Inference on Predictive and Causal Effects in High DimensionalLinear Regression Models

Alexander Quispe

September 25, 2021

Introduction

- We discuss inference on predictive effects using **double lasso methods**, where we use lasso twice to residualize outcomes and a target covariate of interest, **whose predictive effect we'd like to infer**.
- The predictive effects coincide with causal structural effects **under random assignment of treatment** conditional on controls.
- The **approach relies on approximate sparsity** of the best linear predictors for outcome and for the target covariate.

Introduction

Remember our main equation

$$Y = \alpha D + \beta' W + \epsilon \tag{1}$$

- How does the predicted value of Y change if a regressor D increases by a unit, while other regressors W remain unchanged? - **Predictive Effect Question**
- If conditioning on W, is sufficient for identification of the structural causal effect of D on Y, then
- how does predicted value of Y change if we intervene and increase D by a unit holding W fixed? - **Causal effect Question**

Inference with Double Lasso

- **Inference on One Coefficient**

The key will be the application of the FWL partialling-out.

$$Y = \alpha D + \beta' W + \epsilon \tag{2}$$

where D is the target regressor and W consists of p controls. After partialling-out,

$$\hat{Y} = \alpha \tilde{D} + \epsilon, \quad E\epsilon\tilde{D} = 0, \tag{3}$$

where the variables with tilde are residuals from taking out the linear effect of W

$$\tilde{D} = D - \gamma'_{DW} W, \quad \gamma_{DW} = \arg \min_{\gamma \in \mathbb{R}^p} E(D - \gamma' W)^2 \tag{4}$$

$$\tilde{Y} = Y - \gamma'_{YW} W, \quad \gamma_{YW} = \arg \min_{\gamma \in \mathbb{R}^p} E(Y - \gamma' W)^2 \tag{5}$$

Inference with Double Lasso

We now consider estimation of in high-dimensional setting. For estimation purposes we have a random sample $(Y_i, X_i)_{i=1}^n$.

- when p/n is mall, then Least Squares
- when p/n is not small, **then Lasso-based methods in the partialling-out steps.**

Inference with Double Lasso

1. Step1. We run the Lasso regressions of Y_i on W_i and D_i on W_i:

$$\hat{\gamma}_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i (Y_i - \gamma' W_i)^2 + \lambda_1 \sum_j \widehat{\psi}_j |\gamma_j|. \tag{6}$$

$$\hat{\gamma}_{DW} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i (D_i - \gamma' W_i)^2 + \lambda_2 \sum_j \widehat{\psi}_j |\gamma_j|. \tag{7}$$

2. Step2. Obtain the resulting residuals:

$$\check{Y}_i = Y_i - \hat{\gamma}'_{YW} W_i \tag{8}$$

$$\check{D}_i = D_i - \hat{\gamma}'_{DW} W_i \tag{9}$$

3. Step3. We run the **least squares** of $\check{Y}_i$ on $\check{D}_i$ to obtain the estimator $\check{\alpha}$:

$$\check{\alpha} = \arg\min_{a \in \mathbb{R}} \mathbb{E}_n(\check{Y}_i - a\check{D}_i)^2 = (\mathbb{E}_n \check{D}_i \check{D}_i)^{-1} \mathbb{E}_n \check{D}_i \check{Y}_i \tag{10}$$

Adaptive Inference with Double Lasso in High-Dimensional Regression

Under the stated approximate sparsity and additional regularity conditions

$$\sqrt{n}(\breve{\alpha} - \alpha) \approx \sqrt{n}\mathbb{E}_n\tilde{D}\epsilon / \mathbb{E}_n\tilde{D}^2 \sim^a N(0, V). \tag{11}$$

$$V = (E\tilde{D}^2)^{-1} E(\tilde{D}^2\epsilon^2)(E\tilde{D}^2)^{-1} \tag{12}$$

$$\sqrt{\hat{V}/n} \ [\breve{\alpha} \pm 2\sqrt{\hat{V}/n}] \tag{13}$$

Application to Testing Convergence Hypothesis

$$Y = \alpha D + \beta' W + \epsilon \tag{14}$$

- Y = economic growth rates
- D = the initial wealth levels in each country
- W = country's institutional, educational, and other similar characteristics
- α speed of convergence/divergence. which predicts the speed at which poor countries catch up.
- poor countries catch up $(\alpha < 0)$, or fall behind $(\alpha > 0)$
- The sample contains 90 countries and about 60 controls. Thus $p = 60$, $n = 90$ and p/n is not small.

Application to Testing Convergence Hypothesis

Figure: OLS and Double Lasso results

	Estimate	Std. Error	95% CI
OLS	-0.009	0.030	[-0.071, 0.052]
Double Lasso	-0.050	0.014	[-0.078, -0.022]

As expected, least squares provides a rather noisy estimate of the speed of convergence, and does not allow us to answer the question about the convergence hypothesis. In sharp contrast, double Lasso provides a more precise estimate. The lasso based point estimate is -5% and the 95% confidence interval for the (annual) rate of convergence is -7.8% to -2.2%. This empirical evidence does support the convergence hypothesis.

Inference on Many Coefficients

Here we consider the model

$$\underbrace{Y}_{Outcome} = \underbrace{\sum_{l=1}^{p_1} \alpha_l D_l}_{InterestingPredictors} + \underbrace{\sum_{j=1}^{p_2} \beta_j \overline{W}_j}_{Controls} + \epsilon \tag{15}$$

- p_1 = number predictors , is very large
- p_2 = number of controls, is also large

When do we want to consider many coefficients of interest?:

- there can be multiple policies whose predictive effect we would like to infer
- **we can be interested in heterogeneous predictive effects across groups**
- we can be interested in nonlinear effects of policies

$$(\bar{X}_l)_{l=1}^{p_1} \ , \ D_l = D_0 \overline{X}_l, \ \ l = 1, ..., p_1. \tag{16}$$

$(\bar{X}_l)_{l=1}^{p_1}$ are known transformations of controls $\overline{W}$, **for example various subgroups.**

One by One Double Lasso for Many Target Parameters.

For each $l = 1, ..., p_1$ we apply Double Lasso method in a one-by-one instance for estimation and inference on the coefficient $\alpha_l in the model$

$$Y = \alpha_l D_l + \gamma_l' W_l + \epsilon, \quad W_l = ((D_k)_{k \neq l}', \overline{W}')'. \tag{17}$$

Wage pay gap, CPS 2012 dataset

Heterogeneity of the wage pay gap on CPS 2012 data set. Here we interact the female indicator with group indicators capturing marital status, education groups, geographical regions, and a third degree polynomial in experience.

Figure: Lasso optimization with many coefficients.

	Estimate.	Std. Error	p-value
female	-0.15	0.05	0.00
female:widowed	0.14	0.09	0.13
female:divorced	0.14	0.02	0.00
female:separated	0.02	0.05	0.66
female:nevermarried	0.19	0.02	0.00
female:hsd08	0.03	0.12	0.82
female:hsd911	-0.12	0.05	0.02
female:hsg	-0.01	0.02	0.50
female:cg	0.01	0.02	0.58
female:ad	-0.03	0.02	0.16
female:mw	-0.00	0.02	0.96
female:so	-0.01	0.02	0.67
female:we	-0.00	0.02	0.84
female:exp1	0.00	0.01	0.53
female:exp2	-0.16	0.05	0.00
female:exp3	0.04	0.01	0.00

Neyman Orthogonality

It states that the target parameter is parameterized in terms of nuisance parameters in such away that small perturbations in terms of biased estimation of these parameters will translate to a negligible effect on estimating the target parameter.

- We have the nuisance projection parameters

$$\eta^0 = (\gamma'_{DW}, \gamma'_{YW})' \tag{18}$$

- The target parameter

$$\alpha(\eta) \tag{19}$$

- is first-order insensitive to local perturbations of these parameters:

$$D = \partial_n \alpha(\eta^0) = 0 \tag{20}$$

Mathematical Proof

$$M(a, \eta) = E[(\tilde{Y}(\eta_1) - a\tilde{D}(\eta_2))\tilde{D}(\eta_2)] = 0 \tag{21}$$

$$n := (\eta'_1, \eta'_2)' = \eta^0 := (\gamma'_{DW}, \gamma'_{YW})' \tag{22}$$

Neyman Orthogonality

$$\tilde{Y}(\eta_1) = Y - \eta_1' W, \quad \tilde{D}(\eta_2) = D - \eta_2' W \tag{23}$$

$$\tilde{Y} = Y - \gamma_{YW}' W, \quad \tilde{D} = D - \gamma_{DW}' W \tag{24}$$

$$D = -\partial_a M(\alpha, \eta^0)^{-1} \partial_n M(\alpha, \eta^0), \tag{25}$$

$$\partial_{n_1} M(\alpha, \eta^0) = E[W\tilde{D}] = 0 \tag{26}$$

$$\partial_{n_2} M(\alpha, \eta^0) = -E[\tilde{Y}W] + 2E[\alpha \tilde{D} W] = 0 \tag{27}$$

$$\widehat{M}(a, \hat{n}) = \mathbb{E}_n[(\check{Y} - a\check{Y})\check{D}] = 0, \quad \check{Y} = \tilde{Y}(\hat{\eta}_1), \quad \check{D} = \tilde{D}(\hat{\eta}_2) \tag{28}$$

Neyman Orthogonality

$$D = \partial_n \alpha(\eta^0) = 0 \ , \ \partial_n M(\alpha, \eta^0) = 0 \tag{29}$$

What happens if we don't have Neyman Orthogonality

(Invalid) Single Selection/Naive Method. In this method one applies Lasso regression of Y on D and W to select relevant covariates W_Y in addition to the covariate of interest, then refit the model by least squares of Y on D and W_Y and carry out conventional inference.