A New Perspective on High-Dimensional Causal Inference

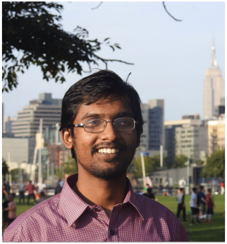
Pragya Sur
Dept. of Statistics
Harvard University

Deep Learning Theory Workshop
Simons Institute for the Theory of Computing
Aug 3, 2022

Collaborators

Kuanhao Jiang

Rajarshi Mukherjee



Subhabrata Sen

Outline

- *Problem: causal effect estimation from observational studies*
- *Goal: understand in high-dimensional settings without sparsity assumptions*

Outline

- *Problem: causal effect estimation from observational studies*
- *Goal: understand in high-dimensional settings without sparsity assumptions*
- *Main result: a new central limit theorem*
- *Some insights into proof techniques*

Outline

- *Problem: causal effect estimation from observational studies*
- *Goal: understand in high-dimensional settings without sparsity assumptions*
- *Main result: a new central limit theorem*
- *Some insights into proof techniques*
- *Opportunities and implications for machine learning*

Causal effect estimation from observational studies

The problem: Observe n i.i.d. samples $(Y_i, A_i, \boldsymbol{X}_i)$.

– Outcome $Y_i \in \mathbb{R}$, treatment $A_i \in \{0, 1\}$, covariates $\boldsymbol{X}_i \in \mathbb{R}^p$.

Causal effect estimation from observational studies

The problem: Observe n i.i.d. samples $(Y_i, A_i, \boldsymbol{X}_i)$.

– Outcome $Y_i \in \mathbb{R}$, treatment $A_i \in \{0, 1\}$, covariates $\boldsymbol{X}_i \in \mathbb{R}^p$.

– *Can we estimate effect of the treatment on the outcome?*

Causal effect estimation from observational studies

The problem: Observe n i.i.d. samples $(Y_i, A_i, \boldsymbol{X}_i)$.

– Outcome $Y_i \in \mathbb{R}$, treatment $A_i \in \{0, 1\}$, covariates $\boldsymbol{X}_i \in \mathbb{R}^p$.

– *Can we estimate effect of the treatment on the outcome?*

- Issue: For unit i, observe only outcome for assigned treatment!

Causal effect estimation from observational studies

The problem: Observe n i.i.d. samples $(Y_i, A_i, \boldsymbol{X}_i)$.
– Outcome $Y_i \in \mathbb{R}$, treatment $A_i \in \{0, 1\}$, covariates $\boldsymbol{X}_i \in \mathbb{R}^p$.
– *Can we estimate effect of the treatment on the outcome?*

- Issue: For unit i, observe only outcome for assigned treatment!
- One approach: The potential outcomes framework (Neyman-Rubin)
  –Denote $Y_i(t)$ to be outcome we would observe if treatment t assigned.

Causal effect estimation from observational studies

The problem: Observe n i.i.d. samples $(Y_i, A_i, \boldsymbol{X}_i)$.

– Outcome $Y_i \in \mathbb{R}$, treatment $A_i \in \{0, 1\}$, covariates $\boldsymbol{X}_i \in \mathbb{R}^p$.

– *Can we estimate effect of the treatment on the outcome?*

- Issue: For unit i, observe only outcome for assigned treatment!
- One approach: The potential outcomes framework (Neyman-Rubin)
  –Denote $Y_i(t)$ to be outcome we would observe if treatment t assigned.
- Causal effects take multiple forms:

  E.g: (1) The average treatment effect: $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$

  (2) Conditional average treatment effect: $\tau(\boldsymbol{x}) = \mathbb{E}[Y_i(1) - Y_i(0) | \boldsymbol{X} = \boldsymbol{x}]$

  $\vdots$

Causal effect estimation from observational studies

The problem: Observe n i.i.d. samples $(Y_i, A_i, \boldsymbol{X}_i)$.

– Outcome $Y_i \in \mathbb{R}$, treatment $A_i \in \{0, 1\}$, covariates $\boldsymbol{X}_i \in \mathbb{R}^p$.

– *Can we estimate effect of the treatment on the outcome?*

- Issue: For unit i, observe only outcome for assigned treatment!
- One approach: The potential outcomes framework (Neyman-Rubin)
  –Denote $Y_i(t)$ to be outcome we would observe if treatment t assigned.
- Causal effects take multiple forms: This talk

  E.g: (1) The average treatment effect: $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$

  (2) Conditional average treatment effect: $\tau(\boldsymbol{x}) = \mathbb{E}[Y_i(1) - Y_i(0) | \boldsymbol{X} = \boldsymbol{x}]$

  $\vdots$

Conditions for ATE identification

- Unidentifiable from observational studies (in general).

Conditions for ATE identification

- Unidentifiable from observational studies (in general).
- Assume structure on observed data distribution so identifiable.
  - *No unmeasured confounding:* $Y(1), Y(0) \perp A|\boldsymbol{X}$
  - *Consistency:* $Y = AY(1) + (1-A)Y(0)$
  - *Positivity*: $\mathbb{P}[A = 1|\boldsymbol{X} = \boldsymbol{x}] > 0$

Conditions for ATE identification

- Unidentifiable from observational studies (in general).
- Assume structure on observed data distribution so identifiable.
  - *No unmeasured confounding:* $Y(1), Y(0) \perp A|\boldsymbol{X}$
  - *Consistency:* $Y = AY(1) + (1-A)Y(0)$
  - *Positivity*: $\mathbb{P}[A = 1|\boldsymbol{X} = \boldsymbol{x}] > 0$

- The ATE can be identified from observational data using

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}(Y|A = 1, \boldsymbol{X}) - \mathbb{E}(Y|A = 0, \boldsymbol{X})]$$

ATE estimation: A well-studied problem

- **General theme:** The estimation problem involves two nuisance functions:

ATE estimation: A well-studied problem

- **General theme:** The estimation problem involves two nuisance functions:
  - *The propensity score:* $\pi = \mathbb{E}[A|\boldsymbol{X}] = \mathbb{P}(A = 1|\boldsymbol{X})$

ATE estimation: A well-studied problem

- **General theme:** The estimation problem involves two nuisance functions:
  - *The propensity score:* $\pi = \mathbb{E}[A|\boldsymbol{X}] = \mathbb{P}(A = 1|\boldsymbol{X})$
  - *The outcome regressions:* $m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}]$, $m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$

ATE estimation: A well-studied problem

- **General theme:** The estimation problem involves two nuisance functions:
  - *The propensity score:* $\pi = \mathbb{E}[A|\boldsymbol{X}] = \mathbb{P}(A = 1|\boldsymbol{X})$
  - *The outcome regressions:* $m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}]$, $m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$

- Multiple approaches exist:

  (1) Inverse Probability Weighting (Horvitz and Thompson '52, Rosenbaum and Rubin '83)

ATE estimation: A well-studied problem

- **General theme:** The estimation problem involves two nuisance functions:
  - *The propensity score:* $\pi = \mathbb{E}[A|\boldsymbol{X}] = \mathbb{P}(A = 1|\boldsymbol{X})$
  - *The outcome regressions:* $m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}]$, $m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$

- Multiple approaches exist:

  (1) Inverse Probability Weighting (Horvitz and Thompson '52, Rosenbaum and Rubin '83)

  (2) G-computation (Robins '86)

ATE estimation: A well-studied problem

- **General theme:** The estimation problem involves two nuisance functions:
  - *The propensity score:* $\pi = \mathbb{E}[A|\boldsymbol{X}] = \mathbb{P}(A = 1|\boldsymbol{X})$
  - *The outcome regressions:* $m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}]$, $m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$

- Multiple approaches exist:

  (1) Inverse Probability Weighting (Horvitz and Thompson '52, Rosenbaum and Rubin '83)

  (2) G-computation (Robins '86)

  (3) Augmented Inverse Probability Weighting (Robins, Rotnitzky, Zhao '94; Rotnitzky, Robins, Scharfstein '98; Scharfstein, Rotnitzky, Robins '99; Bang and Robins '05); Targeted MLE (van der Laan and Rubin '06)

ATE estimation: A well-studied problem

- **General theme:** The estimation problem involves two nuisance functions:
  - *The propensity score:* $\pi = \mathbb{E}[A|\boldsymbol{X}] = \mathbb{P}(A = 1|\boldsymbol{X})$
  - *The outcome regressions:* $m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}]$, $m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$

- Multiple approaches exist:

  (1) Inverse Probability Weighting (Horvitz and Thompson '52, Rosenbaum and Rubin '83)

  (2) G-computation (Robins '86)

  (3) Augmented Inverse Probability Weighting (Robins, Rotnitzky, Zhao '94; Rotnitzky, Robins, Scharfstein '98; Scharfstein, Rotnitzky, Robins '99; Bang and Robins '05); Targeted MLE (van der Laan and Rubin '06)

The Augmented Inverse Probability Weighting

- Recall our nuisance functions:
  $\pi = \mathbb{P}(A = 1|\boldsymbol{X}), \quad m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}], \quad m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$
- The AIPW computes estimates for the above, then uses a plug-in principle.

The Augmented Inverse Probability Weighting

- Recall our nuisance functions:
  $\pi = \mathbb{P}(A = 1|\boldsymbol{X}), \quad m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}], \quad m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$
- The AIPW computes estimates for the above, then uses a plug-in principle.
- Suppose $\hat{\pi}_i, \hat{m}_i^{(1)}, \hat{m}_i^{(0)}$ denote estimates for i-th sample.
- The AIPW is given by $\hat{\tau}_{\mathsf{AIPW}} = \hat{\tau}_{\mathsf{AIPW},1} - \hat{\tau}_{\mathsf{AIPW},0}$, where

The Augmented Inverse Probability Weighting

- Recall our nuisance functions:
  $\pi = \mathbb{P}(A = 1|\boldsymbol{X}), \quad m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}], \quad m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$
- The AIPW computes estimates for the above, then uses a plug-in principle.
- Suppose $\hat{\pi}_i, \hat{m}_i^{(1)}, \hat{m}_i^{(0)}$ denote estimates for i-th sample.
- The AIPW is given by $\hat{\tau}_{\mathsf{AIPW}} = \hat{\tau}_{\mathsf{AIPW},1} - \hat{\tau}_{\mathsf{AIPW},0}$, where

$$\hat{\tau}_{\mathsf{AIPW},1} = \frac{1}{n} \sum_i \Big[\frac{A_i Y_i}{\hat{\pi}_i} - \frac{A_i - \hat{\pi}_i}{\hat{\pi}_i} \hat{m}_i^{(1)} \Big]$$

$$\hat{\tau}_{\mathsf{AIPW},0} = \frac{1}{n} \sum_i \Big[\frac{(1 - A_i) Y_i}{1 - \hat{\pi}_i} + \frac{A_i - \hat{\pi}_i}{1 - \hat{\pi}_i} \hat{m}_i^{(0)} \Big]$$

The Augmented Inverse Probability Weighting

- Recall our nuisance functions:

  $\pi = \mathbb{P}(A = 1|\boldsymbol{X}), \quad m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}], \quad m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$

- The AIPW computes estimates for the above, then uses a plug-in principle.
- Suppose $\hat{\pi}_i, \hat{m}_i^{(1)}, \hat{m}_i^{(0)}$ denote estimates for i-th sample.
- The AIPW is given by $\hat{\tau}_{\text{AIPW}} = \hat{\tau}_{\text{AIPW},1} - \hat{\tau}_{\text{AIPW},0}$, where

$$\hat{\tau}_{\text{AIPW},1} = \frac{1}{n}\sum_i \left[\frac{A_i Y_i}{\hat{\pi}_i} - \frac{A_i - \hat{\pi}_i}{\hat{\pi}_i}\hat{m}_i^{(1)}\right]$$

$$\hat{\tau}_{\text{AIPW},0} = \frac{1}{n}\sum_i \left[\frac{(1 - A_i)Y_i}{1 - \hat{\pi}_i} + \frac{A_i - \hat{\pi}_i}{1 - \hat{\pi}_i}\hat{m}_i^{(0)}\right]$$

$\rightsquigarrow$ Combines best of both worlds.

The Augmented Inverse Probability Weighting

- Recall our nuisance functions:
  $\pi = \mathbb{P}(A = 1|\boldsymbol{X}), \quad m^{(1)} = \mathbb{E}[Y|A = 1, \boldsymbol{X}], \quad m^{(0)} = \mathbb{E}[Y|A = 0, \boldsymbol{X}]$
- The AIPW computes estimates for the above, then uses a plug-in principle.
- Suppose $\hat{\pi}_i, \hat{m}_i^{(1)}, \hat{m}_i^{(0)}$ denote estimates for i-th sample.
- The AIPW is given by $\hat{\tau}_{\text{AIPW}} = \hat{\tau}_{\text{AIPW},1} - \hat{\tau}_{\text{AIPW},0}$, where

$$\hat{\tau}_{\text{AIPW},1} = \frac{1}{n} \sum_i \left[\frac{A_i Y_i}{\hat{\pi}_i} - \frac{A_i - \hat{\pi}_i}{\hat{\pi}_i} \hat{m}_i^{(1)} \right]$$
$$\hat{\tau}_{\text{AIPW},0} = \frac{1}{n} \sum_i \left[\frac{(1 - A_i) Y_i}{1 - \hat{\pi}_i} + \frac{A_i - \hat{\pi}_i}{1 - \hat{\pi}_i} \hat{m}_i^{(0)} \right]$$

$\rightsquigarrow$ Combines best of both worlds.

The double robustness property

Remains consistent, under classical fixed dimensions, large sample asymptotics, even if one of the outcome regression or propensity score models misspecified (Scharfstein, Rotnitzky, Robins '99, Bang and Robins '05)

Extensions to high dimensions

- High-dimensional data increasingly common in practice.
  – Holds promise for alleviating issues with "no unmeasured confounding".
- Extensive recent works in high dimensions: rate double robustness, model double robustness (Belloni, Chernozhukov, Hansen '14; Farrell '15; Bloniarz, Liu, Zhang, Sekhon, Yu '16; Wager, Du, Taylor, Tibshirani '16; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins '17; Athey, Imbens, Wager '18; Bradic, Wager, Zhu '19; Smucler Rotnitzky, Robins '19; Wang and Shah '20; Ning, Sida, Imai '20, Tan '20a, '20b . . .)

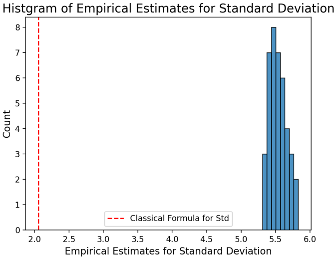
Extensions to high dimensions

- High-dimensional data increasingly common in practice.
  – Holds promise for alleviating issues with "no unmeasured confounding".

- Extensive recent works in high dimensions: rate double robustness, model double robustness (Belloni, Chernozhukov, Hansen '14; Farrell '15; Bloniarz, Liu, Zhang, Sekhon, Yu '16; Wager, Du, Taylor, Tibshirani '16; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins '17; Athey, Imbens, Wager '18; Bradic, Wager, Zhu '19; Smucler Rotnitzky, Robins '19; Wang and Shah '20; Ning, Sida, Imai '20, Tan '20a, '20b . . .)

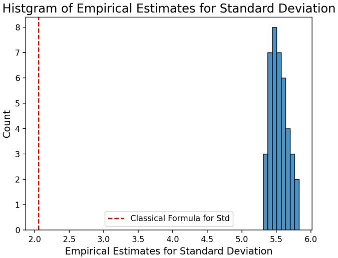
- Typically requires at least one of the propensity score/outcome regression models to be highly sparse.

Issue: Fails to capture certain high-dimensional phenomena



- $p = 700, n = 10000$. Plot examines a version of the AIPW.
- Existing theory fails to capture true variability even in moderate dim.
- Such var. inflation known in high-dim regression context. (Bean, Bickel, El Karoui, Yu '13, El Karoui, Bean, Bickel, Lim, Yu '13, El Karoui '13, Donoho and Montanari '13, Cattaneo, Jansson and Newey, '15, S. and Candes '18) & for causal inference (Yadlowsky '22+)

Issue: Fails to capture certain high-dimensional phenomena



- $p = 700, n = 10000$. Plot examines a version of the AIPW.
- Existing theory fails to capture true variability even in moderate dim.
- Such var. inflation known in high-dim regression context. (Bean, Bickel, El Karoui, Yu '13, El Karoui, Bean, Bickel, Lim, Yu '13, El Karoui '13, Donoho and Montanari '13, Cattaneo, Jansson and Newey, '15, S. and Candes '18) & for causal inference (Yadlowsky '22+)

This talk

Can we analyze a commonly used version of the AIPW estimator in a high-dimensional regime, without assuming any sparsity-type conditions?

Versions of AIPW in high dimensions

- If all nuisances and ATE estimate calculated from full data, estimator is intractable in high dimensions.

Versions of AIPW in high dimensions

- If all nuisances and ATE estimate calculated from full data, estimator is intractable in high dimensions.
- Typical fix: sample split then cross-fit (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins '17; Newey and Robins '18; Smucler Rotnitzky, Robins '19)

Versions of AIPW in high dimensions

- If all nuisances and ATE estimate calculated from full data, estimator is intractable in high dimensions.

- Typical fix: sample split then cross-fit (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins '17; Newey and Robins '18; Smucler Rotnitzky, Robins '19)

  – Version 1: Split sample into 3 parts, calculate PS/OR estimates and ATE estimate from separate folds, switch role of folds, then average.

Versions of AIPW in high dimensions

- If all nuisances and ATE estimate calculated from full data, estimator is intractable in high dimensions.

- Typical fix: sample split then cross-fit (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins '17; Newey and Robins '18; Smucler Rotnitzky, Robins '19)

  – Version 1: Split sample into 3 parts, calculate PS/OR estimates and ATE estimate from separate folds, switch role of folds, then average.

  – Version 2: Split into 2 parts, calculate all nuisances from one part and ATE estimate from other part, switch role of folds then average.

Versions of AIPW in high dimensions

- If all nuisances and ATE estimate calculated from full data, estimator is intractable in high dimensions.

- Typical fix: sample split then cross-fit (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins '17; Newey and Robins '18; Smucler Rotnitzky, Robins '19)

  – Version 1: Split sample into 3 parts, calculate PS/OR estimates and ATE estimate from separate folds, switch role of folds, then average.

  – Version 2: Split into 2 parts, calculate all nuisances from one part and ATE estimate from other part, switch role of folds then average.

- Existing theory: cross-covariances asymptotically negligible at $\sqrt{n}$ scale.

  – Recovers efficiency loss due to sample splitting

  (in regimes studied in previous literature).

Versions of AIPW in high dimensions

- If all nuisances and ATE estimate calculated from full data, estimator is intractable in high dimensions.
- Typical fix: sample split then cross-fit (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins '17; Newey and Robins '18; Smucler Rotnitzky, Robins '19)

  – Version 1: Split sample into 3 parts, calculate PS/OR estimates and ATE estimate from separate folds, switch role of folds, then average. This talk!

  – Version 2: Split into 2 parts, calculate all nuisances from one part and ATE estimate from other part, switch role of folds then average.
- Existing theory: cross-covariances asymptotically negligible at $\sqrt{n}$ scale.

  – Recovers efficiency loss due to sample splitting

  (in regimes studied in previous literature).

The cross-fit 3-split AIPW

- Split the data into three equal parts S_a, S_b, S_c.
- Estimate propensity score from S_a: $\hat{\pi}_i^{S_a}$.
- Estimate outcome regression models from S_b: $\hat{m}_i^{(1),S_b}, \hat{m}_i^{(0),S_b}$.

The cross-fit 3-split AIPW

- Split the data into three equal parts S_a, S_b, S_c.
- Estimate propensity score from S_a: $\hat{\pi}_i^{S_a}$.
- Estimate outcome regression models from S_b: $\hat{m}_i^{(1),S_b}, \hat{m}_i^{(0),S_b}$.
- Use S_c to obtain the final estimator $\hat{\tau}^{S_c}_{\mathsf{AIPW}} = \hat{\tau}^{S_c}_{\mathsf{AIPW},1} - \hat{\tau}^{S_c}_{\mathsf{AIPW},0}$

$$\hat{\tau}^{S_c}_{\mathsf{AIPW},1} = \frac{1}{n/3} \sum_{i \in S_c} \Big[\frac{A_i Y_i}{\hat{\pi}_i^{S_a}} - \frac{A_i - \hat{\pi}_i^{S_a}}{\hat{\pi}_i^{S_a}} \hat{m}_i^{(1),S_b} \Big]$$

$$\hat{\tau}^{S_c}_{\mathsf{AIPW},0} = \frac{1}{n/3} \sum_{i \in S_c} \Big[\frac{(1 - A_i) Y_i}{1 - \hat{\pi}_i^{S_a}} + \frac{A_i - \hat{\pi}_i^{S_a}}{1 - \hat{\pi}_i^{S_a}} \hat{m}_i^{(0),S_b} \Big]$$

The cross-fit 3-split AIPW

- Split the data into three equal parts S_a, S_b, S_c.
- Estimate propensity score from S_a: $\hat{\pi}_i^{S_a}$.
- Estimate outcome regression models from S_b: $\hat{m}_i^{(1),S_b}, \hat{m}_i^{(0),S_b}$.
- Use S_c to obtain the final estimator $\hat{\tau}_{\mathsf{AIPW}}^{S_c} = \hat{\tau}_{\mathsf{AIPW},1}^{S_c} - \hat{\tau}_{\mathsf{AIPW},0}^{S_c}$

$$\hat{\tau}_{\mathsf{AIPW},1}^{S_c} = \frac{1}{n/3} \sum_{i \in S_c} \Big[\frac{A_i Y_i}{\hat{\pi}_i^{S_a}} - \frac{A_i - \hat{\pi}_i^{S_a}}{\hat{\pi}_i^{S_a}} \hat{m}_i^{(1),S_b} \Big]$$

$$\hat{\tau}_{\mathsf{AIPW},0}^{S_c} = \frac{1}{n/3} \sum_{i \in S_c} \Big[\frac{(1 - A_i) Y_i}{1 - \hat{\pi}_i^{S_a}} + \frac{A_i - \hat{\pi}_i^{S_a}}{1 - \hat{\pi}_i^{S_a}} \hat{m}_i^{(0),S_b} \Big]$$

- Switch roles of $S_a, S_b, S_c \rightsquigarrow$ yields $3!$ estimators, average these.
  Call resulting estimator $\hat{\tau}_{\mathsf{cf}}$.

So, what are the main hurdles?

- Characterize asymptotic distribution of estimators pre-cross-fit.
- Challenges: without sparsity assumptions, upto p signals allowed.

So, what are the main hurdles?

- Characterize asymptotic distribution of estimators pre-cross-fit.
- Challenges: without sparsity assumptions, upto p signals allowed.
- If p diverges higher than $o(n)$, already entails a high-dimensional problem.

So, what are the main hurdles?

- Characterize asymptotic distribution of estimators pre-cross-fit.
- Challenges: without sparsity assumptions, upto p signals allowed.
- If p diverges higher than $o(n)$, already entails a high-dimensional problem.
- Track joint distribution between pre-cross-fit estimators.
- Involves tracking the variance & cross-covariances.

So, what are the main hurdles?

- Characterize asymptotic distribution of estimators pre-cross-fit.
- Challenges: without sparsity assumptions, upto p signals allowed.
- If p diverges higher than $o(n)$, already entails a high-dimensional problem.
- Track joint distribution between pre-cross-fit estimators.
- Involves tracking the variance & cross-covariances.
- Need theoretical tools for all of these, that applies for dense as well as sparse signals, in high dimensions.

Our formal setting

- Logistic propensity scores, linear outcome regression models:

$$
\begin{aligned}
&A_i \sim \mathsf{Ber}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta})) \\
&y_i = \alpha^{(A_i)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(A_i)} + \epsilon_i^{(A_i)}, \quad \epsilon_i^{A_i} \sim \mathcal{N}(0, \{\sigma^{(A_i)}\}^2)
\end{aligned}
$$

Our formal setting

- Logistic propensity scores, linear outcome regression models:

$$A_i \sim \mathsf{Ber}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$$
$$y_i = \alpha^{(A_i)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(A_i)} + \epsilon_i^{(A_i)}, \quad \epsilon_i^{A_i} \sim \mathcal{N}(0, \{\sigma^{(A_i)}\}^2)$$

- High-dimensional setting: $p, n \to \infty, p/n \to \kappa > 0$

Our formal setting

- Logistic propensity scores, linear outcome regression models:

$$A_i \sim \mathsf{Ber}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$$
$$y_i = \alpha^{(A_i)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(A_i)} + \epsilon_i^{(A_i)}, \quad \epsilon_i^{A_i} \sim \mathcal{N}(0, \{\sigma^{(A_i)}\}^2)$$

- High-dimensional setting: $p, n \to \infty, p/n \to \kappa > 0$
- Covariate distribution: $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}/n)$

Our formal setting

- Logistic propensity scores, linear outcome regression models:

$$A_i \sim \text{Ber}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$$
$$y_i = \alpha^{(A_i)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(A_i)} + \epsilon_i^{(A_i)}, \quad \epsilon_i^{A_i} \sim \mathcal{N}(0, \{\sigma^{(A_i)}\}^2)$$

- High-dimensional setting: $p, n \to \infty, p/n \to \kappa > 0$
- Covariate distribution: $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}/n)$
- Signal moments:

$$\frac{\|\boldsymbol{\beta}\|^2}{p} \to \gamma^2, \ \frac{\|\boldsymbol{\beta}^{(0)}\|^2}{p} \to \sigma_{0\beta}^2, \ \frac{\|\boldsymbol{\beta}^{(1)}\|^2}{p} \to \sigma_{1\beta}^2,$$
$$\frac{\langle \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)} \rangle}{p} \to \rho_{01}\sigma_{0\beta}\sigma_{1\beta}$$

- Recall ATE: $\tau = \mathbb{E}[Y(1) - Y(0)]$

Our formal setting

- Logistic propensity scores, linear outcome regression models:

$$
\begin{aligned}
A_i &\sim \mathsf{Ber}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta})) \\
y_i &= \alpha^{(A_i)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(A_i)} + \epsilon_i^{(A_i)}, \quad \epsilon_i^{A_i} \sim \mathcal{N}(0, \{\sigma^{(A_i)}\}^2)
\end{aligned}
$$

- High-dimensional setting: $p, n \to \infty, p/n \to \kappa > 0$
- Covariate distribution: $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}/n)$
- Signal moments:

$$
\frac{\|\boldsymbol{\beta}\|^2}{p} \to \gamma^2, \ \frac{\|\boldsymbol{\beta}^{(0)}\|^2}{p} \to \sigma_{0\beta}^2, \ \frac{\|\boldsymbol{\beta}^{(1)}\|^2}{p} \to \sigma_{1\beta}^2,
$$
$$
\frac{\langle \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)} \rangle}{p} \to \rho_{01}\sigma_{0\beta}\sigma_{1\beta}
$$

- Recall ATE: $\tau = \mathbb{E}[Y(1) - Y(0)]$

The proportional scaling regime

- High-dimensional statistics: Johnstone and Lu ('09); Donoho, Maleki, Montanari ('09); Bayati and Montanari ('11); Bean, Bickel, El Karoui, Yu ('13); El Karoui, Bean, Bickel, Lim, Yu ('13); El Karoui ('13); Javanmard and Montanari ('14); Stojnic ('13); Thrampoulidis, Omyak, Hassibi ('15); Dobriban and Wager ('15); Lei et al. ('16); Su, Bogdan, Candés ('17); S., Chen, Candés ('17); Weinstein, Barber, Candés ('17); Thrampoulidis, Abbasi, Hassibi ('18); El Alaoui and Jordan ('18); S. and Candés ('18); Bellec and Zhang ('18); Miolane and Montanari ('18); Bu, Klusowski, Rush, Su ('19); Hastie, Montanari, Rosset, Tibshirani ('19); Zhao, S., Candés ('20); Javanmard, Soltanolkotabi, Hassani ('20); Wang, Weng, Maleki ('20); Celentano, Montanari, Wei ('20); Celentano and Montanari ('21); Feng, Venkataramanan, Rush, Samworth ('21), Patil, Wei, Rinaldo, Tibshirani ('21), Yadlowsky ('22) ...
- Econometrics: Cattaneo, Jansson, Newey '18, Anatolyev '18, Cattaneo, Jansson, Ma '19, Kline et al. '20 ...
- Machine learning: Wang, Mattingly, Lu '17; Mei, Montanari, Nguyen '18; Mei, Misiakiewicz, Montanari '19; Hastie, Montanari, Rosset, Tibshirani '19, Deng, Kammoun, Thrampoulidis '19; Montanari, Ruan, Sohn, Yan '19, Ali, Kolter, Tibshirani '19, Ali, Dobriban, Tibshirani '20, Adlam and Pennington '20, Advani, Saxe, Sompolinsky '20, Liang and S. '20, Liang, Sen, S. '22 ...

The proportional scaling regime

- High-dimensional statistics: Johnstone and Lu ('09); Donoho, Maleki, Montanari ('09); Bayati and Montanari ('11); Bean, Bickel, El Karoui, Yu ('13); El Karoui, Bean, Bickel, Lim, Yu ('13); El Karoui ('13); Javanmard and Montanari ('14); Stojnic ('13); Thrampoulidis, Omyak, Hassibi ('15); Dobriban and Wager ('15); Lei et al. ('16); Su, Bogdan, Candés ('17); S., Chen, Candés ('17); Weinstein, Barber, Candés ('17); Thrampoulidis, Abbasi, Hassibi ('18); El Alaoui and Jordan ('18); S. and Candés ('18); Bellec and Zhang ('18); Miolane and Montanari ('18); Bu, Klusowski, Rush, Su ('19); Hastie, Montanari, Rosset, Tibshirani ('19); Zhao, S., Candés ('20); Javanmard, Soltanolkotabi, Hassani ('20); Wang, Weng, Maleki ('20); Celentano, Montanari, Wei ('20); Celentano and Montanari ('21); Feng, Venkataramanan, Rush, Samworth ('21), Patil, Wei, Rinaldo, Tibshirani ('21), Yadlowsky ('22) ...
- Econometrics: Cattaneo, Jansson, Newey '18, Anatolyev '18, Cattaneo, Jansson, Ma '19, Kline et al. '20 ...
- Machine learning: Wang, Mattingly, Lu '17; Mei, Montanari, Nguyen '18; Mei, Misiakiewicz, Montanari '19; Hastie, Montanari, Rosset, Tibshirani '19, Deng, Kammoun, Thrampoulidis '19; Montanari, Ruan, Sohn, Yan '19, Ali, Kolter, Tibshirani '19, Ali, Dobriban, Tibshirani '20, Adlam and Pennington '20, Advani, Saxe, Sompolinsky '20, Liang and S. '20, Liang, Sen, S. '22 ...

– Much older roots in statistical physics and probability theory!

The proportional scaling regime

- High-dimensional statistics: Johnstone and Lu ('09); Donoho, Maleki, Montanari ('09); Bayati and Montanari ('11); Bean, Bickel, El Karoui, Yu ('13); El Karoui, Bean, Bickel, Lim, Yu ('13); El Karoui ('13); Javanmard and Montanari ('14); Stojnic ('13); Thrampoulidis, Omyak, Hassibi ('15); Dobriban and Wager ('15); Lei et al. ('16); Su, Bogdan, Candés ('17); S., Chen, Candés ('17); Weinstein, Barber, Candés ('17); Thrampoulidis, Abbasi, Hassibi ('18); El Alaoui and Jordan ('18); S. and Candés ('18); Bellec and Zhang ('18); Miolane and Montanari ('18); Bu, Klusowski, Rush, Su ('19); Hastie, Montanari, Rosset, Tibshirani ('19); Zhao, S., Candés ('20); Javanmard, Soltanolkotabi, Hassani ('20); Wang, Weng, Maleki ('20); Celentano, Montanari, Wei ('20); Celentano and Montanari ('21); Feng, Venkataramanan, Rush, Samworth ('21), Patil, Wei, Rinaldo, Tibshirani ('21), Yadlowsky ('22) ...
- Econometrics: Cattaneo, Jansson, Newey '18, Anatolyev '18, Cattaneo, Jansson, Ma '19, Kline et al. '20 ...
- Machine learning: Wang, Mattingly, Lu '17; Mei, Montanari, Nguyen '18; Mei, Misiakiewicz, Montanari '19; Hastie, Montanari, Rosset, Tibshirani '19, Deng, Kammoun, Thrampoulidis '19; Montanari, Ruan, Sohn, Yan '19, Ali, Kolter, Tibshirani '19, Ali, Dobriban, Tibshirani '20, Adlam and Pennington '20, Advani, Saxe, Sompolinsky '20, Liang and S. '20, Liang, Sen, S. '22 ...

– Much older roots in statistical physics and probability theory!

The main result

Recall $\frac{\|\boldsymbol{\beta}\|^2}{p} \to \gamma^2$, $\frac{\|\boldsymbol{\beta}^{(0)}\|^2}{p} \to \sigma_{0\beta}^2$, $\frac{\|\boldsymbol{\beta}^{(1)}\|^2}{p} \to \sigma_{1\beta}^2, \kappa = \lim p/n$, $\frac{\langle \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)} \rangle}{p} \to \rho_{01}\sigma_{0\beta}\sigma_{1\beta}$

Theorem (Jiang, Mukherjee, Sen, S. '22+)

Under convergence of empirical distribution of the signals, suppose either
(i) MLE used for estimating both nuisances (restrict to regime where MLEs exist, whenever using them) or
(ii) MLE used for OR estimation and ridge regularization used for PS estimation with tuning parameter λ, then

$$\sqrt{n}(\hat{\tau}_{cf} - \tau) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_{cf}^2),$$

where $\sigma_{cf}^2 = \left((\sigma^{(0)})^2 + (\sigma^{(1)})^2\right) f(\kappa, \gamma^2, \lambda) + \kappa\left(\sigma_{0\beta}^2 + \sigma_{1\beta}^2 - 2\rho_{01}\sigma_{0\beta}\sigma_{1\beta}\right).$

The main result

Recall $\frac{\|\boldsymbol{\beta}\|^2}{p} \to \gamma^2$, $\frac{\|\boldsymbol{\beta}^{(0)}\|^2}{p} \to \sigma_{0\beta}^2$, $\frac{\|\boldsymbol{\beta}^{(1)}\|^2}{p} \to \sigma_{1\beta}^2, \kappa = \lim p/n$, $\frac{\langle \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)} \rangle}{p} \to \rho_{01}\sigma_{0\beta}\sigma_{1\beta}$

Theorem (Jiang, Mukherjee, Sen, S. '22+)

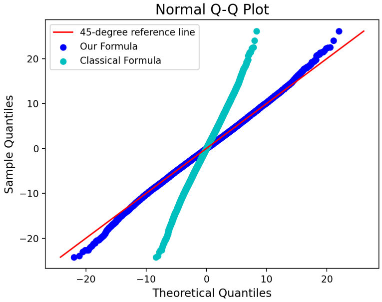
Under convergence of empirical distribution of the signals, suppose either
(i) MLE used for estimating both nuisances (restrict to regime where MLEs exist, whenever using them) or
(ii) MLE used for OR estimation and ridge regularization used for PS estimation with tuning parameter λ, then

$$\sqrt{n}(\hat{\tau}_{cf} - \tau) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_{cf}^2),$$

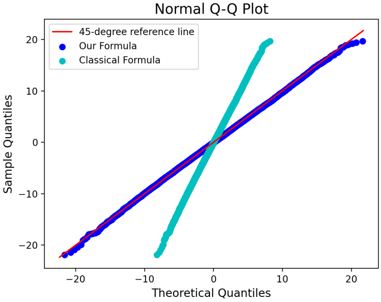
where $\sigma_{cf}^2 = \left((\sigma^{(0)})^2 + (\sigma^{(1)})^2\right) f(\kappa, \gamma^2, \lambda) + \kappa\left(\sigma_{0\beta}^2 + \sigma_{1\beta}^2 - 2\rho_{01}\sigma_{0\beta}\sigma_{1\beta}\right)$.

- σ_{cf}^2 much higher than classical variance or previous ultra-high-dim lit. variance.

Upshot 1: Variance plot - Theory vs empirical

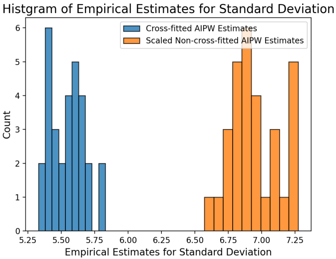


Non-trimmed



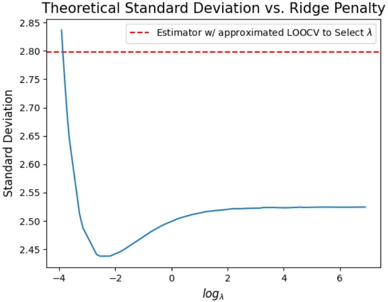
Trimmed at $0.005 \leq \sigma(\cdot) \leq 0.995$

Upshot 2: Cross-fit versus non-cross-fit



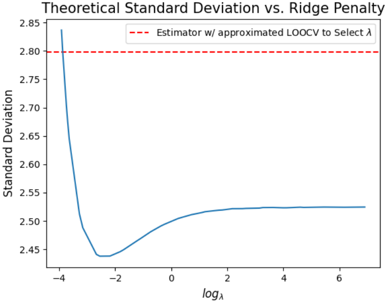
- Cross-covariances are asymptotically non-zero, even at $\sqrt{n}$-scale.
- Stark difference in behavior in our regime.

Upshot 3: Effects of regularization



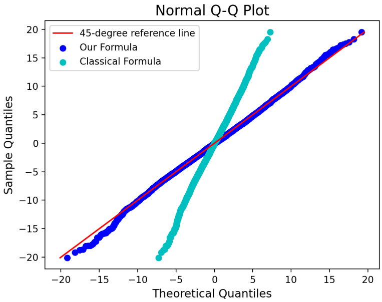
- Minimum variance along the λ-path closer to classical variance.
- (Approximate) LOOCV when optimizing the prediction error (typically done in practice) fails to capture the min-variance λ.

Upshot 3: Effects of regularization



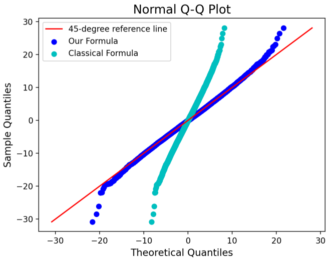
- Minimum variance along the λ-path closer to classical variance.
- (Approximate) LOOCV when optimizing the prediction error (typically done in practice) fails to capture the min-variance λ.
- Need better tuning parameter selection approaches
- Maybe through a good variance estimator?

Robustness to assumptions: Beyond Gaussianity I

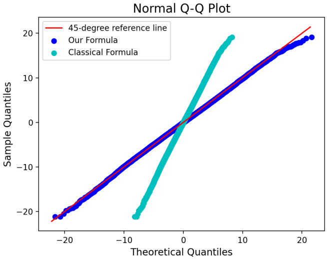


- Covariates i.i.d. Uniform, appropriately scaled.

Robustness to assumptions: Beyond Gaussianity II



Non-trimmed



Trimmed at $0.005 \leq \sigma(\cdot) \leq 0.995$

- Covariates inspired by genetics applications.
- j-th feature takes values $\{0, 1, 2\}$ w.p. $p_j^2, 2p_j(1-p_j), (1-p_j)^2$.
- Appropriately centered and scaled.

The theoretical workhorses

- *Leave-one-out*: Helps decorrelate dependencies in various terms. (Known as cavity method in statistical physics: Mezard, Parisi Virasoro ('87); Statistics ref with linear models: Bean, Bickel, El Karoui, Yu ('13); El Karoui, Bean, Bickel, Lim, Yu ('13); El Karoui ('13); Statistics ref. with GLMs: S., Chen, Candès ('17), S. and Candès ('18); Spectral methods: Chen Chi Fan Ma ('21), ...)

The theoretical workhorses

- *Leave-one-out*: Helps decorrelate dependencies in various terms. (Known as cavity method in statistical physics: Mezard, Parisi Virasoro ('87); Statistics ref with linear models: Bean, Bickel, El Karoui, Yu ('13); El Karoui, Bean, Bickel, Lim, Yu ('13); El Karoui ('13); Statistics ref. with GLMs: S., Chen, Candès ('17), S. and Candès ('18); Spectral methods: Chen Chi Fan Ma ('21), ...)

- *Approximate Message Passing Theory*: Helps track properties of estimators in the logistic model, which don't have closed forms. (Donoho, Maleki, Montanari ('09); Bayati and Montanari ('11); Rangan ('11); Javanmard and Montanari ('14); S. and Candès ('19), Barbier Krzakala, Macris, Miolane, Zdeborova ('19), ...)

The theoretical workhorses

- *Leave-one-out*: Helps decorrelate dependencies in various terms. (Known as cavity method in statistical physics: Mezard, Parisi Virasoro ('87); Statistics ref with linear models: Bean, Bickel, El Karoui, Yu ('13); El Karoui, Bean, Bickel, Lim, Yu ('13); El Karoui ('13); Statistics ref. with GLMs: S., Chen, Candès ('17), S. and Candès ('18); Spectral methods: Chen Chi Fan Ma ('21), . . .)

- *Approximate Message Passing Theory*: Helps track properties of estimators in the logistic model, which don't have closed forms. (Donoho, Maleki, Montanari ('09); Bayati and Montanari ('11); Rangan ('11); Javanmard and Montanari ('14); S. and Candès ('19), Barbier Krzakala, Macris, Miolane, Zdeborova ('19), . . .)

- *Deterministic equivalents*: Helps track quadratic forms of random matrices by connecting to more tractable deterministic matrices. ((Hachem et al. ('07); Couillet et al. ('11); Girko ('12))

Quick peek into Leave-one-out in our setting

- Recall cross-fit AIPW involves $\boldsymbol{X}_i$ from S_a and $\hat{\boldsymbol{\beta}}_{S_a}$. Dependence complicated.

- Note that $\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}$ is independent of $\boldsymbol{X}_i$.

Quick peek into Leave-one-out in our setting

- Recall cross-fit AIPW involves $\boldsymbol{X}_i$ from S_a and $\hat{\boldsymbol{\beta}}_{S_a}$. Dependence complicated.

- Note that $\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}$ is independent of $\boldsymbol{X}_i$.

  – Connect $\hat{\boldsymbol{\beta}}_{S_a}$ to $\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}$ through their 1st order stationary conditions.

$$\Longrightarrow \hat{\boldsymbol{\beta}}_{S_a} = f(\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}, \boldsymbol{X}_i, y_i),$$

  f depends on σ crucially!

Quick peek into Leave-one-out in our setting

- Recall cross-fit AIPW involves $\boldsymbol{X}_i$ from S_a and $\hat{\boldsymbol{\beta}}_{S_a}$. Dependence complicated.

- Note that $\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}$ is independent of $\boldsymbol{X}_i$.
  – Connect $\hat{\boldsymbol{\beta}}_{S_a}$ to $\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}$ through their 1st order stationary conditions.

$$\Longrightarrow \hat{\boldsymbol{\beta}}_{S_a} = f(\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}, \boldsymbol{X}_i, y_i),$$

  f depends on σ crucially!

- Often, need to track $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{S_a}^{(-i)} \sim \mathcal{N}(\mathbf{0}, \|\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}\|^2/n)$, conditional on $\hat{\boldsymbol{\beta}}_{S_a}^{(-i)}$.

Approximate Message Passing

The MLE:

- $\nabla \ell(\hat{\boldsymbol{\beta}}) = \mathbf{0}$
- No closed form
- Hard to track in high dimensions

$\Longrightarrow$

Replace by a tractable surrogate

Approximate Message Passing

The MLE:

- $\nabla \ell(\hat{\boldsymbol{\beta}}) = \mathbf{0}$
- No closed form
- Hard to track in high dimensions

$\Longrightarrow$

Replace by a tractable surrogate

– $\hat{\boldsymbol{\beta}}$ can be any minimizer of a convex loss, not just the MLE.

Approximate Message Passing

The MLE:

- $\nabla \ell(\hat{\boldsymbol{\beta}}) = \mathbf{0}$
- No closed form
- Hard to track in high dimensions

$\Longrightarrow$

Replace by a tractable surrogate

– $\hat{\boldsymbol{\beta}}$ can be any minimizer of a convex loss, not just the MLE.

An algorithmic route

- Introduce a 'suitable' iterative algorithm
- Analyze asymptotic behavior of the iterates $\hat{\boldsymbol{\beta}}^t$
- Establish $\hat{\boldsymbol{\beta}}^t \to \hat{\boldsymbol{\beta}}$ in an appropriate limiting sense

Rich history in statistical physics—Approximate Message Passing—DMM ('09), BM ('11), JM ('13), BLM ('15)

Basic structure for regression problems

Tracks two sets of iterates:
– $\{\hat{\boldsymbol{\beta}}_t\}$, proxy for $\hat{\boldsymbol{\beta}}$
– $\{\boldsymbol{R}_t\}$, proxy for $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or $\boldsymbol{X}\hat{\boldsymbol{\beta}}$

Basic structure for regression problems

Tracks two sets of iterates:
– $\{\hat{\boldsymbol{\beta}}_t\}$, proxy for $\hat{\boldsymbol{\beta}}$
– $\{\boldsymbol{R}_t\}$, proxy for $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or $\boldsymbol{X}\hat{\boldsymbol{\beta}}$

The algorithm

$$\begin{aligned} \hat{\boldsymbol{R}}_t &= \boldsymbol{X}\boldsymbol{f}(\hat{\boldsymbol{\beta}}_t) - b_t \boldsymbol{g}(\boldsymbol{R}_{t-1}) \\ \hat{\boldsymbol{\beta}}_{t+1} &= \boldsymbol{X}^\top \boldsymbol{g}(\boldsymbol{R}_t) - d_t \boldsymbol{f}(\hat{\boldsymbol{\beta}}_{t-1}) \end{aligned}$$

Basic structure for regression problems

Tracks two sets of iterates:
– $\{\hat{\boldsymbol{\beta}}_t\}$, proxy for $\hat{\boldsymbol{\beta}}$
– $\{\boldsymbol{R}_t\}$, proxy for $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or $\boldsymbol{X}\hat{\boldsymbol{\beta}}$

The algorithm

$$\begin{aligned} \hat{\boldsymbol{R}}_t &= \boldsymbol{X}\boldsymbol{f}(\hat{\boldsymbol{\beta}}_t) - b_t \boldsymbol{g}(\boldsymbol{R}_{t-1}) \\ \hat{\boldsymbol{\beta}}_{t+1} &= \boldsymbol{X}^\top \boldsymbol{g}(\boldsymbol{R}_t) - d_t \boldsymbol{f}(\hat{\boldsymbol{\beta}}_{t-1}) \end{aligned}$$

Very special forms: $b_t \approx \mathrm{tr}[\boldsymbol{\nabla}(\boldsymbol{g})], \quad d_t \approx \mathrm{tr}[\boldsymbol{\nabla}(\boldsymbol{f})]$

Basic structure for regression problems

Tracks two sets of iterates:
– $\{\hat{\boldsymbol{\beta}}_t\}$, proxy for $\hat{\boldsymbol{\beta}}$
– $\{\boldsymbol{R}_t\}$, proxy for $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or $\boldsymbol{X}\hat{\boldsymbol{\beta}}$

The algorithm

$$\begin{aligned} \hat{\boldsymbol{R}}_t &= \boldsymbol{X}\boldsymbol{f}(\hat{\boldsymbol{\beta}}_t) - b_t \boldsymbol{g}(\boldsymbol{R}_{t-1}) \\ \hat{\boldsymbol{\beta}}_{t+1} &= \boldsymbol{X}^\top \boldsymbol{g}(\boldsymbol{R}_t) - d_t \boldsymbol{f}(\hat{\boldsymbol{\beta}}_{t-1}) \end{aligned}$$

Very special forms: $b_t \approx \mathrm{tr}[\boldsymbol{\nabla}(\boldsymbol{g})], \quad d_t \approx \mathrm{tr}[\boldsymbol{\nabla}(\boldsymbol{f})]$

- Roughly iterates score equation for estimator of choice.

Basic structure for regression problems

Tracks two sets of iterates:
– $\{\hat{\boldsymbol{\beta}}_t\}$, proxy for $\hat{\boldsymbol{\beta}}$
– $\{\boldsymbol{R}_t\}$, proxy for $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or $\boldsymbol{X}\hat{\boldsymbol{\beta}}$

The algorithm

$$\begin{aligned}\hat{\boldsymbol{R}}_t &= \boldsymbol{X}\boldsymbol{f}(\hat{\boldsymbol{\beta}}_t) - b_t\boldsymbol{g}(\boldsymbol{R}_{t-1}) \\ \hat{\boldsymbol{\beta}}_{t+1} &= \boldsymbol{X}^\top\boldsymbol{g}(\boldsymbol{R}_t) - d_t\boldsymbol{f}(\hat{\boldsymbol{\beta}}_{t-1})\end{aligned}$$

Very special forms: $b_t \approx \mathrm{tr}[\boldsymbol{\nabla}(\boldsymbol{g})], \quad d_t \approx \mathrm{tr}[\boldsymbol{\nabla}(\boldsymbol{f})]$

- Roughly iterates score equation for estimator of choice.
- Known as Onsager correction term

  –Tracks dependence between iterations. Fundamentally important quantity!

High-level idea

The MLE:

- $\nabla \ell(\hat{\boldsymbol{\beta}}) = \mathbf{0}$
- No closed form
- Hard to track in high dimensions

$\Rightarrow$

Replace by a tractable surrogate

An algorithmic route

- Introduce a 'suitable' iterative algorithm
- Analyze asymptotic behavior of the iterates $\hat{\boldsymbol{\beta}}^t$
- Establish $\hat{\boldsymbol{\beta}}^t \to \hat{\boldsymbol{\beta}}$ in an appropriate limiting sense

Rich history in statistical physics—Approximate Message Passing (AMP) algorithms—DMM ('09), BM ('11), JM ('13), BLM ('15)

Tracking algorithm iterates

Tracking algorithm iterates

State evolution formalism

Under moment conditions, for any pseudo-Lipschitz function ψ, for any t.

$$\lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\hat{\beta}_{t,j}, \beta_j) = \mathbb{E}[\psi(\sigma_t Z, \tilde{\beta})],$$

where $Z \sim \mathcal{N}(0,1)$, $\tilde{\beta}$ drawn from limiting empirical distribution of the regression vectors; $Z, \tilde{\beta}$ independent. Asymptotic variance σ_t can be precisely characterized.

Tracking algorithm iterates

State evolution formalism

Under moment conditions, for any pseudo-Lipschitz function ψ, for any t.

$$\lim_{n\to\infty} \frac{1}{p}\sum_{j=1}^{p} \psi(\hat{\beta}_{t,j}, \beta_j) = \mathbb{E}[\psi(\sigma_t Z, \tilde{\beta})],$$

where $Z \sim \mathcal{N}(0,1)$, $\tilde{\beta}$ drawn from limiting empirical distribution of the regression vectors; $Z, \tilde{\beta}$ independent. Asymptotic variance σ_t can be precisely characterized.

- Upshot: Can characterize any separable function of $\hat{\boldsymbol{\beta}}_t$,
  - Important examples: $\|\hat{\boldsymbol{\beta}}_t\|^2/p,\ \|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|^2/p$.

Tracking algorithm iterates

State evolution formalism

Under moment conditions, for any pseudo-Lipschitz function ψ, for any t.

$$\lim_{n\to\infty} \frac{1}{p}\sum_{j=1}^{p} \psi(\hat{\beta}_{t,j}, \beta_j) = \mathbb{E}[\psi(\sigma_t Z, \tilde{\beta})],$$

where $Z \sim \mathcal{N}(0,1)$, $\tilde{\beta}$ drawn from limiting empirical distribution of the regression vectors; $Z, \tilde{\beta}$ independent. Asymptotic variance σ_t can be precisely characterized.

- Upshot: Can characterize any separable function of $\hat{\boldsymbol{\beta}}_t$,
  – Important examples: $\|\hat{\boldsymbol{\beta}}_t\|^2/p$, $\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|^2/p$.
- Caveat: Crucially uses the form of the AMP algorithm.

High-level idea

The MLE:

- $\nabla \ell(\hat{\boldsymbol{\beta}}) = \mathbf{0}$
- No closed form
- Hard to track in high dimensions

$\Longrightarrow$

Replace by a tractable surrogate

An algorithmic route

- Introduce a 'suitable' iterative algorithm
- Analyze asymptotic behavior of the iterates $\hat{\boldsymbol{\beta}}^t$
- Establish $\hat{\boldsymbol{\beta}}^t \to \hat{\boldsymbol{\beta}}$ in an appropriate limiting sense

Rich history in statistical physics—Approximate Message Passing (AMP) algorithms—DMM ('09), BM ('11), JM ('13), BLM ('15)

The final convergence step

- Recall $\hat{\boldsymbol{\beta}}$ was our original estimator of interest.
- Final step: Construct η s.t. $\eta(\hat{\boldsymbol{\beta}}_t) \approx \hat{\boldsymbol{\beta}}$ under appropriate limits.

The final convergence step

- Recall $\hat{\boldsymbol{\beta}}$ was our original estimator of interest.
- Final step: Construct η s.t. $\eta(\hat{\boldsymbol{\beta}}_t) \approx \hat{\boldsymbol{\beta}}$ under appropriate limits.

Formally, show that

$$\lim_{n\to\infty} \sum_{j=1}^{p} \psi(\hat{\beta}_j, \beta_j) = \lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\eta(\hat{\beta}_{t,j}), \beta_j)$$

,

The final convergence step

– Recall $\hat{\boldsymbol{\beta}}$ was our original estimator of interest.

– Final step: Construct η s.t. $\eta(\hat{\boldsymbol{\beta}}_t) \approx \hat{\boldsymbol{\beta}}$ under appropriate limits.

Formally, show that

$$\begin{aligned}\lim_{n\to\infty} \sum_{j=1}^{p} \psi(\hat{\beta}_j, \beta_j) &= \lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\eta(\hat{\beta}_{t,j}), \beta_j) \\ &= \lim_{t\to\infty} \mathbb{E}[\psi(\eta(\sigma_t Z), \tilde{\beta})],\end{aligned}$$

The final convergence step

– Recall $\hat{\boldsymbol{\beta}}$ was our original estimator of interest.

– Final step: Construct η s.t. $\eta(\hat{\boldsymbol{\beta}}_t) \approx \hat{\boldsymbol{\beta}}$ under appropriate limits.

Formally, show that

$$\begin{aligned}\lim_{n\to\infty} \sum_{j=1}^{p} \psi(\hat{\beta}_j, \beta_j) &= \lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\eta(\hat{\beta}_{t,j}), \beta_j) \\ &= \lim_{t\to\infty} \mathbb{E}[\psi(\eta(\sigma_t Z), \tilde{\beta})],\end{aligned}$$

- The algorithm must be constructed so that $\eta(\cdot)$ can be found.

The final convergence step

– Recall $\hat{\boldsymbol{\beta}}$ was our original estimator of interest.

– Final step: Construct η s.t. $\eta(\hat{\boldsymbol{\beta}}_t) \approx \hat{\boldsymbol{\beta}}$ under appropriate limits.

Formally, show that

$$\lim_{n\to\infty} \sum_{j=1}^{p} \psi(\hat{\beta}_j, \beta_j) = \lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\eta(\hat{\beta}_{t,j}), \beta_j)$$
$$= \lim_{t\to\infty} \mathbb{E}[\psi(\eta(\sigma_t Z), \tilde{\beta})],$$

- The algorithm must be constructed so that $\eta(\cdot)$ can be found.
- All info. about the AMP iterates can be transferred to our estimator $\hat{\boldsymbol{\beta}}$!

Caveats

- Need an algorithm in AMP form
  with fixed points satisfying same KKT conditions as estimator of interest.

Caveats

- Need an algorithm in AMP form
  with fixed points satisfying same KKT conditions as estimator of interest.

- Construction highly non-trivial, case-specific.

Caveats

- Need an algorithm in AMP form
  with fixed points satisfying same KKT conditions as estimator of interest.

- Construction highly non-trivial, case-specific.

- The final convergence step also problem-specific.
  – Relies on properties of loss the estimator minimizes.

Causal inference uncovers novel challenges

- Our cross-fit 3-split AIPW involves (among others) *all of*

$$\{\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{S_k}\}_{i \in S_a \text{ or } S_b \text{ or } S_c;\ k \text{ equals one of the others}}.$$

Causal inference uncovers novel challenges

- Our cross-fit 3-split AIPW involves (among others) *all of*

$$\{\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{S_k}\}_{i \in S_a \text{ or } S_b \text{ or } S_c;\ k \text{ equals one of the others}}.$$

- Leave-one-out helps decorrelate pairwise dependencies.

Causal inference uncovers novel challenges

- Our cross-fit 3-split AIPW involves (among others) *all of*

$$\{\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{S_k}\}_{i \in S_a \text{ or } S_b \text{ or } S_c;\ k \text{ equals one of the others}}.$$

- Leave-one-out helps decorrelate pairwise dependencies.
- For our CLT, need track many of these dependencies simultaneously.

Causal inference uncovers novel challenges

- Our cross-fit 3-split AIPW involves (among others) *all of*

$$\{\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{S_k}\}_{i \in S_a \text{ or } S_b \text{ or } S_c;\ k \text{ equals one of the others}}.$$

- Leave-one-out helps decorrelate pairwise dependencies.
- For our CLT, need track many of these dependencies simultaneously.
- The error terms cumulate!

Causal inference uncovers novel challenges

- Our cross-fit 3-split AIPW involves (among others) *all of*

$$\{\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{S_k}\}_{i \in S_a \text{ or } S_b \text{ or } S_c;\ k \text{ equals one of the others}}.$$

- Leave-one-out helps decorrelate pairwise dependencies.
- For our CLT, need track many of these dependencies simultaneously.
- The error terms cumulate!
- Tracking these requires very careful analysis.

Interactions with machine learning

Recall $\frac{\|\boldsymbol{\beta}\|^2}{p} \to \gamma^2$, $\frac{\|\boldsymbol{\beta}^{(0)}\|^2}{p} \to \sigma_{0\beta}^2$, $\frac{\|\boldsymbol{\beta}^{(1)}\|^2}{p} \to \sigma_{1\beta}^2, \kappa = \lim p/n$, $\frac{\langle \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)} \rangle}{p} \to \rho_{01}\sigma_{0\beta}\sigma_{1\beta}$

Theorem (Jiang, Mukherjee, Sen, S. '22+)

Under convergence of empirical distribution of the signals, suppose either (i) MLE used for estimating both nuisances (restrict to regime where MLEs exist, whenever using them) or (ii) MLE used for OR estimation and ridge regularization used for PS estimation with tuning parameter λ, then

$$\sqrt{n}(\hat{\tau}_{cf} - \tau) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_{cf}^2),$$

where $\sigma_{cf}^2 = \left((\sigma^{(0)})^2 + (\sigma^{(1)})^2\right) f(\kappa, \gamma^2, \lambda) + \kappa \left(\sigma_{0\beta}^2 + \sigma_{1\beta}^2 - 2\rho_{01}\sigma_{0\beta}\sigma_{1\beta}\right).$

Interactions with machine learning

Recall $\frac{\|\boldsymbol{\beta}\|^2}{p} \to \gamma^2$, $\frac{\|\boldsymbol{\beta}^{(0)}\|^2}{p} \to \sigma_{0\beta}^2$, $\frac{\|\boldsymbol{\beta}^{(1)}\|^2}{p} \to \sigma_{1\beta}^2, \kappa = \lim p/n$, $\frac{\langle \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)} \rangle}{p} \to \rho_{01}\sigma_{0\beta}\sigma_{1\beta}$

Theorem (Jiang, Mukherjee, Sen, S. '22+)

Under convergence of empirical distribution of the signals, suppose either
(i) MLE used for estimating both nuisances (restrict to regime where MLEs exist, whenever using them) or
(ii) MLE used for OR estimation and ridge regularization used for PS estimation with tuning parameter λ, then

$$\sqrt{n}(\hat{\tau}_{cf} - \tau) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_{cf}^2),$$

where $\sigma_{cf}^2 = \left((\sigma^{(0)})^2 + (\sigma^{(1)})^2\right) f(\kappa, \gamma^2, \lambda) + \kappa \left(\sigma_{0\beta}^2 + \sigma_{1\beta}^2 - 2\rho_{01}\sigma_{0\beta}\sigma_{1\beta}\right).$

- Double/Debiased Machine Learning (Chernozhukov et al. '16): Modern ML methods regularly used for nuisance estimation
- Can we develop analogues here?

Wrapping Up

Summary:

–CLT in high-dimensional setting for cross-fit AIPW estimator

–Without sparsity assumptions

–Quantification of variance inflation

–Non-trivial cross-fit covariances, a new high-dimensional phenomena

Wrapping Up

Summary:

–CLT in high-dimensional setting for cross-fit AIPW estimator

–Without sparsity assumptions

–Quantification of variance inflation

–Non-trivial cross-fit covariances, a new high-dimensional phenomena

Next Steps?

- Towards inference (exploring)
- Formalizing theory beyond covariate distribution assumptions
- More general nuisance estimation
- Analyze other estimators

$\vdots$

$\vdots$

Thank you!

Thanks to NSF DMS and the William F. Milton Fund Award

Contact: *pragya@fas.harvard.edu*