A Bayesian Semiparametric Approach to Intermediate Variables in Causal Inference

Scott L. Schwartz [1] , Fan Li [2], Fabrizia Mealli [3]

[1] Department of Statistics, Texas A&M University;

[2] Department of Statistical Science, Duke University;

[3] Dipartimento di Statistica, Università di Firenze.

ABSTRACT

In causal inference studies, treatment comparisons often need to be adjusted for confounded post-treatment variables. Principal stratification (PS) is a framework to deal with such variables within the potential outcome approach to causal inference. Continuous intermediate variables introduce inferential challenges to PS analysis. Existing methods either dichotomize the intermediate variable, or assume a fully parametric model for the joint distribution of the potential intermediate variables. However, the former is subject to information loss and arbitrary choice of the cutoff point and the latter is often inadequate to represent complex distributional and clustering features. We propose a Bayesian semiparametric approach that consists of a flexible parametric model for the potential outcomes and a Bayesian nonparametric model for the potential intermediate outcomes using a Dirichlet process mixture (DPM) model. The DPM approach provides flexibility in modeling the possibly complex joint distribution of the potential intermediate outcomes and offers better interpretability of results through its clustering feature. The Gibbs sampling based posterior inference is developed. We illustrate the method by two applications: one concerning partial compliance in a randomized clinical trial, and one concerning the causal mechanism between physical activity, body mass index and cardiovascular disease in the observational Swedish National March Cohort study.

KEY WORDS: Bayesian nonparametrics, causal inference, compliance, intermediate variables, Dirichlet process, mixture model, principal stratification.

1 Introduction

In causal inference studies, treatment comparisons often need to be adjusted for intermediate variables, i.e., post-treatment variables potentially affected by the treatment and also affecting the response. In some randomized trials, for example, intermediate variables are present in the form of non or partial compliance to assigned treatment, surrogate endpoints, unintended missing outcome data, or truncation by death of primary outcomes. More generally, in both experimental and observational studies, researchers are interested in knowing not only if the treatment has an effect on the primary outcome, but also to what extent this effect is mediated by some intermediate variables.

It is well documented that directly applying standard methods of pre-treatment variable adjustment, such as regression methods, to intermediate variables can result in estimates that generally lack causal interpretation (e.g., Rosenbaum, 1984). In this paper, we address these problems using the potential outcome approach to causal inference, also known as the Rubin Causal Model (RCM) (Rubin, 1974, 1978). In this perspective, a causal inference problem is viewed as a problem of missing data, where the assignment mechanism is specified as a stochastic rule for revealing the observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects. A commonly invoked identifying assumption is unconfoundedness (Rosenbaum and Rubin, 1983), which usually holds by design in randomized experiments. However, even under such an assumption, inference on causal effects may be invalidated due to the presence of the above mentioned post-treatment complications. Noncompliance to the protocol, for example, renders the two groups of subjects receiving and not receiving the active treatment no longer comparable, and adjustment for the possibly confounded compliance behavior is required. Another example is when outcomes of interest are not well-defined for all units. This problem has been dubbed *truncation by death* (Rubin, 2006), borrowing the term from medical clinical trials where the outcome – e.g., quality of life – is undefined for patients who die. In this setting the intermediate variable

is the indicator of outcome *truncation*.

Within the RCM, a relatively recent approach to deal with these complications is Principal Stratification (PS) (Frangakis and Rubin, 2002). A PS regarding an intermediate variable is a cross-classification of subjects into latent classes defined by the joint potential values of that intermediate variable under each of the treatments being compared. Principal strata comprise units having the same joint values of the intermediate potential outcomes, and so are not affected by treatment assignment. This means that comparisons of potential outcomes under different treatment levels within a principal stratum, or union of principal strata, are well-defined causal effects in the sense of Rubin (1978). These comparisons are called principal causal effects.

Much of the literature discusses settings with binary intermediate variables; if the treatment is also binary, there are at most four principal strata (e.g., Angrist et al., 1996). On the other hand, continuous intermediate outcomes lead to an infinite number of possible principal strata, introducing substantial complications to both inference and interpretation. Few papers have dealt with many (i.e., Mattei and Mealli, 2007; Jin and Rubin, 2009) or continuous principal strata (i.e., Jin and Rubin, 2008; Bartolucci and Grilli, 2011), and one common method is to dichotomize continuous intermediate variables (e.g., Sjölander et al., 2009). However, dichotomization is often subject to arbitrary choice of cutoff points and results can be sensitive to this choice. It is also subject to loss of scientifically important information regarding relevant underlying structure, and *a priori* coarsening of the intermediate outcomes may also lead to violations of some identifying assumptions. Much of the work on continuous PS has been discussed in the context of partial compliance in randomized clinical trials and applied to the data in Efron and Feldman (1991): Jin and Rubin (2008) proposed Bayesian parametric models for both the outcome and the intermediate variables; Bartolucci and Grilli (2011) developed a novel frequentist semiparametric approach consisting of a parametric model for the outcome and a nonparametric model for the intermediate variables based on copula. Restricting models to a single parametric family, as in Jin and Rubin (2008), can potentially be inadequate for

complex data distributions involving, e.g., outliers, skewness and multi-modality, that are common in real applications. This is a serious concern because modeling the association between intermediate potential outcomes plays a crucial role in inference, since it implicitly defines the latent structure that drives PS analysis. The copula-based method, though flexible, is not easily extended to directly include covariates, which are crucial for analyzing observational studies. These issues together have motivated us to develop the flexible approach proposed here.

Inferences in PS generally involve techniques for incomplete data, because at most one potential intermediate outcome is observed for each subject and thus principal strata are only partially observed. Bayesian approaches appear to be particularly useful to deal with the large amount of missing data characterizing PS (and in general causal inference) problems. From a Bayesian perspective, the unobserved potential outcomes are no different than unknown parameters. Inferences across models with different parametric structures can be compared directly because they are all driven by the posterior distribution of the same causal estimands defined by the potentially observable outcomes. The Bayesian approach also clarifies what can be learned when causal estimands are intrinsically not *fully* identified, but only *weakly* identified (Imbens and Rubin, 1997), i.e., the likelihood function can be flat around its maximum. In particular, issues of identification are quite different from those in the frequentist perspective because with proper prior distributions, posterior distributions are always proper. The effect of adding or dropping assumptions is directly addressed in the Bayesian approach by examining how the posterior distributions for causal estimands change; this provides a natural framework to assess the sensitivity of causal conclusions to the different assumptions, as we shall show in our empirical illustrations.

In order to model the infinite principal strata generated by the continuous intermediate variables, we propose a Bayesian nonparametric model for the principal strata based on the Dirichlet process mixture (DPM) model (e.g., Escobar and West, 1995). The resulting principal strata model has support over the space of all mixed continuous distributions, and thus greatly mitigates model flexibility and appropriateness concerns, as well as concerns on the

sensitivity of inference relative to the specification of the joint distribution of intermediate outcomes. In addition, because DPM models exhibit clustering properties, they seem a particularly appropriate choice for the PS setting for two reasons. First, clustering encourages information sharing. In the PS setting, with at least half of the potential outcomes being missing, sharing information across subjects with similar characteristics may be particularly desirable. Second, principal strata are latent classes, thus clustering provides opportunities to model, explore, and potentially interpret the latent structure of the data. This *ex-post* coarsening does not have the drawbacks of the *a priori* dichotomization coarsening mentioned above.

The remainder of the article is organized as follows. In Section 2, we introduce the PS framework and present a general Bayesian semiparametric model to conduct inference within this framework; a procedure for posterior inference for model parameters is also developed. In Section 3, we illustrate the method by estimating the effect of a drug in the randomized clinical trial with partial compliance presented by Efron and Feldman (1991) and compare the results with those from previous studies. We then apply the method to an observational study – the Swedish National March Cohort (NMC) – to investigate the effect of physical activity on coronary heart disease as it relates to BMI in Section 4. Section 5 concludes with a discussion.

2 General model and inference

2.1 Basic setup, definitions and assumptions

Consider a large population of units, each of which can potentially be assigned a treatment indicated by t, with $t = 1$ for active treatment and $t = 0$ for control, i.e., no active treatment. A random sample of n units from this population comprises the participants in a study, designed to evaluate the effect of T (active treatment vs no active treatment) on some outcome Y. At baseline, a set of p background characteristics of the n units are collected and define the $n \times p$ matrix $\mathbf{X}$. Let T_i be the binary variable indicating whether subject i $(i = 1, ..., n)$ is assigned to the treatment group ($T_i = 1$) or to the control group ($T_i = 0$), and let $\mathbf{T} = (T_1, ..., T_n)'$.

We adopt the potential outcome framework to define causal effects. Assuming SUTVA (Stable Unit Treatment Value Assumption) (Rubin, 1980), we define, for each unit i and each post-treatment variable, two potential outcomes, each associated with one of the two treatment levels that unit i can potentially receive. SUTVA implies that potential outcomes for unit i are unaffected by the treatment assignments of other units and that for each unit no different possible *doses* of either treatment are contemplated, so that each treatment defines a single outcome for each unit. SUTVA allows us to write the potential outcomes for unit i as a function of T_i rather than the entire vector $\mathbf{T}$. Thus, for the post-treatment response Y, it is postulated that each unit i has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, representing the hypothetical outcomes under its assignment to treatment or to control. Only one potential outcome is observed for unit i, $Y_i^{obs} = Y_i(T_i)$; the other potential outcome, $Y_i^{mis} = Y_i(1 - T_i)$, is not observed. Yet, the causal effect of T on Y is defined, on a single unit i, as a comparison between $Y_i(1)$ and $Y_i(0)$, e.g., $Y_i(1)$-$Y_i(0)$. More generally a causal effect is any comparison of the potential outcomes under treatment versus control on a common set of units.

An intermediate variable D is a post-treatment variable, and so it too has two potential versions $D_i(0)$ and $D_i(1)$ for each unit i, with $D_i^{obs} = D_i(T_i)$. Formalizing a causal problem by means of potential outcomes helps clarifying, why, in general, it is improper to adjust for intermediate variables by conditioning on their observed values. For example, we cannot compare quality of life of the *survived* treated units with that of *survived* control units, because these two groups are obtained by conditioning on the *truncation* indicator $D_i^{obs} = 0$, and so on different variables for units under treatment and under control, $D_i(1)$ and $D_i(0)$ respectively. This comparison lacks causal interpretation because it estimates summaries of $Y_i(1)$ and $Y_i(0)$ on different sets of units.

A possible way to overcome this problem is to introduce the notion of principal stratification (PS). The *basic* PS, as defined in Frangakis and Rubin (2002), regarding the post-treatment variable D is a partition of units, whose sets are defined by the joint values $(D_i(0), D_i(1))$. More generally, a PS regarding the intermediate variable D is a partition of units, whose sets

are unions of sets in the basic PS. The principal stratum membership $S_i = (D_i(0), D_i(1))$ is not affected by treatment assignment by definition, so it only reflects characteristics of subject i, and can be regarded as a (latent) covariate, which is only partially observed in the sample, as $D_i(0)$ and $D_i(1)$ can never be observed jointly. Comparisons of summaries of $Y(1)$ and $Y(0)$ within a principal stratum, the so-called principal causal effects (PCEs), have a causal interpretation because they compare quantities defined on a common set of units.

Inference on PCEs is challenging due to the latent nature of the principal strata. As in all causal inference problems, we need to first pose an assignment mechanism. The distribution of $\mathbf{T}$ conditional on the potential outcomes and observed covariates defines the assignment mechanism - a probabilistic model that determines which of the two potential outcomes will be observed. Throughout the paper, we will assume that treatment assignment for unit i does not depend on the outcomes and assignment for other units, i.e.,

$$\Pr(\mathbf{T}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1), \mathbf{X}) = \prod_{i}^{n} \Pr(T_i|Y_i(0), Y_i(1), D_i(0), D_i(1), \mathbf{X}_i), \qquad (1)$$

where the bold indicates column vectors of the corresponding unit-level variables (e.g., $\mathbf{Y}(0) = (Y_1(0), ..., Y_n(0))'$) and $\Pr(\cdot|\cdot)$ is generic notation for a conditional distribution. We will also assume that treatment assignment is *conditionally unconfounded* given $\mathbf{X}_i$, meaning that:

$$\Pr(T_i|Y_i(0), Y_i(1), D_i(0), D_i(1), \mathbf{X}_i) = \Pr(T_i|\mathbf{X}_i), \qquad (2)$$

and it is *probabilistic*, i.e., all the unit-level probabilities are between 0 and 1, so that all units have a chance of receiving each of the treatments: $0 < \Pr(T_i|\mathbf{X}_i) < 1$, $i = 1, ..., n$. An unconfounded probabilistic assignment mechanism is called *strongly ignorable* in Rosenbaum and Rubin (1983), a stronger version of an ignorable assignment mechanism (Rubin, 1978). If true, strong ignorability guarantees that the comparisons of treated and control units with the same value of $\mathbf{X}$ have a causal interpretation because $\Pr(Y_i(t)|\mathbf{X}_i) = \Pr(Y_i^{obs}|T_i = t, \mathbf{X}_i)$. Unconfoundedness implies that $\Pr(T_i|Y_i(0), Y_i(1), D_i(0), D_i(1), \mathbf{X}_i) = \Pr(T_i|D_i(0), D_i(1), \mathbf{X}_i)$, so that treatment assignment is independent of the potential outcomes given the principal

strata: while it is in general improper to condition on D_i^{obs}, treated and control units can instead be compared conditional on $S_i = (D_i(0), D_i(1))$. Unconfoundedness also implies that $\Pr(T_i|D_i(0), D_i(1), \mathbf{X}_i) = \Pr(T_i|\mathbf{X}_i)$, i.e., S_i is guaranteed to have the same distribution in both treatment arms, within cells defined by covariates $\mathbf{X}$.

To show how Bayesian inference proceeds in the PS framework, consider the six quantities associated with each sampled unit: $(Y_i(0), Y_i(1), D_i(0), D_i(1), T_i, \mathbf{X}_i)$. These quantities are considered as a joint draw from the population distribution, i.e., Bayesian inference considers the observed values of these quantities to be realizations of random variables and the unobserved values to be unobserved random variables. The joint probability (density) function of all random variables is:

$$\begin{aligned}\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1), \mathbf{T}, \mathbf{X}) &= \Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1)|\mathbf{T}, \mathbf{X}) \Pr(\mathbf{T}|\mathbf{X}) \Pr(\mathbf{X}) \\ &= \Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1)|\mathbf{X}) \Pr(\mathbf{T}|\mathbf{X}) \Pr(\mathbf{X}),\end{aligned}$$

where the last equality follows from (2), and allows us to *ignore* $\Pr(\mathbf{T}|\mathbf{X})$ which, as a consequence, does not need to be modeled. In what follows, we will condition on the observed distribution of covariates, so that also $\Pr(\mathbf{X})$ does not need to be modeled. With essentially no loss of generality, assuming that the distribution $\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1)|\mathbf{X})$ is unit exchangeable, that is, is invariant under a permutation of the unit indexes, by appealing to de Finetti's theorem (Rubin, 1978), we can rewrite $\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1)|\mathbf{X})$ as:

$$\begin{aligned}&\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1)|\mathbf{X}) \\ = &\int \prod_i \Pr(Y_i(0), Y_i(1)|D_i(0), D_i(1), \mathbf{X}_i, \boldsymbol{\theta}) \Pr(D_i(0), D_i(1)|\mathbf{X}_i, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ = &\int \prod_i \Pr(Y_i(0), Y_i(1)|S_i, \mathbf{X}_i, \boldsymbol{\theta}) \Pr(S_i|\mathbf{X}_i, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) d\boldsymbol{\theta}, \end{aligned} \tag{3}$$

for the global parameter $\boldsymbol{\theta}$ with prior distribution $\Pr(\boldsymbol{\theta})$.

A principal causal effect (PCE) can be any comparison of $Y(0)$ and $Y(1)$ conditional on S; this article focuses on the population average PCE,

$$\text{PCE}(d_0, d_1, \mathbf{x}; \boldsymbol{\theta}) \equiv E(Y_i(1) - Y_i(0)|S_i = (d_0, d_1), \mathbf{X}_i = \mathbf{x}; \boldsymbol{\theta}) \tag{4}$$

Based on (4), one can also easily calculate the overall PCE by averaging over the distribution of covariates: $\text{PCE}(d_0, d_1; \boldsymbol{\theta}) = \int \text{PCE}(d_0, d_1, \mathbf{X}; \boldsymbol{\theta}) \Pr(\mathbf{X}|d_0, d_1; \boldsymbol{\theta}) d\mathbf{X}$. The population average PCEs (4) have subtle but important differences from the finite population average estimands for the n sample units. Specifically, population average PCEs (4) do not depend on the association parameters (e.g., correlation) between $Y_i(0)$ and $Y_i(1)$, say ρ_{01}. Thus their posterior distribution of the remaining parameters will not depend on the prior distribution on ρ_{01} as long as ρ_{01} is a priori independent of the remaining parameters in $\boldsymbol{\theta}$. On the contrary, inference for the finite population causal estimands (as those considered in Jin and Rubin (2008)) for the sample units in the study would follow from the posterior distribution of Y_i^{mis} by predictive Bayesian inference, which generally involves association parameters (see Rubin (1990), Section 7, for a specific example, and Imbens and Rubin (1997)). Thus, by focusing on the population average estimands we can ignore the association between $Y_i(0)$ and $Y_i(1)$, and so consider $\boldsymbol{\theta}$ that does not include association parameters for $Y_i(0)$ and $Y_i(1)$.

2.2 A Bayesian semiparametric model

Equation (3) suggests that model-based PS inference usually involves two sets of models: One for the distribution of potential outcomes $Y(0)$ and $Y(1)$ conditional on the principal strata and covariates (hereafter referred to as the Y-model) and one for the distribution of principal strata conditional on the covariates (hereafter referred as the S-model). Since $D(0)$ and $D(1)$ are never jointly observed, the information on their association (and thus on the principal strata) is only implicitly imbedded in a Y-model where both $D(0)$ and $D(1)$ appear. Such a Y-model provides the structure to recover the relationship between Y_i^{obs} and D_i^{mis} given the observed D_i^{obs} and $\mathbf{X}_i$ (this is also evident from the data-augmentation step - step 1 in Section 2.3 - of our MCMC algorithm, where the distribution from which D_i^{mis} is drawn depends on D_i^{obs} through the distribution of Y_i^{obs}). The key association information of $D(0)$ and $D(1)$ induced from the Y-model can be potentially complex, which may not be adequately represented by a fully parametric S-model. This has motivated us to propose a flexible Bayesian nonparametric S-model

via the DPM models. Besides flexibility, an important advantage of the DPM is its characteristic clustering property, which gives opportunities to systematically coarsen and interpret the intermediate variables, as we shall illustrate later. Meanwhile, we can see that the Y-model is crucial to PS inferences and PCE estimates can be sensitive to its specification. We adopt flexible parametric Y-models, allowing for interactions, quadratic terms and heteroscedasticity and, in addition, carry out sensitivity analysis to model assumptions.

Parametric Y-models for potential outcomes. For the reasons explained at the end of Section 2.1, we specify $Y_i(0)$ and $Y_i(1)$ separately as $\Pr(Y_i(t)|S_i, \mathbf{X}_i; \boldsymbol{\beta}_t^Y)$ for $t = 0, 1$, where $\boldsymbol{\beta}_t^Y$ are the corresponding parameters.

Bayesian nonparametric S-models for principal strata. The potential intermediate outcomes need to be modeled jointly to be compatible with the underlying association implicitly imposed by the Y-model. Flexible S-models with strong structuring features are desirable here to capture subtle information from the possibly complex distribution of S_i. A Dirichlet process mixture (DPM) provides such an S-model:

$$\Pr(S_i|\mathbf{X}_i; \boldsymbol{\beta}^D) = \int K(D_i(0), D_i(1)|\mathbf{X}_i; \boldsymbol{\beta}^D, \boldsymbol{\gamma}) dG(\boldsymbol{\gamma}), \quad \text{with} \quad G \sim \text{DP}(\alpha G_0), \tag{5}$$

where the kernel $K(D_i(0), D_i(1)|\mathbf{X}_i, \boldsymbol{\beta}^D, \boldsymbol{\gamma})$ is a bivariate distribution (described later) and the probability measure G is generated from a Dirichlet process, DP(αG_0), with scalar concentration parameter α and base measure G_0 (Ferguson, 1974). A random probability measure G is sampled from DP(αG_0) if, for any Borel set partition of the space A, $\{A_1, \cdots A_k\}$, where G_0 (and G) are defined, the distribution of the realized probabilities $\{\Pr_G(A_1), \cdots, \Pr_G(A_k)\}$ follows a Dirichlet distribution, Dir$(\alpha \Pr_{G_0}(A_1), \ldots, \alpha \Pr_{G_0}(A_k))$. Small values of α imply less variation of a realized distribution G from the base measure G_0, where $E(G) = G_0$ in the sense that $E(G(A_1)) = G_0(A_1)$ for any Borel set A_1 in A.

The stick-breaking (SB) representation of the DP (Sethurman, 1994) shows that $G \sim$ DP(αG_0) may be constructed as

$$\text{G}(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\boldsymbol{\gamma}_h}(\cdot), \quad \boldsymbol{\gamma}_h \overset{iid}{\sim} G_0, \quad w_h = w'_h \prod_{k<h}(1 - w'_k), \; w'_h \overset{iid}{\sim} \text{Be}(1, \alpha), \tag{6}$$

where $\boldsymbol{\gamma}_h$ are called atoms and w_h are probabilities that sum up to 1, and δ_x denotes a point mass at x. The stick-breaking nature of the DP encourages decreasing weights, $w_i > w_j$ for $i < j$, *a priori* since $E[w_h] = 1/(1+\alpha) * (\alpha/(1+\alpha))^{h-1}$. Small α corresponds to sparser models, i.e., models with fewer nontrivial weights that provide a coarser approximation to G_0.

The SB representation shows that samples from a DP are discrete distributions, so the DP cannot be directly used as a prior distribution for continuous data models. However, the discrete atoms and associated weights may be used to define an infinite mixture of continuous distributions, as in (5). In the setting of continuous potential intermediate variables a convenient choice for the kernel of the mixture, $K(D_i(0), D_i(1)|\mathbf{X}_i; \boldsymbol{\beta}^D, \boldsymbol{\gamma})$ in (5) is a (truncated) bivariate Gaussian distribution,

$$K(D_i(0), D_i(1)|\mathbf{X}_i; \boldsymbol{\beta}^D, \boldsymbol{\gamma}) \propto \mathrm{N}((\eta_0 + \mathbf{X}_i\boldsymbol{\beta}_0^D, \eta_1 + \mathbf{X}_i\boldsymbol{\beta}_1^D)', \boldsymbol{\Sigma})1_A,$$

where $\boldsymbol{\gamma} = (\eta_0, \eta_1, \boldsymbol{\Sigma})$, and A is the support of $(D_i(0), D_i(1))$ in the possibly truncated distribution, e.g., $A = R^2$, or $A = [0,1] \times [0,1]$. Using the SB representation (6), (5) is equivalent to

$$\Pr(S_i|\mathbf{X}_i; \boldsymbol{\beta}^D) = \sum_{h=1}^{\infty} w_h c_h \mathrm{N}((\eta_{0h} + \mathbf{X}_i\boldsymbol{\beta}_0^D, \eta_{1h} + \mathbf{X}_i\boldsymbol{\beta}_1^D)', \boldsymbol{\Sigma}_h)1_A, \tag{7}$$

where the atoms $\boldsymbol{\gamma}_h = (\eta_{0h}, \eta_{1h}, \boldsymbol{\Sigma}_h)$ and associated weights (mixture probabilities w_h) are nonparametrically specified via $\mathrm{DP}(\alpha G_0)$, and c_h is the normalizing constant resulting from the truncation to support space A. The coefficients $\boldsymbol{\beta}^D$ are assumed to be common across mixture components, but this may be relaxed. This specification results in a flexible nonparametric mixture structure for the distribution of the principal strata that has support on a very large space of continuous bivariate distributions defined on A. In addition to flexibility, a natural byproduct of the mixture structure of the DPM is clustering, which is very appealing in the PS context. Clustering allows information to be shared locally between the S_i in the same cluster: Increased local information sharing is encouraged because the DPM naturally promotes sparse clustering. Parsimonious clustering provides opportunities for meaningful interpretation of the principal strata. Principal strata are essentially latent classes of subjects, only partially

observed, and the DPM allocates similar subjects into the same clusters. As will be illustrated in the applications, this automatic latent structure recovery may be treated as the continuous analogue to discrete PS analysis. The natural grouping of subjects in terms of principal strata and the subsequent interpretability is one essential advantage of the non-parametric DPM over existing parametric and semiparametric approaches.

The Bayesian model is completed by specifying prior distributions for the parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}^Y, \boldsymbol{\beta}^D, \alpha, G_0\}$. Specification of the DP concentration parameter α is important to the model's performance, as α directly governs the number of active components (i.e., components with nontrivial weights) in the DPM. Instead of fixing α to a specific value, we assume a flexible Gamma prior, Ga(a, b) with hyperparameters a, b, which has been widely used in the literature (e.g. Escobar and West (1995); Ishwaran and Zarepour (2000)). In the PS context, we prefer DPM models with smaller active components for better interpretability. This prior preference can be incorporated by setting a, b to smaller values. For example, we choose $a = b = 1$ in our applications. Sensitivity to the specification is then examined by repeating the analysis with different a, b values. The choice of G_0 suggests the support of $\boldsymbol{\gamma}$ to be explored, and generally depends on the range of the data being modeled. Standard choices for G_0 are conjugate inverse-Wishart IW$(q, \boldsymbol{\Sigma}_0)$ for $\boldsymbol{\Sigma}_h$ with small q and diagonal $\boldsymbol{\Sigma}_0$ expressing prior ignorance, and normal N(m, v^2) or uniform Unif(A) for (η_{0h}, η_{1h}). For the coefficients $\boldsymbol{\beta}^D, \boldsymbol{\beta}^Y$, we use a standard diffuse normal prior distribution N$(0, s^2 I)$ with very large variance s^2.

2.3 Posterior inference

All the PCEs are functions of the model parameters $\boldsymbol{\theta}$, so full Bayesian inference for the PCEs is based on the posterior distribution of $\boldsymbol{\theta}$ conditional on the observed data, which under SUTVA and strong ignorability, can be written as,

$$\Pr(\boldsymbol{\theta}|\mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{T}, \mathbf{X}) \;\propto\; \Pr(\boldsymbol{\theta}) \int\int \prod_i \Pr(Y_i(0), Y_i(1), D_i(0), D_i(1)|\mathbf{X}_i, \boldsymbol{\theta}) dY_i^{mis} dD_i^{mis},$$

where $\Pr(\boldsymbol{\theta})$ is the prior distribution of the parameters. Direct inference from the above distribution is in general not available due to the integrals over D_i^{mis} and Y_i^{mis}. But both $\Pr(\boldsymbol{\theta}|\mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{D}^{mis}, \mathbf{X})$ and $\Pr(\mathbf{D}^{mis}|\mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{X}, \boldsymbol{\theta})$ are generally tractable. $\Pr(\boldsymbol{\theta}|\mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{D}^{mis}, \mathbf{X})$ is the *complete-data* posterior distribution of $\boldsymbol{\theta}$, which is proportional to

$$\Pr(\boldsymbol{\theta}) \prod_{i=1}^{n} \Pr\left(Y_i(0)|S_i, \mathbf{X}_i; \boldsymbol{\beta}_0^Y\right)^{(1-T_i)} \Pr\left(Y_i(1)|S_i, \mathbf{X}_i; \boldsymbol{\beta}_1^Y\right)^{T_i} \Pr\left(S_i|\mathbf{X}_i; \boldsymbol{\beta}^D, \boldsymbol{\gamma}\right)$$

where *complete-data* here means *complete intermediate variable data*, i.e., considering the principal strata as observed, but not the full set of potential outcomes. The joint posterior distribution, $\Pr(\boldsymbol{\theta}, \mathbf{D}^{mis}|\mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{T}, \mathbf{X})$, can be obtained using a data augmentation approach for D^{mis} (Tanner and Wong, 1987). Inference for the joint posterior distribution then provides inference for the marginal posterior distribution $\Pr(\boldsymbol{\theta}|\mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{T}, \mathbf{X})$. Further discussion on Bayesian inference in PS can be found in Imbens and Rubin (1997) for the special case of all-or-none compliance.

Efficient Markov Chain Monte Carlo (MCMC) simulation for general DPM models was first discussed in Escobar (1994) and MacEachern (1994). Exact sampling from the DPM models can be obtained using the algorithms recently developed in, among others, Walker (2007) and Papaspiliopoulos and Roberts (2008). An approximate blocked Gibbs sampler based on a truncation of the SB representation of the DP is proposed in Ishwaran and Zarepour (2000). We adopt this sampling scheme in the posterior inference in this article, mainly for its conceptual simplicity and ease in implementation. Specifically, we first choose a conservative upper bound, $H < \infty$ on the number of mixture components potentially occupied by the subjects in the sample. Let $\boldsymbol{w} = \{w_1, \cdots, w_H\}$ denote the weights of all components. Then latent class indicators Z_i ($\in \{1, \cdots, H\}$) with a multinomial distribution, $Z_i \sim \text{MN}(\boldsymbol{w})$ are introduced to associate each observation i with a cluster h of the DPM. The marginal distribution implied by integrating out Z is the original approximation (based on H) to (7), so this augmentation expands the parameter space but does not change the original model specification. It does,

however, greatly simplify (7), so that for each individual i, conditional on $Z_i = h$,

$$\Pr(S_i|\mathbf{X}_i, Z_i = h; \boldsymbol{\beta}^D, \boldsymbol{\gamma}) = c_h \mathbf{N}((\eta_{0h} + \mathbf{X}_i\boldsymbol{\beta}_0^D, \eta_{1h} + \mathbf{X}_i\boldsymbol{\beta}_1^D)', \boldsymbol{\Sigma}_h)1_A.$$

As shown in Ishwaran and James (2001), an accurate approximation to the infinite limit (i.e., the exact DP) is obtained as long as H is chosen sufficiently large. Intuitively, this approximation is justified through the sparsity property of the DPM, which effectively provides an automatic selection mechanism for the number of active components $H^* < \infty$ in the SB representation, i.e., the number of nontrivial w_h. When the sample size is fixed, and only a small number of w_h are nonzero, the nonparametric behavior of the DPM can be approximated with a finite mixture model that truncates the SB representation at some large $H^* < H$. In our analysis, we use the following simple pragmatic procedure to choose H. First choose an initial upper bound, say $H = 10$, and then monitor the maximum index of the occupied components in the MCMC iterations. If all the iterations have a maximum index several units below H, then H is sufficiently high; otherwise, rerun the MCMC with an increased H (say, $H = 25, 50, 100...$). Repeat this process until a sufficient large H is found.

Using the DPM approximation, random draws from the posterior distribution of all the parameters in the complete-data model, $\boldsymbol{\theta}$ can be obtained via a Gibbs sampler with data augmentation (Tanner and Wong, 1987; Gelfand and Smith, 1990) as follows:

1. Given $\boldsymbol{\theta}$ and Z, draw each D_i^{mis} from

$$\Pr(D_i^{mis}|-) \propto \Pr\left(Y_i^{obs}|S_i, \mathbf{X}_i; \boldsymbol{\beta}_0^Y\right)^{1-T_i} \Pr\left(Y_i^{obs}|S_i, \mathbf{X}_i; \boldsymbol{\beta}_1^Y\right)^{T_i} \Pr(S_i|\mathbf{X}_i, \boldsymbol{\beta}^D, \boldsymbol{\gamma}_{Z_i}).$$

2. Given $\boldsymbol{\beta}^D$, w and S, draw each Z_i from a multinomial distribution with

$$\Pr(Z_i = h|-) \propto w_h \Pr\left(S_i|\mathbf{X}_i, Z_i = h; \boldsymbol{\beta}^D, \boldsymbol{\gamma}_h\right).$$

3. Given Z, set $w'_H = 1$, and for each $h \in \{1, \cdots, H-1\}$ draw w'_h from

$$\Pr(w'_h|-) = \mathbf{Be}\left(1 + \sum_{i:Z_i=h} 1, \alpha + \sum_{i:Z_i>h} 1\right),$$

and update $w_h = w'_h \prod_{k<h}(1 - w'_k)$.

4. Given Z, draw α from

$$\Pr(\alpha|-) \propto \Pr(\alpha) \prod_{h=1}^{H} \text{Be}\left(1 + \sum_{i:Z_i=h} 1, \alpha + \sum_{i:Z_i>h} 1\right).$$

5. Given $\boldsymbol{\beta}^D$, Z, and S, draw each $\boldsymbol{\gamma}_h$ from

$$\Pr(\boldsymbol{\gamma}_h|-) \propto G_0(\boldsymbol{\gamma}_h) \prod_{i:Z_i=h} \Pr\left(S_i|\mathbf{X}_i, Z_i; \boldsymbol{\beta}^D, \boldsymbol{\gamma}_h\right).$$

6. Given $\boldsymbol{\theta}$, Z, and S, draw $\boldsymbol{\beta}^D$ from

$$\Pr(\boldsymbol{\beta}^D|-) \propto \Pr(\boldsymbol{\beta}^D) \prod_{i=1}^{n} \Pr\left(S_i|\mathbf{X}_i, Z_i; \boldsymbol{\beta}^D, \boldsymbol{\gamma}_{Z_i}\right).$$

7. Given S, draw each $\boldsymbol{\beta}_t^Y$ $(t = 0, 1)$ from:

$$\Pr(\boldsymbol{\beta}_t^Y|-) \propto \Pr(\boldsymbol{\beta}_t^Y) \prod_{i:T_i^{obs}=t} \Pr\left(Y_i^{obs}|S_i, \mathbf{X}_i; \boldsymbol{\beta}_t^Y\right).$$

Cycling through the above steps provides correlated draws whose stationary distribution upon convergence is the joint posterior distribution of $\boldsymbol{\theta}$. Since the PCEs are functions of $\boldsymbol{\theta}$, posterior distribution of the PCEs can then be obtained by substituting $\boldsymbol{\theta}$ by their posterior draws, and point and interval estimates can be obtained by their empirical posterior counterparts.

3 Application to randomized trial with partial compliance

3.1 Data and models

To compare with the existing approaches, in this section we apply the proposed Bayesian semi-parametric model to the frequently studied data set from the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) in Efron and Feldman (1991), hereafter EF. The LRC-CPPT was a placebo-controlled double-blind randomized clinical trial that assigned 164 men to receive cholestyramine and 171 men to receive a placebo, and sought to examine the effect of cholestyramine in lowering cholesterol. Compliance with treatment assignment was

not enforced, but was monitored, and percent of compliance over the approximately 7 years of the study was reported. The cholesterol level of each subject was recorded before and after the study, and the outcome was the decrease in cholesterol level over the course of the study. No covariate information is available. An interesting question arising from this experiment concerns the causal effect of different doses of cholestyramine. Because cholestyramine has side-effects, the compliance distributions are different in the drug and placebo arms. Compliance to drug and compliance to placebo are different subject characteristics. Thus, comparisons between experimental arms within a given level of observed compliance to drug and placebo do not define proper causal effects and so do not capture drug efficacy or any dose-response relationship.

To handle the obvious violation of the *perfect blind assumption*, which asserts that compliance to placebo is identical, unit by unit, to compliance to drug, EF assumed a monotonically increasing deterministic function relating the compliances in the two groups and, therefore, deterministically *imputed* drug compliance for each control group member as a deterministic function of his placebo compliance. This assumption was believed to be overly restrictive by Jin and Rubin (2008), hereafter JR, who showed how to replace it with weaker and more plausible assumptions. Two analyses of the EF data were presented by JR. The first is a PS analysis that estimates finite population average PCEs for every combination of placebo and treatment compliance using a fully parametric Bayesian approach, where a side-effect monotonicity assumption was made. The second is a dose-response analysis which imposes additional assumptions to estimate dose-response curves. Here, we limit comparison to their first analysis, leaving the use of our approach to the estimation of dose-response relationships to future research. Bartolucci and Grilli (2011), hereafter BG, conducted a PS analysis on the same data, proposing a more flexible S-model based on a Plackett copula that does not require the side-effect monotonicity assumption and model the marginal distributions of the intermediate variables nonparametrically, using the empirical distribution function as an estimate of the marginal distribution of $D(0)$ and $D(1)$. They provide a likelihood based method to estimate

the model, and compare several alternative flexible parametric Y-models.

Let $D_i(0)$ denote the percentage of placebo taken by subject i when assigned to control and $D_i(1)$ denote the percentage of active treatment taken by subject i when assigned to treatment[1].

The Y-model specification should be based on subject-matter knowledge, exploratory analysis of the observed data, and practical consideration of model fitting given the large amount of missing data and the complexity of the model. Specifically, as shown in JR and BG (thus omitted here), the scatterplots of observed Y and D show a linear relationship between Y and D in the control group and a curved relationship in the treatment group. Within the class of parametric Y-models proposed by BG, we compare the implications of a range of models including those selected by BG and JR (detailed in Section 3.3), and choose the following Y-model,

$$Y_i(t)|D_i(0), D_i(1) \sim \mathbf{N}[\mu_t(D_i(0), D_i(1)), \exp(\sigma_t^2(D_i(0), D_i(1)))], \quad t = 0, 1, \tag{8}$$

with the mean and variance depending on $D_i(0)$ and $D_i(1)$ as follows,

$$\begin{aligned}
\mu_0(D_i(0), D_i(1)) &= \beta_{00}^Y + \beta_{01}^Y D_i(0); \\
\mu_1(D_i(0), D_i(1)) &= \beta_{10}^Y + \beta_{11}^Y D_i(0) + \beta_{12}^Y D_i(1) + \beta_{13}^Y D_i(0) D_i(1); \\
\sigma_0^2(D_i(0), D_i(1)) &= \lambda_0; \quad \text{and} \quad \sigma_1^2(D_i(0), D_i(1)) = \lambda_0 + \lambda_1 D_i(1).
\end{aligned}$$

The key model assumption lies in the mean function of the potential outcome $Y_i(1)$, μ_1, where both $D_i(1)$ and $D_i(0)$ enter as regressors and thus indirectly specify the unobserved association between $D_i(1)$ and $D_i(0)$. Unlike BG, here we allow the intercept terms β_{00}^Y and β_{01}^Y to differ in μ_0 and μ_1, and so do not rule out some placebo effects. But as BG, we impose the constraint $\beta_{01}^Y = \beta_{11}^Y$, assuming that compliance to placebo is a baseline characteristics that affects the

[1]It is worth noting the slight abuse of notation here: Variable D does not have the same meaning under treatment and under control. For this reason, JR use a different notation for the two compliance variables. Also because of this, standard instrumental variable-noncompliance exclusion restriction-type assumptions, that would rule out any effect of assignment on individuals with $D_i(0) = D_i(1)$, are not plausible. The only exclusion restriction assumption that may be plausible is $Y_i(0) = Y_i(1)$ for $\{i : D_i(0) = D_i(1) = 0\}$, i.e., for individuals who take no placebo and no active treatment, assignment to treatment should not have any effect on the outcome.

mean potential cholesterol under treatment and control in the same way, and we allow compliance to placebo to have an interaction with compliance to treatment. Alternative Y-model specifications are compared in Section 3.3.

We assume the following diffuse prior distributions for the parameters:

$$\Pr(\lambda_0) = \mathrm{N}(3, 1), \quad \Pr(\lambda_1) = \mathrm{N}(0, 1.5^2), \quad \Pr(\boldsymbol{\beta}^Y) = \mathrm{N}(\mathbf{0}, 20^2 I_5),$$

where $\boldsymbol{\beta}^Y = (\beta^Y_{00}, \beta^Y_{01}, \beta^Y_{10}, \beta^Y_{12}, \beta^Y_{13})$ and I_5 is a 5-dimensional identity matrix. The specifications for λ_0 and λ_1 are vague, and will allow recovery of the variances observed in the two treatment arms.

For the S-model, the DPM model (7) without covariates is assumed, with DP parameters

$$A \equiv [0, 1] \times [0, 1], \quad G_0 = \mathrm{N}((.5, .5)', .25^2 I_2), \mathrm{IW}(2, I_2) \quad \Pr(\alpha) = \mathrm{Ga}(1, 1).$$

As in BG, JR's side-effect monotonicity assumption, $D_i(1) < D_i(0)$, is not imposed in our S-model. To express prior ignorance, the prior covariance matrix of the Inverse-Wishart prior for G_0 is set to be an identity matrix and the prior sample size is set to be the minimal possible integer, 2. Complete details of posterior sampling, based on the above model specification, are available as the online supplementary material at:
`http://www.stat.duke.edu/~ fl35/PSDP/.`

3.2 Results

The parametric Y-model (8) and nonparametric S-model (7) were jointly fitted to the LRC-CPPT data. Five parallel MCMC chains of 205,000 iterations with the first 5,000 as burn-in period were run, each having different starting values. None of the chains showed signs of adverse mixing and all chains lead to highly similar posterior summary statistics. Using the pragmatic procedure introduced in Section 2.3, the truncation level $H = 30$ is deemed to be adequate for DPM approximation in this analysis.

Table 1 provides the posterior medians and 95% credible intervals for the coefficients in the Y-model (8), and the corresponding MLE and standard errors under the copula-likelihood

approach of BG. The point estimates of $\beta^Y_{00}, \beta^Y_{01}, \lambda_0, \lambda_1$ are similar between the two methods, as their estimation are mostly based on the observed marginal distributions of $Y(0)$ and $Y(1)$, with DPM providing slightly tighter intervals than the copula approach. However, there is a seemingly large discrepancy in the point estimates of $\beta^Y_{12}, \beta^Y_{13}$, with the DPM-based interval estimates being half of those produced by the copula. The sum of $\beta^Y_{12} + \beta^Y_{13}$ is, instead, comparable between methods. Notice that the term of $\beta^Y_{12} D_i(1) + \beta^Y_{13} D_i(0) D_i(1)$ in the μ_1 function equals $\beta^Y_{12} + \beta^Y_{13}$ when $D_i(1) = D_i(0) = 1$. Since the majority of the subjects have high compliance (close to 1) under both assignments (as shown in Figure 1(a)), this suggests that the marginal distributions of $Y_i(0)$ and $Y_i(1)$ are similarly estimated by both DPM and copula methods, but the estimates of individual coefficients from the DPM have less uncertainty than those from the copula. The improved precision is unlikely to result from the diffused prior on $\boldsymbol{\beta}^Y$ as the prior variances on $\boldsymbol{\beta}^Y$ are several times larger than the posterior ones.

	DPM		Copula	
coefficient	post. median	95% CI	MLE	95% CI
β^Y_{00}	-0.71	(-5.19, 3.74)	-0.27	(-5.72, 4.89)
β^Y_{10}	-0.69	(-6.31, 4.69)	-0.27	(-5.72, 4.89)
$\beta^Y_{01} = \beta^Y_{11}$	11.87	(5.95, 17.74)	11.24	(4.70, 18.05)
β^Y_{12}	22.30	(9.36, 35.17)	-21.88	(-102.2, 9.4)
β^Y_{13}	23.02	(8.40, 37.65)	73.36	(38.33, 155.84)
λ_0	5.28	(5.09, 5.48)	5.26	(5.05, 5.43)
λ_1	1.35	(0.96, 1.74)	1.16	(0.19, 1.56)

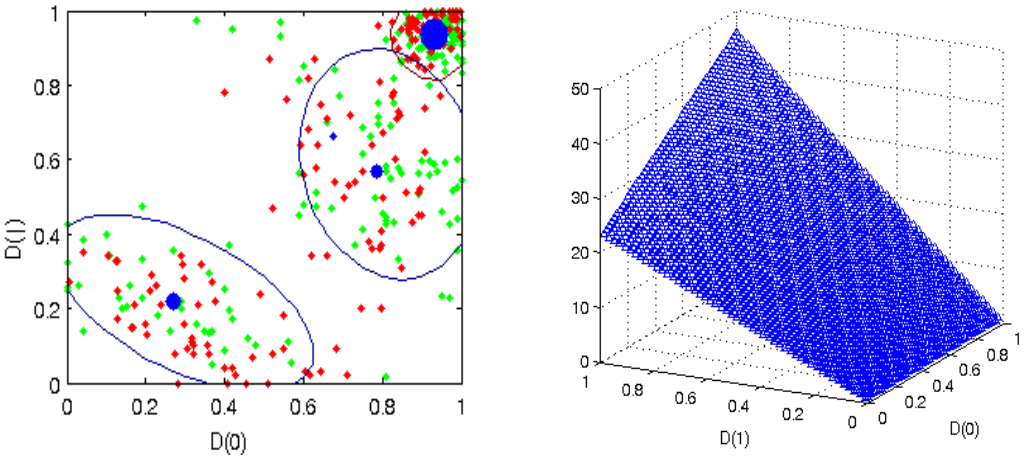
Table 1: Coefficients in the outcome Y-model (8) as estimated using the Bayesian DPM S-model, where posterior medians and 95% credible intervals are shown; and the frequentist copula S-model of BG, where MLE and bootstrap intervals are shown.

To further understand the improved precision in the Y-model, it is useful to look at the results of the DPM S-model. As shown in the scatter plot of a representative posterior draw of the

principal strata $(D_i(0), D_i(1))$ in Figure 1(a), there are three predominant clusters: the largest cluster (45% of all units) in the upper right corner, a second largest cluster (30%) in the right middle, and a third (25%) in the lower left corner. Interestingly, these latent clusters have some analogy to the principal strata in the binary PS classification (Angrist et al., 1996), but with slight different interpretation since $D(0)$ is placebo compliance and $D(1)$ is drug compliance. Here, the cluster in the upper right corner comprises the *full* compliers, i.e., units who take the amount prescribed by the protocol, both under treatment and under control. The cluster in the lower left corner includes the never-takers, units who never take what they are assigned to take, neither under treatment, nor under control. Their noncompliance to the protocol is most likely due to behavioral reasons rather than to reasons related to possible side-effects of the active treatment. Units in the right middle cluster (with $D(1) < D(0) < 1$) are the most difficult to classify with an analogue in the binary case; we will call them *partial* compliers. They are units who are generally willing to comply to the protocol if they have no side effects (i.e., under control), but probably experience negative side-effects under treatment and thus do not take the prescribed amount of the drug.

The same cluster structure was consistently observed in all the MCMC chains, and the majority of imputed D_i^{mis} maintained a single cluster membership. Thus, for the LRC-CPPT data, there was strong evidence of relevant latent structure recovery, and this information was used to inform D_i^{mis} locally on the basis of cluster membership. The reduced variability in estimating the unobserved D_i^{mis} lead to more precise estimates of the Y-model. Our S-model did not assume side-effect monotonicity, but appears to support the assumption. Figure 1(b) shows the posterior medians of all the PCEs over the entire $(D(0), D(1))$ space. The PCE surface is smoothly increasing as the compliance increases in both assignment arms, suggesting better compliance behavior is associated to larger overall reduction in cholesterol level.

Comparison with the results of JR and BG is made on the estimated PCE at four selected principal strata $S = (d_0, d_1)$, including the stratum of "median complier" under both assignments, $S = (0.89, 0.70)$. The comparison is displayed in Table 2 which includes the posterior



(a) (b)

Figure 1: (a) A single presentative posterior draw of $(D(0), D(1))$ based on the DPM S-model. The green and red dots correspond to units in control and treatment groups, respectively. Each component is labeled with a blue dot that is proportional to the components mass contribution. (b) Median PCE over the entire $(D(0), D(1))$ space.

medians and 95% credible intervals for the PCEs under the DPM approach, and the corresponding estimates from the fully Bayesian parametric approach of JR and the copula approach of BG. Interval estimates were obtained by a bootstrap algorithm in BG: They only reported the confidence interval for stratum (0.89,0.7), and the PCEs are 0 for strata (1,0) and (0,0) by their model specification. The results are comparable across methods for the strata where most of the observed compliance are presented, while the DPM approach provides tighter estimate intervals than the fully parametric and copula approaches, again highlighting the improved precision that results from the clustering structure imposed by the DP. Slight discrepancies are observed at extreme compliance levels (e.g., (0,0), (1,0)), with our approach pointing more clearly to a likely absence of placebo effects.

$S = (d_0, d_1)$	DPM		Parametric (JR)		Copula (BG)	
$(1, 1)$	45	(38, 52)	50	(39, 59)	51	-
$(0.89, 0.70)$	29	(25, 34)	24	(17, 30)	30	(22, 39)
$(1, 0)$	0	(-6, 6)	-13	(-42, 27)	0	-
$(0, 0)$	0	(-6, 6)	5	(-6, 16)	0	-

Table 2: Estimated PCE for selected principal stratum (D_0, D_1) using the DPM approach, the fully parametric approach of JR, and the copula approach of BG.

3.3 Sensitivity analysis

We examined the sensitivity of our analysis to a) the specification of the prior distributions, and b) the parametric specification of the Y-model.

For the hyper prior Ga(a, b) of α, the strength parameter of the DP that controls the number of active components, we refit the models using hyperparameters for (a, b) of (2,2), (5,5), (1,5), (5,1). The resulting estimates for coefficients and PCEs showed only minimal difference from our initial (1,1) specification. We prefer DP models with small number of active components, and within this general preference, the exact specification was not crucial. As for the priors for the regression coefficients, we use standard diffused priors that do not drive analysis results.

For the Y-model, we conduct a sensitivity analysis to the key specification of the mean function μ in (8), denoted as M_0. We consider the following alternative models:

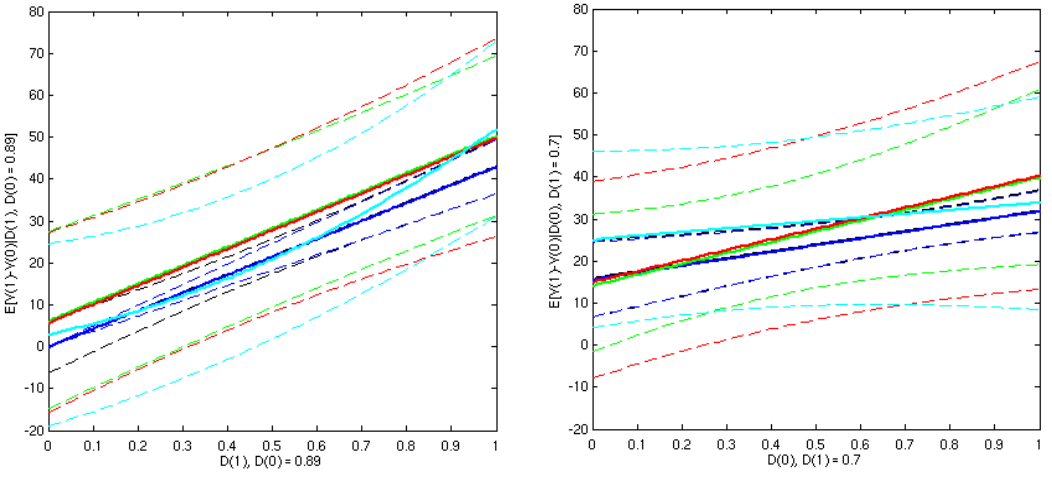
1. M_1: Same as M_0, but impose $\beta^Y_{00} = \beta^Y_{10}$. M_1 implies $\mu_0(d, 0) = \mu_1(d, 0)$, for $d \in [0, 1]$, that is, it rules out any placebo effect for subjects who take zero dose of the active treatment, which maybe questionable in practice. This is the Y-model adopted by BG.

2. M_2: Same as M_0, but allow $\beta^Y_{01} \neq \beta^Y_{11}$

3. M_3: Add a $D(1)$ term to the μ_0 function in M_2 and allow the coefficients $\beta^Y_{02} \neq \beta^Y_{12}$. This model allows $Y(0)$ to be affected also by $D(1)$, and thus by the basic principal strata.

4. M_4: Change the interaction term in μ_1 in M_3 to a quadratic term. This is essentially the Y-model adopted by JR. In addition, impose the side-effect monotonicity $D_i(1) < D_i(0)$ as JR.

We focus on comparing the causal estimands of primary interest - the population PCEs defined in (4) under the above models[2]. Figure 2 gives the posterior means (solid lines) and 95% credible intervals (dashed lines) of cross sections of PCE surfaces estimated from the above models, with each model being labeled by a different color: Panel (a) displays the estimates of PCE$(0.89, d)$ with d varying from 0 to 1, where 0.89 is the median of the observed $D(0)$'s, and panel (b) displays those of PCE$(d, 0.70)$, where 0.70 is the median of the observed $D(1)$'s.



(a) (b)

Figure 2: Posterior median PCEs (solid lines) and 95% credible intervals (dashed lines) under different Y-models for the EF data, labeled by different colors: black M_0, blue M_1, green M_2, red M_3, turquoise M_4. Figure (a) fixed $D(0) = 0.89$; (b) fixed $D(1) = 0.70$.

Figure 2 shows that the posterior medians of both PCE$(0.89, d)$ and PCE$(d, 0.70)$ are highly similar across the models, while the credible interval gets wider as the Y-model gets more flexible, as expected. Thus, it is reasonable to conclude from the sensitivity plots that the proposed PCE analysis for the EF data is not very sensitive to the parametric Y-model specification. One

[2]Parameter estimates are available from the authors upon request.

possible reason for the insensitivity is that the LRC-CPPT is a randomized trial, therefore the key unconfoundedness assumption is likely to hold, implying that the distribution of the principal strata is the same, in expectation, in the two treatment arms. As we shall see in the next application, results can be more sensitive to model specification in observational studies, where unconfoundedness may be more questionable.

4 Application to the Swedish National March Cohort

4.1 Data and models

This section examines the effect of physical activity (PA) on cardiovascular disease (CVD) as it relates to body mass index (BMI), using the observational Swedish National March Cohort (NMC). The NMC was established in year 1997, when 300,000 Swedes participated in a national fund-raising event organized by the Swedish Cancer Society. Every participant was asked to fill in a questionnaire that included items on known or suspected risk factors for cancer and CVD. Questionnaire data were obtained on over 43,880 individuals. Using the Swedish patient registry, these individuals were followed for the period from year 1997 to 2004, and each CVD event was recorded. Further details on the NMC can be found in Lagerros et al. (2009).

The question of scientific interest here is the extent of a causal effect of PA on CVD risk mediated or not mediated through BMI. The principal stratum with respect to the intermediate variable BMI is the joint potential value of BMI for an individual under high and low exercise regimes. The PCEs in the principal strata consisting of individuals whose BMI remains the same, or approximately so, regardless of exercise can be interpreted as the principal strata *direct* effect of exercise on CVD not mediated through BMI. Similarly, we can define the principal strata *mediated* effect as the PCEs in the principal strata of individuals whose BMI would change due to exercise[3].

[3]Once the underlying continuous structure has been estimated, one could look at average PCEs in some union

Sjölander et al. (2009), hereafter SJ, analyzed the NMC data using PS, where each subject was classified as either a "low-level exerciser" (T = 0) or a "high-level exerciser" (T = 1) based on self-reported history of PA; obese (D = 1) or not obese (D = 0) based on baseline BMI in year 1997 dichotomized at cutoff point 30; and "with disease" (Y = 1) or "without disease" (Y = 0) based on if the subject had at least one CVD event recorded during follow-up. Age is a strong confounder in this setting, and is the sole covariate reported in SJ. In what follows, in order to compare with SJ analysis, we also assume that T is ignorable given age. The principal strata *direct* effects are the main causal estimands in SJ, and they found evidence for beneficial *direct* effects in strata where $D(0) = D(1)$. In the current analysis, we follow the definition of T and Y in SJ, but analyze D (BMI) in its original continuous scale and let $\mathbf{X}_i = (X_{i1}, X_{i2})$ be the centered age and square of age. In addition, we will investigate PCEs in strata where $D(0) = D(1)$ and in strata where $D(0) \neq D(1)$. Of the participants, 38,349 were "high-exercisers" and 2,956 were "low-exercisers". The former included 2,262 cases of CVD, while the latter included 172 cases. The distribution of age is similar in both arms, with the $T = 1$ arm being slightly older on average. The median age is 49.3 years in $T = 0$ arm and 52.2 in $T = 1$. The distribution of BMI is right skewed in both arms, with a heavier tail in the low-exercise arm, as seen in Figure 3 top left panel. The median BMI in the $T = 0$ and $T = 1$ arms are 24.8 and 24.0, respectively. A quantile-quantile plot of BMI in the two arms (Figure 3 top right panel) suggests the high-exercisers have lower BMI than the low-exercisers on average. There appears to be a strong positive correlation between CVD incidence and BMI (Figure 3 bottom right panel), with a larger variance of CVD in the low-exerciser arm. Likelihood ratio tests on the observed marginal distribution of $Y(0)$ and $Y(1)$ suggest that both age and BMI are significant predictors of CVD risk in both arms.

Based on the above exploratory data analysis, and assuming that it is not only the absolute

of the strata where $|D(0) - D(1)| < \epsilon$ and interpret those as principal strata *direct* effects, i.e., PCEs in unions of principal strata where PA has *little* effect of BMI. Because of the varying and arbitrary choice of ϵ, and to limit exposition, we only report PCEs where for $D(0) = D(1) = s$ for different values of s.

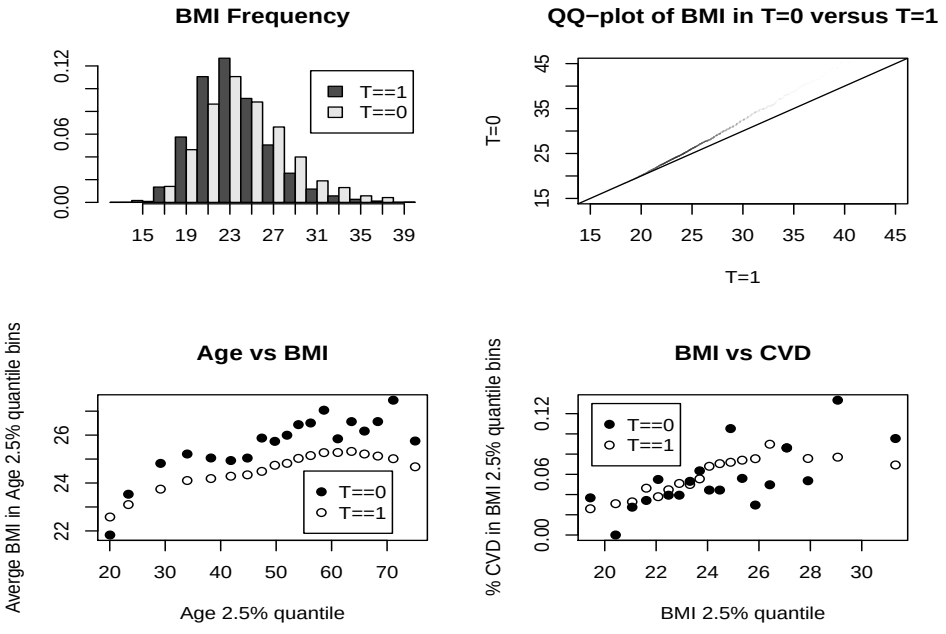


Figure 3: Top left: frequency of BMI in $T = 0, 1$ arms. Top right: quantile-quantile plot of BMI in $T = 0$ versus $T = 1$ arms. Bottom left: scatter plot of age versus BMI in $T = 0, 1$ arms. Bottom right: scatter plot of BMI versus CVD incidence in $T = 0, 1$ arms.

decrease (or increase) of BMI due to PA that matters, but also its relative size, we choose the following generalized linear Y-model,

$$\begin{aligned} \text{logit}\{\Pr(Y_i(0) = 1|S_i, \mathbf{X}_i)\} &= \beta_{00}^Y + \beta_{01}^Y X_{i1} + \beta_{02}^Y D_i(0) \\ \text{logit}\{\Pr(Y_i(1) = 1|S_i, \mathbf{X}_i)\} &= \beta_{10}^Y + \beta_{11}^Y X_{i1} + \beta_{12}^Y D_i(0) + \beta_{13}^Y D_i(1) + \beta_{14}^Y \frac{D_i(0)}{D_i(1)}, \end{aligned} \quad (9)$$

where the constraint $\beta_{02} = \beta_{12}$ is imposed. Here, the $Y(0)$ model specifies the relationship between CVD and BMI when an individual does not exercise. In the $Y(1)$ model, the term $\beta_{12}^Y D_i(0) + \beta_{13}^Y D_i(1)$ can be re-written as $(\beta_{12}^Y + \beta_{13}^Y) D_i(1) + \beta_{12}^Y (D_i(0) - D_i(1))$. Thus, $\beta_{12}^Y + \beta_{13}^Y$ can be interpreted as the baseline effect of BMI on CVD when an individual exercises, while β_{12}^Y and β_{14}^Y represent the additional linear and nonlinear (ratio) effect of the change in BMI due to exercise on CVD, respectively. The prior distribution for $\boldsymbol{\beta}^Y = (\beta_{00}^Y, \beta_{01}^Y, \beta_{02}^Y, \beta_{10}^Y, \beta_{11}^Y, \beta_{13}^Y, \beta_{14}^Y)$ is set to be a diffused normal $\Pr(\boldsymbol{\beta}^Y) = \text{N}(\mathbf{0}, 25^2 I_7)$.

Scatterplot of average BMI versus age shows a positive and curvilinear relationship between age and BMI (Figure 3 bottom left panel). Likelihood ratio tests on the observed marginal distribution of $D(1)$ and $D(1)$ suggest that both age and square of age are significant predictors of BMI in both arms. Thus we assume the following DPM S-model,

$$\Pr((D_i(0), D_i(1))|\mathbf{X}_i, \boldsymbol{\beta}^D) = \sum_{h=1}^{\infty} w_h c_h \mathrm{N}\left(\begin{pmatrix} \eta_{0h} + X_{i1}\beta_{01}^D + X_{i2}\beta_{02}^D \\ \eta_{1h} + X_{i1}\beta_{11}^D + X_{i2}\beta_{12}^D \end{pmatrix}, \boldsymbol{\Sigma}_h\right) 1_A, \quad (10)$$

with $A = \{(D_i(0), D_i(1)) : 0 < D_i(0), D_i(1) < 100\}$. Specifications for the DP are $G_0 = \mathrm{N}((25, 25)', \boldsymbol{\Sigma}_0)\mathrm{IW}(2, 3^2 I_2)$, with $\sigma_0^2 = \sigma_1^2 = 5^2$, and $\Pr(\alpha) = \mathrm{Ga}(1, 1)$. The prior distribution for $\boldsymbol{\beta}^D$ is a diffused normal $\Pr(\boldsymbol{\beta}^D) = \mathrm{N}(\mathbf{0}, 10^2 I_4)$. Similarly to the partial compliance application, the posterior distribution of the parameters in the complete data model can be obtained via a Gibbs sampler with data augmentation, details of which are thus omitted.

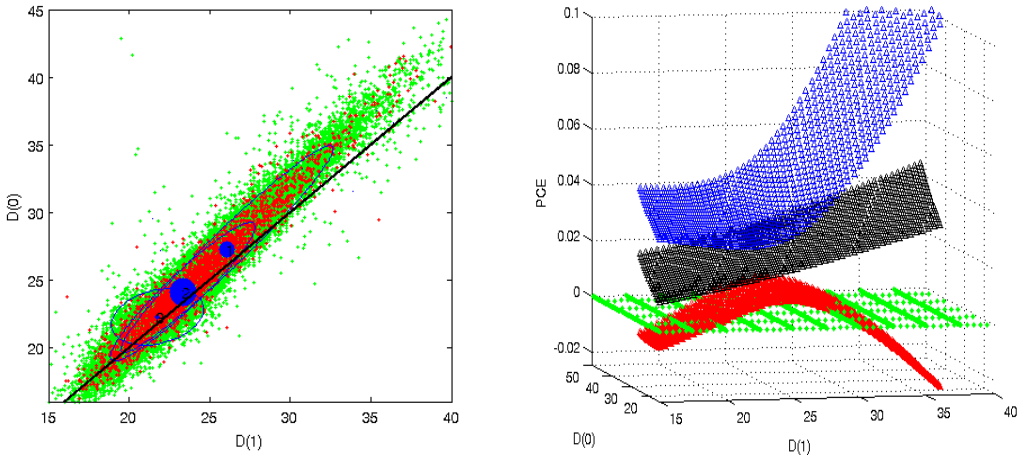
4.2 Results

The Y-model (9) and DPM S-model (10) are jointly fitted to the NMC data. Similarly as before, five parallel MCMC chains of 205,000 iterations with different starting values were run, with the first 5,000 as burn-in. Mixing of the chains was determined to be adequate and all chains lead to similar posterior summary statistics. Using the procedure discussed in Section 2.3, the truncation level $H = 30$ is deemed to be adequate for DP approximation here.

Estimates of coefficients in models (9) and (10) (not displayed here but available from the authors upon request) suggest that a positive association between CVD incidence and both age and baseline BMI without exercise, as well as an sizable reduction in CVD associated with a reduction in BMI due to PA (a negative posterior median of β_{14}^Y).

A representative posterior draw of principal strata $(D_i(0), D_i(1))$ along with the DPM configuration from the S-model (10) is displayed in Figure 4(a). There are two predominant clusters that are consistently found throughout all analyses: Component 1 in the middle of the 45^o line that consists of around 80% individuals whose BMI is stable regardless of PA; and

component 2 above the 45^o line that consists of around 15% individuals whose BMI decreases with PA. The precise configuration of remaining components varies in different MCMC chains, but still consistently suggest that the remaining individuals are people who have a even larger reduction in BMI as a result of PA.



(a) (b)

Figure 4: (a) A representative posterior draw of principal strata S_i under the DPM S-model. Each component is labeled with a number and a blue dot representing its mass contribution with its size being proportional to the corresponding component weight. The solid line is the 45^o line. (b) Median surface and point-wise 95% credible intervals for the PCE, over the relevant space of $(D(0), D(1))$ for individuals 10 years above the median age (60 years old). The green surface is the reference surface of PCE = 0.

Figure 4(b) shows posterior medians and 95% credible intervals for PCEs over the plausible range of $D(0)$ and $D(1)$ for individuals who are 60 years old. The PCE surface increases smoothly with both $D(0)$ and $D(1)$, suggesting that the causal effect of exercise in reducing the probability of developing CVD increases as one's BMI increases. Table 3 provides PCEs for principal strata $S_i = (d_0, d_1)$ that are of scientific interest. BMI values of 18.5, 25, 30, 35 are the standard cutoff points of underweight, overweight, class I obese and class II obese, respectively. Since there are very few individuals with BMI below 18.5, we present the PCEs

for BMI equal 20 rather than 18.5. Evidence on a *direct* effect of PA on CVD can be drawn from PCEs on the 45^o line. We can see that these PCEs increase as age and BMI increase. For example, for a person whose BMI is 20 no matter whether he exercises, the reduction in CVD risk due to exercise is 0.53% and 1.33% when he is 50 and 60 years old, respectively, and the reduction increases to 0.77% and 1.87% respectively if his BMI is always 30. This means that even if exercise does not reduce the BMI, it does reduce the risk of CVD, and the benefit is bigger for older and heavier population. Our results also suggest that PCEs of PA on CVD are even bigger if mediated through BMI. For example, for a person whose BMI reduces from 30 to 25 as a result of exercise, the reduction in CVD risk is 0.92% and 2.23% when he or she is 50 and 60 years old, respectively; and the corresponding reduction is 1.09% and 2.57% respectively for a person whose BMI reduces from 35 to 30 due to exercise.

	Age $= 50$		Age $= 60$	
$S = (d_0, d_1)$	median	95% cred. ints.	median	95% cred. ints.
(20, 20)	0.53	(0.01, 1.18)	1.33	(0.03, 2.93)
(25, 25)	0.64	(0.23, 1.11)	1.58	(0.57, 2.72)
(30, 30)	0.77	(-0.01, 1.64)	1.87	(-0.02, 3.91)
(35, 35)	0.94	(-0.64, 2.81)	2.21	(-1.53, 6.48)
(25, 20)	0.79	(0.09, 1.48)	1.95	(0.23, 3.65)
(30, 25)	0.92	(0.34, 1.60)	2.23	(0.83, 3.82)
(35, 30)	1.09	(0.08, 2.45)	2.57	(0.20, 5.61)

Table 3: Posterior medians and 95% credible intervals for the percent PCE, $E(Y(0)-Y(1)|S = (d_0, d_1)) \times 100$, for selected principal strata S at age of 50 and 60 years.

The PCE results here do not contradict the findings in SJ. However, analysis of PCE based on continuous BMI offers a more refined picture of the causal mechanism among PA, BMI and CVD risk than that based on dichotomized BMI. In addition, our continuous analysis does not rely on some standard identifying assumptions made in in the case of binary D, as the

monotonicity assumption (i.e., $D_i(0) > D_i(1)$) imposed in SJ. Even if approximately 18% of the posterior draws of $(D_i(0), D_i(1))$ do not adhere to monotonicity when it was not enforced (Figure 4(a)), these draws are relatively close to the 45^o line, and the PCE surface estimate for the NMC data are rather robust to deviations from this assumption.

4.3 Sensitivity analysis

As in the previous application, we examined the sensitivity of our analysis to the prior distribution and specification of the Y-model. We considered the the same values that we used in Section 3.3 for the hyperparameters (a, b) in the Gamma hyper prior for α, Ga(a, b). The PCE estimates were again robust to these specifications. Denote the Y-model (9) as M_0; we fit the following alternative Y-models to the NMC data:

1. M_1: Same as M_0, but allow $\beta^Y_{01} \neq \beta^Y_{11}$.

2. M_2: Add a $D(1)$ term to the μ_0 function in M_1, allowing $Y(0)$ to be affected by $D(1)$.

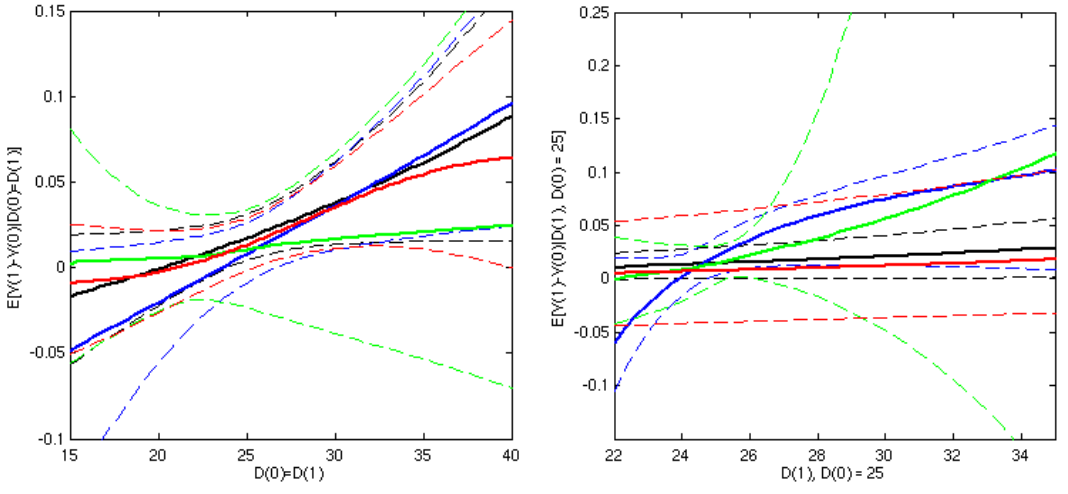
3. M_3: Change the ratio term $D(1)/D(0)$ in μ_1 in M_2 to an interaction term $D(0)D(1)$ and also add a term $D(0)D(1)$ in μ_0.

The population average PCEs defined in (4) are compared. Figure 5 gives the posterior means (solid lines) and 95% credible intervals (dashed lines) of two cross sections of the PCE surfaces estimated from the above models, with each model being labeled by a different color: Panel (a) displays the estimates of PCE$(D(0) = D(1) = d)$ with d varying from the common BMI range of 18 to 35, and panel (b) displays those of PCE$(D(0) = 25, d)$ where 25 is the median of the observed $D(0)$.

The sensitivity plots show that the PCE estimates from M_0, M_1 and M_3 are very similar, while larger variability is observed in M_2. In particular, despite the similarity in the posterior median of the PCEs between M_2 and other models, M_2 results in much wider credible intervals for subjects with larger $D(1)$ (> 26). Moreover, the MCMCs from M_1 and M_2 display poor mixing. This is an evidence that the NMC data contain rather weak information to

identify the causal estimands, thus reasonable parsimonious model constraints are necessary for credible estimation. Our preference for M_0 over M_3 was mainly due to its interpretability, as $D(1)/D(0)$ has the natural meaning of the ratio of BMI change due to PA; as said above, it appears to be rather plausible that the relative decrease (or increase) of BMI should be an important determinant of CVD.

Overall, the PS analysis in the NMC data displays certain sensitivity to the Y-model specification. One likely reason is the possible violation of the uncounfoundedness assumption: the NMC is an observational study and the only covariate information that is available to our analysis is age, but there can be many remaining important confounders, such as sex, diet, smoking behavior, etc. Fitting the models with more covariates is expected to reduce the sensitivity. The sensitivity to deviations from the unconfoundedness assumption in the PS framework has been formally explored elsewhere (Schwartz et al., 2011).



(a) (b)

Figure 5: Posterior median PCEs (solid lines) and 95% credible intervals (dashed lines) of subjects at age 60 under different Y-models for the NMC data, labeled by different colors: black M_0, blue M_1, green M_2, red M_3. Figure (a) is with $D(1) = D(0)$; (b) with $D(0) = 25$, which is the median of the observed $D(0)$.

5 Conclusion

We have proposed a Bayesian semiparametric approach to conduct causal inference in the presence of continuous intermediate variables using the PS framework, where causal estimands are defined on latent subgroups of units, namely the principal strata. Even though never jointly observed, modeling the joint distribution of the two potential intermediate outcomes is critical for PS inference. The key element in our proposal is to model the principal strata by the nonparametric DPM model, which has proved to offer a new approach to causal inference with intermediate variables that has advantages in both inference and interpretability. Specifically, it allowed us to a) overcome the limits of the dichotomization approach, b) exploit the clustering properties of DPMs to explore and interpret the latent structure of the data, c) easily quantify posterior uncertainty on PCEs, without relying on asymptotic approximations.

We have illustrated our approach using the randomized LRC-CPPT data, obtaining comparable point estimates but more precise interval estimates than those from existing approaches. We also applied our proposal to the more challenging observational NMC study to investigate the causal mechanism between PA, BMI and CVD risk. As reflected in these two applications, PS is a general framework that can be used to represent and tackle intrinsically different problems. While some PS analyses may be mathematically equivalent, they can differ on fundamental issues of study design, on interpretation, on the specific (union of) principal strata of interest, and on the potential identifying structural and modeling assumptions. Principal stratification is one of the possible ways to conceptualize the mediatory role of an intermediate variable in the treatment-outcome relationship. An alternative approach focuses on what would happen to the treatment-outcome relationship under interventions on the intermediate variable, and defines direct and indirect causal effects by using the concept of so-called *a priori counterfactual* values of outcomes that would have been observed under assignment to a given treatment level and if the intermediate variable were somehow simultaneously forced to attain a predetermined value. Using these concepts, Robins and Greenland (1992) and Pearl

(2001) give definitions for controlled direct effects and natural direct and indirect effects based on hypothetical interventions on the intermediate variable. While in some contexts one can at least conceive the idea of manipulating the intermediate variable, in some others, as with BMI, it is not that obvious how an experiment could be conceived, where the values of the intermediate variable could be controlled. This motivated our preference for a PS analysis for the NMC data.

Extension of our method to relax some of the assumptions in the Y-model by, e.g., specifying a DPM also for the conditional distribution of the potential outcomes, can be explored. Bayesian model selection or averaging methods can also be applied to the model building process.

Acknowledgments

Scott Schwartz is postdoctoral fellow in Department of Statistics Texas A&M University, Fan Li is assistant professor in Department of Statistical Science, Duke University (email:fli@stat.duke.edu), Fabrizia Mealli is professor in Dipartimento di Statistica, Università di Firenze, Italy. We thank Hal Stern, the associate editor and three reviewers for their constructive comments and suggestions that helped to improve the manuscript significantly, and Don Rubin and Leonardo Grilli for stimulating discussions. We also thank Brad Efron for providing the LRC-CPPT data, and Olof Nyren, Rino Bellocco and Arvid Sjölander for providing the Swedish NMC data.

References

JD Angrist, GW Imbens, and DB Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

F Bartolucci and L Grilli. Modeling partial compliance through copulas in the principal stratification framework. *Journal of the American Statistical Association*, 106:469–479, 2011.

B Efron and D Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86:9–17, 1991.

MD Escobar. Estimating normal means with a Dirchlet process prior. *Journal of the American Statistical Association*, 89:268–277, 1994.

MD Escobar and M West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

TS Ferguson. Prior distributions on spaces of probability measures. *Annals of Statistics*, 2: 615–29, 1974.

CE Frangakis and DB Rubin. Principal stratification in causal inference. *Biometrics*, 58(1): 21–29, 2002.

AE Gelfand and AFM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

GW Imbens and DB Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305–327, 1997.

H Ishwaran and LF James. Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

H Ishwaran and M Zarepour. Markov chain Monte Carlo in approximate Dirichlet and Beta two parameter process hierarchical models. *Biometrika*, 87:371–290, 2000.

H Jin and DB Rubin. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103:101–111, 2008.

H Jin and DB Rubin. Public schools versus private schools: Causal inference with partial compliance. *Journal of Educational and Behavioral Statistics*, 34(1):24–45, 2009.

YT Lagerros, R Bellocco, H-O Adami, and O Nyren. Measures of physical activity and their correlates: The swedish national march cohort. *European Journal of Epidemiology*, 24: 161–169, 2009.

SN MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 23:727–741, 1994.

A Mattei and F Mealli. Application of the prinicipal stratification approach to the faenza randomized experiment on breast self-examination. *Biometrics*, 63(2):437–46, 2007.

O Papaspiliopoulos and GO Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95:169–186, 2008.

J Pearl. Direct and indrect effects. In *Proceedings of the 17th Conference on Uncertainy in artificial intelligence*, pages 411–420, San Francisco, CA, 2001. Morgan Kaufmann.

JM Robins and S Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

PR Rosenbaum. The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series B*, 147(5):656–666, 1984.

PR Rosenbaum and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

DB Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(1):688–701, 1974.

DB Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

DB Rubin. Comment on 'Randomization analysis of experimental data: The fisher randomization test' by D. Basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

DB Rubin. Comment on 'Neyman (1923) and causal inference in experiments and observational studies'. *Statistical Science*, 5:472–480, 1990.

DB Rubin. Causal inference through potential outcomes and principal stratification: application to studies with censoring due to death. *Statistical Science*, 91:299–321, 2006.

SL Schwartz, F Li, and JP Reiter. Sensitivity analysis for unmeasured confounding in principal stratification. Technical report, Department of Statistical Science, Duke University, 2011.

J Sethurman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

A Sjölander, K Humphreys, S Vansteelandt, R Bellocco, and J Palmgren. Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics*, 65(2):514–520, 2009.

MA Tanner and WH Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

SG Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics*, 36:45–54, 2007.