On Detecting Discrimination in Treatment Assignment

Debopam Bhattacharya,
University of Oxford

First draft February, 2009.
This draft: July, 2010.

Abstract

In real-life, individuals are often assigned to binary treatments based on established covariate-based protocols. Direct or implicit taste-based discrimination would make such protocols economically inefficient in that the expected gain from treatment would be smaller for a subset of the currently treated than the currently untreated. We present a framework for detecting such inefficiency using a partial identification approach which continues to work when the decision-maker observes more covariates than us. We also propose a novel way of inferring the relevant counterfactual distributions by combining observational datasets with experimental estimates. The method can be extended to (partially) infer risk-preferences of the decision-maker, under which observed allocations are efficient. The most risk neutral solution may be obtained via maximizing entropy. We outline the theory of inference and study the efficacy of our methodology using a simulation exercise. Our methods apply when individuals cannot alter their potential treatment outcomes in response to the decision-maker's actions unlike the case of law enforcement (c.f., Knowles, Persico and Todd (2001)).

1 Introduction

In many real-life situations, external decision-makers assign individuals to treatments. Examples include banks approving mortgage and business loans, doctors referring patients

to surgery and firms hiring interns. However, existing protocols for deciding treatment status may be economically inefficient in that the expected gain from treating a subgroup of those currently being treated is smaller than that from treating a subset of the untreated individuals, the expectation being taken with respect to the true underlying distribution of the random variables concerned. This situation implies that the treatment, to be thought of as a scarce resource, is being assigned among individuals in a way that does not maximize its overall productivity. A leading cause of such inefficiency is prejudice against specific demographic groups– either direct or implicit– but inefficiency can also arise from the failure to condition on relevant covariates or systematic biases in the decision-maker's (DM, henceforth) subjective expectation. When such misallocation occurs, we say that there is non-statistical discrimination against the untreated subgroup with the higher potential gain. We present a framework for detecting and analyzing such misallocation using a partial identification approach. Unless one has access to exactly the same variables as the DM, one can at best detect misallocation but cannot in general infer if misallocation has occurred because of prejudice or some other, perhaps more innocuous, reason.[1]

The present paper focuses on the case where the treatment in question is binary and the outcome of interest either binary or continuous. We assume that an experienced DM observes for each individual a set of covariates and assigns him/her to treatment based on the expected gains from treatment, conditional on these covariate values. In this set-up, a necessary condition for the DM's assignment to be productively efficient is that in every observable covariate group, the expected net benefit of treatment (relative to cost) to the marginal treatment recipient, i.e., the "last person" to have received treatment is equal to or greater than[2] a common threshold which, in turn, is greater than or equal to the expected net benefit of the "first person" to have been denied treatment. Here, last and first refer to the types of individuals with the smallest and the largest expected benefit from treatment, respectively, where types are defined in terms of the characteristics observed by the DM. Misallocation occurs when the benefit to the marginal recipient differs significantly across covariates and it can arise if, for instance, different thresholds are used for different covariate groups, which would happen if the DM was prejudiced against certain groups. The DM's assignment results in an observational dataset, where for each individual, we observe her treatment status, her outcome conditional on her

[1] Thus, here we are making a conscious distinction between "non-statistical discrimination" and "prejudice"– a point on which we elaborate below.

[2] Strict inequality can occur at the margin if, for instance, all covariates are discrete,

treatment status and a set of covariate values. The problem is to detect misallocation of treatment from this dataset.

Detecting inefficient treatment allocation from such an observational dataset alone is complicated by two reasons. The first is that the DM can base treatment assignment on characteristics that are not observed by us. This makes it is hard, if not impossible, to know who are the "marginal" treatment recipient and non-recipients– a problem already recognized in the literature (c.f., Heckman (1998), Persico (2009)) and labeled "inframarginality". Further, benefits are also hard to measure with a single observational dataset because, as is well known, counterfactual means are not observed.

In this paper, we discuss a new approach to detecting misallocation in such situations. The key idea is to use the notion of partial identification, motivated by the implication of efficient allocation that expected net benefits in every subset of the treated group must weakly exceed expected net benefits in every subset of the untreated group– a (conditional) moment inequality condition. These moment inequalities for subsets defined by covariates that the DM observes have testable implications for the (cruder) subsets based on the covariates that we observe. These implications can therefore be tested. Secondly, we propose a novel way to identify the necessary counterfactual means by combining an observational dataset with experimental estimates on subjects drawn from the same population. The latter supplements existing methods of identifying counterfactuals using, say, instrumental variables.

The bulk of our analysis rests upon three assumptions. The first is that the DM is experienced in the sense that he can form correct expectations. The second is that the DM observes and can condition treatment allocation on all the characteristics (and possibly more than) those that we observe. Third, we observe the same *outcomes* whose expectations, taken by the DM, determine treatment assignment in the observational dataset.

The third assumption simply clarifies that the definition of productivity (with respect to which inefficiency is defined) must be unanimously agreed upon and this common measure of productivity should be observable and verifiable. The second assumption seems quite natural; but we discuss below in section 4 when it might fail and the implications thereof. The first assumption–a "rational expectations" idea– is part of our definition of efficiency, i.e., we are testing the joint hypothesis that the DM can calculate correct expectations *and* is allocating treatment efficiently, based on those calculations. In other

words, we cannot distinguish between the cases where the DM's subjective expectations deviate from the true expectations in a specific, systematic way and lead to the inefficiency observed and the case where the DM can form correct expectations but ends up with an inefficient allocation due to, say, prejudice against specific demographic groups. Thus, rational expectations can be assumed here without loss of generality since our ultimate aim is to detect misallocation–whatever its cause.

The rest of the paper is organized as follows. Section 2 discusses the contribution of the present paper in relation to the existing literature in economics and econometrics. Section 3 presents the partial identification methodology, discusses how counterfactuals may be identified via data combination and demonstrates how the analysis is robust to failure of a key identical distribution assumption which underlies the data combination method. Section 4 discusses some, albeit subtle, difficulties which prevent one from pinpointing the cause of misallocation even when it has been detected. Section 5 discusses the complementary problem of inferring a DM's underlying risk-preferences which would justify the current allocations as efficient. Section 6 briefly outlines the theory of inference. Section 7 presents a simulation study and section 8 concludes.

We would like to end this section by re-emphasizing that our methodology can detect inefficiency of treatment assignment and identify the demographic groups that are suffering as a result of it but it cannot distinguish between the various causes which might lead a DM to such an allocation. The existing literature in economics sometimes uses the term "statistical discrimination" to mean an efficient allocation that leads to disparities in treatment rates but, in our reading, it is unclear about whether "non-statistical" discrimination can arise only from prejudice. The present paper analyzes this latter point in greater details than previously attempted and presents some examples to illustrate the subtleties involved in the analysis.

2 Literature and contributions

It is useful at this stage to contrast our work with some existing empirical approaches within economics to detect "discrimination" in the assignment of a binary treatment. Recall the key identification problem, viz., that it is impossible to tell if the group receiving less treatment does so because it is less endowed with those determinants of productivity which we do not observe but the DM does or because the DM is prejudiced against

that group. To get around the problem of confounding unobservables, Bertrand and Mullainathan (2004) randomly assigned black and white names to fake resumes and found that resumes with black names are less favorably regarded by potential employers. Similarly, Goldin and Rouse compared blind auditions for musicians with non-blind ones and found that significantly more women were selected in blind auditions. If we *assume* that race (gender) has no additional information about the relevant productivity over and above the resume (the audition performance), then the two cited studies can be regarded as having detected prejudice. It is, however, conceivable that race or gender do play an important role in general in the eventual outcomes that DMs care about. Thus, without actually observing the final outcomes of interest, one cannot distinguish between statistical and non-statistical sources of the disparities observed.

In a series of studies, Knowles, Persico, Todd (2001) and several other authors have examined the problem of detecting taste-based prejudice separately from statistical discrimination in the context of vehicle search by the police, using data on final outcomes (hit rates). The key insight is that in law-enforcement contexts, potential treatment recipients can alter their behavior– and thus their potential outcome upon being treated– in response to the treator's behavior. This argument enables them to solve the problem of "inframarginality" and test for the economic efficiency of the existing allocation. While their approach applies to many situations of interest, especially ones involving law enforcement, it is not applicable to all situations of treatment assignment where misallocation is a concern. For example, it is very difficult– if not impossible– for patients to alter their potential health outcomes with and without surgery in response to the nature of treatment protocols used by doctors.

In independent and ongoing work, Chandra and Staiger (2010) consider the problem of identifying provider prejudice in intensive treatments for heart- attacks. They use an instrumental variable approach to identify counterfactuals and attempt to test equality of treatment thresholds under a strong high-level assumption on the distribution of unobservables for the two groups being compared (e.g., males and females). Essentially, their method works if either the c.d.f. of benefit distribution of one group is simply that of the other group, translated by a fixed amount, or if the unobservable distribution is identical for the two groups, conditional on the observables–leading to a "single index" structure (c.f., Powell (1994)). But the latter essentially assumes away the inframarginality problem which is the central source of difficulty in detecting inefficiency.

On the econometric side, our paper links the discrimination literature in economics with the partial identification approach, pioneered by Manski, that has been receiving a lot of recent attention. Our paper attempts to show how one can use the partial identification idea to make progress in solving an important and difficult detection problem long recognized in the economics literature. In particular, we analyze the problem of inferring the set of (sub)-utility functions of the DM which would rationalize an observed allocation pattern. We also show that a maximum entropy solution in this case gives us that admissible utility which is closest to risk neutrality and this solution is easy to compute and report rather than the entire identified set. This focus on finding specific parameter values which have intuitive or economic interpretation among all the ones in an identified set appears to be novel. It also raises interesting inferential problems pertaining to optimization with stochastic constraints which, to our knowledge, have not been explored before.

A series of papers in the forecasting literature propose testing rationality of forecasts made by central agencies (c.f. Elliott, Komunjer and Timmerman (2005), Patton and Timmerman (2007) etc.). The idea is to (point) estimate parameters of the loss-function which rationalize the observed forecasts. The set-up in that literature assumes that the action (i.e., the forecast) has no effect on the distribution of the future outcome. In contrast, the key issue in our set-up is that the action (treatment status) fundamentally affects the distribution of the outcome and so the methodology of forecast rationality tests cannot be used in our problem.

Lastly, our identification approach uses counterfactual means and we propose a novel way of obtaining them by combining the observational dataset with estimates from an experimental study where individuals were randomized into and out of treatment.[3] This method works best when the experimental group is drawn from the same population as the observational one. However, the method works even in the case where the individuals willing to be randomized are worse (i.e., have worse outcomes with and without the treatment) along unobserved dimensions, which is sometimes the case in medical trials. The cost of this complication is that it will be harder for us to detect misallocation relative to the case where the two distributions are identical.

[3]In the context of Bertrand and Mullainathan (2004) for example, our proposed method would be to randomly offer jobs to individuals and record individual productivity and then combine this with a dataset where employers were allowed to choose among applicants and the resulting productivities were recorded.

3 Methodology

Denote outcome with and without treatment by Y_0 and Y_1, respectively and let $\Delta Y = Y_1 - Y_0$. Analogously, define C_1 and C_0 as the potential costs corresponding to treatment and no treatment, respectively. Let $W = (X, Z)$ denote the covariates observed by the DM, where the component Z is not observed by us. Let $\mathcal{X}^j$ denote the support of X for the subpopulation who would be assigned $D = j$, $j = 0, 1$ by the DM and E^{sub} denote expectations taken w.r.t. the DM's subjective probability distributions, P^{sub}.

Consider the DM's optimization problem

$$\max_A \left\{ \int 1\{w \in A\} y_1 dP^{sub}_{Y_1,W}(y_1, w) + \int 1\{w \in A^c\} y_0 dP^{sub}_{Y_0,W}(y_0, w) \right\}$$

s.t.

$$\int 1\{w \in A\} c_1 dP^{sub}_{C_1,W}(c_1, w) + \int 1\{w \in A^c\} c_0 dP^{sub}_{C_0,W}(c_0, w) \leq c.$$

The solution, as shown in the appendix, is of the form

$$\begin{aligned} A^* &= \{w : \beta(w) < \gamma\}, \text{ with} \\ \beta(w) &\equiv \frac{E^{sub}(\Delta C | W = w)}{E^{sub}(\Delta Y | W = w)}, \\ c &= \int_{w \in \boldsymbol{w}} 1(\beta(w) < \gamma) dP^{sub}_W(w). \end{aligned} \tag{1}$$

Although this solution is intuitive, a formal proof is needed because other criteria like $E[\frac{\Delta C}{\Delta Y} | W = w] < \gamma$, or $E[\frac{C_1}{Y_1} - \frac{C_0}{Y_0} | W = w] < \gamma$, etc., which seem intuitively just as sensible, do not solve the problem!

Since the DM's subjective expectations are assumed to be consistent with true distributions in the population, we must have that w.p.1,

$$\begin{aligned} \gamma E(\Delta Y | X, Z, D = 1) &\geq E(\Delta C | X, Z, D = 1), \\ \gamma E(\Delta Y | X, Z, D = 0) &\leq E(\Delta C | X, Z, D = 0). \end{aligned} \tag{2}$$

Given the allocation procedure leading to (2), as $W = (X, Z)$ varies, γ remains fixed but treatment rates $\Pr(D = 1 | W)$ will in general vary, giving rise to efficient or statistical– as opposed to inefficient or taste-based– discrimination. In contrast, taste-based discrimination will be said to occur if γ varies by W. Another equivalent interpretation is that a w-type individuals will be treated if $\gamma E(\Delta Y | W = w) - E(\Delta C | W = w)$ exceeds zero. Thus γ may be interpreted as the weight being put on the benefit of a w-type person while

the same threshold of zero is being applied to all w. If γ varies by w, then the benefits of different covariate groups are being weighed differently–which would be regarded as discriminatory.

Since we do not observe Z, the inequalities in (2) are not of immediate use to us. However, an implication of (2) is potentially useful for detecting inefficiency. Indeed, (2) implies that

$$\begin{aligned} & \gamma \int E\left(\Delta Y | X, Z, D=1\right) dF_{Z|X,D=1}\left(z|X, D=1\right) \\ \geq & \int E\left(\Delta C | X, Z, D=1\right) dF_{Z|X,D=1}\left(z|X, D=1\right), \end{aligned}$$

i.e.

$$\frac{E\left[\Delta C | D=1, X_1=a\right]}{E\left[\Delta Y | D=1, X_1=a\right]} \leq \gamma, \text{ for all } a \in \mathcal{X}^1, \tag{3}$$

and similarly

$$\frac{E\left[\Delta C | D=0, X_1=a\right]}{E\left[\Delta Y | D=0, X_1=a\right]} > \gamma, \text{ for all } a \in \mathcal{X}^0. \tag{4}$$

In words, if the DM is acting rationally, then the ratio of average gain from treatment and average increase in cost for every subgroup (that the DM can observe) among the treatment recipients must exceed the treatment threshold. Since this would have to hold for every subgroup among the treated, it must also hold for groups (observed by us) constructed by aggregating these subgroups and averaging the gain across those subgroups. This leads to (3) and analogously for (4). This reasoning lets us overcome the problem posed by the DM observing more covariates than us and preserves the inequality needed for inference.

It follows now that if for some $a \neq b$, we have that

$$\frac{E\left[\Delta C | D=0, X_1=b\right]}{E\left[\Delta Y | D=0, X_1=b\right]} < \frac{E\left[\Delta C | D=1, X_1=a\right]}{E\left[\Delta Y | D=1, X_1=a\right]},$$

then we conclude that there is misallocation and too few people of type b are being teated.

One example of X_1 in the case of medical treatment is health insurance status. To judge whether providers are discriminating against the uninsured, we need to test the above inequality with $X_1 = b$ denoting the uninsured and $X_1 = a$ denoting the insured. In this case, C_1 and C_0 can denote either total cost of the two treatments or the out-of-pocket cost, borne by the patient. For a loan application example, where $D = 1$ is approving the loan, Y_1 would denote the return on that loan if approved, $Y_0 \equiv 0$,

$C_0 \equiv 0$, and C_1 is the amount of the loan plus administrative costs involved in managing the money-lending procedure. In this latter case, the constraint would be imposed by a regulatory ceiling on how much in aggregate the bank can lend.

To be able to use the above inequalities to learn about γ, we need to identify the counterfactual mean outcomes $E(Y_0|X, D=1)$ and $E(Y_1|X, D=0)$ and the counterfactual mean costs $E(C_0|X, D=1)$ and $E(C_1|X, D=0)$. The econometric literature on treatment effect estimation has proposed a variety of ways to point-identify or provide bounds on these counterfactual means. We propose a new and simple way to point identify these means, viz., we supplement the observational dataset with estimates from an experiment, where individuals are randomized in and out of treatment. If the observational and the experimental samples are drawn from the same population, then combining them will yield the necessary counterfactual distributions. To see this, notice that for any $x \in \mathcal{X}^1$,

$$\begin{aligned}
\underbrace{P(Y_0 < y|X=x)}_{\text{known from expt}} &= P^{obs}(Y_0 < y|X=x) \\
&= P^{obs}(Y_0 < y|D=1, X=x) \times \underbrace{P^{obs}(D=1|X=x)}_{\text{known from obs}} \\
&\quad + \underbrace{P^{obs}(Y_0 < y|D=0, X=x)}_{\text{known from obs}} \times \underbrace{P^{obs}(D=0|X=x)}_{\text{known from obs}}. \qquad (5)
\end{aligned}$$

Similarly for any $x \in \mathcal{X}^0$,

$$\begin{aligned}
\underbrace{\Pr(Y_1 < y|x)}_{\text{known from expt}} &= \Pr(Y_1 < y|D=0, x) \times \underbrace{\Pr(D=0|x)}_{\text{known from obs}} \\
&\quad + \underbrace{\Pr(Y_1 < y|D=1, x)}_{\text{known from obs}} \times \underbrace{\Pr(D=1|x)}_{\text{from obs}}. \qquad (6)
\end{aligned}$$

Thus the two equalities above yield the counterfactual distributions $P(Y_0 < y|D=1, x)$ on $\mathcal{X}^1$ and $P(Y_1 < y|D=0, x)$ on $\mathcal{X}^0$. When we know the means but not the distribution of Y_1 and Y_0 from the experiment, we have to replace the c.d.f.'s in the previous displays by the corresponding means, giving us, for instance, for any $x \in \mathcal{X}^0$,

$$\begin{aligned}
\underbrace{E(Y_1|x)}_{\text{known from expt}} &= E(Y_1|D=0, x) \times \underbrace{\Pr(D=0|x)}_{\text{known from obs}} \\
&\quad + \underbrace{E(Y_1|D=1, x)}_{\text{known from obs}} \times \underbrace{\Pr(D=1|x)}_{\text{from obs}}.
\end{aligned}$$

Combining (3), (4), (5) and (6) yield the following bounds on γ:

$$\gamma_{ub} = \inf_{x \in \mathcal{X}^0} \left(\frac{\underbrace{E\left(C_1 | X = x, D = 0\right)}_{\text{from (5)}} - \underbrace{E\left(C_0 | X = x, D = 0\right)}_{\text{from obs data}}}{\underbrace{E\left(Y_1 | X = x, D = 0\right)}_{\text{from (5)}} - \underbrace{E\left(Y_0 | X = x, D = 0\right)}_{\text{from obs data}}} \right),$$

$$\gamma_{lb} = \sup_{x \in \mathcal{X}^1} \left(\frac{\underbrace{E\left(C_1 | X = x, D = 1\right)}_{\text{from obs data}} - \underbrace{E\left(C_0 | X = x, D = 1\right)}_{\text{from (6)}}}{\underbrace{E\left(Y_1 | X = x, D = 1\right)}_{\text{from obs data}} - \underbrace{E\left(Y_0 | X = x, D = 1\right)}_{\text{from (6)}}} \right). \tag{7}$$

The bounds derived above essentially replace a minimum over finer subgroups (observed by the DM) by the minimum over groups (observed by us) of the subgroup averages. So one would expect the bounds to be wider when (i) the unobserved covariates have larger support making the average across subgroups further from the minimum or maximum across subgroups, and (ii) the observed covariates are correlated with the unobserved ones to a lesser extent. The bounds would collapse to a singleton if we observe the same covariates as the DM. In the loan example, if all that the DM sees is the application form which is also made available to the econometrician, then Z is null and $X = W$, leading to point identification of γ.

Alternative designs and data issues: There are two different ways to perform the data combination exercise. In the first, the observational micro-data are combined with estimates obtained from an experimental study, conducted by other researchers. In practical terms, due to data protection conventions, it is much easier to access experimental estimates than it is to access the raw micro-data from trials which were used to calculate those estimates. However, one has to make sure that the observational group and the experimental group were drawn from the same population and the same covariates were recorded in both cases.

The better but practically harder option is to actually run an experiment, which can also be done in two ways. In the first, a sample of individuals is randomly divided into an experimental arm and a non-experimental one. The experimental arm individuals are randomly assigned to treatment and the observational arm ones are handed over to a DM who uses his/her discretion. The second way is as follows. First, present all the individuals to the DM and record his recommendations for treatment. This recommendation is recorded as $D = 1$ when recommended to have treatment and as $D = 0$, otherwise. Then we randomize actual approval across all applications (ignoring the DM's recommenda-

tion) and observe the outcomes for each individual. The counterfactual $P(Y_0|D=1,X)$ can then be obtained from the outcomes of those who are approved by the DM but were randomized out of treatment. Conversely for $P(Y_1|D=0,X)$.

The experimental approach requires significantly more work to implement but gives us the ideal set-up where the experimental and observational groups are ex-ante identical and the same variables can be recorded for both groups. The first method, where experimental results from existing studies are used instead of actually running an experiment, is applicable in many more situations. However, one is somewhat constrained by the outcomes and covariates that the original researchers had chosen. For the exercise of inferring risk preferences (see section 5, below) in the case of non-binary outcomes, one would need the full experimental approach because trial studies rarely report marginal distributions of Y_0 and Y_1 (rather than means and medians) which are needed to conduct this exercise.

3.1 Misallocation

The bounds analysis presented above can be used to test whether there is misallocation of treatment both within and between demographic groups. To fix ideas, suppose $X = (X_1, female)$ and we are interested in testing if there is treatment misallocation within males and within females and then we want to test if treatment misallocation between males and females occurs in a way that hurts, say, females.

To do these tests, perform the above analysis separately for females and males and get the bounds

$$\Gamma_{fem} = \begin{pmatrix} \sup_{x \in Supp(X_1|fem=1,D=1)} \frac{E[\Delta C|X_1=x,fem=1,D=1]}{E[\Delta Y|X_1=x,fem=1,D=1]}, \\ \inf_{x \in Supp(X_1|fem=1,D=0)} \frac{E[\Delta C|X_1=x,fem=1,D=0]}{E[\Delta Y|X_1=x,fem=1,D=0]} \end{pmatrix}$$

and analogously Γ_{male}. Now, if Γ_{fem} (or Γ_{male}) is empty, then we conclude that there is misallocation within females (males). Further, if $\Gamma_{fem} \cap \Gamma_{male}$ is empty, then it implies that different thresholds were used for females and males and thus there is misallocation between males and females.

Intuition: Why empty sets imply misallocation can be best understood by ignoring X_1 for the time being. Notice that $\Gamma_{fem} \cap \Gamma_{male} = \phi$ means that either

$$\frac{E[\Delta Y|fem=0,D=1]}{E[\Delta C|fem=0,D=1]} < \frac{E[\Delta Y|fem=1,D=0]}{E[\Delta C|fem=1,D=0]} \tag{8}$$

or

$$\frac{E[\Delta Y|fem=1,D=1]}{E[\Delta C|fem=1,D=1]} < \frac{E[\Delta Y|fem=0,D=0]}{E[\Delta C|fem=0,D=0]}. \tag{9}$$

The first inequality (8) means that the expected (rise in) benefit relative to (rise in) cost of treatment among treated males is less than that among untreated females– i.e., females are being under-treated. Equivalently, females face a smaller γ. Similarly, (9) means that males are being under-treated.

Notice that the inequalities (**??**) or (8) can be interpreted and used directly without reference to a specific model of optimization or treatment allocation such as (1) or (**??**). However, the link with (1) and (**??**) gives our analysis a firm grounding in classical economic theory of choice under uncertainty.

3.2 Nonidentical distributions

We now consider the possibility that the observational sample and the experimental sample were drawn from different subsets of the population. For example, sometimes it is the case in medical trials that inherently sicker patients agree to be randomized. In this case, it is reasonable to expect that $E^{\exp}(Y_0|x) \leq E^{obs}(Y_0|x)$ and $E^{\exp}(Y_1|x) \leq E^{obs}(Y_1|x)$. Similarly, $E^{\exp}(C_0|x) > E^{obs}(C_0|x)$ and $E^{\exp}(C_1|x) > E^{obs}(C_1|x)$. Using the same steps as those leading to (5), one gets that

$$\begin{aligned} E^{obs}(Y_0|D=1,x) &= \frac{E^{obs}(Y_0|x) - P^{obs}(D=0|x) \times E^{obs}(Y_0|D=0,x)}{P^{obs}(D=1|x)} \\ &\geq \frac{E^{\exp}(Y_0|x) - P^{obs}(D=0|x) \times E^{obs}(Y_0|D=0,x)}{P^{obs}(D=1|x)} \\ &\equiv \bar{E}(Y_0|D=1,x), \end{aligned}$$

and similarly,

$$\begin{aligned} E^{obs}(Y_1|D=0,x) &= \frac{E^{obs}(Y_1|x) - P^{obs}(D=0|x) \times E^{obs}(Y_1|D=0,x)}{P^{obs}(D=1|x)} \\ &\geq \frac{E^{\exp}(Y_1|x) - P^{obs}(D=0|x) \times E^{obs}(Y_1|D=0,x)}{P^{obs}(D=1|x)} \\ &\equiv \bar{E}(Y_1|D=0,x). \end{aligned}$$

The quantities $\bar{E}(Y_1|D=0,x)$ and $\bar{E}(Y_0|D=1,x)$ are clearly identified. An analogous set of inequalities hold with Y replaced by C and the inequality sign reversed (since the

experimental group, being sicker will be more expensive to treat). These bounds can be used to detect misallocation. For instance, if it is the case that

$$\begin{aligned} & \frac{E^{obs}\left(Y_1|D=1,male\right)-\bar{E}\left(Y_0|D=1,male\right)}{E^{obs}\left(C_1|D=1,male\right)-\bar{E}\left(C_0|D=1,male\right)} \\ \leq\ & \frac{\bar{E}\left(Y_1|D=0,female\right)-E^{obs}\left(Y_0|D=0,female\right)}{\bar{E}\left(C_1|D=0,female\right)-E^{obs}\left(C_0|D=0,female\right)}, \end{aligned} \tag{10}$$

then it follows that

$$\begin{aligned} \frac{1}{\gamma_{male}} &< \frac{E^{obs}\left(\Delta Y|D=1,male\right)}{E^{obs}\left(\Delta C|D=1,male\right)} \\ &\leq \frac{E^{obs}\left(Y_1|D=1,male\right)-\bar{E}\left(Y_0|D=1,male\right)}{E^{obs}\left(C_1|D=1,male\right)-\bar{E}\left(C_0|D=1,male\right)} \\ &\leq \frac{\bar{E}\left(Y_1|D=0,female\right)-E^{obs}\left(Y_0|D=0,female\right)}{\bar{E}\left(C_1|D=0,female\right)-E^{obs}\left(C_0|D=0,female\right)} \\ &\leq \frac{E^{obs}\left(\Delta Y|D=0,female\right)}{E^{obs}\left(\Delta C|D=0,female\right)} \\ &\leq \frac{1}{\gamma_{fem}}. \end{aligned}$$

Thus, γ_{female} is smaller, meaning that the outcomes of females are being weighed less relative to males. However, since (10) implies (8), it will be harder to detect misallocation here compared to when the experimental and observational data came from identical populations.

4 Alternative mechanisms leading to misallocation

We now discuss three alternative allocation mechanisms which can potentially lead to empty identified sets and thus suggest misallocation. We define and make the distinction between wilful or prejudicial discrimination, inadvertent discrimination and implicit discrimination– all of which will lead to misallocation that we can potentially detect with our bounds-based analysis. Instead of presenting general models, we describe specific scenarios to outline the subtleties which make it hard to move from detection of misallocation to discerning its source. For simplicity of exposition, we will assume that ΔC is a constant k (i.e., does not vary with any component of W), so that the optimal decision criterion will be

$$D=1 \Longleftrightarrow E\left(\Delta Y|W\right)>\lambda,$$

where $\lambda = k/\gamma$.

4.1 Inadvertent Discrimination

Suppose individuals are characterized by race (black/white) and gender (male/female). Suppose it is the case that

$$\begin{aligned} E(\Delta Y|fem, black) &> E(\Delta Y|male, White) > E(\Delta Y|male, black) \\ &> \lambda \\ &> > E(\Delta Y|fem, white). \end{aligned} \tag{11}$$

Suppose that the fraction of whites among women is high enough that

$$E(\Delta Y|male) > \lambda > E(\Delta Y|female). \tag{12}$$

That is, black females benefit a lot from treatment while while females benefit the least. If white females are a much larger group than black females, then on average, females benefit less from treatment and hence (12) holds.

Now suppose the DM ignores race and allocates treatment, based only on gender. Then $D = 1$ iff the individual is male and so it must be the case that

$$\begin{aligned} E(\Delta Y|D &= 0, Black) \\ &= E(\Delta Y|female, Black) \\ &> E(\Delta Y|male, White), \text{ by (11)} \\ &= E(\Delta Y|D = 1, White). \end{aligned}$$

Thus, we would conclude that there is misallocation which works against blacks *precisely because the DM is race-blind in his decision-making.*

Notice that this violates a key assumption we started with, viz., that the DM uses all covariates that we observe plus possibly more. Here we observe race but the DM does not take into account race in making the allocation. This works against black females because they are treated the same as white females because of their gender and the inability or unwillingness of the DM to condition on race. The scenario described above is quite stark in that we are detecting misallocation by race precisely because the DM is *not* taking race into account in making the allocation. It would thus be dramatically wrong to conclude from (**??**) that there is *prejudice* against blacks. Notice that this "mistake" is very different from and more subtle than the mistake of interpreting statistical discrimination as taste-based discrimination.

4.2 Wilful Discrimination

This is the simplest case where different thresholds are being used for the different demographic groups. In our gender example above, females are discriminated against if $\lambda_{fem} > \lambda_{male}$. Notice that such wilful discrimination has no implications for the rate of treatment in the two groups. That is, it is certainly possible that $\lambda_{fem} = \lambda_{male}$ and $\Pr(D = 1|fem = 0) > \Pr(D = 1, fem = 1)$. Conversely, it is also possible that $\lambda_{fem} > \lambda_{male}$ but $\Pr(D = 1|fem = 0) \leq \Pr(D = 1, fem = 1)$. Whether the rate of treatment is equal across demographic groups depends on the fraction of individuals within that group whose expected benefits from treatment are above the threshold for that group. So an efficient allocation, using a common threshold for all demographic groups, may be one where a larger fraction of males are treated if a larger fraction of males have expected benefits from treatment above the common threshold than females.

4.3 Implicit discrimination

Suppose the DM discriminates by race. It is possible that we will conclude there is misallocation which works against the poor. The following scenario illustrates the point. Suppose it is the case that

$$\begin{aligned} E(\Delta Y|black, rich) &> \lambda_{bl} > E(\Delta Y|black, poor) \\ &> E(\Delta Y|white, rich) > E(\Delta Y|white, poor) > \lambda_{wh}. \end{aligned}$$

Suppose the DM observes both race and wealth status and thus assigns the rich blacks and all whites to treatment. Then we have that

$$\begin{aligned} E(\Delta Y|poor, D = 0) &= E(\Delta Y|poor, black) \\ E(\Delta Y|rich, D = 1) &= E(\Delta Y|rich, black) \times \Pr(black|rich) \\ &\quad + E(\Delta Y|rich, wh) \times \Pr(wh|rich) \\ &\simeq E(\Delta Y|rich, wh) \text{ if } \Pr(wh|rich) \simeq 1. \end{aligned}$$

Since it is the case that

$$E(\Delta Y|black, poor) > E(\Delta Y|white, rich),$$

we will conclude that

$$E(\Delta Y|poor, D = 0) > E(\Delta Y|rich, D = 1),$$

i.e., that there is misallocation which works against the poor. This will happen even if the DM is not explicitly discriminating against the poor. The root is of course the high positive correlation between being white and rich. Pope and Sydnor (2008) in somewhat different contexts have discussed such "implicit profiling".

In all three cases listed above, we would potentially detect misallocation based on some covariate(s). However, this misallocation could be a result of prejudicial discrimination based on that particular covariate, inadvertent discrimination from ignoring that covariate in the allocation or implicit discrimination on a positively correlated covariate. While the exact form of discrimination cannot be pinpointed, one can conclude that there has been misallocation of treatment, leading some demographic groups to receive less and some others to receive more amounts of treatment than what economic efficiency would dictate. In our terminology, the former group has been subjected to non-statistical discrimination.

5 Broadening the model

We now extend the analysis to include risk averse behavior by the DM and transform the problem of detecting misallocation for a specific outcome to the problem of detecting the extent of risk aversion which justify the observed allocation as an efficient one.

5.1 Risk Aversion: Parametric

In this part of the analysis we ask what risk-averse utility function(s) are consistent with efficient allocation, given the data. To do this we consider a family of risk averse utility functions $u(\cdot,\theta)$, indexed by a finite dimensional parameter θ and the corresponding allocation rule which is a generalization of (**??**)

$$D=1 \text{ iff } \frac{E\left(u\left(Y_1,\theta\right)|X,Z\right)-E\left(u\left(Y_0,\theta\right)|X,Z\right)}{E\left(C_1|X,Z\right)-E\left(C_0|X,Z\right)}>\lambda. \tag{13}$$

Examples of such utility functions are CRRA $u(Y,\theta)\equiv\frac{Y^{1-\theta}}{1-\theta}$ for $\theta\in(0,1)$ and CARA $u(Y,\theta)\equiv -e^{\theta Y}$ for $\theta\geq 0$. Let $\Delta Y(\theta)\equiv u\left(Y_1,\theta\right)-u\left(Y_0,\theta\right)$.

When the DM's subjective expectations are consistent with true distributions in the population, we have that

$$\frac{E\left(u\left(Y_1,\theta\right)|X,D=1\right)-E\left(u\left(Y_0,\theta\right)|X,D=1\right)}{E\left(C_1|X,D=1\right)-E\left(C_0|X,D=1\right)}>\lambda, \text{ w.p.1.}$$

As before, we do the analysis separately for males and females to get the bounded sets in terms of θ:

$$[L_f(\theta), U_f(\theta)] = \left\{ \theta : \begin{array}{c} \sup_{x \in Supp(X_1|fem=1,D=0)} \frac{E[\Delta Y(\theta)|X_1=x,fem=1,D=0]}{E[\Delta C|X_1=x,fem=1,D=0]} \\ < \lambda \\ \leq \inf_{x \in Supp(X_1|fem=1,D=1)} \frac{E[\Delta Y(\theta)|X_1=x,fem=1,D=1]}{E[\Delta C|X_1=x,fem=1,D=1]} \end{array} \right\}$$

and similarly, $[L_m(\theta), U_m(\theta)]$.

So the values of θ which are consistent with efficient allocation within gender are the ones for which

$$L_f(\theta) \leq U_f(\theta) \text{ and } L_m(\theta) \leq U_m(\theta). \tag{14}$$

Further, the values of θ which are consistent with efficient allocation across demographic groups are the ones for which

$$\max\{L_f(\theta), L_m(\theta)\} \leq \min\{U_f(\theta), U_m(\theta)\}. \tag{15}$$

If the set of $\theta \geq 0$ for which both (14) and (15) hold turns out to be empty, then no member of the corresponding family of utility functions will justify the observed allocation as an efficient one.

5.2 Risk Aversion: nonparametric

Now consider a general differentiable Bernoulli utility function $u(\cdot)$ which will be the ingredient of a VnM utility defined over lotteries. In order for such a utility function to rationalize the observed treatment choice, we must have that for all x, x'

$$\frac{E[u(Y_1) - u(Y_0)|D=1, X=x]}{E[\Delta C|D=1, X=x]} \geq \frac{E[u(Y_1) - u(Y_0)|D=0, X=x']}{E[\Delta C|D=0, X=x']}. \tag{16}$$

Here, we focus on the case where both Y and X are discrete. The continuous case is treated as a separate subsection. So assume that Y_1 and Y_0 are discrete, with union support equal to $\{a_1 ... a_m\}$.The above condition reduces to: for all x, x':

$$\begin{aligned} & \sum_{j=1}^{m} u(a_j) \underbrace{\left\{ \frac{\Pr(Y_1 = a_j|x, D=1) - \Pr(Y_0 = a_j|x, D=1)}{E[C_1 - C_0|D=1, X=x]} \right\}}_{=h_1(a_j,x), \text{ say}} \\ \geq & \sum_{j=1}^{m} u(a_j) \underbrace{\left\{ \frac{\Pr(Y_1 = a_j|x, , D=0) - \Pr(Y_0 = a_j|x', D=0)}{E[C_1 - C_0|D=0, X=x,]} \right\}}_{h_0(a_j,x'), \text{ say}}. \end{aligned}$$

Letting $u(a_j) = u_j$ and $q_k(x, x') = h_1(a_k, x) - h_0(a_k, x')$, the previous display reduces to a set of linear restrictions

$$\begin{aligned} u_1 &= 0,\ u_m = 1 \text{ (affine normalization)}, \\ u_{k+1} &\geq u_k,\ k = 1, ...m-1 \text{ (monotonic)}, \\ \frac{u_{k+1} - u_k}{a_{k+1} - a_k} &\geq \frac{u_{k+2} - u_{k+1}}{a_{k+2} - a_{k+1}},\ k = 1, ...m-2 \text{ (concave)}, \\ \sum_{k=1}^{m} u_k q_k(x, x') &\geq 0 \text{ for all } x, x'. \end{aligned} \tag{17}$$

When X is also discrete, the above inequalities define a finite-dimensional polyhedron. There exist algorithms for finding extreme points of a polyhedron defined through inequality constraints. The identified set of u_k's are the convex hull of those extreme points and one can base a test of DM rationality on whether the identified set of u's is empty.

5.2.1 Equivalent conditions:

At this point, it is meaningful to ask the following question. Suppose we find that for $u(y) = y$, i.e., allocations based on expected gains, the corresponding set of γ's is empty– suggesting misallocation. Then under what conditions shall we always (never) find a nondecreasing concave utility function under which the observed allocations will be efficient under the utility function? In other words, is *every* observed allocation justifiable as an efficient one for *some* choice of $u(\cdot)$? The following proposition provides the answer in the case where Y takes on finite positive values.

Suppose w.l.o.g. Y takes values in the finite set $0 = a_1 \leq a_2 \leq ... \leq a_m = 1$. For two subgroups 1 and 2, let

$$\mu_{lbc}(j) = \frac{\Pr(Y_l = a_j | D = b, G = c)}{E(\Delta C | D = b, G = c)},$$

for $j = 1, ...m$, $l = 0, 1$, $b = 0, 1$ and $c = 1, 2$. Suppose that we have detected inefficiency whereby group 2 is being under-treated, viz.,

$$\begin{aligned} \frac{E(\Delta Y | D = 1, G = 1)}{E(\Delta C | D = 1, G = 1)} &< \frac{E(\Delta Y | D = 0, G = 2)}{E(C_1 - C_0 | D = 0, G = 2)}, \text{ i.e.,} \\ \sum_{j=1}^{m} a_j \left[\mu_{111}(j) - \mu_{011}(j) - \mu_{102}(j) + \mu_{002}(j)\right] &< 0. \end{aligned} \tag{18}$$

Let

$$r_j = \underbrace{\mu_{111}(j) - \mu_{011}(j)}_{=r_{1j}} - \underbrace{(\mu_{102}(j) - \mu_{002}(j))}_{=r_{2j}},$$

and observe that by definition, $\sum_{j=1}^{m} r_j = 0$ and $\sum_{j=1}^{m} a_j r_j < 0$. The question is: can we necessarily find $u(\cdot)$ nondecreasing and concave, such that

$$\begin{aligned} \frac{E(u(Y_1) - u(Y_0)|D=1, G=1)}{E(\Delta C|D=1, G=1)} &\geq \frac{E(u(Y_1) - u(Y_0)|D=0, G=2)}{E(\Delta C|D=0, G=2)}, \text{ i.e.,} \\ \sum_{j=1}^{m} r_j u(a_j) &\geq 0. \end{aligned} \tag{19}$$

The following proposition provides a characterization.

Define $R_k = \sum_{j=1}^{k} r_j$, $S_l = \sum_{k=1}^{l-1} R_k (a_{k+1} - a_k)$, for $l = 2,...m$. Note that

$$S_m = \sum_{k=1}^{m-1} R_k (a_{k+1} - a_k) = -\sum_{j=1}^{m} r_j a_j > 0.$$

Proposition 1 *Suppose* $\{r_j\}$ *is such that* $\sum_{j=1}^{m} r_j = 0$ *and* $\sum_{j=1}^{m} r_j a_j < 0$. *The following conditions are equivalent:*

(i) $S_l \geq 0$, *for every* $l = 1, ..., m-1$.

(ii) there does not exist any nondecreasing and concave $u(\cdot)$, *such that (19) holds.*

Condition (i) can be checked directly before we try to find the set of solutions. Note that this proposition is of a similar flavour to the equivalence of second order stochastic dominance and dominance in terms of every concave and monotone sub-utility function, but applicable to the case where the r_j's are more complicated than just probabilities and the support points are not equally spaced.

Proof. See appendix. ■

5.2.2 Maximum entropy solution

The methodology outlined above (c.f., (17)) gives a whole set of utility functions which may be difficult to report because it will generically be an infinite set. We therefore consider a variant of the problem where, instead of trying to find the entire set of admissible utilities, we find the one among them which is closest to a specific utility function, such as

the risk neutral one $u(y) = y$ or a specific risk-averse one, e.g., $u(y) = \sqrt{y}$. This objective can be achieved through the use of entropy maximization, which we describe now.

Recall the constraints (17). Define $v_1 = u_1 = 0$ and $v_k = u_k - u_{k-1}$ for $k = 2, ...K$. In matrix notation,

$$\underbrace{\begin{bmatrix} v_1 \\ v_2 \\ ... \\ v_k \end{bmatrix}}_{v} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ -1 & 1 & 0 & ... & 0 \\ 0 & -1 & 1 & 0... & 0 \\ . & . & . & . & . \\ 0 & ... & 0 & -1 & 1 \end{bmatrix}}_{S} \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_k \end{bmatrix},$$

where S is nonsingular. Also, for fixed x, x', let $q(x, x')$ denote the k-vector whose kth entry is $q_k(x, x')$. Then the constraints (17) can be rewritten as

$$\begin{aligned} v_k &\geq 0,\ k = 1, ...K-1, \\ \sum_{k=1}^{k} v_k &= 1, \\ \frac{v_k}{a_k - a_{k-1}} &\geq \frac{v_{k+1}}{a_{k+1} - a_k},\ k = 1, ...K-1 \\ v' \left[S^{-1} q(x, x')\right] &\geq 0 \text{ for all } x, x'. \end{aligned} \tag{20}$$

Given the form of the constraints, one can apply the principle of maximum entropy and solve

$$\max \left\{ -\sum_{k=1}^{K} \left(\frac{v_k}{a_k - a_{k-1}} \right) \ln \left(\frac{v_k}{a_k - a_{k-1}} \right) \right\}, \text{ s.t. (20).}$$

If there were no q-constraints, then the solution would be $v_k = a_k - a_{k-1}$. This corresponds to the risk-neutral situation $u(a) = a$. Therefore maximizing the entropy s.t. the constraints corresponds to finding the most risk-neutral $u(\cdot)$ which satisfies the constraints. Standard software can be used to perform these calculations since the problem is strictly concave. Once the v's are obtained, one can find the corresponding u's by using $u = S^{-1}v$.

To get the utility function closest to $u(y) = \sqrt{y}$, one would solve

$$\max \left\{ -\sum_{k=1}^{K} \frac{v_k}{\sqrt{a_k} - \sqrt{a_{k-1}}} \ln \left(\frac{v_k}{\sqrt{a_k} - \sqrt{a_{k-1}}} \right) \right\}, \text{ s.t. (20).}$$

In the absence of the q-constraints, the solution would be $v_k = \sqrt{a_k} - \sqrt{a_{k-1}}$, i.e. $u(y) = \sqrt{y}$, as desired.

In contrast to the set-identified situation, the maximum entropy problem will either have no solution (if the constraint set is empty, for instance) or a unique solution, which would make it easy to report. This unique solution will have a meaningful interpretation as the admissible utility function closest to a specific utility function (e.g., a risk-neutral one). Moreover, when the q's are estimated, one can, in principle, construct confidence intervals for both the solution and the value function for the above problems, using the distribution theory for the estimated q's.

5.2.3 Inference

Testing whether the existing allocation is efficient for a given utility function, reduces essentially to testing a set of (conditional) moment inequalities (c.f., (8) or (9) above). There is an existing and expanding literature in econometrics, dealing with such tests. For example, one can adopt the method of Andrews and Soares (2009) to conduct such tests and calculate confidence intervals for the difference in treatment thresholds between demographic groups. This corresponds to inference on the true parameters, rather than inference on the identified set.

Inferring utility parameters consistent with efficient allocation is an estimation problem where the parameters of interest are defined via conditional moment inequalities. The test of rationality thereof is analogous to specification testing in GMM problems but now with inequality constraints. For the parametric case or the nonparametric case with discrete outcome and covariates, the utility parameters are finite-dimensional and we are interested in inferring the entire feasible *set* of utility parameters. So inference can be conducted using, e.g., CHT (2005). Tests of rationality again amount to checking emptiness of confidence sets, which can be done using Andrews and Soares.

Inference for the maximum entropy solution, to our knowledge, is nonstandard. Essentially, the inference problem is to find the distribution theory for the solution and value

function for the problem

$$\begin{aligned} & \max\left\{-\sum_{k=2}^{K} v_k \ln(v_k)\right\}, \text{ s.t.} \\ v_k &\geq 0,\ k=2,...K,\ \sum_{k=1}^{K} v_k = 1, \\ \frac{v_k}{a_k - a_{k-1}} &\geq \frac{v_{k+1}}{a_{k+1} - a_k},\ k=2,...K-1 \\ v'\left[S^{-1}\hat{q}(x,x')\right] &\geq 0 \text{ for all } x, x'. \end{aligned} \tag{21}$$

The problem (21) is simply (20) with q replaced by its estimate.

We now outline a method of solving (21) via a penalty method and conduct inference on the solution and value function thereof. To do this, focus on the case where X is also discrete and rewrite the last set of inequalities in the previous display as $\sum_{k=1}^{K} \hat{g}_{jk} v_k \leq 0$ for $j = 1, ...J$. Now consider the problem

$$\begin{aligned} & \min_{\{v_k\}}\left\{\sum_{k=1}^{K} v_k \ln(v_k) + r_n \sum_{j=1}^{J}\left(\max\left\{0, \sum_{k=1}^{K} \hat{g}_{jk} v_k\right\}\right)^2\right\}, \text{ s.t.} \\ v_k &\geq 0,\ k=2,...K-1,\ \sum_{k=1}^{K} v_k = 1, \\ \frac{v_k}{a_k - a_{k-1}} &\geq \frac{v_{k+1}}{a_{k+1} - a_k},\ k=2,...K-1. \end{aligned} \tag{22}$$

This is a penalized version of (21), where the penalty term

$$r_n \sum_{j=1}^{J}\left(\max\left\{0, \sum_{k=1}^{K} \hat{g}_{jk} v_k\right\}\right)^2$$

is positive whenever for one or more j, the term(s) $\sum_{k=1}^{K} \hat{g}_{jk} v_k$ are positive. Here r_n will be a sequence of positive numbers increasing with n, so that the penalty for constraint violation is higher when the sample size is large. Note that (22) is an optimization problem with convex objective function and linear constraints. So finding a solution is easy. The remaining question is: how to choose the sequence r_n to gurantee desirable properties of the resulting solution and value– in particular, consistency and a tractable asymptotic distribution.

Define

$$\begin{aligned}
\hat{Q}_n(v) &= \left\{ \sum_{k=1}^{K} v_k \ln(v_k) + r_n \sum_{j=1}^{J} \left(\max\left\{ 0, \sum_{k=1}^{K} \hat{g}_{jk} v_k \right\} \right)^2 \right\}, \\
Q(v) &= \sum_{k=1}^{K} v_k \ln(v_k), \\
A &= \left\{ \begin{array}{c} v : v_k \geq 0,\ k = 2, ...K,\ \sum_{k=1}^{K} v_k = 1, \\ \frac{v_k}{a_k - a_{k-1}} \geq \frac{v_{k+1}}{a_{k+1} - a_k},\ k = 2, ...K-1. \end{array} \right\}, \\
B &= \left\{ v : \sum_{k=1}^{K} g_{jk} v_k \leq 0,\ j = 1, ...J \right\}.
\end{aligned}$$

We use the standard convention that $v_k \ln(v_k) = 0$ if $v_k = 0$.

Proposition 2 (Consistency) *Assume that* $\sqrt{n} vec(\hat{g} - g) \rightsquigarrow N(0, \Sigma)$. *Choose* r_n *such that* $r_n = o_p(\sqrt{n})$ *and* $r_n \to \infty$ *as* $n \to \infty$. *Then*

$$p \lim_{n \to \infty} \left\{ \arg\min_{v \in A} \hat{Q}_n(v) \right\} = \arg\min_{v \in A \cap B} Q(v).$$

Proof. See appendix. ■

Corollary 1 (Consistency of value function) *Under the same conditions, as the previous proposition,*

$$p \lim_{n \to \infty} \left\{ \hat{Q}_n \left(\arg\min_{v \in A} \hat{Q}_n(v) \right) \right\} = Q \left(\arg\min_{v \in A \cap B} Q(v) \right).$$

Proof. See appendix. ■

5.3 Continuous case

Inequalities: When Y is continuously distributed with support $[0, 1]$, the condition (16) reduces to

$$\begin{aligned}
& \int u(y) \underbrace{\left\{ \frac{p_{Y_1}(y|x, D=1) - p_{Y_0}(y|x, D=1)}{E[C_1 - C_0 | D=1, X=x]} \right\}}_{=h_1(y,x), \text{ say}} dy \\
\geq & \int u(y) \underbrace{\left\{ \frac{p_{Y_1}(y|x', D=0) - p_{Y_0}(y|x', D=0)}{E[\Delta C | D=0, X=x']} \right\}}_{h_0(y,x'), \text{ say}} dy.
\end{aligned}$$

So the question is: does there exist a function $u(.)$ s.t.

$$\begin{aligned} u(0) &= 0,\ u(1) = 1 \text{ (affine normalization)} \\ u'(\cdot) &\geq 0 \text{ (monotonic)} \\ u''(\cdot) &\leq 0 \text{ (concave)} \\ \int u(y) h_1(y,x)\, dy &\geq \int u(y) h_0(y,x')\, dy \text{ for all } x, x'. \end{aligned}$$

Maximum Entropy: Define $v(y) = u'(y)$ and

$$H(a; x, x') \equiv \int_0^a [h_1(y,x) - h_0(y,x')]\, dy.$$

Then the constraints (16) reduce to

$$\begin{aligned} v(y) &\geq 0,\ \int_0^1 v(y)\, dy = 1 \\ -v'(y) &\geq 0 \text{ for all } y \in [0,1] \\ -\int v(y) H(y; x, x')\, dy &\geq 0 \text{ for all } x, x'. \end{aligned} \tag{23}$$

The last inequality follows by applying integration by parts to $\int u(y) [h_1(y,x) - h_0(y,x')]\, dy$ and recognizing that for any x, it follows from definition of the h functions that $\int_0^1 h_1(y,x)\, dy = 0$. The maximum entropy solution is then given by

$$\max -\int_0^1 v(y) \ln\left(\frac{v(y)}{v_0(y)}\right) dy, \text{ s.t. } (23),$$

where $v_0(y)$ corresponds to a reference utility function. For example, when $v_0(y) = 1$ we find the most risk-neutral utility function satisfying the constraints whereas $v_0(y) = 1/y$ corresponds to finding the utility function closest to the risk-averse sub-utility $u(y) = \ln(y)$ which satisfies the constraints.

Inference: Consider the case where Y is continuous and recall the restrictions:

$$\begin{aligned} u(0) &= 0,\ u(1) = 1 \text{ (affine normalization)} \\ u'(\cdot) &\geq 0 \text{ (monotonic)} \\ u''(\cdot) &\leq 0 \text{ (concave)} \\ \int u(y) h_1(y,x)\, dy &\geq \int u(y) h_0(y,x')\, dy \text{ for all } x, x'. \end{aligned}$$

Suppose X is discrete and takes values $a_1, ...a_K$. Let $h_{jk}(y) \equiv h_1(y, a_j) - h_0(y, a_k)$. Define the criterion function as

$$Q(u) = \sum_j \sum_k \left(\min \left\{ 0, \int u(y) h_{jk}(y) dy \right\} \right)^2$$

and its estimated analog (with $h_{jk}(\cdot)$ replaced by their estimates) as $Q_n(u)$. At this point, it is necessary to approximate the functions $u(\cdot)$ via a basis (a sieve) and imposing the restrictions implied by the structure of utility functions and the efficiency requirement.

Typically, we would consider a sieve basis $\{p_1, ..., p_{J_n}\}$ and consider approximating $u(\cdot)$ by the sum $\sum_{l=1}^{J_n} \beta_l p_l(\cdot)$ and choosing the coefficients such that the monotonicity and concavity are satisfied. One convenient choice of basis are cardinal B-splines (see, Chen (2005)) for which monotonicity and concavity are equivalent to the coefficients $\{\beta_l\}$, $l = 1, 2, ...$, satisfying simple linear inequalities– say, $A_{(J_n)}\beta \geq 0$, for an appropriate matrix $A_{(J_n)}$. Then the identified set for $u(\cdot)$ can be approximated by the identified set for the β-coefficients. The latter can be obtained by using the criterion function

$$\tilde{Q}_n(\beta) = \sum_j \sum_k \left(\min \left\{ 0, \int \left\{ \sum_{l=1}^{J_n} \beta_l p_l(y) \right\} h_{jk}(y) dy \right\} \right)^2$$

and its estimated analog (with $h_{jk}(\cdot)$ replaced by their estimates) denoted by $\breve{Q}_n(\beta)$. The CI for the identified set of approximating β's is then given by

$$\hat{C}_n = \left\{ \beta : A_{(J_n)}\beta \geq 0;\ \breve{Q}_n(\beta) \leq c_\alpha \right\},$$

for an appropriately chosen c_α. A detailed analysis is left to future research.

The maximum entropy solution corresponding to the sample values of $\hat{q}(x, x')$ will be random simply because the are estimated. Finding the asymptotic distribution of the resulting $\hat{v}$'s would require us to first decide on which constraints are (approximately) binding. One approach is to first calculate the sample solution for the v's by solving the optimization problem with the estimated value of $\hat{q}(x, x')$. Then drop those constraints for which the the constraint function evaluated at the sample estimates of the v's exceeds τ_n–a decreasing function of the sample size n (such as $\ln(n)^{-1/2}$). These are the estimated nonbinding constraints with zero Lagrange multipliers and these will not appear in the final solution. Then use the remaining constraints to solve the optimization problem explicitly, resulting in a set of first order equations involving (possibly nonlinearly) the

v's, the Lagrange multipliers for the binding constraints and the $\hat{q}$'s. Finding the asymptotic distribution of the resulting estimated v's will simply be an application of the delta method. The only nontrivial issue involves figuring out how the first stage decision on which constraints are binding will impact the asymptotic distribution of the estimated v's. It is reasonable to expect that when the first-stage cut-off is chosen to be of higher order than $n^{-1/2}$, the first stage decision has no impact on the asymptotic distribution, i.e., it is as if we knew which constraints bind.

6 A simulation exercise

We report simulation results for the following linear regression model linking outcome Y with regressors Z, f and the treatment indicator D as follows.

$$\begin{aligned} Y &= 1 + 0.2Z + \beta \times DZ + 0.4f + 0.5Df \\ E(\Delta Y|Z, f) &= \beta \times Z + 0.5 \times f. \end{aligned}$$

We generate (Z, f^*) from a bivariate normal with mean zero, correlation 0.5 and variances $(\sigma_z^2, 1)$. The coefficient β is chosen to be positive. The variable f is a dummy for female and is generated as the indicator for $f^* > 0$. We generate $2N$ observations this way and randomly divide them into an observational and an experimental group.

Within the experimental group, the binary treatment D was generated randomly. Corresponding to a realization of (D, Z, f), the corresponding Y was generated according to the model.

In the observational group, the DM was assumed to calculate $E(Y|Z, f, D = 1)$ and $E(Y|Z, f, D = 0)$ using the actual model coefficients and then assign the males $(f = 0)$ to treatment if the difference exceeds $\gamma_{male} = 0$ and assign the females $(f = 1)$ to treatment if the difference exceeds $\gamma_{fem} = 0.5$. Given the model, this means that for both males and females, $D = 1$ iff $Z > 0$.

In both the experimental and observational samples, the econometrician observes f, Y but not Z. Additionally, the econometrician observes a noisy signal of Z. This is given by the variable $X = 1 + 1(\rho Z + \epsilon > 0)$ with $\varepsilon \tilde{} N(0, 1)$. The DM's rule implies the following

expressions.

$$\begin{aligned}
E\left(\Delta Y|D=1,f=1,X=x\right) &= \beta\times E\left\{Z|Z>0,f=1,X=x\right\}+0.5,\\
E\left(\Delta Y|D=0,f=1,X=x\right) &= \beta\times E\left\{Z|Z<0,f=1,X=x\right\}+0.5,\\
E\left(\Delta Y|D=1,f=0,X=x\right) &= \beta\times E\left\{Z|Z>0,f=0,X=x\right\},\\
E\left(\Delta Y|D=0,f=0,X=x\right) &= \beta\times E\left\{Z|Z<0,f=0,X=x\right\}.
\end{aligned}$$

The true bounds are then given by

$$\begin{aligned}
&\max_{x\in\{1,2\}}\beta\times E\left\{Z|Z<0,f=1,X=x\right\}+0.5\\
\leq\ &\gamma_f\\
\leq\ &\min_{x\in\{1,2\}}\beta\times E\left\{Z|Z>0,f=1,X=x\right\}+0.5,
\end{aligned}$$

and

$$\begin{aligned}
&\max_{x\in\{1,2\}}\beta\times E\left\{Z|Z<0,f=0,X=x\right\}\\
\leq\ &\gamma_m\\
\leq\ &\min_{x\in\{1,2\}}\beta\times E\left\{Z|Z>0,f=0,X=x\right\},
\end{aligned}$$

and these can be easily simulated.

The values of ρ, β and variance of Z were varied in the experiment. A larger absolute value of ρ indicates that observing X rather than Z is less of a handicap and so this should lead to narrower bounds on the thresholds. A larger σ_z^2 implies wider support for Z, which, ceteris paribus, will widen the bounds because it would lead to a larger difference between the minimum (or maximum) over the support of Z and the mean over the distribution of Z. Bounds will be narrower when β is close to zero. The intuition is that when β is close to zero, the omitted variable Z plays a smaller role in treatment assignment and not knowing Z is less of a handicap for knowing γ_f and γ_m.

For each sample size and each choice of ρ, β, σ_z^2, we ran 100 replications of the experiment. In each replication, we calculated sample-based bounds on γ_{male} and γ_{fem} and constructed confidence intervals for the difference $\gamma_{fem}-\gamma_{male}$, using the steps outlined in section 3. We report two sets of bounds–one that is simply the sample analog of the population inequalities and the other is a weighted average of bounds across values of X where weight equals the inverse of the square-root of $p*(1-p)$ and p is the probability

in the observational sample that $D = 1$ for that value of X. The results are shown in table 1.

Second, we calculated bounds on the risk aversion parameter corresponding to the parametric case. Third, we calculated bounds on the (monotone and concave) utility functions corresponding to the nonparametric case. One would expect the test to reject rationality more often and the bounds to be narrower when the sample size is large.

7 Conclusion

Alternatives

Get DM's decision on applications but randomize–cleanest

Randomized allocation followed by DM's allocation– e.g. students into classes

TBA...

References

[1] Andrews and Soares

[2] Bertrand, Marianne and Mullainathan, Sendhil (2002). ""Are Emily and Jane More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination,"". American Economic Review 94: 991.

[3] Bhattacharya & Dupas (2010): Inferring Efficient Treatment Assignment under Budget Constraints", NBER working paper number

[4] Chandra and Staiger (2009): Identifying Provider Prejudice in Healthcare," manuscript, March 2008, downloadable from: http://www.hks.harvard.edu/fs/achandr/Provider%20Prejudice%204March%202008.pdf

[5] Chernozhukov, Hong and Tamer

[6] Elliott, Komunjer and Timmerman (2005): Estimation and Testing of Forecast Rationality under Flexible Loss. Review of Economic Studies, 72, pp. 1107-1125.

[7] Goldin, Claudia and Rouse, Cecilia (2000): "Orchestrating Impartiality", The American Economic Review, 90(4), pp. 715 - 41.

[8] Heckman, J. (1998): Journal of Economic Perspectives-Volume 12, Number 2, Pages 101-116.

[9] Knowles, Persico and Todd (2001): Racial Bias in Motor Vehicle Searches: Theory and Evidence, Journal of Political Economy, 2001 V109 (11) pages 203-229.

[10] Manski, C. F. (1990): Nonparametric Bounds on Treatment Effects, The American Economic Review, Vol. 80, No. 2, pp. 319-323.

[11] Patton, Andrew J. & Timmermann, Allan (2007): "Testing Forecast Optimality Under Unknown Loss," Journal of the American Statistical Association, vol. 102, pages 1172-1184.

[12] Persico, N (2009): "Racial Profiling? Detecting Bias Using Statistical Evidence", Annual Review of Economics, volume 1.

8 Appendix: Proofs

Derivation of (1): The solution to the problem

$$\max_{A}\left\{\int 1\left\{w\in A\right\}y_1 dP^{sub}_{Y_1,W}(y_1,w)+\int 1\left\{w\in A^c\right\}y_0 dP^{sub}_{Y_0,W}(y_0,w)\right\}$$

s.t.

$$\int 1\left\{w\in A\right\}c_1 dP^{sub}_{C_1,W}\left(c_1,w\right)+\int 1\left\{w\in A^c\right\}c_0 dP^{sub}_{C_0,W}\left(c_0,w\right)\leq c,$$

is of the form $A^*=\{w:\beta\left(w\right)<\gamma\}$, with

$$\beta\left(w\right)\equiv\frac{E^{sub}\left(\Delta C|W=w\right)}{E^{sub}\left(\Delta Y|W=w\right)};\ c=\int_{w\in\boldsymbol{w}}1\left(\beta\left(w\right)<\gamma\right)dP^{sub}_{W}\left(w\right).$$

Proof. The welfare resulting from a generic choice of A, satisfying the budget con-

straint, differs from the welfare from using A^* by an amount given by

$$\begin{aligned}
& G(A) - G(A^*) \\
= & \int 1\{w \in A\} y_1 dP^{sub}_{Y_1,W}(y_1,w) + \int 1\{w \in A^c\} y_0 dP^{sub}_{Y_0,W}(y_0,w) \\
& - \left\{ \int 1\{w \in A^*\} y_1 dP^{sub}_{Y_1,W}(y_1,w) + \int 1\{w \in A^{*c}\} y_0 dP^{sub}_{Y_0,W}(y_0,w) \right\} \\
= & \int \begin{bmatrix} 1\{w \in A\} \times 1\{w \in A^{*c}\} \\ -1\{w \in A^*\} \times 1\{w \in A^c\} \end{bmatrix} y_1 dP^{sub}_{Y_1,W}(y_1,w) \\
& + \int \begin{bmatrix} 1\{w \in A^c\} \times 1\{w \in A^*\} \\ -1\{w \in A\} \times 1\{w \in A^{*c}\} \end{bmatrix} y_0 dP^{sub}_{Y_0,W}(y_0,w). \\
= & \int 1\{w \in A\} \times 1\{w \in A^{*c}\} \times \left[\int y_1 dP^{sub}_{Y_1|W}(y_1|w) - \int y_0 dP^{sub}_{Y_0|W}(y_0|w) \right] dF(w) \\
& - \int 1\{w \in A^c\} \times 1\{w \in A^*\} \times \left[\int y_1 dP^{sub}_{Y_1|W}(y_1|w) - \int y_0 dP^{sub}_{Y_0|W}(y_0|w) \right] dF(w).
\end{aligned}$$

Now, note that $w \in A^{*c}$ implies that $\frac{E^{sub}(\Delta C|W=w)}{\gamma} \geq E^{sub}(\Delta Y|W=w)$, and $w \in A^*$ implies that $\frac{E^{sub}(\Delta C|W=w)}{\gamma} \leq E^{sub}(\Delta Y|W=w)$. Consequently, the previous display

$$\begin{aligned}
\leq & \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A\} \times 1\{w \in A^{*c}\} dF(w) \\
& - \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A^c\} \times 1\{w \in A^*\} dF(w) \\
= & \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A\} dF(w) \\
& - \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A\} \times 1\{w \in A^*\} dF(w) \\
& - \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A^*\} dF(w) \\
& + \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A\} \times 1\{w \in A^*\} dF(w) \\
= & \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A\} dF(w) \\
& - \frac{1}{\gamma} \int E^{sub}(\Delta C|W=w) \times 1\{w \in A^*\} dF(w) = \frac{c}{\gamma} - \frac{c}{\gamma} = 0.
\end{aligned}$$

The last but one step follows from the fact that both A and A^* must satisfy the budget constraint. Since $G(A) \leq G(A^*)$, and A is any set satisfying the budget constraint, it follows that A^* must be the optimal one. ■

Proposition 1:

Proof. (i) implies (ii). Notice that

$$\begin{aligned}
-\sum_{j=1}^{m} r_j u(a_j) &= \sum_{j=1}^{m-1} r_j (u(a_m) - u(a_j)) = \sum_{j=1}^{m-1} r_j \sum_{k=j}^{m-1} (u(a_{k+1}) - u(a_k)) \\
&= \sum_{k=1}^{m-1} (u(a_{k+1}) - u(a_k)) R_k \\
&= \sum_{k=1}^{m-1} \frac{u(a_{k+1}) - u(a_k)}{a_{k+1} - a_k} \times R_k (a_{k+1} - a_k) \\
&= \sum_{k=1}^{m-1} \left(\sum_{l=k+1}^{m-1} \left\{ \frac{u(a_l) - u(a_{l-1})}{a_l - a_{l-1}} - \frac{u(a_{l+1}) - u(a_l)}{a_{l+1} - a_l} \right\} \right) \times R_k (a_{k+1} - a_k) \\
&\quad + \frac{u(a_m) - u(a_{m-1})}{a_m - a_{m-1}} \sum_{k=1}^{m-1} R_k (a_{k+1} - a_k) \\
&= \sum_{l=2}^{m-1} \left\{ \frac{u(a_l) - u(a_{l-1})}{a_l - a_{l-1}} - \frac{u(a_{l+1}) - u(a_l)}{a_{l+1} - a_l} \right\} \times \sum_{k=1}^{l-1} R_k (a_{k+1} - a_k) \\
&\quad + \frac{u(a_m) - u(a_{m-1})}{a_m - a_{m-1}} \sum_{k=1}^{m-1} R_k (a_{k+1} - a_k) \\
&= \sum_{l=2}^{m-1} \left\{ \frac{u(a_l) - u(a_{l-1})}{a_l - a_{l-1}} - \frac{u(a_{l+1}) - u(a_l)}{a_{l+1} - a_l} \right\} \times S_l + \frac{u(a_m) - u(a_{m-1})}{a_m - a_{m-1}} S_m
\end{aligned}$$

By concavity of $u(\cdot)$, we have that

$$u(a_l) \geq \frac{a_{l+1} - a_l}{a_{l+1} - a_{l-1}} u(a_{l-1}) + \frac{a_l - a_{l-1}}{a_{l+1} - a_{l-1}} u(a_{l+1}),$$

whence it follows that for every l:

$$\frac{u(a_l) - u(a_{l-1})}{a_l - a_{l-1}} > \frac{u(a_{l+1}) - u(a_l)}{a_{l+1} - a_l}.$$

This plus $S_l \geq 0$, for every $l = 1, ..., m-1$, implies that $\sum_{j=1}^{m} r_j u(a_j) \leq 0$ for every concave and nondecreasing $u(\cdot)$.

(ii) implies (i). Suppose $S_k < 0$ for some $k \in \{2, ...m-1\}$. We will show that there exists a nondecreasing concave $u(\cdot)$ such that $-\sum_{j=1}^{m} r_j u(a_j) \leq 0$. Recall that

$$\begin{aligned}
-\sum_{j=1}^{m} r_j u(a_j) &= \sum_{l=2}^{m-1} \left\{ \frac{u(a_l) - u(a_{l-1})}{a_l - a_{l-1}} - \frac{u(a_{l+1}) - u(a_l)}{a_{l+1} - a_l} \right\} \times S_l \\
&\quad + \frac{u(a_m) - u(a_{m-1})}{a_m - a_{m-1}} S_m
\end{aligned}$$

Consider a utility function of the form

$$u(a) = \frac{a}{a_k} \times 1(a \leq a_k) + 1 \times 1(a \geq a_k).$$

It is obvious that this is a nondecreasing concave continuous function. Now, for this utility function,

$$\begin{aligned} \frac{u(a_m) - u(a_{m-1})}{a_m - a_{m-1}} &= 0, \\ \frac{u(a_l) - u(a_{l-1})}{a_l - a_{l-1}} - \frac{u(a_{l+1}) - u(a_l)}{a_{l+1} - a_l} &= \frac{1}{a_k} \times 1(l = k), \end{aligned}$$

implying that $-\sum_{j=1}^{m} r_j u(a_j) = S_k/a_k < 0$. ■

Proposition 2: Assume that $\sqrt{n}vec(\hat{g} - g) \rightsquigarrow N(0, \Sigma)$. Choose r_n such that $r_n = o_p(\sqrt{n})$ and $r_n \to \infty$ as $n \to \infty$. Then

$$p\lim_{n\to\infty} \left\{ \arg\min_{v \in A} \hat{Q}_n(v) \right\} = \arg\min_{v \in A \cap B} Q(v).$$

Proof. Let

$$\begin{aligned} Q_n(v) &= \left\{ \sum_{k=1}^{K} v_k \ln(v_k) + r_n \sum_{j=1}^{J} \left(\max\left\{ 0, \sum_{k=1}^{K} g_{jk} v_k \right\} \right)^2 \right\}, \\ v^{(n)} &= \arg\min_{v \in A} \hat{Q}_n(v). \end{aligned}$$

If $v^{(n)}$ is non-unique, we choose any of those values (c.f., Amemiya (1985), page 103)).

Fix $\varepsilon > 0$ and assume that for at least one j, we have $\sum_{k=1}^{K} g_{jk} v_k^{(n)} > \varepsilon$. Then,

$$\begin{aligned} \hat{Q}_n(v_n) &= \sum_{k=1}^{K} v_k^{(n)} \ln\left(v_k^{(n)}\right) + r_n \sum_{j=1}^{J} \left(\max\left\{ 0, \sum_{k=1}^{K} \hat{g}_{jk} v_k^{(n)} \right\} \right)^2 \\ &= \sum_{k=1}^{K} v_k^{(n)} \ln\left(v_k^{(n)}\right) + r_n \sum_{j=1}^{J} \left(\max\left\{ 0, \sum_{k=1}^{K} (\hat{g}_{jk} - g_{jk}) v_k^{(n)} + \sum_{k=1}^{K} g_{jk} v_k^{(n)} \right\} \right)^2 \\ &= \sum_{k=1}^{K} v_k^{(n)} \ln\left(v_k^{(n)}\right) + \sum_{j=1}^{J} \left(\max\left\{ 0, \frac{r_n}{\sqrt{n}} \sum_{k=1}^{K} \left\{ \sqrt{n}(\hat{g}_{jk} - g_{jk}) \right\} v_k^{(n)} + r_n \sum_{k=1}^{K} g_{jk} v_k^{(n)} \right\} \right)^2 \\ &> \sum_{k=1}^{K} v_k^{(n)} \ln\left(v_k^{(n)}\right) + \sum_{j=1}^{J} \left(\max\left\{ 0, \frac{r_n}{\sqrt{n}} \sum_{k=1}^{K} \left\{ \sqrt{n}(\hat{g}_{jk} - g_{jk}) \right\} v_k^{(n)} + r_n \varepsilon \right\} \right)^2 \\ &= \sum_{k=1}^{K} v_k^{(n)} \ln\left(v_k^{(n)}\right) + J r_n^2 \varepsilon^2 + o_p(1), \text{ by hypothesis.} \end{aligned}$$

Thus, by choosing n (and thus r_n) large enough, $\hat{Q}_n\left(v^{(n)}\right)$ can be made arbitrarily large, but for any $v \in A \cap B$, $\hat{Q}_n(v) = \sum_{k=1}^{K} v_k \ln(v_k)$ remains finite. This contradicts that $v^{(n)} = \arg\min_{v \in A} \hat{Q}_n(v)$. Since ε is arbitrary, it must be that $\Pr\left(v^{(n)} \in B\right) \to 1$. Therefore, for any $\delta > 0$,

$$\Pr\left(\left\|\arg\min_{v \in A \cap B} \hat{Q}_n(v) - \arg\min_{v \in A} \hat{Q}_n(v)\right\| > \delta\right) \leq \Pr\left(v^{(n)} \notin B\right) \to 0,$$

implying

$$\left\|\arg\min_{v \in A \cap B} \hat{Q}_n(v) - \arg\min_{v \in A} \hat{Q}_n(v)\right\| = o_p(1). \tag{24}$$

Next, note that for any two real numbers a, b:

$$|\max\{0, a\} - \max\{0, b\}| \leq |a - b|. \tag{25}$$

Therefore,

$$\begin{aligned}
& \left|\hat{Q}_n(v) - Q_n(v)\right| \\
= \; & r_n \left|\sum_{j=1}^{J} \left(\max\left\{0, \sum_{k=1}^{K} \hat{g}_{jk} v_k\right\}\right)^2 - \sum_{j=1}^{J} \left(\max\left\{0, \sum_{k=1}^{K} g_{jk} v_k\right\}\right)^2\right| \\
\leq \; & r_n \sum_{j=1}^{J} \left|\left(\max\left\{0, \sum_{k=1}^{K} \hat{g}_{jk} v_k\right\}\right)^2 - \left(\max\left\{0, \sum_{k=1}^{K} g_{jk} v_k\right\}\right)^2\right| \\
= \; & r_n \sum_{j=1}^{J} \left\{ \begin{array}{c} \left|\left(\max\left\{0, \sum_{k=1}^{K} \hat{g}_{jk} v_k\right\}\right) - \left(\max\left\{0, \sum_{k=1}^{K} g_{jk} v_k\right\}\right)\right| \\ \times \left|\left(\max\left\{0, \sum_{k=1}^{K} \hat{g}_{jk} v_k\right\}\right) + \left(\max\left\{0, \sum_{k=1}^{K} g_{jk} v_k\right\}\right)\right| \end{array} \right\} \\
\leq \; & r_n \sum_{j=1}^{J} \left\{ \begin{array}{c} \left|\left(\max\left\{0, \sum_{k=1}^{K} \hat{g}_{jk} v_k\right\}\right) - \left(\max\left\{0, \sum_{k=1}^{K} g_{jk} v_k\right\}\right)\right| \\ \times \left\{\left|\sum_{k=1}^{K} \hat{g}_{jk} v_k\right| + \left|\sum_{k=1}^{K} g_{jk} v_k\right|\right\} \end{array} \right\} \\
\leq \; & r_n \sum_{j=1}^{J} \left\{ \begin{array}{c} \left|\sum_{k=1}^{K} (\hat{g}_{jk} - g_{jk}) v_k\right| \\ \times \left\{\left|\sum_{k=1}^{K} \hat{g}_{jk} v_k\right| + \left|\sum_{k=1}^{K} g_{jk} v_k\right|\right\} \end{array} \right\}, \text{ by (25)} \\
= \; & \frac{r_n}{\sqrt{n}} \sum_{j=1}^{J} \left\{ \begin{array}{c} \left|\sum_{k=1}^{K} \sqrt{n} (\hat{g}_{jk} - g_{jk}) v_k\right| \\ \times \left\{\left|\sum_{k=1}^{K} \hat{g}_{jk} v_k\right| + \left|\sum_{k=1}^{K} g_{jk} v_k\right|\right\} \end{array} \right\}.
\end{aligned}$$

By hypothesis and the fact that $A \cap B$ is a compact set, we get that $\sup_{v \in A \cap B} \left|\hat{Q}_n(v) - Q_n(v)\right| = o_p(1)$. But because $Q_n(v) = Q(v)$ for $v \in A \cap B$, it follows that

$$\sup_{v \in A \cap B} \left|\hat{Q}_n(v) - Q(v)\right| = o_p(1).$$

Note also that $A \cap B$ is compact and $\hat{Q}_n(v)$ is continuous in v. Finally, since $A \cap B$ is compact and $Q(v)$ is strictly convex in v, it follows that $\arg\min_{v \in A \cap B} Q(v)$ is unique. Thus all the conditions for consistency of M-estimators (e.g., Amemiya (1985), theorem 4.1.1) are satisfied, and it follows that

$$p \lim_{n \to \infty} \left\{ \arg \min_{v \in A \cap B} \hat{Q}_n(v) \right\} = \arg \min_{v \in A \cap B} Q(v).$$

The final result follows from the previous display and (24). ■

Corollary (Consistency of value function): Under the same conditions, as the previous proposition,

$$p \lim_{n \to \infty} \left\{ \hat{Q}_n \left(\arg \min_{v \in A} \hat{Q}_n(v) \right) \right\} = Q \left(\arg \min_{v \in A \cap B} Q(v) \right).$$

Proof. Let $v^* = \arg\min_{v \in A \cap B} Q(v)$. By triangle inequality,

$$\begin{aligned}
& \Pr\left\{ \left| \hat{Q}_n(v^n) - Q(v^*) \right| > \varepsilon \right\} \\
< \; & \Pr\left(\left| \hat{Q}_n(v^n) - Q\left(v^{(n)}\right) \right| > \varepsilon/2 \right) + \Pr\left(\left| Q(v^n) - Q(v^*) \right| > \varepsilon/2 \right) \\
= \; & \Pr\left(\left| \hat{Q}_n(v^n) - Q\left(v^{(n)}\right) \right| > \varepsilon/2 \right) + o(1), \text{ by continuous mapping theorem} \\
= \; & \Pr\left(\left| \hat{Q}_n(v^n) - Q\left(v^{(n)}\right) \right| > \varepsilon/2, v^n \in B \right) + o(1), \text{ since } \Pr\left(v^n \notin B\right) \to 0 \\
\leq \; & \Pr\left(\sup_{v \in A \cap B} \left| \hat{Q}_n(v) - Q(v) \right| > \varepsilon/2 \right) + o(1) \\
= \; & o(1), \text{ by uniform convergence on } A \cap B.
\end{aligned}$$

■