This paper is the post-print version of the following paper:
Bartolucci F. & Grilli L. (2011) Modelling partial compliance through copulas in a principal stratification framework. Journal of the American Statistical Association, 106, pp. 469-479. DOI: 10.1198/jasa.2011.ap09094. http://amstat.tandfonline.com/doi/abs/10.1198/jasa.2011.ap09094

Modeling partial compliance through copulas in a principal stratification framework

F. Bartolucci and L. Grilli

March 9, 2011

Abstract

Within the principal stratification framework for causal inference, modeling partial compliance is challenging because the continuous nature of the principal strata raises subtle specification issues. In this context, we propose an approach based on the assumption that the joint distribution of the degree of compliance to the treatment and the degree of compliance to the control follows a Plackett copula, so that their association is modeled in a flexible way through a single parameter. Moreover, given the two compliances, the distribution of the outcomes is parameterized in a flexible way through a regression model which may include interaction and quadratic terms and may also be heteroscedastic. In order to estimate the parameters of the resulting model, and then the causal effect of the treatment, we adopt a maximum likelihood approach via the EM algorithm. In applying this approach, the marginal distributions of the two compliances are estimated by their empirical distribution functions, so that no constraints are posed on these distributions. Since the two compliances cannot be jointly observed, there is not direct empirical support for the association parameter. We describe a strategy for studying this parameter by a profile likelihood method and discuss an analysis of the sensitivity of the causal inference results to its value. We apply the proposed approach to data previously analyzed by Efron and Feldman (1991) and Jin and Rubin (2008). Estimated causal effects are in line with those of previous analyses, but the pattern of association between the compliances is qualitatively different, apparently due to the flexibility of the copula and to allowing regression equations in the proposed method to include interactions and heteroscedasticity.

Keywords: causal inference, compliance, EM algorithm, profile likelihood, sensitivity analysis.

1 Introduction

In clinical trials, non-compliance to the assigned treatment is a frequent and tricky problem. In the case of all-or-none compliance, the Instrumental Variable (IV) estimator represents a standard solution to estimate the causal effect of receiving the treatment. Angrist *et al.* (1996) showed that, under certain assumptions, the IV estimate can be interpreted as the complier average causal effect (CACE), a fundamental quantity within the potential-outcome approach to causal inference (Rubin, 1978). Subsequently, Frangakis and Rubin (2002) outlined a general framework, known as principal stratification, to address causal inference in the presence of an intermediate variable (i.e., a variable that is measured after treatment has been assigned). Compliance status is an example of an intermediate variable, and the CACE is a particular principal causal effect, corresponding to the average effect of treatment in the principal stratum of compliers.

In the potential-outcome approach to causal inference, any post-treatment variable has one potential version for each level of the treatment. In the case of binary treatment, such as drug versus placebo, and a binary intermediate variable, such as all-or-none compliance, there are four possible configurations: "always-takers" (who take the whole dose under both drug and placebo assignment), "compliers" (who receive the whole dose when assigned to drug treatment but no dose when assigned to placebo treatment), "never-takers" (who receive no drug dose under either treatment assignment), and "defiers" (who receive the opposite treatment of the treatment they were assigned). The potential-outcome framework makes it clear that it is generally a mistake to condition on a post-treatment concomitant variable (Cochran, 1957; Rubin, 2005); the principal stratification approach works by viewing an individual's vector of potential outcomes as an immutable characteristic of the individual that is unaffected by treatment. In a randomized study, principal strata can be expected to be balanced across treatment arms.

Since only one potential outcome is observed on an individual, stratum membership is not directly observable. However, for any individual the admissible strata are only a subset of all strata: for example, an individual assigned to drug treatment who actually takes the drug can be either a "complier" or an "always-taker"; accordingly, the implied statistical model may be seen as a constrained latent class model (Grilli, 2011).

Sometimes, an individual may take a portion of the drug or placebo, and then we say that the clinical trial is affected by partial compliance. In this case, the intermediate variable for

compliance status can be thought of as continuous. The principal stratification framework is still valid, but with a continuum of compliance behaviors, a latent-class approach with discrete classes cannot be applied directly. A solution proposed by Jin and Rubin (2008), hereafter JR, is to use a more structured model by specifying functional relationships between potential outcomes and potential values of the intermediate variables. The method was illustrated through re-analysis of the data investigated by Efron and Feldman (1991), hereafter EF, involving partial compliance with a medication aiming to control cholesterol levels.

The analysis carried out by EF relies on an assumption that drug and placebo compliances are linked by a deterministic function that allows one, in effect, to impute missing compliance status. Specifically, EF used an equipercentile equating function implying that if the proportion of drug doses taken by an individual assigned to active treatment is at a given percentile of the distribution of observed compliance behavior, then it is assumed that the individual's placebo compliance behavior would be at the same percentile of observed placebo compliances. The same rule can be used to determine the corresponding drug compliance status for an individual assigned to the placebo arm. JR (Sections 1 and 2.3) argued that "the EF assumption of a deterministic relation between drug and placebo compliances is overly restrictive", in particular because it "denies the possibility that two patients who take the same amount of placebo under control may take different amounts of drug under treatment, possibly because of different tolerances to the drug's side effects". Instead, JR cast the problem in a principal stratification framework and formulated the weaker assumption of *negative side effect monotonicity*, according to which the drug compliance is no larger than the placebo compliance. Such an assumption seems plausible in placebo-controlled experiments where the drug has some negative side effects. However, this kind of monotonicity is violated if, for instance, there exists a subset of patients having positive side effects from reported cholesterol reductions after periodic blood tests, a possibility noted by JR (Section 2.4).

Relaxing the EF assumption that drug and placebo compliances are linked by a deterministic function requires specifying a model for the compliances. To this end, JR specified a parametric model for compliance behavior based on beta distributions, while for the potential outcomes they utilized a regression parameterization consistent with that adopted by EF. The Bayesian analysis performed by JR suggested that a strict equipercentile equating assumption was not essential and was arguably too restrictive.

We argue that the specification of a parametric model for the drug and placebo compliances is a critical point. In fact, the two compliances cannot be jointly observed and we have empirical

evidence on their relationship only indirectly through the model on the potential outcomes. The scarcity of the empirical support on this relationship, however, may be masked in a fully parametric Bayesian analysis, with results sensitive to the specification of the model and to the choice of the prior distributions. Moreover, the assumption of negative side effect monotonicity might be unduly restrictive.

To investigate the above issues, we propose an approach in which the marginal distributions of the two compliances are left unspecified and their joint distribution is modeled through the Plackett copula (Nelsen, 2006), avoiding in this way any monotonicity assumption. The association between the two compliances is thus summarized by a single parameter which can be studied via profile likelihood. Specifically, we first estimate the marginal distribution of each type of compliance by the empirical distribution function and then, for a given value of the Plackett association parameter, we obtain the maximum likelihood (ML) estimate of the model parameters through the EM algorithm (Dempster et al., 1977). The computational task is performed by a series of MATLAB functions which are available from the authors upon request. In the application, the method we propose turns out to be flexible and straightforward to implement, yielding an alternative way of modeling and interpreting such data.

Similarly to JR, we aim at drawing causal inference in a principal stratification framework and the use of different modes of inference (Bayesian in JR versus ML here) is more a matter of convenience than a substantial issue. Our analysis differs from the JR's essentially because we allow more flexibility in the potential outcome regressions (interaction terms and heteroscedasticity) and in the joint distribution of the compliances.

The reminder of this paper is organized as follows. Section 2 describes the EF data and shows the results of a preliminary model for these data. Section 3 outlines the proposed model and Section 4 explains the estimation strategy. The next two sections illustrate the results of fitting the proposed model to data from a randomized study of a cholesterol medication; in particular, Section 5 summarizes the model selection procedure, whereas Section 6 illustrates the main inferential results. Section 7 reports a final discussion. Details on the maximization of the likelihood via the EM algorithm are given in the Appendix.

2 Motivating application and preliminary analysis

In this paper, we analyze the data set previously analyzed by EF and JR, which is a subset of data from a placebo-controlled double-blinded randomized clinical trial designed to study

the effectiveness of a certain drug, cholestyramine, for lowering cholesterol levels. The data set includes three variables: the binary indicator for the treatment assignment, the proportion of compliance rounded to the second decimal, and one continuous variable measuring the average decrease in the cholesterol level during the study (visits at two-month intervals with an average length of 7.3 years). Participants were assigned packets of drug or placebo and they were asked to return unused packets at each visit. Thus, compliance is computed as the proportion of assigned packets not returned, averaged over all visits. Although it seems conceivable to investigate the dependence of compliance behavior at a given time on previous observations, the data set available to us does not contain any information on missing visits or temporal trends among the variables, so we focus our attention on analyzing average compliance over time. Following ER and JR, we remove two outliers and then analyze data on 335 men: 164 assigned to active drug and 171 assigned to placebo.

Figure 1 helps to understand the role of compliance in the cholestyramine study. The left panel refers to the treatment group, and the right panel refers to the control group. In both panels, the y-axis reports the observed outcome of cholesterol reduction and the x-axis reports the proportion of packets not returned as a reflection of medication compliance. The rising trend in the left panel reveals that a higher compliance to drug is associated with a larger reduction in the cholesterol level; therefore, the drug seems to be effective. However, the right panel also shows an increasing, though less pronounced, trend. Since the placebo is not thought to have a chemical impact on cholesterol levels, such a trend would seem to be due to compliance. The effect of the observed compliance to drug is likely the combination of a "genuine" effect, due to the chemical action of the drug, and a "collateral" effect, due to the correlation between the degree of compliance and some unobserved characteristics of the patients which affect the cholesterol level, such as the propensity to eat healthy food or to exercise.

The analysis would be straightforward if one could assume that compliance to placebo is identical, unit by unit, to compliance to drug (what EF called the *perfect blind assumption*). However, the quantile-quantile plot (see JR's Figure 3) shows that the observed compliance to placebo is considerably larger than the observed compliance to drug, presumably due to some adverse side effects of the drug. The crucial methodological issue is then to properly model the relationship between the two compliances in order to impute the missing placebo compliance to patients in the treatment arm, and viceversa.

We now introduce some basic notation. For any individual i, with $i = 1, \ldots, n$, we define Z_i to be a binary variable equal to 1 if the subject is assigned to the treatment arm (drug)

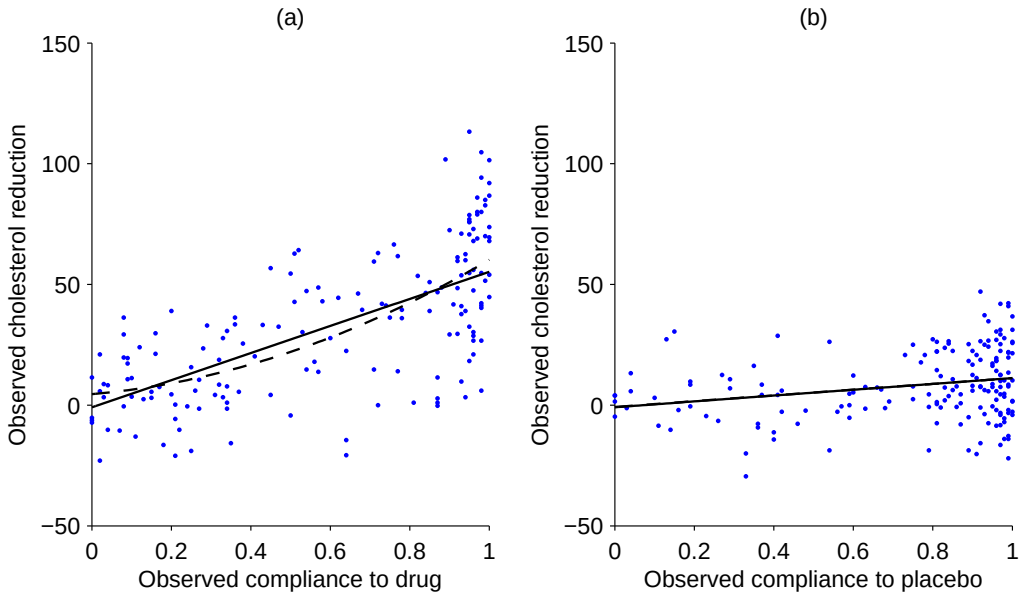

Figure 1: *Cholesterol reduction against degree of compliance observed in the treatment arm (a) and control arm (b) for the cholestyramine data. The straight lines correspond to linear regressions and the dashed curve corresponds to a quadratic regression.*

and equal to 0 if he is assigned to the control arm (placebo). The potential compliances are represented by the pair (d_i, D_i), where d_i denotes the placebo compliance if subject i is assigned to the control arm and D_i denotes the drug compliance if he is assigned to the treatment arm. Both d_i and D_i are proportions and thus lie on the unit interval. The outcome variable has two potential versions, denoted by $Y_i^{(0)}$ (outcome under placebo) and $Y_i^{(1)}$ (outcome under drug), reflecting cholesterol reduction, with higher values implying more favorable outcomes.

Before developing a full model for the cholestyramine data, we fit regression models separately for the patients assigned to drug ($Z_i = 1$) and patients assigned to placebo ($Z_i = 0$). The results of this preliminary fitting are reported in Table 1, which reports for each model the maximum log-likelihood, the number of parameters, and the value of the likelihood ratio (LR) test statistic with respect to the previous and the initial model, together with the corresponding p-value.

Following EF, we consider the regression model in which $Y_i^{(0)}$ depends on d_i whereas $Y_i^{(1)}$ depends on D_i and D_i^2. Since the plots show patterns of increasing variance, we allow for heteroscedasticity with $Var(Y_i^{(0)}|d_i, Z_i = 0) = \exp(\delta_{00} + \delta_{01} d_i)$ and $Var(Y_i^{(1)}|D_i, Z_i = 1) = \exp(\delta_{10} + \delta_{11} D_i)$. This model, denoted as *Initial* in Table 1, gives rise to a log-likelihood of -1426.33 with 9 parameters. Building on this model, we follow a procedure of model selection

Table 1: *Preliminary model selection: separate regressions for the two treatment arms.*

			Against previous		Against initial	
Model	log-lik.	#par.	LR stat.	p-value	LR stat.	p-value
Initial*	-1426.33	9	-	-	-	-
Previous + ($\delta_{00} = \delta_{10}$, $\delta_{01} = 0$)	-1427.78	7	2.90	0.234	2.90	0.234
Previous + ($\beta_{00} = \beta_{10}$)	-1428.18	6	0.82	0.366	3.72	0.294
Previous + ($\beta_{12} = 0$)	-1429.42	5	2.47	0.116	6.19	0.186

*Initial model: $Y_i^{(0)} \mid d_i, Z_i = 0 \sim N[\beta_{00} + \beta_{01} d_i,\, \exp(\delta_{00} + \delta_{01} d_i)]$, $Y_i^{(1)} \mid D_i, Z_i = 1 \sim N[\beta_{10} + \beta_{11} D_i + \beta_{12} D_i^2,\, \exp(\delta_{10} + \delta_{11} D_i)]$

in which we first try to simplify the variance structure. A reasonable restriction is $\delta_{00} = \delta_{10}$ and $\delta_{01} = 0$, so that heteroscedasticity is only in the second equation and $Y_i^{(1)}$ has the same variance as $Y_i^{(0)}$ when $D_i = 0$. The model can be further simplified by assuming that the two equations have the same intercept and dropping the quadratic term from the equation for $Y_i^{(1)}$. The equality of the intercepts implies that patients with no compliance at all have the same outcome regardless of the treatment arm; EF found further empirical evidence in favor of such a restriction.

Compared to the homoscedastic model with quadratic term for $Y_i^{(1)}$, adopted by JR, the final model of Table 1 has two fewer parameters and a higher log-likelihood. The estimated conditional distributions are

$$Y_i^{(0)} \mid d_i, Z_i = 0 \;\sim\; N\left[-0.869 + 12.081 d_i,\, \exp(5.294)\right], \tag{1}$$

$$Y_i^{(1)} \mid D_i, Z_i = 1 \;\sim\; N\left[-0.869 + 56.106 D_i,\, \exp(5.294 + 1.366 D_i)\right]. \tag{2}$$

The corresponding regression lines are drawn in Figure 1, where the left panel also reports a dashed curve for the quadratic regression.

3 Proposed copula model

We now describe a proposed copula model for the random variables defined in the previous section, namely the treatment indicator Z_i, the potential placebo compliance d_i, the potential drug compliance D_i, the potential outcome under placebo $Y_i^{(0)}$, and the potential outcome under drug $Y_i^{(1)}$.

In line with JR (Section 2.2), we rely on two standard assumptions, namely the *Stable Unit Treatment Value Assumption* (SUTVA) and *ignorable treatment assignment*. These assumptions simplify the model formulation by avoiding the need to worry about interference between units or to model the assignment mechanism. Even if these assumptions are untestable, they are

reasonable in carefully designed randomized trials such as the cholestyramine study discussed in Section 2. Also in line with JR (Section 2.3), we assume *strong access monotonicity*, namely that the drug compliance is null for patients assigned to placebo and placebo compliance is null for patients assigned to drug, which we regard as plausible in the present context since the trial prevented patients assigned to drug from taking placebo, and vice versa.

Under strong access monotonicity, the compliance behavior can be represented by the pair (d_i, D_i) instead of the four variables that would be needed in an "extended partial compliance" framework (JR, Section 2.1).

EF and JR made further assumptions about the relationship between drug and placebo compliances: EF assumed equipercentile equating of the compliances (a deterministic relationship), whereas JR assumed negative side effect monotonicity (a stochastic relationship with $D_i \leq d_i$); the approach here does not invoke either of these assumptions.

As in JR, we adopt a principal stratification approach where the principal strata are defined by the values of the pair (d_i, D_i). Therefore, the principal causal effect (PCE) is defined as the average difference $Y_i^{(1)} - Y_i^{(0)}$ for individuals belonging to the same principal stratum:

$$PCE\,(d_i, D_i) = E\left(Y_i^{(1)} - Y_i^{(0)} \mid d_i, D_i\right). \tag{3}$$

By representing the PCE as a function of d_i and D_i, we can think of principal causal effects on what Gilbert and Hudgens (2008) refer to as the causal effect predictiveness (CEP) surface.

Within a principal stratum, we adopt the following specification for the conditional distribution of each potential outcome given the pair of potential compliance values:

$$Y_i^{(z)} \mid d_i, D_i \sim N\left[\mu_z(d_i, D_i), \sigma_z^2(d_i, D_i)\right], \quad z = 0, 1, \tag{4}$$

with $\mu_z(d_i, D_i) = \boldsymbol{b}_z(d_i, D_i)'\boldsymbol{\beta}_z$ and $\sigma_z^2(d_i, D_i) = \exp[\boldsymbol{c}_z(d_i, D_i)'\boldsymbol{\gamma}_z]$, where $\boldsymbol{b}_z$ and $\boldsymbol{c}_z$ are functions to be appropriately chosen. The implied principal causal effect is $PCE\,(d_i, D_i) = \boldsymbol{b}_1(d_i, D_i)'\boldsymbol{\beta}_1 - \boldsymbol{b}_0(d_i, D_i)'\boldsymbol{\beta}_0$. In JR (Section 3.2) the mean under control is linear in d_i and D_i, the mean under treatment is linear in d_i and quadratic in D_i, whereas the variances are both constant; the model here is a generalization allowing comparison of different specifications.

It is important to note that the identification of the PCE relies on the modeling assumptions specified by (4), especially those concerning the conditional means $\mu_0(d_i, D_i)$ and $\mu_1(d_i, D_i)$. Unfortunately, the functional forms of the conditional means cannot be tested separately since d_i and D_i are never jointly observed. The same issue concerns the variance functions. The normality assumption, while also relevant, is expected to play a minor role and can be checked by studying alternative Box-Cox transformations (see Section 6).

A further issue concerns the correlation between $Y_i^{(0)}$ and $Y_i^{(1)}$ conditional on the stratum. This correlation is not estimable and the standard solution is setting it to zero. In their approach, JR performed a sensitivity analysis by setting the correlation to arbitrary values and found no important changes in the inferential conclusions when it varies; therefore, we do not pursue this matter anymore.

For the joint distribution of the compliances, JR (Section 3.2) assumed that d_i has distribution $Beta(\alpha_1, \alpha_2)$ and (conditional on d_i) the ratio D_i/d_i has distribution $Beta(\alpha_3, \alpha_4)$, which implies $D_i \leq d_i$ (*negative side effect monotonicity*). Even if the compliances d_i and D_i cannot be jointly observed, empirical evidence on their correlation is indirectly induced by the equations for the outcomes, see (4), where both compliances enter as regressors. Indeed, the specification of JR allows d_i and D_i to be correlated, although in a particular way that should not be viewed as the only possible way.

On this issue we consider the sensitivity of inference for the causal effect of treatment on cholesterol outcomes to the model for the compliances. We propose to use copulas, which are families of functions that can be applied to the marginal distributions of two random variables to obtain a corresponding joint distribution (Nelsen, 2006), focusing on copulas that can be characterized by a single parameter corresponding to a measure of association between the two random variables. Here we choose to model the joint distribution of d_i and D_i by the Plackett copula (Plackett, 1965), which is briefly illustrated in the following.

Let X and Y be two random variables with distribution functions F_X and F_Y, respectively, and let $u = F_X(x)$ and $v = F_Y(y)$. Then the Plackett copula of X and Y is the function

$$C_\psi(u,v) = \begin{cases} uv & \text{if } \psi = 1, \\ \frac{[1+(\psi-1)(u+v)] - \left\{[1+(\psi-1)(u+v)]^2 - 4\psi(\psi-1)uv\right\}^{\frac{1}{2}}}{2(\psi-1)} & \text{if } \psi > 0,\ \psi \neq 1. \end{cases} \tag{5}$$

By Sklar's theorem (Nelsen, 2006), $C_\psi(F_X(x), F_Y(y))$ is a joint distribution function with marginal distributions F_X and F_Y. The parameter ψ is a measure of association between the two random variables: $\psi = 1$ corresponds to independence, $\psi < 1$ to negative association, and $\psi > 1$ to positive association. The parameter ψ is related to the Spearman correlation coefficient ρ by the function $\rho = (\psi+1)/(\psi-1) - 2\psi \log \psi/(\psi-1)^2$.

The copula has the merit of allowing us to study the association between the compliances d_i and D_i without specifying a model for their marginal distributions, which are estimated by their empirical distribution functions.

The choice of the type of copula is expected to have minor consequences on the results. Indeed, we also fitted the model using the Gaussian copula, noting that it is more computationally

intensive but gives similar results (see Section 6.2).

4 Estimation

The model outlined in the previous section is based on three sets of parameters: (i) the parameters for the outcome under placebo $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\gamma}_0')'$; (ii) the parameters for the outcome under treatment $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1', \boldsymbol{\gamma}_1')'$; (iii) the Plackett association parameter ψ. For fixed ψ, we first estimate the joint distribution of d_i and D_i by applying the copula to the empirical distribution functions and then we estimate $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ by maximizing the likelihood of the observed data. Inference for ψ is based on the profile likelihood for this parameter.

First of all, we outline how the joint distribution of (d_i, D_i) is estimated from the observed marginal distributions of the two compliances, given the value of the association parameter ψ. Let $\mathcal{D}_0 = \{d_1^*, \ldots, d_{k_0}^*\}$ be the ordered set of distinct values of d_i in the data; similarly, let $\mathcal{D}_1 = \{D_1^*, \ldots, D_{k_1}^*\}$ be the ordered set of distinct values of D_i in the data, where k_0 is the number of distinct compliance levels observed under control and k_1 is the number of distinct compliance levels observed under treatment. At each discontinuity point, the empirical distribution functions of d_i and D_i are, respectively,

$$\hat{F}_{0h_0} = \frac{\sum_{i:Z_i=0} 1\{d_i \leq d_{h_0}^*\}}{\sum_{i=1}^n (1 - Z_i)}, \quad h_0 = 1, \ldots, k_0,$$
$$\hat{F}_{1h_1} = \frac{\sum_{i:Z_i=1} 1\{D_i \leq D_{h_1}^*\}}{\sum_{i=1}^n Z_i}, \quad h_1 = 1, \ldots, k_1,$$

where $1\{\cdot\}$ denotes the indicator function. Then, the joint distribution of (d_i, D_i) is estimated by a discrete distribution having support set $\mathcal{D}_0 \times \mathcal{D}_1$ and probability masses defined by linking the empirical distribution functions via the Plackett copula; see equation (5). Specifically, each of the $k_0 k_1$ support points $(d_{h_0}^*, D_{h_1}^*)$ has probability

$$\hat{p}_{h_0 h_1}(\psi) = C(\hat{F}_{0h_0}, \hat{F}_{1h_1}; \psi) - C(\hat{F}_{0,h_0-1}, \hat{F}_{1h_1}; \psi) - C(\hat{F}_{0h_0}, \hat{F}_{1,h_1-1}; \psi) + C(\hat{F}_{0,h_0-1}, \hat{F}_{1,h_1-1}; \psi),$$

where $F_{00} \equiv 0$ and $F_{10} \equiv 0$. Consequently, the conditional distribution of d_i given $D_i = D_{h_1}^*$ is estimated by a discrete distribution with support set $\mathcal{D}_0$ and probabilities

$$\hat{p}_{0h_0|h_1}(\psi) = \frac{\hat{p}_{h_0 h_1}(\psi)}{\sum_{j=1}^{k_0} \hat{p}_{j h_1}(\psi)}, \quad h_0 = 1, \ldots, k_0. \tag{6}$$

Similarly, the estimated conditional distribution of D_i given $d_i = d_{h_0}^*$ has support set $\mathcal{D}_1$ and probabilities

$$\hat{p}_{1h_1|h_0}(\psi) = \frac{\hat{p}_{h_0 h_1}(\psi)}{\sum_{j=1}^{k_1} \hat{p}_{h_0 j}(\psi)}, \quad h_1 = 1, \ldots, k_1. \tag{7}$$

Since (d_i, D_i) are not jointly observed, the distributions of the potential outcomes in model (4) cannot be directly used to define the model likelihood. However, we can use the distributions (6) and (7) to integrate out the missing compliance. Therefore, the density of $Y_i^{(0)}$ given d_i is

$$\hat{f}(Y_i^{(0)}|d_i;\boldsymbol{\theta}_0,\psi) = \sum_{h_1=1}^{k_1} f(Y_i^{(0)}|d_i, D_{h_1}^*;\boldsymbol{\theta}_0,\psi)\hat{p}_{1h_1|h_{0i}}(\psi),$$

where h_{0i} is an index between 1 and k_0 such that $d_i = d_{h_{0i}}^*$ and the density of $Y_i^{(0)}$ given the two compliances is defined by model (4). The conditional density $\hat{f}(Y_i^{(0)}|d_i;\boldsymbol{\theta}_0,\psi)$ can be used to define the likelihood since $Y_i^{(0)}$ and d_i are jointly observed for subjects in the control arm. Similarly, for subjects in the treatment arm we use the density of $Y_i^{(1)}$ given D_i

$$\hat{f}(Y_i^{(1)}|D_i;\boldsymbol{\theta}_1,\psi) = \sum_{h_0=1}^{k_0} f(Y_i^{(1)}|d_{h_0}^*, D_i;\boldsymbol{\theta}_1,\psi)\hat{p}_{0h_0|h_{1i}}(\psi),$$

where h_{1i} is an index between 1 and k_1 such that $D_i = D_{h_{1i}}^*$.

For given ψ, the *likelihood* of the model is thus

$$L_\psi(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) \propto \prod_{i:Z_i=0} \hat{f}(Y_i^{(0)}|d_i;\boldsymbol{\theta}_0,\psi) \prod_{i:Z_i=1} \hat{f}(Y_i^{(1)}|D_i;\boldsymbol{\theta}_1,\psi), \tag{8}$$

where we have omitted the component concerning the marginal distribution of the two compliances $\prod_{i:Z_i=0}(F_{0h_{0i}} - F_{0,h_{0i}-1}) \prod_{i:Z_i=1}(F_{1h_{1i}} - F_{1,h_{1i}-1})$. In fact, such a component, which recalls the saturated likelihood used in the empirical likelihood approach (Owen, 2001), is not relevant for making inference on the parameters of interest.

Details on the maximization of the likelihood via the EM algorithm are given in the Appendix. Note that the model under consideration relies on regressors d_i and D_i which are never jointly observed: estimation through the EM algorithm is feasible because of the structure imposed by the copula on the pair (d_i, D_i). For any given value of the Plackett association parameter ψ, the EM algorithm yields estimates of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, which we denote by $\hat{\boldsymbol{\theta}}_0(\psi)$ and $\hat{\boldsymbol{\theta}}_1(\psi)$ respectively. The *profile log-likelihood* function for ψ is then

$$\ell(\psi) = \log L_\psi(\hat{\boldsymbol{\theta}}_0(\psi), \hat{\boldsymbol{\theta}}_1(\psi)).$$

This function can be numerically maximized with respect to ψ to get the ML estimate $\hat{\psi}$ of the association parameter and the corresponding ML estimate of the other parameters. A graphical representation of $\ell(\psi)$ helps to realize how the values of ψ are supported by the data and to check for the presence of local maxima.

5 Model selection

We choose among models defined by the distributions (4) for $Y_i^{(z)} \,|\, d_i, D_i$, $z = 0, 1$, and the copula (5) for (d_i, D_i). We begin the model selection procedure from model M_0, which specifies the functions $\boldsymbol{b}_z(d_i, D_i)$ and $\boldsymbol{c}_z(d_i, D_i)$ so that the distributions of the potential outcomes in (4) have means

$$\mu_0(d_i, D_i) = \beta_{00} + d_i\beta_{01} + D_i\beta_{02} + d_iD_i\beta_{03}, \tag{9}$$

$$\mu_1(d_i, D_i) = \beta_{10} + d_i\beta_{11} + D_i\beta_{12} + D_i^2\beta_{13} + d_iD_i\beta_{14}, \tag{10}$$

and variances

$$\sigma_z^2(d_i, D_i) = \exp(\gamma_{z0} + d_i\gamma_{z1} + D_i\gamma_{z2}), \quad z = 0, 1. \tag{11}$$

This initial model is more general than the one adopted by JR because we include the interaction between d_i and D_i in both regression equations and we allow for heteroscedasticity. After running the estimation algorithm, it turns out that this model has maximum log-likelihood equal to -1420.82 with 16 parameters (9 for the regression parameters and 6 for the variance parameters in addition to the Plackett association parameter).

We then consider models incorporating constraints on the parameters $\boldsymbol{\beta}_z$ and $\boldsymbol{\gamma}_z$. The results of the selection procedure are summarized in Table 2, with comparisons between nested models carried out using likelihood ratio (LR) testing at the 5% significance level. Note that the conclusions would be unchanged if we relied on the Bayesian Information Criterion (BIC) of Schwarz (1978).

Table 2: *Model selection (model M_k incorporates hypotheses H_1 to H_k).*

				Against M_{k-1}		Against M_0	
Model	log-lik.	par.	BIC	LR stat.	p-value	LR stat.	p-value
M_0: initial model *	-1420.82	16	2934.67	-	-	-	-
M_1: $M_0 + H_1(\beta_{00} = \beta_{10})$	-1420.85	15	2928.91	0.06	0.804	0.06	0.804
M_2: $M_1 + H_2(\gamma_{00} = \gamma_{10})$	-1422.24	14	2925.88	2.78	0.095	2.84	0.241
M_3: $M_2 + H_3(\beta_{01} = \beta_{11})$	-1423.24	13	2922.06	2.00	0.157	4.84	0.184
M_4: $M_3 + H_4(\gamma_{01} = \gamma_{02} = \gamma_{11} = 0)$	-1425.71	10	2909.56	4.94	0.176	9.78	0.134
M_5: $M_4 + H_5(\beta_{13} = 0)$	-1426.05	9	2904.42	0.68	0.410	10.46	0.164
M_6: $M_5 + H_6(\beta_{03} = 0)$	-1426.78	8	2900.07	1.46	0.227	11.92	0.155
M_7: $M_6 + H_7(\beta_{02} = 0)$	-1427.21	7	2895.12	0.86	0.354	12.78	0.173
M_8: $M_7 + H_8(\psi = 1)$	-1433.69	6	2902.27	12.97	0.000	25.75	0.004

* Initial model defined by equations (4) and (5), with means and variances specified in (9), (10), and (11).

The structure of the cholestyramine study suggests some possible simplifying hypotheses: $H_1 : \beta_{00} = \beta_{10}$, implying $\mu_0(0, 0) = \mu_1(0, 0)$, namely the expected outcome of a subject with

null placebo and drug compliances is not affected by the treatment assignment; $H_2 : \gamma_{00} = \gamma_{10}$, implying $\sigma_0^2(0,0) = \sigma_1^2(0,0)$, namely the variance of the outcome of a subject with null placebo and drug compliances is not affected by the treatment assignment; $H_3 : \beta_{01} = \beta_{11}$, implying (together with H_1) $\mu_0(d_i, 0) = \mu_1(d_i, 0)$, which means that, for subjects with null drug compliance, the dependence of the expected outcome on the placebo compliance is not affected by the treatment assignment. Hypotheses H_1 to H_3 are sequentially incorporated in model M_0, obtaining model M_3.

One could also try to simplify the variance equations. Under parameterization (11), the restrictions considered in Section 2 may be written as $H_4 : \gamma_{01} = \gamma_{02} = \gamma_{11} = 0$, implying that the conditional variance of $Y_i^{(0)}$ given (d_i, D_i) is constant, whereas that of $Y_i^{(1)}$ only depends on D_i. Including hypothesis H_4 into model M_3 yields model M_4.

Simplifying the regression equations by setting some of the parameters in $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ to zero while taking care to retain a hierarchical parameterization gives rise to $H_5 : \beta_{13} = 0$, $H_6 : \beta_{03} = 0$, and $H_7 : \beta_{02} = 0$, which are sequentially introduced in model M_4 to yield models M_5, M_6, and M_7.

Finally, in order to test the hypothesis of independence between d_i and D_i, we fit model M_8 which includes the restriction $H_8 : \psi = 1$. The LR test statistic comparing model M_8 with model M_7 is equal to 12.97 with a p-value smaller than 0.001; beyond this, we did not find evidence requiring additional flexibility in model parameters and take M_7 as the final model. This model has maximum log-likelihood equal to -1427.21 with 7 parameters. The estimates of the parameters of model M_7 are reported in Table 3, together with the intervals obtained through the non-parametric bootstrap (Efron and Tibshirani, 1994; Davison and Hinkley, 1997). In particular, we re-sampled the observed individuals with replacement 1,000 times and fitted the model for every sample obtained in this way; then, we computed 95% intervals using the 2.5% and 97.5% quantiles of the distribution of these estimates.

The use of bootstrap intervals is motivated by the skewness of the sampling distribution of some parameters, in particular the Plackett parameter and the interaction between the two compliances. We found bootstrap intervals to be in substantial agreement with profile likelihood intervals. While acknowledging that there have been debates about the interpretation of bootstrap intervals in complex problems (Schenker, 1985), we proceeded by interpreting bootstrap intervals as confidence intervals. The bootstrap method also can be applied to transformations of the parameters; in particular, we exploit it to produce interval estimates for principal causal effects (see Section 6.1).

Table 3: *Parameter estimates for the selected model M_7.*

Parameter	Estimate	Bootstrap interval
$\beta_{00} = \beta_{10}$: intercept for $E(Y_i^{(z)}\|d_i, D_i)$, $z = 0, 1$	-0.269	(-5.272, 4.887)
$\beta_{01} = \beta_{11}$: slope of d_i on $E(Y_i^{(z)}\|d_i, D_i)$, $z = 0, 1$	11.243	(4.704, 18.053)
β_{12}: slope of D_i on $E(Y_i^{(1)}\|d_i, D_i)$	-21.878	(-102.206, 9.407)
β_{14}: slope of $d_i D_i$ on $E(Y_i^{(1)}\|d_i, D_i)$	73.359	(38.331, 155.840)
$\gamma_{00} = \gamma_{10}$: intercept for $\log Var(Y_i^{(z)}\|d_i, D_i)$, $z = 0, 1$	5.260	(5.053, 5.429)
γ_{12}: slope of D_i on $\log Var(Y_i^{(1)}\|d_i, D_i)$	1.161	(0.194, 1.560)
$\log \psi$: log of the Plackett association parameter	2.875	(0.914, 16.744)

The assumptions of model M_7 imply that the conditional means of $Y_i^{(0)}$ and $Y_i^{(1)}$ are equal when $d_i = D_i = 0$ and the slope of d_i is the same in both equations when $D_i = 0$. The estimated distributions of the potential outcomes are

$$Y_i^{(0)} \,|\, d_i, D_i \;\sim\; N\left[-0.269 + 11.243 d_i,\, \exp(5.260)\right], \tag{12}$$

$$Y_i^{(1)} \,|\, d_i, D_i \;\sim\; N\left[-0.269 + 11.243 d_i - 21.878 D_i + 73.359 d_i D_i,\, \exp(5.260 + 1.161 D_i)\right]. \tag{13}$$

The finding that d_i has a positive and significant coefficient on the conditional mean of both potential outcomes seems to confirm that placebo compliance is a proxy for better cholesterol control. On the other hand, the negative estimate $\hat{\beta}_{12} = -21.878$ for the regression coefficient for D_i may be unexpected. However, this coefficient does not result to be significantly different from 0 and, more importantly, it cannot be interpreted separately from the estimate of the coefficient β_{14} for the interaction $d_i D_i$. Indeed, the estimated slope for D_i in $E(Y_i^{(1)}|d_i, D_i)$ is $-21.878 + 73.359 d_i$, which is negative only when $d_i < 0.298$; this happens for 12.3% of the subjects in the control arm. Also note that the positive association between d_i and D_i (due to $\widehat{\psi} > 1$) implies that the slope of D_i tends to increase with D_i itself. This pattern is compatible with the quadratic relationship between outcome and compliance in the treatment arm (left panel of Figure 1), which prompted both EF and JR to specify nonlinear functions. Thus, our linear model with an interaction between drug and placebo compliances provides an alternative explanation.

6 Inference

Inference under the selected model (model M_7 in Table 3) requires some care since the estimates of the regression coefficients in (12) and (13) are related to the estimate of the parameter ψ measuring the association between the drug compliance and the placebo compliance.

The uncertainty about ψ is evident in Figure 2, where the profile log-likelihood has values close to the maximum for a wide interval of values of $\log \psi$. In fact, the 95% profile confidence interval for $\log \psi$, which may be found as $\{\log \psi : \ell(\psi) > -1429.13\}$, is equal to $(1.017, \infty)$ corresponding to the interval $(2.765, \infty)$ for ψ. In general, it is not advisable to base the inference exclusively on the point estimate of ψ. Specifically, it is important to assess the sensitivity of the estimate of the PCE, as defined in (3), to the point estimate of ψ.

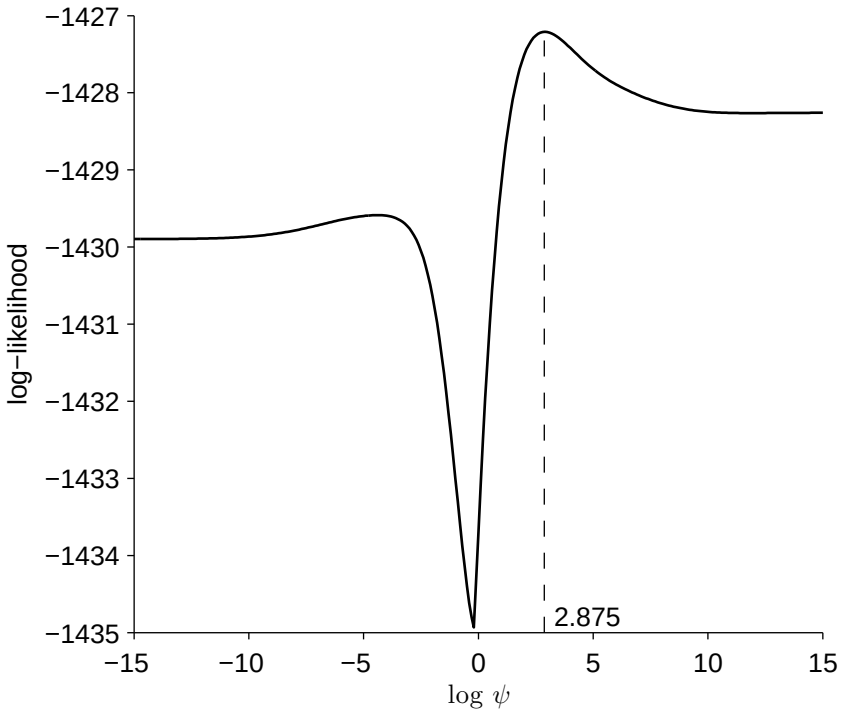

Figure 2: *Profile log-likelihood for* $\log \psi$ *under model* M_7.

For clarity of exposition and to help the comparison with the previous analyses of EF and JR, in the following we first illustrate the inference drawn under model M_7 with ψ at its ML estimate and then we perform a sensitivity analysis on the estimated PCE by letting ψ vary on a suitable interval and considering alternative specifications of the adopted model.

6.1 Inference based on the maximum likelihood estimate of ψ

First of all, it is interesting to consider the relationship between the placebo compliance d_i and the drug compliance D_i. Under model M_7, the estimate of $\log \psi$ is equal to 2.875, corresponding to $\widehat{\psi} = 17.727$, and the hypothesis of independence ($\psi = 1$) is rejected by the LR test; see Table 3. The resulting joint distribution of the two compliances d_i and D_i is represented in Figure 3

by a scatter plot of 1000 random draws, together with the curve of the estimated conditional mean of D_i given d_i.

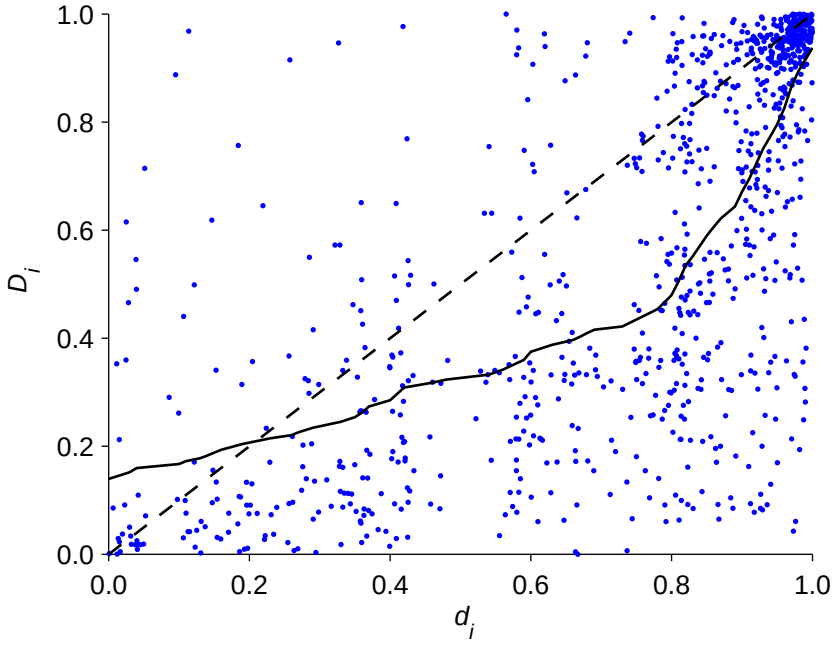

Figure 3: *Scatter plot of 1000 draws from the estimated joint distribution of d_i and D_i. The solid curve represents $E(D_i|d_i)$.*

Similarly to JR, we conclude that the equipercentile equating assumption of EF is not appropriate since the relationship between the two compliances is evidently not deterministic. Direct calculation based on the estimated bivariate distribution yields a Pearson linear correlation of 0.689 and a Spearman rank correlation of 0.726, which is high, but far from the perfect rank correlation implied by the equipercentile equating assumption.

The joint distribution of the drug and placebo compliances is also notably different from the JR's. In fact, the scatter plot in Figure 3 should be compared with one of the scatter plots of JR's Figure 5 reporting MCMC draws from (d_i, D_i). The plots of JR show an accumulation of points on the upper right angle (corresponding to high values of both d_i and D_i) and along the bisectrix (corresponding to $D_i = d_i$). Since the negative side effect monotonicity assumed by JR amounts to $D_i \leq d_i$, the bisectrix seems to act as a barrier and the accumulation of points along this line suggests that the monotonicity assumption may be too restrictive. Indeed, 21.6% of the points in our scatter plot goes beyond the bisectrix ($D_i > d_i$), even if in most cases D_i is only slightly larger than d_i; these points correspond to individuals with positive side effects. It may be that the population of patients is a mixture of three subpopulations of patients experiencing negative side effects ($D_i < d_i$), no side effects ($D_i = d_i$), and positive side effects ($D_i > d_i$).

This structure, which is scientifically meaningful, could be explicitly modeled via latent classes, but such an attempt would need a larger data set in order to provide reliable results.

We now consider the inference on the principal causal effects. Equations (12) and (13) with ψ at its ML estimate imply that

$$PCE(d_i, D_i) = (-21.878 + 73.359 d_i) D_i.$$

It is worth noting that, differently from JR, the dependence of the PCE on the dose of the taken treatment is stronger at higher levels of the placebo compliance due to the interaction term $d_i D_i$ entering the equation for $Y_i^{(1)}$. A consequence of the presence of this interaction term is that the slope for D_i on the PCE is negative for $d_i < 0.298$, but, as discussed above, this is not a major issue.

The results in terms of estimated PCE are summarized by the surface in Figure 4, which can be compared with the surface in the JR's Figure 4.

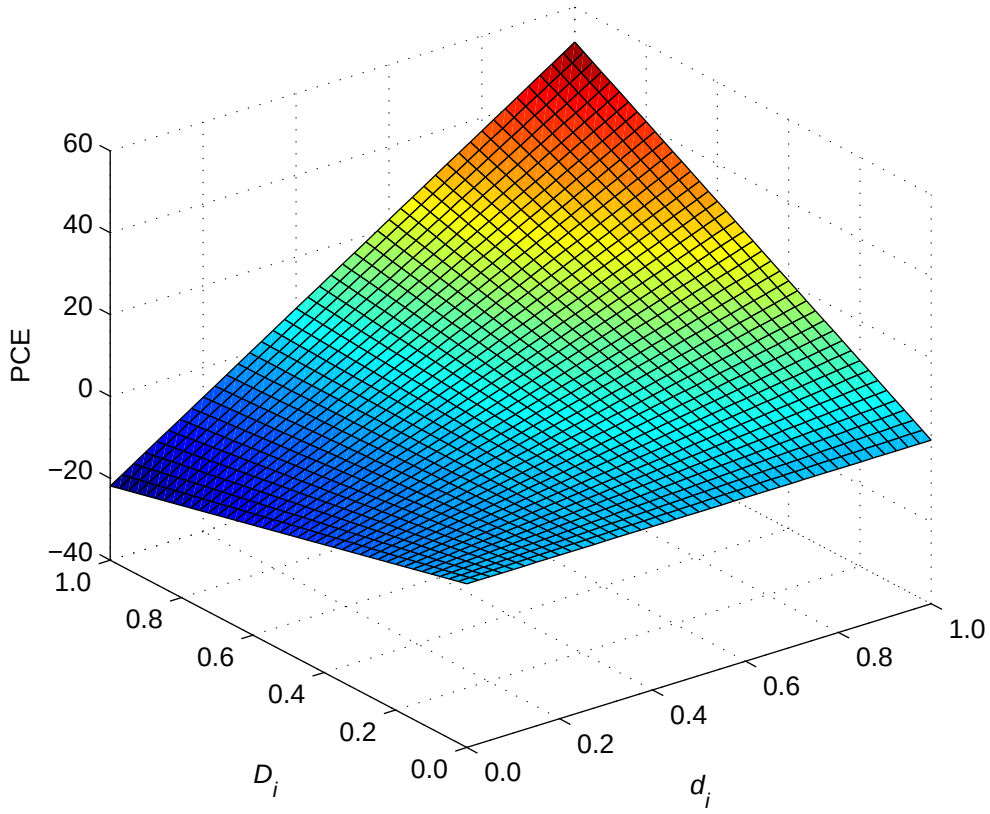

Figure 4: *Estimated PCE surface under model M_7.*

A synthetic comparison is obtained by contrasting the Bayesian posterior median of JR with our copula-based ML estimate at some interesting points (d_i, D_i). In particular, we have:

- 5 (JR) vs. 0 (copula-based ML) at $(0, 0)$, i.e. null compliance in both treatment arms;

- -13 (JR) vs. 0 (copula-based ML) at $(1, 0)$, i.e. full placebo compliance and null drug compliance;
- 24 (JR) vs 30 (copula-based ML) at $(0.89, 0.70)$, i.e. median placebo compliance and median drug compliance;
- 50 (JR) vs 51 (copula-based ML) at $(1, 1)$, i.e. full compliance in both treatment arms.

Our estimates of the PCE are thus in good agreement with those of JR.

Confidence intervals for the PCE are easily obtained via the non-parametric bootstrap, which was discussed in Section 5 in connection with the bootstrap intervals for the model parameters reported in Table 3. When the compliances are at their median values ($d_i = 0.89, D_i = 0.70$), the distribution of PCE across the 1,000 bootstrap samples has a median of 30.4 (identical to the first decimal to the ML point estimate) and the 95% bootstrap interval is (22.5, 39.2). As another example, when the compliances are at their first quartiles ($d_i = 0.59, D_i = 0.27$), the 95% bootstrap interval of the PCE is (-3.1, 9.9) and thus there is not convincing evidence of a positive effect.

Finally, we consider the problem of the prediction of the PCE as a function of the sole compliance under treatment D_i. This is of practical relevance since the placebo compliance d_i is unknown and it must be integrated out from the joint distribution of (d_i, D_i) in order to predict the PCE for any level of drug compliance. To this end, Figure 5 shows the predicted $Y_i^{(1)}$ and PCE given D_i only, together with the 95% bootstrap interval. These predictions are obtained by replacing d_i with its conditional mean given D_i, which is computed on the basis of the conditional distribution in expression (6) with ψ at the ML estimate. As one would expect, the predicted PCE is an increasing function of D_i. Also note that the expectation of $Y_i^{(1)}$ (dashed curve) has a shape similar to that of the quadratic regression curve we fitted for the patients in the treatment arm (left panel of Figure 1). This confirms the remark at the end of Section 5 about the non-linearity induced by the interaction between d_i and D_i.

6.2 Sensitivity analysis

The fragile identification of the association parameter ψ prompts us to perform a sensitivity analysis to assess how the estimated PCE depends on the value of this parameter. *Sensitivity bounds* for PCE should always accompany the point estimate and should be the only reported result when the profile likelihood is rather flat or multimodal. To compute these bounds we estimate the PCE for every value of ψ on a suitable grid and we take the lower bound as the

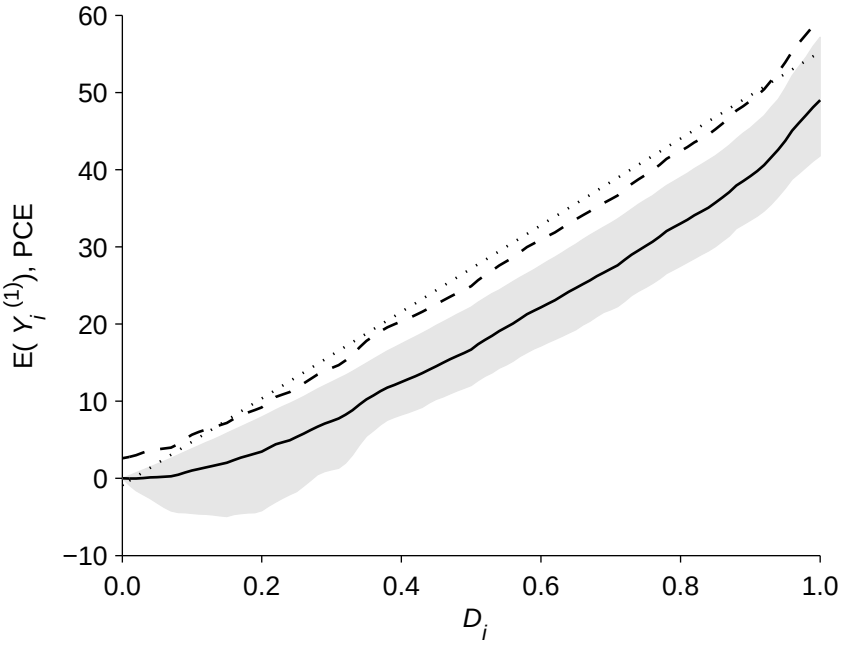

Figure 5: *Expectation of* $Y_i^{(1)}$ *and PCE with respect to* D_i*. Dotted line: expectation of* $Y_i^{(1)}$ *under the preliminary model of equations (1) and (2); dashed line: expectation of* $Y_i^{(1)}$ *under model* M_7 *of Section 5; the continuous line is the estimated PCE under the same model* M_7 *and the grey region represents the corresponding 95% pointwise bootstrap intervals.*

smallest estimate and the upper bound as the largest one. In particular, we consider a set of values of ψ corresponding to the 95% profile confidence interval found on the basis of the plot in Figure 2. The sensitivity bounds for the PCE are reported in Table 4 for d_i and D_i at their minimum, first quartile, median, third quartile, and maximum.

Table 4: *Sensitivity bounds for the PCE for* ψ *such that* $\ell(\psi) > -1429.13$ *for* d_i *and* D_i *at their minimum, first quartile, median, third quartile, and maximum.*

	$d_i = 0.00$	$d_i = 0.59$	$d_i = 0.89$	$d_i = 0.97$	$d_i = 1.00$
$D_i = 0.00$	$(0.0, 0.0)$	$(0.0, 0.0)$	$(0.0, 0.0)$	$(0.0, 0.0)$	$(0.0, 0.0)$
$D_i = 0.27$	$(-9.1, -1.6)$	$(4.0, 8.4)$	$(10.6, 13.4)$	$(12.3, 14.8)$	$(13.0, 15.3)$
$D_i = 0.70$	$(-23.5, -4.1)$	$(10.3, 21.7)$	$(27.4, 34.8)$	$(31.9, 38.3)$	$(33.6, 39.6)$
$D_i = 0.95$	$(-31.9, -5.6)$	$(14.0, 29.5)$	$(37.2, 47.3)$	$(43.3, 52.0)$	$(45.6, 53.8)$
$D_i = 1.00$	$(-33.6, -5.9)$	$(14.7, 31.0)$	$(39.1, 49.8)$	$(45.6, 54.8)$	$(48.0, 56.6)$

The PCE is stable at least for values (d_i, D_i) in the middle of the data, i.e. around the medians of the observed compliances, whereas wide intervals at unlikely compliance levels are not necessarily cause for concern. In particular, Table 4 shows that the PCE at the median point (0.89, 0.70) is quite stable, ranging from 27.4 to 34.8. At points where D_i is larger than d_i, such as $d_i = 0.59$ (the first quartile) and $D_i = 0.95$ (the third quartile), the PCE shows a

greater variation, ranging from 14.0 to 29.5. Note that to summarize the empirical evidence on the existence of a positive PCE one should also account for the sampling variance, which can be considerable at unlikely compliance levels: for example, the 95% bootstrap interval for the PCE at $d_i = 0.59$ and $D_i = 0.95$ is $(-10.8, 34.7)$ and then the causal effect is not significant even without considering its sensitivity to ψ.

The sensitivity bounds for the PCE in Table 4 are negative when $d_i = 0$ and $D_i > 0$. We interpret this pattern as a consequence of an extrapolation since in the fitted joint distribution of the compliances (Figure 3) such configuration has a small probability. In general, one should be very cautious when drawing inference on the PCE at an unlikely pair of compliances.

A sensitivity analysis can also be performed to assess how the estimate of the PCE depends on the assumption of normality for the conditional distributions of the potential outcomes in model (4). To this end, we exploit the Box and Cox (1964) transformation $y^* = [(y+30)^\lambda - 1]/\lambda$, where we added 30 to avoid negative values. Note that $\lambda = 1$ corresponds to the identity function (up to a translation), whereas for $\lambda = 0$ the transformation is defined as $\log(y+30)$. We compute the ML estimate of the parameters, including the association parameter ψ, for a grid of values of λ in the $[0, 2]$ interval. There is not evidence against the normality assumption since the LR test does not reject the hypothesis $\lambda = 1$ (p-value=0.123). The estimated PCE is fairly stable across the values of λ: for example, when the compliances are at their median values $(d_i = 0.89, D_i = 0.70)$, the estimate ranges from 29.4 to 32.9. As expected, the robustness is lower at points farther from the bulk of the data, such as points with D_i larger than d_i: for example, when $d_i = 0.59$ (the first quartile) and $D_i = 0.95$ (the third quartile), the estimate of the PCE ranges from 12.9 to 22.8.

Finally, to evaluate the dependence of the results on the type of copula, we replace the Plackett copula with a Gaussian copula (Nelsen, 2006). Due to the use of the bivariate normal distribution function, the maximization of the likelihood is much more computationally demanding, but the results are very close. In particular, the Pearson linear correlation between d_i and D_i turns out to be 0.680 (compared to 0.689), whereas the estimated PCE at $d_i = 0.89$ and $D_i = 0.70$ is 30.8 (compared to 30.4) and the estimated PCE at $d_i = 0.59$ and $D_i = 0.95$ is 18.2 (compared to 20.3).

7 Final remarks

Principal stratification is a general and effective framework for casting problems of causal inference. However, building statistical models within such a framework is difficult since the suitable models tend to be highly complex and their specification requires several assumptions for which there may be little empirical support. Indeed, the specification process is usually assisted by a set of assumptions with a clear subject-matter meaning to be justified by theoretical arguments or external information. Assumptions are specially important when the principal strata are continuous, such as in experimental studies with partial compliance. The approach illustrated in this paper aims at relaxing the assumptions needed to model continuous principal strata, exploiting the information in the data as much as possible.

The core of our approach is to link the observed marginal distributions of drug compliance and placebo compliance by a copula, so that their association is modeled in a flexible way through a single parameter. Since the marginal distributions are estimated by their empirical distribution functions, the copula does not impose any restriction on them. The point is that the two compliances are never jointly observed and the only empirical evidence about their relationship comes from the regression equations for the outcomes; therefore, it is crucial to represent their association with a single parameter to be estimated separately from the observed marginal distributions. This allows us to study the association parameter via profile likelihood so as to evaluate how inferential conclusions are supported by the data and to study the sensitivity of the resulting causal inference.

The approach proposed in the paper has been applied to the cholestyramine data with the main aim to make a comparison with the results of JR, who adopted the same principal stratification framework, but developed a different model. The estimated principal causal effects are in line with those estimated by JR, but the overall picture is somewhat different. Notably, our method yields an estimated joint distribution of the drug and placebo compliances that raises doubts on the appropriateness of JR's negative side effect monotonicity assumption. Moreover, the sensitivity analysis shows that the Principal Causal Effects are reliably estimated at drug and placebo compliance levels near the sample medians, whereas inference at unlikely compliance levels appears to be unduly affected by model assumptions.

The implementation of the EM algorithm for the proposed model turned out to be computationally simple and allowed us to easily compare, through likelihood ratio tests, a variety of specifications for the conditional mean and variance of the potential outcomes given the two

compliances. The selection process ended with a specification that differs from those of EF and JR especially in the conditional distribution of the potential outcome under treatment. In fact, the model for the mean includes an interaction term between the drug and placebo compliances, whereas the variance has a heteroscedastic form. Heteroscedasticity and, in particular, the interaction suggest alternative interesting ways of interpreting the cholestyramine data.

The analysis performed in this paper aims at estimating the causal effects within principal strata defined by both drug compliance and placebo compliance, as in JR's Section 3. Therefore, the results should not be directly used to draw a dose-response function, since the effect of the dose of the drug is mixed with the effect of the unobserved features associated with the degree of compliance. The estimation of a dose-response function, carried out in JR's Section 4, requires further problematic assumptions such as the latent ignorability of the dosage assignment mechanism at any level of placebo compliance. In a similar fashion, our model could be extended to estimate a dose-response function.

Acknowledgements

We thank Prof. Bradley Efron for giving us access to the Efron-Feldman data and Prof. Fabrizia Mealli for several insightful comments. We also thank the Associate Editor and two reviewers whose suggestions contributed to improve the paper. The research was partially supported by grants funded by the Italian Government (PRIN 2007 XECZ7L003 and PRIN 2008 WKHJPK003). Francesco Bartolucci also acknowledges the financial support from the Einaudi Institute for Economics and Finance (EIEF), Rome.

Appendix

In order to maximize the likelihood (8) with respect to $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ for given ψ, we exploit the EM algorithm of Dempster et al. (1977). This algorithm relies on the likelihood of the *complete data*, which correspond to the triple $(d_i, D_i, Y_i^{(0)})$ for every i such that $Z_i = 0$ and $(d_i, D_i, Y_i^{(1)})$ for every i such that $Z_i = 1$. Note that D_i is missing in the first case, whereas d_i is missing in the second case. For given ψ, the likelihood has expression

$$L_\psi^{(C)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \propto \prod_{i:Z_i=0} f(Y_i^{(0)}|d_i, D_i; \boldsymbol{\theta}_0)\hat{p}_{h_{0i}h_{1i}}(\psi) \prod_{i:Z_i=1} f(Y_i^{(1)}|d_i, D_i; \boldsymbol{\theta}_1)\hat{p}_{h_{0i}h_{1i}}(\psi).$$

However, $\hat{p}_{h_{0i}h_{1i}}(\psi)$ does not depend on the parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Therefore, for given ψ and up to a constant, the corresponding log-likelihood is

$$\begin{aligned}\ell_\psi^{(C)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= \sum_{i:Z_i=0} \sum_{h_1=1}^{k_1} W_{ih_1} \log f(Y_i^{(0)}|d_i, D_{h_1}^*; \boldsymbol{\theta}_0) \\ &+ \sum_{i:Z_i=1} \sum_{h_0=1}^{k_0} w_{ih_0} \log f(Y_i^{(1)}|d_{h_0}^*, D_i; \boldsymbol{\theta}_1),\end{aligned}$$

where $W_{ih_1} = 1\{D_i = D_{h_1}^*\}$ and $w_{ih_0} = 1\{d_i = d_{h_0}^*\}$.

The EM algorithm alternates two steps until convergence in $L_\psi(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$: at the E-step it computes the expected value of $\ell_\psi^{(C)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ given the observed data and the current value of the parameters; at the M-step it maximizes the above expected value with respect to $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$.

The E-step is performed by computing the *posterior probabilities*

$$\tilde{w}_{ih_0} = \frac{f(Y_i^{(1)}|d_{h_0}^*, D_i; \boldsymbol{\theta}_1)\hat{p}_{0h_0|h_{1i}}}{\hat{f}(Y_i^{(1)}|D_i; \boldsymbol{\theta}_1, \psi)}, \quad h_0 = 1, \ldots, k_0,$$

for every i such that $Z_i = 1$, and

$$\tilde{W}_{ih_1} = \frac{f(Y_i^{(0)}|d_i, D_{h_1}^*; \boldsymbol{\theta}_0)\hat{p}_{1h_1|h_{0i}}}{\hat{f}(Y_i^{(0)}|d_i; \boldsymbol{\theta}_0, \psi)}, \quad h_1 = 1, \ldots, k_1,$$

for every i such that $Z_i = 0$. The posterior probabilities are substituted to the binary variables w_{ih_0} and W_{ih_1} to obtain the expected value of the complete-data log-likelihood, denoted by $\tilde{\ell}_\psi^{(C)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$.

The M-step consists in updating the parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ by maximizing $\tilde{\ell}_\psi^{(C)}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. This is equivalent to fitting the regression models in (4), once a suitable weight has been attached to each observation. In practice, the M-step may be performed as follows:

- for given $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$, update the parameter vector $\boldsymbol{\beta}_z$ as

$$\boldsymbol{\beta}_z = [\sum_{i:Z_i=z} (\boldsymbol{X}_i^{(z)})' \mathrm{diag}(\boldsymbol{w}_i^{(z)}) \boldsymbol{X}_i^{(z)}]^{-1} \sum_{i:Z_i=z} (\boldsymbol{X}_i^{(z)})' \boldsymbol{w}_i^{(z)} Y_i^{(z)}, \quad z = 0, 1,$$

  where $\boldsymbol{X}_i^{(0)}$ is a design matrix with rows $\boldsymbol{b}_0(d_i, D_{h_1}^*)'$ and the column vector $\boldsymbol{w}_i^{(0)}$ has elements $\tilde{W}_{ih_1}/\sigma_0^2(d_i, D_{h_1}^*)$, whereas $\boldsymbol{X}_i^{(1)}$ is a design matrix with rows $\boldsymbol{b}_1(d_{h_0}^*, D_i)'$ and the column vector $\boldsymbol{w}_i^{(1)}$ has elements $\tilde{w}_{ih_0}/\sigma_1^2(d_{h_0}^*, D_i)$.

- for given $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, use a standard numerical optimizer to update $\boldsymbol{\gamma}_0$ by maximizing

$$-\sum_{i:Z_i=0} \sum_{h_1=1}^{k_1} \tilde{W}_{ih_1} \{\log \sigma_0^2(d_i, D_{h_1}^*) + [Y_i^{(0)} - \mu_0(d_i, D_{h_1}^*)]^2/\sigma_0^2(d_i, D_{h_1}^*)\}$$

and update $\boldsymbol{\gamma}_1$ by maximizing

$$-\sum_{i:Z_i=1}\sum_{h_0=1}^{k_0}\tilde{w}_{ih_0}\{\log\sigma_1^2(d^*_{h_0},D_i)+[Y_i^{(1)}-\mu_1(d^*_{h_0},D_i)]^2/\sigma_1^2(d^*_{h_0},D_i)\}.$$

The M-step can be modified in a suitable way in order to take into account constraints on the model parameters.

An important point concerns the choice of the starting values for the EM algorithm. We choose the starting values for $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ by regressing $Y_i^{(0)}$ on $\boldsymbol{b}_0(d_i,\tilde{D}_i(\psi))$ for subjects in the control arm, where $\tilde{D}_i(\psi)$ is the conditional expected value of D_i given d_i computed on the basis of the conditional probabilities in (7). Given model (4), this preliminary fitting is based on a heteroscedastic model with variance function $\sigma_0^2(d_i,\tilde{D}_i(\psi))$ depending on $\boldsymbol{\gamma}_0$. We choose the starting values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$ in a similar way. These starting values are randomly perturbed in order to try different initializations of the EM algorithm. This safeguards against the multimodality of the likelihood.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables (with discussion), *Journal of the American Statistical Association*, **91**, 444–472.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, **26**, 211–252.

Cochran, W.G. (1957) Analysis of covariance: Its nature and uses, *Biometrics*, **13**, 261–281.

Davison, A. C. and Hinkley D. V. (1997) *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, UK.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

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

Efron, B. and Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*, Chapman & Hall/CRC, London.

Frangakis, C. E. and Rubin, D. B. (2002) Principal stratification in causal inference, *Biometrics*, **58**, 21–29.

Gilbert, P. B. and Hudgens, G. (2008) Evaluating Candidate Principal Surrogate Endpoints, *Biometrics*, **64**, 1146-1154.

Grilli, L. (2011) Causal inference through principal stratification: a special type of latent class modelling. In: B. Fichet, D. Piccolo, R. Verde, M. Vichi (Eds) *Classification and Multivariate Analysis for Complex Data Structures*, Springer, 265–270.

Jin, H. and Rubin, D. B. (2008) Principal Stratification for Causal Inference With Extended Partial Compliance, *Journal of the American Statistical Association*, **103**, 101–111.

Nelsen, R.B. (2006) *An introduction to Copulas*, 2nd edition, Springer, New York.

Owen, A. (2001) *Empirical Likelihood*, Chapman & Hall/CRC, Boca Raton, Florida.

Plackett, R. L.(1965) A class of bivariate distributions, *Journal of the American Statistical Association*, **60**, 516–522.

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

Rubin, D.B. (2005) Causal inference using potential outcomes: Design, modeling, decisions, *Journal of the American Statistical Association*, **100**, 322–331.

Schenker, N. (1985) Qualms about bootstrap confidence intervals, *Journal of the American Statistical Association*, **80**, 360–361.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.