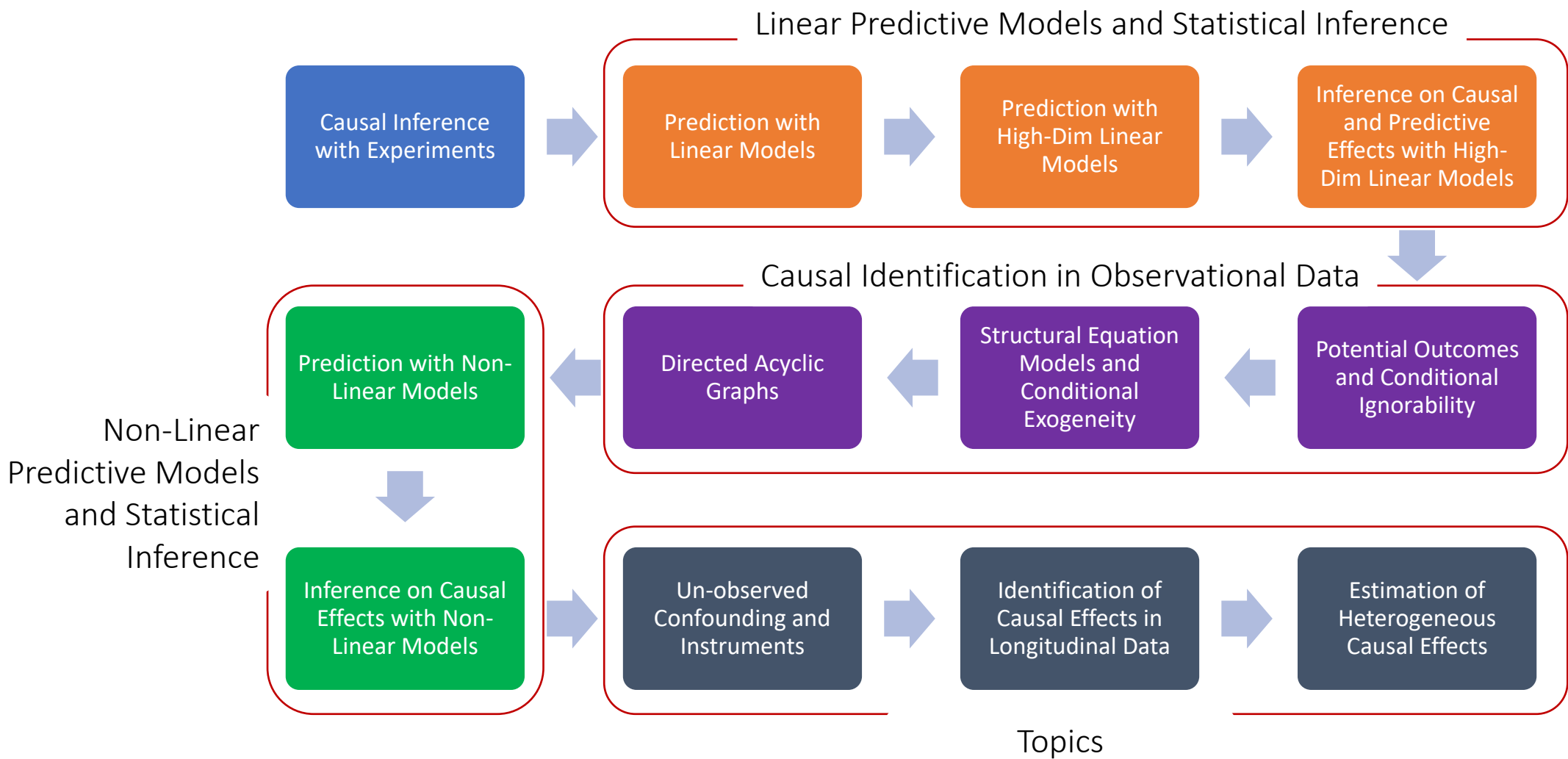
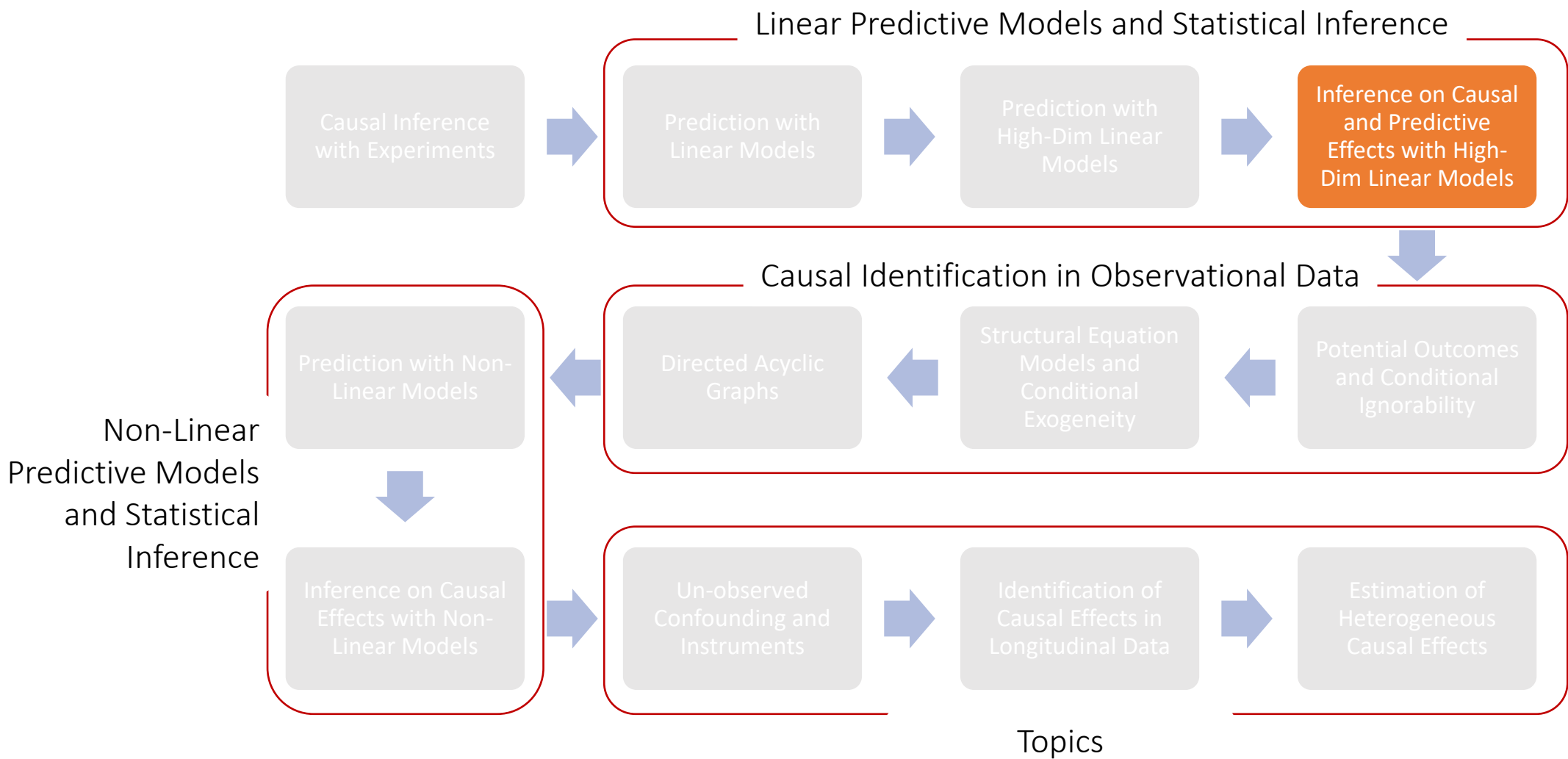


MS&E 228: Inference in High-Dimensional Linear Models

Vasilis Syrgkanis

MS&E, Stanford





Recap of Previous Lecture

High-dimensionality

$$p \gg n$$

Inherent: rich dataset with many covariates



Fabricated (engineered features): for flexible modeling that better approximates the true CEF

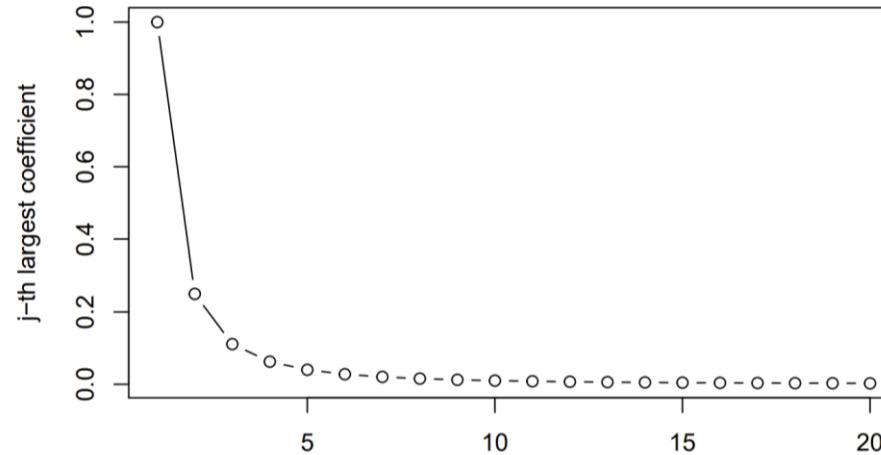


Not imposing any restrictions or biases on the parameters can lead to un-stable estimation in finite samples



Solution: add penalty terms to your estimation that induce biases towards solutions we a prior believe are more probable

Willing to believe that most of the parameters β are roughly zero



Then, penalize finite sample solutions $\hat{\beta}$ that have many non-zero and large coefficients:



$$\min_b \frac{1}{2} E_n[(Y - b'X)^2] + \lambda \|b\|_1, \quad (\text{LASSO})$$

Intuition: a covariate needs to introduce a large improvement in predictive performance to be included in the solution



$$\left| \partial_{b_j} E_n \left[(Y - \hat{\beta}' X)^2 \right] \right| \geq \lambda$$



Marginal benefit in prediction



Marginal increase in penalty

To avoid asymmetrically penalizing different features;
make sure you standardize your features

$$\tilde{X} = \frac{(X - E_n[X])}{\sqrt{Var_n(X)}}$$



Make sure you don't penalize the intercept (e.g. center the outcome and covariates before passing to a package if the package penalizes the intercept)



Theoretically driven penalty specification

$$\lambda = \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left(1 - \frac{\alpha}{2p} \right) \approx \sigma \sqrt{\frac{\log(p/\alpha)}{n}}$$

Under approximate sparsity, *restricted isometry condition (RIP)*, with probability approaching $1 - \alpha$:

$$\sqrt{E_X \left[(\beta' X - \hat{\beta}' X)^2 \right]} \leq \text{const} \cdot \sqrt{E[\epsilon^2]} \sqrt{\frac{s \log(p \vee n)}{n}}$$



s is roughly the number of non-zero (large) coefficients



Practical way to choose penalty: cross-validation

Watch out:

Post (Lasso-CV) OLS \neq (Post Lasso OLS)-CV

Different inductive biases lead to different penalties

Dense coefficients (many small)

$$\min_b \frac{1}{2} E_n[(Y - b'X)^2] + \lambda \|\beta\|_2^2, \quad (\text{Ridge})$$

Dense or Sparse coefficients

$$\min_b \frac{1}{2} E_n[(Y - b'X)^2] + \lambda \left((1 - \alpha) \|b\|_2^2 + \alpha \|b\|_1 \right), \quad (\text{ElasticNet})$$



Dense + Sparse coefficients

$$\min_{b=\gamma+\delta} \frac{1}{2} E_n[(Y - b'X)^2] + \lambda_1 \|\gamma\|_2^2 + \lambda_2 \|\delta\|_1, \quad (\text{LAVA})$$

Confidence Intervals in High Dimensions

Inference on Predictive Effect

- Partition $X = (D, W)$, for some *target* regressor D of interest

$$Y = \alpha D + \beta' W + \epsilon$$

- Construct a confidence interval for the predictive effect/coefficient α
- Even when W is high-dimensional $p \gg n$

Revisit Partialling-Out Interpretation of OLS

Understanding α

- Consider the following partialling out operation
- For any random variable V , let \tilde{V} be the residual of V after subtracting the part of V that is linearly predictable from W

$$\tilde{V} = V - \gamma'_V W, \quad \gamma_W \in \operatorname{argmin}_{\gamma} E[(V - \gamma'W)^2]$$

- Note that we can also write the standard decomposition:

$$V = \gamma'_V W + \tilde{V}, \quad E[\tilde{V}W] = 0$$

Frisch-Waugh-Lovell (FWL) Theorem!

- The population linear regression coefficient α can be recovered from the population linear regression of \tilde{Y} on \tilde{D}

$$\alpha = \operatorname{argmin}_a E \left[(\tilde{Y} - a \tilde{D})^2 \right] = \frac{E[\tilde{Y} \tilde{D}]}{E[\tilde{D}^2]}$$

- We made the assumption that $E[\tilde{D}^2] > 0$, i.e. D is not perfectly linearly predictable from W

Predictive effect α of *target variable* is the coefficient in a *simple one variable regression*



$$\left(\begin{array}{c} \text{part of outcome} \\ \text{(un-explained by other)} \end{array} \right) \sim \left(\begin{array}{c} \text{part of target} \\ \text{(un-explained by other)} \end{array} \right)$$

FWL in samples for low dimensions $p \ll n$

Coefficient $\hat{\alpha}$ of D in $\text{OLS}(y \sim D, W)$ is mathematically equivalent in samples to

$y_{\text{res}} = y - \text{OLS}(y \sim W).\text{predict}(W)$

$D_{\text{res}} = D - \text{OLS}(D \sim W).\text{predict}(W)$



$\hat{\alpha}$ is coefficient of D_{res} in $\text{OLS}(y_{\text{res}} \sim D_{\text{res}})$

What can we do for $p \gg n$!?

FWL in samples for high dimensions $p \gg n$

Coefficient of D in $\text{OLS}(y \sim D, W)$ is mathematically
equivalent in samples to

$y_{\text{res}} = y - \text{Lasso}(y \sim W).\text{predict}(W)$
 $D_{\text{res}} = D - \text{Lasso}(D \sim W).\text{predict}(W)$



$\hat{\alpha}$ is coefficient of D_{res} in $\text{OLS}(y_{\text{res}} \sim D_{\text{res}})$

Mathematically

- For any random variable V , let \check{V} be the residual of V after subtracting the part of V that is linearly predictable from W in sample with Lasso

$$\check{V} = V - \hat{\gamma}'_V W, \quad \hat{\gamma}_V \in \operatorname{argmin}_{\gamma} E_n[(V - \gamma'W)^2] + \lambda_V \|\gamma\|_1$$

- An estimate $\hat{\alpha}$ of the predictive effect α can be recovered from the sample linear regression of \check{Y} on \check{D}

$$\hat{\alpha} = \operatorname{argmin}_a E_n [(\check{Y} - a \check{D})^2] = \frac{E_n[\check{Y}\check{D}]}{E_n[\check{D}^2]}$$

Adaptive Inference

- Under regularity conditions, if effective dimension s of γ_D, γ_Y is $\ll \sqrt{n}$, the estimation error in \check{D}_i, \check{Y}_i has no first-order effect on the asymptotic stochastic behavior of $\hat{\alpha}$

$$\sqrt{n} (\hat{\alpha} - \alpha) \approx \sqrt{n} \frac{E_n[\epsilon \tilde{D}]}{E_n[\tilde{D}^2]}$$

- By application of LLN and CLT

$$\sqrt{n}(\hat{\alpha} - \alpha) \overset{a}{\sim} N(0, V), \quad V = \frac{E[\epsilon^2 \tilde{D}^2]}{E[\tilde{D}^2]^2}$$

If we want an interval that roughly contains the predictive effect with probability 95%, we can use



$$CI := \hat{\alpha} \pm 1.96 \sqrt{\hat{V}/n}, \quad \hat{V} := \frac{E_n[\check{\epsilon}^2 \check{D}^2]}{E_n[\check{D}^2]^2}$$

Why Partialling-Out Works: Neyman Orthogonality

Target and Nuisance Parameters

- In the double lasso we have a **target parameter** of interest α
- But we also have other parameters that we estimate $\eta^o = (\gamma_Y, \gamma_D)$
- We don't care about these parameters and their estimation error
- We will call any such parameter that we need to estimate but don't care about it in its own shake a “**nuisance parameter**”

Target Estimate Parameterized by Nuisances

- Useful to write target parameter estimate as a function of nuisances

$$\hat{\alpha}(\eta)$$

- e.g. for double lasso, for any value η we can define the residuals

$$\left(\check{Y}_i(\eta), \check{D}_i(\eta) \right) = (Y_i - \eta'_1 W_i, D_i - \eta'_2 W_i)$$

- then estimate $\hat{\alpha}(\eta)$ is the solution to

$$M_n(a, \eta) := E_n \left[\left(\check{Y}(\eta) - a \check{D}(\eta) \right) \check{D}(\eta) \right] = 0$$

Population Analogue of Estimation Process

- The estimation process typically has a population analogue, that expresses the target parameter as a function of nuisances in the population limit, and which closely approximates the sample process

$$\alpha(\eta)$$

- e.g. for double lasso, for any value η we can define the residuals

$$\left(\tilde{Y}(\eta), \tilde{D}(\eta) \right) = (Y - \eta'_1 W, D - \eta'_2 W)$$

- Then $\alpha(\eta)$ is the solution to

$$M(a, \eta) := E \left[\left(\tilde{Y}(\eta) - a \tilde{D}(\eta) \right) \tilde{D}(\eta) \right] = 0$$

Insensitivity to Nuisances

- The estimation process is **Neyman orthogonal to the nuisances** if the population analogue $\alpha(\eta)$ of our estimation procedure is first-order insensitive to perturbations of the nuisances around their true value

$$\partial_{\eta} \alpha(\eta^0) = 0, \quad (\text{Neyman Orthogonality})$$

- For any parameter defined as the solution to an equation $M(a, \eta) = 0$
- By implicit function theorem

$$\partial_{\eta} \alpha(\eta^0) = -\partial_a M(\alpha^0, \eta^0)^{-1} \partial_{\eta} M(\alpha^0, \eta^0)$$

- It suffices that $\partial_{\eta} M(\alpha^0, \eta^0) = 0$

Double Lasso is Neyman Orthogonal

- For the double lasso

$$M(a, \eta) := E \left[\left(\tilde{Y}(\eta) - a\tilde{D}(\eta) \right) \tilde{D}(\eta) \right]$$

- And note

$$\tilde{Y}(\eta^o) \equiv \tilde{Y} = Y - \eta_1^o{}' W, \quad \tilde{D}(\eta^o) \equiv \tilde{D} = D - \eta_2^o{}' W$$

- We can verify orthogonality

$$\partial_{\eta_1} M(\alpha^o, \eta^o) = E[-W \tilde{D}] = 0$$

$$\partial_{\eta_2} M(\alpha^o, \eta^o) = E[-W(\tilde{Y} - \alpha^o \tilde{D})] + E[W \alpha^o \tilde{D}] = 0$$

Estimation process for parameter of interest α that depends on nuisance parameters η with true values η^0 is Neyman orthogonal if

$$\partial_{\eta} a(\eta^0) = 0$$

If α solution to an equation $M(a, \eta) = 0$ then, equation is Neyman orthogonal if

$$\partial_{\eta} M(\alpha^0, \eta^0) = 0$$



Lasso is not Neyman Orthogonal

- Suppose we run a single Lasso of Y on (D, W)
- If we fix the rest of the coefficients η that correspond to W
- Lasso in the limit and as $\lambda \rightarrow 0$ can be thought as finding α by solving the Normal equation

$$M(\alpha, \eta) := E[(Y - \alpha D - \eta'W) D]$$

- This is the population analogue of the single lasso process parameterized by the nuisances η

$$\partial_{\eta} M(\alpha, \eta^0) = E[-WD] \neq 0$$

- Exception: if D is from RCT (e.g. independent of W) and de-meanded

Lasso is not Neyman Orthogonal

- Population nuisance parameterized process is the same
- Single lasso approach selects primarily strong predictors of the outcome
- But can fail to omit strong predictors of the target
- Thus can partially omit “confounders”
- The double Lasso approach controls for both strong predictors of the target and strong predictors of the outcome; by running the two prediction problems separately

Inference on Many Coefficients

Inference on Many Predictive Effects

- We want to do inference on predictive effects of many *target* regressors D_1, \dots, D_m of interest

$$Y = \underbrace{\sum_{\ell=1}^m \alpha_{\ell} D_{\ell}}_{\text{target predictors}} + \underbrace{\beta' W}_{\text{controls}} + \epsilon$$

- Construct joint confidence intervals for the predictive effects
- Even when W is high-dimensional $p \gg n$

Motivating Examples

- Multiple treatment policies to be evaluated (e.g. 5 treatments in Re-employment bonus experiment)
- Inference on treatment effect heterogeneity; how do different factors modify the effect; can be accomplished by interactions $D_\ell = D \cdot W_\ell$
- Non-linear effects of policies with continuous treatments, e.g. pricing

Joint Confidence Intervals

- We want to produce a set of intervals $[\underline{\alpha}_\ell, \bar{\alpha}_\ell]$ for each α_ℓ
- With probability 95%

$$\forall \ell: \alpha_\ell \in [\underline{\alpha}_\ell, \bar{\alpha}_\ell]$$

- If we just construct confidence intervals for each α_ℓ separately as usual, then this only guarantees that this probability holds if a priori we fixed which coordinate we want to examine
- If we want to be able to examine “after the fact” all coordinates and pay attention to one that is interesting and look at that entry’s CI, then we need a joint confidence interval so that all coordinates are covered simultaneously (multiple hypotheses testing)

Estimation: One-by-One Double Lasso

For $\ell = 1, \dots, m$ apply the Double Lasso process to infer α_ℓ

- Treat D_ℓ as the target regressor
- Treat other treatments $(D_k)_{k \neq \ell}$ and original controls W as controls
- Estimate BLP in the decomposition

$$Y = \alpha_\ell D_\ell + \gamma'_\ell X_\ell + \epsilon, \quad X_\ell = ((D_k)_{k \neq \ell}, W)$$

- Using the Double Lasso process

Joint Adaptive Inference

- Under regularity conditions, if effective dimension s of γ_D, γ_Y is $\ll \sqrt{n}$, and the Restrict Isometry Property holds, the estimation error $\tilde{D}_{i,\ell}, \check{Y}_{i,\ell}$ has no first-order effect on the asymptotic stochastic behavior of $\hat{\alpha}_\ell$.
- If we don't have exponentially many target coefficients, $\log(m)^5 \ll n$ the vector $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_m)$ is approximately jointly Gaussian

$$\sqrt{n}(\hat{\alpha} - \alpha) \overset{a}{\sim} N(0, V), \quad V_{\ell k} = \frac{E[\tilde{D}_\ell \tilde{D}_k \epsilon^2]}{E[\tilde{D}_\ell^2] \cdot E[\tilde{D}_k^2]}$$

Joint Confidence Interval

- Joint asymptotic normality means that for the set of all hyper-rectangles \mathcal{R}

$$\sup_{A \in \mathcal{R}} |\Pr(\sqrt{n}(\hat{\alpha} - \alpha) \in A) - \Pr(N(0, V) \in A)| \rightarrow 0$$

- We can construct joint confidence intervals of the form:

$$CR = \times_{\ell} \left[\hat{\alpha}_{\ell} \pm c \sqrt{\hat{V}_{\ell\ell}/n} \right]$$

- Note that

$$\Pr(a \in CR) = \Pr \left(\max_{\ell} \left| \frac{\sqrt{n}(a_{\ell} - \hat{a}_{\ell})}{\sqrt{\hat{V}_{\ell\ell}}} \right| \leq c \right)$$

- By Gaussian approximation, for $\Delta = \text{diag}(V)$

$$\Pr \left(\max_{\ell} \left| \frac{\sqrt{n}(a_{\ell} - \hat{a}_{\ell})}{\sqrt{\hat{V}_{\ell\ell}}} \right| \leq c \right) \approx \Pr \left(\|N(0, \Delta^{-1/2} V \Delta^{-1/2})\|_{\infty} \leq c \right)$$

By Gaussian approximation, choose c as the $1 - \alpha$ quantile of the maximum entry in a gaussian vector drawn with covariance $\Delta^{-1/2} V \Delta^{-1/2}$

$$\Delta := \text{diag}(V) = \begin{bmatrix} V_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & V_{mm} \end{bmatrix}$$



For 95% confidence interval, c slightly larger than 1.96