Evaluating Treatment Protocols using Data Combination*

Debopam Bhattacharya,
University of Oxford

First draft February, 2009, this draft: November, 2010.

Abstract

In real-life, individuals are often assigned by external planners to binary treatments. Taste-based allocation by planners would make such assignments productively inefficient in that the expected returns to treatment for the marginal treatment recipient would vary across covariates and be larger for discriminated groups. This cannot be directly tested if a planner observes more covariates than the researcher, because then the marginal treatment recipient is not identified. We present (i) a partial identification approach to detecting such inefficiency which is robust to selection on unobservables and (ii) a novel way of point-identifying counterfactual distributions needed to calculate treatment returns by combining observational datasets with experimental estimates. Our methods can also be used to (partially) infer risk-preferences of the planner, which can rationalize the observed data. The most risk neutral solution may be obtained via maximizing entropy. We illustrate our methods using survival data from the Coronary Artery Surgery Study which combined experimental and observational components. Such data combination can be useful even when outcome distributions are partially known. Collecting such data is no harder than running field experiments and its use is analogous to using validation data for measurement error analysis. Our methods apply when individuals cannot alter their potential treatment outcomes in response to the planner's actions, unlike in the case of law enforcement.

*Address for correspondence: debobhatta@gmail.com. I am grateful to seminar participants at CEMMAP, Cambridge and Uppsala and especially to Andrew Chesher for comments and to Amitabh Chandra for pointing me to the CASS dataset. All errors are mine.

1 Introduction

In many real-life situations, external agents (planners, henceforth) assign individuals to treatments using covariate based protocols. For example, welfare agencies assign the unemployed to job-training based on employment record, doctors refer patients to surgical or medical treatment based on clinical test results, colleges admit student applicants to academic programs based on test scores and so on. When protocols are chosen to maximize a functional (say, mean) of the marginal distribution of the resulting outcome subject to cost constraints, the protocol can be said to be "outcome-based". In the above examples, the outcomes can be post-program earnings, days of survival and performance in final examination, respectively. In all these cases, optimal protocol choice will seek to equate the returns to treatment for the marginal treatment recipient across covariate groups but this will typically cause average treatment rates to vary between groups. This is a situation of "statistical discrimination". If, on the other hand, protocols are chosen to maximize a covariate weighted mean of the outcome, the protocol can be said to be covariate-based and the resulting between-group disparities in treatment rates at the optimal choice be regarded as "taste-based" or non-statistical discrimination. For example, consider the case where the treatment is assigning the unemployed to a job-training program and the outcome is post-program earning. Then, an outcome oriented protocol choice will aim to maximize mean earning.[1] In contrast, a covariate-oriented protocol choice will seek to maximize mean weighted earnings where the weights vary with covariate– rather than outcome– values such as race or gender. In most situations of alleged discrimination involving binary treatments, e.g., hiring, medical treatment, college admissions etc., the problem reduces to discerning which of these two mechanisms had led to the observed disparities.

Non-statistical discrimination 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, where productivity is measured in terms of the outcome. This idea has a long history in economics (Becker, 1957, Arrow, 1973) and suggests that distinguishing between statistical and taste-based discrimination may be based on testing inefficiency of treatment assignment using outcomes data. Detecting such inefficiency in practice, however, is difficult because planners typically observe more characteristics than us, the

[1] or more generally, a weighted mean of earnings, where weights depend on earning alone–for instance a logarithmic weight will incorporate risk aversion–but not on covariates like gender or race.

researchers (Heckman, 1998). This makes it hard to rule out the possibility that the subgroup receiving seemingly sub-optimal levels of treatment does so because they are less endowed with some unobservable (to us) qualities which lower their expected outcome from treatment as perceived by the planner. The purpose of this paper is to show how a partial identification approach can be used in this situation to test implications of efficient treatment assignment and, more generally, to infer which welfare functionals, defined on the marginal distribution of outcome, can rationalize observed treatment assignments by the planner.

We focus on the case where the treatment in question is binary but allow the outcome of interest to be either binary or continuous. We assume that an experienced planner 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 and subject to an overall cost-constraint. In this set-up, a necessary condition for the planner's assignment to be productively efficient is that in every observable covariate group, the expected net benefit of treatment to the marginal treatment recipient(s) is weakly greater than a common threshold which, in turn, is weakly greater than the expected net benefit of the marginal treatment non-recipient and where marginal is defined in terms of the characteristics observed by the planner. The planner's assignment results in an observational dataset, where for each individual, we observe her treatment status, her outcome and costs conditional on her treatment status and a set of covariate values. The problem is to test taste-based treatment assignment from these data.

Typically, a single observational dataset is inadequate for this purpose for two reasons. The first, already noted above, is that the planner can base treatment assignment on characteristics that are not observed by us. This makes it 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). Secondly, benefits are also hard to measure using observational data alone because counterfactual means are not observed. In contrast, when we observe all the characteristics observed by the planner and the outcome value without treatment is known (e.g., the final exam score of non-admitted students is zero), then neither of these two problems exist and the observational dataset may be adequate (c.f., sec. 3.4 below). In this case, the minimum of predicted gains from treatment– with predictions calculated using the commonly observed covariates and the minimum taken over their support– for the treated members of each covariate group

represents the gain to the marginal treatment recipient in that group. One can then test if these minima are equal across the groups. Such a strategy does not work if there is selection on unobservables because the researcher– observing fewer covariates– cannot replicate the predicted gains from treatment calculated by the planner.

In this paper, we discuss a new approach to detecting taste-based allocation in such situations using the notion of partial identification. Our approach is motivated by the implication of outcome-based allocation that expected net benefits in every subset of treated individuals must weakly exceed expected net benefits in every subset of untreated individuals– a (conditional) moment inequality condition. These moment inequalities for subsets defined by covariates that the planner observes have testable implications for the (cruder) subsets based on the covariates that we observe and intend to test for.[2] These implications can therefore be tested if we can measure the relevant counterfactual means. We propose a novel way to measure the relevant counterfactuals by combining the observational dataset with experimental or quasi-experimental evidence on treatment effects on subjects drawn from the same population. The latter appears to be of independent interest because such combined data can be used to learn about features of the treatment assignment process in more general settings. We discuss one such setting in section 5, below. The data combination proposed here is analogous to using validation data for measurement error analysis and collecting such combined data involves small incremental effort beyond running a field experiment, as explained below.

Substantive assumptions: We now state the substantive assumptions which define our set-up. The first is that the planner is experienced in the sense that he can form correct expectations. The second is that the planner observes and can condition treatment allocation on all the characteristics (and possibly more than) those that we observe. Third, we observe the same outcomes and costs whose expectations– taken by the planner– should logically determine (productively efficient) treatment assignment in the observational dataset. Fourth, there are negligible externalities, i.e. where treating one individual has a significant impact on the outcome of another individual (c.f., Angelucci et al, 2009) and maximizing the overall outcome needs to take this into account.

The fourth assumption is credible in, say, the case of job-training, mortgage approval or treatment of non-infectious diseases such as heart attack but less so in, say, academic

[2] Which covariates we should test on is guided by the problem at hand– e.g., for gender disparities we analyze expected returns for treated and untreated males and for treated and untreated females.

settings or treatment of infections such as AIDS or malaria. Bhattacharya, 2009 considers roommate assignment in college where peer effects play a crucial role. The third assumption simply clarifies that the notion of productivity (with respect to which inefficiency is defined) must be fixed beforehand and it should be observable and verifiable. The second assumption defines the "selection on unobservables" problem. The first assumption–a "rational expectations" idea is standard for analyzing choice under uncertainty in applied microeconomics (c.f., the KPT paper cited below). It is part of our *definition* of efficiency, i.e., we are testing the joint hypothesis that the planner can calculate correct expectations *and* is allocating treatment efficiently, based on those calculations. This has been termed "accurate statistical discrimination" elsewhere in the literature (c.f., Pope and Sydnor, 2008, Persico, 2009, page 250). Correct expectations are more reasonable for treatments that are fairly routine– such as college admissions to well-established academic programs and less tenable for treatments that are relatively new, e.g., admission to a relatively new academic program.[3] Concerns for misallocation, especially along discriminatory lines, are more frequently voiced for routine treatments and therefore, it makes sense to concentrate on those for the purpose of the present paper. Notice that here we are describing the beliefs of a large central planner who is experienced, rather than small individuals making one-time choice decisions. It is presumably less contentious to expect correct beliefs in the former case than in the latter.

Plan of the paper: 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, describes how a bounds analysis can help detect misallocation and also discusses some extensions and caveats. Section 4 analyzes the complementary problem of inferring a planner's underlying risk-preferences which would justify the current allocations as efficient. Given that the identified set of admissible preferences is typically large dimensional and can therefore be hard to report, we introduce a method of choosing those elements in the identified set that are "close" to some reference ones with specific economic meaning and outline inference theory for such estimated elements. Section 5 sketches an alternative use of data combination in a set-up where correct beliefs are

[3] We will be concerned with expectations conditional on covariate values and so correct expectation is more credible the cruder the conditioning set. In our application, we consider a two-covariate conditioning set.

not assumed but treatment is assigned based on commonly observed covariates. Section 6 presents the empirical illustration and section 7 concludes. The appendix contains further details of proofs/statements mentioned in the main text.

2 Literature

Persico, 2009, provides a comprehensive survey of existing empirical approaches to the detection of taste-based discrimination in general settings. The approaches are varied and their applicability is usually context-specific. Here, we focus on detecting evidence of taste-based assignment of a binary treatment where the treator can be expected to observe more characteristics than the researcher. Our approach is based on using outcome data. In that sense, it is thematically close to Knowles, Persico, Todd, 2001 (KPT, henceforth) who examined the problem of detecting taste-based prejudice separately from statistical discrimination in the context of vehicle search by the police, using data on the search-outcome (hit rates). KPT's 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 implies that equilibrium hit rates should be equalized across *observed* demographic groups under efficient search– a testable prediction. If hit rates are higher for one group, then the police is better-off searching that group more intensively and hence the group is better-off reducing the contraband activity. While the KPT 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 another outcome-based approach, Pope and Sydnor (2008) seek to detect taste-based discrimination in peer-to-peer lending programs. PS use the facts that in these lending programs, (i) the researchers observe all the characteristics that the planners (lenders) observe and (ii) a competitive auction among lenders for funding each individual application drives interest rates so that every approved loan is at the "marginal" level of (expected) return. PS observe the actual returns on the approved loans and can test efficiency by comparing mean (and thus marginal) returns across race for approved loans. The peer-to-peer lending situation is different from job-training, medical treatment etc.,

where the same treatment protocol is used for all applicants and/or treatments are not allocated via a competitive bid, so that the PS approach cannot be used here (c.f., page 11 of PS).

A second aim of the present paper is to infer what outcome-based objectives can rationalize observed treatment disparities across demographic groups. In that sense, it has some substantive similarities to a series of papers in the time-series forecasting literature which propose testing rationality of forecasts made by central agencies (c.f. Elliott, Komunjer and Timmerman (2005), Patton and Timmerman (2007) and references cited therein). The idea there is to (point) estimate parameters of a loss-function which rationalize the observed forecasts. The set-up in that literature assumes that the action (i.e., the forecast) itself has no effect on the distribution of the realized future outcome. In contrast, the key issue in our set-up is that the action (the imposed treatment status) fundamentally determines which distribution the eventual outcome will be drawn from and so the methodology of forecast rationality tests cannot be used in our problem.

A recent set of papers in the econometrics literature have addressed the issue of how treatments should be assigned when only finite sample information is available to the *planner* regarding treatment effectiveness. This is relevant to those treatments that are relatively new, so that the planner is unlikely to know the actual distribution of outcomes with or without treatment– a situation usually termed "ambiguity" in the decision theory literature. See, for instance, Dehejia, 2005, Manski, 2004, 2005, Hirano and Porter, 2008, Stoye, 2006 and Bhattacharya and Dupas, 2010. The present paper may be described as addressing the reverse problem. That is, when the treatment in question is routine and the planner can be expected to know the true outcome distributions (or at least able to form correct expectations), can *we* assess efficiency of the treatment assignment protocols using finite sample evidence, allowing for the possibility of selection on unobservables?

3 Methodology

Using the Neyman-Rubin terminology, denote outcome with and without treatment by Y_0 and Y_1, respectively and let $\Delta Y = Y_1 - Y_0$. We will allow for treatment effects to be negative, i.e., $\Pr(Y_1 - Y_0 < 0)$ may be positive. 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 planner, where the component Z is

not observed by us. Let $\mathcal{W}$ denote the support of W. Let E denote expectations taken w.r.t. the planner's subjective probability distributions, which are assumed to be identical to the true probability distributions in the population. We will assume that all variables defined here have finite expectations. The planner's treatment allocation gives rise to the observational dataset, where for each individual, we observe her treatment status ($D = 1$ or 0), her outcome, Y and cost C which are respectively (Y_1, C_1) or (Y_0, C_0) depending on whether $D = 1$ or 0, and the set of covariates X. For any random variables U, V, let $F_{U|V}(u|v)$ denote the conditional C.D.F. of U at u given $V = v$ and $F_U(\cdot)$ denote the marginal c.d.f. of U.

From the planner's perspective, a treatment protocol is a function $p : \mathcal{W} \to [0, 1]$, specifying the probability of treatment for individuals with $W = w$. Each such protocol will give rise to a distribution of outcome Y, given by

$$F^p(y) = \int \left[\int p(w) F_{Y_1|W}(y|w) + \int \{1 - p(w)\} F_{Y_0|W}(y|w) \right] dF_W(w).$$

An outcome-based criterion is one where protocol p is preferred over protocol q if and only if $F^p(\cdot)$ is preferred over $F^q(\cdot)$. The latter preference could be captured by expected utility i.e. $U(p) = \int u(y) dF(y|p)$ or quantile utility $U(p) = F^{-1}(\tau|p)$ for some $\tau \in [0, 1]$ etc. The important point here is that the planner's preferences are over the *marginal* distribution of Y resulting from the protocol and not the distribution of Y, conditional on W and hence the term "outcome-based". For example, if the treatment is a job-training program and Y is post-program earning, an outcome oriented protocol choice will aim to maximize mean earnings. In contrast, a covariate-oriented protocol choice will seek to maximize mean weighted earnings where the weights vary with covariate values, such as race or gender.

The planner's mean maximization problem in the outcome oriented case is:

$$\max_{p(\cdot)} \left\{ \int \left[\int p(w) y dF_{Y_1|W}(y|w) + \int \{1 - p(w)\} y dF_{Y_0|W}(y|w) \right] dF_W(w) \right\}, \quad (1)$$

s.t.

$$\int \left[\int p(w) s dF_{C_1|W}(s|w) + \int \{1 - p(w)\} s F_{C_0|W}(s|w) \right] dF_W(w) \leq c. \quad (2)$$

The solution, as shown in the appendix part A, is of the form[4][5]

$$\begin{aligned} p^*(w) &= 1\{E(\Delta Y - \gamma \Delta C | W = w) \geq 0\}, \text{ with} \\ c &= \int_{w \in \mathcal{W}} 1\{E(\Delta Y - \gamma \Delta C | W = w) \geq 0\} dF_W(w). \end{aligned} \tag{3}$$

To keep the problem realistic, we will assume that the constraint is such that not all individuals with positive expected treatment effect can be treated, i.e.,

$$c << \int_{w \in \mathcal{W}} 1\{E(\Delta Y | W = w) \geq 0\} dF_W(w),$$

which will imply that the the constraint binds at the optimum.

In contrast, in the covariate-oriented case, the planner will maximize

$$\max_{p(\cdot)} \left\{ \int_{w \in \mathcal{W}} h(w) \left[\int p(w) y dF_{Y_1|W}(y|w) + \int \{1 - p(w)\} y dF_{Y_0|W}(y|w) \right] dF_W(w) \right\}, \tag{4}$$

subject to (2), where $h(w)$ represents "taste-based weights" used by the planner for inflating the outcome of individuals with covariate equal to w. In this case, the solution will be of the form

$$\begin{aligned} p^*(w) &= 1\{E(h(w)\Delta Y - \Delta C | W = w) \geq 0\}, \text{ with} \\ c &= \int_{w \in \mathcal{W}} 1\{E(h(w)\Delta Y - \Delta C | W = w) \geq 0\} dF_W(w). \end{aligned}$$

Denoting the net expected benefit from treatment by

$$\beta(w) = \frac{E[\Delta Y | W = w]}{E[\Delta C | W = w]},$$

it follows that in the outcome oriented case, type w is treated when $\beta(w)$ exceeds the fixed threshold γ but for the taste-based case, the corresponding threshold $\frac{1}{h(w)}$ varies by w and is lower for those w's whose outcomes are more important to the planner. In either case, the threshold represents the return to treatment for the marginal treatment recipient; in the outcome-based case, it stays constant across covariates W but in the

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

[5] Also, in the (essentially "measure zero") case of a tie– viz., where the budget constraint is such that some but not all individuals of the marginal group can be treated, we implicitly assume that the treatment is randomized among members of the marginal group.

taste-based case, it varies with W.[6]Thus, a test of taste-based assignment can be based on comparing the treatment thresholds for different covariate-groups and testing if they are equal. However, due to selection on the unobservables Z, the marginal treatment recipient and consequently the treatment threshold cannot be identified. We now show how certain inequalities implied by efficient treatment assignment may be useful for detecting taste-based allocation.

Testable Inequalities: In the outcome-oriented case, since the planner 's subjective expectations are assumed to be consistent with true distributions in the population, we must have that w.p.1,

$$\begin{aligned} E(\Delta Y|X,Z,D=1) &\geq \gamma E(\Delta C|X,Z,D=1), \\ E(\Delta Y|X,Z,D=0) &\leq \gamma E(\Delta C|X,Z,D=0). \end{aligned} \tag{5}$$

Given the allocation procedure leading to (5), 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 discrimination.

Since we do not observe Z, even the inequalities in (5) are not of immediate use to us. However, an implication of (5) is potentially useful for detecting taste-based allocation. Let $\mathcal{X}^j$ denote the support of X for the subpopulation who would be assigned $D=j$, $j=0,1$ by the planner. Indeed, (5) implies that

$$\begin{aligned} &\int_{z\in \text{supp}(Z|D=1,X)} E(\Delta Y|X,Z,D=1)\, dF_{Z|X,D=1}(z|X,D=1) \\ \geq\ & \gamma \int_{z\in \text{supp}(Z|D=1,X)} E(\Delta C|X,Z,D=1)\, dF_{Z|X,D=1}(z|X,D=1), \end{aligned}$$

i.e.

$$E[\Delta Y - \gamma \Delta C|D=1, X=a] \geq 0, \text{ for all } a\in\mathcal{X}^1, \tag{6}$$

and similarly

$$E[\Delta Y - \gamma \Delta C|D=0, X=a] \leq 0, \text{ for all } a\in\mathcal{X}^0. \tag{7}$$

[6]If $\beta(w)$ equals the threshold for some value(s) of w, then it represents the return to the marginal treatment recipient; if not– e.g. if all elements of W are discrete– then it is a lower bound on the return to the marginal treatment non-recipient and an upper bound on the return to the marginal treatment non-recipient. But in either case,

$$\min_{w:D=1} \beta(w) \geq threshold \geq \max_{\omega:D=0} \beta(w),$$

which is what we work off.

In words, if the planner is outcome-oriented, then the net benefit from treatment for every subgroup (that the planner can observe) among the treatment recipients must weakly exceed the treatment threshold, i.e. $\frac{E[\Delta Y|D=1,W=w]}{E[\Delta C|D=1,W=w]} \geq \gamma$. 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, i.e. $\frac{E[\Delta Y|D=1,X=x]}{E[\Delta C|D=1,X=x]} \geq \gamma$. This leads to (6) and analogously for (7). This reasoning lets us overcome the problem posed by the planner 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

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

then we conclude that there is misallocation in terms of the mean outcome in a way that hurts type b people.

Counterfactuals: To be able to use the above inequalities to learn about γ, we need to identify the counterfactual mean outcomes $E\left(Y_0|X,D=1\right)$ and $E\left(Y_1|X,D=0\right)$ and the counterfactual mean costs $E\left(C_0|X,D=1\right)$ and $E\left(C_1|X,D=0\right)$. 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 $x \in \mathcal{X}^1$,

$$\begin{aligned}
\underbrace{P\left(Y_0 \leq y|X=x\right)}_{\text{known from expt}} &= P^{obs}\left(Y_0 \leq y|X=x\right) \\
&= P^{obs}\left(Y_0 \leq y|D=1,X=x\right) \times \underbrace{P^{obs}\left(D=1|X=x\right)}_{\text{known from obs}} \\
&\quad + \underbrace{P^{obs}\left(Y_0 \leq y|D=0,X=x\right)}_{\text{known from obs}} \times \underbrace{P^{obs}\left(D=0|X=x\right)}_{\text{known from obs}}. \qquad (8)
\end{aligned}$$

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

$$\begin{aligned}
\underbrace{\Pr\left(Y_1 \leq y|x\right)}_{\text{known from expt}} &= \Pr\left(Y_1 \leq y|D=0,x\right) \times \underbrace{\Pr\left(D=0|x\right)}_{\text{known from obs}} \\
&\quad + \underbrace{\Pr\left(Y_1 \leq y|D=1,x\right)}_{\text{known from obs}} \times \underbrace{\Pr\left(D=1|x\right)}_{\text{from obs}}. \qquad (9)
\end{aligned}$$

Thus the two equalities above yield the counterfactual distributions $P(Y_0 \leq y|D=1,x)$ on $\mathcal{X}^1$ and $P(Y_1 \leq y|D=0,x)$ 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 $x \in \mathcal{X}^0$,

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

Bounds: Combining (6), (7), (8) and (9) yield the following bounds on γ:

$$\begin{aligned} \gamma_{lb} &= \sup_{x \in \mathcal{X}^0} \left(\frac{\underbrace{E(Y_1|X=x,D=0)}_{\text{from (8)}} - \underbrace{E(Y_0|X=x,D=0)}_{\text{from obs data}}}{\underbrace{E(C_1|X=x,D=0)}_{\text{from (8)}} - \underbrace{E(C_0|X=x,D=0)}_{\text{from obs data}}} \right), \\ \gamma_{ub} &= \inf_{x \in \mathcal{X}^1} \left(\frac{\underbrace{E(Y_1|X=x,D=1)}_{\text{from obs data}} - \underbrace{E(Y_0|X=x,D=1)}_{\text{from (9)}}}{\underbrace{E(C_1|X=x,D=1)}_{\text{from obs data}} - \underbrace{E(C_0|X=x,D=1)}_{\text{from (9)}}} \right). \end{aligned} \tag{10}$$

The bounds derived above essentially replace a minimum over finer subgroups (observed by the planner) 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.

Simplified calculation: Observe that the lower bound calculation, suppressing x, reduces to

$$\begin{aligned} & E(\Delta Y|D=0) \\ = & \frac{E(Y_1) - E(Y_1|D=1)\Pr(D=1)}{\Pr(D=0)} - E(Y_0|D=0) \\ = & \frac{E(Y_1) - E(Y_1|D=1)\Pr(D=1) - \Pr(D=0)E(Y_0|D=0)}{\Pr(D=0)} \\ = & \frac{E(Y_1) - E(DY_1) - E((1-D)Y_0)}{\Pr(D=0)} = \frac{E(Y_1) - E(Y)}{\Pr(D=0)}. \end{aligned}$$

Similarly, for the upper bound,

$$E(\Delta Y|D=1) = \frac{E(Y) - E(Y_0)}{\Pr(D=1)}.$$

The bounds are then easily calculated as

$$\begin{aligned}\gamma_{ub} &= \inf_{x\in\mathcal{X}^1}\left\{\frac{E^{obs}\left(Y|X=x\right)-E^{\exp}\left(Y_0|X=x\right)}{E^{obs}\left(C|X=x\right)-E^{\exp}\left(C_0|X=x\right)}\right\}\\ \gamma_{lb} &= \sup_{x\in\mathcal{X}^0}\left\{\frac{E^{\exp}\left(Y_1|X=x\right)-E^{obs}\left(Y|X=x\right)}{E^{\exp}\left(C_1|X=x\right)-E^{obs}\left(C|X=x\right)}\right\}.\end{aligned}$$

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 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 planner who uses his/her discretion. This design was used in the CASS (1981) study in the US for studying the efficacy of coronary artery surgery. This is the set-up used to derive (8) and (9) above, which is motivated by our empirical application.[7] The second way is as follows. First, present all the individuals to the planner 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 planner 's recommendation) and observe the outcomes for each individual. The counterfactual $P\left(Y_0|D=1,X\right)$ can then be obtained directly (i.e. without using (8) and (9)) from the outcomes of those who are approved by the planner but were randomized out of treatment. Conversely for $P\left(Y_1|D=0,X\right)$.

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

[7] In the appendix, we present a brief outline of how our methods need to be modified if the experimental sample has worse outcomes and/or higher costs than the observational sample, which may happen in medical trials. In brief, this would widen the bounds and make it harder to detect inefficiency. But if inefficiency is detected with wider bounds, then it would also have been detected with narrower bounds.

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

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

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

$$\frac{E[\Delta Y|fem=0,D=1]}{E[\Delta C|fem=0,D=1]} < \frac{E[\Delta Y|fem=1,D=0]}{E[\Delta C|fem=1,D=0]}, \tag{11}$$

or

$$\frac{E[\Delta Y|fem=1,D=1]}{E[\Delta C|fem=1,D=1]} < \frac{E[\Delta Y|fem=0,D=0]}{E[\Delta C|fem=0,D=0]}. \tag{12}$$

The first inequality (11) means that the return to treatment among treated males is less than that among untreated females– i.e., females are being under-treated. Equivalently, females face a higher threshold. Similarly, (12) means that males are being under-treated.

Notice that the inequalities (12) or (11) can be interpreted and used directly without reference to a specific model of optimization or treatment allocation such as (3). However, the link with (3) gives our analysis a firm grounding in classical economic theory of choice under uncertainty.

3.2 Implicit discrimination

Suppose the two groups of interest are the rich and the poor. Assume identical treatment costs for now and suppose we detect an inequality of type (11):

$$E[\Delta Y|poor = 0,D = 1] < E[\Delta Y|poor = 1,D = 0],$$

which suggests that there is taste-based treatment assignment that hurts the poor. It is possible that this is brought about by a planner who practices taste-based discrimination against blacks but is not necessarily biased against the poor. The following scenario illustrates the point. Suppose it is the case that for two constants $\lambda_{bl} > \lambda_{wh}$, we have

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

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

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

Since it is the case that

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

we will conclude that

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

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. This shows that detecting misallocation that hurts a group we chose to test may not imply that the planner is practising taste-based allocation where taste dictates him to be biased for or against the characteristics which define the chosen group– it could arise from intentional discrimination against a positively correlated characteristic.

3.3 Selection on observables

In some situations, a researcher may have access to exactly the same set of covariates W that the planner had observed prior to making the decision. Examples include written application for loans or student admissions where applications are scored and the application forms and scores respectively are made available to the researcher. Even in this case, the treatment threshold may not be point-identified. To see this, consider the situation where there is a single covariate–say gender– that is observed both by the planner and us and no other covariate is observed by anyone else. Also assume that $\Delta C = \eta$, a known fixed cost of treatment. Suppose the planner assigns individuals to treatment only if $E(\Delta Y - \gamma\eta | gender) \geq 0$. Now suppose the expected benefit of treatment to women is δ_f and that to men is δ_m and they satisfy $\delta_m < \gamma\eta < \delta_f$. Then

$$E(\Delta Y|D=1) = E(\Delta Y|fem) = \delta_f > \delta_m = E(\Delta Y|male) = E(\Delta Y|D=0)$$

and all we know is that $\gamma \in \left(\frac{\delta_m}{\eta}, \frac{\delta_f}{\eta}\right)$. Thus, even when selection into treatment is based only on observables, the treatment threshold may not be point-identified.

4 Broadening the model

We now extend the analysis to include risk averse behavior by the planner 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.

4.1 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 (3)

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

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 planner '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:

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

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

So the values of θ 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 gender are the ones for which

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

If the set of θ 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.

4.2 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'

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

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':

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

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

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

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

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 planner rationality on whether the identified set of u's is empty.

4.2.1 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.,

$$\frac{E(\Delta Y|D=1,G=1)}{E(\Delta C|D=1,G=1)} \leq \frac{E(\Delta Y|D=0,G=2)}{E(C_1 - C_0|D=0,G=2)}, \text{ i.e.,}$$

$$\sum_{j=1}^{m} a_j \left[\mu_{111}(j) - \mu_{011}(j) - \mu_{102}(j) + \mu_{002}(j)\right] \leq 0. \tag{18}$$

Let

$$r_j = \mu_{111}(j) - \mu_{011}(j) - (\mu_{102}(j) - \mu_{002}(j)),$$

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

$$\frac{E(u(Y_1) - u(Y_0)|D=1,G=1)}{E(\Delta C|D=1,G=1)} \geq \frac{E(u(Y_1) - u(Y_0)|D=0,G=2)}{E(\Delta C|D=0,G=2)},$$

$$\text{i.e., } \sum_{j=1}^{m} r_j u(a_j) > 0. \tag{19}$$

The following propositions provide 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 $l = 2, ...m$.

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

(i) $S_l \geq 0$, *for every* $l = 2, ..., m$.

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

Proposition 2 *Suppose* $\{r_j\}$ *is such that* $\sum_{j=1}^{m} r_j = 0$. *The following conditions are equivalent:*

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

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

The conditions in these propositions can be checked directly before we try to find the set of solutions. Note that these two propositions are similar in spirit to the equivalence of second (first) order stochastic dominance and dominance in terms of every concave and monotone (resp., 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 part B. ■

4.2.2 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 construct confidence intervals for both the solution and the value function for the above problems, using the distribution theory for the estimated q's.

4.2.3 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., (11) or (12) 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., Chernozhukov et al (2007). Tests of rationality again amount to checking emptiness of confidence sets, which can be done using Andrews and Soares (2007), for instance.

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

problem where the q functions in (20) are replaced by their estimate. Notice that this problem is distinct from M-estimation problems with deterministic parameter constraints. Here the constraints involve estimated terms and the objective function is deterministic which is the opposite of, say, inequality constrained least squares (c.f., Wolak (1989)).

We now outline a method of solving the sample analog of (20) and conduct inference on the solution thereof. To do this, focus on the case where both Y and X are discrete, set $a_k = k/K$ for all k (to save on notation) and rewrite the last set of inequalities in the previous display as $\sum_{k=1}^{K} \hat{g}_{jk} v_k \leq 0$ for $j = 1, ...J$. We drop the concavity requirement on the sub-utility function and simply impose that it be non-decreasing. This allows for more "behavioral" features such as loss aversion. Define

$$Q(v) = \sum_{k=1}^{K} v_k \ln(v_k), \; k = 0, 1, ...K.$$

Denote the solution

$$v^0 = \arg\min_v Q(v) \text{ s.t. } v \geq 0, \; v'1 = 1 \text{ and } \sum_{k=1}^{K} g_{jk} v_k \leq 0, \; j = 1, ...J.$$

Consider the estimator $\hat{v}$, defined as

$$\hat{v} = \arg\min_v Q(v) \text{ s.t. } v \geq 0, \; v'1 = 1 \text{ and } \sum_{k=1}^{K} \hat{g}_{jk} v_k \leq \delta_n, \; j = 1, ...J.$$

Proposition 3 *(Consistency) Assume that* $\sqrt{n} vec(\hat{g} - g) \rightsquigarrow N(0, \Sigma)$. *Choose* δ_n *such that* $\delta_n \to 0$ *and* $\sqrt{n}\delta_n \to \infty$, *as* $n \to \infty$. *Then* $p\lim \hat{v} = v^0$.

Proof. See appendix part C. ■

The asymptotic distribution for $\hat{v}$ is interesting in that it entails an asymptotic bias term. Here we consider the problem without concavity constraints on the sub-utility functions but maintain the non-decreasing property. Define λ_m to be the Lagrange multiplier corresponding to the mth g-constraint for the population problem. Suppose constraints $1, ...J_1$ bind at v^0, i.e. $g_j' v_0 = 0$ for $j = 1, ...J_1$ and $g_j' v^0 < 0$ for $j = J_1 + 1, ...J$. Let

$$\begin{aligned} A_{J_1 \times J_1} &= \begin{bmatrix} \sum_k g_{1k}^2 \left(e^{\sum_{m=1}^{J_1} \lambda_m g_{mk}} \right) & ... & \sum_k g_{1k} g_{J_1 k} \left(e^{\sum_{m=1}^{J_1} \lambda_m g_{mk}} \right) \\ ..... & .... & .... \\ \sum_k g_{1k} g_{J_1 k} \left(e^{\sum_{m=1}^{J_1} \lambda_m g_{mk}} \right) & ... & \sum_k g_{J_1 k}^2 \left(e^{\sum_{m=1}^{J_1} \lambda_m g_{mk}} \right) \end{bmatrix}, \\ g_{\cdot k}' &= (g_{1k}, ..., g_{J_1 k}), \; k = 1, ...K, \\ 1_{J_1} &= (1...1)'_{1 \times J_1}. \end{aligned}$$

Proposition 4 *Assume that* $\sqrt{n}vec\left(\hat{g}-g\right) \rightsquigarrow N\left(0,\Sigma\right)$. *Choose* δ_n *such that* $\delta_n \to 0$, $\sqrt{n}\delta_n \to \infty$, *as* $n \to \infty$ *and* $\sqrt{n}\delta_n^2 \to 0$ *as* $n \to \infty$. *Then for any* $k = 1, 2, ...K$,

$$\begin{aligned}
&\sqrt{n}\left(\ln \hat{v}_k - \ln v_k^0 - \delta_n \left(\sum_{k=1}^{K} e^{\lambda' g_{\cdot k}}\right) g_{\cdot k}' A^{-1} 1_{J_1}\right) \\
= \; & -g_{\cdot k}' A^{-1}\sqrt{n}\sum_{k=1}^{K}\left\{\hat{g}_{\cdot k} e^{\lambda'\hat{g}_{\cdot k}} - g_{\cdot k} e^{\lambda' g_{\cdot k}}\right\} \\
& +\lambda'\sqrt{n}\left\{\left(\hat{g}_{\cdot k} - g_{\cdot k}\right) - \frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}}\left(\hat{g}_{\cdot l} - g_{\cdot l}\right)}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}}\right\} + o_p\left(1\right).
\end{aligned}$$

The distribution of $\hat{v}$ *follows by the standard delta-method.*

Proof. See appendix part C ■

There are three points to note about the distribution stated above. First, the asymptotic distribution of $\ln\left(\hat{v}\right)$ and, consequently $\hat{v}$, will have an asymptotic bias term, proportional to δ_n which, when multiplied by $\sqrt{n}$ will go to $\pm\infty$. Hence, constructing a CI for v_k will require us to remove this bias and a standard resampling-based bias correction, where the bias in $\hat{v}_k^* - \hat{v}_k$ calculated by averaging across repeated resamples is subtracted from $\hat{v}_k - v_k$ to calculate the CI, will lead to (first order) correct coverage asymptotically. Second, the asymptotic distribution depends on which constraints bind. If none of the constraints bind, i.e., $J_1 = 0$, then $\Pr\left(\hat{v} \neq v\right) \to 0$ as $n \to \infty$, so that the point-estimate $\hat{v}$ can be regarded as a degenerate 100% CI for v. More generally, a CI can be constructed by first testing which constraints bind. The CI is thus of a pre-test variety and adjustments need to be made to the critical values at the first stage testing to guarantee correct coverage for the eventual CI for v. To test if a particular constraint binds, i.e., $H_0 : g'v^0 = 0$ vs $H_1 : g'v^0 < 0$, we can use the criterion: reject if $\hat{g}'\hat{v} \leq a_n$ where $a_n = o\left(\delta_n\right)$. We show in the appendix part C that this will give us a consistent test of $g'v^0 = 0$.

An alternative is to construct conservative CI's which are valid no matter which constraints bind. The latter are likely to have large volumes but can be relatively easy to construct as follows. We now let $g_{\cdot k}$ and A to have dimension equal to J and let λ_j

determine which ones appear in the expressions, i.e.

$$\begin{aligned} A_{J\times J} &= \begin{bmatrix} \sum_k 1(\lambda_1>0) g_{1k}^2 \left(e^{\sum_{m=1}^J \lambda_m g_{mk}}\right) & ... & \sum_k 1(\lambda_1,\lambda_J>0) g_{1k}g_{Jk} \left(e^{\sum_{m=1}^J \lambda_m g_{mk}}\right) \\ ..... & .... & .... \\ \sum_k 1(\lambda_1,\lambda_J>0) g_{1k}g_{Jk} \left(e^{\sum_{m=1}^J \lambda_m g_{mk}}\right) & ... & \sum_k 1(\lambda_J>0) g_{Jk}^2 \left(e^{\sum_{m=1}^J \lambda_m g_{mk}}\right) \end{bmatrix}, \\ g_{\cdot k}' &= (1(\lambda_1>0) g_{1k}, ..., 1(\lambda_J>0) g_{Jk}), \ k=1,...K, \\ 1_J &= (1...1)_{1\times J}'. \end{aligned}$$

Then it follows from the previous display that

$$\begin{aligned} w_n &= \sqrt{n}\left(\ln \hat{v}_k - \ln v_k^0 - \delta_n \left(\sum_{k=1}^K e^{\lambda' g_{\cdot k}}\right) g_{\cdot k}' A^{-1} 1_{J_1}\right) \\ &= \sum_{j=1}^J 1(\lambda_j>0) Z_{1j} + \sum_{j=1}^J 1(\lambda_j>0)\lambda_j Z_{2j} + o_p(1), \end{aligned}$$

where Z_{1j}, Z_{2j} are asymptotically normal random variables. This implies that w.p.a.1,

$$\min_j \{Z_{1j}\} + \min_j \{\lambda_j Z_{2j}\} \leq w_n \leq \max_j \{Z_{1j}\} + \max_j \{\lambda_j Z_{2j}\}$$

and the distribution of the bounding random variables can be simulated by replacing λ_j by its consistent estimate $\hat{\lambda}_j$.

5 Inferring uncertainty aversion from treatment choice

Our data combination method can be used in treatment assignment situations, where the planner may not have correct expectations to start with and is ambiguity averse, i.e., for two treatment protocols with the same (subjective) expected outcomes, the planner would prefer the one for which there is less "parameter uncertainty". In this section we outline this problem and show how data combination is useful for learning the planner's aversion to parameter uncertainty.

In this set-up, there is an experimental dataset where the treatment was randomly assigned, the entire dataset is available to both the planner and us. Next, there is an observational dataset where the planner observed the characteristics of the subjects and assigned them to treatment, using his knowledge of treatment effects from the experimental data alone. In the observational data, we observe the characteristics of the subjects and their treatment status, as determined by the planner. From this, we try to infer the

planner's extent of uncertainty aversion. Thus, we drop the assumption that the planner knows the true outcome distributions but impose that the planner and we observe (i) the same experimental data and (ii) the same covariates for the observational subjects.

To fix ideas, consider the case with discrete X taking values 1, 2,...J. Let $\pi_{1j} = \Pr(Y_1 = 1|X = j)$, $\pi_{0j} = \Pr(Y_0 = 1|X = j)$, $q_j = \Pr(X = j)$. The following exposition is motivated by the job-training example, where the cost of training a candidate does not vary by covariate values but generalization to that case is not hard. Further, we will assume that q_j, the fraction of unemployed with $X = j$ in the observational study is known to the planner at the time of deciding on the protocol. We will maintain these assumptions throughout this section.

If the planner uses protocol $p : x \longmapsto [0, 1]$, his expected outcome is

$$\sum_{j=1}^{J} \{p_j \pi_{1j} + (1 - p_j) \pi_{0j}\} q_j.$$

Since the π's are now assumed unknown to the planner, the above expectation cannot be calculated directly. But, based on the experimental dataset, the planner can construct a posterior distribution for the π's, given his priors. We will assume that the planner chooses p to maximize

$$\begin{aligned} W(p) &= \int V\left(\sum_{j=1}^{J} \{p_j \pi_{1j} + (1 - p_j) \pi_{0j}\} q_j; \alpha\right) dF_{pos}(\pi|Z^n), \\ \text{s.t.} \sum_{j=1}^{J} q_j p_j &= c, \end{aligned}$$

where $F_{pos}(\pi|Z^n) = \Pi_{j=1}^{J} dF_{pos}(\pi_{1j}, \pi_{0j}|Z^n)$ denotes the posterior for π, Z^n denotes the experimental data observed by us as well as the planner and V is a known class of utility functions, strictly concave in the first argument and indexed by the uncertainty aversion parameter α known to the planner but unknown to us. Examples include $V(t; \alpha) = -e^{-\alpha t}$ or $V(t; \alpha) = \frac{t^{\alpha-1}}{\alpha-1}$ etc. For each α, the planner's optimization problem written above, due to the strict concavity of V, will have a unique solution $\{p_j(\alpha)\}_{j=1,...J}$. Since we observe the realized values of treatment d_i following the planner's assignment of each individual i in the observational data, we can estimate α by maximizing the likelihood:

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{N} \left\{ d_i \ln\left(\sum_{j=1}^{J} 1(X_i = j) p_j(\alpha)\right) + (1 - d_i) \ln\left(\sum_{j=1}^{J} 1(X_i = j) \{1 - p_j(\alpha)\}\right)\right\}.$$

Here the asymptotics is on the number N of individuals in the observational data and the distribution theory of interest is that of $\sqrt{N}\,(\hat{\alpha} - \alpha)$, from which a confidence interval for α can be constructed. For this problem, flat priors on (π_{1j}, π_{0j}) may be specified through copulas whence the marginal prior on π_{1j} and on π_{0j} are uniform [0,1].

6 Empirical illustration

We now present an empirical illustration of the methodology developed in section 3. The illustration is based on the Coronary Artery Surgery Study (CASS), conducted in early 1980's in the US. A detailed description of the study design and its findings is provided in the CASS paper cited below. Here we provide a brief overview. The purpose of the present illustration is to show how our method performs in a real dataset that has the data combination flavor. A more substantive empirical analysis of these data is being conducted by the present author in an ongoing collaborative project (c.f., reference 7 below).[8]

The goal of the CASS study was to evaluate the effectiveness of coronary artery surgery versus medical therapy in patients with mild to moderate angina. Patients with severe angina were excluded from the study–bypass surgery was already known to improve longevity in such patients. The design involved dividing the patients into a trial arm where patients were randomized into or out of surgery and a non-trial arm where they were assigned to surgery by physician discretion. The stated goal of this design, deduced ex-post by the present author from the research paper cited below, was to check if outcomes with and without the treatment were different in the experimental arm from that in the observational arm. The study did not find any appreciable difference and it is unclear to the present author as to what this conclusion implies. Nonetheless, the study design is ideal for the objective of the present paper and provides a useful dataset for illustrating the usefulness of the methodology developed above.

Specifically, in the CASS study, all patients undergoing coronary angiography in participating sites and who showed indication of suspected or proven coronary artery disease were entered into a registry (about n = 25,000). Out of these 2,100 were medically eligible for randomization (< 65 years, mild to moderate angina, etc.) Out of these 2100, about

[8] The CASS data may be obtained through online request at https://biolincc.nhlbi.nih.gov/studies/cass/?q=CASS.

	Experimental		Observational		t-test
Variable	Mean	Std. Dev.	Mean	Std. Dev	p-value
death	0.36	0.48	0.34	0.47	
treatment	0.50	0.50	0.43	0.50	
unemployed	0.29	0.44	0.28	0.45	0.63
age	51.10	7.31	50.87	7.82	0.63
lvscor	7.55	2.90	7.46	2.96	0.54
previousmi	0.62	0.49	0.59	0.49	0.16
diabetes	0.09	0.28	0.06	0.24	0.04
stroke	0.01	0.12	0.01	0.10	0.52
smoking	0.40	0.49	0.42	0.47	0.46
N	704		1192		

Figure 1:

1320 patients were not randomized and are referred to as randomizable patients and they constitute our observational group. 780 patients were evenly randomized into medical or surgical arms– the "randomized" patients constituting the experimental group. The specific surgical (medical) therapy given to a surgical (medical) patient was decided by the physician attending to the case. The primary endpoints of the study included death and myocardial infarction (hear attack), and secondary endpoints included evaluation of angina and quality of life. About 17 years of follow-up data for vital status were included. Due to some cross-over in the long-run,[9] we will refer to being assigned to surgery as the treatment. Also, we choose only males for our analysis. Females constitute less than 10% of the study sample and race is not recorded.

Summary statistics for some key variables is provided in the table marked "figure 1". The variable lvscor is an index for how well the heart functions, with $5-8$ being a normal range; previousmi is a dummy for whether the patient had a previous heart attack and smoking is a dummy for whether the patient is currently smoking. Our outcome variable of interest, labeled "death", is the binary indicator for whether the patient died within 17 years from the date of treatment assignment.In the ideal situation, the enrollment into the experimental arm should be random. However, in the CASS case, this seems to

[9] 31 of the 390 patients randomized into the surgical group refused surgery and utilized medical therapy instead and about $\frac{1}{4}$ of the 390 patients in the medical arm elected to undergo surgical therapy in the long run.

have been influenced to some extent by the physicians who were treating the patients. In terms of most observable characteristics however, the two groups seem very similar. They are very slightly different in terms of prior incidence of heart-attack and diabetes, for which the experimental group seems slightly sicker. As explained in the appendix part D, labelled "Non-identical distributions", this means that our bounds are still valid but wider. When we do detect different thresholds using these wider bounds, we would also have detected different thresholds under narrower bounds which would result if enrollment into the experimental arm were random.

A drawback of this dataset is that no cost figures are available. So we focus on testing efficiency in terms of the survival outcome alone without regard to costs. However, we do have a variable recording employment status. Since all individuals in this dataset are under 65 (and hence not covered by Medicare), we may regard employment status as a crude proxy for health insurance coverage. In this case, cost-criteria might lead the non-employed to receive the treatment less frequently than their health outcomes alone might dictate.

Without cost numbers, our bounds are

$$\begin{aligned}
\gamma_{lb} &= \sup_{x\in\mathcal{X}^0} E\left(Y_1 - Y_0 | D=0, X=x\right) = \sup_{x\in\mathcal{X}^1} \left\{ \frac{E^{\exp}\left(Y_1|X=x\right) - E^{obs}\left(Y|X=x\right)}{\Pr^{obs}\left(D=0|X=x\right)} \right\}, \\
\gamma_{ub} &= \inf_{x\in\mathcal{X}^1} E\left(Y_1 - Y_0 | D=1, X=x\right) = \inf_{x\in\mathcal{X}^0} \left\{ \frac{E^{obs}\left(Y|X=x\right) - E^{\exp}\left(Y_0|X=x\right)}{\Pr^{obs}\left(D=1|X=x\right)} \right\}.
\end{aligned}$$

We first consider the case where the groups of interest are the unemployed versus employed and then we will consider smokers versus non-smokers and use q quantiles of age to narrow the bounds. That is, calculate the sample analog of the following bounds: for unemployed,

$$\begin{aligned}
\gamma_{lb}^{unem} &= \max_{x\in(1,\ldots,q)} \left\{ \frac{E^{\exp}\left(Y_1|age_q=x, unem=1\right) - E^{obs}\left(Y|age_q=x, unem=1\right)}{\Pr^{obs}\left(D=0|age_q=x, unem=1\right)} \right\}, \\
\gamma_{ub}^{unem} &= \min_{x\in(1,\ldots,q)} \left\{ \frac{E^{obs}\left(Y|age_q=x, unem=1\right) - E^{\exp}\left(Y_0|age_q=x, unem=1\right)}{\Pr^{obs}\left(D=1|age_q=x, unem=1\right)} \right\}.
\end{aligned}$$

Similarly, for employed:

$$\begin{aligned}
\gamma_{lb}^{emp} &= \max_{x\in(1,\ldots,q)} \left\{ \frac{E^{\exp}\left(Y_1|age_q=x, unem=0\right) - E^{obs}\left(Y|age_q=x, unem=0\right)}{\Pr^{obs}\left(D=0|age=x, unem=1\right)} \right\}, \\
\gamma_{ub}^{emp} &= \min_{x\in(1,\ldots,q)} \left\{ \frac{E^{obs}\left(Y|age_q=x, unem=0\right) - E^{\exp}\left(Y_0|age_q=x, unem=0\right)}{\Pr^{obs}\left(D=1|age_q=x, unem=0\right)} \right\}.
\end{aligned}$$

Instead of reporting the estimated max or min, we report the simple average, e.g., instead of

$$\hat{\gamma}_{lb}^{unem} = \max_{x\in(1,...,q)} \left\{ \frac{\hat{E}^{\exp}\left(Y_1|age\_q = x, unem = 1\right) - \hat{E}^{obs}\left(Y|age\_q = x, unem = 1\right)}{\hat{P}^{obs}\left(D = 0|age\_q = x, unem = 1\right)} \right\},$$

we report

$$\tilde{\gamma}_{lb}^{unem} = \frac{1}{q}\sum_{x=1}^{q} \left\{ \frac{\hat{E}^{\exp}\left(Y_1|age\_q = x, unem = 1\right) - \hat{E}^{obs}\left(Y|age\_q = x, unem = 1\right)}{\hat{P}^{obs}\left(D = 0|age\_q = x, unem = 1\right)} \right\}.^{10}$$

Because the max will seek out those observations for which the probability in the denominator is small, the estimate $\hat{\gamma}_{lb}^{unem}$ is likely to be very noisy. Since

$$\gamma_{lb}^{unem} \leq \frac{E^{\exp}\left(Y_1|age\_q = x, unem = 1\right) - E^{obs}\left(Y|age\_q = x, unem = 1\right)}{\Pr^{obs}\left(D = 0|age\_q = x, unem = 1\right)},$$

for each x, it must also be smaller than the simple average of the RHS over $x \in (1, ..., q)$. So $\tilde{\gamma}_{lb}^{unem}$ is a valid (potentially conservative) but less noisy estimate. Also, one can bootstrap $\tilde{\gamma}$'s to construct p-values, unlike the case of max or mins for which there is a boundary-value problem (Andrews, 2000). The estimated differences between the upper bound of threshold for one group less the lower bound for threshold of the other group are reported in the following tables for two different choices of groups. A negative difference suggests that the second group is facing a higher threshold for treatment– i.e., there is taste-based allocation against the second group. In the table, marked "figure 2", we report the results corresponding to $q = 2$ and $q = 10$. The (one-sided) p-values were computed by bootstrapping the end-points jointly. When taste-based allocation is detected, we highlight the corresponding entry.This table suggests that the treatment threshold for the non-employed is *lower* than that for the employed, contrary to our original hypothesis. Why that should be the case cannot be answered convincingly within the confines of this illustrative section. One possibility is that invasive procedures require longer recovery periods which may be easier to implement when the person is not in employment. Further, as we hinted above, not being employed does not necessarily mean that the individual

[10] Notice that this is not the same as

$$\frac{\hat{E}^{\exp}\left(Y_1|unem = 1\right) - \hat{E}^{obs}\left(Y|unem = 1\right)}{\hat{P}^{obs}\left(D = 0|unem = 1\right)}.$$

q	Nonem_ub-Emp_lb	Emp_ub-Nonem_lb
2	-0.081 (pvalue=0.083)	0.282
10	-0.023 (pvalue=0.094)	0.221

Figure 2:

q	Smoke_ub-Non_lb	Non_ub-Smoke_lb
2	0.104	-0.032 (pvalue=0.016)
10	0.152	-0.116 (pvalue=0.019)

Figure 3:

has no health insurance coverage– they may receive Medicaid or be covered through a spouse's job.

We then repeat the analysis where unemployment status is replaced by smoking status. The hypothesis of interest is that smokers are set a higher threshold for treatment. The results are reported in the table marked "figure 3". It appears that smokers are indeed set a higher treatment threshold than non-smokers. It is conceivable that this happens due to worse "quality" of life for smokers or because they are likely to suffer from heart-attack in the future again, raising costs. But it is hard pinpoint the exact reason without further investigation.[11]

[11] Further investigations are being pursued in a separate and independent ongoing project in collaboration with Amitabh Chandra.

7 Conclusion

We have defined and analyzed the problem of detecting taste-based allocation of a binary treatment via a partial identification approach using a novel data combination method. The latter, though nonstandard, is somewhat similar in spirit to using validation data in measurement error analysis. As explained in the text, we believe that such data combination exercises should entail little additional logistical costs beyond running a field experiment and are potentially useful for evaluating treatment assignment processes that produce observational datasets in a variety of settings. Much of the treatment effects literature in econometrics has, justifiably, focused on identifying effects of the treatment from observational studies. In contrast, evaluating the treatment procols which give rise to such observational data seems to be an interesting but less-researched topic to which the present paper has attempted to contribute.

Our analysis in this paper is based on an expected utility framework. In the case of non-binary outcomes, there are alternative approaches that are worth investigating. One would be to consider the notion of loss aversion. A crucial feature of the analysis presented above is that inequalities in terms of variables that the planner observes are preserved when aggregated across unobservables– a version of the law of iterated expectations for inequalities. This feature holds for expected utilities–including the case where the sub-utility function exhibits loss aversion– but may not be shared by all the alternative criteria for treatment assignment, e.g., if the goal were to minimize the variance of outcome in the population. There are also possible extensions of the analysis that would relax the assumption of correct expectations and incorporate some form of learning by the planner. Altonji et al, 2001 consider such a case where the researcher initially has more information than the planner who is supposed to "catch up" with increasing experience if he is simply practising statistical discrimination initially. Another generalization is to consider heterogeneity in assignment protocols across different planners. On the econometric front, it is worth extending the inference methodology to the continuous outcome case, using a sieve-type approach. We leave the exploration of these issues to future research.

References

[1] Altonji, J. and C Pierret (2001): Employer learning and statistical discrimination, Quarterly Journal of Economics, 116, pp. 313-50.

[2] Andrews, D.W.K. & Gustavo Soares (2010): Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection, Econometrica, vol. 78(1), pages 119-157.

[3] Angelucci, M & G. De Giorgi (2009): "Indirect Effects of an Aid Program: How do Cash Injections Affect Ineligibles' Consumption?", American Economic Review, 99(1), pp. 486-508.

[4] Arrow, K. (1973): The theory of discrimination, in "Discrimination in labor markets", Princeton.

[5] Becker, Gary (1957): The economics of discrimination, University of Chicago Press.

[6] Bhattacharya, D. (2009): "Inferring Optimal Peer Assignment from Experimental Data", Journal of the American Statistical Association, Jun 2009, Vol. 104, No. 486: 486–500.

[7] Bhattacharya, D and A. Chandra (in progress): Detecting discrimination in treatment assignment: evidence from CASS trials for coronary artery surgery.

[8] Bhattacharya, D. & Dupas, P. (2010): Inferring Efficient Treatment Assignment under Budget Constraints", NBER working paper number 14447.

[9] CASS (1984): Circulation, Journal of American College of Cardiology, vol. 3, pp.114-128, published by the American College of Cardiology Foundation.

[10] Chen, X (2007): Large Sample Sieve Estimation of Semi-Nonparametric Models, in J.J. Heckman & E.E. Leamer (ed.) Handbook of Econometrics, Elsevier, chapter 76.

[11] Chernozhukov, V., Han Hong & E. Tamer (2007): Estimation and Confidence Regions for Parameter Sets in Econometric Models, Econometrica, Econometric Society, vol. 75(5), pages 1243-1284.

[12] Dehejia, Rajeev H (2005): Program Evaluation as a decision Problem, *Journal of Econometrics*, vol. 125, no. 1-2, pp. 141-73.

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

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

[15] Hirano, K. and J. Porter (2009): "Asymptotics for Statistical Treatment Rules", *Econometrica,* vol. 77(5), pages 1683-1701.

[16] Knowles, Persico and Todd (2001): Racial Bias in Motor Vehicle Searches: Theory and Evidence, Journal of Political Economy,109 (11) pp. 203-229.

[17] Manski, C. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, vol. 72, no. 4, pp. 1221-46.

[18] Manski, C. (2005): Social choice with partial knowledge of treatment response, Princeton University Press.

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

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

[21] Pope, D. and Sydnor (2008): What's in a Picture? Evidence of Discrimination from Prosper.com, forthcoming in Journal of Human Resources, 2010.

[22] Stoye, J. (2008): Minimax Regret Treatment Choice with Finite Sample, Journal of Econometrics, (forthcoming).

8 Appendix: Proofs

A. Derivation of (3): The solution to the problem

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

s.t.

$$\begin{aligned} \int p\left(w\right) c_1 dP_{C_1,W}\left(c_1,w\right) + \int \left\{1-p\left(w\right)\right\} c_0 dP_{C_0,W}\left(c_0,w\right) &\leq c, \text{ i.e.,} \\ E\left(C_1\right) - \int \left\{1-p\left(w\right)\right\} E\left(\Delta C|W=w\right) dF_W(w) &\leq c, \end{aligned}$$

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

$$\beta(w) \equiv \frac{E(\Delta Y|W=w)}{E(\Delta C|W=w)}; \; c = \int_{w \in \boldsymbol{w}} 1(\beta(w) \geq \gamma) \, dF_W(w).$$

Proof. Since both p and A^* satisfy the budget constraint, we must have that

$$\begin{aligned} & \int 1(\beta(w) < \gamma) \, E(\Delta C|W=w) \, dF_W(w) \\ = \; & E(C_1) - c \\ = \; & \int \{1 - p(w)\} \, E(\Delta C|W=w) \, dF_W(w). \end{aligned} \tag{21}$$

Then the welfare resulting from a generic choice of p, differs from the welfare from using A^* by

$$\begin{aligned} W(p) - W(A^*) &= \int [p(w) - 1(\beta(w) \geq \gamma)] \, \beta(w) \, E(\Delta C|W=w) \, dF_W(w) \\ &= \int 1(\beta(w) < \gamma) \, p(w) \, \beta(w) \, E(\Delta C|W=w) \, dF_W(w) \\ &\quad - \int 1(\beta(w) \geq \gamma) \{1 - p(w)\} \, \beta(w) \, E(\Delta C|W=w) \, dF_W(w) \\ &\leq \gamma \int 1(\beta(w) < \gamma) \, p(w) \, E(\Delta C|W=w) \, dF_W(w) \\ &\quad - \gamma \int 1(\beta(w) \geq \gamma) \{1 - p(w)\} \, E(\Delta C|W=w) \, dF_W(w) \end{aligned}$$

$$\begin{aligned} = \; & \underbrace{\gamma \int 1(\beta(w) < \gamma) \, E(\Delta C|W=w) \, dF_W(w)}_{=\gamma\{E(C_1)-c\}} \\ & - \gamma \int 1(\beta(w) < \gamma) \{1 - p(w)\} \, E(\Delta C|W=w) \, dF_W(w) \\ & - \underbrace{\gamma \int \{1 - p(w)\} \, E(\Delta C|W=w) \, dF_W(w)}_{=\gamma\{E(C_1)-c\}} \\ & + \gamma \int 1(\beta(w) < \gamma) \{1 - p(w)\} \, E(\Delta C|W=w) \, dF_W(w) \\ = \; & 0. \end{aligned}$$

■

B. Proof of 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 for every l:

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

This plus $S_l \geq 0$, for every l 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 $-\sum_{j=1}^{m} r_j u(a_j) = S_k/a_k < 0$. ■

Proof of proposition 2:

Define $R_k = \sum_{j=1}^{k} r_j$, for $k = 2,...m$.

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

$$\begin{aligned} -\sum_{j=1}^{m} r_j u(a_j) &= \sum_{j=1}^{m-1} r_j (u(a_m) - u(a_j)) = \sum_{j=1}^{m-1} r_j \sum_{k=j}^{m-1} (u(a_{k+1}) - u(a_k)) \\ &= \sum_{k=1}^{m-1} (u(a_{k+1}) - u(a_k)) R_k \geq 0. \end{aligned}$$

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

$$-\sum_{j=1}^{m} r_j u(a_j) = \sum_{l=1}^{m-1} (u(a_{l+1}) - u(a_l)) R_l$$

Consider a utility function of the form

$$u(a) = 0 \times 1(a \leq a_k) + \frac{a - a_k}{a_{k+1} - a_k} 1(a_k < a \leq a_k) + 1 \times 1(a > a_k).$$

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

$$u(a_{k+1}) - u(a_k) = 1 \text{ and } u(a_{l+1}) - u(a_l) = 0 \text{ for all } l \neq k$$

and therefore $-\sum_{j=1}^{m} r_j u(a_j) = \sum_{l=1}^{m-1} (u(a_{l+1}) - u(a_l)) R_l = R_k < 0$. ■

C. Proof of consistency for max entropy solution

Solution

$$v_0 = \arg\min_v Q(v) \text{ s.t. } v \geq 0,, v'1 = 1 \text{ and } \sum_{k=1}^{K} g_{jk} v_k \leq 0,\ j = 1, ...J.$$

Estimator

$$\hat{v} = \arg\min_{v} Q(v) \text{ s.t. } v \geq 0,, v'1 = 1 \text{ and } \sum_{k=1}^{K} \hat{g}_{jk} v_k \leq \delta_n,\ j = 1, ...J,$$

$\delta_n \to 0$ and $\sqrt{n}\delta_n \to \infty$.

Proof. For any $j = 1, ...J$, since $g_j' v_0 \leq 0$,

$$\begin{aligned} \Pr\left(\hat{g}_j' v_0 \leq \delta_n\right) &= \Pr\left(\left(\hat{g}_j - g_j\right)' v_0 \leq \delta_n - g_j' v_0\right) \\ &= \Pr\left(\sqrt{n}\left(\hat{g}_j - g_j\right)' v_0 \leq \sqrt{n}\delta_n + \sqrt{n}\left|g_j' v_0\right|\right) \\ &\to 1, \text{ by hypothesis.} \end{aligned}$$

Next, for any $\varepsilon > 0$, and for any $j = 1, ...J$, since $\hat{g}_j'\hat{v} \leq \delta_n$,

$$\begin{aligned} \Pr\left(g_j'\hat{v} > \varepsilon\right) &= \Pr\left(\left(g_j' - \hat{g}_j\right)'\hat{v} > \varepsilon - \hat{g}_j'\hat{v}\right) \\ &\leq \Pr\left(\left(g_j' - \hat{g}_j\right)'\hat{v} > \varepsilon - \delta_n\right) \\ &\leq \Pr\left(\left|\left(g_j' - \hat{g}_j\right)'\hat{v} + \delta_n\right| > \varepsilon\right) \to 0 \end{aligned}$$

because $\left(g_j' - \hat{g}_j\right) = o_p(1)$, $\hat{v}$ belongs to the unit simplex with probability 1 and $\delta_n \to 0$, by hypothesis. The two previous displays imply that for each j,

$$\Pr\left(\hat{g}_j' v_0 \leq \delta_n\right) \to 1 \text{ and } \Pr\left(g_j'\hat{v} \leq 0\right) \to 1.$$

Consider for some $\alpha \in (0,1)$, $\tilde{v} = \alpha\hat{v} + (1-\alpha)v_0$. Then for each j,

$$\begin{aligned} \hat{g}_j'\tilde{v} &= \alpha\hat{g}_j'\hat{v} + (1-\alpha)\hat{g}_j' v_0 \leq \delta_n \text{ w.p.a.1, by first part of previous display, and} \\ g_j'\tilde{v} &= \alpha g_j'\hat{v} + (1-\alpha) g_j' v_0 \leq 0 \text{ w.p.a.1, by second part of previous display.} \end{aligned}$$

These imply that $\tilde{v}$ belongs to the constraint set of both the population and the sample problems w.p.a.1. But by strict convexity of Q,

$$Q(\tilde{v}) < \max\left\{Q(v_0), Q(\hat{v})\right\}.$$

Therefore, if $\hat{v}$ stays away from v_0, then $\tilde{v}$ will remain distinct from both $\hat{v}$ and v_0 and therefore $Q(\tilde{v})$ will be smaller than at least one of $Q(v_0), Q(\hat{v})$, contradicting the definition of v_0 or $\hat{v}$. ■

Proof of proposition 3:

Proof. Solution to the population problem is unique due to strict convexity of $Q(\cdot)$ so that first order conditions are necessary and sufficient.

$$\begin{aligned} L &= Q(v) - \sum_{j=1}^{J} \lambda_j \left(g_j' v\right) + \sum_{k=1}^{K} \alpha_k v_k + \eta \left(1 - \sum_{k=0}^{K} v_k\right) \\ 0 &= 1 + \ln(v_k) - \sum_{j=1}^{J} \lambda_j g_{jk} - \eta \\ 1 &= \sum_k v_k = \sum_k e^{\left(\sum_{j=1}^{J} \lambda_j g_{jk}\right) + \eta - 1},\ e^{1-\eta} = \sum_k e^{\sum_{j=1}^{J} \lambda_j g_{jk}} \\ 0 &= 1 + \ln(\hat{v}_k) - \sum_{j=1}^{J} \hat{\lambda}_j \hat{g}_{jk} - \hat{\eta}. \end{aligned}$$

This implies that

$$v_k = \frac{e^{\left(\sum_{j=1}^{J} \lambda_j g_{jk}\right)}}{\sum_{l=1}^{K} e^{\left(\sum_{j=1}^{J} \lambda_j g_{jk}\right)}},\ \hat{v}_k = \frac{e^{\left(\sum_{j=1}^{J} \hat{\lambda}_j \hat{g}_{jk}\right)}}{\sum_{l=1}^{K} e^{\left(\sum_{j=1}^{J} \hat{\lambda}_j \hat{g}_{jk}\right)}}.$$

First consider the case where $J = 1$. Then $g'1 \leq 0 \Leftrightarrow v_k = \frac{1}{K}$, $\lambda = 0$. Then

$$\begin{aligned} \Pr\left(\hat{\lambda} = 0\right) &= \Pr\left(\hat{g}'1 \leq \delta_n\right) \\ &= \Pr\left(\sqrt{n}\left(\hat{g} - g\right)' 1 \leq \sqrt{n}\delta_n - \sqrt{n} g'1\right) \\ &= \Pr\left(\sqrt{n}\left(\hat{g} - g\right)' 1 \leq \sqrt{n}\delta_n + \sqrt{n}\left|g'1\right|\right) \end{aligned}$$

which converges to 1 since $\sqrt{n}\delta_n \to \infty$. Conversely, when $g'1 > 0$, $\lambda > 0$ is defined via $\sum_k g_k e^{\lambda g_k} = 0$ and $v_k = e^{\lambda g_k} / \sum_k e^{\lambda g_k}$. Then

$$\Pr\left(\hat{\lambda} = 0\right) = \Pr\left(\hat{g}'1 \leq \delta_n\right) = \Pr\left(\left(\hat{g} - g\right)' 1 - \delta_n \leq -g'1\right) \to 0$$

since $\delta_n \to 0$. So $\lambda > 0$ implies $\hat{\lambda} > 0$ w.p.a.1. So when $\lambda > 0$, we have that $\hat{g}'\hat{v} = \delta_n$ w.p.a.1.

$$1 = \sum_{k=1}^{K} \hat{v}_k = \sum_{k=1}^{K} e^{\hat{\lambda}\hat{g}_k + \hat{\eta} - 1} \Longrightarrow e^{1-\hat{\eta}} = \sum_{k=1}^{K} e^{\hat{\lambda}\hat{g}_k} \Longrightarrow \hat{v}_k = \frac{e^{\hat{\lambda}\hat{g}_k}}{\sum_{l=1}^{K} e^{\hat{\lambda}\hat{g}_l}}.$$

Then

$$\begin{aligned}
\delta_n &= \sum_{k=1}^{K} \hat{g}_k \hat{v}_k = \frac{\sum_{k=1}^{K} \hat{g}_k e^{\hat{\lambda}\hat{g}_k}}{\sum_{k=1}^{K} e^{\hat{\lambda}\hat{g}_k}} \\
&= \frac{\sum_{k=1}^{K} \hat{g}_k e^{\lambda \hat{g}_k} + \sum_{k=1}^{K} \hat{g}_k^2 e^{\lambda \hat{g}_k} \left(\hat{\lambda} - \lambda\right)}{\sum_{l=1}^{K} e^{\hat{\lambda}\hat{g}_l}}, \text{ implying} \\
\left(\hat{\lambda} - \lambda\right) &= \frac{\delta_n \sum_{l=1}^{K} e^{\hat{\lambda}\hat{g}_l} - \left(\sum_{k=1}^{K} \hat{g}_k e^{\lambda \hat{g}_k} - \sum_k g_k e^{\lambda g_k}\right)}{\sum_{k=1}^{K} \hat{g}_k^2 e^{\lambda \hat{g}_k} = O_p(1)}.
\end{aligned}$$

Rewriting,

$$\sqrt{n}\left(\hat{\lambda} - \lambda - \underbrace{\delta_n \sum_{l=1}^{K} e^{\lambda g_l}}_{bias}\right) = \frac{\underbrace{\sqrt{n}\left(\sum_{l=1}^{K} \hat{g}_l e^{\lambda \hat{g}_l} - \sum_{l=1}^{K} g_l e^{\lambda g_l}\right)}_{=O_p(1)}}{\sum_{l=1}^{K} g_l^2 e^{\lambda g_l}} + o_p(1).$$

Finally,

$$\hat{\eta} - \eta = \ln\left(\sum_{k=1}^{K} e^{\lambda g_k}\right) - \ln\left(\sum_{k=1}^{K} e^{\hat{\lambda}\hat{g}_k}\right).$$

Putting all of this together and applying the delta method,

$$\begin{aligned}
&\sqrt{n}\left\{\ln(\hat{v}_k) - \ln(v_k) - \underbrace{1(g'v = 0)\, g_k \delta_n \sum_{l=1}^{K} e^{\lambda g_l}}_{bias}\right\} \\
= \; & g_k \frac{\sqrt{n}\left(\sum_{l=1}^{K} \hat{g}_l e^{\lambda \hat{g}_l} - \sum_{l=1}^{K} g_l e^{\lambda g_l}\right)}{\sum_{l=1}^{K} g_l^2 e^{\lambda g_l}} 1(g'v = 0) \\
& - \frac{\lambda 1(g'v = 0)}{\sum_{l=1}^{K} e^{\lambda g_l}} \left[\sum_{l \neq k} e^{\lambda g_l} (\hat{g}_l - g_l) \sqrt{n}\right] + o_p(1).
\end{aligned}$$

Applying the delta method,

$$\begin{aligned}\sqrt{n}\left\{\hat{v}_1 - v_1\right\} &= v_1\sqrt{n}\left(\ln\left(\hat{v}_1\right) - \ln\left(v_1\right)\right) + \frac{v_1}{2}\sqrt{n}\left(\ln\left(\hat{v}_1\right) - \ln\left(v_1\right)\right)^2 + o_p\left(1\right) \\ &= v_1\sqrt{n}\left(\ln\left(\hat{v}_1\right) - \ln\left(v_1\right) - g_1\delta_n\sum_{l=1}^{K} e^{\lambda g_l}\right) \\ &\quad + \frac{v_1}{2}\sqrt{n}\left(\ln\left(\hat{v}_1\right) - \ln\left(v_1\right) - g_1\delta_n\sum_{l=1}^{K} e^{\lambda g_l}\right)^2 + o_p\left(1\right) \\ &\quad + v_1\sqrt{n}\delta_n g_1\sum_{l=1}^{K} e^{\lambda g_l} + O\left(\sqrt{n}\delta_n^2\right).\end{aligned}$$

So if $\sqrt{n}\delta_n^2 \to 0$, then

$$\sqrt{n}\left\{\hat{v}_i - v_i - \delta_n g_i\sum_{l=1}^{K} e^{\lambda g_l}\right\} = v_i\sqrt{n}\left(\ln\left(\hat{v}_i\right) - \ln\left(v_i\right) - g_i\delta_n\sum_{l=1}^{K} e^{\lambda g_l}\right) + o_p\left(1\right).$$

By exactly analogous arguments, when $J \geq 2$, and $g_j'v = 0$ for $j = 1, ...J_1$ and $g_j'v < 0$ for $j = J_1 + 1, ...J$, we will get that

$$\begin{aligned}&\sqrt{n}\left(\begin{bmatrix} \hat{\lambda}_1 - \lambda_1 \\ ... \\ \hat{\lambda}_{J_1} - \lambda_{J_1} \end{bmatrix} - \delta_n\left(\sum_{k=1}^{K} e^{\sum_{m=1}^{J_1}\lambda_m g_{mk}}\right)A^{-1}1_{J_1}\right) \\ &= -A^{-1}\sqrt{n}\begin{Bmatrix} \sum_{k=1}^{K}\hat{g}_{1k}e^{\sum_{m=1}^{J_1}\lambda_m\hat{g}_{mk}} - \sum_{k=1}^{K} g_{1k}e^{\sum_{m=1}^{J_1}\lambda_m g_{mk}} \\ ... \\ \sum_{k=1}^{K}\hat{g}_{J_1k}e^{\sum_{m=1}^{J_1}\lambda_m\hat{g}_{mk}} - \sum_{k=1}^{K} g_{J_1k}e^{\sum_{m=1}^{J_1}\lambda_m g_{mk}} \end{Bmatrix} + o_p\left(1\right) \\ &= -A^{-1}\sqrt{n}\sum_{k=1}^{K}\left\{\hat{g}_{\cdot k}e^{\lambda'\hat{g}_{\cdot k}} - g_{\cdot k}e^{\lambda' g_{\cdot k}}\right\} + o_p\left(1\right). \qquad (22)\end{aligned}$$

where

$$A_{J_1\times J_1} = \begin{bmatrix} \sum_k g_{1k}^2\left(e^{\sum_{m=1}^{J_1}\lambda_m g_{mk}}\right) & ... & \sum_k g_{1k}g_{J_1k}\left(e^{\sum_{m=1}^{J_1}\lambda_m g_{mk}}\right) \\ ..... & .... & .... \\ \sum_k g_{1k}g_{J_1k}\left(e^{\sum_{m=1}^{J_1}\lambda_m g_{mk}}\right) & ... & \sum_k g_{J_1k}^2\left(e^{\sum_{m=1}^{J_1}\lambda_m g_{mk}}\right) \end{bmatrix}.$$

From first-order conditions, we get that

$$\begin{aligned}
& \sqrt{n}\left(\ln \hat{v}_k - \ln v_k\right) \\
= \; & \sqrt{n}\left[\sum_{m=1}^{J_1} \hat{\lambda}_m \hat{g}_{mk} - \sum_{m=1}^{J_1} \lambda_m g_{mk} - \ln\left(\sum_{k=1}^{K} e^{\sum_{m=1}^{J_1} \hat{\lambda}_m \hat{g}_{mk}}\right) + \ln\left(\sum_{k=1}^{K} e^{\sum_{m=1}^{J_1} \lambda_m g_{mk}}\right)\right] \\
= \; & g_{\cdot k}'\left(\hat{\lambda} - \lambda\right)\sqrt{n} + \lambda'\left(\hat{g}_{\cdot k} - g_{\cdot k}\right)\sqrt{n} - \frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}}\left\{g_{\cdot l}'\left(\hat{\lambda} - \lambda\right)\sqrt{n} + \lambda'\left(\hat{g}_{\cdot l} - g_{\cdot l}\right)\sqrt{n}\right\}}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}} + o_p(1) \\
= \; & \left\{g_{\cdot k} - \frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}} g_{\cdot l}}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}}\right\}'\left(\hat{\lambda} - \lambda\right)\sqrt{n} + \lambda'\left\{\left(\hat{g}_{\cdot k} - g_{\cdot k}\right)\sqrt{n} - \frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}}\left(\hat{g}_{\cdot l} - g_{\cdot l}\right)\sqrt{n}}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}}\right\} + o_p(1).
\end{aligned}$$

Plugging in (22), we get

$$\begin{aligned}
& \sqrt{n}\left(\ln \hat{v}_k - \ln v_k - \delta_n\left\{g_{\cdot k} - \frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}} g_{\cdot l}}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}}\right\}'\left(\sum_{k=1}^{K} e^{\lambda' g_{\cdot k}}\right) A^{-1} 1_{J_1}\right) \\
= \; & -\left\{g_{\cdot k} - \frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}} g_{\cdot l}}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}}\right\}' A^{-1}\sqrt{n}\sum_{k=1}^{K}\left\{\hat{g}_{\cdot k} e^{\lambda' \hat{g}_{\cdot k}} - g_{\cdot k} e^{\lambda' g_{\cdot k}}\right\} \\
& + \lambda'\sqrt{n}\left\{\left(\hat{g}_{\cdot k} - g_{\cdot k}\right) - \frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}}\left(\hat{g}_{\cdot l} - g_{\cdot l}\right)}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}}\right\} + o_p(1)
\end{aligned}$$

Since $g_{j\cdot}' v = 0$ for $j = 1, ..J_1$, we have that $\frac{\sum_{l=1}^{K} e^{\lambda' g_{\cdot l}} g_{\cdot l}}{\sum_{l=1}^{l} e^{\lambda' g_{\cdot l}}} = 0$. Hence the result. ■

Inference: Consistency of test

$H_0 : g'v^0 = 0$ vs $H_1 : g'v^0 < 0$. Reject if $\hat{g}'\hat{v} \leq a_n$. Then

$$\begin{aligned}
& \Pr\left(\hat{g}'\hat{v} \leq a_n | g'v^0 = 0\right) \\
= \; & \Pr\left(\sqrt{n}\left(\hat{g} - g\right)'\hat{v} + \sqrt{n} g'\left(\hat{v} - v_0\right) \leq -c_n\sqrt{n} | g'v^0 = 0\right) \\
= \; & \Pr\left(\begin{array}{c} \sqrt{n}\left(\hat{g} - g\right)'\hat{v} + \sqrt{n} g'\left(\hat{v} - v_0 - \left(\begin{array}{c} g_1 v_1 \\ ... \\ g_k v_k \end{array}\right)\delta_n \sum_{l=1}^{K} e^{\lambda g_l}\right) \\ \leq a_n\sqrt{n} - \sqrt{n}\delta_n g'\left(\left(\begin{array}{c} g_1 v_1 \\ ... \\ g_k v_k \end{array}\right)\sum_{l=1}^{K} e^{\lambda g_l}\right) | g'v^0 = 0 \end{array}\right) \to 0
\end{aligned}$$

if $a_n = o(\delta_n)$. So by choosing a_n to be of smaller order than δ_n, we will get a consistent test of $g'v^0 = 0$.

D. Nonidentical distributions

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. This may be suspected in the CASS data where the incidence of prior heart attack and incidence of diabetes are slightly larger (significant but with numerically small difference in point-estimates) in the experimental group. 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, $E^{\exp}(C_0|x) \geq E^{obs}(C_0|x)$ and $E^{\exp}(C_1|x) \geq E^{obs}(C_1|x)$. Using the same steps as those leading to (8), 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 still be used to detect misallocation. For instance, if it is the case that

$$\begin{aligned} &\frac{E^{obs}(Y_1|D=1,male) - \bar{E}(Y_0|D=1,male)}{E^{obs}(C_1|D=1,male) - \bar{E}(C_0|D=1,male)} \\ \leq\ &\frac{\bar{E}(Y_1|D=0,female) - E^{obs}(Y_0|D=0,female)}{\bar{E}(C_1|D=0,female) - E^{obs}(C_0|D=0,female)}, \end{aligned} \tag{23}$$

then it follows that

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

Thus, females are facing a larger threshold relative to males. However, since (23) implies (11), it will be harder to detect misallocation here compared to when the experimental and observational data came from identical populations.