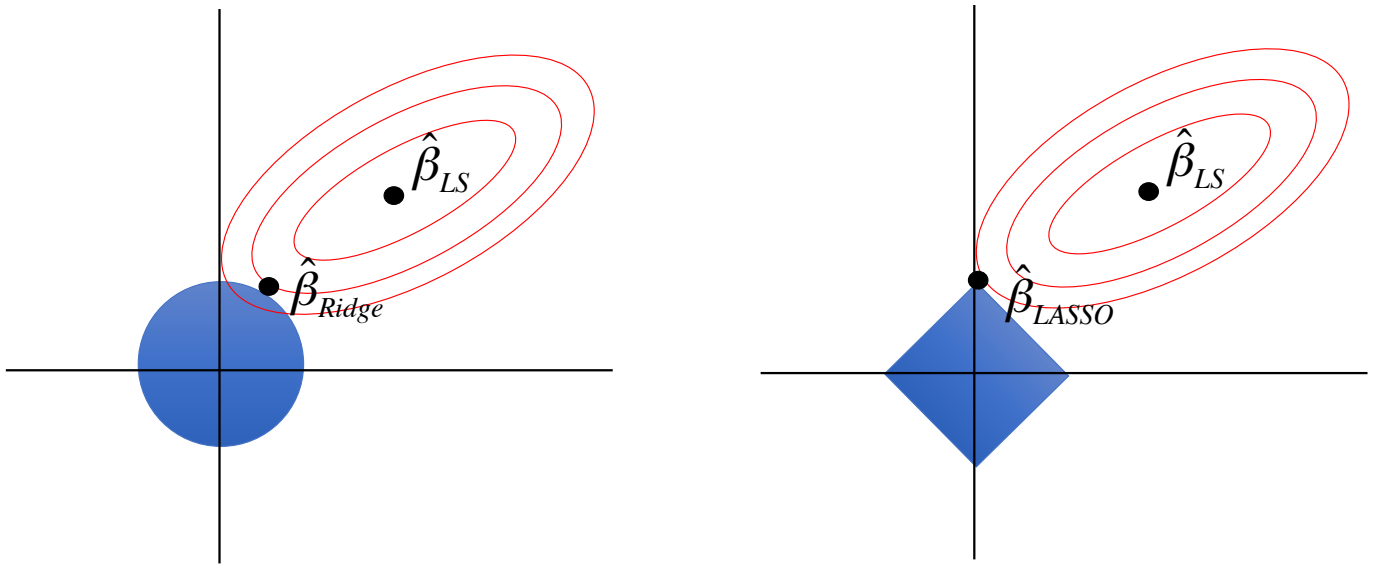
$$\hat{\beta}_{LASSO} = \text{argmin}\left\{ \| \underline{Y} - \underline{X}\beta \|^2 + \lambda \| \beta \|_{(1)} \right\}$$

$$\| \beta \|_{(1)} = \sum_{j=1}^{p} | \beta_j | \quad \text{L1 Norm}$$

Note: while LS is scale equivariant, Ridge and LASSO are not – important to perform them after scaling features (e.g., divide each by its sd, so all will have the same unitary spread)

Cartoons with p=2

the diamond shape of the L1 constraint makes it more likely that the solution lies in a corner



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

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

size constraint

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

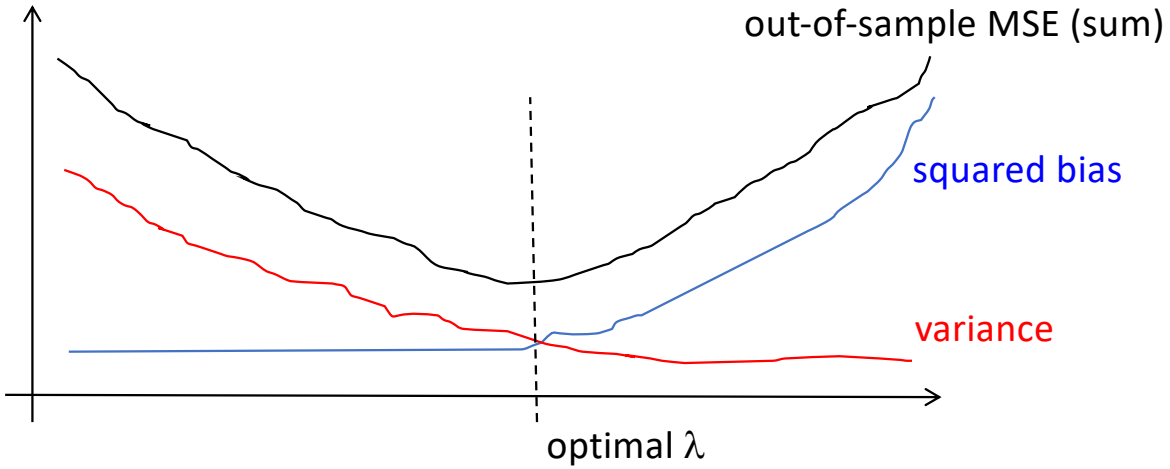
$$\sum_{j=1}^{p} | \beta_j | \leq s_\lambda$$

(maximal) VALUE OF THE SIZE CONSTRAINT (PENALTY PARAMETER) FIXED…

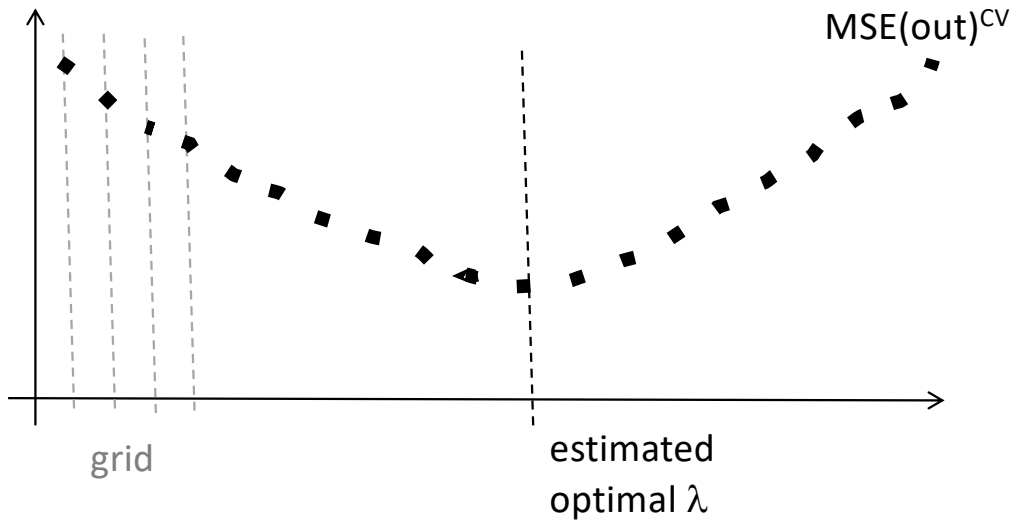
Extremely efficient algorithms exist to minimize the objective function with penalty and find the constrained solutions (convex optimization). LASSO is of course more computationally expensive that LS (with its close form solution), but not prohibitive also for very large p.

The critical part is ***selecting an appropriate λ (s_λ)***.

Ridge and LASSO can improve over LS because they reduce the sample variability of the coefficient estimators and emiliorate overfitting. But they create (if LS is unbiased) or increase a bias component. As λ (and thus the weight of the penalty vs the RSS term in the objective function) grows, the variance decreases but the bias increases – in the limit for infinite λ all coefficient estimates will be 0, with no variance! We want the "sweet spot" where the out-of-sample MSE is minimized:



In applications we don't know how the out-of-sample MSE varies as a function of λ. But we can estimate it on a grid of values:



The selection of an appropriate λ substantially adds to the computational burden (we need to iterate the optimization algorithm many times) but it is essential for an effective use of Ridge and LASSO.

… and algorithms for running LASSO are very fast!

Note: we have introduced the geometry of the constraints taking λ as fixed, then discussed the selection of λ.

One could imagine selecting (e.g., by cross-validation) also some aspects of the geometry of the constraints? For instance, L1 or L2 norm? A combination of norms?

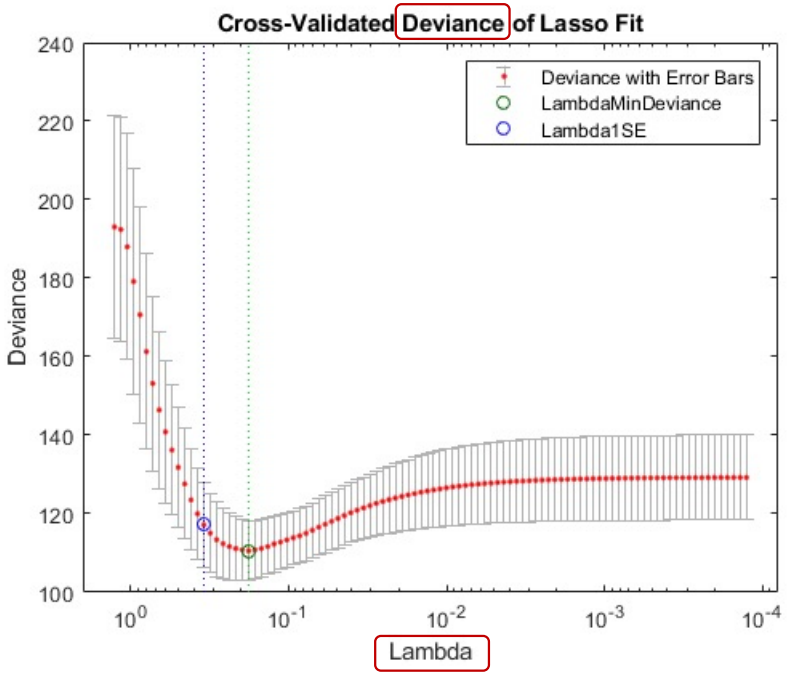
Elastic Nets – combine Ridge and LASSO penalization (not in ISLR).

Implementation for Linear (Gaussian) as well as Generalized Linear Models in the R package **GLMnet** (Ridge, LASSO, and Elastic Nets in between)

https://cran.r-project.org/web/packages/glmnet/glmnet.pdf
https://cran.r-project.org/web/packages/glmnet/index.html

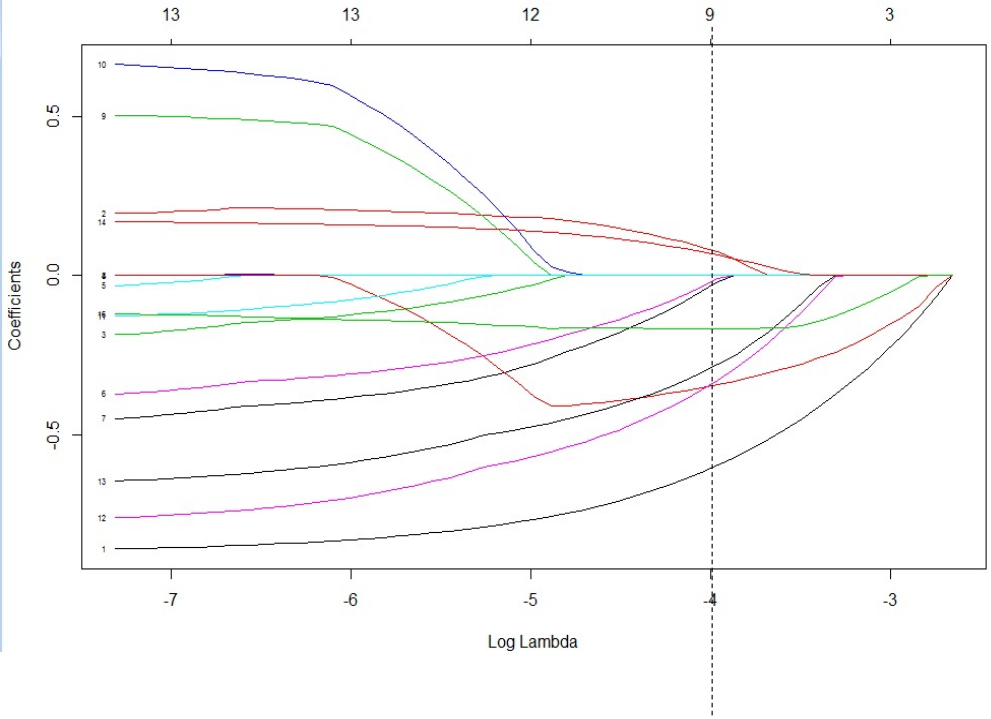
Choosing the right penalty parameter

or MSE



or Log Lambda

How coefficient estimates shrink with penalization



at a given level of penalization, how many and which coefficients are non-0?