Bayesian Double Machine Learning for Causal Inference

Francis J. DiTraglia[1] Laura Liu[2]

[1]University of Oxford

[2]University of Pittsburgh

February 26th, 2025

My Research Interests

Econometrics

Causal Inference, Spillovers, Measurement Error, Model Selection, Bayesian Inference

Applied Work

Childhood Lead Exposure, Pawn Lending in Mexico City, Colombian Civil Conflict

Overview of Today's Talk

- Causal inference is hard, especially when there are many controls.
- Bayesian approach is appealing, but doesn't work out-of-the-box
- Find a way to combine the advantages of Bayes with good Frequentist properties (bias / variance / coverage probability)
- Related to Frequentist literature on "Double Machine Learning" but aims to improve on finite-sample performance.
- Workshop on Bayesian Causal Inference this Friday: email me for a link!

The Problem / Model

$$Y_i = \alpha D_i + X_i'\beta + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | D_i, X_i] = 0, \quad i = 1, \ldots, n$$

- Learn effect α of treatment D_i (not necessarily binary)
- Selection-on-observables: p-vector of controls X_i
- OLS: unbiased and consistent estimator of α, but noisy if p is large
- Drop control $X_i^{(j)}$ that is correlated with $D_i \Rightarrow$ biased estimate of α if $\beta^{(j)} \neq 0$.

Naïve Shrinkage Estimator: Ridge Regression

Assume everything de-meaned, X scale-normalized

Unique, closed-form solution even if $p > n$

$$\begin{bmatrix} \widehat{\alpha}_{\text{naive}} \\ \widehat{\beta}_{\text{naive}} \end{bmatrix} = \left[\begin{pmatrix} D'D & D'X \\ X'D & X'X \end{pmatrix} + \begin{pmatrix} 0 & 0_p' \\ 0_p & \lambda \mathbb{I}_p \end{pmatrix} \right]^{-1} \begin{pmatrix} D'Y \\ X'Y \end{pmatrix}, \quad \lambda \equiv \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}.$$

Frequentist Interpretation

Minimize $(Y - \alpha D - X\beta)'(Y - \alpha D - X\beta) + \lambda \beta' \beta$

Bayesian Interpretation

Posterior mean: σ_ε known, flat prior on α, independent Normal$(0, \sigma_\beta^2)$ priors on β_j

Regularization-Induced Confounding (RIC)

Term coined by Hahn et al. (2018)

If $\lambda > 0$, bias from correlation between D and residuals:

$$\text{Bias}(\widehat{\alpha}_{\text{naive}}) = \widehat{\omega}' \left[\mathbb{I}_p - (R + \lambda \mathbb{I}_p)^{-1} R\right] \beta$$

$$\text{Var}(\widehat{\alpha}_{\text{naive}}) = \sigma_\varepsilon^2 \left[(D'D)^{-1} + \widehat{\omega}'(R + \lambda \mathbb{I}_p)^{-1} R (R + \lambda \mathbb{I}_p)^{-1} \widehat{\omega}\right]$$

Notation

$$\widehat{\omega}_j = (D'D)^{-1} D' X_j, \quad \widehat{E}_j = X_j - \widehat{\omega}_j X_j, \quad R = \widehat{E}'\widehat{E}$$

Problem

For $\lambda > 0$, bias depends crucially on $\widehat{\omega}$ and β; strong confounding $\Rightarrow$ large bias

Adding a First-Stage

Just a Projection

$$Y = \alpha D + X'\beta + \varepsilon, \quad \mathbb{E}[\varepsilon|X, D] = 0$$
$$D = X'\gamma + V, \quad \mathbb{E}[V|X] = 0$$

Implied by Casual Assumption

$$\text{Cov}(\varepsilon, V) = \text{Cov}(\varepsilon, D - X'\gamma) = \text{Cov}(\varepsilon, D) - \text{Cov}(\varepsilon, X')\gamma = 0.$$

Idea

Maybe adding this regression allows us to learn the degree of counfounding.

Adding the D on X regression has no effect!

"Bayes Ignorability" – Linero (2023; JASA)

Bayes' Theorem

$$\pi(\theta|Y, D, X) \propto f(Y, D|X, \theta) \times \pi(\theta)$$

$\text{Cov}(\varepsilon, V) = 0 \Rightarrow$ no common parameters!

$$f(Y, D|X, \theta) = f(Y|D, X, \theta)f(D|X, \theta) = f(Y|D, X, \alpha, \beta, \sigma_\varepsilon^2) \times f(D|X, \gamma, \sigma_V^2)$$

Problem

Unless prior treats β and γ as dependent, adding the D on X regression has no effect!

Our Solution: Bayesian Double Machine Learning (BDML)

From Structural to Reduced Form

$$Y_i = \alpha D_i + X_i'\beta + \varepsilon_i = X_i'(\alpha\gamma + \beta) + (\varepsilon_i + \alpha V_i) = X_i'\delta + U_i$$

$$\begin{aligned} Y_i &= X_i'\delta + U_i \\ D_i &= X_i'\gamma + V_i \end{aligned} \qquad \begin{bmatrix} U_i \\ V_i \end{bmatrix} \Bigg| X_i \sim \text{Normal}_2(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 + \alpha^2\sigma_V^2 & \alpha\sigma_V^2 \\ \alpha\sigma_V^2 & \sigma_V^2 \end{bmatrix}$$

BDML Algorithm

1. Place "standard" priors on reduced form parameters (δ, γ, Σ)
2. Draw from posterior $(\delta, \gamma, \Sigma)|(X, D, Y)$
3. Posterior draws for $\Sigma \implies$ posterior draws for $\alpha = \sigma_{UV}/\sigma_V^2$

BDML versus Frequentist Double Machine Learning (FDML)

e.g. Chernozhukov et al. (2018; Econometrics J.)

FDML Optimizes

Plug in "Machine Learning" estimators of reduced form parameters: $(\widehat{\delta}_{\text{ML}}, \widehat{\gamma}_{\text{ML}})$

$$\widehat{\alpha}_{\text{FDML}} = \frac{\sum_{i=1}^{n}(Y_i - X_i'\widehat{\delta}_{\text{ML}})(D_i - X_i'\widehat{\gamma}_{\text{ML}})}{\sum_{i=1}^{n}(D_i - X_i'\widehat{\gamma}_{\text{ML}})^2}.$$

BDML Marginalizes

Posterior for α averages over posterior uncertainty about γ and δ

Theoretical Results

$$Y_i = X_i'\delta + U_i \qquad D_i = X_i'\gamma + V_i$$

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \Bigg| X_i \sim \text{Normal}_2(0, \Sigma)$$

$$\pi(\Sigma, \delta, \gamma) \propto \pi(\Sigma)\pi(\delta)\pi(\gamma)$$

$$\Sigma \sim \text{Inverse-Wishart}(\nu_0, \Sigma_0)$$

$$\delta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\delta)$$

$$\gamma \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\gamma)$$

Naïve Approach

Analogous but with single structural equation and $\beta \sim \text{Normal}(0, \mathbb{I}_p/\tau_\beta)$

Asymptotic Framework

Fixed true parameters $(\Sigma^*, \delta^*, \gamma^*)$; $n \to \infty$ (large sample); $p \to \infty$ (many controls)

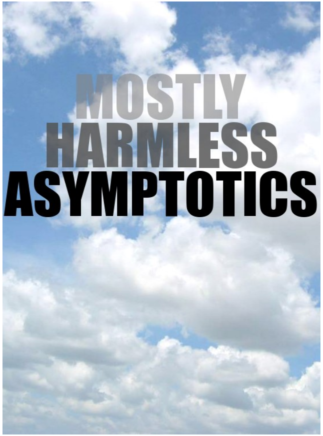
Our asymptotic framework ensures bounded R-squared.

Rate Restrictions

(i) sample size dominates # of controls: $p/n \to 0$

(ii) sample size dominate prior precisions: $\tau/n \to 0$

(iii) precisions of same order as # controls: $\tau \asymp p$



Regularity Conditions

(i) $p < n$

(ii) $\text{Var}(X) \equiv \Sigma_X$ "well-behaved" as $p \to \infty$

(iii) $\lim_{p\to\infty} \sum_{j=1}^{p} (\delta_j^*)^2 < \infty, \quad \lim_{p\to\infty} \sum_{j=1}^{p} (\gamma_j^*)^2 < \infty$

(iv) iid errors/controls, $\mathbb{E}(X_i) = 0$, finite & p.d. Σ^*

Selection Bias in the Limit

When p and n are large, what are our implied beliefs about selection bias?

$$\text{SB} \equiv [\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0)] - \alpha = [\mathbb{E}(X_i|D_i = 1) - \mathbb{E}(X_i|D_i = 0)]'\beta$$

Naïve Model

Degenerate prior centered at zero: $\text{SB} = \dfrac{\gamma'\Sigma_X\beta}{\sigma_V^2 + \gamma'\Sigma_X\gamma} \to_p 0$

BDML

Non-degenerate prior centered at zero: $\text{SB} \to_p \dfrac{\sigma_{UV}}{\sigma_V^2 + \gamma'\Sigma_X\gamma}$

Summary of Asymptotic Results

Consistency

Naïve, BDML and FDML all provide consistent estimators of α.

Asymptotic Bias

BDML and FDML have bias of order p^2/n^2 compared to p/n for Naïve.

$\sqrt{n}$-Consistency

Naïve requires $p/\sqrt{n} \to 0$; BDML and FDML require only $p/n^{3/4} \to 0$.

Why do we focus on variance?

Bias dominates: if $p/\sqrt{n} \to 0$, all three have the same AVAR.

Simulation Experiment

$$Y_i = \alpha D_i + X_i'\beta + \varepsilon_i$$
$$D_i = X_i'\gamma + V_i$$

$$\{X_i\}_{i=1}^n \sim \text{iid Normal}_p(0, \mathbb{I}_p)$$
$$\{(\varepsilon_i, V_i)\}_{i=1}^n \mid X \sim \text{iid Normal}_2\left(0, \text{diag}\left\{\sigma_\varepsilon^2, 1\right\}\right)$$
$$\beta \mid (X, \varepsilon, V) \sim \text{Normal}_p\left(\mu_\beta, \sigma_\beta^2 \mathbb{I}\right).$$

Linero's (2023) "Fixed" Design

$$\alpha = 2, \quad \gamma = \iota_p/\sqrt{p}, \quad \mu_\beta = -\gamma/2, \quad \sigma_\beta^2 = 1/p, \quad n = 200, \quad p = 100$$

Two Versions of BDML

Both Versions

LKJ(4) Prior on Corr(U, V); Independent Cauchy$(0, 2.5)$ priors on SD(U) and SD(V)

Basic Version

Independent Normal$(0, 5^2)$ priors on the elements of δ and γ.

Hierarchical Version

- Independent Normal$(0, \sigma_\delta^2)$ priors on the elements of δ
- Independent Normal$(0, \sigma_\gamma^2)$ priors on the elements of γ
- Independent Inverse-Gamma$(2, 2)$ priors on $\sigma_\delta, \sigma_\gamma$.

Two-Step "Plug-in" Bayesian Approaches

Preliminary Regression

$\widehat{D}_i \equiv X_i' \widehat{\gamma}_{\text{prelim}} \leftarrow$ estimate from Bayesian regression of D on X.

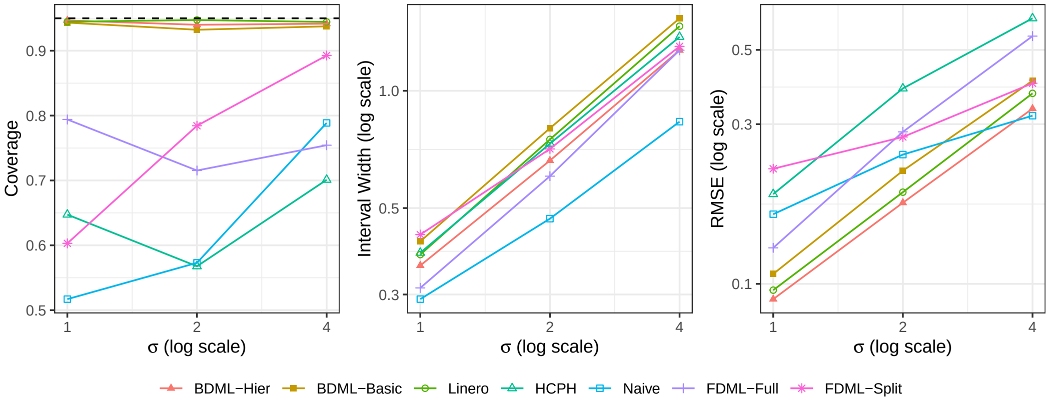
HCPH (Hahn et al, 2018; Bayesian Analysis)

1. Bayesian linear regression of Y on $(D - \widehat{D})$ and X
2. Estimation / inference for α from posterior for $(D - \widehat{D})$ coefficient.

Linero (2023; JASA)

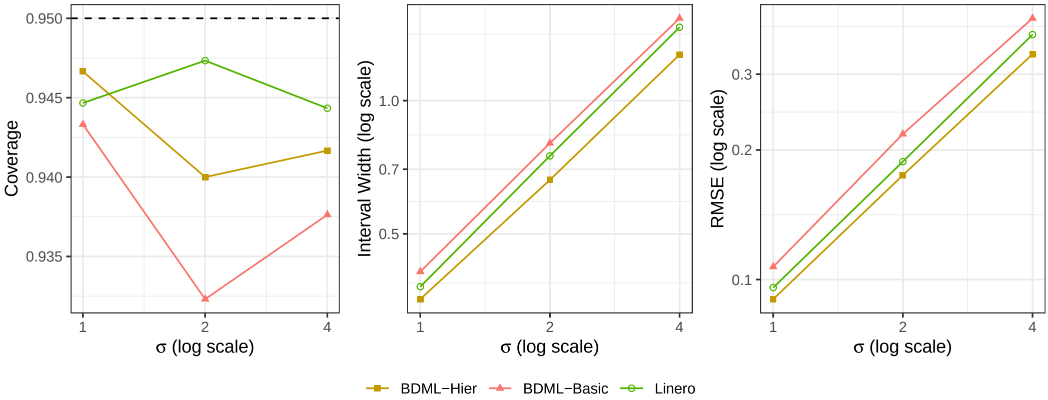
1. Bayesian linear regression of Y on $(D, \widehat{D}, X)$.
2. Estimation / inference for α from posterior the D coefficient.

Simulation Results – 3000 Replications



Only BDML and Linero have correct coverage (Left); Also lower RMSE (Right)

Zooming In: BDML versus Linero



Coverage of Linero & BDML-Hier comparable; BDML-Hier: shortest intervals & lowest RMSE

Thanks for listening!

Summary

- Simple, fully-Bayesian causal inference in a workhorse linear model with many controls.
- Avoids RIC; Excellent Frequentist Properties

In Progress

- More Simulations; Empirical Examples
- Good “default” prior choices?
- Extensions: partially linear model; treatment interactions; instrumental variables?