Journal of Statistical Planning and Inference 225 (2023) 171–187

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Semiparametric Bayesian doubly robust causal estimation

Yu Luo [a,*], Daniel J. Graham [b], Emma J. McCoy [c]

[a] *Department of Mathematics, King's College London, United Kingdom*
[b] *Department of Civil and Environmental Engineering, Imperial College London, United Kingdom*
[c] *Department of Statistics, London School of Economics and Political Science, United Kingdom*

ARTICLE INFO

Article history:
Received 13 December 2021
Received in revised form 30 August 2022
Accepted 28 December 2022
Available online 30 December 2022

MSC:
62D20
62F15
62G05

Keywords:
Causal inference
Bayesian estimation
Empirical likelihood
Double robustness
Semiparametric inference
Propensity score adjustments

ABSTRACT

Frequentist semiparametric theory has been used extensively to develop doubly robust (DR) causal estimation. DR estimation combines outcome regression (OR) and propensity score (PS) models in such a way that correct specification of just one of two models is enough to obtain consistent parameter estimation. An equivalent Bayesian solution, however, is not straightforward as there is no obvious distributional framework to the joint OR and PS model, and the DR approach involves a semiparametric estimating equation framework without a fully specified likelihood. In this paper, we develop a fully semiparametric Bayesian framework for DR causal inference by bridging a nonparametric Bayesian procedure with empirical likelihood via semiparametric linear regression. Instead of specifying a fully probabilistic model, this procedure is only realized through relevant moment conditions. Crucially, this allows the posterior distribution of the causal parameter to be simulated via Markov chain Monte Carlo methods. We show that the posterior distribution of the causal estimator satisfies consistency and the Bernstein–von Mises theorem, when either the OR or PS is correctly specified. Simulation studies suggest that our proposed method is doubly robust and can achieve the desired coverage rate. We also apply this novel Bayesian method to a real data example to assess the impact of speed cameras on car collisions in England.

1. Introduction

Causal inference is concerned with estimation of the relationship between a treatment and an outcome. In observational studies, assignment of the treatment may be not random and this can lead to potential biases due to confounding effects. To infer cause–effect relationships in these settings, we require an appropriate causal inferential framework for identification and statistical modeling methods appropriate to the data. Propensity score (PS) and outcome regression (OR) have been extensively used to reduce confounding bias in estimating causal effects (Rosenbaum and Rubin, 1983). The PS is defined as the conditional probability of the treatment assignment given confounding covariates. The treatment–outcome relationship can be adjusted via estimated PSs in a variety of ways to reduce biases in the OR (see for example Hahn, 1998; Imbens, 2000; Hirano et al., 2003). The OR model performs regression on a set of confounders in addition to the treatment to reduce confounding bias (e.g. via linear regression). Double robust (DR) causal estimators, as pioneered in the past few decades in frequentist inference, have great appeal as they combine OR and PS models such that one only needs to correctly specify either the OR or the PS model to obtain consistent causal estimation. Specifically, Robins et al. (1992) proposed doubly robust estimation via regression by including the estimated PSs in the OR, while Robins et al.

[*] Corresponding author.
E-mail address: yu.luo@kcl.ac.uk (Y. Luo).

https://doi.org/10.1016/j.jspi.2022.12.005

(1994) and Scharfstein et al. (1999) introduced an augmented inverse probability-weighted estimating equation approach based on regression for incomplete data, which also possesses the DR property. Bang and Robins (2005) extended this approach to longitudinal marginal structural models. Cao et al. (2009) proposed alternative doubly robust estimators to improve the robustness of previous methods when some estimated PSs are close to zero. Little and An (2004) achieved the DR property by adding a flexible function of the PS in the OR. Recently, Chernozhukov et al. (2018) developed a debiased approach via double machine learning to estimate the treatment effect which allows that both the OR and the PS are estimated by flexible machine learning models.

A major challenge in the frequentist DR method is that the uncertainty in the estimated PS has been ignored, and there have been a lot of efforts to derive the theoretical approximation to the variance (see for example Rubin and Thomas, 1996). Abadie and Imbens (2016) showed that the estimator of the average treatment effect (ATE) and the parameters in the PS model are jointly normal distributed, and suggested to adjust the variance of the ATE downward to account for the estimated PS (also see Henmi and Eguchi, 2004). These results indicate that the Bayesian method can be an appealing alternative in the PS analysis as it naturally quantifies the uncertainty via a probabilistic statement. Bayesian inference avoids unknown deterministic quantities to studying distributions, which has proven to be increasingly powerful in a large spectrum of applications over the decades. In the context of causal inference, the Bayesian approach offers natural solutions of quantifying uncertainty with multiple component models, predicting complex causal quantities in terms of probabilistic statements and incorporating prior information when expert knowledge becomes available. For example, the posterior predictive distribution on treatment effects may provide more intuition for policy making. There is growing literature on the Bayesian paradigm of causal inference recently, but the primary focus is on the complicated procedures for modeling the OR. However, such approaches utilize flexible and non-parametric modeling to avoid the misspecification in the OR, which is not the primary focus of this paper. Specifically, we will address issues of how to accommodate the two-stage Bayesian causal analysis via the PS adjustment, in the context of misspecification, and in addition develop fully Bayesian strategies to reflect this. Rubin (1985) discussed the importance of PSs as randomized probabilities, and should be of great interest to the applied Bayesian practitioner. The uncertainty of the PS can be well calibrated via Bayesian inference if the PS is an adequate summary of confounders that can be obtained and retains strong ignorability (Rubin, 1985). Although a well-calibrated Bayesian argument is appealing in the context of PS analysis to incorporate all sources of uncertainty, there are some debates on the actual estimation of the PS or the OR from a fully Bayesian perspective. In a conventional Bayesian causal analysis under the PS adjustment, it is natural to fit a joint model for the treatment and outcome (McCandless et al., 2009); however, Rubin (2007) argued that PSs should be estimated without knowing the outcome in order to have balancing properties. Thus, joint estimation is not a true Bayesian PS adjustment approach as the posterior distribution of the PS depends on the outcome (Saarela et al., 2015), leading to a biased estimator of the treatment effect. To avoid this dependence, McCandless et al. (2010) suggested to cut the feedback from the outcome model into the PS model in a fully Bayesian setting. In addition, an alternative to this cut feedback approach is to assume a complete separation between the PS and OR models, and conduct a two-step procedure (Kaplan and Chen, 2012), using the Bayesian propensity score model in the first step, followed by a Bayesian outcome model in the second step. This approach has been shown to provide a preferred estimation procedure (Stephens et al., 2022). While frequentist inference has provided well-established theory for DR causal estimation; the Bayesian counterpart, however, is not straightforward as the standard Bayesian inference is based on the assumption that the distribution of the data belongs to the chosen model class. Therefore, a small violation of this assumption can have a large impact on the outcome of a Bayesian procedure. Recently, there have been some attempts for Bayesian DR estimation (Gustafson, 2012; Graham et al., 2016; Saarela et al., 2016; Luo et al., 2021), due to the advancement in Bayesian non-parametric inference for models formulated through moment restrictions (Chamberlain and Imbens, 2003; Bornn et al., 2019). Specifically, Bayesian DR estimation has been attempted via the Bayesian bootstrap strategy (Rubin, 1978; Newton and Raftery, 1994), with causal quantities derived from posterior predictive distributions. Stephens et al. (2022) has shown that the combination of two-step estimation with a Bayesian bootstrap gives a fully Bayesian procedure with good frequentist properties.

On the frequentist side, a nonparametric analogue of the usual likelihood theory, empirical likelihood, has been developed, which has been proven to possess many properties of conventional parametric likelihood inference (Owen, 2001). Qin and Lawless (1994) linked empirical likelihood and estimating equations for parameter estimation so that it can also satisfy certain moment conditions. This framework offers a different perspective to construct Bayesian methods for models formulated through moment restrictions. Schennach (2005) proposed a computationally convenient representation as an empirical-likelihood-type likelihood where the probability weights are obtained via exponential tilting, admitting a nonparametric limit of a Bayesian procedure for moment condition models. Chib et al. (2018) extended this representation to incorporate misspecified moment condition models. Yiu et al. (2020) used this framework to study unequal probability sampling; however, they focused an augmented inverse probability weighted estimator, resulting in a higher bias in estimating the causal effect in small sample cases. In addition, it is still questionable to deploy plug-in approaches in conventional prior-to-posterior updating as a fully Bayesian approach (Robins and Wasserman, 2000; Robins et al., 2015).

In light of these considerations, in this paper, we attempt to develop a fully semiparametric Bayesian estimation procedure for DR causal inference which bridges the nonparametric Bayesian method with empirical likelihood theory via semiparametric linear regression. Specifically, we employ the conventional Bayesian paradigm using the product of a prior and a likelihood defined by moment constraints, and develop the asymptotic results as the semiparametric frequentist M-estimator. We specifically achieve the following:

1. Improving robustness — by introducing an augmented, rather than weighted regression model, which is less sensitive to estimation bias from extreme values in the PS distribution and reduces finite sample bias;
2. Developing fully Bayesian DR inference — via a two-step procedure for Bayesian regression adjustment. As discussed previously, the two-step procedure has been shown to yield a fully Bayesian procedure with good frequentist properties (Stephens et al., 2022);
3. Investigating posterior consistency and asymptotic normality under model misspecification. We develop asymptotic results when either the PS or the OR model is misspecified, which provides a valid uncertainty quantification for the ATE via the posterior (predictive) distribution.

Although it is not widely agreed that fully Bayesian doubly robust inference is possible, it is argued that the semiparametric efficient adjustment via the fitted PS (see for example Robins et al., 1992) or augmented inverse probability weighting (Bang and Robins, 2005) necessarily involves a plug-in strategy, and as such cannot be regarded as fully Bayesian. This issue can be overcome using Bayesian non-parametric calculations and an alternative view of Bayesian estimation that adopts a decision theoretic standpoint; furthermore, the Bayesian bootstrap, and its extensions, can again be justified as a (Monte Carlo-based) exact Bayesian inference procedure, which offers an alternative view of Bayesian estimation via an exact Bayesian nonparametric calculation (Stephens et al., 2022; Luo et al., 2021). It is under this standpoint that our approach offers an appealing alternative paradigm to the previous Bayesian causal inference strategy via the Bayesian bootstrap. Unlike the Bayesian predictive inference approach in Saarela et al. (2016) and Luo et al. (2021), the posterior distribution of the causal parameter can be derived via the conventional prior-to-posterior update and simulated using Markov chain Monte Carlo (MCMC) methods. We also show the asymptotic behavior of the posterior distribution of the causal parameter in terms of consistency and asymptotic normality. Our proposed method opens up the possibility of developing Bayesian causal inference in other settings, such as marginal structural models and dynamic treatment regimes. Our approach incorporates the usual Bayesian prior-to-posterior updating framework, which provides a means of informed and coherent decision making in the presence of uncertainty.

The rest of the paper is organized as follows. Section 2 recaps DR causal inference via semiparametric linear regression. Section 3 introduces the empirical-likelihood-based Bayesian representation for causal inference using regression with the estimated PS, followed by the asymptotic behavior of the posterior distribution of the causal parameter for misspecifying either the PS or OR model. We provide simulation studies to compare the proposed method with other Bayesian approaches and the frequentist double machine learning method in Section 5, followed by a real example in Section 6. Finally, Section 7 presents some concluding remarks and future research directions.

2. Background

2.1. The average treatment effect

In a causal inference setting, let $Z_i = (Y_i, D_i, X_i)$, $i = 1, \ldots, n$, where for the ith unit of observation Y_i denotes a response, D_i a binary treatment (or exposure) received, and X_i a vector of pre-treatment covariates or confounder variables. Under an appropriate experimental design, we know that $D_i \perp\!\!\!\perp X_i$. Therefore, the average treatment effect (ATE) can be directly estimated via

$$\theta = \mathbb{E}\left[Y_i \,|D_i = 1\right] - \mathbb{E}\left[Y_i \,|D_i = 0\right].$$

This causal quantity of interest measures the expected change of Y when D is changed from 0 to 1. However, in observational studies, $D_i \not\perp\!\!\!\perp X_i$ and appropriate adjustment is required to estimate causal effects. The PS, which is defined as the conditional probability of the treatment assignment given confounding covariates, $f_D\left(d\,|x\right) \equiv e\left(x\right)$, has been increasingly used to reduce confounding bias in estimating causal effects. As shown in Fig. 1, it can block the backdoor path from $X \to D$, which effectively results in $X \perp\!\!\!\perp D$ given the PS under certain assumptions (Rosenbaum and Rubin, 1983). The treatment–outcome relationship can be adjusted via an estimated PS either by weighting, matching, stratification, or regression adjustment. In a frequentist regression setting, as $e(X)$ is a balancing score, the ATE can be evaluated as

$$\theta = \int_x \left\{\mathbb{E}\left[Y\,|d = 1, x, e(x)\right] - \mathbb{E}\left[Y\,|d = 0, x, e(x)\right]\right\} f(x)dx.$$

Suppose we specify the PS model parameterized by γ, i.e., $e\left(x; \gamma\right)$. Typically, γ is estimated from the observed data on x and d. In this setting, $e\left(x; \gamma\right)$ is a balancing score only if the PS model is correctly specified. Given the existence of the true value γ_0 of γ, the estimator of γ is consistent for γ_0 under frequentist likelihood theory or the Bernstein–von Mises theorem, justifying holding the balancing property asymptotically.

2.2. Doubly robust causal inference via semiparametric linear regression.

The DR estimator indicates that we can obtain a consistent estimator of the causal effect from D to Y if at least one of the OR model and the PS model is correctly specified. Scharfstein et al. (1999) proposed an augmented regression which

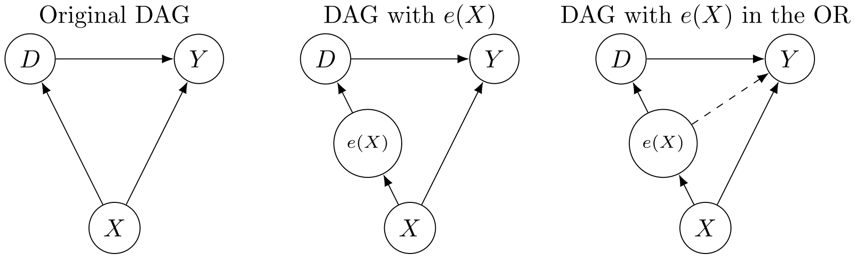


Fig. 1. Directed acyclic graphs (DAGs) with the adjustment via the propensity score. The dashed arrow from $e(X)$ to Y indicates the modeling dependence in the outcome regression model rather than the true data generating mechanism.

has a bias correction property, and the ATE possesses the DR property when estimates of θ are obtained by solving the Q score equations

$$\sum_{i=1}^{n} \hat{\kappa}_i (x_i, d_i) \frac{\partial u (d_i, x_i; \xi)}{\partial \xi^\top} [y_i - u (d_i, x_i; \xi)] = 0 \tag{2.1}$$

where $\hat{\kappa}_i (X_i, D_i) = \frac{I_1(D_i)}{e(X_i;\hat{\gamma})} + \frac{1-I_1(D_i)}{1-e(X_i;\hat{\gamma})}$ and $u (D_i, X_i; \xi)$ is the mean of the conditional density of outcome, Y, given covariates, X and D, and parameterized by ξ. The weights, $\hat{\kappa}_i (x_i, d_i)$, in (2.1) become extremely large when there are extreme values of PSs, leading to skewed and highly variable sampling distributions of those estimators (Robins and Wang, 2000). Therefore, we retain our focus on the DR estimator constructed via semiparametric linear regression (Robins et al., 1992). Suppose the true OR model is

$$Y_i = \theta_0 D_i + h_0(X_i, \beta_0) + \epsilon_i, \quad \forall i = 1, 2, \ldots, n \tag{2.2}$$

where ϵ_i is a random residual error term with $\mathbb{E}(\epsilon_i|D_i, X_i) = 0$ and $Var(\epsilon_i|D_i, X_i) = \sigma^2 < \infty$. The function $h_0(\cdot)$ is some unknown component to describe the relationship between X and Y. In this setting, θ_0 is the ATE and parameter of interest. Given that we estimate the PS either via maximum likelihood estimation or a fully Bayesian procedure, one strategy to estimate the ATE based on covariate adjustments is to add the estimated PS into the OR as

$$Y_i = \theta D_i + h_1(X_i, \beta) + \phi e\left(X_i; \hat{\gamma}\right) + \epsilon_i, \quad \forall i = 1, 2, \ldots, n. \tag{2.3}$$

This model has the advantage of double robustness. If the $h_1(x, \beta)$ component of the OR model is correctly specified and reflects the data-generating mechanism as $h_0(x, \beta_0)$, then the estimator of θ will be consistent regardless of whether the PS model was correctly specified as $\phi = 0$. On the other hand, if the PS model is correctly specified, as demonstrated in the causal diagram (Fig. 1), it blocks the backdoor path from $X \to D$, which leads to $X \perp\!\!\!\perp D \mid e(X)$. Therefore, the OR model based on (2.3) will still yield a consistent causal estimand, i.e., θ, even if the $h_1(x)$ component of the OR model was incorrectly specified. Alternatively, this can be also explained using the semiparametric efficiency theory. The estimating system in (2.3) becomes

$$\sum_{i=1}^{n} \mathbf{U}_i (\beta, \theta, \phi) = \sum_{i=1}^{n} \begin{pmatrix} d_i \\ \frac{\partial h_1(x_i,\beta)}{\partial \beta} \\ e\left(x_i; \hat{\gamma}\right) \end{pmatrix} [Y_i - \theta d_i - h_1(x_i, \beta) - \phi e\left(x_i; \hat{\gamma}\right)] = 0. \tag{2.4}$$

Subtracting the third equation from the first yields the equivalent form of the estimating equation

$$\sum_{i=1}^{n} (d_i - e\left(x_i; \hat{\gamma}\right)) \left(Y_i - \theta d_i - h_1(x_i, \beta) - \phi e\left(x_i; \hat{\gamma}\right)\right) = 0$$

with solution

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \left(d_i - e\left(x_i; \hat{\gamma}\right)\right) (Y_i - h_1(x_i, \hat{\beta}) - \phi e\left(x_i; \hat{\gamma}\right))}{\sum_{i=1}^{n} \left(d_i - e\left(x_i; \hat{\gamma}\right)\right) d_i}.$$

This is the generalized estimator (or G-estimator) proposed in Robins et al. (1992), resulting in consistent inference for θ and robustness to misspecification. Accordingly, double robustness is achieved via the regression approach. For simplicity, we assume that there is a linear relationship between Y and D in (2.2) without interaction terms between D and X; however, it can be extended to include interactions in (2.2) with interaction terms between $e\left(x_i; \hat{\gamma}\right)$ and X included in (2.3) to achieve double robustness. We show an example with interaction terms in Appendix A.

As demonstrated above, least square estimation of the conditional mean model, $\mathbb{E}\left[Y_i \,|x_i, d_i, \theta, \beta, \phi\right]$, is equivalent to solving a set of estimating equations, $\sum_{i=1}^{n} \mathbf{U}_i (\beta, \theta, \phi) = \mathbf{0}$, defined in (2.4). This semiparametric procedure yields a frequentist DR estimator for θ and the variance–covariance matrix can be computed using the theory of estimating equations (Tsiatis, 2007). Stephens et al. (2022) extensively discussed how a parametric PS should be incorporated in the Bayesian setting, which is not the focus of this paper. Therefore, we adopt a two-step approach which assumes a complete separation between the PS and OR models and focus only on the estimating equations deriving from the OR model.

3. Bayesian doubly robust causal inference

3.1. The Bayesian bootstrap

In Bayesian inference, the objective is to draw a posterior sample of the parameter $\xi := (\beta, \theta, \phi)$ based on the observed data $z_i = (y_i, d_i, x_i)$ for $i = 1, \ldots, n$, i.e.,

$$\pi\left(\xi \left| z_1, \ldots, z_n\right.\right) \propto \pi_0\left(\xi\right) \prod_{i=1}^{n} p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right)$$

where $\pi_0\left(\xi\right)$ is the prior distribution. To bridge the data generating mechanism, $p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right)$, with the set of estimating equations in (2.4), one can specify a non-parametric distribution for $p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right)$. If we assume that the data points are realizations a multinomial model on the finite set $\{y_1, \ldots, y_n\}$ with unknown probability $\varpi = (\varpi_1, \ldots, \varpi_n)$ and a priori of $(\varpi_1, \ldots, \varpi_n) \sim \text{Dirichlet}\,(1, \ldots, 1)$, then $p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right) = \sum_{k=1}^{n} \varpi_k \delta_{y_k}\left(y_i\right)$ and the equivalent estimating equation becomes

$$\sum_{k=1}^{n} \varpi_k \mathbf{U}_k\left(\xi\right) = \mathbf{0}.$$

Therefore, we can repeatedly draw a single set of random weights $(\varpi_1, \ldots, \varpi_n)$ from $\text{Dirichlet}\,(1, \ldots, 1)$ to obtain the posterior distribution of ξ under the limiting Dirichlet process specification (Rubin, 1981; Newton and Raftery, 1994), and the prior information can be incorporated via the sampling-importance resampling method (Newton and Raftery, 1994). Bayesian causal inference based on this framework has been extensively elaborated by Chamberlain and Imbens (2003), Saarela et al. (2015), Graham et al. (2016), Saarela et al. (2016) and Stephens et al. (2022).

3.2. Exponentially tilted empirical likelihood

On the other hand, empirical likelihood provides another form of non-parametric likelihood which seeks to reweight the sample so that it can also satisfy moment conditions (Qin and Lawless, 1994). This has opened up possibilities for dealing with models motivated via the estimating equation framework. Lazar (2003) proposed a heuristic strategy of incorporating empirical likelihood in a Bayesian framework for the data distribution. Subsequently, Schennach (2005) proposed a nonparametric likelihood method via exponential tilting to include moment conditions, which is similar to the empirical likelihood representation. Specifically, $p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right), i = 1 \ldots, n$, can be specified as nonparametric probabilities, $(p_1, \ldots, p_n)$, that minimize the Kullback–Leibler (KL) divergence between the probabilities $(p_1, \ldots, p_n)$ assigned to each sample observation and the empirical probabilities $(\frac{1}{n}, \ldots, \frac{1}{n})$, which can be summarized as the following constrained optimization problem

$$\begin{aligned} &\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} -p_i \log(np_i) \\ &\text{subject to} \sum_{i=1}^{n} p_i = 1 \text{ and } \sum_{i=1}^{n} p_i \mathbf{U}_i\left(\xi\right) = \mathbf{0}. \end{aligned} \tag{3.1}$$

Then the solution for p_i can be expressed via a Lagrange multiplier λ as a dual representation

$$p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right) = \frac{\exp\left(\hat{\lambda}(z, \xi)^{\top} \mathbf{U}_i\left(\xi\right)\right)}{\sum_{j=1}^{n} \exp\left(\hat{\lambda}(z, \xi)^{\top} \mathbf{U}_j\left(\xi\right)\right)} \tag{3.2}$$

where $\hat{\lambda}(z, \xi) = \arg\min_{\lambda} \frac{1}{n} \sum_{i=1}^{n} \exp\left(\lambda^{\top} \mathbf{U}_i\left(\xi\right)\right)$. Therefore,

$$\pi(\xi \left| z_1, \ldots, z_n\right.) \propto \pi_0(\xi) \times \prod_{i=1}^{n} \frac{\exp\left(\hat{\lambda}(z, \xi)^{\top} \mathbf{U}_i\left(\xi\right)\right)}{\sum_{j=1}^{n} \exp\left(\hat{\lambda}(z, \xi)^{\top} \mathbf{U}_j\left(\xi\right)\right)}.$$

The function $\pi(\xi \left| z_1, \ldots, z_n\right.)$ is the posterior distribution for ξ via the exponentially tilted empirical likelihood, and we can use MCMC to obtain the sample from this posterior distribution. Suppose we have an independent proposal distribution, $q(\xi|z)$, for example, a multivariate Student's t distribution where the location parameter is the maximum likelihood estimate of empirical likelihood $\prod_{i=1}^{n} p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right)$ and the dispersion matrix is the estimated variance–covariance matrix of the fitted model in (2.3). With this independent proposal, we can draw a sample of ξ as follows, with initial values ξ^0 and superscript $i = 1, 2, \ldots$ denoting the iteration number:

1. Propose ξ^p from the proposal, $q(\xi|z)$, and evaluate the quantity in (3.2), that is, solving the optimization problem in (3.1).
2. Calculate the Metropolis–Hastings probability of the move from ξ^{i-1} to ξ^p,
$$\alpha\left(\xi^{i-1} \to \xi^p\right) = \min\left\{1, \frac{\pi(\xi^p|z)}{\pi(\xi^{i-1}|z)} \times \frac{q(\xi^{i-1}|z)}{q(\xi^p|z)}\right\}.$$
3. Set $\xi^i = \xi^p$ if $R < \alpha\left(\xi^{i-1} \to \xi^p\right)$ where $R \sim Unif[0, 1]$. Otherwise, let $\xi^i = \xi^{i-1}$.

Thus, we have introduced a fully Bayesian procedure which satisfies the set of estimating equations in (2.4).

4. Asymptotic properties: Double robustness

In this section, we establish the large sample properties of the posterior distribution of the ATE parameter under model misspecification. All proofs are relegated to the Appendix. The Bayesian consistency is defined as the posterior distribution of θ being concentrated around the true data generating parameter θ_0. The following definition formulates the Bayesian consistency with the frequentist analogy (Walker and Hjort, 2001).

Definition 4.1 (*Consistency*)**.** As data $z_1, z_2, \ldots, z_n$ accumulate from some unknown underlying $f(z|\theta_0)$, then the posterior mass for the model assigning to a set A is given by

$$\Pi^n(A) = \pi\left(\theta \in A \left| z_1, \ldots, z_n\right.\right) = \frac{\int_A R_n(\theta)\, \pi_0(d\theta)}{\int R_n(\theta)\, \pi_0(d\theta)}$$

where $R_n(\theta) = \prod_{i=1}^{n} \frac{f(z_i|\theta)}{f(z_i|\theta_0)}$ and $\pi_0(\theta)$ is the prior density for θ. If $A_\epsilon = \{\theta : d(\theta, \theta_0) > \epsilon\}$ where $d(\theta, \theta_0)$ is some distance measure, Bayesian consistency is defined as

$$\Pi^n(A_\epsilon) \to 0 \qquad \text{almost surely.}$$

Throughout this section, we assume that the true data generating outcome mean model is

$$u(d, x; \theta_0, \beta_0) = \theta_0 d + h_0(x, \beta_0)$$

or equivalently with $\phi_0 = 0$

$$u(d, x; \theta_0, \beta_0, \phi_0) = \theta_0 d + h_0(x, \beta_0) + \phi_0 e\left(x; \hat{\gamma}\right),$$

and therefore $\xi_0 := (\theta_0, \beta_0, \phi_0)$ is the true parameter. In addition, we assume the true PS model as $e(x; \gamma_0)$. Since we use the estimated PS in the OR model, $\hat{\gamma} \to \gamma_0$ when the PS model is correctly specified, and it has the balancing property as demonstrated in Fig. 1. The log-empirical-likelihood function, $\log p\left(y_i \left| x_i, d_i, e\left(x_i; \hat{\gamma}\right), \xi\right.\right) := \ell_i(\xi)$, can be written as

$$\ell_i(\xi) = \log \frac{\exp\left(\hat{\lambda}(z, \xi)^\top \mathbf{U}_i(\xi)\right)}{\sum_{j=1}^{n} \exp\left(\hat{\lambda}(z, \xi)^\top \mathbf{U}_j(\xi)\right)}.$$

We start with the M-estimator $\hat{\xi}_0$ in frequentist inference, which is defined as the solution to the empirical estimating equation $\sum_{i=1}^{n} \mathbf{U}_i(\xi) = \mathbf{0}$. The M-estimator $\hat{\xi}_0$ maximizes the empirical likelihood in (3.1) (Owen, 2001). Based on previous work (Owen, 2001; Yiu et al., 2020), the empirical likelihood decays superpolynomially to zero outside of any neighborhood of $\hat{\xi}_0$, i.e., under certain regularity conditions, for any $\epsilon > 0$ there exists a $\delta > 0$,

$$\sup_{\left\|\xi - \hat{\xi}_0\right\| > \epsilon} \exp\left[\sum_{i=1}^{n}\left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right] \le \exp\left[-\delta(n-1)^{1/2}\right].$$

If $\hat{\xi}_0$ belongs to the support of the prior distribution, then the convergence rate of the Bayesian exponentially tilted empirical likelihood to the point mass $\hat{\xi}_0$ will be superpolynomially as well. In the next two sections, we attempt to construct the posterior consistency and asymptotic normality under two different scenarios.

4.1. Misspecified PS and correctly specified OR

When the PS model is misspecified, $\hat{\gamma} \not\rightarrow \gamma_0$ and $e\left(x; \hat{\gamma}\right)$ does not have the balancing property anymore; we can still achieve consistency if the part of $h_1(\cdot)$ in the OR is correctly specified as $h_0(\cdot)$. Thus, the OR is specified as

$$u\left(d, x, e\left(x; \hat{\gamma}\right); \theta, \beta, \phi\right) = \theta d + h_0(x, \beta) + \phi e\left(x; \hat{\gamma}\right)$$

regardless of the value of $e\left(x; \hat{\gamma}\right)$ because $\phi \rightarrow \phi_0$ as $n \rightarrow \infty$, and $\phi_0 = 0$ in this case. We use the following notations to show asymptotic results of the proposed method. We denote $\xi \in \Xi$ with Ξ product space for the parameter space of (θ, β, ϕ), which is assume to be compact and connected. In addition, we state the assumptions that are used for the theorems regarding the misspecified PS case.

Assumption 4.1. $\xi_0 = (\theta_0, \beta_0, \phi_0) \in \Xi$ is the unique solution to $\mathbb{E}\mathbf{U}(\xi) = \mathbf{0}$, and ξ_0 is in the interior of Ξ.

Assumption 4.2. Assume that the data generating parameter ξ_0 belongs to the support of the prior distribution, i.e. $\pi_0 (\xi \in U) > 0$, for every neighborhood U of ξ_0. π_0 is also a continuous probability measure with respect to the Lebesgue measure.

Assumption 4.3. Suppose the outcome variables, $Y_1, \ldots, Y_n$, take value in the space $\mathbb{Y} \subset \mathbb{R}$.

Assumption 4.4. $\mathbf{U}_i(\xi)$ is continuous $\forall \xi \in \Xi$ and $\forall i = 1, \ldots, n$.

Assumption 4.5. $\mathbb{E}\left[\sup_{\xi \in \Xi} \|\mathbf{U}(\xi)\|^{\alpha}\right] < \infty$ for some $\alpha > 2$.

Assumption 4.6. $\mathcal{I}$ non-singular and $\mathcal{J}$ is full rank, i.e., $\text{rank}(\mathcal{J}) = p$, with

$$\mathcal{I} = \mathbb{E}[\mathbf{U}(\xi_0)\mathbf{U}(\xi_0)^{\top}] \qquad \mathcal{J} = -\mathbb{E}[\dot{\mathbf{U}}(\xi_0)]$$

both $(p \times p)$ matrices, and $\dot{\mathbf{U}}(\xi_0) = \left.\dfrac{\partial \mathbf{U}(\xi)}{\partial \xi^{\top}}\right|_{\xi=\xi_0}$.

Assumption 4.7. Suppose Ξ^* is a neighborhood of ξ_0 where $\mathbf{U}(\xi)$ is continuously differentiable, and $\mathbb{E}[\sup_{\xi \in \Xi^*} \left\|\dot{\mathbf{U}}(\xi)\right\|_F] < \infty$, with $\|\cdot\|_F$ denoting as the Frobenius norm.

Alongside with Assumptions 4.1–4.7, the semiparametric theory for a $p \times 1$ system of estimating equations with $\sum_{i=1}^{n} \mathbf{U}_i(\hat{\xi}_0) = \mathbf{0}$ and $\mathbb{E}[\mathbf{U}(\xi_0)] = \mathbf{0}$, under some regularity conditions, we have that M-estimator $\hat{\xi}_0$ is consistent and asymptotically normally distributed (Van der Vaart, 2000), that is

$$\sqrt{n}(\hat{\xi}_0 - \xi_0) \xrightarrow{d} \mathit{Normal}_p(\mathbf{0}, \mathbf{V})$$

where $\mathbf{V} = \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-\top}$. We can construct the similar result for the proposed Bayesian empirical likelihood approach under Bayesian consistency.

Theorem 4.1 (*Consistency-misspecified PS*). *Under Assumptions 4.1–4.7 and* $U = \{\xi : \|\xi - \xi_0\|_2 < \epsilon\}$,

$$\Pi^n(U) = \frac{\int_U \exp\left[\sum_{i=1}^{n} \left(\ell_i(\xi) - \log f(y_i|d_i, x_i, \xi_0)\right)\right] \pi_0(d\xi)}{\int_{\Xi} \exp\left[\sum_{i=1}^{n} \left(\ell_i(\xi) - \log f(y_i|d_i, x_i, \xi_0)\right)\right] \pi_0(d\xi)} \rightarrow 1 \ \textit{almost surely}.$$

The frequentist counterpart of consistency also follows under a well-specified outcome model. Another aspect of the asymptotic property in frequentist inference is the limiting behavior of the estimator in terms of the probability law, i.e., the asymptotic normality. The Bayesian analogy is the Bernstein–von Mises theorem, which establishes the asymptotic normality of the posterior density, $\pi\left(\sqrt{n}(\xi - \hat{\xi}_0)\,|z_1, \ldots, z_n\right)$. To derive this result, We need some additional assumptions that coincide with some regularity conditions in the frequentist semiparametric theory.

Assumption 4.8. There exists a neighborhood $\mathcal{B}$ of ξ_0 on which, with probability approaching 1, $\ell_i(\xi)$ is non-zero, that is there exists a function $\hat{\lambda}_n$ of ξ which is the minimizer of $\frac{1}{n}\sum_{i=1}^{n} \exp\left(\lambda^{\top}\mathbf{U}_i(\xi)\right)$, for all $\xi \in \mathcal{B}$.

Assumption 4.9. For all values of z, $\mathbf{U}(\xi)$ is twice differentiable with respect to ξ in a neighborhood of ξ_0, and the second derivative satisfies the Lipschitz condition

$$\left\|\ddot{\mathbf{U}}(\xi) - \ddot{\mathbf{U}}(\xi')\right\|_o \leq K(z)\left\|\xi - \xi'\right\|_2$$

for an integrable function $K(\cdot)$, where $\ddot{\mathbf{U}}(\xi) = \dfrac{\partial \dot{\mathbf{U}}(\xi)}{\partial \xi^{\top}}$ and $\|\cdot\|_o$ represents the operator norm.

Assumption 4.10. For all values of z, there exists a neighborhood of $(0, \xi_0)$, in $\mathbb{R}^p \times \Xi$, in which the function $a(\lambda, \xi) = \exp\{\lambda^\top \mathbf{U}(\xi)\}\mathbf{U}(\xi)$ and all of its first and second partial derivatives are dominated by an integrable function.

With these additional assumptions, we have the following Bernstein–von Mises theorem under the misspecified PS case.

Theorem 4.2. *If $\int \|\xi\|_2\, \pi_0(\xi)d\xi < \infty$ and Assumptions 4.1–4.10 hold, the posterior mass for the model assigning to an arbitrary set $A \subseteq \Xi$ converges to a normal distribution, i.e.,*

$$\sup_A \left|\pi\left(\sqrt{n}(\xi - \xi_0) \in A\,|z_1, \ldots, z_n\right) - Normal_p(\mathbf{0}, \mathbf{V}) \in A\right| \xrightarrow{p} 0.$$

The proof of this theorem is shown in Appendix C, which is based on Chib et al. (2018), Ghosh and Ramamoorthi (2003) and Van der Vaart (2000). This theorem essentially shows the limiting posterior distribution of ξ concentrates on a $\sqrt{n}$-ball centered at the true value of ξ_0 with the same variance–covariance matrix as the M-estimator.

4.2. Misspecified OR and correctly specified PS

When the OR model is misspecified but the PS model is correctly specified, according to Fig. 1, $X \perp\!\!\!\perp D\,|e(x;\gamma_0)$ and $\hat{\gamma} \to \gamma_0$. Therefore, $e\left(x;\hat{\gamma}\right)$ is an asymptotic balancing score. Suppose we specified the OR as

$$u\left(d, x, e\left(x;\hat{\gamma}\right); \theta, \beta, \phi\right) = \theta d + h_1(x, \beta) + \phi e\left(x;\hat{\gamma}\right). \tag{4.1}$$

Therefore, assuming that there is no unmeasured confounding, we can find β^* and ϕ^* such that

$$u\left(d, x, e\left(x;\hat{\gamma}\right); \theta, \beta, \phi\right) - h_1(x, \beta^*) - \phi^* e\left(x;\hat{\gamma}\right) = \theta_0 d$$

that is, the dependence of the mean model on D is correctly specified, and the effect of D is captured via θD. Therefore, we have a pseudo-true value, $\xi^* = (\theta_0, \beta^*, \phi^*)$. The pseudo-true value of β, ϕ is determined by $(\beta^*, \phi^*) = \arg\min_{\beta,\phi} \mathcal{K}\left(f_0, f_1\right)$, where $\mathcal{K}\left(f_0, f_1\right)$ is the KL divergence between the true model f_0 and the misspecified model f_1. With the following assumptions on the identification of the pseudo-true value, we can construct the same asymptotic results as the misspecified PS case by replacing ξ_0 with ξ^*.

Assumption 4.11. Assume that the data generating parameter ξ^* belongs to the support of the prior distribution, i.e. $\pi_0\left(\xi^* \in U\right) > 0$, for every neighborhood U of ξ^*.

In our empirical likelihood analysis case, we approximate the misspecified model f_1 by minimizing the KL divergence between the probabilities $(p_1, \ldots, p_n)$ assigned to each sample observation and the empirical probabilities $(\frac{1}{n}, \ldots, \frac{1}{n})$. It might not be the solution to the dual problem

$$\max_{\xi\in\Xi} \min_{\lambda\in\mathbb{R}^p} \left[\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda^\top \mathbf{U}_i(\xi))\right].$$

Sueishi (2013) explained that this is because the probability measures in misspecified models might not have common support with the true model, f_0, and therefore gave a condition of the identification for the pseudo-true value and the validity of the dual theorem. In addition, Chib et al. (2018) stated that the dual representation is guaranteed to hold for (β^*, ϕ^*) defined as the KL minimizer, if there exists $Q \in P_{(\beta,\theta,\phi)} = \left\{Q : \mathbb{E}_Q\left[\mathbf{U}(\beta, \theta, \phi)\right] = 0\right\}$ such that Q is mutually absolutely continuous with respect to the true data generating model, f_0, where $\mathbb{E}_{f_0}\left[\mathbf{U}(\beta_0, \theta_0, \phi_0)\right] = 0$. Therefore, we have the following assumption, which implies that the set of all misspecified models is in a nonempty set. It also indicates the identification of the pseudo-true value as the KL minimizer, which replaces Assumption 4.1.

Assumption 4.12. For a fixed $\xi \in \Xi$, there exists $Q \in \{f_1 :$ all misspecified models in (2.3)$\}$ such that Q is mutually absolutely continuous with respect to f_0, where f_0 is defined as the true data-generating process as in (2.2).

Assumption 4.12 is required since the ATE in semiparametric linear regression is a single or a linear combination of parameter(s). Hence, the identification of the pseudo-true value implies that the true ATE can be recovered under model misspecification. With these additional assumptions, we have the following corollaries of the asymptotic results for the misspecified OR case.

Corollary 4.1 (*Consistency-misspecified OR*)**.** Under *Assumptions* 4.3–4.7 and 4.11–4.12 and $U = \{\theta : \|\xi - \xi^*\|_2 < \epsilon\}$,

$$\Pi^n\left(U\right) = \frac{\int_U \exp\left[\sum_{i=1}^{n}\left(\ell_i(\xi) - \log f(y_i|d_i, x_i, \xi^*)\right)\right]\pi_0\left(d\xi\right)}{\int_\Xi \exp\left[\sum_{i=1}^{n}\left(\ell_i(\xi) - \log f(y_i|d_i, x_i, \xi^*)\right)\right]\pi_0\left(d\xi\right)} \to 1 \text{ almost surely.}$$

Corollary 4.2. If $\int \|\xi\|_2\, \pi_0(\xi)d\xi < \infty$ and *Assumptions* 4.3–4.12 hold, the posterior mass for the model assigning to an arbitrary set $A \subseteq \Xi$ converges to a normal distribution, i.e.,

$$\sup_A \left|\pi\left(\sqrt{n}(\xi - \xi^*) \in A\,|z_1, \ldots, z_n\right) - Normal_p(\mathbf{0}, \mathbf{V}) \in A\right| \xrightarrow{p} 0.$$

Table 1
Example 1: Simulation results of the marginal causal contrast, with true value equal to 1, on 5000 simulation runs on generated datasets of size 1000. BEL-DR represents our Bayesian doubly robust approach via empirical likelihood, and BS-DR represents results from doubly robust inference via the Bayesian bootstrap approach in Saarela et al. (2016)'s simulation study.

Estimator	Scenario I			Scenario II		
	Mean	Variance	Coverage	Mean	Variance	Coverage
BEL-DR	1.00	0.005	94.7%	1.00	0.005	94.8%
BS-DR	1.00	0.007	94.1%	1.00	0.005	94.8%

5. Examples

The methodology proposed above can be applied in various causal inference problems. In this section, we illustrate some simulated examples, including verifying the double robustness property and the robustness property under some extreme PS distributions. Finally, we compare this proposed method with some recently developed flexible modeling approaches, such as Bayesian causal forests and the double machine learning method.

5.1. Example 1: Double robustness

We consider the simulation study by Saarela et al. (2016) to compare with their doubly robust inference via the Bayesian bootstrap approach. The data are simulated as the following hierarchy:

$$\begin{aligned} X_1, X_2, X_3, X_4 &\sim \mathcal{N}(0,1) \\ U_1 &= \frac{|X_1|}{\sqrt{1-2/\pi}} \\ D\,|U_1, X_2, X_3 &\sim \text{ Bernoulli}\left(\text{expit}\left(0.4U_1 + 0.4X_2 + 0.8X_3\right)\right) \\ Y\,|D, U_1, X_2, X_4 &\sim \mathcal{N}\left(D - U_1 - X_2 - X_4, 1\right). \end{aligned}$$

In this case, two scenarios are considered:

- Scenario I: Misspecify the outcome model using covariates (x_1, x_2, x_4) and correctly specify the treatment assignment model using covariates (u_1, x_2, x_3).
- Scenario II: Correctly specify the outcome model using covariates (u_1, x_2, x_4) and misspecify the treatment assignment model using covariates (x_1, x_2, x_3).

In both scenarios, normal priors with mean 0 and standard deviation 1000 are placed on all the parameters. Table 1 shows the results of 5000 Monte Carlo replicates of the averages of the posterior means, variances and coverage rates for θ based on 5000 MCMC samples with 1000 burn-in iterations in each replicate. The coverage rates are computed by constructing, for each data replicate, a 95% credible interval for θ from the posterior sample, with the 2.5% and 97.5% posterior sample quantiles. The results indicate that our proposed method is doubly robust with unbiased estimation and correct coverage at the nominal level in both scenarios, and has a similar empirical variance compared with the Bayesian bootstrap approach used in Saarela et al. (2016). We notice that when the PS is correctly specified, the variance in the Bayesian bootstrap is slightly higher compared to the case with a misspecified PS model. These results indicate that the posterior means in both methods center at the true ATE, and it verifies the theoretical result of double robustness for the proposed approach, showing that both approaches yield the similar posterior samples.

5.2. Example 2: Propensity score distribution

In this example, we examine the performance of our method under some extreme PS distributions. It corresponds to Stephens et al. (2022)'s simulation study with binary exposure, where there is no treatment effect. The data are simulated as the following hierarchy:

$$\begin{aligned} X_1, X_2 &\sim \mathcal{N}(1,1) \\ X_3, X_4 &\sim \mathcal{N}(-1,1) \\ D\,|X_1, X_2, X_3, X_4 &\sim \text{ Bernoulli}\left(\text{expit}\left(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4\right)\right) \\ Y\,|D, X_1, X_2, X_3, X_4 &\sim \mathcal{N}\left(0.25X_1 + 0.25X_2 + 0.25X_3 + 0.25X_4 + 1.5X_3X_4, 1\right). \end{aligned}$$

In the analyses, we investigate how the PS distribution affects the estimation of the treatment effect. Therefore, different PS distributions are considered using three scenarios with different values of γ:

- Scenario I: $\gamma = (0.00, 0.30, 0.80, 0.30, 0.80)$, generating a nearly uniform distribution of PSs.
- Scenario II: $\gamma = (0.50, 0.50, 0.75, 1.00, 1.00)$, having a greater density of lower scores.
- Scenario III: $\gamma = (0.00, 0.45, 0.90, 1.35, 1.80)$, having very few high scores.

Table 2
Example 2: Simulation results of the marginal causal contrast, with true value equal to 0, on 1000 simulation runs on generated datasets of size n. BEL-WDR used the estimating equation in (2.1), and BB represents results from the Bayesian two-step approach via the Bayesian bootstrap in Stephens et al. (2022)'s simulation study. Bayes-OR represents standard Bayesian inference for the correctly specified OR with non-informative priors.

n	Scenario I				Scenario II				Scenario III			
	100	200	500	1000	100	200	500	1000	100	200	500	1000
Mean												
BEL-DR	0.023	0.008	0.010	−0.002	0.025	−0.004	0.000	0.003	0.031	0.006	0.013	0.002
BEL-WDR	−0.032	0.001	0.010	0.004	−0.058	−0.052	0.004	−0.001	−0.375	−0.311	−0.215	−0.186
BB	−0.011	0.006	−0.003	−0.007	−0.029	−0.015	−0.020	−0.004	−0.018	−0.025	−0.013	−0.012
Bayes-OR	0.000	−0.001	0.002	−0.004	0.008	0.001	−0.004	−0.007	−0.005	0.002	0.000	0.002
Variance												
BEL-DR	0.171	0.080	0.030	0.015	0.188	0.087	0.033	0.016	0.278	0.128	0.046	0.023
BEL-WDR	0.301	0.157	0.069	0.032	0.375	0.244	0.143	0.083	0.676	0.633	0.542	0.371
BB	0.155	0.071	0.030	0.015	0.163	0.080	0.032	0.016	0.233	0.112	0.043	0.023
Bayes-OR	0.055	0.026	0.010	0.006	0.066	0.031	0.012	0.006	0.087	0.040	0.017	0.008

In this example, the PS model is correctly specified while we fit the OR by only including the treatment indicator and the PS as a covariate. We place zero-mean normal priors with the standard deviation 1000 for the parameters in the OR model. Table 2 summarizes the estimates of θ over the 1000 Monte Carlo replicates for the scenarios described above. We also investigate the fully Bayesian model via a correctly specified outcome probability model. In each replicate, we generate 5000 MCMC samples with 1000 burn-in iterations. We implement the proposed Bayesian approach via semiparametric linear regression and the weighting method in (2.1). The results suggest that the shape of the true distribution of the PS affects the magnitude of the bias. In general, the correct specified Bayesian OR approach yields smallest variances compared to other approaches as there is no model misspecification. For our proposed method, both the variance and bias increase in Scenarios II and III when there are more extreme PSs, but they decrease as the sample size increases. The weighting method performs well in Scenario I when the PS distribution is uniform. When the PS distribution becomes more extreme, the biases and empirical variances increase significantly, compared to other regression approaches. This indicates the regression method is less sensitive to the extreme PS. We also notice that, in all cases, our method has similar variances compared to the two-step approach based on the Bayesian bootstrap. According to Luo et al. (2021), the Bayesian bootstrap and its extension have asymptotic sandwich variance, which is confirmed by these numerical results. However, the proposed approach employs the prior-to-posterior framework, and more straightforward to incorporate informative prior whenever the information is available.

5.3. Example 3: Comparison with frequentist regression estimation

In this example, we consider a simulation study with more misspecified forms, and aim to compare with the existing frequentist DR regression-based approaches. This simulation scenario is constructed under the observational setting described in and follows the design of Kang and Schafer (2007), where the data are simulated as the following hierarchy:

$$\begin{aligned} X_1, X_2, X_3, X_4 &\sim \mathcal{N}(0, 1) \\ D\,|X_1, X_2, X_3, X_4 &\sim \text{ Bernoulli}\left(\text{expit}\left(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4\right)\right) \\ Y\,|D, X_1, X_2, X_3, X_4 &\sim \mathcal{N}\left(210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4, 1\right). \end{aligned}$$

In addition to the correctly specified models, the covariates actually observed are assumed to be $Z_1 = \exp(X_1/2), Z_2 = X_2/(1 + \exp(X_1)) + 10, Z_3 = (X_1X_3/25 + 0.6)^3$ and $Z_4 = (Z_2 + Z_4 + 20)^2$. Under misspecification, the covariates X_i are replaced with Z_i. In this case, four scenarios are considered:

- Scenario I: Misspecify the outcome model using covariates (z_1, z_2, z_3, z_4) and correctly specify the treatment assignment model using covariates (x_1, x_2, x_3, x_4).
- Scenario II: Correctly specify the outcome model using covariates (x_1, x_2, x_3, x_4) and misspecify the treatment assignment model using covariates (z_1, z_2, z_3, z_4).
- Scenario III: Correctly specify both the outcome model and treatment assignment model using covariates (x_1, x_2, x_3, x_4).
- Scenario IV: Misspecify both the outcome model and treatment assignment model using covariates (z_1, z_2, z_3, z_4).

We compare the performance with regression-based frequentist DR methods discussed in Kang and Schafer (2007). Specifically, we consider frequentist regression estimators based on inverse-propensity weighted least squares (WLS) and propensity-covariate (as a single covariate) regression (π-cov). Fig. 2 shows the boxplots of this study under different estimation procedures across $n = 20, 200, 1000$. As discussed in Yiu et al. (2020) and Luo et al. (2021), an informative prior yields more stabilized results with higher shrinkage towards the prior mean when the sample size is small. Therefore,

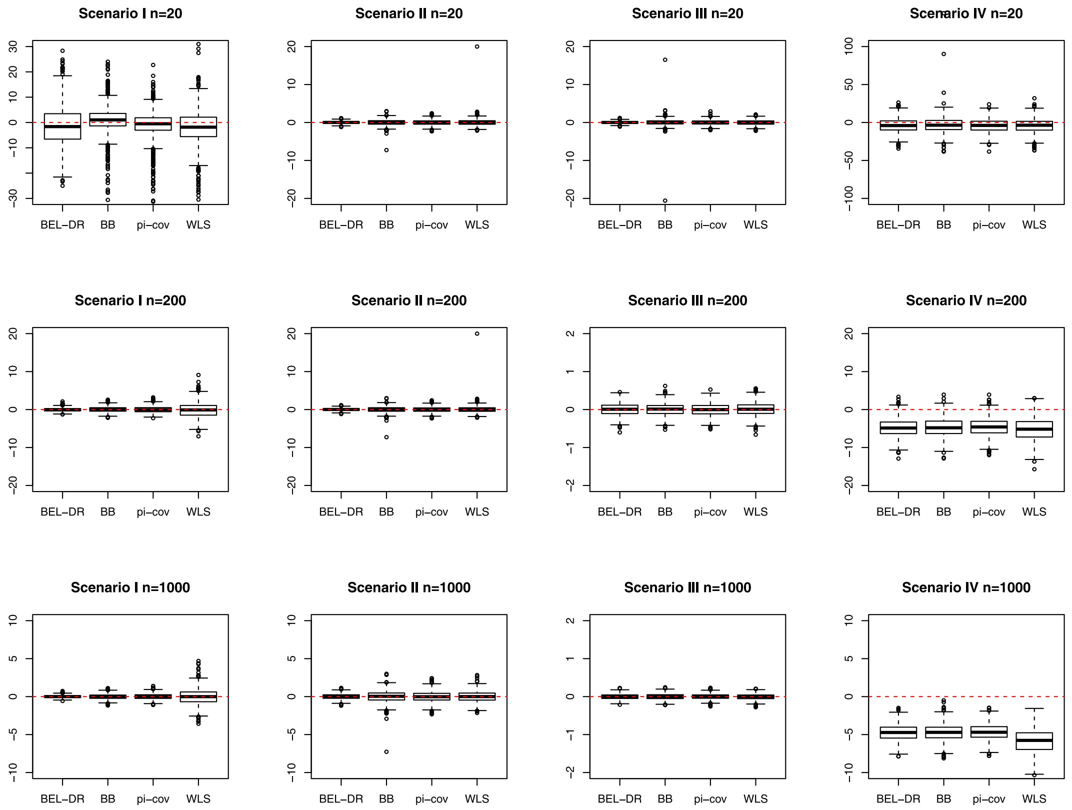


Fig. 2. Example 3: Simulation results of the marginal causal contrast, with true value equal to 0, on 1000 simulation runs on generated datasets of size n using the set up in Kang and Schafer (2007). pi-cov represents the least squared estimation for propensity-covariate regression. WLS represents inverse-propensity weighted estimation for the OR.

we place an informative normal prior for θ with mean zero the and standard deviation 2 when $n = 20$. For $n = 200, 1000$, we place zero-mean normal priors with the standard deviation 1000 for the parameters in the OR model. When $n = 20$, BEL-DR yield more stable results with fewer extreme values compared to other approaches. BB and WLS have more extreme estimates due to the sensitivity to extreme weights with small sample sizes. We notice that when the OR model is correctly specified, BEL-DR has the smallest empirical variance compared to other methods. In all other scenarios when $n = 200, 1000$, the WLS approach yields higher biases and significantly large empirical variances as expected. In Scenarios I and II when the either of PS or OR is misspecified, our proposed method has a smaller variance compared to other approaches. While in Scenarios III and IV, all three methods yield similar results in terms of empirical mean and variance estimates, except the WLS method having a greater empirical variance. This example demonstrates that the Bayesian prior-to-posterior update yields a more stable results when using the informative prior, compared to the frequentist methods in small sample cases.

5.4. Example 4: Comparison with flexible modeling

In this example, we seek to compare the proposed approach with existing flexible/machine learning causal estimation approaches. We first consider the following data generating mechanism:

$$\begin{aligned} X_1, X_2 &\sim \mathcal{N}(1,1), X_3, X_4 \sim \mathcal{N}(-1,1) \\ D\,|X_1, X_2, X_3, X_4 &\sim \text{ Bernoulli}\left(\operatorname{expit}\left(0.3X_1 + 0.9X_2 - 1.25X_3 + 1.5X_4\right)\right) \\ Y|D, X_1, X_2, X_3, X_4 &\sim \mathcal{N}(\mu_0(D, X; \beta), 1) \\ \mu_0(D, X) &= D + 2DX_1 + X_1 + X_2 + X_3 + X_4 + 0.25X_1^2 + 0.75X_2X_4 + 0.75X_3X_4. \end{aligned}$$

The ATE in this case is $\mathbb{E}[\mu_0(1, X)] - \mathbb{E}[\mu_0(0, X)] = 1 + 2 \times \mathbb{E}[X_1] = 3$. Since there are interaction terms in the OR, we specify the mean of the treatment-effect model as

$$\beta + (\theta + \theta_1 x_1)d + \phi_1 e\left(x; \hat{\gamma}\right) + \phi_2 x_1 e\left(x; \hat{\gamma}\right).$$

Table 3
Example 4: Comparison of results for the proposed Bayesian empirical likelihood, Bayesian causal forests (BCFs) and frequentist double machine learning estimator (FDML). Summary of 1000 simulation runs. Rows correspond to the bias, root mean square error (RMSE), and coverage rates.

	Method	n		
		200	1000	2000
Bias	BEL-DR	−0.007	−0.006	0.001
	BCF	0.021	−0.004	0.001
	FDML-Tree	−0.348	−0.366	−0.255
	FDML-Forest	0.252	−0.060	0.010
	FDML-Boosting	−0.147	−0.116	−0.102
	FDML-Nnet	−0.090	−0.089	−0.088
	FDML-Ensemble	−0.022	0.018	−0.067
	FDML-Best	−0.111	0.020	−0.089
RMSE	BEL-DR	0.339	0.143	0.101
	BCF	0.293	0.115	0.080
	FDML-Tree	0.448	0.390	0.272
	FDML-Forest	0.401	0.153	0.100
	FDML-Boosting	0.343	0.179	0.139
	FDML-Nnet	0.341	0.168	0.134
	FDML-Ensemble	0.317	0.158	0.116
	FDML-Best	0.346	0.169	0.135
Coverage rate	BEL-DR	89.9	93.3	94.5
	BCF	92.3	90.8	89.5
	FDML-Tree	91.0	42.3	45.4
	FDML-Forest	93.2	94.0	95.4
	FDML-Boosting	93.6	90.8	85.2
	FDML-Nnet	95.1	92.0	86.4
	FDML-Ensemble	96.0	92.8	90.7
	FDML-Best	94.4	92.0	86.0

As discussed in Appendix A, this model yields a consistent estimate for the ATE, and is fitted through the proposed Bayesian exponentially tilted empirical likelihood approach. We also consider the Bayesian causal forests (BCFs) method in Hahn et al. (2020). The BCF is a flexible approach for the OR using the Bayesian additive regression trees (BARTs) to infer the individual treatment effects, and it is based on linear predictor

$$\mu(d, x) = h(x, e\left(x; \hat{\gamma}\right)) + t\left(x, e\left(x; \hat{\gamma}\right)\right) d$$

with assumed normal errors. The functions $h(\cdot, \cdot)$ and $t(\cdot, \cdot)$ are estimated via the BCFs. In this analysis, we assume the PS model is correctly specified and estimated via a parametric logistic regression in BCFs and the proposed Bayesian approach. Finally, we consider a frequentist double machine learning (FDML) approach proposed in Chernozhukov et al. (2018). In their method, the ATE estimator, θ, is the solution to $\mathbb{E}[\psi(Z; \theta, \mu, e(X))] = 0$, where

$$\psi(Z; \theta, \mu, e(X)) = \mu(1, X) - \mu(0, X) + \frac{D(Y - \mu(1, X))}{e(X)} - \frac{(1 - D)(Y - \mu(0, X))}{1 - e(X)} - \theta$$

and $\mu(\cdot, \cdot)$ is the treatment-effect model and $e(\cdot)$ is the PS. Both of them are estimated via various machine learning approaches. Specifically, we use the FDML estimator in Definition 3.2 in Chernozhukov et al. (2018), for which the data are partitioned into K groups. The functions $\hat{\mu}_k(\cdot, \cdot)$ and $\hat{e}_k(\cdot)$ are estimated using the all the data excluding the kth group. Then the FDML estimator for ATE the solution to $1/K \sum_{k=1}^{K} \mathbb{E}_k[\psi(Z; \theta, \hat{\mu}_k, \hat{e}_k(X))] = 0$, where $\mathbb{E}_k(\cdot)$ is the empirical expectation over the kth fold of the data.

The results of this analysis are presented in Table 3. In the FDML, we used the methods described in Chernozhukov et al. (2018), i.e., regression tree (CART), random forest, boosting (tree-based), and neural network (two neuros) to estimate the $\mu(\cdot, \cdot)$ and $e(\cdot)$. There are two hybrid methods. 'Ensemble' represents the optimal combination of boosting and random forest and neural network, while 'Best' represents the best methods for estimating each of $\mu(\cdot, \cdot)$ and $e(\cdot)$ based on the average out-of-sample prediction for the ATE associated with each of $\mu(\cdot, \cdot)$ and $e(\cdot)$ estimates obtained from the previous machine learning approaches. The BCFs display certain bias in the small size, but have relatively a small variance, and therefore a lower RMSE compared to the BEL-DR approach. All FDML results show quite significant biases, especially using the regression tree approach. However, the variances are relatively smaller than other approaches, and therefore the RMSEs are not far off from the other two methods. In terms of the coverage rate, the coverage of BCFs and FDML is decreasing as n increases and ultimately is below the nominal level, except the FDML-Forest. FDML-Tree results in a large bias, and its coverage rate is eventually below 50%, while the proposed approach yields the coverage rate at the nominal level in all cases. It should be noted that the comparison is not completely fair as the flexible approaches such as the BCFs and FDML do not assume a known functional form for the OR model, which is robust to mis-specification of the treatment effect model. Finally, it should also be stressed that the BCFs and FDML methods require on average more computational expenditure compared to the proposed BEL-DR method.

Table 4
Summary statistics for the posterior predictive distribution of the percentage change of the average treatment effect for the UK speed camera data. ABDR represents results from the Bayesian bootstrap doubly robust approach from Graham et al. (2019). Gibbs-Post represents results from the Bayesian approach via loss functions from Luo et al. (2021).

	Posterior mean	SD	95% Credible interval
BEL-DR	−18.375	2.377	(−23.102, −13.710)
Gibbs-Post (B-spline)	−16.412	1.851	(−20.009, −12.780)
Gibbs-Post (GAM)	−16.213	1.847	(−19.779, −12.569)
ABDR	−14.359	3.605	(−21.841, −7.352)

6. Application: UK speed camera data

In this section, we follow the real data example used in Graham et al. (2019) and Luo et al. (2021) to assess the causal effect of the installation of speed cameras on the number of car related personal injury collisions. The data were collected from eight English administrative districts, including Cheshire, Dorset, Greater Manchester, Lancashire, Leicester, Merseyside, Sussex and the West Midlands, on the location of fixed speed cameras for 771 camera sites. The control group is a random sample of 4787 points on the network within the eight administrative districts. The outcome of interest is the number of personal injury collisions per kilometer as recorded from the location with or without speed cameras. The confounders include variables, such as the site length, the number of fatal and serious collisions (FSCs) and the number of personal injury collisions (PICs), the annual average daily flow (AADF), road types, speed limit, and the number of minor junctions within site length.

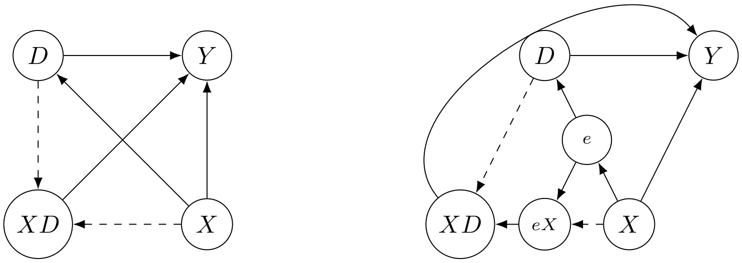
Graham et al. (2019) estimated the PS with a generalized additive model (GAM) by including smooth functions on the AADF, the number of minor junctions, and we adopt this model to estimate the PS. For the OR model, we include all the confounders and the estimated PS. We place zero-mean normal priors with standard deviation 1000 on the parameters. Table 4 shows the summary statistics of the posterior predictive distribution of the percentage change of the average treatment effect based on 50,000 MCMC samples with 10,000 burn-in iterations. Our model indicates that there is around an 18% reduction in road traffic collisions due to the presence of the speed camera, indicating a stronger causal relationship between the installation of the speed camera and the number of road traffic collisions. Compared to the approach Bayesian estimation approach (Graham et al., 2019), we obtained narrower 95% credible intervals. However, we obtained a slightly greater standard deviation and a stronger causal effect, compared to the Gibbs posterior approach (see Luo et al., 2021). The majority part of the 95% credible intervals overlapped among all methods, indicating an agreement in estimation of the percentage change of the ATE. Therefore, there is robust evidence of the reduction of road traffic collisions associated with the presence of speed cameras.

7. Discussion

In this article, we developed a fully semiparametric Bayesian DR causal estimation procedure via the non-parametric likelihood. This methodology was constructed by incorporating empirical likelihood for the data distribution and reweighting the sample so that it can also satisfy the moment condition. Specifically, we employed the exponentially tilted empirical likelihood strategy (Schennach, 2005, 2007), which connects the non-parametric Bayesian method and exponential tilting empirical likelihood estimator. We have also shown in this paper that the posterior distribution of the causal parameter is consistent and asymptotically normal, when either the OR or PS model is correctly specified. Simulation studies and the real data example demonstrated that our method is doubly robust, robust to the PS distribution, and also more stable when the sample size is small compared to frequentist approaches.

Our approach is appealing as it offers a prior-to-posterior update for causal inference which connects with frequentist DR inference based on estimating equations. This has been an obstacle to the widespread update of Bayesian methods in causal inference as a full probabilistic model is required in the usual Bayesian setting. To our knowledge, in current causal inference research, only the Bayesian bootstrap and its extension can adopt the estimating equation framework, and therefore our method provides an attractive alternative. This is because, unlike the Bayesian bootstrap method which requires resampling to include the prior distribution, our proposed method is ready to implement via MCMC. Therefore, it offers a means of informed and coherent decision making in the presence of uncertainty. In our analyses, we considered noninformative priors with a relatively large sample, which has yielded a smaller empirical variance compared to frequentist approaches when either of the PS or OR is misspecified. For smaller sample sizes, we suggest the usage of an informative prior so that the ATE is more stable and less sensitive to extreme PSs.

In our illustrative semiparametric regression model, we focus on the ATE as a single parameter, which can be extended to include further interaction terms (see Appendix A). The principles presented in this paper can be applied in much more general settings. For example, other flexible outcome regression models, such as BARTs, can be incorporated within the proposed model. Moreover, there are other versions of frequentist DR estimators (Kang and Schafer, 2007; Cao et al., 2009), which can also be implemented in this framework. The proposed methodology can be also widely applied in other causal settings when the traditional Bayesian set-up requires overspecifying the model, clashing with moment



Original DAG with interactions Including PS, e, to block the open paths

Fig. A.3. DAGs with the true data generating mechanism, including the interactions.

restrictions. For example, in finding optimal dynamic treatment regimes, decision making based on previous treatment covariate history and baseline covariates is required, and the estimation procedure is often based on semiparametric methods (Murphy, 2003; Moodie et al., 2007). Also in longitudinal settings, G-estimation is commonly used to study the causal effects of time-varying exposures from a structural nested mean model (Robins and Hernán, 2009), which involves solving a set of estimating equations to estimate parameters of interest. It will be interesting further research to link the proposed Bayesian approach with those directions, which will offer a different perspective of uncertainty quantification and prediction in causal inference.

Acknowledgment

The authors thank Professor David A. Stephens who shared the early draft of their paper entitled "Causal inference under mis-specification: adjustment based on the propensity score". This research has been supported by the Lloyd's Register Foundation and Alan Turing Institute, United Kingdom through the Data Centric Engineering programme. We also thank Prof Stephens for pointing us to the formulations in Fig. 1 and Fig. A.3 in the Appendix, and the equivalence between propensity score regression and G-estimation discussed in Section 2.2.

Appendix A. Semiparametric linear regression with the interactions

Suppose the true OR model as demonstrated in Fig. A.3 is

$$\mathbb{E}(Y|D,X) = \theta_0 D + \theta_{0D} XD + h_0(X,\beta_0).$$

In this case, the ATE is $\mathbb{E}(Y|D=1,X) - \mathbb{E}(Y|D=0,X)$. The OR model must then contain terms that block the open backdoor paths that pass through the interaction, i.e., specifying the expectation of the treatment-effect model

$$\mathbb{E}(Y|D,X) = D(\theta + \theta_d X) + h_1(X,\beta) + \phi_1 e\left(X;\hat{\gamma}\right) + \phi_2 X e\left(X;\hat{\gamma}\right).$$

This model can achieve double robustness. As argued in the main paper, if the $h_1(x,\beta)$ component of the OR model is correctly specified, reflecting the data-generating mechanism as $h_0(x,\beta_0)$, then the estimators of θ and θ_d will be consistent regardless of whether the PS model was correctly specified as $\phi_1 = \phi_2 = 0$. From the causal diagram in Fig. A.3, conditioning on $e(X)$ and $e(X)X$, it blocks the open backdoor paths between X to D and X to XD, which achieves the orthogonality of X and D, leading the consistent estimation of the ATE.

Appendix B. Proof of Theorem 4.1

Proof. From Theorem 1 in Schennach (2005), under Assumption 4.1, the posterior distribution obtained under the empirical likelihood approach in (3.2) is the same as under the true data generating mechanism, $f(y|d,x,\xi_0)$, which satisfies $\int \mathbf{U}(\xi_0) f(y|d,x,\xi_0)dy = \mathbf{0}$. Moreover, by consistency of $\hat{\xi}_0$, it is equivalent to show

$$\Pi_E^n(U) = \frac{\int_U \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right]\pi_0(d\xi)}{\int_\Xi \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right]\pi_0(d\xi)} \to 1 \text{ almost surely.}$$

From Theorem 1 in Yiu et al. (2020), under Assumptions 4.1–4.7, then for any $\epsilon > 0$ there exists a $\delta > 0$, we have

$$\sup_{\|\xi-\xi_0\|>\epsilon} \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right] \le \exp\left[-\delta(n-1)^{1/2}\right].$$

Therefore, as $\pi_0(\xi \in U) > 0$, for every neighborhood U of ξ_0,

$$\Pi_E^n\left(U^c\right) \le \frac{\exp\left[-\delta(n-1)^{1/2}\right]}{\exp\left[-\delta(n-1)^{1/2}\right] + \int_U \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right]\pi_0(d\xi)}.$$

We need to show that with probability approaching to 1,

$$\frac{\int_U \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right]\pi_0(d\xi)}{\exp\left[-\delta(n-1)^{1/2}\right]} \to \infty \quad \text{as} \quad n \to \infty$$

which implies $\int_U \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right]\pi_0(d\xi) \to C$ as $n \to \infty$, where C is some constant. As the M-estimator, $\hat{\xi}_0$, maximizes the likelihood in (3.1), $\exp\left[\sum_{i=1}^n \ell_i(\hat{\xi}_0)\right] = 1$. Then, by the consistency of $\hat{\xi}_0$, $\hat{\xi}_0$ will also lie inside of the ball $U = \{\xi : \|\xi - \xi_0\|_2 < \epsilon\}$,

$$\begin{aligned}\int_U \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right]\pi_0(d\xi) &= \int_U \exp\left(\sum_{i=1}^n \ell_i(\xi)\right)\pi_0(d\xi)\\ &\le \int_U \exp\left(\sum_{i=1}^n \ell_i(\hat{\xi}_0)\right)\pi_0(d\xi) = 1.\end{aligned}$$

Therefore, $\int_U \exp\left[\sum_{i=1}^n \left(\ell_i(\xi) - \ell_i(\hat{\xi}_0)\right)\right]\pi_0(d\xi)$ is bounded between 0 and 1. This implies that $\Pi_E^n(U) \to 1$ almost surely. □

Appendix C. Proof of Theorem 4.2

Proof. As $\int \|\xi\|_2\, \pi_0(\xi)d\xi < \infty$ and Assumptions 4.1–4.10 hold, from Theorem 3 in Yiu et al. (2020), then we have

$$n^{1/2}(\hat{\xi}_0 - \xi^*) \to 0 \quad \text{and} \quad n^{1/2}(\xi^* - \xi_0) \to \mathcal{N}(0, \mathbf{V}).$$

Therefore, it is equivalent to show

$$\sup_A \left|\pi\left(\sqrt{n}(\xi - \hat{\xi}_0) \in A\,|z_1, \dots, z_n\right) - Normal_p(\mathbf{0}, \mathbf{V}) \in A\right| \xrightarrow{p} 0.$$

Denote $w = \sqrt{n}(\xi - \hat{\xi}_0)$ and the posterior distribution

$$\pi_n(w) := \pi\left(\sqrt{n}(\xi - \hat{\xi}_0)\,|z_1, \dots, z_n\right).$$

Then by the change of variable

$$\pi_n(w) = C_n^{-1}\pi_0\left(\hat{\xi}_0 + \frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^n\left(\ell_i\left(\hat{\xi}_0 + \frac{w}{\sqrt{n}}\right) - \ell_i\left(\hat{\xi}_0\right)\right)\right]$$

where $C_n = \int \pi_0\left(\hat{\xi}_0 + \frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^n\left(\ell_i\left(\hat{\xi}_0 + \frac{w}{\sqrt{n}}\right) - \ell_i\left(\hat{\xi}_0\right)\right)\right]dw$. Then in order to show that

$$\sup_A \left|\pi_n(w \in A) - Normal_p(\mathbf{0}, \mathbf{V}) \in A\right| \xrightarrow{p} 0.$$

It is equivalent to show

$$\int_A \left|C_n^{-1}\pi_0\left(\hat{\xi}_0 + \frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^n\left(\ell_i\left(\hat{\xi}_0 + \frac{w}{\sqrt{n}}\right) - \ell_i\left(\hat{\xi}_0\right)\right)\right] - (2\pi)^{-\frac{p}{2}}\,|\mathbf{V}|^{-\frac{1}{2}}\,\mathrm{e}^{-\frac{1}{2}(w^{\top}\mathbf{V}^{-1}w)}\right|dw \xrightarrow{p} 0. \tag{C.1}$$

Note that (C.1) can be written as

$$\begin{aligned}
&\int_A \left| C_n^{-1}\pi_0\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^{n}\left(\ell_i\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)-\ell_i\left(\hat{\xi}_0\right)\right)\right]-C_n^{-1}\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right.\\
&\left.+\,C_n^{-1}\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}-(2\pi)^{-\frac{p}{2}}\left|\mathbf{V}\right|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right|dw\\
&\le C_n^{-1}\int_A\left|\pi_0\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^{n}\left(\ell_i\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)-\ell_i\left(\hat{\xi}_0\right)\right)\right]-\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right|dw\\
&+\int_A\left|C_n^{-1}\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}-(2\pi)^{-\frac{p}{2}}\left|\mathbf{V}\right|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right|dw.
\end{aligned} \tag{C.2}$$

Therefore, it is equivalent to show

$$\int_A\left|\pi_0\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^{n}\left(\ell_i\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)-\ell_i\left(\hat{\xi}_0\right)\right)\right]-\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right|dw\xrightarrow{p}0$$

as this also implies that

$$C_n=\int\pi_0\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^{n}\left(\ell_i\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)-\ell_i\left(\hat{\xi}_0\right)\right)\right]dw\xrightarrow{p}\pi_0\left(\xi_0\right)(2\pi)^{\frac{p}{2}}\left|\mathbf{V}\right|^{\frac{1}{2}}.$$

Then the second term in (C.2) becomes

$$\int_A\left|C_n^{-1}\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}-(2\pi)^{-\frac{p}{2}}\left|\mathbf{V}\right|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right|dw\xrightarrow{p}0. \tag{C.3}$$

For the first term in (C.2), by Lemma C.1 in Chib et al. (2018) and Lemma 2 in Yiu et al. (2020) with Assumptions 4.1–4.10, it can be shown that

$$\int_A\left|\pi_0\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^{n}\left(\ell_i\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)-\ell_i\left(\hat{\xi}_0\right)\right)\right]-\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right|dw\xrightarrow{p}0$$

by breaking the domain of integration A into different regions, i.e., $A_1=\left\{w:\|w\|<c\log\sqrt{n}\right\}$, $A_2=\left\{w:c\log\sqrt{n}\le\|w\|\le\delta\sqrt{n}\right\}$ and $A_3=\left\{w:\|w\|>\delta\sqrt{n}\right\}$ for any given c and δ. Therefore,

$$C_n^{-1}\int_A\left|\pi_0\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)\exp\left[\sum_{i=1}^{n}\left(\ell_i\left(\hat{\xi}_0+\frac{w}{\sqrt{n}}\right)-\ell_i\left(\hat{\xi}_0\right)\right)\right]-\pi_0\left(\xi_0\right)\mathrm{e}^{-\frac{1}{2}\left(w^{\mathsf{T}}\mathbf{V}^{-1}w\right)}\right|dw\xrightarrow{p}0. \tag{C.4}$$

Combining the results in (C.4) and (C.3), we can obtain the result of the theorem. i.e.,

$$\sup_A\left|\pi_n\left(w\in A\right)-Normal_p(\mathbf{0},\mathbf{V})\in A\right|\xrightarrow{p}0.\quad\square$$

References

Abadie, A., Imbens, G.W., 2016. Matching on the estimated propensity score. Econometrica 84, 781–807.

Bang, H., Robins, J.M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–973.

Bornn, L., Shephard, N., Solgi, R., 2019. Moment conditions and Bayesian non-parametrics. J. R. Stat. Soc. Ser. B Stat. Methodol. 81, 5–43.

Cao, W., Tsiatis, A.A., Davidian, M., 2009. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96, 723–734.

Chamberlain, G., Imbens, G.W., 2003. Nonparametric applications of Bayesian inference. J. Bus. Econom. Statist. 21, 12–18.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21, C1–C68.

Chib, S., Shin, M., Simoni, A., 2018. Bayesian estimation and comparison of moment condition models. J. Amer. Statist. Assoc. 113, 1656–1668.

Ghosh, J.K., Ramamoorthi, R., 2003. Bayesian Nonparametrics. Springer Science & Business Media.

Graham, D.J., McCoy, E.J., Stephens, D.A., 2016. Approximate Bayesian inference for doubly robust estimation. Bayesian Anal. 11, 47–69.

Graham, D.J., Naik, C., McCoy, E.J., Li, H., 2019. Do speed cameras reduce road traffic collisions? PLoS One 14, e0221267.

Gustafson, P., 2012. Double-robust estimators: slightly more Bayesian than meets the eye. Int. J. Biostat. 8, 1–15.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315–331.

Hahn, P.R., Murray, J.S., Carvalho, C.M., 2020. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal. 15, 965–1056.

Henmi, M., Eguchi, S., 2004. A paradox concerning nuisance parameters and projected estimating functions. Biometrika 91, 929–941.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.

Imbens, G.W., 2000. The role of the propensity score in estimating dose–response functions. Biometrika 87, 706–710.

Kang, J.D., Schafer, J.L., 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22, 523–539.

Kaplan, D., Chen, J., 2012. A two-step Bayesian approach for propensity score analysis: Simulations and case study. Psychometrika 77, 581–609.

Lazar, N.A., 2003. Bayesian empirical likelihood. Biometrika 90, 319–326.

Little, R., An, H., 2004. Robust likelihood-based analysis of multivariate data with missing values. Statist. Sinica 14, 949–968.

Luo, Y., Stephens, D.A., Graham, D.J., McCoy, E.J., 2021. Bayesian doubly robust causal inference via loss functions. arXiv preprint, arXiv:2103.04086.

McCandless, L.C., Douglas, I.J., Evans, S.J., Smeeth, L., 2010. Cutting feedback in Bayesian regression adjustment for the propensity score. Int. J. Biostat. 6.

McCandless, L.C., Gustafson, P., Austin, P.C., 2009. Bayesian propensity score analysis for observational data. Stat. Med. 28, 94–112.

Moodie, E.E., Richardson, T.S., Stephens, D.A., 2007. Demystifying optimal dynamic treatment regimes. Biometrics 63, 447–455.

Murphy, S.A., 2003. Optimal dynamic treatment regimes. J. R. Stat. Soc. Ser. B Stat. Methodol. 65, 331–355.

Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. Ser. B Stat. Methodol. 56, 3–26.

Owen, A.B., 2001. Empirical Likelihood. CRC Press.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Ann. Statist. 22, 300–325.

Robins, J.M., Hernán, M.A., 2009. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (Eds.), Longitudinal Data Analysis. Chapman & Hall, Boca Raton, FL, pp. 553–599.

Robins, J.M., Hernán, M.A., Wasserman, L., 2015. Discussion of on bayesian estimation of marginal structural models. Biometrics 71, 296–299.

Robins, J.M., Mark, S.D., Newey, W.K., 1992. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics 48, 479–495.

Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846–866.

Robins, J.M., Wang, N., 2000. Inference for imputation estimators. Biometrika 87, 113–124.

Robins, J.M., Wasserman, L.A., 2000. Conditioning, likelihood, and coherence: A review of some foundational concepts. J. Amer. Statist. Assoc. 95, 1340–1346.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rubin, D.B., 1978. Bayesian inference for causal effects: The role of randomization. Ann. Statist. 6, 34–58.

Rubin, D.B., 1981. The Bayesian bootstrap. Ann. Statist. 9, 130–134.

Rubin, D.B., 1985. The use of propensity scores in applied Bayesian inference. Bayesian Stat. 2, 463–472.

Rubin, D.B., 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Stat. Med. 26, 20–36.

Rubin, D.B., Thomas, N., 1996. Matching using estimated propensity scores: relating theory to practice. Biometrics 52, 249–264.

Saarela, O., Belzile, L.R., Stephens, D.A., 2016. A Bayesian view of doubly robust causal inference. Biometrika 103, 667–681.

Saarela, O., Stephens, D.A., Moodie, E.E., Klein, M.B., 2015. On Bayesian estimation of marginal structural models. Biometrics 71, 279–288.

Scharfstein, D.O., Rotnitzky, A., Robins, J.M., 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J. Amer. Statist. Assoc. 94, 1096–1120.

Schennach, S.M., 2005. Bayesian exponentially tilted empirical likelihood. Biometrika 92, 31–46.

Schennach, S.M., 2007. Point estimation with exponentially tilted empirical likelihood. Ann. Statist. 35, 634–672.

Stephens, D.A., Nobre, W.S., Moodie, E.E.M., Schmidt, A.M., 2022. Causal inference under mis-specification: adjustment based on the propensity score. Bayesian Anal. 1–24, Advance Publication.

Sueishi, N., 2013. Identification problem of the exponential tilting estimator under misspecification. Econom. Lett. 118, 509–511.

Tsiatis, A., 2007. Semiparametric Theory and Missing Data. Springer Science & Business Media.

Van der Vaart, A.W., 2000. Asymptotic Statistics. Cambridge University Press.

Walker, S.G., Hjort, N.L., 2001. On Bayesian consistency. J. R. Stat. Soc. Ser. B Stat. Methodol. 63, 811–821.

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