Electronic Journal of Statistics
Vol. 20 (2026) 3546–3581
ISSN: 1935-7524
https://doi.org/10.1214/26-EJS2568

Bayesian sensitivity analysis for a missing data model

B. Eggen[1], S.L. van der Pas[2] and A.W. van der Vaart[1]

[1]*Delft Institute of Applied Mathematics, Delft University of Technology, e-mail:* b.eggen@tudelft.nl; a.w.vandervaart@tudelft.nl

[2]*Dept. of Mathematics, Vrije Universiteit Amsterdam, e-mail:* s.l.vander.pas@vu.nl

Abstract: In causal inference, it is important to study the sensitivity of the conclusions to key assumptions. We perform sensitivity analysis of the assumption that missing outcomes are missing completely at random. We follow a Bayesian approach, which is nonparametric for the outcome distribution and can be combined with an informative prior on the sensitivity parameter. We give insight in the posterior and provide theoretical guarantees in the form of Bernstein-von Mises theorems for estimating the mean outcome. We study different parametrisations of the model involving Dirichlet process priors on the distribution of the outcome and on the distribution of the outcome conditional on the subject being treated. We show that these parametrisations incorporate a prior on the sensitivity parameter in different ways and discuss the relative merits. A key result on which the above theorems rely will be a general Bernstein-von Mises theorem for the normalised extended gamma process. We also present a simulation study, showing the performance of the methods in finite sample scenarios.

MSC2020 subject classifications: Primary 62G20; secondary 62D10, 62D20.
Keywords and phrases: Bernstein-von Mises, missing outcomes, causal inference, Dirichlet process, extended gamma process, normalised completely random measures.

Received November 2023.

1. Introduction

Drawing causal conclusions from missing data usually rests on unverifiable assumptions. Sensitivity analysis allows to assess the robustness of study conclusions to these assumptions. A Bayesian approach allows to incorporate prior beliefs on parameters that govern the sensitivity, and can lead to a single summary through the posterior distribution. However, theoretical support for many Bayesian methods has been lacking.

In this paper we follow Scharfstein, Daniels and Robins (2003) in performing sensitivity analysis for the assumption that missing outcomes are missing completely at random (MCAR), i.e. the assumption that observing an individual's outcome is independent of that outcome (Little and Rubin, 2019). Under MCAR the observed outcomes can be considered representative for the missing outcomes, and an ordinary analysis of the observed outcomes yields unbiased

conclusions for the full population. Yet in many applications, MCAR is not plausible. For example, in HIV research, patients who drop out tend to have worse outcomes than those who stay on the study (see e.g. Hammer et al. (1996), Balzer et al. (2020), Scharfstein, Daniels and Robins (2003)). A naive analysis of only the observed outcomes would tend to result in biased, overly optimistic conclusions.

With sensitivity analysis, we aim to answer the question: how will our conclusions change under plausible deviations from MCAR? For example, how does our estimate of drug effectiveness change, if data is missing not at random? To get a mathematical grip on this question, we need to quantify a 'deviation from MCAR'. We do so by introducing a sensitivity parameter. Ideally, the sensitivity parameter has a clear interpretation, so that we can be precise about what deviations from MCAR can be considered 'plausible'. Moreover, we wish to impose as little structure as possible on the distribution of the observed data.

In this paper we denote the distributions of the observed and unobserved outcomes by P_1 and P_0, respectively, and are interested in the distribution P of the outcome in the full population, which is the mixture $P = (1-p)P_0 + pP_1$, for $1-p$ the probability that an observation is missing. Under MCAR the distributions P_0 and P_1 coincide, and hence the distribution of interest P agrees with the distribution P_1 of the observed outcomes. However, without further assumptions, the distribution P is not identifiable from the observed data. Following Scharfstein, Daniels and Robins (2003), we adopt the working model

$$dP_0(y) \propto e^{q(y)}\, dP_1(y),$$

where q is a given function, referred to as the sensitivity function. Any given function q allows to identify the distribution P from the observed data. The MCAR assumption corresponds to the special case that the function q is constant. Sensitivity analysis may consist of performing the statistical analysis for every function q in a given parametrised collection $\{q_\alpha : \alpha \in A\}$ of functions, and compare the results for different values of the parameter α, referred to as the sensitivity parameter. Here it is not assumed that the working model is correct, or that there exists a 'true' sensitivity function or parameter.

Typically, and certainly if we adopt a non-parametric model for P, the sensitivity parameter will not be identifiable from the data. Further inference must then rely on subjective knowledge of the substantive scientist. He or she might deem particular values of α plausible for their particular study and conduct their data analysis accordingly. In this subjective approach, the Bayesian paradigm is natural, as this allows a range of possible values accompanied by weights, as an alternative to assigning a single, fixed value to the sensitivity parameter.

In Scharfstein, Daniels and Robins (2003) a non-parametric Bayesian approach was followed, putting the classical Dirichlet process prior on P. They conjectured a Bernstein-von Mises theorem for estimating the mean of P, conditional on the sensitivity parameter. One aim of the present paper is to prove

the validity of their conjecture. For this purpose, we first show that the Dirichlet prior on P leads to an extended Gamma type prior on the observed data distribution. Next we develop novel theory for posterior distributions corresponding to such priors.

We also introduce an alternative non-parametric Bayesian modelling strategy and compare the results of the approaches. It turns out that the two approaches are asymptotically equivalent given the sensitivity function, but differ when the sensitivity function is also equipped with a prior distribution. In both approaches the final inference is a mixture of the inferences for known sensitivity functions, but the two approaches differ in the mixing distribution. In one approach this is just the prior over the sensitivity function, but in the other approach the mixing distribution is dependent on the estimated data distribution. This is true, even though in both cases the sensitivity parameter is modelled a priori independent from the other parameters. We discuss this somewhat surprising finding in Section 5, when comparing the two approaches.

Our theoretical analysis is 'frequentist Bayesian' in nature in the sense that we adopt prior modelling as a method to derive posterior distributions, but perform theoretical analysis under the assumption that the observations are sampled from a fixed distribution H_0. This observational distribution is arbitrary and independent of the sensitivity parameter.

The paper is organised as follows. After presenting the model in Section 2, we investigate two different ways of putting a prior on the model, and derive asymptotic distributions relating to these priors in Sections 3 and 4. In Section 5 we compare the results of the two approaches and discuss the merits of the two priors. In Section 6, we present a simulation study and in Section 8 we discuss open questions and further work. In particular, we discuss extension of our results to general missing at random (MAR) models and causal inference, which both require that covariates are included in the analysis. There we also briefly review other Bayesian approaches to causal inference.

Section 7 contains the main technical part of the paper, which consists of an analysis of the posterior process based on the extended gamma normalised completely random measure, that arises when using one of our two priors. Part of the simulations can be found in Appendix A, the derivation of the asymptotic information bound for the model can be found in Appendix B and several proofs can be found in Appendix C.

1.1. Notation

We use operator notation $Pg = \int g\, dP$ for the expectation of a function g under a probability measure P. The distribution of a random variable X or its conditional distribution given a random variable Y are denoted by $\mathcal{L}(X)$ or $\mathcal{L}(X|Y)$. We write $X|Y \sim P$ if $\mathcal{L}(X|Y) = P$. We also write $X \perp\!\!\!\perp Y|Z$ if the variables X and Y are conditionally independent given Z. Furthermore we write $\mathbb{P}_n$, $\mathbb{H}_n$ or $\mathbb{P}_{1,n}$ for the empirical distribution of a sample of observations, while $\mathbb{B}_P$ is a P-Brownian bridge process (defined below). We write $X_n|Y_n \rightsquigarrow Z$ if

the conditional distribution of X_n given Y_n converges in distribution to Z, in probability or almost surely. The latter concept is defined formally as convergence to zero, in probability or almost surely, of the bounded Lipschitz distance between $\mathcal{L}(X_n|Y_n)$ and $\mathcal{L}(Z)$, as explained in the beginning of Section 7.

2. Model specification

In this section we define the missing outcomes model and introduce sensitivity functions.

Let the 'full data' consist of a random variable Y with values in a measurable space $(\mathfrak{Y}, \mathcal{Y})$, and a Bernoulli variable R, with values in $\{0,1\}$, which registers whether the full data is observed or not. Instead of (Y,R), we only observe the pair (X,R), for X defined by, for some arbitrary symbol $*$, not contained in $\mathfrak{Y}$,

$$X = \begin{cases} Y, & R=1, \\ *, & R=0. \end{cases} \tag{2.1}$$

Thus (X,R) takes values in $\mathfrak{X} \times \{0,1\}$, for $\mathfrak{X} = \mathfrak{Y} \cup \{*\}$. In the case that $\mathfrak{Y} = (0,\infty)$, it is customary to choose $*$ equal to $0 \in \mathbb{R}$, and we then have the identity $X = RY$, but this special structure is irrelevant for the following. In fact, the results in the following are valid for $(\mathfrak{Y}, \mathcal{Y})$ a general Polish space with its Borel σ-field, and $\mathfrak{X}$ equipped with the σ-field generated by $\mathfrak{Y}$ and the point $*$. (Since $X = *$ if and only if $R = 0$, the variable R in the observation (X,R) is redundant, but we shall keep it for ease of notation.)

We assume that we observe a sample $(X_1,R_1),\ldots,(X_n,R_n)$ of n i.i.d. copies of (X,R), and are interested in estimating characteristics of the distribution of Y. In particular, we consider estimating the expectation $\mathbb{E}_P g(Y)$, for a given measurable function $g : \mathfrak{Y} \to \mathbb{R}$, for instance the mean outcome $\mathbb{E}_P Y$ or the distribution function $\mathbb{E}_P 1_{Y \le y}$, in the case that $\mathfrak{Y} = (0,\infty)$.

Denote the distribution of Y by P, and denote the conditional distributions of Y given $R=0$ and $R=1$ by P_0 and P_1. Then $P = pP_1 + (1-p)P_0$, for $p = \Pr(R=1)$ the success probability of R. The distribution P_1 is identifiable from X, but P_0 and hence P are not. Following Scharfstein, Daniels and Robins (2003), we assume as a working hypothesis, that for a given measurable function $q : \mathfrak{Y} \to \mathbb{R}$ such that $\int e^q \, dP_1 < \infty$,

$$P_0(A) = \frac{\int_A e^q \, dP_1}{\int e^q \, dP_1}, \qquad A \in \mathcal{Y}. \tag{2.2}$$

In particular, the measures P_0 and P_1 possess the same support. The identity $P = (1-p)P_0 + pP_1$ now implies the relationship

$$P(A) = p \int_A \left(1 + \frac{1-p}{p \int e^q dP_1} e^q\right) dP_1, \qquad A \in \mathcal{Y}. \tag{2.3}$$

Since p and P_1 are identifiable from the distribution of the data, it is then clear that, for a given function q, the distribution P is identifiable as well.

Another way to interpret model (2.2) is through the propensity score, the conditional probability of (not) being observed given the outcome. Bayes's theorem gives that (2.2) is equivalent to

$$\text{logit } \Pr(R = 0|\, Y) = \eta + q(Y), \tag{2.4}$$

where

$$e^{\eta} = \frac{1-p}{p} \frac{1}{\int e^{q}\, dP_1}. \tag{2.5}$$

Because of this relationship, the function q will be referred to as the *selection bias* or the *sensitivity function*. When $q = 0$ (or constant), the propensity score is independent of the outcome and therefore the model is MCAR. In the other case the missingness depends on the outcomes, and q models deviations from the MCAR assumption. The function q can be modelled parametrically, as in the following example, or non-parametrically.

Example (Scharfstein, Daniels and Robins (2003))**.** Let $\mathfrak{Y} = \mathbb{R}^+$ and let q range over all functions of the form $q_\alpha(y) = \alpha \log y$, where α ranges over $\mathbb{R}$. Think of a clinical trial, with $R = 0$ indicating a dropout of the study. For $\alpha > 0$, the function $y \mapsto P(R = 0|\, Y = y)$ increases monotonically from 0 to 1 as y increases from 0 to infinity. Thus for $\alpha > 0$, subjects with higher outcomes are more likely to drop out of the study. For $\alpha < 0$, this is reversed. For $\alpha = 0$, the dropout is MCAR.

We do not assume that the model (2.2) or (2.4) is correct, or that there is a 'true' frequentist parameter q_0 that generates the observations. An example situation where q_0 does not exist is when the true P_0^0 and P_1^1 do not have the same support. The main purpose is to assess the magnitude of the changes in the study conclusions when the function q varies, or is equipped with a prior.

We are interested in estimating, for a given measurable function $g : \mathfrak{Y} \to \mathbb{R}$,

$$\mathbb{E}_P g(Y) = \int g\, dP = p \int g(y) \left(1 + \frac{1-p}{p} \frac{e^{q(y)}}{\int e^{q}\, dP_1}\right) dP_1(y). \tag{2.6}$$

The right side of the equation shows that this functional is estimable from the observed data, for any given function q. We might for instance be interested in all functions q such that this functional is above or below a certain threshold, or to give confidence or credible intervals on this functional for varying q. A prior on q in addition to a prior on the data model will lead to a posterior distribution on the functional. It is reasonable to combine non-informative priors on the data distribution with an informative prior on q.

The full model can be parametrised by the triplet (p, P_1, q), but also by the triplet (η, P, q). The first triple determines the marginal distribution of R through p, and next the conditional distribution of Y given $R = 1$ through P_1, and given $R = 0$ through q in view of equation (2.2). The second triple gives the marginal distribution of Y through P and next the conditional distribution

of R given Y through (η, q), in view of equation (2.4). This suggests various possibilities for prior modelling and inference, which we discuss in the next sections.

3. Parametrisation by p, P_1

The distribution of the observation X in (2.1) is given by

$$H = (1-p)\delta_* + pP_1,$$

for δ_* the Dirac measure at the missingness indicator $*$. Since there is a bijective relation between H and the pair (p, P_1), the model can also be parametrised by the pair (H, q). The standard non-parametric prior on the distribution H is the Dirichlet process prior $\mathrm{DP}(a)$, which is given by a base measure a, a finite Borel measure on $\mathfrak{X}$ (see Ferguson (1973) or Chapter 4 in Ghosal and Van der Vaart (2017) for a review). It is well known that the Dirichlet process prior is conjugate: if $H \sim \mathrm{DP}(a)$ and $X_1 \ldots, X_n | H \sim H$, then a version of the posterior distribution is given by $H | X_1, \ldots, X_n \sim \mathrm{DP}(a + n\mathbb{H}_n)$, where $\mathbb{H}_n = \sum_{i=1}^n \delta_{X_i}/n$ is the empirical measure of the observations. Furthermore, the functional Bernstein-von Mises theorem for the Dirichlet process (Lo (1983, 1986) or Ghosal and Van der Vaart (2017), p. 364) gives that

$$\sqrt{n}\,(H - \mathbb{H}_n) \,|\, X_1, \ldots, X_n \rightsquigarrow \mathbb{B}_{H_0}, \qquad H_0^\infty - \text{a.s.} \tag{3.1}$$

Here $\mathbb{B}_{H_0}$ is an H_0-Brownian bridge process, and the convergence can be understood to mean that, for every finite set of measurable functions $g_j : \mathfrak{X} \to \mathbb{R}$ with $(H_0 + a)g_j^2 < \infty$, and for almost every realisation of the i.i.d. sequence $X_1, X_2, \ldots$ with law H_0, the posterior distribution of

$$\left(\sqrt{n}\,(H - \mathbb{H}_n)\, g_1, \ldots, \sqrt{n}\,(H - \mathbb{H}_n)\, g_k\right) | \, X_1, \ldots, X_n$$

converges to the distribution of the random vector $(\mathbb{B}_{H_0} g_1, \ldots, \mathbb{B}_{H_0} g_k)$: a multivariate normal distribution with mean zero and covariances $\mathrm{cov}(\mathbb{B}_{H_0} g_i, \mathbb{B}_{H_0} g_j) = H_0 g_i g_j - H_0 g_i H_0 g_j$. The convergence is also true in the space $\ell^\infty(\mathcal{G})$, for every H_0-Donsker class $\mathcal{G}$ with measurable envelope function G such that $(H_0 + a)G^2 < \infty$. (See e.g. Dudley (2014) or Van der Vaart and Wellner (1996) for the definition of Donsker classes.)

The posterior distribution of H induces a posterior distribution on the pair (p, P_1), and hence for a fixed function q, a posterior distribution on the parameter of interest (2.6). In fact, when the domain of q is extended to $\mathfrak{X}$ such that $\delta_*(e^q) = 0$, the parameter of interest can be expressed as a function of H as

$$\mathbb{E}_P g(Y) = \frac{H\,(ge^q)\,H\{*\}}{He^q} + Hg =: \chi(H, q). \tag{3.2}$$

Here $H\{*\}$ is the mass of H at the point $*$. We may use the (conditional) delta-method to deduce the limit behaviour of the posterior distribution of this quantity.

Theorem 3.1 (Bernstein-von Mises H). *If* $H|\, X_1, \ldots, X_n \sim DP(a + n\mathbb{H}_n)$ *and* $(H_0 + a)\,(g^2 e^{2q} + g^2 + e^{2q}) < \infty$, *then, conditional on* q, $H_0^\infty - a.s.$,

$$\sqrt{n}\big(\chi(H,q) - \chi(\mathbb{H}_n, q)\big)|\, X_1, \ldots, X_n, q$$
$$\rightsquigarrow \frac{\mathbb{B}_{H_0}(1_{\{*\}})H_0(ge^q)}{H_0 e^q} + \frac{H_0\{*\}\mathbb{B}_{H_0}(ge^q)}{H_0 e^q} - \frac{H_0\{*\}H_0(ge^q)\mathbb{B}_{H_0}(e^q)}{(H_0 e^q)^2} + \mathbb{B}_{H_0} g.$$

Proof. The functional can be written $\chi(H,q) = \phi\,(H\{*\}, H(ge^q), He^q, Hg)$, for the function $\phi : \mathbb{R}^4 \to \mathbb{R}$, defined by

$$\phi(x_1, x_2, x_3, x_4) = \frac{x_1 x_2}{x_3} + x_4.$$

The theorem therefore follows from the delta-method for conditional distributions (Van der Vaart and Wellner (1996), Section 3.9.3, or more precisely Van der Vaart and Wellner (2023), Theorem 3.10.13) combined with the functional Bernstein-von Mises theorem (3.1). The limit variable arises as the inner product

$$\nabla\phi(\theta_0) \cdot \big(\mathbb{B}_{H_0}(1_{\{*\}}), \mathbb{B}_{H_0}(ge^q), \mathbb{B}_{H_0} e^q, \mathbb{B}_{H_0} g\big),$$

where $\nabla\phi(\theta_0)$ is the gradient of the function ϕ evaluated at the vector $\theta_0 = (H_0\{*\}, H_0(ge^q), H_0 e^q, H_0 g)$. □

By the multivariate central limit theorem (or Donsker's theorem), the empirical process $\sqrt{n}(\mathbb{H}_n - H_0)$ converges (unconditionally) to the same H_0-Brownian bridge process $\mathbb{B}_{H_0}$. Therefore, the (ordinary) delta-method gives that the sequence $\sqrt{n}\big(\chi(\mathbb{H}_n, q) - \chi(H_0, q)\big)$ tends in distribution to the same limit variable as in the theorem. Thus the Bayesian and non-Bayesian estimation procedures of the parameter merge asymptotically. In the non-parametric situation that the distribution H of the observations is completely unknown, the empirical distribution $\mathbb{H}_n$ is asymptotically efficient for estimating H_0. Because efficiency is retained under the delta-method (van der Vaart (1991a)), the estimator $\chi(\mathbb{H}_n, q)$ is asymptotically efficient for estimating the parameter of interest $\chi(H,q) = Pg$. This implies that the posterior distribution of $\chi(H,q)$ has minimal concentration as well. We give an explicit derivation of the lower bound theory in Appendix B.

However, these results refer to the situation that the function q is known, while our purpose here is sensitivity analysis. The unknown function q can be equipped with a prior as well and be incorporated in a joint Bayesian analysis. This prior would ordinarily be elicited from experts in the field of research. In the present setup this will typically lead to mixing the limit distribution over the prior, as we now argue.

If the sensitivity parameter is specified to be a priori independent from H, and we maintain that $X|\,(H,q) \sim H$, then it follows that $(X_1, \ldots, X_n) \perp\!\!\!\perp q|\, H$, and hence $q|\, X_1, \ldots, X_n, H \sim q|\, H \sim q$. Thus the data provide no information about q and by conditioning on q, the posterior distribution of the parameter can be seen to satisfy

$$\mathcal{L}\big(\chi(H,q)|\, X_1, \ldots, X_n\big) = \int \mathcal{L}\big(\chi(H,q)|\, X_1, \ldots, X_n, q\big)\, d\Pi(q). \tag{3.3}$$

Here the integrand is the distribution considered in the preceding theorem and Π is the prior of q. We conclude that putting a prior on the sensitivity function leads simply to averaging the inferences for fixed sensitivity function with respect to the prior. At first sight this seems natural, but it is special to the present prior modelling, as we shall see in the next section.

In any case, given a prior on the sensitivity function, the distribution of the functional will not be asymptotically normal, but a mixture of normal distributions. This is a consequence of the fact that the sensitivity function is not identifiable from the data, so that its posterior does not contract to a Dirac measure.

We finish this section with two more technical observations on non-parametric modelling with the Dirichlet process.

Remark 1. Although the Dirichlet process is the canonical non-parametric prior for the unknown distribution H, in the present case it induces a possibly undesirable relation between the priors on the probability $1-p = H\{*\}$ of a missing observation and the prior on $P_1 = H_{|\mathfrak{Y}}/p$. If $H \sim \mathrm{DP}(a)$, then $1-p \sim B\big(a\{*\}, a(\mathfrak{Y})\big)$ and is independent of $P_1 \sim \mathrm{DP}(a_{|\mathfrak{Y}})$, where $B(\alpha,\beta)$ denotes the Beta distribution and is to be interpreted as a point mass at 0 if $0 = \alpha < \beta$ (see Ghosal and Van der Vaart (2017), Theorem 4.5). Thus the specification $H \sim \mathrm{DP}(a)$ with a single base measure a links the parameters of the Beta prior for p through the 'prior precision' $a(\mathfrak{Y} \cup \{*\})$ of the base measure. An alternative would be to combine a general beta prior on p with a general Dirichlet process prior on P_1. It can be verified that this does not change the asymptotic distribution of the posterior distribution, although it does change the finite sample behaviour.

Remark 2. Choosing a base measure a in $H \sim \mathrm{DP}(a)$ without an atom at the point $*$ leads to the Dirac distribution at 0 as a prior on $H\{*\} = 1-p$. In the preceding theorem the negative effect of this extreme prior is not seen, because the theorem considers the standard $\mathrm{DP}(a+n\mathbb{H}_n)$-posterior distribution. However, this is only a *version* of the posterior distribution, while a posterior distribution is only unique up to null sets under the Bayesian marginal distribution of the data. If $X \,|\, P \sim P$ and $P \sim DP(a)$, then $X \sim a$. Therefore, in the present case non-uniqueness of the posterior arises for a set of positive frequentist probability if $a\{*\} = 0 < H_0\{*\}$. Since it is likely that $H_0\{*\} > 0$, it is reasonable to put a prior $H \sim \mathrm{DP}(a)$ with $a\{*\} > 0$.

4. Parametrisation by η, P

The distribution P of the full data variable Y is the parameter of prime interest, and is intuitively more fundamental than the distribution H of the observed data variable X. Therefore, as is also argued in Scharfstein, Daniels and Robins (2003), it is reasonable to put a Dirichlet process prior on P rather than H. To obtain a full prior on the observation model, we combine this with an independent prior on the parameter (η, q). The resulting posterior distribution is more

complicated than the one in the preceding section, but can, for fixed (η, q), be studied using the theory of completely random measures.

Also in the parametrisation by the triplet (η, P, q), the sensitivity parameter q is unidentifiable from the observed data. Nevertheless, due to the functional relations between the parameters, the posterior distribution of q under the present prior modelling will be dependent on the data, even if q is modelled a priori independent of (η, P). We discuss this intriguing situation further at the end of the section, after first analysing the posterior distribution for fixed q.

In view of (2.5) and (2.3), we have $dP = p(1+e^{\eta+q})\,dP_1$. This can be inverted to give

$$P_1(A) = \frac{1}{p}\int_A \frac{1}{1+e^{\eta+q}}\,dP, \qquad A \in \mathcal{Y}. \tag{4.1}$$

Here the parameter $p = \Pr(R = 1)$ acts as the normalising constant to the probability distribution on the right and can be expressed in (η, P, q) as

$$p = \int \frac{1}{1+e^{\eta+q}}\,dP. \tag{4.2}$$

The distribution P_1 is the distribution of the observed outcome X. In the following lemma we show that a Dirichlet process prior on P leads, for fixed (η, q), to a normalised extended gamma process prior on P_1 (Dykstra and Laud (1981)). We briefly review this terminology, referring to Kingman (1967, 1975) or Appendix J of Ghosal and Van der Vaart (2017) for extended discussion.

A completely random measure is a random element U taking values in the set of Borel measures on a Polish space $\mathfrak{Y}$ such that $U(A_1), \ldots, U(A_k)$ are independent random variables, for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{Y}$. Every such measure can be represented as the sum of a deterministic measure and a measure of the form

$$U(A) = \int_A \int_0^\infty s\, N^c(dy, ds) + \sum_{j=1}^\infty U_j \delta_{y_j}(A), \tag{4.3}$$

where N^c is a Poisson process on $\mathfrak{Y} \times (0, \infty]$, the variables $U_1, U_2, \ldots$ are arbitrary independent, nonnegative variables independent of N^c, and $y_1, y_2, \ldots$ is an arbitrary sequence of elements of $\mathfrak{Y}$. The weights U_i and locations y_i together form the 'fixed atoms' of U. The distribution of U is determined by the intensity measure of N^c and the distributions and locations of the fixed atoms, which we denote by

$$\begin{aligned} \nu^c(A \times B) &= \mathbb{E}N^c(A \times B), \\ \nu^d(\{y_j\} \times B) &= \Pr(U_j \in B). \end{aligned}$$

We can think of ν^d as another measure on $\mathfrak{Y} \times (0, \infty]$ concentrated on the union $\cup_j \{y_j\} \times [0, \infty]$. The pair (ν^c, ν^d), or equivalently the sum $\nu^c + \nu^d$, is known as the intensity measure of U.

If $0 < U(\mathfrak{Y}) < \infty$ almost surely, then U can be normalised to the random probability measure $U/U(\mathfrak{Y})$, which can serve as a prior for a probability measure. The posterior distribution of this measure given an i.i.d. sample $X_1, \ldots, X_n$ from the prior $U/U(\mathfrak{Y})$, can be characterised as a mixture of normalised, completely random measures (see James, Lijoi and Prünster (2009); Lijoi and Prünster (2010) or Theorem 14.56 in Ghosal and Van der Vaart (2017)).

The Dirichlet process is a normalised completely random measure derived from the Gamma process U. If the process $\big(U(A) : A \in \mathcal{Y}\big)$ consists of random variables such that $U(A_1), \ldots, U(A_k)$ are independent variables with $U(A_i) \sim \mathrm{Ga}\,(a(A_i), 1)$, for every i and every finite collection of disjoint sets $A_1, \ldots, A_k \in \mathcal{Y}$, then $P = U/U(\mathfrak{Y})$ follows the Dirichlet process $\mathrm{DP}(a)$. In this case, in view of (4.1),

$$P_1(A) = \frac{\int_A \frac{1}{1+e^{\eta+q}}\, dU}{\int \frac{1}{1+e^{\eta+q}}\, dU}, \qquad A \in \mathcal{Y}. \tag{4.4}$$

This exhibits P_1 also as a normalised completely random measure. Its intensity measure can be obtained from the intensity measure of the gamma process. This process is a generalisation of the extended gamma process on $\mathbb{R}^+$.

Lemma 4.1. *If $P \sim DP(a)$ for an atomless finite Borel measure a on $\mathfrak{Y}$, then the measure P_1 given in (4.1) is a normalised completely random measure with intensity measure given by $\nu^d = 0$ and*

$$\nu^c(dy, ds) = s^{-1} e^{-s\left(1+e^{\eta+q(y)}\right)}\, ds\, da(y). \tag{4.5}$$

Proof. The gamma process U has no fixed atoms and continuous intensity measure $\nu^c(dy, ds) = s^{-1}e^{-s}\, ds\, da(y)$. Consider the completely random measure defined by $P_1'(A) = \int_A 1/\left(1 + e^{\eta+q}\right)\, dU$. For any non-negative measurable function $f : \mathfrak{Y} \to \mathbb{R}$, we have that $\int f\, dP_1' = \int f/\left(1 + e^{\eta+q}\right)\, dU$, as $P_1' \ll U$ with Radon-Nikodym derivative $1/\left(1 + e^{\eta+q}\right)$. As in (J.6) on p. 599 of Ghosal and Van der Vaart (2017), the minus log Laplace transform of P_1' is then given by

$$\begin{aligned} -\log \mathbb{E}\left(e^{-\theta \int f dP_1'}\right) &= -\log \mathbb{E}\left(e^{-\theta \int \frac{f}{1+e^{\eta+q}} dU}\right) \\ &= \int\int_0^\infty \left(1 - e^{-\theta s \frac{f(y)}{1+e^{\eta+q(y)}}}\right) s^{-1} e^{-s}\, ds\, da(y). \end{aligned}$$

To identify the intensity measure ν' of P_1', the right side must be written in the form $\int\int_0^\infty \left(1 - e^{-\theta s f(y)}\right)\, \nu'(dy, ds)$. By the substitution $s' = s/(1 + e^{\eta+q(y)})$, we obtain

$$\begin{aligned} &\int\int_0^\infty \left(1 - e^{-\theta s \frac{f(y)}{1+e^{\eta+q(y)}}}\right) s^{-1} e^{-s}\, ds\, da(y) \\ &\qquad = \int\int_0^\infty \left(1 - e^{-\theta s' f(y)}\right) (s')^{-1} e^{-s'\left(1+e^{\eta+q(y)}\right)}\, ds'\, da(y). \end{aligned}$$

As the Laplace functional determines the completely random measure, the proof is complete. □

Remark 3. Unlike in the preceding section, the base measure a of the DP(a)-prior is now a measure on $\mathfrak{Y}$, and not on $\mathcal{X} = \mathfrak{Y} \cup \{*\}$. To apply the theory on normalised completely random measures, it will be convenient to take it without atoms. This is natural in this situation, unlike in the preceding section, where an atom at the special point $\{*\}$ is natural, as noted in Remark 2. (Priors with point masses are considered in Canale et al. (2023).)

The non-missing observations Y_i (the values X_i with $R_i = 1$, or equivalently $X_i \neq *$) form a random sample from the distribution P_1, of random size

$$N_n = \sum_{i=1}^{n} R_i.$$

The variable N_n follows a Binomial (n, p) distribution, and will tend to infinity with n. In the Bayesian setup with a Dirichlet DP(a) prior on P, the data may be thought of as having been generated by the Bayesian hierarchy (with every step conditioned on the outcomes of all preceding steps):

prior:
- generate (η, q) from a prior.
- generate P_1 as a normalised completely random measure with intensity measure (4.5).
- compute $p = 1/\int (1 + e^{\eta+q})\, dP_1$.

data:
- generate $N_n \sim \text{binom}(n, p)$.
- generate $\bar{X}_1, \ldots, \bar{X}_{N_n}$ i.i.d. from P_1.
- randomly permute $\bar{X}_1, \ldots, \bar{X}_{N_n}$ and $n - N_n$ copies of $*$ to define $X_1, \ldots, X_n$.

If we also compute P by inverting (4.1), then the variables $(\eta, q, p, P, X_1, \ldots, X_n)$ will have the same joint distribution as in the scheme with prior $(\eta, q) \perp\!\!\!\perp P \sim$ DP(a). Therefore, the resulting posterior distributions of P or P_1 given (η, q, p) and $(X_1, \ldots, X_n)$ are the same.

In the Bayesian hierarchy the conditional distribution of P_1 given $N_n = k, \bar{X}_1, \ldots, \bar{X}_k, \eta, q, p$ is the same as the posterior distribution of P_1 based on a sample of size k from the normalised completely random measure with intensity measure (4.5). The asymptotics of this posterior distribution are obtained in Theorem 7.1 and imply the following theorem.

The theorem gives the limit of the posterior distribution under the assumption that the observations are an i.i.d. sample from the distribution of (R, X) given by $\Pr(R = 1) = p_0$ and $X|\, R = 1 \sim P_{1,0}$, for given $(p_0, P_{1,0})$, or equivalently $X_1, \ldots, X_n$ are i.i.d. copies of $X \sim H_0$, where $H_0 = (1 - p_0)\delta_* + p_0 P_{1,0}$, as in Theorem 3.1. Denote the empirical distribution of the observed outcomes by $\mathbb{P}_{n,1} = N_n^{-1} \sum_{i=1}^{n} \delta_{X_i} 1_{R_i=1}$.

Theorem 4.2 (Functional Bernstein-von Mises P_1)**.** *The posterior distribution of P_1 under the prior $P \sim DP(a)$ independent of (η, q) satisfies, H_0^∞-almost surely,*

$$\sqrt{n}(P_1 - \mathbb{P}_{n,1})|\, X_1, \ldots, X_n, \eta, q, p \rightsquigarrow \sqrt{\frac{1}{p_0}}\, \mathbb{B}_{P_{1,0}},$$

uniformly in η and q that are uniformly bounded above.

Proof. Given $X_1, \ldots, X_n$, define N_n as the number of X_i unequal to $*$ and let $\bar{X}_1, \ldots, \bar{X}_{N_n}$ be the X_i that are not equal to $*$. The variables $N_n, \bar{X}_1, \ldots, \bar{X}_{N_n}$ represent the same information as $X_1, \ldots, X_n$, apart from ordering. Thus for any measurable function h,

$$\begin{aligned}\mathbb{E}h\big(\sqrt{n}(P_1 - \mathbb{P}_{n,1})\big|\, X_1, \ldots, X_n, \eta, q, p\big) \\ = \mathbb{E}h\big(\sqrt{n}(P_1 - \mathbb{P}_{n,1})\big|\, N_n = k, \bar{X}_1, \ldots, \bar{X}_k, \eta, q, p\big).\end{aligned}$$

The Bayesian hierarchy shows that given η, q, p, the variable N_n follows a binomial distribution with probability p, while given $N_k = k, \eta, q, p$ the variables $\bar{X}_1, \ldots, \bar{X}_k$ are an i.i.d. sample from the distribution P_1, which follows an extended gamma normalised completely random measure without fixed atoms and continuous intensity given by (4.5). Thus the distribution of P_1 given $N_n = k, \bar{X}_1, \ldots, \bar{X}_k, \eta, q, p$ in the present notation is the same as the distribution in the notation of Section 7 of P_k given $X_1, \ldots, X_k$, with b in the latter section taken equal to $b = 1 + e^{\eta+q}$ and P_0 in the latter section taken equal to the present $P_{1,0}$.

By the law of large numbers $N_n/n \to p_0$, almost surely. Hence the theorem follows from Theorem 7.1. □

For posterior inference on the functional of interest, we need the posterior distribution of p next to the one of P_1. The preceding theorem shows that P_1 is asymptotically independent of p under the posterior distribution. Because the parameters p and P_1 are related under the prior, this is not true for finite n, and it appears that the posterior distribution of p may depend on the values $\bar{X}_1, \ldots, \bar{X}_{N_n}$ generated later in the Bayesian hierarchy, besides on the variable N_n. The asymptotic independence suggests that little is lost by ignoring $\bar{X}_1, \ldots, \bar{X}_{N_n}$ in posterior inference on p and base this on the observational model $N_n \sim \text{binom}(n, p)$ only. This so-called *cut-posterior* Jacob et al. (2017); Moss and Rousseau (2022) (which cuts out the later data) will satisfy the usual Bernstein-von Mises theorem for the binomial distribution.

Lemma 4.3. *Suppose that the prior of η possesses a bounded, continuous density on a compact interval in $\mathbb{R}$ and that $|q|$ is uniformly bounded. Then the posterior distribution of p based on the observation $N_n \sim \text{binom}(n, p)$ and prior on p induced by $P \sim DP(a) \perp\!\!\!\perp \eta$, almost surely under p_0,*

$$\sqrt{n}\Big(p - \frac{N_n}{n}\Big)\big|\, N_n, q \rightsquigarrow N\big(0, p_0(1-p_0)\big).$$

Proof. The lemma follows from the classical Bernstein-von Mises theorem (e.g. Van der Vaart (1998), Chapter 10), provided that the prior density of p is continuous.

By (4.2) $p = \int \Psi(\eta + q)\, dP =: \phi_P(\eta)$, where $\Psi(v) = 1/(1 + e^v)$ is the logistic function, and $P \sim \text{DP}(a)$ is independent of η. Because the function $\eta \mapsto \phi_P(\eta)$

is strictly increasing, it follows that the cumulative distribution function of p can be written

$$\Pr\big(\phi_P(\eta) \le t\big) = \mathbb{E}_P \int_0^{\phi_P^{-1}(t)} h(\eta)\, d\eta,$$

for h the density of η. The function Ψ is differentiable with derivative $\phi_P'(\eta) = \int \psi(\eta + q)\, dP$, where $\psi(v) = \Psi(v)\big(1 - \Psi(v)\big)$ is bounded away from zero if v is restricted to a compact set in $\mathbb{R}$. It follows that the cumulative distribution function in the preceding display is continuously differentiable with derivative $t \mapsto \mathbb{E}_P h \circ \phi_P^{-1}(t)/\phi_P' \circ \phi_P^{-1}(t)$. □

Remark 4. The Bayesian hierarchy shows that $\bar{X}_1, \ldots, \bar{X}_{N_n} \perp\!\!\!\perp p \,|\, N_n, P_1, q$, whence $p \,|\, (\bar{X}_1, \ldots, \bar{X}_{N_n}, N_n, P_1, q) \sim p \,|\, N_n, P_1, q$. Thus the posterior distribution of p given the full data and P_1, q is obtainable from the observational model $N_n \,|\, p, P_1, q \sim \text{binom}(n, p)$ with prior $p \,|\, P_1, q$. If the latter conditional prior is sufficiently smooth, then the posterior distribution of p given the full data and P_1, q will satisfy the ordinary binomial Bernstein-von Mises theorem, where the Gaussian approximation will be independent of the prior and hence coincide with the cut-posterior in the preceding lemma. We conjecture that this is true, but verifying the required smoothness seems not trivial due to the infinite-dimensional nature of P_1.

The parameter of interest can be expressed in the parameters (P_1, η, p, q) as

$$\mathbb{E}_P g(Y) = (1-p)\frac{P_1(ge^q)}{P_1 e^q} + pP_1 g =: \kappa(p, P_1, q).$$

Therefore, combining the preceding theorem and lemma gives the following Bernstein-von Mises theorem for the parameter of interest.

Theorem 4.4 (Bernstein-von Mises η, P)**.** *If P_1 is as in Theorem 4.2 and p is as in Lemma 4.3, then, conditional on q, H_0^∞-a.s.,*

$$\sqrt{n}\big(\kappa(p, P_1, q) - \kappa(N_n/n, \mathbb{P}_{1,n}, q)\big) \,|\, X_1, \ldots, X_n, q \rightsquigarrow Z_{H_0},$$

where the limit variable Z_{H_0} has the same distribution as the limit in Theorem 3.1.

Proof. The functional can be written $\kappa(p, P_1, q) = \phi\big(p, P_1(ge^q), P_1 e^q, P_1 g\big)$, for the function $\phi : \mathbb{R}^4 \to \mathbb{R}$, defined by $\phi(x_1, x_2, x_3, x_4) = (1 - x_1)x_2/x_3 + x_1 x_4$. The theorem therefore can be derived with the help of the delta-method for conditional distributions (see Van der Vaart and Wellner (1996), Section 3.9.3, or more precisely Van der Vaart and Wellner (2023), Theorem 3.10.13) combined with the results of Theorem 4.2 and Lemma 4.3. The limit variable Z_{H_0} arises as the inner product

$$Z_{H_0} = \nabla\phi(\theta_0) \cdot \Big(Z_0, \sqrt{\frac{1}{p_0}}\mathbb{B}_{P_{1,0}}(ge^q), \sqrt{\frac{1}{p_0}}\mathbb{B}_{P_{1,0}} e^q, \sqrt{\frac{1}{p_0}}\mathbb{B}_{P_{1,0}} g\Big),$$

for $Z_0 \sim N\big(0, p_0(1-p_0)\big)$ independent of the Brownian bridge $\mathbb{B}_{P_{1,0}}$ and $\nabla\phi(\theta_0)$ the gradient of ϕ evaluated at the vector $\theta_0 = \big(p_0, P_{1,0}(ge^q), P_{1,0}e^q, P_{1,0}g\big)$. We

can check by explicit calculation of the variance that Z_{H_0} possesses the same centred normal distribution as the limit variable in Theorem 3.1. (Alternatively, we can apply Corollary 4.1 below to see that the posterior distributions of $\big(H\{*\}, H_{|\mathcal{Y}}/H(\mathfrak{Y})\big)$ and $(1-p, P_1)$ in the two theorems are the same, together with the identity $(1-N_n/n)\delta_* + (N_n/n)\mathbb{P}_{1,n} = \mathbb{H}_n$.) □

Corollary 4.1. *If P_1 is as in Theorem 4.2 and p is as in Lemma 4.3 and $H = (1-p)\delta_* + pP_1$, then the sequence $\sqrt{n}\big(H - \mathbb{H}_n\big)$ has the same asymptotic conditional distribution given $X_1, \ldots, X_n$ as the sequence $\sqrt{n}(H - \mathbb{H}_n)$ in Theorem 3.1.*

Proof. Since $\mathbb{H}_n = (1-N_n/n)\delta_* + (N_n/n)\mathbb{P}_{n,1}$, it follows that $H = \phi(p, P_1)$ and $\mathbb{H}_n = \phi(N_n/n, \mathbb{P}_{1,n})$, for the map $\phi : [0,1] \times \ell^\infty(\mathcal{F}) \to \ell^\infty(\mathcal{F})$ given by $\phi(s,Q) = (1-s)\delta_* + sQ$. By Lemma 4.3 and Theorem 4.2, the sequence $\sqrt{n}(p - N_n/n, P_1 - \mathbb{P}_{1,n})$ converges in distribution given $X_1, \ldots, X_n$ in the space $[0,1] \times \ell^\infty(\mathcal{F})$ to the pair $(Z_0, \mathbb{B}_{P_{1,0}})$ of a normal variable $Z_0 \sim N\big(0, p_0(1-p_0)\big)$ and an independent $P_{1,0}$-Brownian bridge process, almost surely. Thus the delta-method for conditional distributions (Van der Vaart and Wellner, 1996, p. 511) gives that, almost surely,

$$\sqrt{n}(H - \mathbb{H}_n) \,|\, X_1, \ldots, X_n \rightsquigarrow \phi'_{p_0, P_{1,0}}\Big(Z_0, \sqrt{\frac{1}{p_0}}\mathbb{B}_{P_{1,0}}\Big),$$

where the derivative is given by $\phi'_{p_0,P_{1,0}}(g,G) = -g\delta_* + gP_{1,0} + p_0 G$. The limit process $f \mapsto Z_0\big(-f(*) + P_{1,0}f\big) + \sqrt{p_0}\mathbb{B}_{P_{1,0}}f$ is centred Gaussian and linear in f with variance function $p_0(1-p_0)\big(P_{1,0}f - f(*)\big)^2 + p_0\big(P_{1,0}f^2 - (P_{1,0}f)^2\big)$. It can be verified that this is equal to $\operatorname{var}\mathbb{B}_{H_0}f$ for a $\mathbb{B}_{H_0}$-Brownian bridge process, which is the limit process in Theorem 3.1. □

Theorem 4.4 gives the posterior distribution of the parameter of interest given the sensitivity function q. Suppose that we also equip the sensitivity function with a prior. In practical situations this might be an informative prior based on substantive knowledge. If the parameter q is independent of (η, P) under the prior, then q will *not* be independent of the distribution $H = (1-p)\delta_* + pP_1$ of the observations. We can write

$$\begin{aligned}\mathcal{L}\big(\kappa(p, P_1, q) \,|\, X_1, \ldots, X_n\big) &= \int \mathcal{L}\big(\kappa(p, P_1, q) \,|\, X_1, \ldots, X_n, q\big)\, d\Pi(q \,|\, X_1, \ldots, X_n) \\ &= \iint \mathcal{L}\big(\kappa(p, P_1, q) \,|\, X_1, \ldots, X_n, q\big)\, d\Pi(q \,|\, H)\, d\Pi(H \,|\, X_1, \ldots, X_n).\end{aligned}$$

In the second step we use that $q \,|\, H, X_1, \ldots, X_n \sim q \,|\, H$, since $X_1, \ldots, X_n \perp\!\!\!\perp q \,|\, H$, as follows from the fact that $X_1, \ldots, X_n \,|\, H, q, \eta \overset{\text{iid}}{\sim} H$. By Corollary 4.1 the posterior distribution $H \,|\, X_1 \ldots, X_n$ satisfies the same Bernstein-von Mises theorem as in Theorem 3.1. In particular, the posterior distribution of H will concentrate near $\mathbb{H}_n$ (equivalently near H_0). Thus informally the preceding display approximates to

$$\int \mathcal{L}\big(\kappa(p, P_1, q) \,|\, X_1, \ldots, X_n, q\big)\, d\Pi(q \,|\, \mathbb{H}_n). \tag{4.6}$$

Thus the observations influence the posterior distribution of q through the prior dependence between q and H. In the next section we compare this to the result of the preceding section.

5. Comparison of the two parametrisations

While the posterior distributions of the parameter of interest given q is the same for the two prior setups in Sections 3 and 4, the distributions differ when q is also (independently) equipped with a prior.

The asymptotic posterior distributions of the parameter for the two priors are given in (3.3) and (4.6). Both expressions can be informally written in the form

$$\chi(H,q)|\, X_1,\ldots,X_n \sim W_n + \frac{1}{\sqrt{n}} Z_n,$$

where the variables on the right satisfy informally, asymptotically, for a latent 'posterior variable' q,

$$W_n \perp\!\!\!\perp Z_n|\, q, \quad W_n = \chi(\mathbb{H}_n, q) \quad Z_n|\, q \sim N(0, \sigma^2_{H_0}),$$

for $\sigma^2_{H_0}$ the variance of the limit variable in Theorem 3.1. The two prior settings differ in the distributions of the latent variable q:

- When using the prior on H as in Section 3, the sensitivity function q is distributed according to its prior.
- When using the prior on (η, P) as in Section 4, the sensitivity function is distributed according to $q|\, H = \mathbb{H}_n$, where $(q, H) \sim (q, (1-p)\delta_* + pP_1)$.

Thus when using the first prior, the inference on the functional is simply mixed over the various values of the sensitivity function, weighted by its prior. When using the second prior, the data will influence our ideas about the sensitivity function.

Note that this is true even though in both cases the sensitivity function is not identifiable from the data. The non-identifiability makes that for both priors the sensitivity function remains a random object, even if the number of observations tends to infinity: its posterior distribution does not contract to a degenerate distribution. However, the two posterior distributions differ. In the first case the posterior is equal to the prior, while in the second it is adapted to the data through the estimation of H.

One can debate which of the two situations is more desirable. From a non-parametric Bayesian view, it seems natural to place the Dirichlet process prior, which in many ways is *the* nonparametric prior, comparable to the empirical distribution, on the most fundamental distribution in the problem. This seems to favour the second prior, the one which makes the posterior distribution learn something about the sensitivity parameter from the data.

From an applied point of view, the choice of prior may be decided based on subject matter experts' beliefs about q in relation to (p, P_1). Only if experts' ideas about the magnitude of the selection bias are unwavering, no matter the

observed values of X, does the first prior seem to be preferable. The second prior is the better choice if experts' beliefs about the extent of selection bias will change if observed outcomes turn out to be higher or lower than expected, or if fewer or more subjects drop out than expected. Scharfstein, Daniels and Robins (2003) provide an example of the latter situation in the context of HIV research. The desired dependence between q and (p, P_1) can then be expressed in the second prior. Since η is a function of q, p and P_1, and p and P_1 can be identified from the observed data, a prior belief on η induces a posterior belief on q which will depend on the data. To set the priors on η and q in practice, relationship (2.4) forms a useful basis for discussion. For several values of η and functions q, the prior expectations of the conditional probability of drop-out can be shown for a realistic range of values for Y. In this process the priors of η and q are intimately linked. As it can be difficult to provide a prior on η, it is interesting to investigate the consequences of an uninformative prior on η, which is the setting of the simulation study below.

The two prior schemes also differ in ease of implementation. The second scheme clearly requires more computational effort.

6. Simulations

In this section we present the results of a simulation study, illustrating the theoretical results of the preceding sections, in particular when a prior is placed on the sensitivity parameter. We focus on the setting where the prior on α is misspecified, as this is where we expect to see differences between the two parametrisations. The code can be found on Gitlab.

We consider a model with sample space $\mathfrak{Y} = (0, \infty)$ and sensitivity function $q(y) = \alpha(\log y - c)$, for $\alpha, c \in \mathbb{R}$. When $c = 0$, the function is equal to the choice in Scharfstein, Daniels and Robins (2003). The reason we introduce this c is to potentially center the function q, meaning that $\mathbb{E}_{P^0} q = 0$. Ideally this will make η and q independent, as the η will not contain the'intercept' of q. More information on c can be found below in Section 6.3. The functional of interest is the mean outcome $\mathbb{E}_P Y$. We take $* = 0$, and therefore the observations can be written as $X = RY \in [0, \infty)$.

6.1. Sampling scheme for the (P_1, p, α) parametrisation

A DP(a) prior on $H = (1-p)\delta_0 + pP_1$ is combined with an independent prior on α. As the observed data $X_1, \ldots X_n$ is a random sample from H, the posterior distribution of H again is a Dirichlet process, but with base measure $a + n\mathbb{H}_n$, where $\mathbb{H}_n = n^{-1} \sum_{i=1}^n \delta_{X_i} = (1 - N_n/n)\delta_0 + n^{-1} \sum_{i=1}^n R_i \delta_{X_i}$. Furthermore, as discussed in Section 3, see (3.3), the posterior distribution of α is equal to its prior, and the two parameters H and α remain independent.

The functional of interest is written as a function of H and α by equation (3.2), with q the function $q(y) = \alpha(\log y - c)$. The posterior distribution of the functional is obtained by inserting the DP($a + n\mathbb{H}_n$)-distribution for H and

the prior distribution for α. We approximated the Dirichlet process through its stick-breaking representation (Sethuraman (1994), or Ghosal and Van der Vaart (2017), Theorem 4.12), where we truncated the infinite sum such that the retained weights summed numerically to one. We sampled 10000 times from the posterior of $(H,\ \alpha)$ and approximated each expectation in $\chi(H,\alpha)$ by the average of the samples.

6.2. Sampling scheme for the $(P,\ \eta,\ \alpha)$ parametrisation

A DP(a) prior on P is combined with independent priors on η and α. In this case the posterior distribution given the observed data does not possess a simple form, but the posterior distribution given the full data does. Following a similar approach to Scharfstein, Daniels and Robins (2003), we implement a Gibbs sampling scheme, which fills in the missing data.

The full data can be generated through the Bayesian hierarchy, for Ψ the logistic function (see (2.4)):

- $P \sim \mathrm{DP}(a) \perp\!\!\!\perp (\eta,\alpha)$.
- $Y_1 \ldots, Y_n | \, P,\eta,\alpha \overset{\text{iid}}{\sim} P$.
- $R_1,\ldots,R_n | \, Y_1,\ldots,Y_n,P,\alpha,\eta \overset{\text{ind}}{\sim} \mathrm{Bernoulli}\big(\Psi(-\eta-\alpha(\log Y_i - c))\big)$.

This hierarchy shows that $P \perp\!\!\!\perp (\eta,\alpha,R_1,\ldots,R_n) | \, (Y_1,\ldots,Y_n)$. This implies that the posterior distribution of P given the full data $(Y_1,\ldots,Y_n,R_1,\ldots,R_n)$ is the same as its posterior distribution given $Y_1,\ldots,Y_n$, which is $\mathrm{DP}(a+\sum_{i=1}^n \delta_{Y_i})$. Moreover, this implies $P \perp\!\!\!\perp (\eta,\alpha) | \, (Y_1,\ldots,Y_n,R_1,\ldots,R_n)$, so that the posterior distributions of P and (η,α) given the full data are independent. Finally, the second posterior distribution $(\eta,\alpha) | \, (Y_1,\ldots,Y_n,R_1,\ldots,R_n)$ is recognized as the posterior distribution of the parameters (η,α) in the logistic linear regression model regressing $R_1,\ldots,R_n$ on the independent variables $\log Y_i$ with intercept $-\eta$, slope $-\alpha$.

The observed data consists of $R_1,\ldots,R_n$ and the $N_n = \sum_{i=1}^n R_i$ values Y_i with $R_i = 1$. Given the parameters (P,η,α), the $n-N_n$ missing values $(Y_i : R_i = 0)$ are distributed according to the measure P_0 given by (combine equations (2.2) and (4.1)):

$$P_0(A) = \frac{\int_A \frac{e^q}{1+e^{\eta+q}}\,dP}{\int \frac{e^q}{1+e^{\eta+q}}\,dP}, \qquad A \in \mathcal{Y}. \tag{6.1}$$

Furthermore, given (P,η,α), the missing values are conditionally independent of each other and of the observed data.

These observations lead to the following Gibbs sampling scheme, which augments the observed data to full data in the first step and next updates the parameters P and (η,α) according to the full data posterior in the second and third steps. In each step we take the observed data as given, as well as the outcomes of the other two steps.

1. Sample $n-N_n$ values from P_0, assign them to the (missing) values $(Y_i : R_i = 0)$.

2. Sample $P \sim \mathrm{DP}(a + \sum_{i=1}^{n} \delta_{Y_i})$.
3. Sample (η, α) from its posterior distribution given $R_1, \ldots, R_n, Y_1, \ldots, Y_n$ in logistic regression model $R_1, \ldots, R_n$ on $-\eta - \alpha\,(\log Y_1 - c), \ldots, -\eta - \alpha\,(\log Y_n - c)$.

The three steps are repeated until convergence, giving a sample P in each iteration. The corresponding mean values $\int y\,dP(y)$, after burn-in, form an approximate sample from the posterior distribution of the functional of interest. The third step of the algorithm was carried out by the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). We used a random walk proposal distribution with steps generated from the normal distribution with mean zero and a given variance. The latter variance was set such that the probability of accepting a new draw was approximately 0.234 (Gelman, Gilks and Roberts, 1997). The acceptance probabilities are computed using the likelihood of the logistic regression model, given by

$$\prod_{i=1}^{n} \left(\frac{1}{1 + e^{\eta + \alpha(\log Y_i - c)}} \right)^{R_i} \left(\frac{e^{\eta + \alpha(\log Y_i - c)}}{1 + e^{\eta + \alpha(\log Y_i - c)}} \right)^{1 - R_i}.$$

6.3. Simulation settings

First, we simulate data under the assumption that the true data distribution belongs to the sensitivity model. Specifically, we fix a true data distribution through choosing $P_1^0 = \mathrm{Gamma}(r, s)$, for some $r, s > 0$, a fraction $1 - p_0$ of missing observations and a sensitivity parameter α_0, where the superscript 0 indicates the truth. We reserve the subscript for denoting the conditional distributions. With the choice of q, the sensitivity model (2.2) imposes $dP_0^0(y) \propto e^{\alpha_0(\log y - c)}\,dP_1^0(y)$, resulting in $P_0^0 = \mathrm{Gamma}(r + \alpha_0, s)$. The full data distribution $P^0 = (1 - p_0)P_0^0 + p_0 P_1^0$ is a mixture of the two Gamma distributions, and the functional of interest is given by $\mathbb{E}_{P^0} Y = p_0 r/s + (1 - p_0)(r + \alpha_0)/s$. Relation (2.5) yields $\eta_0 = \log\left((1 - p_0)/p_0\right) + \alpha_0\,(\log s + c) - \log\left(\Gamma(r + \alpha_0)/\Gamma(r)\right)$.

We want to observe the performance of the model in a natural, but somewhat challenging setting by letting a substantial part of the data be missing and choosing α_0 such that there is a deviation from MCAR, while keeping the groups comparable. Specifically, we chose $r = 2$, $s = 1$, $p_0 = 0.6$, $\alpha_0 = 2$ and $c \approx 0.42$, giving $\eta_0 \approx -1.35$ and $\mathbb{E}_{P^0} Y = 2.8$. A default choice is $c = 0$, but a different interesting choice would be $c = \mathbb{E}_{P^0}[\log Y] = p_0\,(\psi(r) - \log(s)) + (1 - p_0)\,(\psi(r + \alpha_0) - \log(s))$, where ψ is the digamma function. Using this choice of c results in a centred q, ideally resulting in independent behaviour of α and η. In practice, data would only be available from P_1^0. Therefore another choice for c could be $\mathbb{E}_{P_1^0}[\log Y]$, which was the choice we made. Note that P_1^0, P_0^0 and the functional of interest do not depend on the choice of c. Choosing an oracle c had similar results.

The posterior distributions are agnostic of the above choices, and depend only on the priors chosen for the parameters. In the first prior scheme, we let

TABLE 1
Coverage of the 90%-credible intervals of the posterior distribution of the functional of interest for the different parametrisations with fixed $\alpha = \alpha_0$

	$n = 100$	$n = 1000$	$n = 10000$
	Coverage | Length	Coverage | Length	Coverage | Length
H	0.784 | 0.877	0.857 | 0.361	0.892 | 0.122
P, η	0.771 | 0.870	0.855 | 0.360	0.892 | 0.122

TABLE 2
Coverage of the 90%-credible intervals of the posterior distribution of the functional of interest for the different parametrisations with misspecified prior $\alpha \sim N(1, 0.25)$

	$n = 100$	$n = 1000$	$n = 10000$
	Coverage | Length	Coverage | Length	Coverage | Length
H, α	0.560 | 0.946	0.328 | 0.700	0.026 | 0.662
P, η, α	0.531 | 0.932	0.304 | 0.687	0.131 | 0.635

$H \sim \text{DP}(a)$, with $a(\{0\}) = 1 - p_0$ and $a_{|\mathbb{R}^+} = \text{Gamma}(2, 1)$. For the other parametrisation, we let $P \sim \text{DP}(\tilde{a})$, with $\tilde{a}$ a mixture between the true P_1^0 with probability p_0 and the true P_0^0 with probability $1 - p_0$. In both parametrisations the prior precision was taken to be equal to one, as to minimise the effect of the base measure. We put a normal prior on α with mean 1 and standard deviation 0.5 and in the second parametrisation a uniform prior on η with minimum -5 and maximum 2. By taking this interval relatively big we hope to investigate the effect of the data on the posterior of α. Our main interest lies in observing this effect in the P, η parametrisation, discussed in Section 5. We found that this effect is visible when the prior on α is misspecified: the mean of α is shifted from α_0. Although it should be expected that practitioners have reasonable insight in this choice, usually the α can't be correctly specified as it is not assumed that the sensitivity model is true.

6.4. Results

The results can be seen in Figure 1. We also calculated the coverage of the 90%-credible intervals. These intervals are defined by the 0.05 and 0.95 quantiles of the posteriors. The results for fixed α can be found in Table 1 and with prior on α in Table 2. Because of large run times, especially for $n = 10000$, the coverage results were obtained using the Delftblue supercomputer (Delft High Performance Computing Centre (DHPC), 2022).

From Figure 1 it can be seen that both methods perform similarly. When n increases the variance of the posteriors decrease, but they will never converge to a Dirac measure due to the uncertainty in α. This effect can be seen in Table 2, when compared to Table 1. The latter illustrates the identifiability of the model when $\alpha = \alpha_0$ is fixed. The posteriors converge to a Dirac and have proper uncertainty quantification in this case of fixed α, as the theory in this paper confirms. From Figure 1 and Table 2 it can be seen that the posteriors of the parametrisations asymptotically behave differently. The length of the credible intervals for the P, η parametrisation are slightly smaller than those for H, but the coverage

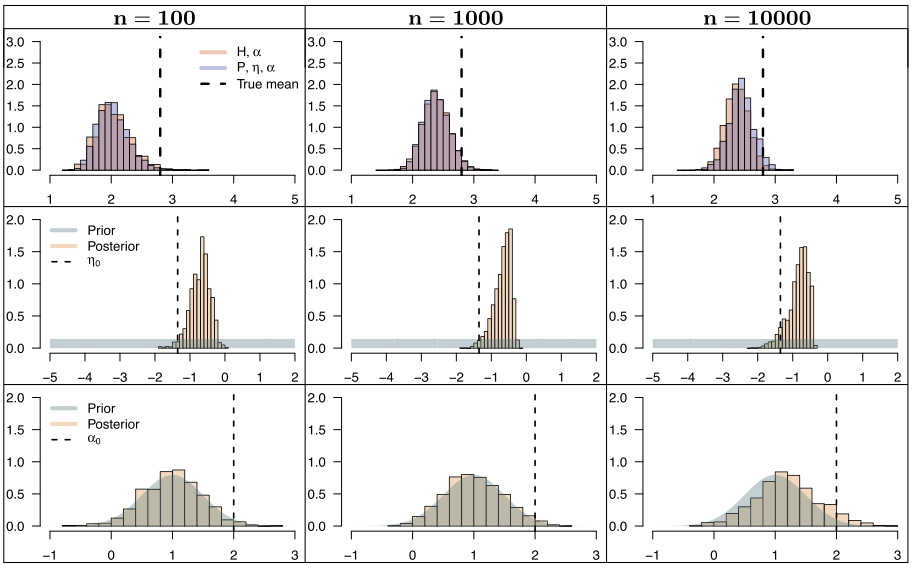

FIG 1. *Posterior distributions with a prior on* α. Row 1*: Histogram of draws of the posterior distribution of the functional of interest for the different parametrisations.* Row 2*: The prior and posterior of* η. Row 3*: The prior and posterior of* α.

is higher. This means the location of this posterior is slightly different than the posterior of the functional in the other parametrisation, which is illustrated in row 1 of Figure 1. In row 3, the posterior of α slightly shifts compared to the prior on α when n increases, indicating the influence of the added data. From the point of view of sensitivity analysis, the P, η parametrisation might therefore be preferable: the credible intervals more often contain the truth when the number of observations is large. If experts don't agree on the location of the prior on α, but do on the direction of the relationship between α, η and P, the data might point the model in the right direction. When α is correctly specified, both parametrisations perform well and give similar results, as can be seen in Appendix A.

7. Extended gamma posterior

In this section we study the posterior distribution for sampling from a prior equal to the extended gamma process. We are interested in the special case (4.1), but in this section adopt a general framework independent of the main paper. We prove a functional Bernstein-von Mises theorem and consistency of the posterior mean in the total variation norm.

For given measurable functions $b : \mathfrak{X} \to (0, \infty)$ and a finite, atomless Borel measure a on the Polish space $\mathfrak{X}$, let P_n be distributed as the normalised completely random measure without fixed atoms and continuous part with intensity

measure $\nu^c(dx, ds) = s^{-1}e^{-sb(x)}\, ds\, da(x)$. Suppose that P_n is used as a prior for the distribution of an i.i.d. sample of observations; hence $X_1, \ldots, X_n | P_n \overset{\text{iid}}{\sim} P_n$. In this section we derive the asymptotics of the conditional distribution of P_n given $X_1, \ldots, X_n$ (i.e. the posterior distribution). In the intended application the function b is equal to $b = (1 + e^{\eta+q})$.

We identify probability distributions P on $\mathfrak{X}$ with the stochastic processes $(Pf : f \in \mathcal{F})$, for a given collection of measurable maps $f : \mathfrak{X} \to \mathbb{R}$, and consider weak convergence of sequences of such processes in $\mathbb{R}^{\mathcal{F}}$ in the case of a finite collection, and more generally in the space $\ell^\infty(\mathcal{F})$ of bounded functions $z : \mathcal{F} \to \mathbb{R}$ equipped with the uniform norm $\|z\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |z(f)|$. Weak convergence of conditional distributions is understood in terms of the bounded Lipschitz metric: we say that $Z_n | X_n \rightsquigarrow Z$ in probability or almost surely if $\sup_{h\in \mathrm{BL}_1} \big|\mathbb{E}\big(h(Z_n)| X_n\big) - \mathbb{E}h(Z)\big| \to 0$, in probability or almost surely. Here BL_1 is the set of functions $h : \ell^\infty(\mathcal{F}) \to [-1, 1]$ such that $|h(z_1) - h(z_2)| \le \|z_1 - z_2\|_{\mathcal{F}}$, for every $z_1, z_2 \in \ell^\infty(\mathcal{F})$. Let $\mathbb{B}_{P_0}$ be a P_0-Brownian bridge process: a tight, Borel measurable, zero-mean Gaussian map into $\ell^\infty(\mathcal{F})$ with covariance function $\operatorname{cov}(\mathbb{B}_{P_0} f, \mathbb{B}_{P_0} g) = P_0(fg) - P_0 f P_0 g$.

Theorem 7.1 (Bernstein-von Mises for extended gamma processes)**.** *Let $\mathcal{F}$ be a P_0-Donsker class with measurable envelope function F such that $\int F\, da < \infty$, and let $b : \mathfrak{X} \to [\underline{b}, \infty)$ be a measurable function, for some constant $\underline{b} > 0$. If there exist $q, r > 0$ with $1/q + 1/r < 1/2$ such that $P_0 b^q + P_0 F^r < \infty$, then for P_0^∞-almost every sequence $X_1, X_2, \ldots$,*

$$\sqrt{n}\,(P_n - \mathbb{P}_n)\,|\,X_1, \ldots, X_n \rightsquigarrow \mathbb{B}_{P_0}, \qquad \text{in } \ell^\infty(\mathcal{F}).$$

The convergence is uniform in measurable functions $b : \mathfrak{X} \to [\underline{b}, \infty)$ such that $\max_{1\le i\le n} b(X_i) = o(n^{1/q})$ and $\mathbb{P}_n b - P_0 b \to 0$, almost surely, uniformly in b. If the latter uniform convergence is true in probability, then the statement is true in probability, uniformly in b.

Proof. Denote the distinct values in $X^{(n)} := (X_1, \ldots, X_n)$ by $\tilde{X}_1, \ldots, \tilde{X}_{K_n}$ and let $N_{1,n}, \ldots, N_{K_n,n}$ be their multiplicities. It was shown in James, Lijoi and Prünster (2009) (alternatively, see Theorem 14.56 of Ghosal and Van der Vaart (2017)), that the posterior distribution of P_n is a mixture over λ of the distributions of the normalised completely random measures $\Phi_\lambda/\Phi_\lambda(\mathfrak{X})$ with intensity measures

$$\nu^c_{\Phi_\Lambda|X^{(n)}}(dx, ds|\,\lambda) = s^{-1}e^{-s\left(\lambda+b(x)\right)}\, ds\, a(dx), \tag{7.1}$$

$$\nu^d_{\Phi_\Lambda|X^{(n)}}(\{\tilde{X}_j\}, s|\,\lambda) \propto s^{N_{j,n}-1}e^{-s\left(\lambda+b(\tilde{X}_j)\right)}\, ds, \tag{7.2}$$

and mixing distribution with Lebesgue density proportional to

$$\pi(\lambda|\,X^{(n)}) \propto \lambda^{n-1}e^{-\psi(\lambda)} \prod_{j=1}^{K_n} \frac{1}{(\lambda + b(\tilde{X}_j))^{N_{j,n}}}, \tag{7.3}$$

where $\psi(\lambda) := \int\int_0^\infty \left(1 - e^{-\lambda s}\right) s^{-1} e^{-sb(x)}\, ds\, da(x)$. For convenience of notation, let Λ be a random variable following the mixing density (7.3), for given $X^{(n)}$. Then (see (4.3)) the posterior distribution can be interpreted as the distribution of $(P_n(A) : A \in \mathcal{X})$ for

$$P_n(A) = \frac{\Psi(A) + \sum_{j=1}^{K_n} W'_{j,K_n} \delta_{\tilde{X}_j}(A)}{\Psi(\mathfrak{X}) + \sum_{j=1}^{K_n} W'_{j,K_n}},$$

where the random measure Ψ is distributed as $\Psi(A) = \int\int \mathbb{1}_A(x) s\, dN^c(x,s)$ conditional on $\Lambda = \lambda$, for a Poisson process N^c on $\mathfrak{X} \times (0,\infty]$ with intensity measure (7.1), and the variables W'_{j,K_n} follow gamma distributions with parameters $\left(N_{j,n}, \lambda + b(\tilde{X}_j)\right)$ (as given in (7.2)), independent of each other and of Ψ. By the properties of the gamma distribution, for any $j \in \{1,\ldots,K_n\}$, we can represent W'_{j,K_n} as the sum of N_{j,K_n} independent exponential random variables with intensity $\lambda + b(\tilde{X}_j)$. This allows to simplify the preceding representation to

$$P_n(A) = \frac{\Psi(A) + \sum_{i=1}^{n} W_i \delta_{X_i}(A)}{\Psi(\mathfrak{X}) + \sum_{i=1}^{n} W_i}, \tag{7.4}$$

where conditional on $\Lambda = \lambda$ the variables $W_1,\ldots,W_n$ are independent exponentially distributed with intensity parameters $\lambda + b(X_i)$, and are independent of Ψ. We denote by $\mathbb{E}_\Psi$ and $\mathbb{E}_W$ the expectations with respect to Ψ and the W_i, for given observations $X^{(n)}$ and given Λ. Furthermore, we denote by $\mathbb{E}_\Lambda$ the expectation relative to the mixture measure with density (7.3), for given $X^{(n)}$. In this notation we need to show that $\left|\mathbb{E}_\Lambda \mathbb{E}_W \mathbb{E}_\Psi h\left(\sqrt{n}(P_n - \mathbb{P}_n)\right) - \mathbb{E}h(\mathbb{B}_{P_0})\right| \to 0$, almost surely (or in probability), uniformly in $h \in \mathrm{BL}_1$. Define a weighted empirical distribution by $\mathbb{P}_n^W = (\sum_{i=1}^n W_i \delta_{X_i})/(\sum_{i=1}^n W_i)$. We shall show that the following two statements are true, almost surely (or in probability), after which an application of the triangle inequality gives the desired result:

$$\sup_{h\in\mathrm{BL}_1} \left|\mathbb{E}_\Lambda \mathbb{E}_W \mathbb{E}_\Psi h\left(\sqrt{n}(P_n - \mathbb{P}_n)\right) - \mathbb{E}_\Lambda \mathbb{E}_W h\left(\sqrt{n}(\mathbb{P}_n^W - \mathbb{P}_n)\right)\right| \to 0, \tag{7.5}$$

$$\sup_{h\in\mathrm{BL}_1} \left|\mathbb{E}_\Lambda \mathbb{E}_W h\left(\sqrt{n}(\mathbb{P}_n^W - \mathbb{P}_n)\right) - \mathbb{E}h(\mathbb{B}_{P_0})\right| \to 0. \tag{7.6}$$

An essential ingredient for proving these statements is that the distributions of Λ will give probability tending to one to the sets $[cn/\log n, \infty)$, for some small $c > 0$, as is shown in Lemma 7.6. This causes that the contribution of Ψ to (7.4) is negligible relative to the contribution of $\mathbb{P}_n^W$, leading to (7.5). Furthermore, it causes that the W_i will be asymptotically i.i.d. exponential variables (with large but almost equal intensities), so that the processes $\mathbb{P}_n^W$ are asymptotically equivalent to the Bayesian bootstrap process, which is known to tend to a Brownian bridge, leading to (7.6). To handle (7.5), we start by noting that, by (7.4) and the definition of $\mathbb{P}_n^W$,

$$P_n - \mathbb{P}_n^W = \frac{\Psi - \Psi(\mathfrak{X})\mathbb{P}_n^W}{\Psi(\mathfrak{X}) + \sum_{i=1}^n W_i}.$$

Because $\|\Psi\|_{\mathcal{F}} \le \Psi F$ and $\|\mathbb{P}_n^W\|_{\mathcal{F}} \le M_n := \max_{1\le i\le n} F(X_i)$, it follows that the left side of (7.5) is bounded above by

$$\mathbb{E}_\Lambda \mathbb{E}_W \mathbb{E}_\Psi \big\|\sqrt{n}(P_n - \mathbb{P}_n^W)\big\|_{\mathcal{F}} \le \sqrt{n}\, \mathbb{E}_\Lambda \mathbb{E}_W \mathbb{E}_\Psi \frac{\Psi F + \Psi(\mathfrak{X}) M_n}{\Psi(\mathfrak{X}) + \sum_{i=1}^n W_i}.$$

The expression on the right side without the leading expectation on Λ is of the form as in Lemma 7.3, with $f = F + 1_{\mathfrak{X}} M_n$ (and with M_n fixed due to the conditioning on $X_1, \ldots, X_n$). Thus the preceding display can be bounded above by

$$\sqrt{n}\, \mathbb{E}_\Lambda \int_0^\infty \Big[e^{\int \log\big((\Lambda+b)/(t+\Lambda+b)\big)\, da} \cdot \int \frac{f}{t+\Lambda+b}\, da \cdot \prod_{i=1}^n \frac{\Lambda + b(X_i)}{t+\Lambda+b(X_i)} \Big]\, dt.$$

The first term of the product in the integrand is bounded above by 1, since the logarithm in the exponent is negative. To bound the second and third terms we split the integral over t in the ranges $[0, P_0 b)$ and $[P_0 b, \infty)$. For t in the range $[0, P_0 b)$ we use the bound $\int f/(t+\Lambda+b)\, da \le af/(\Lambda + \underline{b})$ on the second term, where $af = \int f\, da$. Furthermore, in this range we bound the third term by 1. This shows that the integral over the range $[0, P_0 b)$ contributes no more than $af\, P_0 b\, \mathbb{E}_\Lambda \big[(\Lambda + \underline{b})^{-1}\big]$. For t in the range $[P_0 b, \infty)$, we bound the second term by $af/(t+\Lambda+\underline{b})$, while for the third term we use Jensen's inequality on the logarithm to see that

$$\prod_{i=1}^n \frac{\lambda + b(X_i)}{t+\lambda+b(X_j)} = e^{n\int \log\left(\frac{\lambda+b}{t+\lambda+b}\right) d\mathbb{P}_n} \le \Big(\mathbb{P}_n \frac{\lambda+b}{t+\lambda+b}\Big)^n \le \Big(\frac{\lambda + \mathbb{P}_n b}{t+\lambda+\underline{b}}\Big)^n.$$

Substituting these bounds in the integral, we find

$$af\, \mathbb{E}_\Lambda \Big[(\Lambda + \mathbb{P}_n b)^n \int_{P_0 b}^\infty \frac{1}{(t+\Lambda+\underline{b})^{n+1}}\, dt\Big] = \frac{af}{n} \mathbb{E}_\Lambda \Big(\frac{\Lambda + \mathbb{P}_n b}{P_0 b + \Lambda + \underline{b}}\Big)^n.$$

Since $\underline{b} > 0$ and $\mathbb{P}_n b - P_0 b \to 0$, almost surely (or in probability), the quotient on the right is bounded above by 1 eventually, almost surely. Thus combining the results for the two ranges, we find that as $n \to \infty$, almost surely,

$$\begin{aligned}\mathbb{E}_\Lambda \mathbb{E}_W \mathbb{E}_\Psi \left[\frac{\Psi f}{\Psi(\mathfrak{X}) + \sum_{i=1}^n W_i}\right] &\le af\Big[\mathbb{E}_\Lambda \frac{P_0 b}{\Lambda + \underline{b}} + \frac{1}{n}\Big] \\ &\le af\Big[\frac{P_0 b}{\underline{b}} \Pi\Big(\Lambda < \frac{cn}{\log n} \,|\, X^{(n)}\Big) + \frac{P_0 b\, \log n}{cn} + \frac{1}{n}\Big].\end{aligned}$$

We substitute $f = F + 1_{\mathfrak{X}} M_n$ in the right side and multiply by $\sqrt{n}$ to obtain a bound on (7.5) of the form

$$\big(aF + M_n a(\mathfrak{X})\big)\Big[\frac{P_0 b}{\underline{b}} \sqrt{n}\, \Pi\Big(\Lambda < \frac{cn}{\log n} \,|\, X^{(n)}\Big) + \frac{P_0 b\, \log n + 1}{c\sqrt{n}}\Big].$$

Here $M_n = O(n^{1/r})$, almost surely, for some $r > 2$, by Lemma 7.4 and the assumption that $P_0 F^r < \infty$. This shows that the second term tends to zero. The first term tends to zero for sufficiently small $c > 0$ in view of Lemma 7.6.

We are left with showing (7.6). Given Λ and $X^{(n)}$, the variables $\tilde{W}_i = W_i(\Lambda + b(X_i))/(\Lambda + \underline{b})$ are i.i.d. exponential variables with parameter $\Lambda + \underline{b}$. Since this is a scale parameter, the normalised variables $\tilde{W}_i / \sum_{i=1}^n \tilde{W}_i$ are equal in distribution to the normalised variables $\xi_i / \sum_{i=1}^n \xi_i$, for $\xi_1, \ldots, \xi_n$ standard exponential variables, still conditionally on Λ and $X^{(n)}$. This shows that the empirical process $\mathbb{P}_n^{\tilde{W}}$ with weights $\tilde{W}$, is equal in distribution to the Bayesian bootstrap process (see Example 3.7.9 in Van der Vaart and Wellner (2023)), conditionally given Λ, but then also unconditionally (still given $X^{(n)}$), as its distribution does not depend on Λ. By Præstgaard and Wellner (1993) (or Theorem 3.6.13 in Van der Vaart and Wellner (1996)), the Bayesian bootstrap process converges to the Brownian bridge process $\mathbb{B}_{P_0}$. Therefore, in view of the triangle inequality it suffices to bound the difference in (7.6) when replacing $\mathbb{P}_n^W$ by $\mathbb{P}_n^{\tilde{W}}$. This difference can be bounded by, with $M_n = \max_{1 \le i \le n} F(X_i)$ as before,

$$\begin{aligned}\mathbb{E}_\Lambda \mathbb{E}_W \sqrt{n} \|\mathbb{P}_n^W - \mathbb{P}_n^{\tilde{W}}\|_{\mathcal{F}} &\le 2M_n \sqrt{n}\, \mathbb{E}_\Lambda \mathbb{E}_W \frac{\sum_{i=1}^n |W_i - \tilde{W}_i|}{\sum_{i=1}^n W_i} \\ &= 2M_n \sqrt{n} \max_{1 \le i \le n} b(X_i)\, \mathbb{E}_\Lambda \left[(\Lambda + \underline{b})^{-1} \right],\end{aligned}$$

because $|W_i - \tilde{W}_i| = W_i\big(b(X_i) - \underline{b}\big)$. We again split the expectation over Λ in two parts to bound this by

$$2M_n \max_{1 \le i \le n} b(X_i) \Big[\frac{\sqrt{n}}{\underline{b}} \Pi\Big(\Lambda < \frac{cn}{\log n} \big| X^{(n)} \Big) + \frac{\log n}{c\sqrt{n}} \Big].$$

Here $M_n = O(n^{1/r})$ and $\max_{1 \le i \le n} b(X_i) = O(n^{1/q})$ almost surely, by Lemma 7.4 and the integrability assumptions on F and b, where $1/q + 1/r < 1/2$. it follows that the leading term is $O(n^{1/2-\epsilon})$, for some $\epsilon > 0$. Together with Lemma 7.6 this shows that the expression tends to zero, for sufficiently small $c > 0$. □

Theorem 7.2 (Posterior mean)**.** *For P_0^∞-almost every sequence $X_1, X_2, \ldots$,*

$$\sup_{A \in \mathcal{X}} \big| \mathbb{E}\big(P_n(A) \big| X_1, \ldots, X_n \big) - \mathbb{P}_n(A) \big| = O(1/n), \qquad a.s.$$

If b is bounded below by a positive constant, this order is sharp.

Lemma 7.3. *Let Ψ be a completely random measure without fixed atoms and continuous intensity measure given by $\nu^c(dx, ds) = s^{-1} e^{-sb(x)}\, ds\, dx$ and let $W_1, \ldots, W_n$ be independent exponential variables with means $b(X_i)$, independent of Ψ. Then, for any measurable function $f : \mathfrak{X} \to [0, \infty)$,*

$$\mathbb{E}_W \mathbb{E}_\Psi \left[\frac{\Psi f}{\Psi(\mathfrak{X}) + \sum_{i=1}^n W_i} \right] = \int_0^\infty \Big[e^{\int \log\left(\frac{b}{t+b}\right) da} \cdot \int \frac{f}{t+b}\, da \cdot \prod_{i=1}^n \frac{b(X_i)}{t + b(X_i)} \Big]\, dt.$$

Lemma 7.4 (Maximum of iid random variables)**.** *If $X_1, \ldots, X_n$ are i.i.d. random variables with $\mathbb{E}|X_i|^r < \infty$ for some $r > 0$, then $\max_{1 \le i \le n} |X_i|\, n^{-1/r} \xrightarrow{as} 0$. If $X_{1,n}, \ldots, X_{n,n}$ are i.i.d. random variables for every n such that the variables $|X_{1,n}|^r$ are uniformly integrable, then the assertion is true in probability.*

Lemma 7.5. *For any nonnegative measurable function $b : \mathfrak{X} \to (0, \infty)$, and any $\lambda > 0$, we have $\int \int_0^\infty (1 - e^{-\lambda s}) s^{-1} e^{-sb}\, ds\, da = \int \log(1 + \lambda/b)\, da$.*

Lemma 7.6 (Concentration of the mixing distribution)**.** *For Λ conditionally distributed given $X^{(n)} = (X_1, \ldots, X_n)$ according to density given in (7.3), for a measurable function $b : \mathfrak{X} \to [\underline{b}, \infty)$ with $P_0 b < \infty$, we have for P_0^∞-almost every sequence $X_1, X_2, \ldots$, for any $d \in \mathbb{R}^+$ and $0 < c < \underline{b}/\big(a(\mathfrak{X}) + 1 + d\big)\big)$,*

$$n^d\, \Pi\left(\Lambda < \frac{cn}{\log n} \middle| X^{(n)}\right) \to 0.$$

The convergence is uniform in functions $b : \mathfrak{X} \to [\underline{b}, \infty)$ such that the sequence $\mathbb{P}_n b$ remains uniformly bounded, almost surely. If the functions are uniformly bounded in probability, then the convergence is true uniformly in probability.

8. Discussion

The results in this paper justify the use of the studied Bayesian sensitivity analyses procedures in practice, with the caveat that the utility of the answers will depend on the ability of subject matter experts to express their ideas about the outcomes for those lost to follow-up through a prior on q. The need to incorporate expert knowledge originates from the unidentifiability of the distribution P_0 of those who dropped out and is thus not special to the Bayesian paradigm. In frequentist sensitivity analysis approaches, expert knowledge is incorporated in a variety of ways, usually through bounds on or specific values for parameters that depend on unknown, P_0-derived quantities (Liu, Kuramoto and Stuart, 2013). Bayesian approaches offer perhaps an advantage by allowing the user to specify their ideas about the sensitivity parameter through a probability distribution, leading to a single summary of the sensitivity analysis.

In practice, more information may be available than assumed in this paper, in the form of covariates Z measured for each individual. In that case, the method can be extended to sensitivity analysis on the Missing At Random (MAR) assumption, by introducing dependence of p, P_1 and q or η, P and q on Z. The approach taken here then becomes a Bayesian implementation of the modelling strategy proposed in Robins, Rotnitzky and Scharfstein (2000). They note that failure of the MAR assumption $Y \perp\!\!\!\perp R|\, Z$ is equivalent to the conditional distribution of $R|\, Z, Y$ not being free of Y. They then propose a sensitivity analysis by fitting a model of the type $\Pr(R = 1|\, Z, Y) = \Psi\big(\eta(Z) + q(Y, Z)\big)$, for a given sensitivity function q, which depends on both Y and Z. This is equivalent to (2.4) within levels of the covariate Z. In a Bayesian approach we need priors for this extended logistic regression model, the conditional distribution of the outcome Y given Z, and the marginal distribution of Z. For a binary outcome,

nonparametric priors for both conditional models can take the form of logistic Gaussian process models, as considered in Ray and van der Vaart (2020), who combine this with a Dirichlet process prior on the marginal distribution of the covariates, and prove a Bernstein-von Mises theorem under the MAR assumption. For survival outcomes, covariates could be added through a Cox proportional hazards model (Cox, 1972), for which several Bernstein-von Mises theorems exist (Kim, 2006; Kim and Lee, 2003). Another option of adding covariates is to use a dependent Dirichlet process prior (Quintana et al., 2020; MacEachern, 1999). It will be of interest to extend these results to the sensitivity setup, or from a different angle, extend the results of the present paper to include covariates.

In this paper we considered estimating a mean in the missing data problem, but this is easily extended to the estimation of an *average causal effect*. In the usual causal model (see Hernán and Robins (2022)), there are two potential outcomes Y^1 and Y^0, corresponding to an individual being treated ($R = 1$) or not ($R = 0$), and it is desired to estimate $\mathbb{E}Y^1 - \mathbb{E}Y^0$, based on observing only Y^1 if $R = 1$ and only Y^0 if $R = 0$. In other words, one of the two potential outcomes is missing. The usual assumption, called *conditional exchangeability* (CE) or *no unmeasured confounders*, is that $Y^r \perp\!\!\!\perp R | Z$, for $r = 0, 1$, which is exactly the MAR assumption considered in the preceding paragraph. A sensitivity analysis may investigate deviations from (CE) in exactly the manner proposed.

Within the causal setting, a different modelling perspective was considered in McCandless, Gustafson and Levy (2007); McCandless et al. (2012), also see Gustafson et al. (2010); Gustafson and McCandless (2018); McCandless and Gustafson (2017), and Dorie et al. (2016). These authors interpret the failure of (CE) as the omission of a covariate in the conditioning. Thus they hypothesize the existence of a unmeasured confounder U such that $Y^r \perp\!\!\!\perp R | Z, U$. Starting from a model for the distribution of (Y^r, R, Z, U), estimation of the parameter of interest is then achieved by considering this as a missing data problem with missing data U. Typically U is assumed to be a binary variable and its effect on the other variables is modelled parametrically, although Dorie et al. (2016) allow a nonparametric relation between the outcomes and covariates Z.

Appendix A: Correctly specified α simulations

In this section we discuss the results of the simulation study when the prior on α is correctly specified.

As discussed in the main paper, one of the differences between the two parametrisations is the posterior of α being dissimilar from the prior in the P, η parametrisation. When the prior on α is misspecified, this effect can be seen in the simulations. This setting is natural, as it is not assumed that the model is correct. However, it is still relevant to understand the behaviour of the posteriors when α is correctly specified. We use the same setting as in Section 6 of the main paper, the only difference being that the mean of the prior on α is taken to be $\alpha_0 = 2$. The same results as in the main paper can be found in

Figure 2 and the coverage for the different priors can be found in Table 3 and Table 4.

TABLE 3
Coverage of the 90%-credible intervals of the posterior distribution of the functional of interest for the different parametrisations with misspecified $\alpha \sim N(1, 0.25)$

	$n = 100$	$n = 1000$	$n = 10000$
	Coverage | Length	Coverage | Length	Coverage | Length
H, α	0.560 | 0.946	0.328 | 0.700	0.026 | 0.662
P, η, α	0.531 | 0.932	0.304 | 0.687	0.131 | 0.635

TABLE 4
Coverage of the 90%-credible intervals of the posterior distribution of the functional of interest for the different parametrisations with $\alpha \sim N(2, 0.25)$

	$n = 100$	$n = 1000$	$n = 10000$
	Coverage | Length	Coverage | Length	Coverage | Length
H, α	0.830 | 1.024	0.991 | 0.735	1.000 | 0.668
P, η, α	0.816 | 1.011	0.986 | 0.724	1.000 | 0.650

Both parametrisations seem to perform similarly, with only negligible differences in coverage and length. The posterior of α in the P, η parametrisation seems to behave like the prior, suggesting the data has minimal influence on it. These results are important, as they confirm the practical usefulness of the model. When an expert has proper knowledge of the difference between the missing and observed population, it should not matter which parametrisation is being used. The choice should be based on the knowledge of the other parameters.

Appendix B: Efficiency theory

In this section we derive the information for estimating the functional of interest when the sensitivity function is known, in the nonparametric model in which the distribution of the observations is completely unknown.

We can parametrise the model by the pair (p, P_1) and are then interested in estimating the function $\chi(p, P_1) = (1-p)P_1(ge^q)/P_1e^q + pP_1g$, for a given measurable function g. We suppose that q is known, so that the functional is identifiable, and are interested in obtaining the minimal attainable asymptotic variance of a sequence of estimators (in the sense of the local minimax or convolution theorem, as explained in e.g. van der Vaart (1988, 1991b)).

The likelihood for a single observation (X, R) can be written as

$$(p, P_1) \mapsto (1-p)^{1-R} p^R p_1(X)^R.$$

A path $(p + th, P_{1,t})$ given by $dP_{1,t} \propto e^{t\phi}\, dP_1$, gives a score function at $t = 0$ given by

$$h\frac{R-p}{p(1-p)} + R(\phi(X) - P_1\phi) =: (B^p_{p,P_1}h + B^{P_1}_{p,P_1}\phi)(R, X).$$

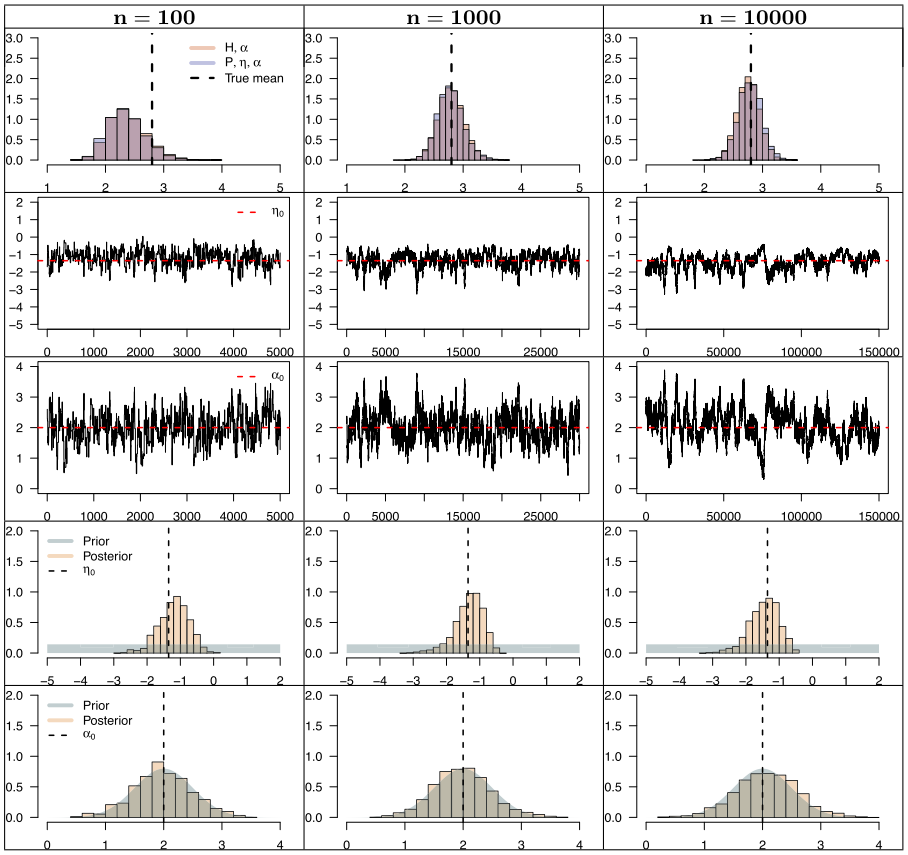

FIG 2. Row 1*: Histogram of the posterior distribution of the functional of interest for the different parametrisations.* Row 2*: The posterior path of* η. Row 3*: The posterior path of* α. Row 4*: The prior and posterior of* η. Row 5*: The prior and posterior of* α.

The tangent space of the model is by the definition the set of all scores, when h varies over $\mathbb{R}$ and ϕ varies of $L_2(P_1)$. Since $dP_{0,t} \propto e^q\, dP_{1,t} \propto e^q e^{\phi}\, dP_1 \propto e^{t\phi}\, dP_0$, the path $P_{0,t}$ has score function $\phi - P_0\phi$. The derivative of the functional of interest is

$$\begin{aligned}\frac{d}{dt}_{|t=0} \int g\, dP_t &= \frac{d}{dt}_{|t=0} \Big[(1-p_t)\int g\, dP_{0,t} + p_t \int g\, dP_{1,t}\Big] \\ &= h\int g\, d(P_1 - P_0) + (1-p)\int g(\phi - P_0\phi)\, dP_0 + p\int g(\phi - P_1\phi)\, dP_1 \\ &= \mathbb{E}_{p,P_1}\left[\Big(\tilde{h}\frac{R-p}{p(1-p)} + R(\tilde{\phi}(X) - P_1\tilde{\phi})\Big)\Big(B^p_{p,P_1}h + B^{P_1}_{p,P_1}\phi\Big)(R,X)\right],\end{aligned}$$

for the 'least favourable directions' given by

$$\tilde{h} = p(1-p)\,(P_1 - P_0)g,$$

$$\tilde{\phi} = \frac{1-p}{p}(g - P_0 g)\frac{e^q}{P_1 e^q} + g - P_1 g = (g - P_0 g)e^{\eta+q} + g - P_1 g.$$

Thus the efficient influence function is equal to

$$B^p_{p,P_1}\tilde{h} + B^{P_1}_{p,P_1}\tilde{\phi} = (R-p)(P_0 g - P_1 g) + R\big((g - P_0 g)e^{\eta+q} + (g - P_1 g)\big).$$

The variance of this function is the least possible asymptotic variance, and is given by

$$\begin{aligned} &p(1-p)(P_0 g - P_1 g)^2 + p\,P_1\big((g - P_0 g)e^{\eta+q} + (g - P_1 g)\big)^2 \\ &\qquad = p(1-p)(P_0 g - P_1 g)^2 + p\,P_1\big(g(1+e^{\eta+q}) - P_1(g(1+e^{\eta+q}))\big)^2. \end{aligned}$$

Appendix C: Proofs

Proof of Theorem 7.2. For $k \in \mathbb{N}$, define $\tau_k(\lambda|\, x) = \int s^{k-1}e^{-s(\lambda+b(x))}\,ds = \Gamma(k)/\big(\lambda + b(x)\big)^k$. Then $\prod_{j=1}^{N_{j,n}} \tau_{N_{j,n}}(\lambda|\, \tilde{X}_j) = \prod_{j=1}^{N_{j,n}} \Gamma(N_{j,n})/\prod_{i=1}^n \big(\lambda + b(X_i)\big)$. Because the posterior mean is equal to the predictive distribution, Ghosal and Van der Vaart (2017), Corollary 14.59 gives that

$$\mathbb{E}\big(P_n(A)|\, X^{(n)}\big) = \sum_{j=1}^{K_n} p_j(X^{(n)})\delta_{\tilde{X}_j}(A) + \int_A p_{K_n+1}(X^{(n)}|\, x)\,d\overline{a}(x),$$

where $\overline{a}$ is the normalisation of a and the p_j are the prediction probability function (PPF) of the exchangeable partitions generated by P_n, given by, for $j = 1, \ldots, K_n$,

$$\begin{aligned} p_j(X^{(n)}) &= \frac{N_{j,n}\int_0^\infty \lambda^{-1}e^{-\psi(\lambda)}\frac{\lambda}{\lambda+b(\tilde{X}_j)}\prod_{i=1}^n \frac{\lambda}{\lambda+b(X_i)}\,d\lambda}{n\int_0^\infty \lambda^{-1}e^{-\psi(\lambda)}\prod_{i=1}^n \frac{\lambda}{\lambda+b(X_i)}\,d\lambda} \\ p_{K_n+1}(X^{(n)}|\, x) &= \frac{\int_0^\infty e^{-\psi(\lambda)}\frac{1}{\lambda+b(x)}\prod_{i=1}^n \frac{\lambda}{\lambda+b(X_i)}\,d\lambda}{n\int_0^\infty \lambda^{-1}e^{-\psi(\lambda)}\prod_{i=1}^n \frac{\lambda}{\lambda+b(X_i)}\,d\lambda}. \end{aligned}$$

Define $p'_1, \ldots, p'_n$ such that $N_{j,n}\,p'_i(X^{(n)}) = p_j(X^{(n)})$ if $X_i = \tilde{X}_j$, and $p'_{n+1}(X^{(n)}|\, x) = p_{K_n+1}(X^{(n)}|\, x)$. (Hence $p'_i(X^{(n)})$ is equal to the quotient in the preceding display that defines $p_j(X^{(n)})$ without the multiplicity $N_{j,n}$, for j the index of the partitioning set to which X_i belongs.) Using this notation, we can simplify the formula for the posterior mean to

$$\mathbb{E}\big(P_n(A)|\, X^{(n)}\big) = \sum_{i=1}^n p'_i(X^{(n)})\delta_{X_i}(A) + \int_A p'_{n+1}(X^{(n)}|\, x)\,d\overline{a}(x).$$

As $\lambda/\big(\lambda + b(x)\big) \le 1$, for all $x \in \mathfrak{X}$, it follows that $p'_i(X^{(n)}) \le 1/n$, for $i = 1, 2 \ldots, n$, and $p'_{n+1}(X^{(n)}|\, x) \le 1/n$, for all $x \in \mathfrak{X}$. Using this, we obtain

$$\sup_{A\in\mathcal{X}} \Big|\mathbb{E}\big(P_n(A)|\, X^{(n)}\big) - \mathbb{P}_n(A)\Big|$$

$$\le \sup_{A\in\mathcal{X}} \sum_{i=1}^n \left| p_i'(X^{(n)}) - \frac{1}{n}\right| \delta_{X_i}(A) + \sup_{A\in\mathcal{X}} \left| \int_A p_{n+1}'(X^{(n)}|\, x)\, d\overline{a}(x)\right|$$
$$= 1 - \sum_{i=1}^n p_i'(X^{(n)}) + \int p_{n+1}'(X^{(n)}|\, x)\, d\overline{a}(x) = 2\int p_{n+1}'(X^{(n)}|\, x)\, d\overline{a}(x) \le \frac{2}{n},$$

where we used that $\overline{a}$ is a probability measure.

To prove the last assertion of the theorem, we consider the continuous part $A \mapsto \int_A p_{n+1}'(X^{(n)}|\, x)\, d\overline{a}(x)$ of the posterior mean. If b is bounded below by $\underline{b}$, then $\prod_{i=1}^n \lambda/\big(\lambda + b(X_i)\big) \le \big(\lambda/(\lambda+\underline{b})\big)^n$. This implies that the probability measures G_n on $(0,\infty)$ defined by

$$dG_n(\lambda) \propto \lambda^{-1} e^{-\psi(\lambda)} \prod_{i=1}^n \frac{\lambda}{\lambda + b(X_i)}\, d\lambda,$$

tend weakly to ∞. Hence $np_{n+1}'(X^{(n)}|\, x) = \int \lambda/\big(\lambda+b(x)\big)\, dG_n(\lambda)$ tends to 1, for every x. This in turn implies that $\liminf n\int p_{n+1}'(X^{(n)}|\, x)\, d\overline{a}(x) \ge \|a\|$, in view of Fatou's lemma. Thus n times the continuous part $A \mapsto \int_A p_{n+1}'(X^{(n)}|\, x)\, d\overline{a}(x)$ of the posterior mean does not tend to zero, and the distance between the posterior mean and the empirical distribution, on for instance the set $A = \{X_1,\dots,X_n\}$, is of no smaller order than $1/n$. □

Proof of Lemma 7.3. We first show the equality for $f = \mathbb{1}_A$ with $A \in \mathcal{X}$. Note that $1/y = \int_0^\infty e^{-ty}\, dt$ for any $y \in \mathbb{R}^+$. This can be used to write

$$\mathbb{E}_W\mathbb{E}_\Psi\left[\frac{\Psi(A)}{\Psi(\mathfrak{X}) + \sum_{i=1}^n W_i}\right] = \mathbb{E}_W\mathbb{E}_\Psi\left[\Psi(A)\int_0^\infty e^{-t(\Psi(\mathfrak{X})+\sum_{i=1}^n W_i)}\, dt\right]$$
$$= \int_0^\infty \mathbb{E}_W\mathbb{E}_\Psi\left[\Psi(A)e^{-t\Psi(A)}\cdot e^{-t\Psi(A^c)}\cdot e^{-t\sum_{i=1}^n W_i}\right] dt,$$

where we used Fubini's theorem in the last equality. Since Ψ is a Poisson process and the sets A and A^c are disjoint, the variables $\Psi(A)$ and $\Psi(A^c)$ are independent. They are also independent of the W_i and these weights are also independent of each other. Therefore the preceding display can be written as

$$\int_0^\infty \mathbb{E}_\Psi\left[\Psi(A)e^{-t\Psi(A)}\right]\cdot \mathbb{E}_\Psi\left[e^{-t\Psi(A^c)}\right]\cdot \prod_{i=1}^n \mathbb{E}_W\left[e^{-tW_i}\right] dt \qquad \text{(C.1)}$$

Each term can be calculated explicitly. We start with the second term in the above display. By Proposition J.6 in (Ghosal and Van der Vaart, 2017, p. 598), we have

$$\mathbb{E}_\Psi\left[e^{-t\Psi(A^c)}\right] = e^{-\int_{A^c}\int_0^\infty (1-e^{-ts})s^{-1}e^{-sb}\, ds\, da} = e^{-\int_{A^c}\log\left(\frac{t+b}{b}\right) da},$$

where the last equality follows from Lemma 8.3 in the main paper, for any positive constant t. In view of this, the first term in equation (C.1) can be

written

$$\mathbb{E}_\Psi\left[\Psi(A)e^{-t\Psi(A)}\right] = -\frac{d}{dt}\mathbb{E}_\Psi\left[e^{-t\Psi(A)}\right] = e^{-\int_A \log\left(\frac{t+b}{b}\right)da} \cdot \int_A \frac{1}{t+b}\,da.$$

The third term in equation (C.1) is

$$\prod_{i=1}^n \mathbb{E}_W\left[e^{-tW_i}\right] = \prod_{i=1}^n \int_0^\infty e^{-tw}\, b(X_i) e^{-wb(X_i)}\,dw = \prod_{i=1}^n \frac{b(X_i)}{t+b(X_i)}.$$

Combining these identities finishes the proof of the lemma for indicator functions.

When $f = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ for disjoint sets $A_1, \ldots, A_k$, the lemma is again true, by the independence of the random variables $\Psi(A_i)$. Any nonnegative measurable function f can be approximated by such linear combinations f_k as above with $f_k \le f_{k+1}$ almost everywhere for each k and $\limsup_{k\to\infty} f_k = f$. Then by the monotone convergence theorem, our claim is also true for any nonnegative measurable function f. □

Proof of Lemma 7.4. For any $x > 0$ we have

$$\max_{1\le i\le n} \frac{|X_i|^r}{n} \le \frac{y^r}{n} + \frac{1}{n}\sum_{i=1}^n |X_i|^r \mathbb{1}_{\{|X_i|>y\}}.$$

As $n \to \infty$ the first term tends to zero, for fixed y, while the second term tends to $\mathbb{E}\,|X_i|^r \mathbb{1}_{\{|X_i|>y\}}$, by the law of large numbers, and can be made arbitrarily small by choice of y. □

Proof of Lemma 7.5. Write $\psi(\lambda)$ for the left side of the equation. The derivative of the function $\lambda \mapsto \psi(\lambda)$ is equal to

$$\psi'(\lambda) = \int\int_0^\infty s e^{-s\lambda} s^{-1} e^{-sb}\,ds\,da = \int \frac{1}{\lambda+b}\,da.$$

Since $\psi(0) = 0$, the fundamental theorem of calculus gives $\psi(\lambda) = \int_0^\lambda \psi'(t)\,dt = \int \log(1+\lambda/b)\,da$, by Fubini's theorem. □

Proof of Lemma 7.6. We start by deriving upper and lower bounds on the posterior density given in equation 8.2 in the main paper. Denote the norming constant of this density by C_n.

For the upper bound, we use that $\psi(\lambda) \ge 0$, $b(\tilde{X}_j) \ge \underline{b}$, and $\lambda + \underline{b} \ge \underline{b}$ to find that

$$\pi(\lambda|\, X^{(n)}) \le \frac{1}{C_n \underline{b}}\left(\frac{\lambda}{\lambda+\underline{b}}\right)^{n-1} \le \frac{1}{C_n \underline{b}} e^{-\underline{b}/\delta},$$

for $\lambda + \underline{b} < \delta(n-1)$, where we used that $1 - x \le e^{-x}$, for every x.

For the lower bound we use that $\psi(\lambda) \le a(\mathfrak{X}) \log(1+\lambda/\underline{b})$, in view of Lemma 8.3 in the main paper, and that $\lambda/(\lambda + b(\tilde{X}_j)) \ge e^{-b(\tilde{X}_j)/\lambda}$ to find that

$$\begin{aligned}\pi(\lambda|\, X^{(n)}) &\ge \frac{1}{C_n\lambda} e^{-a(\mathfrak{X})\log(1+\lambda/\underline{b})} e^{-\sum_{j=1}^{K_n} N_{j,K_n} b(\tilde{X}_j)/\lambda} \\ &= \frac{1}{C_n\lambda}\Big(1+\frac{\lambda}{\underline{b}}\Big)^{-a(\mathfrak{X})} e^{-n\mathbb{P}_n b/\lambda} \\ &\ge \frac{(\underline{b}/2)^{a(\mathfrak{X})}}{C_n\lambda^{1+a(\mathfrak{X})}} e^{-\mathbb{P}_n b/\epsilon},\end{aligned}$$

for $\lambda > n\epsilon \vee \underline{b}$ and any $\epsilon > 0$, so that $1 + \lambda/\underline{b} \le 2\lambda/\underline{b}$.

By combining the upper and lower bounds, we obtain, for $n\epsilon > \underline{b}$,

$$\begin{aligned}\frac{\Pi(\Lambda < \delta n - \delta - \underline{b}|\, X^{(n)})}{\Pi(\Lambda > \epsilon n|\, X^{(n)})} &\le \frac{\int_0^{\delta n} \underline{b}^{-1} e^{-\underline{b}/\delta}\, d\lambda}{\int_{\epsilon n}^{\infty} \lambda^{-1-a(\mathfrak{X})} (\underline{b}/2)^{a(\mathfrak{X})} e^{-\mathbb{P}_n b/\epsilon}\, d\lambda} \\ &= \delta n^{1+a(\mathfrak{X})} e^{-\underline{b}/\delta}\, e^{\mathbb{P}_n b/\epsilon} a(\mathfrak{X}) \epsilon^{a(\mathfrak{X})} \underline{b}^{-a(\mathfrak{X})-1} 2^{a(\mathfrak{X})}.\end{aligned}$$

Since the denominator on the left side is smaller than 1, the probability $\Pi(\Lambda < \delta n - \delta - \underline{b}|\, X^{(n)})$ is also bounded by the right side. Here $\mathbb{P}_n b \to P_0 b$, by the strong law of large numbers, and $e^{-\underline{b}/\delta} = n^{-\underline{b}/c}$, for $\delta = c/\log n$. Then n^d times the right side tends to zero if $1 + a(\mathfrak{X}) + d \le \underline{b}/c$.

This is true uniformly in b provided $\mathbb{P}_n b$ remains bounded. □

Funding

The first and third author were supported in part by a Spinoza prize awarded to author three by the Netherlands Organisation for Scientific Research (NWO).

The second author was supported in part by grant VI.Veni.192.087 by the Netherlands Organisation for Scientific Research (NWO).

Supplementary Material

Supplementary Material for "Bayesian sensitivity analysis for a missing data model"

(doi: 10.1214/26-EJS2568SUPP; .zip).

Simulations.R

Generates the posterior distributions of the functional of interest for both parametric approaches. Uses all cpu cores – 1, but this can be changed at the top of the file. Outputs one .csv with the data and 5 images in the working directory. Image 1 contains histograms of the posteriors, Images 2 and 3 contain the trajectories of the posteriors of eta and alpha and images 4 and 5 contain the priors and posteriors of eta and alpha. Also shows the accept rate of the random walk Metropolis Hastings step in the Gibbs sampling scheme.

Coverage.R
Calculates the coverage of the described method in the paper. It is set-up for running 10 sessions at the same time, each running on multiple cores for use on a cluster. It outputs the credible sets and their lengths into a CSV file (only 1/10 of it). For the full coverage and length, add the values in the CSV of 10 runs of the program. Easy implementation, without any dependencies except parallel. Made to run on a single node on multiple cpu cores. Perform 10 separate runs on different nodes.

Coverage_local.R
Does the same as Coverage.R, but is made for local runs. Not recommended for high values of n due to long runtimes. For $n < 1000$ it is viable if one has no access to multiple nodes.

References

Balzer, L. B., Ayieko, J., Kwarisiima, D., Chamie, G., Charlebois, E. D., Schwab, J., van der Laan, M. J., Kamya, M. R., Havlir, D. V. and Petersen, M. L. (2020). Far from MCAR: Obtaining Population-level Estimates of HIV Viral Suppression. *Epidemiology* **31**.

Canale, A., Lijoi, A., Nipoti, B. and Prünster, I. (2023). Inner spike and slab Bayesian nonparametric models. *Econom. Stat.* **27** 120–135. https://doi.org/10.1016/j.ecosta.2021.10.017 MR4608341

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. With discussion by F. Downton, R. Peto, D. Bartholomew, D. Lindley, P. Glassborow, D. Barton, S. Howard, B. Benjamin, J. Gart, L. Meshalkin, A. Kagan, M. Zelen, R. Barlow, J. Kalbfleisch, R. Prentice and N. Breslow, and a reply by D. R. Cox. MR0341758 (49 ##6504)

Delft High Performance Computing Centre (DHPC) (2022). DelftBlue Supercomputer (Phase 1). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1.

Dorie, V., Harada, M., Carnegie, N. B. and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine* **35** 3453-3470. https://doi.org/10.1002/sim.6973 MR3537215

Dudley, R. M. (2014). *Uniform central limit theorems*, second ed. *Cambridge Studies in Advanced Mathematics* **142**. Cambridge University Press, New York. MR3445285

Dykstra, R. L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9** 356–367. MR606619

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1** 209–230. https://doi.org/10.1214/aos/1176342360 MR0350949

Gelman, A., Gilks, W. R. and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Ap-*

plied Probability **7** 110–120. https://doi.org/10.1214/aoap/1034625254 MR1428751

Ghosal, S. and Van der Vaart, A. W. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. MR3587782

Gustafson, P. and McCandless, L. C. (2018). When is a sensitivity parameter exactly that? *Statist. Sci.* **33** 86–95. https://doi.org/10.1214/17-STS632 MR3757506

Gustafson, P., McCandless, L. C., Levy, A. R. and Richardson, S. (2010). Simplified Bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics* **66** 1129–1137. https://doi.org/10.1111/j.1541-0420.2009.01377.x MR2758500

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S. and Merigan, T. C. (1996). A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with CD4 Cell Counts from 200 to 500 per Cubic Millimeter. *New England Journal of Medicine* **335** 1081–1090. PMID: 8813038. https://doi.org/10.1056/NEJM199610103351501

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57** 97–109. MR3363437

Hernán, M. and Robins, J. (2022). Causal Inference: What If.

Jacob, P. E., Murray, L. M., Holmes, C. C. and Robert, C. P. (2017). Better together? Statistical learning in models made of modules. arXiv:1708.08719.

James, L., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97. https://doi.org/10.1111/j.1467-9469.2008.00609.x MR2508332 (2010g:62020)

Kim, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700. https://doi.org/10.1214/009053606000000533 MR2283713 (2008k:62101)

Kim, Y. and Lee, J. (2003). Bayesian bootstrap for proportional hazards models. *Ann. Statist.* **31** 1905–1922. https://doi.org/10.1214/aos/1074290331 MR2036394 (2004k:62095)

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21** 59–78. MR0210185

Kingman, J. (1975). Random discrete distribution. *J. Roy. Statist. Soc. Ser. B* **37** 1–22. With a discussion by S. J. Taylor, A. G. Hawkes, A. M. Walker, D. R. Cox, A. F. M. Smith, B. M. Hill, P. J. Burville, T. Leonard and a reply by the author. MR0368264 (51 ##4505)

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian nonparametrics. Camb. Ser. Stat. Probab. Math.* **28** 80–136. Cambridge Univ. Press, Cambridge. MR2730661

Little, R. J. A. and Rubin, D. B. (2019). *Statistical analysis with missing data*, third ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. https://doi.org/10.1002/

9781119482260 MR1925014

Liu, W., Kuramoto, S. J. and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science* **14** 570–580.

Lo, A. Y. (1983). Weak Convergence for Dirichlet Processes. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* **45** 105–111. MR0749358

Lo, A. Y. (1986). A Remark on the Limiting Posterior Distribution of the Multiparameter Dirichlet Process. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* **48** 247–249. MR0905464

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science* 50–55. American Statistical Association, pp. 50–55, Alexandria, VA.

McCandless, L. C., Gustafson, P. and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat. Med.* **26** 2331–2347. https://doi.org/10.1002/sim.2711 MR2368419

McCandless, L. C. and Gustafson, P. (2017). A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Stat. Med.* **36** 2887–2901. https://doi.org/10.1002/sim.7298 MR3670397

McCandless, L. C., Gustafson, P., Levy, A. R. and Richardson, S. (2012). Hierarchical priors for bias parameters in Bayesian sensitivity analysis for unmeasured confounding. *Stat. Med.* **31** 383–396. https://doi.org/10.1002/sim.4453 MR2879811

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21** 1087–1092. https://doi.org/10.1063/1.1699114

Moss, D. and Rousseau, J. (2022). Efficient Bayesian estimation and use of cut posterior in semiparametric hidden Markov models. arXiv:2203.06081. MR4736272

Præstgaard, J. and Wellner, J. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301 (94k:60054)

Quintana, F. A., Mueller, P., Jara, A. and MacEachern, S. N. (2020). The Dependent Dirichlet Process and Related Models. https://doi.org/10.48550/ARXIV.2007.06129 MR4371095

Ray, K. and van der Vaart, A. (2020). Semiparametric Bayesian causal inference. *Ann. Statist.* **48** 2999–3020. https://doi.org/10.1214/19-AOS1919 MR4152632

Robins, J. M., Rotnitzky, A. and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997). IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. https://doi.org/10.1007/978-1-4612-1284-3_1 MR1731681

Scharfstein, D. O., Daniels, M. J. and Robins, J. M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials

with missing outcomes. *Biostatistics* **4** 495–512.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4** 639–650. MR1309433

VAN DER VAART, A. W. (1988). *Statistical estimation in large parameter spaces. CWI Tract* **44**. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam. MR927725

VAN DER VAART, A. (1991a). Efficiency and Hadamard differentiability. *Scand. J. Statist.* **18** 63–75. MR1115183

VAN DER VAART, A. (1991b). On differentiable functionals. *Ann. Statist.* **19** 178–204. https://doi.org/10.1214/aos/1176347976 MR1091845 (92i:62100)

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. MR1652247

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York, NY. MR1385671

VAN DER VAART, A. W. and WELLNER, J. A. (2023). *Weak Convergence and Empirical Processes, 2nd edition*. Springer, New York, NY. MR4628026