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

Vasilis Syrgkanis

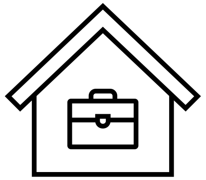
MS&E, Stanford

Recap of Previous Lecture

High-dimensionality

$$p \gg n$$

Inherent: rich dataset with many covariates

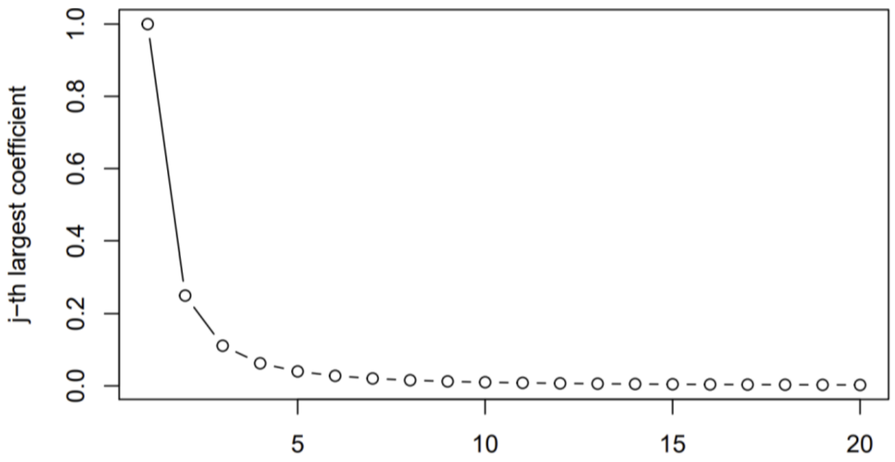


Fabricated (engineered features): for flexible modeling that better approximates the true CEF

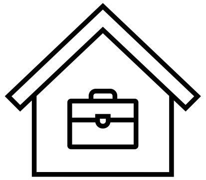
Not imposing any restrictions or biases on the parameters can lead to un-stable estimation in finite samples

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:

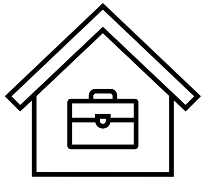


$$\min_b \frac{1}{2} E_n\left[(Y - b'X)^2\right] + \lambda \|b\|_1 , \qquad \text{(LASSO)}$$

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

$$\underbrace{\left|\partial_{b_j} E_n\left[\left(Y - \hat{\beta}' X\right)^2\right]\right|}_{\text{Marginal benefit in prediction}} \geq \underbrace{\lambda}_{\text{Marginal increase in penalty}}$$

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



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

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 $1-\alpha$:

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

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)

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

Dense or Sparse coefficients

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

Dense + Sparse coefficients

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

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

$$\tilde{V} = V - \gamma_V' W, \qquad \gamma_W \in \mathrm{argmin}_\gamma E\left[(V \; - \gamma' W)^2\right]$$

- Note that we can also write the standard decomposition:

$$V = \gamma_V' W + \tilde{V}, \qquad E\left[\tilde{V} W\right] = 0$$

Frisch-Waugh-Lovell (FWL) Theorem!

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

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

- We made the assumption that $E\left[\widetilde{D}^2\right] > 0$, 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*

$$\begin{pmatrix}\text{part of outcome}\\ \text{un-explained by other}\end{pmatrix} \sim \begin{pmatrix}\text{part of target}\\ \text{un-explained by other}\end{pmatrix}$$

FWL in samples for low dimensions p<<n

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

yres = y - OLS(y∼W).predict(W)
Dres = D - OLS(D∼W).predict(W)

$\hat{\alpha}$ is coefficient of Dres in OLS(yres∼Dres)

What can we do for p>>n!?

FWL in samples for high dimensions p>>n

Coefficient of D in OLS(y~D,W) is mathematically equivalent in samples to

Double Lasso

yres = y - Lasso(y~W).predict(W)
Dres = D - Lasso(D~W).predict(W)

$\hat{\alpha}$ is coefficient of Dres in OLS(yres~Dres)

Coding Example

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

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

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

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

Adaptive Inference

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

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

- By application of LLN and CLT

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

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

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

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

$$\left(\check{Y}_i(\eta), \breve{D}_i(\eta)\right) = (Y_i - \eta_1' W_i, D_i - \eta_2' W_i)$$

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

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

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

$$\left(\tilde{Y}(\eta), \widetilde{D}(\eta)\right) = (Y - \eta_1' W, D - \eta_2' W)$$

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

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

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

$$D := \partial_\eta \alpha(\eta^o) = 0, \qquad \text{(Neyman Orthogonality)}$$

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

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

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

Double Lasso is Neyman Orthogonal

- For the double lasso

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

- And note

$$\tilde{Y}(\eta^o) \equiv \tilde{Y} = Y - \gamma_Y' W, \qquad \widetilde{D}(\eta^o) \equiv \widetilde{D} = D - \gamma_D' W$$

- We can verify orthogonality

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

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

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

$$D \coloneqq \partial_\eta a(\eta^o) = 0$$

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

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

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 $\lambda \to 0$ can be thought as finding α by solving the Normal equation

$$M(a, \eta) \coloneqq E[(Y - aD \; - \eta' W)\, D]$$

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

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

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

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 target; by running the two prediction problems separately

Coding Example

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 $[\underline{\alpha}_\ell, \overline{\alpha}_\ell]$ for each α_ℓ
- With probability 95%

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

- 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 entries 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

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

- 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 $\breve{D}_{i,\ell}, \breve{Y}_{i,\ell}$ 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 $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_m)$ is approximately jointly Gaussian

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

Joint Confidence Interval

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

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

- 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 $D=\mathrm{diag}(V)$

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

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

$$\mathrm{D} \coloneqq \mathrm{diag}(\mathrm{V}) = \begin{bmatrix} V_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & V_{mm} \end{bmatrix}$$

For 95% confidence interval, c slightly larger than 1.96