Double Robust Bayesian Inference on Average Treatment Effects*

Christoph Breunig[†] Ruixuan Liu[‡] Zhengfei Yu[§]

October 6, 2022

Abstract

We study a robust Bayesian method for the average treatment effect (ATE) under unconfoundedness. This Bayesian procedure involves a correction term to prior distributions adjusted by the propensity score. We prove asymptotic equivalence of the robust Bayesian estimator and efficient frequentist estimators by establishing a new semiparametric Bernstein-von Mises theorem under double robustness, i.e., the lack of smoothness of regression functions can be compensated by high regularity of the propensity score and vice versa. Consequently, the resulting Bayesian point estimator enjoys the debiasing feature with the frequentist-type doubly robust estimator and the Bayesian credible sets form confidence intervals with asymptotically exact coverage probability. In simulations, we find that this corrected Bayesian procedure leads to significant bias reduction of point estimation and accurate coverage of confidence intervals, especially when the dimensionality of covariates is large relative to the sample size and the underlying functions are complex. We illustrate our method in an application to the National Supported Work Demonstration.

Key words: Average Treatment Effect, unconfoundedness, doubly robust, Nonparametric Bayesian Inference, semiparametric Bernstein–von Mises Theorem, Gaussian processes, Dirichlet process.

*We would like to thank Xiaohong Chen, Yanqin Fan, Essie Maasoumi, Yichong Zhang, and seminar participants at Toulouse School of Economics for helpful comments and discussions.

[†]Department of Economics, Bonn University, cbreunig@uni-bonn.edu

[‡]Department of Decision Sciences and Managerial Economics, CUHK Business School, Chinese University of Hong Kong, ruixuanliu@cuhk.edu.hk

[§]Faculty of Humanities and Social Sciences, University of Tsukuba, yu.zhengfei.gn@u.tsukuba.ac.jp

1 Introduction

In recent years, Bayesian approaches have become increasingly popular in the causal inference and program evaluation due to their excellent performance in finite samples. By assigning nonparametric priors to the function-valued parameters in the model, modern Bayesian inference fully utilize the flexibility of powerful machine learning algorithms. Related constructions using Gaussian Processes (GP) and Bayesian additive regression trees (BART) have both been shown to have excellent empirical performance (Ray and Szabó, 2019; Hahn, Murray, and Carvalho, 2020). In Bayesian analysis, two fundamental aims can be achieved at the same time: point estimation and uncertainty quantification. Researchers can directly read off quantities including both the posterior means and credible sets, once they have draws from the posterior distribution. One remarkable feature is that the Bayesian approach is able to incorporate prior knowledge and adapt to the presence of many covariates. Also, Bayesian approach have traditional appeal in the missing data literature, besides their recent popularity.

This paper establishes the double-robustness for Bayesian inference on the average treatment effect (ATE) under unconfoundedness given a set of pretreatment covariates. Despite the recent success of Bayesian approaches, the literature on the asymptotic properties of the average treatment effect estimation is mainly frequentist based. Indeed, early work on semiparametric Bayesian approaches to the missing data problem produced negative results, proving that many common classes of priors, or more generally likelihood-based procedures, produce inconsistent estimates assuming no smoothness on the underlying parameters; see the results and discussion in Robins and Ritov (1997) or Ritov, Bickel, Gamst, and Kleijn (2014). In contrast, once the prior distribution is corrected via the propensity score, Ray and van der Vaart (2020) establish asymptotic equivalence between the Bayesian procedure and efficient semiparametric estimators via the so called Bernstein-von Mises (BvM) theorem.[1] They show that their novel prior correction[2] significantly reduces the smoothness requirements on the propensity score function, but it still requires differentiability of the order $p/2$ at minimum for the conditional mean in the outcome equation, where p denotes the dimensionality of covarites.

[1]Strictly speaking, the main objective in Ray and van der Vaart (2020) is about the mean response in a missing data model, which is equivalent to observing one arm (either the treatment or control) of the causal setup.

[2]In an earlier unpublished working paper, Yang, Cheng, and Dunson (2015) suggested a related data-dependent prior, which makes certain adjustment through the least favorable direction for partial linear models. Their original purpose is to simplify the verification of the prior stability condition used in proving the BvM theorem. However, Yang, Cheng, and Dunson (2015) did not explore the bias reduction or double robustness property of this procedure.

In this paper, we show that Bayesian estimators with propensity score adjusted priors satisfy the semiparametric Bernstein-von Mises theorem under much less restrictive smoothness assumptions. Our assumptions take the double-robust form, that is, lack of smoothness of regression functions can be compensated by high regularity of the propensity score and vice versa. The proof of this result relies on important insights from the frequentists' study on the Riesz represantor, see Chernozhukov, Newey, and Singh (2020a) or Hirshberg and Wager (2021). Specifically, we are able to show that a correction term for the prior, which depends on the propensity score and forms a Riesz representer, leads to a centered term which can be controlled by elementary methods rather than by the more stringent stochastic equicontinuity. In addition, when we examine the prior stability condition, we tighten the maximal inequality used by Ray and van der Vaart (2020) by exploiting the product structure in the problem, so that the order of a negligible term is determined by the product of the convergence rates of the outcome and selection equations.

Although this paper focuses on the average treatment effect due to its popularity in empirical economics, the methodology per se is more general in nature and could be implemented beyond the ATE example. We establish novel Bayesian procedures that build on alternative corrections of the priors for other causal parameters such as the average treatment effect on the treated (ATT) and the average derivative (AD).[3] Similar to ATE, the prior correction used for other parameters of interest are also closely related to the Riesz represantors and the so-called "least favorable direction" : For ATT the correction term consists of the treated proportion and the propensity score, and for AD it involves a conditional density and its derivative.

Our theoretical results have appealing consequences for practitioners. Our robust Bayesian inference procedure corrects priors based on propensity scores and thus follows the idea of calibrated Bayes methodology advocated by Rubin (1984). The resulting credible interval is Bayesianly justifiable, as it makes use of posterior distribution conditional on the data, we also refer to Imbens (2021) for the preference of using Bayesian posteriors to quantify the estimation uncertainty. Our Bernstein van-Mises Theorem justifies a sound Bayesian inference procedure with prior correction, which internalizes bias correction and delivers asymptotically valid confidence interval. In our Monte Carlo simulations, we find that the prior correction through the estimated propensity score significantly reduces the

[3] The AD is an important semiparametric estimand in its own right and it has been regained the popularity as a structural parameter with causal interpretation when the treatment status is continuous. For instance, Chernozhukov, Newey, and Singh (2020a,b) advocated that the particular coordinate of the AD with respect to the continuous treatment status represents an approximation of the effect of policy that shifts the distribution of covariates through this particular direction.

bias of the Bayesian point estimator, which is consistent with our theory about its asymptotic equivalence with the frequentist doubly robust estimator. Also, the method leads to substantial improved empirical coverage probabilities, in particular, in the presence of many covariates relative to the sample size. Its computation can be implemented by existing software with the simple adjustment on the prior, so it offer greater flexibility for practitioners to apply state-of-the-art Bayesian algorithms that can lead to valid inference with minimal assumptions on the underlying functional classes.

Related Literature Our paper fits into a broader literature on the debiased or double robust inference. Scharfstein et al. (1999) noted that an estimator originally developed and identified as the locally efficient estimator in the class of augmented inverse probability weighted (AIPW) estimators in missing data models in Robins et al. (1994), was double-robust [4]. Since then, many estimators with the double-robust property have been proposed. In the literature of mean regression with missing data, AIPW is a popular method, where both the missingness probability (encoded by the propensity score) and the data distribution (or the conditional mean function) are modeled. In the earlier development, the focus is typically on developing working parametric models for either the propensity score or the conditional mean function. However, implausible parametric assumptions on the data generating process are of limited applicability to complex phenomena in economics and social sciences. Recent advance in the double machine learning literature have led to a number of important developments in causal inference, utilizing flexible nonparametric or machine learning algorithms. In this context, the double robustness the possibility to trade off the estimation accuracy between nuisance functions. We refer readers to Chernozhukov, Newey, and Singh (2020a) for a comprehensive survey of the recent development.

While the Bernstein-von Mises theorem for parametric Bayesian models is well established (van der Vaart, 1998), the semiparametric version is still being studied very actively when nonparametric priors are used. The area has received an enormous amount of attention (Castillo, 2012; Castillo and Rousseau, 2015; Norets, 2015; Yang, Cheng, and Dunson, 2015; Florens and Simoni, 2019; Ray and van der Vaart, 2020). Admitted, the technical arguments in the aforementioned work all build on the so-called “no-bias” condition. This is in the same spirit of the frequentist counterpart (van der Vaart, 1998), which generally leads to harsh smoothness restrictions and may not be satisfied when the dimensionality

[4] An estimator is said to be doubly robust if it is consistent for the target parameter of interest when any one of two nuisance parameters is consistently estimated. This property gives doubly robust estimators a natural appeal: any possible inconsistency in the estimation of one nuisance parameter may be mitigated by the consistent estimation of the other.

increases. To the best of our knowledge, our new Bernstein-von Mises theorem is the first one that possesses the double robustness property. We would like to mention the current research area about the Bayesian inference in econometrics which are robust to partial or weak identification (Chen, Christensen, and Tamer, 2018; Giacomini and Kitagawa, 2020; Andrews and Mikusheva, 2022). The framework and the approach we take is different. Nonetheless, they share the same scope of robustifying the Bayesian inference procedure.

A couple of recent papers present doubly robust Bayesian recipes. While sharing a common goal of correcting for bias with a Bayesian lens, consensus has not reached on how to conduct inference with propensity score adjustment. Ray and van der Vaart (2020); Ray and Szabó (2019) and our study can be interpreted as Empirical Bayes which draws on data dependent priors. Saarela, Belzile, and Stephens (2016) consider a Bayesian procedure based on an analog of the double robust frequentist estimator given in (2.9), replacing the empirical measure with the Bayesian bootstrap measure. Saarela, Belzile, and Stephens (2016) also suggested that initial estimands for the outcome and selection equations should be obtained by similar parametric weighted M-estimators using Bayesian bootstrap weights. There is no formal BvM theorem presented in Saarela, Belzile, and Stephens (2016). Another recent paper by Yiu, Goudie, and Tom (2020) explored Bayesian exponentially tilted empirical likelihood with the set of moment constraints that are of the double-robust type. They proved a BvM theorem for the posterior constructed from the resulting exponentially tilted empirical likelihood under parametric specifications. It is not clear how to extend their analysis to incorporate flexible nonparametric modeling strategies.

The remainder of this paper is organized as follows. Section 2 presents the setup and the semiparametric Bayesian inference procedure. In Section 3 we derive the least favorable direction and presents a main result: a doubly robust version of Bernstein-von Mises Theorem with the implication of asymptotically exact confidence sets. Section 4 provides an illustration using Gaussian priors. We provide numerical illustrations on both synthetic and real data to demonstrate the practical implications of our theoretical results in Section 5. Proofs of main theoretical results can be found in Appendix A. Appendix B establishes Lemmas used in the proof of the main findings. Auxiliary results can be found in Appendix C. Additional simulation results are provided in Appendix D.

2 Setup and Implementation

This section provides the main setup of the average treatment effect and motivates the methodology. Subsection 2.3 extends our framework to average treatment effects on the treated and the average derivatives.

2.1 Setup

Our standpoint remains frequentist, so there is a true data generating process (DGP) denoted by P_0. It is indexed by a fixed (and possibly infinite dimensional) parameter $\eta_0 \in \mathcal{H}$ such that $P_0 = P_{\eta_0}$. A thorough frequentist analysis validates the insensitivity of prior choices and confirms that the data can wash off the influence from priors as sample size increases.

We consider the potential outcome framework of causal inference: for individual i, consider a treatment indicator $D_i \in \{0,1\}$. The observed outcome Y_i is determined by $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ where $(Y_i(1), Y_i(0))$ are the potential outcomes of individual i associated with $D_i = 1$ or 0. This paper focuses on the binary outcome case where both $Y_i(1)$ and $Y_i(0)$ take values in $\{1, 0\}$. Let X be a vector of covariates with the distribution F and the density f. Let $\pi(x) = \Pr(D_i = 1|X_i = x)$ denote the propensity score and $m(d,x) = \Pr(Y_i = 1|D_i = d, X_i = x)$ for the conditional mean. Suppose that the researcher observe an i.i.d. sample of $O_i = (Y_i, D_i, X_i)$ for $i = 1, \ldots, n$. The parameter of interest is the average treatment effect (ATE) $\chi_0 = \mathbb{E}[Y_i(1) - Y_i(0)]$. For its identification, we impose the following standard assumption of unconfoundness and overlap.

Assumption 1. (i) $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid X_i$ and (ii) there exists $\bar{\pi} > 0$ such that $\bar{\pi} < \pi(x) < 1 - \bar{\pi}$ for all x in the support of F.

Since outcome and treatment are binary the joint density of $O_i = (Y_i, D_i, X_i)$ can be written as

$$p_{\pi,m,f}(o) = \pi(x)^d(1 - \pi(x))^{1-d} m(d,x)^y (1 - m(d,x))^{(1-y)} f(x). \tag{2.1}$$

The observed data O_i can be described by the triple (π, m, f). For prior construction it will be useful to transform the parameters (π, m) by a link function and we choose the logistic function $\Psi(t) = 1/(1 + e^{-t})$ here. Specifically, we consider the reparametrization of (π, m, f) given by $\eta = (\eta^\pi, \eta^m, \eta^f)$ where

$$\eta^\pi = \Psi^{-1}(\pi), \quad \eta^m = \Psi^{-1}(m), \quad \eta^f = \log f. \tag{2.2}$$

Below, we write $m_\eta = \Psi(\eta^m)$ and $\pi_\eta = \Psi(\eta^\pi)$ to make the dependence on η explicit. We are interested in the ATE, which under Assumption 1 is identified by

$$\chi_\eta = \mathbb{E}\left[m_\eta(1, X) - m_\eta(0, X)\right]. \tag{2.3}$$

The efficient influence function (see Hahn (1998); Hirano, Imbens, and Ridder (2003)) is given by

$$\widetilde{\chi}_\eta(o) = m_\eta(1, x) - m_\eta(0, x) + \gamma_\eta(d, x)(y - m_\eta(d, x)) - \chi_\eta \tag{2.4}$$

for some Riesz representor γ_η which is given by

$$\gamma_\eta(d, x) = \frac{d}{\pi_\eta(x)} - \frac{1-d}{1-\pi_\eta(x)}. \tag{2.5}$$

Consequently, we can write asymptotically efficient estimators $\widehat{\chi}$ with the following linear representation:

$$\widehat{\chi} = \chi_0 + \frac{1}{n}\sum_{i=1}^{n} \widetilde{\chi}_0(O_i) + o_{P_0}(n^{-1/2}). \tag{2.6}$$

2.2 Bayesian Point Estimators and Credible Sets for the ATE

Our doubly robust inference procedure builds on a nonparametric Bayesian prior specification for m that depends on a preliminary estimator for γ_0. A pilot estimator for the propensity score π_0 is denoted by $\widehat{\pi}$ based on an auxiliary sample. We consider a plug-in estimator for the Riesz representor γ_0 given by

$$\widehat{\gamma}(d, x) = \frac{d}{\widehat{\pi}(x)} - \frac{1-d}{1-\widehat{\pi}(x)}.$$

The use of an auxiliary data for the estimation of the propensity simplifies the technical analysis and is common in the related Bayesian literature, see Ray and van der Vaart (2020) or Ignatiadis and Wager (2022). In practice, we use the full data twice and do not sample-split; we have not observed any over-fitting or loss of coverage thereby.

In order to obtain the Bayesian point estimator and the credible set from the posterior distribution of χ_η through simulation draws, our procedure builds on the following three steps:

1. Compute the propensity score-dependent prior on m:

$$m_\eta(d, x) = \Psi\left(\eta^m(d, x)\right) \qquad \text{and} \qquad \eta^m = W^m + \lambda\,\widehat{\gamma} \tag{2.7}$$

where W^m is a continuous stochastic process independent of the random variable λ, which follows a prior $N(0, \sigma_n^2)$ for some $\sigma_n > 0$. The tuning parameter λ governs the influence strength of the propensity score on the prior distribution of m. Smaller σ_n allow for larger choices of λ but we may also set $\sigma_n = 1$, see Section 4 for data driven choices of such tuning parameters. Then we draw the posterior of $\eta^m(d, x)$ and thus $m_\eta(d, x)$ using Gaussian process classification, also see Section 4 for more details. Denote the $m_\eta^b(\cdot)$ as a generic random function drawn from this posterior, for $b = 1, \ldots, B$.

2. Generate Bayesian bootstrap weights $M_{n1}^b, \ldots, M_{nn}^b$ where $M_{ni}^b = e_i^b / \sum_{i=1}^n e_i^b$ and e_i^b's are independently and identically drawn from the exponential distribution Exp(1) for $b = 1, \ldots, B$. A generic draw from the posterior distribution for the ATE χ_η admits the following representation:

$$\chi_\eta^b = \sum_{i=1}^n M_{ni}^b \big(m_\eta^b(1, X_i) - m_\eta^b(0, X_i) \big), \quad b = 1, \ldots, B. \tag{2.8}$$

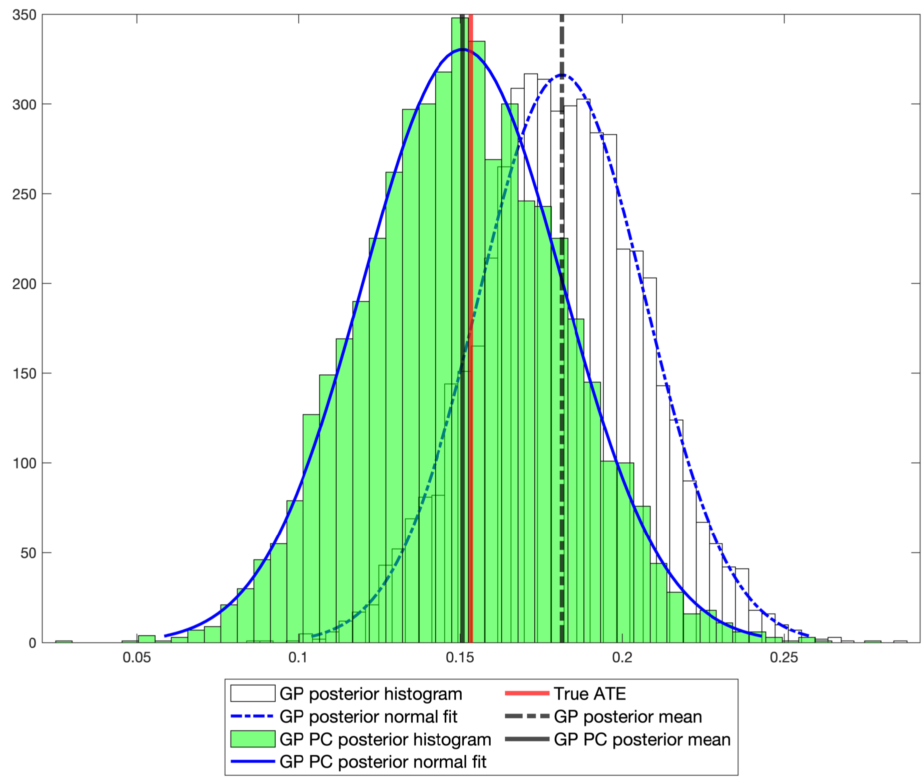
3. Our $100 \cdot (1-\alpha)\%$ confidence set for the ATE parameter χ_0 is then given by

$$\mathcal{C}_n(\alpha) = \big\{ \chi : q_n(\alpha/2) \leqslant \chi \leqslant q_n(1-\alpha/2) \big\},$$

where $q_n(a)$ denotes with the a quantile of $\{\chi_\eta^b : b = 1, \ldots, B\}$. Additionally, we may compute the Bayesian point estimator by the posterior mean: $\overline{\chi}_\eta = \frac{1}{B} \sum_{b=1}^B \chi_\eta^b$.

Example 2.1 (Simulation of Prior Correction)**.** We illustrate the role of prior correction via propensity score adjustment in finite samples. Figure 1 plots a posterior sample of χ_η^b's in (2.15) with $B = 5,000$. For comparison, it also plots the posterior from the conventional Gaussian process approach without the prior correction, that is, $\eta^m = W^m$ in (2.7). It shows that the prior correction based on the (estimated) Riesz representor shifts the center of the posterior distribution towards the true ATE. As a result, the prior corrected algorithm described above would yield smaller bias for the point estimator (posterior mean) as well as more accurate coverage probability for the confidence interval. This illustrative simulation exercise is in line with our Monte Carlo simulation results in Section 5

Figure 1: Plots of $5,000$ posterior draws from Bayesian inference method based on Gaussian process without prior correction (GP) and the one with prior correction described above (GP PC). Data from Design I in the simulation section, $p = 15$, sample size $= 1,000$.



Remark 2.1 (Bayesian bootstrap)**.** *Under unconfoundedness and the reparametriztation in (2.2), the ATE can be written as* $\chi_\eta = \int [\Psi(\eta^m(1,x)) - \Psi(\eta^m(0,x))] dF(x)$. *We put a prior probability distribution* Π *on the function-valued parameters and consider the posterior distribution* $\Pi(\cdot|\boldsymbol{O}^{(n)})$ *based on the observations* $\boldsymbol{O}^{(n)} = (O_1, O_2, \cdots, O_n)$. *This induces a posterior distribution on the functional of interest, i.e. ATE. We consider independent priors on* η^m *and* F*, we have the factorization of posteriors for* η^m *and* F *given that the likelihood function also factorizes into two products. In short, we can consider the posterior for* η^m *and* F *separately. We consider a Dirichlet process prior for* F *(see, for instance, Chamberlain and Imbens (2003)). When we restrict the base measure of the Dirichlet prior to be zero, the posterior law of* F *coincides with the Bayesian bootstrap (Rubin, 1981). One key advantage of the Bayesian bootstrap is that it allows us to incorporate a broad class of DGPs whose posterior can be easily sampled via Bootstrap algorithm. That is, we can avoid*

an additional model for the marginal density of covariates with computationally intensive MCMC algorithms.[5]

Remark 2.2 (Comparison with frequentist robust estimation)**.** *For the average treatment effect, the perhaps most popular method for asymptotic efficient inference is given by the double-robust estimator*

$$n^{-1}\sum_{i=1}^{n}\big(\widehat{m}(1,X_i)-\widehat{m}(0,X_i)\big)+n^{-1}\sum_{i=1}^{n}\widehat{\gamma}(D_i,X_i)\big(Y_i-\widehat{m}(D_i,X_i)\big) \tag{2.9}$$

based on frequentist-type pilot estimators $\widehat{m}$ of the regression function m_0 and $\widehat{\gamma}$ of the Riesz representer γ_0, see Newey (1994), Robins and Rotnitzky (1995), Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2017)). More recently, Chernozhukov, Newey, and Singh (2020b) extend this approach to the high-dimensional case (using the so called Danzig selector). Hirshberg and Wager (2021) use the minimax linear approach with a focus to debias a plugin estimator rather (without being explicitly designed to be double-robust).

2.3 Extension to other Causal Parameters

2.3.1 Bayesian Point Estimators and Credible Sets for the ATT

We now extend the methodology to average treatment effects for the treated (ATT) given by $\mathbb{E}[Y(1)-Y(0)|D=1]$. Under unconfoundedness and the reparametrization in (2.2), the ATT parameter can be written as

$$\chi_\eta^T=\mathbb{E}[Y-m_\eta(0,X)|D=1]. \tag{2.10}$$

Again following Hahn (1998); Hirano, Imbens, and Ridder (2003), the efficient influence function for the ATT parameter under unconfoundedness is given by

$$\widetilde{\chi}_\eta^T(o)=\gamma_\eta^T(d,x)(y-m_\eta(d,x))+\frac{d}{\pi_\eta}\left(m_\eta(1,x)-m_\eta(0,x)-\chi_\eta^T\right)$$

for some Riesz representor γ_η^T which is given by

$$\gamma_\eta^T(d,x)=\frac{d}{\pi_\eta}-\frac{1-d}{\pi_\eta}\frac{\pi_\eta(x)}{1-\pi_\eta(x)} \tag{2.11}$$

[5]We also note that replacing F by the standard empirical cummulative distribution function does not provide sufficient randomization of F as it yields underestimation of the asymptotic variance, see (Ray and van der Vaart, 2020, Remark 2).

We now propose a novel Bayesian estimator for the ATT under unconfoundedness. Based on an initial estimator $\widehat{\pi}(\cdot)$ and $\widehat{\pi}$ of the propensity score $\pi(\cdot)$ and the proportion π, we consider a plug-in estimator for the Riesz representor γ_0^T given by

$$\widehat{\gamma}^T(d,x) = \frac{d}{\widehat{\pi}} - \frac{1-d}{\widehat{\pi}}\frac{\widehat{\pi}(x)}{1-\widehat{\pi}(x)}.$$

We consider generic draws from the ATT parameter χ_η^T by

$$\chi_\eta^{T,b} = \frac{\sum_{i=1}^n M_{ni}D_i\big(Y_i - m_\eta(0,X_i)\big)}{\sum_{i=1}^n M_{ni}D_i}, \quad b = 1,\dots,B, \tag{2.12}$$

where $m_\eta(d,x) = \Psi\left(\eta^m(d,x)\right)$ and $\eta^m = W^m + \lambda\,\widehat{\gamma}^T$ and M_{ni} the Bayesian bootstrap weights introduced in the previous section. Our $100\cdot(1-\alpha)\%$ confidence set for the ATT parameter χ_0^T is then given by

$$\mathcal{C}_n^T(\alpha) = \big\{\chi : q_n^T(\alpha/2) \leqslant \chi \leqslant q_n^T(1-\alpha/2)\big\},$$

where $q_n^T(a)$ denotes with the a quantile of $\{\chi_\eta^{T,b} : b = 1,\dots,B\}$. Our Bayesian point estimator for the ATT is $\overline{\chi}_\eta^T = \frac{1}{B}\sum_{b=1}^B \chi_\eta^{T,b}$.

2.3.2 Bayesian Point Estimators and Credible Sets for the AD

Upon proper change of notations, our analysis can be easily applied to average directional derivative (Chernozhukov, Newey, and Singh, 2020b) and more generally, linear functionals of conditional mean (Hirshberg and Wager, 2021). Considering the average directional derivative, if one estimates the asymptotic variance of the influence function by frequentist methods, it involves analytical or numerical function-valued parameters or their derivatives. In contrast, the nonparametric Bayesian inference requires neither estimation of additional nonparametric elements nor evaluation of the derivatives.

Consider the case of continuous treatment variable D. The average derivative is then given by

$$\chi_\eta^{AD} = \mathbb{E}\left[\partial_d m_\eta(D,X)\right] \tag{2.13}$$

where $\partial_d m$ denotes the partial derivatives of m with respect to the continuous treatment D. The efficient influence function is

$$\widetilde{\chi}_\eta^{AD}(o) = \partial_d m_\eta(d,x) - \mathbb{E}[\partial_d m_\eta(d,x)] + \gamma_\eta^{AD}(d,x)(y - m_\eta(d,x))$$

for some Riesz representor γ_η^{AD} which is given by

$$\gamma_\eta^{AD}(d,x) = \frac{\partial_d \pi_\eta(d,x)}{\pi_\eta(d,x)}, \tag{2.14}$$

where here π_η stands for the conditional density function of D given X.

We now propose a novel estimator for the AD. Based on an initial estimator $\widehat{\pi}_\eta$ of the conditional density π_η, we consider a plug-in estimator for the Riesz representor γ_0^{AD} given by

$$\widehat{\gamma}^{AD}(d,x) = \frac{\partial_d \widehat{\pi}_\eta(d,x)}{\widehat{\pi}_\eta(d,x)}.$$

We consider generic draws from the AD parameter χ_η^{AD} by

$$\chi_\eta^{AD,b} = \sum_{i=1}^n M_{ni}^b \partial_d m_\eta(D_i, X_i), \quad b = 1, \dots, B, \tag{2.15}$$

where $m_\eta(d,x) = \Psi\left(\eta^m(d,x)\right)$ and $\eta^m = W^m + \lambda\,\widehat{\gamma}^{AD}$ and M_{ni}^b the Bayesian bootstrap weights introduced in the previous section. Our $100 \cdot (1-\alpha)\%$ confidence set for the AD parameter χ_0^{AD} is based on quantiles of the bootstrap sample (2.15). We also propose the Bayesian point estimator for the AD by $\overline{\chi}_\eta^{AD} = \frac{1}{B}\sum_{b=1}^B \chi_\eta^{AD,b}$.

3 Main Theoretical Results

Confidence or credible sets are standard means of describing uncertainty about model parameters from a frequentist or Bayesian point of view, respectively. The classical Bernstein-von Mises (BvM) Theorem validates Bayesian approaches from a frequentist point of view and bridges the gap between a Bayesian and a frequentist. By virtue of the BvM Theorem, the following distributions

$$\sqrt{n}(\chi_\eta - \widehat{\chi})|\boldsymbol{O}^{(n)} \quad \text{and} \quad \sqrt{n}(\widehat{\chi} - \chi_\eta)|\eta = \eta_0$$

are asymptotically equivalent under the underlying sampling distribution. As a consequence, so are the resulting credible and confidence sets[6]. In the above display, the first

[6]On a different note, Bayesians use these BvM type results to show that standard frequentist procedures are nearly Bayesian. So not much is lost by confining attention to Bayes procedures. And frequentists can advocate that their inferential procedures also have desirable conditional properties as the limit of the Bayesian counterparts.

one is the posterior, which is of interest to Bayesians, and the second one is of interest to frequentists in asymptotic analysis. The sequence $\sqrt{n}(\widehat{\chi}_n - \chi_0)$ is then asymptotically normal with mean zero and variance

$$\mathrm{v}_0 = P_0[\widetilde{\chi}_0^2] = \mathbb{E}\left[\frac{Var(Y(1)|X)}{\pi_0(X)} + \frac{Var(Y(0)|X)}{1-\pi_0(X)} + \big(m_0(1,X) - m_0(0,X) - \chi_0\big)^2\right],$$

which is the smallest variance possible by the efficiency bound of Hahn (1998).

Consider the one-dimensional submodel $t \mapsto \eta_t$ defined by the path

$$\pi_t(x) = \Psi(\eta^\pi + t\mathfrak{p})(x), \quad m_t(d,x) = \Psi(\eta^m + t\mathfrak{m})(d,x), \quad f_t(x) = f(x)e^{t\mathfrak{f}(x)}/\mathbb{E}[e^{t\mathfrak{f}(X)}], \tag{3.1}$$

for the given direction $(\mathfrak{p}, \mathfrak{m}, \mathfrak{f})$ with $\mathbb{E}[\mathfrak{f}(X)] = 0$. The difficulty of estimating the parameter χ_{η_t} for the submodels depends on the direction $(\mathfrak{p}, \mathfrak{m}, \mathfrak{f})$. Among them, let $\xi_\eta = (\xi_\eta^\pi, \xi_\eta^m, \xi_\eta^f)$ be the *least favorable direction* that is associated with the most difficult submodel, i.e., gives rise to the largest asymptotic optimal variance for estimating χ_{η_t}.[7]

Lemma 3.1. *Consider the submodel* (3.1)*. Under Assumption 1, the least favorable direction for estimating the ATE parameter in* (2.3) *is:*

$$\xi_\eta = (\xi_\eta^\pi, \xi_\eta^m, \xi_\eta^f) := (0, \gamma_\eta(D,X), m_\eta(1,X) - m_\eta(0,X) - \chi_\eta), \tag{3.2}$$

where the Riesz representer γ_η satisfies (2.5)*. Under Assumption 1(i) and if $\bar{\pi} < \pi(x)$ for all x in the support of F, then the least favorable direction for estimating the ATT parameter in* (2.10) *is:*

$$\xi_\eta^T = \left(0, \gamma_\eta^T(D,X), \frac{D}{\pi_\eta}\big(m_\eta(1,X) - m_\eta(0,X) - \chi_\eta^T\big)\right), \tag{3.3}$$

where the Riesz representer γ_η^T satisfies (2.11)*.*

In the setup of AD (see Section 2.3.2), consider the submodel $t \mapsto \eta_t$ defined by the path $m_t(d,x) = \Psi(\eta^m + t\mathfrak{m})(d,x)$, $f_t(d,x) = f(d,x)e^{t\mathfrak{f}(d,x)}/\mathbb{E}[e^{t\mathfrak{f}(D,X)}]$, with $\mathbb{E}[\mathfrak{f}(D,X)] = 0$. The least favorable direction for estimating the AD parameter in (2.13) *is:*

$$\xi_\eta^{AD} = \big(\gamma_\eta^{AD}(D,X), \partial_d m(D,X) - \mathbb{E}[\partial_d m(D,X)]\big), \tag{3.4}$$

where the Riesz representer γ_η^{AD} satisfies (2.14)*.*

[7] See the proof of Lemma 3.1 in the appendix for a formal definition of the least favorable direction that follows Ghosal and Van der Vaart (2017, p.370).

From Lemma 3.1 we see that the least favorable direction is invariant under a shift of the nonparametric component of propensity score π. This reflects conditions for the semiparametric Bernstein-von Mises theorem to hold, see Ghosal and Van der Vaart (2017). Our prior correction, which takes the form of the (estimated) least favorable direction, exactly provides such an invariance by giving the prior an explicit component in this direction. It provides additional robustness against posterior inaccuracy in the 'most difficult direction', i.e., the one inducing the largest bias in the ATE.

We now provide additional notation and assumptions provided for the derivation of our semiparametric Bernstein-van Mises Theorem. The posterior distribution plays an important role in the following analysis and is given by

$$\Pi\left((\pi, m) \in A, F \in B|\boldsymbol{O}^{(n)}\right) = \int_B \frac{\int_A \prod_{i=1}^n p_{(\pi,m)}(O_i) d\Pi(\pi, m)}{\int \prod_{i=1}^n p_{(\pi,m)}(O_i) d\Pi(\pi, m)} d\Pi(F|\boldsymbol{O}^{(n)}).$$

We write $\mathcal{L}_\Pi(\sqrt{n}(\chi_\eta - \widehat{\chi})|\boldsymbol{O}^{(n)})$ for the marginal posterior distribution of $\sqrt{n}(\chi_\eta - \widehat{\chi})$. Because the factorization of the likelihood function and the fact that χ_η does not depend on η^π, it is unnecessary to further discuss a prior or posterior distribution on η^ϕ.

We first introduce assumptions, which are high-level and discuss primitive conditions for those in the next section. Below, we consider some measurable sets $\mathcal{H}_n^m$ of functions η^m such that $\Pi(\eta^m \in \mathcal{H}_n^m|\boldsymbol{O}^{(n)}) \to_{P_0} 1$.

Assumption 2. [Rates of Convergence] The functional components satisfy

$$\|\widehat{\gamma} - \gamma_0\|_{L^2(F_0)} \leqslant r_n \quad \text{and} \quad \sup_{\eta \in \mathcal{H}_n^m} \|m_\eta(d, \cdot) - m_0(d, \cdot)\|_{L^2(F_0)} \leqslant \varepsilon_n \text{ for } d = 1, 0,$$

where $\max\{\varepsilon_n, r_n\} \to 0$ and $\sqrt{n}\varepsilon_n r_n \to 0$. Further, $\|\widehat{\gamma}\|_\infty = O_{P_0}(1)$.

We adopt the standard empirical process notation as follows. For a function h of a random vector $O = (Y, D, X^\top)^\top$ that follows distribution P, we let $P[h] = \int h(o) dP(o), \mathbb{P}_n[h] = n^{-1}\sum_{i=1}^n h(O_i)$, and $\mathbb{G}_n[h] = \sqrt{n}\left(\mathbb{P}_n - P\right)[h]$.

Assumption 3. [Complexity] For $\mathcal{G}_n = \{m_\eta(1, \cdot) - m_\eta(0, \cdot) : \eta \in \mathcal{H}_n\}$ we assume that $\sup_{m \in \mathcal{G}_n} |\mathbb{P}_n m - P_0 m| = o_{P_0}(1)$. We further impose that

$$\sup_{\eta \in \mathcal{H}_n^m} \left|\mathbb{G}_n\left[\left(\widehat{\gamma} - \gamma_0\right)\left(m_\eta - m_0\right)\right]\right| = o_{P_0}(1).$$

Recall the propensity score-dependent prior on m given in (2.7), i.e., $m(\cdot) = \Psi\left(W^m(\cdot) + \lambda\widehat{\gamma}(\cdot)\right)$. Below, we restrict the behavior for λ through its hyperparameter $\sigma_n > 0$.

Assumption 4. [Prior Stability] W^m is a continuous stochastic process independent of the normal random variable $\lambda \sim N(0, \sigma_n^2)$, where $n\sigma_n^2 \to \infty$. The following two conditions are imposed: (i)

$$\Pi\left(\lambda : |\lambda| \leqslant u_n \sigma_n^2 \sqrt{n} \mid \boldsymbol{O}^{(n)}\right) \to_{P_0} 1,$$

for some deterministic sequence $u_n \to 0$ and (ii) for any $t \in \mathbb{R}$:

$$\Pi\left((w, \lambda) : w + (\lambda + tn^{-1/2})\widehat{\gamma} \in \mathcal{H}_n^m \mid \boldsymbol{O}^{(n)}\right) \to_{P_0} 1$$

Discussion of Assumptions: Assumption 2 imposes sufficiently fast convergence rates for the estimators for regression function m_0 and the propensity score π_0. In practice, one can explore the recent proposals from Chernozhukov, Newey, and Singh (2020b) and Hirshberg and Wager (2021). The posterior convergence rate for the conditional mean can be derived by modifying the classical results of Ghosal, Ghosh, and van der Vaart (2000) by accommodating the propensity score adjusted prior, in the same spirit of Ray and van der Vaart (2020). The rate restriction is easier to satisfy if one function is easier to estimate, which resembles Theorem 1 conditions (i) and (ii) of Farrell (2015). Remark 4.1 illustrates that under classical smoothness assumptions, this condition is less restrictive than plug-in method of Ray and van der Vaart (2020) or other approaches for semiparametric estimation of ATEs Chen, Hong, and Tarozzi (2008) or Farrell, Liang, and Misra (2021). Assumption 4 is imposed to check the prior invariance property.

In contrast to our Assumption 3, Ray and van der Vaart (2020) (see their Assumption (3.12)) require a stochastic equicontinuity condition[8] $\sup_{\eta \in \mathcal{H}_n^m} \mathbb{G}_n\left[m_\eta - m_0\right] = o_{P_0}(1)$.[9] In comparison, a condition similar to our Assumption 3 is also used in the frequentist literature for ATE inference under unconfoundness; see Section 2 of Benkeser, Carone, Laan, and Gilbert (2017) or Assumption 3(a)-(c) in Farrell (2015). We argue that our formulation significantly weakens the requirement from Ray and van der Vaart (2020) and allows for double robustness under Hölder smoothness classes (see Remark 4.1). Hence, the complexity of the functional class $(m - m_0)$ can be compensated by certain high regularity of the corresponding Riesz representor and vice versa. Assumption 4(i) can be established along similar lines as in Lemma 4 of Ray and van der Vaart (2020), combined with the Gaussian

[8] In a different context for kernel-based semiparametric estimation, Cattanoe and Jansson (2018) relaxed the stochastic equicontinuity condition which takes into account the slow convergence rate of the kernel estimands, due to the small bandwidth.

[9] If one translates their missing data setup to the current ATE setup.

tail bound for the prior mass of $\{\lambda : |\lambda| > u_n \sigma_n^2 \sqrt{n}\}$. Referring to Assumption 4(ii), one can argue that this set hardly differs from the set $\mathcal{H}_n^m$.

We now establish a semiparametric Bernstein–von Mises theorem, which establish asymptotic normality of the posterior distribution. This asymptotic equivalence result is established using the so called *bounded Lipschitz distance* defined as follows. For two probability measures P, Q defined on a metric space $\mathcal{X}$ with a metric $d(\cdot,\cdot)$, we define the bounded Lipschitz distance as

$$d_{BL}(P,Q) = \sup_{f \in BL(1)} \left| \int_{\mathcal{O}} f(dP - dQ) \right|, \tag{3.5}$$

where

$$BL(1) = \left\{ f : \mathcal{O} \mapsto \mathbb{R}, \sup_{o \in \mathcal{O}} |f(o)| + \sup_{o \neq o'} \frac{|f(o) - f(o')|}{\|o - o'\|_{\ell_2}} \leqslant 1 \right\}.$$

Our main result is to show this sequence of marginal posteriors converges in the bounded Lipschitz distance to a normal distribution under weaker conditions than Ray and van der Vaart (2020).

Theorem 3.1. *Let Assumptions 1–4 hold. Then we have*

$$d_{BL}\left(\mathcal{L}_\Pi(\sqrt{n}(\chi_\eta - \widehat{\chi})|\boldsymbol{O}^{(n)}), N(0, \mathrm{v}_0)\right) \to_{P_0} 0.$$

We now show how Theorem 3.1 can be used to give a frequentist justification of Bayesian methods to construct the point estimator and the confidence sets. Recall that $\widehat{\chi}_\eta^{\mathrm{Bayes}}$ represents the posterior mean. Introduce a Bayesian credible set $\mathcal{C}_n(\alpha)$ for χ_η, which satisfies $\Pi(\chi_\eta \in \mathcal{C}_n(\alpha)|\boldsymbol{O}^{(n)}) = 1 - \alpha$ for a given nominal level $\alpha \in (0,1)$. The next result shows that $\mathcal{C}_n(\alpha)$ also forms a confidence interval in the frequentist sense for the ATE parameter whose coverage probability under P_0 converges to $1 - \alpha$.

Corollary 3.1. *Let Assumptions 1–4 hold. Then under P_0, we have*

$$\sqrt{n}\left(\widehat{\chi}_\eta^{Bayes} - \chi_0\right) \Rightarrow N(0, \mathrm{v}_0). \tag{3.6}$$

Also, for any $\alpha \in (0,1)$ we have

$$P_0\big(\chi_0 \in \mathcal{C}_n(\alpha)\big) \to 1 - \alpha.$$

Our estimation and inferential procedures achieve the semiparametric efficiency in theory. Practically, it can accommodate high-dimensional covariates or complex covariate functions, given its robustness to estimation of nuisance functional components.

4 Illustration with Gaussian Process Priors

We illustrate the general methodology with Gaussian process (GP) prior modeling on the conditional mean function for η^m. The GP regression is a popular Bayesian procedure for learning an infinite-dimensional function by specifying a GP as the prior measure. It has been extensively used among the machine learning community (Rassmusen and Williams, 2006) and it been shown to have remarkable adaptive properties with respect to the smoothness, dimensionality, or sparsity pattern of the underlying functions (van der Vaart and van Zanten, 2008, 2009, 2011; Yang and Tokdar, 2015; Yang and Dunson, 2016). Our study further strengthened the appealing features of this modern Bayesian toolkit, incorporating the PS-dependent prior adjustment. We provide primitive conditions used in our main results in the previous section. In addition, we provide details on implementation of GP priors and discuss data driven choices of tuning parameters.

4.1 Inference based on Gaussian Process Priors

Let $(W_t : t \in \mathbb{R}^p)$ be a centered, homogeneous Gaussian random field with covariance function of the form, for a given continuous function $\phi : \mathbb{R}^p \mapsto \mathbb{R}$,

$$\mathbb{E}[W_s W_t] = \phi(s - t). \tag{4.1}$$

We consider W_t as a Borel measurable map in the space of $C([0,1]^p)$, equipped with the uniform norm $\|\cdot\|_\infty$. By Bochner's theorem, there exists a finite Borel measure μ on $\mathbb{R}^p$, the spectral measure of W, s.t. $\phi(t-s) = \int_{\mathbb{R}^d} e^{i\lambda^\top (t-s)}\mu(d\lambda)$. The well-known squared exponential process (Rasmussen and Williams, 2006) comes with a Gaussian spectral measure, i.e. $\mu(\lambda) = 2^{-p}\pi^{-p/2}\exp\left(-\|\lambda\|^2/4\right)$. The covariance function of a squared exponential process takes the simple form $\mathbb{E}[W_s W_t] = \exp(-\|s-t\|^2)$, as its name suggests. We also consider a rescaled Gaussian process $(W_{a_n t} : t \in [0,1]^p)$. Intuitively speaking, a_n^{-1} can be thought as a bandwidth parameter. For a large a_n (or a small bandwidth), the prior sample path $t \mapsto W_{a_n t}$ is obtained by shrinking the long sample path $t \mapsto W_t$. Hence, it employs more randomness and becomes suitable as a prior model for less regular functions.

Below, $\mathcal{C}^s([0,1]^p)$ denotes a Hölder space with smoothness index s. Considering the

Hölder class, when we take

$$a_n \asymp n^{1/(2s_m+p)}(\log n)^{-(1+p)/(2s_m+p)}, \tag{4.2}$$

the posterior contraction rate for the conditional mean function is the minimax rate (up to some logarithm factor). Specifically, $\varepsilon_n = n^{-s_m/(2s_m+p)}(\log n)^{s_m(1+p)/(2s_m+p)}$; see Section 11.5 of Ghosal and van der Vaart (2017).

Proposition 4.1 (Squared Exponential Process Priors)**.** *Let $\widehat{\gamma}$ be an independent estimator satisfying $\|\widehat{\gamma}\|_\infty = O_{P_0}(1)$ and $\|\widehat{\gamma} - \gamma_0\|_\infty = O_{P_0}\big((n/\log n)^{-s_\pi/(2s_\pi+p)}\big)$ for some $s_\pi > 0$. Suppose $m_0 \in \mathcal{C}^{s_m}([0,1]^p)$ for some $s_m > 0$ with $\sqrt{s_\pi\, s_m} > p/2$. Consider the propensity score-dependent prior on m given by $m(d,x) = \Psi\left(W^m_{d,x} + \lambda\,\widehat{\gamma}(d,x)\right)$ where $W^m_{d,x}$ is the rescaled squared exponential process. If a_n is of the order as specified in* (4.2) *and*

$$\left(\frac{n}{\log n}\right)^{-s_m/(2s_m+p)} \ll \sigma_n \lesssim 1. \tag{4.3}$$

then the posterior distribution satisfies Theorem 3.1.

Remark 4.1. *Proposition 4.1 requires $\sqrt{s_\pi\, s_m} > p/2$ which is a trade-off between the smoothness requirement for m_0 and π_0. In particular, we obtain double robustness, i.e., a lack of smoothness of the regression function m_0 can be mitigated by exploiting regularity of the propensity score and vice versa. Referring to the Hölder class $\mathcal{C}^{s_m}([0,1]^p)$, its complexity measured by the bracketing entropy of size ε is of order $\varepsilon^{-2\upsilon}$ for $\upsilon = d/(2s_m)$. One can show that the key stochastic equicontinuity assumption in Ray and van der Vaart (2020), i.e. their condition (3.5) is violated by exploring the Sudkov lower bound (Han, 2021), when $\upsilon > 1$ or equivalently when $s_m < p/2$. It turns out that this restriction is also sufficient to verify Assumption 3 in the proof of Proposition 4.1. In contrast, our framework accommodates this non-Donsker regime as long as $\sqrt{s_\pi\, s_m} > p/2$, which enables us to exploit the product structure and a fast convergence rate for estimating the propensity score.*

Remark 4.2. *We have focused on the case where the tuning parameter a_n depends on the smoothness level of the underlying functional class. This is not necessary. An active line of research has demonstrated adaptiveness of nonparametric Bayesian methods when one assigns a prior on a_n; see van der Vaart and van Zanten (2009); Ghosal and van der Vaart (2017). When it comes to the corresponding BvM theorems (Rivoirard and Rousseau, 2012; Castillo and Rousseau, 2015), the technical proof utilizes the mixed Gaussian process structure by first conditioning on the a_n and then averaging over the random tuning parameter.*

We believe this line argument can also be adapted to our case. Nonetheless, a detailed verification is beyond the scope of the current paper, and will be pursued elsewhere.

Remark 4.3. *We have focused on the squared exponential Gaussian process, given its popularity among practitioners. Researchers can explore other GPs depending on different applications. For instance, when the derivative function is of interest, the sieve priors using the B-spline basis becomes more convenient. There are also other non-Donsker regimes, in which the posterior convergence rate for various GPs are available. If $m_0 \in \mathcal{C}^{s_m}[0,1]$ for $s_m \leqslant 1/2$, it is known that the posterior convergence rate using a Brownian motion prior is $n^{-\alpha/2}$ (Ghosal and van der Vaart, 2017), which does not pass the standard threshold $o_p(n^{-1/4})$ for semiparametric applications. One can certainly adapt the doubly robust version to this model. The power of a Bayesian approach to handle this functional class provides nice complementary options to frequentist methods. Note that the theoretical framework of AK requires Lipschitz continuity, which is not satisfied by the aforementioned class.*

4.2 Implementation of Gaussian Process Priors

We will place the Gaussian process (GP) prior on the function $\eta^m = \Psi^{-1}(m)$ and provide details on implementation of propensity score adjustments. For the computation of posterior distribution we apply standard a binary Gaussian classifier that uses Laplace approximation (Rassmusen and Williams, 2006).

Covariance kernel: We place on η^m a zero-mean GP prior with a data-driven covariance kernel described below. A benchmark kernel K is the commonly used squared exponential (SE) covariance function (Rasmussen and Williams, 2006, p.83) and with automatic relevance determination. For $(d,x),(d',x') \in \mathbb{R}^{p+1}$:

$$K\left((d,x),(d',x')\right) \;=\; \nu^2 \exp\left(\frac{-(d-d')^2}{2\lambda_0^2}\right)\exp\left(-\sum_{l=1}^{p}\frac{(x_l-x_l')^2}{2\lambda_l}\right), \tag{4.4}$$

where the hyperparameter ν^2 is the kernel variance, $\lambda_0,\cdots,\lambda_p$ are characteristic length-scales that reflect the relevance of D and each covariate in predicting η^m. In practice, they can be obtained by maximizing the log marginal likelihood.

Following Ray and van der Vaart (2020), we incorporate a correction term on the kernel function K. The resulting corrected covariance kernel K_c has an additional term based on the (estimated) Riesz representer $\widehat{\gamma}$:

$$K_c\left((d,x),(d',x')\right) = K\left((d,x),(d',x')\right) + \sigma_n^2 \widehat{\gamma}(d,x)\,\widehat{\gamma}(x',x'), \tag{4.5}$$

where $\widehat{\gamma}(d,x) = d/\widehat{\pi}(x) - (1-d)/(1-\widehat{\pi}(x))$. To obtain $\widehat{\pi}(x)$, we apply a logistic regression to $\{D_i, X_{i1}, \ldots, X_{ip}\}$. The standard deviation σ_n governs the weight of the prior correction. Later in the numerical exercise we choose σ_n such that the rate condition in Assumption 4 is satisfied. Our simulation results also suggest that the performance of our approach is stable with respect to σ_n

We describe the algorithm in the following. Let $\boldsymbol{W} = [\boldsymbol{X}, \boldsymbol{D}] \in \mathbb{R}^{n\times(p+1)}$ be the matrix for the training data, and $\boldsymbol{W}^* \in \mathbb{R}^{2n\times(p+1)}$ for the testing data:

$$\boldsymbol{W}^* = \begin{bmatrix} \boldsymbol{X}, & \boldsymbol{1}_n \\ \boldsymbol{X}, & \boldsymbol{0}_n \end{bmatrix},$$

and $\boldsymbol{\eta}_n^*$ be a $2n$-vector that gives latent function values at testing points:

$$\boldsymbol{\eta}^* = \left[\eta^m(1,X_1),\cdots,\eta^m(1,X_n),\eta^m(0,X_1),\cdots,\eta^m(0,X_n)\right]^\top.$$

Let $\boldsymbol{\eta} = \left[\eta^m(D_1,X_1),\ldots,\eta^m(D_n,X_n)\right]^\top$ denote the n-vector of latent function values at training points. For matrices $\boldsymbol{W}^*$ and $\boldsymbol{W}$, we define $K_c(\boldsymbol{W}^*,\boldsymbol{W})$ as a $2n\times n$ matrix whose (i,j)-th element is $K_c(W_i^*, W_j)$ where W_i^* is the i-th row of $\boldsymbol{W}^*$ and W_j is the j-th row of $\boldsymbol{W}$. Analogously, $K_c(\boldsymbol{W},\boldsymbol{W})$ is an $n\times n$ matrix with the (i,j)-th element being $K_c(W_i,W_j)$, and $K_c(\boldsymbol{W}^*,\boldsymbol{W}^*)$ is a $2n\times 2n$ matrix with the (i,j)-th element being $K_c(W_i^*,W_j^*)$.

Under the GP prior with mean 0 and covariance kernel K_c, the posterior of $\boldsymbol{\eta}^*$ is Gaussian with mean $\bar{\boldsymbol{\eta}}^*$ and covariance $\mathrm{V}(\boldsymbol{\eta}^*)$ can be obtained by routine procedures in Laplace approximation, see Rasmussen and Williams (2006, Chapters 3.3 to 3.5) for details. To be specific,

$$\begin{aligned} \bar{\boldsymbol{\eta}}^* &= K_c(\boldsymbol{W}^*,\boldsymbol{W})K_c^{-1}(\boldsymbol{W},\boldsymbol{W})\,\widehat{\boldsymbol{\eta}}, \\ \mathrm{V}(\boldsymbol{\eta}^*) &= K_c(\boldsymbol{W}^*,\boldsymbol{W}^*) - K_c(\boldsymbol{W}^*,\boldsymbol{W})\left(K_c(\boldsymbol{W},\boldsymbol{W}) + \boldsymbol{\nabla}^{-1}\right)^{-1} K_c^\top(\boldsymbol{W}^*,\boldsymbol{W}), \end{aligned}$$

where $\widehat{\boldsymbol{\eta}} = \mathrm{argmax}_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\boldsymbol{W},\boldsymbol{Y})$ maximizes the posterior $p(\boldsymbol{\eta}|\boldsymbol{W},\boldsymbol{Y})$ on the latent $\boldsymbol{\eta}$ and $\boldsymbol{\nabla} = -\frac{\partial^2 \log p(\boldsymbol{Y}|\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}$ is a $n\times n$ diagonal matrix with the i-th diagonal entry being $-\frac{\partial^2 \log p(\boldsymbol{Y}|\boldsymbol{\eta})}{\partial\boldsymbol{\eta}_i^2}$.

We use the Matlab toolbox GPML for computation.[10]

For each posterior sample, the draw from the mean $\bar{\boldsymbol{\eta}}^*$ and covariance $\mathrm{V}(\boldsymbol{\eta}^*)$ Gaussian distribution contains the posterior draw of $[\eta^m(1, X_1), \cdots, \eta^m(1, X_n)]^\top$ and $[\eta^m(0, X_1), \cdots, \eta^m(0, X_n)]^\top$. The posterior draw of $\eta^m(D_i, X_i) = D_i\eta^m(1, X_i) + (1 - D_i)\eta^m(0, X_i)$. Then one can compute the posterior draw of ATE by equation (2.15) with $m(d, X_i) = \Psi(\eta^m(d, X_i))$ where $d \in \{0, 1\}$ and $m(X_i) = \Psi(\eta^m(D_i, X_i))$.

5 Numerical Results

The aims of the following section are mainly two-fold; (i) to validate the preceding theoretical results and (ii) to demonstrate the proposed inferential procedure in a counterfactual analysis.

5.1 Monte Carlo Simulations

Consider following data generating process:

$$
\begin{aligned}
X_i = (X_{i1}, \ldots, X_{ip})^\top \quad \text{where} \quad X_{i1}, \ldots, X_{ip} &\overset{i.i.d.}{\sim} \text{Uniform}(-1, 1), \\
D_i \mid X_i &\sim \text{Bernoulli}\left(\Psi\left[g(X_i)\right]\right), \\
Y_i \mid X_i, D_i &\sim \text{Bernoulli}\left(\Psi\left[\mu(X_i) + D_i\tau(X_i)\right]\right),
\end{aligned}
$$

where $g(x) = \sum_{j=1}^p x_j/\sqrt[3]{j}$ and $\mu(x) = -2 + 0.2\sum_{j=1}^p x_j$. We set $p = 15$ and 30. We consider two designs for the $\tau(x)$: one is linear in x and the other is nonlinear.

Design I: $\tau(x) = 1 + 0.1\sum_{j=1}^5 x_j$.

Design II: $\tau(x) = \sum_{j=1}^3 \cos(x_j)/j$.

The true ATE is given by $\chi_0 = \mathbb{E}\left[\Psi\left(\mu(X_i) + \tau(X_i)\right) - \Psi\left(\mu(X_i)\right)\right]$. The observed variables are $\{D_i, X_{i1}, \ldots, X_{ip}\}$. To implement the Bayesian approach, we estimate the propensity score $\widehat{\pi}(x)$ by the logistic regression and estimate $m(d, x) = \Psi\left(\mu(x) + d\tau(x)\right)$ by a Gaussian process classifier. We compare the finite sample performance of the following inference procedures.

GP: The usual Bayesian approach based on Gaussian process, which corresponds to (2.15) and estimate $m(d, x)$ using the Gaussian process prior in (4.4) without correction.

[10] The GPML toolbox is developed by Rasmussen and Williams and can be downloaded from http://gaussianprocess.org/gpml/code/matlab/doc/.

GP PC: Bayesian approach with prior correction, which is calculated in the same way as GP, but incorporates the prior correction in (4.5) to the estimation of $m(d, x)$.

Matching/Matching BC : The covariate matching estimator (one to one matching with replacement) and its bias-corrected version that adjusts for the difference in covariate values by regression (Abadie and Imbens, 2011). Both are computed using the R package Matching (Sekhon, 2011).

DR TMLE: Benkeser's doubly robust targeted minimum loss-based estimator (Benkeser, Carone, Laan, and Gilbert, 2017). The nuisance parameters $g(x)$ and $m(x)$ are estimated by a super learner which combines generalized linear regression and regression splines. DR TMLE is computed by the R package drtmle (Benkeser, 2022).[11]

Tables 1 and 2 present finite sample performance of the above approaches for Designs I and II. The Bias and RMSE columns show the performance of the ATE point estimator[12] while the CP and CIL columns report the coverage rate and the average length of the 95% credible/confidence interval for ATE. The number of Monte Carlo iterations is $1,000$ and the posterior sample size is $5,000$. For GP PC, the variance of the prior correction $\sigma_n = \sqrt{pn \log n} / \sum_{i=1}^n |\widehat{\gamma}(D_i, X_i)|$. This choice of σ_n satisfies Assumption 4 and allows σ_n to increase with p. Appendix D presents additional simulation evidence to show that the performance of GP PC is stable with respect to the choice of σ_n, as long as the latter is not too small.

We make the following observations regarding Tables 1 and 2. First, the bias of GP PC is substantially smaller than that of GP, which shows that the prior correction successfully reduces the bias of the point estimator (posterior mean). Second, the confidence interval obtained from GP undercovers. On the other hand, GP PC significantly improves the coverage probability and in most cases restores it to the nominal level. This highlights the role of prior correction in Bayesian inference. Third, the matching estimator does not yield valid confidence intervals, which is not surprising given the relative large dimension of covariates.[13] The bias-corrected matching estimator substantially improves the coverage rate, but still does not restore it to the nominal level. Fourth, Benkeser's DR TMLE yields confidence interval that slightly undercovers especially when the sample size is small. Overall, as far as the validity of inference (coverage probability) is concerned, GP PC performs the best among all methods considered.

[11]See https://github.com/benkeser/drtmle.

[12]For Bayesian approaches, ATE point estimator is calculated by the posterior mean.

[13]Abadie and Imbens (2006)[p.245] noted that if p is large enough, the asymptotic distribution of a matching estimator is dominated by the bias term.

Table 1: Performance of ATE inference for Design I: GP = Gaussian process estimation of m without any correction, GP PC = with prior correction, Matching = matching estimator, Matching BC: bias-corrected matching estimator, DR TMLE = Benkeser's doubly robust targeted minimum loss-based estimator. True ATE $\chi_0 \approx 0.15$.

		$p = 15$				$p = 30$			
n	Methods	Bias	RMSE	CP	CIL	Bias	RMSE	CP	CIL
500	GP	0.0200	0.0462	0.854	0.1381	-0.0725	0.1121	0.282	0.0757
	GP PC	0.0023	0.0423	0.965	0.1787	0.0085	0.0544	0.941	0.2041
	Matching	0.0413	0.0590	0.822	0.1658	0.0728	0.0842	0.593	0.1696
	Matching BC	0.0021	0.0505	0.896	0.1670	-0.0006	0.0544	0.898	0.1739
	DR TMLE	0.0059	0.0488	0.910	0.1628	0.0095	0.0550	0.886	0.1794
1,000	GP	0.0218	0.0376	0.819	0.1007	0.0440	0.0645	0.478	0.0982
	GP PC	0.0006	0.0306	0.958	0.1230	-0.0004	0.0343	0.952	0.1367
	Matching	0.0358	0.0461	0.782	0.1177	0.0717	0.0781	0.349	0.1210
	Matching BC	-0.0001	0.0339	0.914	0.1179	0.0006	0.0399	0.861	0.1214
	DR TMLE	0.0021	0.0316	0.934	0.1180	0.0069	0.0378	0.916	0.1289
2,000	GP	0.0172	0.0286	0.782	0.0735	0.0400	0.0455	0.430	0.0745
	GP PC	0.0001	0.0207	0.956	0.0858	-0.0027	0.0253	0.935	0.0931
	Matching	0.0319	0.0384	0.689	0.0834	0.0678	0.0714	0.139	0.0861
	Matching BC	-0.0016	0.0243	0.908	0.0834	-0.0015	0.0290	0.865	0.0860
	DR TMLE	0.0000	0.0226	0.935	0.0840	0.0012	0.0268	0.912	0.0924

Table 2: Performance of ATE inference for Design I: GP = Gaussian process estimation of m without any correction, GP PC = with prior correction, Matching = matching estimator, Matching BC: bias-corrected matching estimator, DR TMLE = Benkeser's doubly robust targeted minimum loss-based estimator. True ATE $\chi_0 \approx 0.26$.

		$p = 15$				$p = 30$			
n	Methods	Bias	RMSE	CP	CIL	Bias	RMSE	CP	CIL
500	GP	0.0208	0.0464	0.887	0.1472	-0.1239	0.1829	0.281	0.0808
	GP PC	-0.0031	0.0462	0.952	0.1854	0.0070	0.0592	0.898	0.1942
	Matching	0.0466	0.0652	0.817	0.1741	0.0807	0.0914	0.572	0.1767
	Matching BC	0.0076	0.0547	0.890	0.1754	0.0088	0.0593	0.870	0.1812
	DR TMLE	0.0071	0.0493	0.923	0.1762	0.0097	0.0606	0.878	0.1943
1,000	GP	0.0230	0.0383	0.827	0.1066	0.0216	0.0748	0.491	0.1003
	GP PC	-0.0025	0.0337	0.948	0.1292	-0.0060	0.0357	0.952	0.1428
	Matching	0.0407	0.0515	0.739	0.1239	0.0770	0.0834	0.338	0.1260
	Matching BC	0.0049	0.0377	0.896	0.1241	0.0049	0.0420	0.863	0.1264
	DR TMLE	0.0024	0.0347	0.935	0.1280	0.0038	0.0401	0.917	0.1400
2,000	GP	0.0229	0.0320	0.761	0.0775	0.0400	0.0451	0.478	0.0783
	GP PC	-0.0007	0.0227	0.952	0.0909	-0.0054	0.0268	0.939	0.0985
	Matching	0.0383	0.0443	0.612	0.0878	0.0749	0.0783	0.092	0.0897
	Matching BC	0.0055	0.0267	0.900	0.0878	0.0067	0.0301	0.865	0.0895
	DR TMLE	0.0007	0.0239	0.941	0.0909	0.0013	0.0270	0.936	0.0992

We investigate the effect of σ_n on the performance of GP PC. For that purpose, we set $\sigma_n = C \times \sqrt{pn \log n} / \sum_{i=1}^n |\widehat{\gamma}(D_i, X_i)|$ and vary the value of C. The results are presented in Tables 4 and 5 of Appendix D, where the performance of GP PC appear stable when C ranges from 0.5 to 50.[14]

Our theory assumes the independence between the estimated propensity score, which appears in the prior correction, and the observations used to obtain the posterior of the conditional mean. This assumption follows Ray and van der Vaart (2020) and simplifies the technical analysis. On the other hand, as Ray and van der Vaart (2020)[p.3008] noted, this independence seems unnecessary in practice. Therefore, to make the implementation as convenient as possible, our simulation exercises so far have used the full sample in estimating the propensity score and drawing the posterior.[15] As Tables 1 and 2 show,

[14]GP PC reduces to GP when $C = 0$. Since GP substantially undercovers in our experiments, C (and thus σ_n) cannot be too small.

[15]This implementation strategy also follows Ray and Szabó (2019), an empirical companion paper to

confidence intervals based on GP PC yield good coverage probabilities even when we do not split the sample. Table 6 in Appendix D presents the results when we apply sample-splitting in implementing the GP PC. In our empirical application below, as the sample size is relatively small ($n = 365$ after trimming out observations with extremely small or large propensity score), we use the full sample for both the propensity score estimation and the posterior draw.

Overall, our simulation results suggest that Bayesian inference with prior correction can be a useful tool for conducting valid ATE inference when the model is fairly complicated.

5.2 Empirical Application

We apply our method to the data from the National Supported Work (NSW) program. The dataset, which has been used by Dehejia and Wahba (1999), Abadie and Imbens (2011) and Armstrong and Kolesár (2021), among others, contains a treated sample of 185 men from the NSW experiment and a control sample of 2490 men from the Panel Study of Income Dynamics (PSID). We also refer readers to Imbens (2004) and Abadie and Cattaneo (2018) for comprehensive reviews related to the data. We slightly depart from previous studies by focusing on a binary outcome Y: the unemployment indicator for year 1978, which is defined as an indicator for zero earnings. The treatment D is the participation in the NSW program. We consider two specifications for the selection of covariates X: Spec I follows Table 1 of Armstrong and Kolesár (2021) and Spec II follows Table 3 of Dehejia and Wahba (2002).

Spec I: Covariates X contain 9 variables: age, education, indicators for black and Hispanic, indicator for marriage, earnings in 1974, earnings in 1975, and unemployment indicators for 1974 and 1975.

Spec II: Covariates X contain 15 variables: the 9 variables in Spec I and their functions: squared age, squared education, squared earnings in 1974, squared earnings in 1975, indicator for no degree (education < 12), indicator for unemployment in 1974 $\times$ indicator for Hispanic.

Table 3 presents the ATE estimates from the Bayesian inference with and without prior correction, matching with and without bias-correction and Benkeser's DR TMLE. As a benchmark, We also include the experimental estimates for the sample where both the

Ray and van der Vaart (2020). A similar strategy is also taken by Ignatiadis and Wager (2022)[p.8] when they construct the confidence interval for nonparametric empirical Bayes analysis.

treated and control subsamples are from the NSW program. Since the treated and control groups for the nonexprimental data are higly unbalanced in covariates, we discard observations with the estimated propensity score outside the range $[0.05, 0.95]$. The numbers of of treated units (n_1) and control units (n_0) after trimming are comparable to the experimental data.

In Table 3, our Bayesian inference with prior correction (GP PC) finds that the job training problem reduced the probability of unemployment by about 21% under Spec I and about 16% under Spec II. Both are statistically significant at 5% level. The results barely change when we vary σ_n. The experimental data also reveals a 5%-level significant effect of the program in reducing unemployment, though the point estimate is smaller (around 11%). There is a considerable overlapping between the 95% credible interval of GP PC and the experimental estimate. Under Spec II, the uncorrected Gaussian process inference (GP) generates a much smaller estimate than other approaches. As our simulation results (Tables 1 and 2) show that GP performs badly when the number of covariates (p) is large and the sample size is small, we suspect that the GP estimate here is not reliable either. The matching estimates with and without bias correction turn out insignificant at 5% level under Spec I but become 5%-level significant under Spec II. Similar to GP PC, Benkeser's DR TMLE yields a negative estimate that is significant at 5% level. We also note that the length of confidence interval is shorter for the GP PC than the frequentist approaches.

Table 3: Nonexperimental and experimental estimates of ATE for the NSW data: n_1 and n_0 are treated and control sample sizes. GP = Gaussian process estimation of m without any correction, GP PC = with prior correction, Matching = matching estimator, Matching BC: bias-corrected matching estimator, DR TMLE = Benkeser's doubly robust targeted minimum loss-based estimator. $\sigma_n = C \times \sqrt{p\,n \log n} / \sum_{i=1}^{n} |\widehat{\gamma}(D_i, X_i)|$. The asterisk denotes 5% statistical significance.

	Spec	n_1	n_0	Methods	ATE	95% CI	CIL
Non. Exper	I	145	220	GP	-.2169*	[-.3109, -.1180]	.1929
				GP PC ($C = 1$)	-.2058*	[-.3169, -.0904]	.2266
				GP PC ($C = 10$)	-.2026*	[-.3172, -.0780]	.2393
				Matching	-.1687	[-.3387, .0013]	.3400
				Matching BC	-.1595	[-.3274, .0084]	.3359
				DR TMLE	-.1656*	[-.3198, -.0114]	.3084
	II	132	220	GP	-.0217	[-.0636, .0194]	.0830
				GP PC ($C = 1$)	-.1633*	[-.2873, -.0020]	.2854
				GP PC ($C = 10$)	-.1684*	[-.2932, -.0047]	.2885
				Matching	-.2043*	[-.3861, -.0225]	.3636
				Matching BC	-.1969*	[-.3770, -.0167]	.3603
				DR TMLE	-.1900*	[-.3549, -.0252]	.3297
Exper.		185	260	Mean diff.	-.1106*	[-.1957, -.0255]	.1701
	I			Reg. with cov.	-.1132*	[-.2006, -.0258]	.1747
	II			Reg. with cov.	-.1045*	[-.1931, -.0158]	.1774

6 Conclusion

There are several directions that would be interesting to pursue in future work. First, the template of our theoretical investigation should also be useful for complicated structural models, where the likelihood functions are computationally intractable. These analytical difficulties can often be alleviated by Bayesian methods, which has proven to be successful in many areas. Second, one can extend our analysis to other interesting causal effects other than ATE. Causal mediation analysis has attract a lot of attention recently. For a particular causal parameter, Diaz et al. (2021) presented an explicit influence function and proposed frequentist type efficient estimators. Because their plug-in type estimator involves multi-dimensional integral, it is desirable to explore Bayesian tools. Further investigation of the possibility to generalize our methodology to nonlinear functionals is needed; see Examples 4.2-4.4 from Castillo and Rousseau (2015). Additionally, it would be beneficial to incorporate some more sophisticated prior such as the Bayesian additive regression trees (BART) and prove the corresponding BvM theorem. The latter prior is shown to be particularly attractive in the high dimensional regime and can effectively conduct variable selection. These topics are beyond the scope of the current paper and will be examined elsewhere.

References

Abadie, A., and G. W. Imbens (2006): "Large sample properties of matching estimators for average treatment effects," *econometrica*, 74(1), 235–267.

——— (2011): "Bias-corrected matching estimators for average treatment effects," *Journal of Business & Economic Statistics*, 29(1), 1–11.

Andrews, I., and A. Mikusheva (2022): "Optimal decision rules for weak gmm," *Econometrica*, 90, 715–748.

Armstrong, T. B., and M. Kolesár (2021): "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," *Econometrica*, 89(3), 1141–1177.

Benkeser, D. (2022): *drtmle: Doubly-Robust Nonparametric Estimation and Inference.*R package version 1.1.1.

Benkeser, D., M. Carone, M. V. D. Laan, and P. Gilbert (2017): "Doubly robust nonparametric inference on the average treatment effect," *Biometrika*, 104(4), 863–880.

Castillo, I. (2012): "A semiparametric Bernstein–von Mises theorem for Gaussian process priors," *Probability Theory and Related Fields*, 152, 53–99.

Castillo, I., and J. Rousseau (2015): "A Bernstein–von Mises theorem for smooth functionals in semiparametric models," *Annals of Statistics*, 43, 2353–2383.

Cattanoe, M., and M. Jansson (2018): "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," *Econometrica*, 86, 955–995.

Chamberlain, G., and G. Imbens (2003): "Nonparametric applications of Bayesian inference," *Journal of Business and Economic Statistics*, 21, 12–18.

Chen, X., T. M. Christensen, and E. Tamer (2018): "Monte Carlo confidence sets for identified sets," *Econometrica*, 86, 1965–2018.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, 36(2), 808–843.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2017): "Locally robust semiparametric estimation," *arXiv preprint*, arXiv:1608.00033.

Chernozhukov, V., W. Newey, and R. Singh (2020a): "Automatic Debiased Machine Learning of Causal and Structural Effects," *arXiv preprint arXiv:1809.05224*.

——— (2020b): "De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers," *arXiv preprint arXiv:1802.08667*.

Dehejia, R. H., and S. Wahba (1999): "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American statistical Association*, 94(448), 1053–1062.

——— (2002): "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and statistics*, 84(1), 151–161.

Farrell, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189(1), 1–23.

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep neural networks for estimation and inference," *Econometrica*, 89(1), 181–213.

Florens, J.-P., and A. Simoni (2019): "Gaussian processes and Bayesian moment estimation," *Journal of Business and Economic Statistics*, forthcoming.

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000): "Convergence rates of posterior distributions," *The Annals of Statistics*, 28, 500–531.

Ghosal, S., and A. Van der Vaart (2017): *Fundamentals of nonparametric Bayesian inference*, vol. 44. Cambridge University Press.

Ghosal, S., and A. van der Vaart (2017): *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

Giacomini, R., and T. Kitagawa (2020): "Robust Bayesian inference for set-identified models," *Econometric*, forthcoming.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, pp. 315–331.

Hahn, P., J. Murray, and C. Carvalho (2020): "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects," *Bayesian Analysis*, 15, 965–1056.

Han, Q. (2021): "Set structured global empirical risk minimizers are rate optimal in general dimensions," *The Annals of Statistics*, 49(5), 2642–2671.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71(4), 1161–1189.

Hirshberg, D. A., and S. Wager (2021): "Augmented minimax linear estimation," *The Annals of Statistics*, 49(6), 3206–3227.

Ignatiadis, N., and S. Wager (2022): "Confidence intervals for nonparametric empirical Bayes analysis," *Journal of the American Statistical Association*, pp. 1–18.

Imbens, G. W. (2021): "Statistical significance, p-values, and the reporting of uncertainty," *Journal of Economic Perspectives*, 35(3), 157–174.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, 62, 1349–1382.

Norets, A. (2015): "Bayesian regression with nonparametric heteroskedasticity," *Journal of Econometrics*, 185, 409–419.

Rasmussen, C. E., and C. K. Williams (2006): *Gaussian processes for machine learning*. MIT press Cambridge, MA.

Rassmusen, C., and C. Williams (2006): *Gaussian processes for machine learning*. MIT.

Ray, K., and B. Szabó (2019): "Debiased Bayesian inference for average treatment effects," *Advances in Neural Information Processing Systems*, 32.

Ray, K., and A. van der Vaart (2020): "Semiparametric Bayesian causal inference," *Annals of Statistics*, 48, 2999–3020.

Ritov, Y., P. J. Bickel, A. C. Gamst, and B. J. K. Kleijn (2014): "The Bayesian Analysis of Complex, High-Dimensional Models: Can It Be CODA?," *Statistical Science*, 29(4), 619–639.

Rivoirard, V., and J. Rousseau (2012): "Bernstein–von Mises theorem for linear functionals of the density," *Annals of Statistics*, 40(3), 1489–1523.

Robins, J. M., and Y. Ritov (1997): "Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models," *Statistics in medicine*, 16(3), 285–319.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90(429), 122–129.

Rubin, D. (1981): "Bayesian bootstrap," *The Annals of statistics*, 9, 130–134.

——— (1984): "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of statistics*, 12, 1151–1172.

Saarela, O., L. Belzile, and D. Stephens (2016): "A Bayesian view of doubly robust causal inference," *Biometrika*, 103, 667–681.

Sekhon, J. S. (2011): "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R," *Journal of Statistical Software*, 42.

van der Vaart, A. (1998): *Asymptotic statistics*. Cambridge University Press.

van der Vaart, A. W., and J. H. van Zanten (2008): "Rates of contraction of posterior distributions based on Gaussian process priors," *The Annals of Statistics*, 36, 1435–1463.

——— (2009): "Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth," *The Annals of Statistics*, 37, 2655–2675.

——— (2011): "Information Rates of Nonparametric Gaussian Process Methods," *Journal of Machine Learning Research*, 12.

Yang, Y., G. Cheng, and D. Dunson (2015): "Semiparametric Bernstein-von Mises Theorem: Second order studies," *working paper*, arxiv:1503.04493v1.

Yang, Y., and D. B. Dunson (2016): "Bayesian manifold regression," *The Annals of Statistics*, 44(2), 876–905.

Yang, Y., and S. Tokdar (2015): "Minimax-optimal nonparametric regression in high dimensions," *Annals of Statistics*, 43, 652–674.

Yiu, A., R. J. Goudie, and B. D. Tom (2020): "Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood," *Biometrika*, 107, 857–873.