Bayesian Double Machine Learning for Causal Inference

Francis J. DiTraglia*[1] and Laura Liu†[2]

[1]Department of Economics, University of Oxford

[2]Department of Economics, University of Pittsburgh

First Version: November 6, 2024 This Version: March 20, 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

*DiTraglia: francis.ditraglia@economics.ox.ac.uk

†Liu: laura.liu@pitt.edu

1 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

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

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.[1] 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

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

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

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

by iterated expectations and $\mathbb{E}[\varepsilon | X_i, D_i] = 0$, 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

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

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

where we have

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

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 if 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

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

which controls it would *suffice* to adjust for without being sure of which are most important.

2 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 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:

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

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:

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

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, $\widehat{\alpha}_\lambda$ 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 $\widehat{\alpha}_\lambda$ as follows under the simplifying assumption that ε is homoskedastic.[2]

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

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

[2] Hahn et al. (2018) derive a special case of Bias$(\widehat{\alpha}_\lambda)$, but their expression contains a minor error.

Assumption 1 implies

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

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

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

The j^{th} element of $\widehat{\rho}$ 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 $\widehat{\xi}$ is the corresponding vector of residuals. Setting $\lambda = 0$ corresponds to $\sigma_\beta^2 \to \infty$ and gives the OLS estimator of α. In this case the bias is zero and the variance simplifies to

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

If σ_β^2 is finite, then $\lambda \neq 0$ and $\widehat{\alpha}_\lambda$ is biased unless $\widehat{\rho}$ 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 $\text{Var}(\widehat{\alpha})$. 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 $\widehat{\alpha}_\lambda$ involves both β and $\widehat{\rho}$. All else equal, larger values of $\widehat{\rho}^{(j)}\beta^{(j)}$ 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 $\text{Bias}(\widehat{\alpha}_\lambda)$ also depends crucially on $\widehat{\rho}$, and λ fails to take this into account.[3] 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

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

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^2_γ quantifies our prior beliefs about the strength of dependence between the treatment and control variables. Unfortunately, σ^2_γ 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 the shorthand θ to refer to all model parameters, this implies that the likelihood factorizes as follows

$$f(Y, D|X, \theta) = f(Y|D, X, \theta)f(D|X, \theta) = f(Y|D, X, \alpha, \beta, \sigma^2_\varepsilon)f(D|X, \gamma, \sigma^2_V). \tag{11}$$

Notice that the first factor involves only $(\alpha, \beta, \sigma^2_\varepsilon)$ while the second involves only (γ, σ^2_V). Therefore, unless we specify a prior in which β and γ Treating σ^2_ε 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 4.2 below.

3 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 2. 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

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

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

$$\begin{aligned} Y_i &= X_i'\delta + U_i \\ D_i &= X_i'\gamma + V_i \end{aligned} \qquad \begin{bmatrix} U_i \\ V_i \end{bmatrix} \Bigg| X_i \sim \text{Normal}(\mathbf{0}, \Sigma). \tag{13}$$

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,

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

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

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

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 2 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 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 $\widehat{\alpha}_\lambda$ 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

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

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.[4] 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 $\widehat{D}_i$.

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

[4] While 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.

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 that adhere to the likelihood principle, all of which yield different posteriors. As we show in our simulations below, accounting for uncertainty in $\widehat{\gamma}$ 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.

4 Asymptotic Properties

4.1 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

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

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

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

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.[5] In particular, we combine (18) with the prior $\pi(\Sigma, B) = \pi(\Sigma)\pi(\delta)\pi(\gamma)$ where

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

and $(\tau_\delta, \tau_\gamma)$ denote the prior precisions for the reduced-form parameters δ and γ.[6]

To permit a clean comparison between our BDML approach based on the reduced form regressions in (13) and the more familiar approach based on the structural equation (4), we define a "naïve" alternative that is as similar as possible to (19), representing a fully-Bayesian equivalent of the Ridge Regression estimator from (8). In particular, our naïve model combines $\varepsilon_i|(X, D) \sim$ iid Normal$(0, \sigma_\varepsilon^2)$ with $\pi(\sigma_\varepsilon^2, \alpha, \beta) \propto \pi(\sigma_\varepsilon^2)\pi(\beta)$ where

$$\sigma_\varepsilon^2 \sim \text{Inverse-Gamma}(a_0, b_0), \quad \beta \sim \text{Normal}_p(0, \mathbb{I}_p/\tau_\beta) \tag{20}$$

and τ_β denotes the prior precision for β. Note that the naïve model places a flat prior on α

Equations 19 and 20 describe the Bayesian *models* 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^**.*

[5] See Zellner (1971) section 8.5 for a textbook treatment.

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

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) $\Sigma^* \equiv Var(E_i)$ *is finite and positive definite.*

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

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) $\Sigma_X \equiv Var(X_i)$ *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* $p \to \infty$*.*

(iv) $\frac{1}{p}\sum_{j=1}^{p} [\lambda_j(\Sigma_X)]^m \to \mu^{(m)} < \infty$ *as* $p \to \infty$ *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 from (19) and (20). 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) $\tau_\delta, \tau_\gamma, \tau_\beta = o(n)$

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

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.[7] Since $p \to \infty$, 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 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.

4.2 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*

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

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

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

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

As mentioned in section 2 above, Linero (2023) shows that any model that factorizes as in (11) and in which and 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 $n, p \to \infty$ the prior from (20) 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*.

4.3 Posterior Properties

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*

$$\hat{\alpha}_{naive} = \alpha^* + \frac{\lambda}{n}\frac{\alpha^*\gamma^{*\prime}\delta^* - \|\gamma^*\|_2^2}{\Sigma_{22}^*}(1 + o(1)) + O_p\left(\frac{1}{\sqrt{n}}\right),$$

where the $O_p(1/\sqrt{n})$ term is mean zero. Therefore, the naive estimator is consistent, but not $\sqrt{n}$-consistent for $p = \omega(\sqrt{n})$ and $\|\gamma^*\|_2^2/\gamma^{*\prime}\delta^* \neq \alpha^*$.

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

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

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

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

where

$$\nu_n = \nu_0 + n,$$
$$\Sigma_n = \Sigma_0 + (W - XB)'(W - XB).$$

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*

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

where

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

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.[8]

As $n \to \infty$, $\tau_\delta, \tau_\gamma = o(n)$, $\tilde{C}_n = o(n)$, 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 esti-*

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

mator is given by

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

where

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

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.*

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

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

where

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

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

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

where

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

Then,

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

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, 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 $\hat{\alpha}_{\text{BDML}}$, $\hat{\alpha}_{\text{FDML}}$, and $\hat{\alpha}_{\text{naive}}$ 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.

4.4 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 $n \to \infty$,*

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

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 $M_X > 0$, $\max_{1 \le j \le p} |X_{i,j}| \le M_X$. 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

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

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 $S_1, \ldots, S_n$ be i.i.d., mean-zero, symmetric random matrices in $\mathbb{R}^{p \times p}$, satisfying that $\|S_i\|_{\psi_1} \le M_S$ for all i. Then, for any $\epsilon \ge 0$,*

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

where $c > 0$ is an absolute constant.

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

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 $0 < \underline{\lambda}_X \le \bar{\lambda}_X < \infty$. Define event*

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

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

5 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

$$\begin{aligned} \{X_i\}_{i=1}^n &\sim \text{iid Normal}_p(0, \mathbb{I}_p) \\ \{(\varepsilon_i, V_i)'\}_{i=1}^n \mid X &\sim \text{iid Normal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_\varepsilon^2 & 0\\ 0 & 1\end{bmatrix}\right) \\ \beta \mid (X, \varepsilon, V) &\sim \text{Normal}_p\left(\mu_\beta, \sigma_\beta^2 \mathbb{I}_p\right). \end{aligned} \tag{21}$$

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

$$\alpha = 2, \quad \gamma = \iota_p/\sqrt{p}, \quad \mu_\beta = -\gamma/2, \quad \sigma_\beta^2 = 1/p, \quad n = 200, \quad p = 100 \tag{22}$$

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$(0, 5^2)$ on the elements of δ and γ. The second, **BDML-Hier**, allows different standard deviations, σ_δ and σ_γ, 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 and σ_γ. Our theoretical derivations suggest that the hierarchical approach should give better performance and, as we see below, this is borne out in our simulations. 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.

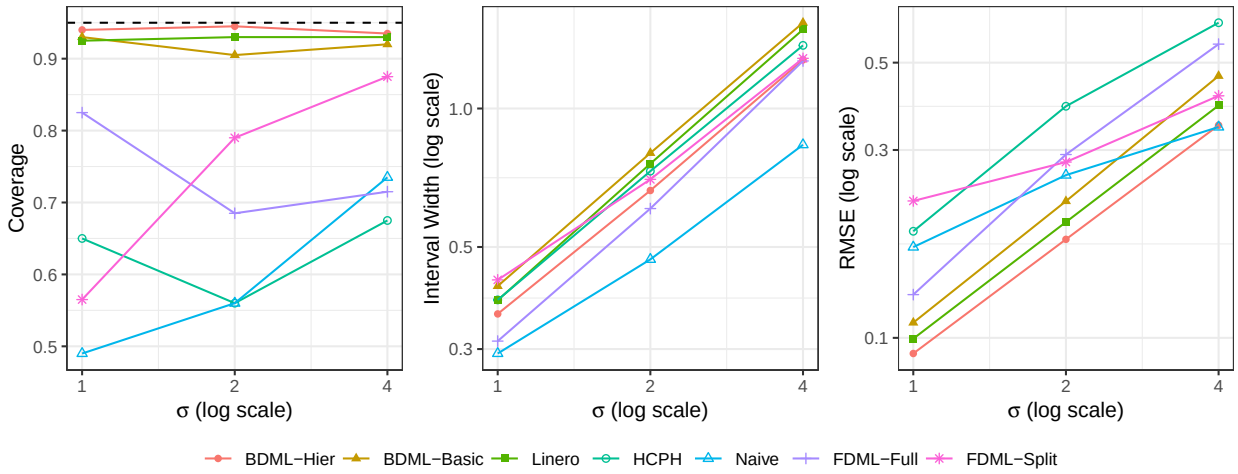
The second group of estimators we consider contains Bayesian competitors to our BDML approach. This group includes three estimators. The first two, **HCPH** and **Linero**, correspond to the proposals from Hahn et al. (2018) and Linero (2023) detailed in Algorithms 2–3 above and rely on a first-stage 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 2 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 with the default parameter settings. Again following Linero (2023) we place a flat prior on α across all three approaches and on κ from Algorithm 3. We likewise place ridge priors on the coefficients on X across all regressions in both stages.

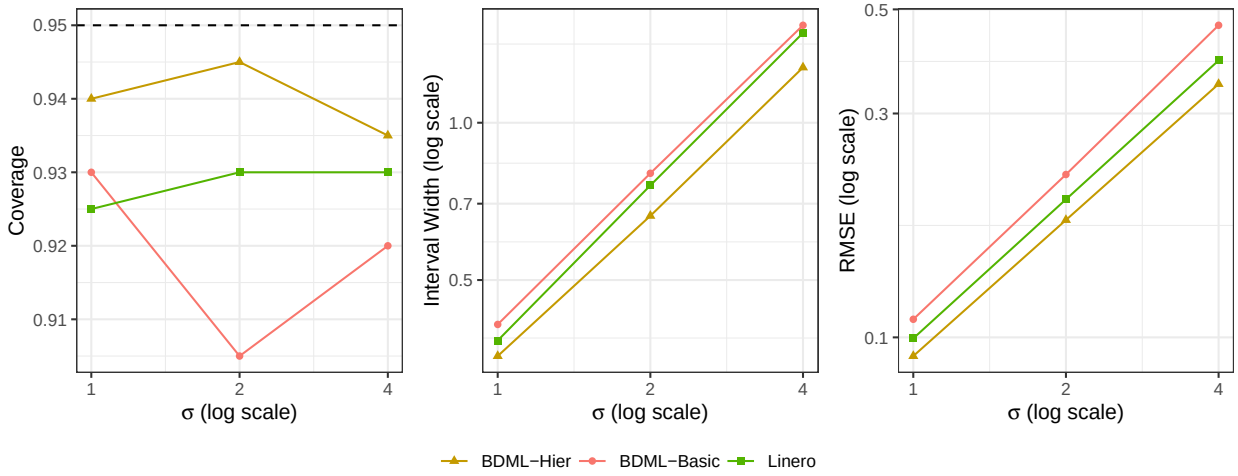
The third and final group of estimators contains frequentist DML competitors to our BDML approach. Both of the approaches in this group estimate α via (16) where the "first-step" estimators $\widehat{\gamma}$ and $\widehat{\delta}$ are constructed via ridge regression.[9] We carry out ridge regression 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 $(\widehat{\delta}_{\text{ML}}, \widehat{\gamma}_{\text{ML}})$ and 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 (21)–(22) 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.

[9] We use ridge rather than LASSO for these preliminary DML regressions because recent work by Kolesár et al. (2024) and Shen and Xiu (2024) suggests that the "double-LASSO" approach (Belloni et al., 2014) can have unfavorable performance due to fragility in variable selection.



(a)



(b)

Figure 1: This figure presents results for the estimators described in section 5 and simulation from (21)–(22) over a grid of values for σ_ε, the error variance in the structural equation given in (4). The panel labled RMSE gives the root mean-squared error of the estimator for α under each method, while those labled 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 betwen 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 5 and simulation from (21)–(22) 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

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.

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.

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.

Kolesár, M., Müller, U.K., Roelsgaard, S.T., 2024. The Fragility of Sparsity. URL: http://arxiv.org/abs/2311.02299. arXiv:2311.02299 [econ].

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.

Shen, Z., Xiu, D., 2024. Can Machines Learn Weak Signals? Technical Report. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4748784.

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

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

$$M^{11} = (D'D)^{-1} - M^{12}\widehat{\rho} \tag{A.2}$$

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

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

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

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

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

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

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

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

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

Similarly, we have

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

The result follows. □

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

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

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

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

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

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

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

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

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

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

$$SNR_Y \to \tau_\delta^{-1} p \frac{\mu^{(1)}}{\Sigma_{11}}, \quad \text{and} \quad SNR_D \to \tau_\gamma^{-1} p \frac{\mu^{(1)}}{\Sigma_{11}}.$$

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

$$R^2_Y \to \frac{\tau_\delta^{-1} p}{\tau_\delta^{-1} p + \Sigma_{11}/\mu^{(1)}} \quad \text{and} \quad R^2_D \to \frac{\tau_\gamma^{-1} p}{\tau_\gamma^{-1} p + \Sigma_{22}/\mu^{(1)}}.$$

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

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

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

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

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,

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

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 4.2 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*

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

Therefore, $\Delta(z) \xrightarrow{p} 0$*, if* $p \to \infty$ *as* $n \to \infty$*, 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. □

Proof of Proposition 3. Denote the posterior mean estimator of α as $\hat{\alpha}_{\text{naive}}$. Denote $\lim_{n\to\infty} p/n = r$. Let

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

For a generic vector θ, let $\theta_Q = Q_X'\theta$. By an argument similar to that of Proposition 1,

$$\hat{\alpha}_{\text{naive}} = \frac{D'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)Y}{D'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)Y} = \frac{(X\gamma^* + V)'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)(X\delta^* + U)}{(X\gamma^* + V)'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)(X\gamma^* + V)}.$$

For the numerator, note that as $\|\delta^*\|_2$ and $\|\gamma^*\|_2$ are bounded and the spectral of Σ_X converge to a limit distribution on $(0, \infty)$, we have that

$$V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)X\delta^*/n \quad \text{and} \quad \gamma^{*\prime}X'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)U/n$$

both are $O_p\left(\frac{1}{\sqrt{n}}\right)$. There are two remaining terms:

$$\begin{aligned}\gamma^{*\prime}X'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)X\delta^*/n &= \frac{\lambda}{n}\sum_{j=1}^{p} \frac{d_j}{d_j + \lambda/n}\gamma_{Q,j}^*\delta_{Q,j}^* + O_p\left(\frac{1}{\sqrt{n}}\right), \\ V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)U/n &= \Sigma_{12}^*\left(\frac{\lambda}{n}\sum_{j=1}^{p}\frac{1}{d_j + \lambda/n} + 1 - r\right) + O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

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

remaining terms:

$$\gamma^{*\prime}X'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)X\gamma^*/n = \frac{\lambda}{n}\sum_{j=1}^{p}\frac{d_j}{d_j+\lambda/n}(\gamma^*_{Q,j})^2 + O_p\left(\frac{1}{\sqrt{n}}\right),$$

$$V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)V/n = \Sigma^*_{22}\left(\frac{\lambda}{n}\sum_{j=1}^{p}\frac{1}{d_j+\lambda/n} + 1 - r\right) + O_p\left(\frac{1}{\sqrt{n}}\right).$$

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

Combining the numerator and denominator, we have

$$\hat{\alpha}_{\text{naive}} - \alpha^* = \frac{\frac{\lambda}{n}\sum_{j=1}^{p}\frac{d_j}{d_j+\lambda/n}\gamma^*_{Q,j}\delta^*_{Q,j} + \Sigma^*_{12}\left(\frac{\lambda}{n}\sum_{j=1}^{p}\frac{1}{d_j+\lambda/n} + 1 - r\right) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\frac{\lambda}{n}\sum_{j=1}^{p}\frac{d_j}{d_j+\lambda/n}(\gamma^*_{Q,j})^2 + \Sigma^*_{22}\left(\frac{\lambda}{n}\sum_{j=1}^{p}\frac{1}{d_j+\lambda/n} + 1 - r\right) + O_p\left(\frac{1}{\sqrt{n}}\right)} - \alpha^* \quad \text{(A.4)}$$

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

$$\begin{aligned}
\hat{\alpha}_{\text{naive}} - \alpha^* &= \frac{1}{(\Sigma^*_{22})^2}\left\{\Sigma^*_{22}\left[\frac{\lambda}{n}\sum_{j=1}^{p}\frac{d_j}{d_j+\lambda/n}\gamma^*_{Q,j}\delta^*_{Q,j} + \Sigma^*_{12}\left(\frac{\lambda}{n}\sum_{j=1}^{p}\frac{1}{d_j+\lambda/n} - r\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\right]\right. \\
&\qquad \left. -\Sigma^*_{12}\left[\frac{\lambda}{n}\sum_{j=1}^{p}\frac{d_j}{d_j+\lambda/n}(\gamma^*_{Q,j})^2 + \Sigma^*_{22}\left(\frac{\lambda}{n}\sum_{j=1}^{p}\frac{1}{d_j+\lambda/n} - r\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\right]\right\} \\
&= \frac{\lambda}{n}\frac{1}{\Sigma^*_{22}}\sum_{j=1}^{p}\frac{d_j}{d_j+\lambda/n}\left(\alpha^*\gamma^*_{Q,j}\delta^*_{Q,j} - (\gamma^*_{Q,j})^2\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \\
&= \frac{\lambda}{n}\frac{\alpha^*\gamma'\delta - \|\gamma\|_2^2}{\Sigma^*_{22}}(1+o(1)) + O_p\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where the $O_p(1/\sqrt{n})$ term is mean zero, so the naïve estimator is consistent. However,

$$\sqrt{n}(\hat{\alpha}_{\text{naive}} - \alpha^*) = \frac{\lambda}{\sqrt{n}}\frac{\alpha^*\gamma^{*\prime}\delta^* - \|\gamma^*\|_2^2}{\Sigma^*_{22}}(1+o(1)) + O_p(1).$$

If $\lambda \asymp p = \omega(\sqrt{n})$ and $\gamma'\delta/\|\gamma\|_2^2 \neq \alpha^*$, the naïve estimator is not $\sqrt{n}$-consistent. □

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:

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

Integrating out B, we get:

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

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

$$V_n = (\Sigma^{-1}\otimes X'X + V_0^{-1})^{-1} = \Sigma\otimes(X'X)^{-1}(1+o_p(1)), \tag{A.5}$$
$$V_n = \Sigma\otimes(X'X)^{-1} - (\Sigma^{-1}\otimes X'X)^{-1}V_0^{-1}(\Sigma^{-1}\otimes X'X)^{-1}(1+o_p(1)) \tag{A.6}$$

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):

$$\begin{aligned}
|V_n|^{1/2} &= |\Sigma\otimes(X'X)^{-1}|^{1/2}\left|I - V_0^{-1}(\Sigma\otimes(X'X)^{-1})\right|^{1/2}(1+o_p(1)) \\
&= |\Sigma|^{p/2}|(X'X)^{-1}|^{1/2}\exp\left(-\frac{1}{4}\text{tr}(V_0^{-1}(\Sigma\otimes(X'X)^{-1}))\right)(1+o_p(1)) \\
&= |\Sigma|^{p/2}|(X'X)^{-1}|^{1/2}\exp\left(-\frac{1}{4}(\tau_\delta\Sigma_{11}+\tau_\gamma\Sigma_{22})\text{tr}((X'X)^{-1})\right)(1+o_p(1)),
\end{aligned}$$

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

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

Second, note that

$$\text{vec}(B_n) = \left[I - V_nV_0^{-1}\right]\text{vec}(\hat{B}).$$

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

$$\begin{aligned}
&\text{vec}(B_n)'V_n^{-1}\text{vec}(B_n) \\
&= \text{vec}(\hat{B})'\left[I - V_nV_0^{-1}\right]'V_n^{-1}\left[I - V_nV_0^{-1}\right]\text{vec}(\hat{B}) \\
&= \text{vec}(\hat{B})'V_n^{-1}\text{vec}(\hat{B}) - 2\text{vec}(\hat{B})'V_0^{-1}\text{vec}(\hat{B}) + \text{vec}(\hat{B})'V_0^{-1}V_nV_0^{-1}\text{vec}(\hat{B}) \\
&= \text{vec}(\hat{B})'(\Sigma^{-1}\otimes X'X)\text{vec}(\hat{B}) - \text{vec}(\hat{B})'V_0^{-1}\text{vec}(\hat{B}) \\
&\quad + \text{vec}(\hat{B})'V_0^{-1}(\Sigma\otimes(X'X)^{-1})V_0^{-1}\text{vec}(\hat{B})(1+o_p(1)),
\end{aligned}$$

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

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

where

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

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

Proof of Proposition 6. Then, the asymptotic variance of $\hat{\alpha}_{\text{BDML}}$ and $\hat{\alpha}_{\text{FDML}}$ is given by

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

Note that $n\bar{\Sigma} \sim \mathcal{W}_2(n-p, \Sigma^*)$. So the result follows from the variance and covariance of the Wishart distribution.

For $\hat{\alpha}_{\text{naive}}$, note that the variance and covariance of

$$V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)U/n \quad \text{and} \quad V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)V/n$$

dominate the variance and covariance across other terms. As

$$\begin{aligned}n\text{Var}[V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)U/n] &\xrightarrow{p} n\text{Var}(\bar{\Sigma}_{12}) \\ n\text{Var}[V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)V/n] &\xrightarrow{p} n\text{Var}(\bar{\Sigma}_{22}) \\ n\text{Cov}[V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)U/n, V'(\mathbb{I}_n - XS_\lambda^{-1}X'/n)V/n] &\xrightarrow{p} n\text{Cov}(\bar{\Sigma}_{12}, \bar{\Sigma}_{22}),\end{aligned}$$

we obtain the same asymptotic variance for $\hat{\alpha}_{\text{naive}}$. □

Proof of Proposition 5. By FOC, as $n \to \infty$, the posterior mode (and mean as well) of Σ is given by

$$\mathbb{E}[\Sigma|X, W] = \frac{\tilde{\Sigma}_n}{n} + \frac{\tilde{\Sigma}_n\tilde{C}_n\tilde{\Sigma}_n}{n^3}(1 + o_p(1)).$$

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 residuals, 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. We can also approximate $\tilde{C}_n$ as

$$\tilde{C}_n = C^*(1 + o_p(1)).$$

Under Assumptions 2–5, $C^* \asymp p^2/n$. Then, the posterior mode and mean of Σ up to the order of $1/n$ and $(p/n)^2$ is[10]

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

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

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

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

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

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

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

Moreover,

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

where the $O_p(1)$ term is mean zero. Then, if $p = o(n^{3/4})$, $\hat{\alpha}_{\text{BDML}}$ is $\sqrt{n}$-consistent. □

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

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 $n \to \infty$.

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

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

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 $D_m, D_\theta > 0$,

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

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

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

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

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

Under event $\mathcal{K}_n$, by equivalence of norm (A.10), the metric on $\mathcal{F}_\nu$ 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

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

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

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

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

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

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

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

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,

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

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.

□

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 mininum is achieved by the quadratic term. We choose

$\epsilon = M_S\sqrt{\frac{2\log p}{\check{c}n}}$, then

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

as $n \to \infty$.

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

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

as $n \to \infty$. □