AME

Week 6: High Dimensional Models II

Sophie Bindslev, October 2022

UNIVERSITY OF COPENHAGEN

Today's Plan

- Inference in High Dimensional Models
- Why "Single" Post Lasso doesn't work
- Post Partialling Out Lasso
- Post Double Lasso
- Your time to shine!

Inference in High Dimensional Models I

- Linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

  where $\mathbf{Y}$ is $N \times 1$, $\mathbf{X}$ is $N \times p$ and $\boldsymbol{\beta}$ is $p \times 1$
- Last time we considered issues with High Dimensional Models for prediction (OLS prediction error)
- Today we consider inference in this context
- Recall that for $p > N$ the rank condition fails, OLS not defined
- Lasso can alleviate this problem if we believe sparsity applies i.e. only a subset of regressors' coefficients are non-zero

Inference in High Dimensional Models II

- Suppose we're interested in the causal effect of D on Y:

$$\mathbf{Y} = \alpha D + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{2}$$

- With plain Lasso no closed-form confidence interval is available and the asymptotic distribution of α is unknown
- Solution: use Post Double Lasso or Post Partialling Out Lasso for inference in High Dimensional Models

Naive Approach: Why Single Post Lasso doesn't work

- Single Post Lasso:
  1) Use Lasso of Y on D and $\mathbf{Z}_p$, forcing D to remain in model
  2) Run OLS of Y on D and $\mathbf{Z}_J \in \mathbf{Z}_p$ selected by Lasso
- Issue with this approach: Lasso's main objectives are prediction and sparsity. It shuts down coefficients on regressors which are highly correlated with one another since adding another one of these will only help marginally in improving predictive power of the model
- Lasso risks excluding exactly those variables correlated with our treatment variable, D, making Single Post Lasso susceptible to omitted variable bias
- The estimated treatment effect, $\hat{\alpha}$, risks capturing effects of excluded variables correlated with D

Post Partialling Out Lasso

- Method: **A)** Lasso Y on $\mathbf{Z}_p$, obtain residuals $\hat{v} = Y - \mathbf{Z}_{J,Y}\hat{\phi}$, **B)** Lasso D on $\mathbf{Z}_p$, obtain residuals $\hat{w} = D - \mathbf{Z}_{J,D}\hat{\delta}$, **C)** Estimate α as

$$\hat{\alpha} = \frac{\sum_{i=1}^{N}(Y_i - \mathbf{Z}_{i,J,Y}\hat{\phi})(D_i - \mathbf{Z}_{i,J,D}\hat{\delta})}{\sum_{i=1}^{N}(D_i - \mathbf{Z}_{i,J,D}\hat{\delta})^2} = \frac{\sum_{i=1}^{N}\hat{v}_i\hat{w}_i}{\sum_{i=1}^{N}\hat{w}_i^2} \tag{3}$$

- Under some conditions Post Partialling Out Lasso satisfies:

$$\sqrt{N}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \sigma^2), \quad \sigma^2 = \frac{E(v^2w^2)}{(E(w^2))^2} \tag{4}$$

- We can then obtain confidence intervals:

$$CI_{1-\xi} = (\hat{\alpha} \pm q_\xi \hat{\sigma}/\sqrt{N}) \tag{5}$$

where $q_\xi = \Phi^{-1}(\xi)$, we have e.g. $q_{0.025} = 1.96$

Post Double Lasso

- Method: **A)** Lasso Y on $\mathbf{Z}_p$ *and* D. Obtain residuals $\hat{u} = Y - \mathbf{Z}_{J,Y}\hat{\phi}$ constructed by *excluding* D and $\hat{v} = Y - \hat{\alpha}D - \mathbf{Z}_{J,Y}\hat{\phi}$, **B)** Lasso D on $\mathbf{Z}_p$, obtain residuals $\hat{w} = D - \mathbf{Z}_{J,D}\hat{\boldsymbol{\delta}}$, **C)** Estimate α as

$$\hat{\alpha} = \frac{\sum_{i=1}^{N}(Y_i - \mathbf{Z}_{i,J,Y}\hat{\phi})(D_i - \mathbf{Z}_{i,J,D}\hat{\boldsymbol{\delta}})}{\sum_{i=1}^{N}(D_i - \mathbf{Z}_{i,J,D}\hat{\boldsymbol{\delta}})D_i} = \frac{\sum_{i=1}^{N}\hat{u}_i\hat{w}_i}{\sum_{i=1}^{N}\hat{w}_i D_i} \tag{6}$$

- Under some conditions Post Double Lasso satisfies:

$$\sqrt{N}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \sigma^2), \quad \sigma^2 = \frac{E(v^2w^2)}{(E(w^2))^2} \tag{7}$$

- NB note that for the variance σ^2 we use $\hat{v}$ and not $\hat{u}$!
- Obtain confidence intervals as for Post Partialling Out Lasso

Post Double Lasso and Post Partialling Out Lasso

- Why do Post Double Lasso and Post Partialling Out Lasso avoid the issue of excluding exactly those regressors that might be a source of omitted variable bias?
- Intuitively, by running the second Lasso of D on controls $\mathbf{Z}_p$ (as well as Y on controls $\mathbf{Z}_p$) we are more likely to control for regressors $\mathbf{Z}_{J,D}$ which are highly correlated with treatment, D. These are exactly the ones that risk confounding our estimate of α
- These alternative estimators guard against omitted variable bias
- Post Double Lasso and Post Partialling Out Lasso are first order equivalent. That is, they have the same probability limit:

$$\sqrt{N}(\hat{\alpha}^{PDL} - \hat{\alpha}^{PPOL}) \to 0, \quad \text{as} \quad N \to \infty \tag{8}$$

Your time to shine!

- Solve the problem set
- Like last time the following functions will come in handy: `sklearn.linear_model.Lasso`, `sklearn.linear_model.LassoCV` and `sklearn.preprocessing.PolynomialFeatures`
- Remember to standardize your variables!
- Features such as `.predict_`, `.alpha_` and `.coef_` are especially useful when e.g. computing residuals which you'll need even more this time
- `norm.ppf` is useful for obtaining the inverse Normal CDF Φ^{-1}