Bayesian Double Machine Learning for Causal Inference

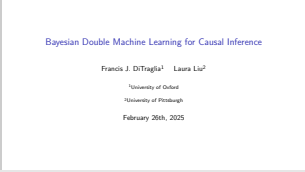
Francis J. DiTraglia[1] Laura Liu[2]

[1]University of Oxford

[2]University of Pittsburgh

February 26th, 2025

Bayesian DML



1. Thanks for inviting me: great to be back in York, one of my favorite cities in the UK
2. Joint work with Laura Liu from University of Pittsburgh
3. New project: first time I'm presenting it.
4. Grateful for your comments and your forbearance!

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

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

1. Start with brief overview of some of my research interests
2. Applied econometrician: work on a mix of theory and application.
3. I do a lot of work on causal inference, particularly with instrumental variables.
4. This includes some work looking at spillover effects, and also measurement error.
5. I also do applied work, mainly in empirical micro, and usually with some meaty econometrics to sink my teeth into.
6. I have an ongoing research agenda on childhood lead exposure; getting ready to pilot an at-home screening test in Leeds; designing sampling plan.
7. Have a working paper on Pawn Lending in Mexico City, and another project I'm hoping to resurrect this summer on the Colombian Civil Conflict of the 1990s and 2000s

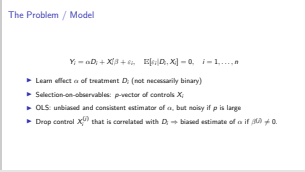
Overview of Today's Talk

- 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 $\beta^{(j)} \neq 0$.



1. Start by introducing the problem: I'll stick with this simple model throughout the talk, although we're working on some extensions
2. Countless examples from empirical micro have this structure.
3. Problem: may need to control for a large number of covariates to make selection-on-observables plausible. And if p is large relative to n, OLS can be extremely noisy. If $p > n$ doesn't even exist!
4. But if we drop controls and they turn out to be correlated with D and predictive of Y, this biases our estimate of α.
5. I've posed this as a classic bias-variance trade-off, so perhaps you're thinking: machine learning to the rescue!

Naïve Shrinkage Estimator: Ridge Regression

Assume everything de-meaned, X scale-normalized

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

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

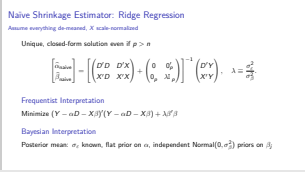
Frequentist Interpretation

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

Naïve Shrinkage Estimator: Ridge Regression



1. Ridge Regression: "shrink" β towards zero; closed-form expression even if $p > n$.
2. λ controls shrinkage: $\lambda = 0$ is OLS with all controls; $\lambda \to \infty$ is OLS of Y on D *only*.
3. Frequentist interpretation: add "penalty" term to least squares objective. Makes it costly to move a coefficient away from zero so we don't "overreact" to small changes in the data.
4. Bayesian interpretation: place a "shrinkage prior" on β (note scale normalization) where σ^2_β represents beliefs about how large the coefficients are expected to be.
5. Link between the two interpretations: $\lambda = \sigma^2_\varepsilon / \sigma^2_\beta$
6. Notice: no shrinkage for α here. This is the effect we're interested in. Bayes: flat prior.

Regularization-Induced Confounding (RIC)

Term coined by Hahn et al. (2018)

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

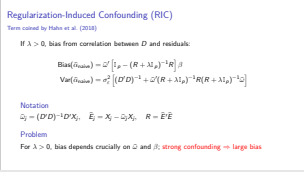
$$\text{Bias}(\widehat{\alpha}_{\text{naive}}) = \widehat{\omega}' \left[\mathbb{I}_p - (R + \lambda \mathbb{I}_p)^{-1} R \right] \beta$$
$$\text{Var}(\widehat{\alpha}_{\text{naive}}) = \sigma_\varepsilon^2 \left[(D'D)^{-1} + \widehat{\omega}' (R + \lambda \mathbb{I}_p)^{-1} R (R + \lambda \mathbb{I}_p)^{-1}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

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



1. So either a Frequentist or a Bayesian might come up with this Ridge approach to estimating α when p is large. Unfortunately it turns out to be a bad idea. From now on: focus on Bayesian interpretation; discuss Frequentist properties.
2. "Regularizing" OLS estimator by setting $\lambda > 0$ causes residuals to violate sample analogue of $\mathbb{E}(D_i \varepsilon_i | X_i) = 0$, the popn moment condition needed to give α a causal interpretation.
3. Explain the Bias and Variance expressions
4. Point about "shrinkage" prior rather than genuine prior beliefs; failing to account for beliefs about confounding; challenging in high dimensions!
5. Also point about how we implicitly take account of the magnitude of β on the Bayesian view of things. But haven't explicitly accounted for ω.

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 a First-Stage

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

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

Idea

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

1. RIC suggests we need to take account of the relationship between D and X so let's do it: add another regression equation to capture "propensity score" / "first-stage"
2. First stage here is "just a projection" and linear for simplicity. Can handle generic non-parametric first stage and partially linear outcome equation.
3. Structural assumption (selection-on-observables) implies that the errors ε and V are uncorrelated
4. Talk through the little derivation

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!

$$f(Y, D|X, \theta) = f(Y|D, X, \theta) f(D|X, \theta) = f(Y|D, X, \alpha, \beta, \sigma^2_\varepsilon) \times f(D|X, \gamma, \sigma^2_V)$$

Problem

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

Adding the D on X regression has no effect!

Adding the D on X regression has no effect!
"Bayes Ignorability" – Linero (2023, JASA)

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

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!

1. Simple normal linear regression model, condition on controls X but now we'll include both the outcome equation *and* the first-stage regression of D on X in our model.
2. Bayes theorem: posterior is proportional to prior times likelihood.
3. But the likelihood *factorizes*. Remember that V and ε are uncorrelated.
4. And the "usual" priors are independent across the two equations.
5. This means we can "integrate out" the D on X regression: no effect on posterior for α!
6. Maybe you're thinking: no problem, let's just add dependence! But remember: β and γ are high-dimensional vectors corresponding to a long list of controls. Challenging to elicit informative prior including dependence between them. (p is big so p^2 is really big!)

Our Solution: Bayesian Double Machine Learning (BDML)

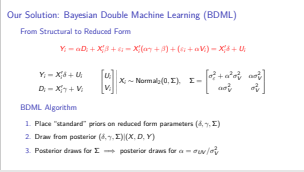
From Structural to Reduced Form

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

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

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 $\alpha = \sigma_{UV}/\sigma_V^2$



1. Our idea: ignore D and work with the reduced form! Substitute the "first-stage" regression of D on X into the structural equation to obtain the following.
2. Remember: ε_i and V_i are uncorrelated. But since U includes both ε and V it *is correlated* with V, in particular: (SHOW ON SLIDE)
3. Key point: Σ contains all the information needed to recover α. Describe the steps.
4. Standard: nothing special needed to avoid RIC: δ and γ can be indep. since Σ prevents likelihood from factorizing; indep priors for (γ, δ) imply dependence for (β, γ).
5. BDML: simple, flexible, and fully-Bayesian. (likelihood principle). Contrast other "Bayesian" DML ideas.
6. Note: we don't place an explicit prior on α here, but there's an implied prior via Σ. Could examine via simulations to help elicit Σ if desired

BDML versus Frequentist Double Machine Learning (FDML)

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

FDML Optimizes

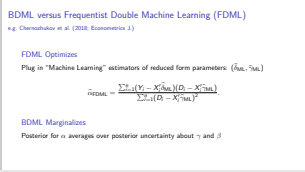
Plug in "Machine Learning" estimators of reduced form parameters: $(\widehat{\delta}_{\text{ML}}, \widehat{\gamma}_{\text{ML}})$

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

BDML Marginalizes

Posterior for α averages over posterior uncertainty about γ and β

BDML versus Frequentist Double Machine Learning (FDML)



1. Point out that RIC is still a problem for Frequentists! Difference between prediction and causal inference (e.g. OLS versus IV – question I always ask my students!)
2. Here is where we want to talk about why BDML could be better than FDML even in terms of Frequentist performance
3. In high-dimensional spaces there is vanishingly little probability near the mode.
4. FDML was an important advance, but some simulation evidence is emerging to suggest that it doesn't always perform well in finite samples and performance can be quite sensitive to the first-step, despite large-sample "generic" results.
5. BDML aims for best of both worlds: not too sensitive to "pragmatic" aspects of prior but allows subject-matter expertise if desired; fully-Bayesian but good Frequentist properties.

Theoretical Results

$$Y_i = X_i'\delta + U_i$$
$$D_i = X_i'\gamma + V_i$$
$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \Bigg| X_i \sim \text{Normal}_2(0, \Sigma)$$

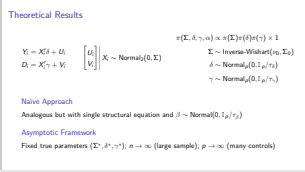
$$\pi(\Sigma, \delta, \gamma, \alpha) \propto \pi(\Sigma)\pi(\delta)\pi(\gamma) \times 1$$
$$\Sigma \sim \text{Inverse-Wishart}(\nu_0, \Sigma_0)$$
$$\delta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\delta)$$
$$\gamma \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\gamma)$$

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)



1. In practice: exact finite sample inference conditional on model (sampling posterior).
2. But to compare and contrast FDML and Naive with BDML present some asymptotics
3. This slide: model specification we use for our derivations.
4. "Vanilla" Bayes model for multivariate regression (Zellner; 1971). EXPLAIN
5. IW mean is $\Sigma_0/(\nu_0 - p - 1)$; $\nu_0 = \#$ of "pseudo-obs" i.e. $\nu_0 \uparrow$ means tighter prior.
6. Parameterize normals in terms of precision τ: 1/Variance. Larger τ means tighter prior.
7. Naive approach: the same idea but only one equation: Y on (D, X) regression with error ε
8. Use priors but consider asymptotics where true parameters are *fixed*. Could also derive results for random coefficients.
9. Asymptotic sequence with a large sample and 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: $\tau \asymp p$

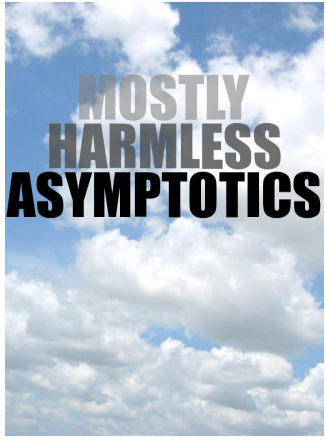
Regularity Conditions

(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^*



Our asymptotic framework ensures bounded R-squared.

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) $\text{Var}(X) = \Sigma_X$ "well-behaved" as $p \to \infty$
(iii) $\lim_{p\to\infty} \sum_{j=1}^{p} (\theta_1^*)^2 < \infty$, $\lim_{p\to\infty} \sum_{j=1}^{p} (\gamma_1^*)^2 < \infty$
(iv) iid errors/controls, $\mathbb{E}(X_i) = 0$, finite & p.d. Σ^*

1. KEY POINT: ensure that R-squared for both reduced form regressions is strictly between zero and one in the limit; aim to capture the finite-sample phenomenon of interest
2. Recall: fixed true pars; asymptotic sequence where #obs and #controls both grow.
3. Intuition for rates: (i) many controls but not too many; (ii) weakly informative priors – data wins; (iii) shrink more when you have more controls
4. $p < n$ isn't need to apply our method but makes it easier to analyze all estimators at once.
5. Here "well-behaved" means: (i) average of e-values bounded, (ii) spread of e-values is bounded, (iii) e-values don't get too small. Last condition: limit version of "strictly pd"
6. Third regularity condition: true reduced form pars "don't explode" as #controls grows. More controls $\Rightarrow$ each matters less on average. "Add most important controls first"
7. Zero mean controls is WLOG

Selection Bias in the Limit

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

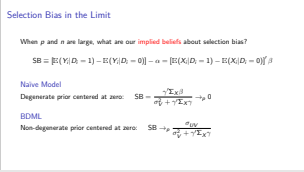
$$\text{SB} \equiv [\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0)] - \alpha = [\mathbb{E}(X_i|D_i = 1) - \mathbb{E}(X_i|D_i = 0)]' \beta$$

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: $\text{SB} \to_p \dfrac{\sigma_{UV}}{\sigma_V^2 + \gamma' \Sigma_X \gamma}$



1. Mention my beliefs paper: prior beliefs are "overdetermined" in that we have beliefs over many aspects of the problem that could contradict one another so it's a useful exercise to see what a particular prior implies about derived quantities we can think about.
2. Explain what selection bias is; in our setting it takes this form
3. This slide: when p and n are large, under our asymptotic approximation, what is the *implied* prior on selection bias?
4. Another way to think about RIC / Bayes Ignorability: in high dimensions, you've "accidentally" ruled out selection bias a priori!
5. BDML avoids this problem

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.

Summary of Asymptotic Results

Consistency
Naive, 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 Naive.

$\sqrt{n}$-Consistency
Naive 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.

1. Remember: $p/n \to 0$
2. Intuitively: BDML and FDML allow for more controls. Can see this both in the bias comparison and the point about root-n consistency.
3. Why do we care about root-n consistency? Measure of quality of estimator: required to use CLT for inference

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

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

Simulation Experiment

$$Y_i = \alpha D_i + X_i'\beta + \varepsilon_i \qquad \{X_i\}_{i=1}^n \sim \text{iid Normal}_p(0, \iota_p)$$
$$D_i = X_i'\gamma + V_i \qquad ((\varepsilon_i, V_i))_{i=1}^n \mid X \sim \text{iid Normal}_2\left(0, \text{diag}\left\{\sigma_\varepsilon^2, 1\right\}\right)$$
$$\beta \mid (X, \varepsilon, V) \sim \text{Normal}_p\left(\mu_\beta, \sigma_\beta^2 \mathrm{I}\right).$$

Linero's (2023) "Fixed" Design

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

1. Goals of the simulation study: compare Frequentist performance (coverage of 95% intervals / RMSE) of various competing estimators.
2. Especially interested in comparison with Frequentist BDML.
3. Compare against Linero (2023) "giving him the home field advantage" i.e. using one of the simulation designs from his paper that was *explicitly chosen* to make his approach look good compared to HCPH.
4. This design generates new control regressors $\mathbf{X}$ and a new parameter vector β in each replication but holds γ and α constant across replications.
5. Following the "fixed" design from Linero (2023), we vary σ_ε over a grid of values from 1 to 4 and set the remaining parameters as indicated, where ι_p denotes a p-vector of ones.

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 $\sigma_\delta, \delta_\gamma$.

Two Versions of BDML

Both Versions
LKJ(4) Prior on $\mathrm{Corr}(U, V)$; Independent Cauchy(0, 2.5) priors on $\mathrm{SD}(U)$ and $\mathrm{SD}(V)$

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 $\sigma_\delta, \delta_\gamma$.

1. Point out that the theory from above was from a plain-vanilla Normal $\times$ IG, but actually we haven't simulated that one yet. (Not sure why: it's actually really easy!)
2. The LJK prior is generally considered to be the "state of the art" weakly informative prior. Admit that LKJ(4) is actually a bit strong; we should probably try 2 next and see if it matters. (This was just the first thing we tried) Point out that LKJ(1) is uniform over correlation matrices.
3. Some of our theoretical derivations–not discussed today?–suggest that the hierarchical version would be expected to perform better.

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.

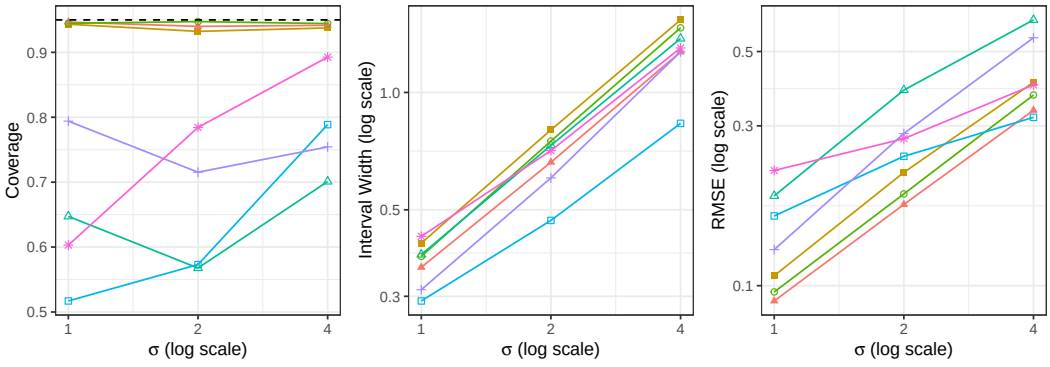
Linero (2023; JASA)

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

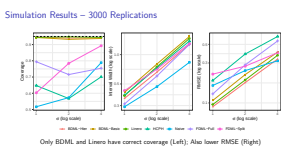
Two-Step "Plug-in" Bayesian Approaches
Preliminary Regression
$\widehat{D}_i \coloneqq X_i^\top \widehat{\gamma}_{\text{pooled}} \leftarrow$ estimate from Bayesian regression of D on X.
HCPH (Hahn et al, 2018; Bayesian Analysis)
1. Bayesian linear regression of Y on $(D - \widehat{D})$ and X
2. Estimation / 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.

1. Could call HCPH "Bayesian Single Machine Learning" since it (approximately) residualizes D with respect to X.
2. Notice that Linero's approach *requires* Bayes / regularization: otherwise there's perfect multicollinearity between X_i and $\widehat{D}_i$. Not obvious why adding this term makes a difference.
3. Point out that a key point of Linero is that HCPH doesn't work well in practice.
4. Explain how these are different from ours: we do do "full luxury Bayes" rather than plug-in; can think of what we're doing is "approximately" residualizing both Y and D with respect to X.

Simulation Results – 3000 Replications

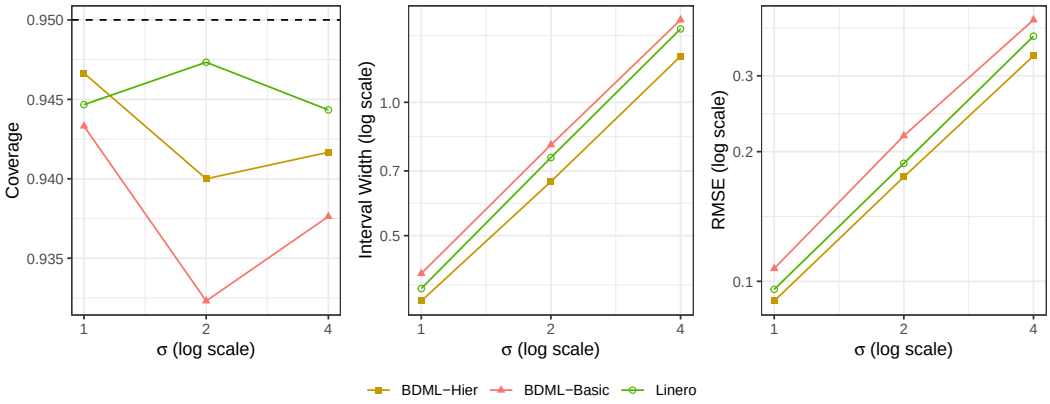


Only BDML and Linero have correct coverage (Left); Also lower RMSE (Right)

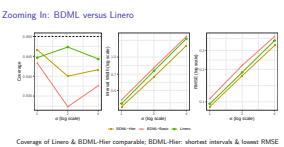


1. Left panel: coverage probability of nominal 95% interval for α; middle is average width of those intervals; right panel is RMSE of estimator of α (posterior mean for Bayes)
2. FDML: ridge regression first step. Two versions: sample-splitting estimates "final" regression on a hold-out dataset. Full re-uses the data.
3. For Naive, HCPH, and Linero carry out estimation and inference *exactly* as in Linero (2023), using the `BLR` package in R.
4. BDML Basic and Hier: STAN (Hamiltonian MC via NUTS) using priors described above.
5. In this DGP, both of the FDML procedures perform poorly.
6. HCPH performs poorly which isn't a surprise: Linero chose this design to make the point that the HCPH approach has problems.
7. BDML and Linero perform best; but this is Linero's design! We didn't tweak or tune.

Zooming In: BDML versus Linero



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



1. Zoom in on best performers
2. Warn that the legend has (unfortunately) changed compared to previous slide: sorry!
3. Point out that difference in coverage difference between BDML-Hier and Linero not being statistically discernible (1 SE difference)
4. In terms of width and RMSE, BDML-Hier has an edge.

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?