TREATMENT EFFECTS: A BAYESIAN PERSPECTIVE*

Econometric Reviews

Special Issue in Honor of Arnold Zellner

James J. Heckman[1], Hedibert F. Lopes[2], and Rémi Piatek[1]

[1]Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA

[2]Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave, Chicago, IL 60637, USA

This draft, February 22, 2012

Abstract

This paper contributes to the emerging Bayesian literature on treatment effects. It derives treatment parameters in the framework of a potential outcomes model with a treatment choice equation, where the correlation between the unobservable components of the model is driven by a low-dimensional vector of latent factors. The analyst is assumed to have access to a set of measurements generated by the latent factors. This approach has attractive features from both theoretical and practical points of view. Not only does it address the fundamental identification problem arising from the inability to observe the same person in both the treated and untreated states, but it also turns out to be straightforward to implement. Formulae are provided to compute mean treatment effects as well as their distributional versions. A Monte Carlo simulation study is carried out to illustrate how the methodology can easily be applied.

JEL classification: C11, C15, C31.

Keywords: Potential Outcomes, Treatment Effects, Bayesian, Counterfactual Distributions.

*This research was supported in part by the American Bar Foundation, the JB & MK Pritzker Family Foundation, Susan Thompson Buffett Foundation, NICHD R37HD065072, R01HD54702, we acknowledge the support of a European Research Council grant hosted by University College Dublin, DEVHEALTH 269874, a grant to the Becker Friedman Institute for Research and Economics from the Institute for New Economic Thinking (INET), and an anonymous funder. We thank the guest editor, Ehsan Soofi, and two anonymous referees for helpful comments. The views expressed in this paper are those of the authors and not necessarily those of the funders or commentators mentioned here.

Contents

1 Introduction

The estimation of treatment effect parameters has attracted a great deal of attention in the econometric literature on program evaluation due to their policy relevance. Frequentist approaches have been the focus of a large number of theoretical papers (Abbring and Heckman, 2007; Heckman and Vytlacil, 2007a,b; Imbens and Wooldridge, 2009), and have been fruitfully applied to address questions in labor, education and health economics.[1] This interest is, however, not reflected in the Bayesian literature, where only a few papers have been published compared to the huge body of research in the classical literature, and where empirical applications of the proposed methods are scarce. This paper discusses several explanations for the lack of interest in treatment effects on the Bayesian side. Technical details are then provided on how to derive and compute a large array of treatment effect parameters in the framework of a potential outcomes factor structure model that directly tackles problems with the existing Bayesian approaches, and is therefore likely to be useful in practice.

The fundamental problem in program evaluation is that it is impossible to directly observe outcome gains for the same person in both states. To deal with this issue, some classical approaches focus only on identifying mean treatment parameters and rely on conditional independence assumptions to solve the selection problem. Matching has become very popular in this field (Cochrane and Rubin, 1973; Rosenbaum and Rubin, 1983; Heckman et al., 1998). See the survey in Heckman and Vytlacil (2007b).

Bayesian approaches to date have been based on different strategies that, unfortunately, turn out to be problematic in practice. Some of these approaches only allow analysts to derive mean treatment effects and are thus of limited relevance for policy analyses, where joint distributions of potential outcomes often provide more insights into the effectiveness of programs. It is often more informative to learn what proportion of a population benefits from a program, or to identify how some target groups are affected by it, rather than just measuring mean effects that do not reveal the heterogeneity of the impact of the program (Heckman et al., 1997; Abbring and Heckman, 2007). Other approaches are based on special assumptions of limited generality in application.

This paper contributes to the current literature by showing how treatment parameters can be computed in a Bayesian fashion for a potential outcomes model and treatment choice equation with a factor structure, where not necessarily perfect measurements on the factors are available from an auxiliary data source. The

[1] See, e.g., Carneiro et al. (2003); Hansen et al. (2004); Heckman et al. (2006, 2010); and Conti et al. (2012).

combination of auxiliary data, the equation determining choice of treatment, and the factor structure model, facilitates identification by bringing extra information to the table. Factor models are widely used to proxy latent measures of ability. (See, e.g., Thurstone, 1934, and the large ensuing literature that built on his work.)

The approach is easy to implement. Latent factors are used to explain the unobserved correlation between the determinants of treatment status and the potential outcomes. As noted by Carneiro et al. (2003), this type of model generalizes the method of matching to account for mismeasured proxy variables. It can be used to derive mean treatment effects as well as distributional versions. We use a standard model to illustrate the main ideas.

The paper is organized as follows. Section 2 briefly introduces the potential outcomes model based on the Generalized Roy model, and discusses the different ways the fundamental identification problem has been addressed in the Bayesian literature. By weighing the advantages and disadvantages of the different approaches, it becomes clear why the potential outcomes factor structure model represents an attractive approach to this problem. Section 3 provides full technical details about the derivation of the treatment effect parameters, with both their mean and distributional versions. We focus our attention on three popular treatment parameters, namely the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT), and the Marginal Treatment Effect (MTE).[2] Section 4 explains how to apply this approach in practice to compute the treatment parameters. A simple Monte Carlo study illustrates and compares the performances of the different proposed estimators. Section 5 finally concludes with some remarks about useful directions for future research.

2 Potential Outcomes Model and Identification Issues

2.1 The baseline model and its fundamental identification problem

The textbook model considered in this paper is an extension of the original Roy model (Roy, 1951; Heckman and Honoré, 1990) and assumes a binary treatment decision D that involves two continuous potential

[2] Applying the proposed methodology to other types of treatment parameters would be straightforward to achieve. Heckman and Vytlacil (1999, 2000) and Carneiro et al. (2010, 2011) show how a variety of treatment parameters can be derived from the MTE.

outcomes Y_1 and Y_0 for the *treated* and *untreated* states, respectively:

$$\begin{aligned} D &= \mathbf{1}(D^\star > 0)\,, \\ D^\star &= Z'\gamma + U_D, \\ Y_1 &= X'\beta_1 + U_1, \\ Y_0 &= X'\beta_0 + U_0, \end{aligned} \tag{1}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function that is equal to 1 if the corresponding condition is fulfilled, and equal to 0 otherwise.[3] The relationship of the Roy model to other models of potential outcomes is discussed in Heckman (2008). The treatment decision is specified as a standard threshold-crossing model, where the utility associated with taking treatment is a linear function of some observed characteristics Z through the vector of slope parameters γ. The two potential outcomes are assumed to linearly depend on a set of covariates X through the vectors of slopes β_1 and β_0. Without any loss of generality, these covariates are assumed to be common across states. The presence of at least one exclusion restriction—or instrument—is assumed in the treatment equation, hence the use of the notation Z to include X and any additional covariates not in X in the outcome equations. Exclusion restrictions would be necessary to achieve nonparametric identification, without specifying any restrictions on functional forms.[4] Note, however, that the assumption of an exclusion restriction is not strictly required in the remainder of this paper, since we rely on a parametric specification of the model. Unobserved heterogeneity is captured by the error terms U_D, U_1 and U_0, which are assumed to have zero means and finite variances.

We adopt the linear-in-parameters specification for the sake of familiarity and computational tractability. As noted below, a much more general model can be identified. Fundamental to our approach in a linear or nonlinear model is the assumption that the dependence among the (U_D, U_1, U_0) are generated by a low-dimensional vector of latent variables. See the analysis in Abbring and Heckman (2007).

Since a person can only be in one specific treatment state, the actual outcome Y is observed according to the following "switching regression" model (see Quandt, 1958):

$$Y = D\,Y_1 + (1 - D)\,Y_0.$$

[3] See Heckman (1990) for one discussion of the more general Roy model and its identification. Heckman and Vytlacil (2007b) provide a general discussion. This model has been used in econometrics since Heckman (1974).

[4] Heckman (1990) and Heckman and Vytlacil (2007b) present conditions for nonparametric identification of the model.

In the program evaluation literature, the main parameter of interest is the *outcome gain* defined as $\Delta \equiv Y_1 - Y_0$. Consequently, for each individual present in state j ($j = 0, 1$), the challenge is to identify her potential outcome in the alternative state. The unobserved outcome is usually called a *counterfactual* outcome, and the estimation of its distribution has received a great deal of interest in econometrics (Heckman et al., 1997; Carneiro et al., 2003; Abbring and Heckman, 2007; Chernozhukov et al., 2012).

A naive approach to solving this missing data problem would be to simply compare outcomes in the two treatment groups. Such an approach could, however, ignore the underlying selection process, and thus yield biased treatment parameters.[5] The observed characteristics X and Z often explain a small part of the heterogeneity of the treatment decision and of the potential outcomes Y_0 and Y_1. Other personal characteristics, unobserved to the econometrician, are captured by the error terms U_D, U_1 and U_0. A prototypical example is personal abilities: individuals with higher abilities often perform better, whatever choice they make. If the treatment decision is achieving higher education, and the outcome of interest is a labor market outcome, high-skilled individuals will more likely achieve higher education than less-skilled individuals and earn higher outcomes, even if they actually fail to succeed in school. This implies a positive correlation between the error terms of the different equations.

To complete the specification of the potential outcomes model, and better understand the consequences of the fundamental identification problem, the covariance structure of the unobserved part of the model is expressed as:

$$\text{Cov}\begin{pmatrix} U_D \\ U_1 \\ U_0 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{D1}\sigma_1 & \rho_{D0}\sigma_0 \\ \rho_{D1}\sigma_1 & \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{D0}\sigma_0 & \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix}, \tag{2}$$

where $\rho_{rs} \equiv \text{Corr}(U_r, U_s)$. The variance of U_D is set to 1 for identification of the binary choice equation. Obviously, since Y_1 and Y_0 can never be observed simultaneously for a given person, the correlation ρ_{10} cannot be directly inferred from the data. Heckman (1992) and Vijverberg (1993) were the first to highlight this problem, which arises in all models with state-dependent outcomes (the Roy model and its extensions). It has been tackled (or ignored) in two different strands in the Bayesian literature.

On the one hand, learning about the non-identified parameter ρ_{10} can take place during the estimation

[5] This approach at best would also only be able to estimate marginal distributions of Y_0 and Y_1.

process and be exploited to derive distributional versions of the treatment effects. On the other hand, the problem can be avoided by estimating the model without the joint distribution of the potential outcomes. This solution bypasses the identification problem, but also makes it impossible to go beyond mean treatment parameters (see Section 2.2 below).

The approach adopted in this paper uses factor structure models, defined in Section 2.3, and enables analysts to estimate a wide array of mean and distributional treatment effects. Although this topic has been extensively treated in the classical literature (Abbring and Heckman, 2007), to the best of our knowledge no attempts have been made so far to propose a formal presentation of the treatment effects in the specific framework of a model with latent factors and from a Bayesian perspective.[6] In principle, it is possible to develop a nonparametric version of our analysis, but we do not do so in this paper.

2.2 Dealing with the non-identified correlation between the potential outcomes

Learning about ρ_{10}. Vijverberg (1993) points out that although the correlation parameter ρ_{10} is not identified, it cannot take any possible values in the interval $[-1, 1]$ since the covariance matrix in Equation (2) has to be positive semidefinite. This constraint is fulfilled if and only if the determinant of the covariance matrix is positive, which leads to the following bounds on the non-identified correlation:

$$\underline{\rho}_{10} \leq \rho_{10} \leq \overline{\rho}_{10}, \qquad \begin{aligned} \underline{\rho}_{10} &= \rho_{D1}\rho_{D0} - [(1-\rho_{D1}^2)(1-\rho_{D0}^2)]^{1/2}, \\ \overline{\rho}_{10} &= \rho_{D1}\rho_{D0} + [(1-\rho_{D1}^2)(1-\rho_{D0}^2)]^{1/2}. \end{aligned} \tag{3}$$

Since the data are informative about the identified correlation parameters ρ_{D1} and ρ_{D0}, Equation (3) provides a vehicle for learning about ρ_{10}. Among other contributions, Heckman et al. (1997) extend these results to nonparametric settings using the Fréchet-Hoeffding bounds.

Koop and Poirier (1997), as well as Poirier (1998), complete Vijverberg (1993)'s analysis by showing that the learning taking place about the non-identified correlation parameter is actually driven by the prior dependence between the identified and non-identified parameters. Prior and posterior marginal distributions need not be identical, because "data information on the identified parameters 'spills-over' to the non-identified parameter if the two groups are *a priori* dependent" (Koop and Poirier, 1997, p. 218). However, prior beliefs about ρ_{10} are not updated by the data *conditional* on ρ_{D1} and ρ_{D0}. The learning process

[6] Papers dealing with Bayesian inference of treatment effects include Poirier and Tobias (2003); Li et al. (2004); Tobias (2006); Li and Tobias (2008), but none of them assume a latent structure with factors.

only operates through the bounds derived in Equation (3).

The support restrictions can be more or less informative about the range of ρ_{10}, depending on how tight the bounds are. Unfortunately, this problem can only be assessed on a case-by-case basis. Three empirical examples drawn from published research are re-analyzed in Vijverberg (1993) and it appears that in some cases, it is possible to actually learn about the magnitude and/or the sign of the unobserved correlation.[7] Poirier and Tobias (2003) present different sets of results from synthetic data showing that the bounds sometimes convey a lot of information and make it possible to accurately approximate ρ_{10}, and sometimes turn out to be completely uninformative about this parameter. Interestingly, they note that "researchers will know when the support restrictions provide information about the unidentified parameter, because this can be determined from the identified correlation coefficients" (Poirier and Tobias, 2003, p. 263).

A further issue concerns the choice of the prior on the restricted support. Since learning only takes place about the bounds of the non-identified correlation, and the data remain silent about the correlation parameter within these bounds, the prior has a clear influence on any inference about ρ_{10}. Vijverberg (1993), Koop and Poirier (1997), and subsequent papers employ a uniform prior over the interval $[\underline{\rho}_{10}, \overline{\rho}_{10}]$ for ρ_{10}, conditional on ρ_{D1} and ρ_{D0}, and a uniform prior over $[-1, 1]$ for these latter parameters. These priors are thus vague within the bounds and guarantee that the covariance matrix is positive semidefinite. Specifying more informative priors might significantly affect the results, "as the prior will affect the location and shapes of the posteriors within the conditional supports" (Tobias, 2006, p. 14). Poirier and Tobias (2003) establish the same conclusion with a simulation study.

The extended Roy model appears as a natural framework within which to explore how constraints on the covariance matrix make it possible to learn about the non-identified correlation between the potential outcomes. This approach can be generalized to any type of model where one or more non-identified parameters are *a priori* not independent of some other identified parameters. For example, similar to the cases derived in Vijverberg (1993)'s appendix, Tobias (2006) uses a four-potential-outcome model and investigates the sources of learning about the four non-identified cross-regime correlation coefficients. He shows that learning about the unobserved correlations can take place through a mechanism very similar to that in the single-outcome case. However, complications arise with the learning about partial correlations, inasmuch as the support of some of these partial correlations cannot be restricted.

Bayesian approaches to estimating the distribution of outcome gains in the presence of non-identified

[7] In the examples in Heckman et al. (1997), the nonparametric bounds are rather wide.

parameters has been the subject of several papers, including Poirier and Tobias (2003); Li et al. (2004); Tobias (2006); Li and Tobias (2008). Once the range of ρ_{10} has been pinned down through the support restrictions, no complications arise in the derivation of the different treatment effects. In cases where the learning mechanism fails to point identify the correlation through information in support restrictions, an informative prior has to be placed on the correlation parameter in order to be able to compute distributional treatment effects. This solution is often unsatisfactory, as it is usually difficult in practice to have well-defined prior beliefs about this unobserved correlation. We address this problem by using a factor structure joined with auxiliary measurements on the factors. With this extra information and extra data in hand, we can point identify the model.

Estimating the model without using the joint distribution of potential outcomes. As noted in Heckman (1990, 1992) and Vijverberg (1993), the likelihood function does not contain any information about the correlation between the potential outcomes (Y_1, Y_0). Only the joint distributions of both (U_D, U_1) and (U_D, U_0) are actually needed to estimate the selection model (Heckman, 1990). Building on this idea, Chib (2007) develops an alternative approach where the joint distribution of the potential outcomes is not required to analyze the treatment model. Since either pair $(D_i = 1, Y_{1i})$ or $(D_i = 0, Y_{0i})$ is observed for each individual i, it is possible to estimate the model by specifying only the joint distributions of (U_D, U_1) and (U_D, U_0). This approach is straightforward to implement using standard MCMC methods,[8] and bypasses the assumptions required to identify the joint distribution of the potential outcomes and the resulting need to specify an informative prior for their non-identified correlation. However, a direct consequence of this approach is that it is impossible to go beyond estimating mean treatment effects and to compute distributional treatment effects.

Distributional treatment parameters have a particular policy relevance. They make it possible to go beyond mean treatment effects by providing a more complete picture of the impact of the treatment. For example, the average treatment effect $\mathrm{E}[\Delta]$ of a given program can be equal to zero in two distinct cases: either the program under study has absolutely no impact, or the proportion of the population that benefits from the program is equal to the proportion that suffers from it (Conti and Heckman, 2010, provide examples of such cases). Computing the distribution of outcome gains would uncover this phenomenon, while mean treatment effects would not reveal it (Heckman et al., 1997; Heckman and Smith, 1998; Aakvik et al.,

[8] For a general introduction to Markov chain Monte Carlo methods see, for example, Gamerman and Lopes (2006).

2005). Identifying fractions of the population that benefit or suffer from a given treatment can be of crucial importance in applications, such as in health economics (for an example, see Conti et al., 2012).

2.3 Potential outcomes model with a latent factor structure

Restricting the structure of $\text{Cov}\begin{pmatrix} U_D \\ U_1 \\ U_0 \end{pmatrix}$ can produce identification. In particular, specifying a model with underlying latent factors for the error terms can solve the problem of constructing the distribution of treatment effects (see Carneiro et al., 2003; Aakvik et al., 2005; Abbring and Heckman, 2007). The latent factors are assumed to explain all of the correlation between the different observed variables, including the correlation between the potential outcomes. Provided that the dimensionality of the number of factors is restricted, the covariance matrix of the error terms is fully identified by the latent structure of the model.

This approach is used in the empirical literature. It solves the identification problem in a very simple way. In addition, in many cases the latent factors play an important role in the model, as they are assumed to capture some specific features of interest. For example, in the growing economic literature on cognitive and noncognitive abilities, such models have been extensively used to capture the effects of latent cognitive ability and personality traits on educational choices, labor market outcomes, social outcomes, and health-related outcomes (Carneiro et al., 2003; Hansen et al., 2004; Heckman et al., 2006, 2010; Conti et al., 2012).

We assume that a vector of K latent factors θ drives the unobserved correlation between the different outcomes of interest in Equation (1):

$$\begin{aligned} D &= \mathbf{1}(D^\star > 0)\,, \\ D^\star &= Z'\gamma + \alpha_D'\theta + \varepsilon_D, \\ Y_1 &= X'\beta_1 + \alpha_1'\theta + \varepsilon_1, \\ Y_0 &= X'\beta_0 + \alpha_0'\theta + \varepsilon_0, \end{aligned} \tag{4}$$

where the error terms are assumed to be centered, $\text{E}[\varepsilon_D] = \text{E}[\varepsilon_1] = \text{E}[\varepsilon_0] = 0$, with finite variances, $V[\varepsilon_D] = 1$, $V[\varepsilon_1] = \sigma_1^2$, $V[\varepsilon_0] = \sigma_0^2$, and mutually independent, $\varepsilon_D \perp\!\!\!\perp \varepsilon_1 \perp\!\!\!\perp \varepsilon_0$. Separability between observables and unobservables is assumed. It is further assumed that $(\theta, \varepsilon) \perp\!\!\!\perp (Z, X)$. The latent factors θ are also assumed to be centered, $\text{E}[\theta] = 0$, and to have a covariance matrix $V[\theta] = \Sigma_\theta$.[9] All of the dependence between the unobservables of the model is driven by the latent factors θ.

[9] In some applications, this restriction is relaxed, and other centerings are used. See Heckman et al. (2011).

A model with discrete potential outcomes would not be more complicated to handle. Consider for instance the dichotomous outcome case treated in Aakvik et al. (2005), where for $j = 0, 1$:

$$\begin{aligned} Y_j &= \mathbf{1}\big(Y_j^\star > 0\big)\,, \\ Y_j^\star &= X'\beta_j + \alpha_j'\theta + \varepsilon_j, \qquad\qquad \varepsilon_j \sim \mathcal{N}(0,1)\,. \end{aligned}$$

In this case, the outcome gain can take three distinct values, indicating if the individual benefits from the treatment ($\Delta = 1$), suffers from it ($\Delta = -1$) or is indifferent ($\Delta = 0$). Other types of discrete outcomes can be considered as well (e.g., Li and Tobias, 2008, focus on ordered potential outcomes).

Further assumptions and restrictions are required to identify the latent structure of the model. In many empirical applications, additional information can be exploited for this purpose. Abbring and Heckman (2007) review different identification strategies, including approaches based on the availability of a single proxy measure (e.g., a test score), of several measurements with or without choice data (Hansen et al., 2004; Heckman et al., 2006), or of panel data on the realized outcomes (Carneiro et al., 2001, 2003; Cunha et al., 2004, 2006). The case where proxy variables are available is of particular interest, since these variables allow the researcher to provide a meaningful interpretation to the factors, depending on their mapping with the measurements.

Assuming that the latent factors θ can be measured by Q continuous variables $M = (M_1, \ldots, M_Q)'$, the following system can be added to the model:

$$M = \mu(X) + \Lambda\,\theta + \varepsilon_M, \tag{5}$$

where $\mu(X)$ is a deterministic function of the covariates X, $\Lambda = (\lambda_1, \ldots, \lambda_Q)'$ is a $(Q \times K)$-dimensional matrix of factor loadings, and ε_M is a Q-dimensional vector of error terms that are assumed to be mutually independent, independent of the factors and of the covariates ($\varepsilon_M \perp\!\!\!\perp \theta \perp\!\!\!\perp X$), to have zero means and finite variances, i.e., $\mathrm{E}[\varepsilon_M] = 0$ and $V[\varepsilon_M] \equiv D_\varepsilon = \mathrm{diag}(\sigma^2_{M_1}, \ldots, \sigma^2_{M_Q})$.

The linearity and separability of the measurement system is strictly *not* required in our approach. Under the conditions presented in Cunha et al. (2010), it is straightforward to establish that a nonlinear, nonseparable (in X, θ, and ε_M) version of this model is nonparametrically identified. We adopt the linear separable model for sake of familiarity and computational convenience.

The covariance structure of the latent part of the measurement system is:

$$\text{Cov}(M|X) = \Lambda \Sigma_\theta \Lambda' + D_\varepsilon,$$

where the covariance matrix of the factors Σ_θ can be very general, allowing for correlated components of the vector θ. The observation of the correlation between the measurements makes it possible to extract $\Lambda \Sigma_\theta \Lambda'$, but how to uniquely identify Λ and Σ_θ? There is a fundamental indeterminacy problem, since for any arbitrary nonsingular matrix R of dimension $(K \times K)$, the covariance structure is unchanged after assigning $\Lambda^\star = \Lambda R$ and $\theta^\star = R^{-1}\theta$. Assuming that the factors are uncorrelated (i.e., Σ_θ diagonal) only partially solves this problem, since the system remains the same by choosing any arbitrary orthogonal matrix such that $R^{-1} = R'$ for the transformation. This is the well-known *rotation* problem.

To solve the rotation problem and achieve identification, some appropriate restrictions have to be made on the factor loadings matrix and/or on the covariance matrix of the latent factors (see, e.g., Anderson and Rubin, 1956). One solution is to assume that Σ_θ is diagonal and that some measures are dedicated, i.e., one component of θ is associated with a block of measurements on M. The scale of each factor is set by fixing one loading to a given value, for example to 1. This solution is particularly appealing in psychology, where factors are often postulated to be uncorrelated. This approach has been widely and successfully implemented in the empirical economic literature on personal abilities and personality traits, where proxy variables provided by psychometric tests are related to well-defined psychological constructs (examples can be found in Carneiro et al., 2001, 2003; Hansen et al., 2004; Heckman et al., 2006, 2010).

Many other solutions exist to the factor rotation problem. For example, Geweke and Zhou (1996) show that a block lower triangular factor loadings matrix, assumed to be of full rank, with strictly positive diagonal elements, and associated with independent standard normal factors ($\Sigma_\theta = I_k$), rule out any rotation. (This solution has been adopted by Lopes and West, 2004; Frühwirth-Schnatter and Lopes, 2010.) With the identification of the latent structure of the measurement system in hand, it is straightforward to identify the factor loadings of the outcome system in Equation (4) from the observed covariance between measurements and outcomes.

The measurement system in Equation (5) can easily be extended to accomodate mixed continuous and discrete measurements (see Carneiro et al., 2003, and Abbring and Heckman, 2007). This is particularly useful in practice, where psychometric tests often provide a small number of possible answers to a given

question (e.g., Likert scales). In this case, the variables in M corresponding to discrete measurements can be regarded as latent, and the observed variables are then obtained through the specification of a threshold-crossing mechanism. This introduces no further complications with identification other than the standard normalization assumptions used in discrete data analysis and is straightforward to implement.

The identification strategy can be extended to the nonparametric case as well. Under some regularity conditions and support assumptions, the distribution of the latent factors θ and of the error terms ε can be nonparametrically identified. See Cunha et al. (2010).

For specificity, normality assumptions are adopted for the distributions of the error terms not arising from θ:

$$\begin{aligned} \varepsilon &= (\varepsilon_D, \varepsilon_1, \varepsilon_0, \varepsilon_M')' \sim \mathcal{N}(0, \Sigma_\varepsilon)\,, \\ \Sigma_\varepsilon &= \text{diag}(1, \sigma_1^2, \sigma_0^2, \sigma_{M_1}^2, \ldots, \sigma_{M_Q}^2). \end{aligned} \tag{6}$$

As for the factors θ, no particular distributional assumptions are made at this point, so as to remain as general as possible. We just assume that they are centered and have bounded covariances, and that their distribution function is parameterized by a vector ψ_θ.[10] The normal case will be considered in Section 3.2, and other more general distributional forms will be discussed in Section 3.1. The overall model consists of the set of parameters $\Gamma = (\gamma', \beta_1', \beta_0', \beta_{M_1}', \ldots, \beta_{M_Q}', \alpha_D', \alpha_1', \alpha_0', \lambda_1', \ldots, \lambda_Q', \sigma_1^2, \sigma_0^2, \sigma_{M_1}^2, \ldots, \sigma_{M_Q}^2, \psi_\theta')'$ and of the latent variables $D^\star$ and θ, for which prior beliefs can be updated by the application of Bayes' rule after the observation of the $data = (D, Y, M, Z, X)$. Standard MCMC schemes for Bayesian inference, such as Gibbs sampling and the Metropolis-Hastings algorithm, can be applied to approximate the posterior distribution of the model parameters as well as to perform model comparison and criticism. See Gamerman and Lopes (2006) for further details on these and other MCMC schemes.

3 Deriving Treatment Effects

The model specified in Equation (4) can be used to estimate both distributional treatment parameters as well as mean treatment impacts. In this paper, we focus on the Average Treatment Effect (ATE), which is the expected outcome gain from the treatment for a randomly chosen individual, the effect of Treatment on the Treated (TT), which is the expected outcome gain for an individual randomly chosen from the subgroup of

[10] E.g., if the factors are assumed to follow a mixture of normals then ψ_θ is the set of mixture means, variances and weights.

people who are actually treated, the Marginal Treatment Effect, which is the expected outcome gain for an individual with a given value of unobservables $U_D = \alpha_D'\theta + \varepsilon_D$, and the fraction of the population that would benefit from treatment. These treatment effects are most commonly invoked and have wide applicability in the empirical literature. They represent a simple case to study in order to explain the methodology that can easily be extended to estimate other treatment parameters.[11]

Let the outcome gain be defined as $\Delta \equiv Y_1 - Y_0$. Distributional treatment effects are formally defined in Aakvik et al. (2005). Following their notation, let $\mathcal{A}$ be any measurable set, and $\mathbf{1}_{\mathcal{A}}(\zeta)$ be the indicator function for the event $\zeta \in \mathcal{A}$. The distributional versions of the treatment effects corresponding to ATE, TT and MTE can then be defined for a given set of covariates x and z, respectively, as:

$$\mathrm{E}[\mathbf{1}_{\mathcal{A}}(\Delta)\,|X = x] = \int \mathbf{1}_{\mathcal{A}}(\Delta)\; p(\Delta|X = x)d\Delta, \tag{7}$$

$$\mathrm{E}[\mathbf{1}_{\mathcal{A}}(\Delta)\,|D = 1, Z = z, X = x] = \int \mathbf{1}_{\mathcal{A}}(\Delta)\; p(\Delta|D = 1, Z = z, X = x)d\Delta, \tag{8}$$

$$\mathrm{E}[\mathbf{1}_{\mathcal{A}}(\Delta)\,|U_D = u, X = x] = \int \mathbf{1}_{\mathcal{A}}(\Delta)\; p(\Delta|U_D = u, X = x)d\Delta. \tag{9}$$

This very general definition of the distributional treatment effects allows the derivation of a broad range of treatment parameters. Depending on the goals of the analyst and on the questions to be addressed, the specification of the set $\mathcal{A}$ determines the types of treatment parameters that can be inferred. For example, if $\mathcal{A} = [0, +\infty)$, Equation (7) becomes $\Pr(\Delta > 0|X = x)$ and measures the proportion of people, at a level of covariates x, who benefit from the treatment. Similarly, $\Pr(\Delta > 0|D = 1, Z = z, X = x)$ obtained from Equation (8) with the same set $\mathcal{A}$ measures the proportion of people *taking the treatment* who benefit from it (see Heckman et al., 1997).

The conventional mean treatment effects are obtained in a similar fashion by averaging the outcome gains:

$$\mathrm{ATE}(x) \equiv \mathrm{E}[\Delta|X = x] = \int \Delta\, p(\Delta|X = x)d\Delta, \tag{10}$$

$$\mathrm{TT}(z, x) \equiv \mathrm{E}[\Delta|D = 1, Z = z, X = x] = \int \Delta\, p(\Delta|D = 1, Z = z, X = x)d\Delta, \tag{11}$$

$$\mathrm{MTE}(u, x) \equiv \mathrm{E}[\Delta|U_D = u, X = x] = \int \Delta\, p(\Delta|U_D = u, X = x)d\Delta. \tag{12}$$

[11] E.g., treatment effects on the untreated (see Heckman et al., 1998), local average treatment effects (Imbens and Angrist, 1994; Heckman and Vytlacil, 1999) and policy-relevant treatment effects (Heckman and Vytlacil, 2001).

Since the different versions of the treatment effects defined in Equations (7) to (12) are derived from the conditional distributions of the outcome gain, one way to identify these parameters is to recover the distributions:

$$p(\Delta|X), \qquad p(\Delta|D=1,Z,X), \qquad p(\Delta|U_D=u,X). \tag{13}$$

Notice, however, because of the linearity of the mean operator, we can identify treatment parameters (10)–(12) using only the marginal distributions of Y_1 and Y_0, i.e. $f_{Y_1}(y_1|X)$, $f_{Y_0}(y_0|X)$ for ATE, $f_{Y_1}(y_1|D=1,Z=z,X=x)$ and $f_{Y_0}(y_0|D=1,Z=z,X=x)$ for TT, and $f_{Y_1}(y_1|U_D=u,X=x)$ and $f_{Y_0}(y_0|U_D=u,X=x)$ for MTE. We need the joint distribution to form the distribution of (Y_1,Y_0) to estimate the proportion of the population that benefits from the policy. In Section 3.1, we derive these distributions and the corresponding treatment effects. Mean treatment parameters are presented along with their distributional versions. More specifically, we work the example $\Pr(\Delta>0|X)$ in the overall population, in the treated population, and also at different margins of taking treatment.

3.1 Deriving treatment parameters

This section introduces general formulae for the derivation of the treatment effects. The framework is very flexible and can be extended to accommodate more general distributional and functional forms. The treatment effects are derived for given sets of covariates Z and X, which can overlap or not. As a consequence, and to simplify notation, X and/or Z will be dropped from the conditioning sets of the following expressions when they are redundant. For expositional simplicity, we use the distribution of $\Delta=Y_1-Y_0$ to identify all treatment effects, although it is not strictly required for ATE, TT, and MTE. The treatment effects can be derived using prior predictive distributions or posterior distributions. We use the former to simplify the exposition but they are not strictly required.

Average Treatment Effect. The average treatment effect and its distributional version are derived from the distribution of the outcome gains. For a given set of covariates x, this distribution is obtained by integrating out the latent factors θ and the model parameters Γ:

$$p(\Delta|x)=\iint p(\Delta|x,\theta,\Gamma)p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma, \tag{14}$$

where $p(\Gamma)$ and $p(\theta|\Gamma)$ are used to denote the distributions of model parameters and of the latent factors in a very generic way. In a posterior approach, $p(\Gamma)$ would be replaced by the posterior distribution $p(\Gamma|data)$, where Γ is updated after observing the $data$ through the application of Bayes' rule. As for the conditional distribution $p(\theta|\Gamma)$, we will see in Section 3.3 that the information available to measure the latent factors determines how these factors are integrated out.

From Equation (14), the distributional and mean treatment effects read:

$$\begin{aligned}
\Pr(\Delta > 0|x) &= \int_0^\infty p(\Delta|x)d\Delta, \\
&= \iint \Pr(\Delta > 0|x,\theta,\Gamma)\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma, \\
\text{ATE}(x) &= \iint \mathrm{E}[\Delta|x,\theta,\Gamma]\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma.
\end{aligned} \tag{15}$$

Effect of Treatment on the Treated. Similarly, the distribution of the gains in the treated state is expressed as:

$$\begin{aligned}
&p(\Delta|D=1,z,x) \\
&\quad = \int \left[\int p(\Delta|D=1,z,x,\theta,\Gamma)p(\theta|D=1,z,x,\Gamma)d\theta \right] p(\Gamma|D=1,z,x)d\Gamma, \\
&\quad = \iint p(\Delta|x,\theta,\Gamma)\frac{\Pr(D=1|\theta,z,\Gamma)\, p(\theta|\Gamma)p(\Gamma)}{\Pr(D=1|z)}d\theta d\Gamma,
\end{aligned} \tag{16}$$

where the last line results, after simplification, from the application of Bayes' rule on the conditional distributions of θ and Γ, respectively. The covariates and the factors explain all the dependence between (Y_0, Y_1) and D, hence $\Delta \perp\!\!\!\perp D \mid X, \theta$. Equation (16) can be viewed as a weighted average of the conditional distribution of the outcome gains:

$$p(\Delta|D=1,z,x) = \iint \omega^{\text{TT}} p(\Delta|x,\theta,\Gamma)p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma,$$

where the weights are:

$$\begin{aligned}
\omega^{\text{TT}} \equiv \omega^{\text{TT}}(z,\theta,\Gamma) &= \frac{\Pr(D=1|z,\theta,\Gamma)}{\Pr(D=1|z)}, \\
&= \frac{\Pr(D=1|z,\theta,\Gamma)}{\iint \Pr(D=1|z,\theta,\Gamma)\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma}.
\end{aligned}$$

Distributional and mean treatment effects conditional on taking treatment are derived in the same fashion as the average treatment effect:

$$\begin{aligned}\Pr(\Delta > 0|D=1,z,x) &= \int_0^\infty p(\Delta|D=1,z,x)d\Delta, \\ &= \iint \omega^{\text{TT}} \Pr(\Delta > 0|x,\theta,\Gamma)\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma, \\ \text{TT}(x) &= \iint \omega^{\text{TT}} \text{E}[\Delta|x,\theta,\Gamma]\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma.\end{aligned} \tag{17}$$

Marginal Treatment Effect. Marginal treatment effects are constructed in the same fashion as ATE and TT, using the following distribution of outcome gains conditional on unobservables U_D:

$$p(\Delta|U_D=u,x) = \iint \omega^{\text{MTE}} p(\Delta|x,\theta,\Gamma)p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma,$$

where the weights are:

$$\begin{aligned}\omega^{\text{MTE}} \equiv \omega^{\text{MTE}}(\theta,\Gamma) &= \frac{\Pr(U_D=u|\theta,\Gamma)}{\Pr(U_D=u)}, \\ &= \frac{\Pr(U_D=u|\theta,\Gamma)}{\iint \Pr(U_D=u|\theta,\Gamma)\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma}.\end{aligned}$$

Distributional and mean treatment effects conditional on unobservables U_D can be written as:

$$\begin{aligned}\Pr(\Delta > 0|U_D=u,x) &= \int_0^\infty p(\Delta|U_D=u,x)d\Delta, \\ &= \iint \omega^{\text{MTE}} \Pr(\Delta > 0|x,\theta,\Gamma)\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma, \\ \text{MTE}(x) &= \iint \omega^{\text{MTE}} \text{E}[\Delta|x,\theta,\Gamma]\, p(\theta|\Gamma)p(\Gamma)d\theta d\Gamma.\end{aligned} \tag{18}$$

Thus, we can compute the proportion of people with a given $X = x$ who benefit from treatment as well as the return to people at the margin of participation in the program (i.e., those indifferent at a given level of $U_D = u$ given z and x).

There are several benefits of our approach, which we now enumerate.

- **From the continuous outcome case to more complicated functional forms.** In Section 2.3, trivari-

ate normality of the potential outcomes and of the treatment status conditional on the factors θ was assumed. Using this analysis, the expressions of the treatment effects derived above, which are all weighted averages of the expected conditional outcome gains and of their distributions, can be further simplified with the following relationships:

$$\begin{aligned}
p(\Delta|x,\theta,\Gamma) &= \frac{1}{\sqrt{\sigma_1^2+\sigma_0^2}}\phi\left(\frac{\Delta - x'(\beta_1-\beta_0)-(\alpha_1-\alpha_0)'\theta}{\sqrt{\sigma_1^2+\sigma_0^2}}\right), \\
\Pr(\Delta>0|x,\theta,\Gamma) &= \Phi\left(\frac{x'(\beta_1-\beta_0)+(\alpha_1-\alpha_0)'\theta}{\sqrt{\sigma_1^2+\sigma_0^2}}\right), \\
\mathrm{E}[\Delta|x,\theta,\Gamma] &= x'(\beta_1-\beta_0)+(\alpha_1-\alpha_0)'\theta, \\
\Pr(D=1|z,\theta,\Gamma) &= \Pr\big(\varepsilon_D > -z'\gamma-\alpha_D'\theta|z,\theta,\Gamma\big) = \Phi\big(z'\gamma+\alpha_D'\theta\big), \\
\Pr(U_D=u|\theta,\Gamma) &= \Pr\big(\varepsilon_D = u-\alpha_D'\theta|\theta,\Gamma\big) = \phi\big(u-\alpha_D'\theta\big),
\end{aligned} \tag{19}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the probability density function and the cumulative distribution function of the standard normal distribution.

The very general form of the formulae derived in the previous section allows us to extend the analysis to more complicated functional forms, such as those for discrete outcomes. For instance, in the dichotomous case (Aakvik et al., 2005), the probabilities of the three values the outcome gain can take are:

$$\begin{aligned}
\Pr(\Delta=1|x,\theta,\Gamma) &= \Pr(Y_1=1,Y_0=0|X=x,\theta,\Gamma), \\
&= \Pr(Y_1=1|X=x,\theta,\Gamma)\Pr(Y_0=0|X=x,\theta,\Gamma), \\
&= \Phi\big(x'\beta_1+\alpha_1'\theta\big)\left[1-\Phi\big(x'\beta_0+\alpha_0'\theta\big)\right], \\
\Pr(\Delta=-1|x,\theta,\Gamma) &= \left[1-\Phi\big(x'\beta_1+\alpha_1'\theta\big)\right]\Phi\big(x'\beta_0+\alpha_0'\theta\big), \\
\Pr(\Delta=0|x,\theta,\Gamma) &= 1-\Pr(\Delta=1|X=x,\theta,\Gamma)-\Pr(\Delta=-1|X=x,\theta,\Gamma).
\end{aligned}$$

These expressions can then be used to derive the different versions of the treatment effects, including their distributional and mean versions. See Li and Tobias (2008) for the case of ordered outcomes.

- **Adding a measurement system facilitates the interpretation of the factors.** In some cases, the latent factors capture some theoretical concepts (e.g., personality traits or cognitive abilities, widely

used in psychology and more recently in economics) and it might be of interest to analyze how the different treatment effects are distributed along the distribution of these factors. Measurements also facilitate identification.

- **Non-normal factors.** The normal case presented in the coming Section 3.2 is standard in factor analysis and represents the baseline model of most empirical analyses. Nevertheless, the normality assumption may be too restrictive in practice and result in biased estimators for the factor loadings, which would affect the estimation of the treatment effects. The general approach presented here is not wedded to the normality assumption (see, e.g., Abbring, 2011). Any distribution that is centered with finite covariance matrix, and that can be easily sampled from, can be used for the latent factors. Common examples include Student-t distributions that allow fatter tails than the normal distribution, and mixtures of normals that make it possible to approximate more peculiar distributions, such as multimodal distributions. Li et al. (2004) derive the predictive distributions of outcome gains for these two alternative distributions of the factors.

The flexibility of our approach comes at a cost. In practice, the precision of the estimators for the treatment effects depends on the quality of the numerical integration of the latent factors. This aspect will be illustrated and discussed in Section 4.

3.2 Treatment effects in the normal case

Assuming that the latent factors are normally distributed,

$$\theta \sim \mathcal{N}(0, \Sigma_\theta)\,,$$

the joint distribution of the treatment index $D^\star$ and of the potential outcomes Y_1 and Y_0 is straightforward to derive. In the wake of the analysis conducted in Section 3.1, the conditional distribution is expressed as[12]

$$\begin{pmatrix} D^\star \\ Y_1 \\ Y_0 \end{pmatrix} | \, \theta \sim \mathcal{N}\left(\begin{bmatrix} Z'\gamma + \alpha_D'\theta \\ X'\beta_1 + \alpha_1'\theta \\ X'\beta_0 + \alpha_0'\theta \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_0^2 \end{bmatrix} \right), \tag{20}$$

[12]The conditioning on the covariates Z and X, as well as and on the model parameters Γ, is kept implicit and therefore omitted.

while the following unconditional joint distribution is implied:

$$\begin{pmatrix} D^\star \\ Y_1 \\ Y_0 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} Z'\gamma \\ X'\beta_1 \\ X'\beta_0 \end{bmatrix}, \begin{bmatrix} \alpha_D'\Sigma_\theta\alpha_D + 1 & \alpha_D'\Sigma_\theta\alpha_1 & \alpha_D'\Sigma_\theta\alpha_0 \\ \alpha_D'\Sigma_\theta\alpha_1 & \alpha_1'\Sigma_\theta\alpha_1 + \sigma_1^2 & \alpha_1'\Sigma_\theta\alpha_0 \\ \alpha_D'\Sigma_\theta\alpha_0 & \alpha_1'\Sigma_\theta\alpha_0 & \alpha_0'\Sigma_\theta\alpha_0 + \sigma_0^2 \end{bmatrix} \right). \tag{21}$$

Either expression can be used for the derivation of the likelihood function. However, there are some cases where the analyst might prefer to work with one version or the other. If the latent factors are required to conduct further posterior analysis, then they should explicitly be incorporated into the model (e.g., to investigate how a treatment parameter depends on a particular factor of interest).

On the other hand, if the factors are of secondary importance, then it is preferable to work with the unconditional distribution of Equation (21). The benefit of the latter approach is in simplification of the computation. Since Markov chain Monte Carlo methods simulate the parameters to bypass the problems inherent in performing complicated integrations, any integration that can be performed analytically before the simulation should be carried out. *Collapsing* of the state space of the algorithm results in more efficient sampling schemes (Liu, 1994; Liu et al., 1994).

In the framework of the potential outcomes factor structure model, the normality of the factors greatly simplifies the problem and leads to compact form solutions for most treatment effect parameters. For this purpose, the joint distribution of the treatment index and of the outcome gain is easier to work with to derive treatment effects and their distribution:

$$\begin{pmatrix} D^\star \\ \Delta \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} Z'\gamma \\ X'(\beta_1 - \beta_0) \end{bmatrix}, \begin{bmatrix} \alpha_D'\Sigma_\theta\alpha_D + 1 & \alpha_D'\Sigma_\theta(\alpha_1 - \alpha_0) \\ \alpha_D'\Sigma_\theta(\alpha_1 - \alpha_0) & \sigma_\Delta^2 \end{bmatrix} \right), \tag{22}$$

where $\sigma_\Delta^2 \equiv (\alpha_1 - \alpha_0)'\Sigma_\theta(\alpha_1 - \alpha_0) + \sigma_1^2 + \sigma_0^2$. Note however that this joint distribution is only useful for the computation of the treatment parameters, and cannot be used for other purposes such as model fitting, due to the loss of information compared to Equation (21). For simplicity, the different versions of the outcome gains distributions and of the treatment parameters are presented conditional on Γ in the sequel.

Distributions of outcome gains. Using general results on the bivariate normal distribution, the distributions of outcome gains in the overall population, conditional on taking treatment and conditional on

unobservables U_D read, respectively:

$$
\begin{aligned}
p(\Delta|x,\Gamma) &= \frac{1}{\sigma_\Delta}\phi\left(\frac{\Delta - x'(\beta_1-\beta_0)}{\sigma_\Delta}\right), \\
p(\Delta|D=1,z,x,\Gamma) &= \frac{\Pr(D=1|\Delta,z,x,\Gamma)\,p(\Delta|x,\Gamma)}{\Pr(D=1|z,\Gamma)}, \\
p(\Delta|U_D=u,x,\Gamma) &= \frac{1}{\xi}\phi\left(\frac{\Delta - x'(\beta_1-\beta_0) - \frac{\alpha_D'\Sigma_\theta(\alpha_1-\alpha_0)}{\alpha_D'\Sigma_\theta\alpha_D+1}u}{\xi}\right), \\
&\text{with} \quad \xi \equiv \sqrt{\sigma_\Delta^2 - \frac{[\alpha_D'\Sigma_\theta(\alpha_1-\alpha_0)]^2}{\alpha_D'\Sigma_\theta\alpha_D+1}}.
\end{aligned} \tag{23}
$$

The ingredients required to derive the distribution of outcome gains conditional on taking treatment are:

$$
\begin{aligned}
\Pr(D=1|\Delta,z,x,\Gamma) &= \Phi\left(\frac{z'\gamma + \frac{\alpha_D'\Sigma_\theta(\alpha_1-\alpha_0)}{\sigma_\Delta^2}[\Delta - x'(\beta_1-\beta_0)]}{\sqrt{\alpha_D'\Sigma_\theta\alpha_D + 1 - \frac{[\alpha_D'\Sigma_\theta(\alpha_1-\alpha_0)]^2}{\sigma_\Delta^2}}}\right), \\
\Pr(D=1|z,\Gamma) &= \Phi\left(z'\gamma/\sqrt{\alpha_D'\Sigma_\theta\alpha_D+1}\right).
\end{aligned}
$$

Distributional treatment effects. The distributional treatment parameters are, conditional on a set of model parameters Γ:

$$
\Pr(\Delta>0|x,\Gamma) = \Phi\left(\frac{x'(\beta_1-\beta_0)}{\sigma_\Delta}\right), \tag{24}
$$

$$
\Pr(\Delta>0|D=1,z,x,\Gamma) = \frac{\int_0^\infty\int_0^\infty \Pr(\Delta,D^\star|z,x,\Gamma)\,d\Delta dD^\star}{\Pr(D=1|z,\Gamma)}, \tag{25}
$$

$$
\Pr(\Delta>0|U_D=u,x,\Gamma) = \Phi\left(\frac{x'(\beta_1-\beta_0) + \frac{\alpha_D'\Sigma_\theta(\alpha_1-\alpha_0)}{\alpha_D'\Sigma_\theta\alpha_D+1}u}{\xi}\right), \tag{26}
$$

where σ_Δ is the standard deviation of the outcome gain and ξ is defined in Equation (23).

Unfortunately, there is no tractable solution for the parameter shown in Equation (25), which represents the fraction of the treated population benefiting from treatment. Note, however, that the double integration can be collapsed to a single integral by reexpressing this parameter as:

$$
\Pr(\Delta>0|D=1,z,x,\Gamma) = \frac{\int_0^\infty \Pr(\Delta>0|D^\star=\eta,z,x,\Gamma)\Pr(D^\star=\eta|z,\Gamma)\,d\eta}{\Pr(D=1|z,\Gamma)}, \tag{27}
$$

where:

$$\begin{aligned}\Pr(\Delta > 0 | D^\star = \eta, z, x, \Gamma) &= \Phi\left(\frac{x'(\beta_1 - \beta_0) + \frac{\alpha_D' \Sigma_\theta (\alpha_1 - \alpha_0)}{\alpha_D' \Sigma_\theta \alpha_D + 1}(\eta - z'\gamma)}{\xi}\right), \\ \Pr(D^\star = \eta | z, \Gamma) &= \frac{\phi\big((\eta - z'\gamma)/\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}\big)}{\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}}.\end{aligned} \tag{28}$$

Standard Monte Carlo integration methods can then be performed to compute this treatment effect.

Mean treatment effects. The mean treatment parameters read, conditional on a set of model parameters Γ:

$$\begin{aligned}\text{ATE}(x, \Gamma) &= \text{E}[\Delta | X = x, \Gamma] \\ &= x'(\beta_1 - \beta_0), \qquad (29) \\ \text{TT}(z, x, \Gamma) &= \text{E}[\Delta | D^\star > 0, X = x, Z = z, \Gamma], \\ &= x'(\beta_1 - \beta_0) + \sigma_\Delta \text{E}\left[\frac{\Delta - x'(\beta_1 - \beta_0)}{\sigma_\Delta} \middle| \frac{D^\star - z'\gamma}{\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}} > -\frac{z'\gamma}{\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}}\right], \\ &= x'(\beta_1 - \beta_0) + \frac{\alpha_D' \Sigma_\theta (\alpha_1 - \alpha_0)}{\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}} \frac{\phi\big(-z'\gamma/\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}\big)}{\Phi\big(z'\gamma/\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}\big)}, \qquad (30) \\ \text{MTE}(u, x, \Gamma) &= \text{E}[\Delta | U_D = u, X = x, \Gamma], \\ &= x'(\beta_1 - \beta_0) + \frac{\alpha_D' \Sigma_\theta (\alpha_1 - \alpha_0)}{\alpha_D' \Sigma_\theta \alpha_D + 1} u, \qquad (31)\end{aligned}$$

where $\phi(t) \,/\, \Phi(-t)$ is the inverse Mills ratio in the formula for TT.

Integrating out model parameters. A naive approach would be to average the different formulae of the treatment parameters derived above as a function of Γ over the posterior distribution $p(\Gamma | data)$. This approximation might produce accurate results, because we can expect the distributions $p(\Gamma | D = 1, data)$ and $p(\Gamma | U_D = u, data)$ to mimic pretty closely $p(\Gamma | data)$. Nevertheless, there is a potential flaw due to the fact that the conditioning on $D = 1$ and $U_D = u$ is omitted in the integration of Γ for TT and MTE, respectively. Li et al. (2004), as well as Tobias (2006), note that since these two events involve parameters of Γ, they should be taken into account when the parameters Γ are integrated out. The integration can thus be performed in the same way as in Section 3.1, with respect to the posterior distribution $p(\Gamma | data)$ and

with the following weights for the distributional versions of TT and MTE:

$$\omega^{\text{TT}} = \frac{\Pr(D=1|z,\Gamma)}{\int \Pr(D=1|z,\Gamma)\, p(\Gamma) d\Gamma}, \tag{32}$$

$$\omega^{\text{MTE}} = \frac{\Pr(U_D=u|\Gamma)}{\int \Pr(U_D=u|\Gamma)\, p(\Gamma) d\Gamma}, \tag{33}$$

where $\Pr(U_D=u|\Gamma) = \phi\big(u/\sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}\big) / \sqrt{\alpha_D' \Sigma_\theta \alpha_D + 1}$.

In some applications, it is of interest to infer the treatment effects unconditional on covariates X and Z. Common practice is to use the empirical distribution of the covariates as an approximation to $p(X, Z|data)$. However, if there is selection on the observables X and Z, the distribution $p(X, Z|data)$ will look different than $p(X, Z|D=1, data)$. In this case, an appropriate correction is required using weights similar to those derived for Equations (32) and (33) for TT and MTE.

3.3 Available information about the factors, out-of-sample and in-sample treatment effects

Depending on the information available to measure the latent factors, different approaches can be adopted for the estimation of the treatment effects. The first one is a posterior predictive approach that consists of using the model to predict outcome gains for future populations (*out-of-sample* approach), while the second one infers the outcome gains for the actual individuals present in the observed sample (*in-sample* approach).

Out-of-sample approach. The posterior predictive approach is the mainstream in Bayesian analysis (see, for example, Poirier and Tobias, 2003; Li et al., 2004). It focuses on the prediction of outcome gains for future populations that are similar to the observed population represented by the $data$, in terms of observables Z and X and unobservables θ. To derive the treatment effects, it is essential to know what type of information is available about the future individual at the time of the prediction. If both outcomes Y and measurements M are unobserved, the only information that can be exploited to infer θ is the one learned about its distribution (the parameters ψ_θ of the factor distribution are updated during the estimation process). The distribution $p(\theta|\Gamma)$ thus reduces to the posterior distribution of θ obtained from the estimation on the observed sample.

An interesting case arises if the measurements M—all or only a subset of them—are observed along with the covariates X and Z. The sampling procedure of the predictive outcome gains can then be modified to account for this information extracted from the measurement system. Economic examples where this sit-

uation happens are numerous in the empirical literature: proxy variables for personality traits and cognitive ability, such as test scores, are often measured *during childhood*. The goal is to measure the impact of a particular policy program on the outcomes of a future *adult* using a predictive approach, given that only her personal characteristics X and Z, as well as her measurements M are available. How can this extra information, of particular policy relevance, be taken into account to compute the treatment effects?

Concretely, the measurements of the future individual are introduced into the conditioning set of the distribution of the outcome gains, and the latent factors are predicted from the posterior distribution, conditional on these observed measures. In other words, the posterior obtained from the observed sample is used as a *future prior* for the factors to compute the treatment effects. Using the subscript f to denote the future, as yet unobserved, outcome gains, personal characteristics, and factors, the predictive distribution of Δ_f is:

$$p(\Delta_f|x_f, m_f) = \iint p(\Delta_f|x_f, \theta_f, \Gamma)p(\theta_f|x_f, m_f, \Gamma)p(\Gamma)d\theta d\Gamma, \tag{34}$$

where, after simplification, m_f only appears in the conditional distribution of θ_f and thus only serves to derive its posterior predictive distribution. Reading Equation (34) from right to left reveals the updating process at stake: prior beliefs about model parameters Γ are updated after observation of the data; these updated parameters, combined with the observed covariates x_f and measurements m_f for the individual *in the future*, make it possible to predict the latent factors θ_f; in turn, all the information derived from these two steps is used to predict the future outcome gain Δ_f. Similar to TT and MTE, this expression can be expressed as a weighted average of the outcome gain distribution:

$$p(\Delta_f|x_f, m_f) = \iint \omega^M p(\Delta_f|x_f, \theta_f, \Gamma)p(\theta_f)p(\Gamma)d\theta_f d\Gamma,$$

with the weights:

$$\omega^M \equiv \omega^M(x_f, m_f, \Gamma) = \frac{p(m_f|x_f, \theta_f, \Gamma)}{\int p(m_f|x_f, \theta_f, \Gamma)p(\theta_f|x_f, \Gamma)d\theta_f}.$$

Intuitively, outcome gain predictions will be given more importance if they are based on sampled latent factors that are good at explaining the observed measurements m_f.

In-sample approach. In contrast with the approach presented above, where the person *in the future* may or may not be in the actual sample but is still assumed to belong to the same population, in this approach we base our analysis on a given person from the observed sample—or even on the whole sample—to compute the treatment effects. In this case, it is possible to make use of her realized treatment status and of her corresponding outcome, rather than predicting them.

This alternative approach is in the same spirit as classical approaches that rely on a three-step procedure to compute the treatment effects: The measurement system is estimated in the first stage, factor scores are then predicted in the second stage of the analysis, and then plugged in to compute the treatment effects in the third step. Special care is then required to correct for errors in measurement that arise from using predicted factors instead of the true factors (Heckman et al., 2011).

4 Estimating Treatment Effects: A Simulation Study

We generate a simple potential outcomes model built to the same design as Equation (4), and compute the different versions of the distributional treatment effects measuring the proportion of the population benefiting from treatment (i.e., in the overall population, in the treated population, and at different margins of taking treatment). The approach conditional on θ (based on Equation (20)), as well as the unconditional approach (based on Equation (21)), are implemented and compared.

4.1 Simulating and estimating an artificial potential outcomes model

The potential outcomes model is simulated with the following parameters for the outcome system:

$$
\begin{array}{lll}
x = (1, \widetilde{x}), & z = (1, \widetilde{x}, \widetilde{z}), & \Sigma_\theta = 1, \\
\widetilde{x} \sim \mathcal{N}(1, 1)\,, & \widetilde{z} \sim \mathcal{N}(-1, 3)\,, & \theta \sim \mathcal{N}(0, 1)\,, \\
\gamma = (0.5, 1.0, 1.0)', & \alpha_D = 0.8, & \\
\beta_1 = (2.0, 1.0)', & \alpha_1 = 0.9, & \sigma_1^2 = 1.0, \\
\beta_0 = (1.0, 0.5)', & \alpha_0 = -0.6, & \sigma_0^2 = 1.0,
\end{array}
$$

and the latent factor θ is proxied by five continuous measurements generated from the system in Equation (5) specified as:

$$\Lambda = (1.00, 0.90, 0.85, 0.80, 0.70)', \qquad D_\varepsilon = \text{diag}(0.25, 0.16, 0.09, 0.16, 0.25).$$

We consider samples of size $N = 500$ and $N = 5,000$. Within this framework, the fraction of the population being treated is equal to 0.58, and the true expected outcome gain is equal to 1.50. Since $\text{Cov}(U_D, U_1 - U_0) = 1.20$, a positive selection on latent gains occurs in this model.

To gain insights into the computation of the treatment effects, we perform a Monte Carlo experiment and estimate the model for 1,000 data sets, using a standard Gibbs sampler with noninformative priors: a flat prior $\mathcal{N}(0, \infty)$ is assumed for the slope parameters and intercept terms, a normal prior $\mathcal{N}(0, 10)$ for the factor loadings, and for the idiosyncratic variance an inverse-gamma prior $\mathcal{G}^{-1}(2.11, 1.11)$ implying a prior mean and standard deviation of 1 and 3, respectively. To improve the mixing of the MCMC chain, parameter expansion is carried out by freeing the variance of the latent factor in the working model (Ghosh and Dunson, 2009), where the same inverse-gamma prior as for the idiosyncratic variances is used. The parameters are then transformed back to the inferential model where the variance of the factor is fixed to 1, and the first factor loading is constrained to be positive to prevent sign-switching. 10,000 iterations are saved for posterior inference and for the computation of the treatment effects, after a burn-in period of 2,000 iterations. The distributions of the posterior means of the parameters are summarized in Figures 1 and 2, where it appears that the parameters are overall precisely estimated, especially the factor loadings and the idiosyncratic variances that play a major role in the calculation of the distributional treatment effects.

4.2 Computing the treatment effects from the MCMC chains

The Gibbs sampling algorithm used to estimate our potential outcomes model generates, after convergence of the Markov chain, a sample of model parameters $\Gamma^{(1)}, \ldots, \Gamma^{(M)}$ from their posterior distributions. This sample can in the end be used to perform a Monte Carlo integration of Γ in the formulae of the treatment effects and of their distributional versions derived in Section 3. For example, the compact forms of the distributional treatment effects in the approach unconditional on θ, which are derived in Equations (24), (26) and (27), can be combined with the corresponding ingredients derived in Equation (28) and the weights in Equations (32) and (33). Averaging over the post-convergence draws, for a given set of covariates $Z = z$

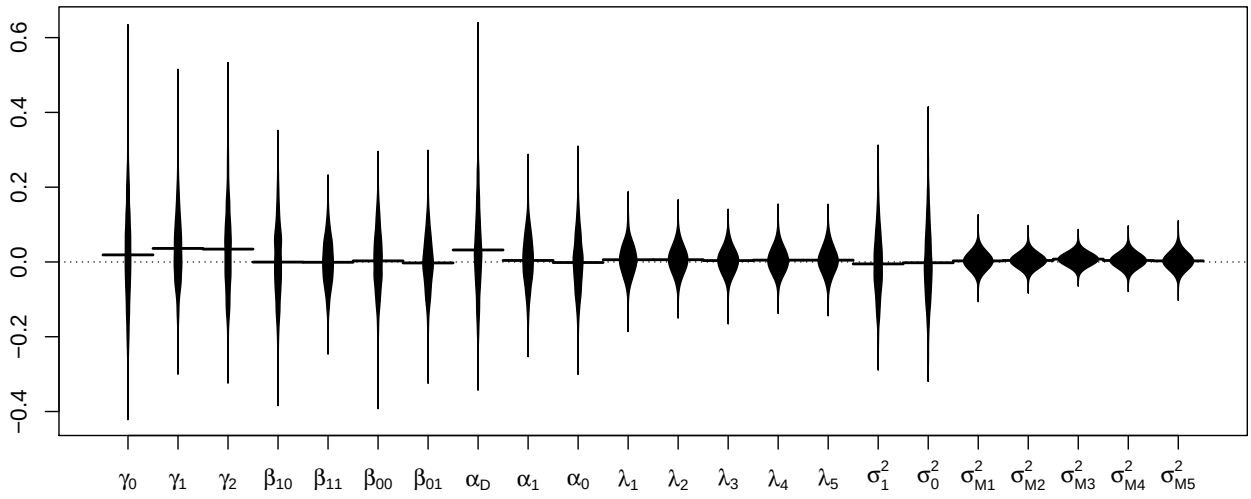

Figure 1: Model with $N = 500$. Distribution of the posterior means of the parameters across the 1,000 Monte Carlo replications. The densities have been demeaned by the true values for ease of comparison.

and $X = x$, provides the following expressions (we call this first approach MC1):

$$\widehat{\Pr}_{\text{MC1}}(\Delta > 0|x) = \frac{1}{M}\sum_{m=1}^{M} \Pr\Big(\Delta > 0|x, \Gamma^{(m)}\Big),$$

$$\widehat{\Pr}_{\text{MC1}}(\Delta > 0|D = 1, z, x) = \frac{\frac{1}{J}\sum_{m=1}^{M}\sum_{j=1}^{J} \Pr\big(\Delta > 0|D^\star = \eta^{(j)}, z, x, \Gamma^{(m)}\big) \Pr\big(D^\star = s^{(j)}|z, x, \Gamma^{(m)}\big) / g(\eta^{(j)})}{\sum_{m=1}^{M} \Phi\left(z'\gamma^{(m)} / \sqrt{\alpha_D^{(m)\prime}\alpha_D^{(m)} + 1}\right)}, \quad (35)$$

$$\widehat{\Pr}_{\text{MC1}}(\Delta > 0|U_D = u, x) = \frac{\sum_{m=1}^{M} \Pr\big(\Delta > 0|U_D = u, x, \Gamma^{(m)}\big) \phi\left(u / \sqrt{\alpha_D^{(m)\prime}\alpha_D^{(m)} + 1}\right) / \sqrt{\alpha_D^{(m)\prime}\alpha_D^{(m)} + 1}}{\sum_{m=1}^{M} \phi\left(u / \sqrt{\alpha_D^{(m)\prime}\alpha_D^{(m)} + 1}\right) / \sqrt{\alpha_D^{(m)\prime}\alpha_D^{(m)} + 1}},$$

where the superscript (m) is used to denote the m^{th} draw from the Gibbs sampler. As shown in Equation (35), the integral in Equation (27) can be approximated through importance sampling by drawing J random values of η from a distribution $g(\eta)$, which has to be truncated to the interval $[0, +\infty)$.[13]

In the same fashion, the distributional treatment effects derived from the approach conditional on θ can

[13] In our simulation study, we sample $1,000$ draws of η from a normal distribution with mean $\frac{1}{M}\sum_m z'\gamma^{(m)}$ and variance $\frac{1}{M}\sum_m \alpha_D^{(m)\prime}\alpha_D^{(m)} + 1$ truncated to the interval $[0, +\infty)$. $g(\eta)$ is therefore the probability density function of the corresponding truncated normal distribution.

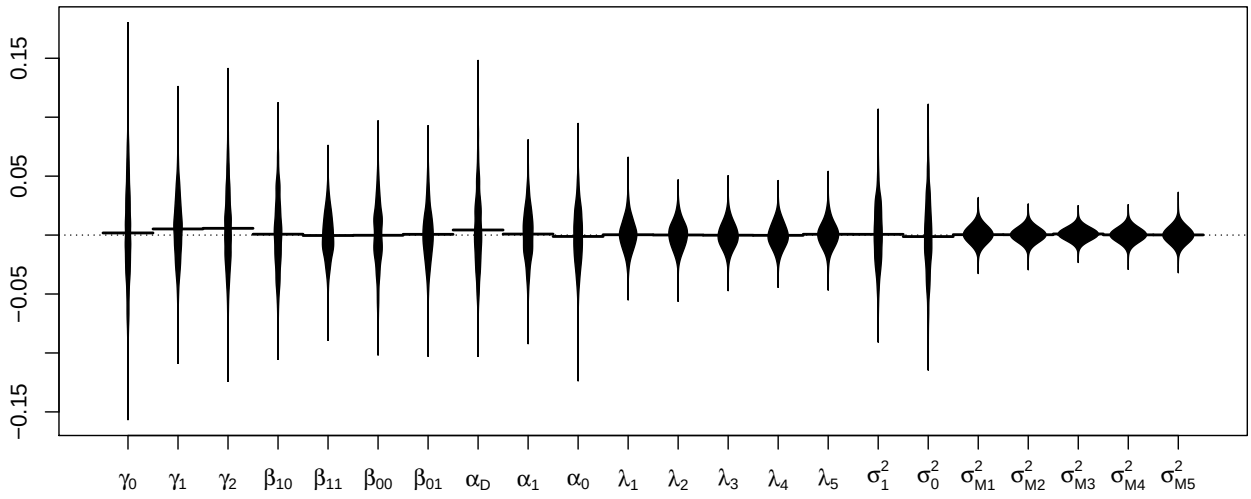

Figure 2: Model with $N = 5,000$. Distribution of the posterior means of the parameters across the 1,000 Monte Carlo replications. The densities have been demeaned by the true values for ease of comparison.

be computed by combining Equations (15) and (17), and (18) with the ingredients in Equation (19), and averaging over the MCMC draws (we call this approach MC2):

$$\widehat{\Pr}_{\text{MC2}}(\Delta > 0|x) = \frac{1}{M}\sum_{m=1}^{M} \Pr\Big(\Delta > 0|x, \theta^{(m)}, \Gamma^{(m)}\Big),$$

$$\widehat{\Pr}_{\text{MC2}}(\Delta > 0|D = 1, z, x) = \frac{\sum_{m=1}^{M} \Pr\big(\Delta > 0|x, \theta^{(m)}, \Gamma^{(m)}\big)\, \Phi\Big(z'\gamma^{(m)} + \alpha_D^{(m)\prime}\theta^{(m)}\Big)}{\sum_{m=1}^{M} \Phi\Big(z'\gamma^{(m)} + \alpha_D^{(m)\prime}\theta^{(m)}\Big)},$$

$$\widehat{\Pr}_{\text{MC2}}(\Delta > 0|U_D = u, x) = \frac{\sum_{m=1}^{M} \Pr\big(\Delta > 0|x, \theta^{(m)}, \Gamma^{(m)}\big)\, \phi\Big(u - \alpha_D^{(m)\prime}\theta^{(m)}\Big)}{\sum_{m=1}^{M} \phi\Big(u - \alpha_D^{(m)\prime}\theta^{(m)}\Big)},$$

where the probability of benefiting from treatment reads:

$$\Pr\Big(\Delta > 0|x, \theta^{(m)}, \Gamma^{(m)}\Big) = \Phi\left(\frac{x'(\beta_1^{(m)} - \beta_0^{(m)}) + (\alpha_1^{(m)} - \alpha_0^{(m)})'\theta^{(m)}}{\sqrt{\sigma_1^{2(m)} + \sigma_0^{2(m)}}}\right).$$

In these expressions, the latent factors θ are integrated out simultaneously with model parameters Γ through Monte Carlo integration. Given the distributional assumptions of our simple model and the posterior predictive approach we adopt, this consists of drawing from the standard normal distribution, $\theta^{(m)} \sim \mathcal{N}(0, 1)$. In an *in-sample* approach, these probabilities would be computed for a given individual i

with covariates z_i and x_i, and the draws of θ_i simulated from their posterior distribution during the Gibbs sampling would be recycled and used for the numerical integration of the factors.

Alternatively, the integration can be achieved sequentially by embedding the integration of the factors in the integration of model parameters (we call this approach MC3):

$$\widehat{\Pr}_{\text{MC3}}(\Delta > 0|x) = \frac{1}{MJ}\sum_{m=1}^{M}\sum_{j=1}^{J}\Pr\Big(\Delta > 0|x,\theta^{(j)},\Gamma^{(m)}\Big),$$

$$\widehat{\Pr}_{\text{MC3}}(\Delta > 0|D=1,z,x) = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J}\Pr\big(\Delta > 0|x,\theta^{(j)},\Gamma^{(m)}\big)\,\Phi\Big(z'\gamma^{(m)} + \alpha_D^{(m)\prime}\theta^{(j)}\Big)}{\sum_{m=1}^{M}\sum_{j=1}^{J}\Phi\Big(z'\gamma^{(m)} + \alpha_D^{(m)\prime}\theta^{(j)}\Big)},$$

$$\widehat{\Pr}_{\text{MC3}}(\Delta > 0|U_D=u,x) = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J}\Pr\big(\Delta > 0|x,\theta^{(j)},\Gamma^{(m)}\big)\,\phi\Big(u - \alpha_D^{(m)\prime}\theta^{(j)}\Big)}{\sum_{m=1}^{M}\sum_{j=1}^{J}\phi\Big(u - \alpha_D^{(m)\prime}\theta^{(j)}\Big)},$$

The number J of random draws determines the accuracy of the Monte Carlo integration of the latent factor.[14] In our simulation study, we compare results obtained from Monte Carlo integration with $J = 100$, $J = 1,000$ and $J = 10,000$ draws.

4.3 Simulation results

The three approaches described above are implemented to compute the distributional treatment effects. A posterior predictive approach is adopted and each of these treatment parameters is evaluated at the sample means $\overline{z}$ and $\overline{x}$ of the covariates Z and X of the future individual, respectively. Results are displayed in Tables 1 and 2, where the means of the distributional treatment effects across the Monte Carlo replications, as well as the standard errors (SE), the biases and the root mean squared errors (RMSE) allow to compare the different estimators. Overall, the three approaches perform very well in estimating the treatment effects, and several interesting facts emerge.

The quality of the results from the approaches based on the integration of the latent factor (MC2 and MC3) depends on the method used for the numerical integration. Integrating the latent factor simultaneously with model parameters (MC2) happens to perform as well as MC1, and better than the sequential integration carried out in MC3 if the number of draws J is not large enough for this latter one. Indeed, 100 draws from the distribution of θ provide the least accurate approximation, and this number has to be increased up to

[14] The same J draws are used across the M iterations to damp numerical instabilities generated by the integration of the latent factor (Lee, 1992). Hence the notation $\theta^{(j)}$ where the superscript m has been dropped.

Table 1: Monte Carlo experiment results — $N = 500$

	True	Mean	SE	Bias	RMSE
MC1: Unconditional approach					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7655	(0.0172)	-0.0011	0.0172
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8529	(0.0163)	0.0005	0.0163
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5102	(0.0324)	0.0022	0.0325
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7929	(0.0183)	-0.0001	0.0183
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9457	(0.0104)	-0.0010	0.0104
MC2: Conditional approach					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7683	(0.0172)	0.0017	0.0173
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8532	(0.0162)	0.0008	0.0162
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5142	(0.0322)	0.0062	0.0328
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7949	(0.0182)	0.0019	0.0183
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9456	(0.0103)	-0.0011	0.0104
MC3: Conditional approach with $J = 100$					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7882	(0.0206)	0.0216	0.0299
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8609	(0.0175)	0.0085	0.0194
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5435	(0.0335)	0.0355	0.0488
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.8100	(0.0195)	0.0170	0.0259
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9437	(0.0108)	-0.0030	0.0112
MC3: Conditional approach with $J = 1,000$					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7634	(0.0182)	-0.0032	0.0185
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8487	(0.0168)	-0.0037	0.0172
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5138	(0.0323)	0.0059	0.0328
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7930	(0.0187)	0.0000	0.0187
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9416	(0.0110)	-0.0052	0.0122
MC3 Conditional approach with $J = 10,000$					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7683	(0.0172)	0.0017	0.0173
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8532	(0.0162)	0.0009	0.0162
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5142	(0.0322)	0.0063	0.0328
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7949	(0.0182)	0.0019	0.0183
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9456	(0.0103)	-0.0011	0.0104

Notes: 1,000 Monte Carlo replications. SE = MCMC standard error, RMSE = MCMC root mean squared error. MC2 = factors integrated out numerically, along with model parameters; MC3 = numerical integration of the factors embedded in the integration of model parameters.

Table 2: Monte Carlo experiment results — $N = 5,000$

	True	Mean	SE	Bias	RMSE
MC1: Unconditional approach					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7664	(0.0055)	-0.0002	0.0055
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8530	(0.0052)	0.0006	0.0052
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5077	(0.0104)	-0.0003	0.0104
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7930	(0.0058)	-0.0000	0.0058
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9467	(0.0031)	0.0000	0.0031
MC2: Conditional approach					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7672	(0.0061)	0.0007	0.0061
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8524	(0.0055)	0.0000	0.0055
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5101	(0.0105)	0.0021	0.0107
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7938	(0.0061)	0.0008	0.0061
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9461	(0.0031)	-0.0006	0.0032
MC3: Conditional approach with $J = 100$					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7751	(0.0191)	0.0086	0.0209
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8519	(0.0134)	-0.0004	0.0134
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5302	(0.0183)	0.0223	0.0289
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.8003	(0.0142)	0.0073	0.0159
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9407	(0.0061)	-0.0060	0.0085
MC3: Conditional approach with $J = 1,000$					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7706	(0.0099)	0.0040	0.0107
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8524	(0.0068)	0.0001	0.0068
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5193	(0.0147)	0.0114	0.0186
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7963	(0.0072)	0.0033	0.0079
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9442	(0.0038)	-0.0025	0.0046
MC3: Conditional approach with $J = 10,000$					
$\Pr(\Delta > 0 \mid X)$	0.7666	0.7672	(0.0061)	0.0007	0.0061
$\Pr(\Delta > 0 \mid D = 1, Z, X)$	0.8524	0.8524	(0.0055)	0.0000	0.0055
$\Pr(\Delta > 0 \mid U_D = -2, X)$	0.5079	0.5101	(0.0105)	0.0021	0.0107
$\Pr(\Delta > 0 \mid U_D = 0, X)$	0.7930	0.7938	(0.0061)	0.0008	0.0061
$\Pr(\Delta > 0 \mid U_D = 2, X)$	0.9467	0.9461	(0.0031)	-0.0006	0.0032

Notes: 1,000 Monte Carlo replications. SE = MCMC standard error, RMSE = MCMC root mean squared error. MC2 = factors integrated out numerically, along with model parameters; MC3 = numerical integration of the factors embedded in the integration of model parameters.

$10,000$ draws to reach the precision of MC2.[15] There is, however, a trade-off between the level of accuracy and computational burden in this case, and in this respect MC2 clearly appears more attractive than MC3.

To summarize, estimators based on the compact forms of the treatment effects, where the latent factors are integrated out analytically, should be favored in practice, whenever possible. They are simple to calculate. If the estimators based on the conditional model with the latent factors are preferred, then attention should be dedicated to the details of numerical integration. Our example shows that nesting the integration of latent factors and model parameters is only worthwhile with a large number of random draws for the Monte Carlo integration. These conclusions might however not be extended to more complicated models with multiple latent factors and/or more complicated functional forms than the simple continuous case presented here. Further research is therefore required to investigate the performances of the different estimators in these more general cases.

5 Conclusion

This paper discusses Bayesian approaches that address a fundamental identification problem arising in the analysis of treatment effects: we do not simultaneously observe the same person in both treated and untreated states. Our procedure breaks the identification problem by using the information in the choice equation supplemented with auxiliary measurements joined with an assumption that a low-dimensional vector of unobservables generates the dependence among the unobservables in the outcome and treatment choice equations.[16] It shows how a variety of treatment parameters can be alternatively computed in the framework of a potential outcomes factor structure model that generates identification of the covariance structure of the model. In this approach, a variety of classical and choice-theoretic mean treatment effects and their distributional versions can be derived. Depending on the needs of the analyst, different estimators can then easily be constructed to approximate these treatment effects. The standard tools of Bayesian inference apply to this model. (See, e.g., Conti et al., 2012.) Our small simulation study illustrates the methodology and provides encouraging results. It also reveals that the choice of the estimator among those suggested might not be neutral. More systematic simulation studies should therefore be carried out to better investigate the

[15] The results of MC2 and MC3 with $J = 10,000$ are virtually identical, to the 4-digit level of precision used to report the results in the table.

[16] As noted in Abbring and Heckman (2007), the "auxiliary measurements" can include vectors of the realized potential outcomes and need not be qualitatively different from the outcomes being studied. In principle, one can also dispense with the choice equation, although the resulting model is difficult to interpret.

behavior of these estimators in various contexts. Using the methods developed in Cunha et al. (2010), a more general nonparametric version of the model can be identified. Implementing this version is a task left for the future.

References

AAKVIK, A., J. J. HECKMAN, AND E. J. VYTLACIL (2005): "Estimating Treatment Effects for Discrete Outcomes when Responses to Treatment Vary: an Application to Norwegian Vocational Rehabilitation Programs," *Journal of Econometrics*, 125, 15–51.

ABBRING, J. H. (2011): "Mixed Hitting-Time Models," *Econometrica, forthcoming*.

ABBRING, J. H. AND J. J. HECKMAN (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6B, chap. 72, 5145–5303.

ANDERSON, T. AND H. RUBIN (1956): "Statistical Inference in Factor Analysis," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 5*, ed. by J. Neyman, Berkeley: University of California Press, 111–150.

CARNEIRO, P., K. T. HANSEN, AND J. J. HECKMAN (2001): "Removing the Veil of Ignorance in assessing the Distributional Impacts of Social Policies," *Swedish Economic Policy Review*, 8, 273–301.

——— (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44, 361–422.

CARNEIRO, P., J. J. HECKMAN, AND E. J. VYTLACIL (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394.

——— (2011): "Estimating Marginal and Average Returns to Education," *American Economic Review*, forthcoming.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2012): "Inference on Counterfactual Distributions," MIT Department of Economics Working Paper No. 08-16, Draft February 10, 2012.

CHIB, S. (2007): "Analysis of Treatment Response Data Without the Joint Distribution of Potential Outcomes," *Journal of Econometrics*, 140, 401–412.

COCHRANE, W. G. AND D. B. RUBIN (1973): "Controlling Bias in Observational Studies: A Review," *Sankhya The Indian Journal Of Statistics Series A*, 35, 417–446.

CONTI, G., S. FRÜHWIRTH-SCHNATTER, J. J. HECKMAN, H. F. LOPES, AND R. PIATEK (2012): "Constructing Economically Justified Aggregates: An Application to the Early Origins of Health," *working paper*.

CONTI, G. AND J. J. HECKMAN (2010): "Understanding the Early Origins of the Education-Health Gradient: A Framework That Can Also Be Applied to Analyze Gene-Environment Interactions," *Perspectives on Psychological Science*, 5, 585–605.

CUNHA, F., J. J. HECKMAN, AND S. NAVARRO (2004): "Separating uncertainty from heterogeneity in life cycle earnings," *Oxford Economic Papers*, 57, 191–261.

——— (2006): "Counterfactual Analysis of Inequality and Social Mobility," in *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, ed. by S. L. Morgan, D. B. Grusky, and G. S. Fields, Stanford University Press, chap. 11, 290–348.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica*, 78, 883–931.

FRÜHWIRTH-SCHNATTER, S. AND H. F. LOPES (2010): "Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown," *working paper*.

GAMERMAN, D. AND H. F. LOPES (2006): *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Boca Raton: Chapman and Hall/CRC, 2nd ed.

GEWEKE, J. F. AND G. ZHOU (1996): "Measuring the Pricing Error of the Arbitrage Pricing Theory," *Review of Financial Studies*, 9, 557–587.

GHOSH, J. AND D. B. DUNSON (2009): "Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis," *Journal Of Computational And Graphical Statistics*, 18, 306–320.

HANSEN, K. T., J. J. HECKMAN, AND K. J. MULLEN (2004): "The Effect of Schooling and Ability on Achivement Test Scores," *Journal of Econometrics*, 121, 39–89.

HECKMAN, J. J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694.

——— (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

——— (1992): "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs*, ed. by C. Manski and I. Garfinkel, Harvard University Press, chap. 5, 201–230.

——— (2008): "Econometric Causality," *International Statistical Review*, 76, 1–27.

HECKMAN, J. J. AND B. E. HONORÉ (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1149.

HECKMAN, J. J., J.-E. HUMPHRIES, S. URZUA, AND G. F. VERAMENDI (2010): "The Effects of Schooling on Labor Market and Health Outcomes," *unpublished manuscript*.

HECKMAN, J. J., H. ICHIMURA, J. SMITH, AND P. E. TODD (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

HECKMAN, J. J., L. MALOFEEVA, R. PINTO, AND P. A. SAVELYEV (2011): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," Unpublished manuscript, University of Chicago, Department of Economics (first draft, 2008). Under revision, *American Economic Review*.

HECKMAN, J. J., J. SMITH, AND N. CLEMENTS (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487.

HECKMAN, J. J. AND J. A. SMITH (1998): "Evaluating the Welfare State," in *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, ed. by S. Strom, New York: Cambridge University Press, 241–318.

HECKMAN, J. J., J. STIXRUD, AND S. URZUA (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24, 411–482.

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

——— (2000): "The Relationship Between Treatment Parameters Within a Latent Variable Framework," *Economics Letters*, 66, 33–39.

——— (2001): "Policy-Relevant Treatment Effects," *American Economic Review*, 91, 107–111.

——— (2007a): "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6B, chap. 70, 4779–4874.

——— (2007b): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6B, chap. 71, 4875–5143.

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G. W. AND J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

KOOP, G. AND D. J. POIRIER (1997): "Learning About the Across-Regime Correlation in Switching Regression Models," *Journal of Econometrics*, 78, 217–227.

LEE, L.-F. (1992): "On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models," *Econometric Theory*, 8, 518–552.

LI, M., D. J. POIRIER, AND J. L. TOBIAS (2004): "Do Dropouts Suffer from Dropping out? Estimation and Prediction of Outcome Gains in Generalized Selection Models," *Journal of Applied Econometrics*, 19, 203–225.

LI, M. AND J. L. TOBIAS (2008): "Bayesian Analysis of Treatment Effects in an Ordered Potential Outcomes Model," in *Modelling and Evaluating Treatment Effects in Econometrics (Advances in Econometrics, Volume 21)*, ed. by T. Fomby, R. C. Hill, D. L. Millimet, J. A. Smith, and E. Vytlacil, Emerald Group Publishing Limited, vol. 21 of *Modelling and Evaluating Treatment Effects in Econometrics (Advances in Econometrics)*, 57–91.

LIU, J. S. (1994): "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966.

LIU, J. S., W. H. WONG, AND A. KONG (1994): "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

LOPES, H. F. AND M. WEST (2004): "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, 14, 41–67.

POIRIER, D. J. (1998): "Revising Beliefs in Nonidentified Models," *Econometric Theory*, 14, 483–509.

POIRIER, D. J. AND J. L. TOBIAS (2003): "On the Predictive Distributions of Outcome Gains in the Presence of an Unidentified Parameter," *Journal of Business and Economic Statistics*, 21, 258–268.

QUANDT, R. E. (1958): "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *Journal of the American Statistical Association*, 53, 873–880.

ROSENBAUM, P. R. AND D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

ROY, A. D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.

THURSTONE, L. L. (1934): "The Vectors of Mind," *Psychological Review*, 41, 1–32.

TOBIAS, J. L. (2006): "Estimation, Learning and Parameters of Interest in a Multiple Outcome Selection Model," *Econometric Reviews*, 25, 1–40.

VIJVERBERG, W. P. (1993): "Measuring the Unidentified Parameter of the Extended Roy Model of Selectivity," *Journal of Econometrics*, 57, 69–89.