Recitation 14

Bruno Kömel

April 17, 2024

Overview

Lasso: Motivation:

- Increasingly, economists work with high-dimensional data.
- High-dimensional: Many characteristics per observation are available
  - e.g. Scanner datasets with transaction level data and text data
- Many statistical methods for constructing prediction models using high-dimensional data (e.g. trees)
- But may lead to incorrect conclusions when inference is the goal
- For today, we consider lasso models in the context of "approximately sparse" regression models
- Approximately sparse: Many potential predictor/control variables but only a few are important at predicting the outcome.

Lasso: Problem

Suppose we have a linear model:

$$y_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \zeta_i$$

where $x_i's$ are the possible regressors[1] and ζ_i is the random error

- Lasso stands for Least Absolute Shrinkage and Selection Operator.
- Idea: Coefficients chosen to minimize sum of squared residuals plus a penalty term that penalizes (size of model) number of nonzero coefficients.

[1] Be sure to **always** standardize the regressors (Stata will do this for you automatically)

Lasso: Problem

Specifically, the lasso estimator can be defined as:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where $\lambda > 0$ is the penalty level.

- The first term is the same value that OLS minimises
- The second term is a penalty that increases in value the more complex the model
- The second term causes lasso to omit variables because of the kink in the absolute value terms $|b_j|$.
  - Why? Imagine if the penalty term consisted of squares $\sum_{j=1}^{p} \beta_j^2$.
  - Then many small coefficients will still be included in the model, since the penalty term is relatively flat near zero.

Lasso: Problem

Specifically, the lasso estimator can be defined as:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where $\lambda > 0$ is the penalty level.

- The penalty level λ controls the degree of penalization.
- When λ is large, the penalty is large and the model has few or no variables. Conversely, if λ is small, more variables can be admitted.

Lasso: Solutions

- Cross-validation
  - Works well for prediction
  - Criterion function is $f(\lambda)$, an estimate of the out-of-sample prediction error
  - Relatively slow
- Adaptive Lasso
  - Works well when the goal is to find parsimonious models
  - Starts by finding the CV solution
  - Then, by using weights on the coefficients in the penalty function, does another lasso and selects a model that has fewer variables

Lasso: Solutions

- Plug-in lasso
  - Faster than CV or adaptive lasso
  - Does not minimize out-of-sample prediction error
  - Uses an iterative formula to calculate the smallest λ that is large enough to dominate the estimation error in the coefficients
  - Produces more parsimonious models
  - Default selection method for inference because of speed, but is not so good for prediction

Lasso: Prediction vs. Inference

- Lasso is useful for obtaining forecasting rules and estimating which variables have a strong association to an outcome in a sparse framework.
- But, procedures are designed for forecasting, not for inference on true model parameters.
- Lasso tends to omit covariates with small coefficients, if those covariates were part of the true model, it can lead to omitted variable bias.
- Lasso can include variables that are highly correlated to the true variables, omitting the true variables themselves.
- In general, variables selected by lasso do not converge to the true set even as $N \to \infty$.

Lasso for Inference

Consider the model:

$$y = d\alpha + x\beta + \epsilon$$

Where d is the covariate of interest and α is the coefficient of interest

- One could try running a lasso on all controls
- But, any variable highly correlated to d will drop out since it doesn't add much to predictive power
- Substantial omitted variable bias if the coefficient is nonzero in β
- So, we want to find methods whereby variables that are great predictors of both y and d are selected
- We outline three such methods

Lasso for inference

$$y = d\alpha + x\beta + \epsilon$$

1. Double Selection method:
    - Run lasso of d on X
    - Run lasso of y on X
    - Let $\tilde{X}$ be the union of selected covariates from Steps 1 and 2
    - Regress y on d and the set of selected covariates $\tilde{X}$

    Idea: Variables in X highly correlated to d would still be included. Good controls are identified in the two lassos.

Lasso for inference

$$y = d\alpha x\beta + \epsilon$$

2. Partialling out method:
   - Run lasso of d on X, let $\tilde{X}_d$ be the set of selected covariates
   - Regress d on $\tilde{X}_d$, and let $\tilde{d}$be the residuals from this regression
   - Run lasso of y on X, let $\tilde{X}_y$ be the set of selected covariates
   - Regress y on $\tilde{X}_y$, let $\tilde{y}$ be the residuals from this regression
   - Regress $\tilde{y}$ on $\tilde{d}$

     Idea: "Partial out" the effects of X in order to estimate the effect of d on y.

Lasso for inference

$$y = d\alpha + x\beta + \epsilon$$

3. Cross-fit partialling out method:

1. Divide the data into equal-sized subsamples 1 and 2
2. In sample 1:
   - Run lasso of d on X, let $\tilde{X}_{d1}$ be the set of selected covariates
   - Regress d on $\tilde{X}_{d1}$, and let $\hat{\beta}$ be the estimated coefficients
   - Run lasso of y on X, let $\tilde{X}_{y1}$ be the set of selected covariates
   - Regress y on $\tilde{X}_{y1}$, let $\hat{\gamma}$ be the estimated coefficients
3. In sample 2:
   - Fill in $\tilde{d} = d - \tilde{X}_{d1}\hat{\beta}_1$
   - Fill in $\tilde{y} = y - \tilde{X}_{y1}\hat{\gamma}_1$

Lasso for inference

$$y = d\alpha + x\beta + \epsilon$$

3. Cross-fit partialling out method:
    4. In sample 2:
        - Run lasso of d on X, let $\tilde{X}_{d2}$ be the set of selected covariates
        - Regress d on $\tilde{X}_{d2}$, and let $\hat{\beta}$ be the estimated coefficients
        - Run lasso of y on X, let $\tilde{X}_{y2}$ be the set of selected covariates
        - Regress y on $\tilde{X}_{y2}$, let $\hat{\gamma}$ be the estimated coefficients
    5. In sample 1:
        - Fill in $\tilde{d} = d - \tilde{X}_{d2}\hat{\beta}_2$
        - Fill in $\tilde{y} = y - \tilde{X}_{y2}\hat{\gamma}_2$
    6. In full sample: Regress $\tilde{y}$ on $\tilde{d}$

Lasso for inference

$$y = d\alpha + x\beta + \epsilon$$

3. Cross-fit partialling out method:
   - Cross-fit partialling has a more relaxed sparsity requirement compared to the other two methods.
   - Subsampling is random.
   - Obtaining coefficients from one subsample and using them in another independent sample adds robustness.

Lasso: Further Reading

- Stata Lasso Manual: https://www.stata.com/manuals/lasso.pdf
- *High-Dimensional Methods and Inference on Structural and Treatment Effects* by Belloni, Chernozhukov and Hansen (2014 JEP)
- *Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain* by Belloni, Chen, Chernozhukov and Hansen (2012 ECTA)