"High-Dimensional Methods and Inference on Structural and Treatment Effects" by Alexandre Belloni, Victor Chernozhukov, and Christian Hansen

Presented by Daniela Urbina Julio

March 1, 2018

Shrinkage Methods: Overview of Lasso Regularization

- **How to deal with complex, high-dimensional data sources ($p >> n$)?**

- Regularization methods reduce this complexity by shrinking or constraining coefficient estimates.This dimension reduction approach helps researchers to see a clearer picture of their data and draw meaningful conclusions.

- In particular, regularization methods constrains coefficient estimates towards zero, exchanging a reduction in variance for inducing some bias.

- Regularizers which draw coefficients to exact zeroes are called sparsity-inducing regularizers.

Mathematical Form of Lasso Regularization

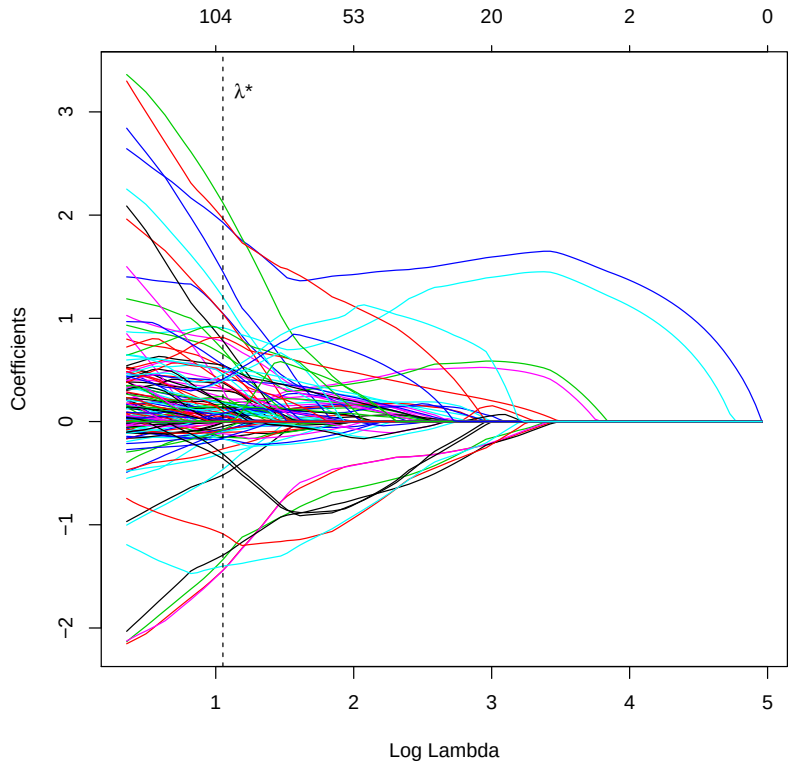
- We can express the lasso problem as an optimization of the following objective function:

$$\hat{\beta}_\lambda = \arg\min \left(||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|_1 \right), \tag{1}$$

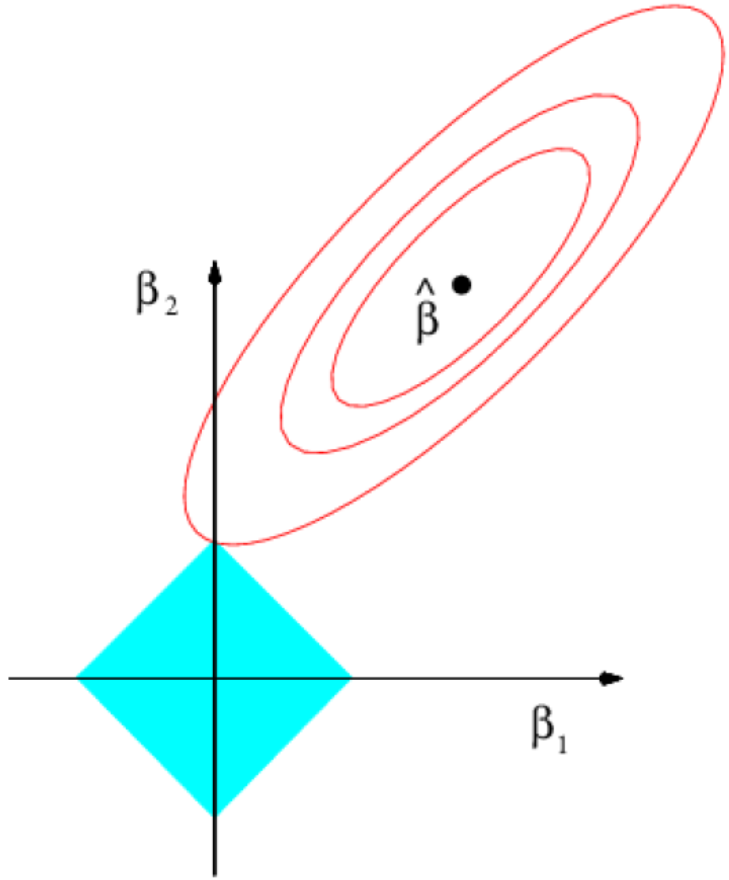
- Penalty (tuning parameter) is often defined through cross-validation with the goal of minimizing the expected out of sample prediction error while preventing over-fitting (Hastie 2009).
- A higher value of λ indicates a lower tolerance for complexity in the fitted model.
- Note the ℓ_1 penalty, which imposes a linear constraint to the absolute value of the magnitude of coefficients, generating corner solutions.

Main findings: Selection via Lasso

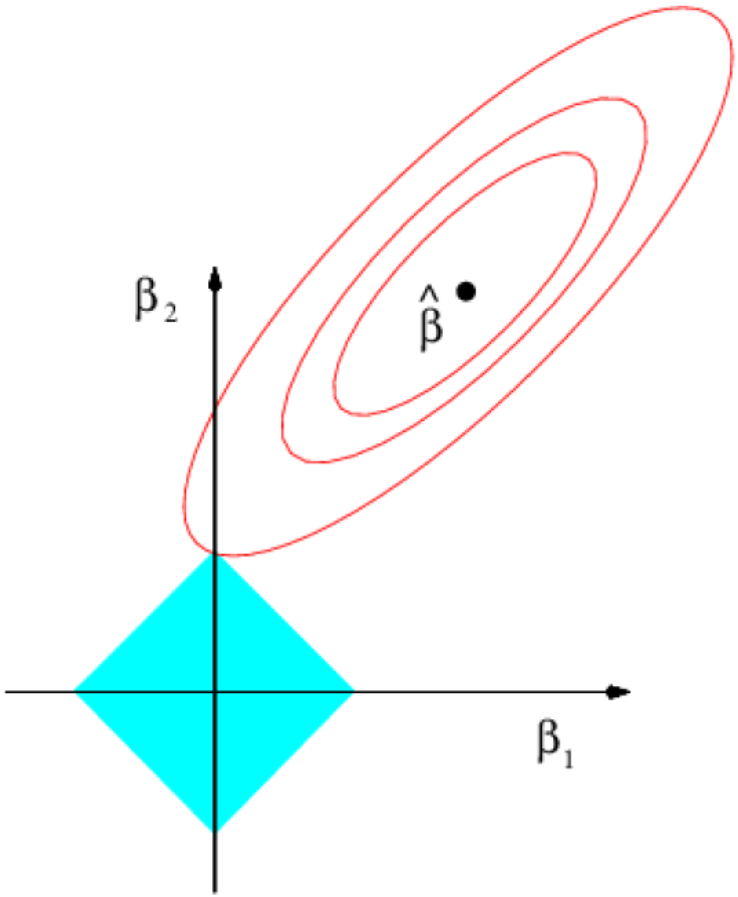
Figure: **N** $\approx$ **55,255**; (Left Panel) Path of Lasso regularized coefficients for the Saturated Model

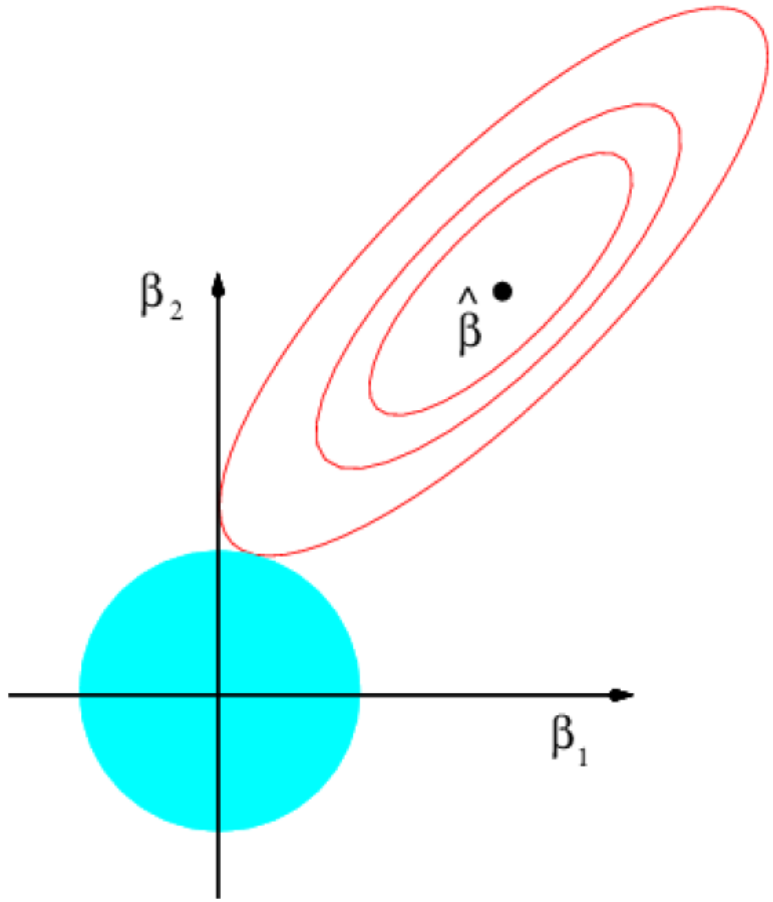


Geometrical Form of Lasso Regularization



Lasso versus Ridge





Bottom Line

- Helpful in a setting where the researcher is considering many explanatory variables and uses data to learn which of the many variables are the most important.
- Lasso regularization reduces over-fitting and increases out-of-sample prediction + nice efficient algorithm.
- **But is Lasso regularization useful if the goal is causal inference?**

  Yes! But we need to be careful. These procedures were designed for forecasting, and not for inference about model parameters.

Belloni, Chernozhukov, and Hansen's Position

- The Problem: Effectively Lasso is an approach that generates omitted variables. As we change our tuning parameter, the nature of these omitted variables change (Athey AER video). Model selection mistakes may occur; some variables may have small but non-zero effects and might not get selected via the Lasso.

- If we care about specific estimates this is a problem as Lasso could be generating omitted variables bias.

- The authors propose to use the Lasso over the predictive parts of the economic problem–reduced forms and first stages (IV)–rather than using model selection in the structural model directly.

Proposed Approach I: Inference with Selection among Many Instruments

$$y_i = \alpha d_i + \epsilon_i \tag{2}$$

$$d_i = z_i' \Pi + r_i + v_i \tag{3}$$

- d_i is a scalar endogenous variable of interest.
- z_i is a p-dimensional vector of instruments
- Where the number of instruments p may be much larger than the number of observations.
- Main idea: Select instrument via the Lasso. Variable selection only takes place in the first stage equation relating the endogenous variable to the instruments, which is a purely predictive relationship.
- Selection mistakes are not a problem as long as other instruments with larger coefficients are selected."Second-stage is immune to variable selection issues".

Procedure for IV regression: Judges and Housing Prices Example

- The effect of taking laws on housing prices. Chen and Yeh (2012) rely on the random assignment of judges to federal appellate panels.
- Where d_i is the judge's decision and z_i are a series of characteristics of judges that satisfy the exclusion restriction; y_i corresponds to property prices.
- Parameter of interest α corresponds to the effect of an additional decision upholding individual property rights on a measure of property prices.
- Use Lasso to identify a set of good instruments from a large set of potential instruments. Lasso used for prediction purposes.

Procedure for IV regression: Judges and Housing Prices Example

- **Step 1**: Choose a reduced set of variables that intuitively could predict d_i. In this case, they choose 147 instruments.
- **Step 2**:Include z_{ij} 's that are significant predictors of d_i as judged by LASSO. In this case, Lasso only selected one instrument: number of panels with one or more members with a JD from a public university squared.
- **Step 3**: Refit the model by two-stage least squares after selection, use standard confidence intervals. Turns out a single judicial decision reinforcing individual property is associated with an effect between 2 and 11 percent higher property prices.

Inference with Selection among Many Controls: Wrong Approach

$$y_i = \alpha d_i + x_i' \theta_y + r_{yi} + \gamma_i \tag{4}$$

- Where d_i is taken as exogenous after conditioning on control variables.
- $E[\gamma_i | d_i, x_i, r_{yi}] = 0$
- x_i is a p-dimensional vector of controls where $p >> n$
- r_{yi} is an approximation error
- And α is our parameter of interest, the treatment effect.
- Naive approach: Select control variables via Lasso, forcing the treatment variable to remain in the model by excluding α from the Lasso penalty. Then run OLS with only the selected controls.

Wrong Approach to Inference with High-Dimensional Methods

- What is the problem? From the standpoint of prediction, any variable highly correlated with the treatment will tend to be dropped.
- This would lead to a substantial omitted variable bias problem.
- Also, this approach is based on a structural model, where the goal is to learn about the treatment given controls, not an equation representing a forecasting rule for y_i given d_i and x_i.

Proposed Approach: "Double Selection"

- To guard against these issues Belloni and colleagues propose we must consider both equations for selection.
- We apply variable selection to each of the two reduced form equations and then use all of the selected controls in the estimation of α.
- Using the variables selected in both reduced form equations ensures that any variables with a large effect for y_i or d_i are included in the model. Any excluded variable are at most mildly associated with y_i and d_i. This is the robustness of the double selection process.

Procedure for Controls: Abortion Laws Example

- Donohue and Levitt (2001) sough to estimate the effect of abortion on crime rates using a differences-in-differences approach.

$$y_{cit} = \alpha_c a_{cit} + w'_{it}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit} \tag{5}$$

- Where y_{cit} change in crime rate c in state i in year t
- a_{it} change in abortion rate in state i in year t
- w_{it} are basic controls (time varying confounding state-level factors p $= 20$).
- δ_{ci} state specific effects that control for any time-invariant state-specific characteristics.
- γ_{ct} are time-specific effects that control for national aggregate trends.

Procedure for Controls: Abortion Laws Example

- **Step 1**: Include w_{it} 's that are significant predictors of y_{cit} crime rate as judged by LASSO.

- **Step 2**: Include w_{it} 's that are significant predictors of a_{it} abortion rate judged by LASSO.

  In this case they propose a series of nonlinear trends interacted with observed state-specific characteristics for the x's.

- **Step 3**: Then use the union of the set of selected variables, including time effects, as controls in a final OLS regression of y_{it} on a_{it}.
- This procedure yields different results than the original findings using first differences with intuitively selected controls.

Double Selection for IV: Institutions on Output

- **Step 1**: Include x_{ij} 's that are significant predictors of y_i as judged by LASSO. (Log GDP per Capita)
- **Step 2**. Include x_{ij} 's that are significant predictors of either d_i (Protection from Expropriation) and z_i (Settler Mortality) as judged by LASSO.
- **Step 3**. Refit the model by two-stage least squares with the union of variables selected from each reduced form.