Università degli Studi di Pisa
Dipartimento di Statistica e Matematica
Applicata all'Economia

Report n. 292

Identifiability and two-steps estimation procedures in casual models with ignorable assignments and non-ignorable compliance

Andrea Mercatanti

Pisa, dicembre 2006
- Stampato in Proprio –

Identifiability and two-steps estimation procedures in causal models with ignorable assignments and non-ignorable compliance

Andrea Mercatanti*
Viale Montegrappa 81, 59100 Prato
Phone: +39-335-5495405 e-mail: mercatan@libero.it

December 2006

Abstract

This paper examines the problem of relaxing the exclusion restriction for the evaluation of causal effects in experiments with non-ignorable compliance to the assignments. The exclusion restriction is a relevant assumption for identifying causal effects by the nonparametric instrumental variables technique, for which the template of a randomized experiment with imperfect compliance can represent a natural parametric extension. However the full relaxation of the exclusion restriction yields likelihood functions characterized by the presence of mixtures of distributions. This complicates a likelihood-based analysis because it implies only partially identified models and more than one maximum likelihood point. We consider the identifiability when the outcome distributions of various compliance statuses are in the same class. Two-steps estimation procedures when the outcomes are normally distributed are also proposed. In these cases we do not need to impose any extra assumptions compared to those usually adopted for the instrumental variables technique. An economic application concerning return to schooling concludes the paper.

*The author thanks the Dep. of Statistics and Applied Mathematics of the University of Pisa, where he spent more than four years in research and teaching activity, for allowing the publication of this working paper.

1 Introduction

The exclusion restriction is crucial in the identification of treatment effects in various causal inference methods. Historically, the assumption appeared in the literature concerning the Instrumental Variables (IV henceforth) method which has a long tradition in econometrics, and that has been applied in the context of causal evaluation, for example, by Heckmann and Robb (1985), Angrist (1990), Angrist and Krueger (1991), Kane and Rouse (1993), Card (1995), and more recently by Ichino and Winter-Ebmer (2004). In particular, Angrist et al. (1996) showed that, under a suitable set of assumptions including the exclusion restriction, the nonparametric IV method can identify causal treatment effects for compliers, the individuals who would receive the treatment only if assigned to it. Under a general approach to causal inference, labeled the Rubin Causal Model by Holland (1986), the exclusion restriction requires that the instrumental variable has not a direct causal effect on the outcome. In terms of a linear regression model this is equivalent to imposing the absence of a probabilistic link between the instrumental variable and the error term.

The connection between a randomized experiment with imperfect compliance and the IV model is in the fact that the former is a template that can be adopted for the identification and estimation of treatment causal effects also in nonexperimental situations. Regarding the IV model, the template is that of a randomized experiment with imperfect compliance in the sense that the particular instrumental variable adopted should have the role of a random assignment for which the treatment does not necessarily comply.

Nonparametric bounds, on the average treatment effects of a randomized experiment with imperfect compliance, over the whole population have been developed by Balke and Pearl (1997) under the exclusion restriction, and supposing a binary treatment and a binary outcome. Their paper was based on the general result of Manski (1990) for nonparametric bounds on treatment effects.

Subsequently, research in causal inference turned from the nonparametric instrumental variables method to parametric models. In particular with the contribution of Imbens and Rubin (1997a) who introduced a suitable likelihood function, and proposed also a weak version of the exclusion restriction requiring that the assignment to treatment has to be unrelated to potential outcomes but only for noncompliers, the individuals that would receive or would not receive the treatment regardless of whether it is offered.

In spite of its importance, the exclusion restriction can often be unrealistic in practice; however relaxing the assumption is not straightforward since it is directly related to the identifiability of the parametric models. A example on a real data set (Imbens and Rubin, 1997a) shows that, without the exclusion restriction and with a binary outcome, the model does not have a unique maximum likelihood point, but rather a region of values at which the likelihood function is maximized. Given this precedent, other studies propose relaxing the assumption by relying on prior distributions in a Bayesian framework and with a binary outcome (Hirano et al., 2000), or by introducing auxiliary information from pretreatment variables under normally distributed outcomes (Jo, 2002).

The current study explores a new option, where in a likelihood-based context we fully relax the exclusion restriction without introducing extra information compared to the usual set of conditions adopted to identify causal effect in the IV framework (Angrist et al., 1996). Supposing a binary treatment and outcome distributions of various compliance statuses in the same class, we show that relaxing the exclusion restriction introduce two mixtures of distributions in the parametric model. Some of the usual difficulties in identifying and estimating mixed distribution models, such as the switching of mixture component indicators, the presence of several local maximum likelihood points and the singularities of the likelihood function (McLachlan and Peel, 2000), complicate our likelihood-based analysis.

This article is briefly organized as follows. Section 2 fixes the conditions for the identifiability of the model when the outcome distributions of various compliance statuses are in the same class. In this context the study of identifiability is driven by the need to attain a right labelling of the mixture components. In Section 3 we propose two-steps estimation procedures when the outcomes are normally distributed. The proposed procedures are based on identifying the efficient likelihood estimate as the solution of the likelihood equations closest to a consistent, but not efficient, estimate of the parameters vector or of a suitable parameters sub-vector. Their relative merits will be investigated by simulation studies in Section 4. Section 5 concludes the paper by proposing an application based on a microeconomic data set; this is suggested by a recent paper of Ichino and Winter-Ebmer (2004) who investigated the long run educational cost of World War II. The results obtained by applying the proposed procedure are compared to those obtained by the IV method.

2 Identifiability

A remarkable contribution to the parametric formalization of the IV technique in identifying and estimating the causal effects is due to Imbens and Rubin (1997a). The authors based the resulting distribution function on the concept of potential quantities: the concept of causality we want to adopt in this paper. Consequently, the population under study can be subdivided in four groups that are characterized by the way the individuals react, from a counterfactual point of view, to the assignment to treatment. These groups are labeled compliance statuses. To clarify, assume the simplest experimental setting where there is only one outcome measure (Y_i), and where the assignment to treatment (Z_i) and the treatment received (D_i) are binary ($Z_i = 1$ =assigned, $Z_i = 0$ =not assigned; $D_i = 1$ =received, $D_i = 0$ =not received). In settings of imperfect compliance with respect to an assigned binary treatment, and on the basis of the concept of potential quantities, the whole population can be subdivided into four subgroups to characterize different compliance behaviors. Units for which $Z_i = 1$ implies $D_i = 1$ and $Z_i = 0$ implies $D_i = 0$ (*compliers*) are induced to take the treatment by the assignment. Units for which $Z_i = 1$ implies $D_i = 0$ and $Z_i = 0$ implies $D_i = 0$ are called *never-takers* because they never take the treatment, while units for which $Z_i = 1$ implies $D_i = 1$ and $Z_i = 0$ implies $D_i = 1$ are called *always-takers* because they always take the treatment. Finally the units for which $Z_i = 1$ implies $D_i = 0$ and $Z_i = 0$ implies $D_i = 1$ do exactly the opposite of the assignment and are called *defiers*. Each of these four groups define a particular *compliance status*.

Let $Y_i(Z_i = z, D_i = d)$ with $z \in \{0, 1\}$ and $d \in \{0, 1\}$ be the potential outcome with respect to the assignment, z, and to the treatment, d. The exclusion restriction implies that $Y_i(Z_i = 1, D_i = d) = Y_i(Z_i = 0, D_i = d)$. In order to achieve a complete relaxation of the assumption, the current study employs a likelihood estimation approach which is known to be often more efficient than the IV framework in the identification and estimation of causal effects for compliers (Imbens and Rubin, 1997a; Little and Yau, 1998; Jo, 2002). At these purposes let introduce this set of assumptions:

Assumption 1 : *S.U.T.V.A. (Stable Unit Treatment Value Assumption)* by which the potential quantities for each unit are unrelated to the treatment status of other units;

Assumption 2 : "*Random assignment to treatment*" by which the proba-

bility to be assigned to the treatment is the same for every unit;

Assumption 3 : *Nonzero average causal effect of* Z_i *on* D_i, imposing the presence of compliers;

Assumption 4 : *"Monotonicity"* imposing the absence of defiers;

Assumption 5 : the outcome distributions of various compliance statuses are in the same parametric class.

Assumptions 1-4 are the necessary set of conditions for identifying the compliers average treatment effect by the IV method, apart from the exclusion restriction (Angrist et al., 1996). The distribution function for a randomized experiment with imperfect compliance, a binary treatment, under the previous 1-5 assumptions, and adopting the parameter set proposed by Imbens and Rubin (1997a), is in the parametric class:

$$\mathcal{F}' = \{f(y_i,\, d_i,\, z_i;\boldsymbol{\theta}) = I_{\varsigma(D_i=1,\, Z_i=0)} \cdot (1-\pi) \cdot \omega_a \cdot g^i_{a0} + I_{\varsigma(D_i=0,\, Z_i=1)} \cdot \pi \cdot \omega_n \cdot g^i_{n1} +$$

$$+ I_{\varsigma(D_i=1,\, Z_i=1)} \cdot \pi \cdot (\omega_a \cdot g^i_{a1} + \omega_c \cdot g^i_{c1}) + I_{\varsigma(D_i=0,\, Z_i=0)} \cdot (1-\pi) \cdot (\omega_n \cdot g^i_{n0} + \omega_c \cdot g^i_{c0}) | \boldsymbol{\theta} \in \Theta\} \tag{1}$$

where

$$\Theta : \left\{ \boldsymbol{\theta} = (\pi, \omega_a, \omega_n, \omega_c, \boldsymbol{\eta}_{a0}, \boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{n0}, \boldsymbol{\eta}_{n1}, \boldsymbol{\eta}_{c0}, \boldsymbol{\eta}_{c1}) | \sum_{t=a,\, n,\, c} \omega_t = 1;\, \omega_t > 0\,, \forall t;\, 0 > \pi > 1 \right\} \tag{2}$$

and where: $I_{(\cdot)}$ is an indicator function; $\varsigma(D_i = d, Z_i = z)$ is the group of the units assuming treatment d and assigned to the treatment z; π is the probability $P(Z_i = 1)$; ω_t is the mixing probability, that is the probability of an individual being in the t group, $t = a$ *(always-takers)*, n *(never-takers)*, c *(compliers)*; the function $g^i_{tz} = g_{tz}(y_i; \boldsymbol{\eta}_{tz})$ is the outcome distribution for a unit in the t group and assigned to the treatment z.

Then (1) factors in four terms, where any term refers to a group $\varsigma(D_i = d, Z_i = z)$ of the units assuming treatment d and assigned to the treatment z. In particular the units in group $\varsigma(D_i = 0, Z_i = 0)$ are a mixture of compliers and never-takers, and the units in group $\varsigma(D_i = 1, Z_i = 1)$ are a mixture

of compliers and always-takers. Mixture models can present particular difficulties with identifiability; consequently the study of identifiability for the parametric class $\mathcal{F}'$, that involves two mixtures, is not straightforward. In order to explain the reasons of these difficulties, let's consider the general class of distribution functions from which the two mixtures are to be formed:

$$\mathcal{G} = \{g(y_i; \boldsymbol{\eta}) | \, \boldsymbol{\eta} \in \Upsilon, \, y_i \in R\}, \tag{3}$$

and the general class of distribution functions of two-components mixtures of (3):

$$\mathcal{F}'' = \left\{ f(y_i, \boldsymbol{\theta}) = \sum_{h=1}^{2} \omega_h \cdot g(y_i; \boldsymbol{\eta}_h) | \, g(\cdot; \boldsymbol{\eta}_h) \in \mathcal{G}, \, \forall h; \, y_i \in R; \, \boldsymbol{\theta} \in \Theta \right\}, \tag{4}$$

where

$$\Theta = \{\boldsymbol{\theta} = (\omega_1, \omega_2, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) | \, (\omega_1 + \omega_2) \leq 1, \, \omega_1 > 0, \, \omega_2 > 0; \, \boldsymbol{\eta} \in \Upsilon\}$$

In general a parametric family of densities $\mathcal{E} = \{e(y; \boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \Lambda, \, y \in R\}$ is identifiable if distinct members of the parameter space Λ always determine distinct members of the family:

$$e(y; \boldsymbol{\lambda}') \equiv e(y; \boldsymbol{\lambda}'') \;\Leftrightarrow\; \boldsymbol{\lambda}' = \boldsymbol{\lambda}''.$$

It is well known (Titterington et al., 1985; McLachlan and Peel, 2000) that (4) is not identifiable, since $f(y; \boldsymbol{\theta})$ is invariant under the two permutations of the component labels h in $\boldsymbol{\theta}$. Indeed, the presence of two densities in the same class, $g(y; \boldsymbol{\eta}_1)$ and $g(y; \boldsymbol{\eta}_2)$, implies that $f(y; \boldsymbol{\theta}) = f(y; \boldsymbol{\theta}^*)$ if the component labels 1 and 2 are interchanged in $\boldsymbol{\theta}^*$ compared to $\boldsymbol{\theta}$. Titterington et al. (1985) propose a weak definition of identifiability for finite mixtures of distribution in the same parametric class by which a class of mixtures is identifiable if distinct members of the parameter vector Θ always determine distinct members of the family up to permutations of the label components. Under their definition, (4) is identifiable if and only if $\mathcal{G}$ is a linearly independent set over the field of real number R. Relevant findings in the literature (for example Titterington et al., 1985; Teicher 1961, 1963; Yakowitz and Spragins 1968; Li and Sedransk 1985) shows that apart from special cases with very simple density function such as finite mixtures of uniformes, or

with finite sample spaces such as mixtures of two Bernoulli distributions, the identifiability up to the permutation of label components of (4) is generally assured.

However, and contrarily to an analysis of the mixture model $f(y_i, \boldsymbol{\theta}) \in \mathcal{F}''$ at cluster purposes, the components labelling matters for $f(y_i,\, d_i,\, z_i\,; \boldsymbol{\theta}) \in \mathcal{F}'$ at causal inference purposes. The causal effects from a counterfactual point of view are indeed defined by the three differences $\Delta_t = (\mu_{t1} - \mu_{t0})$, where $t = a, n, c$. Consequently, the right labelling of all the components now matters in order to identify Δ_t. For example, let's consider a point $\hat{\boldsymbol{\theta}}$, for which the component labels of the mixture $\varsigma(D_i = 1, Z_i = 1)$, composed by assigned always-takers and assigned compliers, permute compared to the true parameter vector $\boldsymbol{\theta}$. In this case the causal effects of the assignment to treatment for always-takers and compliers are not identified because of the permutation of component labels in $\hat{\boldsymbol{\theta}}$. Indeed, the causal effect for compliers Δ_c in $\hat{\boldsymbol{\theta}}$ would be wrongly identified by $(\mu_{a1} - \mu_{c0})$ instead of $(\mu_{c1} - \mu_{c0})$, and the causal effect for always-takers Δ_a would be wrongly identified by $(\mu_{c1} - \mu_{a0})$ instead of $(\mu_{a1} - \mu_{a0})$.

In order to study the identifiability of parametric class (1), let's consider this is a member of the more general class:

$$\mathcal{M} = \big\{m(y, \mathbf{x}; \boldsymbol{\theta}) = I_{(\mathbf{x}\in A1)}\, m_1(y; \boldsymbol{\theta}) + I_{(\mathbf{x}\in A2)}\, m_2(y; \boldsymbol{\theta}) + \cdots + I_{(\mathbf{x}\in A_j)}\, m_j(y; \boldsymbol{\theta}) + \cdot\cdot$$

$$\cdot\cdot + I_{(\mathbf{x}\in Ak)}\, m_k(y; \boldsymbol{\theta}) | y \in R,\, \mathbf{x} \in A \subseteq R^d,\, A = \cup_j A_j,\, \cap_j A_j = \emptyset\big\} \qquad (5)$$

where the k distributions $m_j(y; \boldsymbol{\theta})$ are not necessarily in the same parametric class. A first useful result is proposed in the following proposition:

Proposition 1 *A necessary and sufficient condition for parametric class (5) to be identifiable is that set $\Xi = \cap_j \Xi_j = \emptyset$; where Ξ_j is the set of pairs $(\boldsymbol{\theta}', \boldsymbol{\theta}'')$, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}'' \in \Theta$ such that $m_j(y; \boldsymbol{\theta}') \equiv m_j(y; \boldsymbol{\theta}'')$.*

Proof (Necessity): suppose that $\Xi = \cap_j \Xi_j \neq \emptyset$, then $m_j(y; \boldsymbol{\theta}') \equiv m_j(y; \boldsymbol{\theta}'')$, $\forall j$ and $\forall\, (\boldsymbol{\theta}', \boldsymbol{\theta}'') \in \Xi$. Consequently $m(y, \mathbf{x}; \boldsymbol{\theta}') = \sum_j I_{(\mathbf{x}\in Aj)}\, m_j(y; \boldsymbol{\theta}') \equiv \sum_j I_{(\mathbf{x}\in Aj)}\, m_j(y; \boldsymbol{\theta}'') = m(y, \mathbf{x}; \boldsymbol{\theta}'')$, $\forall\, (\boldsymbol{\theta}', \boldsymbol{\theta}'') \in \Xi$, which implies that (5) is not identifiable.

Proof (Sufficiency): If $\Xi = \cap_j \Xi_j = \emptyset$, then $\nexists$ pairs $(\boldsymbol{\theta}', \boldsymbol{\theta}'')$, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}'' \in \Theta$ such that $m_j(y; \boldsymbol{\theta}') \equiv m_j(y; \boldsymbol{\theta}'')$, $\forall j$. Consequently $\exists\, y$ such that $m(y, \mathbf{x}; \boldsymbol{\theta}') =$

$\sum_j I_{(\mathbf{x}\in Aj)}\, m_j(y;\boldsymbol{\theta}') \neq \sum_j I_{(\mathbf{x}\in Aj)}\, m_j(y;\boldsymbol{\theta}'') = m(y,\mathbf{x};\boldsymbol{\theta}'')$ which implies that (5) is identifiable•

Parametric class (1) is a particular case of (5), with $k = 4$. Proposition 2 identifies the set Ξ for (1) under the assumption that parametric class of the outcome distributions is a linearly independent set over the field of real number R:

Proposition 2 *If, in (1), the parametric class of outcome distributions $\mathcal{G}$ is a linearly independent set over the field of real number R, then one of the following conditions holds for any pair $(\boldsymbol{\theta}', \boldsymbol{\theta}'') \in \Xi \neq \emptyset$, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}'' \in \Theta$:*

$$\omega'_a = \omega'_c = \omega''_a = \omega''_c,$$

or

$$\omega'_n = \omega'_c = \omega''_n = \omega''_c,$$

or

$$\omega'_a = \omega'_c = \omega'_n = \omega''_a = \omega''_c = \omega''_n.$$

Proof: Given $\mathcal{G}$ is a linearly independent set over R, the mixture in $\varsigma(D_i = 1,\, Z_i = 1)$ is identifiable up to permutations of the label components in the parametric sub-vector $(\omega_a,\, \omega_c,\, \boldsymbol{\eta}_{a1},\, \boldsymbol{\eta}_{c1})$. The pairs $(\boldsymbol{\theta}', \boldsymbol{\theta}'')$, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}'' \in \Theta$ in $\Xi_{\varsigma(D_i=1,\, Z_i=1)}$ are such that $\boldsymbol{\theta}'$ is an element of the set

$$\left\{\boldsymbol{\theta}' : (\omega'_a,\, \omega'_c,\, \boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}) \times \{\omega_n,\, \boldsymbol{\eta}_{n0},\, \boldsymbol{\eta}_{c0}\} \times \{\boldsymbol{\eta}_{a0}\} \times \{\boldsymbol{\eta}_{n1}\} \,|\, \sum_t \omega_t = 1,\, \omega_t > 0,\, \forall t\right\}$$

and $\boldsymbol{\theta}''$ is an element of the set

$$\left\{\boldsymbol{\theta}'' : (\omega''_a,\, \omega''_c,\, \boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1}) \times \{\omega_n,\, \boldsymbol{\eta}_{n0},\, \boldsymbol{\eta}_{c0}\} \times \{\boldsymbol{\eta}_{a0}\} \times \{\boldsymbol{\eta}_{n1}\} \,|\, \sum_t \omega_t = 1,\, \omega_t > 0,\, \forall t\right\},$$

where $(\omega'_a,\, \omega'_c,\, \boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}) = (\omega''_a,\, \omega''_c,\, \boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1})$ up to permutations of the label components.

Again, given $\mathcal{G}$ is a linearly independent set over R, we cannot have $\omega'_a \boldsymbol{\eta}'_{a0} = \omega''_a \boldsymbol{\eta}''_{a0}$ unless $\boldsymbol{\eta}'_{a0} = \boldsymbol{\eta}''_{a0}$ and $\omega'_a = \omega''_a$. Consequently, permutations of the label components in $\varsigma(D_i = 1,\, Z_i = 1)$ are restricted to the case $\omega'_a =$

$\omega'_c = \omega''_a = \omega''_c$. The pairs $(\boldsymbol{\theta}', \boldsymbol{\theta}'')$, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}'' \in \Theta$ in $\Xi_{(D_i=1,\, Z_i=1)} \cap \Xi_{(D_i=1,\, Z_i=0)}$ are such that $\boldsymbol{\theta}'$ is an element of the set

$$\left\{ \boldsymbol{\theta}' : (\omega'_a,\, \omega'_c,\, \boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}, \boldsymbol{\eta}'_{a0}) \times \{\omega_n,\, \boldsymbol{\eta}_{n0},\, \boldsymbol{\eta}_{c0}\} \times \{\boldsymbol{\eta}_{n1}\} \,|\, \sum_t \omega_t = 1,\, \omega_t > 0,\, \forall t \right\}$$

and $\boldsymbol{\theta}''$ is an element of the set

$$\left\{ \boldsymbol{\theta}'' : (\omega''_a,\, \omega''_c,\, \boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1}, \boldsymbol{\eta}''_{a0}) \times \{\omega_n,\, \boldsymbol{\eta}_{n0},\, \boldsymbol{\eta}_{c0}\} \times \{\boldsymbol{\eta}_{n1}\} \,|\, \sum_t \omega_t = 1,\, \omega_t > 0,\, \forall t \right\},$$

where:

$(\omega'_a,\, \omega'_c,\, \boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}, \boldsymbol{\eta}'_{a0}) = (\omega''_a,\, \omega''_c,\, \boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1}, \boldsymbol{\eta}''_{a0})$, if $\omega'_a \neq \omega'_c$,

or

$(\boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}) = (\boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1})$ up to permutations in the label components, and $(\omega'_a,\, \omega'_c, \boldsymbol{\eta}'_{a0}) = (\omega''_a,\, \omega''_c, \boldsymbol{\eta}''_{a0})$, if $\omega'_a = \omega'_c = \omega''_a = \omega''_c$.

Given the constraint $\sum_t \omega_t = 1$ we have $\omega'_n = 1 - \omega'_a - \omega'_c = 1 - \omega''_a - \omega''_c = \omega''_n$. Given the linear independency of the elements of $\mathcal{G}$, we cannot have $\omega'_n \boldsymbol{\eta}'_{n1} = \omega''_n \boldsymbol{\eta}''_{n1}$ unless $\boldsymbol{\eta}'_{n1} = \boldsymbol{\eta}''_{n1}$ and $\omega'_n = \omega''_n$. This implies the pairs $(\boldsymbol{\theta}', \boldsymbol{\theta}'')$, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}'' \in \Theta$ in $\Xi_{(D_i=1,\, Z_i=1)} \cap \Xi_{(D_i=1,\, Z_i=0)} \cap \Xi_{(D_i=0,\, Z_i=1)}$ are such that $\boldsymbol{\theta}'$ is an element of the set

$$\left\{ \boldsymbol{\theta}' : (\omega'_a,\, \omega'_c, \omega'_n,\, \boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}, \boldsymbol{\eta}'_{a0}, \boldsymbol{\eta}'_{n1}) \times \{\, \boldsymbol{\eta}_{n0},\, \boldsymbol{\eta}_{c0}\} \,|\, \sum_t \omega_t = 1,\, \omega_t > 0,\, \forall t \right\}$$

and $\boldsymbol{\theta}''$ is an element of the set

$$\left\{ \boldsymbol{\theta}'' : (\omega''_a,\, \omega''_c, \omega''_n,\, \boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1}, \boldsymbol{\eta}''_{a0}, \boldsymbol{\eta}''_{n1}) \times \{\, \boldsymbol{\eta}_{n0},\, \boldsymbol{\eta}_{c0}\} \,|\, \sum_t \omega_t = 1,\, \omega_t > 0,\, \forall t \right\},$$

where:

$(\omega'_a,\, \omega'_c, \omega'_n,\, \boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}, \boldsymbol{\eta}'_{a0}, \boldsymbol{\eta}'_{n1}) = (\omega''_a,\, \omega''_c, \omega''_n,\, \boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1}, \boldsymbol{\eta}''_{a0}, \boldsymbol{\eta}''_{n1})$, if $\omega'_a \neq \omega'_c$,

or

$(\boldsymbol{\eta}'_{a1},\, \boldsymbol{\eta}'_{c1}) = (\boldsymbol{\eta}''_{a1},\, \boldsymbol{\eta}''_{c1})$ up to permutations in the label components, and $(\omega'_a,\, \omega'_c, \omega'_n, \boldsymbol{\eta}'_{a0}, \boldsymbol{\eta}'_{n1}) = (\omega''_a,\, \omega''_c, \omega''_n, \boldsymbol{\eta}''_{a0}, \boldsymbol{\eta}''_{n1})$, if $\omega'_a = \omega'_c = \omega''_a = \omega''_c$.

Finally, given $\mathcal{G}$ is a linearly independent set over R, the mixture in $\varsigma(D_i = 0,\, Z_i = 0)$ is identifiable up to permutations of the label components

in the parametric sub-vector $(\omega_n, \omega_c, \eta_{n0}, \eta_{c0})$. This implies that the pairs $(\boldsymbol{\theta}', \boldsymbol{\theta}'')$, $\boldsymbol{\theta}' \neq \boldsymbol{\theta}'' \in \Theta$ in Ξ are such that one of the following conditions holds:

$(\eta'_{a1}, \eta'_{c1}) = (\eta''_{a1}, \eta''_{c1})$ up to permutations in the label components, and $(\omega'_a, \omega'_c, \omega'_n, \eta_{n0}, \eta_{c0}, \eta'_{a0}, \eta'_{n1}) = (\omega''_a, \omega''_c, \omega''_n, \eta_{n0}, \eta_{c0}, \eta''_{a0}, \eta''_{n1})$, if $\omega'_a = \omega'_c = \omega''_a = \omega''_c$,

or

$(\eta'_{n0}, \eta'_{c0}) = (\eta''_{n0}, \eta''_{c0})$ up to permutations in the label components, and $(\omega'_a, \omega'_c, \omega'_n, \eta_{a1}, \eta_{c1}, \eta'_{a0}, \eta'_{n1}) = (\omega''_a, \omega''_c, \omega''_n, \eta_{a1}, \eta_{c1}, \eta''_{a0}, \eta''_{n1})$, if $\omega'_n = \omega'_c = \omega''_n = \omega''_c$,

or

$(\eta'_{a1}, \eta'_{c1}) = (\eta''_{a1}, \eta''_{c1})$ and $(\eta'_{n0}, \eta'_{c0}) = (\eta''_{n0}, \eta''_{c0})$ up to permutations in the label components, and $(\omega'_a, \omega'_c, \omega'_n, \eta'_{a0}, \eta'_{n1}) = (\omega''_a, \omega''_c, \omega''_n, \eta''_{a0}, \eta''_{n1})$, if $\omega'_a = \omega'_c = \omega'_n = \omega''_a = \omega''_c = \omega''_n$●

Given Propositions 1 and 2, a distribution function $f(y_i, d_i, z_i; \boldsymbol{\theta})$ in (1) is identifiable unless: $\omega_a = \omega_c$, or $\omega_n = \omega_c$, or $\omega_a = \omega_n = \omega_c$. This is a set of less restrictive conditions compared to simple mixture models where identifiability is assured only up to permutations of the label components.

The restriction on the parametric class of the outcome distributions $\mathcal{G}$, imposed in Proposition 2, rules out the case of a binary outcome. The parametric class of binomials $Bi(N, \theta)$, $0 < \theta < 1$, is indeed a linearly independent set on R if and only if $N \geq 2T - 1$, where N is the number of independent trials for each observation (Teicher 1961, 1963; Titterington et al. 1985). Given $T = 2$ for the two mixtures in (1), the condition on N is not satisfied for a binary outcome, where $N = 1 < 2T - 1 = 3$. This implies that for a binary outcome Ξ could be greater than under $N \geq 2T - 1$. This is confirmed by an application to data from a randomized community trial of the impact of vitamin A supplements on children's survival (Imbens and Rubin, 1997a). The authors made a likelihood analysis of this randomized experiment with noncompliance, a binary outcome, in absence of always-takers and removing the exclusion restriction. There was no a unique solution, rather the resulting likelihood function had a set-valued maximizer.

3 Estimation issues

We have showed in Section 2 that in a likelihood-based analysis of a randomized experiment without exclusion restriction the parameter vector $\boldsymbol{\theta}$ is only partially identified. In recent years, some methods for relaxing the exclusion

restriction based on exploiting extra information compared to the assumptions 1-5 of Section 2 were proposed. For example, Hirano et al. (2000) that worked in a Bayesian context adopting a relatively diffuse but proper prior distribution, or more recently Jo (2002) that studied alternative model specifications allowing the identification of causal effects in the presence of observed pretreatment information. In this Section the case of normally distributed outcomes, that is posing $g^i_{tz} = N(y_i\,;\mu_{tz},\,\sigma_{tz})$ in (4), is considered. An alternative approach based on identifying the solution of the likelihood equations closest to the method of moments estimate of the parameter vector will be proposed.

A first problem associated with a likelihood analysis of $\boldsymbol{\theta}$ in (1) arises from the possibility to have multiple roots for the likelihood equations. This is due to the two mixtures of distributions involved, indeed the likelihood function for a mixture model will generally have multiple roots[1] (McLachlan and Peel, 2000). The presence of multiple roots when the likelihood is based only on the units in one of the two mixtures, that is $\sum_{i\in\varsigma(D_i=1,\,Z_i=1)} \log f(y_i,\,d_i,\,z_i\,;\omega_a,\omega_c,\boldsymbol{\eta}_{a1},\boldsymbol{\eta}_{c1})$ or $\sum_{i\in\varsigma(D_i=0,\,Z_i=0)} \log f(y_i,\,d_i,\,z_i\,;\omega_n,\omega_c,\boldsymbol{\eta}_{n0},\boldsymbol{\eta}_{c0})$, is sufficient to have multiple roots for the likelihood equations based on the entire sample, given the particular factorial structure of the distribution (1). A proof is in Appendix A. In general when the likelihood equations have multiple roots, the consistency of the MLE is guaranteed only for those class of distributions satisfying Wald's conditions (1949).

Supposing normally distributed outcomes, additional problems arise from the unboundedness of the likelihood. These are due to the fact that likelihood function for a mixture of normal distributions is unbounded, Day (1969). Again, the unboundedness of the likelihood function based only on the units in one of the two mixtures, $i \in \varsigma(D_i = 1,\, Z_i = 1)$ or $i \in \varsigma(D_i = 0,\, Z_i = 0)$, implies the unboundedness of the likelihood function based on the entire sample, given the particular factorial structure of the distribution (1). A proof is in Appendix B. The consequence is that an efficient estimator could not exist as a global likelihood maximizer. The existence of a consistent and efficient likelihood equation root is guaranteed by the satisfaction of the multivariate extension of the Cramer conditions. Simple but tedious

[1]The local maximum points that do not correspond to the consistent maximizer are usually indicated as "spurious" maximum points in the mixture models literature. In particular, for normally distributed outcomes the spurious maximum points corresponding to parameter points having at least one variance component very close to zero are generated by groups of few outliers (Day, 1969).

checks show the existence of the first, second and third derivatives of the likelihood. Each of these derivatives has a factorial structure where each factor is a derivative of the type showed by Kiefer (1978) in proving the existence of a consistent and efficient likelihood root for a mixture of two normal distributions. This guarantees the boundedness of them and the positive definiteness of the dispersion matrix.

Given the presence of multiple roots for the likelihood equations and the unboundedness of the likelihood function, an approach to identify the consistent and efficient estimate can be based on finding the root closest to a consistent estimate of the parameter vector, typically that resulting by the method of moments (Lehmann and Casella, 1998). In the present case the method of moments estimate of the parameter vector, $\tilde{\boldsymbol{\theta}}$, are obtainable by:

- equating the first three moments of $f(d_i,\, z_i\,;\, \omega_a, \omega_n, \pi)$ to its first three sample moments; we obtain $\tilde{\omega}_a = \sum_i I_{(D_i=1,\, Z_i=0)} / \sum_i I_{(Z_i=0)}$ (the proportion of treated units in the group of not assigned units), $\tilde{\omega}_n = \sum_i I_{(D_i=0,\, Z_i=1)} / \sum_i I_{(Z_i=1)}$ (the proportion of untreated units in the group of assigned units), $\tilde{\pi} = \sum_i I_{(Z_i=1)}/N$, and $\tilde{\omega}_c$ as the difference $\tilde{\omega}_c = 1 - \tilde{\omega}_a - \tilde{\omega}_n$;

- equating the first two moments of $I_{\varsigma(D_i=1,\, Z_i=0)}\, N(y_i; \mu_{a0},\, \sigma_{a0})$, and $I_{\varsigma(D_i=0,\, Z_i=1)}\, N(y_i;$ to their first two sample moments respectively. We obtain: $\tilde{\mu}_{a0}$ and $\tilde{\sigma}_{a0}$ as the sample mean and sample variance of y_i for $i \in \varsigma(D_i = 1,\, Z_i = 0)$, $\tilde{\mu}_{n1}$ and $\tilde{\sigma}_{n1}$ as the sample mean and sample variance of y_i for $i \in \varsigma(D_i = 0,\, Z_i = 1)$;

- equating the first five moments of $I_{\varsigma(D_i=1,\, Z_i=1)}\, N(y_i; \omega_{c|11},\, \mu_{a1}, \mu_{c1},\, \sigma_{a1},\, \sigma_{c1})$, and $I_{\varsigma(D_i=0,\, Z_i=0)}\, N(y_i; \omega_{c|00},\, \mu_{n0}, \mu_{c0},\, \sigma_{n0}, \sigma_{c0})$ to their first five sample moments; where $\omega_{t|dz}$ is the conditional mixing probability $P(C_i = t \,|\, D_i = d,\, Z_i = z)$. We know the two mixtures are identifiable only up to the permutation of their label components. A way to check the labelling for the mixture $\varsigma(D_i = 1,\, Z_i = 1)$ can be proposed by comparing the resulted estimate $\tilde{\omega}_{c|11}$ to a simple transformation of $\tilde{\omega}_a$ and $\tilde{\omega}_c$: $\tilde{\omega}_c/(\tilde{\omega}_a + \tilde{\omega}_c)$; the latter is indeed a consistent estimate of $\omega_{c|11}$ then a good term of reference to compare $\tilde{\omega}_{c|11}$. The proposal is to check the distance between $\tilde{\omega}_{c|11}$ and $\tilde{\omega}_c/(\tilde{\omega}_a + \tilde{\omega}_c)$; then to switch the tern $(\tilde{\omega}_{c|11},\, \tilde{\mu}_{c1}, \tilde{\sigma}_{c1})$ to $(1 - \tilde{\omega}_{c|11},\, \tilde{\mu}_{a1},\, \tilde{\sigma}_{a1})$ if $\left|\tilde{\omega}_{c|11} - \tilde{\omega}_c/(\tilde{\omega}_a + \tilde{\omega}_c)\right| > \left|(1 - \tilde{\omega}_{c|11}) - \tilde{\omega}_c/(\tilde{\omega}_a + \tilde{\omega}_c)\right|$. Analogous arguments hold for the other mixture.

However, there is no guarantee to obtain an unique real solution for the two mixtures without imposing the equal variances conditions: $\sigma_{a1} = \sigma_{c1}$ and $\sigma_{n0} = \sigma_{c0}$ (Quandt and Ramsey, 1978; Lindsay and Basak, 1993). Under these two homoscedastic conditions, the likelihood analysis can be performed in a first step by calculating $\tilde{\boldsymbol{\theta}}$, then detecting the root of the likelihood equations closest to $\tilde{\boldsymbol{\theta}}$. In order to make the detection less time consuming, it can be limited to a neighborhood of $\tilde{\boldsymbol{\theta}}$: $\Omega_h^{\tilde{\theta}}$ (where h is the radius).

An empirical procedure can be proposed also for the unrestricted (heteroscedastic) case where the component variances for the two mixtures are unequal. Given the method of moments estimates of the mixing probabilities, $\tilde{\boldsymbol{\omega}} = (\tilde{\omega}_a, \tilde{\omega}_n, \tilde{\omega}_c)$, are not affected by restrictions on the variance components, then the second step can be limited to detect the root $\tilde{\boldsymbol{\theta}}$ whose subvector $\hat{\boldsymbol{\omega}} = (\hat{\omega}_a, \hat{\omega}_n, \hat{\omega}_c)$ is closest to $\tilde{\boldsymbol{\omega}}$. Again, the detection can be limited to a neighborhood of $\tilde{\boldsymbol{\omega}}$ and of radius h: $\Omega_h^{\tilde{\omega}}$. From a theoretical point of view, the procedure guarantees only the detection of the efficient likelihood estimate for $\boldsymbol{\omega} = (\omega_a, \omega_n, \omega_c)$; however the simulation-based analysis in next section will show the conditions under which the method can have a good performance in detecting the efficient likelihood estimate for the entire parameter vector $\boldsymbol{\theta}$.

From a computational point of view, the EM algorithm can make the inference relatively straightforward. The EM algorithm is indeed attractive in making likelihood inference because if the compliance status C_i was known for all units, the likelihood would not involve mixtures. The compliance status of the units in any of the two mixtures can be indeed considered as a missing information whose imputation produces the so-called augmented likelihood. Moreover, in our context the augmented log-likelihood function is linear in the missing information, so the EM algorithm corresponds to fill-in missing data and then updating parameter estimates. The imputation of the unobserved compliance status is handled by the E-step; it requires the calculation of the conditional expectation of C_i given the observed data and the current fit for $\boldsymbol{\theta}$. The compliance status C_i can be represented by a three component indicator $t = c$ *(complier), n (never-taker), a (always taker).* At the k-iteration, the conditional probability of subject i being type t given the observed data and a current value of the vector $\boldsymbol{\theta}$, $\tau_{it}^{(k)}(\hat{\boldsymbol{\theta}}^{(k-1)})$, is obtainable by a ratio of two quantities. The numerator of the ratio is the corresponding Table 3.1 entry and the denominator is the corresponding row total, where $\hat{g}_{tz}^{i\,(k-1)}$ is the outcome distribution for a unit in the t group and assigned to

the treatment z, based on the estimated parameter vector updated at the $(k-1)$ iteration, $\hat{\boldsymbol{\theta}}^{(k-1)}$.

Table 3.1. *Inputs for calculating the conditional probabilities* $\tau_{it}^{(k)}(\hat{\boldsymbol{\theta}}^{(k-1)})$.

D_i	Z_i	Subject type t		
		$t=a$	$t=n$	$t=c$
0	0	0	$\hat{\omega}_n^{(k-1)} \cdot \hat{g}_{n0}^{i\,(k-1)}$	$\hat{\omega}_c^{(k-1)} \cdot \hat{g}_{c0}^{i\,(k-1)}$
0	1	0	1	0
1	0	1	0	0
1	1	$\hat{\omega}_a^{(k-1)} \cdot \hat{g}_{a1}^{i\,(k-1)}$	0	$\hat{\omega}_c^{(k-1)} \cdot \hat{g}_{c1}^{i\,(k-1)}$

The subsequent M-step then maximizes the log-likelihood function based on the augmented data set, that is the data set created by merging the observed and the imputed data. This is equivalent to a weighted maximization of the log-likelihood function, where subjects are differently classified in the different compliance groups, t, with weights equal to the conditional probabilities of being in t calculated in the E-step. The output is the update estimated vector $\hat{\boldsymbol{\theta}}^{(k)}$.

In particular, for the normal distributions case the updates of the component means, $\hat{\mu}_{tz}^{(k)}$, and component variances, $(\hat{\sigma}_{tz}^{(k)})^2$, are given by:

$$\hat{\mu}_{tz}^{(k)} = \sum_{i=1}^{n} \left\{ \tau_{it}^{(k)}(\hat{\boldsymbol{\theta}}^{(k-1)}) \cdot y_i \cdot I(Z_i = z) \right\} / \sum_{i=1}^{n} \left\{ \tau_{it}^{(k)}(\hat{\boldsymbol{\theta}}^{(k-1)}) \cdot I(Z_i = z) \right\},$$

$$(\hat{\sigma}_{tz}^{(k)})^2 = \sum_{i=1}^{n} \left\{ \tau_{it}^{(k)}(\hat{\boldsymbol{\theta}}^{(k-1)}) \cdot (y_i - \hat{\mu}_{tz}^{(k)})^2 \cdot I(Z_i = z) \right\} / \sum_{i=1}^{n} \left\{ \tau_{it}^{(k)}(\hat{\boldsymbol{\theta}}^{(k-1)}) \cdot I(Z_i = z) \right\}.$$

4 Examples based on artificial data sets

This Section proposes some simulation analyses based on artificial samples from hypothetical distributions satisfying the assumptions 1-5 presented in Section 2; we are therefore fully relaxing the exclusion restriction. The aim is to study the relative advantages of the two-steps procedures proposed in Section 3.

We start by analyzing the homoscedastic case. At this purpose we consider a set of six hypothetical populations with equal distributions apart

from the parameter μ_{c0} for which we choose a set of values ranging between $\mu_{c0} = 1.2$ and $\mu_{c0} = 6$. The mean for the compliers not assigned is indeed posed $\mu_{c0} = 1.2, 2, 3, 4, 5, 6$. We impose the equal variances condition for any of the two mixtures: $\sigma_{a1} = \sigma_{c1} = 1.2$ and $\sigma_{n0} = \sigma_{c0} = 0.85$. The common parameters values for the six hypothetical distributions are shown in Table 4.1.

Table 4.1. *Hypothetical populations distributions under the homostedastic conditions: common parameters values.*

t	ω_t	(μ_{t0}, σ_{t0})	(μ_{t1}, σ_{t1})
a	0.4	(0, 1)	(1, 1.2)
n	0.25	(1, 0.85)	(2, 1)
c	0.35	(., 0.85)	(7, 1.2)
$\pi = P(Z_i = 1) = 0.25$			

In order to evaluate the performance of the two-steps procedure restricted to $\Omega_h^{\tilde{\theta}}$, we drew 100 samples each of size 10000 from any of these six distributions and for any of the proposed value of h (0.25, 0.15, and 0.04). For each sample, we started 50 times the EM algorithm with random values of $\boldsymbol{\theta}$, and we detect the root closest to $\tilde{\boldsymbol{\theta}}$ in $\Omega_h^{\tilde{\theta}}$. Table 3 shows that, for the considered samples, the two-steps procedure does not always converge to the solution corresponding to the consistent maximizer[2]. Indeed, it can converge also to spurious solutions, or to points on the boundary of $\Omega_h^{\tilde{\theta}}$. However, we note that, for any of the proposed value of h, the frequencies of convergence to the consistent solution increase with the value of μ_{c0}, that is with the difference in means for the mixture of compliers and never-takers not assigned: $|\mu_{c0} - \mu_{n0}| = |\mu_{c0} - 1|$.

[2]Like in Hataway (1986), the local maximum point that corresponds to the consistent maximizer is taken to be the limit of the EM algorithm using the true parameter values as a starting point.

Table 4.2. *Performances of the two-steps procedure restricted to* $\Omega_h^{\tilde{\theta}}$ *for some values of* h *and* μ_{c0}*; homoscedastic case*.*

h	μ_{c0}	Convergence to the consistent solution	Convergence to a spurious solution	Convergence on the boundary of $\Omega_h^{\tilde{\theta}}$
0.25	6	100	0	0
	5	100	0	0
	4	100	0	0
	3	100	0	0
	2	30	14	56
	1.2	12	12	76
0.15	6	100	0	0
	5	100	0	0
	4	100	0	0
	3	97	0	3
	2	22	6	72
	1.2	0	3	97
0.04	6	46	0	54
	5	34	0	67
	4	20	0	80
	3	24	0	76
	2	0	0	100
	1.2	0	0	100

*: 100 replications for any value of h and μ_{c0}; size: 10000 for each sample.

The simulation analysis continues by removing the equal variances conditions, then considering the two-steps procedure restricted to $\Omega_h^{\tilde{\omega}}$. We repeat the simulation analysis with six hypothetical distributions whose parameters assume the same values of the previous ones apart from σ_{n0}, that now is posed $\sigma_{n0} = 1.15$, and σ_{c1}, that now is posed $\sigma_{c1} = 0.7$. The common parameters values for these hypothetical distributions are shown in Table 4.3. Table 4.4 shows that, like in the homoscedastic case, the two-steps procedure does not always converge to the solution corresponding to the consistent maximizer. Again, the frequencies of convergence to the consistent solution increase with

the difference $|\mu_{c0} - \mu_{n0}|$ for any of the proposed value of h (0.03, 0.01 and 0.005).

Table 4.3. *Hypothetical populations distributions (heterostedastic case): common parameters values.*

t	ω_t	(μ_{t0}, σ_{t0})	(μ_{t1}, σ_{t1})
a	0.4	(0, 1)	(1, 1.2)
n	0.25	(1, 1.15)	(2, 1)
c	0.35	(., 0.85)	(7, 0.7)
$\pi = P(Z_i = 1) = 0.25$			

Table 4.4. *Performances of the two-steps procedure restricted to $\Omega_h^{\tilde{\omega}}$ for some values of h and μ_{c0}; heteroscedastic case*.*

h	μ_{c0}	Convergence to the consistent solution	Convergence to a spurious solution	Convergence on the boundary of $\Omega_h^{\tilde{\omega}}$
0.03	6	100	0	0
	5	100	0	0
	4	100	0	0
	3	67	33	0
	2	54	46	0
	1.2	48	52	0
0.01	6	60	0	40
	5	59	0	41
	4	75	0	25
	3	50	31	19
	2	42	35	23
	1.2	38	38	24
0.005	6	44	0	56
	5	46	0	54
	4	42	0	58
	3	30	25	45
	2	27	23	50
	1.2	22	23	55

*: 100 replications for any value of h and μ_{c0}; size: 10000 for each sample.

Table 4.5 presents the average Allocation Rates, AR (McLachlan and Basford, 1988), calculated for the consistent solution over 100 replications from the hypothetical distributions of Table 4.3 for each of the proposed values of μ_{c0}. The AR is a useful indicator for quantifying a mixture disentanglement and is calculated by averaging the higher imputation probability[3] for any unit observed at convergence of the EM algorithm.: $AR = \left\{\sum_{i=1}^{N} \max_t \tau_{it|dz}^{(k)}(\hat{\boldsymbol{\theta}}^{(k-1)})\right\}/N$. The AR takes the upper value 1 only if the mixtures are perfectly disentangled, otherwise AR is less than 1 but positive. The lower bound for AR is $1/p$, where p is the number of mixture components ($AR \geq 0.5$ in our cases). Low AR values correspond to bad mixtures disentanglements, and vice-versa. Table 4.5 shows an increasing trend: that the overall average AR increases with the difference $|\mu_{c0} - \mu_{n0}|$. In particular, while the average AR is substantially stable over the six populations concerning mixture $\varsigma(D_i = 1, Z_i = 1)$, the decreasing value of the overall AR is due to the bad disentaglement of $\varsigma(D_i = 0, Z_i = 0)$.

The simulation-based analysis suggests the identification of the consistent solution with the proposed two-steps procedures is feasible when a good disentanglement of both the mixtures happens as indicated by the average AR values. The procedure under the homoscedastic conditions appear to perform slightly better than in the heteroscedastic case. The former indeed does not converge to spurious solutions even for the samples where $\mu_{c0} = 3$, other than when μ_{c0} is posed equal to 4, 5, and 6. The negative effect of getting near the means of a mixture has been sufficient in order to increase the frequencies of converging to a spurious solution both for the homoscedastic and heteroscedastic case. A practical suggestion then can be to check the overall AR and to compare the distances to $\tilde{\boldsymbol{\theta}}$ (or $\tilde{\boldsymbol{\omega}}$) for the solutions detected in the neighborhood $\Omega_h^{\tilde{\theta}}$ (or $\Omega_h^{\tilde{\omega}}$). A low overall AR and the presence of solutions with no appreciable different distances to $\tilde{\boldsymbol{\theta}}$ (or $\tilde{\boldsymbol{\omega}}$), can be considered as a signal for the necessity to introduce a restriction on the difference $|\mu_{c0} - \mu_{n0}|$ and/or $|\mu_{c1} - \mu_{a1}|$.

[3]The imputation probability is the conditional probability of unit i being compliance status t given that the unit is in group $\varsigma(D_i = d, Z_i = z)$, Mercatanti (2005).

Table 4.5. *Average Allocation Rates (AR) for the consistent solutions for some values of* μ_{c0}.*

μ_{c0}	Average AR	
1.2	Overall	0.8327
	for $\varsigma(D_i = 1, Z_i = 1)$	0.9990
	for $\varsigma(D_i = 0, Z_i = 0)$	0.6269
2.0	Overall	0.8760
	for $\varsigma(D_i = 1, Z_i = 1)$	0.9991
	for $\varsigma(D_i = 0, Z_i = 0)$	0.7249
3.0	Overall	0.9346
	for $\varsigma(D_i = 1, Z_i = 1)$	0.9991
	for $\varsigma(D_i = 0, Z_i = 0)$	0.8550
4.0	Overall	0.9724
	for $\varsigma(D_i = 1, Z_i = 1)$	0.9993
	for $\varsigma(D_i = 0, Z_i = 0)$	0.9391
5.0	Overall	0.9900
	for $\varsigma(D_i = 1, Z_i = 1)$	0.9993
	for $\varsigma(D_i = 0, Z_i = 0)$	0.9782
6.0	Overall	0.9971
	for $\varsigma(D_i = 1, Z_i = 1)$	0.9992
	for $\varsigma(D_i = 0, Z_i = 0)$	0.9940

*: 100 replications for each μ_{c0}; size: 10000 for each sample.

In order to evaluate the relative merits of the two-steps procedures, we continue our analysis by drawing 100 samples of size 10000 from two hypothetical populations. One of these population satisfies the homoscedatic conditions and has the parameter values listed in Table 4.1; the other one does not satisfy the homoscedatic conditions and has the parameter values listed in Table 4.3; μ_{c0} is posed equal to 6 for both the hypothetical populations.

The efficient likelihood estimate, ELE, interior to $\Omega_h^{\tilde{\theta}}$ has been identified, running the EM algorithm and posing $h = 0.25$, for each sample from the hypothetical population satisfying the homoscedastic conditions. Analogously for the samples from the other hypothetical population, where h is posed equal to 0.03 in $\Omega_h^{\tilde{\omega}}$. Tables 4.6 and 4.7 report mean biases, root mean

squared errors, coverage rates of 95% confidence intervals, and mean widths of the intervals, for the repeated estimates of some parameters. The results are also compared to other standard procedures: (i) the maximum likelihood method under the weak exclusion restriction, by imposing: $\mu_{a1} = \mu_{a0}$, $\mu_{n1} = \mu_{n0}$, $\sigma_{a1} = \sigma_{a0}$, $\sigma_{n1} = \sigma_{n0}$; (ii) the C.A.C.E. (Compliers Average Causal Effect), $\mu_{c1} - \mu_{c0}$, obtained by the instrumental variables method.

Tables 4.6 and 4.7 show that the estimations of the compliers parameters based only on imposing the weak version of the exclusion restriction systematically present absolute mean biases and root MSEs higher than those calculated by the two-steps procedures. The C.A.C.E. estimations obtained by the instrumental variables method, that have very high coverage rates but at the cost of dramatically higher mean widths of associated 95% intervals are even worse. It is to be put in evidence that the maximum likelihood analyses under the weak exclusion restriction do not produce unique solutions on the artificial samples. For this reason, the analyses under the weak exclusion restriction have been restricted to a neighborhood of $\tilde{\boldsymbol{\omega}}_t = (\tilde{\omega}_a, \tilde{\omega}_n, \tilde{\omega}_c)$, posing the radius equal to 0.01, for both the hypothetical populations.

Table 4.6. *Operating characteristics of various procedures for replications from an hypothetical distribution (homoscedastic case).*

Parameter	Estimator	Mean bias	Root MSE	95% Interval	
				Coverage Rate	Mean width
$\mu_{c0} = 6$	ELE interior to $\Omega_h^{\tilde{\theta}}$	-0.001	0.025	0.93	0.097
	MLE under the exclusion restriction	0.109	0.136	0.33	0.093
$\mu_{c1} = 7$	ELE interior to $\Omega_h^{\tilde{\theta}}$	-0.004	0.046	0.95	0.182
	MLE under the exclusion restriction	1.254	1.461	0.00	0.330
$\sigma_{c0} = 0.85$	ELE interior to $\Omega_h^{\tilde{\theta}}$	-0.000	0.008	1.00	0.048
	MLE under the exclusion restriction	0.019	0.022	0.67	0.048
$\sigma_{c1} = 1.2$	ELE interior to $\Omega_h^{\tilde{\theta}}$	0.001	0.019	1.00	0.129
	MLE under the exclusion restriction	-0.987	1.204	0.02	0.340
C.A.C.E.= $\mu_{c1} - \mu_{c0} = 1$	ELE interior to $\Omega_h^{\tilde{\theta}}$	0.004	0.045	0.94	0.180
	MLE under the exclusion restriction	1.145	1.412	0.04	0.336
	IVE	-1.802	1.817	1.00	16.07

Table 4.7. *Operating characteristics of various procedures for replications from an hypothetical distribution (heteroscedastic case).*

Parameter	Estimator	Mean bias	Root MSE	95% Interval	
				Coverage Rate	Mean width
$\mu_{c0} = 6$	ELE interior to Ω_h^{ω}	-0.001	0.025	0.94	0.098
	MLE under the exclusion restriction	0.211	0.213	0.00	0.096
$\mu_{c1} = 7$	ELE interior to Ω_h^{ω}	-0.004	0.030	0.97	0.118
	MLE under the exclusion restriction	0.253	0.255	0.00	0.118
$\sigma_{c0} = 0.85$	ELE interior to Ω_h^{ω}	-0.001	0.011	0.98	0.050
	MLE under the exclusion restriction	0.034	0.036	0.26	0.049
$\sigma_{c1} = 0.7$	ELE interior to Ω_h^{ω}	-0.001	0.016	0.94	0.068
	MLE under the exclusion restriction	-0.009	0.021	0.86	0.066
C.A.C.E.= $\mu_{c1} - \mu_{c0} = 1$	ELE interior to Ω_h^{ω}	-0.003	0.030	0.97	0.115
	MLE under the exclusion restriction	0.041	0.050	0.70	0.115
	IVE	-1.844	1.857	1.00	15.99

5 An illustrative application: return to schooling in Germany and Austria

In microeconomic literature, the IV method has been widely used in evaluating return to schooling. The method provided indeed a good strategy for solving the selection bias problem that arises when an individual's choice of educational attainment is related to the potential earnings (Card,1999). Some previous studies provide examples of various choices of the instrumental variable such as: quarter of birth (Angrist and Krueger, 1991), college proximity (Card, 1995; Kling, 2001), education policy reform (Denny and

Harmon, 2000), presence of any sisters (Deschenes, 2002), place of childhood (Becker and Siebern-Thomas, 2004).

In particular, two remarkable studies have been recently proposed by Ichino and Winter-Ebmer (IW henceforth) in 1999 and 2004. In both papers the authors investigated the causal effect of education on earnings: the first paper (1999) intended for estimating lower and upper bounds of returns to schooling in Germany, the second (2004) for quantifying the long run educational cost of World War Two in Germany and Austria. In particular the basic idea characterizing the IW 2004 paper relies on the fact that individuals who were about ten years old during or immediately after the war, were damaged in their educational choices compared to individuals in the immediately previous or subsequent cohorts. War physical disruptions and related consequences indeed made harder to achieve the desired level of education for most of the schooling age population in these two countries. Moreover the authors show, using the IV method, that individuals whose education was affected by the war (compliers) suffered a significant earning loss about forty years after the end of the war. For this purpose the IW causal analysis was supported by several instruments; in particular, given the date of birth can be reasonably supposed to be a random event, cohort of birth was adopted as an instrumental variable for both countries[4]. The authors had to assume the exclusion restriction, other than the assumptions 1-4 of Section 2, for identifying and evaluating the average causal effect for compliers by the IV method.

In order to show an example of fully relaxing the exclusion restriction and consequently estimating causal effects also for noncompliers, the previously proposed procedure will be here applied to the same economic context of the IW (2004) paper. The data are from Mikrozensus 1981 for Austria (a 1% sample of the Austrian population), and from the Socio-Economic Panel, wave 1986, for Germany. We are considering males born between 1925 and 1949 for both countries.

Log hourly earnings for employed workers are observed about 40 years after the end of the war. Like IW, and in order to consider the increasing trend of individual earnings respect to age, the outcome Y_i is defined as the residual of a regression of log hourly earnings on a cubic polynomial in age. An

[4]Other two significant instrumental variables were adopted for Germany: an indicator of the father educational background and an indicator of the father's serving in the military during the war.

increasing trend respect to age also characterized the candidate treatment, that is the individual years of education; for this reason the residuals of a regression of years of education on a cubic polynomial in age are calculated[5]. But in order to apply the previously proposed procedure, the treatment has to be a binary variable. Then we define the treatment, D_i, equal to one if the individual residual is smaller than the residuals sample average and equal to zero if the individual residual is greater than the residuals sample average. In this way we are considering individuals having $D_i = 1$ as low educated, and individuals having $D_i = 0$ as high educated. The cohort of birth is used as an instrumental variable, Z_i, having the role of a random assignment to treatment. For this purpose, Z_i has to be necessarily equal to one for people assigned to being low educated, and equal to zero for people assigned to being high educated. Table 5.1 shows that both the estimated mean years of education and the estimated mean residuals of the years of education[6] are smaller for individuals in the cohort 1930-39[7] than for people in the cohort obtained merging 1925-29 and 1940-49 cohorts. These results suggest defining $Z_i = 1$ for individuals born during 1930-39, and $Z_i = 0$ for individuals born during 1925-29 or 1940-49.

Table 5.1. *Estimated mean years of education and estimated mean residual of years of education per country and cohort of birth.*

Country	Cohort of birth	Num. observ.	Years of education	Residuals of years of educ.
Germany	1930-39	633	11.36 (0.091)	-0.243 (0.091)
	1925-29 ∪ 1940-49	893	11.86 (0.084)	0.099 (0.083)
Austria	1930-39	11765	9.18 (0.017)	-0.134 (0.017)
	1925-29 ∪ 1940-49	17383	9.49 (0.015)	0.073 (0.015)

Standard errors in parenthesis.

We apply the likelihood analysis presented in the Section 3 with no restrictions on the variance components. At this purpose the first step will be

[5]Like IW, these residuals are calculated by considering individuals born between 1910 and 1960, and by including two dummies (1949, 1952) in order to consider the increases in the minimal school leaving age in Austria.

[6]For Germany, the units having missing values in the years of education have been dropped, and the resulting sample size is 1526. There are no missing years of education for the 29148 units in the Austrian sample.

[7]The individuals in 1930-39 cohort were in schooling age during World War Two.

limited to estimate the mixing probabilities by the method of moments, $\tilde{\omega}$; the second step to detect the root of the likelihood equations closest to $\tilde{\omega}$ in $\Omega_h^{\tilde{\omega}}$. The outcomes are assumed to be normally distributed[8]. Table 5.2[9] presents the method of moments estimate of the mixing probabilities for the two countries $\tilde{\omega} = (\tilde{\omega}_a, \tilde{\omega}_n, \tilde{\omega}_c)$.

Table 5.2. *Estimated mixing prob.* $\tilde{\omega}_t$ *per country,* $t = a, n, c$.

Country	$\tilde{\omega}_a$	$\tilde{\omega}_n$	$\tilde{\omega}_c$
Germany	0.7309	0.2219	0.0470
Austria	0.7798	0.1519	0.0682

The value $\tilde{\omega}_c$ in Table 5.2, estimating the probability of an individual being in the group of compliers, can also be obtained as the difference between the average treatment under $Z_i = 1$ and $Z_i = 0$. A simple t-test on $\tilde{\omega}_c$ informs about the causal effect of the supposed randomized instrument on the treatment; we obtain a highly significant result for the t-test on $\tilde{\omega}_c$ for Austria (t: 10.58, *s.e.*: 0.0062, p-value: 0.000); for Germany the t-test on $\tilde{\omega}_c$ assumes a value of 1.83 corresponding to a p-value of 0.067 (s.e.: 0.0254), then a significant effect but at a level of at least 6.7%.

We have seen in the Section 2 that the parameter vector $\boldsymbol{\theta}$, in is identified unless $\omega_a = \omega_c$ and/or $\omega_n = \omega_c$; these conditions on the mixing probabilities has been largely refused by likelihood ratio tests, based on $f(d_i, z_i; \omega_a, \omega_n)$, for both the countries. Table 5.3 presents the results of the two-steps procedure posing $h = 0.03$ in $\Omega_h^{\tilde{\omega}}$.

For Germany the proposed method produces a unique nonspurious solution interior to $\Omega_h^{\tilde{\omega}}$, $\hat{\boldsymbol{\theta}}_{\text{Ger}}$, whose elements are all significantly different from zero apart from the outcome means for compliers, $\hat{\mu}_{c0}$ and $\hat{\mu}_{c1}$.

[8]This assumption is made accordingly to Imbens and Rubin (1997b) who estimated the return to high school in the United States with quarter to birth as an instrumental variable. Normality for the log of weekly earning was there assumed in order to present a parametric MLE alternative to the standard IV method. Other than the exclusion restriction, the authors imposed also that the variance for not assigned compliers equals that for never-takers and the variance for assigned compliers equals that for always-takers.

[9]Units having missing values in the years of education and/or in the hourly earning have been dropped. The resulting sample size is 15434 individuals for Austria, and 1160 for Germany.

For Austria, the procedure does not identify a unique nonspurious interior solution; we obtain indeed two roots interior to $\Omega_h^{\tilde{\omega}}$: $\hat{\boldsymbol{\theta}}_{\text{Aus},1}$ and $\hat{\boldsymbol{\theta}}_{\text{Aus},2}$, for which all the parameters are significantly different from zero apart from the outcome mean for assigned compliers, $\hat{\mu}_{c1}$.

Table 5.3. *Results from the two-steps procedure restricted to* $\Omega_h^{\tilde{\omega}}$ *per country;* $h = 0.03$.

	Germany	Austria	
	$\hat{\boldsymbol{\theta}}_{\text{Ger}}$	$\hat{\boldsymbol{\theta}}_{\text{Aus},1} : \mu_{c1} > \mu_{a1}$ $\mu_{n0} < \mu_{c0}$	$\hat{\boldsymbol{\theta}}_{\text{Aus},2} : \mu_{c1} > \mu_{a1}$ $\mu_{n0} > \mu_{c0}$
$\hat{\omega}_a$	0.7236 (0.0253)	0.7769 (0.0075)	0.7764 (0.0075)
$\hat{\omega}_n$	0.2221 (0.0150)	0.1489 (0.0044)	0.1481 (0.0044)
$\hat{\omega}_c$	0.0543 (0.0110)	0.0740 (0.0058)	0.0753 (0.0058)
$\hat{\mu}_{a0}$	-0.0872 (0.0317)	-0.0740 (0.0032)	-0.0740 (0.0032)
$\hat{\mu}_{a1}$	-0.1484 (0.0154)	-0.0802 (0.0042)	-0.0803 (0.0042)
$\hat{\mu}_{n0}$	0.2243 (0.0256)	0.2806 (0.0132)	0.3213 (0.0149)
$\hat{\mu}_{n1}$	0.3761 (0.0514)	0.3502 (0.0123)	0.3502 (0.0123)
$\hat{\mu}_{c0}$	0.3559 (0.2334)	0.3395 (0.0282)	0.2589 (0.0214)
$\hat{\mu}_{c1}$	0.2795 (0.2922)	-0.0437 (0.0326)	-0.0435 (0.0323)
$\hat{\sigma}_{a0}$	0.5324 (0.0083)	0.2780 (0.0019)	0.2780 (0.0019)
$\hat{\sigma}_{a1}$	0.2709 (0.0116)	0.2464 (0.0032)	0.2462 (0.0032)
$\hat{\sigma}_{n0}$	0.2650 (0.0205)	0.2883 (0.0096)	0.4063 (0.0088)
$\hat{\sigma}_{n1}$	0.4653 (0.0219)	0.3779 (0.0080)	0.3779 (0.0080)
$\hat{\sigma}_{c0}$	0.9858 (0.1577)	0.4669 (0.0169)	0.2349 (0.0163)
$\hat{\sigma}_{c1}$	1.4304 (0.3420)	0.5030 (0.0205)	0.5012 (0.0202)
# Obs.	1160	15434	
LogLik.	-2140.4	-20799.4	-20798.0
$d(\hat{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}})$	0.0092	0.0071	0.0087
AR	0.9722	0.9354	0.9284

Stand. err. in parenthesis are calculated by the asymptotic covariance matrices of consistent roots.
$d(\hat{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}) = \sqrt{\sum_t (\hat{\omega}_t - \tilde{\omega}_t)^2}$

The last row of Table 5.3 shows the values of the Allocation Rate (AR) for each solution. We observe the unique solution for Germany obtains a higher AR value compared to those for Austria. This result can be explained by the univocal identification of the consistent solution being feasible when

a good mixtures disentanglement of both the mixtures happens as indicated by the AR values.

Table 5.4 shows the difference in variances for the two mixtures are significantly different from zero for any of the considered roots. These results do not support to continue the likelihood analysis by assuming the homoscedastic conditions and detecting the likelihood root closest to $\tilde{\boldsymbol{\theta}}$ in $\Omega_h^{\tilde{\theta}}$.

Table 5.4. *Estimated difference in variances for the two mixtures from the two-steps procedure restricted to* $\Omega_h^{\tilde{\omega}}$.

	Germany	Austria	
	$\hat{\boldsymbol{\theta}}_{\text{Ger}}$	$\hat{\boldsymbol{\theta}}_{\text{Aus},1} : \mu_{c1} > \mu_{a1}$ $\mu_{n0} < \mu_{c0}$	$\hat{\boldsymbol{\theta}}_{\text{Aus},2} : \mu_{c1} > \mu_{a1}$ $\mu_{n0} > \mu_{c0}$
$\hat{\sigma}_{n0} - \hat{\sigma}_{c0}$	-0.7208 (0.1557)	-0.1786 (0.0214)	0.1714 (0.0201)
$\hat{\sigma}_{a1} - \hat{\sigma}_{c1}$	-1.1595 (0.3403)	-0.2566 (0.0209)	-0.2550 (0.0205)

Standard errors in parenthesis are calculated by the asymptotic covariance matrices of the consistent roots.

Table 5.5 presents the estimated causal effect for each compliance status compared to the estimated causal effect for compliers obtained by applying the IV method under the exclusion restriction (LATE: Local Average Treatment Effect).

Table 5.5. *Estimated causal effects for each compliance status from the two-steps procedure restricted to* $\Omega_h^{\tilde{\omega}}$, *and estimated LATE per country.*

	Germany	Austria	
	$\hat{\boldsymbol{\theta}}_{\text{Ger}}$	$\hat{\boldsymbol{\theta}}_{\text{Aus},1} : \mu_{c1} > \mu_{a1}$ $\mu_{n0} < \mu_{c0}$	$\hat{\boldsymbol{\theta}}_{\text{Aus},2} : \mu_{c1} > \mu_{a1}$ $\mu_{n0} > \mu_{c0}$
$\hat{\mu}_{a1} - \hat{\mu}_{a0}$	-0.0612 (0.0302)	-0.0062 (0.0053)	-0.0063 (0.0053)
$\hat{\mu}_{n1} - \hat{\mu}_{n0}$	+0.1518 (0.0574)	+0.0696 (0.0180)	+0.0289 (0.0194)
$\hat{\mu}_{c1} - \hat{\mu}_{c0}$	-0.0764 (0.3737)	-0.3832 (0.0432)	-0.3024 (0.0387)
LATE	-0.1538 (0.6565)	-0.3006 (0.0720)	

Standard errors in parenthesis are calculated by the asymptotic covariance matrices of consistent roots and IV estimators.

For Germany, the estimated LATE assumes a value of -0.1538 but not significantly different from zero (s.e.: 0.6565). Relaxing the exclusion restriction is not sufficient to obtain a significant compliers average causal effect, but produces significant effects for both the noncompliers types; in particular

we observe a negative effect for always-takers (-0.0612), and a positive effect for never-takers (+0.1518).

The resulting significant effects for noncompliers can be explained by general equilibrium considerations. In a recent remarkable paper Card and Lemieux (2001), using a model with imperfect substitution between similarly educated workers in different cohort of birth, argued that shifts in the college-high school wage gap reflect changes in the relative supply of highly educated workers across cohorts. The authors argued that the increase in wage gap for younger men in U.S.A., U.K. and Canada in the past two decades is due to the rising of relative demand for college educated labor, coupled with the slowdown in the rate of growth of the relative supply of college educated workers. Tables 5.6 and 5.7 confirm these relations for our two countries. Both the estimated mean of log hourly earnings and the estimated mean of the residuals of log hourly earnings differences between high, $(D_i = 0)$, and low, $(D_i = 1)$, educated individuals are indeed greater for the cohort 1930-39, $(Z_i = 1)$, than for the cohort obtained merging 1925-29 and 1940-49 cohorts, $(Z_i = 0)$.

Table 5.6. *Estimated mean log hourly earnings per country, educational level (D_i), and cohort of birth (Z_i).*

Country	Z_i	Num. observ.	$D_i = 0$	$D_i = 1$	Difference
Germany	$Z_i = 1$	491	3.428 (0.044)	2.940 (0.025)	0.488 (0.053)
	$Z_i = 0$	669	3.317 (0.035)	2.984 (0.024)	0.333 (0.045)
Austria	$Z_i = 1$	6214	4.509 (0.124)	4.077 (0.004)	0.432 (0.108)
	$Z_i = 0$	9220	4.467 (0.008)	4.089 (0.003)	0.378 (0.007)

Standard errors in parenthesis.

Table 5.7. *Estimated mean residual of log hourly earnings per country, educational level (D_i), and cohort of birth (Z_i).*

Country	Z_i	Num. observ.	$D_i = 0$	$D_i = 1$	Difference
Germany	$Z_i = 1$	491	0.376 (0.044)	-0.113 (0.025)	0.489 (0.053)
	$Z_i = 0$	669	0.247 (0.035)	-0.087 (0.024)	0.334 (0.045)
Austria	$Z_i = 1$	6214	0.350 (0.012)	-0.077 (0.003)	0.427 (0.010)
	$Z_i = 0$	9220	0.300 (0.007)	-0.074 (0.003)	0.374 (0.007)

Standard errors in parenthesis.

Even if Card and Lemieux's (2001) conclusions do not regard causal relationships but only observed wage gap between cohorts, these general equilibrium considerations can justify the violation of the exclusion restriction in our cases. The lower average education in the 1930-39 cohort, as indicated in Table 5.1, can indeed explain both the positive return to education for never-takers, individuals always high educated under the two different assignments, and the negative return to education for always-takers, individuals always low educated under the two different assignments. Indeed, the exclusion restriction states the instrumental variable has to have only a treatment mediated effect. But given our definition of the variables Z_i and D_i, we know that the different educational levels between cohorts are due only to the compliers behavior. Consequently the value of the instrumental variable, other than providing information regarding the compliers educational choices, also gives information on the relative supplies of differently educated workers in different cohorts. For example considering the individuals born in the 1930-39 period, we know that compliers born in that cohort will be low educated. Therefore, given the invariant educational behaviors of noncompliers, it is reasonable to suppose a decrease in the relative supply of high educated workers compared to the other cohort (1925-29 $\cup$ 1940-49). Consequently it is reasonable to think never-takers would exploit less competitive labor market conditions then increasing their mean outcome, and on the contrary always-takers would experience worse labor market conditions then decreasing their mean outcome.

For Austria, the estimated nonparametric LATE assumes a significantly different from zero value of -0.3006 (s.e.: 0.0720). Relaxing the exclusion restriction produces two nonspurious interior solutions characterized by different orders of the means of the mixture composed by not assigned never-takers and compliers, $\varsigma(D_i = 0, Z_i = 0)$. Indeed, we observe $\hat{\mu}_{n0} < \hat{\mu}_{c0}$ for $\hat{\boldsymbol{\theta}}_{\text{Aus},1}$, and $\hat{\mu}_{n0} > \hat{\mu}_{c0}$ for $\hat{\boldsymbol{\theta}}_{\text{Aus},2}$. Solution $\hat{\boldsymbol{\theta}}_{\text{Aus},1}$ is characterized by a more pronounced significant estimated causal effect for compliers ($\hat{\mu}_{c1} - \hat{\mu}_{c0}$: -0.3832) compared to the LATE, and by a significant positive effect for never-takers ($\hat{\mu}_{n1} - \hat{\mu}_{n0}$: +0.0696). For solution $\hat{\boldsymbol{\theta}}_{\text{Aus},2}$, on the contrary, the estimated compliers average causal effect ($\hat{\mu}_{c1} - \hat{\mu}_{c0}$: -0.3024) is very close to the estimated LATE, and the estimated noncompliers average causal effects are both not significantly different from zero. Then introducing the further restriction for Austria produces equivalent results to estimating the LATE that is based on imposing the exclusion restriction.

The choice of the particular solution depends on both statistical evidence and economic considerations. Solution $\hat{\boldsymbol{\theta}}_{\text{Aus},1}$ obtains slightly better statistical performances concerning log-likelihood (-20799.4 compared to -20798.0), the distance $d(\hat{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}})$ (0.0071 compared to 0.0087), and the AR value (0.9354 compared to 0.9284). Both the two solutions for Austria (like the unique interior solution for Germany) present a plausible order of mean of the mixture composed by assigned always-takers and compliers, $\varsigma(D_i = 1, Z_i = 1)$. Indeed, compliers can be considered more motivated and able compared to always-takers, individuals never educated from a counterfactual point of view. It is then reasonable to think that the outcome mean for compliers is greater than the outcome mean for always-takers in the relevant mixture. Choice in the order of means in the other mixture is more problematic; compliers can be again considered at least more motivated individuals. But never-takers are always high educated under the two different assignments, so presumably in better social conditions and then exploiting more advantages and opportunities in the labor market. For these reasons the choice of the sign for the difference $(\mu_{c0} - \mu_{n0})$ is more questionable, and depends on a deeper and more specific analysis of the Austrian social-economical context of this period. Anyway, the two interior solutions for Austria share a not significant effect for always-takers, and a negative remarkable effect for compliers.

6 Conclusions

Identification and estimation issues in analyzing a randomized experiment with imperfect compliance without exclusion restriction have been considered. The main difficulties in this task are due to the presence of mixtures of distributions that implies both the partial identifiability of the models and the possibility to have multiple roots for the likelihood equations.

Supposing the outcome distributions of various compliance statuses are in the same parametric class, the model is identifiable unless the equality of at least two of the mixing probabilities: $\omega_a = \omega_c$, or $\omega_n = \omega_c$, or $\omega_a = \omega_n = \omega_c$. This is a set of less restrictive conditions compared to simple mixture models where identifiability is assured only up to permutations of the label components. Furthermore this set of equality conditions for the mixing probabilities are easily testable given the usual assumptions for identifying causal effects by the IV method.

Supposing normally distributed outcomes and taking into account both

the possibility to have multiple roots and the unboundedness of the likelihood, statistical theory guarantees the efficient estimate can be detect by the root closest to the method of moments estimate of the parameter vector. However an unique method of moments estimate can be obtained only by imposing the homoscedastic conditions for the two mixtures components. In the heteroscedastic case the detection can be restricted to the root closest to the method of moments estimate of the mixing probabilities. A simulation based analysis proves the detection of the efficient likelihood estimate is feasible when a good mixtures disentaglement of both the mixtures happens. For computational purposes and for exploiting the particular incomplete structure of the likelihood an EM algorithm can be easily developed.

An empirical microeconomic example has also been proposed. Supposing normal distributions for the outcome, we estimate the noncompliers cohort of birth effects on earnings (other than the compliers average causal effect) for individuals born in Germany and Austria between 1925 and 1949. The microeconomic context has been suggested by a recent paper of Ichino and Winter-Ebmer (2004).

7 Appendix A

If $(\hat{\omega}_a, \hat{\omega}_c, \hat{\boldsymbol{\eta}}_{a1}, \hat{\boldsymbol{\eta}}_{c1})$ is one of the multiple roots for the likelihood equations based only on the units $i \in \varsigma(D_i = 1,\ Z_i = 1)$ then

$$\partial \sum_{i \in \varsigma(D_i=1,\, Z_i=1)} \log\, f(y_i,\, d_i,\, z_i; \boldsymbol{\theta}) / \partial(\boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{c1}) \Big|_{\boldsymbol{\eta}_{a1}=\hat{\boldsymbol{\eta}}_{a1},\, \boldsymbol{\eta}_{c1}=\hat{\boldsymbol{\eta}}_{c1}} = 0,$$

where $f(y_i,\, d_i,\, z_i; \boldsymbol{\theta})$ is in the parametric class (1).

A root of the likelihood equations based on the entire sample satisfies $\partial \sum_i \log\, f(y_i,\, d_i,\, z_i; \boldsymbol{\theta}) / \partial(\boldsymbol{\theta}) = 0$

$$\partial \sum_{i \notin \varsigma(D_i=1,\, Z_i=1)} \log\, f(y_i,\, d_i,\, z_i; \boldsymbol{\theta}) + \sum_{i \in \varsigma(D_i=1,\, Z_i=1)} \log\, f(y_i,\, d_i,\, z_i; \boldsymbol{\theta}) / \partial(\boldsymbol{\theta}) = 0;$$

this implies

$$\partial \sum_{i \notin \varsigma(D_i=1,\, Z_i=1)} \log\, f(y_i,\, d_i,\, z_i;\, \boldsymbol{\theta}) / \partial(\pi, \omega_a, \omega_n, \omega_c, \boldsymbol{\eta}_{a0}, \boldsymbol{\eta}_{n0}, \boldsymbol{\eta}_{n1}, \boldsymbol{\eta}_{c0}) = 0$$

$$\partial \sum_{i \in \varsigma(D_i=1,\, Z_i=1)} \log\, f(y_i,\, d_i,\, z_i;\, \boldsymbol{\theta}) / \partial(\pi, \omega_a, \omega_c) = 0$$

$$\partial \sum_{i \in \varsigma(D_i=1,\, Z_i=1)} \log f(y_i,\, d_i,\, z_i\,;\, \boldsymbol{\theta}) / \partial(\boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{c1}) = 0$$

Consequently $(\hat{\boldsymbol{\eta}}_{a1}, \hat{\boldsymbol{\eta}}_{c1})$ is also a sub-vector of a root of the likelihood equations based on the entire sample. Analogous arguments hold for a root of the likelihood equations based only on the units $i \in \varsigma(D_i = 0,\, Z_i = 0)$.

8 Appendix B

Let's define the set $\mathcal{S}(\mathbf{y})$ as:

$$\mathcal{S}(\mathbf{y}) = \left\{ \boldsymbol{\theta} \in \bar{\Theta} | \exists\, tz \in \{a1, c1, n0, c0\}\,,\, n \in \{1, ..., N\}\,,\, \mu_{tz} = y_n,\, \sigma_{tz} = 0 \right\}$$

where $\bar{\Theta}$ is the closure of Θ.

Proposition 3 *For any i.i.d. sample* $(\mathbf{y}, \mathbf{d}, \mathbf{z})$ *of* N *units, the likelihood function* $L(\boldsymbol{\theta})$ *degenerates at every point of* $\mathcal{S}(\mathbf{y})$*:*

$\forall \mathbf{y},\, \forall \boldsymbol{\theta}^* \in \mathcal{S}(\mathbf{y}),\, \exists \left(\boldsymbol{\theta}^{(k)} \in \Theta,\, k = 1, 2, ... \right)$ such that $\lim_{k \to \infty} \boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^*$ and $\lim_{k \to \infty} L(\boldsymbol{\theta}) = \infty$.

Proof: suppose that $\sigma_{a1} = 0$ or $\sigma_{c1} = 0$ in $\boldsymbol{\theta}^*$. The likelihood can be written:

$$L(\boldsymbol{\theta}) = \prod_i f(y_i,\, d_i,\, z_i\,;\boldsymbol{\theta}) = \prod_{i \in \varsigma(D_i=1,\, Z_i=1)} f(y_i,\, d_i,\, z_i\,;\pi, \omega_a, \omega_c, \boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{c1}) \cdot$$

$$\cdot \prod_{i \notin \varsigma(D_i=1,\, Z_i=1)} f(y_i,\, d_i,\, z_i\,;\boldsymbol{\theta} \backslash \boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{c1}) = L_1(\pi, \omega_a, \omega_c, \boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{c1}) \cdot L_2(\boldsymbol{\theta} \backslash \boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{c1})$$

where the first factor of $L(\boldsymbol{\theta})$ is the likelihood for a mixture of two normal distributions:

$$L_1(\boldsymbol{\theta}) = \prod_{i \notin \varsigma(D_i=1,\, Z_i=1)} \left[\omega_a \cdot N\left(y_i;\, \mu_{a1}, \sigma_{a1}^2\right) + \omega_c \cdot N\left(y_i;\, \mu_{c1}, \sigma_{c1}^2\right) \right].$$

This factor degenerates if $\sigma_{a1} \to 0$ and $\mu_{a1} \to y_n$, or if $\sigma_{c1} \to 0$ and $\mu_{c1} \to y_n$, Day(1969). Given $L_2(\boldsymbol{\theta} \backslash \boldsymbol{\eta}_{a1}, \boldsymbol{\eta}_{c1})$ does not depends on σ_{a1} and σ_{c1}, this implies the degeneracy of the overall $L(\boldsymbol{\theta})$. Analogous arguments hold if $\sigma_{n0} = 0$ or $\sigma_{c0} = 0$ in $\boldsymbol{\theta}^*$•

9 References

Angrist J.D. (1990); Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records; *American Economic Review,* **80**, 313-335.

Angrist J.D., A.B. Krueger (1991); Does compulsory schooling attendance affect schooling and earnings?*; Quarterly Journal of Economics,* **106**, 979-1014.

Angrist J.D., G.W. Imbens, D.B. Rubin (1996); Identification of causal effects using instrumental variables; *J.A.S.A.*, **91**, 444-455.

Balke A., J. Pearl (1997); Bounds of treatment effects from studies with imperfect compliance; *J.A.S.A.*, **92**, 1171-1176.

Becker S.O., F. Siebern-Thomas (2004); Supply of schools, educational attainment and earnings; http://www.lrz-muenchen.de/~sobecker/returns.pdf.

Card D. (1995); Earnings, schooling, and ability revisited; *Research in labor economics,* **14**, 23-48.

Card D. (1999); The causal effect of education on earnings; in Ashenfelter O. and D. Card eds., *Handbook of Labour Economics,* Vol. 3A, Elsevier Science, North-Holland, 1801-1863.

Day N.E. (1969); Estimating the components of a mixture of normal distributions; *Biometrika,* **56**, 463-474.

Denny K.J., C.P. Harmon (2000); Education policy reform and the return to schooling from instrumental variables; *Working Paper 00/07, The Institute for Fiscal Studies,* London.

Deschenes O. (2002); Estimating the effects of family background on the return to schooling; *Working Paper 10-02, Dep. of Economics University of California, Santa Barbara.*

Fowlkes E.B. (1979); Some methods for studying the mixture of two normal (lognormal) distributions; *J.A.S.A.*, **74**, 561-575.

Gan L., J. Jiang (1999); A test for a global maximum; *J.A.S.A.*, **94**, 847-854.

Hataway R.J. (1985); A constrained formulation of maximum-likelihood estimation for normal mixture distributions; *The Annals of Statistics,* **13**, 795-800.

Hataway R.J. (1986); A constrained EM algorithm for univariate normal mixtures; *J. Statist. Comput. Simul.,* **23**, 211-230.

Heckmann J., R. Robb (1985); Alternative methods for evaluating the impact of interventions; in *Longitudinal Analysis of Labor Markets Data* (J. Heckmann and B. Singer, eds.). Cambridge University Press.

Hirano K., G.W. Imbens, D.B. Rubin, X. Zhou (2000); Assessing the effect of an influenza vaccine in an encouragement design; *Biostatistics,* **1**, 69-88.

Holland (1986); Statistics and Causal Inference; *J.A.S.A.,* **81**, 945-970.

Ichino A., R. Winter-Ebmer (2004); The long run educational cost of World War II; *Journal of Labor Economics,* **22**, 57-86.

Ichino A., R. Winter-Ebmer (1999); Lower and upper bounds of returns to schooling: an exercise in IV estimation with different instruments; *European Economic Review,* **43**, 889-901.

Imbens G.W., D.B. Rubin (1997a); Bayesian inference for causal effects in randomized experiments with noncompliance; *The Annals of Statistics,* **25**, 305-327.

Imbens G.W., D.B. Rubin (1997b); Estimating outcome distributions for compliers in instrumental variable models; *Review of Economic Studies,* **64**, 555-574.

Jo B. (2002); Estimation of intervention effects with non compliance: alternative model specifications; *Journal of Educational and Behavioral Statistics,* **27**, 385-409.

Kane T.J., C.E. Rouse (1993); Labor market returns to two and four-years colleges: is a credit and do degrees matter?; *Princeton University Industrial Relations Section, Working Paper n.311.*

Kiefer N.M. (1978); Discrete parameter variation: efficient estimation of a switching regression model; *Econometrica,* **46**, 427-434.

Kling J.R. (2001); Interpreting instrumental variables estimates of the returns to schooling; *Journal of Business and Economic Statistics,* **19**, 358-364.

Lehmann E.L., G. Casella (1998); Theory of point estimation; *Springer.*

Li L.A., N. Sedransk (1985); Mixtures of distributions: a topological approach; *The Annals of Statistics,* **16**, 1623-1634.

Lindsay B.G., P. Basak (1993); Multivariate normal mixtures: a fast consistent metohod of moments; *J.A.S.A.,* **88**, 468-476.

Little R.J.A., L.H.Y. Yau (1998); Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model; *Psychological Methods, 3*, 147-159.

Manski C.F. (1990); Nonparametric bounds on treatment effects; *American Economic Review, Papers and Proceedings,* **80**, 319-323.

McLachlan G.J., K.E. Basford (1988); Mixture models, inference and applications to clustering; *Marcel Dekker, Inc.*

McLachlan G.J., D. Peel (2000); Finite mixture models; *John Wiley and Sons, Inc.*

Mercatanti A. (2005); A constrained likelihood maximization for relaxing the exclusion restriction in causal inference; *Report n.263, Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa.*

Quandt R.E., J.B. Ramsey (1978); Estimating mixtures of normal distributions and switching regressions; *J.A.S.A.,* **73**, 730-738.

Redner R.A., H.F. Walker (1984); Mixture densities, maximum likelihood, and the EM algorithm; *SIAM Rev.*, **26**, 195-239.

Ridolfi A., J. Idier (2002); Penalized maximum likelihood estimation for normal mixture distributions; *EPFL, School of Computer and Information Sciences, Tec. Report 200285.*

Teicher H. (1963); Identifiability of mixtures; *Annals of Mathematical Statistics*, **31**, 244-248.

Teicher H. (1968); Identifiability of finite mixtures; *Annals of Mathematical Statistics*, **34**, 1265-1269.

Titterington D.M., A.F.M. Smith, U.E. Makov (1985); Statistical analysis of finite mixture distributions; *John Wiley and Sons, Inc.*

Wald A. (1949); Note on the consistency of the maximum likelihood estimate; *Annals of Mathematical Statistics,* **20**, 595-601.

Yakowitz S.J., J.D. Spragins (1968); On the identifiability of finite mixtures; *Annals of Mathematical Statistics,* **39**, 209-214.