Outline: (Linear) Models in Large Feature Spaces
Introducing Penalizations: Ridge and LASSO

(F. Chiaromonte)

Introduction to Statistical Learning

Ridge & LASSO: Chapter 6 Section 2, Lab 2

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 reducing variance – so overall accuracy improves.

Additionally, constraining may result in a smaller, more *parsimonious and interpretable model* (some features are eliminated, or the focus is shifted 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.

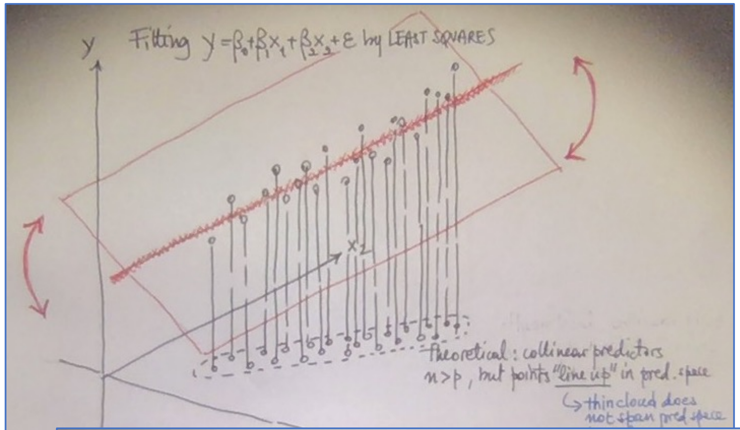
Ridge (‘60-’70s): fitting procedure to *overcome multicollinearity* in regressions with highly interdependent features.

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

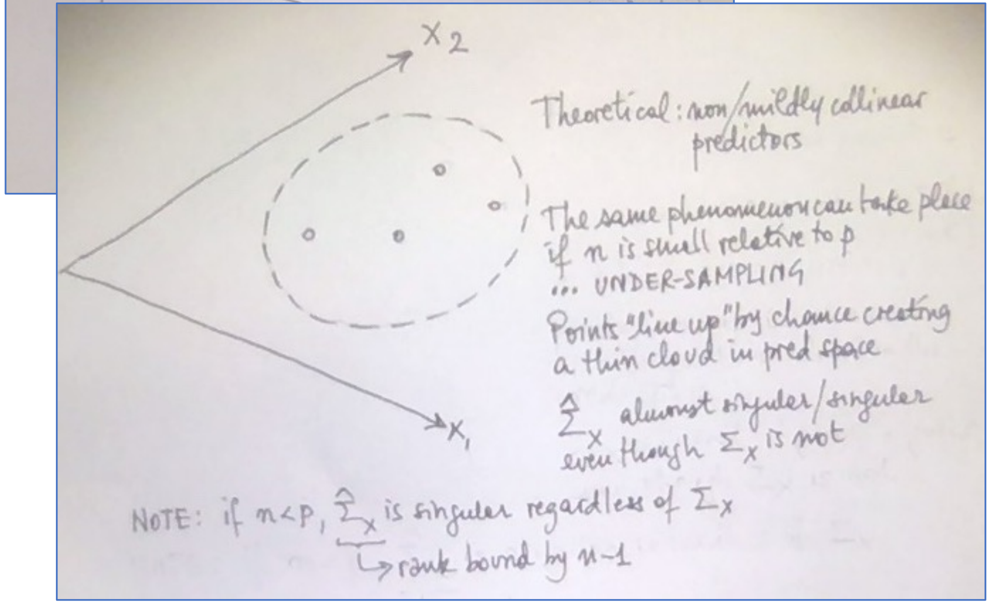


Fitting $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ by LEAST SQUARES

Theoretical: collinear predictors

$n > p$, but points "line up" in pred. space

↳ thin cloud does not span pred space



Theoretical: non/mildly collinear predictors

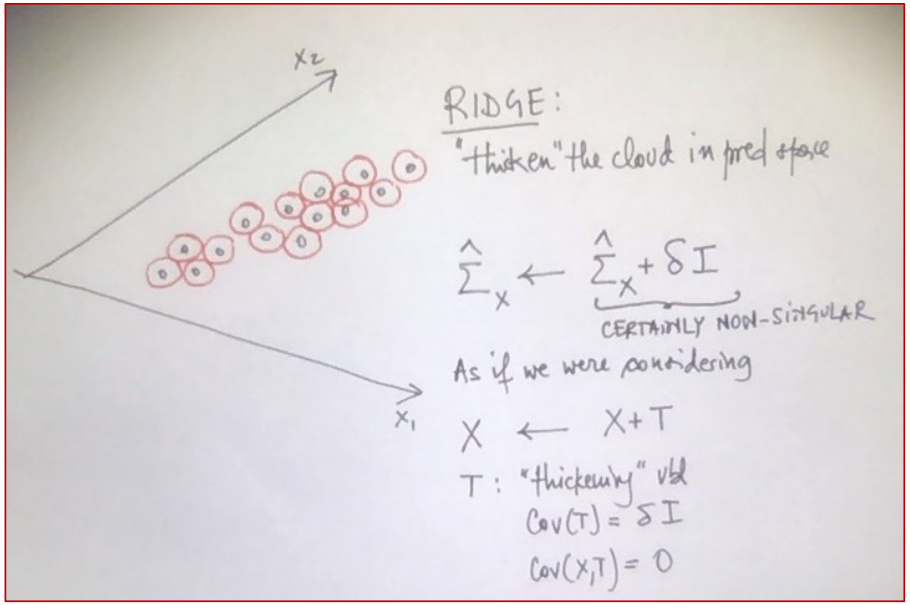
The same phenomenon can take place if n is small relative to p ... UNDER-SAMPLING

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 $n<p$, $\underbrace{\hat{\Sigma}_X}$ is singular regardless of Σ_X

↳ rank bound by $n-1$



RIDGE:

"thicken" the cloud in pred space

$$\hat{\Sigma}_X \leftarrow \underbrace{\hat{\Sigma}_X + \delta I}_{\text{CERTAINLY NON-SINGULAR}}$$

As if we were considering

$$X \leftarrow X + T$$

T: "thickening" vbl

$Cov(T) = \delta I$

$Cov(X, T) = 0$

Constrained Least Squares (using different norms)

Ridge

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

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

LASSO

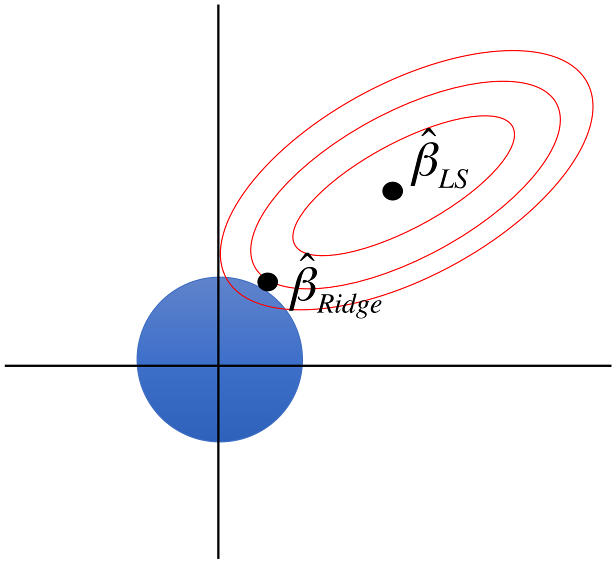
$$\hat{\beta}_{LASSO} = \operatorname{argmin}\left\{\| \underline{Y} - \underline{X}\beta \|^2 + \lambda \| \beta \|_{(1)}\right\}$$

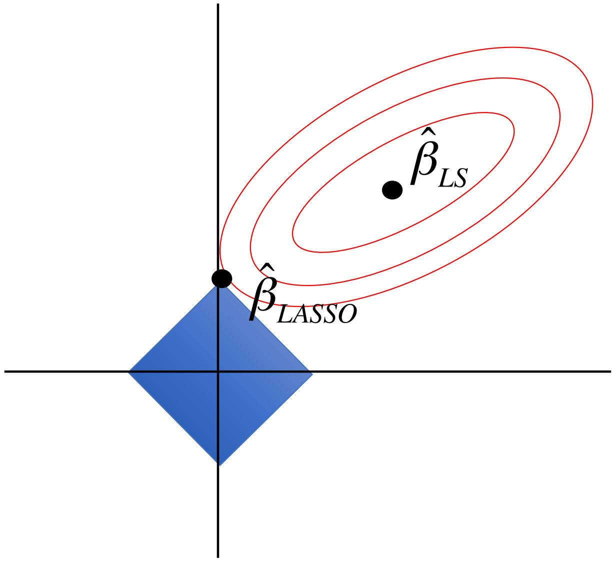
$$\| \beta \|_{(1)} = \sum_{j=1}^{p} | \beta_j | \quad \text{L1 Norm}$$

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





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

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

← size constraint →

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

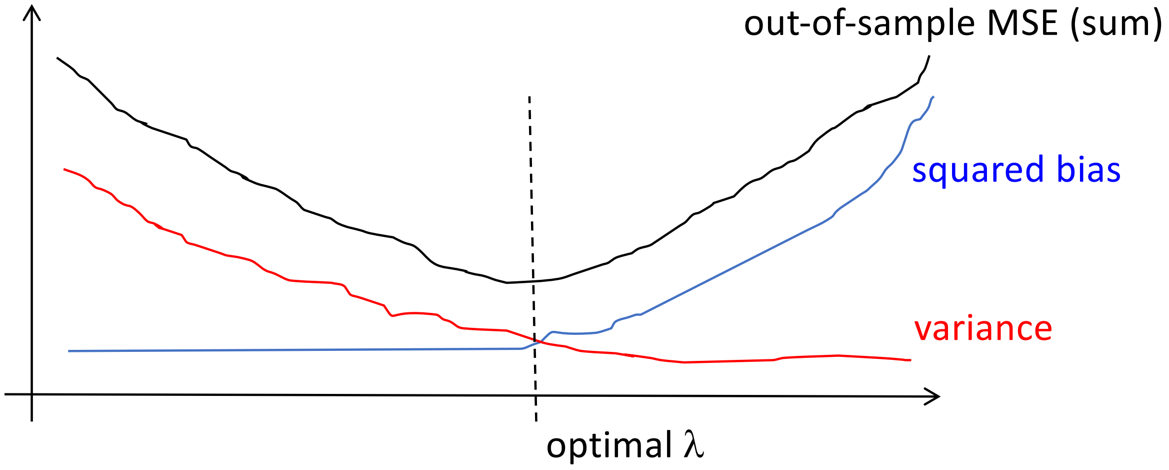
$$\sum_{j=1}^{p} | \beta_j | \le s_\lambda$$

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

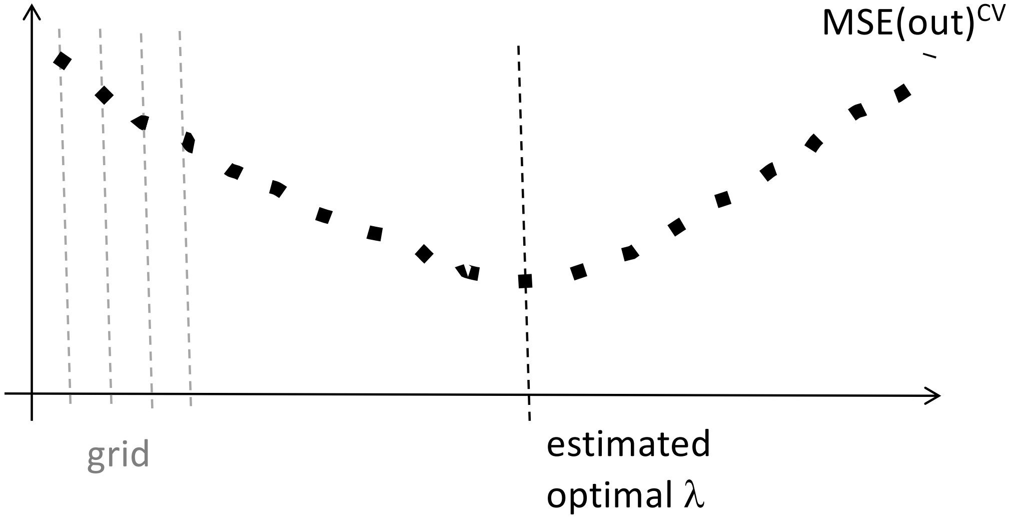
Extremely efficient algorithms exist to minimize the objective function with penalty and find the constrained solutions (convex optimization). LASSO is of course more computationally expensive that LS (with its close form solution), but not prohibitive also for very large p.

The critical part is ***selecting an appropriate*** λ *(i.e., 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, e.g., using cross-validation:



The selection of an appropriate λ by cross-validation 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!
(implemented in standard statistical software)

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).

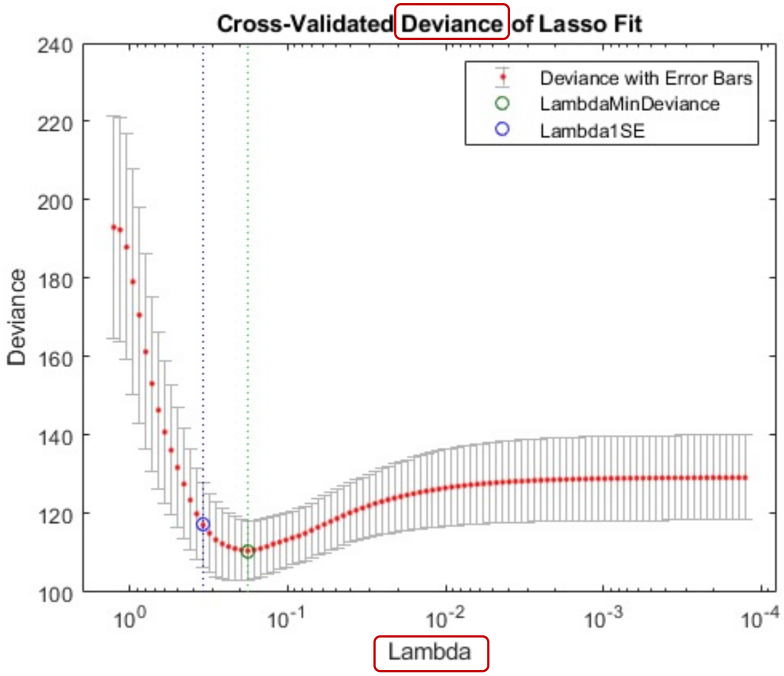
Implementation for Linear (Gaussian) as well as Generalized Linear Models in the R package **GLMnet** (Ridge, LASSO, and Elastic Nets in between)

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

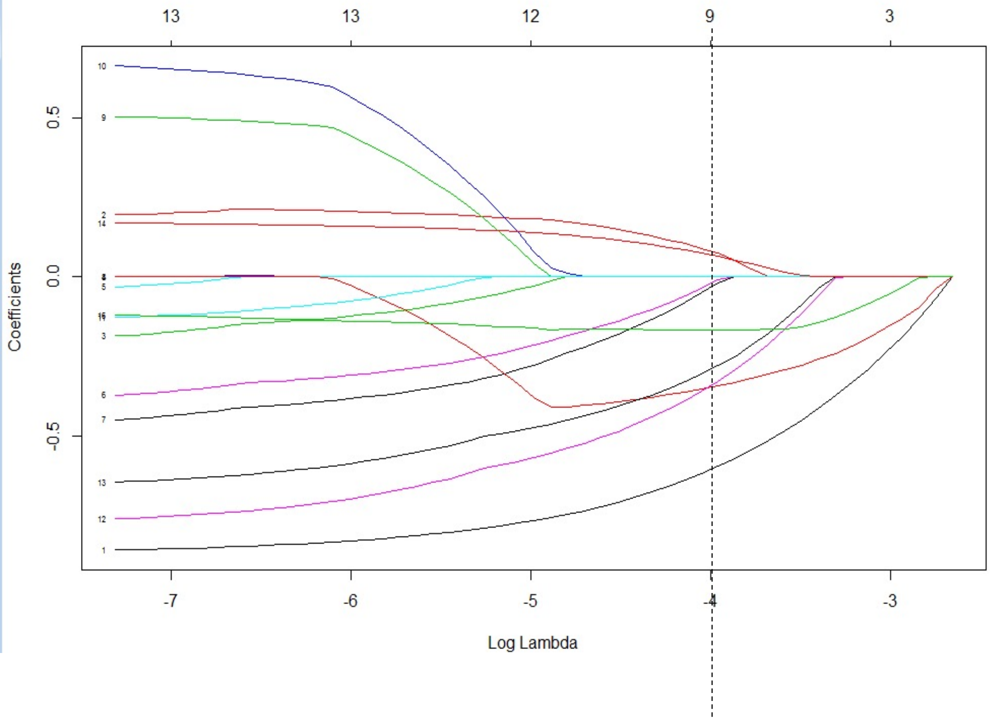
Choosing the right penalty parameter

How coefficient estimates shrink with penalization

or MSE



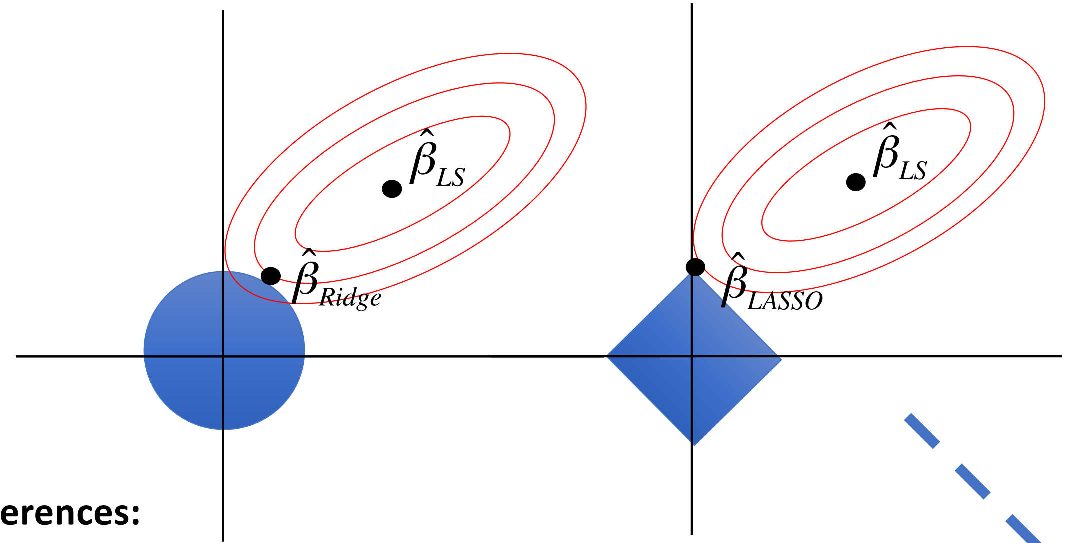
or Log Lambda



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

SOME ADDITIONS: NON-CONVEXITY AND ISSUES WITH THE LASSO

FURTHER CONSIDERATIONS AND REFERENCES ABOUT PENALIZATIONS



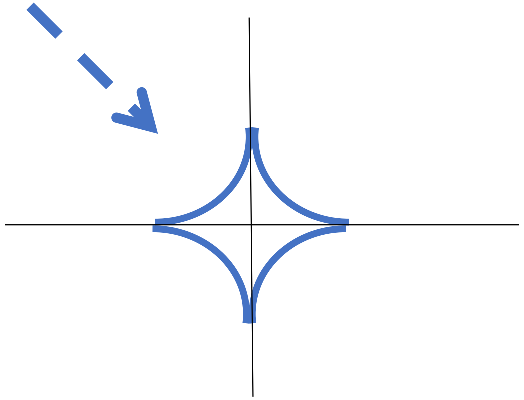
CONVEX constrained optimization
very efficient computational approaches.

Estimates are biased (even when the LS for the full model is not)… but **more stable**, and **sparse** (L1).

Very broad literature, lots of variants, including those to **incorporate group or order structure** for features.

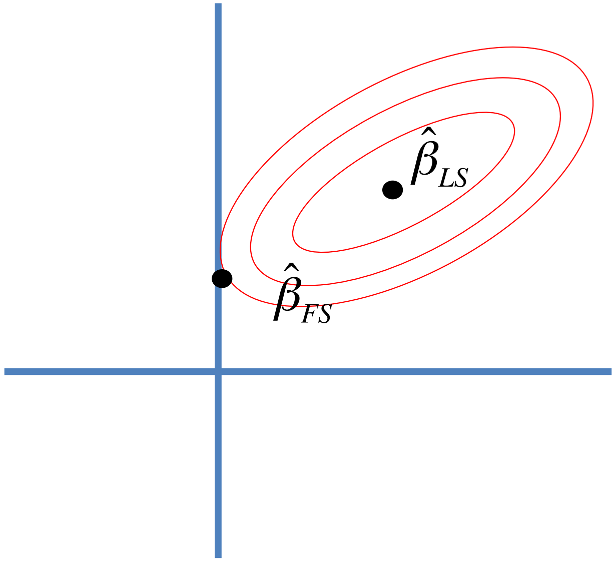
Some references:

- Hastie T., Tibshirani R., Friedman J. (2009). Elements of Statistical learning 2nd ed. Springer.
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. JRSS B, 58(1) 267-288.
- Zou H. Hastie T. (2005) Regularization and variable selection via the elastic net. JRSS B, 67(2) 301–320.
- Tibshirani R., Saunders M. (2005) Sparsity and smoothness via the fused lasso. JRSS B, 67(1) 91–108
- Yuan M., Lin Y. (2006) Model selection and estimation in regression with grouped variables. JRSS B, 68(1) 49–67.
- Fan J. Li R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. JASA, 96(456) 1348-1360



… abandon CONVEXITY
reduced bias, but harder computational problem.

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.

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

$$\sum_{j=1}^{p} Ind(\beta_j \neq 0) \leq c$$

size constraint

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

Important: can also add further "integer" constraints to capture structure.

MIP-BOOST: Efficient and Effective L_0 Feature Selection for Linear Regression

Abstract: Recent advances in mathematical programming have made mixed integer optimization a competitive alternative to popular regularization methods for selecting features in regression problems. The approach exhibits unquestionable foundational appeal and versatility, but also poses important challenges. Here, we propose MIP-BOOST, a revision of standard mixed integer programming feature selection that reduces the computational burden of tuning the critical sparsity bound parameter and improves performance in the presence of feature collinearity and of signals that vary in nature and strength. The final outcome is a more efficient and effective L_0 feature selection method for applications of realistic size and complexity, grounded on rigorous cross-validation tuning and exact optimization of the associated mixed integer program. Computational viability and improved performance in realistic scenarios is achieved through three independent but synergistic proposals.

First: a novel bisection procedure designed specifically for tuning the sparsity bound, which significantly reduces the needed number of evaluations, cutting the computational burden of the MIP approach.

Second: a novel cross-validation scheme that exploits the structure of the MIP and of the simplex algorithm used for its solution to significantly reduce the computational effort of repeating calculations across folds (also warm starts and surrogate lower bounds).

Third: whitening, a pre-processing step that, by handling feature collinearities, can both reduce computational burden and improve solution quality. Whitening can be applied prior to any feature selection technique but it benefits MIP more than it does other approaches.

More important remarks:

- Even when introducing some Ridge for stabilization through an Elastic Net, collinearities can hinder the performance of LASSO (as of any other feature selection approach). Strong associations, especially between relevant and irrelevant features, can deteriorate the ability of any feature selection approach to identify the former.

Zhao P., Yu B. (2006) On model selection consistency of LASSO. Journal of Machine Learning Research, 7, 2541-2563.

Abstract: Sparsity or parsimony of statistical models is crucial for their proper interpretations, as in sciences and social sciences. Model selection is a commonly used method to find such models, but usually involves a computationally heavy combinatorial search. Lasso (Tibshirani, 1996) is now being used as a computationally feasible alternative to model selection. Therefore it is important to study Lasso for model selection purposes. In this paper, we prove that a single condition, which we call the Irrepresentable Condition, is almost necessary and sufficient for Lasso to select the true model both in the classical fixed p setting and in the large p setting as the sample size n gets large. Based on these results, sufficient conditions that are verifiable in practice are given to relate to previous works and help applications of Lasso for feature selection and sparse representation. This Irrepresentable Condition, which depends mainly on the covariance of the predictor variables, states that Lasso selects the true model consistently if and (almost) only if the predictors that are not in the true model are "irrepresentable" (in a sense to be clarified) by predictors that are in the true model. Furthermore, simulations are carried out to provide insights and understanding of this result.

- Addressing biases and post-selection inference issues for penalized methods such as the LASSO.

Belloni A., Chernozhukov V. (2013) Least squares after model selection in high-dimensional sparse models. Bernoulli, 19(2), 521 – 547.

Abstract: In this article we study post-model selection estimators that apply ordinary least squares (OLS) to the model selected by first-step penalized estimators, typically Lasso. It is well known that Lasso can estimate the nonparametric regression function at nearly the oracle rate, and is thus hard to improve upon. We show that the OLS post-Lasso estimator performs at least as well as Lasso in terms of the rate of convergence, and has the advantage of a smaller bias. Remarkably, this performance occurs even if the Lasso-based model selection “fails” in the sense of missing some components of the “true” regression model. By the “true” model, we mean the best s-dimensional approximation to the nonparametric regression function chosen by the oracle. Furthermore, OLS post-Lasso estimator can perform strictly better than Lasso, in the sense of a strictly faster rate of convergence, if the Lasso-based model selection correctly includes all components of the “true” model as a subset and also achieves sufficient sparsity. In the extreme case, when Lasso perfectly selects the “true” model, the OLS post-Lasso estimator becomes the oracle estimator. An important ingredient in our analysis is a new sparsity bound on the dimension of the model selected by Lasso, which guarantees that this dimension is at most of the same order as the dimension of the “true” model. Our rate results are nonasymptotic and hold in both parametric and nonparametric models. Moreover, our analysis is not limited to the Lasso estimator acting as a selector in the first step, but also applies to any other estimator, for example, various forms of thresholded Lasso, with good rates and good sparsity properties. Our analysis covers both traditional thresholding and a new practical, data-driven thresholding scheme that induces additional sparsity subject to maintaining a certain goodness of fit. The latter scheme has theoretical guarantees similar to those of Lasso or OLS post-Lasso, but it dominates those procedures as well as traditional thresholding in a wide variety of experiments.