CEIS Tor Vergata

RESEARCH PAPER SERIES

Vol. 9, Issue 5, No. 194 – April 2011

Principal Stratification in sample selection problems with non normal error terms

Giovanni Mellace and Roberto Rocci

Principal Stratification in sample selection problems with non normal error terms

Giovanni Mellace[*,1] and Roberto Rocci[†,1]

[1] *University of Rome "Tor Vergata", via Columbia 2, 00133 Rome, Italy*

Abstract

The aim of the paper is to relax distributional assumptions on the error terms, often imposed in parametric sample selection models to estimate causal effects, when plausible exclusion restrictions are not available. Within the principal stratification framework, we approximate the true distribution of the error terms with a mixture of Gaussian. We propose an EM type algorithm for ML estimation. In a simulation study we show that our estimator has lower MSE than the ML and two-step Heckman estimators with any non normal distribution considered for the error terms. Finally we provide an application to the Job Corps training program.

Keywords: causal inference, principal stratification, mixture models, EM algorithm, sample selection.

JEL classification: C10, C13, C31, C34, C38.

We have benefited from comments by Martin Huber, Micheal Lechner, Franco Peracchi and seminar participants in St. Gallen and Tor Vergata.

[*]E-mail: giovanni.mellace@uniroma2.it, Tel: (+39) 06 7259 5625

[†]E-mail: roberto.rocci@uniroma2.it, Tel: (+39) 06 7259 5920

1 Introduction

In many fields where treatment effect evaluations are conducted, such as labor, health, and educational economics, the outcome of interest may be observed only for a non-randomly selected subpopulation. This problem is known in the literature as sample selection and may flaw causal analysis (see for instance Gronau, 1974 and Heckman, 1974). Indeed, even a randomized experiment cannot guarantee that treatment and control individuals will be comparable conditional on being selected.

For example, if we want to estimate the wage effect of a training program, often, only a selective subgroup of training participants and non-participants finds a job which is a condition for observing earnings. Similar problems are inherent in clinical trials when some of the participants in a medical treatment pass away ("truncation by death") before the health outcome is measured. As a final example, consider the effect of randomly provided private schooling on college entrance examinations. The sample selection problem arises when only a non-random subgroup of students takes the exam.

Principal stratification (PS hereafter, see Frangakis and Rubin, 2002), provides a natural framework to characterize sample selection problems, as it allows defining populations (i.e., principal strata) in terms of their behavior w.r.t. selection under different treatment states. This is useful because the selection problem does not arise within a particular stratum consisting of individuals with the same selection behavior, i.e., being of the same "type". Thus, the treatment effects identified are causal if the imposed assumptions and the data imply that individuals belonging to the same stratum may be observed both under treatment and non-treatment. Therefore, the principal stratification framework enables us to explicitly state under which assumptions identification works and for which latent population. Without strong and often unreasonable assumptions, when the treatment effects are heterogeneous, point identification is possible only for the always selected, i.e., those who are selected regardless of the treatment assignment. Indeed, only for this population it is possible to observe units in both the treatment arms (for partial

identification on different populations see Lechner and Melly, 2010 and Huber and Mellace, 2010).

In general, without the availability of a continuous instrument for the selection (see Das et al., 2003 and Huber, 2009), point identification can be achieved only by imposing strong distributional assumptions. In most of the applied works in the principal stratification framework, parametric point identification is achieved by means of finite mixture (see McLachlan and Peel, 2001) of Gaussian distributions. These models assume that the error terms are normally distributed among the strata. Similarly, in the econometric literature, it is well known that parametric sample selection models heavily rely on distributional assumptions. In particular, it is often assumed joint normality between the errors of the selection and the outcome equations. It is important to stress the fact that a failure in the joint normality assumption leads to inconsistent estimates, for this reason several semi-parametric and non-parametric estimators have been proposed in the literature (see Vella, 1998). However all this methods require additional exclusion restrictions assumptions for identification. Moreover, most of them are at least as restrictive as our, in the second step.

The idea of our paper is to relax distributional assumptions imposed on the error terms, when plausible exclusion restrictions are not available. Although, our model is still parametric, we are able to allow for heterogeneous treatment effects adding interactions between the treatment and the covariates. Moreover, it is often difficult or even impossible to find valid continuous instruments for the selection (variables that are relevant for the selection but not for the outcome). This is true, in particular, when the analysis regards the wage effect of a training program, as in our application.

It is well known that any distribution can be approximated by a mixture of normal distributions. Starting from this result, Bartolucci and Scaccia (2005), have shown how fitting a regression model with error terms distributed as a mixture of Gaussians, may improve OLS when the true distribution is not normal. Our idea is to extend this approach within the principal stratification framework. We will show that the particular structure of the problem allows us to identify a mixture of mixture model,

which is not identifiable in general. Maximum likelihood estimation is then carried out by means of an EM type algorithm (Dempster et al., 1977).

In order to study the performances of our estimator, we run a Monte Carlo simulation where data are generated from a standard Heckman sample selection model, and the performances of our estimator are compared with those of the Heckman maximum likelihood and two-step estimators. The results show that our approach performs better in terms of mean square error (MSE) when the true distribution of the error terms is not Gaussian (See Mealli and Pacini, 2008b for the Gaussian case).

Finally, we re-evaluate the long term wage effect of the Job Corps training program, using the same dataset analyzed in Lee (2009).

The paper is organized as follows. In section 2 we introduce the causal problem and we briefly compare the parametric Heckman sample selection and principal stratification models. In section 3 we present our idea and the main steps of the EM algorithm. In section 4 we show the simulation results. In section 5 we report the results of the application. Section 6 concludes.

2 Causal inference in presence of sample selection problems

Suppose we want to estimate the effect of a binary treatment $T = 1, 0$, on an outcome Y, at a specific time after assignment. Using the potential outcome framework advocated, among many others, by Rubin (1977), we will denote by $Y_i(1)$ and $Y_i(0)$, the two potential outcomes that an individual would receive under treatment and non-treatment, and by $\Delta_i = Y_i(1) - Y_i(0)$, the individual treatment effect. Even under randomization of the treatment, post-treatment complications might introduce selection bias and flaw causal inference. One particular form of post-treatment complications is sample selection, implying that the outcome of interest is only observed for a non-random subpopulation. To address this issue let $Q_i \in \{1, 0\}$, be an observed binary post-treatment selection indicator which is 1 if the outcome of individual i is observed and 0 otherwise and we denote by $Q_i(1)$ and $Q_i(0)$, the two potential selec-

tion states.

Throughout the discussion, the so-called Stable Unit Treatment Value Assumption (SUTVA, e.g., Rubin, 1990) will be maintained, ruling out interference between units as well as general equilibrium effects of the treatment.

Assumption 1 (SUTVA):

$Y_i(t_i) \perp t_j \ \forall j \neq i,$

$Q_i(t_i) \perp t_j \ \forall j \neq i.$

Where "$\perp$" denote independence. SUTVA implies that not only the potential outcomes but also the potential post-treatment variables for each subject i are unrelated to the treatment status of other individuals.

Causal inference requires the specification of the treatment assignment mechanism. If the treatment is randomly assigned it will be independent from the post-treatment variables Q, Y and from their potential values. However, in observational studies, randomization is assumed to hold conditional on the observed pre-treatment variables X. This assumption is known in the literature as conditional independence assumption (CIA), also referred to as "selection on observables" or "unconfoundedness", (see for instance Imbens, 2004 and Imbens and Wooldridge, 2009). It implies that the potential outcomes and selection states are independent of the treatment conditional on the pre-treatment variables.

In the sample selection framework, Lee (2009), Mealli and Pacini (2008a) and Mealli and Pacini (2008b), among others, assume that the joint distribution of the potential post treatment variables is independent of the treatment given X. This assumption can be formalized as:

Assumption 2 (Unconfoundness):

$(Y(1), Y(0), Q(1), Q(0)) \perp t | X = x \quad \forall x \in \mathcal{X}.$

Where $\mathcal{X}$ denotes the support of X.

A more formal representation of the causal problem is given by the following struc-

tural model (Huber, 2009 , Imbens, 2006 and Mealli and Pacini, 2008a)

$$\begin{aligned} y_i &= \chi(t_i, x_i, \epsilon_i) \\ q_i &= \varsigma(t_i, x_i, \nu_i) \\ t_i &= \psi(x_i, \zeta_i) \end{aligned} \tag{1}$$

where $\epsilon_i, \nu_i \perp \zeta_i | X = x$ by unconfoundness, thus the third equation can be ignored.

2.1 Heckman sample selection model and structural model

The standard parametric Heckman sample selection model imposes linearity in the first two equations of model (1), and can be written as

$$\begin{aligned} y_i^* &= \beta_0 + \beta_1 t_i + \beta_2^T x_i + \epsilon_i, \\ q_i &= I(\alpha_0 + \alpha_1 t_i + \alpha_2^T x_i + \nu_i > 0), \\ y_i &= y_i^* q_i, \end{aligned}$$

where $\begin{pmatrix} \epsilon_i \\ \nu_i \end{pmatrix} \sim \mathcal{N}_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon\nu} \\ \sigma_{\epsilon\nu} & \sigma_\nu^2 \end{pmatrix} \right]$ and $I(\cdot)$ is the indicator function. The idea is to adjust for the bias that arises from the correlation between the regressors of the outcome equation and its error term which operates through the relationship between ϵ_i and ν_i. This model can be estimated parametrically either by maximum likelihood or by the popular two-step estimator. The latter is based on the fact that, thanks to the joint normality of the error terms we can write

$$E(y_i | t_i, q_i = 1, x_i) = \beta_0 + \beta_1 t_i + \beta_2^T x_i + \beta_3 \lambda(\alpha_0 + \alpha_1 t_i + \alpha_2^T x_i),$$

where $\beta_1 = E(\Delta_i | t_i = 1, q_i = 1, x_i)$ is the average treatment effect (ATT) for the respondents(ATTR), $\lambda(\alpha_0 + \alpha_1 t_i + \alpha_2^T x_i)$ is the inverse Mills ratio and $\beta_3 = \sigma_{\epsilon\nu}/\sigma_\nu^2$. This also clarify that under the model assumptions, we are able to identify the ATTR.

This specification implicitly assume that no interactions between the treatment variable and the pretreatment characteristics are relevant, this would be plausible if the treatment effect is homogeneous, but it can lead to inconsistency when the effects are heterogeneous. One solution could be to add interactions between the treatment variable and the pre-treatment variables. In the simulation study we consider the case with no interactions between the treatment and the pre-treatment variables, then the comparison between the two models refer to this case. In section 3 we will discuss how our model can allow for heterogeneous effects.

As it has already been pointed out in the introduction, this model heavily rely on the joint normality of the error terms. Several estimators in the literature try to relax this assumption but additional exclusion restriction are needed (among many others: Ahn and Powell, 1993, Cosslett, 1991, Das et al., 2003, Ichimura and Lee, 1991, Ichimura, 1993, Lee, 1994, Li and Wooldridge, 2002, Newey, 2009, Newey, 1990, Newey et al., 1990, Powell et al., 1989 and Robinson, 1988).

2.2 Principal stratification

The Principal stratification approach suggests to stratify the units, within each cell defined by the values of the covariates, into four latent principal strata, according to the joint values of $(Q_i(1), Q_i(0))$. Frangakins and Rubin (2002) give the following definition

Definition

The basic principal stratification P_0 with respect to post-treatment variable Q is the partition of units $i = 1, \ldots, n$ such that, within any set of P_0, all units have the same vector $(Q_i(1), Q_i(0))$. In our case P_0 is given by

$$\begin{aligned} 11 &= \{i : Q_i(1) = Q_i(0) = 1\} \\ 10 &= \{i : Q_i(1) = 1,\ Q_i(0) = 0\} \\ 01 &= \{i : Q_i(1) = 0,\ Q_i(0) = 1\} \\ 00 &= \{i : Q_i(1) = Q_i(0) = 0\} \end{aligned}$$

Let $S_i \in \{11, 10, 01, 00\}$ represent the principal stratum to which subject i belongs. S_i is not affected by the treatment assignment by definition and can be seen as a covariate, only partially observed in the sample. Unconfoundedness guarantees to have the same distribution in both treatment arms, within cells defined by pre-treatment variables, it implies that $Y(0), Y(1) \perp t | Q(0), Q(1), X = x$, the potential outcomes are, therefore, independent of the treatment given the principal strata. Thus, any effect defined conditional on a principal stratum, is a well defined causal effect. In some sense, we can state that principal strata play a similar role of control functions in deriving independence conditions, even if not derived from a model (Mealli and Pacini, 2008a).

In our setting direct information on the causal effect can be found only in the 11 stratum of the always respondents, because only for units belonging to this stratum one can consistently compare $Y(1)$ and $Y(0)$. Since Q represents non response, in fact, only in this stratum we have both treated or control units, so that the causal effect which can be estimated, without any further restriction, is an effect within stratum 11.

The following correspondence between the observed values of T and Q and the latent strata holds

$o(1,1) = \{i : t_i = 1, Q_i(t_i) = 1\}$ subject i belongs either to 11 or to 10,

$o(1,0) = \{i : t_i = 1, Q_i(t_i) = 0\}$ subject i belongs either to 01 or to 00,

$o(0,1) = \{i : t_i = 0, Q_i(t_i) = 1\}$ subject i belongs either to 11 or to 01,

$o(0,0) = \{i : t_i = 0, Q_i(t_i) = 0\}$ subject i belongs either to 10 or to 00.

In each observed group we have a mixture of two principal strata, then, it is not possible to point-identify the strata proportions, as well as the distribution of Y within the strata.

To improve identification it is often assumed, that Q is monotone in T

Assumption 3 (monotonicity of selection):

$\Pr(Q(1) \geq Q(0)) = 1$.

This requires that the potential selection state never decreases in the treatment and, thus, rules out the existence of stratum 01[1].

In the standard sample selection model monotonicity is imposed by construction, since for example $\alpha_1 > 0$ implies $Q(1) \geq Q(0)$. When $\alpha_1 > 0$, the following correspondence between the specified selection model and the underlying latent strata holds (see Vytlacil, 2002)

$$\begin{aligned}
\{i : \nu_i > -\alpha_0 - \alpha_2^T x_i\} &\equiv \{i : Q_i(1) = Q_i(0) = 1\}, \\
\{i : -\alpha_0 - \alpha_1 - \alpha_2^T x_i < \nu_i < -\alpha_0 - \alpha_2^T x_i\} &\equiv \{i : Q_i(1) = 1,\ Q_i(0) = 0\}, \\
\{i : \nu_i < -\alpha_0 - \alpha_1 - \alpha_2^T x_i\} &\equiv \{i : Q_i(1) = Q_i(0) = 0\}.
\end{aligned}$$

Under assumption 3 we have

$o(1,1) = \{i : t_i = 1, Q_i(t_i) = 1\}$ subject i belongs either to 11 or to 10,

$o(1,0) = \{i : t_i = 1, Q_i(t_i) = 0\}$ subject i belongs to 00,

$o(0,1) = \{i : t_i = 0, Q_i(t_i) = 1\}$ subject i belongs to 11,

$o(0,0) = \{i : t_i = 0, Q_i(t_i) = 0\}$ subject i belongs either to 10 or to 00.

Monotonicity allows to point-identify at least the strata proportions, but again is not possible to disentangle the distribution of Y between strata 11 and 10. In this case only nonparametric bounds can be derived, unless some parametric distributional assumptions are introduced. Non parametric point identification, again rely on exclusion restrictions, more precisely on the availability of a valid instrument for the selection mechanism, as discussed for example in Mealli and Pacini (2008a).

Parametric identification is achieved by means of finite mixture models. Although others specification are possible, the proportions of units belonging to each stratum in the cell $X = x$ defined as $\pi_{11|x}$, $\pi_{10|x}$, and $\pi_{00|x} = 1 - \pi_{11|x} - \pi_{10|x}$ are often modeled

[1] A symmetric result can be obtained by assuming $\Pr(Q(0) \geq Q(1)) = 1$ which implies that stratum 10 does not exist. As Huber and Mellace (2010) have shown only one kind of monotonicity can be consistent with the data.

as a multinomial logit

$$\begin{aligned}
\pi_{11|x} &= \frac{\exp\{B_{11,0}+B_{11,1}^T x\}}{1+\exp\{B_{11,0}+B_{11,1}^T x\}+\exp\{B_{10,0}+B_{10,1}^T x\}}, \\
\pi_{10|x} &= \frac{\exp\{B_{10,0}+B_{10,1}^T x\}}{1+\exp\{B_{11,0}+B_{11,1}^T x\}+\exp\{B_{10,0}+B_{10,1}^T x\}}, \\
\pi_{00|x} &= 1-\pi_{11|x}-\pi_{10|x},
\end{aligned}$$

and the distributions of Y conditionally on the principal strata are assumed to be

$$\begin{aligned}
y_i|q_i = 1, x_i, 11 &\sim \mathcal{N}(\beta_0 + \beta_1 t_i + \beta_2^T x_i, \sigma_{11}^2), \\
y_i|q_i = 1, x_i, 10 &\sim \mathcal{N}(\delta_0 + \delta_1 t_i + \delta_2^T x_i, \sigma_{10}^2),
\end{aligned} \tag{2}$$

where we set $\delta = \delta_0 + \delta_1 t_i$ since t_i is always equal to 1 for this units.

In this specification $\beta_1 = E(\Delta_i|t_i = 1, x_i, 11) \equiv \Delta_{11}$ the ATT for the always respondents (ATTAR) [2]. In order to see this, from 2 we can write

$$E(y_i|t_i = 1, x_i, 11) = \beta_0 + \beta_1 t_i + \beta_2^T x_i$$

and

$$E(y_i|t_i = 0, x_i, 11) = \beta_0 + \beta_2^T x_i$$

Subtracting the two we have

$$\Delta_{11} = E(y_i|t_i = 1, x_i, 11) - E(y_i|t_i = 0, x_i, 11) = E(Y_i(1) - Y_i(0)|t_i = 1, x_i, 11) = \beta_1.$$

Notice that under the assumptions of the Heckman sample selection model ATTR=ATTAR. Indeed, even though it accounts for the correlation between ϵ_i and ν_i imposing joint normality, the effect is constant among the strata. This is a possible explanation of the relative bad performance of the ML Heckman estimator when data are generated under the principal stratification model found in Mealli and Pacini (2008b).

Finally, it is interesting to rephrase principal stratification in terms of the structural

[2] Same consideration as before for heterogeneous effects. For a fully saturated model with normal error terms see Zhang et al. (2009).

model described above. With no loss of generality and to avoid extra notation, suppose we are already within cells defined by observed pretreatment variables. From the structural model we have

$$\begin{aligned} y_i &= \chi(t_i, \epsilon_i) \\ q_i &= \varsigma(t_i, \nu_i) \end{aligned}$$

and y_i is observed only if $q_i = 1$, the endogeneity of Q depends on the relationship between ϵ and ν. Of course the endogeneity problem disappears whenever we are able to condition on ν, because in that case Q and ϵ will be conditionally independent. Obviously we will never observe ν, but we can find a function of it, say $S(\nu)$, called *type of unit* (Imbens, 2006) such that

$$\epsilon \perp Q | S(\nu).$$

As in the propensity score literature, the *type* function should have a small variation, i.e. it should be constant on sets of values of ν such that for all value of T lead to the same value of Q.

Principal stratification is the choarest choice of $S(\cdot)$, because if $S(\nu)$ represents the stratum S then $\epsilon \perp Q | S(\nu)$ by unconfoundness and

$$\begin{aligned} S(\nu) &= S(\nu') \qquad \text{if} \quad \varsigma(t,\nu) = \varsigma(t,\nu') \quad \forall\, t, \\ S(\nu) &\neq S(\nu') \qquad \text{if} \quad \varsigma(t,\nu) \neq \varsigma(t,\nu') \quad \text{for some } t. \end{aligned}$$

As an example in the Heckman sample selection model with $\alpha_1 > 0$, $S(\nu) \in \{I(\nu > -\alpha_0 - \alpha_2^T x) = 11, I(-\alpha_0 - \alpha_1 - \alpha_2^T x < \nu < -\alpha_0 - \alpha_2^T x) = 10, I(\nu < -\alpha_0 - \alpha_1 - \alpha_2^T x) = 00\}$, clearly satisfies the conditions above, because Q is constant and then independent from ϵ given $S(\nu) = 11, 00$, while $Q = T$ and then independent from ϵ by unconfoundness given $S(\nu) = 10$.

Within the PS approach, the comparison of y between treated and controls is possible only for some values of S, in particular those with $\varsigma(0, \nu) = 1$ and $\varsigma(1, \nu) = 1$ (Mealli and Pacini, 2008a, Zhang et al., 2008). This limitation, however, is created by the selection mechanism, and is not a drawback of principal stratification.

3 Our Proposal

In this section we present the main ideas underling our approach. In the first model we allow for heterogeneous effects w.r.t. to the observable covariates. Maintaining assumptions 1, 2 and 3, we assume that if the i-th individual belongs to stratum 11 its outcomes equation is $y_i = \beta_0 + \beta_1 t_i + \beta_2^T x_i + \beta_3^T x_i * t_i + \epsilon_i$, while if it belongs to stratum 10, then $y_i = \delta_0 + \delta_2^T x_i + \epsilon_i$. In the two equations we allow for effect heterogeneity by adding interactions between the treatment and the covariates.

The main difference of our model and others proposed in the literature is that we approximate the true distribution of ϵ_i as

$$f(\epsilon_i) = \sum_{g=1}^{G} \tau_g \phi(\epsilon_i; \mu_g, \sigma_k^2), \quad k = 11, 10,$$

where $\tau_g \geq 0,\ g = 1, \ldots, G$, $\sum_{g=1}^{G} \tau_g = 1$, $\sum_{g=1}^{G} \tau_g \mu_g = 0$[3] and $\phi(\epsilon_i, \mu, \sigma^2)$ denotes the density at ϵ_i of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. This implies that

$$f(y_i|t_i, x_i, 11) = \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \beta_0 + \beta_1 t_i + \beta_2^T x_i + \beta_3^T x_i * t_i, \sigma_{11}^2), \tag{3}$$

$$f(y_i|t_i, x_i, 10) = \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \delta_0 + \delta_2^T x_i, \sigma_{10}^2). \tag{4}$$

When $q_i = 1$ and $t_i = 0$, i.e. $i \in o(0, 1)$, then individual i belongs to stratum 11 and the distribution of y_i is given by 3, while when $q_i = 1$ and $t_i = 1$, i.e. $i \in o(1, 1)$, individual i belongs either to stratum 10 or to 11. In the latter case the distribution of y_i is given by the following mixture of mixtures

$$f(y_i|t_i = 1, q_i = 1, x_i) = \pi_{11|x_i} \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \beta_0 + \beta_1 t_i + \beta_2^T x_i + \beta_3^T x_i * t_i, \sigma_{11}^2) + \\ + \pi_{10|x_i} \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \delta_0 + \delta_2^T x_i, \sigma_{10}^2) \tag{5}$$

The proportions of units belonging to each stratum are still modeled as the multi-

[3] This condition is needed only if the intercepts are parameters of interest, indeed in this case we can identify, for example, β_0 as $(\beta_0 = [\sum_{g=1}^{G} \tau_g(\mu_g + \beta_0)])$.

nomial logit described before. In the simulation study we will use a more restrictive version of this model, which is obtained assuming that the treatment effect is homogeneous and that the distribution of the potential outcomes in stratum 11 stochastic dominate the distribution of the potential outcomes in stratum 10. As shown in Zang, Rubin and Mealli (2008) stochastic dominance implies that $\beta_2 = \delta_2$, $\beta_3 = 0$ and $\sigma^2_{11} = \sigma^2_{10}$. The only difference is that the distributions of the potential outcomes within the strata become

$$\begin{aligned} f(y_i|t_i, x_i, 11) &= \textstyle\sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \beta_0 + \beta_1 t_i + \beta_2^T x_i, \sigma^2_{11}), \\ f(y_i|t_i, x_i, 10) &= \textstyle\sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \delta_0 + \beta_2^T x_i, \sigma^2_{11}). \end{aligned}$$

3.1 Identification

Identifiability of finite mixture models has been proved for some important class of distributions such as gamma or multivariate Gaussian (see Teicher, 1963; Yakowitz and Spragins, 1968). Hennig (2000) provides sufficient conditions under which mixtures of Gaussian regression models are identified. As Henning pointed out, these conditions are rather mild if at least one regressor is continuous. In what follows, as in the standard PS model, we assume that the prior probabilities $\pi_{S_i|x_i}$ are identified.

Notice that $f(y_i|t_i = 1, q_i = 1, x_i)$ in 5 is a mixture of two mixtures of Gaussians whose parameters are not identifiable in general. Indeed, it may exist two different sets of parameters, say Θ and $\tilde{\Theta}$, such that

$$\begin{aligned} &\pi_{11|x_i} \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \beta_0 + \beta_1 + (\beta_2 + \beta_3)^T x_i, \sigma^2_{11}) + \pi_{10|x_i} \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \delta_0 + \delta_2^T x_i, \sigma^2_{10}) = \\ &= \tilde{\pi}_{11|x_i} \sum_{g=1}^{\tilde{G}} \tilde{\tau}_g \phi(y_i; \tilde{\mu}_g + \tilde{\beta}_0 + \tilde{\beta}_1 + (\tilde{\beta}_2 + \tilde{\beta}_3)^T x_i, \tilde{\sigma}^2_{11}) + \tilde{\pi}_{10|x_i} \sum_{g=1}^{\tilde{G}} \tilde{\tau}_g \phi(y_i; \tilde{\mu}_g + \tilde{\delta}_0 + \tilde{\delta}_2^T x_i, \tilde{\sigma}^2_{10}) \end{aligned} \tag{6}$$

for every y_i and x_i.

To see this notice that, equation 5 can be seen as a mixture of $2G$ normal linear

regressions and can be rewritten as

$$\sum_{h=1}^{2G} \xi_h \phi(y_i; \gamma_h^T x_i, \sigma_h^2) = \sum_{h=1}^{2\tilde{G}} \tilde{\xi}_h \phi(y_i; \tilde{\gamma}_h^T x_i, \tilde{\sigma}_h^2),$$

where

$$\begin{bmatrix} \xi_1 & \xi_{G+1} \\ \xi_2 & \xi_{G+2} \\ \vdots & \vdots \\ \xi_G & \xi_{2G} \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_G \end{bmatrix} \times \left[\pi_{11|x_1} \pi_{10|x_1}\right]$$

Identifiability of mixtures of Gaussians implies that $G = \tilde{G}$ and that for every h there exists a unique l such that

$$(\xi_h, \gamma_h^T, \sigma_h^2) = (\tilde{\xi}_l, \tilde{\gamma}_l^T, \tilde{\sigma}_l^2). \tag{7}$$

this does not guarantee identification of the other parameters. As an example, let us suppose that

$$\begin{bmatrix} \xi_1 & \xi_5 \\ \xi_2 & \xi_6 \\ \xi_3 & \xi_7 \\ \xi_4 & \xi_8 \end{bmatrix} = \frac{1}{60} \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \\ 6 & 12 \end{bmatrix} = \frac{1}{12} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 6 \end{bmatrix} \times \frac{1}{5} \left[1\ 2\right].$$

Those parameters can be also written as

$$\begin{bmatrix} \tilde{\xi}_1 & \tilde{\xi}_5 \\ \tilde{\xi}_2 & \tilde{\xi}_6 \\ \tilde{\xi}_3 & \tilde{\xi}_7 \\ \tilde{\xi}_4 & \tilde{\xi}_8 \end{bmatrix} = \frac{1}{60} \begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 2 & 6 \\ 4 & 12 \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \end{bmatrix} \times \frac{9}{60} \left[1\ 3\right].$$

In this example, 7 is true but our model is not identified if we just consider only the observed strata $o(1,1)$. However, we note that the probabilities τ_g can be identified in the observed strata $o(1,0)$. Since the prior probabilities $\pi_{S_i|x_i}$ are identified, we have

$\pi_{11|x_i} = \tilde{\pi}_{11|x_i}$ and $\pi_{10|x_i} = \tilde{\pi}_{10|x_i}$, while there exists a relabeling of the $\tilde{\tau}_g$ such that $\tau_g = \tilde{\tau}_g$. This guarantees identification.

Therefore the ATTAR denoted by Δ_{11} is identified as

$$\Delta_{11} = \int_{\mathcal{X}} E(Y|T=1,X,11)dF(X) - \int_{\mathcal{X}} E(Y|T=0,X,11)dF(X),$$

where $F(X)$ denotes the distribution of the covariates. Identification of the restricted model can be proved in a similar way, thus it will be skipped.

3.2 EM-algorithm

Without loss of generality, and to simplify the exposition, we discuss the estimation for the model without interaction under stochastic dominance.

The proportions of units belonging to each stratum are still modeled as the multinomial logit described above. Therefore the likelihood becomes

$$\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i\in o(1,1)} \left[\pi_{11|x_i} \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \breve{\omega}_i, \sigma^2) + \pi_{10|x_i} \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \breve{\eta}_i, \sigma^2)\right] \\
&\times \prod_{i\in o(1,0)} \pi_{00|x_i} \\
&\times \prod_{i\in o(0,1)} \pi_{11|x_i} \sum_{g=1}^{G} \tau_g \phi(y_i; \mu_g + \breve{\omega}_i, \sigma^2) \\
&\times \prod_{i\in o(0,0)} \left[\pi_{10|x_i} + \pi_{00|x_i}\right].
\end{aligned}$$

where $\theta = (B_{11,0}, B_{11,1}, B_{10,0}, B_{10,1}, \tau_1, \ldots, \tau_g, \mu_1, \ldots, \mu_G, \beta_0, \delta, \beta_1, \beta_2, \sigma^2)$, $\breve{\omega}_i = \beta_0 + \beta_1 t_i + \beta_2^T x_i$ and $\breve{\eta}_i = \delta_0 + \beta_2^T x_i$.

The maximum likelihood estimate of θ can be obtained iterating until convergence the EM type algorithm described below.

First of all the log-likelihood of the model can be written in compact notation as

$$\ell(\theta) = \sum_i ln \left(\sum_k \pi_{ik} \sum_g \tau_g \phi_{ikg}\right)$$

where $k = 00, 10, 11 = 1, 2, 3$ and $\phi_{ikg} = \begin{cases} 1 & \text{if} \quad k = 1 \\ q_i \phi(y_i; \mu_g + \breve{\eta}_i, \sigma^2) + (1 - q_i) & \text{if} \quad k = 2 \\ \phi(y_i; \mu_g + \breve{\omega}_i, \sigma^2) & \text{if} \quad k = 3 \end{cases}$.
Maximizing the log-likelihood is equivalent to maximize the "fuzzy" function (Hathaway, 1986)

$$\begin{aligned} \ell_f(\theta) &= \sum_{ikg} u_{ikg} ln(\pi_{ik} \tau_g \phi_{ikg}) - \sum_{ikg} u_{ikg} ln(u_{ikg}) \\ &= \sum_{ikg} u_{ikg} ln(\pi_{ik}) + \sum_{i:q_i=1} \sum_{k=2}^{3} \sum_{g} u_{ikg} ln(\tau_g) + \sum_{i:q_i=1} \sum_{k=2}^{3} \sum_{g} u_{ikg} ln(\phi_{ikg}) \\ &- \sum_{ikg} u_{ikg} ln(u_{ikg}) \end{aligned}$$

where $u_{ikg} \geq 0$ and $\sum_k \sum_g u_{ikg} = 1$.

The algorithm we adopt, in each step, maximizes the objective function ℓ_f with respect to a subset of parameters, given the current values of the others. In this way each parameter, or subset of parameters, is in turn updated increasing the value of the objective function at each iteration. The algorithm stops whenever the increment between two consecutive iterations is lower than a given threshold.

Before analyzing the fundamental steps of our algorithm, to simplify the exposition, we introduce some notation. First of all, we let n_{11}, n_{10}, n_{01} and n_{00} be the number of individuals in each observed subgroup defined by the value of t and q. Let y_{11} and y_{01} be the outcomes at $o(1, 1)$ and $o(0, 1)$, respectively. In the same way we define x_{11}, x_{01}, t_{11} and t_{01}. Let I_G be the identity matrix of dimension G, ι_G and ι be $G \times 1$ and $(n_{01} + 2n_{11}) \times 1$ vectors with all elements equal to 1. ι_j and O_j, are $n_j \times 1$ vectors $(j = 01, 11)$ with all elements equal to 1 and 0 respectively. Let

$$Y_0 = \begin{pmatrix} y_{01} \\ y_{11} \\ y_{11} \end{pmatrix} \qquad X_0 = \begin{pmatrix} O_{01} & t_{01} & x_{01} \\ \iota_{11} & O_{11} & x_{11} \\ O_{11} & t_{11} & x_{11} \end{pmatrix}$$

and

$$Y_G = \iota_G \otimes Y_0 \qquad X_G = \begin{pmatrix} I_G \otimes \iota & \iota_G \otimes X_0 \end{pmatrix}.$$

Let $u_{01,3,g},\ g = 1, \ldots, G$ and $u_{11,k,g},\ k = 2, 3,\ g = 1, \ldots, G$, be vectors which elements are the values of the u's at $o(1,1)$ and $o(0,1)$, respectively. Consider the following $(Gn_{01} + 2Gn_{11})$ vector

$$w = \begin{pmatrix} \sqrt{u_{01,3,1}} \\ \sqrt{u_{11,2,1}} \\ \sqrt{u_{11,3,1}} \\ \sqrt{u_{01,3,2}} \\ \sqrt{u_{11,2,2}} \\ \sqrt{u_{11,3,2}} \\ \vdots \\ \sqrt{u_{01,3,G}} \\ \sqrt{u_{11,2,G}} \\ \sqrt{u_{11,3,G}} \end{pmatrix}.$$

Finally let ι_{G+2+J} be a $(G + 2 + J)$ vector with all elements equal to 1, we define

$$\widetilde{Y} = w \odot Y_G \quad \text{and} \quad \widetilde{X} = (\iota_{G+2+J}^T \otimes w) \odot X_G$$

where $\odot$ is the element wise product.

The fundamental steps of our algorithm are

(a) *Update of* u_{ikg}: It can be easily shown that ℓ_f attains a maximum with respect to the u's when

$$u_{ikg} = \frac{\pi_{ik}\tau_g\phi_{ikg}}{\sum_{kg}\pi_{ik}\tau_g\phi_{ikg}}.$$

(b) *Update of* $B = B_{11,0}, B_{11,1}, B_{10,0}, B_{10,1}$: By rewriting

$$\ell_f = \sum_{ikg} u_{ikg} ln(\pi_{ik}) + const.,$$

where const. indicates a term that does not depend on the B's, they are updated

fitting a multinomial logit of the current update of the u's on the x's.

(c) *Update of* τ_g: By rewriting

$$\ell_f = \sum_{i:q_i=1} \sum_{k=2}^{3} \sum_{g} u_{ikg} ln(\tau_g) + const.,$$

where const. indicates a term that does not depend on the τ's, we will achieve a maximum when

$$\tau_g = \frac{\sum_{i:q_i=1} \sum_{k=2}^{3} u_{ikg}}{n_1}.$$

where n_1 is the number of subjects for which $q = 1$

(d) *Update of* $\beta = (\mu_1 + \beta_0, \ldots, \mu_g + \beta_0, \delta - \beta_0, \beta_1, \beta_2)^T$: By rewriting

$$\ell_f = -\frac{1}{2} \sum_{i:q_i=1} \sum_{k=2}^{3} \sum_{g} u_{ikg} \frac{(y_i - \gamma_g - \gamma_0 - \beta_2^T x_i)^2}{\sigma^2} + const.,$$

where $\gamma_g = \mu_g + \beta_0, \quad g = 1, \ldots, G$, $\gamma_0 = \begin{cases} \delta - \beta_0 & \text{if } k = 2 \\ \beta_1 t_i & \text{if } k = 3 \end{cases}$ and const. indicates a term that does not depend on the β's. Can be shown that the objective function is maximized at

$$\beta = (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \widetilde{Y}.$$

(e) *Update of* σ^2: By rewriting

$$\ell_f = -\frac{1}{2} \sum_{i:q_i=1} \sum_{k=2}^{3} \sum_{g} u_{ikg} \left(ln(\sigma^2) + \frac{(y_i - \gamma_g - \gamma_0 - \beta_2^T x_i)^2}{\sigma^2} \right) + const.,$$

where const. indicates a term that does not depend on the σ^2's, the maximum is achieved at

$$\sigma^2 = (w^T w)^{-1} (\widetilde{Y} - \widetilde{X}\beta)^T (\widetilde{Y} - \widetilde{X}\beta).$$

Keribin (1998) has shown that the Bayesian information criterion (BIC) give a consistent estimate of G, therefore, we choose the number of components, according to this criterion.

Finally notice that the algorithm above can be easily modified to relax the stochastic dominance assumption, by substituting X_0 with

$$X_0' = \begin{pmatrix} O_{01} & t_{01} & x_{01} & O_{01} \\ \iota_{11} & O_{11} & O_{11} & x_{11} \\ O_{11} & t_{11} & x_{11} & O_{11} \end{pmatrix}.$$

A modified EM-algorithm that allows for heterogeneous effects is available from the authors upon request.

3.2.1 Starting values

In maximum likelihood estimation, it is well known, that the log-likelihood may have several local maxima, thus, the choice of the starting values of the EM algorithm may be crucial. However, several strategies are available to overcome this problem. We propose the following.

The π's are initialized drawing at random from a Uniform distribution in $[0, 1]$ and rescaled such that $\sum_{k=1}^{3} \pi_{ik} = 1$. Consistent estimates of $\mu_1 + \beta_0, \ldots, \mu_g + \beta_0, \beta_2, \sigma^2$ and the τ's can be obtained estimating a mixture of G Gaussians in $o(0, 1)$. Finally for $\delta - \beta_0$ and β_1, a weighted regression of Y_0 on X_0 and a constant in the subsample in which $q = 1$, with weights the corresponding values of the π's, can be run or they can be estimated in the same sub-sample by OLS, or can be used the intercept and the coefficient of t of a two-step Heckman estimator.

Although, this seems to be reasonable, in some cases, starting from completely random points can leads to higher values of the log-likelihood. For this reason we suggest to try many random starting points, as well as the ones proposed above, and choose the solution corresponding to the highest value of the log-likelihood.

Many others approach are available in the literature, e.g. simulated annealing, none of them, however, seems to be optimal (See for example Ingrassia, 1991, 1992, Everitt, 1984, Davenport et al., 1988, Lindsay and Basak, 1993, and Aitkin and Aitkin, 1996).

4 Simulation results

We carry out a simulation study in which our approach is compared with Heckman's maximum likelihood and two-step sample selection estimators. We draw 500 samples of 1000 observations at each simulation step in which we simulate under the following Heckman sample selection model

$$\begin{aligned} y_i^* &= 2 + 5t_i + 10x_i + \epsilon_i \\ q_i &= I(-0.2 + 0.3t_i + 0.5x_i + \nu_i > 0) \\ y_i &= y_i^* q_i. \end{aligned}$$

In this model the treatment effect is constant and equal to 5. Let $e_i \sim \mathcal{N}(0,1)$ at each simulation the error terms are distributed such that $corr(\epsilon_i, \nu_i) \equiv \rho = 0.5$, then we have

- $\epsilon_i \sim 0.3\mathcal{N}(7,1) + 0.7\mathcal{N}(-3,1)$, a mixture of two normal distributions, and to induce correlation between the two error terms $\nu_i = 0.1231\epsilon_i + e_i$
- $\epsilon_i \sim \mathcal{T}(3)$, a Student's t with 3 degrees of freedom, and $\nu_i = 1/3\epsilon_i + e_i$
- $\epsilon_i \sim \mathcal{EV}(0,1) - \varphi$, ($\varphi \cong 0.57721$ is the Euler-Mascheroni constant) an extreme values whit location 0 and scale 1, and $\nu_i = 0.4502\epsilon_i + e_i$
- $\epsilon_i \sim \mathcal{GEV}(0,1,-.6) - \dfrac{\overbrace{\Gamma(1.6) - 1}^{0.1774}}{0.6}$, ($\Gamma(\cdot)$ is the Gamma function) a generalized extreme values distribution with location 0, scale 1 and shape -0.6, and $\nu_i = 0.6289\epsilon_i + e_i$
- $\epsilon_i \sim Log\mathcal{N}(0,1) - \exp(0.5)$, a lognormal distribution, and $\nu_i = 0.2671\epsilon_i + e_i$

In table 1 it is reported the MSE of the three estimators. For our estimator we reports the result for the optimal G [4], notice that for G=1 we have the standard principal stratification model.

[4] The G is chosen estimating our model in 5 simulated sample and then it is kept fixed for all the others. It might be the case that choosing the optimal G in each sample can improve the performance of our estimator.

[Insert Table 1 here]

According to the simulation result our estimator seems to have a lower MSE than the two Heckman estimators, for any distributions we consider.

5 An application to the Job Corps training program

In this section we present the results of an application to the Job Corps program, which is one of the largest job training programs in the U.S. and is aimed to help young people residents that belong to a low-income household from 16 to 24 years old. The program is described as "*the nation's largest career technical training and education program for young people at least 16 years of age. A voluntary program administered by the U.S. Department of Labor, Job Corps provides eligible young men and women with an opportunity to gain the experience they need to begin a career or advance to higher education*"[5]. In the mid nineties, In order to evaluate the effectiveness of the program, eligible applicants where randomly assigned to participate or rejected. In this section we will analyze the wage effect of the program 208 weeks after assignment.

5.1 The dataset and previous applications

The Job Corps public available data base has already been analyzed by Lee (2009) and Zhang et al. (2009). The former provides bound on the treatment effect under monotonicity, the latter uses a principal stratification mixture model approach to provide a point estimate.

The data sets used in the two papers differ essentially in the way in which missing values are treated, but as pointed out in Zhang et al. (2009) the imputation procedure used apparently do not affect the results.

Since the aim of this application is just to illustrate how our procedure can be effectively applied we will only focus on the imputed data base of Lee (2009). We will

[5] http://www.jobcorps.gov/faq.aspx

assume either perfect compliance to treatment assignment or that we just estimate an "intent to treat" effect.

Because of the random treatment assignment, it is not necessary to include covariates in our analysis, however, has argued in Zhang et al. (2009), including covariates may improve efficiency.

From November 1994 to December 1995, the 80,883 individuals who were eligible were randomly assigned either to enroll as usual ("treatment group") or they were embargoed from the program for 3 years ("control group"). The control group consisted of 5,997 individuals, from the remaining treated 9,409 applicants were randomly selected to be followed for the data collection, thus the total sample was of 15,386 individuals. In the dataset of Lee (2009), which is the one that we use, all the missing values due to non-response as well as to attrition are discarded then the final sample size is of only 9,145 individuals. The sample selection problem considered here is just due to unemployment. Since some subpopulation were randomized into the program group with known probabilities, design weights denoted by *we* must be included in the analysis. Finally, since we will use exactly the same variables as Lee (2009), an exhaustive description of the data can be found in that paper.

5.2 Estimation results

In this section we estimate the long run effect of the program on the logarithm of hourly wages 208 week after the treatment assignment. Since the treatment were assigned at random unconfoundness hold by sample design. The first assumption that we make is that there are no individuals that would have been employed if non treated and unemployed if treated (monotonicity). This assumption would be violated if there are people that four years after the end of the program, were still waiting for a better offer because of participation. However, we can reasonable argue that four years is a sufficient amount of time to prevent this possibility. We assume that the probability to belong to a given stratum depend only on the baseline characteristics, and we model it as a multinomial logit. We assume that the error terms are distributed as a mixture of Gaussians in both stratum 10 and 11. We let the coefficients of the covariates vary

between the two strata (no stochastic dominance). In order to compare our results with the ones in Lee (2009) we assume that the treatment effect is constant and we do not include any interaction. Notice that because of randomization covariates should not matter, then this simplifying assumption should hold.

The table belove report the results of our point estimate of the wage effect as well as the results in Lee (2009). Moreover we estimate the effect of the program on unemployment as the share of people belonging to stratum 10 given by

$$\pi_{10} = \frac{\sum_{i=1}^{n}(\pi_{10|x_i} we_i)}{\sum_{i=1}^{n} we_i}.$$

[Insert Table 2 here]

Before commenting on our finding it is important to stress that the program can be seen as a human capital investment of 1 year of schooling. For this reason, as Lee (2009) has pointed out, if the program had literally no effect we would find a strong negative impact of the program even 4 years after the assignment. Lee suggests that the lost labour market experience for these young applicants that are on the steep part of their wage profile should be around -0.58.

As reported in table 2 (PS G=7), differently from the standard principal stratification mixture model, which under monotonicity estimates a strong positive effect, we cannot reject the hypothesis of a zero effect of the program on the log hourly wages 208 week after assignment (notice that our point estimate lies within the bounds derived in Lee, 2009). However, this result must be interpret, in the light of the considerations made above, as a positive effect of the program. In particular, it can be argued that four years after assignment applicants are able to offset 100% of the lost labour market experience due to participation to the program.

Finally, according to our result 3.37% of the individuals have found a job because of the program.

6 Conclusions

In this paper we propose a new principal stratification approach to identify and estimate causal effects in the presence of sample selection when the error terms are not normally distributed and any plausible instrument for the selection mechanism is at hand. Our approach is based on the fact that any distribution can be approximated with a mixture of Gaussians, a known result in the mixture models literature. Given the particular structure of the problem we are able to identify a mixture of mixtures model which allows us to take into account the non-normality of the error terms.

Even if our model is fully parametric we show how to allow for heterogeneous effects, interacting the treatment variable with all the covariates available.

Under basically the same assumptions as the standard parametric principal stratification models that have been proposed in the literature we are able to identify the average treatment effect on the latent stratum of subject who always respond. This, however, seems a limitation created by the selection mechanism, rather than a drawback of the model.

Simulating under a standard sample selection model, we show that our estimator has always lower MSE than the two Heckman's estimators considered, with all the four different non normal distributions assumed for the error terms.

Finally we present an application to the long run wage effect of the Job Corps training program. The results are consistent with the bounds derived by Lee (2009).

References

Ahn H, Powell JL. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* **58**: 3–29.

Aitkin M, Aitkin I. 1996. A hybrid em/gauss-newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing* **6**: 127–130.

Bartolucci F, Scaccia L. 2005. The use of mixtures for dealing with non-normal regression errors. *Computational Statistics and Data Analysis* **48**: 821–834.

Cosslett S. 1991. Semiparametric estimation of a regression model with sample selectivity. In Barnett WA, Powell J, Tauchen G (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge university press.

Das M, Newey WK, Vella F. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* **70**: 33–58.

Davenport W J, Pierce A M, Hathaway J R. 1988. A numerical comparison of em and quasi-newton type algorithms for computing mle's for a mixture of normal distributions. In *Computer Science and Statistics: Proceedings of the 20th Symposium on the Interface*. American Statistical Association, 410–415.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B* **39**: 1–38.

Everitt BS. 1984. Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithms. *Journal of the Royal Statistical Society. Series D* **33**: 205–215.

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

Gronau R. 1974. Wage comparisons-a selectivity bias. *Journal of Political Economy* **82**: 1119–1143.

Hathaway J R. 1986. Another interpretation of the em algorithm for mixture distributions. *Statistics and Probability Letters* **4**: 53 – 56.

Heckman JJ. 1974. Shadow prices, market wages and labor supply. *Econometrica* **42**: 679–694.

Hennig C. 2000. Identifiablity of models for clusterwise linear regression. *Journal of Classification* **17**: 273–296.

Huber M. 2009. Treatment evaluation in the presence of sample selection. *Discussion paper 09-07* Department of Economics, University of St. Gallen `http://www.alexandria.unisg.ch/export/DL/69710.pdf`.

Huber M, Mellace G. 2010. Sharp bounds on average treatment effects under sample selection. *mimeo* University of St. Gallen `http://www.alexandria.unisg.ch/export/DL/70308.pdf`.

Ichimura H. 1993. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics* **58**: 71–120.

Ichimura H, Lee L. 1991. Semiparametric least squares of multiple index models: Single equation estimation. In Barnett WA, Powell J, Tauchen G (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge university press.

Imbens GW. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* **86**: 4–29.

Imbens GW, Wooldridge JM. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**: 5–86.

Imbens W G. 2006. Nonadditive models with endogenouus regressors. *mimeo* University of Chicago `http://ws1.ad.economics.harvard.edu/faculty/imbens/files/wc_06feb28.pdf`.

Ingrassia S. 1991. Mixture decomposition via the simulated annealing algorithm. *Applied Stochastic Models and Data Analysis* **7**: 317–325.

Ingrassia S. 1992. A comparison between the simulated annealing and the em algorithms in normal mixture decompositions. *Statistics and Computing* **2**: 203–211.

Keribin C. 1998. Consistent estimate of the order of mixture models. *Sankya: The Indian journal of statistics* **62**: 49–66.

Lechner M, Melly B. 2010. Partial identification of wage effects of training programs. *Brown University Economics Working Paper 2010-8* Brown University, Department of Economics, `http://www.brown.edu/Departments/Economics/Papers/2010/2010-8_paper.pdf`.

Lee DS. 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies* **76**: 1071–1102.

Lee L. 1994. Semiparametric instrumental variable estimation of simultaneous equation sample selection models. *Journal of Econometrics* **63**: 341 – 388.

Li Q, Wooldridge JM. 2002. Semiparametric estimation of partially linear models for dependent data with generated regressors. *Econometric Theory* **18**: 625–645.

Lindsay BG, Basak P. 1993. Multivariate normal mixtures: A fast consistent method of moments. *Journal of the American Statistical Association* **88**: 468–476.

McLachlan G, Peel D. 2001. *Finite Mixture Models*. Wiley series in probability and statistics.

Mealli F, Pacini B. 2008a. Causal inference with nonignorably missing outcomes: instrumental variables and principal stratification. *mimeo* University of Florence, `http://www.ds.unifi.it/mealli/pubblicazioni/mealli_pacini_IV_short_11_11_08.pdf`.

Mealli F, Pacini B. 2008b. Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. *Computational Statistics and Data Analysis* **53**: 507–516.

Newey W. 1990. Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**: 99–135.

Newey WK. 2009. Two-step series estimation of sample selection models. *Econometrics Journal* **12**: 217–229.

Newey WK, Powell JL, Walker J. 1990. Semiparametric estimation of selection models: Some empirical results. *American Economic Review* **80**: 324–328.

Powell JL, Stock J, Stoker T. 1989. Semiparametric estimation of index coefficients. *Econometrica* **57**: 1403–1430.

Robinson P. 1988. Root-n consistent semiparametric regression. *Econometrica* **56**: 931–954.

Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* **2**: 1–26.

Rubin DB. 1990. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* **25**: 279–292.

Teicher H. 1963. Identifiability of finite mixtures. *The Annals of Mathematical Statistics* **34**: 1265–1269.

Vella F. 1998. Estimating models with sample selection bias: A survey. *The Journal of Human Resources* **33**: 127–169.

Vytlacil E. 2002. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica* **70**: 331–341.

Yakowitz SJ, Spragins JD. 1968. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics* **39**: 209–214.

Zhang J, Rubin DB, Mealli F. 2008. Evaluating the effects of job training programs on wages through principal stratification. In Millimet D, Smith J, Vytlacil E (eds.) *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics*, volume 21. Elsevier Science Ltd., 117–145.

Zhang JL, Rubin DB, Mealli F. 2009. Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association* **104**: 166–176.

Table 1: MSE and bias of the estimates of β_1.

	PS G=2	PS G=1	ML	Two-step
Mixture				
MSE β_1	**0.0105**	3.7025	1.0115	0.3989
Bias	(-0.0113)	(-1.2597)	(-0.5913)	(0.0350)
Student's t				
MSE β_1	**0.0201**	0.0266	0.0434	0.0379
Bias	(0.0011)	(0.0355)	(-0.0016)	(-0.0009)
Extreme values				
MSE β_1	**0.0145**	0.0490	0.0436	0.0174
Bias	(0.0173)	(0.1909)	(0.1157)	(0.0042)
GEV				
MSE β_1	**0.0065**	0.0360	0.0285	0.0091
Bias	(-0.0149)	(0.1708)	(0.0863)	(-0.0003)
Lognormal				
MSE β_1	**0.0093**	0.0636	0.4111	0.0500
Bias	(-0.0120)	(-0.0918)	(-0.1499)	(-0.0005)

Note: The optimal G has been chosen according to the BIC. The lowest MSE is reported in bold.

Table 2: Estimation results

Method	Parameters	Estimates
Lee(2009)	Bounds	(-0.019,0.093)
	S.e.	[0.0179, 0.0130]
	Worst c. i.	{-.055,0.119}
Heckman two step	ATE	0.0148
	S.e.	0.0117
Das et al. (2003)	ATE	0.0140
	S.e.	0.0122
PS (G=1)	ATE	0.0503
	LR-test p-value	0.0005
	π_{10}	0.0686
PS (G=7)	ATE	0.0065
	LR-test p-value	0.2334
	π_{10}	0.0337

Note:the optimal G has been chosen according to the BIC.