

Bayesian Double Machine Learning for Causal Inference*

Francis J. DiTraglia^{†1} and Laura Liu^{‡2}

¹Department of Economics, University of Oxford

²Department of Economics, University of Pittsburgh

First Version: November 6, 2024 This Version: August 18, 2025

Abstract

This paper proposes a simple, novel, and fully-Bayesian approach for causal inference in partially linear models with high-dimensional control variables. Off-the-shelf machine learning methods can introduce biases in the causal parameter known as regularization-induced confounding. To address this, we propose a Bayesian Double Machine Learning (BDML) method, which modifies a standard Bayesian multivariate regression model and recovers the causal effect of interest from the reduced-form covariance matrix. Our BDML is related to the burgeoning frequentist literature on DML while addressing its limitations in finite-sample inference. Moreover, the BDML is based on a *fully generative probability model* in the DML context, adhering to the likelihood principle. We show that in high dimensional setups the naïve estimator implicitly assumes no selection on observables—unlike our BDML. The BDML exhibits lower asymptotic bias and achieves asymptotic normality and semiparametric efficiency as established by a Bernstein–von Mises theorem, thereby ensuring robustness to misspecification. In simulations, our BDML achieves lower RMSE, better frequentist coverage, and shorter confidence interval width than alternatives from the literature, both Bayesian and frequentist.

Keywords: Causal Inference, Regularization-induced Confounding, Bayesian Methods, Double Machine Learning

JEL Codes: C11, C14, C55

*We thank Otilia Boldea, Neal Shephard, and seminar participants at York and the Frank Diebold 65+35 Conference for helpful comments and suggestions. We thank Lasse von der Heydt for excellent research assistance.

[†]DiTraglia: francis.ditraglia@economics.ox.ac.uk

[‡]Liu: laura.liu@pitt.edu

1 Introduction

Two decades into the “credibility revolution” popularized by *Mostly Harmless Econometrics*, much work in applied microeconomics continues to rely on relatively simple reduced-form models to untangle cause-and-effect from observational data. In the simplest case, and the one we consider below, a researcher argues that a treatment D is as good as randomly assigned after adjusting for a set of control variables X and goes on to report the coefficient on D from a regression of Y on (D, X) as the causal effect of interest. Naturally this selection-on-observables approach is only as good as the choice of controls; researchers who rely on it typically devote much of their paper to defending the choice of X in their application.

All else equal, a selection-on-observables strategy is more convincing when it includes more observables. For this reason the dimension of X can be very large relative to sample size. Even if there are relatively few “underlying” control variables, fully adjusting for them may require us to include a large number of interactions and other transformations in our regression. Applied researchers thus face a difficult trade-off: including more controls reduces the risk of bias but can lead to extremely imprecise estimates of the causal effect of interest.

In this paper, we provide a simple, fully Bayesian approach to navigating this bias-variance trade-off in a workhorse partially linear model: Bayesian Double Machine Learning (BDML). While our methods are Bayesian, the resulting point and interval estimators are designed to appeal to a broad audience. For Bayesians, BDML offers a natural way to bring the full Bayesian toolbox to bear on high-dimensional causal inference problems in applied microeconomics. For Frequentists, BDML provides superior finite-sample performance in practice—lower root mean-squared error and improved coverage of confidence intervals—while delivering the same attractive asymptotic guarantees as existing Frequentist approaches.

Any attempt to tame the estimation uncertainty that arises from including many controls relies on *regularization*, also known as shrinkage. Rather than “projecting out” X from Y and D using ordinary least squares (OLS), a regularized estimator *constrains* the high-dimensional vector β of coefficients on X in the regression of Y on (D, X) . Bayesian approaches do so by placing a prior on β ; machine learning (ML) approaches do so either by adding a penalty term to the objective function to favor “smaller” values for the elements of β , as in LASSO or Ridge, or by penalizing model complexity through implicit algorithmic constraints, as in random forests and gradient boosting.

Whether Bayesian or Frequentist, implicit or explicit, any application of regularization to causal inference problems faces a key challenge: *regularization-induced confounding* (RIC).¹ Unlike OLS, regularized estimators only partially adjust D and Y for the observed con-

¹This term was coined by [Hahn et al. \(2018\)](#).

founders X . This induces a correlation between D and the regularized regression residuals, biasing the estimated causal effect. Granted, researchers must be prepared to accept at least some increase in bias in exchange for reduced estimation uncertainty. The problem of RIC is that this bias can be *extremely sensitive* to the precise way in which regularization is carried out, depending on the relationship between D and X . Indeed, naïve approaches to regularized estimation can easily *increase* the root mean-squared error (RMSE) of our causal effect estimator, adding a bias that far outweighs the reduction in variance. Crucially, our BDML approach side-steps this problem completely: using a simple re-parameterization of the model, we avoid RIC while relying on standard methods for Bayesian linear regression.

From a subjective Bayesian point of view, RIC may seem puzzling. If we believe our prior and likelihood, Bayes optimality *requires* us to base estimation and inference on the implied posterior. In practice, however, priors are often chosen for pragmatic reasons, and as [Sims \(2012\)](#) notes, such priors “can unintentionally imply dogmatic beliefs about parameters of interest” when expanded “unthinkingly to high dimensions.” [Linero \(2023\)](#) shows that the causal inference problem we study here has this problem in spades. It is common, and convenient in practice, to place independent priors on γ and β , the coefficients on X in the “propensity score” and “outcome” regressions, respectively. In low dimensions, this is innocuous. In high dimensions, on the other hand, it implies a *dogmatic* prior belief that there is no selection bias. If we truly believed this, of course, there would be no reason to adjust for X in the first place!

To avoid prior dogmatism without the need to elicit a dependent prior for (γ, β) , BDML replaces the outcome regression of Y on (D, X) with a *reduced form regression* of Y on X . This leaves us with a bivariate, seemingly unrelated regression of (Y, D) on X . Under the selection-on-observables assumption, the error terms of these two equations are correlated, and their covariance matrix contains all the information needed to carry out estimation and inference for the causal effect of D on Y . We show that placing independent priors on the coefficients of the seemingly unrelated regression equations runs no risk of prior dogmatism, regardless of the dimension of X . Any standard prior on the error covariance matrix implies a plausible range for the *a priori* selection bias.

We call our method BDML to highlight its connection to the Frequentist Double Machine Learning (FDML) approach that has become extremely popular in recent years (see for example [Chernozhukov et al., 2018](#)). Like our proposal, FDML can be viewed as a means of avoiding RIC when carrying out causal inference in high dimensions. In the partially linear regression model we study here, FDML likewise considers a bivariate reduced form regression of (Y, D) on X . After estimating the coefficients of these two regressions using an ML approach, FDML recovers the causal effect of interest using a further regression of the Y -

residuals on the D -residuals. This approach is computationally convenient and can be given a rigorous large-sample justification. In practice, however, it amounts to treating the reduced form coefficients as if they were known—that is, FDML *profiles* out the nuisance parameters by plugging in point estimates. When the dimension of X is large relative to sample size there may be considerable sampling uncertainty in the reduced form coefficients. BDML explicitly acknowledges this by *marginalizing* over all sources of uncertainty in the problem, leading to improved small-sample performance as measured by traditional Frequentist metrics, e.g. root mean-squared error (RMSE) and coverage probability.

Crucially, the finite-sample advantages of BDML do not come at the expense of large sample guarantees. To show this, we study a fixed coefficient asymptotic sequence in which the number of control regressors p grows with sample size n , matching the approach of the FDML literature. We consider three estimators: BDML, FDML, and a “naïve” Bayesian alternative that ignores RIC. Under regularity conditions, we show that all three yield consistent estimators of the causal effect of interest. When they are \sqrt{n} -consistency, we likewise show that all three have the same asymptotic variance. The asymptotic bias of BDML and FDML, however, is strictly smaller than that of the naïve Bayesian approach: p^2/n^2 compared to p/n . Moreover, the naïve estimator is only \sqrt{n} -consistent if the number of controls is much smaller than sample size: $p/\sqrt{n} \rightarrow 0$. BDML and FDML, in contrast, allow for a much larger number of controls, requiring only $p/(n^{3/4}) \rightarrow 0$. Asymptotically, BDML improves on the naïve Bayesian approach while matching the performance of FDML.

Continuing our asymptotic analysis, we build on results from Walker (2025) to establish a Bernstein-von Mises (BvM) theorem for BDML when $(\log n)^2 \prec p \prec \sqrt{n}$.² This result further strengthens the Frequentist appeal of BDML, showing that posterior credible sets constructed using our approach are valid Frequentist confidence intervals. Our BvM result further establishes the semi-parametric efficiency of BDML while imposing weaker assumptions on the prior than the recent Bayesian literature (e.g. Breunig et al., 2024; Luo et al., 2023). While we prove the BvM theorem under the assumption of Gaussian errors, these results are robust to mis-specification of the error distribution.

We conclude by comparing the finite-sample performance of BDML against a range of Bayesian and Frequentist competitors in a comprehensive simulation study and a number of empirical examples drawn from the recent literature in applied microeconomics. Our results show that BDML provides the “best of both worlds”, matching the asymptotic properties of Frequentist alternatives while delivering dramatically improved finite-sample performance. Compared existing Bayesian alternatives (e.g. Hahn et al., 2020; Linero, 2023), BDML maintains its impressive performance across a much wider range of data-generating processes

²We use the symbol $a \prec b$ to indicate that $\lim_{n \rightarrow \infty} a/b \rightarrow 0$.

(DGPs). Since BDML is fully-Bayesian, it correctly accounts for all sources of sampling uncertainty, smoothly adapting to key features of the DGP.

This paper relates to a large recent literature on (Frequentist) double machine learning (FDML), also known as “double-debiased machine learning”. See [Ahrens et al. \(2025\)](#) for a comprehensive practical introduction and [Chernozhukov et al. \(2018\)](#) for further theoretical details. FDML provides a general framework for estimating a low-dimensional target parameter using moment conditions that depend on a high-dimensional nuisance parameter. The naïve plug-in approach—constructing a preliminary ML estimate of the nuisance parameter and plugging it into the original moment condition—has two serious shortcomings. The first is regularization bias, a more general term for RIC. The second is overfitting bias from using the same data twice—first to estimate the nuisance parameter and then to estimate the target parameter. By *orthogonalizing* the second-step moment condition with respect to the nuisance parameter, FDML inoculates against regularization bias to first order. By using sample splitting, it likewise avoids overfitting bias.

Two key advantages of FDML are its simplicity and generality: whenever we can derive the appropriate orthogonalized moment condition, FDML justifies treating the estimated nuisance parameter as if it were known while providing theoretical guarantees across a wide range of first-step ML estimators. These attractive properties have led applied researchers in economics to employ FDML in a range of empirical settings, especially the “post-double selection” (PDS) LASSO approach of [Belloni et al. \(2014\)](#). Recent papers that apply FDML as their primary method in an observational setting include [Dube et al. \(2020\)](#) who estimate monopsony power in online labor markets, and [Yang et al. \(2020\)](#) who estimate the “Big N” audit quality effect. While its primary motivation is to reduce bias in observational settings, FDML is also widely applied in experimental settings to flexibly incorporate covariates, either as a robustness check ([Alsan et al., 2023](#); [Armand et al., 2020](#)) or with the aim of improving estimation precision ([Chioda et al., 2021](#); [Cotti et al., 2025](#); [Delfino, 2024](#); [Hussam et al., 2022](#)). FDML has also been used to evaluate, and ultimately question, the plausibility of the selection-on-observables assumption, either by comparing against an instrumental variables (IV) benchmark ([Dufflo et al., 2021](#)) or by creating a “simulated” observational dataset from a collection of randomized controlled trials ([Gordon et al., 2023](#)).

Although FDML is indisputably a valuable tool in applied economics, some recent work suggests that its finite sample performance can be sensitive to details of the first-step ML procedure in ways that may lead it to underperform in practice ([Angrist and Frandsen, 2022](#); [Bach et al., 2024](#); [Wüthrich and Zhu, 2023](#)). [Ahrens et al. \(2025\)](#) in particular stress that “poorly tuned or ill-suited nuisance estimators can produce severely misleading point estimates”. As such, they encourage researchers using FDML to “carefully choose and tune their

nuisance function estimators and consider a rich set of estimation approaches.” While by no means an insurmountable challenge, the need for careful tuning makes FDML more complicated to implement in practice than its theoretical simplicity may suggest. In our view, this challenge makes Bayesian alternatives to FDML particularly appealing. By marginalizing over uncertainty in the nuisance parameter, BDML provides a fuller accounting for finite-sample estimation uncertainty. Moreover, our method can provide robust performance across a wider range of data generating processes through the use of hierarchical priors, which adapt the degree of shrinkage to the data at hand.

Our work also relates to a growing literature on Bayesian approaches to causal inference: see [Heckman et al. \(2014\)](#), [Li et al. \(2023\)](#), and [Linero and Antonelli \(2023\)](#) for broad discussions of this field. The studies closest to ours are [Hahn et al. \(2018, 2020\)](#) and [Linero \(2023\)](#), who consider the same partially linear model that we analyze below.³ Compared to BDML, which marginalizes over all sources of uncertainty, these papers adopt a “hybrid” strategy: profiling over some parts of the problem and marginalizing over others. This hybrid approach is computationally convenient and can be given a sound Bayesian justification. Our simulation studies reveal, however, that the full marginalization provided by BDML is crucial for obtaining inferences that have good frequentist properties across a wide range of data generating properties. As mentioned above, our BvM results draw heavily on [Walker \(2025\)](#). Recent Bayesian work exploring issues related to those considered here, albeit in different models, includes [Saarela et al. \(2016\)](#), [Ray and Van Der Vaart \(2020\)](#), [Antonelli et al. \(2022\)](#), [Luo et al. \(2023\)](#), [Breunig et al. \(2024\)](#) and [Breunig et al. \(2025\)](#).

The remainder of this paper is organized as follows. Section 2 introduces our model and notation while Section 3 discusses the problem of RIC in detail. Section 4 presents our Bayesian Double Machine Learning approach and describes how it relates to both Bayesian and Frequentist alternatives, while Section 5 presents a comparison of the asymptotic properties of BDML, FDML, and a naïve Bayesian alternative. Finally, Section 6 presents a simulation study comparing the finite-sample performance of BDML against a range of Bayesian and Frequentist competitors. Proofs appear in Section A.

2 The Model

We consider the problem of estimating and carrying out inference for the parameter α in a partially linear structural model of the form

$$Y_i = \alpha D + g(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | D_i, X_i] = 0 \quad (1)$$

³See [Belloni et al. \(2014\)](#) for a frequentist analysis of the partially linear model.

where Y_i is the outcome of interest, D_i is the “treatment” or “policy” variable, and X_i is a vector of control variables.⁴ We do not assume that D_i is binary; if it is, (1) can be viewed as a special case of the “selection on observables” framework. For reasons that will become clear below, we introduce an additional *reduced form* model, namely

$$D_i = m(X_i) + V_i, \quad \mathbb{E}[V_i|X_i] = 0 \quad (2)$$

where $m(X_i) \equiv \mathbb{E}[D_i|X_i]$ so that (2) holds by construction and we have

$$\text{Cov}(\varepsilon_i, V_i) = \mathbb{E}[\varepsilon_i V_i] = \mathbb{E}[\varepsilon_i D_i] - \mathbb{E}[\varepsilon_i m(X_i)] = 0 \quad (3)$$

by iterated expectations and $\mathbb{E}[\varepsilon_i|X_i, D_i] = 0$, since ε_i and V_i are mean zero.

A leading example of (1)–(3) assumes that $g(X_i) = X_i' \beta$ and $m(X_i) = X_i' \gamma$ where β and γ are p -vectors of unknown regression coefficients. This simplification gives

$$Y_i = \alpha D_i + X_i' \beta + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i|X_i, D_i] = 0 \quad (4)$$

$$D_i = X_i' \gamma + V_i, \quad \mathbb{E}[V_i|X_i] = 0 \quad (5)$$

where we have

$$\text{Cov}(\varepsilon_i, V_i) = \text{Cov}(\varepsilon_i, D_i - X_i' \gamma) = \text{Cov}(\varepsilon_i, D_i) - \text{Cov}(\varepsilon_i, X_i') \gamma = 0. \quad (6)$$

While the methods we develop in this paper are easily extended to the general case of (1)–(3), we focus on the special case of (4)–(6) to simplify the exposition. We assume throughout that the dimension p of X_i is *large relative to sample size* n . If the dimension of X_i exceeds the number of observations, then there is no unique ordinary least squares (OLS) estimator of α based on (4). Even if there are more observations than regressors, the OLS estimator of α is likely to be extremely noisy if X_i contains many “relatively unimportant” control variables—variables that, taken individually, have little predictive power for D_i and Y_i . This situation is arguably quite common in practice. Applied researchers may know which controls it would *suffice* to adjust for without being sure of which are most important.

3 Regularization-Induced Confounding (RIC)

As is well-known, the OLS estimator for α corresponds to the Bayesian posterior mean under (4) with a normal distribution for ε_i and flat priors on (α, β) . Accordingly, in a situation

⁴Our model and parameter of interest match those of [Belloni et al. \(2014\)](#).

where we expect OLS to perform poorly because the dimension of X_i is large, a natural idea would be to “shrink” our estimates by introducing an informative prior on β . While there are many possibilities for the choice of shrinkage prior, the message of this section is that *none of them* is likely to perform well in practice. To illustrate this point, suppose we retain a flat prior for α but place independent normal priors on the elements of β :

$$\beta^{(1)} \dots \beta^{(p)} \sim \text{iid Normal}(0, \sigma_\beta^2). \quad (7)$$

Under the simplifying assumption that σ_ε^2 is known, the Bayes estimator for (α, β) under squared error loss has a closed-form solution:

$$\begin{bmatrix} \hat{\alpha}_\lambda \\ \hat{\beta}_\lambda \end{bmatrix} = \left[\begin{pmatrix} D'D & D'X \\ X'D & X'X \end{pmatrix} + \begin{pmatrix} 0 & 0'_p \\ 0_p & \lambda \mathbb{I}_p \end{pmatrix} \right]^{-1} \begin{pmatrix} D'Y \\ X'Y \end{pmatrix}, \quad \lambda \equiv \frac{\sigma_\varepsilon^2}{\sigma_\beta^2}. \quad (8)$$

As seen from (8), the Bayes estimator in this setting corresponds to a modified version of Ridge Regression in which α is unpenalized. As usual in Ridge Regression, we implicitly assume that D_i , X_i and Y_i have all been de-meaned so that the regression requires no intercept. If (7) represents genuine subjective researcher beliefs over β then, from a Bayesian point of view, $\hat{\alpha}_\lambda$ is the mean-squared error optimal estimator of α . In practice, however, shrinkage priors like (7) are often used for pragmatic reasons: when X_i is high-dimensional, it may be difficult if not impossible to elicit a fully-informative prior over β . For the purposes of *predicting* Y_i this is usually innocuous, provided that σ_β^2 is chosen in a reasonable way. But when the goal is to understand the *causal effect* of D_i , the situation changes dramatically. To illustrate this point, we derive the (Frequentist) bias and variance of $\hat{\alpha}_\lambda$ as follows under the simplifying assumption that ε is homoskedastic.⁵

Assumption 1. $Y = \alpha D + X'\beta + \varepsilon$ where $\mathbb{E}[\varepsilon|X, D] = 0$ and $\text{Var}(\varepsilon|X, D) = \sigma^2 \mathbb{I}_p$.

Proposition 1. Let $\hat{\rho}' \equiv (D'D)^{-1}D'X$, and $R \equiv \hat{\xi}'\hat{\xi}$ where $\hat{\xi} \equiv [\mathbb{I}_p - D(D'D)^{-1}D']X$. Then *Assumption 1* implies

$$\text{Bias}(\hat{\alpha}_\lambda|X, D) = \hat{\rho}' [\mathbb{I}_p - (R + \lambda \mathbb{I}_p)^{-1}R] \beta \quad (9)$$

$$\text{Var}(\hat{\alpha}_\lambda|X, D) = \sigma_\varepsilon^2 [(D'D)^{-1} + \hat{\rho}'(R + \lambda \mathbb{I}_p)^{-1}R(R + \lambda \mathbb{I}_p)^{-1}\hat{\rho}] \quad (10)$$

where $\hat{\alpha}_\lambda$ is as defined in (8).

The j^{th} element of $\hat{\rho}$ from *Proposition 1* is the slope coefficient from a regression of $X_i^{(j)}$ on D_i without an intercept (both have been de-meaned) while the j^{th} column of $\hat{\xi}$ is the

⁵Hahn et al. (2018) derive a special case of $\text{Bias}(\hat{\alpha}_\lambda)$, but their expression contains a minor error.

corresponding vector of residuals. Setting $\lambda = 0$ corresponds to $\sigma_\beta^2 \rightarrow \infty$ and gives the OLS estimator of α . In this case the bias is zero and the variance simplifies to

$$\text{Var}(\hat{\alpha}_{\text{OLS}}) = \sigma_\varepsilon^2 [(D'D)^{-1} + \hat{\rho}' R^{-1} \hat{\rho}].$$

If σ_β^2 is finite, then $\lambda \neq 0$ and $\hat{\alpha}_\lambda$ is biased unless $\hat{\rho}$ or β equals zero. Of course this in and of itself is not a convincing argument against shrinkage. In general, higher values of λ imply lower values of $\text{Var}(\hat{\alpha})$. Even if $n > p$ so that OLS remains feasible, we should still prefer to set $\lambda > 0$ provided that the decrease in variance outweighs the increase in (squared) bias.

The problem isn't the presence of bias but rather its dependence on other aspects of the model. We include covariates on the right-hand side of (4) precisely because we believe that α would lack a causal interpretation without them. This implies an *a priori* belief that the control variables covary with both the outcome and treatment. As seen from (9), for a fixed value of λ the bias of $\hat{\alpha}_\lambda$ involves both β and $\hat{\rho}$. All else equal, larger values of $\hat{\rho}^{(j)} \beta^{(j)}$ imply a higher bias. Our beliefs about the magnitude of $\beta^{(j)}$ are already accounted for, at least in principle. Since $\lambda = \sigma_\varepsilon^2 / \sigma_\beta^2$, a belief that $|\beta^{(j)}|$ will be large on average leads us to shrink *less*, using a smaller value of λ and keeping the bias in check. But $\text{Bias}(\hat{\alpha}_\lambda)$ also depends crucially on $\hat{\rho}$, and λ fails to take this into account.⁶ If the treatment and controls are strongly correlated, even a reasonable choice for σ_β^2 can produce a highly biased estimator of α , a phenomenon that Hahn et al. (2018) call **regularization-induced confounding (RIC)**. The name alludes to the fact that “regularizing” the OLS estimator by setting $\lambda > 0$ causes the residuals to violate the sample analogue of $\mathbb{E}[D_i \varepsilon_i | X_i] = 0$, the population moment condition that gives α a causal interpretation. While RIC is typically equated with unacceptably high bias, numerical experiments using (9)–(10) confirm that RIC implies an unfavorable bias-variance *trade-off*. This is the sense in which we use the term.

The trouble, it seems, is failing to account for our prior beliefs about γ , the reduced form coefficient from (5) that reflects the dependence between D_i and X_i . A natural solution, then, would be to incorporate these beliefs into our model. Doing so requires us to add an equation and work with a *bivariate* regression including both (4) and (5). Staying as close as possible to the single-equation example from above, suppose we specify a normal likelihood for (ε_i, V_i) and place independent normal shrinkage priors on the elements of β and γ , all centered at zero. In this model, an additional hyperparameter σ_γ^2 quantifies our prior beliefs about the strength of dependence between the treatment and control variables. Unfortunately, σ_γ^2 will have *no effect whatsoever* on our estimates or inferences for α . Recall from (6) above that the causal assumptions of our model require $\text{Cov}(\varepsilon_i, V_i) = 0$. Adopting

⁶The stated priors on (α, β) along with the fixed known value of σ_ε^2 *implicitly* encode beliefs about the likely magnitude of $\hat{\rho}$, as we explain further below.

the shorthand θ to refer to all model parameters, this implies that the likelihood factorizes as follows

$$f(Y, D|X, \theta) = f(Y|D, X, \theta)f(D|X, \theta) = f(Y|D, X, \alpha, \beta, \sigma_\varepsilon^2)f(D|X, \gamma, \sigma_V^2). \quad (11)$$

Notice that the first factor involves only $(\alpha, \beta, \sigma_\varepsilon^2)$ while the second involves only (γ, σ_V^2) . Therefore, unless we specify a prior in which β and γ Treating σ_ε^2 as known, the bias and variance expressions from (9)–(10) continue to hold and the problem of RIC remains. Linero (2023) calls a causal model in which the likelihood factorizes as in (11) and β and γ are *a priori* independent a **Bayesian-ignorable** model, and points out that such models imply a highly dogmatic prior for the extent of selection bias when p is large. This is another lens through which to view RIC: it arises from “accidentally” imposing a strong prior belief that there is little or no selection bias. We explore this phenomenon in more detail in subsection 5.2 below.

4 Bayesian Double Machine Learning (BDML)

In principle, specifying an informative prior that incorporates dependence between β from (4) and γ from (5) would solve the problem of RIC introduced in section 3. Eliciting such a prior in practice, however, may be extremely challenging, especially when the number of control variables is large. This motivates us to introduce an alternative approach, one that is inspired by ideas from the recent Frequentist literature on so-called “double-debiased machine learning”, often referred to as “double machine learning” for short. Substituting (5) into (4) gives the reduced form regression

$$Y_i = X_i'(\alpha\gamma + \beta) + (\varepsilon_i + \alpha V_i) = X_i'\delta + U_i \quad (12)$$

where $U_i \equiv \varepsilon_i + \alpha V_i$ and $\delta \equiv \alpha\gamma + \beta$. Since since $\mathbb{E}[\varepsilon_i|X_i, D_i] = 0$ by (4) and X_i is uncorrelated with V_i by (5), X_i is uncorrelated with U_i . Assuming normal errors for simplicity, this gives the following *bivariate* reduced form regression model

$$\begin{aligned} Y_i &= X_i'\delta + U_i \\ D_i &= X_i'\gamma + V_i \end{aligned} \quad \left[\begin{array}{c} U_i \\ V_i \end{array} \right] \bigg| X_i \sim \text{Normal}(\mathbf{0}, \Sigma). \quad (13)$$

Given that our goal is to estimate and carry out inference for α , it may seem odd that we have eliminated this parameter from the regression equation for Y . Crucially, however, the errors (U_i, V_i) in (13) are *correlated* whenever $\alpha \neq 0$. This is what will allow us to learn the

causal effect of interest. By (6) we have $\text{Cov}(\varepsilon, V) = 0$ and thus

$$\begin{aligned}\text{Var}(U_i) &= \text{Var}(\varepsilon_i + \alpha V_i) = \text{Var}(U_i) + \alpha^2 \text{Var}(V_i) = \sigma_\varepsilon^2 + \alpha^2 \sigma_V^2 \\ \text{Cov}(U_i, V_i) &= \text{Cov}(\varepsilon_i + \alpha V_i, V_i) = \alpha \sigma_V^2\end{aligned}$$

by the definition of U . Therefore,

$$\Sigma \equiv \text{Var} \begin{bmatrix} U_i \\ V_i \end{bmatrix} = \begin{bmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{bmatrix} = \begin{bmatrix} \sigma_\varepsilon^2 + \alpha^2 \sigma_V^2 & \alpha \sigma_V^2 \\ \alpha \sigma_V^2 & \sigma_V^2 \end{bmatrix}. \quad (14)$$

Notice that knowledge of Σ immediately implies knowledge of α via the simple transformation

$$\frac{\text{Cov}(U, V)}{\text{Var}(V)} = \frac{\sigma_{UV}}{\sigma_V^2} = \frac{\alpha \sigma_V^2}{\sigma_V^2} = \alpha. \quad (15)$$

The relationship in (15) suggests the following approach to estimation and inference for α .

Algorithm 1 (Bayesian Double Machine Learning).

1. Place priors on (δ, γ, Σ) in the reduced form model given by (13).
2. Sample from the joint posterior $(\delta, \gamma, \Sigma) | (X, D, Y)$ where Σ is defined in (14).
3. Transform the posterior draws for Σ to obtain a posterior for $\alpha = \sigma_{UV} / \sigma_V^2$.

The procedure given in Algorithm 1 is simple, flexible, and fully Bayesian. At the same time, it allows us to avoid the problem of inadvertently specifying a dogmatic prior on selection bias, pointed out by Linero (2023) and explained in section 3 above. First, because we work with a pair of reduced form regressions rather than a structural regression for Y_i and a reduced form regression for D_i , our likelihood does not factorize: (U_i, V_i) are correlated whereas (ε_i, V_i) are not. Second, because $\delta = (\alpha\gamma + \beta)$, placing independent priors on the reduced form regression coefficients δ and γ does *not* imply independent priors for β and γ . Hence, (13) in general satisfies *neither* of the two conditions required for Bayesian-ignorability introduced by Linero (2023). Naturally, the properties of point and interval estimates based on Algorithm 1 will depend on the specific choice of priors for (δ, γ, Σ) . We explore this point in detail below and compare the theoretical and practical performance of different priors. The key insight of our approach, however, is that re-writing the model from (4)–(6) in the form given by (13) allows us to side-step the problem of RIC using a standard Bayesian multivariate regression model with conditionally conjugate priors. As far as we are aware, this observation is new to the literature.

Because the estimator $\hat{\alpha}_\lambda$ from (8) can be viewed from a Frequentist perspective as an instance of Ridge Regression, (9)–(10) imply that RIC is just as much of a challenge for frequentists who apply machine learning methods to high-dimensional causal inference problems as it is for Bayesians. The increasing recognition of this problem over the past decade has led to the development of a framework called (frequentist) double machine learning (FDML). See Chernozhukov et al. (2018) for an overview. The simplest FDML approach to estimating α in (4) proceeds by running three regressions. First, let $\hat{\delta}_{\text{ML}}$ be the estimated coefficient vector from some unspecified “machine learning” (i.e. regularized) regression of Y_i on X_i , e.g. ridge regression, LASSO, random forests, etc. Second, let $\hat{\gamma}_{\text{ML}}$ be the estimated coefficient vector from some unspecified regularized regression of D_i on X_i . The FDML estimator is then obtained by regressing the residuals from the Y_i on X_i regression on the residuals from the D on X regression, in other words

$$\hat{\alpha}_{\text{FDML}} = \frac{\sum_{i=1}^n (Y_i - X_i' \hat{\delta}_{\text{ML}})(D_i - X_i' \hat{\gamma}_{\text{ML}})}{\sum_{i=1}^n (D_i - X_i' \hat{\gamma}_{\text{ML}})^2}. \quad (16)$$

Like (13) from above, (16) involves reduced form regressions of Y and D on X_i . Because of this similarity, we refer to the approach from Algorithm 1 as **Bayesian Double Machine Learning (BDML)**. From a Bayesian perspective, BDML provides a way to guard against RIC while using a fully generative model and conducting inference in a manner that respects the likelihood principle. This gives researchers access to the full arsenal of Bayesian tools, from hierarchical modeling to posterior predictive checks. But even for researchers who are primarily concerned with frequentist performance, BDML offers two key advantages. First, it allows practitioners to incorporate subject matter expertise into estimation through the choice of prior, potentially yielding sizeable efficiency gains in small samples. To show how this can be achieved in practice, we propose a simple and effective way of incorporating researcher beliefs concerning the R-squared of the reduced form regressions, building on the so-called “R2D2” prior of (Zhang et al., 2022). Second, rather than simply “plugging in” point estimates of δ and γ , as in FDML, our BDML approach marginalizes over these high-dimensional parameters, correctly accounting for all sources of estimation uncertainty for α . As such, BDML may provide more accurate frequentist inference while avoiding the need for cumbersome and computationally-expensive “cross-fitting” procedures, as are commonly employed in applications of FDML.

While ours is not the only Bayesian proposal for avoiding RIC in a model like (10), BDML differs in important ways from existing methods. The two closest to our approach are the methods of Hahn et al. (2018, 2020) on the one hand and Linero (2023), on the other hand. Both begin with a preliminary Bayesian linear regression of D_i on X_i to yield a Bayes

estimator of γ that is then substituted into a second Bayesian regression.⁷ The approaches differ in the precise regression model that they estimate in the second step. [Hahn et al. \(2018\)](#) rely on the same substitution as our (12) but, rather than absorbing αV_i into a reduced form error term, replace it with its point estimate \widehat{V}_i based on a first-step regression.

Algorithm 2 (HCPH). *Obtain the marginal posterior for α from the linear regression*

$$Y_i = \alpha \widehat{V}_i + X_i' \delta_2 + \nu_i, \quad \widehat{V}_i \equiv (D_i - X_i' \widehat{\gamma})$$

where $\widehat{\gamma}$ is a first-step Bayes-estimator of γ from (5).

Note that the parameter $\delta_2 = \delta + \alpha \widehat{\gamma}$ from [Algorithm 2](#) does *not* in general coincide with our δ from (12). [Linero \(2023\)](#) generalizes the second-stage regression from [Hahn et al. \(2018\)](#). Defining $\widehat{D}_i \equiv X_i' \widehat{\gamma}$, we can re-express the regression from [Algorithm 2](#) as

$$Y_i = \alpha D_i - \alpha \widehat{D}_i + X_i' \delta_2 + \nu_i.$$

[Linero \(2023\)](#) estimates the parameters of this regression *without the restriction* that the coefficient on \widehat{D}_i equals the negative of the coefficient on D_i .

Algorithm 3 (Linero). *Obtain the marginal posterior for α from the linear regression*

$$Y_i = \alpha D_i + \kappa \widehat{D}_i + X_i' \delta_3 + \nu_i, \quad \widehat{D}_i \equiv X_i' \widehat{\gamma}$$

where $\widehat{\gamma}$ is a first-step Bayes-estimator of γ from (5).

Notice that the regression from [Algorithm 3](#) has an identification problem: $\widehat{D}_i = X_i' \widehat{\gamma}$ is *perfectly collinear* with X_i . We write δ_3 as the vector of regression coefficients for X_i to emphasize that, when estimated *without* the restriction that $\kappa = -\alpha$, this procedure will not coincide with [Algorithm 2](#).

Compared to the procedures in Algorithms 2–3, our BDML approach differs in two important ways. First, our inferences for α incorporate posterior uncertainty over *both* γ and δ . [Hahn et al. \(2020\)](#) argue that, since Bayesian inference in a regression model *conditions* on the realizations of the regressors, a two-step approach can be viewed as fully-Bayesian: any Bayes-estimator of γ from the first-step regression is necessarily a measurable function of (D, X) , so plugging this estimate into a second-step regression does not violate the likelihood principle. We agree with this argument, as far as it goes. But there are *countless* approaches

⁷While [Hahn et al. \(2018\)](#) takes a more involved joint estimation approach, their followup paper [Hahn et al. \(2020\)](#) advocates the simpler two-step approach that we describe here.

that adhere to the likelihood principle, all of which yield different posteriors. As we show in our simulations below, accounting for uncertainty in $\hat{\gamma}$ is crucial for obtaining inferences that have good frequentist properties in addition to a sound Bayesian justification. Second, by residualizing *both* D_i and Y_i with respect to X_i , BDML achieves robustness properties similar to those of frequentist double machine learning in a fully-Bayesian model. In contrast [Hahn et al. \(2018\)](#) residualizes only D_i . Alluding to similar robustness concerns raised by [Zigler \(2016\)](#), [Linero \(2023\)](#) suggests that his two-step approach could give added robustness when either (4) or (5) may be misspecified by eliminating “feedback” from the estimation problem. As we show below, however, our approach performs well in practice even in the simulation designs from [Linero \(2023\)](#). This is because our approach, like FDML, is “double de-biased” where as Algorithms 2–3 are only “single de-biased”, a point we expand upon below.

5 Asymptotic Properties

5.1 Assumptions

We now explore the large-sample properties of the BDML approach from [Algorithm 1](#) and contrast them with those of the frequentist alternative, FDML, as well as those of a “naïve” Bayesian approach analogous to the Ridge Regression estimator from (8). We first introduce some notation. Let X be the $(n \times p)$ matrix of control regressors from above and define the $(n \times 2)$ matrix of “outcomes” W , the $(p \times 2)$ matrix of coefficients B , and the $(n \times 2)$ matrix of errors E according to

$$W \equiv \begin{bmatrix} Y & D \end{bmatrix}, \quad B \equiv \begin{bmatrix} \delta & \gamma \end{bmatrix}, \quad E \equiv \begin{bmatrix} U & V \end{bmatrix} \quad (17)$$

where $E'_i \equiv (U_i, V_i)$ denotes a specified row of E . Then, we can express the reduced form regression model from (13) in matrix form as

$$W = XB + E, \quad E_i|X \sim \text{iid Normal}_2(0, \Sigma). \quad (18)$$

We consider an asymptotic sequence in which the number of control regressors p grows with sample size n and the true parameter values are *fixed*. We write these fixed true values as Σ^* and $B^* \equiv \begin{bmatrix} \delta^* & \gamma^* \end{bmatrix}$ to distinguish them from the random variables Σ and B that represent prior (or posterior) uncertainty in a Bayesian model. Analogously, when we have occasion to refer to the true values of the structural parameters from (4), we write them as $(\alpha^*, \beta^* \sigma_\varepsilon^*)$.

Both for simplicity and to underscore the fact that BDML does not require the researcher

to specify unusual or elaborate priors, we work with the classic conditionally conjugate prior for seemingly unrelated regression (SUR) models.⁸ In particular, we combine (18) with the prior $\pi(\Sigma, B) = \pi(\Sigma)\pi(\delta)\pi(\gamma)$ where

$$\Sigma \sim \text{Inverse-Wishart}(\nu_0, \Sigma_0), \quad \delta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\delta), \quad \gamma \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\gamma) \quad (19)$$

and $(\tau_\delta, \tau_\gamma)$ denote the prior precisions for the reduced-form parameters δ and γ .⁹ For purposes of comparison, we also consider a naïve Bayesian estimator given by the Ridge Regression estimator from (8). This approximation treats the error variance σ_ε^2 as a known constant and imposes the prior $\beta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\beta)$, where $\lambda = \tau_\beta \sigma_\varepsilon^2$

Equation 19 describes the Bayesian *model* we study. We now describe the assumptions that we place on the true data-generating process. For ease of notation, we state these in terms of the reduced form model. Combining these assumptions with (4)–(5) implies closely related assumptions for the structural model. The causal structure, in turn, gives a meaningful causal interpretation to $\alpha = \sigma_{UV}/\sigma_V^2$. We first assume that the researcher observes a sample of n iid observations from the reduced form model.

Assumption 2 (Random Sampling). *The random variables $\{(X_i, E_i)\}_{i=1}^n$ are iid across i and $W_i = (Y_i, D_i)$ is generated according to (18) with true parameters B^* and Σ^* .*

While Assumption 2 imposes normality of the errors E_i , as we explain further below this assumption turns out to be innocuous. Our next assumption concerns the true reduced form parameters Σ^* and (δ^*, γ^*) . As mentioned above, we consider an asymptotic sequence in which the number of control regressors, p , grows with sample size. This implies that the dimensions of δ and γ likewise grow. To accommodate this, we consider δ^* and γ^* to be square-summable infinite *sequences* of parameters, i.e. elements of the sequence space $\ell^2(\mathbb{N})$. In finite samples only the first p elements of each are “active”, but as we obtain more control regressors, the number of regression coefficients increases to match. Our assumption ensures that these coefficients do not “explode” as the number of regressors increases. Intuitively, each additional control matters “less on average” the more that we include in our model. Because the dimension of Σ^* does not change with sample size, we need only assume that this variance-covariance matrix is well-behaved.

Assumption 3 (True Reduced Form Parameters).

- (i) $\Sigma^* \equiv \text{Var}(E_i)$ is finite and positive definite.

⁸See Zellner (1971) section 8.5 for a textbook treatment.

⁹As we explain further below, our use of a normal likelihood for E_i is innocuous.

(ii) $\|\delta^*\|_2$ and $\|\gamma^*\|_2$ are bounded as $p \rightarrow \infty$

Our next assumption concerns the control regressors X_i . For simplicity, and without loss of generality, we assume that X_i has been centered around its mean. To make it easier to compare our BDML procedure against its frequentist cousin, we further assume that there are more observations than controls: $p < n$. The remaining conditions we impose on X_i restrict the variances and covariances of the control regressors as p grows. To state them, let $\Sigma_X \equiv \text{Var}(X_i)$ and $\lambda_j(\Sigma_X)$ be the j -th eigenvalue of Σ_X , where these values are arranged in descending order. Further let $F_p(t) \equiv \frac{1}{p} \sum_{j=1}^p \mathbb{1}\{\lambda_j(\Sigma_X) \leq t\}$ be the empirical spectral distribution of Σ_X .

Assumption 4 (Control Regressors).

(i) $\mathbb{E}(X_i) = 0$, without loss of generality, and $p < n$.

(ii) $\Sigma_X \equiv \text{Var}(X_i)$ is finite and positive definite for any p

(iii) The empirical spectral distribution $F_p(t) \equiv \frac{1}{p} \sum_{j=1}^p \mathbb{1}\{\lambda_j(\Sigma_X) \leq t\}$ converges to a limit distribution $F(t)$ supported on $(0, \infty)$ for all continuity points of F as $p \rightarrow \infty$.

(iv) $\frac{1}{p} \sum_{j=1}^p [\lambda_j(\Sigma_X)]^m \rightarrow \mu^{(m)} < \infty$ as $p \rightarrow \infty$ for $m = -1, 2, 2$ and $2 + \eta$ for some $\eta > 0$.

Parts (iii) and (iv) of **Assumption 4** ensure that the spectrum of Σ_X remains “well-behaved” as the dimension of X_i grows. The final ingredient of our asymptotic framework is an assumption about the relative rates of n , p , and the prior precisions. To state it we use the notation $a_n \asymp b_n$ to indicate that two sequences are “of the same order”, i.e. that $a_n = O(b_n)$ and $b_n = O(a_n)$.

Assumption 5 (Rate Restrictions).

(i) $p = o(n)$

(ii) $\tau_\delta, \tau_\gamma, \lambda = o(n)$

(iii) $\tau_\delta, \tau_\gamma, \lambda \asymp p$

Part (i) of **Assumption 5** states that the sample size dominates the number of controls and is the asymptotic equivalent of $p < n$ from **Assumption 4** (i) above.¹⁰ Since $p \rightarrow \infty$, this means that there are many controls but not *so many* that they overwhelm the available data. Part (ii) of **Assumption 5** states that the sample size dominates each of the prior

¹⁰Formally **Assumption 5** (i) is redundant given (ii) and (iii), but we state it explicitly to emphasize a key feature of our asymptotic framework.

precisions. Intuitively, this means that the priors are only “weakly informative” in the sense that they will be dominated by the data in large samples. Along this asymptotic sequence, the shrinkage bias of the Ridge estimator from (8) vanishes in the limit. An $o(n)$ rate for the prior precisions also accords with the typical assumption from the frequentist literature on double machine learning. Finally, part (iii) asserts that the prior precisions are of the same order as the number of controls. This means that our prior imposes *more shrinkage* when there are more controls. Taken together, assumptions [Assumption 2](#)–[Assumption 5](#) implied that, under the prior from (19), the implied R-squared values from the regressions of D on X and Y on X remain well-behaved in the limit: they do not diverge to zero or one. See [Proposition A.1](#) in the Appendix for details.

5.2 Prior Properties

Our BDML approach places priors on the reduced form coefficients (Σ, B) . We now consider the prior that (19) *implies* for other quantities of interest. We begin with the causal effect of interest, α . Partition Σ into blocks Σ_{11}, Σ_{12} and Σ_{22} and likewise partition Σ_0 into $\Sigma_{0,11}$ and so on. Then $\alpha = \Sigma_{12}/\Sigma_{22}$. Under (19), our prior for α follows a location-scale t-distribution.

Lemma 1 (Induced prior on α). *Under (18) and (19), the induced marginal prior on α is a location-scale t-distribution, in other words*

$$\alpha \sim \frac{\Sigma_{0,12}}{\Sigma_{0,22}} + \frac{|\Sigma_0|^{1/2}}{\sqrt{\nu_0} \Sigma_{0,22}} t_{\nu_0},$$

where t_{ν_0} denotes a Student-t random variable with ν_0 degrees of freedom.

If the prior scale matrix Σ_0 is diagonal, as is typical in practice, then $\Sigma_{0,12}/\Sigma_{0,22} = 0$ so the implied prior for α is symmetric around zero. Choosing a small value for ν_0 gives this prior heavy tails, making it compatible with both very large positive and negative values for the effect of interest. This is reassuring: although we have placed a prior on the reduced form rather than the structural model, we can easily avoid imposing unintentionally strong prior beliefs on α by making Σ_0 diagonal.

We now consider the implied prior for a different but equally important quantity, namely the *selection bias*. This idea is most easily explained in the case where D_i is binary. For the general version, see [Propositions A.2–A.3](#) in the Appendix. Following the usual convention in applied microeconomics, we define the selection bias as the difference between the quantity identified by a simple comparison of mean outcomes, treated minus untreated, and the true

causal effect. Under (4), this becomes

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

As mentioned in section 3 above, Linero (2023) shows that any model that factorizes as in (11) and in which β and γ are *a priori* independent implies a highly dogmatic prior for the extent of selection bias when p is large. This is another lens through which to view RIC: it arises from “accidentally” imposing a strong prior belief that there is little or no selection bias. The following result shows that the naïve estimator suffers from this problem whereas our BDML estimator does not.

Proposition 2 (Selection bias: naïve estimator vs BDML). *Suppose that Assumptions 2–5 hold. Then, as $n, p \rightarrow \infty$ the naïve (Ridge) prior implies*

$$SB_{naive} = \frac{\gamma' \Sigma_X \beta}{\Sigma_{22} + \gamma' \Sigma_X \gamma} \xrightarrow{p} 0,$$

whereas the prior from (19) implies

$$SB_{BDML} \xrightarrow{p} \frac{\Sigma_{12}}{\Sigma_{22} + \gamma' \Sigma_X \gamma}.$$

When n and p are large, the naïve approach in effect *assumes away selection bias*: it implies a prior that is concentrated in a small neighborhood around zero. But if there is no selection bias, then there is no need to adjust for X_i in the first place: it is as if D_i had been randomly assigned! Intuitively, in high-dimensional spaces randomly chosen vectors tend to be nearly orthogonal to each other. Since Σ_X is positive definite, the numerator $\gamma' \Sigma_X \beta$ from SB_{naive} converges in probability to zero. In contrast, the numerator of SB_{BDML} is Σ_{12} , which has a proper prior that is unaffected by the sample size or the number of control regressors. This is a clear advantage of working with the reduced form: stating a prior in which δ and γ are independent is innocuous because the implied prior for δ and $\beta = \delta - \alpha\gamma$ are *not independent*.

5.3 Posterior Properties

Let $\beta^* = \delta^* - \alpha^* \gamma^*$. We have the following result on the posterior mean of α under the naïve estimator.

Proposition 3 (Posterior mean: naïve estimator). *Assume that the model and prior are as*

above. Suppose that Assumptions 2–4 hold. Then, the posterior mean of α is given by

$$\hat{\alpha}_{naive} = \alpha^* + \frac{\lambda \gamma^{*'} \beta^*}{n \Sigma_{22}^*} + o\left(\frac{p}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right),$$

where the $O_p(1/\sqrt{n})$ term is mean zero. Therefore, the naive estimator is consistent, but not \sqrt{n} -consistent for $\lim_{n \rightarrow \infty} p^2/n = \infty$ and $\gamma^{*'} \beta^* \neq 0$.

Note that a similar (in)consistency argument can be established for the cases with $\lambda, p \asymp n$, as (A.4) still holds.

Conditional posterior Note that by Assumption 3, the true Σ^* is finite and positive definite, so Σ^{-1} is also finite and positive definite in the limit, almost surely in the posterior. Also by Assumption 4, Σ_X is finite with a positive proportion of positive eigenvalues. If $X'X$ does not have full rank, its inverse is given by the Moore-Penrose pseudoinverse.

The conditional posterior of B given Σ is

$$p(\text{vec}(B) \mid X, W, \Sigma) \propto \exp\left(-\frac{1}{2} \text{vec}(B - B_n)' V_n^{-1} \text{vec}(B - B_n)\right),$$

where

$$\begin{aligned} V_n &= [\Sigma^{-1} \otimes X'X + V_0^{-1}]^{-1}, \\ \text{vec}(B_n) &= V_n \left[(\Sigma^{-1} \otimes X'X) \text{vec}(\hat{B}) + V_0^{-1} \text{vec}(B_0) \right], \end{aligned}$$

and $\hat{B} = [\hat{\delta}, \hat{\gamma}] = (X'X)^{-1} X'W$ is the OLS estimator of B . And the conditional posterior of Σ given B is

$$\Sigma \mid X, W, B \sim \text{Inverse-Wishart}(\nu_n, \Sigma_n),$$

where

$$\begin{aligned} \nu_n &= \nu_0 + n, \\ \Sigma_n &= \Sigma_0 + (W - XB)'(W - XB). \end{aligned}$$

Below, for notation simplicity, let $B_0 = 0$.

Remark 1 (Posterior: BDML, finite sample). We cannot derive the exact posterior distribution of α in the finite sample case, as there is no closed form for the marginal posterior distribution of Σ . However, we can derive the posterior of α in the limit.

Proposition 4 (Marginal posterior distribution of Σ). *Assume that the model and prior are as above. Suppose that Assumptions 2–5 hold. We can approximate the marginal posterior of Σ as*

$$p(\Sigma|X, W) \propto |\Sigma|^{-(\tilde{\nu}_n+3)/2} \exp\left(-\frac{1}{2}\text{tr}(\tilde{\Sigma}_n \Sigma^{-1})\right) \exp\left(\frac{1}{2}\text{tr}(\tilde{C}_n \Sigma)\right) (1 + o_p(1)),$$

where

$$\begin{aligned} \tilde{\nu}_n &= \nu_0 + n - p, \\ \tilde{\Sigma}_n &= \Sigma_0 + (W - X\hat{B})'(W - X\hat{B}), \\ \tilde{C}_n &= \begin{bmatrix} \tau_\delta^2 \hat{\delta}'(X'X)^{-1} \hat{\delta} - \tau_\delta \text{tr}((X'X)^{-1}) & \tau_\delta \tau_\gamma \hat{\delta}'(X'X)^{-1} \hat{\gamma} \\ \tau_\delta \tau_\gamma \hat{\gamma}'(X'X)^{-1} \hat{\delta} & \tau_\gamma^2 \hat{\gamma}'(X'X)^{-1} \hat{\gamma} - \tau_\gamma \text{tr}((X'X)^{-1}) \end{bmatrix}. \end{aligned}$$

Remark 2 (Intuition on the posterior of Σ). The $\exp\left(\frac{1}{2}\text{tr}(\tilde{C}_n \Sigma)\right)$ term introduces exponential tilting to the Inverse-Wishart distribution. To visualize this effect, we could think of the tilting as “pulling” the distribution of Σ in the direction of matrices that have a structure similar to \tilde{C}_n (or $-\tilde{C}_n^{-1}$), if \tilde{C}_n is positive (or negative) definite.¹¹

As $n \rightarrow \infty$, $\tau_\delta, \tau_\gamma = o(n)$, $\tilde{C}_n = o(n)$, and $\tilde{\Sigma}_n = O(n)$ so the IW part would dominate the tilting part.

Proposition 5 (Posterior mean: BDML). *Assume that the model and prior are as above. Suppose that Assumptions 2–5 hold. The posterior mean of $\alpha|X, W$ under the BDML estimator is given by*

$$\begin{aligned} \hat{\alpha}_{BDML} &= \mathbb{E}[\alpha|X, W] \\ &= \alpha^* + \frac{\alpha^*}{n} \left(\frac{\Sigma_{0,12} + (\Sigma^* C^* \Sigma^*)_{12}}{\Sigma_{12}^*} - \frac{\Sigma_{0,22} + (\Sigma^* C^* \Sigma^*)_{22}}{\Sigma_{22}^*} \right) \\ &\quad + o_p\left(\frac{1}{n}\right) + o_p\left(\left(\frac{p}{n}\right)^2\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

where

$$C^* = \frac{1}{n} \begin{bmatrix} \tau_\delta^2 \delta^{*'} \Sigma_X^{-1} \delta^* - \tau_\delta \text{tr}(\Sigma_X^{-1}) & \tau_\delta \tau_\gamma \delta^{*'} \Sigma_X^{-1} \gamma^* \\ \tau_\delta \tau_\gamma \gamma^{*'} \Sigma_X^{-1} \delta^* & \tau_\gamma^2 \gamma^{*'} \Sigma_X^{-1} \gamma^* - \tau_\gamma \text{tr}(\Sigma_X^{-1}) \end{bmatrix},$$

and the $O_p\left(\frac{1}{\sqrt{n}}\right)$ term is mean zero. Therefore, $\hat{\alpha}_{BDML}$ is consistent. Moreover, if $p = o(n^{3/4})$, $\hat{\alpha}_{BDML}$ is \sqrt{n} -consistent.

¹¹If \tilde{C}_n is negative definite, then X follows a multivariate Generalized Inverse Gaussian distribution.

Remark 3 (BDML vs. naïve estimator). Compared with Proposition 3 for the naïve estimator, we see that when p is between the order of $n^{1/2}$ and $n^{3/4}$, the BDML estimator is \sqrt{n} -consistent, while the naïve estimator is not; this is in line with the argument for FDML in Chernozhukov et al. (2018, EctJ). In other words, BDML and FDML have bias of order p^2/n^2 compared to p/n for Naïve.

Remark 4 (BDML vs. FDML). For the frequentist DML based on ridge, for the first stage with $\tau_\delta, \tau_\gamma = o(n)$,

$$\text{vec}(\hat{B}_{\text{FDML}} - \hat{B}) = -V_{n,\text{FDML}}V_0^{-1}\text{vec}(\hat{B}),$$

where

$$V_{n,\text{FDML}} = \begin{bmatrix} [(X'X)^{-1} + \tau_\delta I]^{-1} & 0 \\ 0 & [(X'X)^{-1} + \tau_\gamma I]^{-1} \end{bmatrix} = I \otimes (X'X)^{-1}(1 + o_p(1)).$$

Then,

$$\hat{B}_{\text{FDML}} - \hat{B} = -(X'X)^{-1}[\tau_\delta \hat{\delta}, \tau_\gamma \hat{\gamma}](1 + o_p(1)).$$

The sample covariance matrix of the residues after the first stage is

$$\begin{aligned} (W - X\hat{B}_{\text{FDML}})'(W - X\hat{B}_{\text{FDML}})/n &= \bar{\Sigma} + (\hat{B}_{\text{FDML}} - \hat{B})'(X'X)(\hat{B}_{\text{FDML}} - \hat{B}) \\ &= \bar{\Sigma} + \tilde{C}_{n,2}/n(1 + o_p(1)), \end{aligned}$$

where

$$\begin{aligned} \bar{\Sigma} &= (W - X\hat{B})'(W - X\hat{B}), \\ \tilde{C}_{n,2} &= \begin{bmatrix} \tau_\delta^2 \hat{\delta}'(X'X)^{-1} \hat{\delta} & \tau_\delta \tau_\gamma \hat{\delta}'(X'X)^{-1} \hat{\gamma} \\ \tau_\delta \tau_\gamma \hat{\gamma}'(X'X)^{-1} \hat{\delta} & \tau_\gamma^2 \hat{\gamma}'(X'X)^{-1} \hat{\gamma} \end{bmatrix}, \end{aligned}$$

Then,

$$\begin{aligned} \hat{\alpha}_{\text{FDML}} &= \frac{\bar{\Sigma}_{12} + \tilde{C}_{n,2,12}/n(1 + o_p(1))}{\bar{\Sigma}_{22} + \tilde{C}_{n,2,22}/n(1 + o_p(1))} \\ &= \alpha^* + \frac{\alpha^*}{n} \left(\frac{(C_2^*)_{12}}{\Sigma_{12}^*} - \frac{(C_2^*)_{22}}{\Sigma_{22}^*} \right) + o_p\left(\frac{1}{n}\right) + o_p\left(\left(\frac{p}{n}\right)^2\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

where $\tilde{C}_{n,2} = C_2^*(1 + o_p(1))$ with

$$C_2^* = \frac{1}{n} \begin{bmatrix} \tau_\delta^2 \delta^{*\prime} \Sigma_X^{-1} \delta^* & \tau_\delta \tau_\gamma \delta^{*\prime} \Sigma_X^{-1} \gamma^* \\ \tau_\delta \tau_\gamma \gamma^{*\prime} \Sigma_X^{-1} \delta^* & \tau_\gamma^2 \gamma^{*\prime} \Sigma_X^{-1} \gamma^* \end{bmatrix},$$

and the $O_p\left(\frac{1}{\sqrt{n}}\right)$ term is mean zero.

Comparing with the BDML case in Proposition 5, the formula takes a similar form. Note that $C_2^* = O(p^2/n)$, so the FDML estimator is also consistent, and \sqrt{n} -consistent if $p = o(n^{3/4})$. The higher order differences being

- BDML has an extra Σ_0/n term from the prior of the error covariance.
- FDML incorporates C_2^* , which is the second component of the tilting term in the BDML case in (A.9). BDML has an extra first component of the tilting term in (A.7), coming from the first order approximation of $|V_n|^{1/2}$.
- In addition, the last term in BDML is based on $\Sigma^* C^* \Sigma^*$. While the variance terms in Σ^* can be absorbed into τ_δ and τ_γ , the correlation is not captured in FDML.

For the bias-variance trade-off, note that the variance of the posterior mean is not the same as the posterior variance. As we only have posterior variance for the BDML estimator, below we focus on deriving and comparing the variance of the posterior mean.

Proposition 6 (Variances of $\hat{\alpha}$). *Assume that the model and prior are as above. Suppose that Assumptions 2–5 hold. The asymptotic variance of $\hat{\alpha}_{naive}$, $\hat{\alpha}_{BDML}$ and $\hat{\alpha}_{FDML}$ are the same, and are given by*

$$\frac{\Sigma_{11}^* \Sigma_{22}^* - (\Sigma_{12}^*)^2}{(\Sigma_{22}^*)^2}.$$

Remark 5 (Comparison of variances). The variances of $\hat{\alpha}_{BDML}$, $\hat{\alpha}_{FDML}$, and $\hat{\alpha}_{naive}$ are the same. This is because the variances of the error terms are the same, which is the dominating term in the variance of $\hat{\alpha}$. Therefore, to minimize MSE, we would focus on the bias of the estimators, where the BDML and FDML estimators are \sqrt{n} -consistent in more general cases.

5.4 Bernstein–von Mises theorem

Under the assumption that σ_ε^{2*} and σ_V^{2*} are known, we can resort to Walker (2024) by verifying his assumptions in our high dimensional regression setup.

Remark 6 (Advantages of BDML: BvM). The procedure is semiparametrically efficient in that it is adaptive, incurring no cost from not knowing the nuisance parameters. It avoids the

need for prior invariance and does not require strict smoothness conditions on the propensity score.

In addition, the method is robust due to its use of a Gaussian profile likelihood; it only requires the mean restriction—similar to ordinary least squares. This property ensures that Bayesian credible intervals remain efficient even under model misspecification.

Let P^* denote the probability measure based on the true DGP, and $\|\cdot\|_{TV}$ denote the total variation distance.

Proposition 7 (Bernstein–von Mises theorem: known σ_ε^{2*} and σ_V^{2*}). *Assume that the model and prior are as above, except that σ_ε^{2*} and σ_V^{2*} are known, the prior for α is continuous and positive over a neighborhood of α^* , and α^* is in the interior of its support. Suppose that Assumptions 2–5 hold. Further suppose that X_i is sub-Gaussian, and the eigenvalues of Σ_X is uniformly bounded by $[\underline{\lambda}_X, \bar{\lambda}_X]$ with $0 < \underline{\lambda}_X \leq \bar{\lambda}_X < \infty$, $p = o(\sqrt{n})$, and $\log n = o(p^{1/2})$. Then, as $n \rightarrow \infty$,*

$$\left\| P(\alpha \in \cdot | X, W) - \text{Normal}\left(\alpha^* + \frac{\tilde{\Delta}_n^*}{\sqrt{n}}, \frac{1}{n} I_n(\gamma^*)^{-1}\right) \right\|_{TV} \xrightarrow{P^*} 0,$$

where $\tilde{I}_n(\gamma) = \frac{1}{n} \sum_{i=1}^n \frac{(D_i - X_i' \gamma)^2}{\sigma_\varepsilon^{2*}}$ and $\tilde{\Delta}_n^* = \tilde{I}_n(\gamma^*)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{[Y_i - \alpha^*(D_i - X_i' \gamma^*) - X_i' \delta^*](D_i - X_i' \gamma^*)}{\sigma_\varepsilon^{2*}}$.

Remark 7 (BvM: empirical L_2 -norm). Here I consider the empirical L_2 -norm in the definition of the shrinking neighborhood of m_0 to accommodate our linear setup with unbounded X_i . Specifically, for generic functions f and g , the empirical L_2 inner product is defined as $\langle f, g \rangle_{n,2} = \frac{1}{n} \sum_{i=1}^n f(X_i)g(X_i)$, and the empirical L_2 -norm is $\|f\|_n = \sqrt{\langle f, f \rangle_{n,2}}$. Then, the neighborhood of m_0 based on the empirical L_2 -norm is defined as

$$B_{\mathcal{M},n,2}(m^*, \nu) = \{m \in \mathcal{M} : \|m - m^*\|_{n,2} \leq \nu\},$$

which is different from $B_{\mathcal{M},\infty}(m^*, \nu)$ in Walker (2024) for the sup-norm. All proofs in Walker (2024) can be adapted to the empirical L_2 -norm: see also the last sentence on page 8 in Walker (2024).

Remark 8 (BvM: alternative conditions). One stronger alternative condition is that the eigenvalues of $X'X/n$ are bounded away from zero and infinity, so we don't need matrix Bernstein and $\mathcal{K}_n = \mathbb{R}^p$ below.

Another alternative condition is that X_i is bounded, i.e., for some constant $M_X > 0$, $\max_{1 \leq j \leq p} |X_{i,j}| \leq M_X$. This is necessary for the sup-norm case, but incurs another \sqrt{p} term in ν_n , and only allows for $p = o(n^{1/4})$.

For a generic matrix S , define the Orlicz ψ_1 norm as

$$\|S\|_{\psi_1} = \inf \left\{ s > 0 : \mathbb{E} \exp \left(\frac{\|S\|}{s} \right) \leq 2 \right\},$$

where $\|\cdot\|$ is the spectral norm. Note that $\|S\|_{\psi_1}$ is finite if and only if S is sub-exponential. Then, the matrix Bernstein inequality is as follows.

Lemma 2 (Matrix Bernstein). *Let S_1, \dots, S_n be i.i.d., mean-zero, symmetric random matrices in $\mathbb{R}^{p \times p}$, satisfying that $\|S_i\|_{\psi_1} \leq M_S$ for all i . Then, for any $\epsilon \geq 0$,*

$$P \left\{ \left\| \sum_{i=1}^n S_i \right\| \geq \epsilon \right\} \leq 2p \cdot \exp \left(-c \cdot \min \left(\frac{\epsilon^2}{nM_S^2}, \frac{\epsilon}{M_S} \right) \right),$$

where $c > 0$ is an absolute constant.

Corollary 1 (Eigenvalue bounds). *Suppose that X_i is sub-gaussian and i.i.d., and the eigenvalues of Σ_X is uniformly bounded by $[\underline{\lambda}_X, \bar{\lambda}_X]$ with $0 < \underline{\lambda}_X \leq \bar{\lambda}_X < \infty$. Define event*

$$\mathcal{K}_n = \left\{ X \in \mathbb{R}^{n \times p} : \lambda_{\min} \left(\frac{X'X}{n} \right) \geq \frac{\underline{\lambda}_X}{2} \text{ and } \lambda_{\max} \left(\frac{X'X}{n} \right) \leq 2\bar{\lambda}_X \right\}.$$

Then, $P(\mathcal{K}_n) \xrightarrow{p} 1$ and $P(\mathcal{K}_n^c) = O(p^{-1})$, as $n \rightarrow \infty$.

6 Simulation Study

We now compare our proposed BDML to a number of alternatives from the literature in a simulation experiment adapted from [Linero \(2023\)](#). In each replication we generate

$$\begin{aligned} \{X_i\}_{i=1}^n &\sim \text{iid Normal}_p(0, \mathbb{I}_p) \\ \{(\varepsilon_i, V_i)'\}_{i=1}^n \mid X &\sim \text{iid Normal} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & 1 \end{bmatrix} \right) \\ \beta \mid (X, \varepsilon, V) &\sim \text{Normal}_p(\mu_\beta, \sigma_\beta^2 \mathbb{I}_p). \end{aligned} \tag{20}$$

We then construct (D_i, Y_i) according to (4)–(5) above. This design generates new control regressors X and a new parameter vector β in each replication but holds γ and α constant across replications. Following the “fixed” design from [Linero \(2023\)](#), we vary σ_ε over a grid of values from 1 to 4 and set

$$\alpha = 2, \quad \gamma = \iota_p / \sqrt{p}, \quad \mu_\beta = -\gamma/2, \quad \sigma_\beta^2 = 1/p, \quad n = 200, \quad p = 100 \tag{21}$$

where ι_p denotes a p -vector of ones.

We consider three groups of estimators. The first group includes two alternative versions of the procedure from [Algorithm 1](#), corresponding to alternative prior structures for (γ, δ) . Both of our BDML approaches place a flat prior on α and construct a weakly informative prior for Σ by placing an LKJ(4) prior on the correlation matrix for (U, V) and independent Cauchy(0, 2.5) priors on both σ_U and σ_V . The first of our BDML approaches, **BDML-Basic**, places independent Normal(0, 5²) on the elements of δ and γ . The second, **BDML-Hier**, allows different standard deviations in the normal shrinkage priors for δ and γ . We achieve this via a hierarchical prior that places independent Inverse-Gamma(2, 2) hyper-priors on σ_δ^2 and σ_γ^2 .¹² We sample from the posteriors for both BDML approaches via Hamiltonian Monte Carlo (HMC), using the no-U-turn (NUTS) sampler as implemented in the Stan probabilistic programming language.

The second group of estimators we consider contains Bayesian competitors to our BDML approach. This group includes three estimators. The first two, **HCPH** and **Linero**, correspond to the proposals from [Hahn et al. \(2018\)](#) and [Linero \(2023\)](#) detailed in Algorithms 2–3 above and rely on a first-stage regression of D on X . The third, **Naïve**, estimates a single-equation linear regression model including (4) only and is a fully-Bayesian variant of the ridge regression approach described in [section 3](#) above. We carry out posterior sampling for these three approaches exactly as in [Linero \(2023\)](#). In particular, we rely on the Gibbs sampler implemented in the BLR() function from the R package BLR for Bayesian linear regression. Again following [Linero \(2023\)](#) we place a flat prior on α across all three approaches and on κ from [Algorithm 3](#) and a Jeffreys prior on all error variances. We likewise place default “ridge priors” on the coefficients on X across all regressions in both stages: independent Normal(0, σ^2) priors on each coefficient and a Jeffreys prior on σ^2 .¹³

The third and final group of estimators contains frequentist DML competitors to our BDML approach. Both of the approaches in this group estimate α via (16) where the “first-step” estimators $\hat{\gamma}$ and $\hat{\delta}$ are constructed via ridge regression.¹⁴ We carry out ridge regression by computing the posterior mean from draws produced using BLR() with the default parameter values, as explained above. We consider two versions of this FDML approach: “full-sample” **FDML-Full** and “split-sample” **FDML-Split**. **FDML-Full** uses the full dataset for both estimation stages whereas **FDML-Split** randomly splits the sample partitions the sample into two halves, the first of which is used to estimate $(\hat{\delta}_{\text{ML}}, \hat{\gamma}_{\text{ML}})$ and

¹²This is equivalent to placing independent Student-t(4) distributions on each of the coefficients δ_j and γ_j .

¹³The resulting prior for the coefficients on X is improper and sharply peaked at zero.

¹⁴We use ridge rather than LASSO for these preliminary DML regressions because recent work by [Kolesár et al. \(2025\)](#) and [Shen and Xiu \(2024\)](#) suggests that the “double-LASSO” approach ([Belloni et al., 2014](#)) perform poorly due to its fragility in variable selection.

the second of which is used to residualize (Y, D) and estimate α .

For each of the seven estimators of α described above, we compute the root mean-squared error (RMSE), along with the frequentist coverage probability and average width of a corresponding (nominal) 95% confidence interval. For the five Bayesian procedures, interval estimates are constructed as equal-tailed 95% posterior credible intervals. For the two frequentist DML procedures, confidence intervals are constructed based on the usual OLS interval for the regression of the residualized Y on the residualized D , following the approach suggested by [Belloni et al. \(2014\)](#) and often used in empirical research. Results based on 200 simulation replications from the data-generating process given in (20)–(21) appear in [Table 1](#) and [Figure 1](#). Coverage probabilities are very poor for all methods except BDML-Basic, BDML-Hier, and Linero. All three of these approaches produce coverage probabilities close to the nominal 95% level, but BDML-Hier is closest. From panel 1b of [Figure 1](#), we also see that BDML-Hier produces the shortest average intervals among the three methods with approximately correct coverage. This holds across all values of σ_ε in our simulation. Turning our attention to RMSE, the three best-performing methods remain BDML-basic, BDML-Hier, and Linero. Once again from panel 1b of [Figure 1](#), we see that BDML-Hier performs best. The fact that BDML out-peforms Linero in this setting is particularly encouraging given that this simulation design was taken directly from [Linero \(2023\)](#) and favors [Algorithm 3](#).

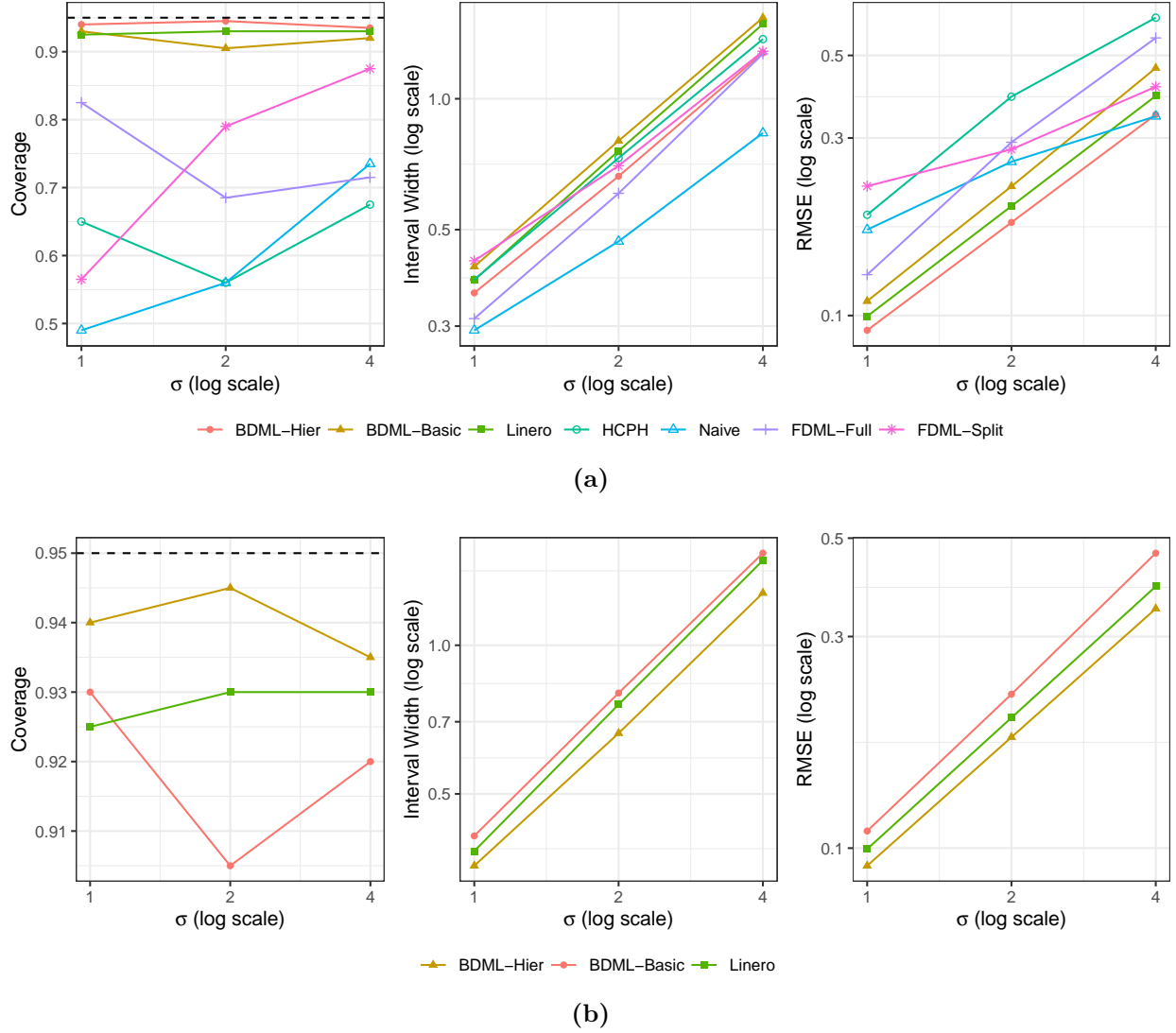


Figure 1: This figure presents results for the estimators described in [section 6](#) and simulation from (20)–(21) over a grid of values for σ_ε , the error variance in the structural equation given in (4). The panel labeled RMSE gives the root mean-squared error of the estimator for α under each method, while those labeled Coverage and Avg. Width give the coverage probability and average width of corresponding (nominal) 95% confidence intervals. Panel 1a plots results on a scale such that all seven approaches are visible; panel 1b zooms in to more clearly show the differences in performance between the three best estimators. Results are based on 200 simulation replications.

Method	σ_ε	n	p	Coverage	RMSE	Avg. Width
BDML-Hier	1	200	100	0.94	0.09	0.36
BDML-Hier	2	200	100	0.94	0.18	0.66
BDML-Hier	4	200	100	0.94	0.35	1.28
BDML-Basic	1	200	100	0.93	0.11	0.41
BDML-Basic	2	200	100	0.91	0.22	0.80
BDML-Basic	4	200	100	0.92	0.46	1.54
Linero	1	200	100	0.93	0.10	0.38
Linero	2	200	100	0.93	0.20	0.76
Linero	4	200	100	0.93	0.39	1.49
HCPH	1	200	100	0.65	0.19	0.38
HCPH	2	200	100	0.56	0.39	0.73
HCPH	4	200	100	0.68	0.63	1.37
Naïve	1	200	100	0.49	0.17	0.29
Naïve	2	200	100	0.56	0.26	0.47
Naïve	4	200	100	0.73	0.34	0.83
FDML-Full	1	200	100	0.82	0.13	0.31
FDML-Full	2	200	100	0.69	0.29	0.61
FDML-Full	4	200	100	0.71	0.56	1.27
FDML-Split	1	200	100	0.56	0.22	0.42
FDML-Split	2	200	100	0.79	0.28	0.70
FDML-Split	4	200	100	0.88	0.41	1.29

Table 1: This table presents results for the estimators described in [section 6](#) and simulation from [\(20\)](#)–[\(21\)](#) over a grid of values for σ_ε , the error variance in the structural equation given in [\(4\)](#). The column RMSE gives the root mean-squared error of the estimator for α under each method, while Coverage and Avg. Width give the coverage probability and average width of corresponding (nominal) 95% confidence intervals. Results are based on 200 simulation replications.

References

- Ahrens, A., Chernozhukov, V., Hansen, C., Kozbur, D., Schaffer, M., Wiemann, T., 2025. An Introduction to Double/Debiased Machine Learning. URL: <http://arxiv.org/abs/2504.08324>, doi:10.48550/arXiv.2504.08324 [econ].
- Alsan, M., Durvasula, M., Gupta, H., Schwartzstein, J., Williams, H., 2023. Representation and extrapolation: Evidence from clinical trials*. The Quarterly Journal of Economics 139, 575–635. URL: <https://doi.org/10.1093/qje/qjad036>, doi:10.1093/qje/qjad036, arXiv:<https://academic.oup.com/qje/article-pdf/139/1/575/57247273/qjad036.pdf>.
- Angrist, J.D., Frandsen, B., 2022. Machine Labor. Journal of Labor Economics 40, S97–S140. URL: <https://www.journals.uchicago.edu/doi/10.1086/717933>, doi:10.1086/717933.
- Antonelli, J., Papadogeorgou, G., Dominici, F., 2022. Causal inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. Biometrics 78, 100–114. URL: <https://academic.oup.com/biometrics/article/78/1/100-114/7459997>, doi:10.1111/biom.13417.
- Armand, A., Coutts, A., Vicente, P.C., Vilela, I., 2020. Does information break the political resource curse? experimental evidence from mozambique. American Economic Review 110, 3431–53. URL: <https://www.aeaweb.org/articles?id=10.1257/aer.20190842>, doi:10.1257/aer.20190842.
- Bach, P., Bach, P., De, U.H., Chernozhukov, V., 2024. Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study. Proceedings of Machine Learning Research 236, 1065–1117.
- Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on Treatment Effects after Selection among High-Dimensional Controls. The Review of Economic Studies 81, 608–650. URL: <https://academic.oup.com/restud/article-lookup/doi/10.1093/restud/rdt044>, doi:10.1093/restud/rdt044.
- Breunig, C., Liu, R., Yu, Z., 2024. Semiparametric Bayesian Difference-in-Differences URL: https://christophbreunig.com/wp-content/uploads/2024/12/bayes_did.pdf.
- Breunig, C., Liu, R., Yu, Z., 2025. Double Robust Bayesian Inference on Average Treatment Effects. Econometrica 93, 539–568. URL: <https://www.econometricsociety.org/doi/10.3982/ECTA21442>, doi:10.3982/ECTA21442.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21, C1–C68. URL: <https://academic.oup.com/ectj/article/21/1/C1/5056401>, doi:10.1111/ectj.12097.
- Chioda, L., Contreras-Loya, D., Gertler, P., Carney, D., 2021. Making Entrepreneurs: Returns to Training Youth in Hard Versus Soft Business Skills. Working Paper 28845.

- National Bureau of Economic Research. URL: <http://www.nber.org/papers/w28845>, doi:10.3386/w28845.
- Cotti, C., Courtemanche, C., Liang, Y., Maclean, J.C., Nesson, E., Sabia, J.J., 2025. The effect of e-cigarette flavor bans on tobacco use. *Journal of Health Economics* 102, 103013. URL: <https://www.sciencedirect.com/science/article/pii/S0167629625000487>, doi:<https://doi.org/10.1016/j.jhealeco.2025.103013>.
- Delfino, A., 2024. Breaking gender barriers: Experimental evidence on men in pink-collar jobs. *American Economic Review* 114, 1816–53. URL: <https://www.aeaweb.org/articles?id=10.1257/aer.20220582>, doi:10.1257/aer.20220582.
- Dube, A., Jacobs, J., Naidu, S., Suri, S., 2020. Monopsony in online labor markets. *American Economic Review: Insights* 2, 33–46. URL: <https://www.aeaweb.org/articles?id=10.1257/aeri.20180150>, doi:10.1257/aeri.20180150.
- Duflo, E., Dupas, P., Kremer, M., 2021. The Impact of Free Secondary Education: Experimental Evidence from Ghana. NBER Working Papers 28937. National Bureau of Economic Research, Inc. URL: <https://ideas.repec.org/p/nbr/nberwo/28937.html>, doi:None.
- Gordon, B.R., Moakler, R., Zettelmeyer, F., 2023. Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science* 42, 768–793. URL: <https://doi.org/10.1287/mksc.2022.1413>, doi:10.1287/mksc.2022.1413.
- Hahn, P.R., Carvalho, C.M., Puelz, D., He, J., 2018. Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis* 13. URL: <https://projecteuclid.org/journals/bayesian-analysis/volume-13/issue-1/Regularization-and-Confounding-in-Linear-Regression-for-Treatment-Effect-Estimation/10.1214/16-BA1044.full>, doi:10.1214/16-BA1044.
- Hahn, P.R., Murray, J.S., Carvalho, C.M., 2020. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis* 15. URL: <https://projecteuclid.org/journals/bayesian-analysis/volume-15/issue-3/Bayesian-Regression-Tree-Models-for-Causal-Inference--Regularization-Confounding/10.1214/19-BA1195.full>, doi:10.1214/19-BA1195.
- Heckman, J.J., Lopes, H.F., Piatek, R., 2014. Treatment Effects: A Bayesian Perspective. *Econometric Reviews* 33, 36–67. URL: <http://www.tandfonline.com/doi/abs/10.1080/07474938.2013.807103>, doi:10.1080/07474938.2013.807103.
- Hussam, R., Kelley, E.M., Lane, G., Zahra, F., 2022. The psychosocial value of employment: Evidence from a refugee camp. *American Economic Review* 112, 3694–3724. URL: <https://www.aeaweb.org/articles?id=10.1257/aer.20211616>, doi:10.1257/aer.20211616.
- Kolesár, M., Müller, U.K., Roelsgaard, S.T., 2025. The Fragility of Sparsity. URL: <http://arxiv.org/abs/2311.02299>, doi:10.48550/arXiv.2311.02299. arXiv:2311.02299 [econ].

- Li, F., Ding, P., Mealli, F., 2023. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 381, 20220153. URL: <https://royalsocietypublishing.org/doi/10.1098/rsta.2022.0153>, doi:10.1098/rsta.2022.0153.
- Linero, A.R., 2023. In Nonparametric and High-Dimensional Models, Bayesian Ignorability is an Informative Prior. *Journal of the American Statistical Association* , 1–14URL: <https://www.tandfonline.com/doi/full/10.1080/01621459.2023.2278202>, doi:10.1080/01621459.2023.2278202.
- Linero, A.R., Antonelli, J.L., 2023. The how and why of Bayesian nonparametric causal inference. *WIREs Computational Statistics* 15, e1583. URL: <https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.1583>, doi:10.1002/wics.1583.
- Luo, Y., Graham, D.J., McCoy, E.J., 2023. Semiparametric Bayesian doubly robust causal estimation. *Journal of Statistical Planning and Inference* 225, 171–187. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0378375822001124>, doi:10.1016/j.jspi.2022.12.005.
- Ray, K., Van Der Vaart, A., 2020. Semiparametric Bayesian causal inference. *The Annals of Statistics* 48. URL: <https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-5/Semiparametric-Bayesian-causal-inference/10.1214/19-AOS1919.full>, doi:10.1214/19-AOS1919.
- Saarela, O., Belzile, L.R., Stephens, D.A., 2016. A Bayesian view of doubly robust causal inference. *Biometrika* 103, 667–681. URL: <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asw025>, doi:10.1093/biomet/asw025.
- Shen, Z., Xiu, D., 2024. Can Machines Learn Weak Signals? Technical Report. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4748784.
- Sims, C.A., 2012. ROBINS-WASSERMAN, ROUND N .
- Walker, C.D., 2025. Parametrization, Prior Independence, and the Semiparametric Bernstein-von Mises Theorem for the Partially Linear Model. URL: <http://arxiv.org/abs/2306.03816>, doi:10.48550/arXiv.2306.03816. arXiv:2306.03816 [math].
- Wüthrich, K., Zhu, Y., 2023. Omitted Variable Bias of Lasso-Based Inference Methods: A Finite Sample Analysis. *Review of Economics and Statistics* 105, 982–997. URL: <https://direct.mit.edu/rest/article/105/4/982/107662/Omitted-Variable-Bias-of-Lasso-Based-Inference>, doi:10.1162/rest_a_01128.
- Yang, J.C., Chuang, H.C., Kuan, C.M., 2020. Double machine learning with gradient boosting and its application to the big n audit quality effect. *Journal of Econometrics* 216, 268–283. URL: <https://www.sciencedirect.com/science/article/pii/S0304407620300245>, doi:<https://doi.org/10.1016/j.jeconom.2020.01.018>. *annals* Issue in honor of George Tiao: Statistical Learning for Dependent Data.

- Zellner, A., 1971. An Introduction to Bayesian Inference in Econometrics. John Wiley & Sons, New York.
- Zhang, Y.D., Naughton, B.P., Bondell, H.D., Reich, B.J., 2022. Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior. *Journal of the American Statistical Association* 117, 862–874. URL: <https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1825449>, doi:10.1080/01621459.2020.1825449.
- Zigler, C.M., 2016. The Central Role of Bayes' Theorem for Joint Estimation of Causal Effects and Propensity Scores. *The American Statistician* 70, 47–54. URL: <https://www.tandfonline.com/doi/full/10.1080/00031305.2015.1111260>, doi:10.1080/00031305.2015.1111260.

A Proofs

The following lemma is a preliminary step in the proof of [Proposition 1](#).

Lemma A.1. *Define the shorthand*

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \equiv \begin{pmatrix} D'D & D'X \\ X'D & X'X + \lambda \mathbb{I}_p \end{pmatrix}, \quad M^{-1} \equiv \begin{pmatrix} M^{11} & M^{12} \\ M^{21} & M^{22} \end{pmatrix}$$

and let $\hat{\rho}' \equiv (D'D)^{-1}D'X$, $\hat{\xi} \equiv [\mathbb{I}_p - D(D'D)^{-1}D']X$ and $R = \hat{\xi}'\hat{\xi}$. Then we have

$$M^{12} = -\hat{\rho}'(R + \lambda I)^{-1} \quad (\text{A.1})$$

$$M^{11} = (D'D)^{-1} - M^{12}\hat{\rho} \quad (\text{A.2})$$

$$\begin{pmatrix} M^{11} & M^{12} \end{pmatrix} \begin{pmatrix} D'D & D'X \\ X'D & X'X \end{pmatrix} = \begin{bmatrix} 1 & (\hat{\rho}' + M^{12}R) \end{bmatrix}. \quad (\text{A.3})$$

Proof of Lemma A.1. By the partitioned matrix inverse formula,

$$M^{-1} = \begin{pmatrix} M^{11} & M^{12} \\ M^{21} & M^{22} \end{pmatrix} = \begin{pmatrix} M_{11}^{-1} + M_{11}^{-1}M_{12}S^{-1}M_{21}M_{11}^{-1} & -M_{11}^{-1}M_{12}S^{-1} \\ -S^{-1}M_{21}M_{11}^{-1} & S^{-1} \end{pmatrix}$$

where we define

$$S \equiv M_{22} - M_{21}M_{11}^{-1}M_{12}.$$

The preceding expression for M^{-1} holds so long as M_{11} and S are invertible. We see that $M_{11} \equiv D'D$ is invertible whenever there is variation in the treatment variable while $M_{22} \equiv X'X + \lambda \mathbb{I}_p$ is invertible when $\lambda > 0$ although possibly not when $\lambda = 0$, since p could be larger than the sample size. Now, defining $P_D \equiv D(D'D)^{-1}D'$, we have

$$\begin{aligned} S &= M_{22} - M_{21}M_{11}^{-1}M_{12} = (X'X + \lambda \mathbb{I}_p) - X'D(D'D)^{-1}D'X \\ &= X'(\mathbb{I}_p - P_D)X + \lambda \mathbb{I}_p \\ &= [(\mathbb{I}_p - P_D)X]'[(\mathbb{I}_p - P_D)X] + \lambda \mathbb{I}_p \\ &= \hat{\xi}'\hat{\xi} + \lambda \mathbb{I}_p \\ &= R + \lambda \mathbb{I}_p \end{aligned}$$

since P_D is a projection matrix and $(\mathbb{I}_p - P_D)$ is its orthogonal complement. Substituting this simplified expression for S into the expression for M^{12} from above gives

$$M^{12} = -M_{11}^{-1}M_{12}S^{-1} = -(D'D)^{-1}D'X(R + \lambda \mathbb{I}_p)^{-1} = -\hat{\rho}'(R + \lambda \mathbb{I}_p)^{-1}.$$

Substituting the expression for M^{12} from the partitioned matrix inverse formula into the expression for M^{11}

$$M^{11} = M_{11}^{-1} + M_{11}^{-1}M_{12}S^{-1}M_{21}M_{11}^{-1} = M_{11}^{-1} - M^{12}M_{21}M_{11}^{-1}.$$

It follows that M^{11} can be written as

$$\begin{aligned}
M^{11} &= M_{11}^{-1} - M^{12} M_{21} M_{11}^{-1} \\
&= (D'D)^{-1} + M^{12} X'D (D'D)^{-1} \\
&= (D'D)^{-1} + M^{12} [(D'D)^{-1} D'X]' \\
&= (D'D)^{-1} - M^{12} \hat{\rho}.
\end{aligned}$$

Now, using this more compact expression for M^{11} , we obtain

$$\begin{aligned}
\begin{bmatrix} M^{11} & M^{12} \end{bmatrix} \begin{bmatrix} D'D \\ X'D \end{bmatrix} &= [\{(D'D)^{-1} - M^{12} \hat{\rho}\} \quad M^{12}] \begin{bmatrix} D'D \\ X'D \end{bmatrix} \\
&= [(D'D)^{-1} - M^{12} \hat{\rho}] D'D + M^{12} X'D \\
&= 1 - M^{12} \hat{\rho} D'D + M^{12} X'D \\
&= 1 - M^{12} (X'D - \hat{\rho} D'D) \\
&= 1 - M^{12} \left[X'D - \{(D'D)^{-1} D'X\}' D'D \right] \\
&= 1 - M^{12} \left[X'D - X'D (D'D)^{-1} D'D \right] \\
&= 1 - M^{12} (X'D - X'D) \\
&= 1.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\begin{bmatrix} M^{11} & M^{12} \end{bmatrix} \begin{bmatrix} D'X \\ X'X \end{bmatrix} &= [\{(D'D)^{-1} - M^{12} \hat{\rho}\} \quad M^{12}] \begin{bmatrix} D'X \\ X'X \end{bmatrix} \\
&= [(D'D)^{-1} - M^{12} \hat{\rho}] D'X + M^{12} X'X \\
&= \hat{\rho}' - M^{12} \hat{\rho} D'X + M^{12} X'X \\
&= \hat{\rho}' + M^{12} (X'X - \hat{\rho} D'X) \\
&= \hat{\rho}' + M^{12} \left[X'X - X'D (D'D)^{-1} D'X \right] \\
&= \hat{\rho}' + M^{12} (X'X - X'P_D X) \\
&= \hat{\rho}' - \hat{\rho}' (R + \lambda \mathbb{I}_p)^{-1} R \\
&= \hat{\rho}' [\mathbb{I}_p - (R + \lambda \mathbb{I}_p)^{-1} R].
\end{aligned}$$

The result follows. □

Proof of Proposition 1. Using the shorthand defined in the statement of Lemma A.1,

$$\hat{\alpha}_\lambda = \begin{bmatrix} M^{11} & M^{12} \end{bmatrix} \begin{bmatrix} D' \\ X' \end{bmatrix} Y = \begin{bmatrix} M^{11} & M^{12} \end{bmatrix} \begin{bmatrix} D' \\ X' \end{bmatrix} \left(\begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \epsilon \right).$$

Since $\mathbb{E}[\varepsilon|X, D] = 0$ it follows that

$$\text{Bias}(\hat{\alpha}_\lambda|X, D) = \begin{bmatrix} M^{11} & M^{12} \end{bmatrix} \begin{bmatrix} D'D & D'X \\ X'D & X'X \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - \alpha.$$

Thus, substituting (A.1) and (A.3),

$$\text{Bias}(\hat{\alpha}_\lambda|X, D) = \begin{bmatrix} 1 & (\hat{\rho}' + M^{12}R) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - \alpha = \hat{\rho}' [\mathbb{I}_p - (R + \lambda\mathbb{I}_p)^{-1}R] \beta.$$

For the variance calculation, since $\text{Var}(\varepsilon|X, D) = \sigma^2\mathbb{I}_p$, we obtain

$$\begin{aligned} \text{Var}(\hat{\alpha}_\lambda|X, D) &= \text{Var} \left\{ (M^{11} \ M^{12}) \begin{pmatrix} D' \\ X' \end{pmatrix} \left[(D \ X) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \varepsilon \right] \middle| X, D \right\} \\ &= \sigma^2 \begin{bmatrix} M^{11} & M^{12} \end{bmatrix} \begin{bmatrix} D'D & D'X \\ X'D & X'X \end{bmatrix} \begin{bmatrix} M^{11} & M^{12} \end{bmatrix}' \\ &= \sigma^2 \begin{bmatrix} 1 & (\hat{\rho}' + M^{12}\hat{\xi}'\hat{\xi}) \end{bmatrix} \begin{bmatrix} (M^{11})' \\ (M^{12})' \end{bmatrix} \\ &= \sigma^2 \left[\{(D'D)^{-1} - M^{12}\hat{\rho}\}' + (\hat{\rho}' + M^{12}R)(M^{12})' \right] \\ &= \sigma^2 \left[(D'D)^{-1} + M^{12}R(M^{12})' \right] \\ &= \sigma^2 \left[(D'D)^{-1} + \hat{\rho}'(R + \lambda\mathbb{I}_p)^{-1}R(R + \lambda\mathbb{I}_p)^{-1}\hat{\rho} \right] \end{aligned}$$

by substituting Lemma A.1, since $R \equiv \hat{\xi}'\hat{\xi}$ is symmetric. \square

Proposition A.1 (Signal-to-noise ratio and R^2 of high-dimensional regression). *Suppose that Assumptions 2–5 hold. As $p \rightarrow \infty$, the signal-to-noise ratios of the Y and D regressions are*

$$\text{SNR}_Y \rightarrow \tau_\delta^{-1}p \frac{\mu^{(1)}}{\Sigma_{11}}, \quad \text{and} \quad \text{SNR}_D \rightarrow \tau_\gamma^{-1}p \frac{\mu^{(1)}}{\Sigma_{11}}.$$

And the R^2 of the Y and D regressions are

$$R_Y^2 \rightarrow \frac{\tau_\delta^{-1}p}{\tau_\delta^{-1}p + \Sigma_{11}/\mu^{(1)}} \quad \text{and} \quad R_D^2 \rightarrow \frac{\tau_\gamma^{-1}p}{\tau_\gamma^{-1}p + \Sigma_{22}/\mu^{(1)}}.$$

Proof of Proposition A.1. As the two regressions are similar, let us focus on the Y regression. Note that $\mathbb{E}[X_i] = 0$. The *prior-induced* variance of the signal (per observation) is

$$\text{Var}(X_i'\delta) = \mathbb{E}[X_i'\text{Var}(\delta)X_i] = \tau_\delta^{-1}\text{tr}(\Sigma_X) \rightarrow \tau_\delta^{-1}p\mu^{(1)}.$$

The per-observation noise variance is simply the error variance for the Y equation, which is Σ_{11} . The signal-noise ratio is then

$$\text{SNR}_Y \rightarrow \tau_\delta^{-1}p \frac{\mu^{(1)}}{\Sigma_{11}}.$$

Then R^2 is the fraction of the total variance that is explained by the signal. For the Y equation, one can write

$$R^2 = \frac{\text{Signal Variance}}{\text{Signal Variance} + \text{Noise Variance}} = \frac{\text{SNR}_Y}{\text{SNR}_Y + 1} = \frac{\tau_\delta^{-1}p}{\tau_\delta^{-1}p + \Sigma_{11}/\mu^{(1)}}.$$

□

Proof of Lemma 1. From the properties of the Inverse-Wishart distribution,

$$\begin{aligned} \alpha | \Sigma_{11 \cdot 2} &\sim \text{Normal} \left(\frac{\Sigma_{0,12}}{\Sigma_{0,22}}, \frac{\Sigma_{11 \cdot 2}}{\Sigma_{0,22}} \right) \\ \Sigma_{11 \cdot 2} &\sim \text{Inverse-Gamma} \left(\frac{\nu_0}{2}, \frac{\Sigma_{0,11 \cdot 2}}{2} \right). \end{aligned}$$

where $\Sigma_{11 \cdot 2} = \Sigma_{11} - \Sigma_{12}^2/\Sigma_{22}$ and $\Sigma_{0,11 \cdot 2} = \Sigma_{0,11} - \Sigma_{0,12}^2/\Sigma_{0,22}$. Thus, the marginal distribution of α is a location-scale t -distribution:

$$\begin{aligned} \alpha &\sim \frac{\Sigma_{0,12}}{\Sigma_{0,22}} + \frac{t_{\nu_0}}{\sqrt{\nu_0 \Sigma_{0,22}/\Sigma_{0,11 \cdot 2}}} \\ &= \frac{\Sigma_{0,12}}{\Sigma_{0,22}} + \frac{|\Sigma_0|^{1/2}}{\sqrt{\nu_0 \Sigma_{0,22}}} t_{\nu_0}. \end{aligned}$$

□

The following two propositions generalize [Proposition 2](#) from [subsection 5.2](#) to the case where D_i may not be binary. In this setting we work with the *confounding bias*, defined as $\Delta(z) = \mathbb{E}[Y_i | D_i = z] - \mathbb{E}[Y_i(z)]$, rather than selection bias but the intuition and proofs are effectively identical.

Proposition A.2 (Confounding bias: naïve estimator). *Assume that the model and naïve estimator are as above. Suppose that Assumptions 2–5 hold. The confounding bias of the naïve estimator is*

$$\sqrt{p}\Delta(z) \xrightarrow{d} \text{Normal} \left(0, \frac{\lambda^{-1}(d^{(1)})^2}{\tau_\gamma^{-1}d^{(2)}} z^2 \right).$$

Therefore, $\Delta(z) \xrightarrow{p} 0$, if $p \rightarrow \infty$ as $n \rightarrow \infty$, and thus the naïve estimator is dogmatic in the limit.

Proof of A.2. See [Linero \(2023\)](#) Supplement S.2. Note that

$$\Delta(z) = z \frac{\gamma' \Sigma_X \beta}{\Sigma_{22} + \gamma' \Sigma_X \gamma}.$$

□

Proposition A.3 (Confounding bias: BDML). *Assume that the model and prior are as above. Suppose that Assumptions 2–5 hold. Then confounding bias of the BDML estimator*

is

$$\Delta(z) \xrightarrow{p} \frac{\Sigma_{12}}{\gamma' \Sigma_X \gamma + \Sigma_{22}} z,$$

which is non-dogmatic in the limit.

Proof of Proposition A.3. Note that the confounding bias is defined as $\Delta(z) = \mathbb{E}[Y_i|D_i = z] - \mathbb{E}[Y_i(z)]$. For the first term, as $\mathbb{E}[X_i] = 0$,

$$\begin{aligned} \mathbb{E}[Y_i|D_i = z] &= \frac{\text{Cov}(Y_i, D_i)}{\text{Var}(D_i)} z = \frac{\text{Cov}(X_i \delta, X_i \gamma) + \text{Cov}(V_i, U_i)}{\text{Var}(X_i \gamma) + \text{Var}(V_i)} z \\ &= \frac{\gamma' \Sigma_X \delta + \Sigma_{12}}{\gamma' \Sigma_X \gamma + \Sigma_{22}} z \xrightarrow{p} \frac{\Sigma_{12}}{\gamma' \Sigma_X \gamma + \Sigma_{22}} z, \end{aligned}$$

where the convergence follows from a similar argument as in Proposition A.2. The distribution of the first term is non-degenerate as $\Sigma \sim \text{Inverse-Wishart}(\nu_0, \Sigma_0)$. The second term equals 0 as D does not directly enter into the Y equation. Combining both terms, we obtain the distribution of the confounding bias under the BDML estimator, which is non-dogmatic in the limit. \square

Proof of Proposition 3. Denote the posterior mean estimator of α as $\hat{\alpha}_{\text{naive}}$. Denote $p/n = r(1 + o(1))$. Let

$$S_X = X'X/n \quad \text{and} \quad S_\lambda = S_X + \lambda/n\mathbb{I}.$$

Recall $\Sigma_X = Q_X \Lambda_X Q_X'$ and Q_X is the matrix of eigenvectors of Σ_X . For a generic vector θ , let $\theta_Q = Q_X' \theta$. By an argument similar to that of Proposition 1,

$$\begin{aligned} \hat{\alpha}_{\text{naive}} &= \frac{D'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)Y}{D'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)D} = \frac{(X\gamma^* + V)'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)(\alpha^* D + X\beta^* + \varepsilon)}{(X\gamma^* + V)'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)(X\gamma^* + V)} \\ &= \alpha^* + \frac{(X\gamma^* + V)'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)(X\beta^* + \varepsilon)}{(X\gamma^* + V)'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)(X\gamma^* + V)}. \end{aligned}$$

For the numerator, note that $\|\delta^*\|_2$ and $\|\gamma^*\|_2$ are bounded, and since $\beta^* = \delta^* - \alpha^* \gamma^*$, it follows that $\|\beta^*\|_2$ is bounded as well. Moreover, the spectral of Σ_X converge to a limit distribution on $(0, \infty)$. Therefore, following the argument in Lemma S.2 of Linero (2023), we have that

$$V'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)X\beta^*/n, \quad \gamma^{*'}X'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)\varepsilon/n, \quad V'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)\varepsilon/n$$

are all $O_p\left(\frac{1}{\sqrt{n}}\right)$. There is only one remaining term

$$\gamma^{*'}X'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)X\beta^*/n = \frac{\lambda}{n} \sum_{j=1}^p \frac{d_j}{d_j + \lambda/n} \gamma_{Q,j}^* \beta_{Q,j}^* + O_p\left(\frac{1}{\sqrt{n}}\right).$$

For the denominator, similarly, we have $V'(\mathbb{I}_n - X S_\lambda^{-1} X'/n)X\gamma^*/n$ is $O_p\left(\frac{1}{\sqrt{n}}\right)$, and two

remaining terms

$$\begin{aligned}\gamma^{*'}X'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)X\gamma^*/n &= \frac{\lambda}{n} \sum_{j=1}^p \frac{d_j}{d_j + \lambda/n} (\gamma_{Q,j}^*)^2 + O_p\left(\frac{1}{\sqrt{n}}\right), \\ V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)V/n &= \Sigma_{22}^* \left(\frac{\lambda}{n} \sum_{j=1}^p \frac{1}{d_j + \lambda/n} + 1 - r \right) + O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

All $O_p(1/\sqrt{n})$ terms are mean zero.

Combining the numerator and denominator, we have

$$\hat{\alpha}_{\text{naive}} - \alpha^* = \frac{\frac{\lambda}{n} \sum_{j=1}^p \frac{d_j}{d_j + \lambda/n} \gamma_{Q,j}^* \beta_{Q,j}^* + O_p\left(\frac{1}{\sqrt{n}}\right)}{\frac{\lambda}{n} \sum_{j=1}^p \frac{d_j}{d_j + \lambda/n} (\gamma_{Q,j}^*)^2 + \Sigma_{22}^* \left(\frac{\lambda}{n} \sum_{j=1}^p \frac{1}{d_j + \lambda/n} + 1 - r \right) + O_p\left(\frac{1}{\sqrt{n}}\right)} \quad (\text{A.4})$$

As $\lambda = o(n)$, and $\|\beta^*\|_2$ and $\|\gamma^*\|_2$ are bounded,

$$\begin{aligned}\hat{\alpha}_{\text{naive}} - \alpha^* &= \frac{1}{\Sigma_{22}^*} \frac{\lambda}{n} \sum_{j=1}^p \frac{d_j}{d_j + \lambda/n} \gamma_{Q,j}^* \beta_{Q,j}^* + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{\lambda}{n} \frac{\gamma^{*'} \beta^*}{\Sigma_{22}^*} + o\left(\frac{p}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

where the $O_p(1/\sqrt{n})$ term is mean zero, so the naïve estimator is consistent. The last line is given by the fact that $\lambda \asymp p$.

$$\sqrt{n}(\hat{\alpha}_{\text{naive}} - \alpha^*) = \frac{\lambda}{\sqrt{n}} \frac{\gamma^{*'} \beta^*}{\Sigma_{22}^*} + o\left(\frac{p}{\sqrt{n}}\right) + O_p(1).$$

As $\lambda \asymp p$, if $\lim_{n \rightarrow \infty} p^2/n = \infty$ and $\gamma^{*'} \beta^* \neq 0$, the naïve estimator is not \sqrt{n} -consistent. \square

Proof of Proposition 4. Now, let's derive the (approximate) marginal posterior distribution of Σ integrating out B . We begin with the joint posterior of B and Σ :

$$\begin{aligned}p(B, \Sigma | X, W) &\propto |\Sigma|^{-(n+\nu_0+3)/2} \exp\left(-\frac{1}{2}\text{tr}(\Sigma_0 \Sigma^{-1})\right) \\ &\quad \times \exp\left(-\frac{1}{2}\text{tr}[(W - XB)'(W - XB)\Sigma^{-1}]\right) \\ &\quad \times \exp\left(-\frac{1}{2}\text{vec}(B)'V_0^{-1}\text{vec}(B)\right).\end{aligned}$$

Integrating out B , we get:

$$p(\Sigma|X, W) \propto |\Sigma|^{-(n+\nu_0+3)/2} |V_n|^{1/2} \exp\left(-\frac{1}{2}\text{tr}((\Sigma_0 + WW')\Sigma^{-1})\right) \\ \times \exp\left(\frac{1}{2}\text{vec}(B_n)'V_n^{-1}\text{vec}(B_n)\right).$$

Based on Assumption 5, $\tau_\delta, \tau_\gamma = o(n)$. Then, for large n , we can approximate V_n :

$$V_n = (\Sigma^{-1} \otimes X'X + V_0^{-1})^{-1} = \Sigma \otimes (X'X)^{-1}(1 + o_p(1)), \quad (\text{A.5})$$

$$V_n = \Sigma \otimes (X'X)^{-1} - (\Sigma^{-1} \otimes X'X)^{-1}V_0^{-1}(\Sigma^{-1} \otimes X'X)^{-1}(1 + o_p(1)) \quad (\text{A.6})$$

where (A.5) and (A.6) give the zeroth and first order approximations, respectively.

First, we approximate $|V_n|^{1/2}$ using the first order approximation (A.6):

$$|V_n|^{1/2} = |\Sigma \otimes (X'X)^{-1}|^{1/2} |I - V_0^{-1}(\Sigma \otimes (X'X)^{-1})|^{1/2} (1 + o_p(1)) \\ = |\Sigma|^{p/2} |(X'X)^{-1}| \exp\left(-\frac{1}{2}\text{tr}(V_0^{-1}(\Sigma \otimes (X'X)^{-1}))\right) (1 + o_p(1)) \\ = |\Sigma|^{p/2} |(X'X)^{-1}| \exp\left(-\frac{1}{2}(\tau_\delta \Sigma_{11} + \tau_\gamma \Sigma_{22})\text{tr}((X'X)^{-1})\right) (1 + o_p(1)),$$

where the second line is given by the fact that $|I - A| \approx 1 - \text{tr}(A) \approx \exp(-\text{tr}(A))$ for small A . Note that

$$(\tau_\delta \Sigma_{11} + \tau_\gamma \Sigma_{22})\text{tr}((X'X)^{-1}) = \text{tr}\left(\begin{bmatrix} \tau_\delta \text{tr}((X'X)^{-1}) & 0 \\ 0 & \tau_\gamma \text{tr}((X'X)^{-1}) \end{bmatrix} \Sigma\right). \quad (\text{A.7})$$

Second, note that

$$\text{vec}(B_n) = (I - V_n V_0^{-1}) \text{vec}(\hat{B}).$$

Now, let's expand $\text{vec}(B_n)'V_n^{-1}\text{vec}(B_n)$:

$$\text{vec}(B_n)'V_n^{-1}\text{vec}(B_n) \\ = \text{vec}(\hat{B})'(I - V_n V_0^{-1})'V_n^{-1}(I - V_n V_0^{-1})\text{vec}(\hat{B}) \\ = \text{vec}(\hat{B})'V_n^{-1}\text{vec}(\hat{B}) - 2\text{vec}(\hat{B})'V_0^{-1}\text{vec}(\hat{B}) + \text{vec}(\hat{B})'V_0^{-1}V_n V_0^{-1}\text{vec}(\hat{B}) \\ = \text{vec}(\hat{B})'(\Sigma^{-1} \otimes X'X)\text{vec}(\hat{B}) - \text{vec}(\hat{B})'V_0^{-1}\text{vec}(\hat{B}) \\ + \text{vec}(\hat{B})'V_0^{-1}(\Sigma \otimes (X'X)^{-1})V_0^{-1}\text{vec}(\hat{B})(1 + o_p(1)),$$

where the last line is by plugging in $V_n^{-1} = \Sigma^{-1} \otimes X'X + V_0^{-1}$ and the zeroth order approximation of V_n in (A.5). The first term $\text{vec}(\hat{B})'(\Sigma^{-1} \otimes X'X)\text{vec}(\hat{B})$ will be absorbed by the OLS part in $\tilde{\Sigma}_n$. The second term $\text{vec}(\hat{B})'V_0^{-1}\text{vec}(\hat{B})$ does not involve Σ . Let $\hat{B}_V = [\tau_\delta \hat{\delta}, \tau_\gamma \hat{\gamma}]$.

Then, $\text{vec}(\hat{B}_V) = V_0^{-1}\text{vec}(\hat{B})$, and the third term

$$\begin{aligned} \text{vec}(\hat{B})'V_0^{-1}(\Sigma \otimes (X'X)^{-1})V_0^{-1}\text{vec}(\hat{B}) &= \text{vec}(\hat{B}_V)'(\Sigma \otimes (X'X)^{-1})\text{vec}(\hat{B}_V) \\ &= \text{tr} \left(\hat{B}_V'(X'X)^{-1}\hat{B}_V\Sigma \right), \end{aligned} \quad (\text{A.8})$$

where

$$\begin{aligned} \hat{B}_V'(X'X)^{-1}\hat{B}_V &= \begin{bmatrix} \tau_\delta \hat{\delta}' \\ \tau_\gamma \hat{\gamma}' \end{bmatrix} (X'X)^{-1} \begin{bmatrix} \tau_\delta \hat{\delta} & \tau_\gamma \hat{\gamma} \end{bmatrix} \\ &= \begin{bmatrix} \tau_\delta^2 \hat{\delta}'(X'X)^{-1}\hat{\delta} & \tau_\delta \tau_\gamma \hat{\delta}'(X'X)^{-1}\hat{\gamma} \\ \tau_\delta \tau_\gamma \hat{\gamma}'(X'X)^{-1}\hat{\delta} & \tau_\gamma^2 \hat{\gamma}'(X'X)^{-1}\hat{\gamma} \end{bmatrix}. \end{aligned} \quad (\text{A.9})$$

Combining (A.7) and (A.8), we obtain the tilting part in \tilde{C}_n . □

Proof of Proposition 5. First, we approximate the posterior mode (and mean) of Σ by the fixed-point iteration solution to the Riccati equation. From the FOC, we have the matrix quadratic equation

$$\Sigma \tilde{C}_n \Sigma - (\tilde{\nu}_n + 3)\Sigma + \tilde{\Sigma}_n = 0.$$

Rearranging into fixed-point form

$$\Sigma = \frac{\tilde{\Sigma}_n}{\tilde{\nu}_n + 3} + \frac{\Sigma \tilde{C}_n \Sigma}{\tilde{\nu}_n + 3}.$$

Taking the zeroth-order approximation $\Sigma^{(0)} = \tilde{\Sigma}_n/(\tilde{\nu}_n + 3)$ and performing one Picard iteration, we have

$$\begin{aligned} \Sigma^{(1)} &= \frac{\tilde{\Sigma}_n}{\tilde{\nu}_n + 3} + \frac{1}{\tilde{\nu}_n + 3} \left(\frac{\tilde{\Sigma}_n}{\tilde{\nu}_n + 3} \right) \tilde{C}_n \left(\frac{\tilde{\Sigma}_n}{\tilde{\nu}_n + 3} \right) \\ &= \frac{\tilde{\Sigma}_n}{n} + \frac{\tilde{\Sigma}_n \tilde{C}_n \tilde{\Sigma}_n}{n^3} (1 + o_p(1)), \end{aligned}$$

where the last line is by the fact that $\tilde{\nu}_n + 3 \sim n$.

Note that

$$\begin{aligned} \frac{\tilde{\Sigma}_n}{n} &= \frac{\Sigma_0}{n} + \frac{(W - X\hat{B})'(W - X\hat{B})}{n} \\ &= \frac{\Sigma_0}{n} + \Sigma^* - \frac{p}{n}\Sigma^* + O_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

where the $O_p\left(\frac{1}{\sqrt{n}}\right)$ term is mean zero. The first line is by the definition of $\tilde{\Sigma}_n$. The second line is by the property of sample variance covariance matrix of the OLS residues, i.e., $\mathbb{E} \left[(W - X\hat{B})'(W - X\hat{B})/n - p \right] = \Sigma^*$, given $p < n$ and $p = o(n)$ in Assumptions 4 and 5.

Under Assumptions 2–5, we can also approximate \tilde{C}_n as

$$\tilde{C}_n = C^*(1 + o_p(1)), \text{ where } C^* \asymp p^2/n.$$

Then, the posterior mode and mean of Σ up to the order of $1/n$ and $(p/n)^2$ is¹⁵

$$\mathbb{E}[\Sigma|X, W] = \Sigma^* + \frac{1}{n}\Sigma_0 - \frac{p}{n}\Sigma^* + \frac{1}{n}\Sigma^*C^*\Sigma^* + o_p\left(\frac{1}{n}\right) + o_p\left(\left(\frac{p}{n}\right)^2\right) + O_p\left(\frac{1}{\sqrt{n}}\right),$$

where the $O_p\left(\frac{1}{\sqrt{n}}\right)$ term is mean zero.

Also note that the posterior variance of Σ is of order $O_p\left(\frac{1}{n}\right)$, and $\Sigma_{22}^* > 0$. Then, the posterior mean of α is approximately

$$\begin{aligned} \hat{\alpha}_{\text{BDML}} = \mathbb{E}[\alpha|X, W] &= \frac{\Sigma_{12}^* + [\Sigma_{0,12} - p\Sigma_{12}^* + (\Sigma^*C^*\Sigma^*)_{12}]/n}{\Sigma_{22}^* + [\Sigma_{0,22} - p\Sigma_{22}^* + (\Sigma^*C^*\Sigma^*)_{22}]/n} \\ &\quad + o_p\left(\frac{1}{n}\right) + o_p\left(\left(\frac{p}{n}\right)^2\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

where the $O_p\left(\frac{1}{\sqrt{n}}\right)$ term is mean zero. Then, we approximate $\hat{\alpha}_{\text{BDML}}$ up to the order of $1/n$ and $(p/n)^2$. Let

$$\Delta = [\Sigma_0 - p\Sigma^* + \Sigma^*C^*\Sigma^*]/n,$$

Then,

$$\begin{aligned} \frac{\Sigma_{12}^* + [\Sigma_{0,12} - p\Sigma_{12}^* + (\Sigma^*C^*\Sigma^*)_{12}]/n}{\Sigma_{22}^* + [\Sigma_{0,22} - p\Sigma_{22}^* + (\Sigma^*C^*\Sigma^*)_{22}]/n} &= \frac{\Sigma_{12}^* + \Delta_{12}}{\Sigma_{22}^* + \Delta_{22}} \\ &= \frac{\Sigma_{12}^*}{\Sigma_{22}^*} \left[1 + \left(\frac{\Delta_{12}}{\Sigma_{12}^*} - \frac{\Delta_{22}}{\Sigma_{22}^*} \right) \left(1 - \frac{\Delta_{22}}{\Sigma_{22}^*} \right) \right] \end{aligned}$$

Note that

$$\begin{aligned} \frac{\Delta_{12}}{\Sigma_{12}^*} - \frac{\Delta_{22}}{\Sigma_{22}^*} &= \frac{\Sigma_{0,12} - p\Sigma_{12}^* + (\Sigma^*C^*\Sigma^*)_{12}}{n\Sigma_{12}^*} - \frac{\Sigma_{0,22} - p\Sigma_{22}^* + (\Sigma^*C^*\Sigma^*)_{22}}{n\Sigma_{12}^*} \\ &= \frac{1}{n} \left(\frac{\Sigma_{0,12} + (\Sigma^*C^*\Sigma^*)_{12}}{\Sigma_{12}^*} - \frac{\Sigma_{0,22} + (\Sigma^*C^*\Sigma^*)_{22}}{\Sigma_{12}^*} \right) \\ &= O\left(\frac{1}{n}\right) + O\left(\left(\frac{p}{n}\right)^2\right), \end{aligned}$$

¹⁵Here we consider $(p/n)^2$ as the p/n order terms will be cancelled out in the approximation for $\hat{\alpha}_{\text{BDML}}$.

where we cancel out the p/n order terms. As $\alpha^* = \Sigma_{12}^*/\Sigma_{22}^*$, we have

$$\begin{aligned}\hat{\alpha}_{\text{BDML}} &= \alpha^* + \frac{\alpha^*}{n} \left(\frac{\Sigma_{0,12} + (\Sigma^* C^* \Sigma^*)_{12}}{\Sigma_{12}^*} - \frac{\Sigma_{0,22} + (\Sigma^* C^* \Sigma^*)_{22}}{\Sigma_{22}^*} \right) \\ &\quad + o_p\left(\frac{1}{n}\right) + o_p\left(\left(\frac{p}{n}\right)^2\right) + O_p\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

where we cancel out common terms in the second equality. Then, $\hat{\alpha}_{\text{BDML}} \xrightarrow{p} \alpha^*$, so the BDML estimator is consistent.

Moreover,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}_{\text{BDML}} - \alpha^*) &= \frac{\alpha^*}{\sqrt{n}} \left(\frac{\Sigma_{0,12} + (\Sigma^* C^* \Sigma^*)_{12}}{\Sigma_{12}^*} - \frac{\Sigma_{0,22} + (\Sigma^* C^* \Sigma^*)_{22}}{\Sigma_{22}^*} \right) \\ &\quad + o_p\left(\frac{1}{\sqrt{n}}\right) + o_p\left(\frac{p^2}{n^{3/2}}\right) + O_p(1),\end{aligned}$$

where the $O_p(1)$ term is mean zero. Then, if $p = o(n^{3/4})$,

$$\frac{\alpha^*}{\sqrt{n}} \left(\frac{\Sigma_{0,12} + (\Sigma^* C^* \Sigma^*)_{12}}{\Sigma_{12}^*} - \frac{\Sigma_{0,22} + (\Sigma^* C^* \Sigma^*)_{22}}{\Sigma_{22}^*} \right) = O\left(\frac{1}{\sqrt{n}}\right) + O\left(\frac{p^2}{n^{3/2}}\right) = o_p(1),$$

then $\hat{\alpha}_{\text{BDML}}$ is \sqrt{n} -consistent. \square

Proof of Proposition 6. Both $\hat{\alpha}_{\text{BDML}}$ and $\hat{\alpha}_{\text{FDML}}$ can be approximated as $\hat{\alpha} \approx \bar{\Sigma}_{12}/\bar{\Sigma}_{22}$ where $\bar{\Sigma} = (W - X\hat{B})'(W - X\hat{B})/n$. Applying the delta method to this ratio, we obtain the leading term of their asymptotic variance

$$\begin{aligned}n\text{Var} \left[\frac{\Sigma_{12}^* \bar{\Sigma}_{22} - \Sigma_{22}^* \bar{\Sigma}_{12}}{(\Sigma_{22}^*)^2} \right] \\ = \frac{n}{(\Sigma_{22}^*)^4} \left[(\Sigma_{12}^*)^2 \text{Var}(\bar{\Sigma}_{22}) + (\Sigma_{22}^*)^2 \text{Var}(\bar{\Sigma}_{12}) - 2\Sigma_{12}^* \Sigma_{22}^* \text{Cov}(\bar{\Sigma}_{12}, \bar{\Sigma}_{22}) \right] (1 + o(1)).\end{aligned}$$

Note that $n\bar{\Sigma} \sim \mathcal{W}_2(n-p, \Sigma^*)$, then we have

$$\text{Cov}(n\bar{\Sigma}_{ij}, n\bar{\Sigma}_{kl}) = (n-p)(\Sigma_{ik}^* \Sigma_{jl}^* + \Sigma_{il}^* \Sigma_{jk}^*).$$

As $p = o(n)$, we have

$$\begin{aligned}n\text{Var}(\bar{\Sigma}_{22}) &\rightarrow 2(\Sigma_{22}^*)^2, \\ n\text{Var}(\bar{\Sigma}_{12}) &\rightarrow \Sigma_{11}^* \Sigma_{22}^* + (\Sigma_{12}^*)^2, \\ n\text{Cov}(\bar{\Sigma}_{12}, \bar{\Sigma}_{22}) &\rightarrow 2\Sigma_{12}^* \Sigma_{22}^*.\end{aligned}$$

Results follow from plugging in the above expressions.

For $\hat{\alpha}_{\text{naive}}$, the leading term of the asymptotic variance is given by

$$n\text{Var} \left[\frac{V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)\varepsilon/n}{\Sigma_{22}^*} \right] = \frac{(\sigma_\varepsilon^*)^2 \Sigma_{22}^*}{(\Sigma_{22}^*)^2} (1 + o(1)) = \frac{(\sigma_\varepsilon^*)^2}{\Sigma_{22}^*} (1 + o(1)),$$

where the first equality is given by the fact that

$$n\text{Var}[V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)\varepsilon/n] \rightarrow \text{Var} \left(\frac{1}{n} \sum_{i=1}^n V_i' \varepsilon_i \right) = (\sigma_\varepsilon^*)^2 \Sigma_{22}^*.$$

Our structural model imposes $\alpha^* = \Sigma_{12}^*/\Sigma_{22}^*$ and $(\sigma_\varepsilon^*)^2 = \Sigma_{11}^* - (\alpha^*)^2 \Sigma_{22}^* = \Sigma_{11}^* - (\Sigma_{12}^*)^2/\Sigma_{22}^*$, so the asymptotic variance expressions coincide. \square

Proof of Lemma 2. See Theorems 2.8.1 and 5.4.1 in Vershynin (2018). \square

Proof of Proposition 7. The proof follows from Walker (2024) by verifying his assumptions in our high dimensional regression setup.

Part 1: Assumptions 1 and 2 in Walker (2024). Assumption 1 in Walker (2024) is satisfied by our model setup and Assumptions 2–5. Our Assumption 4 also implies that $\mathbb{E}[X_i X_i']$ is non-singular, which ensures identification.

Assumption 2 in Walker (2024) is satisfied by our prior together with the additional assumption that σ_ε^{2*} and σ_V^{2*} are known, and the prior for α is continuous and positive over a neighborhood of α^* .

Part 2: Assumptions 3 and 4 in Walker (2024). For notation simplicity, let $\theta_1 = \delta$, $\theta_2 = \gamma$, and $\theta = (\theta_1', \theta_2')'$. By Corollary 1, since X_i is sub-gaussian and i.i.d., and the eigenvalues of Σ_X is uniformly bounded by $[\underline{\lambda}_X, \bar{\lambda}_X]$ with $0 < \underline{\lambda}_X \leq \bar{\lambda}_X < \infty$. For event

$$\mathcal{K}_n = \left\{ X \in \mathbb{R}^{n \times p} : \lambda_{\min} \left(\frac{X'X}{n} \right) \geq \frac{\underline{\lambda}_X}{2} \text{ and } \lambda_{\max} \left(\frac{X'X}{n} \right) \leq 2\bar{\lambda}_X \right\},$$

we have that $P(\mathcal{K}_n) \xrightarrow{P} 1$ as $n \rightarrow \infty$.

Define $\mathcal{M}_n = \mathcal{M}$, then Walker (2024)'s Assumption 3.1 is automatically satisfied.

For Assumption 3.2 in Walker (2024), the posterior contracts at rate

$$\|\theta - \theta^*\|_2 = O_p \left(\sqrt{\frac{p}{n}} \right)$$

uniformly over $X \in \mathcal{K}_n$. By the equivalence of norms, in event \mathcal{K}_n , we have

$$\frac{\underline{\lambda}_X}{2} \|\theta - \theta^*\|_2^2 \leq \|m - m^*\|_{n,2}^2 \leq 2\bar{\lambda}_X \|\theta - \theta^*\|_2^2, \quad (\text{A.10})$$

posterior contraction in ℓ_2 -norm implies contraction in the $L_2(P_n)$ norm for the functions.

Therefore, let $\nu_n = \sqrt{p/n}$, for some constants $D_m, D_\theta > 0$,

$$\begin{aligned} & \mathbb{E}_{P^*} [P(\mathcal{M}_n \cap B_{\mathcal{M},n,2}(m^*, D_m \nu_n)^c \mid X, W)] \\ & \leq \mathbb{E}_{P^*} [P(\|\theta - \theta^*\|_2 > D_\theta \nu_n \mid X, W, \mathcal{K}_n)] + P(\mathcal{K}_n^c) \rightarrow 0. \end{aligned}$$

For Assumption 4, as $p = o(\sqrt{n})$, $\nu_n = \sqrt{p/n} = o(n^{-1/4})$.

Part 3: Assumption 5 in Walker (2024). Assumption 5.1 considers the empirical process

$$G_n^{(1)}(m) = \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i [m_1(X_i) - m_1^*(X_i)]$$

be uniformly small over the set $B_{\mathcal{M},n,2}(m^*, D_m \nu_n)$.

In our linear model, write $m_1(X_i) - m_1^*(X_i) = X_i'(\delta - \delta^*)$. Thus, define the function class

$$\mathcal{F}_\nu = \{f(X) = X'(\delta - \delta^*) : \|\delta - \delta^*\|_2 \leq \nu\}.$$

The empirical L_2 norm is defined by

$$\|f\|_{n,2} = \left[\frac{1}{n} \sum_{i=1}^n (X_i'(\delta - \delta^*))^2 \right]^{1/2}.$$

Under event \mathcal{K}_n , by equivalence of norm (A.10), the metric on \mathcal{F}_ν is equivalent to the Euclidean norm on $\delta - \delta^*$. Then, the covering number of the ball $B_{\delta, D_\theta \nu_n} = \{\delta : \|\delta - \delta^*\|_2 \leq D_\theta \nu_n\}$ in \mathbb{R}^p satisfies

$$N(\epsilon, B_{\delta, D_\theta \nu_n}, \|\cdot\|_2) \leq \left(\frac{3D_\theta \nu_n}{\epsilon} \right)^p,$$

for $\epsilon < D_\theta \nu_n$. Thus, the covering number for $\mathcal{F}_{D_m \nu_n}$ with respect to $\|\cdot\|_{n,2}$ follows

$$\log N(\epsilon, \mathcal{F}_{D_m \nu_n}, \|\cdot\|_{n,2}) \lesssim p \log \frac{3D_m \nu_n}{\epsilon}.$$

Dudley's inequality states that there is a constant $K > 0$ such that

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}_{D_m \nu_n}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i f(X_i) \right| \cdot 1_{\mathcal{K}_n} \right] & \leq K \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}_{D_m \nu_n}, \|\cdot\|_{n,2})} d\epsilon \\ & = K \int_0^{3D_m \nu_n} \sqrt{\log N(\epsilon, \mathcal{F}_{D_m \nu_n}, \|\cdot\|_{n,2})} d\epsilon, \end{aligned}$$

where the second line is by the fact that the covering number is one for $\epsilon > 3D_m \nu_n$. Then the entropy integral becomes

$$\int_0^{3D_m \nu_n} \sqrt{p \log \frac{3D_m \nu_n}{\epsilon}} d\epsilon = O(\nu_n \sqrt{p}) = o(1).$$

The first equality follows from that the integral $\int_0^1 \sqrt{\log(1/u)} du$ is finite, and the second

equality is by $\nu_n = O(\sqrt{p/n})$ and $p = (\sqrt{n})$.

Under event \mathcal{K}_n^c , we do not have the norm equivalence. However, for any $f \in \mathcal{F}_{D_m \nu_n}$, we have that

$$\begin{aligned} |G_n^{(1)}(m)| &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^n |V_i| \|X_i\|_2 D_\theta \nu_n \\ &\leq D_\theta \nu_n \sqrt{n} \cdot \max_{1 \leq i \leq n} |V_i| \cdot \max_{1 \leq i \leq n} \|X_i\|_2, \end{aligned}$$

Under our sub-Gaussian assumptions for V_i and X_i , Exercise 2.5.10 in Vershynin (2018) yield

$$\max_{1 \leq i \leq n} |V_i| = O_p(\sqrt{\log n}) \text{ and } \max_{1 \leq i \leq n} \|X_i\|_2 = O_p(\sqrt{\log n}).$$

Hence, on \mathcal{K}_n^c ,

$$\sup_{f \in \mathcal{F}_{D_m \nu_n}} |G_n^{(1)}(m)| \leq D_\theta \nu_n \sqrt{n} \cdot O_p(\log n).$$

Then,

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}_{D_m \nu_n}} |G_n^{(1)}(m)| 1_{\mathcal{K}_n^c} \right] &\leq D_\theta \nu_n \sqrt{n} \cdot O(\log n) \cdot P(\mathcal{K}_n^c) \\ &= D_\theta \nu_n \sqrt{n} \cdot O(\log n) \cdot O(p^{-1}) \\ &= O\left(\frac{\log n}{\sqrt{p}}\right), \end{aligned}$$

where the second line follows from $P(\mathcal{K}_n^c) = O(p^{-1})$ by Corollary 1, and the last line is by $\nu_n = \sqrt{p/n}$.

Thus, the overall bound on the supremum is given by

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}_{D_m \nu_n}} |G_n^{(1)}(m)| \right] \leq O\left(\frac{p}{\sqrt{n}}\right) + O\left(\frac{\log n}{\sqrt{p}}\right).$$

Hence, provided that $p = o(\sqrt{n})$ and $\log n = o(p^{1/2})$, the overall supremum tends to zero.

This completes the verification of Walker's Assumption 5.1 in the $L_2(P_n)$ metric. Assumption 5.2 can be verified in a similar manner. \square

Proof of Corollary 1. As X_i is sub-gaussian, $S_i = X_i X_i' - \Sigma_X$ is sub-exponential, and its Orlicz ψ_1 norm is bounded, i.e., there exists a constant $M_S > 0$ such that $\|S_i\|_{\psi_1} \leq M_S$, for all i . By Lemma 2, for any $\epsilon \geq 0$,

$$P \left\{ \left\| \frac{X'X}{n} - \Sigma_X \right\| \geq \epsilon \right\} = P \left\{ \left\| \sum_{i=1}^n S_i \right\| \geq tn \right\} \leq 2p \cdot \exp \left(-cn \cdot \min \left(\frac{\epsilon^2}{M_S^2}, \frac{\epsilon}{M_S} \right) \right).$$

Typically, ϵ is small, so the minimum is achieved by the quadratic term. We choose

$\epsilon = M_S \sqrt{\frac{2 \log p}{cn}}$, then

$$P \left\{ \left\| \frac{X'X}{n} - \Sigma_X \right\| \geq \epsilon \right\} \leq 2p \cdot \exp(-2 \log p) = \frac{2}{p} = O(p^{-1}) \rightarrow 0,$$

as $n \rightarrow \infty$.

Define event $\mathcal{E}_n = \{X \in \mathbb{R}^{n \times p} : \left\| \frac{X'X}{n} - \Sigma_X \right\| < \epsilon\}$. For any $X \in \mathcal{E}_n$, Weyl's inequality implies that for every eigenvalue in decreasing order, $|\lambda_j(\frac{X'X}{n}) - \lambda_j(\Sigma_X)| < \epsilon$. For large n , $\epsilon < \underline{\lambda}_X/2 < \bar{\lambda}_X$, so $\lambda_{\min}(\frac{X'X}{n}) \geq \underline{\lambda}_X/2$ and $\lambda_{\max}(\frac{X'X}{n}) \leq 2\bar{\lambda}_X$, and thus $X \in \mathcal{K}_n$. Therefore,

$$\begin{aligned} P(\mathcal{K}_n) &\geq P(\mathcal{E}_n) \xrightarrow{p} 1, \\ P(\mathcal{K}_n^c) &\leq P(\mathcal{E}_n^c) = O(p^{-1}), \end{aligned}$$

as $n \rightarrow \infty$. □