Sampling Variation, Monotone Instrumental Variables and the Bootstrap Bias Correction

Qian, Hang
Iowa State University

August 2011

Sampling Variation, Monotone Instrumental Variables and the Bootstrap Bias Correction

Hang Qian

Abstract

This paper discusses the finite sample bias of analogue bounds under the monotone instrumental variables assumption. By analyzing the bias function, we first propose a conservative estimator which is biased downwards (upwards) when the analogue estimator is biased upwards (downwards). Using the bias function, we then show the mechanism of the parametric bootstrap correction procedure, which can reduce but not eliminate the bias, and there is also a possibility of overcorrection.This motivates us to propose a simultaneous multi-level bootstrap procedure so as to further correct the remaining bias. The procedure is justified under the assumption that the bias function can be well approximated by a polynomial. Our multi-level bootstrap algorithm is feasible and does not suffer from the curse of dimensionality. Monte Carlo evidence supports the usefulness of this approach and we apply it to the disability misreporting problem studied by Kreider and Pepper (2007).

Keywords: Monotone instrumental variables, Bootstrap, Bias correction.

1. Introduction

Proposed by Manski and Pepper (2000), Monotone instrumental variables (MIV) is a powerful tool for treatment response identification. The MIV assumption weakens the traditional instrumental variable assumption by a weak inequality of mean response across sub-populations. As a result, the MIV sharp lower bound invariably involves a supremum operator and the upper bound contains an infimum operator.

However, when sampling variation is taken into account, the bounds themselves assume randomness since the population moments or probabilities are replaced by their analogues. Though the analogue estimates are still consistent, finite sample bias is a serious concern. As is noted by Manski and Pepper (2009, p.211), "the sup and inf operations ... significantly complicate the bounds under other MIV assumptions, rendering it difficult to analyze the sampling behavior of analogue estimates."[1] The major statistical problem is that the analogue estimate of the lower bound is biased upwards and upper bound biased downwards, resulting in the estimates narrower than the true bounds.

To address this concern, two major lines of research are present in the literature to our best knowledge. One is direct adjustment. Chernozhukov et al. (2009) develop an inference method on intersection bounds with a continuum of inequalities. Their estimator maximizes or minimizes the precision-

[1]The bounds under the monontone treatment selection assumption have simple forms, but under other MIV assumptions the supremum and infimum operators will appear in the bounds.

corrected curve defined by the analogue estimates plus a critical value multiplied by pointwise standard errors. Another solution is bootstrap adustment. Kreider and Pepper (2007) propose a heuristic bootstrap bias correction and applied this approach to their employment gap identification problems. Though Monte Carlo experiments in Manski and Pepper (2009) provide evidence on the effectiveness of bias reduction, theoretical foundation has not been established to justify the bootstrap correction. In addition, the simulation results of Manski and Pepper (2009) show that in some cases moderate biases remain after the correction.

The goal of this paper is to justify the bootstrap bias correction. Traditionally, the improvement of the corrected estimator is in the sense of asymptotic refinement. That is, we expect the bootstrap corrected estimator has a bias going to zero at a faster rate than the uncorrected estimator. However, there are difficulties applying asymptotic expansion techniques to our problem, since the bounds under the MIV assumption are not differentiable. In this paper, we take an innovative, and perhaps more direct, approach to study bootstrap bias reduction. We rely on asymptotic normality of the estimators to derive our results. Given normally distributed variates, we bound the magnitude of the upward bias induced by the $\max(\cdot)$ operator and show how the one-level bootstrap reduces this upward bias but cannot eliminate it. In some circumstances, one-level bootstrap may over-correct the bias. Then under an assumption that the bias function can be approximated by a polynomial, we show the mechanism of the multi-level bootstrap bias correction, which successively lower the order of the polynomial towards unbiasness. Lastly, to make multi-level bootstrap computationally feasible,

we propose a simultaneous bootstrap procedure which conducts many levels of bootstraps at affordable computational costs.

For convenience, we discretize every random variable so that we can use a categorical distribution of several dimensions to characterize their joint distribution, which makes easier the statistical properties of the analogue MIV bounds. For this problem discretization is not unreasonable. First, the treatment variable is discrete, usually binary, in most applications. Second, the MIV identification requires the response variable is bounded below and above. Otherwise the MIV has no identification power unless it is used together with the monotone treatment selection. Finite-valued discrete distribution by nature has a lower and upper bound. Third, to compute the analogue estimates for each subpopulation classified by MIV, we usually group the values of the MIV so as to ensure sufficient sample size. Therefore, we model treatments, responses and MIVs as finite-valued discrete random variables.

2. The mathematical structure of MIV bounds

Manski and Pepper (2000, 2009) use the MIV to help bound counterfactual outcomes, while Kreider and Pepper (2007) consider MIV identification in a partial misreporting problem. Though the derived MIV bounds look different, they share the same mathematical structure, so the same bias correction procedure can be applied to both problems. In this section, we summarize their common structure.

The counterfactual outcomes identification problem can be raised as follows. Let $D \in \{d_1, ..., d_{n_D}\}$ be a treatment variable. The n_D varieties of treatments generate n_D types of latent responses, denoted as $Y_t \in \{y_1, ..., y_{n_Y}\}$,

$t = 1, ..., n_D$. Since a person cannot receive all these treatments simultaneously, the only observable outcome is $Y = \sum_{t=1}^{n_D} Y_t \cdot I(D = d_t)$, where $I(\cdot)$ is an indicator function. Let $Z \in \{z_1, ...z_{n_Z}\}$ be a MIV such that for any two realizations $z_i \leq z_j$,

$$E(Y_t | Z = z_i) \leq E(Y_t | Z = z_j), \forall t = 1, ..., n_D.$$

Without loss of generality, discrete values of Y_t and Z are sorted in an increasing order: $y_1 \leq y_2 ... \leq y_{n_Y}$, $z_1 \leq z_2 ... \leq z_{n_Z}$.

Consider $E(Y_t | Z = z_j)$ for some $t = 1, ..., n_D$, $j = 1, ..., n_Z$. It is bounded below by $\sup_{1 \leq i \leq j} E(Y_t | Z = z_i)$ and above by $\inf_{j \leq i \leq n_Z} E(Y_t | Z = z_i)$. Since the MIV is discretized, we can replace $\sup(\cdot)$ by $\max(\cdot)$, and $\inf(\cdot)$ by $\min(\cdot)$. Furthermore, $E(Y_t | Z = z_i)$ can be dissembled into an observable part $E(Y | Z = z_i, D = d_t)$ and an unobservable part $E(Y_t | Z = z_i, D \neq d_t)$. The latter need to be replaced by the worse-case lower bound y_1 and upper bound y_{n_Y}, which yield the sharp bounds under the MIV assumption alone:

$$\max_{1 \leq i \leq j} E(Y | Z = z_i, D = d_t) \cdot P(D = d_t | Z = z_i) + y_1 \cdot P(D \neq d_t | Z = z_i) \tag{1}$$

$$\leq E(Y_t | Z = z_j) \leq$$

$$\min_{j \leq i \leq n_Z} E(Y | Z = z_i, D = d_t) \cdot P(D = d_t | Z = z_i) + y_{n_Y} \cdot P(D \neq d_t | Z = z_i).$$

To make notations compact, let us define

$p_{ikm} \equiv P(Z = z_i, Y = y_k, D = d_m)$,

$i = 1, ..., n_Z$, $k = 1, ..., n_Y$, $m = 1, ..., n_D$,

$p_{i\cdot\cdot} \equiv \sum_{k=1}^{n_Y} \sum_{m=1}^{n_D} p_{ikm}$,

$\mathbf{p} \equiv vec\left(\{p_{ikm}\}_{i=1,k=1,m=1}^{n_Z, n_Y, n_D}\right)$,

$\mathbf{p}_i \equiv vec\left(\{p_{ikm}\}_{k=1,m=1}^{n_Y,n_D}\right)$.

Here $vec(\cdot)$ is an operator that vectorizes a multi-dimension array into a long column vector. For instance, $vec\left(\{p_{ikm}\}_{i=1,k=1,m=1}^{n_Z,n_Y,n_D}\right)$ turns a $n_Z \times n_Y \times n_D$ array to a $n_Z n_Y n_D \times 1$ vector. Also assume $p_{i\cdot\cdot} > 0$, $\forall i = 1, ..., n_Z$. Then we can rewrite Eq. (1) as

$$\max_{1\leq i\leq j} f_L(\mathbf{p}_i) \leq E(Y_t | Z = z_j) \leq \min_{j\leq i\leq n_Z} f_U(\mathbf{p}_i), \tag{2}$$

where

$$f_L(\mathbf{p}_i) = \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} [y_k \cdot I(m = t) + y_1 \cdot I(m \neq t)],$$
$$f_U(\mathbf{p}_i) = \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} [y_k \cdot I(m = t) + y_1 \cdot I(m \neq t)].$$

The misreporting identification problem in Kreider and Pepper (2007) uses respondents' self-reported health information to bound the effects of (true) disability on employment. Let $L \in \{0, 1\}$ be observed employment status, $X \in \{0, 1\}$ and $W \in \{0, 1\}$ be the reported and true disability status respectively, and $Z \in \{z_1, ...z_{n_Z}\}, z_1 \leq z_2 ... \leq z_{n_Z}$ be a MIV (namely negative age in their paper) such that

$$P(L = 1 | W, Z = z_i) \leq P(L = 1 | W, Z = z_j), \text{ if } i \leq j.$$

Respondents are classified into two groups, namely the verified ($Y = 1$) and the unverified ($Y = 0$), on the basis of researchers' prior information on their accurate reporting rate. Taking this accuracy rate as given, Kreider and Pepper (2007) derive the sharp bounds of $P(L = 1 | W = 1)$. For simplicity, we consider an extreme case that the verified group has a 100% truth-telling

rate, while the unverified has an accuracy rate $\geq 0\%$ (i.e., no information). For each $j = 1, ..., n_Z$, we have

$$\begin{aligned} &\max_{1\leq i\leq j} \frac{P\left(L=1, X=1, Y=1 \left| Z=z_i\right.\right)}{P\left(X=1, Y=1 \left| Z=z_i\right.\right) + P\left(L=0, Y=0 \left| Z=z_i\right.\right)} \\ &\leq P\left(L=1 \left| W=1, Z=z_j\right.\right) \leq \\ &\min_{j\leq i\leq n_Z} \frac{P\left(L=1, X=1, Y=1 \left| Z=z_i\right.\right) + P\left(L=1, Y=0 \left| Z=z_i\right.\right)}{P\left(X=1, Y=1 \left| Z=z_i\right.\right) + P\left(L=1, Y=0 \left| Z=z_i\right.\right)} \end{aligned} \tag{3}$$

Readers are referred to Proposition 2, corollary 1 in Kreider and Pepper (2007, p.436) for the derivation. Note that when the accuracy rate is not as extreme as 100% and 0%, the bounds will be more cumbersome. However, what remain unchanged are all the probabilities are conditional on $Z = z_i$. This feature makes the mathematical structure of the MIV bounds (see below) unchanged.

Define a set of symbols similar to what we defined in the previous problem.

$p_{ijkl} \equiv P\left(Z=z_i, L=j, X=k, Y=l\right)$, $i = 1, ..., n_Z$, $j, k, l = 0, 1$,

$p_{i\cdots} \equiv \sum_{j=0}^{1}\sum_{k=0}^{1}\sum_{l=0}^{1} p_{ijkl}$,

$\mathbf{p} \equiv vec\left(\left\{p_{ijkl}\right\}_{i=1,j=0,k=0,k=0}^{n_Z,1,1,1}\right)$,

$\mathbf{p}_i \equiv vec\left(\left\{p_{ijkl}\right\}_{j=0,k=0,k=0}^{1,1,1}\right)$.

Then Eq (3) can be written as

$$\max_{1\leq i\leq j} f_L\left(\mathbf{p}_i\right) \leq P\left(L=1 \left| W=1, Z=z_j\right.\right) \leq \min_{j\leq i\leq n_Z} f_U\left(\mathbf{p}_i\right), \tag{4}$$

where

$$\begin{aligned} f_L\left(\mathbf{p}_i\right) &= \frac{p_{i111}}{p_{i111} + p_{i011} + p_{i010} + p_{i000}} \\ f_U\left(\mathbf{p}_i\right) &= \frac{p_{i111} + p_{i110} + p_{i100}}{p_{i111} + p_{i011} + p_{i110} + p_{i100}} \end{aligned}$$

Comparing Eq. (2) with Eq. (4), we see the MIV bounds of the two problems have some features in common:

First, the theoretical bounds are determined by $\mathbf{p}$, the parameter vector summarizing the joint probability of observable variates. In other words, the observable variates follows a categorical distribution of multiple dimensions, which is equivalent to a long single-dimension categorical distribution with parameters $\mathbf{p}$.

Second, MIV bounds take the form $\max_{1\leq i\leq j} f_L(\mathbf{p}_i)$ and $\min_{j\leq i\leq n_Z} f_U(\mathbf{p}_i)$, where $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{n_Z}$ form a partition of $\mathbf{p}$ according to the possible values of the MIV.

Third, both $f_L(\mathbf{p}_i)$ and $f_U(\mathbf{p}_i)$ are homogeneous functions of degree zero. Eq. (1) and Eq. (3) involves probabilities conditional on $Z = z_i$, which is the ratio of the joint and the marginal probabilities. Since a constant cancels in the nominator and denominator, $f_L(\mathbf{p}_i)$ and $f_U(\mathbf{p}_i)$ in Eq. (2) and Eq. (4) always satisfy degree-zero homogeneity.

3. Sampling Variation

In applications, the probability vector $\mathbf{p}$ need to be estimated from the data. Let $\{\mathbf{v}_s\}_{s=1}^n$ be i.i.d. draws from the categorical distribution with parameters $\mathbf{p}$. Conceptually, this means there are n persons taking the survey which asks for each respondent's realized choice of (Z, Y, D) or (Z, L, X, W). All possible choices of (Z, Y, D) define $n_Z n_Y n_D$ categories and that of (Z, L, X, W) define $8n_Z$ categories. So the length of the vector $\mathbf{v}_s$ is $n_Z n_Y n_D$ and $8n_Z$ respectively. The person s choose a category, so the component in $\mathbf{v}_s$ corresponding to that realized category will be coded as 1 with other elements in

$\mathbf{v}_s$ being 0.

By construction, the sample analogue of $\mathbf{p}$ can be expressed as

$$\widehat{\mathbf{p}} = \frac{1}{n}\sum_{s=1}^{n}\mathbf{v}_s.$$

Proposition 1. *$\widehat{\mathbf{p}}$ is a consistent estimate of $\mathbf{p}$, and the asymptotic distribution is*

$$\sqrt{n}\left(\widehat{\mathbf{p}}-\mathbf{p}\right) \xrightarrow{d} N\left[\mathbf{0}, diag\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'\right],$$

where $diag\left(\mathbf{p}\right)$ refers to a diagonal matrix with the main diagonal being the vector $\mathbf{p}$.

Proofs of propositions in this paper are provided in the appendix.

Suppose the length of $\mathbf{p}$ is r, then $diag\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'$ is a positive semidefinite matrix of reduced rank $r-1$. The linear combination $\iota'\widehat{\mathbf{p}}$, where ι is a vector of ones, have the mean of one and variance of zero. Therefore, the analogue probability estimates always sum up to one. In addition, the elements of $\widehat{\mathbf{p}}$ are negatively correlated since they are subject to the aggregation constraint.

Proposition 1 suggests that the large-sample approximating distribution of $\widehat{\mathbf{p}}$ is $N\left[\mathbf{p}, \frac{1}{n}diag\left(\mathbf{p}\right)-\frac{1}{n}\mathbf{p}\mathbf{p}'\right]$. Of course, it is understood that $\widehat{\mathbf{p}}$ is a bounded random vector since each component must fall in the unit interval.

Partition $\widehat{\mathbf{p}}$ into $\widehat{\mathbf{p}}_1, ..., \widehat{\mathbf{p}}_{n_Z}$ in the same way we partition $\mathbf{p}$ into $\mathbf{p}_1, ..., \mathbf{p}_{n_Z}$. Now we consider the asymptotic distribution of $f_L\left(\widehat{\mathbf{p}}_i\right)$, $f_U\left(\widehat{\mathbf{p}}_i\right)$, $i = 1, ..., n_Z$.

Proposition 2. *Let $f_L\left(\cdot\right)$ be a real differentiable function satisfying homogeneity of degree zero, that is, $f_L\left(c\mathbf{x}\right) = f_L\left(\mathbf{x}\right), \forall c > 0$. Then $f_L\left(\widehat{\mathbf{p}}_1\right), ..., f_L\left(\widehat{\mathbf{p}}_{n_Z}\right)$ are asymptotically independent and for each $i = 1, ..., n_Z$,*

$$\sqrt{n}\left[f_L\left(\widehat{\mathbf{p}}_i\right)-f_L\left(\mathbf{p}_i\right)\right] \xrightarrow{d} N\left[\mathbf{0}, \mathbf{G}_i \cdot diag\left(\mathbf{p}_i\right) \cdot \mathbf{G}_i'\right],$$

where $\mathbf{G}_i$ *is a row vector such that*

$$\mathbf{G}_i = \frac{\partial f_L\left(\widehat{\mathbf{p}}_i\right)}{\partial \widehat{\mathbf{p}}_i'} \mid_{\widehat{\mathbf{p}}_i = \mathbf{p}_i} .$$

The asymptotic distribution of $f_U\left(\widehat{\mathbf{p}}_i\right)$ can be derived similarly with the subscript L replaced by U in Proposition 2.

The zero-degree homogeneity of $f_L\left(\cdot\right)$ plays an important role in Proposition 2 since Euler's Theorem implies that $\mathbf{G}_i\mathbf{p}_i = 0$, $i = 1, ..., n_Z$, resulting in both zero covariances and simplified variances of the normal variates. Theoretically, Proposition 2 provides a unified asymptotic distribution of $f_L\left(\cdot\right)$ for any identification problem with the MIV, as long as $f_L\left(\cdot\right)$ can be written as a differentiable function of the population probabilities conditional on the MIV. Proposition 2 will be also used to justify the assumptions of the bootstrap bias correction in the next section. Practically, Proposition 2 can be used to compute the asymptotically variance of $f_L\left(\widehat{\mathbf{p}}_i\right)$ if we are willing to calculate the cumbersome gradients. However, for a specific problem, there might be some better way to compute the finite-sample variance. For instance, once we recognize that the $f_L\left(\mathbf{p}_i\right)$ in Eq. (2) can be represented as a conditional expectation, the finite-sample variance of $f_L\left(\widehat{\mathbf{p}}_i\right)$ is readily given in the next proposition.

Proposition 3. $f_L\left(\mathbf{p}_i\right)$ *in Eq. (2) takes the following form:*

$$f_L\left(\mathbf{p}_i\right) = E\left(Q \mid Z = z_i\right),$$

where

$$Q = Y \cdot I\left(D = d_t\right) + y_1 \cdot I\left(D \neq d_t\right).$$

Conditional on the positive analogue $p_{i\cdot\cdot}$, the finite-sample variance of $f_L(\widehat{\mathbf{p}}_i)$ is given by

$$Var\left[f_L\left(\widehat{\mathbf{p}}_i\right)\right] = \left[\sum_{r=1}^{n} \frac{1}{r} \frac{\binom{n}{r}\left(p_{i\cdot\cdot}\right)^r\left(1-p_{i\cdot\cdot}\right)^{n-r}}{1-\left(1-p_{i\cdot\cdot}\right)^n}\right] \cdot Var\left(Q\left|Z=z_i\right.\right),$$

where

$$\begin{aligned} Var\left(Q\left|Z=z_i\right.\right) &= E\left(Q^2\left|Z=z_i\right.\right) - \left[E\left(Q\left|Z=z_i\right.\right)\right]^2 \\ &= \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} q_{km}^2 - \left[\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} \frac{p_{ikm}}{p_{i\cdot\cdot}} q_{km}\right]^2, \end{aligned}$$

and

$$q_{km} = y_k \cdot I\left(d_m = d_t\right) + y_1 \cdot I\left(d_m \neq d_t\right).$$

4. Estimating the MIV bounds

Proposition 2 indicates that the large-sample approximating distribution of $f_L(\widehat{\mathbf{p}}_i)$ is $N\left[f_L\left(\mathbf{p}_i\right), \frac{1}{n}\mathbf{G}_i \cdot diag\left(\mathbf{p}_i\right) \cdot \mathbf{G}_i'\right]$. To estimate the MIV bounds as in Eq. (2) and Eq. (4), we need to find an estimator for $\max_{1\leq i\leq j} f_L(\mathbf{p}_i)$. An naive choice is $\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)$. Though $f_L(\widehat{\mathbf{p}}_i)$ is an asymptotically unbiased estimator for $f_L(\mathbf{p}_i)$, $\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)$ is not an unbiased estimator for $\max_{1\leq i\leq j} f_L(\mathbf{p}_i)$ in the finite sample. It is biased upwards simply because $\max(\cdot)$ is convex and Jensen's inequality implies $E\left[\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)\right] > \max_{1\leq i\leq j} f_L(\mathbf{p}_i)$. Similarly, $\min_{j\leq i\leq n_Z} f_U(\widehat{\mathbf{p}}_i)$ has a downward bias if it is used to estimate $\min_{j\leq i\leq n_Z} f_U(\mathbf{p}_i)$. This is unfavorable from the perspective of decision making in that the estimated bounds are narrower than the true bounds. Kreider and Pepper (2007) propose a heuristic bootstrap bias correction. The Monte Carlo evidence in Manski and Pepper (2009) indicates

the bias can be considerably reduced, but not eliminated after the correction. In this section, we will analyze the biases of a series of estimators and provide a justification for the bootstrap correction. We will also suggest a feasible approach to conduct several levels of bootstraps simultaneously. We will focus on the bias correction of $\max_{1\leq i\leq j} f_L(\widehat{\mathbf{p}}_i)$, and the same principle can be applied to the case of $\min_{j\leq i\leq n_Z} f_U(\widehat{\mathbf{p}}_i)$ as well.

To make our notations compact, define

$\mu_i \equiv f_L(\mathbf{p}_i)$, $\sigma_i^2 \equiv \frac{1}{n}\mathbf{G}_i \cdot diag(\mathbf{p}_i) \cdot \mathbf{G}_i'$, $X_i \equiv f_L(\widehat{\mathbf{p}}_i)$, $i = 1, ..., j$.

$\boldsymbol{\mu} \equiv (\mu_1, ..., \mu_j)'$, $\boldsymbol{\sigma}^2 \equiv diag\left(\sigma_1^2, ..., \sigma_j^2\right)$, $\boldsymbol{X} \equiv (X_1, ..., X_j)'$.

Let $\mathbf{x}$ be a realization of $\boldsymbol{X}$. That is, the only one realized $\mathbf{x}$ is what we obtained from the data.

Essentially our task is to propose a good estimator for $\max(\boldsymbol{\mu})$ by observing $\mathbf{x}$. To that end, we need to make some assumptions.

Assumption 1: $\mathbf{X} \sim N(\boldsymbol{\mu},\boldsymbol{\sigma}^2)$.

Assumption 2: $\boldsymbol{\sigma}^2$ is known.

The rationale for the first assumption is Proposition 2, which suggests $X_1, ..., X_j$ are asymptotically independent normal variates. The second assumption is arguable. In practice, the variances of those variates are unknown, and we at best can provide a consistent estimator for the variances, say $\widehat{\boldsymbol{\sigma}}^2$, using Proposition 2 or Proposition 3. It is true that each σ_i^2 is positively related to the magnitude of the upward bias (which is most apparent if we assume the convex function is differentiable and examine the Taylor expansion). However, we do not know whether $E(\widehat{\sigma}_i^2)$ is larger or smaller than σ_i^2 in the finite sample, so at best we can argue that the upward bias derived with $\widehat{\sigma}_i^2$ will be close to the true upward bias determined by σ_i^2. In

this sense, we view Assumption 2 as a working assumption.

4.1. Bias function and a conservative estimator

A naive estimator is the maximum of the sample.

$$T_1(\mathbf{x}) = \max(\mathbf{x}).$$

By Jensen's inequality, $E[T_1(\mathbf{X})] > \max(\boldsymbol{\mu})$. So the estimator is biased upwards. Define the first-level bias function $B_1 : \mathbb{R}^j \to \mathbb{R}$ such that

$$B_1(\boldsymbol{\mu}) = E[T_1(\mathbf{X})] - \max(\boldsymbol{\mu}).$$

$B_1(\cdot)$ is a function of $\boldsymbol{\mu}$ since $\mathbf{X} \sim N(\boldsymbol{\mu},\boldsymbol{\sigma}^2)$. Of course, it is also a function of $\boldsymbol{\sigma}^2$, which is assumed to be known and therefore suppressed.

The first-level bias function has a useful property stated below.

Proposition 4 (Bounds of the bias function)**.** *$B_1(\cdot)$ is bounded by $0 < B_1(\boldsymbol{\mu}) \leq M$, $\forall\ \boldsymbol{\mu} \in \mathbb{R}^j$, where*

$$M = E[\max(\mathbf{X}_0)],$$
$$\mathbf{X}_0 \sim N(\mathbf{0},\boldsymbol{\sigma}^2).$$

Note that the upper bound M is computable, at least by simulation. For the special case of $j = 2$, we have analytic results. See Clark (1961), Cain (1994) for derivations.

$$B_1(\boldsymbol{\mu}) = \omega\mu_1 + (1-\omega)\mu_2 + \sigma_0\phi\left(\frac{\mu_1 - \mu_2}{\sigma_0}\right) - \max(\mu_1, \mu_2),$$
$$M = \sigma_0\phi(0),$$

where $\phi(\cdot)$, $\Phi(\cdot)$ is the standard normal p.d.f. and c.d.f. respectively, and

$$\omega = \Phi\left(\frac{\mu_1 - \mu_2}{\sigma_0}\right),$$
$$\sigma_0 = \sqrt{\sigma_1^2 + \sigma_2^2}.$$

For $j = 2$, we may plot a 3-D graph of $B_1(\cdot)$, with μ_1, μ_2 on the x, y axis and B_1 on the z axis (see Figure 1). It is a ridge-shaped function. Along the 45° line on the x, y plane, $B_1(\cdot)$ attains the same maximum value $\sigma_0\phi(0)$. Off the 45° line, $B_1(\cdot)$ gradually decreases towards zero.

Proposition 4 shows that the bias of the naive estimator $\max(\mathbf{X})$ is bounded above, so we can propose a conservative estimator for $\max(\boldsymbol{\mu})$.

$$T_c(\mathbf{x}) = \max(\mathbf{x}) - M.$$

By construction, T_c is biased downwards. We call it a conservative estimator because we can use the same principle to propose an upward biased estimator for $\min_{j \le i \le n_Z} f_U(\widehat{\mathbf{p}}_i)$, and then we will obtain bounds wider than the true bounds. For decision making, perhaps we would rather have too wide bounds than too narrow bounds. Also note that if we allow $\boldsymbol{\sigma}^2 \to \mathbf{0}$, M will also decrease to zero, so that T_c will converge to $\max(\boldsymbol{\mu})$. Therefore, if T_c is applied to the MIV bounds, it is still a consistent estimator. Furthermore, since T_1 is biased upwards and T_c is biased downwards, they themselves bound the unbiased estimator of the MIV bounds.

4.2. Bootstrap bias correction

Clearly, T_c over-corrects the bias. Is it possible to find an estimator "being just right"? Kreider and Pepper (2007) proposed a heuristically motivated bootstrap bias corrected estimator. This subsection aims to provide a rationale for this correction.

The idea of bootstrap bias correction is to use the bias function to correct the naive estimator. Define

$$T_2^* (\mathbf{x}) = T_1 (\mathbf{x}) - B_1 (\boldsymbol{\mu}),$$
$$T_2 (\mathbf{x}) = T_1 (\mathbf{x}) - B_1 (\mathbf{x}).$$

If T_2^* were an estimator, it would be unbiased by construction. That is, $E[T_2^* (\mathbf{X})] = \max(\boldsymbol{\mu})$. However, since T_2^* contains the unknown $\boldsymbol{\mu}$, it is not computable. The bootstrap treats the sample as if it represents the bootstrap population, evaluating the bias as $E\left[T_1\left(\widetilde{\mathbf{X}}\right)\right] - \max(\mathbf{x})$, where $\widetilde{\mathbf{X}} \sim N(\mathbf{x}, \boldsymbol{\sigma}^2)$. Analytically, this is equivalent to replacing $B_1(\boldsymbol{\mu})$ with $B_1(\mathbf{x})$, so that T_2 is the bootstrap bias corrected estimator. Unfortunately, T_2 is not unbiased unless we have

$$E[B_1 (\mathbf{X})] = B_1 (\boldsymbol{\mu}).$$

To further analyze the bias, define the second-level bias function $B_2 \colon \mathbb{R}^j \to \mathbb{R}$ such that

$$B_2 (\boldsymbol{\mu}) = E[T_2 (\mathbf{X})] - \max(\boldsymbol{\mu}).$$

$B_2(\cdot)$ has the following property:

Proposition 5. $B_2 (\boldsymbol{\mu}) < B_1 (\boldsymbol{\mu})$, $\forall \boldsymbol{\mu} \in \mathbb{R}^j$.

Proposition 5 justifies the usage of the bootstrap bias correction since the upward bias of T_1 will be reduced after the bootstrap correction. However, in general it cannot eliminate the bias. It is helpful to consider the case when $\mu_1 = ... = \mu_j$. As suggested in the proof of Proposition 4, $B_1(\boldsymbol{\mu})$ has already attained its maximum, while $E[B_1(\mathbf{X})]$ is the weighted average of $B_1(\cdot)$ evaluated at every realization of $\mathbf{X}$ with the weight given by the normal p.d.f. $\phi(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\sigma}^2)$. So we have $B_2(\boldsymbol{\mu}) = B_1(\boldsymbol{\mu}) - E[B_1(\mathbf{X})] > 0$. In that case, positive bias still exists after the bootstrap. Furthermore, it is possible that the bootstrap over-corrects the upward bias since $B_1(\boldsymbol{\mu})$ might be smaller than $E[B_1(\mathbf{X})]$ for some $\boldsymbol{\mu}$. For illustration, Figure 2 plots the two levels of bias functions when $j = 2$. We set $\sigma_1^2 = 1$, $\sigma_2^2 = 1$. Since only the difference between μ_1 and μ_2 matters, we normalize $\mu_1 = 0$ and plot B_1, B_2 against different values of μ_2. As we can see, i) when μ_2 goes to infinity or minus infinity, both $B_1$and B_2 approach zero; ii) the largest bias occurs when $\mu_2 = 0$; iii) the B_2 curve always lies below the B_1 curve; iv) though B_1 is always positive, there is a region that B_2 is slightly negative, which implies there is a possibility that the one-level bootstrap may over-correct the bias.

4.3. Multi-level bootstrap correction

Since one level of bootstrap estimator T_2 does not eliminate the bias, a natural extension is using its bias B_2 to further correct T_2. Define

$$T_3^*(\mathbf{x}) = T_2(\mathbf{x}) - B_2(\boldsymbol{\mu}),$$
$$T_3(\mathbf{x}) = T_2(\mathbf{x}) - B_2(\mathbf{x}).$$

Again, If T_3^* were an estimator, it would be unbiased by construction. However, our inability to evaluate $B_2(\cdot)$ at the right point, namely $\boldsymbol{\mu}$, forces

us to compute $B_2(\mathbf{x})$ instead. In essence, we treat the sample $\mathbf{x}$ as the bootstrap population and evaluate $B_2(\mathbf{x}) = B_1(\mathbf{x}) - E\left[B_1\left(\widetilde{\mathbf{X}}\right)\right]$, where $\widetilde{\mathbf{X}} \sim N(\mathbf{x}, \boldsymbol{\sigma}^2)$. Since evaluating $B_1(\cdot)$ is equivalent to one level of bootstrap, evaluating $B_2(\cdot)$ can be viewed as doubling the bootstrap. Clearly, the estimator T_3 is not unbiased unless we have

$$E[B_2(\mathbf{X})] = B_2(\boldsymbol{\mu}).$$

The effect of bias reduction depends on the functional form of the bias function as well as the discrepancy between $\mathbf{x}$ and $\boldsymbol{\mu}$. The latter is unknown, and we cannot expect the realization $\mathbf{x}$ happens to be $\boldsymbol{\mu}$ in the finite sample. However, in some sense the bias function is under control. Note that if $B_1(\cdot)$ were a linear function, T_2 would be unbiased regardless of the unknown $\boldsymbol{\mu}$. Similarly, if $B_2(\cdot)$ were a linear function, T_3 would be unbiased. We double the bootstrap because we hope $B_2(\cdot)$ ensembles more linearity. This raises two questions: Is $B_2(\cdot)$ flatter than $B_1(\cdot)$? If we proceed to higher level of the bootstrap, will we eventually obtain an unbiased estimator?

Define the higher-level bias function and bias corrected estimator as

$$\begin{aligned} B_i(\boldsymbol{\mu}) &= E[T_i(\mathbf{X})] - \max(\boldsymbol{\mu}) \\ &= B_{i-1}(\boldsymbol{\mu}) - E[B_{i-1}(\mathbf{X})], \\ T_{i+1}(\mathbf{x}) &= T_i(\mathbf{x}) - B_i(\mathbf{x}), \end{aligned}$$

for $i = 3, 4, 5, ...$

If we are willing to make an additional assumption, we have an answer to the above two questions.

Assumption 3: $B_1(\boldsymbol{\mu})$ can be well approximated by a polynomial.

There is a need to justify this assumption. Note that $B_1(\boldsymbol{\mu})$ is a continuous, but not differentiable function in that $\max(\cdot)$ is not differentiable. The Taylor theorem of polynomial approximation does not apply. However, in Eq. (2) and Eq. (4), $f_L(\mathbf{p}_i)$ is bounded by $[y_1, y_{n_Y}]$ and $[0,1]$ respectively. Therefore, $\boldsymbol{\mu}$ is bounded. By Stone-Weierstrass theorem, the bias function $B_1(\boldsymbol{\mu})$ can be uniformly approximated by a polynomial. A limitation of our study is that we are unable to quantify the precision of the approximation. We will designate a polynomial of large order and assume the approximation error is negligible.

Proposition 6. *Suppose $B_1(\boldsymbol{\mu})$ is a polynomial of order d, where $d \geq 2$, then $B_2(\boldsymbol{\mu})$ is a polynomial of order $d-2$. Each level of bootstrap will reduce the polynomial order by 2 successively. Bias can be eliminated after $\left[\frac{d}{2}\right]$ levels of bootstraps, where $[\cdot]$ refers to the operator of taking integers.*

Let us illustrate this property with a numerical example. Consider two independent normal variates $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$. Assume $B_1(\boldsymbol{\mu}) = 2\mu_1^5\mu_2^6$, a polynomial of order 11.

$$\begin{aligned} E[B_1(\mathbf{X})] &= 2E\left(X_1^5\right) E\left(X_2^6\right) \\ &= 2\left(\mu_1^5 + 10\sigma_1^2\mu_1^3 + 15\sigma_1^4\mu_1\right) \cdot \left(\mu_2^6 + 15\sigma_2^2\mu_2^4 + 45\sigma_2^4\mu_2^2 + 15\sigma_2^6\right) \end{aligned}$$

When $B_1(\boldsymbol{\mu}) - E[B_1(\mathbf{X})]$, the leading term $2\mu_1^5\mu_2^6$ cancels, and there are no terms of order 10 like $\mu_1^5\mu_2^5$, $\mu_1^4\mu_2^6$. Therefore, $B_2(\boldsymbol{\mu})$ is reduced to a polynomial of order 9. If we forward the bootstrap to higher levels, then $B_3(\boldsymbol{\mu})$ will be a polynomial of order 7, and $B_4(\boldsymbol{\mu})$ of order 5, etc. Eventually $B_i(\boldsymbol{\mu})$ will be of order one or zero. $E[B_i(\mathbf{X})] = B_i(\boldsymbol{\mu})$ is satisfied, and

$T_{i+1}(\mathbf{x})$ becomes an unbiased estimator. In other words, d rounds of the bootstraps can correct the bias for polynomial $B_1(\boldsymbol{\mu})$ of order up to $2d$.

4.4. Simultaneous bootstrap

The upper level bias function $B_i(\cdot)$ is constructed by the expectation of the lower level bias function $E[B_{i-1}(\cdot)]$, which has to be evaluated with simulation. The nested, iterative simulation suffers from the curse of dimensionality, and practically we are unable to proceed beyond double or triple bootstraps. To resolve the computational difficulty, we propose a simultaneous bootstrap algorithm which can conduct many level of bootstrap correction with affordable computational costs. Davidson and MacKinnon (2002, 2007) provide a similar procedure which they refer to as "fast double bootstrap".

The rationale for the simultaneous bootstrap comes from the identity

$$E_{\xi}\left\{E_{\eta|\xi}\left[g(\xi,\eta)\right]\right\} = E_{\xi,\eta}\left[g(\xi,\eta)\right],$$

for arbitrary random variables ξ, η and real valued function $g\colon \mathbb{R}^2 \to \mathbb{R}$, where the subscript in $E(\cdot)$ explicitly indicates random variables that expectation operator applies to.

Suppose $E(\cdot)$ must be evaluated with simulation. The left hand side of that identity prescribes a nested procedure. In the first step we draw a ξ. Conditional on this value of ξ, we draw thousands of η, and then average $g(\xi,\eta)$. In the second step, we repeat the first step with thousands of ξ, and then average the averaged $g(\xi,\eta)$. However, the right hand side prescribes a simultaneous procedure such that we draw (ξ,η) from their joint distribution, and take the average of $g(\xi,\eta)$.

Given the same computational costs measured as the number of visits to $g(\xi, \eta)$, the latter procedure provides a more accurate approximation. This is because in the simultaneous simulation procedure draws of the pair (ξ, η) are independent, while in the nested simulation the same draw of ξ needs to be used for multiple times, which induces positive correlation and larger variance. To formalize this idea, we present the following proposition.

Proposition 7 (Efficiency of simultaneous simulation)**.** *Let the simulator for* $E_{\xi,\eta}[g(\xi, \eta)]$ *be*

$$S_1 = \frac{1}{N^2} \sum_{i=1}^{N^2} g(\xi_i, \eta_i),$$

where $\{\xi_i, \eta_i\}_{i=1}^{N^2}$ *are i.i.d. draws from the joint distribution of* (ξ, η)*.*

Let the simulator for $E_\xi \left\{ E_{\eta|\xi}[g(\xi, \eta)] \right\}$ *be*

$$S_2 = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{1}{N} \sum_{k=1}^{N} g(\xi_j, \eta_{j,k}) \right],$$

where $\{\xi_j\}_{j=1}^{N}$ *are i.i.d. draws from the marginal distribution of* ξ*, while* $\{\eta_{j,k}\}_{k=1}^{N}$ *are i.i.d. draws from the conditional distribution of* $\eta | (\xi = \xi_j)$*,* $j = 1, ..., N$*.*

Then we have

$$E(S_1) = E(S_2),$$
$$Var(S_1) \leq Var(S_2),$$

with equality of variance iff $E_{\eta|\xi}[g(\xi, \eta)] = E_{\xi,\eta}[g(\xi, \eta)]$ *for all realizations of* ξ*.*

To illustrate the efficiency of the simultaneous simulation relative to the nested simulation, consider a simple numerical example.

Let $(\xi, \eta) \sim N(0, 0, 1^2, 1^2, 0.5)$, $g(\xi, \eta) = \xi + \eta$, $N = 10$.
Then $Var(S_1) = Var\left[\frac{1}{100}\sum_{i=1}^{100}(\xi_i + \eta_i)\right] = \frac{3}{100}$,
but $Var(S_2) = Var\left[\frac{1}{100}\sum_{j=1}^{10}\sum_{k=1}^{10}(\xi_j + \eta_{j,k})\right] = \frac{21}{100}$.

We see that the nested simulation has a variance seven times larger than the simultaneous procedure, given 100 visits to $g(\xi, \eta)$ in both procedures. Even if we change the correlation of (ξ, η) from 0.5 to 0, nested simulation still has a larger variance. In that case, we have $Var(S_1) = \frac{2}{100}$, and $Var(S_2) = \frac{11}{100}$. The inflation of variance is due to the fact that the same draw of ξ_j has to be used 10 times in nested simulation.

Generally speaking, the simultaneous simulation will substantially improve the quality of the simulator. The case of no improvement is rare. It happens only when the conditional expectation is identical to the unconditional expectation for all realizations of the variable being conditioned on. To give a example, consider $(\xi, \eta) \sim N(0, 0, 1^2, 1^2, 0.5)$ with $g(\xi, \eta) = \xi\eta$. In that case, $Var(S_1) = Var(S_2) = \frac{5}{1000}$. However, once (ξ, η) have non-zero means, there will be improvement.

The results can be extended to multivariate and vector-valued random variables. We have the identity

$$E_{\xi_1} E_{\xi_2|\xi_1} \ldots E_{\xi_n|\xi_{n-1}\ldots\xi_1} g(\xi_1, ..., \xi_n) = E_{\xi_1,\ldots,\xi_n}[g(\xi_1, ..., \xi_n)],$$

for arbitrary vector-valued random variables $\xi_1, ..., \xi_n$ and real valued function g.

Again, the left hand side prescribes a multi-level nested simulation procedure, while the right hand side suggests a simultaneous simulation algorithm. The inefficiency of the nested procedure comes from the multiple usage of the same draw of ξ_{n-1}, and of ξ_{n-2}, ..., and worst of all, of ξ_1.

Multi-level bootstrap bias correction is a direct application of the above results.

Though $B_1(\cdot)$ might be evaluated by analytic formula or deterministic quadrature, $B_2(\cdot)$, $B_3(\cdot)$, etc. are better evaluated by simulation. For example, consider evaluating $B_3(\mathbf{x})$:

$$\begin{aligned} B_3(\mathbf{x}) &= B_2(\mathbf{x}) - E_{\mathbf{X}} B_2(\mathbf{X}) \\ &= [B_1(\mathbf{x}) - E_{\mathbf{X}} B_1(\mathbf{X})] - E_{\mathbf{X}}\left[B_1(\mathbf{X}) - E_{\widetilde{\mathbf{X}}|\mathbf{X}} B_1\left(\widetilde{\mathbf{X}}\right)\right] \\ &= E_{\mathbf{X}} E_{\widetilde{\mathbf{X}}|\mathbf{X}}\left\{[B_1(\mathbf{x}) - B_1(\mathbf{X})] - \left[B_1(\mathbf{X}) - B_1\left(\widetilde{\mathbf{X}}\right)\right]\right\} \\ &= E_{\mathbf{X},\widetilde{\mathbf{X}}}\left[g\left(\mathbf{X},\widetilde{\mathbf{X}}\right)\right] \end{aligned}$$

where $\mathbf{X} \sim N(\mathbf{x}, \boldsymbol{\sigma}^2)$, $\widetilde{\mathbf{X}}|(\mathbf{X} = \mathbf{y}) \sim N(\mathbf{y}, \sigma^2)$. $g\left(\mathbf{X},\widetilde{\mathbf{X}}\right) = [B_1(\mathbf{x}) - B_1(\mathbf{X})] - \left[B_1(\mathbf{X}) - B_1\left(\widetilde{\mathbf{X}}\right)\right]$.

The simultaneous procedure for $B_3(\mathbf{x})$ takes the following steps:

First, sample a pair $(\mathbf{y}, \mathbf{z})$ from the joint distribution of $\left(\mathbf{X},\widetilde{\mathbf{X}}\right)$. The easiest way is the method of composition, that is, to sample $\mathbf{y}$ from $N(\mathbf{x}, \sigma^2)$, and then sample $\mathbf{z}$ from $N(\mathbf{y}, \boldsymbol{\sigma}^2)$.

Second, evaluate $g(\mathbf{y}, \mathbf{z})$, which is a difference of differenced $B_1(\cdot)$.

Third, repeat the first and second step, and average the results.

Higher order bias function $B_i(\cdot)$, $i > 3$ can be simultaneously simulated in the same way. The first step is a hierarchical sampling of normal variates. The second step is a multiple difference of $B_1(\cdot)$ evaluated at the obtained sample.

From the perspective of computation, instead of being evaluated directly, $B_1(\cdot)$ may be treated as another level (that is, the bottom level) of the simultaneous simulation. It is less precise, but much faster. The saved com-

putation time can be used for a larger scale simulation, which improves the precision of all levels of bootstraps. Given the same computation costs measured in CPU time, whether the gains outweighs the loss is largely a practical issue.

5. Monte Carlo evidence

In this section, we replicate the Monte Carlo experiment in Manski and Pepper (2009), with multi-level bootstrap added to further reduce the bias. The experiment simulates the MIV lower bound of the treatment response $E(Y_t|Z=z_j)$ as in Eq. (1). The joint distribution of (Y,D,Z) is specified in the identical way as in Manski and Pepper (2009). The MIV Z has a categorical distribution with M equal-probability mass points $\left\{\frac{1}{M},\frac{2}{M},...,1\right\}$. The treatment variable $D=I(Z+\varepsilon>0)$, where $\varepsilon\sim N(0,1)$. The response variable Y follows $N(0,\sigma^2)$ censored to $(-1.96,1.96)$. With a random sample of n observations, we evaluate the Monte Carlo distribution of the analogue MIV bound for $E(Y_1|Z=1)$ with 1000 repetitions.

Our bootstrap correction algorithm assumes normality as well as fixed variances. The finite-sample variances are computed from the analogue version of the formula in Proposition 3. Note that there is no need to discretize Y when we apply that formula since analogue conditional variance can be used. This is advantageous to the asymptotic variances given by Proposition 2, where we have to discretize every variable and calculate the gradients. Nevertheless, the computed variances are close no matter whatever approach in use.

Once we obtained the variances, we apply the simultaneous multi-level

bootstrap procedure to correct the bias. 100000 draws are used to evaluate up to four levels of bootstraps. The simulation results are presented in Table 1. Each column is an experiment with selected values of M, σ^2, n. The fourth row displays the biases of raw analogue estimator (T_1), which are comparable to Table 1 in Manski and Pepper (2009). The fifth row shows the biases of first-level bootstrap corrected estimator (T_2), comparable to Table 2 in Manski and Pepper (2009). The following rows show the biases of second, third, fourth levels of bootstrap corrected estimators (T_3, T_4, T_5). The last row presents the biases of the conservative estimator (T_c), which is supposed to be biased downwards.

Our results of the biases of T_1 and T_2 are very close to what reported by Manski and Pepper (2009). The slight difference might due to the fact that they used nonparametric bootstrap (resample from the empirical distribution) and we use parametric bootstrap (resample from the normal distribution with estimated variance). The most important new results are T_3, T_4, T_5 has smaller biases. For example, in the setting $M = 8$, $\sigma^2 = 25$, $n = 100$, T_1 has a huge bias of 0.55. T_2 reduces it to 0.22, but the bias is still relatively large. As predicted by Proposition 6, higher level of bootstrap can further improve the estimator. T_3, T_4, T_5 have biases 0.15, 0.11, 0.09 respectively. In fact, in most M, σ^2, n settings the simulated biases are monotone decreasing as the bootstrap is forwarded to higher level.

Also note that when the bias has already achieved a tiny level (compared to the numerical standard errors of simulation), further bootstrap may not improve the estimator any more, but there is also no sign of deterioration. This observation is in line with Proposition 6, which indicates that d rounds

of bootstraps can correct the bias for polynomial $B_1(\boldsymbol{\mu})$ of order up to $2d$. After that, the bias function becomes a constant, and no improvement afterwards. This happens mostly in settings where $n = 1000$. In those cases, since the raw analogue estimator is consistent, the finite sample bias of T_1 is already small. We cannot expect multi-level bootstrap will eliminate the bias because high dimensional simulation itself introduces non-negligible error. As a practical suggestion, we recommend more levels of bootstrap correction when the sample size is small, but one or two levels of bootstrap may suffice for a large dataset. Of course, increasing simulation draws will make higher level bootstrap bias correction more reliable, if we can afford the computation costs.

The simulation results also suggest the usefulness of the conservative estimator T_c. If we prefer some wider, but not narrower, bounds than the true bounds, and are not willing to resort to any bootstrap correction, we may use the conservative estimator. For $M = 4$, the magnitude of downward bias induced by T_c is relatively larger than the magnitude of upward bias caused by T_1, though still on the same scale. For $M = 8$, the absolute size of bias are similar between T_c and T_1. Furthermore, as n becomes larger, T_c decreases as well, which suggests that in large sample T_c offers a cheap but effective solution to the problematic analogue MIV bounds.

6. An application to disability misreporting identification

In this section, we reconsider the empirical study of Kreider and Pepper (2007) on the employment gap between the disabled and non-disabled person.

The employment gap is defined as

$$P\left(L=1\,|W=1\right)-P\left(L=1\,|W=0\right)$$
$$=\sum P\left(Z=z_j\right)\cdot\left[P\left(L=1\,|W=1,Z=z_j\right)-P\left(L=1\,|W=0,Z=z_j\right)\right],$$

where the MIV bounds of $P\left(L=1\,|W=1,Z=z_j\right)$ is given by Eq. (4), and that of $P\left(L=1\,|W=0,Z=z_j\right)$ can be formulated similarly.

Kreider and Pepper (2007) analyze two datasets: 1992-93 Health and Retirement Study (HRS) and 1996 Survey of Income and Program Participation (SIPP) with the sample size 12503 and 29807 respectively. Respondents' employment status (L), reported disability status (X) and grouped age (Z) can be directly read from the data. As for the verification status (Y), it depends on how researchers use prior information to classify the verified group. They consider five different ways to define the verified subpopulation: a) disability beneficiaries; b) those verified in Wave 2; c) gainfully employed workers; d) those claiming no disability in the current wave; e) all of the above. Readers are referred to Kreider and Pepper (2007, p.435) for the detailed definition of subgroups.

From the data, the analogue joint probability of (L,X,Y,Z) are obtained, and then the analogue bounds of employment gap are computed. Then we use simultaneous multi-level bootstraps to correct the biases. The estimated bounds are presented in Table 2. T_1 and T_2 are the raw analogue bounds and first-level bootstrap corrected bounds respectively. Our results are almost identical to what reported by Kreider and Pepper (2007) in their Table 4, despite that they used the standard non-parametric bootstrap and we use normal distribution with estimated variances to correct the biases. This is because the current sample size is large, and the estimated probability vector

is well approximated by the multivariate normal variates. As a result, our parametric bootstrap works well.

In the finite sample, the raw analogue bounds are narrower than the true bounds on average. After the bootstrap correction, the bounds are enlarged. It seems that the first-level bootstrap does not fully remove the bias since higher order bootstraps further enlarge the estimated bounds. This is most apparent for the HRS data. For example, in the beneficiaries verification scenario the analogue bounds are $[-0.959, 0.809]$, first-level bootstrap magnify the bounds to $[-0.971, 0.830]$, and further bootstraps expand them to $[-0.975, 0.836]$ and $[-0.978, 0.839]$, and so on. Of course the speed of expand decreases with the level of bootstraps. As an empirical guide, when the expansion mitigates, it is better to stop increasing the bootstrap levels. For the SIPP data, the sample size is twice as large as that of the HRS data. Therefore, the speed of bounds expansion are modest. It seems that one or two level of bootstraps suffice to remove most of the biases.

It is worth mentioning that the conservative estimator T_c provides widest bounds. This is not surprising since the conservative lower (upper) bound is biased downwards (upwards). However, it is not too wide to be informative. Whenever the raw analogue bounds and bootstrap corrected bounds are indecisive on the sign of the employment gap, so are conservative bounds. Only in the last case, the analogue estimator indicates the employment gap in SIPP data is negative and bounded by $[-0.413, -0.224]$. Three levels of bootstraps enlarge the bounds to $[-0.447, -0.199]$, and the conservative estimator also suggests the gap is negative and bounded by $[-0.482, -0.131]$.

7. Conclusion

To reduce the finite sample bias of the MIV analogue estimator, the bootstrap correction turns out to be an effective method. Under the asymptotic normality and known variance assumptions, we unveil the mechanism of that correction, not in terms of asymptotic refinement but a direct reduction of the upward bias induced by the $\max(\cdot)$ operator. This reduction can be justified by comparing the bias functions before and after the bootstrap correction. Furthermore, since the bias function is bounded above, we can propose a conservative estimator which is biased downwards instead. This offers a cheap solution to practitioners' serious concern over the too-narrow MIV analogue bounds. Monte Carlo evidence suggests the conservative estimator yields a reasonable magnitude of downward bias, so the estimated bounds are not too wide to be informative. Since the bias of the conservative estimator also decays with the increasing sample size, it is most useful when the practitioners have access to a large sample but limited computational resources.

The analysis of bias functions reveals that one level of the bootstrap cannot eliminate the bias in general, and there is also a possibility of over-correction, which can be seen by examining the maximum of two normal variates as their difference in mean varies. The inadequacy of the single bootstrap leaves room for higher level bootstraps, which are shown to be able to further reduce the bias if we assume the bias function can be well approximated by a polynomial function. Mostly importantly, higher level bootstraps do not necessarily suffer from the curse of dimensionality, since a simultaneous simulation strategy can be used to make multi-level bootstraps computationally feasible. Monte Carlo evidence supports our simultaneous

multi-level bootstraps procedure, since we observe the remaining bias does shrink with the order of the bootstrap. For practitioners, once analogue estimates as well as associated standard errors are provided in accordance with Proposition 2 or 3, our Matlab routine can perform the rest of the bias correction.

Appendix A. Proof of Proposition 1

By the properties of the categorical distribution,

$$E(\mathbf{v}_s) = \mathbf{p}$$
$$Cov(\mathbf{v}_s) = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'$$

Since $\widehat{\mathbf{p}} = \frac{1}{n}\sum_{s=1}^{n}\mathbf{v}_s$, it is a strongly consistent estimator of $\mathbf{p}$, and the central limit theorem implies

$$\sqrt{n}\left(\widehat{\mathbf{p}} - \mathbf{p}\right) \xrightarrow{d} N\left[\mathbf{0}, diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right].$$

■

Appendix B. Proof of Proposition 2

The Delta Method implies that

$$\sqrt{n}\left\{\begin{bmatrix} f_L(\widehat{\mathbf{p}}_1) \\ ... \\ f_L(\widehat{\mathbf{p}}_{n_Z}) \end{bmatrix} - \begin{bmatrix} f_L(\mathbf{p}_1) \\ ... \\ f_L(\mathbf{p}_{n_Z}) \end{bmatrix}\right\} \xrightarrow{d} N\left\{\mathbf{0}, \mathbf{G}\left[diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right]\mathbf{G}'\right\},$$

where $\mathbf{G}$ is a block diagonal matrix such that

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & & \\ & \ddots & \\ & & \mathbf{G}_{n_Z} \end{pmatrix}.$$

Since f_L is homogeneous of degree zero, Euler's theorem implies that $\mathbf{G}_i\mathbf{p}_i = 0,\ i = 1, ..., n_Z$. It follows that $\mathbf{G}\mathbf{p}\mathbf{p}'\mathbf{G}' = \mathbf{0}$. As a result, the

n	100	100	100	100	100	100
M	4	4	4	8	8	8
σ^2	1	4	25	1	4	25
T1	0.10	0.15	0.20	0.31	0.42	0.53
T2	0.01	0.03	0.06	0.09	0.14	0.21
T3	0.00	0.01	0.03	0.04	0.07	0.13
T4	-0.01	0.00	0.02	0.02	0.03	0.09
T5	-0.01	-0.01	0.01	0.00	0.01	0.07
Tc	-0.15	-0.16	-0.17	-0.22	-0.23	-0.23
n	500	500	500	500	500	500
M	4	4	4	8	8	8
σ^2	1	4	25	1	4	25
T1	0.02	0.02	0.04	0.08	0.12	0.15
T2	0.00	-0.01	-0.01	0.01	0.03	0.04
T3	-0.01	-0.02	-0.02	0.00	0.01	0.01
T4	-0.01	-0.02	-0.02	0.00	0.00	0.00
T5	-0.01	-0.02	-0.02	-0.01	0.00	0.00
Tc	-0.09	-0.11	-0.12	-0.14	-0.15	-0.16
n	1000	1000	1000	1000	1000	1000
M	4	4	4	8	8	8
σ^2	1	4	25	1	4	25
T1	0.00	0.01	0.02	0.04	0.07	0.09
T2	-0.01	-0.01	0.00	0.00	0.01	0.02
T3	0.00	-0.01	0.00	0.00	0.01	0.01
T4	0.00	-0.01	0.00	-0.01	0.00	0.01
T5	0.00	-0.01	0.00	-0.01	0.00	0.01
Tc	-0.07	-0.09	-0.09	-0.11	-0.12	-0.13

T1 is the average bias of the naive estimator (maximum of the sample). T2 is the average bias of first-level bootstrap corrected estimator. T3, T4, T5 are biases of second-, third-, fourth- level bootstrap corrected estimators. Tc is the bias of the (downward biased) conservative estimator. Two decimals are retained since the average numerical standard error is 0.007 (maximum 0.022, minimum 0.002)

Table 1: Bias of analogue estimate of the MIV lower bound with the bootstrap correction

HRS	Beneficiaries	Wave 2	Workers	No disability	All of above
T1	[-0.959, 0.809]	[-0.741, 0.645]	[-0.811, 0.350]	[-0.760, 0.350]	[-0.402, -0.341]
T2	[-0.971, 0.830]	[-0.760, 0.672]	[-0.824, 0.358]	[-0.767, 0.358]	[-0.430, -0.307]
T3	[-0.975, 0.836]	[-0.763, 0.681]	[-0.826, 0.359]	[-0.766, 0.359]	[-0.434, -0.302]
T4	[-0.978, 0.839]	[-0.764, 0.688]	[-0.826, 0.359]	[-0.766, 0.359]	[-0.434, -0.300]
Tc	[-0.980, 0.857]	[-0.794, 0.704]	[-0.847, 0.383]	[-0.788, 0.383]	[-0.492, -0.217]

SIPP	Beneficiaries	Wave 2	Workers	No disability	All of above
T1	[-0.967, 0.908]	[-0.793, 0.869]	[-0.784, 0.318]	[-0.781, 0.318]	[-0.413, -0.224]
T2	[-0.974, 0.915]	[-0.804, 0.880]	[-0.794, 0.322]	[-0.785, 0.322]	[-0.437, -0.202]
T3	[-0.977, 0.916]	[-0.808, 0.882]	[-0.795, 0.322]	[-0.786, 0.322]	[-0.444, -0.199]
T4	[-0.978, 0.917]	[-0.811, 0.883]	[-0.795, 0.322]	[-0.786, 0.322]	[-0.447, -0.199]
Tc	[-0.982, 0.925]	[-0.820, 0.900]	[-0.816, 0.346]	[-0.797, 0.346]	[-0.482, -0.131]

Beneficiaries, Wave 2, Workers, No disability are defined identically as in Kreider and Pepper (2007). T1 is the raw analogue estimator,that is,maximum of the sample, comparable to Table 4 in Kreider and Pepper (2007). T2 is first-level bootstrap corrected estimator, comparable to Table 4 in Kreider and Pepper (2007). T3 is second-level bootstrap corrected estimator. T4 is the third-level bootstrap corrected estimator. The upper panel shows the results for the HRS dataset, and the lower panel for SIPP dataset.

Table 2: MIV bounds of employment gap with the bootstrap correction

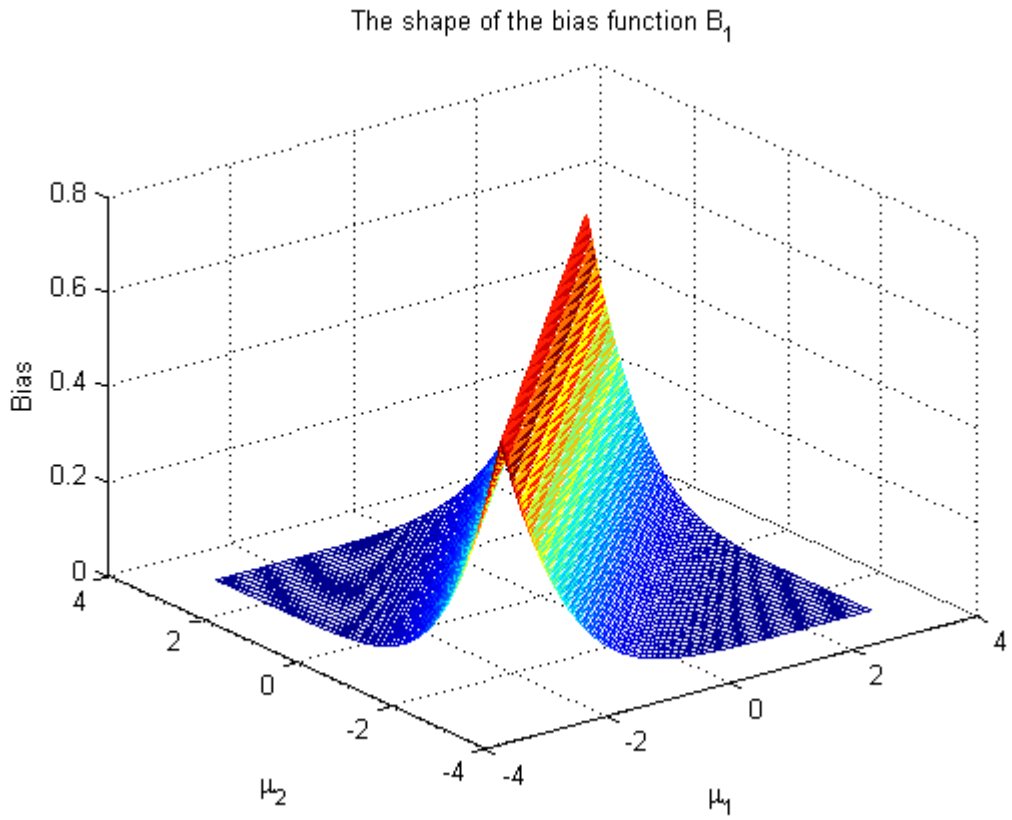

Figure 1: The first-level bias (B_1) is plotted for the case of two normal variates. The two arguments of B_1 function is the mean of the two normal variates. We set $\sigma_1^2 = 1$, $\sigma_2^2 = 1$.

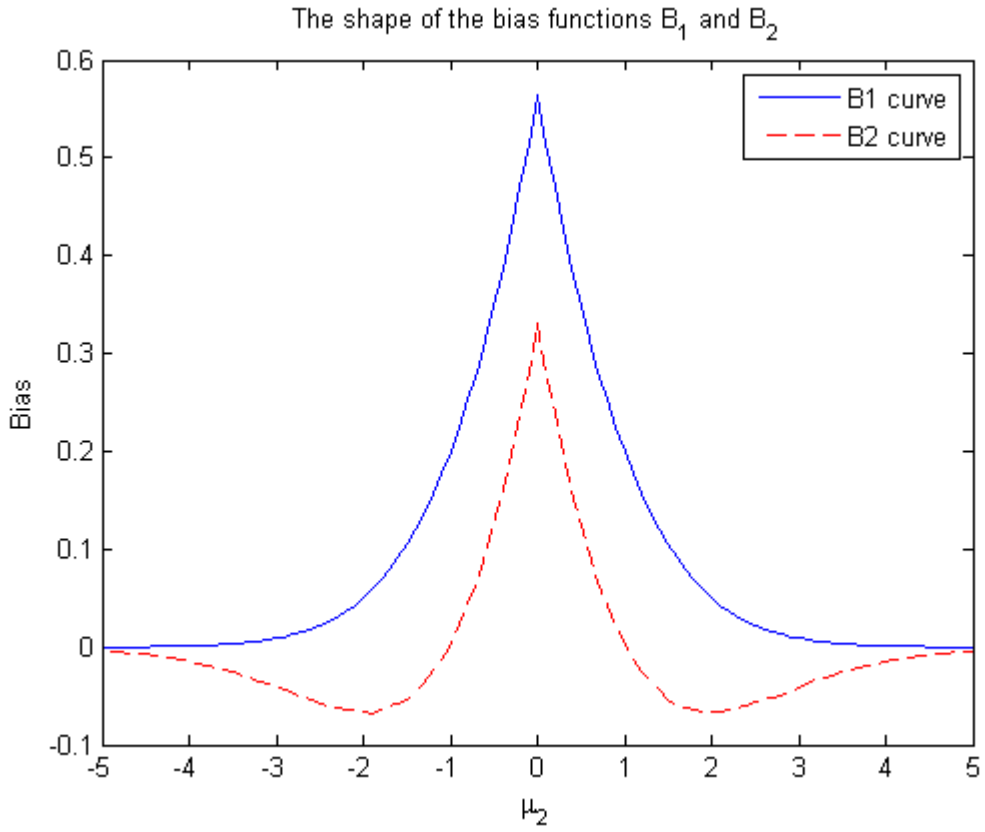

Figure 2: The first level (B_1) and second level (B_2) of the bias functions are plotted for the case of two normal variates. We set $\sigma_1^2 = 1$, $\sigma_2^2 = 1$. Since only the difference in mean matters, μ_1 is normalized to zero. As μ_2 moves, the magnitude of the first-level bias and the second-level bias change accordingly. However, the B_1 curve always lies above the B_2 curve. Though B_1 is always positive, there is a region where B_2 falls below zero.

covariance matrix simplifies to

$$\mathbf{G}\left[diag\left(\mathbf{p}\right)-\mathbf{pp}'\right]\mathbf{G}'=\begin{pmatrix}\mathbf{G}_1\cdot diag\left(\mathbf{p}_1\right)\cdot\mathbf{G}_1' & & \\ & \ddots & \\ & & \mathbf{G}_{n_Z}\cdot diag\left(\mathbf{p}_{n_Z}\right)\cdot\mathbf{G}_{n_Z}'\end{pmatrix}.$$

In the case of the multivariate normal distribution, zero covariance implies independence. ■

Appendix C. Proof of Proposition 3

From Eq. (1),

$$\begin{aligned}f_L\left(\mathbf{p}_i\right)&=E\left(Y\left|Z=z_i,D=d_t\right.\right)\cdot P\left(D=d_t\left|Z=z_i\right.\right)+y_1\cdot P\left(D\neq d_t\left|Z=z_i\right.\right)\\&=E\left[Y\cdot I\left(D=d_t\right)\left|Z=z_i\right.\right]+y_1\cdot E\left[I\left(D\neq d_t\right)\left|Z=z_i\right.\right]\\&=E\left(Q\left|Z=z_i\right.\right)\\&=\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}\frac{p_{ikm}}{p_{i\cdot\cdot}}q_{km}.\end{aligned}$$

The last equality is consistent with Eq. (2).

Now consider sampling variations. Previously in the paper, we use the encoded vectors $\{\mathbf{v}_s\}_{s=1}^n$ to summarize the sample, which defines $\widehat{\mathbf{p}}$ and $\widehat{\mathbf{p}}_1,...,\widehat{\mathbf{p}}_{n_Z}$ as well as $f_L\left(\widehat{\mathbf{p}}_i\right)$ accordingly. We can equivalently use i.i.d. $\{Z_s,Y_s,D_s\}_{s=1}^n$ to denote the sample, where the law of (Z_s,Y_s,D_s) is identical to the representative triple (Z,Y,D). Also define

$$Q_s=Y_s\cdot I\left(D_s=d_t\right)+y_1\cdot I\left(D_s\neq d_t\right).$$

When $\widehat{p}_{i\cdot\cdot} = \frac{1}{n}\sum_{s=1}^{n} I\left(Z_s = z_i\right) > 0$, the analogue probability estimator $f_L\left(\widehat{\mathbf{p}}_i\right)$ is well-defined and can be written as

$$\begin{aligned}
f_L\left(\widehat{\mathbf{p}}_i\right) &= \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}\left[\frac{\frac{1}{n}\sum_{s=1}^{n} I\left(Z_s = z_i, Y_s = y_k, D_s = d_m\right)}{\frac{1}{n}\sum_{s=1}^{n} I\left(Z_s = z_i\right)} q_{km}\right] \\
&= \sum_{k=1}^{n_Y}\sum_{m=1}^{n_D}\left[\frac{\sum_{s=1}^{n} I\left(Z_s = z_i, Q_s = q_{km}\right)}{\sum_{s=1}^{n} I\left(Z_s = z_i\right)} q_{km}\right] \\
&= \frac{\sum_{s=1}^{n}\left[\sum_{k=1}^{n_Y}\sum_{m=1}^{n_D} q_{km} I\left(Q_s = q_{km}\right)\right]\cdot I\left(Z_s = z_i\right)}{\sum_{s=1}^{n} I\left(Z_s = z_i\right)} \\
&= \frac{\sum_{s=1}^{n} Q_s \cdot I\left(Z_s = z_i\right)}{\sum_{s=1}^{n} I\left(Z_s = z_i\right)} \equiv \widetilde{f_L}\left(\mathbf{p}_i\right).
\end{aligned}$$

Note that $\widetilde{f_L}\left(\mathbf{p}_i\right)$ is simply the analogue moment estimator for $E\left(Q\,|Z = z_i\right)$. It indicates that whether we use analogue probability or analogue moment, the functional form of the estimator is the same. Working on the variance of $\widetilde{f_L}\left(\mathbf{p}_i\right)$ is easier than directly computing the variance of $f_L\left(\widehat{\mathbf{p}}_i\right)$.

To make notations compact, denote $\theta \equiv f_L\left(\mathbf{p}_i\right)$, $\widetilde{\theta} \equiv \widetilde{f_L}\left(\mathbf{p}_i\right) = f_L\left(\widehat{\mathbf{p}}_i\right)$, $\gamma \equiv Var\left(Q\,|Z = z_i\right)$.

From here to the end of the proof, when we write $E\left(\cdot\right)$, we leave implicit that the expectation is conditional on $\widehat{p}_{i\cdot\cdot} > 0$.

Using the law of iterated expectations, we have

$$\begin{aligned}
E\left(\widetilde{\theta}\right) &= E\left[E\left(\widetilde{\theta}\,|\{Z_s\}_{s=1}^{n}\right)\right] \\
&= E\left[\frac{\sum_{s=1}^{n}\theta I\left(Z_s = z_i\right)}{\sum_{s=1}^{n} I\left(Z_s = z_i\right)}\right] \\
&= \theta.
\end{aligned}$$

Then the variance of $\widetilde{\theta}$ equals

$$\begin{aligned}
Var\left(\widetilde{\theta}\right) &= E\left[E\left(\widetilde{\theta}^2 \left|\{Z_s\}_{s=1}^n\right.\right)\right] - \theta^2 \\
&= E\left\{\frac{\sum_{a=1}^n \sum_{b=1}^n E\left(Q_a Q_b \left|\{Z_s\}_{s=1}^n\right.\right) I\left(Z_a = z_i\right) I\left(Z_b = z_i\right)}{\sum_{a=1}^n \sum_{b=1}^n I\left(Z_a = z_i\right) I\left(Z_b = z_i\right)}\right\} - \theta^2 \\
&= E\left\{\frac{\sum_{a=1}^n \sum_{b=1}^n \theta^2 I\left(Z_a = z_i\right) I\left(Z_b = z_i\right) + \sum_{a=1}^n \gamma I\left(Z_a = z_i\right)}{\sum_{a=1}^n \sum_{b=1}^n I\left(Z_a = z_i\right) I\left(Z_b = z_i\right)}\right\} - \theta^2 \\
&= E\left[\frac{1}{\sum_{a=1}^n I\left(Z_a = z_i\right)}\right] \cdot \gamma \\
&= \left[\sum_{r=1}^n \frac{1}{r} \frac{\binom{n}{r}\left(p_{i\cdot\cdot}\right)^r \left(1 - p_{i\cdot\cdot}\right)^{n-r}}{1 - \left(1 - p_{i\cdot\cdot}\right)^n}\right] \cdot \gamma
\end{aligned}$$

Note that in the second and third equality, $E\left(Q_a Q_b \left|\{Z_s\}_{s=1}^n\right.\right)$ itself does not equal to $E\left(Q_a Q_b \left| Z_a = z_i, Z_b = z_i \right.\right)$. However, $E\left(Q_a Q_b \left|\{Z_s\}_{s=1}^n\right.\right) I\left(Z_a = z_i\right) I\left(Z_b = z_i\right)$ equals $E\left(Q_a Q_b \left| Z_a = z_i, Z_b = z_i \right.\right) I\left(Z_a = z_i\right) I\left(Z_b = z_i\right)$. For $a \neq b$, $E\left(Q_a Q_b \left| Z_a = z_i, Z_b = z_i \right.\right) = \theta^2$; for $a = b$, $E\left(Q_a Q_b \left| Z_a = z_i, Z_b = z_i \right.\right) = \theta^2 + \gamma$. The results follows. ■

Appendix D. Proof of Proposition 4

Jensen's inequality implies $B_1\left(\boldsymbol{\mu}\right)$ is bounded below by zero. To show it is also bounded above, we first show $E\left[T_1\left(\mathbf{X}\right)\right]$ is strictly increasing in each μ_i. As the maximum of j normal variates, $T_1\left(\mathbf{X}\right)$ has the c.d.f.

$$F\left(c; \boldsymbol{\mu}\right) = \prod_{i=1}^j P\left(X_i \leq c\right) = \prod_{i=1}^j \Phi\left(c - \mu_i; 0, \sigma_i^2\right).$$

Since the normal c.d.f. is a strictly increasing function, $F\left(c; \boldsymbol{\mu}\right)$ is strictly decreasing in $\boldsymbol{\mu}$. To evaluate the expectation, we use the formula, as is suggested by David (1981) and Ross (2010),

$$E\left[T_1\left(\mathbf{X}\right)\right] = \int_0^\infty \left[1 - F\left(c; \boldsymbol{\mu}\right) - F\left(-c; \boldsymbol{\mu}\right)\right] dc,$$

It follows that $E[T_1(\mathbf{X})]$ is strictly increasing in each μ_i. Also note that $\max(\boldsymbol{\mu})$ is merely non-decreasing in each μ_i. Therefore, to maximize $B_1(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, a necessary condition is $\mu_a = \mu_b$, $\forall a, b = 1, ..., j$. Otherwise, consider $\mu_a < \mu_b$, for some a, b. Let $\Delta' = \mu_b - \mu_a$, then increasing μ_a by Δ' will increase $E[T_1(\mathbf{X})]$ while leaving $\max(\boldsymbol{\mu})$ unchanged. A contradiction to the maximum.

Lastly, by the property of the $\max(\cdot)$ function,

$$\begin{aligned} B_1(\boldsymbol{\mu}+c\cdot\boldsymbol{\iota}) &= E[T_1(\mathbf{X})+c]-[\max(\boldsymbol{\mu})+c] \\ &= B_1(\boldsymbol{\mu}), \end{aligned}$$

$\forall c \in \mathbb{R}$, where $\boldsymbol{\iota}$ is a vector of ones. This implies as long as $\mu_a = \mu_b \equiv \mu_0$, $\forall a, b = 1, ..., j$, $B_1(\cdot)$ does not depend on specific choice of μ_0. We pick $\mu_0 = 0$, and $B_1(\mathbf{0})$ attains the maximum $E[\max(\mathbf{X}_0)]$.■

Appendix E. Proof of Proposition 5

$$\begin{aligned} B_2(\boldsymbol{\mu}) &= E[T_1(\mathbf{X}) - B_1(\mathbf{X})] - \max(\boldsymbol{\mu}) \\ &= B_1(\boldsymbol{\mu}) - E[B_1(\mathbf{X})]. \end{aligned}$$

Proposition 4 indicates that $B_1(\boldsymbol{\mu}) > 0$, $\forall \boldsymbol{\mu} \in \mathbb{R}^j$, so that $E[B_1(\mathbf{X})] > 0$. So we have $B_2(\boldsymbol{\mu}) < B_1(\boldsymbol{\mu})$. ■

Appendix F. Proof of Proposition 6

To show the proposition, we first introduce a lemma.

Lemma: The n^{th} (uncentered) moment of $N(\mu, \sigma^2)$ is a polynomial of order n with respect to μ. The leading coefficient (that of μ^n) is one, and the next leading coefficient (that of μ^{n-1}) is zero.

Proof: It is well known that the central moment of $N(\mu, \sigma^2)$ has a closed-form expression.

$$E\left[(X-\mu)^n\right] = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \sigma^d (n-1)!! & \text{if } n \text{ is even} \end{cases},$$

where $(n-1)!!$ is the double factorial. This implies that $E\left[(X-\mu)^n\right]$ is a constant with respect to μ. To find the raw moment $E(X^n)$, we expand $E\left[(X-\mu)^n\right]$ with the formula

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k.$$

Put $a = 1$, $b = -1$, we have $\sum_{k=0}^{n} \binom{n}{k} (-1)^k = 0$, or $\sum_{k=1}^{n} \binom{n}{k} (-1)^k = -1$. We will show the lemma by induction. Clearly, the it holds for $n = 1$. Suppose it is true for the first $n-1$ raw moments, we want to show it holds for the n^{th} raw moment. Note that

$$E\left[(X-\mu)^n\right] = E(X^n) + \sum_{k=1}^{n} \binom{n}{k} (-\mu)^k E\left(X^{n-k}\right).$$

As is assumed, $E\left(X^{n-k}\right)$ is a polynomial of order $n-k$, the leading coefficient is one and the next leading coefficient is zero, hence $\sum_{k=1}^{n} \binom{n}{k} (-\mu)^k E\left(X^{n-k}\right)$ is a polynomial of order n, the leading coefficient is $\sum_{k=1}^{n} \binom{n}{k} (-1)^k = -1$, and the next leading coefficient is zero. It follows that $E(X^n)$ is a polynomial of order n, with the leading coefficient being one and the next leading coefficient being zero. This proves the lemma.

Now put $r = 2$ and consider $B_r(\boldsymbol{\mu}) = B_{r-1}(\boldsymbol{\mu}) - E[B_{r-1}(\mathbf{X})]$. Since $B_{r-1}(\boldsymbol{\mu})$ is a polynomial of order d w.r.t. $\boldsymbol{\mu}$, so the leading term takes the form $\prod_{i=1}^{j}\mu_i^{a_i}$, where $\sum_{i=1}^{j}a_i = d$. The corresponding term in $E[B_{r-1}(\mathbf{X})]$ takes the form $E\left(\prod_{i=1}^{j}X_i^{a_i}\right) = \prod_{i=1}^{j}E(X_i^{a_i})$. By the lemma, $E(X_i^{a_i})$ is a polynomial of order a_i w.r.t. μ_i, and the coefficient of the leading term $\mu_i^{a_i}$ is one, and the coefficient of the next leading term $\mu_i^{a_i-1}$ is zero. This implies that $\prod_{i=1}^{j}E(X_i^{a_i})$ is a polynomial of order d w.r.t. $\boldsymbol{\mu}$, with the leading term (of order d) coefficient one and next leading terms (of order $d-1$) zero. As a result, when $B_{r-1}(\boldsymbol{\mu})$ is subtracted by $E[B_{r-1}(\mathbf{X})]$, the terms corresponding to order d and $d-1$ are canceled, so the order of the polynomial is reduced by 2. The same arguments can be applied to $r = 3, 4, 5$, etc. ■

Appendix G. Proof of Proposition 7

Let $A \equiv E[g(\xi_i, \eta_i)] = E[g(\xi_j, \eta_{j,k})]$, $B \equiv Var[g(\xi_i, \eta_i)] = Var[g(\xi_j, \eta_{j,k})]$, $\forall i = 1, ..., N^2$, $j = 1, ..., N$, $k = 1, ..., N$. The two equalities hold because $(\xi_j, \eta_{j,k})$ are drawn by the method of composition, the joint distribution of $(\xi_j, \eta_{j,k})$ is the same as that of directly sampled (ξ_i, η_i). Clearly, $E(S_1) = E(S_2) = A$, $Var(S_1) = \frac{1}{N^2}B$. When we compute $Var(S_2)$, we

need to consider the covariance terms as well.

$$\begin{aligned} Var\left(S_2\right) &= \frac{1}{N} Var\left[\frac{1}{N}\sum_{k=1}^{N} g\left(\xi_1, \eta_{1,k}\right)\right] \\ &= \frac{1}{N}\frac{1}{N^2}\sum_{k=1}^{N}\sum_{h=1}^{N} cov\left[g\left(\xi_1, \eta_{1,k}\right), g\left(\xi_1, \eta_{1,h}\right)\right] \\ &= \frac{1}{N^2} B + \frac{1}{N^3}\sum_{k=1}^{N}\sum_{h=1, h\neq k}^{N} cov\left[g\left(\xi_1, \eta_{1,k}\right), g\left(\xi_1, \eta_{1,h}\right)\right]. \end{aligned}$$

Now we show each of those covariance terms is non-negative.

$$\begin{aligned} & cov\left[g\left(\xi_1, \eta_{1,k}\right), g\left(\xi_1, \eta_{1,h}\right)\right] \\ &= E\left\{\left[g\left(\xi_1, \eta_{1,k}\right) - A\right] \cdot \left[g\left(\xi_1, \eta_{1,h}\right) - A\right]\right\} \\ &= E_{\xi_1}\left\{E_{\eta_{1,k}|\xi_1}\left[g\left(\xi_1, \eta_{1,k}\right) - A\right] \cdot E_{\eta_{1,h}|\xi_1}\left[g\left(\xi_1, \eta_{1,h}\right) - A\right]\right\} \\ &= E_{\xi_1}\left[c^2\left(\xi_1\right)\right] \geq 0, \end{aligned}$$

where $c\left(\xi_1\right) \equiv E_{\eta_{1,k}|\xi_1}\left[g\left(\xi_1, \eta_{1,k}\right) - A\right] = E_{\eta_{1,h}|\xi_1}\left[g\left(\xi_1, \eta_{1,h}\right) - A\right]$. It follows that $Var\left(S_1\right) \leq Var\left(S_2\right)$.

Note that in the above proof, $Var\left(S_1\right) = Var\left(S_2\right)$ only if $E_{\eta_{1,k}|\xi_1}\left[g\left(\xi_1, \eta_{1,k}\right)\right] = A$ for all realizations of ξ_1. The independency of ξ and η does not necessarily imply $Var\left(S_1\right) = Var\left(S_2\right)$. When we take conditional expectation of $g\left(\xi_1, \eta_{1,k}\right)$, ξ_1 should be treated as a constant and in general $c\left(\xi_1\right) \neq 0$, even if for independent variates. ■

Cain, M., 1994. The moment-generating function of the minimum of bivariate normal random variables. The American Statistician 48, 124–125.

Chernozhukov, V., Lee, S. S., Rosen, A., 2009. Intersection bounds: estimation and inference. CeMMAP working papers CWP19/09.

Clark, C. E., 1961. The greatest of a finite set of random variables. Operations Research 9, 145–162.

David, H. A., 1981. Order Statistics. Wiley.

Davidson, R., MacKinnon, J., 2002. Fast double bootstrap tests of nonnested linear regression models. Econometric Reviews 21 (4), 419–429.

Davidson, R., MacKinnon, J. G., 2007. Improving the reliability of bootstrap tests with the fast double bootstrap. Computational Statistics & Data Analysis 51 (7), 3259–3281.

Kreider, B., Pepper, J. V., 2007. Disability and employment: Reevaluating the evidence in light of reporting errors. Journal of the American Statistical Association 102, 432–441.

Manski, C. F., Pepper, J. V., 2000. Monotone instrumental variables, with an application to the returns to schooling. Econometrica 68 (4), 997–1012.

Manski, C. F., Pepper, J. V., 2009. More on monotone instrumental variables. Econometrics Journal 12 (s1), S200–S216.

Ross, A., 2010. Computing bounds on the expected maximum of correlated normal variables. Methodology and Computing in Applied Probability 12, 111–138.