May 2009
Revised January 16, 2010

BAYESIAN ASPECTS OF TREATMENT CHOICE

Gary Chamberlain
Harvard University

ABSTRACT

This paper considers an individual making a treatment choice. The individual has access to data on other individuals, with values for a list of characteristics, treatment assignments, and outcomes. The individual knows his value for the list of characteristics. The goal is to use this data set to guide his treatment choice. The role of treatment assignment is developed, and how it affects the specification of prior distributions. The likelihood function is the same for random assignment and for selection on observables, but the prior distributions differ. A question here is whether there is a value in knowing the propensity score. The propensity score does not appear in the likelihood, but it does appear in the prior distribution. So there is a value to knowing the propensity score if the prior is not dominated by the data. In particular, the list of measured characteristics may be of high dimension, and the paper considers prior distributions that may be effective in this case.

The paper also considers selection on unobservables and the use of instrumental variables. The prior distribution is not dominated by the data. We make a particular suggestion, in which the undominated part of the prior shows up in the choice of a functional form, which is then combined with a maximum-likelihood approximation to obtain a decision rule. We discuss the role of extrapolation in this decision rule by making a connection with compliers, always-takers, and never-takers in the local average treatment effect developed by Imbens and Angrist (1994).

May 2009
Revised January 16, 2010

BAYESIAN ASPECTS OF TREATMENT CHOICE

1. INTRODUCTION

An individual is making a choice between two treatments. Data are available in which other individuals have received one of the treatments and an outcome was recorded. In addition, the data set contains various characteristics of the individuals. For example, the choice could be between two medical treatments. The characteristics could include gender, age, body mass index, blood pressure, lipid profile, smoking history, medical history, and some information on medical history of relatives. The individual making the choice knows his values on all of these variables.

We shall consider three types of data. They differ in how the treatments were assigned to the individuals in the data set. The first case is simple random assignment. For example, there could be a clinical trial in which a group of subjects is selected and then a coin flip determines whether an individual is assigned treatment 0 or treatment 1. A central question here is the use of the possibly extensive data on individual characteristics. One possibility is that the decision maker look only at the subset of the data that exactly matches his values on these characteristics. There may, however, be only a few such matches, even if the data set has a large number of individuals.

In the second type of data, the assignment probability may depend upon individual characteristics, and we work with the assumption of random assignment conditional on the measured characteristics. Within a group of individuals with the same measured characteristics, the assumption is that assignment is as if determined by a coin flip, where the probability of heads may depend upon the measured characteristics. This could correspond to an observational study in which data are collected on characteristics, treatment, and outcome for a group of individuals. This conditional assignment probability was called "propensity score" by Rosenbaum and Rubin (1983). They showed that random assignment conditional on the measured characteristics implies random

assignment conditional on the propensity score. Suppose that the propensity score is known. If the object of interest is an average treatment effect, averaging over the measured characteristics, then the Rosenbaum-Rubin result suggests the possibility of a simpler analysis, in which one conditions only on the propensity score and not on the full set of measured characteristics. But the counterpart of a treatment effect for our decision maker is a conditional treatment effect, conditional on the measured characteristics being equal to her values, so there is a question of what role, if any, a known propensity score should play. A related issue is that it may be appropriate to do the analysis conditional on treatment assignment and on the measured characteristics. This analysis can resemble a classical regression problem, where one might argue that the distribution of the regressors is irrelevant, perhaps appealing to a notion of ancillarity. Then the joint distribution of assignment and measured characteristics would not be relevant, and so the propensity score, which gives the assignment distribution conditional on measured characteristics, would not be relevant. We shall be working in a likelihood framework, and the role of the propensity score is a central question for us.

In the third type of data, treatment assignment depends upon unobservables, but an instrumental variable is available. In a clinical trial, for example, an intended treatment could be determined by simple random assignment, but individuals may not comply with the intended treatment. Then the intended treatment could serve as an instrumental variable. A key issue here is the lack of identification, and how to deal with it in the context of a decision maker who has to make a choice.

Now I shall introduce some notation and describe the problem in more detail. Data are available on N individuals. For individual i, we observe a vector of discrete characteristics, which is coded as $X_i \in \{1, \ldots, K\}$. There is assignment to one of two treatments: $D_i \in \{0, 1\}$. There is a discrete outcome, which is coded as $Y_i \in \{1, \ldots, J\}$. Let $Z_i = (X_i, D_i, Y_i)$ and let $Z = (Z_1, \ldots, Z_N)$. So Z is observed.

An individual, call him α, needs to choose between $D_\alpha = 0$ and $D_\alpha = 1$. This individual knows his value for the characteristics X_α. Let $Y_{\alpha 0}$ denote the outcome if $D_\alpha = 0$, and let $Y_{\alpha 1}$ denote the outcome if $D_\alpha = 1$. The uncertainty the individual faces is over the values of the decision outcomes

$Y_{\alpha 0}$ and $Y_{\alpha 1}$. The goal of this paper is to provide guidance on how to advise the individual on making his choice.

We shall work in an expected utility framework:

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad E[u_1(Y_{\alpha 1}) \,|\, Z] > E[u_0(Y_{\alpha 0}) \,|\, Z]. \tag{1}$$

The notation allows for additional consequences of the choice that are known to the decision maker, such as costs c_0 and c_1. Then we could have $u_l(\cdot) = u(\cdot, c_l)$ for $l = 0, 1$. We shall take the utility function $(u_0(\cdot), u_1(\cdot))$ for α as given, and focus on the expectation. So we need to construct conditional distributions for $Y_{\alpha 0}$ and for $Y_{\alpha 1}$, conditional on the observation Z.

An immediate issue is that we do not observe (Y_{i0}, Y_{i1}) for the individuals in the data set $(i = 1, \ldots, N)$. Suppose that i is sufficiently similar to α so that we can think about i making a choice between the two treatments, even though the actual assignment of D_i may not have been through such a choice. For example, the actual value of D_i may have been randomly assigned. Define Y_{i0} as the outcome if i chooses treatment 0, and Y_{i1} as the outcome if i chooses treatment 1. We think about these decision outcomes in the same way that we think about $Y_{\alpha 0}$ and $Y_{\alpha 1}$ in the decision-maker's problem. Now it becomes a key assumption that if $D_i = 0$, then $Y_i = Y_{i0}$, regardless of how this assignment of $D_i = 0$ came to be. So we are saying that the outcome that is observed when i is assigned to treatment 0 is the same as the outcome that would be observed if i had chosen treatment 0. With a corresponding assumption for treatment 1, we have

$$Y_i = (1 - D_i)Y_{i0} + D_i Y_{i1}.$$

To appreciate the force of these assumptions, suppose that immediately after treatment assignment, individual i can at some cost change the assignment. This is not relevant for unconstrained choice, but if i is assigned to treatment 0 when he would have chosen treatment 1, then this option may become relevant and the observed outcome may differ from Y_{i0}. More generally, if i is assigned a treatment that differs from what he would have chosen, then various actions may become relevant, even though they are not relevant for the unconstrained choice of the decision-maker α.

In the literature on evaluating treatment effects, the term "potential outcome" is often used, in the sense that if $D_i = 0$, the problem is missing data on the potential outcome that would have been observed under $D_i = 1$. Likewise, if $D_i = 1$, the problem is missing data on the potential outcome that would have been observed under $D_i = 0$. We could refer to Y_{i0} and Y_{i1} as potential outcomes, but I want to stress that the problem, if $D_i = 0$, is not just that Y_{i1} is not observed. It is a key assumption that the decision outcome, Y_{i0}, defined to correspond to $Y_{\alpha 0}$ for the decision maker, is observed under $D_i = 0$. Likewise, it is a key assumption that the decision outcome Y_{i1} is observed under $D_i = 1$.

Section 2 develops a likelihood function for the case of random assignment conditional on observed characteristics, and then makes the stronger assumption of simple random assignment in setting up prior distributions. If the number of values K for X_i is large, with a small number of observations in a typical $X_i = k$ cell, then the prior distribution plays an important role and is not dominated by the data. Section 3 relaxes the assumption of simple random assignment, maintaining selection on observables. This does not affect the likelihood function but does affect the prior distributions. A question here is whether there is a value in knowing the propensity score. When we condition on X and D, the propensity score does not appear in the likelihood, but it does appear in the prior distribution. So there is a value to knowing the propensity score if the prior is not dominated by the data. This would be the case if the number of values for X_i is large, with a small number of observations in a typical $X_i = k$ cell.

Section 4 considers selection on unobservables. We simplify notation by dropping the X variable. Either there are no observable characteristics for the data individuals or we work with a subset that matches the decision maker. There is, however, an additional vector of discrete variables, which is coded as $W_i \in \{1, \ldots, M\}$. W_i plays the role of an instrumental variable. The prior distribution is not dominated by the data. We make a particular suggestion, in which the undominated part of the prior shows up in the choice of a functional form, which is then combined with a maximum-likelihood approximation to obtain a decision rule. We discuss the role of extrapolation in this decision rule by making a connection with compliers, always-takers, and

never-takers in the local average treatment effect developed by Imbens and Angrist (1994).

Section 5 makes connections with the literature.

2. SIMPLE RANDOM ASSIGNMENT

Let $Z_i^* = (X_i, D_i, Y_{i0}, Y_{i1})$. Suppose that the label i conveys no information, so that the joint distribution of $(Z_1^*, \ldots, Z_N^*)$ is exchangeable. If this is to hold for arbitrary N, then conditional on some distribution F^*, the Z_i^* are independent and identically distributed according to F^*. We can decompose F^* into a distribution for (Y_{i0}, Y_{i1}) conditional on (X_i, D_i), a distribution for D_i conditional on X_i, and a marginal distribution for X_i. We shall condition throughout on $X = (X_1, \ldots, X_N)$, so the marginal distribution for X_i will not play a role. The observation is $Z = (Z_1, \ldots, Z_N)$, with $Z_i = (X_i, D_i, Y_i)$. In order to form a likelihood function for the observation Z, we do not need the joint distribution of (Y_{i0}, Y_{i1}) conditional on (X_i, D_i), just the two margins: Y_{i0} conditional on (X_i, D_i) and Y_{i1} conditional on (X_i, D_i). Because the distributions are discrete, we can use the following notation:

$$\Pr(Y_{i0} = j \,|\, X_i = x_i, D_i = d_i; \pi, \eta) = \pi_{0j}(x_i, d_i),$$
$$\Pr(Y_{i1} = j \,|\, X_i = x_i, D_i = d_i; \pi, \eta) = \pi_{1j}(x_i, d_i) \qquad (j = 1, \ldots, J),$$

where the functions π_{0j} and π_{1j} map $\{1, \ldots, K\} \times \{0, 1\}$ to the interval $[0, 1]$ and satisfy

$$\sum_{j=1}^{J} \pi_{0j}(k, d) = 1, \quad \sum_{j=1}^{J} \pi_{1j}(k, d) = 1 \qquad (k = 1, \ldots, K;\ d = 0, 1).$$

Our notation for the distribution of D_i conditional on X_i is

$$\Pr(D_i = 1 \,|\, X_i = x_i; \pi, \eta) = 1 - \Pr(D_i = 0 \,|\, X_i = x_i; \pi, \eta) = \eta(x_i),$$

where the function η maps $\{1, \ldots, K\}$ to the interval $[0, 1]$. So the parameter space is $\Theta = \Theta_1 \times \Theta_2$, with

$$\Theta_1 = \{\pi : \pi_l(k, d) = (\pi_{l1}(k, d), \ldots, \pi_{lJ}(k, d)) \in \mathcal{S}_{J-1};\ l = 0, 1;\ k = 1, \ldots, K;\ d = 0, 1\} = \mathcal{S}_{J-1}^{4K},$$
$$\Theta_2 = \{\eta : \eta(k) \in [0, 1],\ k = 1, \ldots, K\} = [0, 1]^K,$$

where $\mathcal{S}_{J-1}$ is the unit simplex of dimension $J-1$ in $\mathcal{R}^J$.

2.1 *Likelihood Function*

Let z denote the realization of the random variable Z, with $z_i = (x_i, d_i, y_i)$ and $z = (z_1, \ldots, z_N)$, and let $\theta = (\pi, \eta)$. The likelihood function for the observation Z is

$$\begin{aligned}
f_{Z\,|\,X}(z\,|\,x;\theta) &= \Pr(Z = z\,|\,X = x;\theta) \\
&= \prod_{i=1}^{N} \Pr(Y_i = y_i\,|\,X_i = x_i, D_i = d_i;\pi) \cdot \Pr(D_i = d_i\,|\,X_i = x_i;\eta) \\
&= \prod_{i=1}^{N}\Bigg(\prod_{k=1}^{K}\prod_{j=1}^{J} \pi_{0j}(k,0)^{1(d_i=0)1(x_i=k)1(y_i=j)}\pi_{1j}(k,1)^{1(d_i=1)1(x_i=k)1(y_i=j)} \\
&\qquad \times \prod_{k=1}^{K}[1-\eta(k)]^{1(d_i=0)1(x_i=k)}\eta(k)^{1(d_i=1)1(x_i=k)}\Bigg) \\
&= \prod_{k=1}^{K}\prod_{j=1}^{J} \pi_{0j}(k,0)^{n(0,k,j)}\pi_{1j}(k,1)^{n(1,k,j)} \times \prod_{k=1}^{K}(1-\eta(k))^{n(0,k)}\eta(k)^{n(1,k)} \\
&= f_{Y\,|\,X,D}(y\,|\,x,d;\pi) f_{D\,|\,X}(d\,|\,x;\eta), \qquad (2)
\end{aligned}$$

where

$$n(0,k,j) = \sum_{i=1}^{N} 1(d_i = 0)1(x_i = k)1(y_i = j), \quad n(1,k,j) = \sum_{i=1}^{N} 1(d_i = 1)1(x_i = k)1(y_i = j),$$

$$n(0,k) = \sum_{i=1}^{N} 1(d_i = 0)1(x_i = k) = \sum_{j=1}^{J} n(0,k,j), \quad n(1,k) = \sum_{i=1}^{N} 1(d_i = 1)1(x_i = k) = \sum_{j=1}^{J} n(1,k,j).$$

The value of X_α for the decision maker is τ. Suppose that the following sequence of random variables is exchangeable:

$$((Y_{\alpha 0}, Y_{\alpha 1}),\, (Y_{i0}, Y_{i1}) : X_i = \tau).$$

Then the F^* distribution of Y_{i0} conditional on $X_i = \tau$ is relevant for the decision maker, and we shall assume that:

$$\begin{aligned}
\Pr(Y_{\alpha 0} = j\,|\,X_\alpha = \tau;\theta) &= \Pr(Y_{i0} = j\,|\,X_i = \tau;\theta) \\
&= (1-\eta(\tau))\pi_{0j}(\tau,0) + \eta(\tau)\pi_{0j}(\tau,1).
\end{aligned}$$

Likewise,

$$\begin{aligned}\Pr(Y_{\alpha 1} = j \,|\, X_\alpha = \tau; \theta) &= \Pr(Y_{i1} = j \,|\, X_i = \tau; \theta) \\ &= (1 - \eta(\tau))\pi_{1j}(\tau, 0) + \eta(\tau)\pi_{1j}(\tau, 1).\end{aligned}$$

Then, conditional on θ, the decision rule is to choose $D_\alpha = 1$ if

$$\sum_{j=1}^{J} u_1(j)[(1 - \eta(\tau))\pi_{1j}(\tau, 0) + \eta(\tau)\pi_{1j}(\tau, 1)] > \sum_{j=1}^{J} u_0(j)[(1 - \eta(\tau))\pi_{0j}(\tau, 0) + \eta(\tau)\pi_{0j}(\tau, 1)]. \quad (3)$$

We need to obtain a distribution on Θ conditional on the observation Z, in order to go from (3) to a decision rule that conditions only on the observation, as in (1).

Note that the likelihood function depends upon π only through $(\pi_{0j}(k, 0), \pi_{1j}(k, 1))$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K$. So there is no direct information in the data on the terms $\pi_{0j}(\tau, 1)$ and $\pi_{1j}(\tau, 0)$ in (3). A tractable special case restricts the F^* distribution so that the treatment assignment D_i is independent of the decision outcomes (Y_{i0}, Y_{i1}) conditional on the measured characteristics X_i. In that case, we have

$$\begin{aligned}&\pi_{0j}(k, 0) = \pi_{0j}(k, 1) \equiv \pi_{0j}(k), \quad (4)\\ &\pi_{1j}(k, 0) = \pi_{1j}(k, 1) \equiv \pi_{1j}(k) \qquad (j = 1, \ldots, J;\ k = 1, \ldots, K),\end{aligned}$$

and

$$\Theta_1 = \{\pi : \pi_l(k) = (\pi_{l1}(k), \ldots, \pi_{lJ}(k)) \in \mathcal{S}_{J-1};\ l = 0, 1;\ k = 1, \ldots, K\} = \mathcal{S}_{J-1}^{2K}. \quad (5)$$

Now the decision rule in (3) becomes: conditional on θ,

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j)\pi_{1j}(\tau) > \sum_{j=1}^{J} u_0(j)\pi_{0j}(\tau). \quad (3')$$

In order to examine the assumption in (4), suppose that

$$D_i = 1(E_i[u_1(Y_{i1})] > E_i[u_0(Y_{i0})]),$$

where the operator E_i provides the expectation with respect to the personal (subjective) distribution of individual i. The assumption in (4) will hold if $(E_i[u_0(Y_{i0})], E_i[u_1(Y_{i1})])$ is independent of (Y_{i0}, Y_{i1}) conditional on X_i and π. For example, we could have

$$D_i = 1(\sum_{j=1}^{J} u_1(j)\pi_{1j}(X_i) > \sum_{j=1}^{J} u_0(j)\pi_{0j}(X_i)).$$

More generally, (4) will hold if the information available to individual i is independent of (Y_{i0}, Y_{i1}) conditional on X_i and π. The assumption in (4) is commonly referred to as "random assignment conditional on X" or "selection on observables." We shall use those terms, but note that although X is observable, we are also conditioning on π. This conditioning on π will play an important role when we discuss prior distributions.

2.2 *Limited Information*

Let $\pi_l(k) = (\pi_{l1}(k), \ldots, \pi_{lJ}(k))$ and $\pi_l = (\pi_l(1), \ldots, \pi_l(K))$ for $l = 0, 1$. The task of specifying a prior distribution will be easier if we can work with the marginal distributions for π_0 and π_1 without specifying the joint distribution for $\pi = (\pi_0, \pi_1)$. We can do this by adopting a limited information approach. Let

$$Y_i^{(0)} = \begin{cases} Y_i, & \text{if } D_i = 0; \\ \text{missing}, & \text{if } D_i = 1 \end{cases}$$

and

$$Y_i^{(1)} = \begin{cases} \text{missing}, & \text{if } D_i = 0; \\ Y_i, & \text{if } D_i = 1 \end{cases}.$$

Let $Z_i^{(l)} = (X_i, D_i, Y_i^{(l)})$ and let $Z^{(l)} = (Z_1^{(l)}, \ldots, Z_N^{(l)})$ for $l = 0, 1$. We shall condition on $Z^{(0)}$ in forming a (predictive) distribution for $Y_{\alpha 0}$ and condition on $Z^{(1)}$ in forming a (predictive) distribution for $Y_{\alpha 1}$. So our limited information decision rule is

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad E[u_1(Y_{\alpha 1}) \,|\, Z^{(1)}] > E[u_0(Y_{\alpha 0}) \,|\, Z^{(0)}]. \tag{1'}$$

The likelihood function for $Z^{(0)}$ is

$$f_{Z^{(0)} \,|\, X}(z^{(0)} \,|\, x; (\pi_0, \eta)) = \Pr(Z^{(0)} = z^{(0)} \,|\, X = x; (\pi_0, \eta))$$

$$= \prod_{k=1}^{K} \prod_{j=1}^{J} \pi_{0j}(k)^{n(0,k,j)} \times \prod_{k=1}^{K} (1-\eta(k))^{n(0,k)} \eta(k)^{n(1,k)}$$

$$= f_{Y^{(0)} \,|\, X,D}(y^{(0)} \,|\, x, d; \pi_0) f_{D \,|\, X}(d \,|\, x; \eta), \tag{6}$$

and the likelihood function for $Z^{(1)}$ is

$$f_{Z^{(1)} \,|\, X}(z^{(1)} \,|\, x; (\pi_1, \eta)) = \Pr(Z^{(1)} = z^{(1)} \,|\, X = x; (\pi_1, \eta))$$

$$= \prod_{k=1}^{K} \prod_{j=1}^{J} \pi_{1j}(k)^{n(1,k,j)} \times \prod_{k=1}^{K} (1-\eta(k))^{n(0,k)} \eta(k)^{n(1,k)}$$

$$= f_{Y^{(1)} \,|\, X,D}(y^{(1)} \,|\, x, d; \pi_1) f_{D \,|\, X}(d \,|\, x; \eta). \tag{7}$$

Next we shall develop prior distributions that allow us to go from (3′) to decision rules that depend only on the observation Z and the prior distribution.

2.3 *Prior Distributions*

We shall begin with a prior distribution that leads to a closed-form expression for the decision rule. Impose the restriction in (4) on the likelihood function, corresponding to random assignment conditional on X. Let

$$\Theta_{10} = \{\pi_0 : \pi_0(k) = (\pi_{01}(k), \ldots, \pi_{0J}(k)) \in \mathcal{S}_{J-1};\ k = 1, \ldots, K\} = \mathcal{S}_{J-1}^K,$$

$$\Theta_{11} = \{\pi_1 : \pi_1(k) = (\pi_{11}(k), \ldots, \pi_{1J}(k)) \in \mathcal{S}_{J-1};\ k = 1, \ldots, K\} = \mathcal{S}_{J-1}^K,$$

so that $\Theta_1 = \Theta_{10} \times \Theta_{11}$. Let T_0 denote a random variable that has the prior distribution on Θ_{10} and let T_1 denote a random variable that has the prior distribution on Θ_{11}. We shall use the limited information approach, so that we only specify the marginal distributions for T_0 and for T_1, and do not specify a joint distribution for $T = (T_0, T_1)$. Let $X = (X_1, \ldots, X_N)$, $D = (D_1, \ldots, D_N)$, and $Y^{(0)} = (Y_1^{(0)}, \ldots, Y_N^{(0)})$. We shall work with the distribution of $Y^{(0)}$ conditional on (X, D), so the distribution of D conditional on X will not play a role, and we can set the parameter space equal to Θ_{10}. We shall combine the distribution of $Y^{(0)}$ conditional on $(X = x, D = d; T_0 = \pi_0)$ with a distribution for T_0 conditional on $(X = x, D = d)$, to obtain a distribution for T_0 conditional on

$(X = x, D = d, Y^{(0)} = y^{(0)})$. Likewise, we shall combine the distribution of $Y^{(1)}$ conditional on $(X = x, D = d; T_1 = \pi_1)$ with a distribution for T_1 conditional on $(X = x, D = d)$, to obtain a distribution for T_1 conditional on $(X = x, D = d, Y^{(1)} = y^{(1)})$.

The distributions for T_0 and T_1 conditional on $(X = x, D = d)$ are restricted to not depend upon (x, d). This corresponds to simple random assignment. The assumption that (Y_{i0}, Y_{i1}) is independent of D_i conditional on X_i is implicitly conditioning on $T = \pi$. If we do not condition on $T = \pi$, then D_i and (Y_{i0}, Y_{i1}) can fail to be independent conditional on X_i because D helps to predict T. For example, if

$$D_i = 1(\sum_{j=1}^{J} u_1(j) T_{1j}(X_i) > \sum_{j=1}^{J} u_0(j) T_{0j}(X_i)),$$

then (Y_{i0}, Y_{i1}) is independent of D_i conditional on $(X = x, T = \pi)$, but D_i is not independent of $T(k)$ conditional on $X_i = k$.

We shall assume simple random assignment, so that D is independent of T_l conditional on X for $l = 0, 1$, and we assume in addition that T_l is independent of X. With

$$\Pr(D_i = 1 \,|\, X_i = x_i; \pi, \eta) = 1 - \Pr(D_i = 0 \,|\, X_i = x_i; \pi, \eta) = \eta(x_i),$$

the key is that the randomization probabilities $(\eta(1), \ldots, \eta(K))$ are fixed by design in such a way that the decision maker is confident in assessing T_l independent of η in his personal distribution; for example, $\eta(k) = 1/2$ for $k = 1, \ldots, K$. The distribution of T_l has density p_l which specifies that $(T_l(k) : k = 1, \ldots, K)$ are mutually independent with distributions in the Dirichlet family, where $T_l(k) = (T_{l1}(k), \ldots, T_{lJ}(k))$ for $l = 0, 1$:

$$p_l(\pi_l \,|\, x, d; \beta_l) = p_l(\pi_l \,|\, \beta_l) = \prod_{k=1}^{K} h_{\mathrm{Dir}}(\pi_{l1}(k), \ldots, \pi_{lJ}(k) \,|\, \beta_{l1}(k), \ldots, \beta_{lJ}(k)), \tag{8}$$

where $\beta_{lj}(k) > 0$ and $h_{\mathrm{Dir}}(\cdot \,|\, \zeta)$ is the Dirichlet density with parameter ζ:

$$h_{\mathrm{Dir}}(w_1, \ldots, w_J \,|\, \zeta_1, \ldots, \zeta_J) = \frac{\Gamma(\sum_{j=1}^{J} \zeta_j)}{\prod_{j=1}^{J} \Gamma(\zeta_j)} \prod_{j=1}^{J} w_j^{(\zeta_j - 1)},$$

for $(w_1, \ldots, w_J)$ in the simplex $\mathcal{S}_{J-1}$ and $\zeta_j > 0$.

The conditional density of T_l given $Z^{(l)} = z^{(l)}$ is

$$\bar{p}_l(\pi_l \,|\, z^{(l)}; \beta_l)$$
$$= f_{Y^{(l)} \,|\, X,D}(y^{(l)} \,|\, x, d; \pi_l) p_l(\pi_l \,|\, x, d; \beta_l) \Big/ \int_{\Theta_{1l}} f_{Y^{(l)} \,|\, X,D}(y^{(l)} \,|\, x, d; \pi_l) p_l(\pi_l \,|\, x, d; \beta_l)\, d\pi_l.$$

Inspecting the product of $f_{Y^{(l)} \,|\, X,D}(y^{(l)} \,|\, x, d; \pi_l)$ from (6) and (7) with $p_l(\pi_l \,|\, x, d; \beta_l)$ from (8) shows that the conditional density is a product of Dirichlet densities:

$$\bar{p}_l(\pi_l \,|\, z^{(l)}; \beta_l) = \prod_{k=1}^{K} h_{\mathrm{Dir}}(\pi_{l1}(k), \ldots, \pi_{lJ}(k) \,|\, \bar{\beta}_{l1}(k), \ldots, \bar{\beta}_{lJ}(k)) \qquad (l = 0, 1), \tag{9}$$

where

$$\bar{\beta}_{lj}(k) = \beta_{lj}(k) + n(l, k, j) \quad \text{with} \quad n(l, k, j) = \sum_{i=1}^{N} 1(d_i = l) 1(x_i = k) 1(y_i = j).$$

Applying iterated expectations, we can go from the decision rule in $(3')$ to the limited information rule in $(1')$, that depends only upon the observation and the prior distribution:

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j) E[T_{1j}(\tau) \,|\, Z^{(1)} = z^{(1)}; \beta_1] > \sum_{j=1}^{J} u_0(j) E[T_{0j}(\tau) \,|\, Z^{(0)} = z^{(0)}; \beta_0]. \tag{10}$$

If $(W_1, \ldots, W_J)$ has a Dirichlet distribution with parameter $(\zeta_1, \ldots, \zeta_J)$, then $E(W_j) = \zeta_j / (\zeta_1 + \ldots + \zeta_J)$. So evaluating the conditional expectations in (10) gives the decision rule:

$$\text{choose } D_\alpha = 1 \text{ if}$$
$$\sum_{j=1}^{J} u_1(j) \frac{\beta_{1j}(\tau) + n(1, \tau, j)}{\sum_{j=1}^{J} [\beta_{1j}(\tau) + n(1, \tau, j)]} > \sum_{j=1}^{J} u_0(j) \frac{\beta_{0j}(\tau) + n(0, \tau, j)}{\sum_{j=1}^{J} [\beta_{0j}(\tau) + n(0, \tau, j)]}. \tag{11}$$

Note that this decision rule uses only the data on the subset of individuals with $X_i = \tau$—the individuals who exactly match the decision maker α on the measured characteristics. This aspect

of the decision rule arises from the prior specification that $\pi_0(k)$ are mutually independent for $k = 1, \ldots, K$ and, likewise, that $\pi_1(k)$ are mutually independent. We shall consider an alternative prior distribution below that relaxes this independence.

A potential approximation to the decision rule in (11) is

$$\text{choose } D_\alpha = 1 \text{ if}$$
$$\sum_{j=1}^{J} u_1(j) \frac{n(1,\tau,j)}{\sum_{j=1}^{J} n(1,\tau,j)} > \sum_{j=1}^{J} u_0(j) \frac{n(0,\tau,j)}{\sum_{j=1}^{J} n(0,\tau,j)}. \tag{12}$$

Here the conditional probabilities that $Y_{0\alpha} = j$ and $Y_{1\alpha} = j$ in (3′) are replaced by sample frequencies. For a given utility function (u_0, u_1), the approximation in (12) coincides with the rule in (11) if $\beta_{lj}(\tau)/n(l,\tau,j)$ is sufficiently small ($l = 0, 1$; $j = 1, \ldots, J$). On the other hand, for given values of $\beta_l(\tau)/n(l,\tau) = \sum_{j=1}^{J} \beta_{lj}(\tau) / \sum_{j=1}^{J} n(l,\tau,j)$, there can, depending on the utility function, be extreme sensitivity to the prior distribution. Suppose, for example, that $n(0,\tau) = n(1,\tau) \equiv n(\tau)$, so that there are the same number of observations with $D_i = 0$ and with $D_i = 1$ in the cell with $X_i = \tau$. Then

$$\frac{\bar{\beta}_{1j}(\tau)}{\bar{\beta}_1(\tau)} - \frac{\bar{\beta}_{0j}(\tau)}{\bar{\beta}_0(\tau)} = \frac{1}{n(\tau)} \left(\frac{\beta_{1j}(\tau) + n(1,\tau,j)}{\frac{\beta_1(\tau)}{n(\tau)} + 1} - \frac{\beta_{0j}(\tau) + n(0,\tau,j)}{\frac{\beta_0(\tau)}{n(\tau)} + 1} \right).$$

If $\beta_0(\tau)/n(\tau)$ and $\beta_1(\tau)/n(\tau)$ are sufficiently small, then the sign of this term is determined by the sign of

$$\beta_{1j}(\tau) - \beta_{0j}(\tau) + n(1,\tau,j) - n(0,\tau,j) \tag{13}$$

(provided that (13) is nonzero). If there are no observations with $X_i = \tau$ and $Y_i = j$, then $n(0,\tau,j) = n(1,\tau,j) = 0$ and, if $\beta_{0j}(\tau) \neq \beta_{1j}(\tau)$, the sign is determined by

$$\beta_{1j}(\tau) - \beta_{0j}(\tau).$$

If the absolute value of $u_1(j)$ is sufficiently large relative to $|u_1(j) - u_0(j)|$, then this sign will determine whether the decision rule in (11) chooses $D_\alpha = 1$ or $D_\alpha = 0$. This could correspond to a rare but catastrophic event.

The prior distribution we have been using specifies that $(\pi_l(1), \ldots, \pi_l(K))$ are mutually independent conditional on β (for $l = 0, 1$). We can relax this independence by following Good (1965, p. 28) in putting a prior distribution on the Dirichlet parameter β. A simple version restricts $\beta_{lj}(k)$ to be constant across k: $\beta_{lj}(k) = \beta_{lj}$. Let $\beta_l = (\beta_{l1}, \ldots, \beta_{lJ})$ for $l = 0, 1$. Following our limited information approach, we shall only need the marginal distribution for β_l and not the joint distribution for (β_1, β_2). As β_l varies, we generate a set of distributions for $Y^{(l)}$ conditional on $(X = x, D = d)$. The densities of these conditional distributions form a Type II likelihood function (in Good's terminology) for β_l:

$$
\begin{aligned}
g_{Y^{(l)} \mid X,D}(y^{(l)} \mid x, d; \beta_l) &= \int_{\Theta_{1l}} f_{Y^{(l)} \mid X,D}(y^{(l)} \mid x, d; \pi_l) p_l(\pi_l \mid x, d; \beta_l)\, d\pi_l \qquad (14) \\
&= \prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{J} \beta_{lj})}{\prod_{j=1}^{J} \Gamma(\beta_{lj})} \frac{\prod_{j=1}^{J} \Gamma(\beta_{lj} + n(l,k,j))}{\Gamma(\sum_{j=1}^{J} (\beta_{lj} + n(l,k,j)))} \\
&= \prod_{k=1}^{K} \frac{\prod_{j=1}^{J} \Big(1(n(l,k,j) = 0) + 1(n(l,k,j) \neq 0) \prod_{m=0}^{n(l,k,j)-1} (\beta_{lj} + m) \Big)}{1(n(l,k) = 0) + 1(n(l,k) \neq 0) \prod_{m=0}^{n(l,k)-1} [(\sum_{j=1}^{J} \beta_{lj}) + m]}.
\end{aligned}
$$

The (Type II) parameter space is

$$
\Lambda_l = \{\beta_l = (\beta_{l1}, \ldots,\ \beta_{lJ}) \in \mathcal{R}_+^J\} = \mathcal{R}_+^J \quad (l = 0, 1),
$$

where $\mathcal{R}_+$ is the positive real line.

Let Q_l denote a random variable that has the prior distribution on Λ_l. Suppose that the prior density ψ_l for Q_l conditional on $(X = x, D = d)$ does not depend upon (x, d): $\psi_l(\beta_l \mid x, d) = \psi_l(\beta_l)$. Then the conditional density for Q_l given $Z^{(l)} = z^{(l)}$ is

$$
\bar{\psi}_l(\beta_l \mid z^{(l)}) = g_{Y^{(l)} \mid X,D}(y^{(l)} \mid x, d; \beta_l) \psi_l(\beta_l) \Big/ \int_{\Lambda_l} g_{Y^{(l)} \mid X,D}(y^{(l)} \mid x, d; \beta_l) \psi_l(\beta_l)\, d\beta_l, \qquad (15)
$$

which can be combined with (11) to form the decision rule

$$
\begin{aligned}
\text{choose } D_\alpha = 1 \quad \text{if} \quad & \sum_{j=1}^{J} u_1(j) \int_{\Lambda_1} \frac{\beta_{1j} + n(1,\tau,j)}{\sum_{j=1}^{J} [\beta_{1j} + n(1,\tau,j)]} \bar{\psi}_1(\beta_1 \mid z^{(1)})\, d\beta_1 \\
& > \sum_{j=1}^{J} u_0(j) \int_{\Lambda_0} \frac{\beta_{0j} + n(0,\tau,j)}{\sum_{j=1}^{J} [\beta_{0j} + n(0,\tau,j)]} \bar{\psi}_0(\beta_0 \mid z^{(0)})\, d\beta_0. \qquad (16)
\end{aligned}
$$

A potential approximation to this rule can be based on the maximum-likelihood (Type II) estimate of β_l:

$$\hat{\beta}_l = \arg \max_{\beta_l \in \Lambda_l} g_{Y^{(l)} \mid X,D}(y^{(l)} \mid x, d; \beta_l).$$

If $\bar{\psi}_l(\cdot \mid z^{(l)})$ is concentrated around $\hat{\beta}_l$, then an approximation to the decision rule in (16) is

$$\text{choose } D_\alpha = 1 \text{ if}$$
$$\sum_{j=1}^{J} u_1(j) \frac{\hat{\beta}_{1j} + n(1,\tau,j)}{\sum_{j=1}^{J}[\hat{\beta}_{1j} + n(1,\tau,j)]} > \sum_{j=1}^{J} u_0(j) \frac{\hat{\beta}_{0j} + n(0,\tau,j)}{\sum_{j=1}^{J}[\hat{\beta}_{0j} + n(0,\tau,j)]}. \tag{17}$$

The restriction that $\beta_{lj}(x_i)$ does not vary with x_i can be replaced by a parametric model: $\beta_{lj}(x_i) = h_{lj}(x_i;\gamma)$, where h_{lj} is a given function and γ is a parameter vector. Suppose, for example, that we start with M binary variables and, allowing for all possible interactions, let X_i take on $K = 2^M$ values. Then we could consider parametric models that allow for main effects but restrict the interactions. The parametric model plays the role of a prior distribution that can be dominated by the data.

3. RANDOM ASSIGNMENT CONDITIONAL ON MEASURED CHARACTERISTICS

Now consider relaxing the restriction that the prior distribution on Θ_{1l} conditional on $(X, D) = (x, d)$ does not depend upon (x, d) for $l = 0, 1$. We will maintain the assumption in (4) of selection on observables, so that the assignment D_i is independent of the potential outcomes (Y_{0i}, Y_{1i}) conditional on the measured characteristics X_i and on π. The parameter space for $\theta_l = (\pi_l, \eta)$ is $\Theta_l = \Theta_{1l} \times \Theta_2 = \mathcal{S}_{J-1}^K \times [0,1]^K$. Let (T_l, S) denote a random variable that has the prior distribution on $\Theta_{1l} \times \Theta_2$. We shall continue to use the limited information approach, in order to avoid having to specify a joint distribution for $T = (T_1, T_2)$. The assumption of selection on observables is implicitly conditioning on π as well as X. If we do not condition on $T = \pi$, then D_i and (Y_{i0}, Y_{i1}) can fail to be independent conditional on X_i because D helps to predict S, which is related to T. D and T are independent conditional on X and S, but in general we want to allow T and S to be correlated.

So we specify that the distribution of T_l conditional on $(X = x, D = d; S = \eta)$ does not depend upon (x, d) but may depend upon η. The distribution has density $p_{T_l \mid S}$ which specifies that $(T_l(k) : k = 1, \ldots, K)$ are mutually independent with distributions in the Dirichlet family, where $T_l(k) = (T_{l1}(k), \ldots, T_{lJ}(k))$:

$$\begin{aligned} p_{T_l \mid S}(\pi_l \mid x, d; \eta, \beta_l) &= p_{T_l \mid S}(\pi_l \mid \eta, \beta_l) \\ &= \prod_{k=1}^{K} h_{\mathrm{Dir}}(\pi_{l1}(k), \ldots, \pi_{lJ}(k) \mid \beta_{l1}(k, \eta(k)), \ldots, \beta_{lJ}(k, \eta(k))) \end{aligned} \tag{18}$$

for $l = 0, 1$, where $\beta_{lj}(k, \cdot)$ is a function mapping $[0, 1]$ into the positive real line, and $h_{\mathrm{Dir}}(\cdot \mid \zeta)$ is the Dirichlet density with parameter ζ.

As above, the conditional density of T_l given $(Z^{(l)} = z^{(l)}; S = \eta)$ is a product of Dirichlet densities:

$$\bar{p}_{T_l \mid S}(\pi_l \mid z^{(l)}; \eta, \beta_l) = \prod_{k=1}^{K} h_{\mathrm{Dir}}(\pi_{l1}(k), \ldots, \pi_{lJ}(k) \mid \bar{\beta}_{l1}(k, \eta(k)), \ldots, \bar{\beta}_{lJ}(k, \eta(k))) \tag{19}$$

for $l = 0, 1$, where

$$\bar{\beta}_{lj}(k, \cdot) = \beta_{lj}(k, \cdot) + n(l, k, j) \quad \text{with} \quad n(l, k, j) = \sum_{i=1}^{N} 1(d_i = l) 1(x_i = k) 1(y_i = j).$$

The corresponding decision rule, given (β_0, β_1, η), is

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j)(\bar{\beta}_{1j}(\tau, \eta(\tau))/\bar{\beta}_1(\tau, \eta(\tau))) > \sum_{j=1}^{J} u_0(j)(\bar{\beta}_{0j}(\tau, \eta(\tau))/\bar{\beta}_0(\tau, \eta(\tau))), \tag{20}$$

with

$$\bar{\beta}_0(k, \cdot) = \sum_{j=1}^{J} \bar{\beta}_{0j}(k, \cdot), \quad \bar{\beta}_1(k, \cdot) = \sum_{j=1}^{J} \bar{\beta}_{1j}(k, \cdot) \qquad (k = 1, \ldots, K).$$

We can still consider using the decision rule in (12) as an approximation, in which the conditional probabilities that $Y_{0\alpha} = j$ and $Y_{1\alpha} = j$ in $(3')$ are replaced by sample frequencies. For a given utility

function (u_0, u_1), the approximation in (12) coincides with the rule in (20) if $\beta_{lj}(\tau, \eta(\tau))/n(l, \tau, j)$ is sufficiently small ($l = 0, 1$; $j = 1, \ldots, J$).

3.1 *Known Propensity Score*

Now suppose that the prior is not dominated by the data. There may, for example, be a large number K of values for X_i, with a small number of observations in a typical $X_i = k$ cell. We shall assume initially that η is known, so that the propensity score,

$$\Pr(D_i = 1 \,|\, X_i = x_i; \eta) = \eta(x_i),$$

is given. Then we can build on the results for this case when we introduce a prior distribution for η.

Suppose that $\beta_{lj}(k, \cdot)$ does not depend upon k and has the following form:

$$\beta_{lj}(k, u) = \beta_{lj}(u) = \exp(\sum_{m=1}^{O} \beta_{lj}^{(m)} r_{lj}^{(m)}(u)) \qquad (l = 0, 1;\ j = 1, \ldots, J;\ k = 1, \ldots, K;\ 0 \leq u \leq 1),$$

where $r_{lj}^{(m)}(\cdot)$ is a given function mapping $[0, 1]$ into $\mathcal{R}$. For example, we could have a polynomial, with $r_{lj}^{(m)}(u) = u^{m-1}$. If O is sufficiently large, then this specification can be very flexible. Let β_l $= \{\beta_{lj}^{(m)} : j = 1, \ldots, J;\ m = 1, \ldots, O\}$ for $l = 0, 1$. We can form a Type II likelihood function for β_l:

$$\begin{aligned} g_{Y^{(l)} \,|\, X,D}(y^{(l)} \,|\, x, d; \eta, \beta_l) &= \int_{\Theta_{1l}} f_{Y^{(l)} \,|\, X,D}(y^{(l)} \,|\, x, d; \pi_l) p_{T_l \,|\, S}(\pi_l \,|\, x, d; \eta, \beta_l)\, d\pi_l \qquad (21) \\ &= \prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{J} \beta_{lj}(\eta(k)))}{\prod_{j=1}^{J} \Gamma(\beta_{lj}(\eta(k)))} \frac{\prod_{j=1}^{J} \Gamma(\beta_{lj}(\eta(k)) + n(l, k, j))}{\Gamma(\sum_{j=1}^{J} (\beta_{lj}(\eta(k)) + n(l, k, j)))} \end{aligned}$$

for $l = 0, 1$. The (Type II) parameter space is

$$\Lambda_l = \{\beta_{lj}^{(m)} \in \mathcal{R} : j = 1, \ldots, J;\ m = 1, \ldots, O\} = \mathcal{R}^{JO}.$$

Given a prior distribution for β_l, we can form a decision rule. A potential approximation to that rule can be based on the maximum-likelihood (Type II) estimate of β_l:

$$\hat{\beta}_l = \arg \max_{\beta_l \in \Lambda_l} g_{Y^{(l)} \,|\, X,D}(y^{(l)} \,|\, x, d; \eta, \beta_l).$$

Let

$$\hat{\pi}_{lj}(k) = \frac{\exp(\sum_{m=1}^{O} \hat{\beta}_{lj}^{(m)} r_{lj}^{(m)}(\eta(k))) + n(l,k,j)}{\sum_{j=1}^{J}[\exp(\sum_{m=1}^{O} \hat{\beta}_{lj}^{(m)} r_{lj}^{(m)}(\eta(k))) + n(l,k,j)]} \qquad (l = 0,1;\ j = 1,\ldots,J;\ k = 1,\ldots,K).$$

The approximation to the decision rule is

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j)\hat{\pi}_{1j}(\tau) > \sum_{j=1}^{J} u_0(j)\hat{\pi}_{0j}(\tau). \tag{22}$$

When the propensity score is given, it does not play a role through the likelihood function, which is based on the conditional density for the distribution of Y given (X, D). That likelihood function depends only upon π. The propensity score enters through the prior distribution for π. If the prior distribution is dominated by the data, then there would not be value in knowing the propensity score. This corresponds to limit results in Hahn (1998). But if the asymptotic approximation has K increasing as well as N, then knowing the propensity score could have value in the limit results.

3.2 *Correlated Random Effects*

This section examines the role of the propensity score in a random effects model with normal distributions for the outcomes and the random effects. Some of the issues raised in Section 3.1 show up here in a particularly simple form.

Suppose that

$$\begin{aligned} Y_{i1} \,|\, X = x, D = d; \pi, \eta, \sigma &\overset{\text{ind}}{\sim} \mathcal{N}(\pi(x_i), \sigma^2), \\ D_i \,|\, X = x; \pi, \eta, \sigma &\overset{\text{ind}}{\sim} \text{Bern}(\Phi(\eta(x_i))) \qquad (i = 1,\ldots,N), \end{aligned}$$

where the discrete characteristics are coded as $X_i \in \{1,\ldots,K\}$ and Φ is the standard normal cdf. Let $Y_i = D_i * Y_{i1}$ and $Z_i = (X_i, D_i, Y_i)$. We observe $Z = (Z_1,\ldots,Z_N)$. (There could be a parallel analysis in which we observe Y_{i0} if $D_i = 0$.)

The correlated random effects model is

$$\begin{pmatrix} \pi(k) \\ \eta(k) \end{pmatrix} \,|\, X = x; \mu, \Sigma \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma) \qquad (k = 1,\ldots,K),$$

with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

The decision maker, α, is interested in the distribution of $Y_{\alpha 1}$ conditional on the data Z. The decision maker knows her value, τ, for the characteristics. Assume that

$$Y_{\alpha 1} \,|\, Z; \pi, \eta, \sigma \quad \sim \quad \mathcal{N}(\pi(\tau), \sigma^2).$$

Suppose that the propensity score is given, so that

$$\eta(k) = \Phi^{-1}(\Pr(D_i = 1 | X_i = k))$$

is given, for $k = 1, \ldots, K$. Then the decision maker can use the distribution of $Y_{\alpha 1}$ conditional on η and Z. Define

$$\rho = \sigma_{12}/\sigma_{22}, \quad \tilde{\mu}_1 = \mu_1 - \rho\mu_2, \quad \tilde{\sigma}_{11} = \sigma_{11} - \sigma_{12}^2/\sigma_{22},$$

and let $\beta = (\sigma, \rho, \tilde{\mu}_1, \tilde{\sigma}_{11})$. Define

$$n(\tau) = \sum_{i=1}^{N} 1(X_i = \tau, D_i = 1), \quad \bar{Y}(\tau) = n(\tau)^{-1} \sum_{i=1}^{N} 1(X_i = \tau, D_i = 1) Y_i \quad (\text{if } n(\tau) \neq 0)$$

and set $\bar{Y}(\tau) = 0$ if $n(\tau) = 0$. Then we have

$$\pi(\tau) \,|\, \eta, Z; \beta \quad \sim \quad \mathcal{N}(c_1(\beta), c_2(\beta)),$$

where

$$c_1(\beta) = \frac{(n(\tau)/\sigma^2)\bar{Y}(\tau) + \tilde{\sigma}_{11}^{-1}[\tilde{\mu}_1 + \rho\eta(\tau)]}{(n(\tau)/\sigma^2) + \tilde{\sigma}_{11}^{-1}},$$
$$c_2(\beta) = [(n(\tau)/\sigma^2) + \tilde{\sigma}_{11}^{-1}]^{-1}.$$

The corresponding distribution for $Y_{\alpha 1}$ is

$$Y_{\alpha 1} \,|\, \eta, Z; \beta \quad \sim \quad \mathcal{N}(c_1(\beta), c_2(\beta) + \sigma^2).$$

Let z denote the realization of the random variable Z, with $z_i = (x_i, d_i, y_i)$ and $z = (z_1, \ldots, z_N)$. The random effects likelihood function is

$$g_{Y \mid X,D}(y|x,d;\eta,\beta) = \prod_{k:n(k)\geq 1} (2\pi)^{-n(k)/2} [\det \Omega(\beta)]^{-1/2}$$

$$\times \exp\{-\frac{1}{2}[y(k) - (\tilde{\mu}_1 + \rho\eta(k))\mathbf{1}]'[\Omega(\beta)]^{-1}[y(k) - (\tilde{\mu}_1 + \rho\eta(k))\mathbf{1}]\},$$

where $n(k)$ is the number of observations with $(x_i = k, d_i = 1)$, $y(k)$ is the $n(k) \times 1$ matrix formed from the y_i with $(x_i = k, d_i = 1)$,

$$\Omega(\beta) = \tilde{\sigma}_{11}\mathbf{1}\mathbf{1}' + \sigma^2 I,$$

$\mathbf{1}$ is a $n(k) \times 1$ matrix of ones, and I is the identity matrix of order $n(k)$.

A potential approximation for the predictive distribution of $Y_{\alpha 1}$ conditional on η and Z is

$$Y_{\alpha 1} \mid \eta, Z \quad \overset{\text{a}}{\sim} \quad \mathcal{N}(c_1(\hat{\beta}), c_2(\hat{\beta}) + \hat{\sigma}^2),$$

where $\hat{\beta}$ maximizes the random-effects likelihood function:

$$\hat{\beta} = \arg\max_{\beta} g_{Y \mid X,D}(y|x,d;\eta,\beta)$$

(and $\hat{\sigma}$ is the first element of $\hat{\beta}$).

We can extend the correlated random effects specification so that the constant mean μ is replaced by a parametric model $h(x_i;\gamma)$:

$$\begin{pmatrix} \pi(k) \\ \eta(k) \end{pmatrix} \mid X = x; \gamma, \Sigma \quad \overset{\text{ind}}{\sim} \quad \mathcal{N}(h(k;\gamma), \Sigma) \qquad (k = 1, \ldots, K),$$

where $h(\,\cdot\,;\,\cdot\,)$ is a given function and γ is a parameter vector. Suppose, for example, that the underlying characteristics for individual i are in the variables $W_i = (W_{i1}, \ldots, W_{iM})$, and that $X_i = k$ corresponds to the value $W_i = w^{(k)} = (w_1^{(k)}, \ldots, w_M^{(k)})$. Then we could have

$$h(k,\gamma) = w_1^{(k)}\gamma_1 + \ldots + w_M^{(k)}\gamma_M$$

(where $w_m^{(k)}$ is scalar and γ_m is 2×1).

3.3 *Unknown Propensity Score*

Now suppose that the propensity score is not given. We return to the model in Section 3.1, where we have prior distributions for T_0 and for T_1 conditional on $X = x, S = \eta$. Suppose that the prior distribution for S conditional on $X = x$ does not depend upon x. It has density p_S which specifies that $(S(k) : k = 1, \ldots, K)$ are mutually independent with distributions in the beta family:

$$p_S(\eta \,|\, x; \gamma) = p_S(\eta \,|\, \gamma) = \prod_{k=1}^{K} h_{\mathrm{Be}}(\eta(k) \,|\, \gamma_1(k), \gamma_2(k)),$$

where $\gamma_1(k) > 0$, $\gamma_2(k) > 0$, and $h_{\mathrm{Be}}(\cdot \,|\, \zeta_1, \zeta_2)$ is the beta density with parameter (ζ_1, ζ_2):

$$h_{\mathrm{Be}}(w \,|\, \zeta_1, \zeta_2) = \frac{\Gamma(\zeta_1 + \zeta_2)}{\Gamma(\zeta_1)\Gamma(\zeta_2)} w^{\zeta_1 - 1}(1 - w)^{\zeta_2 - 1},$$

for $w \in [0, 1]$ and $\zeta_1 > 0$, $\zeta_2 > 0$.

For $l = 0, 1$, this prior distribution for S can be combined with the prior distribution for T_l conditional on (X, S) to obtain the prior distribution for (T_l, S) conditional on $X = x$ (which in fact does not depend upon x). This prior distribution for (T_l, S) conditional on X can be combined with the joint distribution for $(Y^{(l)}, D)$ conditional on $(X = x; T_l = \pi_l, S = \eta)$ in (6) and (7) to obtain the posterior distribution for (T_l, S) conditional on $Z^{(l)} = z^{(l)}$. The posterior density factors over k, so that $((T_l(k), S(k)) : k = 1, \ldots, K)$ are mutually independent conditional on $Z^{(l)} = z^{(l)}$:

$$\bar{p}_{T_l, S}(\pi_l, \eta \,|\, z^{(l)}; \beta_l, \gamma) = \prod_{k=1}^{K} \bar{p}_{T_l(k), S(k)}(\pi_l(k), \eta(k) \,|\, z^{(l)}; \beta_l, \gamma),$$

and

$$\bar{p}_{T_l(k), S(k)}(\pi_l(k), \eta(k) \,|\, z^{(l)}; \beta_l, \gamma) = \bar{p}_{T_l(k) \,|\, S(k)}(\pi_l(k) \,|\, z^{(l)}; \eta(k), \beta_l) \bar{p}_{S(k)}(\eta(k) \,|\, z^{(l)}; \beta_l, \gamma).$$

The posterior density for $T_l(k)$ conditional on $S(k) = \eta(k)$ is

$$\bar{p}_{T_l(k) \,|\, S(k)}(\pi_l(k) \,|\, z^{(l)}; \eta(k), \beta_l) = h_{\mathrm{Dir}}(\pi_{l1}(k), \ldots, \pi_{lJ}(k) \,|\, \bar{\beta}_{l1}(k, \eta(k)), \ldots, \bar{\beta}_{lJ}(k, \eta(k))),$$

where

$$\bar{\beta}_{lj}(k,u) = \beta_{lj}(u) + n(l,k,j)$$

with

$$\beta_{lj}(u) = \exp(\sum_{m=1}^{O} \beta_{lj}^{(m)} r_{lj}^{(m)}(u)) \quad \text{and} \quad n(l,k,j) = \sum_{i=1}^{N} 1(d_i = l)1(x_i = k)1(y_i = j).$$

The posterior density for $S(k)$ is

$$\bar{p}_{S(k)}(\eta(k) \,|\, z^{(l)}; \beta_l, \gamma) = \frac{\Gamma(\sum_{j=1}^{J} \beta_{lj}(\eta(k)))}{\prod_{j=1}^{J} \Gamma(\beta_{lj}(\eta(k)))} \frac{\prod_{j=1}^{J} \Gamma(\beta_{lj}(\eta(k)) + n(l,k,j))}{\Gamma(\sum_{j=1}^{J}(\beta_{lj}(\eta(k)) + n(l,k,j)))}$$
$$\times\, h_{\mathrm{Be}}(\eta(k) \,|\, \bar{\gamma}_1(k), \bar{\gamma}_2(k))/c^{(l)}(k; \beta_l, \gamma),$$

where

$$\bar{\gamma}_1(k) = \gamma_1(k) + \sum_{i=1}^{N} 1(d_i = 1)1(x_i = k), \quad \bar{\gamma}_2(k) = \gamma_2(k) + \sum_{i=1}^{N} 1(d_i = 0)1(x_i = k),$$

and

$$c^{(l)}(k; \beta_l, \gamma) = \int_{[0,1]} \Bigg(\frac{\Gamma(\sum_{j=1}^{J} \beta_{lj}(u))}{\prod_{j=1}^{J} \Gamma(\beta_{lj}(u))} \frac{\prod_{j=1}^{J} \Gamma(\beta_{lj}(u) + n(l,k,j))}{\Gamma(\sum_{j=1}^{J}(\beta_{lj}(u) + n(l,k,j)))} \tag{23}$$
$$\times\, h_{\mathrm{Be}}(u \,|\, \bar{\gamma}_1(k), \bar{\gamma}_2(k)) \Bigg) du \qquad (l = 0, 1).$$

To evaluate the decision rule, we can use iterated expectations:

$$E(T_{lj}(k) \,|\, Z^{(l)} = z^{(l)}; \beta_l, \gamma) = E[E(T_{lj}(k) \,|\, Z^{(l)} = z^{(l)}; S(k), \beta_l) \,|\, Z^{(l)} = z^{(l)}; \beta_l, \gamma]$$
$$= \int_{[0,1]} \Bigg(\Big[\bar{\beta}_{lj}(k, \eta(k)) / \sum_{j=1}^{J} \bar{\beta}_{lj}(k, \eta(k)) \Big] \bar{p}_{S(k)}(\eta(k) \,|\, z^{(l)}; \beta_l, \gamma) \Bigg) d\eta(k).$$

This only requires one-dimensional numerical integration, which can be done by quadrature. Then, given (β_l, γ), the decision rule is

choose $D_\alpha = 1$ if

$$\sum_{j=1}^{J} u_1(j) E(T_{1j}(\tau) \,|\, Z^{(1)} = z^{(1)}; \beta_1, \gamma) > \sum_{j=1}^{J} u_0(j) E(T_{0j}(\tau) \,|\, Z^{(0)} = z^{(0)}; \beta_0, \gamma). \tag{24}$$

Let $L^{(l)}(\beta_l, \gamma)$ denote the Type II likelihood function for (β_l, γ):

$$\begin{aligned} L^{(l)}(\beta_l, \gamma) &= g_{Y^{(l)}, D \,|\, X}(y^{(l)}, d \,|\, x; \beta_l, \gamma) \\ &= \int_{\Theta_{1l}} \int_{\Theta_2} f_{Y^{(l)} \,|\, X, D}(y^{(l)} \,|\, x, d; \pi_l) f_{D \,|\, X}(d \,|\, x; \eta) p_{T_l \,|\, S}(\pi_l \,|\, \eta, \beta_l) p_S(\eta \,|\, \gamma) \, d\pi_l \, d\eta. \end{aligned}$$

It is given by

$$L^{(l)}(\beta_l, \gamma) = \prod_{k=1}^{K} c^{(l)}(k; \beta_l, \gamma) \frac{\Gamma(\gamma_1(k) + \gamma_2(k))}{\Gamma(\gamma_1(k))\Gamma(\gamma_2(k))} \frac{\Gamma(\bar{\gamma}_1(k))\Gamma(\bar{\gamma}_2(k))}{\Gamma(\bar{\gamma}_1(k) + \bar{\gamma}_2(k))},$$

where $c^{(l)}(k; \beta_l, \gamma)$ is in (23). The evaluation of this likelihood at any point (β_l, γ) only requires the calculation of one-dimensional numerical integrals (there are K of them), which can be done by quadrature.

Suppose that $\gamma_1(k)$ and $\gamma_2(k)$ do not vary with k:

$$\gamma_1(k) = \gamma_1, \quad \gamma_2(k) = \gamma_2 \qquad (k = 1, \ldots, K).$$

Then the (Type II) parameter space is

$$\Lambda_l = \{(\beta_l, \gamma) : \beta_{lj}^{(m)} \in \mathcal{R},\ j = 1, \ldots, J;\ m = 1, \ldots, O;\ (\gamma_1, \gamma_2) \in \mathcal{R}_+ \times \mathcal{R}_+\} = \mathcal{R}^{JO} \times \mathcal{R}_+^2,$$

which has dimension $JO + 2$. A prior distribution on Λ_l can be combined with the (Type II) likelihood function $L^{(l)}(\beta_l, \gamma)$ to obtain a posterior distribution. Then we can integrate

$$M_j^{(l)}(\beta_l, \gamma) \equiv E(T_{lj}(\tau) \,|\, Z^{(l)} = z^{(l)}; \beta_l, \gamma)$$

with respect to this posterior distribution to obtain

$$E(T_{lj}(\tau) \,|\, Z^{(l)} = z^{(l)}) \qquad (l = 0, 1),$$

and the decision rule

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j) E(T_{1j}(\tau) \,|\, Z^{(1)} = z^{(1)}) > \sum_{j=1}^{J} u_0(j) E(T_{0j}(\tau) \,|\, Z^{(0)} = z^{(0)}). \qquad (25)$$

A potential approximation to this decision rule can be based on the maximum-likelihood (Type II) estimate of (β_l, γ):

$$(\hat{\beta}_l, \hat{\gamma}^{(l)}) = \arg \max_{(\beta_l, \gamma) \in \Lambda_l} L^{(l)}(\beta_l, \gamma) \qquad (l = 0, 1).$$

The approximation is

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j) \hat{\pi}_{1j}(\tau) > \sum_{j=1}^{J} u_0(j) \hat{\pi}_{0j}(\tau) \tag{26}$$

with $\hat{\pi}_{lj}(\tau) = M_j^{(l)}(\hat{\beta}_l, \hat{\gamma}^{(l)})$.

Another possibility is to use a separate limited-information approach for S, basing its posterior distribution on (X, D), so that

$$\bar{p}_S(\eta \,|\, x, d; \gamma) = \prod_{k=1}^{K} h_{\mathrm{Be}}(\eta(k) \,|\, \gamma_1 + n(1,k), \gamma_2 + n(0,k)),$$

where

$$n(l,k) = \sum_{i=1}^{N} 1(d_i = l) 1(x_i = k).$$

We can form a Type II likelihood function for γ:

$$\begin{aligned} g_{D \,|\, X}(d \,|\, x; \gamma) &= \int_{\Theta_2} f_{D \,|\, X}(d \,|\, x; \eta) p_S(\eta \,|\, x; \gamma) \, d\eta \\ &= \prod_{k=1}^{K} \frac{\Gamma(\gamma_1 + \gamma_2)}{\Gamma(\gamma_1)\Gamma(\gamma_2)} \frac{\Gamma(\gamma_1 + n(1,k))\Gamma(\gamma_2 + n(0,k))}{\Gamma(\gamma_1 + \gamma_2 + n(1,k) + n(0,k))}. \end{aligned}$$

The maximum-likelihood (Type II) estimate of $\gamma = (\gamma_1, \gamma_2)$ is

$$\hat{\gamma} = \arg \max_{\gamma \in \mathcal{R}_+^2} g_{D \,|\, X}(d \,|\, x; \gamma). \tag{27}$$

Given γ, we can form a Type II likelihood function for β_l, based on the distribution of $Y^{(l)}$ conditional on (X, D):

$$L^{(l)}(\beta_l) = g_{Y^{(l)} \,|\, X, D}(y^{(l)} \,|\, x, d; \beta_l, \gamma)$$

$$= \int_{\Theta_{1l}} \int_{\Theta_2} f_{Y^{(l)} \mid X,D}(y^{(l)} \mid x, d; \pi_l) p_{T_l \mid S}(\pi_l \mid \eta, \beta_l) \bar{p}_S(\eta \mid x, d; \gamma) \, d\pi_l \, d\eta$$

$$= \prod_{k=1}^{K} c^{(l)}(k; \beta_l, \gamma) \qquad (l = 0, 1),$$

where $c^{(l)}(k; \beta_l, \gamma)$ is in (23). Set $\gamma = \hat{\gamma}$ from (27) and obtain

$$\hat{\beta}_l^* = \arg \max_{\beta_l \in \mathcal{R}^{JO}} \prod_{k=1}^{K} c^{(l)}(k; \beta_l, \hat{\gamma}).$$

Then a potential approximation for the decision rule in (25) is

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j) M_j^{(1)}(\hat{\beta}_1^*, \hat{\gamma}) > \sum_{j=1}^{J} u_0(j) M_j^{(0)}(\hat{\beta}_0^*, \hat{\gamma}). \tag{28}$$

4. SELECTION ON UNOBSERVABLES

Now we are going to drop the assumption of selection on observables. For individual i, we observe a vector of discrete variables, which is coded as $W_i \in \{1, \ldots, M\}$. W_i will play the role of an instrumental variable. As before, there is assignment to one of two treatments: $D_i \in \{0, 1\}$. There is a discrete outcome, which is coded as $Y_i \in \{1, \ldots, J\}$. We shall simplify notation by dropping the X variable. Either there are no observable characteristics for the data individuals, or we work with the subset that match the decision maker. Using individuals who do not match the decision maker involves issues similar to those discussed above. Let $Z_i = (W_i, D_i, Y_i)$ and let $Z = (Z_1, \ldots, Z_N)$. So Z is observed.

Let $W = (W_1, \ldots, W_N)$. We shall condition throughout on W, so its distribution will not play a role. The model for treatment assignment uses a latent variable $V = (V_1, \ldots, V_N)$. Conditional on $W = w = (w_1, \ldots, w_N)$, we have

$$D_i = 1(\lambda(w_i) - V_i > 0),$$

where the function λ maps $\{1,\ldots,M\}$ into $[0,1]$ and the V_i are independently and identically distributed with a uniform distribution on the interval $[0,1]$. So the distribution for D_i conditional on W is

$$\Pr(D_i = 1 \,|\, W = w; \lambda) = 1 - \Pr(D_i = 0 \,|\, W = w; \lambda) = \Pr(V_i \leq \lambda(w_i)) = \lambda(w_i).$$

This distribution is unrestricted if λ is unrestricted, so that $\lambda(m)$ can be any value in the interval $[0,1]$ $(m = 1,\ldots,M)$.

For example, we could have

$$D_i = 1(c(w_i) + E_i[u_1(Y_{i1})] - E_i[u_0(Y_{i0})] > 0),$$

where the operator E_i provides the expectation with respect to the personal (subjective) distribution of individual i, and the function c maps $\{1,\ldots,M\}$ into $\mathcal{R}$. Let $U_i = E_i[u_0(Y_{i0})] - E_i[u_1(Y_{i1})]$ and suppose that $(U_1,\ldots,U_N)$ are independent and identically distributed with distribution function G, which is continuous and strictly increasing. Then

$$V_i = G(U_i), \quad \lambda(w_i) = G(c(w_i)), \tag{29}$$

and we want to allow V_i to be correlated with (Y_{i0}, Y_{i1}).

We shall assume that W_i is randomly assigned in that (Y_{i0}, Y_{i1}, V_i) is independent of W_i. Then (Y_{i0}, Y_{i1}) is independent of (W_i, D_i) conditional on V_i. As before, in order to form a likelihood function, we do not need the joint distribution of (Y_{i0}, Y_{i1}), just the two margins. We shall use the following model:

$$\begin{aligned}
&\Pr(Y_{i0} = j \,|\, W_i = w_i, D_i = d_i, V_i = v_i; \beta) = h_{0j}(v_i; \beta), \\
&\Pr(Y_{i1} = j \,|\, W_i = w_i, D_i = d_i, V_i = v_i; \beta) = h_{1j}(v_i; \beta) \qquad (j = 1,\ldots,J),
\end{aligned}$$

where $h_{0j}(\cdot\,;\beta)$ and $h_{1j}(\cdot\,;\beta)$ are functions that map $[0,1]$ into $[0,1]$ and satisfy

$$\sum_{j=1}^{J} h_{0j}(v;\beta) = 1, \quad \sum_{j=1}^{J} h_{1j}(v;\beta) = 1 \qquad (v \in [0,1]).$$

The functions h_{0j} and h_{1j} are given up to a parameter β. We could specify

$$h_{l1}(v;\beta) = \frac{1}{1+\sum_{j=2}^{J}\exp(\sum_{k=1}^{O}\beta_{lj}^{(k)}r_{lj}^{(k)}(v))}, \tag{30}$$

$$h_{lj}(v;\beta) = \frac{\exp(\sum_{k=1}^{O}\beta_{lj}^{(k)}r_{lj}^{(k)}(v))}{1+\sum_{j=2}^{J}\exp(\sum_{k=1}^{O}\beta_{lj}^{(k)}r_{lj}^{(k)}(v))} \qquad (l=0,1;\ j=2,\ldots,J),$$

where $r_{lj}^{(k)}(\cdot)$ is a given function mapping $[0,1]$ into $\mathcal{R}$. For example, we could have a polynomial, with $r_{lj}^{(k)}(v) = v^{k-1}$. If O is sufficiently large, then this specification can be very flexible. The parameter space is $\Theta = \Theta_1 \times \Theta_2$ with

$$\Theta_1 = \{\beta : \beta_{lj} = (\beta_{lj}^{(1)},\ldots,\beta_{lj}^{(O)}) \in \mathcal{R}^O;\ l=0,1;\ j=2,\ldots,J\} = \mathcal{R}^{2O(J-1)},$$

$$\Theta_2 = \{\lambda : \lambda(m) \in [0,1],\ m=1,\ldots,M\} = [0,1]^M.$$

Let z denote the realization of the random variable Z, with $z_i = (w_i, d_i, y_i)$, and let $\theta = (\beta, \lambda)$. The likelihood function for the observation Z is

$$\begin{aligned} f_{Z\,|\,W}(z\,|\,w;\theta) &= \Pr(Z=z\,|\,W=w;\theta) \qquad (31)\\ &= \prod_{i=1}^{N}\int_0^1 \Pr(Y_i=y_i\,|\,W_i=w_i, D_i=d_i, V_i=v_i;\beta)\cdot\Pr(D_i=d_i\,|\,W_i=w_i, V_i=v_i;\lambda)\,dv_i \\ &= \prod_{i=1}^{N}\int_0^1 \Bigg(\prod_{j=1}^{J} h_{0j}(v_i;\beta)^{1(d_i=0)1(y_i=j)}h_{1j}(v_i;\beta)^{1(d_i=1)1(y_i=j)} \\ &\qquad \times \prod_{m=1}^{M} 1(\lambda(m)-v_i\le 0)^{1(d_i=0)1(w_i=m)}1(\lambda(m)-v_i>0)^{1(d_i=1)1(w_i=m)}\Bigg)\,dv_i \\ &= \prod_{l=0}^{1}\prod_{m=1}^{M}\prod_{j=1}^{J} q(l,m,j;\beta,\lambda)^{n(l,m,j)}, \end{aligned}$$

where

$$q(0,m,j;\beta,\lambda) = \int_{\lambda(m)}^{1} h_{0j}(v;\beta)\,dv,$$

$$q(1,m,j;\beta,\lambda) = \int_{0}^{\lambda(m)} h_{1j}(v;\beta)\,dv,$$

and

$$n(l,m,j) = \sum_{i=1}^{N} 1(d_i = l)1(w_i = m)1(y_i = j).$$

Suppose that the decision maker α is exchangeable with the data individuals in that the following sequence of random variables is exchangeable:

$$(Y_{\alpha 0}, Y_{\alpha 1}), (Y_{10}, Y_{11}), \ldots, (Y_{N0}, Y_{N1}).$$

Then the marginal distributions of Y_{i0} and Y_{i1} are relevant for the decision maker, and we shall assume that:

$$\Pr(Y_{\alpha 0} = j \mid \beta) = \Pr(Y_{i0} = j \mid \beta) = \int_0^1 h_{0j}(v;\beta)\, dv,$$

$$\Pr(Y_{\alpha 1} = j \mid \beta) = \Pr(Y_{i1} = j \mid \beta) = \int_0^1 h_{1j}(v;\beta)\, dv.$$

Then, conditional on β, the decision rule is to choose $D_\alpha = 1$ if

$$\sum_{j=1}^{J} u_1(j) \int_0^1 h_{1j}(v;\beta)\, dv > \sum_{j=1}^{J} u_0(j) \int_0^1 h_{0j}(v;\beta)\, dv. \tag{32}$$

More generally, the decision maker could use some other distribution Q_α and the decision rule

$$\text{choose } D_\alpha = 1 \quad \text{if} \quad \sum_{j=1}^{J} u_1(j) \int_0^1 h_{1j}(v;\beta)\, dQ_\alpha(v) > \sum_{j=1}^{J} u_0(j) \int_0^1 h_{0j}(v;\beta)\, dQ_\alpha(v). \tag{33}$$

Suppose that i has a personal distribution H for (Y_{i0}, Y_{i1}, R_i), observes the signal R_i that is related to (Y_{i0}, Y_{i1}), and then forms $E_i[u_0(Y_{i0})]$ and $E_i[u_1(Y_{i1})]$ by using H to form the conditional distribution of Y_{i0} given R_i and the conditional distribution of Y_{i1} given R_i. Suppose this holds for $i = 1, \ldots, N$ and for the decision maker $i = \alpha$, with the same personal distribution H. Before conditioning on Z, the decision maker observes R_α and forms $E_\alpha[u_0(Y_{\alpha 0})]$ and $E_\alpha[u_1(Y_{\alpha 1})]$ by using H to form the conditional distribution of $Y_{\alpha 0}$ given R_α and the conditional distribution of $Y_{\alpha 1}$ given R_α. Suppose that (Y_{i0}, Y_{i1}, R_i) are independent and identically distributed according to

some (unknown) distribution P, for $i = \alpha, 1, \ldots, N$. This implies that $U_i \equiv E_i[u_0(Y_{i0})] - E_i[u_1(Y_{i1})]$ is independent and identically distributed, and as above we let G denote the distribution function and assume it is continuous and strictly increasing. Let $V_i = G(U_i)$ for $i = \alpha, 1, \ldots, N$. Then, conditional on P, we have (Y_{i0}, Y_{i1}, V_i) independent and identically distributed, with V_i uniform on $[0, 1]$. P implies a conditional distribution for (Y_{i0}, Y_{i1}) given V_i, and we assume, as above, that this implies

$$\begin{aligned} &\Pr(Y_{i0} = j \,|\, V_i = v_i; \beta) = h_{0j}(v_i; \beta), \\ &\Pr(Y_{i1} = j \,|\, V_i = v_i; \beta) = h_{1j}(v_i; \beta) \qquad (i = \alpha, 1, \ldots, N) \end{aligned} \tag{34}$$

for some $\beta \in \Theta_1$. Then Q_α would be the decision maker's posterior distribution for V_α. Note that V_α depends upon U_α and G. The decision maker knows U_α but he does not know G. Furthermore, G does not appear in the likelihood function, so there is no direct information on G in the data. In addition, this approach requires a detailed specification of what people knew and when they knew it. So a limited information approach may be appropriate, simply setting Q_α equal to the uniform distribution.

A prior distribution for (β, λ) can be combined with the likelihood function in (31) to obtain a posterior distribution. The corresponding decision rule is obtained by integrating both sides of the inequality in (32) with respect to the posterior distribution for β. In general the prior distribution will not be dominated by the data. Nevertheless, it may be useful to have a reference decision rule that does not involve a numerical specification for the prior. One way to do this is to replace β in (32) by $\hat{\beta}$, a maximum-likelihood estimate:

$$(\hat{\beta}, \hat{\lambda}) = \arg \max_{(\beta,\lambda) \in \Theta_1 \times \Theta_2} f_{Z \,|\, W}(z \,|\, w; (\beta, \lambda)).$$

Note that

$$\sum_{j=1}^{J} q(0, m, j; \beta, \lambda) = 1 - \lambda(m), \quad \sum_{j=1}^{J} q(1, m, j; \beta, \lambda) = \lambda(m).$$

Then $q(l, m, j; \beta, \lambda) \geq 0$ and

$$\sum_{l=0}^{1} \sum_{j=1}^{J} q(l, m, j; \beta, \lambda) = 1$$

imply that

$$\prod_{l=0}^{1}\prod_{j=1}^{J} q(l,m,j;\beta,\lambda)^{n(l,m,j)} \le \prod_{l=0}^{1}\prod_{j=1}^{J} [n(l,m,j)/n(m)]^{n(l,m,j)} \qquad (m=1,\ldots,M),$$

where

$$n(m) = \sum_{l=0}^{1}\sum_{j=1}^{J} n(l,m,j).$$

Hence

$$\max_{(\beta,\lambda)\in\Theta_1\times\Theta_2} f_{Z\,|\,W}(z\,|\,w;(\beta,\lambda)) \le \prod_{l=0}^{1}\prod_{m=1}^{M}\prod_{j=1}^{J} [n(l,m,j)/n(m)]^{n(l,m,j)}.$$

So, if we can solve the following equations, we will obtain maximum-likelihood estimates:

$$q(l,m,j;\hat\beta,\hat\lambda) = n(l,m,j)/n(m) \qquad (l=0,1;\ m=1,\ldots,M;\ j=1,\ldots,J).$$

An equivalent set of equations is

$$\begin{aligned}
\hat\lambda(m) &= \frac{n(1,m)}{n(m)},\\
\frac{q(0,m,j;\hat\beta,\hat\lambda)}{1-\hat\lambda(m)} &= \frac{1}{1-\hat\lambda(m)}\int_{\hat\lambda(m)}^{1} h_{0j}(v;\hat\beta)\,dv = \frac{n(0,m,j)}{n(0,m)},\\
\frac{q(1,m,j;\hat\beta,\hat\lambda)}{\hat\lambda(m)} &= \frac{1}{\hat\lambda(m)}\int_{0}^{\hat\lambda(m)} h_{1j}(v;\hat\beta)\,dv = \frac{n(1,m,j)}{n(1,m)},
\end{aligned}$$

where

$$n(l,m) = \sum_{j=1}^{J} n(l,m,j) \qquad (l=0,1;\ m=1,\ldots,M;\ j=1,\ldots,J).$$

Here we obtain $\hat\lambda(m)$ by forming the subgroup whose value for the instrumental variable is $W_i = m$, and then calculate the fraction of this subgroup for which the treatment assignment is $D_i = 1$. Then form the subgroup with $W_i = m$ and $D_i = l$, and calculate the fraction of this subgroup with $Y_i = j$. With $\hat\lambda(m)$ already determined, we try to solve for $\hat\beta$ by matching the model's probabilities to these fractions.

Consider, for example, the specification for h_{lj} in (30). We set $\hat{\lambda}(m) = n(1,m)/n(m)$ and try to solve

$$\frac{1}{1-\hat{\lambda}(m)} \int_{\hat{\lambda}(m)}^{1} \frac{\exp(\sum_{k=1}^{O} \hat{\beta}_{0j}^{(k)} r_{0j}^{(k)}(v))}{1+\sum_{j=2}^{J} \exp(\sum_{k=1}^{O} \hat{\beta}_{0j}^{(k)} r_{0j}^{(k)}(v))}\, dv = \frac{n(0,m,j)}{n(0,m)},$$

$$\frac{1}{\hat{\lambda}(m)} \int_{0}^{\hat{\lambda}(m)} \frac{\exp(\sum_{k=1}^{O} \hat{\beta}_{1j}^{(k)} r_{1j}^{(k)}(v))}{1+\sum_{j=2}^{J} \exp(\sum_{k=1}^{O} \hat{\beta}_{1j}^{(k)} r_{1j}^{(k)}(v))}\, dv = \frac{n(1,m,j)}{n(1,m)} \qquad (m=1,\ldots,M;\ j=2,\ldots,J).$$

For each $l \in \{0,1\}$ and $j \in \{2,\ldots,J\}$ there are M equations and O unknowns in $(\hat{\beta}_{lj}^{(1)},\ldots,\hat{\beta}_{lj}^{(O)})$. So we do not expect a unique solution when O is greater than M. The nonuniqueness does not affect how well we "fit" the data, but different solutions for β imply different decision rules.

One possibility is to set $\beta_{lj}^{(k)} = 0$ for $M < k \leq O$. Then the prior is reflected in a careful choice of the basis elements $r_{lj}^{(k)}(\cdot)$ for $k = 1,\ldots,M$. It may be useful to consider prior distributions on the coefficients $\beta_{lj}^{(k)}$, particularly if some of the cell counts $n(l,m,j)$ are small. Also, a prior distribution on the coefficients could downweight the contribution of later basis elements, without the need for a sharp cutoff that sets $\beta_{lj}^{(k)} = 0$ for $k > M$. We shall leave the development of such prior distributions on the coefficients for future work. The main point here is that given the lack of identification, there will be aspects of the prior that are not dominated by the data.

To get a sense of the extrapolation that the model provides, we can make a connection with the role of compliers, always-takers, and never-takers in the local average treatment effect developed by Imbens and Angrist (1994). Note that

$$\Pr(D_i = 0, Y_i = j \,|\, W_i = m; \beta, \lambda) = \int_{\lambda(m)}^{1} h_{0j}(v;\beta)\, dv.$$

Suppose that $\lambda(m') < \lambda(m'')$. Then we have

$$\begin{aligned} &\frac{\Pr(D_i = 0, Y_i = j \,|\, W_i = m'; \beta, \lambda) - \Pr(D_i = 0, Y_i = j \,|\, W_i = m''; \beta, \lambda)}{\Pr(D_i = 0 \,|\, W_i = m'; \lambda) - \Pr(D_i = 0 \,|\, W_i = m''; \lambda)} \qquad (35) \\ &\quad = \frac{1}{\lambda(m'') - \lambda(m')} \int_{\lambda(m')}^{\lambda(m'')} h_{0j}(v;\beta)\, dv \\ &\quad = \Pr(Y_{i0} = j \,|\, \lambda(m') < V_i < \lambda(m''); \beta, \lambda). \end{aligned}$$

The condition that $\lambda(m') < V_i < \lambda(m'')$ corresponds to the compliers. There is a direct estimate of the probability that $Y_{i0} = j$ for compliers, in which the probabilities of observable events in (35) are replaced by sample frequencies:

$$\frac{\frac{n(0,m',j)}{n(m')} - \frac{n(0,m'',j)}{n(m'')}}{\frac{n(0,m')}{n(m')} - \frac{n(0,m'')}{n(m'')}}.$$

We also have

$$\begin{aligned}\Pr(Y_i = j \mid D_i = 0, W_i = m''; \beta, \lambda) &= \frac{1}{1-\lambda(m'')} \int_{\lambda(m'')}^{1} h_{0j}(v;\beta)\, dv \\ &= \Pr(Y_{i0} = j \mid \lambda(m'') < V_i; \beta, \lambda).\end{aligned}$$

The condition that $\lambda(m'') < V_i$ corresponds to the never-takers. There is a direct estimate of the probability that $Y_{i0} = j$ for never-takers, using sample frequencies :

$$\frac{n(0,m'',j)}{n(0,m'')}.$$

No direct estimate, however, is available for

$$\frac{1}{\lambda(m')} \int_{0}^{\lambda(m')} h_{0j}(v;\beta)\, dv, \tag{36}$$

which is the probability that $Y_{i0} = j$ for always-takers. The role of the model is to provide an extrapolation for this term. An estimate of β is obtained by fitting the sample frequencies, and then $h_{0j}(v;\hat{\beta})$ can be used to evaluate the integral in (36). Likewise, no direct estimate is available for

$$\frac{1}{1-\lambda(m'')} \int_{\lambda(m'')}^{1} h_{1j}(v;\beta)\, dv, \tag{37}$$

which is the probability that $Y_{i1} = j$ for never-takers. The model provides an extrapolation for this term, using $h_{1j}(v;\hat{\beta})$ to evaluate the integral in (37).

5. CONNECTIONS WITH THE LITERATURE

Dehejia (2005) applies Bayesian decision theory to program evaluation. The Greater Avenues for Independence (GAIN) program began operating in California in 1986 with the aim of increasing employment and earnings among welfare (AFDC) recipients. Dehejia considers a caseworker choosing whether to assign a welfare recipient into GAIN or AFDC. The caseworker knows a list of characteristics of the individual including age, ethnicity, educational attainment, score on reading and mathematics tests, sex, an indicator for previous participation in other training programs, and pre-assignment earnings history. The caseworker has access to data on welfare recipients in which half were randomly assigned into the GAIN program and the other half were assigned to a control group that was prohibited from receiving GAIN services. An earnings outcome is observed for the treatment group and the control group, as well as the list of characteristics. So the caseworker's decision problem resembles the one I have developed in Section 2, using random assignment. Dehejia uses diffuse priors for the parameters of his model. In the discrete data case I consider, this could correspond to the decision rule in (12). Dehejia goes on to consider the implications for social welfare of different assignment mechanisms, such as making all assignments to GAIN, or all to AFDC, or having the caseworker make assignments for each individual based on comparing the individual's (predictive) distribution of future earnings under GAIN with his distribution of future earnings under AFDC.

Manski (2004) considers a planner who wants to maximize population mean welfare. The planner observes a list of discrete covariates for each person and can design treatment rules that differentiate between persons based on their covariate values. The planner has access to a data set in which individuals were randomly assigned a treatment, and values were recorded for covariates, treatment, and outcome. Manski focuses on conditional empirical success rules, in which conditional expectations are replaced by sample averages and treatments are chosen to maximize empirical success. He notes that conditioning tends to diminish the statistical precision of sample averages and that conditioning on only some part of the observed covariates may be preferable when making

treatment choices. He uses a minimax regret criterion and develops bounds which give sufficient conditions on sample size in order for it to be optimal to condition treatment choices on all observed covariates. This corresponds to my decision rule in (12), which does not use Good's (1965) Type II likelihood function. Stoye (2009) obtains exact results on minimax regret rules. In assigning treatment to an individual with covariate value x, only the subset of the data which matches that covariate value is used, no matter how small the subset. As the number of values that the discrete covariate can take on increases, a minimax regret rule approaches a no data rule.

Angrist and Hahn (2004) are motivated by the result in Hahn (1998) that knowledge of the propensity score does not lower the semiparametric efficiency bound for the average treatment effect. They say that (page 58): "In short, conventional asymptotic arguments would appear to offer no justification for anything other than full control for covariates in estimation of average treatment effects." They argue (page 58) that "...because covariate cells may be small or empty, in finite samples there is a cost to covariate matching, even if covariates are discrete and exact matching is feasible." They work with a multinomial covariate that takes K possible values, and they develop an alternative asymptotic approximation where cell sizes are fixed but the number of cells becomes infinitely large. They refer to this as "panel asymptotics," because of the similarity to large cross-section, small time-series asymptotics used with panel data models. Their treatment-assignment mechanism has a constant propensity score, so random assignment. They consider an estimator with full control for covariates (covariate matching) and one which ignores the covariates (matching on the propensity score, which is constant). In analogy with random effects estimators for panel data, they also consider a linear combination of these estimators, that is more efficient than either one under their asymptotic sequence. Their focus is on estimating an average treatment effect, whereas the decision problem in my paper is more related to a treatment effect for a particular covariate cell. Nevertheless, their panel data analogy is relevant for my paper and the Type II likelihood function can be given a random effects interpretation, as I have done in Section 3.2.

Hirano and Porter (2009) develop an asymptotic theory explicitly for treatment choice. They establish an asymptotic optimality for Manski's (2004) conditional empirical success rule, in the

case where a multinomial covariate takes on K possible values, where K is fixed as sample size N tends to infinity. It would be of interest to have results here under the Angrist-Hahn (2004) asymptotic sequence, where cell sizes are fixed and K tends to infinity.

Rubin (1978) discusses the role of randomization in Bayesian inference for causal effects. In the case of selection on observables, where the treatment assignment D_i is independent of the decision outcomes (Y_{i0}, Y_{i1}) conditional on the measured characteristics X_i (as in (4)), it follows from Rosenbaum and Rubin (1983) that this independence holds conditional on the propensity score. Rubin (1985) notes that (page 463): "It has often been argued that randomization probabilities in surveys or experiments are irrelevant to a Bayesian statistician." This issue is also discussed in Robins and and Ritov (1997). Rubin (1985) argues for a limited information approach in which the analysis proceeds as if only the propensity score and not X_i had been observed. Robins and Ritov (1997) relate the use of the propensity score to a minimax criterion. They have a discussion of dependent priors (Section 6) that relates to the dependence on η that I allow for in specifying a prior for π in (18). Sims (2006) discusses an example from Wasserman (2004). The arguments that Sims gives for dependence in Section III of his paper (Dependence: Direct Approach) are similar to my motivation for allowing for dependence on η in the prior for π when there is selection on observables. Sims (2006) also discusses a limited information approach.

The latent variable model of selection in Section 4 follows Heckman and Vytlacil (1999, 2005). The connection with the Imbens and Angrist (1994) model is developed in Vytlacil (2002). Manski (1990, 1996) discusses the lack of point identification for average treatment effects, and Manski (2000) discusses implications for decision making.

My use of expected utility maximization is motivated by the Savage (1972) axioms for rational behavior. Some of the decision rules I have provided can be used as "automatic" reference rules in a range of contexts, without needing any additional specification. So risk functions can be calculated for these rules. Chamberlain (2000) discusses the role of risk robustness and regret risk in decision making.

REFERENCES

Angrist, J. and J. Hahn (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *The Review of Economics and Statistics*, 86, 58–72.

Chamberlain, G. (2000): "Econometrics and Decision Theory," *Journal of Econometrics*, 95, 255–283.

Dehejia, R. (2005): "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141–173.

Good, I.J. (1965): *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge: The M.I.T. Press.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Heckman, J. and E. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

Heckman, J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

Hirano, K. and J. Porter (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701.

Imbens, G. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Manski, C. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review Papers and Proceedings*, 80, 319–323.

Manski, C. (1996): "Learning About Treatment Effects from Experiments with Random Assignment of Treatments," *The Journal of Human Resources*, 31, 709–733.

Manski, C. (2000): "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics*, 95, 415–442.

Manski, C. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

Robins, J. and Y. Ritov (1997): "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319.

Rosenbaum, P. and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rubin, D. (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

Rubin, D. (1985): "The Use of Propensity Scores in Applied Bayesian Inference," in *Bayesian Statistics 2*, ed. by J. Bernardo, M. DeGroot, D. Lindley, and A. Smith. Amsterdam: North-Holland.

Savage, L.J. (1972): *The Foundations of Statistics*. New York: Dover Publications.

Sims, C. (2006): "On an Example of Larry Wasserman," unpublished manuscript, Department of Economics, Princeton University.

Stoye, J. (2009): "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70–81.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

Wasserman, L. (2004): *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer.