Machine Learning Methods for Economists
Penalized Regression

Stephen Hansen
University of Oxford

Review

A simple supervised learning model from machine learning is already familiar to us all: the linear regression model.

Recall that we choose a vector of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^P$ to minimize

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_i (y_i - \mathbf{x_i}^T \boldsymbol{\beta})^2$$

and that the problem has solution

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

When P is large, OLS is at risk of overfitting the data.

When $P > N$, the OLS solution is not well defined.

Penalized Regression

A popular approach to correcting these problems is to add some penalty term $\lambda h(\boldsymbol{\beta})$ to RSS, called *penalized* or *regularized* regression.

The λ parameter controls the strength of the penalty, while $h(\cdot)$ controls its shape. Keep in mind these are separate choices.

Generally h is (weakly) increasing in each coefficient so that we penalize model complexity, but it need not be symmetric.

The statistical motivation for penalty terms is at times unclear; when we fit a penalized regression, we cannot appeal directly to MLE theory.

Popular Penalties

1. L0 norm: $h(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_0$, i.e. the number of non-zero coefficients.
2. L1 norm: $h(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$, aka LASSO.
3. L2 norm: $h(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_j \beta_j^2$, aka ridge regression.
4. L1 and L2: $h(\boldsymbol{\beta}) = \sum_j |\beta_j| + \alpha \sum_j \beta_j^2$, aka elastic net regression.

. . . and many variations in the literature.

Measuring Error

To further motivate the use of penalization, consider a model $y_i = f(\mathbf{x}_i) + \varepsilon_i$ where:

1. $\mathrm{Var}(\varepsilon_i) = \sigma^2$
2. $\mathbb{E}(\varepsilon_i) = 0$.
3. ε_i are independent across i.
4. ε_i and $\mathbf{x}_i$ are independent.

Suppose we fit a mapping $\hat{f}$ from the observed N data points, and use it to predict a value for y_0 at some $\mathbf{x}_0$.

The error in the prediction is $y_0 - \hat{f}(\mathbf{x}_0)$, and the mean squared error is $\mathbb{E}[(y_0 - \hat{f}(\mathbf{x}_0))^2]$.

Note that the randomness is generated by the data we used to fit $\hat{f}$—$\mathbf{x}_0$ is given.

Decomposing the MSE

The MSE expands as $\underbrace{\mathbb{E}[y_0^2]}_{(i)} + \underbrace{\mathbb{E}[\hat{f}^2(\mathbf{x}_0)]}_{(ii)} - 2\,\underbrace{\mathbb{E}[y_0\hat{f}(\mathbf{x}_0)]}_{(iii)}$.

(i) expands as $\mathbb{E}[(f(\mathbf{x}_0) + \varepsilon_0)^2] = \mathbb{E}[f^2(\mathbf{x}_0)] + \sigma^2$.

(iii) expands as $\mathbb{E}[(f(\mathbf{x}_0) + \varepsilon_0)\hat{f}(\mathbf{x}_0)] = \mathbb{E}[f(\mathbf{x}_0)\hat{f}(\mathbf{x}_0)]$.

If we add and subtract $\mathbb{E}[\hat{f}(\mathbf{x}_0)]^2$ from the MSE expansion above, we then obtain

$$\mathrm{MSE} = \underbrace{\mathbb{E}[\hat{f}^2(\mathbf{x}_0)] - \mathbb{E}[\hat{f}(\mathbf{x}_0)]^2}_{\text{Variance}} + (\underbrace{\mathbb{E}[\hat{f}(\mathbf{x}_0)] - f(x_0)}_{\text{Bias}})^2 + \sigma^2.$$

Bias-Variance Trade-off

The trade-off between bias and variance is fundamental in machine learning.

Adding any variable to a regression will increase variance of each estimated coefficient. Adding relevant variables reduces bias.

With many covariates, OLS is unbiased but has high variance.

Idea behind penalized regression: introduce bias into coefficient estimates, but reduce variance. Can outperform OLS when P is high—see simulation below.

L0 Norm

With an appropriately selected λ, the L0 norm essentially chooses $\boldsymbol{\beta}$ according to AIC/BIC criteria, which have strong statistical foundations.

The optimization problem is difficult, and becomes computationally infeasible when P is large.

Stepwise forward regression: start with no covariates; add whichever single covariate improves the fit most; continue until no covariate significantly improves fit.

Stepwise backward regression: start with all covariates; drop single covariate that leads to the lowest reduction in fit; continue to drop until there is a significant reduction in fit.

Once we select the relevant covariates, we can perform OLS but standard errors are not correct since we do not condition on model selection.

L2 Norm

One can show that the ridge regression solution is $\hat{\beta}_{\lambda}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_P)^{-1}\mathbf{X}^T\mathbf{y}$.

The coefficients are shrunk towards zero, with the degree of shrinkage determined by λ.

However the ridge regression model does not enforce sparsity since $\hat{\beta}^{\text{ridge}} \gg 0$.

Unlike stepwise regression, we do not estimate a relevant subset of covariates and so there is no model selection.

L1 Norm

LASSO was introduced in the statistics literature by Tibshirani (1996) and is becoming increasingly popular in economics.

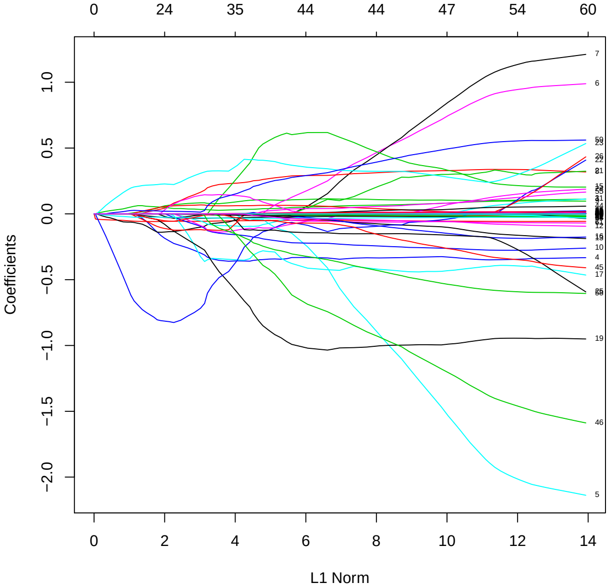
The solution to LASSO is both easy to compute (via algorithms, not in analytic form) and enforces sparsity, which makes it a popular tool for model selection.

Efficient implementation available in glmnet package in R.

Example with Barro and Lee (1994),[1] who collect data on growth rates from 138 countries from 1965-1985 with $P = 62$ predictors ($N = 90$ complete observations). Data accessed from hdm package in R.

[1] Carnegie-Rochester Conference Series on Public Policy, "Sources of economic growth".

LASSO Output



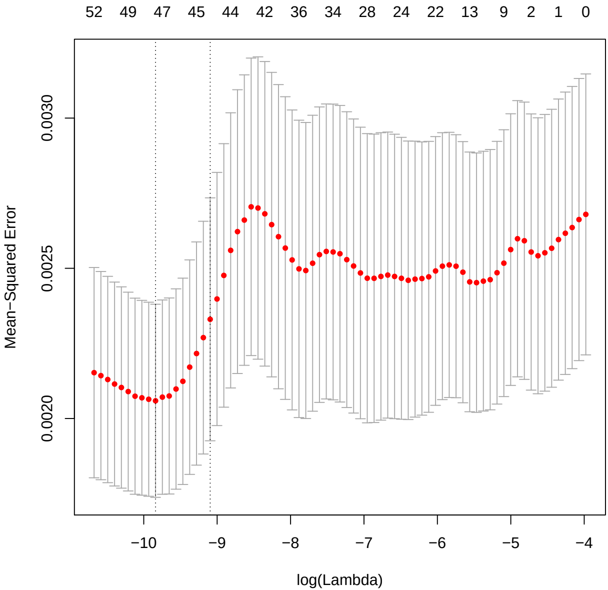
Which Penalty?

Clearly the results of LASSO are sensitive to the choice of the penalty.

A common default is to choose λ via cross-validation. Two options: choose λ to minimize average error of held-out data; choose largest λ such that error is within one standard deviation of minimum.

To assess the predictive performance of this estimator, ensure that test data is NOT part of the sample on which model selection takes place.

Example of CV

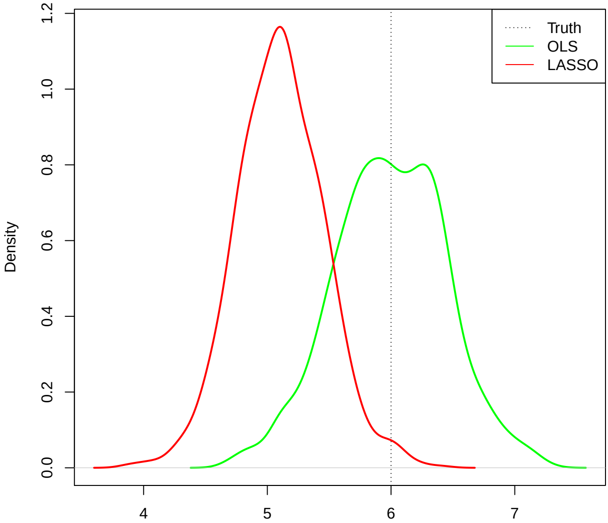


OLS vs LASSO Simulation

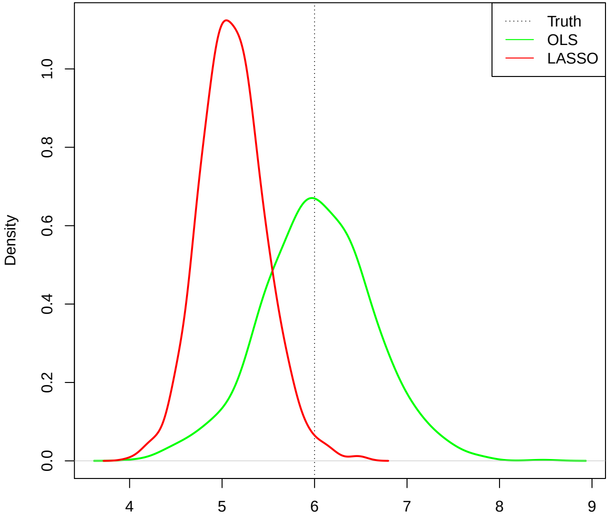
```
beta <- c(2, 1, rep(0, p))
mu <- rep(0, length(beta))

rho <- 0.5
sigma <- diag(length(beta))
sigma[upper.tri(sigma)] <- rho
sigma[lower.tri(sigma)] <- rho

x <- mvrnorm(100, mu=mu, Sigma=sigma)
y <- x %*% matrix(beta, ncol=1) + rnorm(nrow(x))
```

Estimate OLS and LASSO for 1,000 draws, plot predicted y at $\mathbf{x}_0 = (2, \ldots, 2)$ where true y is 6.

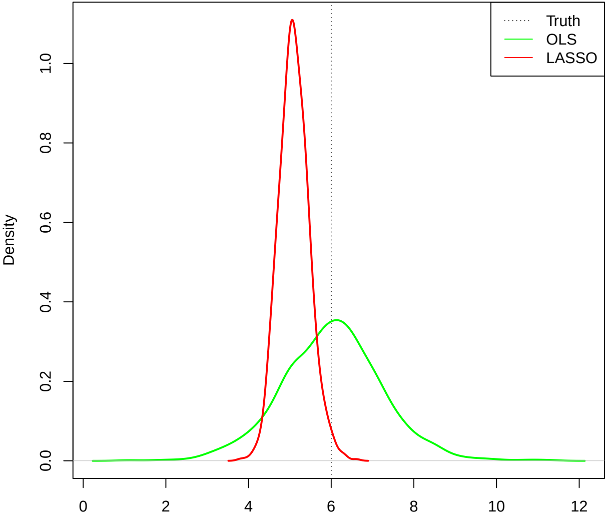
p=50



p=70



p=90



Adaptive LASSO

The LASSO penalty is important for screening out irrelevant variables, but the problem is that it also biases relevant variables.

Ideally we would like to soften (strengthen) the penalty on relevant (irrelevant) covariates.

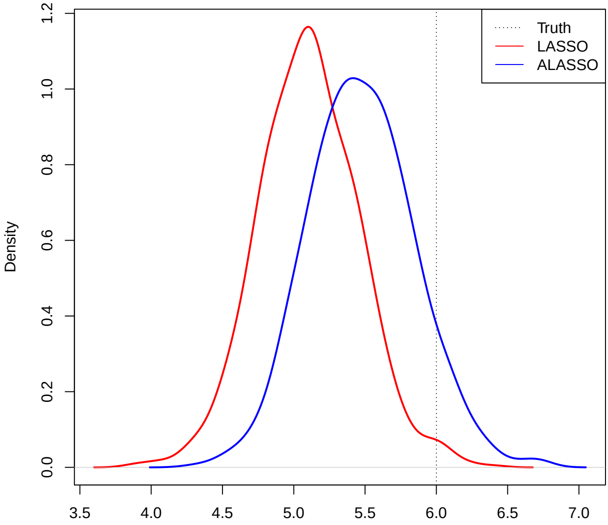
One idea in the literature is the adaptive LASSO (Zou 2006[2]) which uses the penalty $h(\boldsymbol{\beta}) = \sum_j \omega_j |\beta_j|$.

To obtain the weights, first obtain OLS coefficients $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, then $\omega_j = 1/|\hat{\beta}_j^{\mathrm{OLS}}|^{\gamma}$.

The attractive computational properties of the LASSO remain.

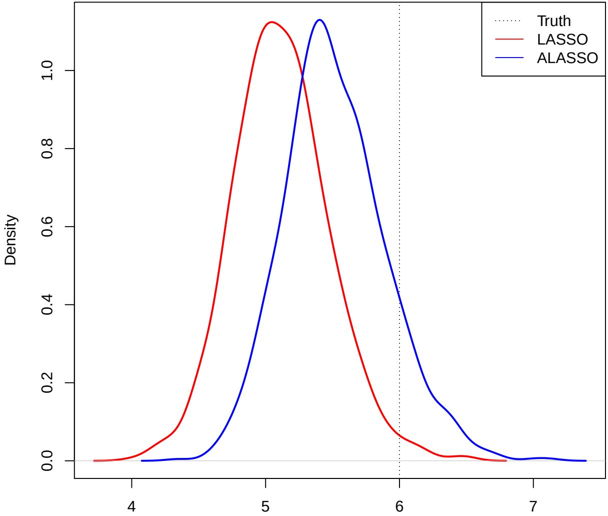
[2] JASA, "The Adaptive Lasso and its Oracle Properties".

p=50

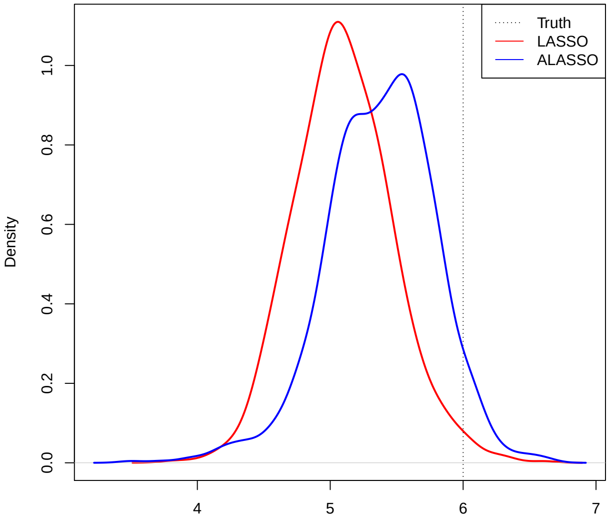


Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Truth
LASSO
ALASSO

p=70



p=90



Model Selection Consistency I

Often in applied work, LASSO is used not for prediction but for model selection. Under what conditions can we recover the “true” model?

Let $\mathbf{X}$ be the design matrix containing all potential variables; $\mathbf{X}_{\mathcal{T}}$ the submatrix formed by columns corresponding to “true” variables; and $\mathbf{X}_{\mathcal{F}}$ the submatrix formed by columns corresponding to “noise” variables.

Proofs for model selection consistency rely on the (necessary) condition

$$\| \left(\mathbf{X}_{\mathcal{T}}'\mathbf{X}_{\mathcal{T}}\right)^{-1}\mathbf{X}_{\mathcal{T}}'\mathbf{X}_{\mathcal{F}}\|_{\infty} \leq \theta \in (0,1).$$

The true and noise variables cannot be too correlated. Fairly stringent requirement, unclear how to test in finite samples.

Model Selection Consistency II

The choice of penalty parameter also affects model selection consistency.

Meinshausen and Bühlmann (2006)[3] show that the prediction-optimal choice of λ is guaranteed NOT to recover the true model asymptotically.

When λ is chosen by cross validation, we tend to overselect variables and include false positives.

We need stronger penalization to guarantee model selection consistency, but in finite samples the appropriate choice of λ is not clear.

Bottom line: in applied work, the LASSO is most useful for variable screening rather than model selection.

[3] Annals of Statistics, "High-Dimensional Graphs and Variable Selection with the LASSO".

Statistical Inference for the LASSO

We might also be interested in performing hypothesis testing on coefficient estimates from the LASSO.

One option is to put selected variables into an OLS regression. Highly problematic because no conditioning on model selection.

For example, the first variable selected by LASSO will be the one with the highest partial correlation with the response.

Suppose P is large and covariates are randomly generated. One of the covariates is likely to be highly correlated with the response; will be selected by LASSO; and have a significant p-value in an OLS regression.

This *post-selection inference* problem is not fully resolved in the statistics literature.

Options

One approach for inference is based on sample splitting. Estimate LASSO on $N/2$ datapoints, perform OLS regression on selected variables in other $N/2$ data points.

Issues:

1. Which $N/2$ observations? If repeated draws, how to treat variables that are sometimes selected and sometimes not?
2. Power for hypothesis tests reduced due to sample size in OLS.

[4] JRSSB, "Stability Selection".

Options

One approach for inference is based on sample splitting. Estimate LASSO on $N/2$ datapoints, perform OLS regression on selected variables in other $N/2$ data points.

Issues:

1. Which $N/2$ observations? If repeated draws, how to treat variables that are sometimes selected and sometimes not?
2. Power for hypothesis tests reduced due to sample size in OLS.

Another approach is based on bootstrapping. Estimate LASSO on bootstrap draws or subsamples, and compute fraction of times each variable is selected.

See Meinshausen and Bühlmann (2010)[4] for formal bounds on false inclusion probabilities.

Bootstrapping can also be seen as a frequentist approximation to Bayesian versions of the LASSO.

[4] JRSSB, "Stability Selection".

Example

In recent work with Michael McMahon and Matthew Tong, we study the impact of the release of the Bank of England's Inflation Report on bond price changes at different maturities.

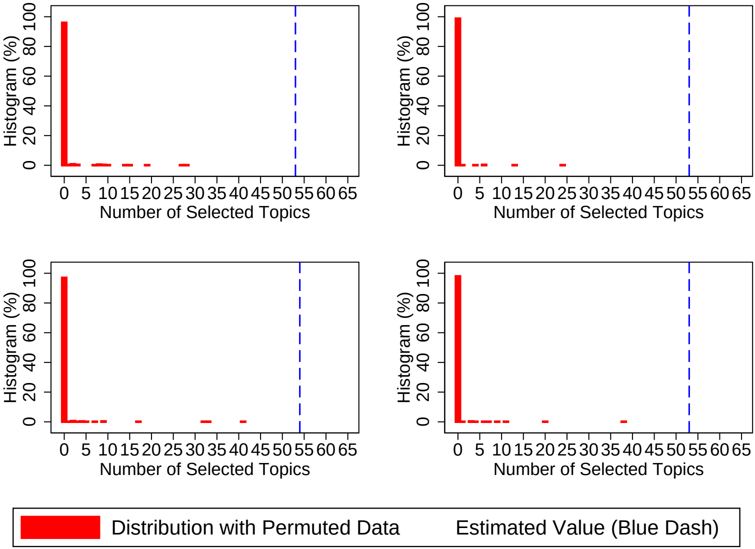
IR contains forecast variables we use as controls: (i) mode, variance, and skewness of inflation and GDP forecasts; (ii) their difference from the previous forecast.

To represent text, we estimate a 30-topic model and represent each IR in terms of (i) topic shares and (ii) evolution of topic shares from previous IR.

First step in the analysis is to partial out the forecast variables from bond price moves and topic shares by constructing residuals.

We are then left with 69 bond price moves (number of IRs in the data) and 60 text features.

LASSO-based Test of Information Content of Narrative



Which Information Matters?

LASSO selects dozens of features at all maturities: standard over-selection problem.

How to identify key topics?

We apply a non-parametric bootstrap to simulate the “inclusion probabilities” of topic features at different maturities.

Draw with replacement from our 69 observations to obtain new sample, perform LASSO, and record whether each feature is included.

Repeat 500 times, and rank topics according to the fraction of bootstrap draws in which they appear.

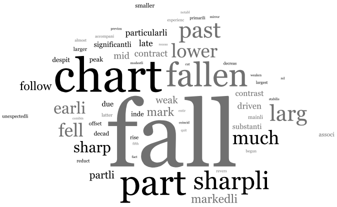
Results: Top Topic

Top Topics for Different Yields (L=Level; D=Change)

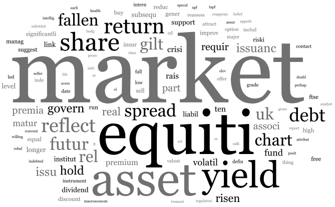
$\|\Delta i_{0:12,t}\|$		$\|\Delta f_{36,t}\|$		$\|\Delta f_{60,t}\|$		$\|\Delta f_{60:120,t}\|$	
Var	Selection %	Var	Selection %	Var	Selection %	Var	Selection %
L25	0.958	D24	0.858	L28	0.876	D17	0.91
D24	0.954	D25	0.844	D17	0.784	D18	0.896
L5	0.932	L28	0.826	D18	0.772	L20	0.836
L26	0.91	D14	0.76	L20	0.722	D13	0.808

Results: Top Topics

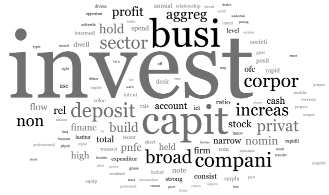
1-Year Spot Rate



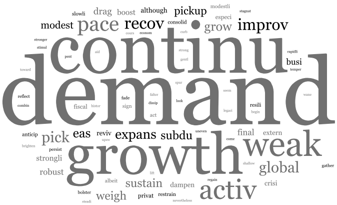
(a) L25



(b) D24



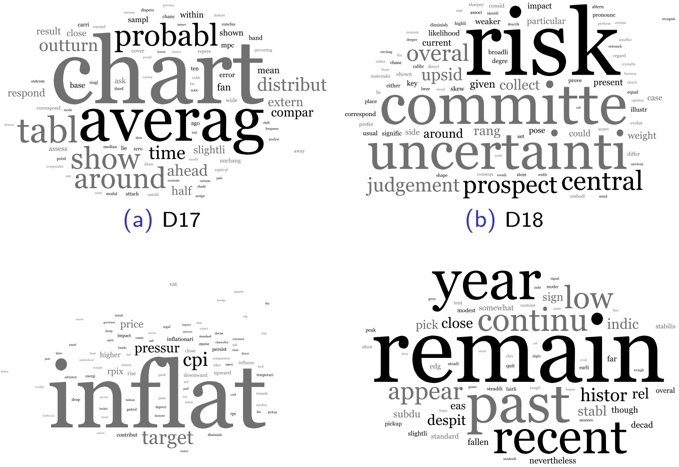
(c) L5



(d) L26

Results: Top Topics

5-Year, 5-Year Forward Rate



(a) D17

(b) D18

(c) L20

(d) D13

Treatment Effect Estimation

We now consider an economic application of the LASSO: treatment effect estimation with high-dimensional controls.

Belloni et. al. (2014)[5] provide an overview of current research on this topic in econometrics.

Here we provide a brief overview of Belloni et. al. (2014)[6]

[5] JEP, "High-Dimensional Methods and Inference on Structural and Treatment Effects."

[6] RESTUD, "Inference on Treatment Effects after Selection among High-Dimensional Controls."

Framework

$$\begin{aligned} y_i &= d_i\alpha_0 + g(\mathbf{z}_i) + \psi_i & \mathbb{E}[\,\psi_i \mid \mathbf{z}_i, d_i\,] = 0 \\ d_i &= m(\mathbf{z}_i) + \nu_i & \mathbb{E}[\,\nu_i \mid \mathbf{z}_i\,] = 0 \end{aligned}$$

d_i is the treatment—single dimensional for simplicity.

$\mathbf{z}_i$ is a vector of confounders and g and m are arbitrary nonlinear functions.

Observations are independent but not necessarily identically distributed.

We are interested on inference about α_0 that is robust to model selection mistakes.

Linear Approximations

We can approximate g and m with linear combinations of control terms $\mathbf{x}_i = P(\mathbf{z}_i)$ as:

$$g(\mathbf{z}_i) = \mathbf{x}_i'\boldsymbol{\beta}_{g0} + r_{gi}$$
$$m(\mathbf{z}_i) = \mathbf{x}_i'\boldsymbol{\beta}_{m0} + r_{mi}$$

where the r terms are approximation error.

$\mathbf{x}_i$ has dimensionality p potentially greater than n.

We can have $\mathbf{x}_i = \mathbf{z}_i$, or the P function could define complex non-linear transformations of $\mathbf{z}_i$.

Approximate Sparsity

Approximate sparsity captures that:

1. There are only a small number of relevant controls:

$$\|\boldsymbol{\beta}_{m0}\|_0 \leq s \text{ and } \|\boldsymbol{\beta}_{g0}\|_0 \leq s$$

2. The non-relevant controls have a high probability of being small:

$$\sqrt{\overline{\mathbb{E}}\left[\, r_{gi}^2 \,\right]} \lesssim \sqrt{s/n} \text{ and } \sqrt{\overline{\mathbb{E}}[\, r_{mi}^2 \,]} \lesssim \sqrt{s/n}$$

The identity of the s relevant variables is unknown.

Naive Procedure

Suppose we regress the outcome on the treatment and controls, and penalize the coefficients on the controls with the L1 norm but not the treatment.

We will tend to drop any control that is highly correlated with the treatment even if the control is moderately correlated with the outcome.

If we then perform an OLS regression of the outcome on the treatment and selected controls, the treatment effect will be contaminated by an omitted variable bias.

Recommended Procedure

The key idea of the paper is to use LASSO (or other variable selection procedures) twice:

1. Outcome is dependent variable and controls are regressors

   $$y_i = \mathbf{x}_i' \overline{\boldsymbol{\beta}}_0 + \overline{r}_i + \overline{\psi}_i$$

   where $\overline{\boldsymbol{\beta}}_0 \equiv \alpha_0 \boldsymbol{\beta}_{m0} + \boldsymbol{\beta}_{g0}$, $\overline{r}_i$ defined similarly, and $\overline{\psi}_i \equiv \alpha_0 \nu_i + \psi_i$.
2. Treatment is dependent variable and controls are regressors

   $$d_i = \mathbf{x}_i' \boldsymbol{\beta}_{m0} + r_{mi} + \nu_i.$$

Then perform an OLS regression of the outcome on the treatment and the union of the set of selected controls in 1. and 2., and use the coefficient estimate on the treatment as the treatment effect.

Any non-selected regressor has a small correlation with the treatment and with the control, which allows for asymptotically valid inference.

Advantage

The most important feature of this method is that it does not rely on the highly unrealistic assumption of perfect model selection which is often invoked to justify inference after model selection.

Post-Double-Selection Estimator

Let $\widehat{I}_1, \widehat{I}_2 \subset \{1, \ldots, p\}$ be the indices of the selected controls for the outcome and treatment, respectively.

The *post-double-selection estimator* is defined to be

$$(\check{\alpha}, \check{\boldsymbol{\beta}}) = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \mathbb{E}_n \big[(y_i - d_i \alpha - x_i' \boldsymbol{\beta})^2 \big] \ : \ \beta_j = 0, \forall j \notin \widehat{I} \right\}$$

Main Result

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}(0,1).$$

Main Result

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0) \xrightarrow{d} \mathcal{N}(0,1).$$

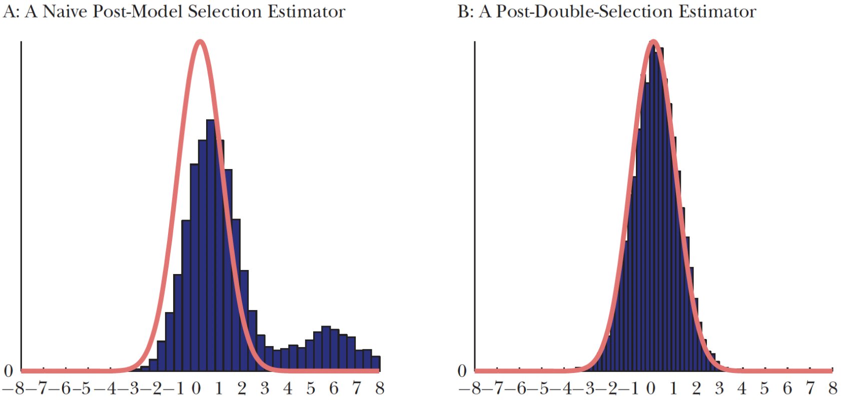
$$\sigma_n^2 = \left[\bar{\mathbb{E}}\left(v_i^2 \right)\right]^{-2} \bar{\mathbb{E}}\left(v_i^2 \psi_i^2 \right).$$

Can use plugin estimator for variance based on residuals

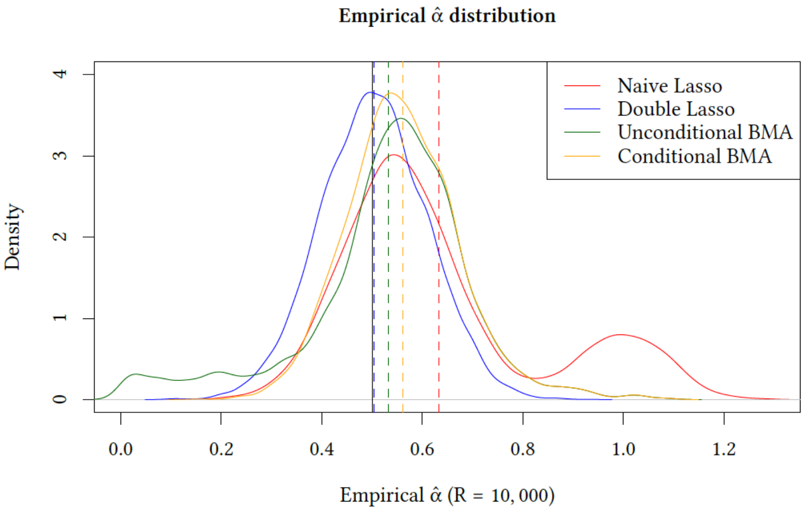
1. $\widehat{\psi}_i = \left(y_i - d_i\check{\alpha} - \mathbf{x}_i'\check{\boldsymbol{\beta}}\right)\left(\frac{n}{n-\hat{s}-1}\right)^{1/2}$
2. $\widehat{v}_i = d_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ where

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \mathbb{E}_n\left[(d_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \right] \ : \ \beta_j = 0, \forall j \notin \widehat{I} \right\}.$$

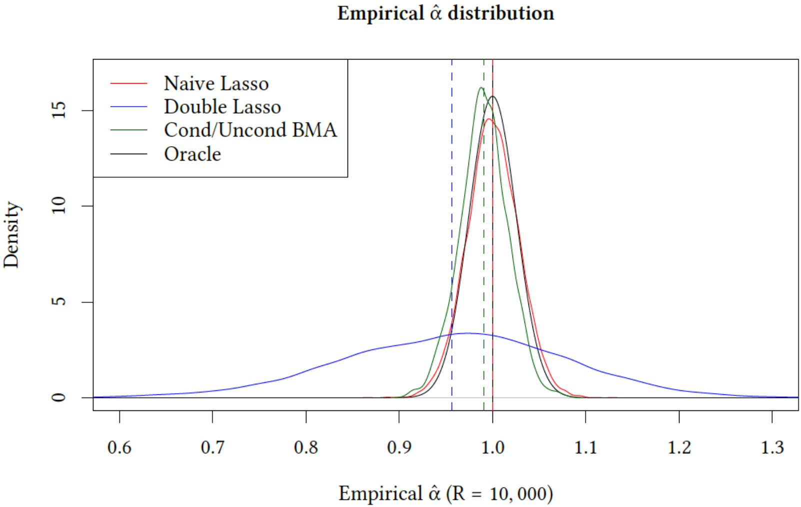
Sampling Distributions



Double ML vs Alternatives I



Double ML vs Alternatives II



Heterogeneous Treatment Effects

Another question of interest is whether a particular treatment effect has heterogeneous impacts according to observed characteristics.

Susan Athey has some papers on this question (see syllabus) that uses a modification of regression trees.

A regression tree is a model that explains an outcome by repeatedly partitioning the covariates.

One can provide conditions under which heterogeneous treatment effects can be estimated by looking at differences in treated and untreated groups in the leaves of the tree.

Conclusion

Penalized regression is a useful tool for settings with a large number of covariates.

LASSO is becoming a popular tool in economics, but important to keep in mind that its performance for consistent model selection is potentially poor.

Moreover, statistical inference for the LASSO is a developing field within statistics without definitive results.

LASSO can nevertheless be a useful tool within econometric models.