Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

In Nonparametric and High-Dimensional Models, Bayesian Ignorability is an Informative Prior

Antonio R. Linero

To cite this article: Antonio R. Linero (06 Nov 2023): In Nonparametric and High-Dimensional Models, Bayesian Ignorability is an Informative Prior, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2278202

To link to this article: https://doi.org/10.1080/01621459.2023.2278202

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2023.2278202

Taylor & Francis
Taylor & Francis Group

In Nonparametric and High-Dimensional Models, Bayesian Ignorability is an Informative Prior

Antonio R. Linero

Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX

ABSTRACT

In problems with large amounts of missing data one must model two distinct data generating processes: the outcome process, which generates the response, and the missing data mechanism, which determines the data we observe. Under the *ignorability* condition of Rubin, however, likelihood-based inference for the outcome process does not depend on the missing data mechanism so that only the former needs to be estimated; partially because of this simplification, ignorability is often used as a baseline assumption. We study the implications of Bayesian ignorability in the presence of high-dimensional nuisance parameters and argue that ignorability is typically incompatible with sensible prior beliefs about the amount of confounding bias. We show that, for many problems, ignorability directly implies that the prior on the selection bias is tightly concentrated around zero. This is demonstrated on several models of practical interest, and the effect of ignorability on the posterior distribution is characterized for high-dimensional linear models with a ridge regression prior. We then show both how to build high-dimensional models that encode sensible beliefs about the confounding bias and also show that under certain narrow circumstances ignorability is less problematic. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2021
Accepted October 2023

KEYWORDS
Bayesian methods; Causal inference; High-dimensional statistics; Missing data

1. Introduction

Dealing with missing data is a fundamental problem in data analysis: it complicates inference in clinical trials (National Research Council 2010) and is inherent in the potential outcomes framework for causal inference (Rubin 2005). A common starting point for addressing missingness is to assume that the mechanism that generated the missingness is *ignorable* (Rubin 1976). Ignorability allows likelihood-based inference to proceed without modeling the missing data mechanism, which can greatly simplify an analysis.

In this article we consider the Bayesian approach to account for missingness. For the sake of specificity, we focus on the *Rubin causal model* (Rubin 1974, 1978) for observational studies, which considers the *potential outcome* $Y_i(a)$ of some outcome under an exposure level $a \in \mathscr{A}$, such that we observe both the received exposure A_i and its associated potential outcome $Y_i \equiv Y_i(A_i)$; in this case, $Y_i(a)$ is regarded as missing for all $a \neq A_i$. Let X_i be a vector of confounders that are predictive of both A_i and $Y_i(a)$. Following Seaman et al. (2013) (see also Little and Rubin 2002, Definition 6.5), we will say that the *exposure model* $f_\phi(A_i \mid X_i)$ is *Bayesian-ignorable*, or simply ignorable, if the following conditions hold:

IG.1 The potential outcomes $\{Y_i(a) : a \in \mathscr{A}\}$ are conditionally independent of A_i given X_i.
IG.2 The parameters β and ϕ are a-priori independent, where β parameterizes the model for the potential outcomes and ϕ parameterizes the missing data mechanism. That is, the prior factors as $\pi(\beta, \phi) = \pi_\beta(\beta)\, \pi_\phi(\phi)$.

We opt for the term Bayesian ignorability to distinguish it from the term *ignorability* (Rosenbaum and Rubin 1983; Imai, Keele, and Tingley 2010) as used in causal inference, which is often taken to be synonymous with the *exchangeability* Assumption 1 (also sometimes referred to as *unconfoundedness*). Condition IG.1 constrains the data generating mechanism and is a type of *missing at random* (MAR) assumption (Rubin 1976), which itself is sometimes conflated with ignorability in the sense of missing data (see Seaman et al. 2013, for a thorough discussion of MAR and ignorability). Condition IG.2, which constrains the prior, is also key to ignorability: it guarantees that the posterior distribution of β given the observed data is proportional to $\pi_\beta(\beta) \prod_i f_\beta\{Y_i(A_i) \mid X_i\}$, which does not depend on the exposure model. Without IG.2 we are still obligated to specify an exposure model.

It has been argued, from a Frequentist perspective, that IG.2 is highly problematic in high-dimensional problems (Robins and Ritov 1997; Robins and Wasserman 2012). We complement this view by studying IG.2 from a Bayesian perspective. That such a perspective is valuable is demonstrated by the fact that Bayesian researchers have run afoul of this problem specifically when attempting to address the examples of Robins and Wasserman (2012). For example, Li (2010) proposes a prior that leads to Bayes estimators that do not appropriately correct for

CONTACT Antonio R. Linero ✉ *antonio.linero@austin.utexas.edu* Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

confounding bias. We argue that, while IG.2 is seemingly an innocuous convenience, it can inadvertently encode strong prior beliefs about the total amount of confounding bias, to the degree that the data has no reasonable chance of overcoming the prior. Following Sims (2012), we refer to priors with this property as *dogmatic* about the confounding bias. We make the following three points.

1. Priors that impose IG.2 are typically dogmatic about the degree of confounding bias, particularly in settings that require informative priors. We illustrate this in the simple, but representative, settings of ridge regression and Gaussian process regression, and argue further that it holds for the spike-and-slab priors. While we note some exceptions, we conclude that IG.2 does not reflect substantive prior knowledge in many cases.
2. By understanding this induced prior on the confounding bias, we are able to identify several highly effective methods for correcting this problem and unify several approaches proposed in the Bayesian causal inference literature that were not motivated by Bayesian considerations. Our remedies take the form of propensity score adjustments, which have typically been recommended in applied Bayesian analysis on the grounds of pragmatism and robustness (see, e.g., Rubin 1985; Li, Ding, and Mealli 2022) rather than subjective Bayesian principles. A limitation of these corrections is that they are derived on a case-by-case basis.
3. We study some relatively narrow settings in which prior dogmatism does not occur, even in high dimensional problems. For example, strong dependence structure in X_i can act as a shield against dogmatism; in our ridge regression example, we use random matrix theory to quantify this behavior (Dobriban and Wager 2018; Dicker 2016). Despite this, we find little benefit to failing to correct for dogmatism in these settings.

Remark 1. Many of our conclusions are reminiscent of D'Amour et al. (2021), who study the assumption of *overlap* in high-dimensional settings; they show that the strict overlap assumption implies that the confounders either (i) are roughly balanced across groups or (ii) are highly correlated. In the same way, our results imply that priors satisfying IG.2 lead to dogmatism unless either (i) there is some dimension-reducing structure in the propensity/outcome regressions or (ii) the confounders are highly correlated. An important difference between these works is that D'Amour et al. (2021) do not make any assumption about the correlation between the propensity score and outcome regression, and consequently their approach would yield bounds on the confounding bias parameters that are weaker than those obtained here.

1.1. Notation

For $i = 1, \dots, N$ we let $Y_i(a)$ denote a potential outcome, $X_i \in \mathbb{R}^P$ denote a vector of confounders, $A_i \in \mathbb{R}$ denote an exposure indicator, and define $Y_i = Y_i(A_i)$. We set $\boldsymbol{Y} = (Y_1, \dots, Y_N)^\top$, $\boldsymbol{A} = (A_1, \dots, A_N)$, and let $\boldsymbol{X}$ denote an $N \times P$ matrix obtained by stacking the row vectors $X_i^\top$. Let β parameterize the distribution of $[Y_i(\cdot) \mid X_i]$, let ϕ parameterize the distribution of $[A_i \mid X_i]$, and let $\theta = (\beta, \phi)$. We invoke IG.1 throughout.

We let $\mathbb{E}_\theta(\cdot)$ denote the expectation operator conditional on θ. If the subscript θ is omitted then $\mathbb{E}(\cdot)$ is the expectation operator with respect to a prior distribution on θ, for example, $\mathbb{E}(Y_i) = \int \mathbb{E}_\theta(Y_i)\, \pi(\theta)\, d\theta$. We use the Big-O notation $W = O_p(V)$ to mean that $|W|/|V|$ is bounded in probability as $P \to \infty$. Finally, we let $\lambda_j(\Sigma)$ denote the jth largest eigenvalue of a covariance matrix Σ; for example, $\lambda_1(\Sigma)$ denotes the largest eigenvalue of Σ.

1.2. Illustrative Problems

We consider two problems to illustrate the existence of dogmatism and how to correct for it. We assume $X_i \sim \text{Normal}(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{P \times P}$ to simplify our analysis. All proofs are deferred to the supplementary material.

High-Dimensional Linear Regression. We posit linear models for the outcome and the exposure models, $Y_i(a) = X_i^\top \beta + \gamma\, a + \epsilon_i(a)$ and $A_i = X_i^\top \phi + \nu_i$ with $\epsilon_i(a) \sim \text{Normal}(0, \sigma_y^2)$ and $\nu_i \sim \text{Normal}(0, \sigma_a^2)$, and allow P to grow with N. The Bayesian ridge regression prior, which satisfies IG.2, takes $\beta \sim \text{Normal}(0, \tau_\beta^2\ \mathrm{I})$, $\phi \sim \text{Normal}(0, \tau_\phi^2\ \mathrm{I})$, and a flat (improper) prior on γ. The parameter of interest is the mean response at a given exposure $\mathbb{E}_\theta\{Y_i(a)\} = \gamma\, a$. Define the *confounding bias parameter* as $\Delta(a) = \mathbb{E}_\theta(Y_i \mid A_i = a) - \mathbb{E}_\theta\{Y_i(a)\}$.

Semiparametric Regression. We posit a semiparametric normal regression model $Y_i(a) \sim \text{Normal}\{\mu(X_i) + a\, \tau(X_i), \sigma^2\}$ with a binary exposure variable $A_i \sim \text{Bernoulli}\{\phi(X_i)\}$. This parameterization was proposed by Hahn, Murray, and Carvalho (2020) for their *Bayesian causal forests* method. The parameter of interest in this problem is the *population average causal effect* $\tau = \int \tau(x)\, F_X(dx)$. For convenience, we will assume that $\mu(\cdot)$ and $\tau(\cdot)$ are both given independent *Gaussian process* priors (Rasmussen and Williams 2006) with covariance function $\kappa(\cdot, \cdot)$, written $\mu, \tau \overset{\text{iid}}{\sim} \text{GP}(0, \kappa)$. For this model, $\beta = (\mu, \tau)$. We define the *confounding bias parameter* for this model to be $\Delta = \mathbb{E}_\theta(Y_i \mid A_i = 1) - \mathbb{E}_\theta(Y_i \mid A_i = 0) - \mathbb{E}_\theta\{Y_i(1) - Y_i(0)\}$.

2. The Induced Prior on the Confounding Bias

The fundamental difficulty with missingness is *confounding bias*. In both of our illustrative examples, this amounts to the fact that $\Delta \neq 0$. Note that the statement $\Delta = 0$ is much stronger than the claim that there are no unmeasured confounders—it instead states that, for the purpose of conducting valid inference, it suffices to *ignore the confounders* (both measured and unmeasured) entirely! More precisely, $\Delta = 0$ implies that the effect of confounding is, on average, 0; hence, inference that is Frequentist-valid under an assumption that $\Delta = 0$ would also be valid under the assumption that there are no confounders. The possibility that $\Delta \neq 0$ is the only issue that makes estimation of average causal effects nontrivial, as otherwise we could ignore the covariates X_i and directly estimate $\mathbb{E}_\theta\{Y_i(a)\}$

by estimating $\mathbb{E}_\theta(Y_i \mid A_i = a)$ nonparametrically. The following proposition gives an expression for Δ in our problems.

Proposition 1. The confounding bias parameter is given by

$$\Delta(a) = a\,\frac{\phi^\top \Sigma \beta}{\sigma_a^2 + \phi^\top \Sigma \phi} \qquad \text{and}$$

$$\Delta = \frac{\mathrm{cov}_\theta\{\mu(X_i), \phi(X_i)\}}{\Pr_\theta(A_i = 1)\Pr_\theta(A_i = 0)} + \frac{\mathrm{cov}_\theta\{\tau(X_i), \phi(X_i)\}}{\Pr_\theta(A_i = 1)},$$

in the high-dimensional linear regression problem and semiparametric regression problem, respectively.

Given the importance of Δ and the working assumption that confounding bias is non-negligible, one would hope that the prior distribution of Δ is relatively diffuse. However, using Proposition 1, we can see this is not the case; for example, for the ridge regression prior we have the following.

Proposition 2. Assume the setup of Proposition 1 for the ridge regression problem and suppose $\beta \sim \mathrm{Normal}(0, \tau_\beta^2\,\mathrm{I})$ and $\phi \sim \mathrm{Normal}(0, \tau_\phi^2\,\mathrm{I})$ independently. Assume $\frac{1}{P}\sum_{j=1}^P \lambda_j(\Sigma)^k$ converges to a positive constant as $P \to \infty$ for $k = 1, 2, 2 + \epsilon$ for some ϵ, and let $\widetilde{\lambda}$ and $\bar{\lambda}^2$ be the limits with $k = 1, 2$. Then $\Delta(a) \overset{\bullet}{\sim} \mathrm{Normal}(0, c/P)$where $c = a^2\,(\tau_\beta^2/\tau_\phi^2)\,(\bar{\lambda}^2/\widetilde{\lambda}^2)$.

Proposition 2 contains several lessons, but the most important is that if confounding bias is a-priori a concern for us then it seems unwise to specify a $\mathrm{Normal}(0, c/P)$ prior for it when P is large. This behavior becomes even more suspect when one considers that the definition of $\Delta(a)$ is completely free of the X_i's, and that there is little reason to expect that the number of confounders we need to control for should change our prior beliefs about $\Delta(a)$. In Section 2.1 we follow up on the inferential consequences of this.

At a high level, the source of the problem in our illustrative examples is the following well-known phenomenon, which we refer to as the *orthogonality principle*.

Principle 1 (The Orthogonality Principle). Let $\widetilde{\beta}$ and $\widetilde{\phi}$ be random unit vectors with mean 0 taking values in some high/infinite dimensional Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$. Then, if $\widetilde{\phi}$ and $\widetilde{\beta}$ are independent and there is no dimension reducing structure in the problem, with high probability, we have $\langle \widetilde{\beta}, \widetilde{\phi} \rangle \approx 0$.

This principle emerges from the geometric properties of high-dimensional spaces (Wegner 2021; Vershynin 2018) and variants of it have proven useful in many problems; one example among many is that, in compressed sensing, random Gaussian ensembles satisfy the restricted isometry property with high probability (Candes and Tao 2007). Examples of "dimension reducing structure" include high degrees of anisotropy of either the distribution of the random vectors or of $\langle \cdot, \cdot \rangle$, which can cause the vectors to behave more like low-dimensional vectors.

The orthogonality principle becomes important when P is large (or in nonparametric problems) because Δ is quantifiable in terms of $\langle \beta, \phi \rangle$ for some suitable inner product (see Proposition 1). If IG.2 holds then the orthogonality principle immediately suggests $\langle \beta, \phi \rangle \approx 0$ with high probability, implying that our prior is dogmatic about the confounding bias.

2.1. Asymptotics for High-Dimensional Ridge Regression

While the dogmatism implied by Proposition 2 is troubling, one might hope that the informative prior on Δ is a theoretical curiosity that is nevertheless swamped by the data. We show that this is not the case, and that the prior concentration on Δ leads to heavily biased inferences if P grows sufficiently quickly with N. We summarize our main results as follows.

- In the regime $P/N \to r$ for some $r \in (0, \infty)$ (i.e., P grows at the same rate as N), the Bayes estimator that takes a flat prior on γ and a Gaussian prior $\beta \sim \mathrm{Normal}(0, \tau^2 P^{-1}\,\mathrm{I})$ is heavily biased. Specifically, when confounding bias is present through the auxiliary covariate $\widehat{A}_i = X_i^\top \phi$, the Bayes estimate will have bias of order $\Delta(1)$.
- In some sense the setting $\Sigma = \mathrm{I}$ is inherently difficult, and the problem is generally easier when the components of X_i are highly correlated. We return to this point in Section 5.

We make two sets of assumptions. The first (high-dimensional asymptotics, or HDA) is used to describe the distribution of the X_i's as $N \to \infty$. The second (random effects model, or REM) describes a particular random effects model for the regression coefficients. This framework modifies the framework of Dobriban and Wager (2018) so that it is suitable for our aims.

HDA.1 The covariates are multivariate normal with $X_i \sim \mathrm{Normal}(0, \Sigma)$.

HDA.2 As $N \to \infty$ we have $P/N \to r$ for some $r \in (0, \infty)$.

HDA.3 The spectral distribution $\sum_{p=1}^P \delta_{\lambda_p}/P$ associated to Σ converges to some limiting distribution H on $[0, \infty)$, where $\lambda_1, \dots, \lambda_P$ are the eigenvalues of Σ and δ_λ denotes a point-mass distribution at λ.

HDA is a standard assumption for understanding the case where P grows like N. HDA.3 allows us to use results from random matrix theory to compute $\lim_{P \to \infty} \mathrm{tr}\{(\boldsymbol{X}^\top \boldsymbol{X} + N\lambda\,\mathrm{I})^{-k}\}$ for $k \in \mathbb{N}$. Under HDA, the empirical distribution of the eigenvalues of $\underline{S} = \boldsymbol{X}\boldsymbol{X}^\top/N$, namely $\widehat{F}(dx) = N^{-1}\sum_{i=1}^N \delta_{\lambda_i(\underline{S})}$, converges to a distribution $F(dx)$ called the *empirical spectral distribution*.

Next, we describe a random effects model (REM) for β and ϕ that we will base our analysis on. Similar models have been used to study both the prediction risk and minimax-optimality of ridge regression (Dicker 2016; Dobriban and Wager 2018). REM is a fruitful assumption for us as it allows exact formulas for the bias to be derived that are free of the particular values of β and ϕ. In Remark 2 we discuss relaxing this assumption.

REM.1 The coefficient vector ϕ is randomly sampled as $\phi \sim \mathrm{Normal}(0, \tau^2 P^{-1}\,\mathrm{I})$.

REM.2 The coefficient vector β is randomly sampled as $\beta \sim \mathrm{Normal}(\omega_0\,\phi, \tau^2 P^{-1}\,\mathrm{I})$.

REM.3 Given β and ϕ, $Y_i \sim \mathrm{Normal}(X_i^\top \beta + A_i\,\gamma_0, 1)$ and $A_i \sim \mathrm{Normal}(X_i^\top \phi, 1)$.

We note that REM.2 is equivalent to setting $Y_i \sim \mathrm{Normal}(X_i^\top b + \omega_0\,\widehat{A}_i + \gamma_0\,A_i, 1)$, where $\widehat{A}_i = X_i^\top \phi = \mathbb{E}(A_i \mid X_i, \phi)$ and $b \sim \mathrm{Normal}(0, \tau^2 P^{-1}\,\mathrm{I})$. REM.2 allows for non-negligible confounding bias to enter the model, and priors based

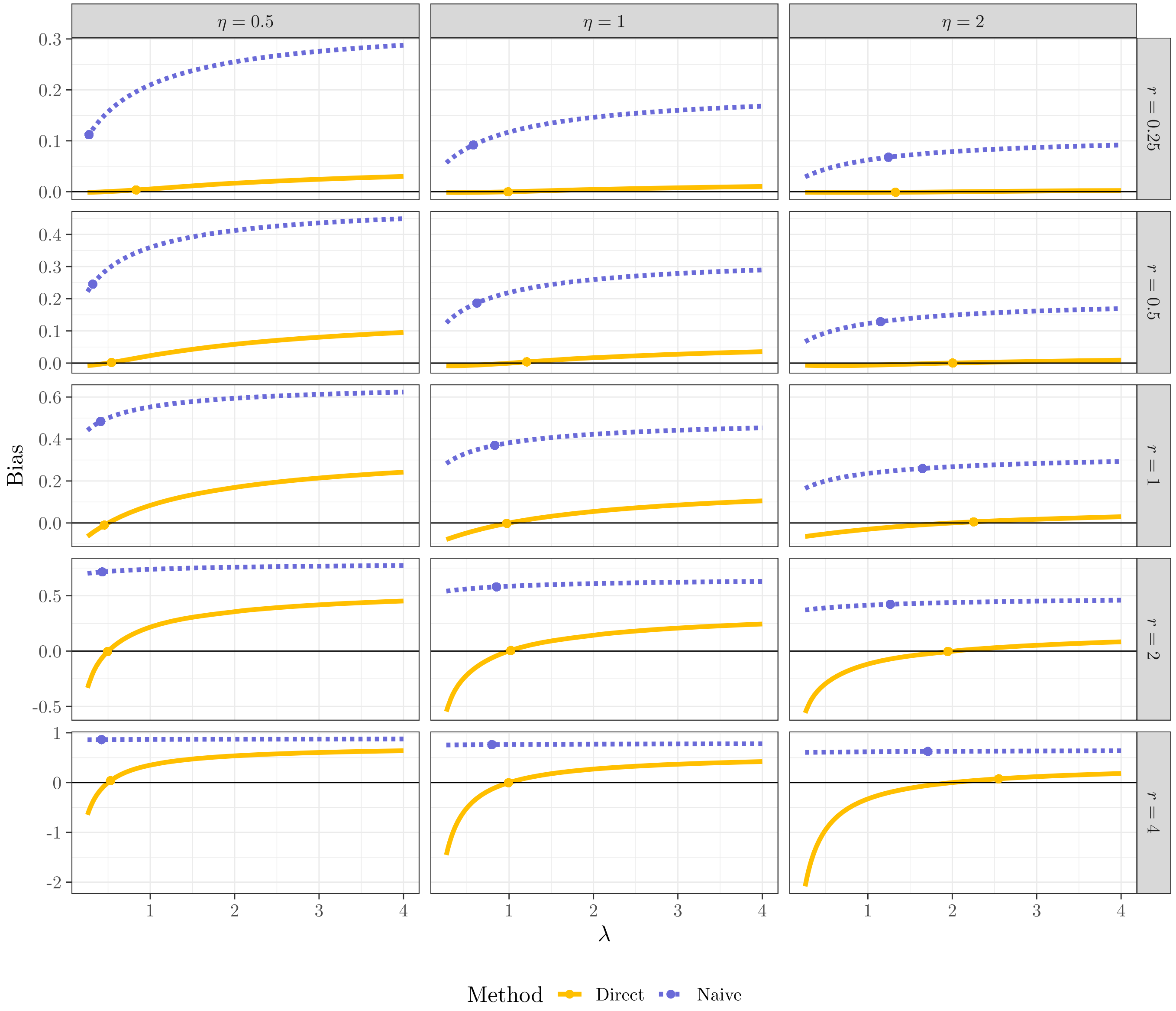

Figure 1. Comparison of the bias of naive ridge regression (dashed blue) to the direct prior (solid, orange) of Section 3.1 for different values of η and r with $\omega_0 \equiv 1$. The points on each line correspond to a value of the ridge parameter λ that obtained from estimating λ via empirical Bayes on a single dataset simulated according to the model and prior.

on this parameterization have been used to account for confounding bias by other researchers (Zigler et al. 2013; Hahn et al. 2018). The parameter ω_0 is closely connected to the confounding bias.

Proposition 3. Suppose that HDA and REM hold and that Σ satisfies the conditions of Proposition 2. Then $\Delta(1) \to \omega_0 \frac{\tau^2 \widetilde{\lambda}}{1+\tau^2 \widetilde{\lambda}}$in probability as $P \to \infty$.

Theorem 1 explicitly computes the bias of the ridge regression estimator under IG.2 when the prior $\beta \sim \text{Normal}(0, N^{-1}\lambda^{-1}\text{I})$ is used, that is, when we apply the usual ridge regression estimator. We sketch a proof of Theorem 1 and verify it numerically in the supplementary material.

Theorem 1. Suppose HDA and REM hold. Let $(\widetilde{\gamma}, \widetilde{\beta}^\top)^\top$ denote the Bayes estimate of $(\gamma, \beta^\top)^\top$ under a prior that takes $\beta \sim \text{Normal}(0, N^{-1}\,\lambda^{-1}\text{I})$ and places a flat prior on γ under IG.2. Then the asymptotic bias of $\widetilde{\gamma}$ is given by

$$\lim_{N,P\to\infty} \mathbb{E}(\widetilde{\gamma} - \gamma_0) = \frac{\omega_0 \int x/(x+\lambda)\; F(dx)}{\int (x+\eta)/(x+\lambda)\; F(dx)} = \omega_0 \times \frac{1 - \lambda\, \nu(-\lambda)}{1 - (\lambda - \eta)\, \nu(-\lambda)} \qquad (1)$$

where $\nu(z) = \int_0^\infty \frac{F(dx)}{x-z}$ is the Stieltjes transform of $F(dx)$ and $\eta = r/\tau^2$.

Ideally we would like the bias to be close to 0 for moderate-to-large values of λ so that we have both small variance and bias; the approach outlined in Section 3.1 *does* accomplish this goal for a properly chosen λ. Figure 1 contrasts this alternative method with standard ridge regression when $\Sigma = \text{I}$ and we see that the bias is quite large for ridge regression unless λ is close to 0 and $r \leq 1$; this latter case corresponds to OLS, which (while unbiased) defeats the purpose of using ridge regression.

Additionally, a data-guided choice of λ (obtained via empirical Bayes on single dataset simulated with $\omega_0 = 1$ and $\gamma_0 = 3$) does not result in a low-bias estimate of γ, whereas a data-guided choice of the tuning parameter of our proposed approach does.

A qualitative observation based on (1) is that a smaller bias is obtained when most of the eigenvalues of $\underline{S}$ are small. For example, unbiasedness is possible (even if $P \gg N$) if $\underline{S}$ is rank deficient, because $F(dx)$ will assign mass to 0, which will cause $\lambda\, v(-\lambda) \to 0$ (by bounded convergence) while $\eta\, v(-\lambda) \to \infty$ as $\lambda \to 0$.

When $P > N$ the only hope for nonnegligible bias is for the eigenvalues of $\underline{S}$ to be heavily concentrated near 0. As $\underline{S}$ has the same nonzero eigenvalues as the sample covariance $S = \boldsymbol{X}^\top \boldsymbol{X}/N$ this means we should hope for strong colinearities among the covariates. A particularly unfavorable setting is $\Sigma = \mathrm{I}$, where the Marchenko-Pastur theorem (see, e.g., Couillet and Debbah 2011, Theorem 2.13) states that if $r \geq 1$ then $F(dx)$ has density $q(\lambda) = \frac{\sqrt{(b-\lambda)(\lambda-a)}}{2\pi\,\lambda} I(a < \lambda < b)$ where $(a, b) = (1 \pm \sqrt{r^{-1}})^2$; this bounds the support of the eigenvalues away from 0. In Section 5 we show that much better results are obtained when the X_i's follow a latent factor model.

Remark 2. REM is a strong assumption, and one might worry that conclusions drawn under REM do not generalize to other settings. In the supplementary material, we study the setting $r < 1$ when REM does not hold and show that the g-prior ($\phi, \beta \sim \mathrm{Normal}(\mathbf{0}, \lambda^{-1} N^{-1} S^{-1})$) leads to inconsistent estimation of γ (Liang et al. 2008) unless $\lambda \to 0$ with N (with the bias proportional to the confounding bias); moreover, the λ used in the g-prior that is *optimal* for prediction purposes can be shown to not converge to 0, suggesting that inconsistency will remain if we place a fixed prior on λ. On the other hand, a g-prior variant of the model described in Section 3.1 is shown to be $\sqrt{N}$-consistent. We argue on this basis that the conclusions drawn under REM are also representative of what occurs when ϕ and β are instead regarded from the Frequentist perspective as fixed-but-unknown parameters, while also providing additional insight into the role that the spectral distribution of Σ plays and being applicable for $r \geq 1$.

Remark 3. As noted by a reviewer, one might also worry that our conclusions are partially driven by the effect of scaling by P in REM.1, and REM.2, or the scaling by N in the prior distribution $\beta \sim \mathrm{Normal}(0, N^{-1}\lambda^{-1}\,\mathrm{I})$. Instead, we might have analyzed the behavior of the ridge regression estimator based on the $\mathrm{Normal}(0, \lambda^{-1}\,\mathrm{I})$ under the REM assumption $\phi \sim \mathrm{Normal}(0, \tau^2\,\mathrm{I})$ and $\beta \sim \mathrm{Normal}(\omega_0\,\phi, \tau^2\,\mathrm{I})$. We feel that this setting is less natural in the high-dimensional setting, as it implies the signal-to-noise ratios (SNRs) $\|\phi\|^2$ and $\|\beta\|^2$ for both the $\boldsymbol{A}$ and $\boldsymbol{Y}$ models diverge; typically, the optimal choice of λ for prediction purposes in our setup is on the same order as the SNRs. In the supplementary material we show that allowing the SNRs to diverge still results in asymptotic bias; specifically, the bias converges to $\omega_0 \times r^\star/\{1 - (1 - r^\star)(1 - \eta/\lambda)\}$ where $r^\star = \min(r, 1)$. We note that, in this regime, the covariance matrix Σ does not play any role.

Remark 4. That the most favorable situation occurs when Σ is nearly low-rank is the *opposite* of the most favorable situation for estimating the regression coefficients ($\Sigma = \mathrm{I}$). A related phenomena is described by Dobriban and Wager (2018), who show that inference and prediction are generally at odds with each other when REM holds.

2.2. *Confounding Bias Dogmatism for Semiparametric Regression*

Recall the semiparametric regression problem described in Section 1.2 with the confounding bias parameter given in Proposition 1. For convenience, we will assume that $\phi(x)$ is known a-priori to be $\phi_0(x)$ and that $\mu, \tau \overset{\text{iid}}{\sim} \mathrm{GP}(0, \kappa)$ (Rasmussen and Williams 2006); the statement $\beta \sim \mathrm{GP}(m, \kappa)$ here means that, for any finite collection $(x_1, \dots, x_M)$, we have $\big(\beta(x_1), \dots, \beta(x_M)\big)^\top \sim \mathrm{Normal}(\boldsymbol{m}, \boldsymbol{K})$ where $\boldsymbol{m} = \big(m(x_1), \dots, m(x_M)\big)^\top$ and $\boldsymbol{K}$ has (j,k)th entry $\kappa(x_j, x_k)$. Gaussian processes have been proposed as priors for causal inference by several authors (Ray and van der Vaart 2020; Ren et al. 2021) and they are convenient to study theoretically.

Figure 2 gives a sense of what to expect. In this figure, β, τ, and $\mathrm{logit}(\phi)$ are sampled from Gaussian processes with squared exponential kernel $\kappa(x, x') = e^{-\|x - x'\|_2^2/2}$. As P increases we see that Δ concentrates around 0. As in the setting of ridge regression, this is troubling both because (i) it will typically violate our prior beliefs about Δ for large P and (ii) given the definition of Δ, there is no reason for our prior beliefs to be dependent on the number of variables that act as confounders.

We apply the orthogonality principle to the Hilbert space $\mathscr{L}_2(F_X)$ of square-integrable functions $\{g : \int g^2 \, dF_X < \infty\}$ with inner product $\langle \beta, \phi \rangle = \int \beta(x)\,\phi(x)\; F_X(dx)$, where F_X denotes the distribution of X_i. Let $\bar{g}(x) = g(x) - \int g(x)\; F_X(dx)$ and $\widetilde{g}(x) = \bar{g}(x)/\|\bar{g}\|$. The following proposition shows that the confounding bias is controlled by $\langle \widetilde{\mu} + \widetilde{\tau}, \widetilde{\phi} \rangle$, implying that the orthogonality principle is in effect.

Proposition 4. Suppose $\mathbb{E}\{\mu(X_i)^2\}$ and $\mathbb{E}\{\tau(X_i)^2\}$ are bounded as $P \to \infty$, and that there exists a $\delta > 0$ such that $\Pr_{\phi_0}(A_i = a) > \delta$ for $a = 0, 1$. Then $\Delta = \frac{\|\bar{\mu}\|\,\|\bar{\phi}\|}{\Pr_{\phi_0}(A=1)\Pr_{\phi_0}(A=0)} \langle \widetilde{\mu}, \widetilde{\phi} \rangle + \frac{\|\bar{\tau}\|\,\|\bar{\phi}\|}{\Pr_{\phi_0}(A=1)} \langle \widetilde{\tau}, \widetilde{\phi} \rangle = O_p(\langle \widetilde{\mu} + \widetilde{\tau}, \widetilde{\phi} \rangle)$

For Gaussian process models, and nonparametric models in general, model complexity tends to scale very rapidly in P. To connect our results for Gaussian processes to ridge regression, we define the effective number of parameters of $f \sim \mathrm{GP}(0, \kappa)$ as

$$D = \frac{\int \kappa(x, x)\; F_X(dx)}{\int \kappa(x, x')\; F_X(dx)\; F_X(dx')} = \frac{\mathrm{var}\{f(X)\}}{\mathrm{var}\{\int f(x)\; F_X(dx)\}}.$$

The intuition behind this definition, which is a natural generalization of Kish's effective sample size to stochastic processes (Kish 1965), is that $\int f(x)\; F_X(dx)$ is an average of infinitely many of the $f(x)$'s, and its variance should be roughly (i) the variance of one of its constituents $f(x_0)$ divided by (ii) the number of "independent" entities being averaged D; to account for the fact that the different evaluations $f(x_0)$ might have different variances, we average the variance against F_X to get $\mathrm{var}\{f(X)\}$ in the numerator.

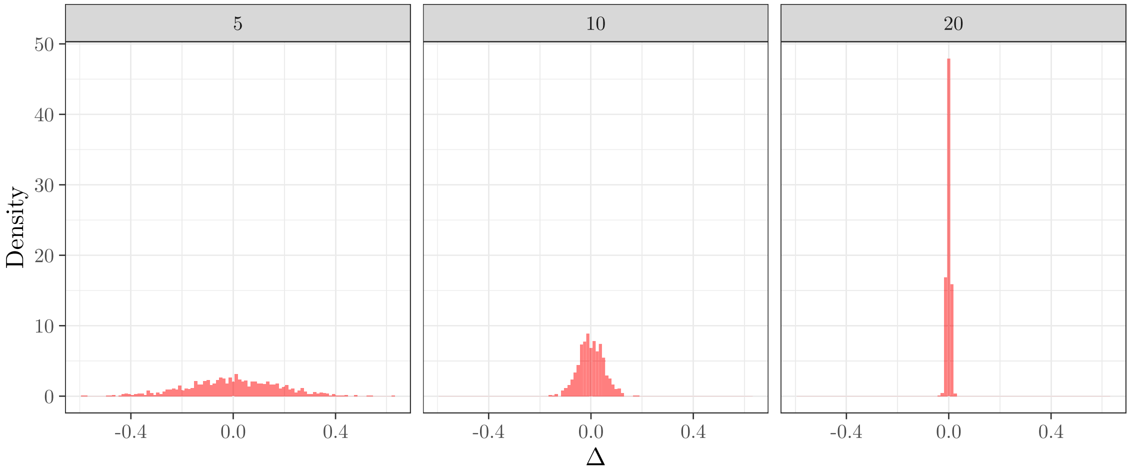

Figure 2. Prior distribution of Δ under a Gaussian process prior in Section 2.2 for $P \in \{5, 10, 20\}$.

We now show that the variance of $\langle\widetilde{\mu} + \widetilde{\tau}, \widetilde{\phi}\rangle$ scales inversely with D and that, for the squared-exponential kernel, D grows quickly in both P and the inverse length-scale.

Proposition 5. Let $\mu, \tau \overset{\text{indep}}{\sim} \text{GP}(0, \tau_\beta^2\,\rho)$ where $\rho(x, x')$ is a correlation function. Then $\Delta \sim \text{Normal}(0, c)$ where $c \le \frac{\mathcal{W}\,\tau_\beta^2}{D}$ for a constant $\mathcal{W}$ that depends only on $\Pr_{\phi_0}(A_i = 1)$ and $\Pr_{\phi_0}(A_i = 0)$.

The conclusion here is similar to the conclusion for ridge regression, with the prior concentration depending inversely on the effective number of parameters D rather than the raw dimensionality P. Next, we consider the specific case of the squared-exponential kernel, which is among the most commonly used kernels for performing Gaussian process regression.

Proposition 6. Let $\kappa(x, x') = \tau_\beta^2 \rho(x, x')$ where $\rho(x, x') = \exp\{-(x - x')^\top H^{-1}(x - x')/2\}$ and H is a covariance matrix, and suppose $X_i \sim \text{Normal}(0, \Sigma)$. Then

$$D = \sqrt{\frac{\det(H + 2\Sigma)}{\det(H)}} \ge \left\{1 + 2\left(\frac{\det(\Sigma)}{\det(H)}\right)^{1/P}\right\}^P.$$

In particular, if $H = \ell^2\Sigma$ then $D = (1 + 2/\ell^2)^{P/2}$, while if $H = \ell^2\,\mathrm{I}$ then $D = \prod_{j=1}^P (1 + 2\lambda_j(\Sigma)/\ell^2)^{1/2}$.

Regarding P as fixed, we see that D grows like ℓ^{-P} as $\ell \to 0$, implying $\text{var}(\Delta) = O(\ell^P)$. Consequently, reducing the length-scale of the process quickly leads to prior dogmatism. As the following corollary shows, letting P diverge makes the problem much worse.

Corollary 1. Under the same conditions as Proposition 6, we have $D \ge \exp(CP)$ for some constant C as $P \to \infty$, provided that $\det(\Sigma)^{1/P}/\det(H)^{1/P}$ is bounded; in particular, this occurs if either $H = \ell^2\Sigma$ and ℓ^2 is bounded, or if $H = \ell^2\mathrm{I}$ and $\det(\Sigma)^{1/P}/\ell^2$ is bounded.

Note that Proposition 5 shows that dogmatism occurs in a *uniform* sense: no matter how favorably $\phi(x)$ is selected, the naive use of Gaussian process priors causes the prior variance on Δ to scale inversely in D. In the case of the squared-exponential kernel, D grows exponentially in P for reasonable choices of H, including the commonly used isotropic ($H = \ell^2\mathrm{I}$) and an anistropic kernel that makes the prior invariant to linear transformations of the predictors ($H = \ell^2\Sigma$).

3. Correcting for Dogmatism

3.1. Direct Priors for Ridge Regression

A simple approach to addressing dogmatism for ridge regression is to make $\beta^\top\Sigma\phi$ large by encouraging β to align with ϕ. For example, we might center β on ϕ by taking $\beta \sim \text{Normal}(\omega\phi, \tau_\beta^2\,\mathrm{I})$. Doing this, we now have $\Delta(a) = a\frac{\phi^\top\Sigma b}{\sigma_a^2+\phi^\top\Sigma\phi} + a\,\omega\frac{\phi^\top\Sigma\phi}{\sigma_a^2+\phi^\top\Sigma\phi}$,where $b \sim \text{Normal}(0, \tau_\beta^2\,\mathrm{I})$. By the same argument as in Proposition 2, the first term is $O_p(P^{-1/2})$; the second term, however, does not tend to 0 as $P \to \infty$, preventing prior dogmatism from taking hold. This allows us to place a *direct prior* on $\Delta(a)$ by placing a prior on ω. For example, following common practice, we might attempt to express "ignorance" about the degree of confounding bias by placing a flat prior on ω.

This approach is related to the targeted maximum likelihood estimation strategy of introducing a "clever covariate" into the outcome model to account for confounding (see, e.g., van der Laan and Rose 2011, sec. 4.2.1). The parameterization $\beta = b + \omega\phi$ gives $Y_i(a) = \beta_0 + X_i^\top b + \omega(X_i^\top\phi) + \gamma\,a + \epsilon_i(a)$,which effectively introduces the new covariate $\widehat{A}_i = X_i^\top\phi$ into the model. A related idea proposed by (Hahn et al. 2018) is to replace a in the outcome model with the residual $(a - \widehat{A}_i)$, which is equivalent to setting $\omega = -\gamma$.

In practice, rather than jointly modeling (β, ϕ) it may be more convenient to set $\widehat{A}_i = X_i^\top\widehat{\phi}$ for some point estimator of ϕ — for example, $\widehat{\phi}$ might be obtained via ridge regression. In addition to being easier to implement, this also reduces the risk of *model feedback* occurring when one of the models is misspecified (Zigler et al. 2013). In the Supplementary Material we show that, when $\widehat{\phi}$ is the Bayes estimator obtained from a g-prior and ω is given a flat prior, this strategy results in a $\sqrt{N}$-consistent estimator of γ if both the exposure and outcome models are correctly specified. We also study the bias induced by ridge regression (rather than the g-prior) under HDA and REM for the direct prior.

Evaluation of the Direct Prior via Simulation

We conduct a simulation study to determine if there is any benefit to using the direct prior relative to either (i) the naive ridge regression prior or (ii) the approach of (Hahn et al. 2018), which we call the "debiased" approach (equivalent to fixing $\omega = -\gamma$). In all cases we set $N = 200$ and $P = 1000$ so that $N \ll P$. We consider a dense model with $\phi = (1, \dots, 1)/\sqrt{P}$, $\beta \sim \text{Normal}(0, P^{-1}\text{I})$, and $\sigma_a \equiv 1$. The methods differ in the treatment effect size γ and the degree to which the coefficients are shifted in the direction of ϕ. We considered four simulation settings.

Fixed We set $\gamma = 2$ and $\omega = -\gamma/4$ so that β is shifted in the direction of ϕ, but not by the amount implied by the debiased approach ($\Delta \approx -1/4$).

Hahn We set $\gamma = 2$ and $\omega = -2$ so that β is shifted in the direction of ϕ by exactly the amount implied by the debiased approach ($\Delta = 1/2$).

Naive We set $\gamma = 2$ and $\omega = 0$ so that the model corresponds precisely to the naive ridge model ($\Delta \approx 0$).

Mixed We set $\gamma = 1$ and $\beta_j \sim \text{Normal}(-\delta_j \phi_j, P^{-1})$, where $\delta_j = 1$ for $j < 500$ and $\delta_j = 0$ otherwise. Note that neither the naive ridge prior nor the direct prior hold under this setting ($\Delta \approx -1/4$).

The simulation was replicated 200 times for each setting and with $\sigma_y =\in \{1, 2, 4, 8, 16\}$. We evaluated each procedure according to the following criteria. **Coverage:** The proportion of nominal 95% credible intervals that capture the true value of γ. **Width:** The average width of the nominal 95% credible interval. **Avg SE:** The average estimated standard error from the model, that is, the posterior standard deviation of γ averaged over all replications. **RMSE:** The root mean squared error in estimating γ with the Bayes estimator $\widehat{\gamma}$.

Results are compiled in Figure 3. The direct and debiased approaches always attain the nominal coverage level, while the naive approach does not come close when the confounding bias is non-negligible unless the signal-to-noise ratio is exceedingly small. We also see that the debiased approach generally requires

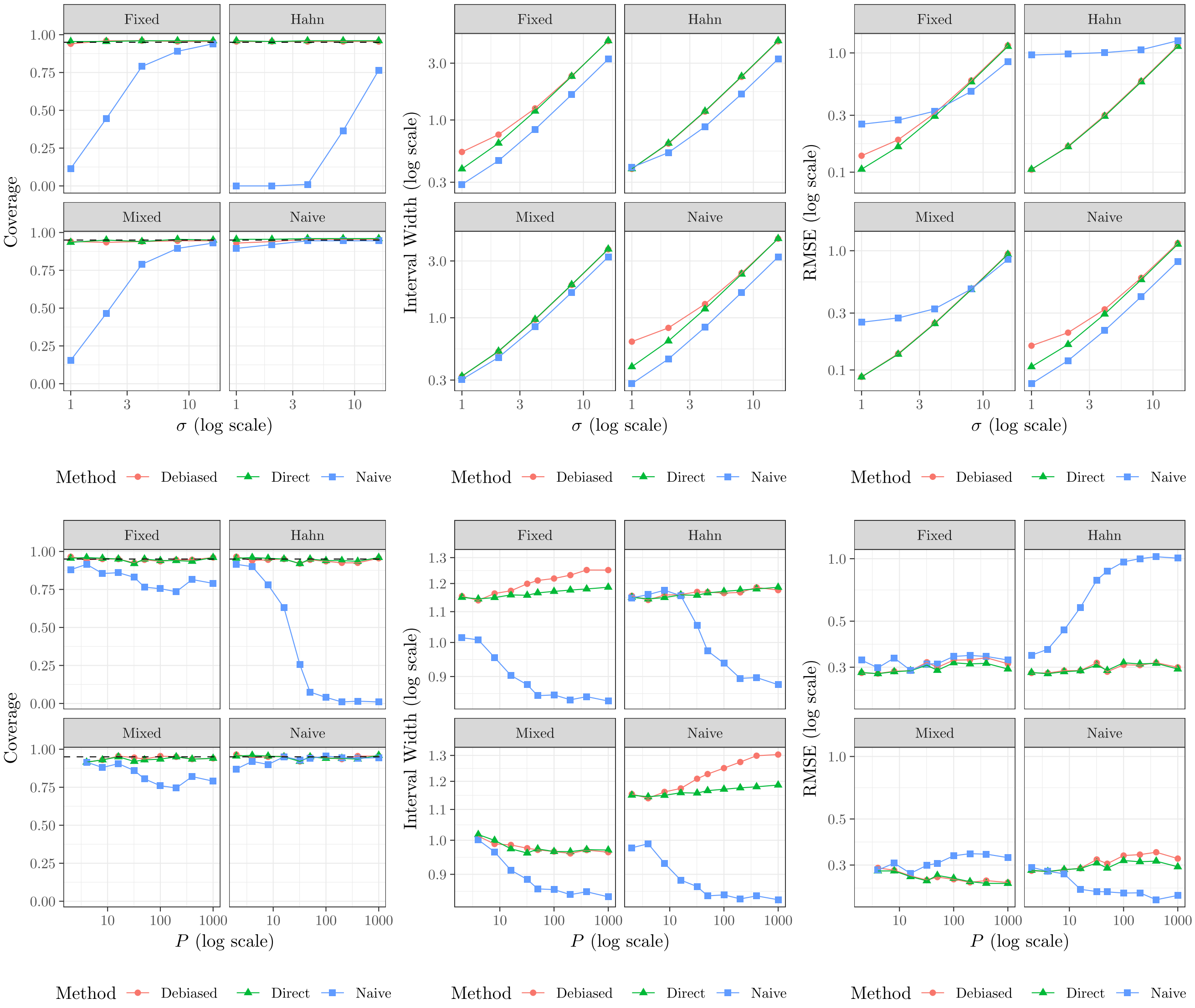

Figure 3. Results for the simulation setting of Section 3.1. The top panels vary σ_y with $P \equiv 1000$ while the bottom panels fix $\sigma_y \equiv 4$ and vary P. Left: the coverage of nominal 95% confidence/credible intervals. Middle: Interval width (log scale). Right: root mean-squared error (log scale) of Bayes estimates.

larger intervals than the direct approach to cover at the appropriate rate. The only exception is under the Mixed and Hahn settings; this is expected because Mixed and Hahn set $\omega = -\gamma$, which is implicitly assumed by the debiased approach. The naive ridge prior only performs well when it is correctly specified ($\omega = 0$), in which case it is unsurprisingly the best method.

Also included in Figure 3 are results for a simulation that fixes $(\sigma, N) \equiv (4, 200)$, with P ranging from 4 to 1000. The results in here broadly agree with our previous conclusions, with the exception that the Naive model performs reasonably well in terms of coverage when $P \ll N$ but performs poorly as P approaches the scale of N.

Additional simulation results that take $\gamma = 0$ with $\Delta = \frac{\omega}{2}$ and σ varying Figure 4. We see that the direct prior and the approach of Hahn, Murray, and Carvalho (2017) are to be preferred unless either Δ is small or σ is large; for small values of Δ this is logical, as the ridge model is only slightly misspecified and is slightly more parsimonious. For reference, the signal-to-noise ratio is roughly $(1 + 2\Delta)^2/\sigma^2$, and we see that even with signal-to-noise ratios as low as 0.25 ($\Delta = 0.6, \sigma = 4$) the coverage of the Bayesian ridge regression model is very poor.

3.2. Semiparametric Regression with Clever Covariates

Mimicking our strategy in Section 3.1, for the semiparametric regression problem we propose setting $\mu(x) = \mu^\star(x) + g\{\phi(x)\}$ and $\tau(x) = \tau^\star(x) + h\{\phi(x)\}$ for some choice of functions $g(\cdot)$ and $h(\cdot)$, with $\mu^\star$ and $\tau^\star$ given independent Gaussian process priors independent of $\phi(\cdot)$. The confounding bias is then given by

$$\begin{aligned}\Delta &= \frac{\operatorname{cov}_\theta[\mu^\star(X_i) + g\{\phi(X_i)\}, \phi(X_i)]}{\Pr_\phi(A_i = 0)\ \Pr_\phi(A_i = 1)} \\ &\quad + \frac{\operatorname{cov}_\theta[\tau^\star(X_i) + h\{\phi(X_i)\}, \phi(X_i)]}{\Pr_\phi(A_i = 1)} \\ &\approx \frac{\operatorname{cov}_\theta[g\{\phi(X_i)\}, \phi(X_i)]}{\Pr_\phi(A_i = 0)\ \Pr_\phi(A_i = 1)} + \frac{\operatorname{cov}_\theta[h\{\phi(X_i)\}, \phi(X_i)]}{\Pr_\phi(A_i = 1)}\end{aligned}$$

by the orthogonality principle. The confounding bias does not concentrate for 0 on this model because $\phi(X_i)$ will generally be highly correlated with $g\{\phi(X_i)\}$ and $h\{\phi(X_i)\}$, even if these functions are modeled nonparametrically.

There are several considerations for choosing g and h. If we are concerned strictly with obtaining good Frequentist properties, an appropriate choice is to take $g(\phi) + a\,h(\phi) = \omega\frac{a-\phi}{\phi(1-\phi)}$ and place a flat prior on ω; when ϕ is known, this guarantees $\sqrt{N}$-consistency. Alternatively, we can set $g, h \sim \mathrm{GP}(0, \kappa_g)$ with the covariance function $\kappa_g(\phi, \phi') = \tau_g^2 \exp\{-(\phi - \phi')^2/(2s_g^2)\}$. This choice of covariance function was noted by Ren et al. (2021) to induce matching on the propensity score: individuals with similar propensity scores have their values of $g(\phi)$ and $h(\phi)$ shrunk together. The penalized-spline-of-propensity approach of Zhou, Elliott, and Little (2019) is similar, except that splines are used instead of Gaussian processes.

Simulation Experiment

We use the simulation setting of Hahn, Murray, and Carvalho (2020, sec. 6.1) to evaluate several different approaches to correcting a Gaussian process prior for dogmatism. We consider the generative model $Y_i(a) = \mu(X_i) + a\,\tau(X_i) + \epsilon_i, \epsilon_i \sim \mathrm{Normal}(0, 1)$with $X_{i2} \sim \mathrm{Bernoulli}(1/2)$, $X_{i4} \sim \mathrm{Uniform}(\{1, 2, 3\})$, and the other covariates iid $\mathrm{Normal}(0, 1)$. We let

$$\tau(x) = \begin{cases} 3 & \text{homogeneous,} \\ 1 + 2\,x_2\,x_5 & \text{heterogeneous,} \end{cases} \quad \text{and}$$
$$\mu(x) = \begin{cases} 1 + g(x_4) + x_1\,x_3 & \text{linear,} \\ -6 + g(x_4) + 6\,|x_3 - 1| & \text{nonlinear,} \end{cases} \tag{2}$$

where $g(1) = 2, g(2) = -1$, and $g(3) = -4$. We then set $A_i \sim \mathrm{Bernoulli}\{\phi(X_i)\}$ with $\phi(x) = 0.8\,\Phi\{3\,\mu(x)/s - 0.5\,x_1\} + 0.1$, where s is the empirical standard deviation of the $\mu(X_i)$'s. In total we consider 16 possible simulation settings, corresponding to a factorial design with $N \in \{250, 500\}$, $P \in \{5, 20\}$, and the four combinations of linear/nonlinear and homogeneous/heterogeneous. We model $\mathbb{E}\{Y_i(a) \mid X_i = x\} = \beta(a, x)$ using a Gaussian process $\beta \sim \mathrm{GP}(0, \kappa)$ with the following choices of $\kappa\big((a, x), (a, x')\big)$

Naive A kernel that makes no correction for dogmatism: $\kappa\big((a, x), (a', x')\big) = 100(1 + a\,a') + \lambda\,\exp\{-b\|(a, x) - (a', x')\|_2^2\}$.

IPW-GP A kernel that incorporates the inverse propensity score linearly as a "clever covariate": $\kappa\big((a, x), (a', x')\big) = 100(1 + a\,a' + w\,w' + z\,z') + \lambda\,\exp\{-b\|(a, x) - (a', x')\|_2^2\}$ where $w = a/\phi(x)$ and $z = (1 - a)/(1 - \phi(x))$.

Spline-of-propensity-GP A kernel that incorporates the propensity score using a spline basis function expansion: $\kappa\big((a, x), (a', x')\big) = 100(1 + a\,a' + \sum_k \psi_k\,\psi_k') + \lambda\,\exp\{-b\|(a, x) - (a', x')\|_2^2\}$ where $\psi_k = \psi_k(x), \psi_k' = \psi(x')$, and $\{\psi_1, \dots, \psi_K\}$ are natural cubic spline basis functions using 10 knots (see Zhou, Elliott, and Little 2019, for related methods).

Spline-of-propensity Same as spline-of-propensity-GP but without the Gaussian kernel.

In order to separate the issue of accurately estimating the propensity scores from the benefit of using them, we assume that $\phi(x)$ is known a-priori. The parameters (λ, b, σ_y) were estimated via empirical Bayes (see Rasmussen and Williams 2006, sec. 5.4.1). The factor of 100 in the various kernels corresponds to including linear terms in the models; for example, in the Naive kernel, inclusion of the term $100(1 + a\,a')$ corresponds to including linear terms $\beta(a, x) = \alpha_0 + \alpha_1\,a + \beta^\star(a, x)$ where $(\alpha_0, \alpha_1) \sim \mathrm{Normal}(0, 100)$ and $\beta^\star(a, x)$ is an independent Gaussian process with kernel $\lambda\,\exp\{-b\|(a, x) - (a', x')\|_2^2\}$.

Our main goals are to (i) determine the extent to which the Naive kernel suffers due to dogmatism, (ii) determine which of the IPW or spline approaches perform better in this case, and (iii) determine whether the propensity score alone is sufficient to produce a good estimator. A subset of the results corresponding to the nonlinear heterogeneous setting with $N = 250$ are given in Figure 5, with the remaining results deferred to the supplementary material. Summarizing these results, we find (i) that the Naive kernel performs well when $P = 5$ where dogmatism is mild, but breaks down completely when $P = 20$; (ii) that the IPW-GP and spline-of-propensity-GP approaches perform comparably in terms of

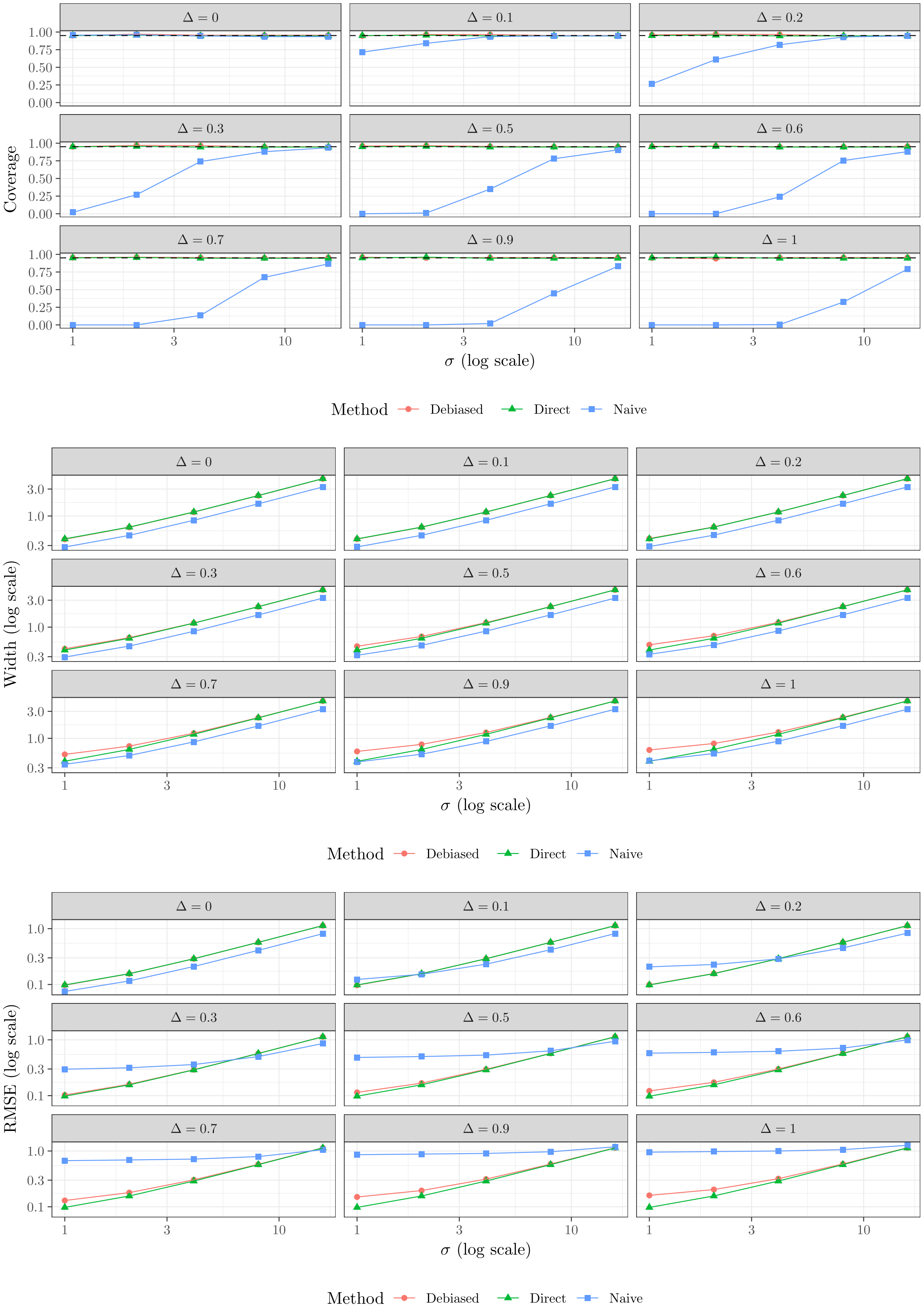

Figure 4. Simulation results for $\gamma = 0$ as Δ and σ_y vary.

coverage, but that the spline-of-propensity-GP generally produces smaller standard errors and RMSEs, suggesting that the spline-of-propensity approach is more stable while accomplishing the same goals as IPW methods; and (iii) that the spline-of-propensity-GP produces smaller standard errors and RMSEs than the spline-of-propensity approach, indicating that there is a benefit to going beyond simply adjusting for the propensity score.

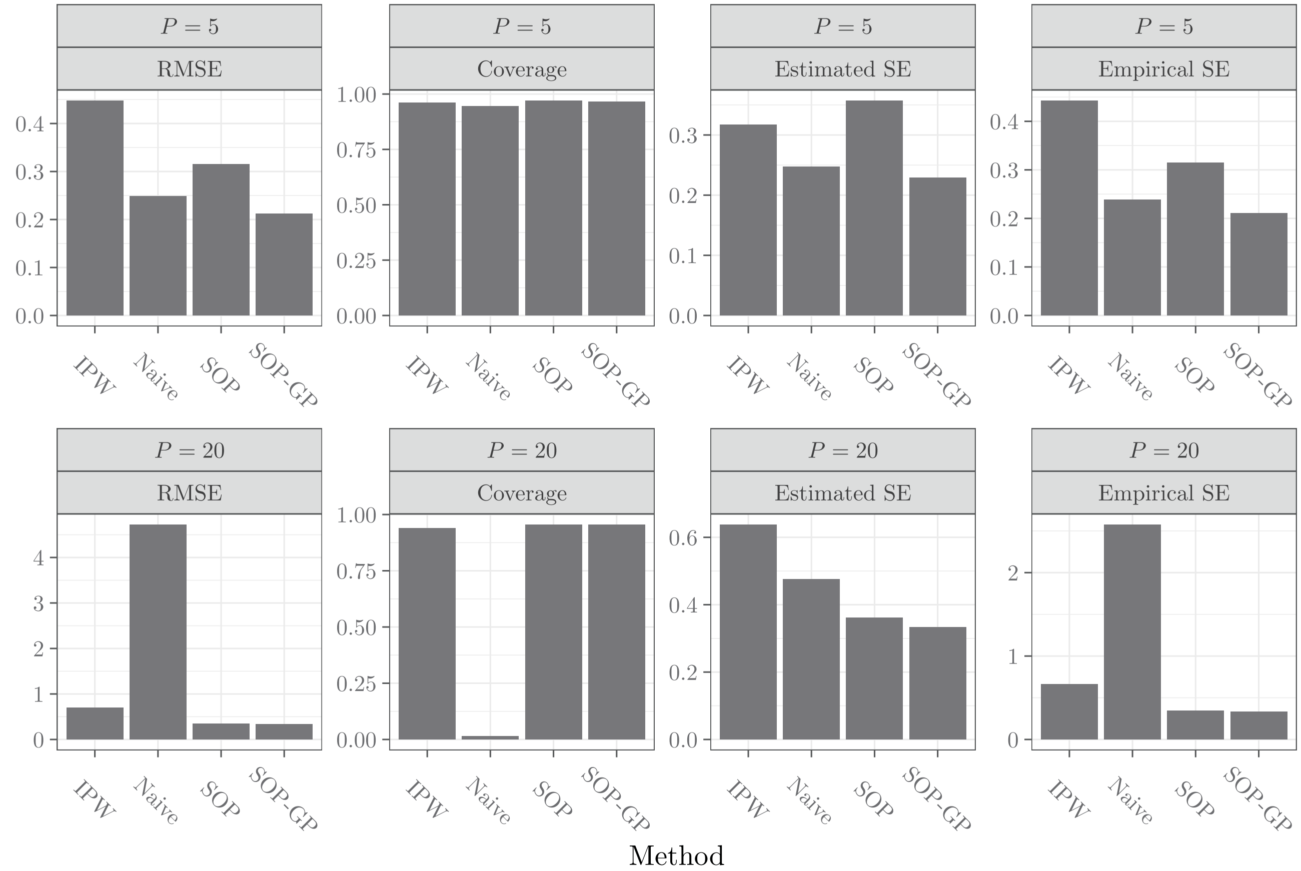

Figure 5. Results for the semiparametric manifold regression problem of Section 5. Bias denotes the average bias of $\widehat{\gamma}$, coverage denotes the coverage of nominal 95% intervals, RMSE denotes the root-mean-squared-error in estimating γ, and SE denotes the average posterior standard deviation of γ.

4. Prior Dogmatism in Other Settings

We now discuss how the proposals given here relate to other existing Bayesian proposals for correcting for confounding bias that have not been discussed above. In particular, we discuss the approaches of *propensity score stratification* and the use of *sparsity-inducing priors.*

4.1. Propensity Score Stratification

Propensity score stratification (PSS) (Imbens and Rubin 2015, chap. 17) is commonly used in applied Bayesian causal inference to robustly control for confounders when inferring average treatment effects (Rubin 1985; Li, Ding, and Mealli 2022). In this section, we relate this method to the Gaussian process methods described in Section 2.2. PSS is motivated by the fact that the propensity score is a *balancing score* in the sense that covariate distribution across treatments is exactly balanced at each level of the propensity score: $[X_i \mid \phi(X_i) = \ell, A_i = 1] \overset{d}{=} [X_i \mid \phi(X_i) = \ell, A_i = 0]$, where $\phi(x)$ denotes the propensity score $\Pr_\phi(A_i = 1 \mid X_i = x)$. For the sake of exposition, we will regard $\phi(x)$ as known, although typically this is replaced with an estimate $\widehat{\phi}(x)$ constructed using only the exposure and confounder data.

PSS stratifies the set of possible confounders $\mathcal{X}$ into J groups $\mathcal{X} = \bigcup_{j=1}^{J} B_j$ where $B_j = \{x : b_{j-1} \le \phi(x) < b_j\}$ $(0 = b_0 < b_1 < \cdots < b_J = 1)$. Separate outcome models are then specified for the different strata, with $[Y_i \mid A_i = a, X_i = x, \beta] \sim f(y \mid a, x, \beta_{B(x)})$ parameterized by $\beta_1, \dots, \beta_J$ where $B(x) = j$ if $x \in B_j$.

The within-strata models are often themselves simple; for example, we might set $[Y_i \mid A_i = a, X_i = x, \beta] \sim \text{Normal}\{\alpha_{B(x)} + \gamma_{B(x)}\, a, \sigma^2\}$, in which case the average treatment effect is $\gamma = \sum_j \Pr(X_i \in B_j)\, \gamma_j$. This can be related to the Gaussian process approach described in Section 3.2, where $\beta(a, x) \sim \text{GP}(0, \kappa)$ and κ is given by

$$\kappa(x, x') = I\{\phi(x) \sim \phi(x')\} \times \{\sigma_\alpha^2 + \sigma_\gamma^2 a\, a'\}. \qquad (3)$$

Here, $I\{\phi(x) \sim \phi(x')\}$ is the indicator that x and x' are in the same strata, while σ_α^2 and σ_γ^2 are the prior variances of α_j and γ_j under a normal prior. From this perspective, it is similar to the spline-of-propensity approach, but differs in that (i) it uses a step function rather than a cubic spline and (ii) it also smooths over the treatment effect rather than treating it as constant in x. This and other PSS approaches can therefore produce estimates equivalent to those obtained from a Gaussian process regression in which the kernel of the Gaussian process depends on the propensity score, and hence the overall approach violates IG.2.

Interestingly, despite violating IG.2, the PSS model described above can still be critiqued from the standpoint of the induced prior on the selection bias; in the supplementary material, we show that the Gaussian process with covariance (3) is itself subject to prior dogmatism when the parameters σ_α^2 and σ_γ^2 are kept fixed as the number of strata grows (specifically, $\text{var}(\Delta) = O(J^{-1})$ under the model we consider); this is not a serious concern for the applications where PSS is most commonly used, as one typically uses flat priors on the (α_j, γ_j)'s and the number of strata is not too large, but is potentially of concern in settings where the within-strata models are complex and require regularization.

4.2. Sparsity Inducing Priors

A common strategy for dealing with the $N \ll P$ setting in linear regression is to use a sparsity inducing spike-and-slab prior (Mitchell and Beauchamp 1988), where the regression coefficients are allowed to be 0 with positive probability. Even when sparsity is imposed, however, serious problems occur for the confounding bias prior. To see this, suppose that $\Sigma = \sigma_x^2 \mathrm{I}$ and let $\eth_j^\beta = I(\beta_j \neq 0)$ and $\eth_j^\phi = I(\phi_j \neq 0)$. Then we can write the confounding bias as

$$\Delta(a) = a \frac{\sum_{j:\eth_j^\beta = \eth_j^\phi = 1} \sigma_x^2 \, \phi_j \, \beta_j}{\sigma_a^2 + \sum_{j:\eth_j^\phi = 1} \sigma_x^2 \, \phi_j^2}.$$

In this case, the denominator will be of order $D_\phi \equiv \sum_j \eth_j^\phi$ while the numerator will be of order $D_{\phi\cap\beta}^{1/2}$ where $D_{\phi\cap\beta} \equiv \sum_j \eth_j^\phi \, \eth_j^\beta$. If we now use independent spike-and-slab priors for β and ϕ that are calibrated to have an on average Q active variables, we expect $D_\phi \approx Q$ while $D_{\phi\cap\beta} \approx Q/P$, so that the confounding bias will be a-priori negligible in high-dimensional sparse settings in which spike-and-slab priors are applied. Hence, even if sparsity is expected (but IG.2 is otherwise in effect) we run into essentially the same problem as with ridge regression: the prior on $\Delta(a)$ regularizes it toward zero.

To correct this issue, it remains valid to include the "clever covariate" $X_i^\top \widehat{\phi}$ in the model to correct for dogmatism. However, an alternative is to make specific use of the variable selection aspect of the model. One possibility, studied by Wang, Parmigiani, and Dominici (2012), is to use *shared variable selection* for the two models; this approach ensures that any variable appearing in the exposure model will have a high probability of also appearing in the outcome model. To implement this, we might set $\phi_j \overset{\text{iid}}{\sim} (1-p_\phi)\,\delta_0 + p_\phi \ \text{Normal}(0, \tau_\phi^2)$ and conditionally set $\beta_j \overset{\text{indep}}{\sim} \{1 - p_\beta(\phi_j)\}\,\delta_0 + p_\beta(\phi_j) \ \text{Normal}(0, \tau_\beta^2)$. Setting $p_\beta(\phi_j) = 1$ if $\phi_j \neq 0$ guarantees that β_j is included whenever ϕ_j is included.

In the supplementary material we conduct a small simulation experiment that verifies that shared variable selection is an effective strategy for combating dogmatism when the exposure and outcome models are both sparse; we also show that using a direct prior on Δ is sufficient for this purpose as well.

5. Factors Mitigating Dogmatism

In demonstrating the issue of dogmatism we made use of the orthogonality principle, which required both that β and ϕ be independent and that there be no "dimension reducing structure." We have already considered violating the independence assumption in Section 3 to control the prior on $\Delta(a)$. Alternatively, we might use dimension reducing structure. One source of possible structure is $\text{var}(X_i) = \Sigma$, which is used to define the inner product $\langle \beta, \phi \rangle = \beta^\top \Sigma \phi$; note that the spectrum of Σ appears prominently in Proposition 2 and Theorem 1. We now examine the role of Σ in the ridge and semiparametric regression problems.

5.1. Dependence Structure and Ridge Regression

We first consider a *latent factor model* that takes $X_i = \Lambda \eta_i + \sigma_x \nu_i$ where $\Lambda \in \mathbb{R}^{P \times L}$ is a matrix of factor loadings and $\eta_i \in \mathbb{R}^L$ is an L-dimensional vector of latent factors for observation i. If $\sigma_x = 0$ in this model then X_i is restricted to be in the L-dimensional subspace $\text{span}(\Lambda)$; similarly, if σ_x is small, then X_i lies very close to $\text{span}(\Lambda)$.

Consider now the induced prior on the confounding bias parameter. Assuming $\eta_i \sim \text{Normal}(0, \mathrm{I})$ and $\nu \sim \text{Normal}(0, \mathrm{I})$, we have $\text{var}(X_i) \equiv \Sigma = \Lambda\Lambda^\top + \sigma_x^2 \mathrm{I}$. Letting $\kappa_1, \dots, \kappa_L$ denote the L nonzero eigenvalues of $\Lambda\Lambda^\top$, Proposition 1 gives

$$\begin{aligned}\Delta(a) &= a \frac{\sum_{j=1}^{L} (\kappa_j + \sigma_x^2) \, W_j \, Z_j}{\sigma_a^2 + \sum_{j=1}^{L} (\kappa_j + \sigma_x^2) \, Z_j^2} + a \frac{\sum_{j=L+1}^{P} \sigma_x^2 \, W_j \, Z_j}{\sigma_a^2 + \sum_{j=L+1}^{P} \sigma_x^2 \, Z_j^2} \\ &\approx a \frac{\sum_{j=1}^{L} \kappa_j \, W_j \, Z_j}{\sigma_a^2 + \sum_{j=1}^{L} \kappa_j \, Z_j^2},\end{aligned}$$

where $W = \Gamma^\top \beta, Z = \Gamma^\top \phi$, and $\Gamma \in \mathbb{R}^{P \times L}$ consists of the L leading eigenvectors of $\Lambda\Lambda^\top$. The approximation holds when σ_x is near zero, so that Σ is approximately low-rank. Because the left-hand-side depends only on L rather than P, we expect $\Delta(a)$ to be roughly of order $L^{-1/2}$ rather than $P^{-1/2}$ for the ridge regression prior. Hence, even if $P \gg N$, we may still avoid dogmatism if $L \ll N$.

Figure 6 confirms the intuition that the naive approach avoids dogmatism when Σ is nearly low-rank; it shows the root mean squared error of the direct approach we proposed in Section 3.1 and the naive ridge regression estimator. To generate this figure, we applied the two approaches under the following conditions: $\sigma_a^2 = \sigma_y^2 = 1$; $L = 5$; $N = P = 200$; $\Lambda_{p\ell} \overset{\text{iid}}{\sim} \text{Normal}(0, 1)$; $\gamma = 1$; and both β and ϕ chosen so that $X_i^\top \beta = X_i^\top \phi = \sum_{\ell=1}^{L} \mathbb{E}(\eta_{i\ell} \mid X_i)$.

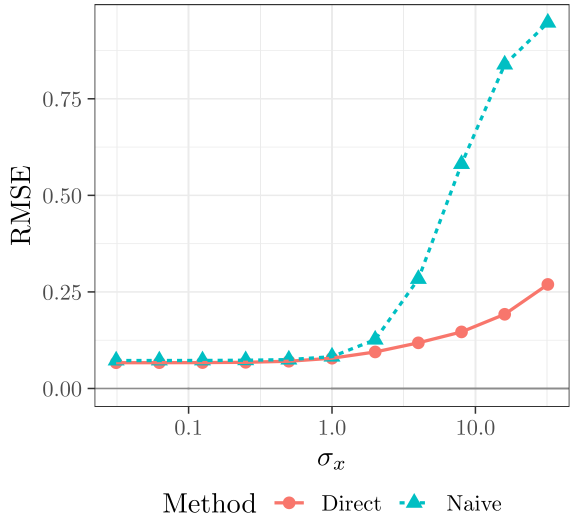

Figure 6. Root mean squared error in estimating γ for direct and naive methods as a function of σ_x.

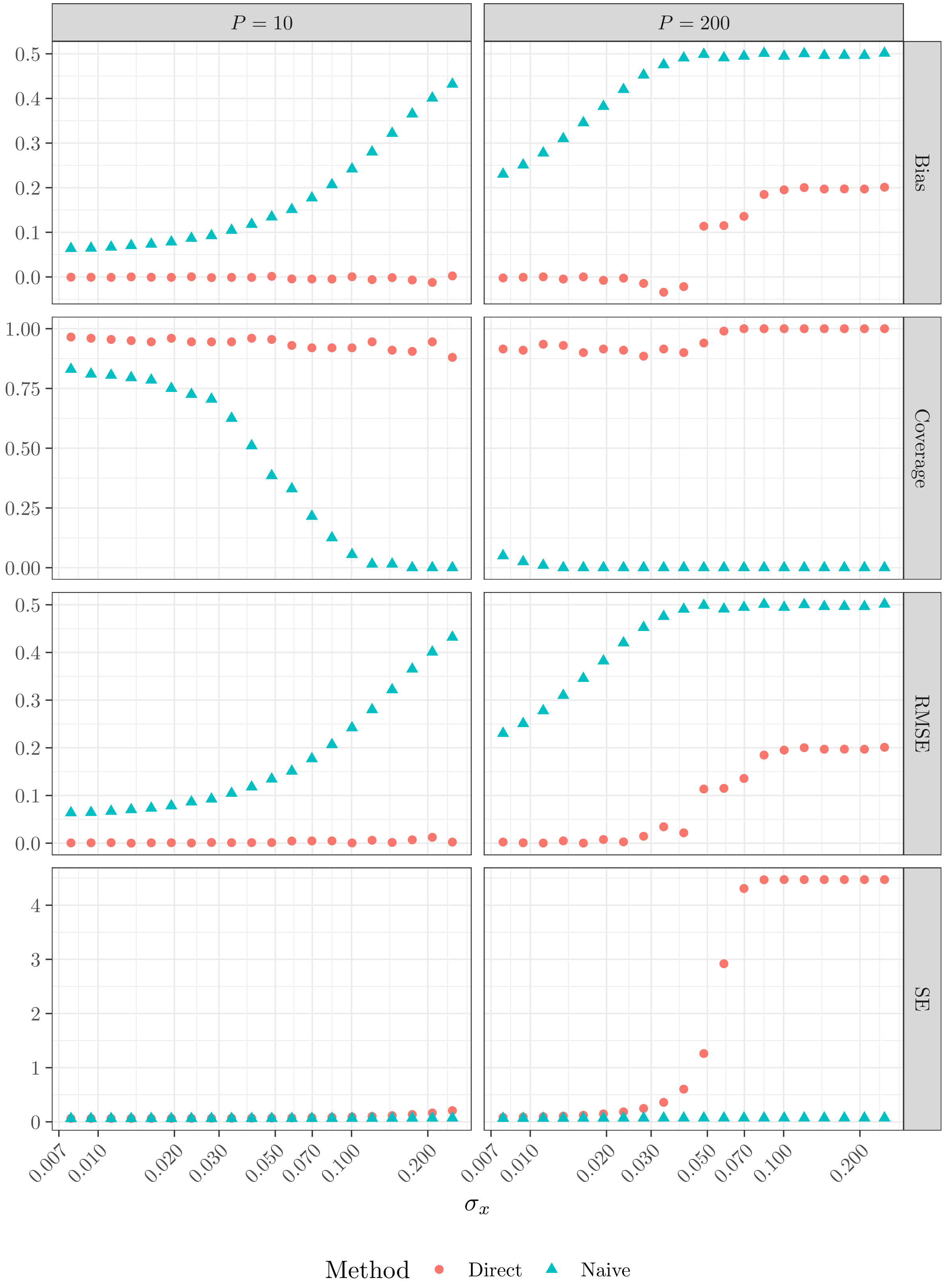

Figure 7. Results for the semiparametric manifold regression problem of Section 5. Bias denotes the average bias of $\widehat{\gamma}$, coverage denotes the coverage of nominal 95% intervals, RMSE denotes the root-mean-squared-error in estimating γ, and SE denotes the average posterior standard deviation of γ.

5.2. Dependence Structure and Semiparametric Regression

We now present evidence for the hypothesis that dimension reducing structure in X_i also shields us from dogmatism in nonparametric problems. Rather than assuming that X_i concentrates near a hyperplane, we instead assume that X_i is concentrated near a manifold of intrinsic dimension L in $\mathbb{R}^P$. Specifically, we take $X_i = \Lambda(\eta_i) + \sigma_x \epsilon_i, \epsilon_i \sim$ Normal$(0, 1)$where $\Lambda : \mathbb{R}^L \to \mathbb{R}^P$ is nonlinear. The function Λ is normalized for each σ_x so that the components of X_i have standard deviation 1; hence, σ_x indexes how close X_i is to $\mathscr{M} = \{x : x = \Lambda(\eta), \eta \in \mathbb{R}\}$. We use a continuous exposure $A_i = r_a(X_i) + \nu_i$ and a continuous outcome $Y_i(a) = r_y(X_i) + a\,\gamma + \epsilon_i(a)$, with $\epsilon_i(a), \nu \overset{\text{iid}}{\sim}$ Normal(0, 1) and $r_y(x) = r_y^\star(x) + r_a(x)$. The parameter of interest is γ, which represents the causal effect of the exposure on the outcome.

To construct ground truths, we generated $r_y^\star, r_a \sim \text{GP}(0, \rho)$ where $\rho(x, x') = \exp\{-\|x - x'\|_2^2\}$. We then constructed $\Lambda(\eta) = \big(\Lambda_1(\eta), \dots, \Lambda_P(\eta)\big)^\top$ by taking $\Lambda_j \sim \text{GP}(0, \kappa)$ independently for $j = 1, \dots, P$, where $\kappa(\eta, \eta') = \exp\{-\|\eta - \eta'\|_2^2\}$. We set $\gamma = 1$, $L = 1$, and consider $P \in \{10, 200\}$, $N = 300$, and $\sigma_x = 2^{-j}$ where the j's are evenly spaced between -7 and -2. For each σ_x and P we generated 200 simulated datasets and applied the Direct and Naive methods to estimate γ and construct a 95% credible interval. We consider two priors.

Naive We impose IG.2, but otherwise use the "true" prior for $r^\star(x)$ using the kernel $2\rho(x, x')$. We specify a Normal$(0, 10^2)$ prior for γ.

Direct We use the model $Y_i(a) = r_y(X_i) + \omega\,\widehat{r}_a(X_i) + \gamma\,a$ where $\widehat{r}_a(x)$ is a pilot estimate of $r_a(x)$ obtained from fitting a Gaussian process to the relationship $A_i = r_a(X_i) + \nu_i$. We specify a Normal$(0, 10^2)$ prior for both γ and ω.

Figure 7 displays the bias, coverage, RMSE, and average standard error in estimating γ. For $P = 10$, the behavior is similar to the ridge regression problem: the Direct approach performs uniformly well, while the naive approach (while never better than the direct approach) performs much better as σ_x is decreased. The $P = 200$ setting, on the other hand, is too difficult for the naive approach, although it does still perform better as σ_x decreased. The behavior of the direct approach at $P = 200$ is very interesting, however. We note a sharp phase transition around $\sigma_x = 0.07$ where the problem essentially goes from infeasible to feasible: the bias, RMSE, and standard error all decrease dramatically at this point. We also see that the direct approach is much more honest in terms of its uncertainty: when the problem is infeasible, the model correctly gives a large posterior standard error. Conversely, the naive model tends to understate the uncertainty in γ, leading to poor coverage.

6. Discussion

The main concrete recommendation we make in this article is that Bayesian ignorability (and in particular IG.2) can encode an informative prior on the degree of confounding bias, and thus should not be imposed in most situations. This tends to regularize confounding bias parameters toward 0, thereby introducing substantial bias in high-dimensional or nonparametric problems. Instead, Bayesians should reject IG.2 by default in favor of a prior that allows for more direct control over the confounding bias, and we have illustrated how to do this in several problems of interest. Of secondary interest, we have noted that features of the design can mitigate prior dogmatism about the confounding bias, and showed that both ridge regression priors and Gaussian process priors possess some degree of adaptivity toward low-dimensional structures in X_i. But this does not change our general recommendation, as we have consistently observed improved performance of priors that reject IG.2 even when such low-dimensional structures exist.

Dogmatism about other features of the model may also have bad inferential consequences. In future work, we will extend our results to other problems in causal inference. Some possibilities include estimation of the conditional average treatment effect (CATE) in observational studies, inference on the average causal effect under sparsity assumptions on (β, ϕ), and estimation of the natural direct and indirect effects in mediation analysis. In the context of mediation analysis, one must control for two different sources of confounding: the effect of the confounders both on the treatment received and on the mediating variable.

While we have presented a number of corrections for dogmatism, we have not presented any coherent framework for *deriving* corrections. This presents an important question: are there any objective Bayes principles that automatically lead to priors which adequately account for dogmatism? Parameterization invariant priors, like Jeffreys' priors, cannot work because they *imply* that IG.2 holds. By contrast, other objective principles that are not parameterization invariant and do not necessarily imply IG.2, such as priors constructed from decision theoretical principles, entropy maximization, and reference priors, have some chance of working (see Kass and Wasserman 1996, for a review). The computational difficulty of implementing these priors can make numeric experimentation difficult. Interestingly, entropy maximization with respect to the distribution of the observed (rather than complete) data can be used to generate models that possess very strong Frequentist properties (Sims 2012), but lack a satisfying justification.

Supplementary Materials

The supplementary materials contain code replicating our results, all proofs, additional theoretical analyses for ridge regression, additional simulation results, a general discussion of Robins-Wasserman-Ritov problems, and a discussion of dogmatism in the context of propensity score stratification.

Acknowledgments

The author thanks the Associate Editor and three anonymous reviewers for their input, which greatly improved the manuscript.

Disclosure Statement

The author does not report any potential competing interest.

Funding

This material is based upon work supported by the National Science Foundation under grant no. DMS-DMS-2144933.

References

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is much Larger than n," *The Annals of Statistics*, 35, 2313–2351. [3]

Couillet, R., and Debbah, M. (2011), *Random Matrix Methods for Wireless Communications*, Cambridge: Cambridge University Press. [5]

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021), "Overlap in Observational Studies with High-Dimensional Covariates," *Journal of Econometrics*, 221, 644–654. [2]

Dicker, L. H. (2016), "Ridge Regression and Asymptotic Minimax Estimation Over Spheres of Growing Dimension," *Bernoulli*, 22, 1–37. [2,3]

Dobriban, E., and Wager, S. (2018), "High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification," *The Annals of Statistics*, 46, 247–279. [2,3,5]

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018), "Regularization and Confounding in Linear Regression for Treatment Effect Estimation," *Bayesian Analysis*, 13, 163–182. [4,6,7]

Hahn, P. R., Murray, J. S., and Carvalho, C. (2017), "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," arXiv preprint arXiv:1706.09523. [8]

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020), "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," (with Discussion), *Bayesian Analysis*, 15, 965–1056. [2,8]

Imai, K., Keele, L., and Tingley, D. (2010), "A General Approach to Causal Mediation Analysis," *Psychological Methods*, 15, 309–334. [1]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press. [10]

Kass, R. E., and Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American statistical Association*, 91, 1343–1370. [13]

Kish, L. (1965), *Survey Sampling*, New York: Wiley. [5]

Li, F., Ding, P., and Mealli, F. (2022), "Bayesian Causal Inference: A Critical Review," arXiv e-prints. arXiv:2206.15460. [2,10]

Li, L. (2010), "Are Bayesian Inferences Weak for Wasserman's Example?" *Communications in Statistics — Simulation and Computation*, 39, 655–667. [1]

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. [5]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Hoboken, NJ: Wiley. [1]

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [11]

National Research Council. (2010), *The Prevention and Treatment of Missing Data in Clinical Trials*, Washington, DC: The National Academies Press. [1]

Rasmussen, C. E., and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, Cambridge, MA: MIT Press. [2,5,8]

Ray, K., and van der Vaart, A. (2020), "Semiparametric Bayesian Causal Inference," *The Annals of Statistics*, 48, 2999–3020. [5]

Ren, B., Wu, X., Braun, D., Pillai, N., and Dominici, F. (2021), "Bayesian Modeling for Exposure Response Curve via Gaussian Processes: Causal Effects of Exposure to Air Pollution on Health Outcomes," arXiv preprint arXiv:2105.03454. [5,8]

Robins, J. M., and Ritov, Y. (1997), "Toward A Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319. [1]

Robins, J. M., and Wasserman, L. (2012), "Robins and Wasserman Respond to a Nobel Prize Winner," Published on *Normal Deviate*. Available at *https://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/* on August 2, 2022. [1]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [1]

——— (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592. [1]

——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58. [1]

——— (1985), "The Use of Propensity Scores in Applied Bayesian Inference," *Bayesian Statistics*, 2, 463–472. [2,10]

——— (2005), "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, 100, 322–331. [1]

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013), "What is Meant by "Missing At Random"?" *Statistical Science*, 28, 257–268. [1]

Sims, C. A. (2012), "Robins-Wasserman, Round N," available at *http://sims.princeton.edu/yftp/WassermanExmpl/WassermanR4a.pdf* on August 17th, 2022. [2,13]

van der Laan, M. J., and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York: Springer. [6]

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science* (1st ed.), Cambridge: Cambridge University Press. [3]

Wang, C., Parmigiani, G., and Dominici, F. (2012), "Bayesian Effect Estimation Accounting for Adjustment Uncertainty," *Biometrics*, 68, 661–671. [11]

Wegner, S.-A. (2021), "Lecture Notes on High-Dimensional Data," arXiv preprint arXiv:2101.05841. [3]

Zhou, T., Elliott, M. R., and Little, R. (2019), "Penalized Spline of Propensity Methods for Treatment Comparison," *Journal of the American Statistical Association*, 114, 1–19. [8]

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013), "Model Feedback in Bayesian Propensity Score Estimation," *Biometrics*, 69, 263–273. [4,6]