Published in final edited form as:
Biometrika. 2020 December ; 107(4): 857–873. doi:10.1093/biomet/asaa028.

Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood

A. Yiu, **R. J. B. Goudie**, **B. D. M. Tom**
Medical Research Council Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Robinson Way, Cambridge CB2 0SR, U.K.

Summary

Fully Bayesian inference in the presence of unequal probability sampling requires stronger structural assumptions on the data-generating distribution than frequentist semiparametric methods, but offers the potential for improved small-sample inference and convenient evidence synthesis. We demonstrate that the Bayesian exponentially tilted empirical likelihood can be used to combine the practical benefits of Bayesian inference with the robustness and attractive large-sample properties of frequentist approaches. Estimators defined as the solutions to unbiased estimating equations can be used to define a semiparametric model through the set of corresponding moment constraints. We prove Bernstein–von Mises theorems which show that the posterior constructed from the resulting exponentially tilted empirical likelihood becomes approximately normal, centred at the chosen estimator with matching asymptotic variance; thus, the posterior has properties analogous to those of the estimator, such as double robustness, and the frequentist coverage of any credible set will be approximately equal to its credibility. The proposed method can be used to obtain modified versions of existing estimators with improved properties, such as guarantees that the estimator lies within the parameter space. Unlike existing Bayesian proposals, our method does not prescribe a particular choice of prior or require posterior variance correction, and simulations suggest that it provides superior performance in terms of frequentist criteria.

Keywords

Bayesian method of moments; Bernstein–von Mises theorem; Double robustness; Exponentially tilted empirical likelihood; M-estimation; Selection bias

1 Introduction

Consider the problem of estimating a population parameter vector in the presence of unequal probability sampling. We investigate two settings. The first, which we refer to as the design setting, assumes a selection mechanism determined by the data collector, but only partial design information in the form of sampling probabilities for the selected individuals is provided to the analyst. This situation is frequently encountered when analysing public-use survey datasets (Si et al., 2015; Zanganeh & Little, 2015; Wang et al., 2017). The second scenario is an observational setting where the selection mechanism is unknown, but assumed to be ignorable conditional on a set of fully observed covariates. Certain problems in missing

Europe PMC Funders Author Manuscripts

data and causal inference, such as estimating an average treatment causal effect with no unmeasured confounders, can be formulated in this framework (Kang & Schafer, 2007).

In both cases, it is common in practice to use semiparametric estimators that incorporate inverse probability weighting. If the selection probabilities are known, weighting methods are simple to implement and require few modelling assumptions for consistency (Horvitz & Thompson, 1952; Hájek, 1971). In the observational setting, inverse probability weights estimated from a selection model can be combined with an imputation model to produce doubly robust estimators that are consistent as long as one of the models is correctly specified (Robins et al., 1994; Scharfstein et al., 1999; Kang & Schafer, 2007; Rotnitzky & Vansteelandt, 2014). However, while the large-sample properties of these estimators are attractive, their reliability for small datasets is less justified theoretically. In particular, the use of inverse probability weighting can lead to disastrous performance in the presence of model misspecification and a practical violation of positivity (Kang & Schafer, 2007). For small-sample inference, the prior distribution in a Bayesian approach can offer both regularization and a systematic way of incorporating informative knowledge into the analysis, with this influence gradually relaxed as the sample size increases. The Bayesian paradigm also provides a convenient framework for evidence synthesis (Ades & Sutton, 2006); repeatedly integrating over the posterior distributions of parameters allows one to propagate uncertainty across multiple data sources within a single analysis.

A significant drawback of Bayesian approaches is the requirement of stronger structural assumptions on the data distribution and the sampling mechanism. In the design setting, one option is to specify a flexible regression imputation model with the sampling probability included as a covariate to adjust for the selection bias. To obtain estimates for the target population, the sampling probability can be integrated out using a sampling probability model conditional on selection (Si et al., 2015; Zanganeh & Little, 2015). Alternatively, one could adopt a sample likelihood approach (Pfeffermann et al., 2006) which truncates the dataset to just the sampled individuals and requires the specification of a conditional selection model. Both approaches involve directly modelling the dependence structure between the incomplete data and the sampling probabilities, rather than using the unavailable design variables specified by the data collector. This is potentially difficult to specify correctly. In the observational setting, Bayesian estimators will generally fail to be doubly robust; either the selection mechanism is ignored or the model parameters are a priori dependent, so that misspecification of just one model can feed back into the other, precluding consistency (Zigler et al., 2013; Robins et al., 2015). Specializing to mean estimation of binary outcomes, Ray & van der Vaart (2019) presented a thorough treatment of assumptions and priors for fully Bayesian approaches to satisfy Bernstein–von Mises theorems, as well as the proposal of novel propensity score-dependent priors that incorporate a preliminary estimate of the propensity score into the prior of the outcome regression model.

There have been a number of Bayes-like proposals that aim to resolve these issues. In a survey inference context, Wang et al. (2017) suggested using an approximate normal likelihood centred at a weighted estimator to match the robustness and simplicity of the frequentist approach. McCandless et al. (2010) and Graham et al. (2016) achieved double

robustness by cutting the feedback between the models, but this necessitates a post hoc variance correction. This idea was reviewed by Saarela et al. (2016), who also proposed a method for doubly robust estimation using a Bayesian bootstrap model. Although their method does not require variance correction and matches the performance of the frequentist estimators, it prescribes a specific choice of noninformative prior. For the estimation of average treatment effects, with a particular focus on settings where the dimension of the covariates is high relative to the sample size, Antonelli & Dominici (2019) proposed a partially Bayesian method. The resulting posterior has the double robustness property, but also requires a variance correction using a method such as the nonparametric bootstrap.

In this paper we develop an inferential framework that offers the practical benefits of Bayesian statistics described above, along with the attractive asymptotic guarantees of frequentist semiparametric estimators. Central to our method is a novel application of Bayesian exponentially tilted empirical likelihood (Schennach, 2005), an approach which forms a posterior by combining a prior with a likelihood function defined by moment conditions. We specialize to the domain of M-estimation, since many proposed semiparametric estimators (e.g., Hájek, 1971; Robins et al., 1994; Scharfstein et al., 1999; Cao et al., 2009; Rotnitzky et al., 2012) are M-estimators, and the unbiased estimating equations they solve are used to define a set of corresponding moment constraints. We prove Bernstein–von Mises theorems showing that the resulting Bayesian exponentially tilted empirical likelihood posterior becomes approximately normal, centred at the chosen estimator with matching asymptotic variance; the choice of prior is unrestricted, apart from continuity and nonzero mass in a neighbourhood of the probability limit of the estimator. Thus, the posterior shares the analogous properties of the estimator, such as double robustness and local efficiency, and the frequentist coverage of any credible set will be approximately equal to its credibility. In particular, the latter implication extends the large-sample posterior properties proved by Chib et al. (2018), and provides an interpretation of the credible sets as regularized or shrinkage estimators of confidence sets, filling a conceptual gap otherwise left empty due to the procedure not being fully Bayesian.

Additionally, we prove that a separation condition, similar to the requirement in Theorem 1 of Chib et al. (2018), is implied under standard assumptions for the consistency of M-estimators. This allows the user to avoid a potentially difficult verification. Schennach (2005) provided an interpretation of Bayesian exponentially tilted empirical likelihood which justifies its use as a Bayesian procedure. However, the conditions of this result are not satisfied in our design setting in § 2.3. We give an alternative interpretation, connecting the likelihood function with a proper likelihood arising from an exponential family of maximum-entropy distributions, and suggest that this paves the way for future work.

Our approach offers the ability to obtain modified versions of existing estimators with improved properties, even in the absence of informative priors. For example, certain proposed estimators (e.g., Cao et al., 2009) may have a nonzero probability of lying outside the parameter space, potentially leading to suboptimal finite-sample performance (Rotnitzky et al., 2012). This problem can be rectified by simply restricting the support of the prior, yielding a new estimator which is population bounded in accordance with the variation of its predecessor and has identical asymptotic behaviour. Having a posterior distribution also

allows the user to have a choice of estimators, such as the mean, median or maximum a posteriori estimator, depending on the situation or the target loss function.

2 Proposal

2.1 Exponentially tilted empirical likelihood

Suppose that D is a random vector drawn from a distribution P_0. The objective is to estimate the parameter $\theta_0 \in \Theta \subset \mathbb{R}^m$, which is assumed to satisfy the moment condition $E_{P_0}\{g(D, \theta_0)\} = 0$, where g is a function mapping into $\mathbb{R}^m$. Thus, the dimension of the moment conditionis assumed to match the dimension of the parameter. The observed data D_i ($i = 1, \ldots, n$) are independent and identically distributed replicates of D with realized values d_i. An M-estimator $\hat{\theta}_n$ solves the estimating equation $n^{-1}\sum_{i=1}^{n} g(d_i, \theta) = 0$ for $\theta \in \Theta$. Many proposed estimators for unequal probability sampling problems take this form, accompanied by a set of regularity assumptions similar to the following.

Assumption 1

- **(i)** The parameter space Θ of θ is compact, and θ_0 lies in the interior of Θ and is the unique solution to $E_{P_0}\{g(D, \theta)\}=0$.
- **(ii)** With probability 1, there is a unique solution $\hat{\theta}_n$ to $n^{-1}\sum_{i=1}^{n} g(D_i, \theta) = 0$ for each n.
- **(iii)** The variance $\Omega_0 = \mathrm{var}_{P_0}\{g(D, \theta_0)\}$ is nonsingular.
- **(iv)** The expectation $E_{P_0}\left\{\sup_{\theta \in \Theta} ||g(D, \theta)||_2^2\right\} < \infty$
- **(v)** With probability 1, $g(D, \theta)$ is continuous at each $\theta \in \Theta$.
- **(vi)** With probability 1, $g(D, \theta)$ is continuously differentiable with respect to θ in a neighbourhood Θ' of θ_0, and $E_{P_0}\{\sup_{\theta' \in \Theta'} ||\partial_\theta g(D, \theta')||_F\} < \infty$ where ∂_θ denotes the partial derivative with respect to θ and F refers to the Frobenius norm.
- **(vii)** The matrix $G_0 = E_{P_0}\{\partial_\theta g(D, \theta_0)\}$ is invertible.

Assumption 1 is sufficient for the M-estimator $\hat{\theta}_n$ to be consistent and asymptotically normally distributed (van der Vaart, 1998),

$$n^{1/2}\left(\hat{\theta}_n - \theta_0\right) \to N(0, \Sigma_0)$$

in the sense of convergence in distribution, with $\Sigma_0 = \left(G_0^{\mathrm{T}}\Omega_0^{-1}G_0\right)^{-1}$ where G_0 and Ω_0 can be consistently estimated by

$$\hat{G}_n = n^{-1}\sum_{i=1}^{n} \partial_\theta g\left(D_i, \hat{\theta}_n\right), \quad \hat{\Omega}_n = n^{-1}\sum_{i=1}^{n} g\left(D_i, \hat{\theta}_n\right) g\left(D_i, \hat{\theta}_n\right)^{\mathrm{T}}, \tag{1}$$

respectively.

The moment condition g can also define a semiparametric model by restricting to distributions P that satisfy $E_P\{g(D, \theta)\} = 0$ for $\theta \in \Theta$. For values of θ such that the origin lies in the convex hull of $\{g(d_i,\theta): i = 1,\ldots,n\}$, the exponentially tilted empirical likelihood (Jing & Wood, 1996; Corcoran, 1998; Lee & Young, 1999; Schennach, 2005) is defined, up to a constant factor, as

$$L_n(\theta) = \prod_{i=1}^{n} np_i(\theta),$$

where the probabilities $p_1(\theta), \ldots, p_n(\theta)$ solve the optimization problem

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} (-p_i \log p_i) \tag{2}$$

subject to

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(d_i, \theta) = 0, \quad p_i \geqslant 0 \quad (i = 1, \ldots, n). \tag{3}$$

For other values of θ, $L_n(\theta)$ is set to 0. The function $p(\theta) = \{p_1(\theta),\ldots,p_n(\theta)\}^{\mathrm{T}}$ is well-defined since: (i) for each value ofθ the constraint set is compact and the objective function is continuous, so if the constraint set is nonempty then the objective function attains the maximum; and (ii) the objective function is strictly concave, so there is a unique maximizer.

One may interpret this likelihood function as being derived from a θ-parameterized set of multinomial distributions supported on the observed data values. For each value of θ, the solution minimizes the Kullback–Leibler divergence from the empirical distribution subject to the constraint $EP\{g(D, \theta)\}=0$. More precisely, the Kullback–Leibler divergence is minimized with the empirical distribution as the second argument; the opposite direction corresponds to the empirical likelihood (Owen, 2001). This connection mirrors the relationship between variational Bayesian methods and expectation propagation (Gelman et al., 2013). The exponentially tilted empirical likelihood is connected to M-estimation in the following manner.

Proposition 1— *The M-estimator* $\hat{\theta}_n$ *maximizes the exponentially tilted empirical likelihood.*

Furthermore, we show that Assumption 1 is sufficient for establishing the following separation property, which says that L_n decays superpolynomially to 0 outside of any ball around $\hat{\theta}_n$.

Theorem 1— *If Assumption 1 is satisfied, then for any* $\delta > 0$ *there exists an* $\varepsilon > 0$ *such that*

$$\sup_{\| \theta - \hat{\theta}_n \|_2 \geqslant \delta} \frac{L_n(\theta)}{L_n(\hat{\theta}_n)} \leqslant \exp\left\{-\varepsilon(n-1)^{1/2}\right\}$$

with probability approaching 1.

2.2 Bayesian exponentially tilted empirical likelihood

From a Bayesian perspective, Schennach (2005) proposed that the exponentially tilted empirical likelihood could be combined with a prior $p(\theta)$ to form a posterior

$$p(\theta \mid d_1, \ldots, d_n) \propto L_n(\theta) p(\theta)$$

and referred to this approach as Bayesian exponentially tilted empirical likelihood. Schennach justified this by proving that if all observed data values are distinct, then $L_n(\theta)$can be represented as a limit

$$L_n(\theta) = \lim_{\varepsilon \to 0} \lim_{B \to \infty} \int \left\{ \prod_{i=1}^{n} p(d_i \mid \xi_B) \right\} p(\xi_B \mid \theta; \varepsilon) \mathrm{d}\xi_B,$$

which suggests that it has a proper probabilistic interpretation as a likelihood derived from a semiparametric model after marginalizing an infinite-dimensional nuisance parameter. The prior for the nuisance parameter $\xi_B = (\xi_{B,1}, \ldots, \xi_{B,B})^{\mathrm{T}}$ conditional on θ and a positive real number ε is a distribution on a grid of values such that the induced mixture of uniform densities centred on the components of ξ_B satisfies the moment restrictions with in a tolerance of ε, favouring mixtures with small support. Conditional on ξ_B, D_i is distributed according to the corresponding mixture of uniform densities. As $B \to \infty$, the spacing of the grid of values tends to zero and the range tends to infinity. Chib et al. (2018) further proved Bernstein–von Mises results, showing that the total variation distance between the posterior distribution of $n^{1/2}(\theta - \theta_0)$ and the normal distribution $N(0, \sum_0)$ tends to zero under correctly specified moment constraints.

We specialize to the domain of M-estimation and prove a Bernstein-von Mises theorem with centring point being the M-estimator $\hat{\theta}_n$. This implies not only that the posterior is consistent and asymptotically normal, but also that frequentist coverage of any credible set will be approximately equal to its credibility, extending the properties implied by the results of Chib et al. (2018). We specify a distinct set of further assumptions.

If $L_n(\theta)$ is nonzero, the optimization problem specified by (2) and (3) can be solved (Schennach, 2007) by considering the dual problem

$$p_i(\theta) = \frac{\exp\left\{ \hat{\lambda}_n(\theta)^{\mathrm{T}} g(d_i, \theta) \right\}}{\sum_{j=1}^{n} \exp\left\{ \hat{\lambda}_n(\theta)^{\mathrm{T}} g(d_j, \theta) \right\}}, \tag{4}$$

where $\hat{\lambda}_n(\theta)$ solves

$$\sum_{i=1}^{n} \exp\left\{ \lambda^{\mathrm{T}} g(d_i, \theta) \right\} g(d_i, \theta) = 0. \tag{5}$$

Assumption 2—There exists a neighbourhood B of θ_0 on which, with probability approaching 1, the exponentially tilted empirical likelihood is nonzero or, equivalently, there exists a function $\hat{\lambda}_n : \mathscr{B} \to \mathbb{R}^m$ such that for all $\theta \in B$,

$$\sum_{i=1}^{n} \exp\left\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g(d_i, \theta)\right\} g(d_i, \theta) = 0.$$

Assumption 3—For almost all values of d, $g(d,\theta)$is twice differentiable with respect to θ in a neighbourhood ofθ_0, and the second derivative satisfies the Lipschitz condition

$$\| \partial_\theta^2 g(d, \theta) - \partial_\theta^2 g(d, \theta') \|_{\mathrm{op}} \leqslant \psi(d) \| \theta - \theta' \| 2$$

for an integrable function ψ, where op refers to the operator norm.

Assumption 4—For almost all values of d, there exists a neighbourhood of $(0,\theta_0)$ contained in $\mathbb{R}_m \times \Theta$ in which the function

$$f(\lambda, \theta) = \exp\left\{\lambda^{\mathrm{T}} g(d, \theta)\right\} g(d, \theta)$$

and all of its first and second partial derivatives are dominated by an integrable function.

These conditions allow us to establish the following intermediate result.

Proposition 2— *If* Assumptions 3 and 4 *are satisfied, then on a neighbourhood of* θ_0 *there exists a unique function* λ_0 *mapping into* $\mathbb{R}^m$ *such that*

$$E_{P_0}\left[\exp\left\{\lambda_0(\theta)^{\mathrm{T}} g(D, \theta)\right\} g(D, \theta)\right] = 0$$

and λ_0 *is twice Lipschitz differentiable*.

Consequently, one can generate an exponential family $\{P\theta\}$ from P_0 via

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}P_0}(d) = \exp\left\{\lambda_0(\theta)^{\mathrm{T}} g(d, \theta) - \kappa(\theta)\right\}$$

locally around θ_0, where $\kappa(\theta) = \log E_{P_0}[\exp\{\lambda_0(\theta)^{\mathrm{T}} g(D, \theta)\}]$ and $EP_\theta\{g(D,\theta)\}=0$. The exponentially tilted distribution P_θ is the I-projection (see Csiszár, 1975) of P_0 onto the set $\{P: EP\{g(D, \theta)\}=0\}$, i.e., the closest element to P_0 in the set in terms of Kullback–Leibler divergence. In this local region of θ_0, the exponentially tilted empirical likelihood is approximately equal to the likelihood generated by this exponential family. This suggests that the exponentially tilted empirical likelihood is a plug-in estimate of a least-favourable family of distributions aimed at reducing the original semiparametric model to a parametric model in a minimally informative way. This provides a general interpretation of the Bayesian exponentially tilted empirical likelihood approach that holds even in certain

situations to which the Schennach (2005) interpretation does not apply, such as the design setting in §2.3 where the set of observed data values may not be distinct.

Theorem 2— *Suppose that Assumptions* 1–4 *hold. Assume also that the prior* $p(\theta)$ *admits a continuous density with respect to the Lebesgue measure andis positive at* θ_0. *Then*

$$\int_{\Theta} \left| p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| \mathrm{d}\theta \to 0$$

in probability, where $p_{\hat{\theta}_n, n^{-1}\Sigma_0}$ *is the density of* $N\left(\hat{\theta}_n, n^{-1}\Sigma_0\right)$.

By centring and scaling and using an alternative form of the total variation distance (Tsybakov, 2009), we obtain the equivalent representation

$$\sup_{B} \left| \mathrm{pr}\left\{ n^{1/2}\left(\theta - \hat{\theta}_n\right) \in B \mid D_1, \ldots, D_n \right\} - N(0, \Sigma_0)(B) \right| \to 0$$

in probability, where B ranges over all elements of the Borel sigma-algebra on $\mathbb{R}^m$. Theorem 2 implies both posterior consistency and asymptotically correct frequentist coverage of credible sets. The following result confirms the first-order equivalence between the posterior mean and $\hat{\theta}_n$, establishing the validity of the methodology as a shrinkage estimation framework that can yield finite-sample gains while matching the asymptotic performance of the standard estimator.

Theorem 3— *Suppose* Assumptions 12013;4 *hold and that* $\int \|\theta\|_2 \, p(\theta) \, \mathrm{d}\theta < \infty$. *Let* $\theta_n^* = \int \theta \, p(\theta \mid d_1, \ldots, d_n) \, \mathrm{d}\theta$ *be the Bayesian exponentially tilted empirical likelihood posterior mean.*

Then

$$n^{1/2}\left(\hat{\theta}_n - \theta_n^*\right) \to 0, \quad n^{1/2}\left(\theta_n^* - \theta_0\right) \to N(0, \Sigma_0),$$

where the convergence is in probability and in distribution, respectively.

2.3 Design setting

We first consider the estimation of a population parameter in a design setting where the selection probabilities are known for the sampled individuals. The data $D_i = (R_i Z_i, R_i, R_i \pi_i)$ ($i = 1, 2026; , n$) are independent and identically distributed from P_0; here R_i is the selection indicator, which is equal to 1 if Z_i is observed and 0 otherwise, $\pi_i = \mathrm{pr}(R_i = 1 \mid W_i)$, and Z_i and R_i are conditionally independent given W_i. The variables $W_1, \ldots, W_n$ are the design variables chosen by the data collector to assign sampling probabilities to individuals in the target population, but they are not included in the dataset. We make the positivity assumption that there exists a $\delta > 0$ such that $\pi_i \geqslant \delta$ with probability 1. The target parameter θ_0 is the unique solution to $E_{P_0}\{u(Z, \theta)\} = 0$ for a function u and $\theta \in \Theta \subset \mathbb{R}^m$. The

full-data estimating function u is adapted below to the estimating function g for the observed data, allowing us to apply Theorem 2.

Example 1 (Outcome mean). We have $Z = Y$ and $u(Z, \theta) = Y - \theta$.

Example 2 (Linear regression). We have $Z = (Y, X)$ and $u(Z, \theta) = X^{\mathrm{T}}(Y - X\theta)$.

Consider the estimator $\hat{\theta}_n$ that solves the estimating equation

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i} u(Z_i, \theta) = 0.$$

To address the technicality that the sampling probabilities are provided as $R_i\boldsymbol{\pi}_i$ in the notation rather than just $\boldsymbol{\pi}_i$, we set $R_i/(R_i\boldsymbol{\pi}_i) = 0$ when $R_i = 0$ so that $R_i/(R_i\boldsymbol{\pi}_i)$ is equivalent to $R_i/\boldsymbol{\pi}_i$. In the case of estimating the population outcome mean, this estimator specializes to the Hájek estimator (Hájek, 1971). For $D = (RY, R, R\boldsymbol{\pi}) \sim P_0$ and $g(D, \theta) = Ru(Z, \theta)/\boldsymbol{\pi}$,

$$\begin{aligned} E_{P_0}\{g(D,\theta)\} &= E_W E_{P_0 \mid W}\left\{\frac{R}{\pi} u(Z,\theta) \mid W\right\} \\ &= E_W\left[\frac{E_{P_0 \mid W}(R \mid W)}{\pi} E_{P_0 \mid W}\{u(Z,\theta) \mid W\}\right] \\ &= E_{P_0}\{u(Z,\theta)\}, \end{aligned}$$

where we have used the conditional independence of R and Z conditional on W and the equality o f $E_{P_0 \mid W}\{R \mid W\}$ and $\boldsymbol{\pi}$. This shows that θ_0 is the unique solutionto $E_{P_0}\{g(D, \theta)\}$=0. Let $L_n(\theta)$ be the exponentially tilted empirical likelihood function corresponding to the moment conditions $E_{P_0}\{g(D, \theta)\} = 0$ for $\theta \in \Theta$. The likelihood function is combined with a user-specified prior $p(\theta)$ to form a posterior

$$p(\theta \mid d_1, \ldots, d_n) \propto L_n(\theta) p(\theta).$$

If Assumptions 1–4 are satisfied and $p(\theta)$ is continuous and nonzero around θ_0, Theorem 2 implies that

$$\int_{\Theta} \left| p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| d\theta \to 0$$

in probability, where $p_{\hat{\theta}_n, n^{-1}\Sigma_0}$ is the density of $N(\hat{\theta}_n, n^{-1}\Sigma_0)$ and $\Sigma_0 = \lim_{n \to \infty} \mathrm{var}_{P_0}(n^{1/2}\hat{\theta}_n)$ as described in §2.1 and §2.2. Since $\hat{\theta}_n$ is a consistent estimator of θ_0, the posterior will concentrate around θ_0 as n gets large. Furthermore, since Σ_0 is equal to the asymptotic variance of $n^{1/2}\hat{\theta}_n$, the frequentist coverage of any credible set will be approximately equal to its credibility.

2.4 Observational setting

We consider estimation of a population parameter when the selection mechanism is unknown. The observed data $D_i = (R_i Z_i, R_i, W_i)(i = 1, \ldots, n)$ are independent and identically distributed from P_0; Z_i and R_i are as before, and W_i is a vector of covariates observed for each i such that Z_i and R_i are conditionally independent given W_i. The target parameter γ_0 is the unique solutionto $E_{P_0}\{u(Z,\gamma)\} = 0$ for a function u and values of γ belonging to a compact real subset Γ. In a missing data context, the conditional independence of Z_i and R_i is sometimes referred to as a missing-at-random assumption. This set-up may also be viewed as one arm of a point exposure causal inference problem in the potential outcomes framework, with the conditional independence corresponding to an assumption of no unmeasured confounders.

Let $\pi_0(W) = \mathrm{pr}(R = 1 \mid W)$ be the true propensity score and let $\varphi_0(W, \gamma) = E_{P_0}\{u(Z, \gamma) \mid W\}$. We make the positivity assumption that there exists a $\delta > 0$ such that $\pi_0(W) \geqslant \delta$ with probability 1. Solving

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i u(Z_i,\gamma)}{\hat{\pi}(W_i)} - \hat{\phi}(W_i,\gamma)\left\{\frac{R_i}{\hat{\pi}(W_i)} - 1\right\}\right] = 0,$$

where $\hat{\pi}$ and $\hat{\phi}$ are estimators of π_0 and φ_0, respectively, leads to a doubly robust estimator of Y; that is, the estimator is consistent and asymptotically normal as long as at least one of $\hat{\pi}$ and $\hat{\phi}$ is consistent. There is a significant body of work regarding choices for $\hat{\pi}$ and $\hat{\phi}$, particularly for population outcome mean estimation, which lead to various favourable efficiency properties. See Kang & Schafer (2007) and Rotnitzky & Vansteelandt (2014) for comprehensive reviews.

If $\hat{\pi}$ and $\hat{\phi}$ are derived from the solutions to unbiased estimating equations, as is often the case in practice, one can exploit this to formulate a set of nested moment constraints for an exponentially tilted empirical likelihood model. We show in Theorem 4 that the resulting marginal posterior distribution of γ is calibrated asymptotically to the behaviour of the selected estimator.

We restrict our attention to parametric working models $\pi(W; \alpha)$ and $\varphi(W, \gamma; \beta)$ for real-valued parameters α and β. Suppose that $(\hat{\alpha}_n, \hat{\beta}_n, \hat{\rho}_n)$ solve the unbiased estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} U_{\alpha,\beta,\rho}(D_i, \alpha, \beta, \rho) = 0,$$

where ρ is a set of additional auxiliary parameters, possibly empty (Rotnitzky & Vansteelandt, 2014). The two parameters (α, β) can be estimated either separately or together. For example, in the case of mean estimation, Robins et al. (1994) estimated α with maximum likelihood for a logistic regression model, and estimated β separately using ordinary least squares. Scharfstein et al. (1999) also used maximum likelihood estimation

for α, but included the reciprocal of the propensity score as a covariate in the outcome regression model.

Let $\hat{\gamma}_n$ be the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i u(Z_i,\gamma)}{\pi(W_i,\hat{\alpha}_n)}-\phi\left(W_i,\gamma;\hat{\beta}_n\right)\left\{\frac{R_i}{\pi(W_i,\hat{\alpha}_n)}-1\right\}\right]=0.$$

Let $\theta=(\alpha,\beta,\rho,\gamma)$ and define $g(D,\theta)=\{U_{\alpha,\beta,\rho}(D,\alpha,\beta,\rho)^{\mathrm{T}}, h(D,\alpha,\beta,\gamma)^{\mathrm{T}}\}^{\mathrm{T}}$, where

$$h(D,\alpha,\beta,\gamma)=\frac{Ru(Z,\gamma)}{\pi(W;\alpha)}-\phi(W,\gamma;\beta)\left\{\frac{R}{\pi(W;\alpha)}-1\right\}.$$

In accordance with Assumption 1(i), we assume that there exists a value $\theta_0=(\alpha_0,\beta_0,\rho_0,\gamma^*)$ which is the unique solution to $E_{P_0}\{g(D,\theta)\}=0$. We say that the working model for the propensity score is correctly specified if $\pi_0(W)=\pi(W;\alpha_0)$, and similarly we say that the model for ϕ is correctly specified if $\phi_0(W,\gamma)=\phi(W,\gamma;\beta_0)$. If at least one of the models is correctly specified, then $\gamma^*=\gamma_0$ and the M-estimator $\hat{\gamma}_n$ consistently estimates the true parameter.

Let $L_n(\theta)$ be the exponentially tilted empirical likelihood function corresponding to the moment conditions $EP\{g(D,\theta)\}=0$. The likelihood function is combined with a user-specified prior $p(\theta)$ to form a posterior

$$p(\theta\mid d_1,\ldots,d_n)\propto L_n(\theta)p(\theta).$$

Let $p(\gamma\mid d1,\ldots,dn)$ be the marginal posterior for γ.

Theorem 4— *Suppose* Assumptions 1–4 *hold and that the prior* $p(\theta)$ *admits a continuous density with respect to Lebesgue measure and is positive at* θ_0. *Then, as* $n\to\infty$,

$$\int_{\Gamma}\left|p(\gamma\mid D_1,\ldots,D_n)-p_{\hat{\gamma}_n,n^{-1}V_0}(\gamma)\right|d\gamma\to 0$$

in P_0-probability, where $p_{\hat{\gamma}_n,n^{-1}V_0}$ is the density of $N(\hat{\gamma}_n,n^{-1}V_0)$ *and* $V_0=\lim_{n\to\infty}\mathrm{var}_{P_0}\left(n^{1/2}\hat{\gamma}_n\right)$.

As stated earlier, $\hat{\gamma}_n$ is, by construction, consistent for estimating γ_0 provided either $\pi_0(W)=\pi(W;\alpha_0)$ or $\phi_0(W,\gamma)=\phi(W,\gamma;\beta_0)$ for all γ or both. Therefore, Theorem 4 implies that the exponentially tilted empirical likelihood posterior shares this double robustness property; the posterior will concentrate around the true value as long as one of the working models is correctly specified. Furthermore, credible sets for γ will asymptotically have nominal frequentist coverage if consistency holds, even if one of the working models is misspecified. If both models are misspecified, the credible sets will have approximately nominal coverage for the probability limit of $\hat{\gamma}_n$, which is possibly different from γ_0.

Europe PMC Funders Author Manuscripts

2.5 Implementation

In this subsection we describe how one can compute $L_n(\theta)$ for a fixed value of θ. To simplify the notation, write $g_i = g(D_i, \theta)$ for each $i = 1, \ldots, n$, suppressing the dependence on θ. To check whether the feasible set of the optimization problem specified by (2) and (3) is nonempty, it is sufficient and computationally convenient, for example by using an `R` (R Development Core Team, 2020) package such as `lpSolve` (Berkelaar, 2015), to check whether there exists a feasible solution to the linear programming problem

$$\begin{aligned} &\text{maximize } 0 \text{ over } \left\{x \in \mathbb{R}^n : 0 \leqslant x_i \leqslant 1, i = 1, \ldots, n\right\} \\ &\text{subject to } g^{\mathrm{T}}x = 0 \text{ and } c^{\mathrm{T}}x = 1, \end{aligned} \tag{6}$$

where $g = (g_1, \ldots, g_n)$ and $c = (1, \ldots, 1)^{\mathrm{T}}$. The objective 0 is suggested here for computational simplicity, but can be replaced by $b^{\mathrm{T}} x$ for arbitrary $b \in \mathbb{R}^n$ as we are concerned only with the feasible set. If the feasible set is empty, $L_n(\theta)$ is set to zero. Otherwise, assuming that the solution to (2) and (3) lies in the interior of the simplex, i.e., all the values of p_i are nonzero, the optimization problem can be solved by considering the dual problem described by (4) and (5).

Assuming that $\sum_{i=1}^{n} g_i g_i^{\mathrm{T}}$ is strictly positive definite, a unique solution to (5) exists and can be found via the Newton–Raphson method. This requires specifying a small convergence tolerance value with respect to a norm of choice. Pseudo-code for evaluating $L_n(\theta)$ is provided in the Supplementary Material. Once we are able to evaluate L_n pointwise, we can perform posterior inference using standard Bayesian computational machinery such as Markov chain Monte Carlo or importance sampling.

3 Simulations

3.1 Mean estimation for binary outcomes

In this simulation, we examine estimation of the population mean of binary outcomes in a design setting. In the notation of § 2.3, $Z = Y$, $u(Z, \theta) = Y - \theta$ and $\theta_0 = E_{P_0}(Y)$. The design variables W_i ($i = 1, \ldots, n$) are independent and identically distributed according to the beta distribution Be(1.5, 3.5), and the outcomes $Y_i \mid W_i \sim \mathrm{Ber}(W_i)$ so that $\theta_0 = 0.3$. The selection variables R_i ($i = 1, \ldots, n$) are independent and identically distributed according to $R_i \mid \boldsymbol{\pi}_i \sim \mathrm{Ber}(\boldsymbol{\pi}_i)$, where $\mathrm{logit}(\boldsymbol{\pi}_i) = W_i$. Thus, Y_i and the selection probability $\boldsymbol{\pi}_i$ are positively correlated, and the selection must be adjusted for in order to estimate θ_0. The data available for analysis are $D_i = (R_i Y_i, R_i, R_i \boldsymbol{\pi}_i)$ ($i = 1, \ldots, n$), so that the design variables are excluded.

Following the approach in §2.3, the M-estimator $\hat{\theta}_n$ is the Hájek estimator which solves

$$\sum_{i=1}^{n} g(D_i, \theta) = \sum_{i=1}^{n} \frac{R_i}{\pi_i}(Y_i - \theta) = 0.$$

We use g to define the exponentially tilted empirical likelihood $L_n(\theta)$, which we combine with three different priors for θ: $\theta \sim \mathrm{Be}(0.5, 0.5)$, $\theta \sim \mathrm{Un}(0, 1)$ and $\theta \sim \mathrm{Be}(1.5, 3.5)$. The

first is Jeffrey's prior. The mean of the Be(1.5, 3.5) prior is equal to θ_0, so we consider this prior informative, while the first two are considered noninformative.

We compare this approach with the proposal inWang et al. (2017, §2). In a survey inference context, Wang et al. (2017) suggested using a Bayesian approach with an approximate normal likelihood

$$\hat{\theta} \mid \theta \sim N(\theta, \hat{V}).$$

where $\hat{\theta}$ is a consistent and asymptotically normal estimator of θ_0 and $\hat{V}$ is a robust estimator of the variance of $\hat{\theta}$. The estimator $\hat{\theta}$ acts as asummary statistic for the data, such that the posterior is

$$p(\theta \mid \hat{\theta}) \propto p(\hat{\theta} \mid \theta) p(\theta),$$

where $p(\hat{\theta} \mid \theta)$ is defined by the normal model above and $p(\theta)$ is a prior for θ. We choose $\hat{\theta}$ to be the Hájek estimator defined above and estimate its variance with the nonparametric bootstrap. We use the same priors as defined above.

Table 1 compares the frequentist estimator $\hat{\theta}_n$ with the Bayesian methods. Each setting was replicated 2000 times. The Bayes point estimators are the posterior means. Coverage rates were computed based on central 95% credible regions. The Bayesian computation was carried out using importance sampling with 5000 particles for each replication, and the tolerance for computing the exponentially tilted empirical likelihood was 10^{-4}. On a standard quad-core laptop computer, a single replication of the Bayesian exponentially tilted empirical likelihood method with $n = 100$ took 0.25 seconds to run, while the method ofWang et al. (2017) took 0.04 seconds.

The Bayesian exponentially tilted empirical likelihood estimators generally have a higher magnitude of bias than the normal approximation when a noninformative prior is used, but have lower bias with the informative prior. This reflects the fact that the exponentially tilted empirical likelihood is less informative than the normal likelihood, resulting in higher shrinkage towards the prior mean. In the case of the two noninformative priors, this causes an upward bias towards the prior mean 0.5. This conservative characteristic leads to the Bayesian exponentially tilted empirical likelihood approach having superior performance in terms of root mean squared error and coverage rate across almost all settings, particularly with the smaller sample sizes for which the normal approximation is less accurate.

3.2 Doubly robust mean estimation with missing data

This simulation is conducted under the observational setting described in § 2.4 and follows the design of Kang & Schafer (2007). For each $i = 1, \ldots, n$, the vector of covariates $W_i = (W_{i1}, W_{i2}, W_{i3}, W_{i4}) \sim N(0, I_4)$ where I_4 is the 4 ×4 identity matrix, the selection indicator $R_i \mid W_i \sim \text{Ber}\{\pi 0(W_i)\}$ where

$$\pi_0(W_i) = expit\left(\alpha_{0,1} + \alpha_{0,2}^{\mathrm{T}} W_i\right), \quad \alpha_{0,1} = 0, \quad \alpha_{0,2} = (-1, 0.5, -0.25, -0.1)^{\mathrm{T}},$$

and $Z = Y$, the outcome, with $Y_i \mid W_i \sim N\{m_0(W_i), 1\}$ where

$$m_0(W_i) = \beta_{0,1} + \beta_{0,2}^{\mathrm{T}} W_i, \quad \beta_{0,1} = 210, \quad \beta_{0,2} = (27.4, 13.7, 13.7, 13.7)^{\mathrm{T}}.$$

Clearly, Y_i and R_i are conditionally independent given W_i. The data are $D_i = (R_i Y_i, R_i, W_i)$ $(i = 1, \ldots, n)$. In addition to the correctly specified models (a) $\pi(w; \alpha) = pr(R = 1 \mid W = w; \alpha) = \text{expit}(\alpha_1 + \alpha_2^{\mathrm{T}} w)$ and (b) $m(w; \beta) = E_{P_0}(Y \mid W = w; \beta) = \beta_1 + \beta_2^{\mathrm{T}} w$, we also consider the misspecified models (c) $\pi(w'; \alpha) = \text{pr}(R = 1 \mid W' = w'; \alpha) = \text{expit}\left(\alpha_1 + \alpha_2^{\mathrm{T}} w'\right)$ and (d) $m(w'; \beta) = E_{P_0}(Y \mid W' = w'; \beta) = \beta_1 + \beta_2^{\mathrm{T}} w'$, where $W_i' = (W_{i1}', W_{i2}', W_{i3}', W_{i4}')$ are transformed covariates with $W_{i1}' = \exp(W_{i1}/2), W_{i2}' = W_{i2}/\{1 + \exp(W_{i1})\} + 10, W_{i3}' = \{(W_{i1} W_{i3})/25 + 0.6\}^3$ and $W_{i4}' = (W_{i2} + W_{i4} + 20)^3$. The target parameter is $\mu_0 = E_{P_0}(Y) = 210$. We write m and μ instead of the φ and γ used in §2.4 to match the notation of Kang & Schafer (2007). For the sake of brevity, the estimators and methods described in the rest of this subsection will be expressed in terms of the correct covariates W_i. Under misspecification, the covariates W_i are replaced with W_i' as appropriate.

The doubly robust augmented inverse probability weighted estimator (Robins et al., 1994), sometimes referred to as the standard doubly robust estimator, is

$$\hat{\mu}_{\mathrm{DR}} = \sum_{i=1}^{n} \frac{1}{n}\left[\frac{R_i Y_i}{\pi(W_i; \hat{\alpha}_n)} - m\left(W_i; \hat{\beta}_n\right)\left\{\frac{R_i}{\pi(W_i; \hat{\alpha}_n)} - 1\right\}\right] \tag{7}$$

with $\hat{\alpha}_n$ and $\hat{\beta}_n$ estimated via maximum likelihood estimation or, equivalently, by solving

$$\frac{1}{n}\sum_{i=1}^{n} U_\alpha(D_i, \alpha) = 0, \quad \frac{1}{n}\sum_{i=1}^{n} U_\beta(D_i, \beta) = 0, \tag{8}$$

where U_α and U_β are the score equations for the logistic and linear regression models, respectively. In this case, the set of additional auxiliary parameters ρ referred to in §2.4 is empty.

Saarela et al. (2016) proposed a Bayesian doubly robust approach using the Bayesian bootstrap (Rubin, 1981). A Dirichlet process model is specified for D_i in the limit of the concentration measure tending to 0. Inference for μ is based on a posterior predictive distribution induced by maximizing expected utility functions. Here, we follow the approach detailed in §6.2 of their paper and choose the utility functions to match the specification of $\hat{\mu}_{\mathrm{DR}}$. More explicitly, the parameters α and β are linked to the Bayesian bootstrap model via

Europe PMC Funders Author Manuscripts

$$\alpha = \underset{\alpha}{\operatorname{argmax}} E\left\{ R\left(\alpha_1 + \alpha_2^{\mathrm{T}} W\right) + \log expit\left(\alpha_1 + \alpha_2^{\mathrm{T}} W\right)\right\},$$

$$\beta = \underset{\beta}{\operatorname{argmax}} E\left\{ R\left(Y - \beta_1 - \beta_2^{\mathrm{T}} W\right)^2\right\},$$

corresponding to maximization of the expected loglikelihoods of, respectively, the propensity score and the outcome regression models under the posterior. The target parameter μ is defined by

$$\mu = E\left[\frac{RY}{\pi(W,\alpha)} - m(W,\beta)\left\{\frac{R}{\pi(W,\alpha)} - 1\right\}\right].$$

In practice, we sample from the posterior predictive distribution by repeatedly generating uniform Dirichlet weights $\omega = (\omega_1, \ldots, \omega_n)$ and computing $\hat{\mu}_{\mathrm{DR}}$ with the fixed uniform weights $(1/n,\ldots,1/n)$ replaced by ω in (7) and (8). Define $\hat{\mu}_{\mathrm{Sa}}$ to be the posterior predictive mean of μ for this method. The Bayesian exponentially tilted empirical likelihood posterior for $\theta = (\alpha, \beta, \mu)$ is obtained by setting $u(Z, \mu) = Y - \mu$ and following the approach described in §2.4. We compare this with the doubly robust augmented inverse probability weighted estimator and the proposal of Saarela et al. (2016).

In Table 2, for both of the Bayesian exponentially tilted empirical likelihood estimators, we use independent, weakly informative generalized Student-t priors for the working models (a)–(d), all with three degrees of freedom: (a)$\alpha \sim t_3\{(0,0,0,0,0)^{\mathrm{T}}, I_5\}$ (b)$\beta \sim t_3\{(210,30,10,10,10)^{\mathrm{T}}, I_5\}$, (c) $\alpha \sim t_3\{(0, 0, 0, 0, 0)^{\mathrm{T}}, I_5\}$, and (d) $\beta \sim t_3\{(35, 50, 0, -135, 0)^{\mathrm{T}}, I_5\}$, where the shape matrices are the 5 × 5 identity matrix so that the prior variance of each component is equal to 3. For $\hat{\mu}_{\mathrm{BETEL},1}$, a weakly informative prior $\mu \sim t_3(210, 1)$ is specified for the target parameter, while $\hat{\mu}_{\mathrm{BETEL},2}$ is equipped with a more informative prior, $\mu \sim N(210,1)$. The three parameters (α,β,μ) are a priori independent across all settings. Sampling from the Saarela et al. (2016) posterior can be implemented directly, as described above. The exponentially tilted empirical likelihood was computed with a tolerance of 10^{-4}, and posterior samples were drawn using a Metropolis–Hastings algorithm with 2000 iterations, with an initial burn-in of 500 iterations. On a standard quad-core laptop computer, a single replication using the Bayesian exponentially tilted empirical likelihood method without parallelization took 55 seconds. Computing the frequentist estimator took 0.05 seconds, and the method of Saarela et al. (2016) took 3.71 seconds.

The results in Table 2 show that in all settings, $\hat{\mu}_{\mathrm{BETEL},2}$ outperforms both $\hat{\mu}_{\mathrm{DR}}$ and $\hat{\mu}_{\mathrm{Sa}}$ in terms of root mean squared error and median absolute error. This illustrates that when substantial prior knowledge of the target parameter is available, use of this information in our proposed approach leads to better overall performance than the other estimators evaluated. When at least one of the working models is correctly specified, $\hat{\mu}_{\mathrm{BETEL},1}$ exhibits a higher magnitude of bias than $\hat{\mu}_{\mathrm{DR}}$, but lower root mean squared error and median

Europe PMC Funders Author Manuscripts

absolute error, demonstrating the regularization benefits of using weakly informative priors with our proposal. When both working models are misspecified, the improvements are more substantial, suggesting that the use of a prior can help to alleviate the effects of misspecification. The patterns of results for $\hat{\mu}_{\mathrm{BETEL},1}$ compared with $\hat{\mu}_{\mathrm{Sa}}$ are less clear for the bias and root mean squared error metrics, but $\hat{\mu}_{\mathrm{BETEL},1}$ has consistently lower median absolute error.

4 Application

We examine the association between blood pressure and sodium, and potassium consumption using data from the National Health and Nutrition Examination Survey 2003–2006. The dataset includes 13 957 individuals with full data on the relevant information who were drawn from the U.S. civilian population, which we have assumed to be constant at 300 million during the time period 2003–2006. Each observation is associated with a weight variable that is assumed to be proportional to the reciprocal of the sampling probability of the individual. This follows the example found in Lumley (2010, § 5.2.4).

We work in the design setting described in § 2.3. The aim is to fit a linear regression model for blood pressure Y on sodium consumption X_1 and potassium consumption X_2. Age X_3 is also included for deconfounding. The moment condition is

$$g(D,\theta) = RW(Y - \theta_0 - X_1\theta_1 - X_2\theta_2 - X_3\theta_3)X,$$

where R is the selection indicator variable, W is the weight variable and $X = (1, X_1, X_2, X_3)^{\mathrm{T}}$. We consider the frequentist M-estimator with standard errors estimated using the sandwich estimator (1). For our Bayesian exponentially tilted empirical likelihood proposal, each regression parameter is assigned an independent prior: $\theta_0 \sim t_3(100,1)$, θ_1 and θ_2 follow half-normal distributions on the positive and negative real numbers, respectively, each with scale parameter 1, and $\theta_3 \sim t_3(0, 1)$. The priors for θ_1 and θ_2 reflect the substantial prior evidence that sodium raises blood pressure in humans while potassium does the opposite. The likelihood was computed with a tolerance of 10^{-4}, and posterior samples were drawn using a Metropolis–Hastings algorithm.

Table 3 compares the frequentist estimates with the Bayesian exponentially tilted empirical likelihood posterior mean estimates. Besides the analysis of the full dataset, an analysis of a random sample of 300 samples was also carried out. With the smaller dataset, the frequentist approach leads to a positive estimated value for the effect of potassium on blood pressure. On the other hand, the Bayesian exponentially tilted empirical likelihood approach gives an estimated value much closer to the values obtained from the full dataset, illustrating the potentially significant impact of using an informative prior. The results of both approaches converge as the sample size increases, in accordance with our theory.

5 Discussion

Empirical likelihood estimators for missing data problems have previously been proposed by Qin & Zhang (2007) and Chan & Yam (2014). Their work provides a convenient framework

for integrating multiple working models into a single analysis, extending the doubly robust property to a multiply robust one. The methods are based on maximizing the conditional empirical likelihood of the outcomes and covariates given selection, and thus differ from ours, which uses the marginal exponentially tilted empirical likelihood.

Lazar (2003) provides a justification for using the empirical likelihood for Bayesian inference based on a criterion proposed by Monahan&Boos (1992). We expect a corresponding justification to hold for the exponentially tilted empirical likelihood because of their similarity. Previous theoretical and empirical comparisons suggest that, in addition to the empirical likelihood, other approximate likelihoods such as the bootstrap likelihood (Davison et al., 1992) and the implied likelihood (Efron, 1993) would yield similar performance.

As suggested in §2.1, the empirical distribution may be viewed as an initial estimate of the true data-generating distribution in the exponentially tilted empirical likelihood. From this perspective, it is natural to wonder whether this initial estimate can be improved. While the empirical distribution can be applied very generally, its use may disregard additional known or assumed structure about the data distribution, such as its support, conditional independencies and smoothness. Nonparametric techniques such as density estimation may offer a way to incorporate this information into the initial estimate. Investigating whether such replacements are advantageous is a topic of further research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

We thank Dr Shaun Seaman and the reviewers for helpful comments and suggestions. This work was funded by the U.K. Medical Research Council.

References

Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision-making: Current approaches. J R Statist Soc. 2006; A 161: 5–35.

Antonelli J, Dominici F. Causal Inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. arXiv: 1805.04899v5. 2019.

Berkelaar, M. lpSolve: Interface to ‘Lp_solve’ v 55 to Solve Linear/Integer Programs. 2015. R package version 5.6.15, available athttps://cran.r-project.org/web/packages/lpSolve/index.html

Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the double robust estimator for a population mean with incomplete data. Biometrika. 2009; 96: 723–34. [PubMed: 20161511]

Chan KCG, Yam SCP. Oracle, multiple robust and multipurpose calibration in a missing response problem. Statist Sci. 2014; 29: 380–96.

Chib S, Shin M, Simoni A. Bayesian estimation and comparison of moment condition models. J Am Statist Assoc. 2018; 113: 1656–68.

Corcoran SA. Bartlett adjustment of empirical discrepancy statistics. Biometrika. 1998; 85: 967–72.

Csiszár I. I-divergence geometry of probability distributions and minimization problems. Ann Prob. 1975; 3: 146–58.

Davison AC, Hinkley DV, Worton BJ. Bootstrap likelihoods. Biometrika. 1992; 79: 113–30.

Efron B. Bayes and likelihood calculations from confidence intervals. Biometrika. 1993; 80: 3–26.

Gelman, A, Carlin, JB, Stern, HS, Rubin, AB. Bayesian Data Analysis. CRC Press; Boca Raton, Florida: 2013.

Graham DJ, McCoy EJ, Stephens DA. Approximate Bayesian inference for doubly robust estimation. Bayesian Anal. 2016; 11: 47–69.

Hájek, J. Foundations of Statistical Inference (Proc Sympos, Univ Waterloo, Ontario, 1970). Rinehart and Winston; Toronto: Holt: 1971. 236

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Statist Assoc. 1952; 47: 663–85.

Jing B-Y, Wood ATA. Exponential empirical likelihood is not Bartlett correctable. Ann Statist. 1996; 24: 365–9.

Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with Discussion). Statist Sci. 2007; 22: 523–39.

Lazar N. Bayesian empirical likelihood. Biometrika. 2003; 90: 319–26.

Lee SMS, Young GA. Nonparametric likelihood ratio confidence intervals. Biometrika. 1999; 86: 107–18.

Lumley, T. Complex Surveys: A Guide to Analysis using R. Wiley; Hoboken, New Jersey: 2010.

McCandless LC, Douglas IJ, Evans SJ, Smeeth L. Cutting feedback in Bayesian regression adjustment for the propensity score. Int J Biostatist. 2010; 6

Monahan JF, Boos DD. Proper likelihoods for Bayesian analysis. Biometrika. 1992; 79: 271–8.

Owen, AB. Empirical Likelihood. Chapman & Hall/CRC; New York: 2001.

Pfeffermann D, Moura FAS, Nascimento-Silva PL. Multi-level modelling under informative sampling. Biometrika. 2006; 93: 943–59.

Qin J, Zhang B. Empirical-likelihood-based inference in missing response problems and its application in observational studies. J R Statist Soc. 2007; B69: 101–22.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2020. ISBN 3-900051-07-0,http://www.R-project.org

Ray K, van der Vaart A. Semiparametric Bayesian causal inference. arXiv: 1808.04246v2. 2019.

Robins JM, Hernán M, Wasserman L. Discussion of ‘On Bayesian estimation of marginal structural models’. Biometrics. 2015; 71: 296–9. [PubMed: 25652314]

Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Statist Assoc. 1994; 89: 846–66.

Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. Biometrika. 2012; 99: 439–56. [PubMed: 23843666]

Rotnitzky, A, Vansteelandt, S. Handbook of Missing Data Methodology. Molenberghs, G, Fitzmaurice, G, Kenward, MG, Tsiatis, A, Verbeke, G, editors. CRC Press; Boca Raton, Florida: 2014. 185–212.

Rubin DB. The Bayesian bootstrap. Ann Statist. 1981; 9: 130–4.

Saarela O, Belzile LR, Stephens DA. A Bayesian view of doubly robust causal inference. Biometrika. 2016; 103: 667–81.

Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with Discussion). J Am Statist Assoc. 1999; 94: 1096–146.

Schennach S. Bayesian exponentially tilted empirical likelihood. Biometrika. 2005; 92: 31–46.

Schennach S. Point estimation with exponentially tilted empirical likelihood. Ann Statist. 2007; 35: 634–72.

Si Y, Pillai N, Gelman A. Bayesian nonparametric weighted sampling inference. Bayesian Anal. 2015; 10: 605–25.

Tsybakov, AB. Introduction to Nonparametric Estimation. Springer; New York: 2009.

van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; Cambridge: 1998.

Wang Z, Kim JK, Yang S. Approximate Bayesian inference under informative sampling. Biometrika. 2017; 105: 91–102.

Zanganeh SZ, Little RJA. Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. J Surv Statist Methodol. 2015; 3: 91–102.

Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA. Model feedback in Bayesian propensity score estimation. Biometrics. 2013; 69: 263–73. [PubMed: 23379793]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Bias, root mean squared error and coverage rate from 2000 Monte Carlo simulations using the Hájek estimator, the normal approximation of Wang et al. (2017) and the proposed Bayesian exponentially tilted empirical likelihood approach

Population size	Prior	Method	Bias (×100)	RMSE (×100)	CR (%)
$n = 25$		Hájek	0.17	11.67	
	Jeffrey's	Normal	-1.57	13.06	88.6
		BETEL	1.19	11.79	92.9
	Uniform	Normal	0.72	11.55	91.5
		BETEL	2.30	11.06	94.5
	Be(1.5, 3.5)	Normal	-1.62	10.22	92.3
		BETEL	-0.11	9.18	96.2
$n = 50$		Hájek	0.08	8.27	
	Jeffrey's	Normal	-0.69	8.92	92.1
		BETEL	0.88	8.29	94.7
	Uniform	Normal	0.39	8.28	91.7
		BETEL	1.45	8.18	94.6
	Be(1.5, 3.5)	Normal	-1.06	7.32	92.6
		BETEL	0.11	7.33	95.7
$n = 100$		Hájek	-0.08	6.01	
	Jeffrey's	Normal	-0.23	6.24	92.1
		BETEL	0.51	6.03	94.8
	Uniform	Normal	-0.11	6.03	92.5
		BETEL	0.57	5.87	94.6
	Be(1.5, 3.5)	Normal	-0.75	5.75	92.8
		BETEL	-0.08	5.56	94.9

RMSE, root mean squared error; CR, coverage rate of 95% credible regions; BETEL, Bayesian exponentially tilted empirical likelihood.

Europe PMC Funders Author Manuscripts

Table 2

Monte Carlo simulations based on 1000 replications using the standarddoublyrobust estimator, the methodofSaarela et al. (2016) andthe Bayesian exponentially tilted empirical likelihood approach

OR correct, PS correct					OR incorrect, PS correct				
Estimator	Bias	RMSE	MAE	ESD	Estimator	Bias	RMSE	MAE	ESD
$\widehat{\mu}_{\mathrm{DR}}$	−0.01	2.55	1.73	2.55	$\widehat{\mu}_{\mathrm{DR}}$	0.27	3.61	2.32	3.60
$\widehat{\mu}_{\mathrm{Sa}}$	0.01	2.57	1.71	2.57	$\widehat{\mu}_{\mathrm{Sa}}$	0.57	3.44	2.31	3.39
$\widehat{\mu}_{\mathrm{BETEL},1}$	0.04	2.23	1.24	2.23	$\widehat{\mu}_{\mathrm{BETEL},1}$	0.31	3.55	2.13	3.54
$\widehat{\mu}_{\mathrm{BETEL},2}$	−0.05	1.95	1.18	1.95	$\widehat{\mu}_{\mathrm{BETEL},2}$	0.29	2.87	1.81	2.85
OR correct, PS incorrect					OR incorrect, PS incorrect				
Estimator	Bias	RMSE	MAE	ESD	Estimator	Bias	RMSE	MAE	ESD
$\widehat{\mu}_{\mathrm{DR}}$	−0.01	2.59	1.73	2.59	$\widehat{\mu}_{\mathrm{DR}}$	−6.44	38.52	3.64	37.97
$\widehat{\mu}_{\mathrm{Sa}}$	−0.09	2.60	1.73	2.60	$\widehat{\mu}_{\mathrm{Sa}}$	−4.81	15.41	3.38	14.64
$\widehat{\mu}_{\mathrm{BETEL},1}$	0.06	2.32	1.33	2.32	$\widehat{\mu}_{\mathrm{BETEL},1}$	−5.11	14.75	3.36	13.84
$\widehat{\mu}_{\mathrm{BETEL},2}$	−0.09	2.10	1.29	2.10	$\widehat{\mu}_{\mathrm{BETEL},2}$	−2.37	4.20	2.51	3.47

RMSE, root mean squared error; MAE, median of absolute errors; ESD, empirical standard deviation; DR, doubly robust; Sa, the method of Saarela et al. (2016); BETEL, Bayesian exponentially tilted empirical likelihood; OR, outcome regression; PS, propensity score; OR correct, use of the correct outcome regression model (a); OR incorrect, use of the model (c); PS correct, use of the correct propensity score model (b); PS incorrect, use of the model (d).

Table 3

Frequentist estimates and standard errors compared with Bayesian exponentially tilted empirical likelihood posterior means and posterior standard deviations

Sample size	Method		θ_0	θ_1	θ_2	θ_3
$n = 300$	Frequentist	Estimate	95.10	0.39	0.85	0.54
		Standard error	2.85	0.63	0.94	0.04
	BETEL	Posterior mean	99.31	0.51	-0.56	0.52
		Posterior s.d.	1.22	0.29	0.39	0.03
$n = 13\,957$	Frequentist	Estimate	99.74	0.80	-0.91	0.50
		Standard error	0.80	0.15	0.19	0.01
	BETEL	Posterior mean	99.82	0.78	-0.89	0.49
		Posterior s.d.	0.39	0.09	0.12	0.01

BETEL, Bayesian exponentially tilted empirical likelihood; s.d., standard deviation.