Sample-constrained partial identification with application to selection bias

BY MATTHEW J. TUDBALL

MRC Integrative Epidemiology Unit, University of Bristol,
Oakfield Grove, Bristol, BS8 2BN, U.K.
matt.tudball@bristol.ac.uk

RACHAEL A. HUGHES, KATE TILLING

MRC Integrative Epidemiology Unit, University of Bristol,
Oakfield Grove, Bristol, BS8 2BN, U.K.
rachael.hughes@bristol.ac.uk kate.tilling@bristol.ac.uk

JACK BOWDEN

College of Medicine and Health, University of Exeter,
Heavitree Road, Exeter, EX1 2LU, U.K.
j.bowden2@exeter.ac.uk

AND QINGYUAN ZHAO

Department of Pure Mathematics and Mathematical Statistics, University of Cambridge,
Wilberforce Road, Cambridge CB3 0WB, U.K.
qyzhao@statslab.cam.ac.uk

SUMMARY

Many partial identification problems can be characterized by the optimal value of a function over a set where both the function and set need to be estimated by empirical data. Despite some progress for convex problems, statistical inference in this general setting remains to be developed. To address this, we derive an asymptotically valid confidence interval for the optimal value through an appropriate relaxation of the estimated set. We then apply this general result to the problem of selection bias in population-based cohort studies. We show that existing sensitivity analyses, which are often conservative and difficult to implement, can be formulated in our framework and made significantly more informative via auxiliary information on the population. We conduct a simulation study to evaluate the finite sample performance of our inference procedure, and conclude with a substantive motivating example on the causal effect of education on income in the highly selected UK Biobank cohort. We demonstrate that our method can produce informative bounds using plausible population-level auxiliary constraints. We implement this method in the `R` package `selectioninterval`.

Some key words: Auxiliary information; Constraint; Partial identification; Selection bias; Sensitivity analysis.

1. INTRODUCTION

1.1. *General problem*

Partial identification problems arise when the observable data are only sufficient to identify a set or interval in which a parameter of interest is contained. A classical example from Manski (2003) is the missing data problem, where Y is a discrete random variable and S is a binary random variable indicating whether Y is observed ($S = 1$) or not ($S = 0$). The distribution of Y can be decomposed into

$$\mathrm{pr}(Y = y) = \mathrm{pr}(Y = y \mid S = 1)\,\mathrm{pr}(S = 1) + \mathrm{pr}(Y = y \mid S = 0)\,\mathrm{pr}(S = 0)$$

for any y in the support of Y. Given that $\mathrm{pr}(Y = y \mid S = 0)$ is unobserved, the smallest value that $\mathrm{pr}(Y = y)$ could take is $\mathrm{pr}(Y = y \mid S = 1)\,\mathrm{pr}(S = 1)$ and the largest value is $\mathrm{pr}(Y = y \mid S = 1)\,\mathrm{pr}(S = 1) + \mathrm{pr}(S = 0)$. Therefore, although $\mathrm{pr}(Y = y)$ itself cannot be point identified, it can be partially identified via the interval corresponding to the smallest and largest possible values.

Many partial identification problems can be formulated as the optimal value of a population objective function, which we write as

$$\nu = \inf\{Q(\theta) : \theta \in \Theta\}, \tag{1}$$

where $Q : \mathbb{R}^p \to \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^p$. In the missing data example, $Q(\theta) = \mathrm{pr}(Y = y \mid S = 1)\,\mathrm{pr}(S = 1) + \theta\,\mathrm{pr}(S = 0)$ and $\Theta = [0, 1]$.

The field of stochastic optimization also considers problems of the form in (1) and has built a large literature on estimation of, and inference to, ν when a sample analogue Q_n is observed instead of Q, where n denotes the sample size. We demonstrate that framing the partial identification problem as a stochastic optimization problem will allow us to draw upon these existing results.

Specifically, in this article we are concerned with the difficult setting where Θ must also be estimated empirically. We consider a setting where Θ is characterized by inequality constraints of the form $\Theta = \{\theta : h_j(\theta) \leqslant 0, j = 1, \ldots, J\}$, where we may only observe corresponding estimators $h_{nj}(\theta)$. Within this setting, our goal is to find a lower confidence bound C_n for any $0 < \alpha < 1$ such that

$$\lim_{n\to\infty} \mathrm{pr}(C_n \leqslant \nu) \geqslant 1 - \alpha, \tag{2}$$

which will suffice to provide useful statistical inference in a wide set of applications.

1.2. *Motivating application*

Our investigation is motivated by an applied question: how will selection bias affect the conclusions of population-based cohort studies? Many statistical analyses begin by selecting a study sample from some population of interest. When the sample is drawn nonrandomly, then valid inference for the population is no longer guaranteed (Bareinboim et al., 2014). Inverse probability weighting could be used to correct this selection bias (Horvitz & Thompson, 1952; Stuart et al., 2011), but data on nonselected observations may be limited or unavailable altogether, such that the weights cannot be estimated. In such settings, there exist approaches to assess the sensitivity of estimates to a range of plausible inverse probability weights (Aronow & Lee, 2013; Thompson & Arah, 2014). However, these approaches could be made more informative via a principled procedure for conducting statistical inference and the inclusion of relevant auxiliary information about the population. We demonstrate that such improvements can be made by casting these sensitivity analyses within the general framework described in § 1.1.

We are specifically motivated by studies conducted in the UK Biobank, which is a large population-based cohort study widely analysed by health researchers. Studies of this cohort are potentially biased since recruited participants are known to differ systematically from the rest of the U.K. population on measures such as education, health status, age and geographical location (Fry et al., 2017; Hughes et al., 2019).

1.3. *Existing literature*

Statistical inference procedures have been developed for some special cases of our general problem in (2). An area of particular focus is the so-called 'sample average approximation' (Shapiro et al., 2009). In this case, Q is the expected value $Q(\theta) = E\{f(\theta, X)\}$ of some function f and Q_n is a sample average $Q_n(\theta) = n^{-1}\sum_{i=1}^{n} f(\theta, X_i)$, where X is some random variable and $X_1, \ldots, X_n$ are independent draws of X.

Statistical inference in the presence of Θ_n has been developed for convex sample average approximations, such that f is convex in θ and $\Theta = \{\theta : h_j(\theta) \leqslant 0, j = 1, \ldots, J\}$, where $h_j(\theta) = E\{g_j(\theta, X)\}$ and $g_j(\theta, X)$ is convex in θ for all j. Shapiro (1991) showed that the plug-in estimator

$$\nu_n^p = \inf\{Q_n(\theta) : \theta \in \Theta_n\} \tag{3}$$

satisfies a central limit theorem under these convexity assumptions, and some additional regularity conditions, where $\Theta_n = \{\theta : \sum_{i=1}^{n} g_j(\theta, X_i) \leqslant 0, j = 1, \ldots, J\}$.

Moving away from convex problems, Wang & Ahmed (2008) considered the special case of minimizing a known function Q subject to a single expected value constraint $\Theta = \{\theta : E\{g(\theta, X)\} \leqslant 0\}$. They proposed an approach for calculating a sample size n so that Θ_n is feasible to some small relaxation of the true problem with high probability.

Our work also overlaps with the partial identification literature in econometrics, much of which considers inference for identified sets characterized by conditional or unconditional moment inequalities, commonly interpreted as the set of minimizers of some criterion function (Chernozhukov et al., 2007; Andrews & Soares, 2010; Andrews & Shi, 2013). A related literature provides inference for parameters lying within partially identified sets, as opposed to inference for the set itself (Imbens & Manski, 2004; Stoye, 2009). For a more comprehensive review of the partial identification literature, see Molinari (2020).

2. Confidence intervals for sample-constrained partial identification

2.1. *Confidence intervals under known constraints*

In this section, we briefly summarize existing results on statistical inference for stochastic optimization when the set Θ is observed, which forms the basis of our generalization to situations where an estimate Θ_n of Θ is observed instead. Suppose that the parameter space is defined by a set of inequality constraints

$$\Theta = \{\theta : h_j(\theta) \leqslant 0, j = 1, \ldots, J\},$$

where an equality constraint for some $h_j(\theta)$ can be introduced by taking the inequality constraints of both $h_j(\theta)$ and $-h_j(\theta)$. Recall that our goal is to provide inference about the infimum $\nu = \inf\{Q(\theta) : \theta \in \Theta\}$.

Much of the literature in stochastic optimization is centred on the statistical properties of the estimator

$$\nu_n = \inf\{Q_n(\theta) : \theta \in \Theta\}. \tag{4}$$

Consistency of optimal values and optimal solutions to such stochastic optimization problems is typically achieved by imposing uniform convergence of $Q_n(\theta)$ to $Q(\theta)$. First-order asymptotic properties are obtained via the functional delta method. The key conditions are that the infimum, viewed as a function of Q, satisfies some notion of differentiability at Q and that $n^{-1/2}(Q - Q_n)$ converges to a Gaussian process; see Shapiro (1991) for further details.

To make the previous discussion more concrete, consider the following four assumptions commonly placed on the stochastic optimization problem described above.

Assumption 1. The set of solutions to (1) is a singleton $\{\theta \in \Theta : Q(\theta) = \nu\} = \{\vartheta\}$.

Assumption 2. Let $B \subseteq \mathbb{R}^p$ denote a compact set and $C(B)$ denote the space of continuous functions on domain B. Then $\Theta \subseteq B$, $Q \in C(B)$ and $Q_n \in C(B)$ with probability 1.

Assumption 3. It holds that $Q_n(\theta)$ converges to $Q(\theta)$ with probability 1 as $n \to \infty$ uniformly on B.

Assumption 4. As $n \to \infty$, the sequence $V_n(\theta) = n^{1/2}\{Q(\theta) - Q_n(\theta)\}$ converges in distribution to a random element $V(\theta) \in C(B)$, where $V(\theta)$ is a Gaussian process with mean 0 and variance $\sigma^2(\theta) \in C(B)$.

These assumptions are jointly sufficient to achieve consistency and asymptotic normality of ν_n, which we state formally in the following two propositions.

Proposition 1. *Let $\vartheta_n \in \arg\min\{Q_n(\theta) : \theta \in \Theta\}$ be a sample solution, and let ν_n be defined as in* (4)*. Under Assumptions* 1*,* 2 *and* 3*, $\nu_n \to \nu$ and $\vartheta_n \to \vartheta$ with probability* 1.

Proposition 1 is identical to Theorem 5.3 of Shapiro et al. (2009) under the condition that ϑ is unique.

Proposition 2. *Under Assumptions* 1*,* 2 *and* 4*,*

$$n^{1/2}(\nu_n - \nu) \to \mathcal{N}\{0, \sigma^2(\vartheta)\}$$

in distribution, where $\sigma^2(\vartheta)$ is the asymptotic variance of ν_n defined in Assumption 4.

Proposition 2 is an immediate consequence of Theorem 3.2 of Shapiro (1991). Although we do not restate the proof here, the intuition is that Assumptions 1 and 2 allow a notion of differentiability of the infimum, and Assumption 4 provides weak convergence of $n^{1/2}(Q - Q_n)$ to a Gaussian process, thus providing the conditions needed for an application of the delta method.

To use Proposition 2 to construct a valid confidence interval, we must take into consideration that both σ^2 and ϑ are unknown. To this end, we state an additional assumption followed by a proposition.

Assumption 5. There exists a uniformly strongly consistent estimator $\sigma_n^2(\theta) \in C(B)$ for $\sigma^2(\theta)$ such that $\sup_{\theta \in \Theta} |\sigma_n^2(\theta) - \sigma^2(\theta)| \to 0$ with probability 1.

Proposition 3. *Under Assumptions 1, 2, 3 and 5, $\sigma_n^2(\vartheta_n) \to \sigma^2(\vartheta)$ with probability* 1.

Assumption 5 applies uniform convergence to an estimator for the asymptotic variance of $Q_n(\theta)$. This strong notion of convergence for $\sigma_n^2(\vartheta_n)$ allows us to construct a confidence bound of the form

$$C_n = \nu_n - Z_\alpha \sigma_n(\vartheta_n) n^{-1/2}, \tag{5}$$

where Z_α is the upper α-quantile of the standard normal distribution. This choice of C_n has asymptotically exact nominal coverage $1 - \alpha$ by Proposition 2, Proposition 3 and Slutsky's theorem.

2.2. *Confidence intervals under sample constraints*

We now consider the more difficult setting where the constraint functions $h_j(\theta)$ need to be estimated as well. We instead observe an estimator $\Theta_n = \{\theta : h_{nj}(\theta) \leqslant 0, j = 1, \ldots, J\}$ comprised of estimators of the constraint functions $h_{nj}(\theta)$. We discuss what properties Θ_n must have to allow valid statistical inference for ν.

It is tempting to follow the approach of the previous section and construct a plug-in estimator for ν by simply replacing Q with Q_n, and Θ with Θ_n, and finding the corresponding infimum. This is the approach taken by Shapiro et al. (2009), given by ν_n^p in (3). A problem with this approach is that it is possible that $\Theta \cap \Theta_n = 0$ with probability 1 even as n becomes large. This means that the true solution ϑ will almost never lie inside Θ_n, prohibiting the construction of a valid confidence interval for ν, as illustrated by the contrived example below.

Example 1. Consider a problem of the form $Q(\theta) = \theta^2 + E(X)$ and $\Theta = \{\theta : \theta = E(X)\}$, where $X \sim \mathcal{N}(1, 1)$ is a normally distributed random variable. The plug-in estimators are $Q_n(\theta) = \theta^2 + \bar{X}_n$ and $\Theta_n = \{\theta : \theta = \bar{X}_n\}$, where $\bar{X}_n$ is the mean of n independent and identically distributed draws of X. It follows that $\nu = 2$ and $\nu_n^p = \bar{X}_n^2 + \bar{X}_n$, where ν_n^p is the plug-in estimator in (3). The asymptotic variance of $Q_n(\theta)$ is $\sigma^2(\theta) = 1$, which we assume is known. The confidence bound in (2) is $C_n = \bar{X}_n^2 + \bar{X}_n - Z_\alpha n^{-1/2}$. A simple Monte Carlo simulation demonstrates that the corresponding 95% confidence interval for $n = 100$ exhibits subnominal coverage of around 70%.

Existing approaches in stochastic optimization address the problem in Example 1 by restricting to sample average approximations, and imposing convexity of both Q and h. To allow inference for a broader class of problems, we propose an intuitive but conservative approach that replaces Θ_n with an appropriate relaxation. In particular, we propose to use the relaxed set

$$\Theta_n^r = \{\theta : h_{nj}(\theta) \leqslant \epsilon_{nj}(\theta), j = 1, \ldots, J\}, \tag{6}$$

where $\epsilon_n(\theta) = \{\epsilon_{n1}(\theta), \ldots, \epsilon_{nJ}(\theta)\}^{\mathrm{T}}$ is some J-dimensional sequence such that $\epsilon_{nj}(\theta) \geqslant 0$ for all $\theta \in B$, chosen so that

$$\lim_{n\to\infty} \mathrm{pr}(\Theta \subseteq \Theta_n^r) \geqslant 1 - \alpha_1 \tag{7}$$

for some $0 < \alpha_1 < 1$. The exact forms of Θ_n^r and $\epsilon_n(\theta)$ are not crucial for our main results, provided (7) holds, which we discuss in more detail toward the end of this section.

Our proposed confidence bound is of the form $C_n(\theta) = Q_n(\theta) - Z_{\alpha_2} \sigma_n(\theta) n^{-1/2}$ for some $0 < \alpha_2 < 1$, where Z_{α_2} is the upper α_2-quantile of the standard normal distribution. We need to select a θ so that (2) is satisfied. This is accomplished by finding the optimal value and solution over the relaxed constraint set,

$$\nu_n^r = \inf\{Q_n(\theta) : \theta \in \Theta_n^r\} \quad \text{and} \quad \vartheta_n^r \in \arg\min\{Q_n(\theta) : \theta \in \Theta_n^r\}, \tag{8}$$

and constructing a confidence bound of the form

$$C_n = C_n(\vartheta_n^r) = \nu_n^r - Z_{\alpha_2}\sigma_n(\vartheta_n^r)n^{-1/2}. \tag{9}$$

We now need to demonstrate that C_n covers ν with known probability in the limit. To this end, we need an additional technical assumption to hold.

Assumption 6. Let $\zeta_n^r \in \arg\min\{C_n(\theta) : \theta \in \Theta_n^r\}$ be the optimal solution of $C_n(\theta)$ over Θ_n^r. Then $|\zeta_n^r - \vartheta_n^r|$ converges to 0 in probability.

Assumption 6 is imposed so that two important quantities become asymptotically close. The first quantity is $C_n(\zeta_n^r)$, which is the infimum over all confidence bounds in Θ_n^r. This confidence bound is important because it provides a lower bound for other quantities with known coverage probabilities, which is a fact we utilize in our main result in Theorem 1. The second quantity is $C_n(\vartheta_n^r)$, which is our main confidence bound proposed in (9).

We argue that Assumption 6 is reasonable in the sense that ζ_n^r and ϑ_n^r are solutions over two objective functions that converge uniformly to the same limit. To make this intuition more concrete, we provide some sufficient conditions for Assumption 6 in the Supplementary Material. Essentially, if Assumption 3 is satisfied, and $h_{nj}(\theta)$ and $\epsilon_{nj}(\theta)$ converge to $h_j(\theta)$ and 0 for all $j = 1, \ldots, J$ uniformly on B with probability 1, then we can show that both ϑ_n^r and ζ_n^r converge to ϑ with probability 1.

We claim that C_n provides an asymptotically valid lower confidence bound.

THEOREM 1. *Suppose that we select a relaxed constraint set* Θ_n^r *as in* (6) *and significance level* $0 < \alpha_1 < 1$ *such that* $\lim_{n\to\infty} \mathrm{pr}(\Theta \subseteq \Theta_n^r) \geqslant 1 - \alpha_1$. *Suppose that we also select a significance level* $0 < \alpha_2 < 1$. *Then, under Assumptions* 1–6,

$$\lim_{n\to\infty} \mathrm{pr}(C_n \leqslant \nu) \geqslant 1 - \alpha_1 - \alpha_2.$$

Here we outline the key steps in the proof. We begin by defining a deterministic sequence $\delta_n = Z_{\alpha_2}\epsilon n^{-1/2}$, where $\epsilon > 0$ is some small constant. We then show that $\mathrm{pr}(C_n \leqslant \nu)$ is bounded from below by the sum of two quantities: $\mathrm{pr}\{C_n(\zeta_n^r) \leqslant \nu - \delta_n\}$ and $\mathrm{pr}\{|\sigma_n(\zeta_n^r) - \sigma_n(\vartheta_n^r)| \leqslant \epsilon\} - 1$. The second quantity converges to 0 under Assumption 6. The remainder of the proof follows a similar argument to the main lemma of Berger & Boos (1994). Whenever $\Theta \subseteq \Theta_n^r$, we know that $C_n(\zeta_n^r)$, which is the infimum over all confidence bounds in Θ_n^r, will cover ν at least as often as $C_n(\vartheta_n)$, which is confidence bound (5). Therefore, $\mathrm{pr}\{C_n(\zeta_n^r) \leqslant \nu, \Theta \subseteq \Theta_n^r\} \geqslant \mathrm{pr}\{C_n(\vartheta_n) \leqslant \nu, \Theta \subseteq \Theta_n^r\}$. We also know that $\mathrm{pr}\{C_n(\vartheta_n) \leqslant \nu, \Theta \subseteq \Theta_n^r\} = \mathrm{pr}\{C_n(\vartheta_n) \leqslant \nu\} - \mathrm{pr}\{C_n(\vartheta_n) \leqslant \nu, \Theta \nsubseteq \Theta_n^r\}$ by the law of total probability. In the limit, the first probability on the right-hand side is equal to $1 - \alpha_2$ by Proposition 2 and the second probability is at most α_1 by assumption. This allows us to arrive at our main result.

The proof sketch also provides some insight into why the naive plug-in estimator ν_n^p defined in (3) may fail to yield a valid confidence interval. A crucial quantity is $\mathrm{pr}(\Theta \subseteq \Theta_n^r)$, which is known under an appropriate choice of $\epsilon_n(\theta)$. The corresponding quantity for the plug-in estimator is $\mathrm{pr}(\Theta \subseteq \Theta_n)$, which could be arbitrarily small. In Example 1, this probability is zero.

It remains to discuss how to construct a relaxed set Θ_n^r. Whenever Θ can be characterized by a set of moment inequalities, such that $h_j(\theta) = E\{m_j(\theta)\}$, the moment inequalities literature summarized in § 1.3 could be used to construct Θ_n^r. A more conservative relaxed set could be constructed via an application of the intersection bound. Suppose that the following assumption holds on the constraint functions.

Assumption 7. For all $\theta \in \Theta$ and $j = 1, \ldots, J$, $n^{1/2}\{h_{nj}(\theta) - h_j(\theta)\} \to \mathcal{N}\{0, \sigma_j^2(\theta)\}$ in distribution and $\sigma_{nj}^2(\theta)$ is a consistent estimator for $\sigma_j^2(\theta)$.

This fairly weak assumption means that $h_{nj}(\theta)$ is pointwise asymptotically normally distributed and that there is a consistent estimator for the variance. This assumption allows us to select

$$\epsilon_n(\theta) = Z_{\alpha_{1j}} \sigma_{nj}(\theta) n^{-1/2},$$

where $\alpha_1 = \alpha_{11} + \alpha_{12} + \cdots + \alpha_{1J}$. It is straightforward to show that this choice of $\epsilon_n(\theta)$ satisfies (7). We could shrink the size of Θ_n^r by assuming that $h_n(\theta) = \{h_{1,n}(\theta), \ldots, h_{nj}(\theta)\}^{\mathrm{T}}$ converges pointwise to a multivariate Gaussian with covariance matrix Σ and consistent estimator Σ_n. This would allow us to construct Θ_n^r as an ellipsoid confidence region.

Remark 1. It remains to discuss how one would select α_1 and α_2. As a rule of thumb, we typically choose the midpoint $\alpha_1 = \alpha_2 = \alpha/2$. It is tempting to select $\alpha = \alpha_1 + \alpha_2$ and choose C_n as the largest confidence bound over all α_1 and α_2 satisfying this equality. This would mean that α_1 and α_2 are sample-dependent quantities and so Theorem 1 will not directly apply. However, we can reason heuristically that the best choice of α_1 and α_2 should lie at an interior point $\alpha_1 > 0$ and $\alpha_2 > 0$. For a fixed sample, as $\alpha_1 \to 0$ and $\alpha_2 \to \alpha$, $\Theta_n^r \to B$, and thus C_n approaches the $100(1-\alpha)\%$ confidence interval over the unconstrained problem. As $\alpha_1 \to \alpha$ and $\alpha_2 \to 0$, $C_n \to -\infty$, and thus the confidence interval becomes arbitrarily wide.

Remark 2. So far, we have focused on inference for the infimum; however, partial identification problems are often characterized by an identified set of the form $I = [\nu^l, \nu^u]$, where $\nu^l = \inf\{Q(\theta) : \theta \in \Theta\}$ and $\nu^u = \sup\{Q(\theta) : \theta \in \Theta\}$ (Imbens & Manski, 2004; Chernozhukov et al., 2013). Suppose that Θ_n^r is chosen so that $\mathrm{pr}(\Theta \subseteq \Theta_n^r) \geqslant 1 - \alpha_1/2$. Moreover, let $\nu_n^{r,l} = \inf\{Q_n(\theta) : \theta \in \Theta_n^r\}$ and $\nu_n^{r,u} = \sup\{Q_n(\theta) : \theta \in \Theta_n^r\}$ denote the optimal values, and let $\vartheta_n^{r,l}$ and $\vartheta_n^{r,u}$ denote the corresponding optimal solutions. The estimated interval can be written as $[\nu_n^{r,l}, \nu_n^{r,u}]$, and we can construct a confidence interval by combining the lower confidence bound for ν^l and the upper confidence bound for ν^u, so that

$$[\nu_n^{r,l} - Z_{\alpha_2/2}\sigma_n(\vartheta_n^{r,l})n^{-1/2}, \nu_n^{r,u} + Z_{\alpha_2/2}\sigma_n(\vartheta_n^{r,u})n^{-1/2}]$$

will cover I with probability at least $1 - \alpha_1 - \alpha_2$. This is the two-sided analogue of the one-sided confidence interval proposed in (9) and Theorem 1.

3. Sensitivity analysis via a logistic model

3.1. *Set-up*

We now return to the motivating example of selection bias in population-based cohort studies briefly described in § 1.2. Specifically, we generalize the sensitivity analysis proposed by Thompson & Arah (2014), who defined a logistic model for the probability of sample selection and proposed to select parameters based on domain knowledge, or enumerate a large number of possible parameters. This approach is challenging to implement in the presence of complicated selection mechanisms with many parameters. Plausible sets of parameters that introduce bias in estimates of interest may be overlooked. Therefore, we begin by framing Thompson & Arah (2014) as an optimization problem over a space of plausible parameters, and describe how relevant auxiliary information could be introduced to further restrict the parameter space and provide more informative bounds, such as survey response rates, population means and negative controls. An additional sensitivity analysis, Aronow & Lee (2013), is generalized in the Supplementary Material.

Consider an independent and identically distributed draw of size N from an infinite population. For concreteness, we can think of this finite draw as the set of individuals who are eligible to enter the sample. Let $S_i \in \{0,1\}$ be a selection indicator for whether individual i enrols in the sample, where $S_i = 1$ indicates sample participation, and let the observed sample size be denoted by $n = \sum_{i=1}^{N} S_i$. For notational convenience, we assume that $S_1 = \cdots = S_n = 1$ and $S_{n+1} = \cdots = S_N = 0$.

Within the observed sample, we observe a vector of variables related to sample selection $W_i \in \mathbb{R}^K$. As in Thompson & Arah (2014), we assume that the probability of sample selection admits a logistic form,

$$e(W_i;\theta) = \mathrm{pr}(S_i = 1 \mid W_i) = \frac{\exp(\theta_0 + \theta_1 W_{i1} + \cdots + \theta_k W_{iK})}{1 + \exp(\theta_0 + \theta_1 W_{i1} + \cdots + \theta_k W_{iK})}, \tag{10}$$

where $\theta = (\theta_0, \theta_1, \theta_2, \ldots, \theta_K)^{\mathrm{T}}$. We further assume that the sample is generated by some true selection probabilities $e(W_i;\theta^*)$ parameterized by θ^*.

For illustration, suppose that our object of interest is the population mean of a random variable X_i. We can write the population mean, and corresponding sample mean, in terms of θ^* as

$$\beta(\theta^*) = E(X_i) = \frac{E\{X_i/e(W_i;\theta^*) \mid S_i = 1\}}{E\{1/e(W_i;\theta^*) \mid S_i = 1\}}, \qquad \beta_n(\theta^*) = \frac{\sum_{i=1}^{n} X_i/e(W_i;\theta^*)}{\sum_{i=1}^{n} 1/e(W_i;\theta^*)}. \tag{11}$$

The expression in (11) relies on $X_i \perp S_i \mid W_i$, which we assume throughout.

Since we only observe X_i when $S_i = 1$, the true parameter θ^* cannot be estimated. Thompson & Arah (2014) proposed to consider a space of plausible values for θ^*, and identified a worst-case lower bound and worst-case upper bound for $\beta_n(\theta^*)$. Inference to $\beta(\theta^*)$ itself was not considered. Formally, we select a parameter space Θ in which we are confident that θ^* resides. We then take the infimum and supremum of $\beta_n(\theta)$ over the space Θ.

3.2. *Sensitivity parameters*

Since we have assumed a logistic form for the selection probabilities (10), we can select sensitivity parameters that have a natural interpretation in terms of odds ratios.

Without loss of generality, suppose that each W_{ik} has mean zero and standard deviation one within the sample. We can then choose a parameter $\Lambda_1 \geqslant 1$ such that

$$\Lambda_1^{-1} \leqslant \exp(\theta_k) \leqslant \Lambda_1 \quad (k = 1, \ldots, K). \tag{12}$$

We can interpret Λ_1 as the change in the conditional odds of sample selection from a one standard deviation increase in W_{ik}, holding all else fixed. When $\Lambda_1 = 1$, sample selection is completely random. Of course, we could select sensitivity parameters Λ_{1k} on a variable-by-variable basis for $k = 1, \dots, K$, although choosing a single $\Lambda_1 = \max_k \Lambda_{1k}$ simplifies the interpretation of the sensitivity analysis.

The intercept term θ_0 also needs to be bounded. We can choose two parameters $\Lambda_0^l, \Lambda_0^u \in (0, 1)$ such that

$$\Lambda_0^l \leqslant \exp(\theta_0) \leqslant \Lambda_0^u, \qquad (13)$$

which can be interpreted as the odds of sample selection among those for whom $W_{ik} = 0$ for all k.

Rearranging (12) and (13) shows that the sensitivity parameters $(\Lambda_0^l, \Lambda_0^u, \Lambda_1)$ characterize a compact subset of $\mathbb{R}^{K+1}$:

$$\theta \in \Theta = [\log(\Lambda_0^l), \log(\Lambda_0^u)] \times [\log(1/\Lambda_1), \log(\Lambda_1)]^K. \qquad (14)$$

From here, the estimand and estimator can be respectively defined for the worst-case lower bound of $\beta(\theta)$ as

$$\nu = \inf\{\beta(\theta) : \theta \in \Theta\}, \qquad \nu_n = \inf\{\beta_n(\theta) : \theta \in \Theta\}.$$

We could of course estimate the worst-case upper bound for $\beta(\theta)$ by taking the supremum of $\beta_n(\theta)$ over Θ; see Remark 2. Naturally, we can also consider estimators other than sample means, such as ordinary least squares or two-stage least squares.

3.3. *Auxiliary information constraints*

We now introduce several common examples, where there may be discordance between known population quantities and quantities implied by the inverse probability weights. In general, provided we can formulate the constraints as a statistical test with a known null distribution, they can be placed within our framework.

Example 2. Suppose that we know the response rate for a survey-based sample $r = E\{e(W_i; \theta^*)\}$. It is straightforward to show that $E\{1/e(W_i; \theta^*) \mid S_i = 1\} = 1/r$. This means that the within-sample expectation of the true inverse selection probabilities is equal to the inverse response rate. We therefore only want to consider parameters θ that imply this inverse response rate. The corresponding constraint can be written as

$$h_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{1/e(W_i; \theta) - 1/r\} \leqslant Z_{\alpha_{1j}/2}\sigma_{nj}(\theta)/n^{1/2},$$

where $\sigma_{nj}(\theta)$ is the sample standard deviation of $1/e(W_i; \theta)$.

Example 3. Suppose that we know the population mean $E(W_{ik})$ of some $W_{ik} \in W_i$. The inverse-probability-weighted sample mean of W_{ik} should therefore equal this mean in expectation, since

$$\frac{E\{W_{ik}/e(W_i; \theta^*) \mid S_i = 1\}}{E\{1/e(W_i; \theta^*) \mid S_i = 1\}} = E(W_{ik}).$$

This is conceptually similar to the raking procedure in survey sampling (Deming & Stephan, 1940), which adjusts sampling weights to match known marginal totals. The covariate mean constraint can be written as

$$h_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{W_{ik} - E(W_{ik})\}/e(W_i;\theta) \leqslant Z_{\alpha_{1j}/2}\sigma_{nj}(\theta)/n^{1/2},$$

where $\sigma_{nj}(\theta)$ is the sample standard deviation of $\{W_{ik} - E(W_{ik})\}/e(W_i;\theta)$.

Example 4. Suppose that we are confident that higher values of W_{ik} are associated with an increased probability of sample selection. For example, W_{ik} could be years of education and we might know from comparisons with representative samples, such as the census, that better educated individuals are more likely to select into the sample, conditional on other selection variables, so that $\theta_j \geqslant 0$ a priori.

Example 5. Suppose we know that two variables W_{ik} and $W_{ik'}$ are uncorrelated in the population. The inverse-probability-weighted correlation between W_{ik} and $W_{ik'}$ should therefore be zero. For example, due to the independent assortment of chromosomes, biological sex and autosomal genetic variants should be independent in the population; however, Pirastu et al. (2021) demonstrated that there is significant correlation within the UK Biobank. This constraint can be formulated in several ways, for example by fixing the regression coefficient of W_{ik} on $W_{ik'}$ to be zero.

Examples 2, 3 and 5 are two-sided constraints such that we also want these inequalities to hold for $-h_{nj}(\theta)$.

Remark 3. In the population means setting, Miratrix et al. (2018) demonstrated how to place shape constraints on the weighted empirical distribution of the response. Their approach involves constructing the worst-case weighted distribution given the Aronow & Lee (2013) bounding assumptions; see the Supplementary Material. This results in a set that contains the oracle weighted distribution with probability approaching 1. Provided we have a valid test, we can implement shape constraints within our framework without the need to characterize the worst-case weighted distribution. In the simplest case, we might want a variable to follow a known distribution in the population. For example, the distribution of IQ scores should be normal with mean 100 and standard deviation 15, which is a stronger constraint than Example 3. This could be formulated as a Kolmogorov–Smirnov test, and the relaxed constraint set could be constructed via the null distribution of that test.

4. Simulations

The aim of these simulations is to provide a brief assessment of the finite sample and limiting properties of the inference procedure described in § 2. For concreteness, we simulate the sensitivity analysis for selection bias described in § 3. Our parameter $\beta(\theta)$ and estimator $\beta_n(\theta)$ are both the coefficient of a weighted linear regression. In particular, a regression of Y_i on X_i for $(X_i, Y_i) \sim \mathcal{N}(0, I_2)$, where I_2 is the identity matrix. The weights take the form in (10) with variables $W_i = (X_i, Y_i)$.

We consider three distinct scenarios for the constraints. In the first scenario, we impose only sensitivity parameters $\Lambda_0^l = 0.11$, $\Lambda_0^u = 0.25$ and $\Lambda_1 = 3$. In the second scenario, we also

Table 1. *Coverage frequency for the three scenarios over 5000 Monte Carlo replications*

	Sample size						
Scenario	10	25	50	100	200	500	1000
1	0.972	0.992	0.995	0.997	0.998	0.996	0.995
2	0.936	0.974	0.981	0.983	0.979	0.966	0.947
3	0.953	0.985	0.991	0.991	0.987	0.986	0.979

impose a direction constraint $\theta_1 \geqslant 0$ as in Example 4. In the third scenario, we impose both the previous direction constraint, and set the response rate equal to 0.15 as in Example 2. In each scenario, we use the discussion in Remark 2 to construct a two-sided 95% confidence interval for the identified set $I = [\nu^l, \nu^u]$, where $\nu^l = \inf\{\beta(\theta) : \theta \in \Theta\}$, $\nu^u = \sup\{\beta(\theta) : \theta \in \Theta\}$ and Θ takes the form in (14). The first and second scenarios have no sample constraints and so the confidence interval corresponds to that in (5). The third scenario employs the confidence interval proposed in (9) and Theorem 1.

Each scenario has distinct properties. In the first scenario, there are two solutions to the population optimization problems, thus violating Assumption 1. In the second scenario, the addition of a direction constraint rules out one of the two solutions and satisfies Assumption 1. In the third scenario, the introduction of a sample constraint necessitates the use of our relaxed confidence bound. In this scenario, we use our rule of thumb from Remark 1 to select $\alpha_1 = \alpha_2 = 0.025$ for both the upper and lower bounds of the two-sided confidence interval.

Table 1 summarizes the results and broadly aligns with our theoretical predictions. The first scenario violates Assumption 1 and the impact of this violation is substantial overcoverage of the confidence interval. Intuitively, this occurs because the sample solution will occur at, or near, the population solution that happens to minimize $\beta_n(\theta)$ in any given sample, which will result in a systematically wider confidence interval. The second scenario satisfies all assumptions for Proposition 2 and therefore converges to exact nominal coverage. The third scenario imposes a sample constraint and exhibits some overcoverage. This overcoverage can occur because our confidence bound in Theorem 1 sidesteps the covariance between the constraints $h_{nj}(\theta)$ and objective function $Q_n(\theta)$, instead imposing a worst-case intersection bound.

In this simulation exercise, the weight model is comprised of two variables. The Supplementary Material contains an additional simulation exploring the computation time of our `R` package `selectioninterval` (R Development Core Team, 2022) as the number of variables in the weight model increases.

5. Applied example: effect of education on income

We consider an instrumental variable design looking at the effect of education on income in the UK Biobank cohort. Our instrument is exposure to an educational reform taking place in England in 1972. Our exposure is whether an individual remained in school until at least age 16, and our outcome is whether an individual earned more than £31 000 per year in 2006. We restrict our sample to individuals who turned 15 within 12 months of September 1972, and we control for sex and month-of-birth indicators. The unweighted estimate is 0.18 (95% confidence interval 0.08–0.28). An in-depth exposition of this design can be found in Davies et al. (2018) and the Supplementary Material.

We first apply the sensitivity analysis described in § 3.2 without auxiliary constraints, where the probability weights contain sex, years of education, income, age, days of physical activity per week, and an interaction term between education and income. We choose sensitivity parameters $\Lambda_0^l = 0.02$, $\Lambda_0^u = 0.25$ and $\Lambda_1 = 2$, so that the average individual in the sample has an odds of

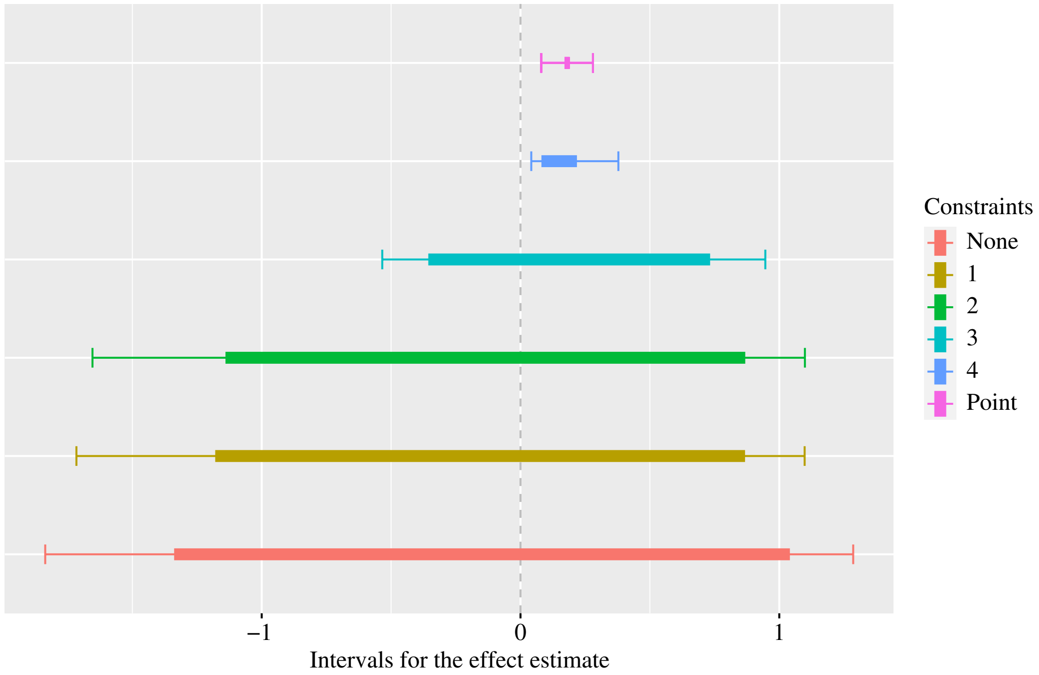

Fig. 1. Estimated intervals (thick lines) and corresponding confidence intervals (thin lines) for effect estimates in the applied example. Point represents the unweighted point estimate. Each constraint is added sequentially. No constraint means that only the sensitivity parameters $\Lambda_0^l = 0.02$, $\Lambda_0^u = 0.25$ and $\Lambda_1 = 2$ are imposed. Constraint 1 sets the response rate equal to 5.5%. Constraint 2 sets the proportion of males in the population to be 49.5%. Constraint 3 sets the proportion of households earning more than £31 000 to be 21%. Constraint 4 sets the average age of individuals to be 48.98 years.

sample selection between 0.02 and 0.25, and each variable in the model can induce a marginal odds of sample selection between 0.5 and 2. Our sensitivity analysis suggests that the effect estimate lies in the interval $[-1.34, 0.94]$ (95% confidence interval $[-1.84, 1.29]$). This interval is completely uninformative as it spans the full range of possible estimates.

One explanation for this conservativeness is that this simple sensitivity analysis does not utilize all of the information available to us on the target population and the sample selection mechanism. The minimizing (maximizing) weights corresponding to this interval imply that the proportion of males in the population is 38.52% (46.6%), and the proportion of households with a gross income greater than £31 000 is 95.84% (95.66%), all of which are inconsistent with known characteristics of the U.K. population.

To address this incongruence, we consider four constraints that are typical of the information available to applied researchers using datasets such as the UK Biobank. The first constraint is the response rate of the UK Biobank (5.5%), which is the proportion of individuals who entered the cohort after receiving an invitation. The second constraint is the proportion of males in the U.K. population within the UK Biobank age range of 40–69 (49.5%). The third is the proportion of U.K. households earning more than £31 000 per year at the date of UK Biobank recruitment in 2006 (21%). The fourth is the average age of individuals within our two-year age bracket (48.98). All statistics were obtained from publicly available records from the U.K.'s Office of National Statistics.

Figure 1 shows the resulting estimated intervals (thicker lines) and their corresponding confidence intervals (thinner lines) where each constraint is added sequentially. The estimated intervals and confidence intervals correspond to those described in Remark 2. We can see that each additional constraint reduces the width of the interval, with constraints 3 and 4, the household income and age constraints, respectively, seemingly having the largest marginal impact. The top interval includes all constraints, and is quite informative for the desired effect, rejecting the null and suggesting an effect estimate in the range 0.08–0.22 (95% confidence interval 0.04–0.38). The unweighted estimate still lies within this interval, but our sensitivity analysis suggests some increased uncertainty in the range of effect estimates. These results also suggest that, despite the potential conservativeness of the confidence interval in Theorem 1, it can still produce informative bounds in practice.

6. Discussion

There has been some existing work on bootstrap inference for Rosenbaum-type sensitivity analyses (Zhao et al., 2019). This approach considers a fixed parameter space Θ. It is unclear how to select the relaxation parameter $\epsilon_n(\theta)$ in a bootstrap analogue of our method under estimated constraints. Simple approaches, such as constructing Θ_n^r via asymptotic approximations and then bootstrapping the distribution of ν_n^r, are plausible, but their statistical properties remain to be explored.

In some instances, including our selection bias application in § 3, the target of inference is $Q(\theta^*)$, where θ^* is some true parameter lying within Θ, rather than ν. Suppose that Θ is known and that we have a two-sided identified set $[\inf_{\theta\in\Theta} Q(\theta), \sup_{\theta\in\Theta} Q(\theta)]$ as in Remark 2; then if $Q(\theta^*)$ lies near the boundary of this set, and the set has positive width, the noncoverage probability of the corresponding confidence interval is effectively one-sided in the limit. A naive two-sided confidence interval constructed around the identified set may be too conservative. Imbens & Manski (2004) discussed approaches for maintaining uniform coverage of $Q(\theta^*)$. The central limit theorem established by Shapiro (1991) for known Θ is amenable to their framework, although, to our knowledge, has not been formally used in this setting; extending this result to sample-constrained problems would be a valuable contribution. Stoye (2009) further extended the work of Imbens & Manski (2004) by developing confidence intervals that exhibit uniform coverage for $Q(\theta^*)$, without relying on assumed superefficiency of the estimated interval width.

A final consideration is the computational burden of our approach. Our general inference procedure in § 2 relies on the optimization problems in (8) being solvable, but the computational complexity of these problems will vary, depending on the application. Our `R` package `selectioninterval` relies on out-of-the-box global and local optimization algorithms. There are no theoretical guarantees of convergence to the global optimum; however, we have not observed a failure of convergence in our simulations.

Acknowledgement

The authors thank the referees and associate editor for their helpful comments, and Dr Shaun Seaman for his insightful comments about our framing of the selection problem in § 3. This research was conducted using the UK Biobank Resource. UK Biobank received ethical approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382). This research was approved as part of application 8786. Tudball acknowledges financial support from the Wellcome Trust. Bowden was supported by an Expanding Excellence in England (E3) award from the University of Exeter. Tilling works in the MRC Integrative Epidemiology Unit, which is supported by the University of Bristol and the Medical Research Council. Hughes was supported by a Sir Henry Dale Fellowship.

Supplementary material

Supplementary Material includes proofs for the results in § 2, further details for the applied example in § 5 and an extension of Aronow & Lee (2013). An R package to implement the sensitivity analysis described in § 3 is available at `https://github.com/matt-tudball/selectioninterval`.

References

Andrews, D. W. K. & Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica* **81**, 609–66.

Andrews, D. W. K. & Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica* **78**, 119–57.

Aronow, P. M. & Lee, D. K. (2013). Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika* **100**, 235–40.

Bareinboim, E., Tian, J. & Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proc. 28th AAAI Conf. Artif. Intel.*, pp. 2410–6. Palo Alto, CA: AAAI Press.

Berger, R. L. & Boos, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *J. Am. Statist. Assoc.* **89**, 1012–16.

Chernozhukov, V., Hong, H. & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica* **75**, 1243–84.

Chernozhukov, V., Lee, S. & Rosen, A. M. (2013). Intersection bounds: estimation and inference. *Econometrica* **81**, 667–737.

Davies, N. M., Dickson, M., Davey Smith, G., Van Den Berg, G. J. & Windmeijer, F. (2018). The causal effects of education on health outcomes in the UK Biobank. *Nature Hum. Behav.* **2**, 117–25.

Deming, W. E. & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11**, 427–44.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R. & Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–34.

Horvitz, D. G. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **44**, 663–85.

Hughes, R. A., Davies, N. M., Davey Smith, G. & Tilling, K. (2019). Selection bias when estimating average treatment effects using one-sample instrumental variable analysis. *Epidemiol.* **30**, 350–7.

Imbens, G. W. & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72**, 1845–57.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer.

Miratrix, L. W., Wager, S. & Zubizarreta, J. R. (2018). Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika* **105**, 103–14.

Molinari, F. (2020). Microeconometrics with partial identification. In *Handbook of Econometrics*, vol. 7, S. N. Durlauf, L. P. Hansen, J. J. Heckman & R. L. Matzkin, eds. Amsterdam: Elsevier, pp. 355–486.

Pirastu, N., Cordioli, M., Nandakumar, P., Mignogna, G., Abdellaoui, A., Hollis, B., Kanai, M., Rajagopal, V. M., Parolo, P. D. B., Baya, N. et al. (2021). Genetic analyses identify widespread sex-differential participation bias. *Nature Genet.* **53**, 663–71.

R Development Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Ann. Oper. Res.* **30**, 169–86.

Shapiro, A., Dentcheva, D. & Ruszczyński, A. (2009). *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica* **77**, 1299–315.

Stuart, E. A., Cole, S. R., Bradshaw, C. P. & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Statist. Soc.* A **174**, 369–86.

Thompson, C. A. & Arah, O. A. (2014). Selection bias modeling using observed data augmented with imputed record-level probabilities. *Ann. Epidemiol.* **24**, 747–53.

Wang, W. & Ahmed, S. (2008). Sample average approximation of expected value constrained stochastic programs. *Oper. Res. Lett.* **36**, 515–19.

Zhao, Q., Small, D. S. & Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J. R. Statist. Soc.* B **81**, 1–27.

[*Received on* 1 *April* 2020. *Editorial decision on* 11 *July* 2022]