


Finite Mixture Models  
Some computational and theoretical developments  
with applications

Paolo Frumento





Università degli Studi di Firenze  
Dipartimento di Statistica “G. Parenti”  
Dottorato di Ricerca in Statistica Applicata  
XXI ciclo

Finite Mixture Models  
Some computational and theoretical developments  
with applications

Paolo Frumento

Tutor: Prof.ssa Fabrizia Mealli

Co-Tutor: Prof.ssa Barbara Pacini

Coordinatore: Prof. Guido Ferrari

*Settore scientifico-disciplinare: SECS-S/01– Statistica*



# Table of Contents

Preface	7
1 General introduction to Finite Mixture Models	11
1.1 Finite Mixture Models: origin and interpretation	11
1.2 Issues in Finite Mixture Models	13
1.3 Estimation of Finite Mixture Models via the EM algorithm	14
2 Optimizing the log-likelihood function	16
2.1 Local maxima, spurious maxima and saddle points	16
2.2 Finite Mixture Models and Genetic Algorithms	21
3 Multivariate models	25
3.1 Mixtures of multivariate Normal distributions	25
3.2 A comparison between multivariate and univariate mixture models	29
4 The <code>mixglm</code> package for the R environment	35
5 Analysis of causal effects of job-training programs	62
5.1 The Rubin Causal Model	63
5.2 Noncompliance in randomized studies	66
5.3 Missing outcomes	72
5.3.1 Analysis of randomized experiments with noncompliance and missing outcomes	73
5.4 Outcomes truncated by death	74
5.4.1 Estimating the causal effect of job-training programs on wages	76
5.5 Estimating the effect on wages with noncompliance and missing outcomes under the MAR assumption	77
5.6 Likelihood approach	81
5.7 Application to the Job Corps Study	88
5.8 Results	90

Concluding remarks	97
Appendix A – Parameters estimates	99
Appendix B – Baseline characteristics	112
References	119

## Preface

Finite Mixture Models have gained an increased popularity in many fields of sciences; the main feature of this class of models is that commonly used density functions are employed as building blocks for more complex distributions: this allows for a great flexibility in statistical modeling and makes finite mixtures adequate to very complicate frameworks.

Very often, the rationale of fitting a mixture model is the presence of  $k$  unobserved subpopulations; however, there are many examples in which the components of a mixture lack any physical sense and may be viewed as latent clusters – according to some meaningful classification. Furthermore, even when there is no reason to believe that a latent structure affects the data-generating process, a mixture model can be fitted with the aim of exploiting its flexibility; as pointed out in McLachlan and Peel (2000), this approach is a good compromise between a fully parametric model ( $k = 1$ ) and a completely nonparametric one: together with a great flexibility, a finite mixture model retains some of the advantages of the parametric approaches, keeping a moderate number of parameters and allowing for a simple interpretation of the estimated component densities.

A finite mixture model may be applied with very different purposes. As pointed out by some authors (see, e.g., Lindsay 1995), a mixture model has a dual usefulness: on the one hand, it enables to study the distribution of the outcome variable ( $Y$ ) when a covariate (the component membership  $Z$ ) is missing; on the other hand, it makes use of a surrogate measure ( $Y$ ) to learn about an unobserved variable  $Z$ ; from this viewpoint, fitting a mixture model can be a valid alternative to using standard clustering methods.

Unfortunately, estimating the parameters of a finite mixture model presents a number of obstacles: first, model identification is not guaranteed; second, estimates are sensitive to the starting values used for the optimization algorithm. In this work, both issues are discussed under a theoretical and computational point of view. The critical points of the log-likelihood function are classified into three main categories (spurious/local optimizers and saddle points), according to their nature. The presence of saddle points has a great relevance in model identification; furthermore, the task of finding the true MLE is complicated by the existence of local maximizers. With a simulation study, we illustrate how difficult it may be to estimate the parameters of a mixture model and we investigate how a Genetic Algorithm may be used for this optimization problem: the choice of the operational parameters is discussed, with a special attention to the trade-off between the effectiveness of the algorithm and the computational effort.

Another relevant topic is constituted by mixtures of multivariate distributions: after a brief presentation of Multivariate Normal mixtures, we explain how the simultaneous modeling of more than one outcome variable may improve the model identification.

In this work, we have also developed and presented a new R package `mixglm` (forthcoming) for fitting finite mixtures of Normal/Poisson/Binomial distributions via the EM algorithm. With respect to other software, `mixglm` allows for a greater flexibility in the model specification; covariates may affect both the mixing distributions and the mixing proportions; any parameter is allowed to vary or to be constant across components, or to be an offset; the package also handles mixtures with partially classified observations. In `mixglm`,

multivariate models are implemented: since the joint distribution is specified as the product of marginal and conditional densities, the outcomes are allowed to belong to different parametric families. Optionally, the starting values for the EM algorithm are provided by a genetic algorithm; the standard errors of the estimates may be computed using the asymptotic covariance matrix (with analytical evaluation of the Hessian of the log-likelihood function) or with a bootstrap approach (parametric or nonparametric); a function for bootstrap-based selection of the optimal number of components is provided; fitted values and conditional membership probabilities are also available.

With some modifications, `mixglm` has been applied to the evaluation of the effects of a randomized job-training program, Job Corps, which stands out as the largest, most comprehensive US education and job training program for disadvantaged youths between the ages of 16 and 24; for our analysis, we use data from the National Job Corps Study, conducted by Mathematica Policy Research, Inc. The study is based on a national random sample of all eligible applicants in late 1994 and 1995. Sampled youths were assigned randomly to a program group or a control group; consistently with the program's aim, key outcomes of interest are: employment status, total earnings, and wages.

In this work, we adopt the general framework of the Rubin Causal Model (Holland, 1986), where, in the case of a binary treatment, for each unit two potential outcomes are defined – one if treated and one if not treated; the causal treatment effect is defined as a comparison of the two potential outcomes. This configures the causal inference as a missing data problem, since only one outcome – corresponding to the actual treatment assignment – is observed for each unit; however, randomization ensures that the sample means are unbiased estimates of expected outcomes in the two groups; as a consequence, the average treatment effect is estimated in a straightforward way. Very often, complications arise and this simple framework must be adapted to more complex settings.

In the study, three complications are present, namely a) compliance with assigned treatment was not perfect, as only 64% of those assigned to the program group effectively enrolled in Job Corps; b) due to attrition, outcome is missing on some participants; c) wages are truncated by death, meaning no wage is defined for those who are not employed. Using the principal stratification approach (Frangakis and Rubin, 2002), we define the average treatment effect of interest as the expected difference between the two potential wages among the compliers-always-employed (units who would comply with the treatment assignment and employed in both treatment and control condition).

Using a likelihood approach, we propose a log-Normal model for wages; we assume that data are missing at random (MAR; Rubin, 1976); both the potential outcomes and the mixing proportions are supposed to depend on pre-treatment covariates. Using the EM algorithm, we estimate the treatment effect on employment and wages for compliers with and without assuming monotonicity of truncation.

The thesis proceeds as follows. In Chapter 1, we present the general framework of finite mixture models, with a brief account of the main issues in estimation and inference; a simple description of the basic EM algorithm is provided. Chapter 2 is devoted to the special issue of the local/spurious optimizers and of the saddle points in the likelihood function; a simple genetic algorithm is presented and applied to a simulated data set. In Chapter 3, mixtures of multivariate distributions are discussed, with a special attention to



the multivariate Normal model; with a simulation study, we illustrate the advantages of using a multivariate mixture. Chapter 4 presents the `mixglm` package. In Chapter 5, we discuss the general approach to the causal inference in presence of noncompliance, missing outcomes and truncation by death; under a likelihood approach, we estimate the average treatment effect on employment and wages for the Job Corps Study. In Chapter 6, results are discussed and concluding remarks are provided, together with some suggestions for further developments.



# 1 General Introduction to Finite Mixture Models

## 1.1 Finite Mixture Models: origin and interpretation

Finite (nonparametric) mixture models represent an advanced and flexible tool in statistical modeling; this class of models have gained increased popularity over the last decades across many fields.

In order to define a mixture model let us present the following scenario. Assume that the random sample  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$  has components having different distributions according to some categorical variable  $Z = (Z_1, \dots, Z_n)$  taking on the values  $z_1, \dots, z_k$  with probabilities  $p_1, \dots, p_k$ , respectively; that is, the conditional density of  $\mathbf{Y}_i$  given  $Z_i = z_j$  is  $f_j(\mathbf{y}_i)$  ( $i = 1, \dots, n$ ,  $j = 1, \dots, k$ ). If in the observed sample  $\mathbf{y}_1, \dots, \mathbf{y}_n$  the subpopulation the observations are coming from is not known, the only distribution we can directly observe is the marginal density:

$$f(\mathbf{y} | \Phi) = f_1(\mathbf{y} | \lambda_1)p_1 + \dots + f_k(\mathbf{y} | \lambda_k)p_k$$

where  $p_1 + \dots + p_k = 1$  and  $\Phi = (p_1, \dots, p_k, \lambda_1, \dots, \lambda_k)$  is the parameters vector of the mixture model. In the above formula,  $f_j$  is the distribution of  $\mathbf{Y}$  when  $Z = z_j$  and  $\lambda_j$  is the related parameters vector; the  $f_j(\cdot)$  are called component densities and the quantities  $p_j = P(Z = z_j)$  are the respective mixing proportions. According to the standard practice, without loss of generality we set  $z_j = j$ ; in most cases, the parametric family is independent of  $Z$ , and we can also suppose  $f_j = f$ .<sup>(1)</sup>

In some settings, the latent variable  $Z$  contains the labels of  $k$  subpopulations that are known a priori to exist; however, there are many examples in which the components lack any physical sense and  $Z$  may be interpreted as a cluster membership, according to some meaningful classification. Furthermore, even when we have no reason to believe that the observed data-generating process is of the form described above, when a mono-component model is unsatisfactory for a given data set, a mixture model can be fitted – with the aim of exploiting its flexibility. As a consequence,  $Z$  will in this case lose any meaning.<sup>(2)</sup>

The latent group-label  $Z$  can be unobserved for a number of reasons: depending on the context,  $Z$  can represent an unobservable quantity or – very often – a variable for which is very hard (or too expensive, in terms of time/money consuming) to obtain a direct meas-

---

<sup>(1)</sup> Mixtures of different families are an interesting topic; see for example the so called “minefield” data set (Dasgupta and Raftery, 1998, Fraley and Raftery, 1998) where the marginal distribution is specified as a mixture of  $g$  bivariate Normal densities plus a Uniform component (that is, a spatial Poisson process) in order to capture a background noise.

<sup>(2)</sup> As pointed out in McLachlan and Peel, 2000, this approach is a good compromise between a fully parametric model ( $k = 1$ ) and a completely nonparametric one (that is a kernel estimate, where  $k = n$  and  $p_j = 1/n$  for each  $j$ ): together with a great flexibility, a finite mixture model retains some of the advantages of the parametric approaches – keeping a moderate number of parameters and allowing for a simple interpretation of the estimated component densities.

urement.<sup>(3)</sup> The number of components ( $k$ ) may be known or unknown: if unknown, it can be treated as an estimand parameter; in many cases,  $k$  is chosen with an ex-post model selection (see Section 1.2 for some details).

It is straightforward to extend the general model to more complicate settings. If a set of observed covariates ( $\mathbf{X}$ ) affects the  $\lambda_j$ , a generalized linear model may be specified for the component densities, according to an appropriate link function. Optionally, a set  $\mathbf{X}_p$  of covariates (possibly overlapping with  $\mathbf{X}$ ) may affect the distribution of the unobserved  $Z$ , with  $p_j = p_j(\mathbf{X}_p)$ : a multinomial (logit, probit etc.) or some non- or semi-parametric model is fitted in this case.

The output of a mixture model is an estimate of  $\Phi$  and an ex-post membership probability for each observation  $\mathbf{y}_i$  ( $i = 1, \dots, n$ ), obtained by Bayes' Theorem: if the component densities  $f_1, \dots, f_k$  are not “too close”, the model will provide a probabilistic clustering of the observed  $\mathbf{Y}$ .

In Figures 1.1-1.2 the shapes of different Gaussian mixtures are displayed, with and without inclusion of covariates for the expected values and for the mixing proportions; different values of the location and scale parameters lead to a great variety of marginal distributions.

Increased spread and popularity of this class of models are due to their wide applicability and their great flexibility. Unfortunately, learning about mixture models can be a very hard task. The next section is devoted to a brief discussion of the main issues in estimation and inference; Section 1.3 concludes this general introduction describing the EM algorithm.

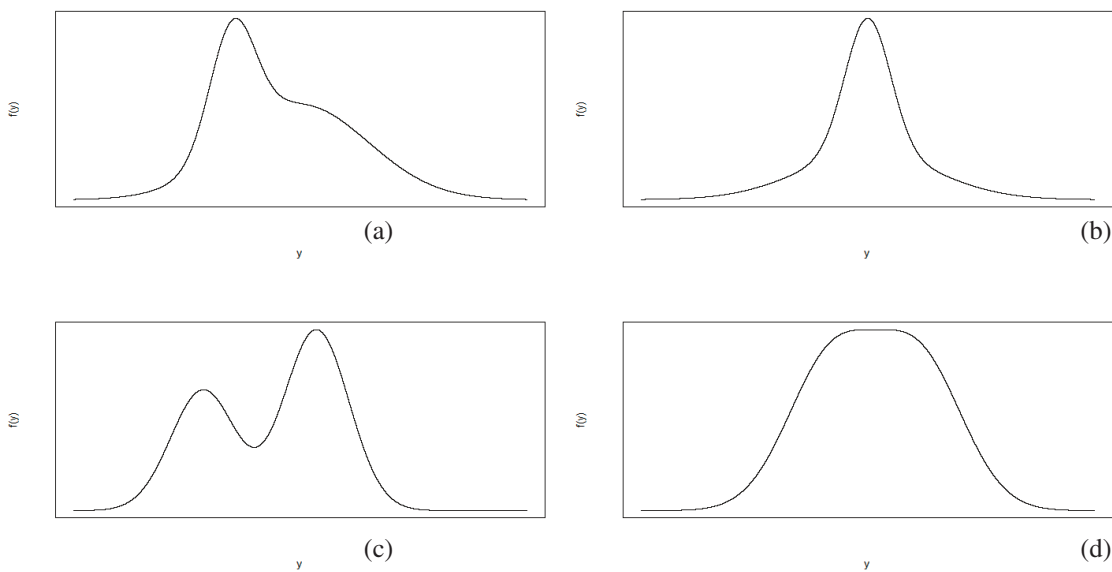


Figure 1.1 *Plots of 2-components Normal mixture densities. The mixing distributions in (a) have different location and scale parameters; in (b) the location parameter is the same for both densities, whereas in (c) and (d) a common scale is used.*

<sup>(3)</sup> It is very important to point out the dual usefulness of a mixture model: on the one hand, it enables to study the distribution of  $\mathbf{Y}$  when a covariate ( $Z$ ) is missing; on the other hand, we can use a surrogate measure ( $\mathbf{Y}$ ) to learn about an unobserved variable  $Z$  (see for example MacDonald and Pitcher, 1979).

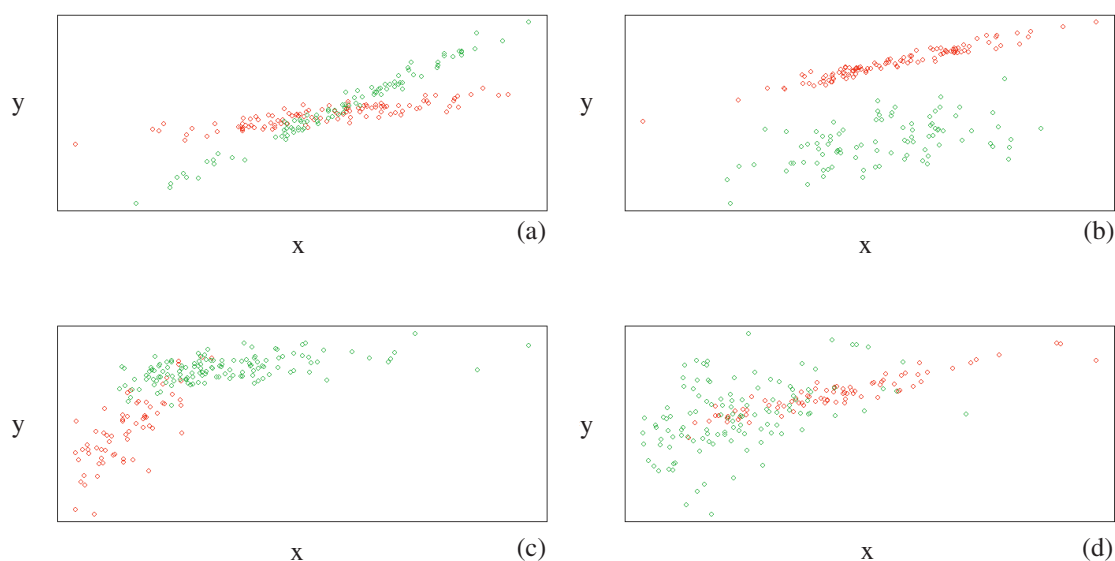


Figure 1.2 *Random samples from different 2-components mixtures of linear regressions with normally distributed errors (different colors denote the true membership of each observation). In (a) and (c) different slopes and intercepts are used, with common variance; in (b) we have different intercepts and variances, with common slope; the scatter in (d) comes from a heteroschedastic model on a common linear trend. In (c) and (d) the mixing proportions vary across the values of  $x$ , according to a logistic model.*

## 1.2 Issues in Finite Mixture Models

In this section, we shortly review the main problems arising in the estimation of mixture models and we focus on some relevant topics in inference.

A great care is needed in the specification and estimation of a mixture model: it is not guaranteed that the parameter space lies in an identifiable set.<sup>(4)</sup> Even when identifiability conditions hold, the identification may be weak: we pay the flexibility of this class of models with an unpredictable and multimodal likelihood surface (posterior distribution in a Bayesian analysis). As a consequence, very different parameters estimates may be obtained according to the starting points for the optimization algorithm; in addition, in models admitting a degenerate distribution in an arbitrary mass point (e.g., the Normal model) the log-likelihood function is unbounded. For the same reasons, the estimates are extremely sensitive to model misspecifications. In Chapter 2, we deepen these aspects from a theoretical and computational point of view; the shape of the log-likelihood function is investigated and some solving strategy is proposed.

Another very important issue is that most algorithms (e.g., the standard EM) assume a known number of components: more complex procedures allow for  $k$  to be an estimand

<sup>(4)</sup>A finite mixture model is unidentifiable when infinitely many parameters vectors lead to the same mixed distribution; to make an example, we cannot estimate a mixture of two or more Bernoulli distributions, without any additional assumptions. An accurate account of this issue is in Titterington et al., 1985.

parameter;<sup>(5)</sup> alternatively, we can choose different values of  $k$  and perform an ex-post model selection ( $k_1$  vs  $k_2$  components) using the standard selection criteria (BIC, AIC, CLC, EIC, LEC etc.) or bootstrapping the LRT statistic, whose asymptotic distribution is generally unknown.<sup>(6)</sup>

Finally, once the model structure has been chosen and a maximum of the log-likelihood function has been found, we have to face the problem of making inference about the model parameters. Also in very simple settings, no closed forms are available for the sampling variances of the estimators, and the Hessian matrix is generally hard to derive; in addition, even with a quite large sample size, the distribution of the estimators may be heavily skewed; a bootstrap approach is preferable in this case.

We conclude this chapter with a brief presentation of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

### 1.3 Estimation of Finite Mixture Models via the EM algorithm

The simplest and most common optimization procedure for the estimation of finite mixture models is the EM (Expectation-Maximization) algorithm (Dempster, Laird and Rubin, 1977). In what follows, we provide a brief description of the algorithm, exploiting the traditional formulation of a mixture model as an incomplete-data structure; for further details, see for example MacLachlan and Peel, 2000.

Given  $n$  independent observations  $\mathbf{y}_1, \dots, \mathbf{y}_n$ , the likelihood and the log-likelihood functions for a  $k$ -components finite mixture model can be written as:

$$L(\Phi) = \prod_{i=1}^n \sum_{j=1}^k f(\mathbf{y}_i | \lambda_j) p_j$$

and

$$l(\Phi) = \sum_{i=1}^n \log \sum_{j=1}^k f(\mathbf{y}_i | \lambda_j) p_j$$

respectively. Let us define the new variable  $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)$  as follows:

$$Z_{ij} = (\mathbf{Z}_i)_j = 1 \quad \text{if } Z_i = j \quad (i = 1, \dots, n; j = 1, \dots, k)$$

That is,  $\mathbf{Z}_i$  represents – for each unit – the indicator function of the cluster membership: we replace the single categorical variable  $Z_i$  with a  $k$ -dimensional component-label vector

<sup>(5)</sup> We can mention the reversible jump (Green, 1995), the IPRA (Iterative Pairwise Replacement Algorithm, Scott and Szewczyk, 2001), the Greedy EM (Vlassis and Likas, 2002), the SMEM (Split-and-Merge EM, Ueda et al., 2000), the Group Membership Function Method (Yang and Liu, 2002).

<sup>(6)</sup> Titterington et al. (1985), Böhning (2000) showed the asymptotic result  $LRT \sim 1/2 \chi_{(0)}^2 + 1/2 \chi_{(d)}^2$  for some simple setting, where  $d = k_2 - k_1$ ; in a more general context, the limiting distribution is a mixture of chi-square densities with unknown mixing proportions (Lindsay, 1995). See for example McLachlan and Peel (2000) for a comparison between different approaches to the model selection.

$\mathbf{Z}_i$ , whose distribution is thus Multinomial(1,  $\mathbf{p}$ ) with  $\mathbf{p} = (p_1, \dots, p_k)$ . Clearly, being  $\mathbf{Z}$  unobserved, also  $\mathbf{Z}$  are latent variables; if  $\mathbf{Z}$  were observed, we could write the complete likelihood and log-likelihood as

$$L_c(\Phi) = \prod_{i=1}^n \prod_{j=1}^k f(\mathbf{y}_i | \lambda_j)^{z_{ij}} p_j^{z_{ij}}$$

and

$$l_c(\Phi) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log f(\mathbf{y}_i | \lambda_j) + \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log(p_j)$$

respectively.

The E-step of the EM algorithm consists in finding an estimate of the unobserved variable  $\mathbf{Z}_i$  for each observation, given the actual parameters vector  $\Phi^{(t)}$ ; by Bayes' rule, we obtain:

$$E(Z_{ij} | \Phi^{(t)}) = P(Z_{ij} = 1 | \Phi^{(t)}) = \frac{f(\mathbf{y}_i | \lambda_j) p_j}{\sum_{h=1}^k f(\mathbf{y}_i | \lambda_h) p_h} = e_{ij}$$

The complete expected log-likelihood is thus:

$$l_e(\Phi) = \sum_{i=1}^n \sum_{j=1}^k e_{ij} \log f(\mathbf{y}_i | \lambda_j) + \sum_{i=1}^n \sum_{j=1}^k e_{ij} \log(p_j)$$

Performing the M-step, that is maximizing  $l_e(\Phi)$ , is straightforward: the first term of the above expression is, for each component ( $j = 1, \dots, k$ ), the weighted log-likelihood of a pure (i.e., mono-component) model for  $\mathbf{y}_i$ , with parameter  $\lambda_j$  and weights  $e_{ij}$ ; the second term is the log-likelihood of a multinomial model for  $e_{ij}$  (estimating the unobserved  $z_{ij}$ ) with parameters  $p_1, \dots, p_k$ . In a more general setting, a regression model can be chosen for  $\mathbf{Y}$  and/or for  $\mathbf{Z}$ : a traditional GLM software can be used in the fitting procedure. The EM algorithm can be summarized as follows:

- step 0: choose a starting vector  $\Phi^{(0)}$ ;
- step 1 (E-step): compute the  $e_{ij}$  given the actual estimates  $\Phi^{(t)}$ ;
- step 2 (M-step): update the parameters vector  $\Phi^{(t)}$  maximizing the expected log-likelihood;
- step 3: go back to the step 1.

The rationale of the EM algorithm is very easy to understand: intuitively, in the E-step we try to assess the component membership of each unit; using this information to weigh the observations, we perform a different maximization for each component of the mixture. This loop is carried on until the stopping condition has been reached. As showed in Dempster et al., the log-likelihood function is not decreasing after an EM iteration. We refer to the final estimate of  $\Phi$  as “nonparametric maximum likelihood estimator” (NPMLE; Böhning, 2000).

## 2 Optimizing the log-likelihood function

In this chapter, we present a very important issue arising in mixture models; we handle the problem of finding the *best* estimate among all the local maxima in the log-likelihood function.<sup>(7)</sup> Many authors (see for example MacLachlan and Peel, 2000) tackle the issue of the local maxima in the log-likelihood function for finite mixtures; with the aid of some simulated data, we illustrate how this question may be relevant in practice.

In the sequel, we illustrate how the solutions of the log-likelihood function may be classified – according to their nature – into three main categories (Section 2.1); then (Section 2.2), we assess how a genetic algorithm performs in searching for the MLE of a mixture model.

### 2.1 Local maxima, spurious maxima and saddle points

When working with finite mixture models, a great number of different estimates may be obtained for the same data, depending on the starting points for the optimization algorithm: believing in the uniqueness of the data-generating process, we have the aim of finding the *best* model for a given data set.

The nature of a stationary point in the log-likelihood function may be disparate; the main partitioning is between spurious maxima, local maxima and saddle points. After a brief account on the well known question of the local and spurious optimizers, we will focus on the latter issue, rarely undertaken in the statistical literature.

#### *Local maxima and spurious solutions*

In a very general way, a local optimizer represents a sub-optimal root of the score function; among all roots, the estimate with the highest value of the log-likelihood may be regarded as the *best* one. However, this very simple decision rule is not always directly applicable. In normal models with unconstrained variances, for example, the log-likelihood function is known to be unbounded: when a component with zero variance and mean equal to an arbitrary data point is fitted, the observed log-likelihood becomes infinite and a singular covariance matrix is obtained. We speak in this case of *spurious* optimizers. In practice, there often exist other solutions which may be regarded as spurious, lying very close to the edge of the parameter space: this happens when a component with very small variance (generalized variance in the multivariate case) is fitted; usually, this component density constitutes a cluster containing a few data points, very close together or almost lying in the same subspace. Such estimate tends to “interpolate” a local pattern and provides a bad fit for the remaining observations; as a consequence, the fitted model is not of practical use in inference. The above arguments hold for any distribution admitting the degenerate case; in such models, a global maximum does not exist and a great number of spurious maxima – usually having a large value of the log-likelihood function – may be found.

---

<sup>(7)</sup> Note that we speak of *best* maximum (rather than *absolute* maximum); this emphasizes that the likelihood function of some mixture models is unbounded and, in this case, the NPMLE corresponds to a local maximum. Further details are provided in the sequel.



In practice, the most common recommendation is to use a variety of starting points for the maximization algorithm: once the spurious maxima have been discarded, the estimate with the highest value of the observed log-likelihood is finally taken as the *true* one. Alternatively, we can use some technique to “escape” from the local maxima during the optimization: it is the case of the Stochastic EM algorithm (Broniatowski, Celeux and Diebolt, 1983). In most cases, these cares are enough to prevent a slip in some “bad” point of the likelihood surface: however, this is not always true.

Model misspecifications generally lead to a substantial bias in the parameters estimates; choosing a wrong number of components also generates spurious and local maxima. The joint estimation of  $k$  and  $\Phi$  may partially obviate to this problem; however, even when the model specification is correct, the likelihood function is likely to have a large number of local maxima: as a consequence, the estimates are very sensitive to the initial parameters. Especially in complex settings, estimating a finite mixture may be very time-consuming: for this reason, the relevant issue is how frequently we have to run the algorithm before obtaining a reliable estimate. To be more specific, we provide some simple guidelines:

- the occurrence of local and spurious optimizers is generally decreasing with the sample size;
- in weakly identified models (e.g., when the component densities are strongly overlapping) it is very frequent to have many different estimates with similar value of the observed log-likelihood;
- adding covariates or specifying a multivariate distribution for the component densities may improve the model identification; a reasonable sample size, however, is needed when a complex model is specified.

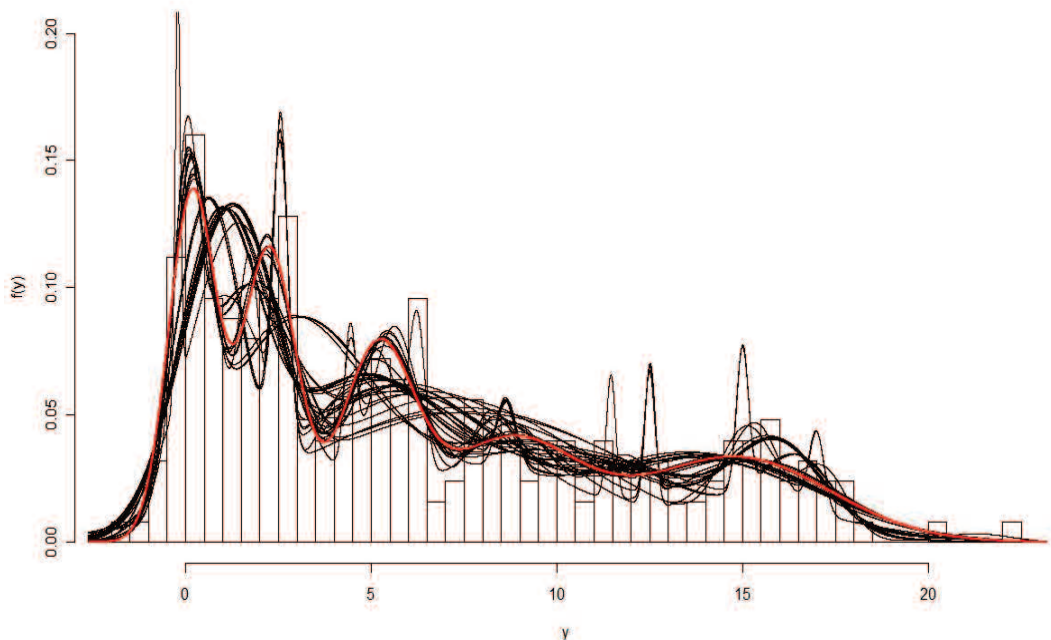


Figure 2.1 *Histogram of a sample from a 5-components Normal mixture with parameters  $\mu = (0, 2, 5, 9, 15)$ ,  $\sigma = (0.7, 1, 1, 2, 2)$  and mixing proportions  $\mathbf{p} = (0.2, 0.2, 0.2, 0.2, 0.2)$  ( $n = 250$ ). The superimposed densities correspond to 40 different local maxima of the log-likelihood function; among them, the red line is the true MLE.*

In Figure 2.1 we propose the histogram of a random sample from a 5-components Normal model ( $n = 250$ ); the superimposed densities correspond to 40 different local maxima found in 1000 runs of the EM with random starting points (we sampled the means from a  $\text{Uniform}(y_{(1)}, y_{(n)})$  distribution – where  $y_{(1)}$  and  $y_{(n)}$  are the sample minimum and maximum; the standard deviations were sampled from a  $\text{Uniform}(0,3)$  density; for the mixing proportions, a  $\text{Dirichlet}(\mathbf{1})$  was used). In 477 cases, an empty component or a spurious optimizer was found; only in 43 runs we found the true MLE (red line in Figure 2.1).

### *Saddle points*

The presence of saddle points in the log-likelihood function is a relevant issue (see, e.g., Wu, 1983; Fukumizu et al., 2003); in what follows, we demonstrate that this feature is common to a wide class of finite mixture models; as showed in the sequel, the feasibility of applying a mixture model to a given data set is also related to this topic.

In a univariate model, we suppose that the distribution of  $y$  – given the unknown cluster membership – belongs to the exponential family; that is, the component densities are of the form:

$$f(y | \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

where  $E(y | \theta) = b'(\theta)$  and  $\text{Var}(y | \theta, \phi) = b''(\theta)a(\phi)$ . In the above formula,  $\theta$  is the canonical parameter ( $\mu$  in the  $\text{Normal}(\mu, \sigma^2)$  density,  $\log(\lambda)$  in the  $\text{Poisson}(\lambda)$ ,  $\log[\pi / (1 - \pi)]$  in the  $\text{Bernoulli}(\pi)$  and so on) and usually  $a(\phi) = \phi$  ( $\sigma^2$  in the Normal case, 1 in one-parameter exponential families). The first derivative with respect to the canonical parameter is of the form:

$$\frac{\partial f(y | \theta, \phi)}{\partial \theta} = f(y | \theta, \phi) \frac{y - b'(\theta)}{a(\phi)}$$

Let us consider a mixture of  $k = 2$  components, both belonging to the same exponential family with known and common dispersion parameter  $\phi$ ; with these settings, each observation  $y_i$  is sampled from the mixed density:

$$\begin{aligned} g(y_i | \theta_1, \theta_2, \phi, p) &= \\ &= p \cdot \exp \left\{ \frac{y_i \theta_1 - b(\theta_1)}{a(\phi)} + c(y_i, \phi) \right\} + \\ &+ (1 - p) \cdot \exp \left\{ \frac{y_i \theta_2 - b(\theta_2)}{a(\phi)} + c(y_i, \phi) \right\} \\ &= \sum_{j=1}^2 p_j f(y_i | \theta_j, \phi) \end{aligned}$$

where  $\phi$  and  $p$  are supposed to be known and, in the last expression,  $f(\cdot)$  is the component density from the exponential family; clearly,  $p_1 = p$  and  $p_2 = 1 - p$ .

The log-likelihood function for a data set containing  $n$  observations can be written as:

$$l(\theta_1, \theta_2) = \sum_{i=1}^n \log \sum_{j=1}^2 p_j f(y_i | \theta_j, \phi)$$

Letting  $L_i = g(y_i | \theta_1, \theta_2, \phi, p)$  – that is, the contribution of the  $i^{\text{th}}$  observation to the likelihood function – the first derivatives with respect to the unknown parameters are of the form:

$$\frac{\partial l(\theta_1, \theta_2)}{\partial \theta_j} = \sum_{i=1}^n \frac{p_j f(y_i | \theta_j, \phi)(y_i - b'(\theta))}{a(\phi)L_i}$$

with  $j = \{1, 2\}$ . The Hessian matrix  $H$  has the following elements on its diagonal:

$$H_{jj} = \frac{\partial^2 l(\theta_1, \theta_2)}{\partial \theta_j^2} = \sum_{i=1}^n \left\{ \frac{p_j f(y_i | \theta_j, \phi)}{L_i^2} \left[ \left( \frac{y_i - b'(\theta_j)}{a(\phi)} \right)^2 (L_i - p_j f(y_i | \theta_j, \phi)) - \frac{b''(\theta_j)L_i}{a(\phi)} \right] \right\}$$

The mixed derivatives are of the form:

$$H_{12} = H_{21} = \frac{\partial^2 l(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2} = - \sum_{i=1}^n \left\{ \frac{p f(y_i | \theta_1, \phi)(y_i - b'(\theta_1)) \cdot (1-p) f(y_i | \theta_2, \phi)(y_i - b'(\theta_2))}{L_i^2 [a(\phi)]^2} \right\}$$

The determinant of the Hessian matrix is  $|H| = H_{11}H_{22} - H_{12}H_{21}$ ; setting  $\theta_1 = \theta_2 = \theta = h(\bar{y})$ , where  $\bar{y}$  is the sample mean of  $\{y_1, \dots, y_n\}$  and  $h(\cdot) = b'^{-1}(\cdot)$  is the canonical link function for the chosen density, we have:

$$\begin{aligned} f(y_i | \theta_1, \phi) &= f(y_i | \theta_2, \phi) = f(y_i | \theta, \phi) \\ L_i &= f(y_i | \theta, \phi) \\ b'(\theta_1) &= b'(\theta_2) = b'(\theta) \end{aligned}$$

It is easy to see that at this point the first derivatives are 0; with some algebra, we can show that the above Hessian matrix has the following determinant:

$$|H|_{\{\theta_1 = \theta_2 = \theta = h(\bar{y})\}} = p(1-p) \frac{nb''(\theta)}{a(\phi)} \left( \frac{nb''(\theta)}{a(\phi)} - \sum_{i=1}^n \left[ \frac{y_i - b'(\theta)}{a(\phi)} \right]^2 \right)$$

Looking at the sign of the above determinant, we can see that the log-likelihood function has a saddle point in  $\theta_1 = \theta_2 = \theta = h(\bar{y})$  (being  $|H| < 0$ ) if

$$\text{Var}(y | \theta = h(\bar{y}), \phi) < \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad [2.1]$$

where

$$\text{Var}(y | \theta, \phi) = b''(\theta)a(\phi)$$

$$\bar{y} = E[y | \theta = h(\bar{y})] = b'(h(\bar{y}))$$

That is, we have a saddle point if the empirical variance is larger than the theoretical one (given  $\phi$ ): working with Poisson or Binomial counts, this is true for every overdispersed data set; in a Normal model, the comparison is between the empirical variance and the known  $\sigma^2$ . If inequality [2.1] does not hold, in  $\theta_1 = \theta_2 = \theta = h(\bar{y})$  we have a *global maximum* of the likelihood function (being  $|H| > 0$  and  $H_{11} < 0$  for every admissible parameters value) and the mixture model cannot be estimated: we can argue that, given the current distributional assumptions, there is no evidence of unobserved heterogeneity.

In Figure 2.2, we depicted the shape of the log-likelihood function of a mixture of two Poisson distributions (with parameters  $\lambda_1$  and  $\lambda_2$  and with known mixing proportions). Figure 2.2a has been obtained from a simulated data set with sample variance greater than the sample mean: the MLE is the couple  $A = (\lambda_1^*, \lambda_2^*)$ ; since the mixing proportions are known, there is no label switching and the couple  $B = (\lambda_1', \lambda_2')$  is a local maximum; finally, there is a saddle point in  $S = (\bar{y}, \bar{y})$ . In Figure 2.2b, the sample mean is greater than the sample variance: the unique MLE is in  $M = (\bar{y}, \bar{y})$  and a mixture of Poisson distributions cannot be estimated.

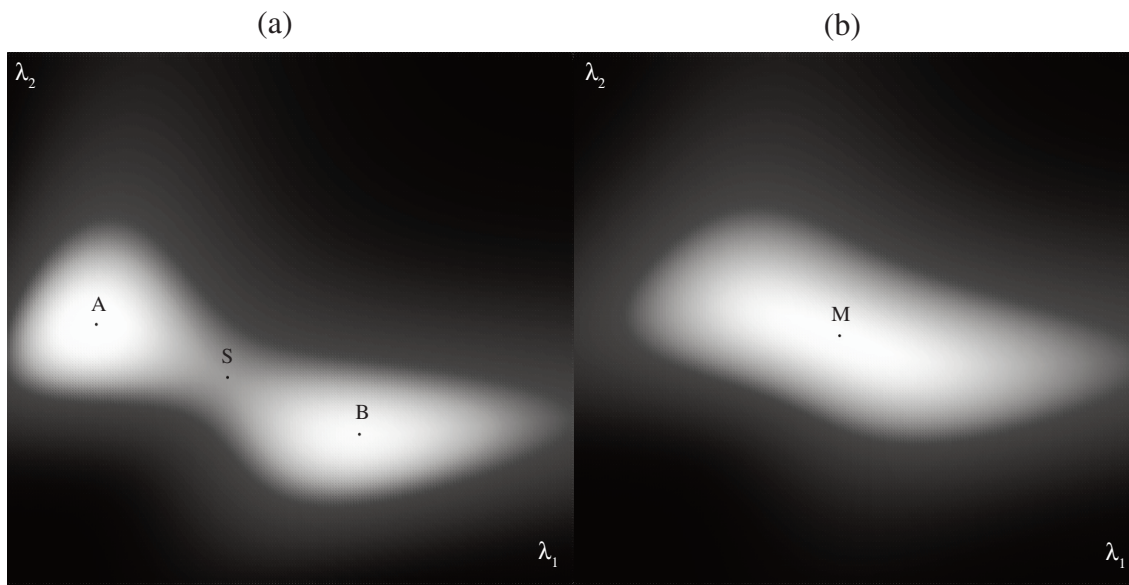


Figure 2.2 Contour plots of the log-likelihood function  $l(\lambda_1, \lambda_2)$  for a 2-components mixture of Poisson distributions with known mixing proportions ( $p = 0.3$ ). In (a) the sample variance is greater than the sample mean: we have a global maximum (A), a local maximum (B) and the saddle point (S) in  $\lambda_1 = \lambda_2 = \bar{y}$ . In (b) the sample mean is greater than the sample variance: as a consequence, the log-likelihood has only one global maximum (M) in  $\lambda_1 = \lambda_2 = \bar{y}$  and a mixture of Poisson densities cannot be estimated.

It is very hard to extend the above arguments to a k-components model with unknown dispersion parameters; however, simulations suggest that the same results hold, with some complications, for more complex models or in presence of covariates. If the mixing proportions are also unknown, when  $\theta_j = \theta_h$  for some (h, j) the Hessian matrix becomes sin-

gular: this is because  $k$  mixing proportions are estimated in a model with  $k - 1$  actual components; that is, the parameter space lies in a not identifiable set: infinitely many couples  $(\mathbf{p}_j, \mathbf{p}_h)$  return the same observed log-likelihood.

In practice, empirical results confirm that a very frequent outcome of the optimization algorithm is of the form

$$\hat{\Phi} = (\hat{\lambda}, \dots, \hat{\lambda}, \hat{\lambda}_h, \dots, \hat{\lambda}_k, \hat{\mathbf{p}})$$

where  $\hat{\lambda}_j = (\hat{\theta}_j, \hat{\phi}_j)$ ,  $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_k)$ ,  $j = 1, \dots, k$  and  $2 < h \leq k$ ; that is, the same estimates are obtained for a subset of two or more component densities (without loss of generality, the first  $h - 1$  in the above notation). This case may be very frequent when the sample size is small and the components are strongly overlapping; in order to find the true MLE, a central role is played by the starting points of the algorithm: in the first step, the component densities (based on the actual parameters vector) should not be too “close” together. At the extreme, if the starting values for the EM algorithm contain *the same* parameters for each component, the conditional membership probabilities in the E-step would be  $e_{ij} = 1/k$ : as a consequence, the EM would converge (after just one iteration) to the parameters vector  $\hat{\Phi} = (\hat{\lambda}, \dots, \hat{\lambda}, \hat{\mathbf{p}})$  with  $\hat{\mathbf{p}} = (1/k, \dots, 1/k)$ .

## 2.2 Finite Mixture Models and Genetic Algorithms

Genetic Algorithms (Holland, 1975) are a very general technique used to find approximate solutions to optimization and search problems.<sup>(8)</sup> The working principle of a Genetic Algorithm (GA) consists in using techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover. In what follows, we will explain how a GA proceeds in optimizing the log-likelihood function of a finite mixture model.

Simplifying, the steps of a GA are the following:

- *Initialization*

Many individual solutions (chromosomes) are randomly generated to form an initial population. In our problem, each individual is composed by  $q$  “genes”, that is the vector of order  $q$  containing the model parameters. The population size ( $N$ , the number of chromosomes) should be adequate for the function to optimize (in a model with many parameters, a bigger population is required). The initial population does not need to cover the entire range of possible solutions: the search space will expand by means of the mutation process. For each individual, the fitness function (in our case, the log-likelihood function) is evaluated.

- *Reproduction and mutation*

From the initial population, the  $N_E$  “best” individuals (the Elitists, that is those with higher fitness) are selected to survive in the new generation; this ensures the monoto-

<sup>(8)</sup> For a complete account on genetic algorithms, see for example Mitchell, 1996, Vose, 1999.

nicity of the algorithm. Afterward,  $(N - N_E)/2$  couples of parents are “selected” from the whole old generation; a possible strategy is to select each individual with probability proportional to some power (say  $\omega$ ) of the observed fitness or of the corresponding order statistic: if  $\omega = 0$ , there is no selection; increasing  $\omega$  will make the selection more and more severe. The parents will breed  $N - N_E$  new individuals by means of some cross-over mechanism. This progeny constitute (together with the Elitist) a new generation of size  $N$ . Optionally, a random sample of individuals (perhaps selected using some fitness criterion) is subject to a mutation; in most cases, the mutation consists in a random alteration of the genes. The mutation process ensures the renewal of the genetic heritage: the aim is to widely explore the parameter space, giving an opportunity to escape from local solutions. These processes ultimately result in the next generation population of chromosomes: the average fitness is expected to be increased, since only the best individuals from the first generation are selected for breeding.

- *Termination*

The generational process is repeated until a termination condition has been reached. Common terminating conditions are:

- a fixed number of generations ( $T$ ) is reached;
- the allocated budget (computation time/money) is reached;
- a fixed number of generation without significant improvements is reached;
- some minimum criterion is satisfied;
- combinations of the above.

The underlying idea of a GA is very simple: combining a stochastic search with the selection mechanism, we try a great number of possible solutions; under a computational point of view, this is less expensive than to do the same number of trials with a complete running of any maximization algorithm. The output of a GA is a number (usually,  $N_E$ ) of starting points available for an optimization routine.

The success of a GA is not guaranteed and depends on the function to be optimized and on the operational parameters: mainly, the population size ( $N$ ), the number of elitists ( $N_E$ ), the terminating condition, the selection criteria and the mutation rate. As a general guideline, in order to have a satisfactory outcome, the algorithm should run for a great number of generations: this ensures the effectiveness of the evolutionary process. With a high mutation rate and a small value of  $\omega$ , the algorithm tends to sacrifice short-term fitness to gain long-term fitness, widely exploring the parameter space and increasing the chance of escaping from a local maximum. Vice versa, a severe selection – together with a small mutation rate – would speed the convergence but may lead to a poor result in terms of fitness.

In order to show how a GA may improve the convergence of the EM algorithm, we conclude this chapter with a simulation study. For this purpose, we take back the data set of Figure 2.1, obtained from a 5-components Normal model ( $n = 250$ ). We also suppose to know the variance parameter of each Gaussian density (this rules out the presence of spurious optimizers): therefore, the estimand parameters are the means  $\mu_1, \dots, \mu_5$  and the mixing proportions  $p_1, \dots, p_5$ . Randomly choosing the starting values for  $\mu_1, \dots, \mu_5$  from a  $\text{Uniform}(y_{(1)}, y_{(n)})$  distribution (where  $y_{(1)}$  and  $y_{(n)}$  are the sample minimum and maximum)



and those for  $p_1, \dots, p_5$  from a Dirichlet(1) density, we ran the EM 2000 times: only in 18 cases (0.9%) the “true” estimate was found; in 447 cases (22.35%) the EM converged to a local maximum, whereas some empty component was found in the remaining 1535 runs (76.75% of cases); on the whole, 26 different local maxima were located. Assuming that the chance of success is 0.9%, in order to find the true MLE with a probability of 95% we should run the EM using 332 different starting values.

In order to evaluate how a genetic algorithm may undertake this estimation problem, we carried out a great number of simulations. In our very simple implementation of a GA, the initial population is randomly generated in the same way as the above starting points; each individual is selected to breed with a probability proportional to the  $\omega^{\text{th}}$  power of the individual’s position after ordering the population by increasing fitness (log-likelihood); given a couple of parents, each gene is selected to switch across parents according to a Bernoulli trial: the crossing-over rate ( $\pi_c$ ) is set to 0.5 (in this phase, the vector of the mixing proportions is taken as a single gene). After breeding, three different mutations operate:

- type 1 mutation: a random alteration of the chromosome (an additive shift from a  $N(0, 0.8)$  for the means; new mixing proportions are sampled from an Uniform Dirichlet distribution). Each chromosome is selected to mutate with a mutation rate  $\pi_1$ ; the mutation is accepted if the fitness has been improved: otherwise, the old individual is restored;
- type 2 mutation: the  $N\pi_2$  lower-fitness chromosomes are selected to be completely replaced by new individuals; this ensures a high renewal rate of the genetic heritage;
- type 3 mutation: with a mutation rate  $\pi_3$ , a random sample of the individuals is selected to be improved with one EM iteration.

The terminating condition is reached after T generations without significant improvements in the population highest fitness (that is, a growth rate less than 0.002 between two successive generations).

We made a great number of trials with different operational parameters: Table 2.2 displays the number of success in 500 trials for different values of T, N,  $\omega$  and mutation rates (with  $N_E = 10$  and  $\pi_c = 0.5$ ); between brackets, the mean computational time is recorded. As we would expect, better results are obtained when  $\omega$  is quite small and the mutation rates are high; furthermore, the proportion of success is generally increasing with the population size (N); the stopping rule (T) is also relevant in determining the chance of success. However, when the mutation rates are too small, a poor performance is observed, irrespective of the value of T; in the same way, if the selection parameter ( $\omega$ ) is high, increasing T does not provide a significant improvement.

In most settings, running the EM from different starting values is enough to find the true MLE; however, the proposed example shows that optimizing the log-likelihood function may be a difficult task. The GAs are a useful tool in escaping from local optimizers; unfortunately, there is not a general rule to set the operational parameters.

Our GA is unsatisfactory in presence of spurious solutions, that may have a larger log-likelihood than the true MLE: the algorithm should be able to “recognize” the spurious maxima, in order to discard them from the number of admissible solutions.

In the next chapter, we will illustrate how a multivariate approach may improve the search for the optimum of the likelihood function; we will also investigate the performance of a multivariate mixture model in terms of standard errors and posterior classification.

T = 50					
		N = 200		N = 100	
		$\omega = 0.4$	$\omega = 1$	$\omega = 0.4$	$\omega = 1$
$\pi_1 = \pi_2 = \pi_3 = 0.4$		85.8% (2'55")	65.4% (2'28")	69.2% (1'28")	48.4% (1'21")
$\pi_1 = \pi_2 = \pi_3 = 0.1$		23.4% (1'3")	19.8% (58")	17.4% (37")	14.2% (35")

T = 200					
		N = 200		N = 100	
		$\omega = 0.4$	$\omega = 1$	$\omega = 0.4$	$\omega = 1$
$\pi_1 = \pi_2 = \pi_3 = 0.4$		99.2% (6'31")	69.4% (6'4")	87.8% (3'13")	48.2% (3'6")
$\pi_1 = \pi_2 = \pi_3 = 0.1$		23.6% (2'9")	21.0% (2'5")	16.8% (1'9")	15.4% (1'8")

Table 2.2 Performance of a GA in finding the true MLE of a data set from a 5-components Normal mixture with  $\boldsymbol{\mu} = (0, 2, 5, 9, 15)$ ,  $\mathbf{p} = (0.2, 0.2, 0.2, 0.2, 0.2)$  and known standard deviations  $\boldsymbol{\sigma} = (0.7, 1, 1, 2, 2)$ ; for each combination of T, N,  $\omega$ ,  $\pi_1$ ,  $\pi_2$  and  $\pi_3$ , the proportion of successes in 500 trials is displayed (between brackets, the mean computational time). In all simulations,  $N_E = 10$  and  $\pi_c = 0.5$ .



### 3 Multivariate models

In this chapter, we present some special features of the multivariate approach in estimating finite mixture models. We denote with  $\mathbf{Y} = (Y_1, \dots, Y_p)$  the  $p$ -variate outcome variable; for ease of notation, we omit the conditioning to an optional set of covariates. The joint distribution of the observed  $\mathbf{y}$  is supposed to have the form:

$$f(\mathbf{y} \mid \Phi) = f(\mathbf{y} \mid \lambda_1)p_1 + \dots + f(\mathbf{y} \mid \lambda_k)p_k$$

where  $f(\mathbf{y} \mid \lambda_j) = f(y_1, \dots, y_p \mid \lambda_j)$ .

In many settings, it could be difficult to completely specify the above distribution, especially when the outcome variables  $Y_1, \dots, Y_p$  are supposed to lie in different parametric families; however, each component density may be specified as the product of the conditional distributions:

$$f(\mathbf{y} \mid \lambda_j) = f(y_1 \mid y_2, \dots, y_p; \lambda_j) f(y_2 \mid y_3, \dots, y_p; \lambda_j) \dots f(y_{p-1} \mid y_p; \lambda_j) f(y_p \mid \lambda_j)$$

In a homogeneous model ( $k = 1$ ) this would factorize the log-likelihood function: that is, the  $p$  conditional models would be fitted separately. When working with a mixture model, this is no longer true: as a consequence, different estimates would be obtained by fitting the  $p$  univariate conditional models and the  $p$ -variate one, even if the joint distribution is specified as the above product.

This chapter proceeds as follows. Section 3.1 is devoted to the special case of the multivariate Normal mixtures; in Section 3.2, we show how the specification of a multivariate mixture – in place of the univariate approach – may be helpful in improving the model identification, with a positive effect on the convergence rate of the EM algorithm and on the standard errors of the estimates.

#### 3.1 Mixtures of multivariate Normal distributions

The Normal distribution (including linear regression and more sophisticated tools) is the most frequent choice when dealing with continuous random variables; if compared to the homogeneous case, a Normal mixture allows for a greater flexibility and can be viewed as a generalization of the basic approach. In this section, we will stress some important features of the multivariate Normal mixture model, with a special care for the marginal and conditional distributions.

A  $k$ -components mixture of  $p$ -variate Normal distributions for the response variable  $\mathbf{y} = (y_1, \dots, y_p)$  has the component densities of the form:

$$f(\mathbf{y} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_j|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j) \right\}$$

In the above formula,  $\boldsymbol{\mu}_j$  and  $\boldsymbol{\Sigma}_j$  are – respectively – the mean and the covariance matrix of  $\mathbf{y}$  in the  $j^{\text{th}}$  subpopulation ( $\boldsymbol{\mu}_j = \{\mu_{1j}, \dots, \mu_{pj}\}$ ,  $j = 1, \dots, k$ ). In the sequel, we will refer to this distribution as  $N_p^{(j)}(\mathbf{y})$ . In Figure 3.1, we illustrate the shape of different bivariate Normal mixtures with some simple graphical example.

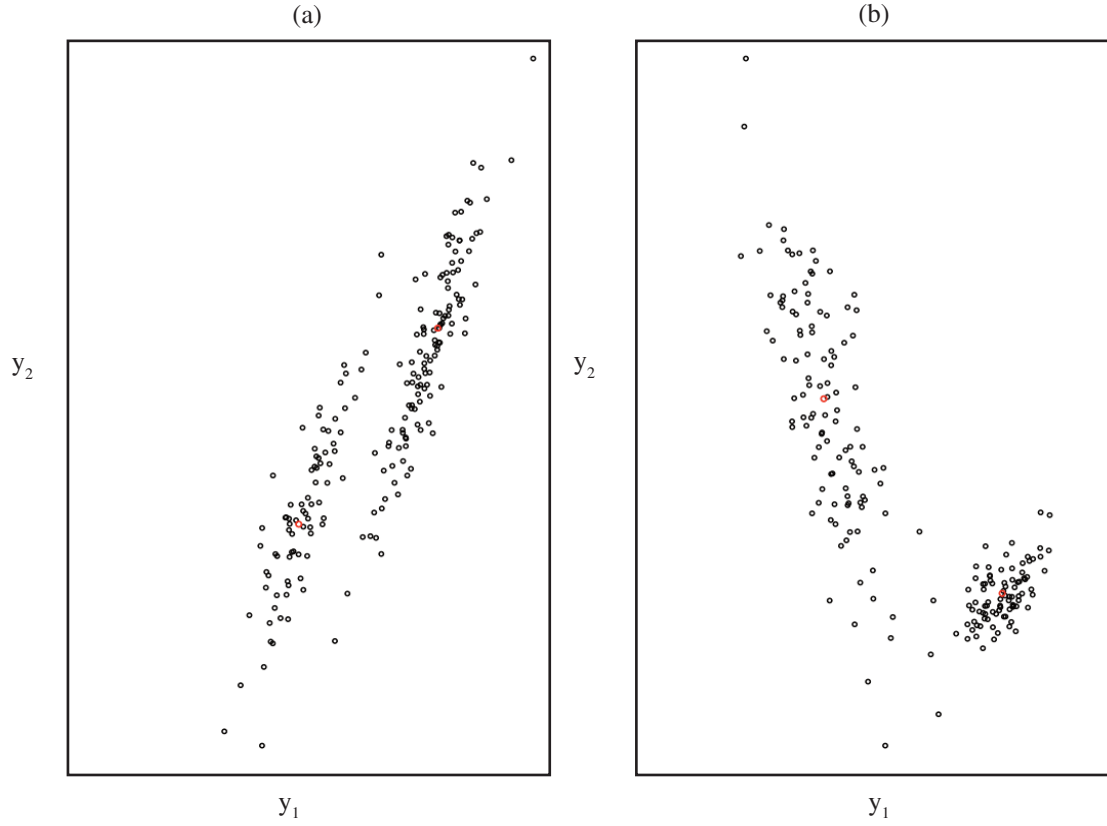


Figure 3.1 Random samples from 2-components mixtures of bivariate Normal distributions; (a) mixture with common covariance matrix; (b) mixture with different covariance matrices.

Without loss of generality, suppose that  $\mathbf{y}$ ,  $\boldsymbol{\mu}_j$  and  $\boldsymbol{\Sigma}_j$  are partitioned in two groups with size  $r$  and  $p - r$ , respectively:

$$\mathbf{y} = (\mathbf{y}_a, \mathbf{y}_b)$$

$$\boldsymbol{\mu}_j = (\boldsymbol{\mu}_{j,a}, \boldsymbol{\mu}_{j,b})$$

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_{j,aa} & \boldsymbol{\Sigma}_{j,ab} \\ \boldsymbol{\Sigma}_{j,ba} & \boldsymbol{\Sigma}_{j,bb} \end{bmatrix}$$

For each component density – that is, within the  $j^{\text{th}}$  subpopulation – the marginal and conditional distributions are, respectively:

$$\mathbf{y}_a \sim N_r(\boldsymbol{\mu}_{j,a}, \boldsymbol{\Sigma}_{j,aa})$$

$$\mathbf{y}_a | \mathbf{y}_b \sim N_r(\boldsymbol{\mu}_j^*, \boldsymbol{\Sigma}_j^*)$$

with

$$\begin{aligned}\boldsymbol{\mu}_j^* &= \boldsymbol{\mu}_{j,a} + \boldsymbol{\Sigma}_{j,ab} \boldsymbol{\Sigma}_{j,bb}^{-1} (\mathbf{y}_b - \boldsymbol{\mu}_{j,b}) \\ \boldsymbol{\Sigma}_j^* &= \boldsymbol{\Sigma}_{j,aa} - \boldsymbol{\Sigma}_{j,ab} \boldsymbol{\Sigma}_{j,bb}^{-1} \boldsymbol{\Sigma}_{j,ba}\end{aligned}$$

For ease of notation, in what follows we will write the above marginal and conditional distributions as  $N_r^{(j)}(\mathbf{y}_a)$  and  $N_r^{(j)}(\mathbf{y}_a | \mathbf{y}_b)$ , respectively.

Being  $p_1, \dots, p_k$  the mixing proportions of the mixture model ( $p_k = 1 - p_1 - \dots - p_{k-1}$ ), the distribution of  $\mathbf{y}$  is given by:

$$g(\mathbf{y}) = p_1 N_p^{(1)}(\mathbf{y}) + \dots + p_k N_p^{(k)}(\mathbf{y})$$

According to the same partition as above,  $\mathbf{y}_a$  has the following marginal density:

$$g(\mathbf{y}_a) = p_1 N_r^{(1)}(\mathbf{y}_a) + \dots + p_k N_r^{(k)}(\mathbf{y}_a)$$

The conditional distribution of  $\mathbf{y}_a$  given  $\mathbf{y}_b$  can be computed in a simple way:

$$\begin{aligned}g(\mathbf{y}_a | \mathbf{y}_b) &= \frac{g(\mathbf{y})}{g(\mathbf{y}_b)} \\ &= \sum_{j=1}^k \frac{p_j N_p^{(j)}(\mathbf{y})}{g(\mathbf{y}_b)} \\ &= \sum_{j=1}^k \frac{p_j N_p^{(j)}(\mathbf{y})}{p_j N_{p-r}^{(j)}(\mathbf{y}_b)} \frac{p_j N_{p-r}^{(j)}(\mathbf{y}_b)}{g(\mathbf{y}_b)} \\ &= \sum_{j=1}^k N_r^{(j)}(\mathbf{y}_a | \mathbf{y}_b) p_j^*(\mathbf{y}_b)\end{aligned}$$

In the above result,  $N_r^{(j)}(\mathbf{y}_a | \mathbf{y}_b)$  is – within the  $j^{\text{th}}$  component – the conditional distribution of  $\mathbf{y}_a$  given  $\mathbf{y}_b$ ;  $p_j^*(\mathbf{y}_b)$  is the ratio between the density of  $\mathbf{y}_b$  in the  $j^{\text{th}}$  subpopulation and its marginal density – that is, the conditional membership probability for the  $j^{\text{th}}$  component. The mixing proportions (as well as the component densities) vary across the values of  $\mathbf{y}_b$ ; if we consider two distinct components ( $j$  and  $h$ ) we have:

$$\begin{aligned}\log \left[ \frac{p_j^*(\mathbf{y}_b)}{p_h^*(\mathbf{y}_b)} \right] &= \\ -\frac{1}{2} \left[ \log \frac{|\boldsymbol{\Sigma}_{j,bb}|}{|\boldsymbol{\Sigma}_{h,bb}|} + 2 \log \left( \frac{p_h}{p_j} \right) + (\mathbf{y}_b - \boldsymbol{\mu}_{j,b})^T \boldsymbol{\Sigma}_{j,bb} (\mathbf{y}_b - \boldsymbol{\mu}_{j,b}) - (\mathbf{y}_b - \boldsymbol{\mu}_{h,b})^T \boldsymbol{\Sigma}_{h,bb} (\mathbf{y}_b - \boldsymbol{\mu}_{h,b}) \right]\end{aligned}$$

This means that a multinomial logistic model holds for the mixing proportions; each ele-

ment of  $\mathbf{y}_b$  ( $\mathbf{y}_{1,b}, \dots, \mathbf{y}_{p-r,b}$ ) gets in the predictor in a linear and quadratic form, including couple interactions  $\mathbf{y}_{m,b}; \mathbf{y}_{m',b}$  ( $m, m' = 1, \dots, p - r$ ).

To make an example, let us consider the simplest case – that is, a mixture of  $k = 2$  bivariate normal distributions. The density of this model is:

$$g(\mathbf{y}_1, \mathbf{y}_2) = pN_2^{(1)}(\mathbf{y}_1, \mathbf{y}_2) + (1 - p)N_2^{(2)}(\mathbf{y}_1, \mathbf{y}_2)$$

where the parameters of  $N_2^{(j)}(\mathbf{y}_1, \mathbf{y}_2)$  are  $\{\mu_{1j}, \mu_{2j}, \sigma_{1j}^2, \sigma_{2j}^2, \rho_j\}$ ,  $j = \{1, 2\}$ . In Table 3.1, we provide the estimand parameters of the bivariate model and of the marginal and conditional models ( $Z$  denotes the component membership label).

Joint distribution $g(\mathbf{y}_1, \mathbf{y}_2)$	Marginal distributions $g(\mathbf{y}_1), g(\mathbf{y}_2)$	Conditional distributions $g(\mathbf{y}_1   \mathbf{y}_2), g(\mathbf{y}_2   \mathbf{y}_1)$
$E[y_1   Z = 1] = \mu_{11}$	$E[y_1   Z = 1] = \mu_{11}$	$E[y_1   y_2, Z = 1] = \mu_{11} + (y_2 - \mu_{21})\rho_1\sigma_{11}/\sigma_{21}$
$E[y_1   Z = 2] = \mu_{12}$	$E[y_1   Z = 2] = \mu_{12}$	$E[y_1   y_2, Z = 2] = \mu_{12} + (y_2 - \mu_{22})\rho_2\sigma_{12}/\sigma_{22}$
$E[y_2   Z = 1] = \mu_{21}$	$E[y_2   Z = 1] = \mu_{21}$	$E[y_2   y_1, Z = 1] = \mu_{21} + (y_1 - \mu_{11})\rho_1\sigma_{21}/\sigma_{11}$
$E[y_2   Z = 2] = \mu_{22}$	$E[y_2   Z = 2] = \mu_{22}$	$E[y_2   y_1, Z = 2] = \mu_{22} + (y_1 - \mu_{12})\rho_2\sigma_{22}/\sigma_{12}$
$\text{Var}[y_1   Z = 1] = \sigma_{11}^2$	$\text{Var}[y_1   Z = 1] = \sigma_{11}^2$	$\text{Var}[y_1   y_2, Z = 1] = \sigma_{11}^2(1 - \rho_1^2)$
$\text{Var}[y_1   Z = 2] = \sigma_{12}^2$	$\text{Var}[y_1   Z = 2] = \sigma_{12}^2$	$\text{Var}[y_1   y_2, Z = 2] = \sigma_{12}^2(1 - \rho_2^2)$
$\text{Var}[y_2   Z = 1] = \sigma_{21}^2$	$\text{Var}[y_2   Z = 1] = \sigma_{21}^2$	$\text{Var}[y_2   y_1, Z = 1] = \sigma_{21}^2(1 - \rho_1^2)$
$\text{Var}[y_2   Z = 2] = \sigma_{22}^2$	$\text{Var}[y_2   Z = 2] = \sigma_{22}^2$	$\text{Var}[y_2   y_1, Z = 2] = \sigma_{22}^2(1 - \rho_2^2)$
$\text{Corr}[y_1, y_2   Z = 1] = \rho_1$	–	–
$\text{Corr}[y_1, y_2   Z = 2] = \rho_2$	–	–
$P(Z = 1) = p$	$P(Z = 1) = p$	$P(Z = 1   y_j)/P(Z = 2   y_j) = \exp\{\gamma_0 + \gamma_1 y_j + \gamma_2 y_j^2\}$

Table 3.1 *Estimand parameters of the joint, marginal and conditional distributions of a mixture of two bivariate Normal densities, with parameters  $(\mu_{11}, \mu_{21}, \sigma_{11}^2, \sigma_{21}^2, \rho_1)$  and  $(\mu_{12}, \mu_{22}, \sigma_{12}^2, \sigma_{22}^2, \rho_2)$ , respectively, and with mixing proportions  $p = P(Z = 1)$  and  $1 - p = P(Z = 2)$ . In the conditional distributions, the mixing proportions follow a logistic model with parameters  $\gamma_0, \gamma_1$  and  $\gamma_2$  (see the text for more details).*

In the conditional distribution of  $y_1$  given  $y_2$ , the logistic model for the mixing proportions has the following parameters:

$$P(Z = 1 | y_2)/P(Z = 2 | y_2) = \exp\{\gamma_0 + \gamma_1 y_2 + \gamma_2 y_2^2\}$$

The analogous result holds for the conditional distribution of  $y_2$  given  $y_1$ . We may express the coefficients  $\gamma_0$ ,  $\gamma_1$  and  $\gamma_2$  as a function of the parameters of the mixing densities:

$$\begin{aligned}\gamma_0 &= \log\left(\frac{p_1}{p_2}\right) - \log\left(\frac{\sigma_{12}}{\sigma_{11}}\right) - \frac{\mu_{11}^2}{2\sigma_{11}^2} + \frac{\mu_{12}^2}{2\sigma_{12}^2} \\ \gamma_1 &= \frac{\mu_{11}}{\sigma_{11}^2} - \frac{\mu_{12}}{\sigma_{12}^2} \\ \gamma_2 &= \frac{1}{2\sigma_{12}^2} - \frac{1}{2\sigma_{11}^2}\end{aligned}$$

The above results show that a mixture of multivariate Normal distributions may be decomposed in a very simple way into the corresponding marginal and conditional models; however, it is very important to understand the impact of different model specification in the inference: we will deeply investigate this issue in the next section.

### 3.2 A comparison between multivariate and univariate mixture models

In this section, a comparison between the univariate approach and the multivariate one is proposed; the differences in the computational time are pointed out; a special attention is directed to the variance of the estimates and to the discriminating power for clustering purposes; the issue of the local optimizers is also of interest. Using a simulation approach, we will illustrate the advantages that a multivariate specification may offer in some special settings.

Intuitively, we can think at every response variable as a criterion to assign each observation ( $i = 1, \dots, n$ ) to the cluster ( $j = 1, \dots, k$ ) which the unit is more likely to belong: when performing the E-step, we assess the cluster membership in a probabilistic way, given the current parameters vector; in the M-step, we provide new estimates for each component of the mixture, weighting the observations based on the above allocation. The response variables can be viewed as an instrument to find out the latent variable  $Z$ : for this reason, using more information will provide a greater discriminating power.

A very simple example of the usefulness of a multivariate approach is given in Figure 3.2, which represents a sample from a mixture of  $k = 2$  bivariate distributions. The red ( $a$ ) and the blue ( $b$ ) data points are clearly belonging to cluster A and B, respectively. Fitting a mixture model on  $y_1$  would imply a weak identification of point  $a$ , which is in an ambiguous place with respect to  $y_1$ . The same consideration holds for  $b$ , if a model for  $y_2$  is fitted. Figure 3.3 displays the marginal distributions of  $y_1$  and  $y_2$  (the blue and the red line represent the  $y_1$ - and  $y_2$ - coordinates of  $a$  and  $b$ , respectively). With a model for the couple  $(y_1, y_2)$  we would get around this problem and both units  $a$  and  $b$  would be unambiguously attributed to the “right” component. As a consequence, a multivariate approach is expected to provide more information and to improve the model identification. It is very important to realize that if  $y_1$  and  $y_2$  are independent given the component membership (as they are in Figure 3.2) the above argument still holds – since the log-likelihood does not factorize.

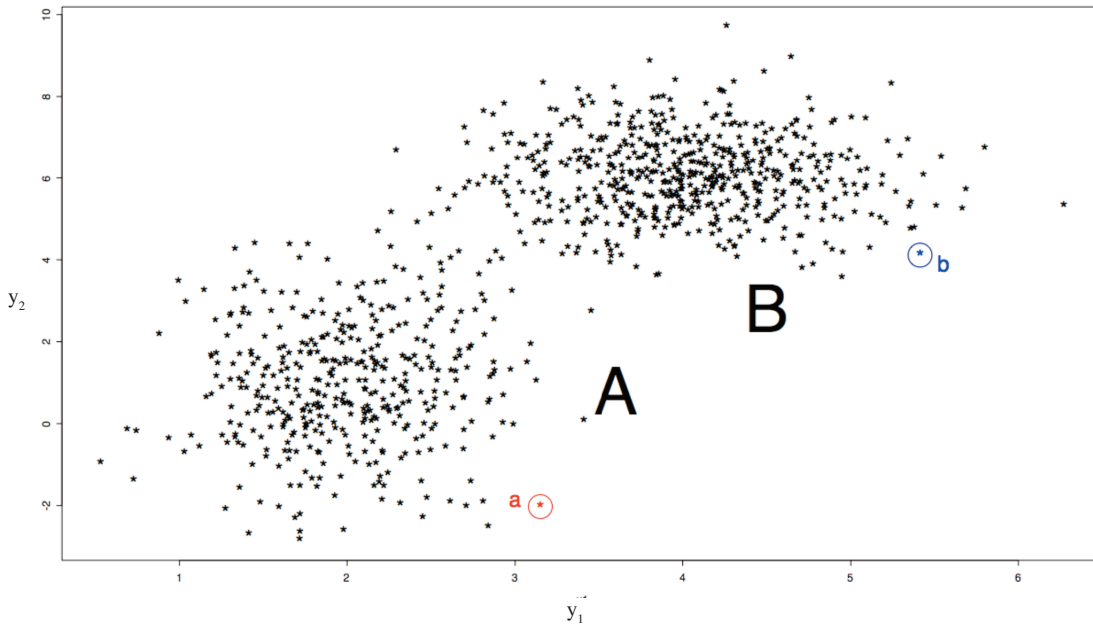


Figure 3.2 Random sample from a 2-components mixture of bivariate distributions. In the two-dimensional space  $(y_1, y_2)$ , there is a strong evidence that the data point a belongs to the component A of the mixture. This is no longer true if we look at the marginal density of  $y_1$ . An analogous argument holds for point b, whose location is “ambiguous” with respect to the marginal distribution of  $y_2$  (see also Figure 3.3).

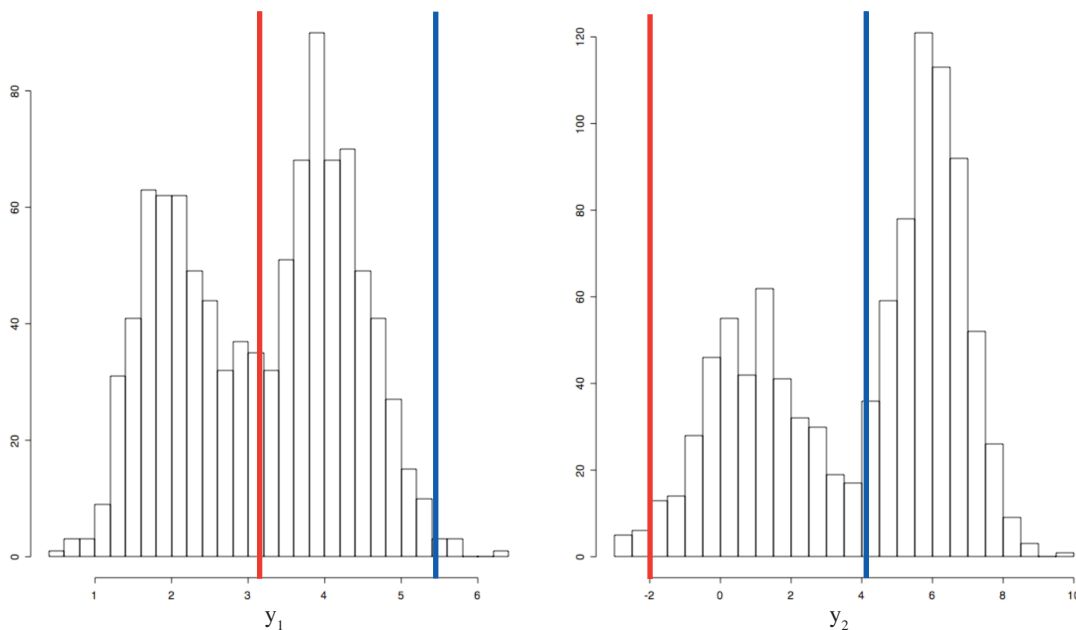


Figure 3.3 Marginal distributions of the bivariate mixture in Figure 3.2. The red and the blue lines represent the coordinates of points a and b in Figure 3.2: it is clearly evident that  $y_1$  cannot be a “good” classifier for point a; in the same way, point b is in an ambiguous location with respect to  $y_2$ .

We conclude with a simple example, whose implications are quite surprising for their strength. To ease the notation, we will write in vector form  $(\theta_1, \theta_2, \dots, \theta_k)^T$  the  $k$  different values that the same parameter  $\theta$  assumes in each component of the mixture.

Let us suppose to have a mixture model with  $k = 2$  latent clusters, with mixing proportions  $p$  and  $1 - p$ , respectively; we observe two variables, say  $U$  and  $V$ , with the following distributional assumptions:

$$U \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}\right)$$

$$V \sim N\left(\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 \\ \tau_2^2 \end{pmatrix}\right)$$

We assume that – given the cluster membership –  $U$  and  $V$  are independent; we can write:

$$U, V \sim N_2\left(\begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_2 \end{pmatrix}\right)$$

where  $\mathbf{M}_1 = (\mu_1, v_1)$ ,  $\mathbf{M}_2 = (\mu_2, v_2)$ ,  $\boldsymbol{\Sigma}_1 = \text{diag}(\sigma_1^2, \tau_1^2)$  and  $\boldsymbol{\Sigma}_2 = \text{diag}(\sigma_2^2, \tau_2^2)$ . If the difference between  $\mathbf{M}_1$  and  $\mathbf{M}_2$  is small compared to the corresponding variances, the component densities are strongly overlapping: as a consequence, the log-likelihood is expected to be very flat and the estimates have a large sampling variance.

For  $\mu_1 = 2$ ,  $\mu_2 = 2.5$ ,  $v_1 = 1$ ,  $v_2 = 1.7$ ,  $\sigma_1 = 0.25$ ,  $\sigma_2 = 0.3$ ,  $\tau_1 = 0.4$ ,  $\tau_2 = 0.35$ ,  $p = 0.5$ , Figure 3.4 displays the marginal distributions of  $U$  and  $V$ . In Figure 3.5, we plotted the joint density and a realized sample from the bivariate distribution ( $n = 10,000$ ). It is clearly visible that the univariate densities are strongly overlapping, while in the joint distribution – which appears to be bimodal – the clustering is much more pronounced.

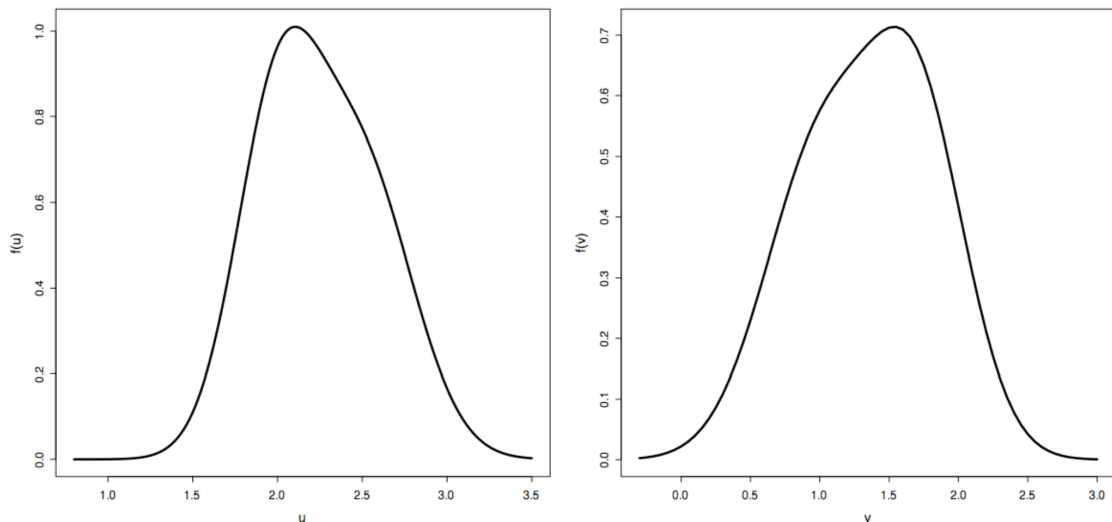


Figure 3.4 *Marginal densities of a mixture of bivariate Normal distributions (the model parameters are in Table 3.2). The component are strongly overlapping: as a consequence, a very flat log-likelihood is found when estimating a univariate model.*

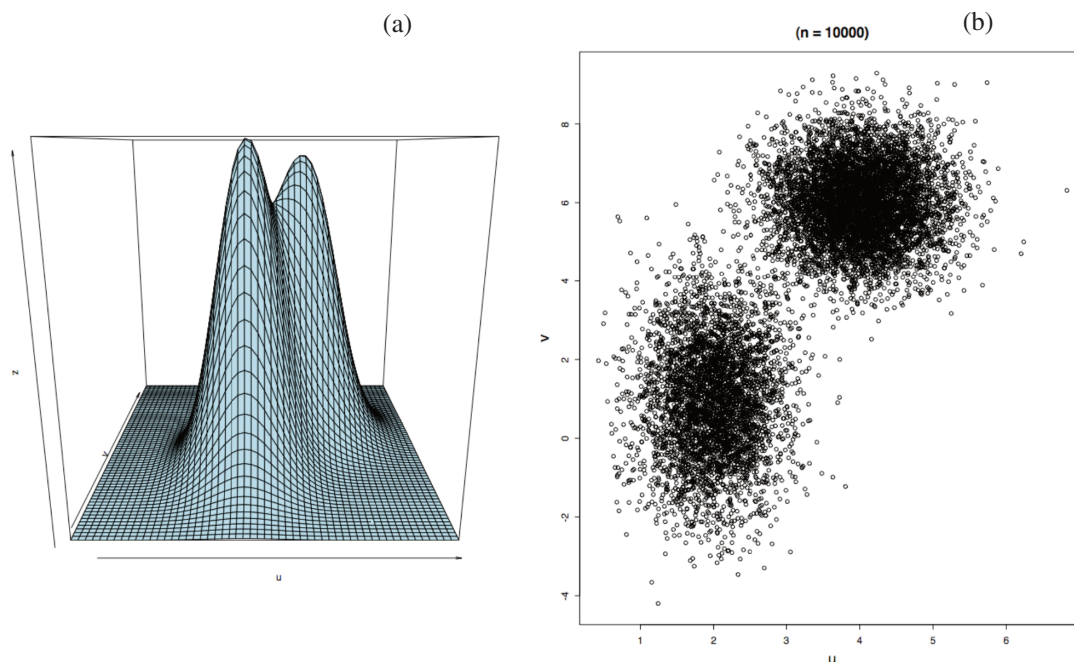


Figure 3.5 (a) joint density of a mixture of bivariate Normal distributions (the model parameters are in Table 3.2); (b) random sample from the distribution in (a) ( $n = 10,000$ ). Unlike the marginal distributions (Figure 3.4) the joint density is bimodal: as a consequence, a better clustering is obtained from a bivariate analysis.

Given a sample from this random variable, the simplest approach is to formulate a univariate model for  $U$  and – separately – for  $V$ . Under a theoretical point of view, introducing a bivariate model leaves the assumptions on the data-generating process unchanged: the estimand parameters will remain the same (given that the covariances are supposed to be null) and we need to compute bivariate normal densities to perform the E-step. In concrete, the strength of a multivariate approach is due to the latter feature: both response variables contribute to the estimation of the unknown cluster membership. Thereby, a bivariate approach is expected to reduce the uncertainty about the model parameters and, ultimately, the variance of estimates.

We drew 1000 samples of size  $n = 400$  from the above mixture distribution: for each sample, both approaches were used to estimate the 9 model parameters (covariances are imposed to be 0; the true values were used as starting points); given the final estimates, each unit was assigned to one of the two latent clusters, according to the conditional membership probability; the computational time was recorded.

Results are in Tables 3.2 and 3.3 (when estimating the univariate models, 2 different estimates of  $p$  are obtained: both standard errors are returned in Table 3.2). The bivariate approach leads to more efficient estimates of all parameters (Table 3.2); from Table 3.3, we can see that the average number of correct allocations (358.4) is much greater than in the univariate models (302.87 and 304.5 using  $U$  and  $V$ , respectively). Finally, Table 3.3 shows a remarkable difference in the computation time between the two approaches: about 18 seconds to estimate separately the univariate models; less than 2 seconds in the bivariate case. This happens because a more peaked log-likelihood function requires less iterations to reach the convergence criterion.



Parameter	True value	Std. Error (univariate model)	Std. Error (bivariate model)
$\mu_1$	2	0.00754	0.00063
$\mu_2$	2.5	0.01713	0.00136
$\nu_1$	1	0.02995	0.00239
$\nu_2$	1.7	0.01565	0.00126
$\sigma_1$	0.25	0.00206	0.00026
$\sigma_2$	0.3	0.00286	0.00054
$\tau_1$	0.4	0.00518	0.00099
$\tau_2$	0.35	0.00435	0.00048
p	0.5	0.03687; 0.03787	0.00238

Table 3.2 *Estimated standard errors (based on 1000 Monte Carlo replications) for the parameters of a mixture of two bivariate Normal distributions ( $n = 400$ ) with diagonal covariance matrix; when analyzing separately the two outcome variables (3<sup>rd</sup> column) greater standard errors are obtained than in the bivariate approach (4<sup>th</sup> column); another drawback of the univariate approach is that two different estimates of p are obtained.*

		U	V	U, V
Number of correct assignments ( $n = 400$ )	mean	302.87	304.50	358.40
	std.deviation	29.96	31.76	7.00
	min	188	181	336
	1 <sup>st</sup> quartile	289	290	354
	median	312	315	359
	3 <sup>rd</sup> quartile	325	327	363
	max	349	348	379
Computational time	total	301' 1"	33' 14"	
	mean	18.06"	1.99"	

Table 3.3 *Comparison between the univariate and the bivariate approach in estimating a mixture of two bivariate Normal distributions with outcome U, V and sample size  $n = 400$  (the model parameters are displayed in Table 3.2): the univariate models are found to be less effective in assessing the cluster membership using Bayes' rule; with respect to the bivariate model, a much greater computational time is required (in the estimation, the standard EM algorithm has been used). The summary statistics are referred to 1000 Monte Carlo replications.*

The improvement in the model identification may have a strong effect on the risk of falling in a local/spurious optimizer. In a multivariate approach, simulation results suggest that the EM algorithm is usually less likely to be attracted in local patterns; it seems reasonable that, as a consequence of the increased discriminating power, the number of local maxima becomes smaller than in the univariate case.

In order to see how often the EM falls in local and spurious optimizers in our example, we drew 100 samples ( $u^*$ ,  $v^*$ ) of size  $n = 400$  from the mixture of bivariate normal distri-

butions; for each dataset, we estimated 100 times the univariate models and the bivariate one, using the same randomly chosen starting points from a Uniform distribution (between  $\min(u^*)$  and  $\max(u^*)$  for  $\mu_1$  and  $\mu_2$ ; between  $\min(v^*)$  and  $\max(v^*)$  for  $v_1$  and  $v_2$ ; between 0 and 1 for the standard deviations and  $p$ ); among these 100 estimates, the local/spurious maximizers were identified and their number was recorded. For the 100 simulated data set, Table 3.4 contains the summary statistics of the proportion of local optimizers: results indicate that a multivariate approach improves the convergence of the EM algorithm to the true MLE; the average proportion of local/spurious optimizers was 8.36% and 11.4% in the two univariate models, versus 0.24% using a bivariate mixture. In the “worst” dataset, the occurrence of local and spurious optimizers was 46% with the univariate approach, only 4% using a bivariate model.

Occurrence of local/spurious optimizers in 100 simulated datasets (%)			
	U	V	U, V
min	0	0	0
1 <sup>st</sup> quartile	1	2.75	0
median	4	6.5	0
mean	8.36	11.4	0.24
3 <sup>rd</sup> quartile	10	18	0
max	46	46	4

Table 3.4 Comparison between the univariate and the bivariate approach in estimating a mixture of two bivariate Normal distributions with outcome  $U, V$  and sample size  $n = 400$  (the model parameters are displayed in Table 3.2). 100 data sets were sampled; for each data set, 100 estimates were obtained and the occurrence of spurious optimizers was recorded; the bivariate model appears to be more likely to find the “true” estimate: in the worst case, the occurrence of local maxima was 4% (versus 46% for the univariate models). In the estimation, the standard EM has been used, with randomly chosen starting points from a Uniform distribution (between 0 and 1 for the standard deviations and  $p$ , in the sample range for the means).

It is impossible to generalize the above statements to any sort of statistical model; however, although different settings can lead to very different results, estimating a multivariate mixture model cannot be detrimental with respect to the univariate approach; adding response variables (as well as the introduction of covariates) brings new information and improves the model identification: the standard errors of the estimates are expected to decrease and the optimization algorithm is more likely to converge to a “good” maximum and with fewer iterations.

## 4 The `mixglm` package for the R environment

We present the documentation of the `mixglm` package for mixtures of Normal/Poisson/Binomial distributions, forthcoming in the R environment. The description of the basic functions is provided, together with some examples. With respect to other software,<sup>(9)</sup> `mixglm` allows for a greater flexibility in the model specification; covariates may affect both the mixing distributions and the mixing proportions; each parameter is allowed to vary or to be constant across components, or to be an offset; the package also handles mixtures with partially classified observations. In `mixglm`, multivariate models are implemented: since the joint distribution is specified as the product of the conditional densities, the  $M$  outcomes are allowed to belong to different parametric families. Optionally, the starting values for the EM algorithm are provided by a genetic algorithm; the standard errors of the estimates may be computed using the asymptotic covariance matrix (with analytical evaluation of the Hessian of the log-likelihood function) or with a bootstrap approach (parametric or nonparametric); a function for bootstrap-based selection of the optimal number of components is provided; fitted values and conditional membership probabilities are also available.

### *Index*

<code>mixglm</code>	Fitting of univariate mixture models
<code>MULTmixglm</code>	Fitting of multivariate mixture models
<code>gen.start</code>	Genetic algorithm for univariate and multivariate mixture models
<code>gen.search</code>	Options for <code>gen.start</code> in <code>mixglm</code> and <code>MULTmixglm</code>
<code>asy.ci</code>	Asymptotic standard errors and confidence intervals for finite mixtures
<code>boot.ci</code>	Bootstrap standard errors and confidence intervals for finite mixtures
<code>simulator</code>	Random numbers generation from mixtures of univariate distributions
<code>mult.simulator</code>	Random numbers generation from mixtures of multivariate distributions
<code>model.choice</code>	Model selection for finite mixtures bootstrapping the LRT statistic

---

<sup>(9)</sup> Within the R environment, other packages are `flexmix`, `mixreg`, `mixdist` and `mixtools`, which provide functions for estimation of different mixture models with bootstrap-based inference; some multivariate models are supported by `mixreg`. We can also mention `mclust` and `normix` (for normal models), `bayesmix` and `vayabelMix` (for bayesian mixture models), `mda` (discriminant analysis), `depmix` (Hidden Markov Models), `mixPHM` (mixtures of proportional hazard models). With respect to `mixglm`, the above packages are generally less flexible; in most of them, covariates cannot affect the mixing proportions and multivariate models are not available; tools for asymptotic/bootstrap inference are rarely implemented.

`mixglm{mixglm}`

## Fitting Univariate Mixture Models

### Description

Univariate mixtures of Normal/Poisson/Binomial distributions via the EM algorithm.

### Usage

```
mixglm(y, x = NULL, x.p = NULL, k, ncomp = NULL, family = "normal",  
       b.0 = NULL, s.0 = NULL, p.0 = NULL, weights = NULL, Z = NULL,  
       offset.b = NULL, offset.s = NULL, offset.p = NULL,  
       maxit = 1000, epsilon = 1e-4, print.level = 1, method = NULL)
```

### Arguments

- `y` The response variable. For `binomial` family, the response can also be specified as a two-column matrix with the columns giving the number of successes and trials (if missing, the default for the number of trials is 1).
- `x` An optional  $n \times p$  matrix of covariates, to be used in the prediction of the expected values (the constant term will be included automatically).
- `x.p` An optional  $n \times q$  matrix of covariates, to be used in the prediction of the cluster membership probabilities (the constant term will be included automatically).
- `k` The number of components of the mixture (allowed  $k = 1$ ).
- `ncomp` A vector of length  $p + 1$  for `binomial` and `poisson` families,  $p + 2$  for `normal` family. For the intercept and for each covariate in `x`, where  $ncomp = k$  the coefficient is allowed to vary across components; where  $ncomp = 1$ , the parameter is the same for all components. In the normal model, the last element of `ncomp` contains the number of different standard deviations to be fitted. By default  $ncomp = (k, k, \dots, k)$ .
- `family` A character string containing the chosen family for the response variable ("normal", "poisson" or "binomial").
- `b.0` Numerical vector of optional starting values for the coefficients of covariates in `x` (if `x` is `NULL`, only the intercepts need to be provided). The

starting points must be ordered by covariate (the intercept first), then by component. For each covariate, if the corresponding `ncomp` is 1, only one starting value must be provided.

<code>s.0</code>	Numerical vector of optional starting values for the standard deviations (ignored if <code>family</code> is not "normal").
<code>p.0</code>	Numerical vector of optional starting values for the mixing proportions. If <code>x.p</code> is <code>NULL</code> , <code>p.0</code> must contain <code>k</code> proportions summing to 1; otherwise, <code>p.0</code> includes the $(q + 1)(k - 1)$ parameters (ordered by component, then by covariate in <code>x.p</code> , including the intercept) of a multinomial logistic model for the mixing proportions (the last component is taken as baseline).
<code>weights</code>	Optional weights to be used in the fitting procedure.
<code>z</code>	Optional vector of length <code>n</code> , including the cluster membership of each observation (a value between 1 and <code>k</code> , 0 if unknown); if <code>NULL</code> , the cluster membership is assumed to be unknown for all observations.
<code>offset.b,</code> <code>offset.s,</code> <code>offset.p</code>	Optional vectors – defined in the same way as <code>b.0</code> , <code>s.0</code> , <code>p.0</code> – containing <code>NA</code> for free parameters, and an offset for parameters whose value is known; if <code>NULL</code> , no offsets are used in the fitting procedure.
<code>maxit</code>	Maximum number of iterations for the EM algorithm (allowed <code>maxit = Inf</code> ).
<code>epsilon</code>	Tolerance for the EM algorithm (see <code>Details</code> ).
<code>print.level</code>	An integer from 0 to 3, indicating how often the procedure must print the progress.
<code>method</code>	The method to be used in the selection of the starting points. Must be <code>NULL</code> (randomly chosen starting values) or <code>gen.search(...)</code> if a genetic algorithm is required. See <code>gen.start</code> for the parameters of the genetic algorithm.

## Details

The standard EM algorithm is used in the estimation, starting from the chosen initial parameters; where the starting points are missing, randomly chosen values are used; if `gen.search` is used as `method`, the starting points are used as `good.subject` and the EM is runned from all the `n.elit` “best” individuals (see `gen.start` for details). The stopping condition of the EM is reached when the maximum absolute change in the parameters

vector is smaller than `epsilon`; otherwise, the procedure is stopped when the maximum number of iterations has been reached. If an empty component is found, the EM is restarted with new starting values. The canonical link for the chosen family is used (note that the parameters are provided in the link scale also in models without covariates).

## Value

An object of class "mixglm". The function `summary` (i.e., `summary.mixglm`) can be used to obtain or print a summary of the results. The accessor function `fitted.mixglm` can be used to extract the fitted  $y$  values (component by component) and the fitted cluster membership probabilities. The functions `post` and `cluster` provide the conditional membership probabilities and the consequent cluster assignment. Confidence intervals on the parameters are provided by `asy.ci` and `boot.ci`.

An object of class "mixglm" is a list containing at least the following elements:

<code>y, x, x.p</code>	The outcome variable and the model matrices.
<code>weights</code>	The specified weights.
<code>z</code>	The specified vector of cluster membership labels.
<code>k</code>	The number of components of the mixture.
<code>ncomp</code>	The number of components, parameter by parameter.
<code>family</code>	The chosen parametric family.
<code>offset.b,</code> <code>offset.s,</code> <code>offset.p</code>	The model offsets.
<code>start</code>	The starting parameters; <code>start = (b.0, s.0, p.0)</code> .
<code>phi</code>	The final parameters estimate, in the same order as in <code>start</code> .
<code>loglik</code>	The observed log-likelihood at the last iteration.
<code>coef</code>	A summary of the parameters of the $k$ component distributions.
<code>p</code>	If <code>x.p = NULL</code> , the estimated mixing proportions. Otherwise, a list of two elements; the first contains the average probability of each cluster, the second is a summary of the logistic model.

In addition, if a normal model is fitted, `s` returns the estimated standard deviations.

## References

- McCullagh P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.
- McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, USA.
- Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.

## See Also

`gen.search` for using `gen.start` in the search of the starting points for the EM.  
`asy.ci/boot.ci` for asymptotic/bootstrap confidence intervals.  
The function `model.choice` can be used for inference on the number of components.  
`simulator` for random sampling from finite mixture distributions.  
`MULTmixglm` for multivariate mixture models.

## Examples

```
# mixture of two poisson distributions with parameters exp(-0.2), exp(4)
# and proportions 0.4 and 0.6

y <- NULL
p <- rbinom(500,1,0.4)
for(i in 1:500){
  if(p[i] == 1){y <- c(y, rpois(1, exp(-0.2)))}
  else{y <- c(y, rpois(1, exp(4)))}
}

m1 <- mixglm(y, k = 2, family = "poisson") # random starting values

m2 <- mixglm(y, k = 2, family = "poisson",
  b.0 = c(-0.2,4), p.0 = c(0.4,0.6)) # true parameters

m3 <- mixglm(y, k = 2, family = "poisson",
  method = gen.search()) # genetic algorithm

# mixture of two simple linear regression models, with common intercept
# and variance and known mixing proportions

y <- NULL
x <- runif(500)
p <- rbinom(500,1,0.5)
```

```
for(i in 1:500){
  if(p[i] == 1){y <- c(y, rnorm(1, 2 + 3*x[i], 0.7))}
  else{y <- c(y, rnorm(1, 2 - 0.5*x[i], 0.7))}
}

m2 <- mixglm(y, x = cbind(x), k = 2, ncomp = c(1,2,1), family = "normal",
  b.0 = c(2,3,-0.5), s.0 = 0.7, offset.p = c(0.5,0.5))
```



MULTmixglm{mixglm}

## Fitting Multivariate Mixture Models

### Description

Multivariate mixtures of Normal/Poisson/Binomial distributions via the EM algorithm.

### Usage

```
MULTmixglm(y1, y2, ..., X = list(), x.p = NULL, k, NCOMP = list(),  
  family = c(), B.0 = list(), S.0 = list(), p.0 = NULL,  
  weights = NULL, Z = NULL,  
  offset.B = list(), offset.S = list(), offset.p = NULL,  
  maxit = 1000, epsilon = 1e-4, print.level = 1, method = NULL)
```

### Arguments

- `y1, y2, ...` The  $M$  response variables. For `binomial` family, the response can also be specified as a two-column matrix with the columns giving the number of successes and trials (if missing, the default for the number of trials is 1).
- `X` An optional list of  $M$  (even different) model matrices to be used in the prediction of the expected values. `X[[h]]` must be `NULL` if no covariates are used for the prediction of the corresponding response variable; the constant term will be included automatically. In the prediction of each response variable, all the subsequent responses are automatically added at the end of the model matrix (see `Details`).
- `x.p` An optional  $n \times q$  matrix of covariates, to be used in the prediction of the cluster membership probabilities (the constant term will be included automatically).
- `k` The number of components of the mixture (allowed  $k = 1$ ). In some case, the parameters of one or more of the  $M$  response variables are unaffected by the latent cluster membership: for this reason, `k` can be a vector of length  $M$ , with `k[h] = 1` if the corresponding response follows a pure model, and `k[h] = k` otherwise.
- `NCOMP` A list of  $M$  `ncomp` vectors (see `ncomp` in `mixglm`); if `NCOMP[[h]]` is `NULL`, the default value will be used for the corresponding response variable. Note that in the prediction of each response variable, all the subsequent responses are used as covariates (their `ncomp` must be declared).

<code>family</code>	A vector of $M$ character strings containing the chosen <code>family</code> for the corresponding response variable ("normal", "poisson" or "binomial"). Mixed families are allowed (see <code>Details</code> ).
<code>B.0</code> , <code>S.0</code> , <code>p.0</code>	Optional starting values for the EM. <code>B.0</code> and <code>S.0</code> are lists of length $M$ , containing the <code>b.0</code> and <code>s.0</code> vectors for the response variables (see the <code>b.0</code> and <code>s.0</code> arguments in <code>mixglm</code> ). <code>B.0[[h]]</code> and <code>S.0[[h]]</code> may be <code>NULL</code> if no starting values are chosen for the corresponding response variable. <code>S.0[[h]]</code> is also <code>NULL</code> if the corresponding <code>family</code> is not "normal". <code>p.0</code> follows the same rules of the corresponding argument in <code>mixglm</code> . Note that in the prediction of each response variable, all the subsequent responses are used as covariates ( <code>B.0</code> must include the respective coefficients). Where the starting points are missing, randomly chosen values are used.
<code>weights</code>	Optional weights to be used in the fitting procedure.
<code>Z</code>	Optional vector of length $n$ , including the cluster membership of each observation (a value between 1 and $k$ , 0 if unknown); if <code>NULL</code> , the cluster membership is assumed to be unknown for all observations.
<code>offset.B</code> , <code>noffset.S</code> , <code>offset.p</code>	Optional arguments – defined in the same way as <code>B.0</code> , <code>S.0</code> , <code>p.0</code> – containing <code>NA</code> for free parameters, and an offset for parameters whose value is known; see also the corresponding arguments <code>offset.b</code> , <code>offset.s</code> , <code>offset.p</code> in <code>mixglm</code> .
<code>maxit</code>	Maximum number of iterations for the EM algorithm (allowed <code>maxit = Inf</code> ).
<code>epsilon</code>	Tolerance for the EM algorithm (see <code>Details</code> ).
<code>print.level</code>	An integer from 0 to 3, indicating how often the procedure must print the progress.
<code>method</code>	The method to be used in the selection of the starting points. Must be <code>NULL</code> (randomly chosen starting values) or <code>gen.search(...)</code> if a genetic algorithm is required. See <code>gen.start</code> for the parameters of the genetic algorithm.

## Details

The standard EM algorithm is used in the estimation, starting from the chosen initial parameters; where the starting points are missing, randomly chosen values are used; if `gen.search` is used as `method`, the starting points are used as `good.subject` and the EM is runned from all the `n.elit` “best” individuals (see `gen.start` for details). The stopping criterion is reached when the maximum absolute change in the parameters vector is smaller

than `epsilon`; otherwise, the procedure is stopped when the maximum number of iterations has been reached. If an empty component is found, the EM is restarted with new starting values. The joint distribution of  $(y_1, \dots, y_M)$  is specified as the product of the conditional densities; for example, if  $M = 3$ :

$$f(y_1, y_2, y_3) = f(y_1 | y_2, y_3) f(y_2 | y_3) f(y_3)$$

Mixed families are allowed; for example:

```
y3 ~ Poisson
y2|y3 ~ Normal
y1|y2, y3 ~ Binomial
```

Each response includes the subsequent in the linear predictor; the canonical link for the chosen family is used (note that the parameters are provided in the link scale also in models without covariates). If a response is supposed to not affect another response, an offset can be used to constrain one or more coefficients to be 0.

## Value

An object of class "mixglm". The function `summary` (i.e., `summary.mixglm`) can be used to obtain or print a summary of the results. The accessor function `fitted.mixglm` can be used to extract the fitted  $y$  values (ordered by response variable, component by component) and the fitted cluster membership probabilities. The functions `post` and `cluster` provide the conditional membership probabilities and the consequent cluster assignment. Confidence intervals on the parameters are provided by `asy.ci` and `boot.ci`. An object of class "mixglm" is a list containing at least the following elements:

<code>y, x</code>	Lists containing the $M$ outcome variables and model matrices.
<code>x.p</code>	The specified matrix of predictors for the mixing proportions.
<code>weights</code>	The specified weights.
<code>z</code>	The specified vector of cluster membership labels.
<code>k</code>	The number of components of the mixture.
<code>ncomp</code>	List containing the number of components, parameter by parameter.
<code>family</code>	The chosen parametric families.
<code>offset.b,</code> <code>offset.s,</code> <code>offset.p</code>	The model offsets.

<code>start</code>	The starting parameters, in the same order as in <code>phi</code> (see below).
<code>phi</code>	The final parameters estimate, ordered by response variable (for each response in $y_1, \dots, y_M$ , the coefficients of the regression model are returned – ordered by covariate, then by component – followed by the standard deviations where <code>family = "normal"</code> ). At the end of <code>phi</code> , the mixing proportions are returned (if <code>x.p</code> is not <code>NULL</code> , the parameters of the logistic model are ordered by component, then by covariate in <code>x.p</code> ; the last component is taken as baseline).
<code>loglik</code>	The observed log-likelihood at the last iteration.
<code>coef</code>	A summary of the parameters of the $k$ component distributions.
<code>p</code>	If <code>x.p = NULL</code> , the estimated mixing proportions. Otherwise, a list of two elements; the first contains the average probability of each cluster, the second is a summary of the logistic model.

In addition, if a normal distribution is assumed for one or more response variables, `s` returns a list containing the estimated standard deviations (`NULL` for the responses with non normal distribution).

## References

- McCullagh P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.
- McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, USA.
- Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.

## See Also

`gen.search` for using `gen.start` in the search of the starting points for the EM.  
`asy.ci/boot.ci` for asymptotic/bootstrap confidence intervals.  
The function `model.choice` can be used for inference on the number of components.  
`mult.simulator` for random sampling from finite mixtures of multivariate distributions.  
`mixglm` for univariate mixture models.

## Examples

```
##### EXAMPLE 1 #####

# mixture of two bivariate normal distributions with diagonal covariance
# matrices.
```

```

y1 <- y2 <- NULL
p <- rbinom(500,1,0.4)
for(i in 1:500){
  if(p[i] == 1){
    y1 <- c(y1, rnorm(1,0,1))
    y2 <- c(y2, rnorm(1,1,0.8))
  }
  else{
    y1 <- c(y1, rnorm(1,2,0.4))
    y2 <- c(y2, rnorm(1,4,0.8))
  }
}

# True parameters

B0 <- list(c(0,2,0,0),c(1,4))
S0 <- list(c(1,0.4),c(0.8,0.8))
p0 <- c(0.4,0.6)

# Unconstrained model (true values used as starting points)

m1 <- MULTmixglm(y1, y2, k = 2, family = c("normal","normal"),
  B.0 = B0, S.0 = S0, p.0 = p0)

# y1 and y2 are known to be independent in both clusters

m2 <- MULTmixglm(y1, y2, k = 2, family = c("normal","normal"),
  B.0 = B0, S.0 = S0, p.0 = p0,
  offset.B = list(c(NA,NA,0,0),NULL))

# y2 has the same standard deviation in both components

m3 <- MULTmixglm(y1, y2, k = 2, family = c("normal","normal"),
  NCOMP = list(NULL,c(2,1)),
  B.0 = B0, S.0 = list(c(1,0.4),c(0.8)), p.0 = p0,
  offset.B = list(c(NA,NA,0,0),NULL))

##### EXAMPLE 2 #####

# y2 ~ Binomial(5,0.4)
# in the first component, y1 | y2, x ~ Poisson(exp(-0.5 + x + 0.2*y2))
# in the second component, y1 | y2, x ~ Poisson(exp(-0.5 + x + 0.6*y2))
# The mixing proportions are 0.4 and 0.6.

```

```

x <- rnorm(500)
y2 <- rbinom(500, 5, 0.4)

y1 <- NULL
p <- rbinom(500,1,0.4)
for(i in 1:500){
  if(p[i] == 1){y1 <- c(y1, rpois(1, exp(-0.5 + x[i] + 0.2*y2[i])))}
  else{y1 <- c(y1, rpois(1, exp(-0.5 + x[i] + 0.6*y2[i])))}
}

# Unconstrained model

m1 <- MULTmixglm(y1, cbind(y2,5), X = list(cbind(x), NULL),
  k = 2, family = c("poisson","binomial"))

# y2 follows a pure model

m2 <- MULTmixglm(y1, cbind(y2,5), X = list(cbind(x), NULL),
  k = c(2,1), family = c("poisson","binomial"))

# In y1, the only mixed parameter is the coefficient of y2

m3 <- MULTmixglm(y1, cbind(y2,5), X = list(cbind(x), NULL),
k = c(2,1), NCOMP = list(c(1,1,2),NULL), family = c("poisson","binomial"))

```

```
gen.start{mixglm}
```

## Genetic Algorithm for the starting values of Finite Mixture Models

### Description

Genetic algorithm for univariate and multivariate finite mixture models.

### Usage

```
gen.start(y, x = cbind(), x.p = cbind(), k, ncomp = NULL, good.subject = NULL,  
  family, offset.b = NULL, offset.s = NULL, offset.p = NULL,  
  weights = NULL, Z = NULL,  
  Min = NULL, Max = NULL, N.pop = NULL, N.gen = 150, print.level = 1,  
  n.elit = 4, p.crossing = 0.5, omega = 0.4,  
  p.mut1 = 0.2, p.mut2 = 0.2, p.mut3 = 0.1,  
  stop.time = 10)
```

### Arguments

- |                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>y</code>                                                                                                                                                                                                                                                | In univariate models, the response variable as in <code>mixglm</code> . In multivariate models, a list of length <code>M</code> containing the response variables.                                                                                                                                                                                                                                                                                                          |
| <code>x</code> , <code>x.p</code> , <code>k</code> ,<br><code>ncomp</code> ,<br><code>family</code> ,<br><code>offset.b</code> ,<br><code>offset.s</code> ,<br><code>offset.p</code> ,<br><code>weights</code> , <code>Z</code> ,<br><code>print.level</code> | In univariate models, the same as the corresponding arguments in <code>mixglm</code> ; in multivariate models, the same as in <code>MULTmixglm</code> , with <code>x</code> , <code>ncomp</code> , <code>offset.b</code> and <code>offset.s</code> corresponding to <code>X</code> , <code>NCOMP</code> , <code>offset.B</code> and <code>offset.S</code> , respectively.                                                                                                   |
| <code>good.subject</code>                                                                                                                                                                                                                                     | A “good” individual (parameters vector), which contributes to the population’s genetic heritage. Must be a vector of length <code>npar</code> (the number of parameters), with the same structure of the <code>phi</code> outcome in <code>mixglm</code> (for univariate models) or <code>MULTmixglm</code> (in the multivariate case).                                                                                                                                     |
| <code>Min</code> , <code>Max</code>                                                                                                                                                                                                                           | Vectors of length <code>npar</code> (the number of parameters). <code>Min</code> and <code>Max</code> represent the range used in the random generation of each gene (parameter) from a Uniform distribution. Default values are $(0, 3)$ for standard deviations, $(0, 1)$ for mixing proportions (if <code>x.p</code> is <code>NULL</code> ), $(-3, 3)$ for other parameters. For the offsets, <code>Min</code> and <code>Max</code> must coincide with the offset value. |

<code>N.pop</code>	The population size. Default is 8 times the number of parameters.
<code>N.gen</code>	Maximum number of generations.
<code>n.elit</code>	Number of elitists (the best individuals of a generation, preserved in the subsequent one; they ensure the monotonicity of the algorithm).
<code>p.crossing</code>	Probability of crossing-over (a couple of parents swaps each gene according to a binomial trial with parameter <code>p.crossing</code> ).
<code>omega</code>	Selection parameter. Each individual is selected for breeding with probability proportional to $(i)^\omega$ , where $(i)$ is the individual's position after ordering the population by increasing fitness (log-likelihood). If <code>omega</code> = 0, there is no selection; as <code>omega</code> increases, the selection becomes more and more severe.
<code>p.mut1</code>	Probability of type 1 mutation: the selected individuals have a random shift in all genes. If the new fitness is smaller than the original, the old individual is restored.
<code>p.mut2</code>	Probability of type 2 mutation: the <code>N.pop*p.mut2</code> worst individuals are completely replaced by new chromosomes.
<code>p.mut3</code>	Probability of type 3 mutation: the selected individuals are improved with one EM iteration.
<code>stop.time</code>	Stopping condition: after <code>stop.time</code> generations without significant (> 0.2%) improvements in the higher population fitness, the algorithm is stopped.

## Details

The log-likelihood function of finite mixture models generally has an unknown number of local maxima. A genetic algorithm is used in solving this optimization problem. This function can be used within the `mixglm` or `MULTmixglm` procedure via `gen.search`. Excluding the operational parameters of the algorithm (`Min`, `Max`, `N.pop`, `N.gen`, `n.elit`, `p.crossing`, `omega`, `p.mut1`, `p.mut2`, `p.mut3`, `stop.time`), the others arguments are the same as in `mixglm` and `MULTmixglm`. A quite small value of `omega`, together with high mutation rates (`p.mut1`, `p.mut2`, `p.mut3`) generally leads to a better result in terms of fitness (log-likelihood).



## Value

`gen.start` returns a list with the following elements:

<code>T</code>	Number of generations reached by the algorithm.
<code>maxfit</code>	The sequence of the maximum observed fitness in the <code>T</code> generations.
<code>chrom</code>	The population at the last iteration, composed by <code>N.pop</code> parameters vectors.
<code>elit</code>	The <code>n.elit</code> elitists at the last iteration.

## References

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.

## See Also

`gen.search` for using `gen.start` in the search of the starting points for the EM within the `mixglm` and `MULTmixglm` functions.

## Examples

```
# Mixture of two poisson distributions; we use the true values as  
# "good.subject".
```

```
y <- NULL  
p <- rbinom(200,1,0.3)  
for(i in 1:200){  
  if(p[i] == 1){y <- c(y, rpois(1,exp(0.2)))}  
  else{y <- c(y, rpois(1, exp(1)))}  
}
```

```
m <- gen.start(y, family = "poisson", k = 2,  
  good.subject = c(0.2,1,0.3,0.7),  
  Min = c(-1,-1,0,0), Max = c(2,2,1,1))
```

```
# the same as above, but with known mixing proportions
```

```
m <- gen.start(y, family = "poisson", k = 2,  
  good.subject = c(0.2,1,0.3,0.7),  
  Min = c(-1,-1,0.3,0.7), Max = c(2,2,0.3,0.7),  
  offset.p = c(0.3,0.7))
```

```
gen.search{mixglm}
```

## Options for `gen.start` in `mixglm` and `MULTmixglm`

### Description

This function can be passed as the `method` argument in `mixglm` and `MULTmixglm`; the EM will start after the `gen.start` procedure. The arguments of `gen.search` are the operational parameters for the genetic algorithm.

### Usage

```
gen.search(Min = NULL, Max = NULL, N.pop = NULL, N.gen = 150,  
  n.elit = 4, p.crossing = 0.5, omega = 0.4,  
  p.mut1 = 0.2, p.mut2 = 0.2, p.mut3 = 0.1,  
  stop.time = 10)
```

### Arguments

The arguments represent the operational parameters of `gen.start`; see `gen.start` for details.

### Details

Calls `gen.start` in `mixglm` and `MULTmixglm`, providing the operational parameters of the genetic algorithm.

### Value

The output is a list containing the values in input, to be transmitted to the `gen.start` procedure.

### See Also

`mixglm`, `MULTmixglm`, `gen.start`.

```
asy.ci{mixglm}
```

## Asymptotic standard errors for Finite Mixture Models

### Description

Asymptotic standard errors and confidence intervals for the parameters of a finite mixture model, via analytical evaluation of the Hessian matrix of the log-Likelihood function.

### Usage

```
asy.ci(model, conf = 0.95)
```

### Arguments

`model` An object of class "mixglm".

`conf` The desired nominal coverage of the confidence intervals.

### Details

The Hessian matrix of the log-Likelihood function is evaluated using analytical derivatives. This ensures an accurate computation of the asymptotic standard errors, even for very complicate models. The confidence intervals are computed assuming that the estimators follow a normal distribution.

### Value

The output is a matrix with 4 columns: the first is the parameters vector; in the second, the estimated standard errors are returned (0 for the offsets); the last two columns contain the lower and upper bounds of the confidence interval, according to the chosen nominal coverage.

### References

McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, USA.

Casella, B., Berger, R.L. (2002). *Statistical Inference*. Duxbury, USA.

### See Also

`mixglm`, `MULTmixglm`; `boot.ci` for bootstrap confidence intervals.

### Examples

```
asy.ci(m1) # m1 is a mixglm object
```

`boot.ci{mixglm}`

## Bootstrap confidence intervals for Finite Mixture Models

### Description

Bootstrap standard errors and confidence intervals for finite mixture models.

### Usage

```
boot.ci(model, B = 100, type = "parametric", conf = 0.95,  
        epsilon = 1e-4, print.level = 1, maxit = 1000)
```

### Arguments

<code>model</code>	An object of class "mixglm".
<code>B</code>	The number of bootstrap samples.
<code>type</code>	The resampling method. Must be "parametric" or "nonparametric".
<code>conf</code>	The desired nominal coverage of the confidence intervals.
<code>epsilon</code>	The tolerance for the EM algorithm (see <code>mixglm</code> and <code>MULTmixglm</code> )
<code>print.level</code>	An integer between 0 and 2, indicating how often the procedure must print the progress. If <code>print.level = 2</code> , histograms of the empirical sampling distributions of the estimators are plotted. See also <code>Value</code> .
<code>maxit</code>	Maximum number of iterations for the EM algorithm.

### Details

The covariance matrix of the estimates is evaluated using a resampling approach. For each of `B` simulated data sets, the same model is estimated (the true values are used as starting points). If the "parametric" bootstrap is chosen, the data sets are obtained as random samples from the estimated model; with "nonparametric" bootstrap, each data set is sampled with replacement from the original one. Different confidence intervals are provided; see `Value` for further details.

## Value

The output is a matrix with 6 columns: the first contains the parameters vector; in the second, the estimated standard errors are returned (0 for the offsets); 3rd and 4th columns contain the lower and upper bounds of the confidence interval, based on the percentiles of the bootstrap sampling distributions. The last two columns return the lower and upper bounds of a HDF (Highest Density Function) confidence interval, obtained fitting a flexible density on the sampling distributions and computing the theoretical percentiles (the couple  $a, b$  such that  $f(a) = f(b)$  and  $P(a \leq t \leq b) = 1 - \alpha$ , where  $t$  is the sample statistic and  $\alpha$  the nominal coverage probability). A three-parameters Gamma is used for coefficients (if the sampling distribution has a negative skewness, the same density is fitted on  $-t$ ); a central Gamma is used for standard deviations and a Dirichlet for mixing proportions. If the superimposed densities provide a good approximation of the empirical sampling distribution, the HDF interval is the shortest among all intervals with the same coverage probability; moreover, it is expected to partially obviate to a small number of bootstrap replications (`B`). Setting `print.level = 2` will generate the histograms of the sampling distributions of the estimators, including the fitted densities used in the HDF interval.

## References

- Casella, B., Berger, R.L. (2002). *Statistical Inference*. Duxbury, USA.
- Efron, B., Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, USA.

## See Also

`mixglm`, `MULTmixglm`.  
`asy.ci` for asymptotic confidence intervals.  
`simulator`, `mult.simulator` for simulating from finite mixture models.

## Examples

```
boot.ci(m1) # m1 is a mixglm object
```

simulator{mixglm}

## Random numbers generation from univariate Finite Mixture Models

### Description

Simulation from univariate mixtures of normal, poisson and binomial distributions.

### Usage

```
simulator(phi, x = NULL, x.p = NULL, k, ncomp = NULL, family = "normal",  
  bin.trial = 1, nrepl = NULL, membership = FALSE)
```

### Arguments

- |           |                                                                                                                                                                                                                                                                                                                            |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| phi       | The parameters vector (the location parameters – ordered by predictor, then by component; if family = "normal", the scale parameters – ordered by component; the mixing proportions); see the analogous output of mixglm for details.                                                                                      |
| x         | The optional $n \times p$ model matrix for the expected values (not including the constant term).                                                                                                                                                                                                                          |
| x.p       | The optional $n \times q$ model matrix for prediction of the mixing proportions (not including the constant term).                                                                                                                                                                                                         |
| k         | The number of components of the mixture.                                                                                                                                                                                                                                                                                   |
| ncomp     | The ncomp vector (see mixglm for details).                                                                                                                                                                                                                                                                                 |
| family    | The chosen family for the response variable ("normal", "poisson" or "binomial").                                                                                                                                                                                                                                           |
| bin.trial | The number of trials (ignored if family is not "binomial"). Must be a positive integer (default is 1) or a vector of length n containing the number of trials to be used for each randomly generated observation.                                                                                                          |
| nrepl     | Number of observations to be generated. By default, nrepl is automatically determined from n (the dimensions of x, x.p, or bin.trial). Otherwise, nrepl can be a multiple of n (in this case, the values in x, x.p, bin.trial are used repeatedly). In model without covariates, where n is unknown, by default nrepl = 1. |

`membership` Logical. If `TRUE`, the realized component membership of each observation is returned.

## Details

According to the mixing proportions, the cluster membership is drawn; for each observation, the  $y$  is sampled from the respective mixing distribution.

## Value

If `membership = FALSE` (default), the realized  $y$  is returned. Otherwise, the output is a list containing the following elements:

`y` The realized `nrepl`-dimensional  $y$  vector.

`cluster` The realized cluster membership for each observation.

## See Also

`mixglm` for the `x`, `x.p`, `k`, `ncomp`, `family` arguments. The parameters vector `phi` has the same structure as in the `mixglm` output.

`mult.simulator` for simulation of multivariate mixture models.

## Examples

```
##### Example 1 #####

# Mixture of 3 binomial distributions. The number of trials varies across
# observations (5 in the first 100, 10 in the remaining 200).
# The probability of success in each trial is (0.2,0.4,0.6) in the three
# components, respectively. The mixing proportions are 0.1,0.3,0.6.

# The parameters in phi must be in the logit scale.

b1 <- log(0.2/(1 - 0.2))
b2 <- log(0.4/(1 - 0.4))
b3 <- log(0.6/(1 - 0.6))

phi <- c(b1, b2, b3, 0.1, 0.3, 0.6)
trials <- c(rep(5,100), rep(10,200))

y <- simulator(phi, k = 3, family = "binomial", bin.trial = trials)
```

```

##### Example 2 #####

# Mixture of two simple linear regressions with common dispersion parameter
# sigma = 0.6.
# In the first component,  $y = 2 + 3x + e$ 
# In the second component,  $y = -1 + 2x + e$ 
#  $e \sim N(0, 0.6^2)$ .

x <- rnorm(300)
y <- simulator(phi = c(2, -1, 3, 2, 0.6, 0.6), x = cbind(x), k = 2,
              family = "normal")

# Or, alternatively:

y <- simulator(phi = c(2, -1, 3, 2, 0.6), x = cbind(x), k = 2,
              ncomp = c(2, 2, 1), family = "normal")

```



mult.simulator{mixglm}

## Random numbers generation from multivariate Finite Mixture Models

### Description

Simulation from multivariate mixtures of normal, poisson and binomial distributions.

### Usage

```
mult.simulator(phi, X = list(), x.p = NULL, k, NCOMP = list(), family,  
  bin.trial = list(), nrepl = NULL, membership = FALSE)
```

### Arguments

phi	The parameters vector, ordered by response variable (the location parameters – ordered by predictor, then by component; where family = "normal", the scale parameters – ordered by component); at the end, the mixing proportions. The joint density is specified as in MULTmixglm; see the analogous output of MULTmixglm for details.
X	The optional list of model matrices for the expected value (not including the constant term); see the analogous argument in MULTmixglm.
x.p	The optional $n \times q$ model matrix for prediction of the mixing proportions (not including the constant term).
k	The number of components of the mixture (see the analogous argument in MULTmixglm).
NCOMP	The NCOMP list (see MULTmixglm for details).
family	Vector of length M containing the chosen families for the response variables ("normal", "poisson" or "binomial"). See the analogous argument in MULTmixglm.
bin.trial	A list of binomial trials vectors (NULL where family is not "binomial"). Each element of bin.trial must be a positive integer (default is 1) or a vector of length n containing the number of trials to be used for each randomly generated observation.
nrepl	Number of observations to be generated. By default, nrepl is automatically determined from n (the number of observations desumed from X,

`x.p`, or `bin.trial`). Otherwise, `nrepl` can be a multiple of `n` (in this case, the values in `X`, `x.p`, `bin.trial` are used repeatedly). In model without covariates, where `n` is unknown, by default `nrepl = 1`.

`membership` Logical. If `TRUE`, the realized component membership of each observation is returned.

## Details

According to the mixing proportions, the cluster membership is drawn; for each observation, the  $y$  are sampled from the respective mixing distribution. The  $M^{\text{th}}$  response variable is generated first; conditioned on the realized  $y_M$ , the  $y_{M-1}$  is simulated, and so on. See also `MULTmixglm` for the model specification.

## Value

If `membership = FALSE` (default), a list containing the realized  $y_1, \dots, y_M$  is returned. Otherwise, the output is a list containing the following elements:

`y` The realized  $M$ -dimensional list of `nrepl`-dimensional  $y$  vectors.

`cluster` The realized cluster membership for each observation.

## See Also

`MULTmixglm` and `mixglm` for the `X`, `x.p`, `k`, `NCOMP`, `family` arguments. The parameters vector `phi` has the same structure as in the `MULTmixglm` output. `simulator` for random sampling from univariate mixture models.

## Examples

```
##### EXAMPLE 1 #####

# mixture of two bivariate normal distributions with diagonal covariance
# matrices.
# In the first component, E(y1) = 0, Sd(y1) = 1; E(y2) = 1, Sd(y2) = 0.8.
# In the second component, E(y1) = 2, Sd(y1) = 0.4; E(y2) = 4, Sd(y2) = 0.8.
# phi has the following composition:
# phi = (E(y1 | y2), Sd(y1 | y2), E(y2), Sd(y2), p) # p are the proportions

phi <- c(0,2,0,0, # linear regression of y1 on y2
        1,0.4,   # Sd(y1 | y2)
        1,4,     # E(y2)
        0.8,0.8, # Sd(y2)
        0.4,0.6) # mixing proportions
```

```

Y <- mult.simulator(phi, k = 2, family = c("normal","normal"), nrepl = 200)

# Alternatively:

Y <- mult.simulator(phi = c(0,2,0, 1,0.4, 1,4, 0.8, 0.4,0.6),
  NCOMP = list(c(2,1,2),c(2,1)), k = 2, family = c("normal","normal"),
  nrepl = 200)

y1 <- Y[[1]]
y2 <- Y[[2]]

##### EXAMPLE 2 #####

# y2 ~ Binomial(5,0.4)
# in the first component, y1 | y2, x ~ Poisson(exp(-0.5 + x + 0.2*y2))
# in the second component, y1 | y2, x ~ Poisson(exp(-0.5 + x + 0.6*y2))
# The mixing proportions are 0.4 and 0.6.

x <- rnorm(500)
phi <- c(-0.5,1,0.2,0.6, log(0.4/(1 - 0.4)), 0.4,0.6)

Y <- mult.simulator(phi, X = list(cbind(x),NULL), k = c(2,1),
  NCOMP = list(c(1,1,2), NULL), family = c("poisson","binomial"),
  bin.trial = list(NULL,5))

y1 <- Y[[1]]
y2 <- Y[[2]]

```

`model.choice{mixglm}`

## Model selection for finite mixtures bootstrapping the Likelihood Ratio Statistic

### Description

Provides a bootstrap approach to the model selection; see `Details`.

### Usage

```
model.choice(model1,model2,...,alpha = 0.05, B = 100,  
  epsilon = 1e-3, maxit = 1000, print.level = 1)
```

### Arguments

<code>model1,</code> <code>model2,...</code>	Objects of class "mixglm".
<code>alpha</code>	The nominal significance level for the test.
<code>B</code>	Number of bootstrap samples.
<code>epsilon</code>	Tolerance criterion for assessing the convergence of the EM (see <code>mixglm</code> ).
<code>maxit</code>	Maximum number of iterations for the EM algorithm.
<code>print.level</code>	An integer between 0 and 2, indicating how often the procedure must print the progress. If <code>print.level = 2</code> , the histograms of the bootstrap samples are plotted.

### Details

Working with finite mixture models, a common problem is testing for the number ( $k$ ) of components. The asymptotic theory of the Likelihood Ratio Test (LRT) is not valid in this case, since the sampling distribution of the LRT statistic does not generally tend to a chi-square distribution. A possible solution is to rely on other selection criteria (AIC, BIC...); alternatively, the sampling distribution of the LRT statistic may be evaluated under a bootstrap approach. The models to be compared should be nested; models with the same value of  $k$  but different constraints on parameters may be compared. Note that `model1,model2,...` should be ordered by increasing observed log-likelihood; otherwise, an automatic reordering is done: in the output of `model.choice`, `model1` is always the one with the smallest observed log-likelihood, and so on.

## Value

The output is a list containing the following elements:

<code>win</code>	The chosen model among <code>model1, model2, ...</code>
<code>k</code>	The number of components of the chosen model.
<code>ncomp</code>	The <code>ncomp</code> vector/list of the chosen model.
<code>x, x.p</code>	The column names for <code>x</code> and <code>x.p</code> in the chosen model.
<code>offset</code>	The offsets of the chosen model.

The procedure automatically prints the key results during the computation.

## References

McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, USA.

Efron, B., Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

## See Also

`mixglm`, `MULTmixglm`.

## Examples

```
# True model: mixture of two normal distributions with means 0 and 2
# and with common variance 1. The mixing proportions are 0.7 and 0.3.
```

```
y <- NULL
p <- rbinom(200, 1, 0.7)
for(i in 1:200){
  if(p[i] == 1){y <- c(y, rnorm(1,0,1))}
  else{y <- c(y, rnorm(1,2,1))}
}

m1 <- mixglm(y, k = 1, family = "normal") # pure model
m2 <- mixglm(y, k = 2, family = "normal", ncomp = c(2,1)) # true model
m3 <- mixglm(y, k = 2, family = "normal") # unconstrained variances
m4 <- mixglm(y, k = 3, family = "normal") # overparametrized

m <- model.choice(m1,m2,m3,m4, B = 50, print.level = 2)
```

## 5 Analysis of causal effects of job-training programs

### Introduction

Estimating causal effects of interventions is often the focus of empirical studies in medicine and the social sciences. The only generally accepted approach for inferring causality requires that treatment receipt is randomized. Experiments, however, and social experiments in particular, often suffer from a number of complications, most notably non-compliance with assigned treatment, missing outcomes, and truncation by death when the outcome is not always well-defined.

The evaluation of the training programs enables the policy-makers to make a cost-benefit analysis, comparing the estimated effects of the programs with their cost to the public. We evaluate the effects of a randomized job-training program, Job Corps, which stands out as the largest, most comprehensive US education and job training program for disadvantaged youths between the ages of 16 and 24; for our analysis, we use data from the National Job Corps Study, conducted by Mathematica Policy Research, Inc. The study is based on a national random sample of all eligible applicants in late 1994 and 1995. Sampled youths were assigned randomly to a program group or a control group. Consistently with the program's aim, key outcomes of interest are: employment status, total earnings, and wages. Usually, the effects on employment and total earnings are of main interest; the effect on wages – unlike the effect on total earnings – reflects the increase in human capital due to the training program. In the empirical analysis, we focus on the effect of the program on wages and employment.

In the study all three complications are present, namely a) compliance with assigned treatment was not perfect, as only 64% of those assigned to the program group effectively enrolled in Job Corps; b) due to attrition, outcome is missing on some participants; c) wages are truncated by death, meaning no wage is defined for those who are not employed.

Previous studies on these data neglected noncompliance, by focusing on intention-to-treat (ITT) effects of being offered participation in Job Corps (Lee, 2008; Zhang et al., 2008a and 2008b; Flores-Lagunes et al., 2007). Being in a all-or-none compliance setting and with access to Job Corps being denied to those assigned to the control group, individuals can be classified as compliers or never-takers (Angrist et al., 1996); in this setting the ITT effect, under a plausible outcome exclusion restriction assumption, can be regarded as being conservative for the effect of treatment receipt, i.e., it is possibly diluted by non-compliance to treatment assignment. This may be a reason why, so far, negligible effects of Job Corps were found on employment and wages, especially in the long run.

Here, we want to take account of three complications, namely noncompliance, truncation of wages and missing outcomes, in order to evaluate the effects of Job Corps on those who were not just assigned but also participated in the program, i.e., the compliers.

The framework we adopt uses potential outcomes to define causal effects regardless of the mode of inference, often referred to as the Rubin Causal Model (RCM; Holland, 1986); causal effects are defined by comparisons of potential outcomes for a common set of units

(Rubin, 1974, 1978, 2005). We apply principal stratification (PS; Frangakis and Rubin, 2002), which was originally introduced to address post-treatment complications, within RCM.

The framework can be applied in various contexts, leading to both parametric and semi(non)parametric inference, depending on the set of assumptions that can be reasonably maintained, as well as whether point (full) or partial identification is to be achieved.

Few papers have dealt with more than one complication simultaneously. In general, the assumptions being considered are more complicated than those in the presence of each of the complications separately.

In this work, we develop a likelihood-based approach to estimate the effect on employment and wages for the compliers. We conduct a likelihood-based analysis using the EM algorithm, proposing different ways of improving computational efficiency and identifiability using the theory of finite mixture models. We maintain the assumption of exclusion restriction, while monotonicity of compliance holds by design. We do not however impose monotonicity of truncation. Thereby, following Frangakis and Rubin (2002), we classify the individuals into six principal strata according to the joint values of the potential compliance and employment status when assigned to be trained and when not assigned to be trained. Our causal estimands are: the average effect on employment for compliers and the average effect on wages for compliers who are employed irrespective of treatment assignment.

Results show that both these effects are positive for compliers and that there is a group of participants for whom participation is detrimental in terms of employment. This group, however, becomes negligible in the long run.

The chapter proceeds as follows. Section 5.1 presents the general framework of the Rubin Causal Model; Section 5.2 discusses the issue of noncompliance with treatment assignment; in Section 5.3, we present the general approach to the missing data problem; Section 5.4 is devoted to causal inference when an outcome is “truncated by death”. Section 5.5 presents the framework needed for simultaneously addressing both issues of noncompliance and truncation by death of potential wages, under the MAR assumption for the missing data mechanism. Section 5.6 illustrates the likelihood approach to the estimation of the average treatment effects on employment and wages. Section 5.7 presents the application to the Job Corps data; In Section 5.8, results are discussed and some concluding remarks are provided.

## 5.1 The Rubin Causal Model

The Rubin Causal Model (RCM; Holland, 1986) is a general framework for causal inference, proposed by Rubin in a series of articles (1974, 1975, 1976, 1977, 1978, 1979, 1980). The RCM represents now the dominant approach to the evaluation of the causal effects of a treatment in experimental and observational settings.

The RCM framework originates from Neyman’s approach (1923), where each unit has two “potential” outcomes, one if the unit is treated and the other if untreated: only one outcome (the one corresponding to the observed treatment assignment) is observed by the researcher. The use of the “potential outcomes” framework constitutes the first element of the RCM: at the unit level, causal effects are defined as a function (typically the difference) of (a) the outcomes that would be observed if the unit received the active treatment, and

(b) the values that would be observed if the unit were assigned to the control group. The second part of the RCM is the assignment mechanism, that is a probabilistic model describing how the  $N$  units are assigned to the treatment or to the control group. A third, optional part of the RCM is a set of distributional assumptions on the potential outcomes, allowing for a model-based inference.

According to the RCM, the potential outcomes are the values that would be observed on the *same* unit at the *same* time, under the two treatment conditions; the use of potential outcomes in defining causal effects has relevant implications in the notion itself of “causality” and characterizes causal inference as a missing data problem: since for each unit only one of two potential outcomes is realized, at least 50% of outcomes is unobserved. With these settings, it is impossible to learn about the causal effects from a single observation, because the causal effect involves the comparison of both potential outcomes. In order to make causal inference, multiple units (exposed to both treatment conditions) must be observed.

In many situations, it is reasonable to assume that the treatment assignment of each unit does not affect the potential outcomes of other units; this is known as the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1978, 1980, 1990); however, there are many examples in which SUTVA is not a plausible assumption, especially in observational settings.

For unit  $i$  ( $i = 1, \dots, N$ ), we denote with  $Y_i$  the observed outcome and with  $Z_i$  the actual treatment assignment (1 = treatment, 0 = control);  $\mathbf{Y}$  and  $\mathbf{Z}$  are the  $N$ -dimensional vectors of observed outcomes and assignment indicators, respectively. We denote as  $Y_i(\mathbf{Z})$  the potential outcome of unit  $i$ , given the vector of treatment assignments  $\mathbf{Z}$ ;  $\mathbf{Y}(\mathbf{Z})$  is the  $N$ -dimensional vector of potential outcomes: with this notation,  $\mathbf{Y}(0)$  represents the outcomes we would observe if all units were assigned to the control group; analogously,  $\mathbf{Y}(1)$  is the vector of outcomes we would observe under the treatment condition: clearly, we never observe the couple  $Y_i(0), Y_i(1)$ .

The SUTVA consists in the following statement:

- *Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)*

The potential outcome of each unit is unrelated with the treatment assignment of any other units, and there are no different versions of the treatment.

The SUTVA states that there is no interference between units: this allows us to write the potential outcome as  $Y_i(Z_i)$  instead of  $Y_i(\mathbf{Z})$ . This means that  $Y_i(1)$  is the outcome of unit  $i$  when assigned to the treatment, and  $Y_i(0)$  is the outcome for unit  $i$  when assigned to the control group. SUTVA also implies that there are no hidden versions of the treatment. When the SUTVA is not a plausible assumption, the interactions between units must be taken in account in drawing inference on the causal effects; making this assumption more realistic is the aim of experimental design; in observational studies, the plausibility of SUTVA must be evaluated according to the specific settings.

Using the potential outcomes notation the role of the assignment mechanism – the second component of the RCM – can be explicitly taken in account. Prior to the Rubin’s work, causal effects were often defined as parameters of a regression model, relating the observed outcome  $Y_i$  to an optional set of covariates ( $\mathbf{X}_i$ ) and to the treatment indicator  $Z_i$  ( $i = 1, \dots, N$ ); this was the standard approach in medical and social sciences: however, using the observed outcome notation completely neglects the role of the assignment mech-



anism (e.g., randomization). In causal inference, the assignment mechanism represents the probabilistic models for the “missing data” process; that is, it specifies the conditional probability of  $\mathbf{Z}$  and defines the design for how some potential outcomes are revealed and some others are unobserved.

In principle, the assignment mechanism may depend on both pre-treatment covariates ( $\mathbf{X}$ ) and potential outcomes  $\mathbf{Y}(1)$  and  $\mathbf{Y}(0)$ ; for example, individuals may optimize a function involving the expectation of potential outcomes (see, e.g., Imbens and Rubin, 2006). A special case is represented by the randomized experiments, where the assignment mechanism is “unconfounded” and “probabilistic”.

Unconfounded assignment mechanisms (Rubin, 1990) are free of dependence on either  $\mathbf{Y}(0)$  and  $\mathbf{Y}(1)$ :

$$P(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = P(\mathbf{Z} \mid \mathbf{X}) \quad [1]$$

Under a probabilistic assignment (Rubin, 1990), each unit has a positive probability of receiving either treatment:

$$0 < P(Z_i = 1 \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1 \quad [2]$$

The assignment mechanism is said to be “strongly ignorable” (Rosenbaum and Rubin, 1983) if satisfies [1] and [2]; if the conditional probability in [1] is free from missing but not from observed potential outcomes, the assignment mechanism is said to be ignorable:

$$P(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = P(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}_{\text{obs}}) \quad [3]$$

where  $\mathbf{Y}_{\text{obs}}$  is the vector of observed outcomes. Clearly, unconfounded mechanisms are ignorable, but an ignorable assignment can be confounded (e.g., in sequential experiments). Very often, the analysis of experiments relies on strong ignorability: this allows for a straightforward estimation of the causal effects; in this case, it is very common to have a “regular” assignment mechanism, where the probability of each assignment vector is proportional to the product of the propensity scores:

$$P(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) \propto \prod_{i=1}^N P(Z_i = 1 \mid \mathbf{X}_i) \quad [4]$$

A special case of randomization is when

$$P(Z_i = 1 \mid \mathbf{X}_i) = P(Z_i = 1) \quad [5]$$

that is, each unit is assigned to the treatment or to the control group according to a binomial trial, with the same probability for all units. This is common in experimental settings, whereas in observational studies (the so-called “natural” experiments) this assumption is generally not plausible. However, under regular assignment, units with the same propensity score are randomized into the two treatment conditions; matching on the propensity score (e.g., Rosenbaum and Rubin, 1984) or subclassifying on it (e.g., Rosenbaum and Rubin, 1985) reproduces the condition of an experimental design and makes the inference straightforward.

In this work, we will assume that both SUTVA and strongly ignorable assignment mechanism hold; in the analysis of the Job Corps Study (Section 5.7), randomization holds

by design and the no-interference between units may be reasonably assumed. For our settings, the randomization assumption may be written as follows:

- *Assumption 2: Random Assignment*

$P(\mathbf{Z} = \mathbf{c}) = P(\mathbf{Z} = \mathbf{c}')$  for all  $\mathbf{c}$  and  $\mathbf{c}'$  such that  $\mathbf{u}^T \mathbf{c} = \mathbf{u}^T \mathbf{c}'$ , where  $\mathbf{u}$  is the  $N$ -dimensional unit vector.

that is, all units have the same probability of being assigned to the treatment group.

We define the individual treatment effect as the difference of the two potential outcomes:

$$\delta_i^{(ZY)} = Y_i(1) - Y_i(0)$$

where the generic  $\delta_i^{(AB)}$  denotes the causal effect of  $A$  on  $B$  for unit  $i$ . Under SUTVA, however, we can define an average treatment effect (ATE) as follows:

$$\Delta^{(ZY)} = E[Y_i(1)] - E[Y_i(0)]$$

where the generic  $\Delta^{(AB)}$  denotes the average causal effect of  $A$  on  $B$ . In an experimental setting,  $\Delta^{(ZY)}$  can be consistently estimated using the difference in the sample means of the treatment and control group.

Very often, complications arise and this simple model must be extended to more complex settings. In Section 5.2 we will discuss how the RCM can be used in addressing the issue of noncompliance; the link between the Instrumental Variable estimator and the Rubin Causal Model will be analyzed.

## 5.2 Noncompliance in randomized studies

In many fields of science, researchers are often interested in evaluating the effectiveness of a new treatment. Experiments are accepted tools to infer on causal effects. The key feature of experiments is that units are randomly assigned to the treatment or to the control group; this ensures that treated and untreated units have the same distribution of the (observed and unobserved) individual characteristics; inference in this case is straightforward, because the sample means are unbiased estimates of expected outcomes in the two groups.

The theory of inference based on randomization (Neyman, 1923; Fisher, 1925) requires that all experimental units comply with the treatment assignment; in practice, noncompliance is a common issue, especially in experiments with human subjects. Different approaches have been proposed to deal with the nonrandom receipt of the treatment due to noncompliance. In the presence of this complication, comparing subjects by treatment received – rather than by treatment assigned – generally leads to a biased estimate of the treatment effect; this is also true with a “per-protocol” approach, where only units who comply with treatment assignment are included in the analysis (Robins and Greenland, 1994; Sheiner and Rubin, 1995; Barnard et al., 1998). For these reasons, the standard approach to noncompliance is to compare average outcomes by assignment – ignoring the compliance behavior – as if compliance had been perfect; this is often referred to as the intention-to-treat (ITT) analysis (Breslow, 1982; Fisher et al., 1990; Lee et al., 1991; Meier, 1991).

The problem of evaluating the effect of a binary treatment has a long history in both

econometrics and statistics. The econometric literature (Ashenfelter, 1978; Ashenfelter and Card, 1985; Heckman and Robb, 1985; Lalonde, 1986; Fraker and Maynard, 1987; Card and Sullivan, 1988; Manski, 1990) focuses on the issue of endogeneity (self-selection) in observational settings; units who choose to enroll in a training program are expected to be different from those who choose not to enroll: for this reason, a simple comparison of the average outcomes by treatment status – even adjusting for covariates – may lead to a wrong inference on the causal effect. A different approach dominates the statistical literature, which originates with the analysis of randomized experiments (Neyman, 1923; Fisher, 1925) and was developed by Rubin (1974, 1978, 2005). We adopt the general framework of potential outcomes; the issue of noncompliance is addressed using a principal stratification approach, where potential outcomes are compared for a common set of units – those who comply with treatment assignment.

PS can be seen as a generalization of the IV model proposed in Angrist et al. (1996) to address noncompliance in randomized studies; in what follows, we describe the relationship between the traditional IV structural equation model and the RCM.

#### *IV estimation in Structural Equation Models*

In a sample of  $N$  units, we suppose to random assign each unit to a treatment or to a control group; we denote with  $Z_i$  the observed assignment for unit  $i$ , and with  $D_i$  the received treatment: only if the compliance is perfect, we observe  $D_i = Z_i$ . As before, we also assume that the treatment is binary (1 = treatment, 0 = control): that is,  $D_i$  and  $Z_i$  are dichotomous; however, it is possible to generalize to more complex treatments and to settings with partial compliance. For all units, we observe an outcome variable  $Y_i$ .

The *dummy endogenous variable model* (see, e.g., Maddala, 1983) is defined as follows:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + \varepsilon_i \\ D_i^* &= \alpha_0 + \alpha_1 Z_i + v_i \end{aligned}$$

where

$$D_i = \begin{cases} 1 & \text{if } D_i^* > 0 \\ 0 & \text{if } D_i^* \leq 0 \end{cases}$$

The parameter of interest is  $\beta_1$ , which represents the causal effect of  $D$  on  $Y$ ; the model identification is driven by the following assumptions:

$$E[Z_i \varepsilon_i] = 0 \quad E[Z_i v_i] = 0$$

The absence of correlation between  $Z$  and  $v$  is typical of standard regression models; the assumption that  $Z$  and  $\varepsilon$  are uncorrelated – together with the absence of  $Z$  in the equation for  $Y$  – reflects the fact that any effect of  $Z$  on  $Y$  is through an effect of  $Z$  on  $D$ .

The latent variable  $D_i^*$  may be interpreted as the expected utility of receiving or not receiving the treatment: we assume  $\alpha_1 \neq 0$ , that is,

$$\text{cov}(D_i, Z_i) \neq 0$$

If  $Z_i$  satisfies the above assumption,  $Z$  is said to be an instrumental variable, in the sense that it determines the compliance status but it does not affect the outcome variable, given the observed treatment  $D_i$ . In the above settings, the IV estimator is the ratio of the sample covariances (Durbin, 1954); from the binary nature of  $D$  and  $Z$ , we have:

$$\hat{\beta}_1^{IV} = \frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \frac{\frac{\sum_{i=1}^N Y_i Z_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N Y_i (1 - Z_i)}{\sum_{i=1}^N (1 - Z_i)}}{\frac{\sum_{i=1}^N D_i Z_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N D_i (1 - Z_i)}{\sum_{i=1}^N (1 - Z_i)}}$$

#### *IV estimation in the Rubin Causal Model*

Let  $\mathbf{Z}$  be the  $N$ -dimensional vector of randomly chosen treatment assignments, with elements  $Z_i = 0$  if unit  $i$  is assigned to the control group and  $Z_i = 1$  if unit  $i$  is assigned to the treatment group. We denote with  $D_i(\mathbf{Z})$  the indicator for whether unit  $i$  would receive the treatment, given the vector  $\mathbf{Z}$ : clearly, if the compliance is perfect,  $D_i(\mathbf{Z}) = Z_i$  for each  $i$ . In a similar way, we define  $Y_i(\mathbf{Z}, \mathbf{D})$  to be the response of the unit  $i$ , given the assignment vector  $\mathbf{Z}$  and the  $N$ -dimensional vector  $\mathbf{D}$  with elements  $D_i(\mathbf{Z})$ . We refer to  $D_i(\mathbf{Z})$  and  $Y_i(\mathbf{Z}, \mathbf{D})$  as “potential outcomes”, because only one value (corresponding to the observed assignment for the unit  $i$ ) can be observed.

With these settings, we can write the SUTVA assumption as follows:

- *Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)*

- If  $Z_i = Z'_i$ , then  $D_i(\mathbf{Z}) = D_i(\mathbf{Z}')$
- If  $Z_i = Z'_i$  and  $D_i = D'_i$ , then  $Y_i(\mathbf{Z}, \mathbf{D}) = Y_i(\mathbf{Z}', \mathbf{D}')$

As before, we also assume randomization of the treatment assignment (Assumption 2). Since the SUTVA assumption implies that the treatment status of each unit does not affect the potential outcomes of the others units, we can write  $Y_i(\mathbf{Z}, \mathbf{D})$  and  $D_i(\mathbf{Z})$  as  $Y_i(Z_i, D_i)$  and  $D_i(Z_i)$ , respectively. Using the same notation of Section 5.1, we define the following causal effects at the individual level:

- Causal effect for unit  $i$  of  $Z$  on  $D$ :  $\delta_i^{(ZD)} = D_i(1) - D_i(0)$
- Causal effect for unit  $i$  of  $Z$  on  $Y$ :  $\delta_i^{(ZY)} = Y_i(1, D_i(1)) - Y_i(0, D_i(0))$

We refer to these causal effects as the individual intention-to-treat (ITT) effects. Clearly, the above quantities are unknown, because only the outcomes corresponding to the actual  $Z_i$  are observed; given the assumption of random assignment, we can obtain an unbiased estimator of the *average* ITT effects. The average causal effects of  $Z$  on  $Y$  and  $D$  – that is,

the expected values of the above  $\delta_i^{(ZY)}$  and  $\delta_i^{(ZD)}$  – may be unbiasedly estimated as

$$\hat{\Delta}^{(ZY)} = \frac{\sum_{i=1}^N Y_i Z_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N Y_i (1 - Z_i)}{\sum_{i=1}^N (1 - Z_i)}$$

and

$$\hat{\Delta}^{(ZD)} = \frac{\sum_{i=1}^N D_i Z_i}{\sum_{i=1}^N Z_i} - \frac{\sum_{i=1}^N D_i (1 - Z_i)}{\sum_{i=1}^N (1 - Z_i)}$$

respectively. We can see that

$$\hat{\beta}_1^{IV} = \frac{\hat{\Delta}^{(ZY)}}{\hat{\Delta}^{(ZD)}}$$

that is, the limit of the IV estimator (the IV estimand) equals the ratio of the average ITT effects.

In order to provide an unbiased estimator of  $\Delta^{(DY)}$  – the average causal effect of D on Y – additional assumptions are required. Typically, the actual receipt of the treatment  $D_i$  is nonignorable: that is,  $\text{cov}(\varepsilon, \nu) \neq 0$  and  $D_i$  (the endogenous regressor in the structural equation model) is correlated with  $\varepsilon_i$ . This implies that the data-generating process of Y, given D, and of D, given Z, are not independent: for this reason, the difference of outcome averages by treatment *received* does not represent an unbiased – or even consistent – estimator of the causal effect of interest – the average effect of D on Y. As showed in Angrist et al. (1996), the key assumption requires that any effect of Z on Y is through an effect of Z on D:

- *Assumption 3: Exclusion Restriction*

$$\mathbf{Y}(\mathbf{Z}, \mathbf{D}) = \mathbf{Y}(\mathbf{Z}', \mathbf{D}) \text{ for all } \mathbf{Z}, \mathbf{Z}' \text{ and for all } \mathbf{D}.$$

The exclusion restriction implies that the potential outcome does not depend on the treatment assignment, given the observed treatment receipt: this allows us to denote the potential outcome as  $Y_i(D_i)$ , being  $\mathbf{Y}(\mathbf{Z}, \mathbf{D}) = \mathbf{Y}(\mathbf{Z}', \mathbf{D}) = \mathbf{Y}(\mathbf{D})$  for all  $\mathbf{Z}, \mathbf{Z}'$  and for all  $\mathbf{D}$ . The causal effect of D on Y for unit  $i$  is  $\delta_i^{(DY)} = Y_i(1) - Y_i(0)$ : this individual effect is only defined for units with  $D_i(1) \neq D_i(0)$ ; for those units, only one term of the above difference (corresponding to the assigned treatment  $Z_i$ ) is observed.

In order to illustrate the relationship between the IV estimator and the causal effect of D on Y, two more assumptions are needed:

- *Assumption 4: Nonzero Average Causal Effect of Z on D*

$$\Delta^{(ZD)} = E[D_i(1) - D_i(0)] \neq 0$$

that is, the treatment assignment has an effect on the average probability of treatment.

- *Assumption 5: Monotonicity of Compliance (Imbens and Angrist, 1994)*

$$D_i(1) \geq D_i(0) \text{ for all } i$$

that is, there is no units that do the opposite of what they are assigned to do. If assumptions 1-5 hold, we say that  $Z$  is an Instrumental Variable for the causal effect of  $D$  on  $Y$ .

Exploiting SUTVA and the exclusion restriction, we can obtain the following relationship between the intention-to-treat effects of  $Z$  on  $Y$  and  $D$  and the causal effect of  $D$  on  $Y$ :

$$\begin{aligned}\delta_i^{(ZY)} &= Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \\ &= Y_i(D_i(1)) - Y_i(D_i(0)) \\ &= [Y_i(1) \cdot D_i(1) + Y_i(0) \cdot (1 - D_i(1))] - [Y_i(1) \cdot D_i(0) + Y_i(0) \cdot (1 - D_i(0))] \\ &= (Y_i(1) - Y_i(0)) \cdot (D_i(1) - D_i(0)) \\ &= \delta_i^{(DY)} \delta_i^{(ZD)}\end{aligned}$$

that is, for unit  $i$  we can express the causal effect of  $Z$  on  $Y$  as the product of the causal effect of  $D$  on  $Y$  and the causal effect of  $Z$  on  $D$ . Taking the expectation, we obtain:

$$\begin{aligned}\Delta^{(ZY)} &= E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] \\ &= E[(Y_i(1) - Y_i(0)) \cdot (D_i(1) - D_i(0))] \\ &= E[Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1] P[D_i(1) - D_i(0) = 1] \\ &\quad - E[Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = -1] P[D_i(1) - D_i(0) = -1]\end{aligned}$$

The monotonicity of compliance rules out the units for which  $D_i(1) - D_i(0) = -1$ , leading to the following simplification:

$$\begin{aligned}\Delta^{(ZY)} &= E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] \\ &= E[Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1] P[D_i(1) - D_i(0) = 1] \\ &= \Delta_*^{(DY)} \Delta^{(ZD)}\end{aligned}$$

where  $\Delta_*^{(DY)}$  is the average treatment effect of  $D$  on  $Y$  for the subpopulation for which  $D_i(1) - D_i(0) = 1$  (that is,  $D_i(1) = 1$  and  $D_i(0) = 0$ );  $E[D_i(1) - D_i(0)] = P[D_i(1) - D_i(0) = 1]$  by virtue of the monotonicity assumption. This ultimately results in the result:

$$\begin{aligned}\hat{\beta}_1^{IV} &= \frac{\Delta^{(ZY)}}{\Delta^{(ZD)}} \\ &= \frac{E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]}{E[D_i(1) - D_i(0)]} \\ &= E[Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1] \\ &= \Delta_*^{(DY)}\end{aligned}$$

The indicators  $D_i(z)$  ( $z = 0, 1$ ) describe the compliance behavior and define four subpopulations: compliers (c), for whom  $D_i(z) = z$ ; never-takers (n), for whom  $D_i(z) = 0$ ; always-takers (a), for whom  $D_i(z) = 1$ ; defiers (d), for whom  $D_i(z) = 1 - z$ . Without additional assumptions, the compliance status of unit  $i$  is never observed; by virtue of randomization, however, the four types have the same distribution in both treatment groups. Table 5.1 contains the causal effect of  $Z$  on  $Y$  for the four subpopulations classified by  $D_i(0)$  and  $D_i(1)$ .

The intention-to-treat analysis compares outcomes according to the assignment  $Z_i$ , ignoring the compliance behavior: this implies that the ITT effects estimate the effect of *assignment*, whereas the parameter of interest is – in most cases – the causal effect of the treatment receipt. The global ITT may be written as the weighted average of the ITT effects across the four subpopulations:

$$ITT = \pi_c ITT_c + \pi_n ITT_n + \pi_a ITT_a + \pi_d ITT_d$$

where  $ITT_j$  is the effect of the treatment assignment on units of type  $j$  and  $\pi_j$  is the proportion of units of type  $j$  ( $j = c, n, a, d$ ).

The exclusion restriction states that  $ITT_n = ITT_a = 0$ : because for never-takers and always-takers the assignment does not affect the receipt of the treatment, it is sometimes reasonable to assume a null effect of  $Z$  on the outcome variable; the monotonicity of compliance rules out the existence of defiers,  $\pi_d = 0$ . These two assumptions allow the identification of the ITT effect for compliers,  $ITT_c = ITT/\pi_c$ . The global ITT may be viewed as a conservative estimate of the treatment effect: with the implicit assumptions that  $\pi_d = 0$  and that both  $ITT_n$  and  $ITT_a$  are strictly less than  $ITT_c$ , it should be expected that  $ITT < ITT_c$ . The proportion of compliers equals the average causal effect of  $Z$  on  $D$ , and the average causal effect of  $Z$  on  $Y$  is proportional to the average causal effect of  $D$  on  $Y$  for the compliers – which is the parameter of interest and corresponds to the IV estimand.

Usually, the no-defiers assumption holds by design, i.e., the access to treatment is denied to those assigned to the control group. The exclusion restriction plays a critical role in separating the distributions of compliers and non-compliers and could be more or less plausible depending on the context; for example, in clinical trials, blinding, double blinding and using placebos justify this critical assumption.

Some testable constraints are implied by the exclusion restriction (Balke and Pearl, 1997; Imbens and Rubin, 1997b), but in order to relax it, it is useful to make additional assumptions. Various strategies have been proposed in the literature to achieve identification in the absence of exclusion restrictions. Little and Yau (1997) and Hirano et al. (2000) extend the analysis of Imbens and Rubin (1997b) to allow for the presence of pre-treatment covariates; if they are available, more modelling options other than strictly forcing the exclusion restriction can be considered to achieve identifiability. In Hirano et al. (2000) relaxing exclusion restrictions (but maintaining monotonicity) within a full Bayesian analysis allows the estimation of the effect of assignment for various subpopulations defined by compliance status. Covariates can also be exploited to achieve identification and improve efficiency. In the framework of principal stratification, plausible behavioral hypotheses within or among groups defined by the values of the covariates can be translated into restrictions on coefficients within or among strata. For some covariates, the same coefficient across strata can



be imposed (Frangakis, 2006), or some interaction terms can be excluded (Jo, 2002).

In the present work we will maintain both monotonicity of compliance and the exclusion restriction on noncompliers, even if the plausibility of the latter can be questioned. Some remarks will be addressed later, giving suggestions for future developments.

		$D_i(0)$	
		0	1
$D_i(1)$	0	$Y_i(1, 0) - Y_i(0, 0) = 0$ Never-takers	$Y_i(1, 0) - Y_i(0, 1)$ $= Y_i(0) - Y_i(1)$ Defiers
	1	$Y_i(1, 1) - Y_i(0, 0)$ $= Y_i(1) - Y_i(0)$ Compliers	$Y_i(1, 1) - Y_i(0, 1) = 0$ Always-takers

Table 4.1 *Two-way classification of the population, according to the compliance behavior; in the cells, the value of the causal effect of Z on Y under the exclusion restriction.*

### 5.3 Missing outcomes

A common complication in observational data is the presence of missing outcomes. As well as noncompliance, missing outcomes are a post-treatment variable: in order to adjust for this complication, a critical role is played by the assumptions on the missing data mechanism; proceeding in inference requires some form of imputation of the missing data, either implicit or explicit. The appropriate set of assumptions, however, depends on the scientific settings.

Randomized experiments often suffer from both complications (noncompliance with the assigned treatment and missing outcomes); a great care is needed in analyzing the causal effects of a treatment in presence of any of these complications. An account of the different approaches proposed in the literature is in Mealli and Rubin (2002).

The standard ITT analysis based on the complete case leads to an unbiased estimate of the treatment effect only under the very restrictive assumption that the data are Missing Completely at Random (MCAR; Rubin, 1976; Little and Rubin, 1987). This assumption has testable implication and is often rejected by the data. A more convenient assumption is that the outcomes are Missing and Random (MAR; Rubin, 1976); this assumption allows the probability of response to depend on observed but not on unobserved quantities. The MCAR is a special case of the MAR model, arising when the response indicator is unrelated with both observed and unobserved variables. If MAR holds and the parameters of the missing data process are distinct from those of the outcome distribution, the missing mechanism is said to be ignorable, meaning that the missing data values are not informative



about the probability of response, given the observed quantities; unfortunately, this very attractive assumption is not testable without auxiliary information, because the data cannot provide any evidence against MAR.

An assumption that links noncompliance with nonresponse is the Latent Ignorability (LI; Frangakis and Rubin, 1999). Under LI, the missing data process would be ignorable if the compliance behavior were known for all units; since the true compliance type is unobserved for those who are assigned to the control group, the missing mechanism is in fact nonignorable. To achieve full identification of the ITT effect for compliers under LI, additional assumptions are required; different forms of response exclusion restriction have been discussed in Frangakis and Rubin (1999) and Mealli and Rubin (2002).

In the next section, we briefly present the most frequent assumptions on the missing data mechanism in presence of noncompliance with the treatment assignment.

### 5.3.1 Analysis of randomized experiments with noncompliance and missing outcomes

As usual,  $Z_i$  and  $D_i$  represent the indicators of the treatment assignment and treatment receipt for unit  $i$ , respectively ( $i = 1, \dots, N$ ); as before, we assume that the monotonicity of compliance holds, such that the population is only composed of compliers and never-takers. For unit  $i$ , we denote as  $R_i(z)$  the potential response indicator, assumed to be dichotomous (1 = respondent, 0 = nonrespondent);  $R_i = R_i(Z_i)$  is the observed response indicator. We denote with  $Y_i$  the (multivariate) outcome for unit  $i$ , which is only observed if  $R_i = 1$ , and with  $X_i$  an optional vector of observed pre-treatment covariates.

The standard model for missing data makes use of the Missing and Random assumption (MAR; Rubin, 1976):

$$R_i \perp Y_i \mid Z_i, D_i, X_i$$

that is,  $P(R_i = 1 \mid Y_i, Z_i, D_i, X_i) = P(R_i = 1 \mid Z_i, D_i, X_i)$ ; under this assumption, which is often relatively plausible, the compliers and the never-takers are allowed to have a different response behavior in the treatment group – since their  $D_i$  would differ – but not in the control group, where the observed  $D_i$  would be the same. A special case of the MAR model is when

$$R_i \perp Z_i, D_i, X_i$$

In this case, the outcome is said to be Missing Completely at Random (MCAR; Rubin, 1976; Little and Rubin, 1987).

Under MCAR, the inference can be performed without conditioning to the observed covariates; the usual IV estimator (computed on the units with observed outcome) provides an unbiased estimate of the treatment effect. However, the MCAR can be viewed as a very restrictive model: for each unit, the response indicator is assumed to be a bernoulli trial, whose parameter is independent of the pre-treatment covariates, the treatment assignment, the treatment receipt and the realized outcome; in other words, the respondent are a random sample of the  $N$  units. The MCAR assumption has testable implication and is often rejected by the data; the MAR assumption is generally more plausible, but is not testable – since

the data cannot provide any evidence against MAR. If the parameters of the MAR model are distinct from those of the outcome distribution, the missing data process is said to be ignorable; in this case, the likelihood function factorizes and the two sets of parameters can be estimated independently. In most cases, the missing data process is not of interest and the probabilities of response are nuisance parameters: an appealing feature of this framework is that it avoids to estimate – and even formulate – a model for the missing data mechanism.

An alternative assumption is Latent Ignorability (LI; Frangakis and Rubin, 1999):

$$R_i \perp Y_i \mid Z_i, U_i, X_i$$

In the above statement,  $U_i = D_i(1)$  is the true compliance type of unit  $i$ : under LI, the missing data process is in fact nonignorable, because the compliance type is unobserved for those who are assigned to the control group; as a consequence, this assumption alone does not lead to full identification of the treatment effect for compliers. Frangakis and Rubin (1999) achieve the full identification by exploiting response exclusion restriction for never-takers and always-takers; using a different rationale, Mealli and Rubin (2002) propose the response exclusion restriction for always-takers and compliers. We will not use such assumptions here; however, the response exclusion restriction for never-takers is often questionable: among never-takers, those who refuse the participation to the program – that is, those who are assigned to the treatment group – may be less willing to reveal their outcomes. On the other hand, since the compliers are willing to follow the protocol they are assigned to, it is reasonable that their response behavior is unaffected by the treatment assignment: this would justify the response exclusion restriction for compliers.

In this work, we assume that the MAR assumption holds; we will provide further details in Section 5.5.

## 5.4 Outcomes truncated by death

The advantages of random assignment may be lost if the outcome is not defined for all sample members. This problem has been dubbed truncation by death (Zhang and Rubin, 2003; Rubin, 2006), borrowing the term from medical clinical trials where the outcome – for example, quality of life – is undefined for those patients who die. This problem also frequently arises in the evaluation of social policy interventions, such as school dropout prevention programs (the students' test scores are only defined if they stay in school), interventions to improve teacher quality (teacher quality is only defined for those who continue teaching) and employment and training programs, designed to affect both the probability of employment and the quality of the employment obtained, such as the wage (wages are observed, and well defined, only for employed individuals).

As pointed out in Rosenbaum (1984), a misleading inference could be obtained by simply comparing employed treated participants and employed controls; this is because the employment status is a post-treatment variable: even under randomization of the treatment assignment, the characteristics of the employed units in the two groups are expected to differ and, in this case, a biased estimate of the causal effect is obtained.

As the intervention had two effects, i.e., it affected the survivor status and it affected the outcome of those who survived, the treatment-control difference-in-means is not informative on the treatment effect. Rubin (2000) and Zhang and Rubin (2003) argue that, because the outcome is undefined for non survivors, the only question that does make sense is what is the impact on the individuals who would survive irrespective of whether they receive the intervention. Rubin (2000) first used the term survivor average causal effect (SACE) for the impact on the sample members whose outcomes are observed whether they are assigned to the treatment or control group. Unfortunately, it is not straightforward to estimate the SACE because we only observe what each couple does either with or without the intervention, but not both. Zhang and Rubin (2003) derived bounds on the possible range of values within which the SACE must lie, which are similar in spirit but narrower than bounds presented in Horowitz and Manski (2000). However, without any assumptions, the Zhang and Rubin bounds on the SACE can still be quite wide. These bounds can be narrowed, however, by specifying additional assumptions; Zhang and Rubin (2003) discuss two particular assumptions that can help narrow the bounds on the SACE, namely monotonicity and stochastic dominance on the outcome distribution. Those bounds have been exploited by Lee (2005) in the analysis of the Job Corps data. In the original paper, finding the bounds involves numerical optimization; a closed form is provided in the recent work of Imai (2007a).

In a parametric setting some authors address this issue using sample selection models (Heckman, 1979); this approach, however, presents the difficulty of finding a set of covariates that are “instruments” in the sense that they determine the survivor status but do not directly affect the outcomes of those who are survived (Heckman and Vytlačil, 1999).

The principal stratification approach allows to explicitly model the truncation process and the outcome (Rubin 2006; Zhang, Rubin, and Mealli, 2008a, 2008b). In this work, we follow this approach, focussing on identification and estimation of intervention’s effect on the subpopulation of the always survivors.

In the evaluation of the effectiveness of government-sponsored training programs, we are interested on the treatment effect on wages: an increase in the expected wage for the treated participants would reflect the raise in the human capital due to the training. The main difficulty in estimating the effect on wages is that wages are observed (and well defined) only for employed individuals.<sup>(10)</sup> Our aim is to evaluate the effect on wages for the subpopulation of always employed, those who would be employed regardless the treatment assignment.

Following the recent work by Zhang, Rubin and Mealli (2008b), we now illustrate how the Rubin Causal Model works in this case. In this section, we assume perfect compliance with the treatment assignment and we only focus on the estimation of the causal effect on wages, assuming that there are not missing outcomes; in Section 5.5 we provide a unified framework to simultaneously address all these issues.

---

<sup>(10)</sup> For those who are unemployed, it could be argued that the wages are also defined but lower than their “reservation wage”: we will not discuss this issue here; under a statistical point of view, the relevant question is that the wages are only observed for employed people.

### 5.4.1 Estimating the causal effect of job-training programs on wages

For unit  $i$ , we denote with  $S_i$  the indicator for the employment status (0 = unemployed, 1 = employed) and with  $W_i$  the observed hourly wage; using the extended space  $\{\mathfrak{R}^+, *\}$ , we define the wages for unemployed people to be  $W_i = *$ . As before, we denote with  $Z_i$  the binary treatment assignment of unit  $i$  and with  $\mathbf{Z}$  the  $N$ -dimensional vector of assignments. We denote as  $W_i(\mathbf{Z})$  and  $S_i(\mathbf{Z})$  the potential outcomes for unit  $i$ .

As usual, before proceeding in the inference, we make the SUTVA and randomization assumptions: in this case, the SUTVA may be written as follows:

- *Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)*

$$\text{If } Z_i = Z_i', \text{ then } S_i(\mathbf{Z}) = S_i(\mathbf{Z}') \text{ and } W_i(\mathbf{Z}) = W_i(\mathbf{Z}')$$

With the SUTVA assumption, we can write the potential outcomes as  $W_i(Z_i)$  and  $S_i(Z_i)$ , respectively; by virtue of the randomization process, we can estimate an average causal effect taking the expected value of the individual effect.

According to the RCM – and using the same notation as in Section 5.1 – we define the causal effects of the treatment for the unit  $i$  as the differences in the potential outcomes:

$$\begin{aligned} \delta_i^{(ZW)} &= W_i(1) - W_i(0) \\ \delta_i^{(ZS)} &= S_i(1) - S_i(0) \end{aligned}$$

We can see that only for units with  $W_i(1) \in \mathfrak{R}^+$  and  $W_i(0) \in \mathfrak{R}^+$  (that is,  $S_i(1) = S_i(0) = 1$ ) we can define in  $\mathfrak{R}^+$  the individual effect on wages  $\delta_i^{(ZW)}$ ; using the same notation as in Zhang et al., we classify the units in the following 4 strata,  $G = \{EE, EN, NE, NN\}$ , according to the values of  $S_i(1)$  and  $S_i(0)$ :

- $EE = \{i: S_i(1) = S_i(0) = 1\}$ , those who would be employed regardless of the treatment assignment; for this stratum,  $W_i(1)$  and  $W_i(0)$  are defined in  $\mathfrak{R}^+$ ;
- $EN = \{i: S_i(1) = 1 \text{ and } S_i(0) = 0\}$ , those who would be employed only under treatment; for this stratum,  $W_i(1) \in \mathfrak{R}^+$  and  $W_i(0) = *$ ;
- $NE = \{i: S_i(1) = 0 \text{ and } S_i(0) = 1\}$ , those who would be employed only if assigned to the control group; for this stratum,  $W_i(1) = *$  and  $W_i(0) \in \mathfrak{R}^+$ ;
- $NN = \{i: S_i(1) = S_i(0) = 0\}$ , those who would be unemployed regardless of the treatment assignment; for this stratum,  $W_i(1) = W_i(0) = *$ .

Only for the EE group we can define the causal effect on wages in a meaningful way; for this reason, the parameter of interest is the average treatment effect (ATE) on wages in the EE group:

$$\Delta_*^{(ZW)} = E[W_i(1) | G_i = EE] - E[W_i(0) | G_i = EE]$$

where the notation  $\Delta_*^{(ZW)}$  instead of  $\Delta^{(ZW)}$  means that the expected value is only taken on the EE group. The average treatment effect on employment is defined on the whole population:

$$\Delta^{(ZS)} = E[S_i(1)] - E[S_i(0)] = P[G_i = EN] - P[G_i = NE]$$

Clearly, we cannot observe the principal stratum  $G_i$  for individual  $i$ : for those assigned to the treatment group, we only observe  $S_i(1)$ ; for those assigned to the control group, we only observe  $S_i(0)$ . This configures our estimation issue as a missing data problem. What we can directly observe are the following groups:

- $O(1,1) = \{i: Z_i = 1 \text{ and } S_i = 1\}$ , those who are assigned to the treatment group and employed; they are a mixture of the principal strata EE and EN;
- $O(1,0) = \{i: Z_i = 1 \text{ and } S_i = 0\}$ , those who are assigned to the treatment group and unemployed; they are a mixture of the principal strata NN and NE;
- $O(0,1) = \{i: Z_i = 0 \text{ and } S_i = 1\}$ , those who are assigned to the control group and employed; they are a mixture of the principal strata EE and NE;
- $O(0,0) = \{i: Z_i = 0 \text{ and } S_i = 0\}$ , those who are assigned to the control group and unemployed; they are a mixture of the principal strata NN and EN.

A common assumption is the following:

- *Assumption 6 (Monotonicity of Truncation):*  $P[G_i = \text{NE}] = 0$

that is, there is no NE group, meaning that the treatment is not harming anyone. This assumption is in general less plausible than the analogous monotonicity of compliance assumption made in Section 5.2 (which rules out the defiers); in this case, such assumption rules out, a priori, a negative treatment effect on the employment: in a short run, it seems plausible that treated individuals choose to be unemployed and wait for a “good” work; as a consequence, in some setting the monotonicity of truncation may have little justification.

Another assumption, considered in Zhang and Rubin (2003) is *stochastic dominance*: the wage distribution for the EE group is assumed to be stochastically larger than the wage distribution for the EN group when trained and the NE group when not trained.

Usually, additional hypotheses are done: a common choice is to specify a parametric model and use the standard mixture analysis in the estimation. In Section 5.5, we propose a complex framework to estimate the effect on wages in presence of noncompliance and missing outcomes; in Section 5.6, we illustrate how to obtain the parameters estimates using the EM algorithm under a likelihood approach.

## 5.5 Estimating the effect on wages with noncompliance and missing outcomes under the MAR assumption

As before, we denote with  $Z_i$  the treatment assignment and with  $D_i$  the treatment receipt for unit  $i$  ( $1 = \text{treatment}$ ,  $0 = \text{control}$ ) whereas  $R_i$  denotes the response indicator ( $1 = \text{respondent}$ ,  $0 = \text{nonrespondent}$ ).<sup>(11)</sup> Among respondent units,  $S_i$  and  $W_i$  represent, respectively, the observed employment status ( $1 = \text{employed}$ ,  $0 = \text{unemployed}$ ) and the wage for individual  $i$  – which is only observed if  $S_i = 1$ ; as usual, we define the wages for unem-

<sup>(11)</sup> For ease of presentation, in what follows we denote as “nonrespondent” each unit whose outcome is missing due to both nonresponse or attrition.

ployed people to be  $W_i = *$ . Nonrespondent units have an unknown value for both  $S_i$  and  $W_i$ . We denote as  $\mathbf{Z}$ ,  $\mathbf{D}$ ,  $\mathbf{R}$ ,  $\mathbf{S}$  and  $\mathbf{W}$  the  $N$ -dimensional vectors with elements  $Z_i$ ,  $D_i$ ,  $R_i$ ,  $S_i$  and  $W_i$ , respectively. The potential outcomes for unit  $i$  are  $D_i(\mathbf{Z})$ ,  $R_i(\mathbf{Z}, \mathbf{D})$ ,  $S_i(\mathbf{Z}, \mathbf{D})$  and  $W_i(\mathbf{Z}, \mathbf{D})$ : we emphasize the fact that the response indicator is a post-treatment measurement, as well as the outcome variables.

In what follows, we will consider that assumptions 1-5 (SUTVA, randomization, exclusion restriction, nonzero average effect of  $Z$  on  $D$ , monotonicity of compliance) hold; for this setting, we can rewrite the SUTVA as follows:

- *Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)*

- i) If  $Z_i = Z'_i$ , then  $D_i(\mathbf{Z}) = D_i(\mathbf{Z}')$
- ii) If  $Z_i = Z'_i$  and  $D_i = D'_i$ , then  $R_i(\mathbf{Z}, \mathbf{D}) = R_i(\mathbf{Z}', \mathbf{D}')$
- iii) If  $Z_i = Z'_i$  and  $D_i = D'_i$ , then  $S_i(\mathbf{Z}, \mathbf{D}) = S_i(\mathbf{Z}', \mathbf{D}')$
- iv) If  $Z_i = Z'_i$  and  $D_i = D'_i$ , then  $W_i(\mathbf{Z}, \mathbf{D}) = W_i(\mathbf{Z}', \mathbf{D}')$

We do not repeat here the assumptions 2-5, which remain unchanged; the exclusion restriction for never-takers is assumed to hold for both  $W$  and  $S$ . However, we do not impose any exclusion restriction on the potential response indicator: as a consequence, units with the same treatment receipt may have a different missing data mechanism, according to their treatment assignment. In virtue of this set of assumptions, we can unambiguously write the potential outcomes as  $D_i(Z_i)$ ,  $R_i(Z_i, D_i)$ ,  $S_i(D_i)$  and  $W_i(D_i)$ .

With respect to the compliance behavior, we assume that the population is only composed of compliers (C) and never-takers (N): assumption 5 (monotonicity of compliance) rules out the defiers; the always-takers are also excluded in this case, because the units assigned to the control group ( $Z_i = 0$ ) are not allowed to enroll in Job Corps. If we ignore the response behavior, units can be cross-classified [ $\{C, N\} \times \{EE, EN, NE, NN\}$ ] in 8 groups:

$$\{C.EE, C.EN, C.NE, C.NN, N.EE, N.EN, N.NE, N.NN\}.$$

By virtue of the exclusion restriction on the employment status, we can cross out the N.EN group and the N.NE group, which would entail a direct effect of  $Z$  on  $S$  for never-takers: the groups reduce to

$$G = \{C.EE, C.EN, C.NE, C.NN, N.EE, N.NN\}.$$

Optionally, if also monotonicity of truncation holds, there is no C.NE group and we can write

$$G = \{C.EE, C.EN, C.NN, N.EE, N.NN\}.$$

As shown in Section 5.2, the causal effect of interest is the differences in the potential outcomes for the subpopulation of compliers; in Section 5.4, we argued that the causal effect on wages is only defined for the EE group. In this more complex framework, a critical role is played by the C.EE group: we will now proceed in defining the causal effects of interest within the common framework of the Rubin Causal Model. The causal effects at the



individual level on D, S and W are defined as follows:

- Causal effect of Z on D:  $\delta_i^{(ZD)} = D_i(1) - D_i(0)$
- Causal effect of D on S:  $\delta_i^{(DS)} = S_i(1) - S_i(0)$
- Causal effect of D on W:  $\delta_i^{(DW)} = W_i(1) - W_i(0)$

Taking back the results of Sections 5.2-5.4, we obtain the following statements:

- the average treatment effect of Z on D equals the proportion of compliers in the population (compare Section 5.2):

$$\Delta^{(ZD)} = E[D_i(1)] - E[D_i(0)] = P[D_i(1) - D_i(0) = 1]$$

- the average treatment effects of D on S and W are, respectively:

$$\begin{aligned} \Delta_*^{(DS)} &= E[S_i(1) \mid D_i(1) - D_i(0) = 1] - E[S_i(0) \mid D_i(1) - D_i(0) = 1] \\ &= P[G_i = C.EN] - P[G_i = C.NE] \end{aligned}$$

$$\Delta_*^{(DW)} = E[W_i(1) \mid G_i = C.EE] - E[W_i(0) \mid G_i = C.EE]$$

where in the above formula we wrote  $\Delta_*^{(DS)}$  and  $\Delta_*^{(DW)}$  instead of  $\Delta^{(DS)}$  and  $\Delta^{(DW)}$ , respectively, to emphasize that the expected values are only taken on a subset of the whole population (the compliers for  $\Delta_*^{(DS)}$  and the C.EE group for  $\Delta_*^{(DW)}$ ).

Without further assumptions on the response behavior (such as monotonicity of response and exclusion restrictions for some subgroup of units), each of the above principal strata is composed of 4 subgroups, according to the couple  $R_i(1, D_i(1)), R_i(0, D_i(0))$  of potential response indicators; the causal effect of Z on R is known to be zero for the always-respondent (RR) and the never-respondent (rr): within each stratum in G, the average treatment effect is defined as the difference between the proportion of the Rr group (units who would respond only under treatment) and the proportion of the rR group (units who would respond only under control). With these settings, 24 latent strata are supposed to exist (20 if the monotonicity of truncation is assumed for the potential employment status); in order to simplify this very general framework, we assume Latent Ignorability (Frangakis and Rubin, 1999):

$$W_i \perp R_i \mid Z_i, U_i, S_i(1), S_i(0), \mathbf{X}_i$$

where  $U_i = D_i(1)$  is the true compliance behavior (1 = complier, 0 = never-taker) and  $\mathbf{X}_i$  is an optional vector of pre-treatment covariates. Under this assumption, given the covariates and the treatment assignment, units with the same compliance behavior and potential employment status have the same expected wage, regardless of the response behavior. The Latent Ignorability implies that  $P(R_i = 1 \mid W_i, G_i, \mathbf{X}_i) = P(R_i = 1 \mid G_i, \mathbf{X}_i)$ ,  $G_i \in G$ : if we knew the group membership ( $G_i$ ) of each unit, the missing data mechanism would be ignorable. This allows us to define the following probabilities:

$$\rho_{i:g,z} = P(R_i = 1 \mid Z_i = z, G_i = g, \mathbf{X}_i)$$

that is,  $\rho_{i:g,z}$  is the probability of observing the outcomes of the unit  $i$ , given that this unit belongs to the  $g^{\text{th}}$  stratum and is assigned to the treatment  $z$  ( $g \in G, z = \{0,1\}, i = 1, \dots, N$ ).

Since the true compliance behavior and the potential employment status are partially unobserved, the missing data process is in fact nonignorable. An alternative missing data model is obtained exploiting the Missing at Random assumption (MAR; Rubin, 1976):

$$S_i, W_i \perp R_i \mid Z_i, D_i, \mathbf{X}_i$$

The MAR requires that  $P(R_i = 1 \mid S_i, W_i, Z_i, D_i, \mathbf{X}_i) = P(R_i = 1 \mid Z_i, D_i, \mathbf{X}_i)$ : the probability of observing the outcomes ( $S$  and  $W$ ) is the same for all units with the same treatment assignment, treatment receipt and pre-treatment covariates; in other words, the missing mechanism is unaffected by the outcomes  $S_i$  and  $W_i$ . If we assume LI and we impose the following restrictions on response probabilities, we can prove that the missing mechanism is ignorable and the MAR assumption holds:

$$\rho_{i:C.EE,1} = \rho_{i:C.EN,1} = \rho_{i:C.NE,1} = \rho_{i:C.NN,1} \quad [5.1]$$

$$\rho_{i:N.EE,1} = \rho_{i:N.NN,1} \quad [5.2]$$

$$\rho_{i:C.EE,0} = \rho_{i:C.EN,0} = \rho_{i:C.NE,0} = \rho_{i:C.NN,0} = \rho_{i:N.EE,0} = \rho_{i:N.NN,0} \quad [5.3]$$

Compliers and never-takers are allowed to have a different response behavior under treatment – since their  $D_i$  would differ – but not under the control condition, where the two groups have the same value of  $D_i$ . The couple of potential outcomes  $S_i(1)$  and  $S_i(0)$  does not affect the response behavior. We will illustrate in the next section how the Latent Ignorability and the MAR assumptions are used in writing the observed likelihood function.

The estimation issue is a missing data problem, because we cannot observe which stratum each unit comes from; among the respondent units, what we can directly observe are the following groups, defined according to different combinations of  $Z, D$  and  $S$ :

- $O(1,1,1) = \{i: Z_i = 1, D_i = 1 \text{ and } S_i = 1\}$ , those who are assigned to the treatment group, compliers with the assignment and employed; they are a mixture of the two principal strata C.EE and C.EN;
- $O(1,1,0) = \{i: Z_i = 1, D_i = 1 \text{ and } S_i = 0\}$ , those who are assigned to the treatment group, compliers and unemployed; they are a mixture of the two principal strata C.NN and C.NE;
- $O(1,0,1) = \{i: Z_i = 1, D_i = 0 \text{ and } S_i = 1\}$ , those who are assigned to the treatment group, noncompliers and employed; they belong to the principal stratum N.EE;
- $O(1,0,0) = \{i: Z_i = 1, D_i = 0 \text{ and } S_i = 0\}$ , those who are assigned to the treatment group, noncompliers and unemployed; they belong to the principal stratum N.NN;
- $O(0,0,1) = \{i: Z_i = 0, D_i = 0 \text{ and } S_i = 1\}$ , those who are assigned to the control group and employed; they are a mixture of the three principal strata C.EE, C.NE, N.EE;
- $O(0,0,0) = \{i: Z_i = 0, D_i = 0 \text{ and } S_i = 0\}$ , those who are assigned to the control group and unemployed; they are a mixture of the three principal strata C.NN, C.EN, N.NN.



For the non respondent, the value of  $S$  is unobserved; according to the couple of indicators  $(Z, D)$ , we observe the following groups:

- $O'(1,1) = \{i: Z_i = 1 \text{ and } D_i = 1\}$ , those who are assigned to the treatment group and compliers; they are a mixture of the four principal strata C.EE, C.EN, C.NE, C.NN;
- $O'(1,0) = \{i: Z_i = 1 \text{ and } D_i = 0\}$ , those who are assigned to the treatment group and noncompliers; they are a mixture of the two principal strata N.EE, N.NN;
- $O'(0,0) = \{i: Z_i = 0 \text{ and } D_i = 0\}$ , those who are assigned to the control group; they are a mixture of all strata in  $G$ .

In the above notation, we denoted with  $O(\cdot)$  the respondent units and with  $O'(\cdot)$  the nonrespondent. For those who are assigned to the treatment group, the compliance behavior is known; the employment status brings information about the couple  $S_i(1), S_i(0)$  and narrows the admissible strata. Among the nonrespondent, the latter information is unavailable: as a consequence, within the treatment group, we can only classify the units as compliers (the  $O'(1,1)$  group) or as never-takers (the  $O'(1,0)$  group), whereas in the control arm (the  $O'(0,0)$  group), when also the compliance behavior is unobserved, it is completely unknown which stratum the units come from.

In the next section, we will illustrate how to obtain the parameters estimates under a likelihood approach, exploiting standard mixture modeling.

## 5.6 Likelihood approach

We now assign a parametric distribution to the potential outcomes; this enables us to consider the effect of pre-treatment covariates  $(\mathbf{X})$ , using a regression model to describe the expected outcomes and the unobserved group membership; a finite mixture model (see, e.g., McLachlan and Peel, 2000) can be fitted using the EM (Expectation-Maximization) algorithm (Dempster, Laird and Rubin, 1977).

We denote with  $G_i$  the unobserved component membership label for unit  $i$ ;  $\mathbf{G} = G_1, \dots, G_N$  is the  $N$ -dimensional vector of unknown group labels. In the present work, we will assume that the exclusion restriction holds: this implies that  $G_i$  is a random draw from the set  $G = \{C.EE, C.EN, C.NE, C.NN, N.EE, N.NN\}$ ; in this case, the number of components of the mixture is  $k = 6$ . If we assume monotonicity of truncation, we cross out the C.NE component and we have  $k = 5$ .

For ease of notation, we assume that  $\mathbf{X}$  includes the constant term – that is, a column containing the unit vector. We assume a multinomial logistic model for the  $k$ -dimensional vector of cluster membership indicators:

$$P(G_i = g) = \frac{\exp\{\mathbf{X}_i \boldsymbol{\alpha}_g\}}{\sum_{h=1}^k \exp\{\mathbf{X}_i \boldsymbol{\alpha}_h\}} = \pi_{i:g}$$

where  $g \in G$  and the  $k^{\text{th}}$  principal stratum (N.NN) is taken as baseline (that is,  $\boldsymbol{\alpha}_{\text{N.NN}} = \mathbf{0}$ ). We denote with  $\pi_{i:g}$  the probability of the stratum  $g$  for unit  $i$ , given the pre-treatment vector of covariates  $\mathbf{X}_i$ .

Assuming a Normal distribution for the log wages, the general model specification is as follows:

- if  $G_i = \text{C.EE}$ ,  $\log[W_i(1, D_i(1))] = \log[W_i(1)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{\text{C.EE},1}, \sigma_{\text{C.EE},1}^2)$   
 $\log[W_i(0, D_i(0))] = \log[W_i(0)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{\text{C.EE},0}, \sigma_{\text{C.EE},0}^2)$
- if  $G_i = \text{C.EN}$ ,  $\log[W_i(1, D_i(1))] = \log[W_i(1)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{\text{C.EN},1}, \sigma_{\text{C.EN},1}^2)$
- if  $G_i = \text{C.NE}$ ,  $\log[W_i(0, D_i(0))] = \log[W_i(0)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{\text{C.NE},0}, \sigma_{\text{C.NE},0}^2)$
- if  $G_i = \text{N.EE}$ ,  $\log[W_i(1, D_i(1))] = \log[W_i(0, D_i(0))] = \log[W_i] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{\text{N.EE}}, \sigma_{\text{N.EE}}^2)$

If unit  $i$  is a complier,  $Z_i = D_i$ ; for never-takers,  $D_i = 0$ , regardless the treatment assignment. For the C.EE group, the parameters of the wage distribution vary across the two treatment levels; in the C.EN group, the wages are only defined if  $Z_i = 1$ ; in the C.NE group, only if  $Z_i = 0$  (optionally, the monotonicity of truncation rules out this group). The exclusion restriction implies that in the N.EE group the parameters of the wage distribution are unaffected by the treatment assignment; at the same time, as showed before, the exclusion restriction rules out the N.EN and N.NE groups. Clearly, for the C.NN and N.NN groups, there are no associated wages.

We denote as  $\xi = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\theta}\}$  the parameters vector of this model, where

$$\begin{aligned}\boldsymbol{\alpha} &= \{\boldsymbol{\alpha}_{\text{C.EE}}, \boldsymbol{\alpha}_{\text{C.EN}}, \boldsymbol{\alpha}_{\text{C.NE}}, \boldsymbol{\alpha}_{\text{C.NN}}, \boldsymbol{\alpha}_{\text{N.EE}}\} \\ \boldsymbol{\beta} &= \{\boldsymbol{\beta}_{\text{C.EE},1}, \boldsymbol{\beta}_{\text{C.EE},0}, \boldsymbol{\beta}_{\text{C.EN},1}, \boldsymbol{\beta}_{\text{C.NE},0}, \boldsymbol{\beta}_{\text{N.EE}}\} \\ \boldsymbol{\sigma} &= \{\sigma_{\text{C.EE},1}, \sigma_{\text{C.EE},0}, \sigma_{\text{C.EN},1}, \sigma_{\text{C.NE},0}, \sigma_{\text{N.EE}}\}\end{aligned}$$

and  $\boldsymbol{\theta}$  is the parameters vector of the probabilistic model for the missing data mechanism. We denote as  $N(\mu, \sigma^2)$  the probability density function of a Normal distribution with mean  $\mu$  and variance  $\sigma^2$  evaluated at  $\log(W_i)$ . We will assume MAR: in order to show the likelihood function under missing at random we first assume Latent Ignorability. If LI holds, the likelihood can be written as:

$$\begin{aligned}L(\xi \mid \mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X}) &= \prod_{i \in O(1,1,1)} \left[ \rho_{i:\text{C.EE},1} \pi_{i:\text{C.EE}} N(\mathbf{X}_i \boldsymbol{\beta}_{\text{C.EE},1}, \sigma_{\text{C.EE},1}^2) + \rho_{i:\text{C.EN},1} \pi_{i:\text{C.EN}} N(\mathbf{X}_i \boldsymbol{\beta}_{\text{C.EN},1}, \sigma_{\text{C.EN},1}^2) \right]^{o_i} \\ &\times \prod_{i \in O(1,1,0)} \left[ \rho_{i:\text{C.NE},1} \pi_{i:\text{C.NE}} + \rho_{i:\text{C.NN},1} \pi_{i:\text{C.NN}} \right]^{o_i} \\ &\times \prod_{i \in O(1,0,1)} \left[ \rho_{i:\text{N.EE},1} \pi_{i:\text{N.EE}} N(\mathbf{X}_i \boldsymbol{\beta}_{\text{N.EE}}, \sigma_{\text{N.EE}}^2) \right]^{o_i} \\ &\times \prod_{i \in O(1,0,0)} \left[ \rho_{i:\text{N.NN},1} \pi_{i:\text{N.NN}} \right]^{o_i}\end{aligned}$$

$$\begin{aligned}
& \times \prod_{i \in O(0,0,1)} \left[ \rho_{i:C,EE,0} \pi_{i:C,EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C,EE,0}, \sigma_{C,EE,0}^2) \right. \\
& \quad + \rho_{i:C,NE,0} \pi_{i:C,NE} N(\mathbf{X}_i \boldsymbol{\beta}_{C,NE,0}, \sigma_{C,NE,0}^2) \\
& \quad \left. + \rho_{i:N,EE,0} \pi_{i:N,EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N,EE}, \sigma_{N,EE}^2) \right]^{\omega_i} \\
& \times \prod_{i \in O(0,0,0)} \left[ \rho_{i:C,EN,0} \pi_{i:C,EN} + \rho_{i:C,NN,0} \pi_{i:C,NN} + \rho_{i:N,NN,0} \pi_{i:N,NN} \right]^{\omega_i} \\
& \times \prod_{i \in O'(1,1)} \left[ (1 - \rho_{i:C,EE,1}) \pi_{i:C,EE} + (1 - \rho_{i:C,EN,1}) \pi_{i:C,EN} + (1 - \rho_{i:C,NE,1}) \pi_{i:C,NE} + (1 - \rho_{i:C,NN,1}) \pi_{i:C,NN} \right]^{\omega_i} \\
& \times \prod_{i \in O'(1,0)} \left[ (1 - \rho_{i:N,EE,1}) \pi_{i:N,EE} + (1 - \rho_{i:N,NN,1}) \pi_{i:N,NN} \right]^{\omega_i} \\
& \times \prod_{i \in O'(0,0)} \left[ \sum_{j \in G} (1 - \rho_{i:j,0}) \pi_{i:j} \right]^{\omega_i}
\end{aligned}$$

where  $\omega_i$  are optional sample weights. The assumptions needed to make the missing data mechanism ignorable (so that the likelihood factorizes and the probabilities of observing the outcomes can be estimated independently of all other parameters) require equalities [5.1]-[5.3] to hold. The likelihood function simplifies as follows under MAR:

$$\begin{aligned}
& L(\xi \mid \mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X}) \\
& \propto \prod_{i \in O(1,1,1)} \left[ \pi_{i:C,EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C,EE,1}, \sigma_{C,EE,1}^2) + \pi_{i:C,EN} N(\mathbf{X}_i \boldsymbol{\beta}_{C,EN,1}, \sigma_{C,EN,1}^2) \right]^{\omega_i} \\
& \times \prod_{i \in O(1,1,0)} \left[ \pi_{i:C,NE} + \pi_{i:C,NN} \right]^{\omega_i} \\
& \times \prod_{i \in O(1,0,1)} \left[ \pi_{i:N,EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N,EE}, \sigma_{N,EE}^2) \right]^{\omega_i} \\
& \times \prod_{i \in O(1,0,0)} \left[ \pi_{i:N,NN} \right]^{\omega_i} \\
& \times \prod_{i \in O(0,0,1)} \left[ \pi_{i:C,EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C,EE,0}, \sigma_{C,EE,0}^2) \right. \\
& \quad + \pi_{i:C,NE} N(\mathbf{X}_i \boldsymbol{\beta}_{C,NE,0}, \sigma_{C,NE,0}^2) \\
& \quad \left. + \pi_{i:N,EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N,EE}, \sigma_{N,EE}^2) \right]^{\omega_i} \\
& \times \prod_{i \in O(0,0,0)} \left[ \pi_{i:C,EN} + \pi_{i:C,NN} + \pi_{i:N,NN} \right]^{\omega_i} \\
& \times \prod_{i \in O'(1,1)} \left[ \pi_{i:C,EE} + \pi_{i:C,EN} + \pi_{i:C,NE} + \pi_{i:C,NN} \right]^{\omega_i} \\
& \times \prod_{i \in O'(1,0)} \left[ \pi_{i:N,EE} + \pi_{i:N,NN} \right]^{\omega_i}
\end{aligned}$$

The units in the  $O'(1,1)$  and  $O'(1,0)$  groups bring information on their compliance behavior and affect the estimates of the  $\pi_{i:g}$  ( $i = 1, \dots, N$ ;  $g \in G$ ); the units in the  $O'(0,0)$  group are uninformative and disappear from the likelihood function (since  $\sum_g \pi_{i:g} = 1$ ). Once again, assuming monotonicity of truncation would imply to set  $\pi_{i:C.NE} = 0$  and proceed with  $k = 5$  instead of  $k = 6$ . The complete-data log-likelihood function may be written as follows:

$$\begin{aligned}
& l(\xi \mid \mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X}, \mathbf{G}) \\
&= \sum_{i \in O(1,1,1)} \omega_i I(G_i = C.EE) \log \left[ \pi_{i:C.EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,1}, \sigma_{C.EE,1}^2) \right] \\
&+ \sum_{i \in O(1,1,1)} \omega_i I(G_i = C.EN) \log \left[ \pi_{i:C.EN} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EN,1}, \sigma_{C.EN,1}^2) \right] \\
&+ \sum_{i \in O(1,1,0)} \omega_i I(G_i = C.NE) \log [\pi_{i:C.NE}] + \sum_{i \in O(1,1,0)} \omega_i I(G_i = C.NN) \log [\pi_{i:C.NN}] \\
&+ \sum_{i \in O(1,0,1)} \omega_i I(G_i = N.EE) \log \left[ \pi_{i:N.EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N.EE}, \sigma_{N.EE}^2) \right] \\
&+ \sum_{i \in O(1,0,0)} \omega_i I(G_i = N.NN) \log [\pi_{i:N.NN}] \\
&+ \sum_{i \in O(0,0,1)} \omega_i I(G_i = C.EE) \log \left[ \pi_{i:C.EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,0}, \sigma_{C.EE,0}^2) \right] \\
&+ \sum_{i \in O(0,0,1)} \omega_i I(G_i = C.NE) \log \left[ \pi_{i:C.NE} N(\mathbf{X}_i \boldsymbol{\beta}_{C.NE,0}, \sigma_{C.NE,0}^2) \right] \\
&+ \sum_{i \in O(0,0,1)} \omega_i I(G_i = N.EE) \log \left[ \pi_{i:N.EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N.EE}, \sigma_{N.EE}^2) \right] \\
&+ \sum_{i \in O(0,0,0)} \omega_i I(G_i = C.EN) \log [\pi_{i:C.EN}] \\
&+ \sum_{i \in O(0,0,0)} \omega_i I(G_i = C.NN) \log [\pi_{i:C.NN}] \\
&+ \sum_{i \in O(0,0,0)} \omega_i I(G_i = N.NN) \log [\pi_{i:N.NN}] \\
&+ \sum_{i \in O'(1,1)} \omega_i I(G_i = C.EE) \log [\pi_{i:C.EE}] + \sum_{i \in O'(1,1)} \omega_i I(G_i = C.EN) \log [\pi_{i:C.EN}] \\
&+ \sum_{i \in O'(1,1)} \omega_i I(G_i = C.NE) \log [\pi_{i:C.NE}] + \sum_{i \in O'(1,1)} \omega_i I(G_i = C.NN) \log [\pi_{i:C.NN}] \\
&+ \sum_{i \in O'(1,0)} \omega_i I(G_i = N.EE) \log [\pi_{i:N.EE}] + \sum_{i \in O'(1,0)} \omega_i I(G_i = N.NN) \log [\pi_{i:N.NN}] \\
&+ l(\boldsymbol{\theta})
\end{aligned}$$

where  $I(\cdot)$  is the general indicator function and  $l(\boldsymbol{\theta})$  contains the parameters of the missing data process; since  $\boldsymbol{\theta}$  is not of interest under the MAR assumption, we focus on the remaining parameters. Once an initial value  $\xi^{(0)}$  for the parameters vector has been chosen, the E-step of the EM algorithm computes the conditional probabilities of each stratum, given the

current estimates  $\xi^{(t)}$ :

for  $i \in O(1,1,1)$ ,

$$\begin{aligned} P^{(t)}(G_i = C.EE) &= \frac{\pi_{i:C.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,1}^{(t)}, \sigma_{C.EE,1}^{2(t)})}{\pi_{i:C.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,1}^{(t)}, \sigma_{C.EE,1}^{2(t)}) + \pi_{i:C.EN}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EN,1}^{(t)}, \sigma_{C.EN,1}^{2(t)})} \\ P^{(t)}(G_i = C.EN) &= \frac{\pi_{i:C.EN}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EN,1}^{(t)}, \sigma_{C.EN,1}^{2(t)})}{\pi_{i:C.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,1}^{(t)}, \sigma_{C.EE,1}^{2(t)}) + \pi_{i:C.EN}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EN,1}^{(t)}, \sigma_{C.EN,1}^{2(t)})} \\ P^{(t)}(G_i = C.NE) &= P^{(t)}(G_i = C.NN) = P^{(t)}(G_i = N.EE) = P^{(t)}(G_i = N.NN) = 0 \end{aligned}$$

for  $i \in O(1,1,0)$ ,

$$\begin{aligned} P^{(t)}(G_i = C.NE) &= \frac{\pi_{i:C.NE}^{(t)}}{\pi_{i:C.NE}^{(t)} + \pi_{i:C.NN}^{(t)}} \\ P^{(t)}(G_i = C.NN) &= \frac{\pi_{i:C.NN}^{(t)}}{\pi_{i:C.NE}^{(t)} + \pi_{i:C.NN}^{(t)}} \\ P^{(t)}(G_i = C.EE) &= P^{(t)}(G_i = C.EN) = P^{(t)}(G_i = N.EE) = P^{(t)}(G_i = N.NN) = 0 \end{aligned}$$

for  $i \in O(1,0,1)$ ,

$$\begin{aligned} P^{(t)}(G_i = N.EE) &= 1 \\ P^{(t)}(G_i = C.EE) &= P^{(t)}(G_i = C.EN) = P^{(t)}(G_i = C.NE) \\ &= P^{(t)}(G_i = C.NN) = P^{(t)}(G_i = N.NN) = 0 \end{aligned}$$

for  $i \in O(1,0,0)$ ,

$$\begin{aligned} P^{(t)}(G_i = N.NN) &= 1 \\ P^{(t)}(G_i = C.EE) &= P^{(t)}(G_i = C.EN) = P^{(t)}(G_i = C.NE) \\ &= P^{(t)}(G_i = C.NN) = P^{(t)}(G_i = N.EE) = 0 \end{aligned}$$

for  $i \in O(0,0,1)$ ,

$$\begin{aligned}
& P^{(t)}(G_i = C.EE) \\
&= \frac{\pi_{i:C.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,0}^{(t)}, \sigma_{C.EE,0}^{2(t)})}{\pi_{i:C.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,0}^{(t)}, \sigma_{C.EE,0}^{2(t)}) + \pi_{i:C.NE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.NE,0}^{(t)}, \sigma_{C.NE,0}^{2(t)}) + \pi_{i:N.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{N.EE}^{(t)}, \sigma_{N.EE}^{2(t)})} \\
& P^{(t)}(G_i = C.NE) \\
&= \frac{\pi_{i:C.NE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.NE,0}^{(t)}, \sigma_{C.NE,0}^{2(t)})}{\pi_{i:C.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,0}^{(t)}, \sigma_{C.EE,0}^{2(t)}) + \pi_{i:C.NE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.NE,0}^{(t)}, \sigma_{C.NE,0}^{2(t)}) + \pi_{i:N.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{N.EE}^{(t)}, \sigma_{N.EE}^{2(t)})} \\
& P^{(t)}(G_i = N.EE) \\
&= \frac{\pi_{i:N.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{N.EE}^{(t)}, \sigma_{N.EE}^{2(t)})}{\pi_{i:C.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.EE,0}^{(t)}, \sigma_{C.EE,0}^{2(t)}) + \pi_{i:C.NE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{C.NE,0}^{(t)}, \sigma_{C.NE,0}^{2(t)}) + \pi_{i:N.EE}^{(t)} N(\mathbf{X}_i \boldsymbol{\beta}_{N.EE}^{(t)}, \sigma_{N.EE}^{2(t)})}
\end{aligned}$$

$$P^{(t)}(G_i = C.EN) = P^{(t)}(G_i = C.NN) = P^{(t)}(G_i = N.NN) = 0$$

for  $i \in O(0,0,0)$ ,

$$\begin{aligned}
P^{(t)}(G_i = C.EN) &= \frac{\pi_{i:C.EN}^{(t)}}{\pi_{i:C.EN}^{(t)} + \pi_{i:C.NN}^{(t)} + \pi_{i:N.NN}^{(t)}} \\
P^{(t)}(G_i = C.NN) &= \frac{\pi_{i:C.NN}^{(t)}}{\pi_{i:C.EN}^{(t)} + \pi_{i:C.NN}^{(t)} + \pi_{i:N.NN}^{(t)}} \\
P^{(t)}(G_i = N.NN) &= \frac{\pi_{i:N.NN}^{(t)}}{\pi_{i:C.EN}^{(t)} + \pi_{i:C.NN}^{(t)} + \pi_{i:N.NN}^{(t)}} \\
P^{(t)}(G_i = C.EE) &= P^{(t)}(G_i = C.NE) = P^{(t)}(G_i = N.EE) = 0
\end{aligned}$$

for  $i \in O'(1,1)$ ,

$$\begin{aligned}
P^{(t)}(G_i = C.EE) &= \frac{\pi_{i:C.EE}^{(t)}}{\pi_{i:C.EE}^{(t)} + \pi_{i:C.EN}^{(t)} + \pi_{i:C.NE}^{(t)} + \pi_{i:C.NN}^{(t)}} \\
P^{(t)}(G_i = C.EN) &= \frac{\pi_{i:C.EN}^{(t)}}{\pi_{i:C.EE}^{(t)} + \pi_{i:C.EN}^{(t)} + \pi_{i:C.NE}^{(t)} + \pi_{i:C.NN}^{(t)}}
\end{aligned}$$

$$\begin{aligned}
P^{(t)}(G_i = C.NE) &= \frac{\pi_{i:C.NE}^{(t)}}{\pi_{i:C.EE}^{(t)} + \pi_{i:C.EN}^{(t)} + \pi_{i:C.NE}^{(t)} + \pi_{i:C.NN}^{(t)}} \\
P^{(t)}(G_i = C.NN) &= \frac{\pi_{i:C.NN}^{(t)}}{\pi_{i:C.EE}^{(t)} + \pi_{i:C.EN}^{(t)} + \pi_{i:C.NE}^{(t)} + \pi_{i:C.NN}^{(t)}} \\
P^{(t)}(G_i = N.EE) &= P^{(t)}(G_i = N.NN) = 0
\end{aligned}$$

for  $i \in O'(1,0)$ ,

$$\begin{aligned}
P^{(t)}(G_i = N.EE) &= \frac{\pi_{i:N.EE}^{(t)}}{\pi_{i:N.EE}^{(t)} + \pi_{i:N.NN}^{(t)}} \\
P^{(t)}(G_i = N.NN) &= \frac{\pi_{i:N.NN}^{(t)}}{\pi_{i:N.EE}^{(t)} + \pi_{i:N.NN}^{(t)}} \\
P^{(t)}(G_i = C.EE) &= P^{(t)}(G_i = C.EN) = P^{(t)}(G_i = C.NE)P^{(t)}(G_i = C.NN) = 0
\end{aligned}$$

The above conditional probabilities are the estimates of the unknown indicator functions  $I(\cdot)$  in the complete-data log-likelihood function; replacing the  $I(G_i = g)$  with the  $P^{(t)}(G_i = g)$  we obtain the expected log-likelihood  $l_E(\xi | \mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X})$ . The M-step consists in optimizing  $l_E(\cdot)$  with respect to the parameters vector  $\xi$ , leading to a new estimate  $\xi^{(t+1)}$ : to update  $\boldsymbol{\beta}$  and  $\boldsymbol{\sigma}$ , a standard routine for linear regression models can be used; a procedure for multinomial logistic models is needed in estimating  $\boldsymbol{\alpha}$ , given the current posterior probabilities. As showed in Dempster et al., iterating this process monotonically increases the likelihood function, or at least leaves it unchanged; the algorithm runs until a stopping criterion has been satisfied.

As in any finite mixture of Normal distributions, the log-likelihood function is unbounded and the EM algorithm may fall in a spurious maximum: in this case, the procedure must be restarted with new starting values. In addition, there often exist other solutions which may be regarded as spurious, lying very close to the edge of the parameter space: this happens when a component with very small variance is fitted; usually, this component density constitutes a cluster containing a few data points, very close together or almost lying in the same subspace. Such estimate tends to “interpolate” a local pattern and provides a bad fit for the remaining observations; as a consequence, the fitted model is not of practical use in inference.

Another complication is that the log-likelihood function presents an unknown number of local solutions: the best one – that is, the one with the higher log-likelihood value – is usually chosen as the MLE. For this reason, a great number of different starting values for the EM algorithm should be used.

Once a parameter estimate has been obtained, the causal effects of interest can be evaluated. In a regression approach, this requires to average on the covariates distribution; moreover, the causal effects are expressed in the natural scale, whereas the model is estimated on the logarithm of the wages. Following Zhang et al. (2007), we estimate the pro-

portion of each stratum as

$$\hat{\pi}_g = \frac{\sum_{i=1}^N \omega_i \hat{\pi}_{i:g}}{\sum_{i=1}^N \omega_i}$$

the causal effect of Z on D (which is not of special interest) is estimated as the proportion of compliers:

$$\hat{\Delta}^{(ZD)} = \hat{\pi}_{C,EE} + \hat{\pi}_{C,EN} + \hat{\pi}_{C,NE} + \hat{\pi}_{C,NN}$$

Consistent estimates of the average treatment effects on wages and employment are obtained as

$$\hat{\Delta}_*^{(DW)} = \frac{\sum_{i=1}^N \omega_i \hat{\pi}_{i:C,EE} \exp\left\{\mathbf{X}_i \hat{\boldsymbol{\beta}}_{C,EE,1} + \frac{\hat{\sigma}_{C,EE,1}^2}{2}\right\}}{\sum_{i=1}^N \omega_i \hat{\pi}_{i:C,EE}} - \frac{\sum_{i=1}^N \omega_i \hat{\pi}_{i:C,EE} \exp\left\{\mathbf{X}_i \hat{\boldsymbol{\beta}}_{C,EE,0} + \frac{\hat{\sigma}_{C,EE,0}^2}{2}\right\}}{\sum_{i=1}^N \omega_i \hat{\pi}_{i:C,EE}}$$

and

$$\hat{\Delta}_*^{(DS)} = \hat{\pi}_{C,EN} - \hat{\pi}_{C,NE}$$

respectively. We can see that the monotonicity of truncation – setting  $\hat{\pi}_{C,NE} = 0$  – forces the estimate of the treatment effect on employment to be positive. Once the asymptotic covariance matrix of the estimates has been obtained, the standard errors of the above quantities may be computed using the Delta method.

In Section 5.7, we present the application of this framework to the Job Corps Study; the model has been estimated on the outcomes of the 45<sup>th</sup>, 135<sup>th</sup> and 208<sup>th</sup> week: very different treatment effects are obtained in the short and in the long run; we will see that the monotonicity of truncation plays a very critical role in the model identification.

## 5.7 Application to the Job Corps Study

The evaluation of government-sponsored job-training programs is a difficult task, undertaken by a number of authors in last decades (Heckman and Hotz, 1989, Lalonde, 1995, Burghardt et al., 2001, Zhang, Rubin and Mealli, 2008a, 2008b).

For our analysis, we use the data from the National Job Corps Study (conducted by Mathematica Policy Research, Inc. for the U.S. Department of Labor) and estimate the effect of the program on employment and wages. The data are from a random sample of all selected applicants (N = 15,386) in 1994 and 1995: among them, a random assignment to



the program or to the control group were done; only those assigned to the program group (9,409 units, about 61%) were admitted to enroll in the Job Corps: among them, 6,039 (64%) complied with the assignment.

For all units, pre-treatment covariates ( $\mathbf{X}$ ) were collected. In principle, including covariates in the analysis is not fundamental in estimating the treatment effects: the covariates effect is not of main interest and – by virtue of the randomization process – the covariates distribution is independent of the treatment assignment. However, the covariates are helpful and necessary for three main reasons: first, they generally improve the model identification and the prediction of the unobserved potential outcomes; second, they allow a more plausible generalization to a population with different characteristics; third, conditioning on covariates is explicitly required by the MAR assumption. The summary statistics of the pre-treatment covariates ( $\mathbf{X}$ ) are displayed in Table 5.2 ( $N = 15,376$ : we removed 2 observations aged more than 30 – all others units are aged 16-24 – and 8 units with overly large ( $> 50,000$ ) values for earnings in the previous year). We imputed the missing values in  $\mathbf{X}$  using the `mice` procedure in R, which generates multiple imputations for incomplete multivariate data by Gibbs Sampling; we used only the baseline covariates as predictor in the chained equations. Linear regression has been used for numerical covariates; binary/multinomial logistic models for dichotomous/polytomous variables. Ten different imputation have been generated, leading to very similar estimates of the model parameters: for this reason, we only present the results from one single imputed data set. In the linear predictor, the education degree (number of scholar years attended by the young and his parents) has been included as a dummy variable (= 1 if greater than the sample median); we also collapsed the information on the marital status in the dummy “partnered”.

In our analysis, we considered as missing all inadmissible outcomes (units with more than 84 weekly hours, employed people with zero weekly earnings or hours). Table 5.3 presents the summary statistics of the outcome variables  $\mathbf{Y}$  (employment, total earnings and weekly hours at 45<sup>th</sup>, 135<sup>th</sup> and 208<sup>th</sup> week after treatment).

For simplicity, we assumed that the treatment assignment for compliers enters in the linear predictor without interactions with the covariates; that is,  $\beta_{C.EE,1}$  and  $\beta_{C.EE,0}$  only differ in the intercept, so that

$$\mathbf{X}_i \beta_{C.EE,1} = \mathbf{X}_i \beta_{C.EE,0} + \gamma$$

for each  $i$ . As in a standard linear model, we also assumed that the treatment receipt in the C.EE group has no effect on the variance: this implies  $\sigma_{C.EE,0}^2 = \sigma_{C.EE,1}^2$ . The exclusion restriction is always maintained – that is, we constrained the causal effects for never-takers to be zero. Violations of the exclusion restriction have no testable consequences and – in this case – we believe that this assumption is plausible; however, units who refused the treatment could regret the vanished opportunity: we do not know if this eventuality would have some consequences in terms of potential employment status and potential wages.

With the above set of assumptions and simplifications, the model has 221 parameters; the monotonicity of truncation assumption rules out the C.NE group and reduces the number of parameters to 175. For the 3 weeks under study, we estimated the model with and without monotonicity assumption; a genetic algorithm was used in the search of the “best” local maximum of the log-likelihood function; the EM algorithm was stopped when the maximum absolute change in the parameters vector between two consecutive iterations

was smaller than 0.0001. For each model, the asymptotic covariance matrix was obtained by analytical evaluation of the Hessian of the log-likelihood function. The causal effects on employment and wages have been computed and approximate standard errors have been obtained by means of the Delta method; also in this case, analytical derivatives have been used.

## 5.8 Results

Tables 5.4 and 5.5 present results without and with monotonicity of truncation assumption, respectively. Without imposing monotonicity (Table 5.4), for week 45 we found evidence of all latent strata; we estimated a negative treatment effect on employment ( $-8.22\%$ ), whereas the effect on wages is found to be positive (about 0.276 \$/hour). For weeks 135 and 208, a positive treatment effect was found on both employment ( $+4.87\%$  and  $+4.85\%$ , respectively) and wages (0.210 and 0.337 \$/hour, respectively). The estimated probability of the C.NE group was found to be very high (15.47%) in week 45, but negligible in the subsequent weeks (1.37% in week 135, 1.35% in week 208).

Because of this lack of evidence of the C.NE group, we also estimated the model imposing monotonicity of truncation; results are displayed in Table 5.5. With respect to Table 5.4, completely different estimates are found for week 45: the effect on employment is constrained to be positive and very different probabilities are obtained for the C.EE, C.EN and C.NN strata. According to the AIC (Akaike's Information Criterion) the monotonicity of truncation should be rejected, whereas the BIC (Bayesian Information Criterion), which generally penalizes models with a great number of parameters, indicates that setting  $p_{i:C.NE} = 0$  causes a nonsignificant reduction of the model fit; however, we have no reason to cross out a component which is known a priori to exist: since we found a strong evidence of the C.NE stratum, we believe that the monotonicity of truncation does not hold for week 45. In week 135 and 208, the estimates under monotonicity are very similar to those of Table 5.4; in both cases, the BIC suggests that the model with monotonicity should be preferred, whereas the opposite conclusion is drawn according to the AIC; however, for weeks 135 and 208 this assumption appears to be quite reasonable.

In the short run, there are trained units that choose to be unemployed and wait for a better job: this results in a negative effect on the employment; in the long run, the C.NE group tends to disappear and a greater employment rate is observed among trained units. These results are consistent with the empirical literature on the effect of active labor market policies, which suggest that almost all programs reduce employment and earnings in the short run. This so-called "lock-in" effect is well documented in many studies and can be also attributed to reduced search intensity of participants or fewer job offers during the program (Lechner and Wunsch, 2007; van Ours, 2004).

The effect on wages is found to be higher in week 45 than in week 135; this corroborates the above arguments: in the short run, treated units generally have a well remunerated job or, otherwise, choose to be unemployed; in the long run, different criteria are used and the job selection become less strict: for this reason, a still positive but lower effect on wages (together with a positive effect on employment) is observed in week 135. However, the gap between treated and untreated units increases over time, as comes out from the estimates of the treatment effect on wages for week 208.

These results also demonstrate how crucial is the monotonicity of truncation in this case: for weeks 135 and 208, it seems reasonable that this assumption holds, because there is a very weak evidence of the C.NE group, whereas in week 45 – according to the former estimates – monotonicity is not a plausible assumption.

The obtained results may be sensitive to our working assumptions; in particular, the exclusion restriction could be questionable, because units who refused the treatment could regret the vanished opportunity; however, removing this assumptions may be detrimental in terms of model identification. A possible strategy to improve identification is to use a multivariate model – e.g., a bivariate normal distribution for the couple  $(\log(W), \log(H))$ , where  $W$  denotes the hourly wage and  $H$  the weekly working hours. With a bivariate approach, an increased efficiency could be achieved; with the aim of decomposing a finite mixture with a great number of component ( $k = 8$  if the exclusion restriction and the monotonicity of truncation are assumed to not hold), using a double classification criterion (wages and hours worked) is also expected to reduce the occurrence of spurious optimizers, which represents a serious problem in the estimation of the univariate model we presented here.

Variable	Treatment			Control			Difference	
	Prop. non-miss.	Mean	Std. Dev.	Prop. non-miss.	Mean	Std. Dev.	Diff.	Std. Err.
Female	0.96	0.41	0.49	0.95	0.40	0.49	0.00	0.01
Age at baseline	0.96	18.85	2.18	0.95	18.82	2.15	0.03	0.04
White, non-Hispanic	1.00	0.30	0.46	1.00	0.30	0.46	0.00	0.01
Black, non-Hispanic	1.00	0.46	0.50	1.00	0.45	0.50	0.01	0.01
Hispanic	1.00	0.17	0.37	1.00	0.17	0.38	0.00	0.01
Other race	1.00	0.07	0.26	1.00	0.07	0.26	0.00	0.00
Never married	0.94	0.91	0.28	0.92	0.91	0.28	0.00	0.00
Married	0.94	0.02	0.14	0.92	0.02	0.14	0.00	0.00
Living together	0.94	0.04	0.20	0.92	0.04	0.20	0.00	0.00
Separated	0.94	0.02	0.16	0.92	0.02	0.14	0.00	0.00
<i>Partnered</i>	<i>0.94</i>	<i>0.06</i>	<i>0.24</i>	<i>0.92</i>	<i>0.06</i>	<i>0.24</i>	<i>0.00</i>	<i>0.00</i>
Has children	0.99	0.17	0.38	0.99	0.17	0.38	0.00	0.01
Number of children	0.99	0.24	0.61	0.98	0.23	0.59	0.01	0.01
Education	0.93	10.06	1.52	0.92	10.07	1.52	-0.01	0.03
Mother's education	0.76	11.52	2.56	0.74	11.53	2.62	-0.01	0.05
Father's education	0.58	11.47	2.87	0.56	11.57	2.84	-0.10	0.06
Ever arrested	0.94	0.26	0.44	0.92	0.26	0.44	0.00	0.01
Household Inc. < 3000	0.59	0.26	0.44	0.59	0.25	0.43	0.01	0.01
3000-6000	0.59	0.20	0.40	0.59	0.21	0.41	-0.01	0.01
6000-9000	0.59	0.11	0.32	0.59	0.11	0.31	0.00	0.01
9000-18000	0.59	0.25	0.43	0.59	0.25	0.43	0.00	0.01
> 18000	0.59	0.19	0.39	0.59	0.19	0.39	0.00	0.01
Personal Inc. < 3000	0.87	0.79	0.41	0.86	0.79	0.40	0.00	0.01
3000-6000	0.87	0.13	0.33	0.86	0.13	0.33	0.00	0.01
6000-9000	0.87	0.05	0.22	0.86	0.04	0.20	0.01	0.00 (*)
> 9000	0.87	0.03	0.18	0.86	0.03	0.18	0.00	0.00
At baseline:								
Have job	0.92	0.21	0.41	0.91	0.21	0.41	0.00	0.01
Had job, prev. yr.	0.94	0.65	0.48	0.92	0.64	0.48	0.01	0.01
Months empl., prev. yr.	0.89	3.77	4.26	0.88	3.75	4.30	0.01	0.07
Earnings, prev. yr.	0.87	2859.89	4210.62	0.86	2868.57	4350.31	-8.69	74.16
N	9409			5977				

Table 5.2 Summary statistics of the pre-treatment covariates; in the last column, (\*) denotes that the difference between the treatment and the control group is statistically significant at 0.05 level (all statistics have been computed before the imputation). All computations use design weights.

Variable	Treatment			Control			Difference	
	Prop. non-miss.	Mean	Std. Dev.	Prop. non-miss.	Mean	Std. Dev.	Diff.	Std. Err.
Week 45								
Employed	0.88	0.35	0.48	0.85	0.43	0.49	-0.08	0.01 (*)
Weekly earnings	0.88	89.19	154.32	0.85	103.39	150.82	-14.19	2.64 (*)
Weekly hours	0.88	14.49	21.83	0.85	17.49	22.52	-3.01	0.38 (*)
Week 135								
Employed	0.76	0.54	0.49	0.76	0.52	0.50	0.03	0.01 (*)
Weekly earnings	0.76	182.16	217.87	0.76	164.24	201.59	17.92	3.88 (*)
Weekly hours	0.76	23.92	24.31	0.76	22.51	24.07	1.41	0.45 (*)
Week 208								
Employed	0.67	0.60	0.49	0.68	0.56	0.50	0.04	0.01 (*)
Weekly earnings	0.67	220.15	240.66	0.68	194.88	219.51	25.27	4.52 (*)
Weekly hours	0.67	26.54	24.12	0.68	24.41	23.86	2.13	0.47 (*)

Table 5.3 Summary statistics of the outcome variables. In the last column, (\*) denotes that the difference between the treatment and the control group is statistically significant at 0.05 level. All computations use design weights.

week	$\hat{\pi}_{C,EE}$	$\hat{\pi}_{C,EN}$	$\hat{\pi}_{C,NE}$	$\hat{\pi}_{C,NN}$	$\hat{\pi}_{N,EE}$	$\hat{\pi}_{N,NN}$	$\hat{\Delta}_*^{(DS)}$	$\hat{\Delta}_*^{(DW)}$	BIC	AIC
45	0.1643 (0.0059)	0.0725 (0.0054)	0.1547 (0.0072)	0.3223 (0.0074)	0.1205 (0.0041)	0.1656 (0.0048)	-0.0822 (0.0072)	0.2757 (0.0523)	28553.4	26864.8
135	0.3328 (0.0057)	0.0624 (0.0049)	0.0137 (0.0016)	0.2894 (0.0051)	0.1636 (0.0045)	0.1381 (0.0045)	0.0487 (0.0050)	0.2099 (0.0576)	28292.9	26604.4
208	0.3789 (0.0060)	0.0620 (0.0049)	0.0135 (0.0015)	0.2549 (0.0052)	0.1713 (0.0051)	0.1194 (0.0050)	0.0485 (0.0051)	0.3374 (0.0668)	25821.2	24132.6

Table 5.4 *Maximum likelihood estimates – adjusted for covariates and without monotonicity of truncation – of the average treatment effects on employment ( $\Delta_*^{(DS)}$ ) and wages ( $\Delta_*^{(DW)}$ ) for week 45, 135 and 208 (asymptotic standard errors between brackets). For each week, we provide the estimated proportion of each principal stratum; the BIC and AIC are returned for a comparison with the results in Table 5.5.*

week	$\hat{\pi}_{C,EE}$	$\hat{\pi}_{C,EN}$	$\hat{\pi}_{C,NN}$	$\hat{\pi}_{N,EE}$	$\hat{\pi}_{N,NN}$	$\hat{\Delta}_*^{(DS)}$	$\hat{\Delta}_*^{(DW)}$	BIC	AIC
45	0.2468 (0.0045)	0.0294 (0.0025)	0.4382 (0.0048)	0.1333 (0.0041)	0.1523 (0.0043)	0.0294 (0.0025)	0.1693 (0.0426)	28505.7	27191.5
135	0.3389 (0.0057)	0.0592 (0.0047)	0.3001 (0.0049)	0.1652 (0.0046)	0.1364 (0.0045)	0.0592 (0.0047)	0.2070 (0.0571)	28127.5	26813.3
208	0.3865 (0.0059)	0.0559 (0.0047)	0.2672 (0.0050)	0.1724 (0.0052)	0.1181 (0.0049)	0.0559 (0.0047)	0.3217 (0.0669)	25696.1	24381.9

Table 5.5 *Maximum likelihood estimates – adjusted for covariates and assuming monotonicity of truncation – of the average treatment effects on employment ( $\Delta_*^{(DS)}$ ) and wages ( $\Delta_*^{(DW)}$ ) for week 45, 135 and 208 (asymptotic standard errors between brackets). For each week, we provide the estimated proportion of each principal stratum; the BIC and AIC are returned for a comparison with the results in Table 5.4.*







## Concluding remarks

In this thesis, the general framework of finite mixture models has been presented. In Chapter 2, the very relevant issue of maximizing the log-likelihood function has been discussed under different viewpoints. With a simulation study, we demonstrated that finding the true MLE of a mixture model can be a very difficult task. In many settings, the EM algorithm has a great risk of falling in a spurious/local optimizer or in a saddle point; moreover, the local maxima are not generally recognizable. For these reasons, running the EM from a variety of starting values is always recommended.

Using a genetic algorithm can be a valid approach to this optimization problem. A wide simulation study has been carried out using the `gen.start` procedure of the `mixglm` package, presented in Chapter 4. Our genetic algorithm is found to be effective: however, very different results are obtained according to the operational parameters. Simulations indicate that a slow selection process, together with a continuous renewal of the genetic heritage, leads to more satisfactory results; the convergence criterion and the population size are also relevant parameters.

A related issue has been presented in Chapter 3, where the advantages of using multivariate mixture models have been discussed. For a simulated data set with two responses variables ( $y_1$  and  $y_2$ ), we estimated the parameters vector using a) the univariate approach, where two independent models are specified for  $y_1$  and  $y_2$ ; b) a bivariate model for the couple  $(y_1, y_2)$ . Results indicate that the bivariate approach leads to more efficient estimates, together with a considerable gain in the computation time; due to the increased discriminating power, a better estimate of the unknown cluster membership is obtained; finally, we found evidence that using a bivariate model decreases the risk of falling in a local/spurious optimizer.

The last part of this work is devoted to the analysis of treatment effects in randomized studies. In our dissertation, we followed the general framework of the Rubin Causal Model; a special attention is devoted to three post-treatment complications, namely noncompliance, missing outcomes, and outcomes truncated by death.

In Chapter 5, we evaluated the effects of the Job Corps training program on employment and wages, using data from a randomized study, the National Job Corps Study, and the principal stratification approach to simultaneously address the issues of noncompliance and truncation of wages – meaning that no wages are observed for those who are unemployed – under the MAR assumption for the missing outcome mechanism. The Principal Stratification approach consists in estimating the causal effects of interest for a common set of units: the average treatment effect on employment is estimated on the subpopulation of compliers; among them, only for the always employed the effect on wages is defined and estimated in a meaningful way.

Pre-treatment covariates were used in the prediction of the outcomes and of the compliance behavior; we focused our analysis on the observed outcomes at 45<sup>th</sup>, 135<sup>th</sup> and 208<sup>th</sup> week after participation in the program. The exclusion restriction for never-takers was maintained, whereas the presence of always-takers and defiers is excluded by design. Some restrictions on covariates reduced the number of parameters and improved model identification.

The treatment effects were found to be increasing in the course of time; the effect on

employment was negative at week 45, whereas the estimated effect on wages is always positive. A critical role is played by the monotonicity of truncation assumption, which rules out those who would be unemployed if treated and employed if not treated: this assumption does not seem to hold at week 45, but becomes more plausible at weeks 135 and 208. We may argue that, in a short run, there are trained units who choose to be unemployed, waiting for a “good” job; in the long run, trained units are more likely to find a job, and a positive treatment effect on employment is found. These results are consistent with the empirical literature on the effect of active labor market policies, which suggest that almost all programs reduce employment and earnings in the short run (the so-called “lock-in” effect). Finally, the effect on wages reflects the increase in the human capital due to the program participation.

The obtained results may be sensitive to our working assumptions; however, relaxing some hypotheses may weaken identification and lead to poor estimates in terms of efficiency. In order to overcome the difficulties inherent in the lack of full identification when the exclusion restriction is relaxed, a possible strategy could be the simultaneous modelling of more than one outcome; indeed, the use of multivariate models generally provides a greater discriminant power in disentangling mixtures.

## APPENDIX A

### Parameters estimates

We provide the parameters estimates for all fitted models. The following covariates have been used in the linear predictors:

age  
female (=1 if the unit is a female)  
evarrst (= 1 if the unit has been arrested one or more times)  
haschld (= 1 if the unit has children)  
nchld (number of children)  
partnered (= 1 if the unit is married or living together a partner)

#### *Education*

educ (= 1 if the unit's education is greater than the sample median [10])  
educ.f (= 1 if the father's education is greater than the sample median [12])  
educ.m (= 1 if the mother's education is greater than the sample median [12])

#### *Race (base = other)*

white (race = white)  
hisp (race = hispanic)  
black (race = black)

#### *Employment*

yr.work1 (= 1 if the unit had job in previous year)  
earn.yr (earnings in previous year, standardized)  
mosinjob (months employed in previous year)  
currjob (= 1 if unit has job at baseline)

#### *Personal income (base: < 3000)*

p.inc 3000-6000  
p.inc 6000-9000  
p.inc > 9000

#### *Household income (base: < 3000)*

h.inc 3000-6000  
h.inc 6000-9000  
h.inc 9000-18000  
h.inc > 18000

The treatment indicator enters in the linear predictors as a dummy variable, without interaction with other covariates; only in the C.EE group the effect is allowed to differ from zero. The estimated models for week 45, 135 and 208 are named as follows: a.45, a.135, a.208 (without monotonicity of truncation); b.45, b.135, b.208 (with monotonicity of truncation). For each model, the estimates of all parameters are returned (between brackets, the estimated standard errors).

**Model: a.45**  
**(week 45, without monotonicity of truncation)**

	$\hat{\beta}_{C,EE}$	$\hat{\beta}_{C,EN,1}$	$\hat{\beta}_{C,NE,0}$	$\hat{\beta}_{N,EE}$
intercept	1.4999 (0.0613)	1.6962 (0.3588)	1.6454 (0.1234)	1.3475 (0.1551)
age	0.0084 (0.0031)	-0.0006 (0.0178)	0.0082 (0.0053)	0.0154 (0.0076)
female	-0.0220 (0.0096)	-0.1892 (0.0661)	-0.0412 (0.0225)	-0.1849 (0.0286)
evarrst	-0.0094 (0.0105)	-0.0078 (0.0654)	0.0893 (0.0299)	-0.0116 (0.0314)
haschld	0.0415 (0.0255)	0.0853 (0.1724)	0.1200 (0.0527)	0.0576 (0.0672)
nchld	-0.0096 (0.0175)	-0.0286 (0.1094)	-0.0677 (0.0299)	0.0154 (0.0420)
partnered	0.0069 (0.0176)	0.1738 (0.1452)	-0.0801 (0.0499)	0.0669 (0.0467)
educ	-0.0079 (0.0112)	0.0518 (0.0678)	0.0457 (0.0219)	0.0309 (0.0304)
educ.f	0.0137 (0.0135)	0.1344 (0.0759)	0.0038 (0.0288)	-0.0004 (0.0357)
educ.m	0.0245 (0.0128)	-0.0678 (0.0727)	0.0023 (0.0264)	0.0307 (0.0349)
white	0.0112 (0.0203)	-0.0232 (0.1172)	-0.1162 (0.0496)	-0.0159 (0.0555)
hispanic	-0.0037 (0.0218)	0.0046 (0.1297)	-0.1121 (0.0473)	0.0256 (0.0591)
black	-0.0007 (0.0199)	-0.0203 (0.1157)	-0.1500 (0.0415)	-0.0057 (0.0544)
yr.work1	-0.0030 (0.0135)	0.1044 (0.0847)	0.0441 (0.0303)	0.0395 (0.0398)
earn.yr	0.0866 (0.0157)	0.1028 (0.0544)	0.0128 (0.0170)	0.0428 (0.0263)
mosinjob	-0.0085 (0.0023)	-0.0259 (0.0126)	-0.0020 (0.0044)	0.0010 (0.0055)
currjob	0.0103 (0.0118)	0.0157 (0.0736)	0.0006 (0.0264)	-0.0148 (0.0317)
p.inc 3000-6000	0.0181 (0.0151)	0.1931 (0.0895)	0.1001 (0.0330)	-0.0402 (0.0383)
p.inc 6000-9000	0.0529 (0.0217)	0.2200 (0.1277)	0.1531 (0.0506)	0.0368 (0.0563)
p.inc > 9000	-0.0049 (0.0321)	0.1576 (0.1471)	0.0861 (0.0491)	-0.0153 (0.0628)
h.inc 3000-6000	0.0173 (0.0149)	0.0969 (0.0898)	-0.0476 (0.0332)	0.0922 (0.0432)
h.inc 6000-9000	-0.0127 (0.0176)	0.1261 (0.1153)	0.0414 (0.0330)	0.0622 (0.0511)
h.inc 9000-18000	0.0117 (0.0132)	0.1683 (0.0919)	0.0660 (0.0400)	0.1257 (0.0416)
h.inc > 18000	0.0065 (0.0153)	0.1674 (0.0916)	0.0128 (0.0315)	0.1529 (0.0440)
Treatment	0.0513 (0.0098)	0 (-)	0 (-)	0 (-)

	$\hat{\sigma}_{C,EE}$	$\hat{\sigma}_{C,EN,1}$	$\hat{\sigma}_{C,NE,0}$	$\hat{\sigma}_{N,EE}$
	0.1442 (0.0060)	0.5806 (0.0252)	0.1655 (0.0089)	0.4936 (0.0094)

	$\hat{\alpha}_{C.EE}$	$\hat{\alpha}_{C.EN}$	$\hat{\alpha}_{C.NE}$	$\hat{\alpha}_{C.NN}$	$\hat{\alpha}_{N.EE}$
intercept	-0.5950 (0.4940)	-2.5135 (0.7561)	-2.7430 (0.5730)	3.1296 (0.5406)	-2.4265 (0.5336)
age	-0.0032 (0.0247)	0.0725 (0.0371)	0.0082 (0.0273)	-0.1152 (0.0283)	0.0710 (0.0266)
female	-0.2143 (0.0885)	-0.5907 (0.1453)	-0.5242 (0.1093)	-0.2624 (0.0886)	-0.2979 (0.0981)
evarrst	-0.3664 (0.0928)	-0.3814 (0.1460)	-0.5729 (0.1165)	-0.4333 (0.0933)	-0.3575 (0.1051)
haschld	0.1565 (0.2113)	-0.2553 (0.3964)	-0.2059 (0.2835)	-0.2481 (0.2032)	-0.0027 (0.2045)
nchld	-0.3192 (0.1387)	-0.2008 (0.2510)	-0.2615 (0.1815)	-0.0558 (0.1240)	-0.2158 (0.1230)
partnered	-0.0478 (0.1484)	-0.5744 (0.2972)	-0.8328 (0.2325)	-0.4027 (0.1847)	-0.0391 (0.1585)
educ	0.0304 (0.0983)	0.1262 (0.1508)	0.0675 (0.1109)	-0.0552 (0.1054)	0.3276 (0.1073)
educ.f	0.0211 (0.1190)	0.1511 (0.1752)	0.0454 (0.1336)	-0.1596 (0.1262)	0.0219 (0.1314)
educ.m	0.0684 (0.1206)	0.3823 (0.1725)	0.3194 (0.1318)	0.2003 (0.1215)	0.2066 (0.1302)
white	0.4175 (0.1809)	0.3207 (0.2748)	-0.0082 (0.1935)	-0.3478 (0.1751)	0.2955 (0.1933)
hisp	-0.0231 (0.1925)	0.0276 (0.2954)	-0.0499 (0.2058)	-0.0927 (0.1748)	0.0825 (0.1907)
black	0.0856 (0.1734)	0.0309 (0.2646)	-0.1943 (0.1878)	-0.0442 (0.1582)	-0.0363 (0.1727)
yr.work1	0.4167 (0.1177)	0.3766 (0.1833)	0.3625 (0.1437)	0.0348 (0.1183)	0.3110 (0.1488)
earn.yr	0.0010 (0.1092)	0.2407 (0.1616)	0.0631 (0.1148)	-0.4503 (0.1660)	0.0917 (0.1206)
mosinjob	0.0449 (0.0205)	0.0051 (0.0309)	0.0619 (0.0223)	0.0165 (0.0263)	0.0508 (0.0228)
currjob	0.2306 (0.1123)	0.1735 (0.1670)	0.1565 (0.1218)	-0.4115 (0.1395)	0.2050 (0.1403)
p.inc 3000-6000	0.0151 (0.1306)	0.0096 (0.2052)	-0.1386 (0.1482)	-0.1260 (0.1554)	0.1515 (0.1419)
p.inc 6000-9000	0.2582 (0.2013)	0.0695 (0.3100)	-0.2090 (0.2329)	0.1750 (0.2476)	0.1561 (0.2197)
p.inc > 9000	0.2388 (0.2975)	0.5379 (0.3852)	0.2544 (0.2964)	0.2824 (0.3887)	0.5709 (0.3031)
h.inc 3000-6000	0.0752 (0.1231)	0.2436 (0.1990)	0.0170 (0.1510)	0.0058 (0.1151)	0.1472 (0.1352)
h.inc 6000-9000	0.1361 (0.1526)	0.0146 (0.2655)	0.1314 (0.1770)	0.1020 (0.1452)	0.1716 (0.1706)
h.inc 9000-18000	0.4058 (0.1230)	0.0881 (0.2126)	-0.0029 (0.1586)	0.2404 (0.1193)	0.3617 (0.1392)
h.inc > 18000	0.2906 (0.1386)	0.4999 (0.2092)	0.5136 (0.1518)	-0.1058 (0.1456)	0.5111 (0.1532)

**Model: a.135**  
**(week 135, without monotonicity of truncation)**

	$\hat{\beta}_{C,EE}$	$\hat{\beta}_{C,EN,1}$	$\hat{\beta}_{C,NE,0}$	$\hat{\beta}_{N,EE}$
intercept	1.7834 (0.0506)	1.9857 (0.4730)	2.6356 (0.0071)	1.7670 (0.1390)
age	0.0056 (0.0025)	0.0100 (0.0238)	-0.0129 (0.0004)	0.0040 (0.0070)
female	-0.0661 (0.0089)	-0.1499 (0.0919)	-0.5128 (0.0012)	-0.1295 (0.0254)
evarrst	-0.0168 (0.0098)	-0.0454 (0.0954)	-0.1047 (0.0011)	0.0765 (0.0278)
haschld	0.0150 (0.0238)	-0.0995 (0.2912)	-0.2428 (0.0037)	0.0405 (0.0574)
nchld	-0.0059 (0.0147)	0.1451 (0.2182)	0.0827 (0.0036)	0.0163 (0.0343)
partnered	0.0231 (0.0210)	-0.3762 (0.1786)	0.1100 (0.0023)	0.0292 (0.0434)
educ	0.0217 (0.0100)	0.0902 (0.0981)	0.0246 (0.0010)	0.1302 (0.0278)
educ.f	0.0188 (0.0117)	0.0217 (0.1240)	0.0708 (0.0020)	0.0549 (0.0338)
educ.m	0.0068 (0.0116)	0.1116 (0.1141)	-0.0550 (0.0013)	0.0776 (0.0332)
white	-0.0084 (0.0183)	-0.0747 (0.1475)	0.1423 (0.0014)	-0.0458 (0.0489)
hisp	0.0160 (0.0197)	-0.0657 (0.1588)	0.3006 (0.0015)	0.0212 (0.0515)
black	-0.0174 (0.0175)	-0.2046 (0.1499)	0.1149 (0.0012)	-0.1162 (0.0479)
yr.work1	0.0240 (0.0120)	0.1182 (0.1208)	0.0109 (0.0015)	0.1016 (0.0347)
earn.yr	0.0327 (0.0128)	0.0296 (0.0699)	0.1546 (0.0008)	0.0475 (0.0248)
mosinjob	-0.0025 (0.0021)	-0.0127 (0.0170)	-0.0197 (0.0002)	-0.0030 (0.0052)
currjob	0.0070 (0.0114)	0.1811 (0.1054)	0.0916 (0.0012)	-0.0141 (0.0305)
p.inc 3000-6000	0.0420 (0.0132)	0.2046 (0.1404)	0.0717 (0.0405)	0.0397 (0.0352)
p.inc 6000-9000	0.0732 (0.0212)	0.3075 (0.1903)	-0.0004 (0.0013)	0.0127 (0.0571)
p.inc > 9000	0.0605 (0.0270)	0.2608 (0.2309)	0.2475 (0.0017)	0.0782 (0.0630)
h.inc 3000-6000	0.0097 (0.0124)	-0.1573 (0.1298)	-0.0662 (0.0013)	-0.0057 (0.0375)
h.inc 6000-9000	0.0071 (0.0159)	-0.1268 (0.1483)	-0.0351 (0.0028)	-0.0081 (0.0438)
h.inc 9000-18000	0.0033 (0.0124)	-0.1191 (0.1205)	-0.0326 (0.0015)	0.0291 (0.0358)
h.inc > 18000	0.0254 (0.0134)	-0.1233 (0.1339)	-0.0509 (0.0015)	0.0518 (0.0397)
Treatment	0.0302 (0.0083)	0 (-)	0 (-)	0 (-)

	$\hat{\sigma}_{C,EE}$	$\hat{\sigma}_{C,EN,1}$	$\hat{\sigma}_{C,NE,0}$	$\hat{\sigma}_{N,EE}$
	0.2093 (0.0039)	0.7093 (0.0324)	0.0025 (0.0003)	0.4867 (0.0084)

	$\hat{\alpha}_{C,EE}$	$\hat{\alpha}_{C,EN}$	$\hat{\alpha}_{C,NE}$	$\hat{\alpha}_{C,NN}$	$\hat{\alpha}_{N,EE}$
intercept	0.2376 (0.4281)	- 1.2779 (0.8549)	- 0.9570 (1.5180)	2.1216 (0.5108)	- 0.7597 (0.5397)
age	0.0313 (0.0217)	0.0442 (0.0431)	- 0.1088 (0.0821)	- 0.0557 (0.0262)	0.0342 (0.0276)
female	- 0.4378 (0.0779)	- 0.5472 (0.1677)	1.2512 (0.3038)	- 0.2034 (0.0904)	- 0.3019 (0.0992)
evarrst	- 0.4669 (0.0801)	- 0.5191 (0.1736)	0.3903 (0.2494)	- 0.3934 (0.0939)	- 0.3923 (0.1050)
haschld	- 0.1803 (0.1658)	0.2274 (0.5020)	3.6714 (0.9129)	- 0.2588 (0.2032)	- 0.0129 (0.1983)
nchld	- 0.1675 (0.0977)	- 0.4498 (0.3643)	- 2.0196 (0.8282)	- 0.1238 (0.1206)	- 0.1130 (0.1141)
partnered	- 0.7811 (0.1342)	- 0.6306 (0.3143)	- 1.0121 (0.3982)	- 0.7969 (0.1782)	- 0.3248 (0.1543)
educ	0.2040 (0.0884)	0.1981 (0.1781)	- 0.2162 (0.2966)	0.0284 (0.1054)	0.4002 (0.1113)
educ.f	- 0.0066 (0.1036)	- 0.2380 (0.2176)	- 1.3513 (0.5063)	- 0.0607 (0.1218)	- 0.0785 (0.1336)
educ.m	0.0779 (0.1029)	0.2283 (0.2053)	0.7380 (0.3216)	0.0630 (0.1202)	0.0315 (0.1307)
white	0.1453 (0.1537)	0.0428 (0.2925)	- 1.1238 (0.3457)	- 0.1067 (0.1849)	0.1887 (0.1922)
hisp	0.0392 (0.1583)	0.0365 (0.3151)	- 1.4018 (0.3870)	0.0568 (0.1874)	0.1279 (0.1916)
black	- 0.0781 (0.1415)	- 0.7010 (0.2964)	- 2.2587 (0.3675)	0.1026 (0.1684)	- 0.2660 (0.1732)
yr.work1	0.1096 (0.1046)	- 0.0492 (0.2138)	0.3852 (0.3688)	- 0.0491 (0.1215)	0.2495 (0.1470)
earn.yr	- 0.1118 (0.0979)	0.1459 (0.1564)	- 0.1324 (0.2212)	- 0.0437 (0.1190)	0.0849 (0.1138)
mosinjob	0.0719 (0.0192)	0.0692 (0.0342)	0.1027 (0.0500)	0.0235 (0.0232)	0.0429 (0.0232)
currjob	- 0.0399 (0.1108)	- 0.0714 (0.2037)	0.1179 (0.2865)	- 0.2400 (0.1340)	- 0.0248 (0.1502)
p.inc 3000-6000	0.1247 (0.1216)	- 0.2861 (0.2518)	- 8.7798 (5.4578)	- 0.1055 (0.1486)	0.1776 (0.1487)
p.inc 6000-9000	- 0.1024 (0.1796)	- 0.5319 (0.3614)	1.8550 (0.4034)	- 0.4735 (0.2455)	- 0.2168 (0.2243)
p.inc > 9000	0.2153 (0.2844)	0.0220 (0.4634)	1.1715 (0.5669)	0.2021 (0.3528)	0.3593 (0.3344)
h.inc 3000-6000	0.0480 (0.1048)	0.1032 (0.2406)	0.4373 (0.3723)	0.0166 (0.1190)	0.1061 (0.1366)
h.inc 6000-9000	- 0.0334 (0.1305)	0.2314 (0.2770)	- 1.0518 (0.5925)	- 0.0180 (0.1477)	0.1627 (0.1652)
h.inc 9000-18000	0.1209 (0.1089)	0.3517 (0.2311)	0.3180 (0.3499)	0.0362 (0.1250)	0.2359 (0.1400)
h.inc > 18000	0.1087 (0.1170)	0.0871 (0.2490)	0.2988 (0.3903)	- 0.2521 (0.1389)	0.1107 (0.1520)

**Model: a.208**  
**(week 208, without monotonicity of truncation)**

	$\hat{\beta}_{C,EE}$	$\hat{\beta}_{C,EN,1}$	$\hat{\beta}_{C,NE,0}$	$\hat{\beta}_{N,EE}$
intercept	1.9167 (0.0537)	2.8806 (0.6339)	2.0168 (0.0016)	1.8335 (0.1570)
age	0.0042 (0.0027)	-0.0459 (0.0334)	0.0024 (0.0001)	0.0082 (0.0078)
female	-0.0741 (0.0094)	-0.2331 (0.1063)	-0.1124 (0.0003)	-0.1201 (0.0289)
evarrst	-0.0156 (0.0108)	-0.0184 (0.1125)	0.0737 (0.0004)	-0.0155 (0.0312)
haschld	0.0012 (0.0246)	0.2071 (0.3923)	-0.1956 (0.0009)	-0.0150 (0.0649)
nchld	0.0050 (0.0163)	-0.0219 (0.2904)	0.1906 (0.0007)	0.0009 (0.0397)
partnered	-0.0166 (0.0209)	0.2831 (0.2641)	0.0003 (0.0005)	0.0044 (0.0471)
educ	0.0327 (0.0105)	0.0823 (0.1240)	-0.0317 (0.0003)	0.0307 (0.0312)
educ.f	-0.0128 (0.0125)	0.1818 (0.1430)	0.3408 (0.0005)	0.0765 (0.0386)
educ.m	-0.0020 (0.0124)	0.2147 (0.1278)	-0.2022 (0.0003)	0.0601 (0.0385)
white	-0.0504 (0.0199)	-0.0047 (0.1697)	-0.0910 (0.0003)	0.0220 (0.0576)
hispanic	-0.0194 (0.0211)	0.0029 (0.1962)	0.3447 (0.0004)	0.0800 (0.0604)
black	-0.0674 (0.0190)	0.0026 (0.1754)	-0.0152 (0.0011)	-0.0359 (0.0564)
yr.work1	0.0326 (0.0126)	0.0682 (0.1402)	-0.0029 (0.0005)	0.1083 (0.0387)
earn.yr	0.0267 (0.0105)	0.1357 (0.1372)	0.1936 (0.0002)	0.0477 (0.0304)
mosinjob	0.0005 (0.0020)	-0.0159 (0.0241)	-0.0026 (0.0001)	0.0006 (0.0061)
currjob	-0.0113 (0.0119)	0.1200 (0.1233)	0.0063 (0.0004)	-0.0649 (0.0345)
p.inc 3000-6000	0.0484 (0.0141)	0.0147 (0.1691)	-0.1112 (0.0005)	-0.0268 (0.0403)
p.inc 6000-9000	0.1034 (0.0244)	0.1030 (0.2176)	-0.1367 (0.0006)	-0.0288 (0.0642)
p.inc > 9000	0.0726 (0.0265)	0.0952 (0.2970)	-0.0752 (0.0007)	-0.0382 (0.0728)
h.inc 3000-6000	0.0026 (0.0127)	-0.1541 (0.1683)	-0.0683 (0.0005)	0.0282 (0.0421)
h.inc 6000-9000	0.0264 (0.0166)	-0.1189 (0.1623)	0.0986 (0.0005)	-0.0660 (0.0497)
h.inc 9000-18000	0.0230 (0.0127)	-0.1488 (0.1488)	0.1368 (0.0004)	-0.0194 (0.0401)
h.inc > 18000	0.0456 (0.0146)	0.1345 (0.1601)	0.3483 (0.0004)	-0.0094 (0.0449)
Treatment	0.0440 (0.0087)	0 (-)	0 (-)	0 (-)

	$\hat{\sigma}_{C,EE}$	$\hat{\sigma}_{C,EN,1}$	$\hat{\sigma}_{C,NE,0}$	$\hat{\sigma}_{N,EE}$
	0.2245 (0.0040)	0.7868 (0.0376)	0.0008 (0.0001)	0.5320 (0.0096)



	$\hat{\alpha}_{C,EE}$	$\hat{\alpha}_{C,EN}$	$\hat{\alpha}_{C,NE}$	$\hat{\alpha}_{C,NN}$	$\hat{\alpha}_{N,EE}$
intercept	1.2391 (0.4777)	0.5403 (0.9834)	- 1.0423 (1.3725)	1.9856 (0.5878)	- 0.0090 (0.6214)
age	- 0.0088 (0.0240)	- 0.0457 (0.0508)	- 0.0404 (0.0715)	- 0.0437 (0.0298)	- 0.0033 (0.0315)
female	- 0.4318 (0.0873)	- 0.1445 (0.1784)	0.1753 (0.2492)	- 0.1507 (0.1046)	- 0.2073 (0.1137)
evarrst	- 0.5439 (0.0894)	- 0.3530 (0.1836)	- 0.2895 (0.2731)	- 0.2778 (0.1072)	- 0.2780 (0.1192)
haschld	0.1410 (0.1812)	0.1117 (0.6597)	1.0653 (0.6429)	- 0.0672 (0.2275)	0.2058 (0.2206)
nchld	- 0.2610 (0.1051)	- 0.4385 (0.4956)	- 0.5636 (0.4747)	- 0.1766 (0.1316)	- 0.1205 (0.1237)
partnered	- 0.5970 (0.1480)	- 0.7882 (0.3940)	- 0.3523 (0.3727)	- 0.5927 (0.2092)	- 0.1210 (0.1774)
educ	0.2212 (0.0987)	0.3828 (0.1968)	0.3489 (0.2770)	- 0.0197 (0.1221)	0.5412 (0.1270)
educ.f	- 0.0795 (0.1148)	- 0.2377 (0.2270)	- 1.2580 (0.4284)	- 0.2025 (0.1400)	- 0.1173 (0.1520)
educ.m	0.1054 (0.1196)	0.5030 (0.2141)	1.0124 (0.3010)	0.2052 (0.1413)	0.0276 (0.1578)
white	0.1985 (0.1855)	- 0.0487 (0.3140)	- 1.0273 (0.3219)	- 0.2572 (0.2293)	0.2798 (0.2432)
hisp	0.0355 (0.1815)	- 0.5984 (0.3472)	- 1.4806 (0.3698)	- 0.0663 (0.2218)	0.1417 (0.2322)
black	- 0.1168 (0.1650)	- 1.0148 (0.3095)	- 6.3867 (4.5485)	0.0290 (0.2018)	- 0.2516 (0.2136)
yr.work1	0.0716 (0.1230)	- 0.0614 (0.2259)	0.3688 (0.4126)	- 0.3263 (0.1470)	- 0.0123 (0.1756)
earn.yr	- 0.0932 (0.1177)	0.0669 (0.2174)	0.0659 (0.1934)	- 0.1269 (0.1617)	0.0037 (0.1435)
mosinjob	0.0602 (0.0226)	0.0200 (0.0418)	0.1249 (0.0480)	0.0197 (0.0294)	0.0498 (0.0281)
currjob	0.1186 (0.1426)	0.3227 (0.2315)	- 0.0976 (0.2932)	- 0.1085 (0.1756)	0.1427 (0.1983)
p.inc 3000-6000	0.0077 (0.1339)	- 0.0926 (0.2701)	0.1291 (0.3646)	- 0.2182 (0.1724)	0.0119 (0.1703)
p.inc 6000-9000	0.1578 (0.2142)	0.1101 (0.4027)	0.9841 (0.4146)	- 0.2925 (0.2884)	0.0229 (0.2707)
p.inc > 9000	0.4147 (0.3401)	0.6326 (0.5238)	0.7984 (0.6249)	0.0691 (0.4328)	0.3693 (0.4119)
h.inc 3000-6000	0.1407 (0.1146)	0.0098 (0.2740)	- 0.3487 (0.4970)	0.1191 (0.1351)	0.1322 (0.1540)
h.inc 6000-9000	0.0773 (0.1444)	0.5389 (0.2795)	0.3539 (0.4231)	- 0.0627 (0.1738)	0.2262 (0.1911)
h.inc 9000-18000	0.2050 (0.1210)	0.3269 (0.2479)	0.1141 (0.3956)	0.2214 (0.1427)	0.3684 (0.1578)
h.inc > 18000	- 0.0324 (0.1287)	0.0206 (0.2711)	0.0543 (0.4133)	- 0.1004 (0.1574)	0.1466 (0.1693)

**Model: b.45**  
**(week 45, with monotonicity of truncation)**

	$\hat{\beta}_{C,EE}$	$\hat{\beta}_{C,EN,1}$	$\hat{\beta}_{N,EE}$
intercept	1.4856 (0.0466)	1.6949 (0.7170)	1.3884 (0.1380)
age	0.0125 (0.0023)	-0.0161 (0.0360)	0.0148 (0.0068)
female	-0.0368 (0.0079)	-0.3167 (0.1333)	-0.1726 (0.0258)
evarrst	0.0012 (0.0089)	0.0464 (0.1268)	0.0066 (0.0279)
haschld	0.0465 (0.0208)	0.1491 (0.3376)	0.0641 (0.0617)
nchld	-0.0201 (0.0139)	-0.0361 (0.2043)	0.0073 (0.0391)
partnered	-0.0236 (0.0152)	0.3358 (0.2794)	0.0481 (0.0435)
educ	0.0029 (0.0089)	0.0675 (0.1359)	0.0387 (0.0275)
educ.f	0.0173 (0.0103)	0.2008 (0.1591)	-0.0054 (0.0323)
educ.m	0.0271 (0.0103)	-0.1553 (0.1525)	0.0245 (0.0317)
white	-0.0261 (0.0172)	0.0716 (0.2420)	-0.0450 (0.0479)
hispanic	-0.0240 (0.0184)	0.0829 (0.2681)	-0.0048 (0.0512)
black	-0.0384 (0.0168)	0.0706 (0.2385)	-0.0425 (0.0470)
yr.work1	0.0062 (0.0111)	0.1888 (0.1686)	0.0439 (0.0363)
earn.yr	0.0626 (0.0108)	0.1529 (0.1015)	0.0379 (0.0239)
mosinjob	-0.0055 (0.0019)	-0.0529 (0.0252)	0.0010 (0.0050)
currjob	0.0000 (0.0095)	0.1278 (0.1502)	-0.0144 (0.0286)
p.inc 3000-6000	0.0414 (0.0123)	0.3781 (0.1859)	-0.0216 (0.0343)
p.inc 6000-9000	0.0631 (0.0179)	0.4716 (0.2532)	0.0490 (0.0511)
p.inc > 9000	0.0303 (0.0252)	0.3329 (0.2740)	-0.0075 (0.0566)
h.inc 3000-6000	0.0016 (0.0118)	0.2662 (0.1789)	0.0672 (0.0395)
h.inc 6000-9000	0.0029 (0.0136)	0.3103 (0.2666)	0.0623 (0.0460)
h.inc 9000-18000	0.0140 (0.0110)	0.3134 (0.1951)	0.1170 (0.0372)
h.inc > 18000	0.0184 (0.0120)	0.3013 (0.1873)	0.1354 (0.0394)
Treatment	0.0299 (0.0075)	0 (-)	0 (-)

$\hat{\sigma}_{C,EE}$	$\hat{\sigma}_{C,EN,1}$	$\hat{\sigma}_{N,EE}$
0.1764 (0.0033)	0.7736 (0.0417)	0.4745 (0.0087)

	$\hat{\alpha}_{C,EE}$	$\hat{\alpha}_{C,EN}$	$\hat{\alpha}_{C,NN}$	$\hat{\alpha}_{N,EE}$
intercept	- 0.8988 (0.4128)	- 2.2234 (1.0198)	1.8857 (0.4209)	- 2.3079 (0.5194)
age	0.0473 (0.0207)	0.0106 (0.0500)	- 0.0212 (0.0214)	0.0808 (0.0260)
female	- 0.3346 (0.0745)	- 0.5327 (0.1895)	- 0.3286 (0.0756)	- 0.3206 (0.0966)
evarrst	- 0.4884 (0.0780)	- 0.2499 (0.1926)	- 0.4711 (0.0791)	- 0.3520 (0.1023)
haschld	0.0690 (0.1680)	- 0.3785 (0.5200)	- 0.2412 (0.1642)	- 0.0263 (0.2027)
nchld	- 0.3220 (0.1054)	- 0.1154 (0.3208)	- 0.1349 (0.0983)	- 0.2271 (0.1231)
partnered	- 0.2704 (0.1327)	- 0.4083 (0.3802)	- 0.5391 (0.1467)	- 0.1164 (0.1586)
educ	0.0110 (0.0839)	0.1998 (0.2011)	- 0.0556 (0.0864)	0.3511 (0.1065)
educ.f	0.0634 (0.1020)	- 0.0172 (0.2356)	- 0.0634 (0.1055)	0.0039 (0.1305)
educ.m	0.1642 (0.1028)	0.3743 (0.2273)	0.2317 (0.1046)	0.2059 (0.1291)
white	0.3471 (0.1503)	0.2760 (0.3654)	- 0.2398 (0.1478)	0.1405 (0.1873)
hisp	0.0232 (0.1550)	0.0335 (0.3960)	- 0.1015 (0.1480)	0.0071 (0.1850)
black	0.0620 (0.1401)	0.0413 (0.3561)	- 0.1128 (0.1332)	- 0.1751 (0.1673)
yr.work1	0.3858 (0.0995)	0.3592 (0.2431)	0.0175 (0.1024)	0.3348 (0.1463)
earn.yr	0.0794 (0.1026)	0.2658 (0.2436)	- 0.0587 (0.1144)	0.0882 (0.1186)
mosinjob	0.0514 (0.0186)	- 0.0050 (0.0434)	0.0212 (0.0201)	0.0609 (0.0224)
currjob	0.2356 (0.1037)	0.2023 (0.2146)	- 0.1368 (0.1123)	0.1940 (0.1413)
p.inc 3000-6000	- 0.0444 (0.1148)	- 0.0477 (0.2860)	- 0.1550 (0.1223)	0.1672 (0.1391)
p.inc 6000-9000	0.0624 (0.1856)	0.3373 (0.3982)	0.0200 (0.1955)	0.1274 (0.2179)
p.inc > 9000	0.2049 (0.2747)	0.8999 (0.4693)	0.3569 (0.2888)	0.6187 (0.3015)
h.inc 3000-6000	0.0697 (0.1001)	0.5132 (0.2625)	- 0.0129 (0.0989)	0.0744 (0.1351)
h.inc 6000-9000	0.1502 (0.1256)	- 0.3130 (0.3986)	0.0479 (0.1265)	0.1517 (0.1674)
h.inc 9000-18000	0.2929 (0.1033)	0.0944 (0.2914)	0.1291 (0.1044)	0.3342 (0.1363)
h.inc > 18000	0.3854 (0.1169)	0.6204 (0.2784)	0.1078 (0.1190)	0.5066 (0.1520)

**Model: b.135**  
(week 135, with monotonicity of truncation)

	$\hat{\beta}_{C,EE}$	$\hat{\beta}_{C,EN,1}$	$\hat{\beta}_{N,EE}$
intercept	1.7905 (0.0502)	1.9894 (0.4887)	1.7698 (0.1376)
age	0.0053 (0.0025)	0.0103 (0.0247)	0.0037 (0.0069)
female	-0.0678 (0.0088)	-0.1454 (0.0960)	-0.1325 (0.0252)
evarrst	-0.0166 (0.0097)	-0.0427 (0.0994)	0.0778 (0.0275)
haschld	0.0100 (0.0232)	-0.0752 (0.2996)	0.0427 (0.0567)
nchld	-0.0044 (0.0146)	0.1360 (0.2198)	0.0161 (0.0340)
partnered	0.0231 (0.0208)	-0.3894 (0.1834)	0.0325 (0.0429)
educ	0.0231 (0.0099)	0.0917 (0.1016)	0.1298 (0.0275)
educ.f	0.0186 (0.0117)	0.0190 (0.1280)	0.0548 (0.0336)
educ.m	0.0072 (0.0114)	0.1152 (0.1186)	0.0756 (0.0329)
white	-0.0053 (0.0177)	-0.0877 (0.1551)	-0.0424 (0.0484)
hisp	0.0219 (0.0191)	-0.0830 (0.1670)	0.0244 (0.0510)
black	-0.0155 (0.0169)	-0.2232 (0.1575)	-0.1128 (0.0474)
yr.work1	0.0239 (0.0118)	0.1170 (0.1254)	0.1004 (0.0344)
earn.yr	0.0362 (0.0124)	0.0229 (0.0675)	0.0483 (0.0243)
mosinjob	-0.0028 (0.0021)	-0.0121 (0.0174)	-0.0027 (0.0051)
currjob	0.0070 (0.0113)	0.1906 (0.1097)	-0.0147 (0.0302)
p.inc 3000-6000	0.0416 (0.0132)	0.2032 (0.1442)	0.0371 (0.0350)
p.inc 6000-9000	0.0777 (0.0207)	0.3345 (0.2027)	0.0158 (0.0560)
p.inc > 9000	0.0673 (0.0263)	0.2735 (0.2414)	0.0737 (0.0623)
h.inc 3000-6000	0.0096 (0.0123)	-0.1616 (0.1342)	-0.0065 (0.0373)
h.inc 6000-9000	0.0056 (0.0158)	-0.1260 (0.1535)	-0.0090 (0.0436)
h.inc 9000-18000	0.0024 (0.0122)	-0.1196 (0.1251)	0.0307 (0.0355)
h.inc > 18000	0.0252 (0.0133)	-0.1281 (0.1388)	0.0525 (0.0394)
Treatment	0.0297 (0.0082)	0 (-)	0 (-)

$\hat{\sigma}_{C,EE}$	$\hat{\sigma}_{C,EN,1}$	$\hat{\sigma}_{N,EE}$
0.2101 (0.0039)	0.7210 (0.0335)	0.4850 (0.0084)

	$\hat{\alpha}_{C,EE}$	$\hat{\alpha}_{C,EN}$	$\hat{\alpha}_{C,NN}$	$\hat{\alpha}_{N,EE}$
intercept	0.3484 (0.4273)	- 1.3895 (0.8661)	2.2787 (0.4960)	- 0.6812 (0.5394)
age	0.0283 (0.0217)	0.0489 (0.0437)	- 0.0583 (0.0256)	0.0318 (0.0276)
female	- 0.3845 (0.0776)	- 0.5979 (0.1703)	- 0.1431 (0.0884)	- 0.2838 (0.0991)
evarrst	- 0.4372 (0.0799)	- 0.5437 (0.1760)	- 0.3566 (0.0917)	- 0.3703 (0.1048)
haschld	- 0.0503 (0.1644)	0.0509 (0.5032)	- 0.0634 (0.1935)	0.0621 (0.1972)
nchld	- 0.2152 (0.0973)	- 0.3608 (0.3586)	- 0.1922 (0.1179)	- 0.1352 (0.1136)
partnered	- 0.8049 (0.1340)	- 0.5634 (0.3145)	- 0.8037 (0.1689)	- 0.3267 (0.1540)
educ	0.1878 (0.0883)	0.1984 (0.1804)	0.0058 (0.1033)	0.3909 (0.1114)
educ.f	- 0.0413 (0.1033)	- 0.2194 (0.2197)	- 0.1205 (0.1200)	- 0.0970 (0.1334)
educ.m	0.0971 (0.1027)	0.2277 (0.2081)	0.0947 (0.1171)	0.0368 (0.1307)
white	0.0772 (0.1533)	0.0799 (0.2998)	- 0.2196 (0.1743)	0.1509 (0.1925)
hisp	- 0.0404 (0.1577)	0.0614 (0.3229)	- 0.0876 (0.1770)	0.0802 (0.1918)
black	- 0.1756 (0.1408)	- 0.6611 (0.3030)	- 0.0733 (0.1577)	- 0.3230 (0.1734)
yr.work1	0.1174 (0.1044)	- 0.0622 (0.2170)	- 0.0474 (0.1197)	0.2555 (0.1472)
earn.yr	- 0.1321 (0.0979)	0.1588 (0.1535)	- 0.0583 (0.1150)	0.0856 (0.1130)
mosinjob	0.0752 (0.0193)	0.0681 (0.0344)	0.0298 (0.0225)	0.0443 (0.0232)
currjob	- 0.0282 (0.1114)	- 0.0818 (0.2070)	- 0.2068 (0.1314)	- 0.0213 (0.1509)
p.inc 3000-6000	0.0786 (0.1210)	- 0.2809 (0.2540)	- 0.1930 (0.1472)	0.1347 (0.1482)
p.inc 6000-9000	0.0227 (0.1808)	- 0.6016 (0.3706)	- 0.1625 (0.2168)	- 0.1229 (0.2243)
p.inc > 9000	0.2773 (0.2861)	- 0.0200 (0.4697)	0.2677 (0.3424)	0.3709 (0.3361)
h.inc 3000-6000	0.0546 (0.1044)	0.0981 (0.2433)	0.0319 (0.1169)	0.1121 (0.1366)
h.inc 6000-9000	- 0.0588 (0.1299)	0.2218 (0.2802)	- 0.0610 (0.1459)	0.1413 (0.1649)
h.inc 9000-18000	0.1311 (0.1087)	0.3333 (0.2341)	0.0563 (0.1230)	0.2492 (0.1400)
h.inc > 18000	0.1187 (0.1171)	0.0796 (0.2521)	- 0.2106 (0.1356)	0.1340 (0.1521)

**Model: b.208**  
(week 208, with monotonicity of truncation)

	$\hat{\beta}_{C,EE}$	$\hat{\beta}_{C,EN,I}$	$\hat{\beta}_{N,EE}$
intercept	1.9225 (0.0531)	3.0103 (0.6993)	1.8635 (0.1558)
age	0.0040 (0.0027)	-0.0531 (0.0370)	0.0071 (0.0078)
female	-0.0735 (0.0093)	-0.2559 (0.1154)	-0.1192 (0.0288)
evarrst	-0.0117 (0.0107)	-0.0265 (0.1217)	-0.0133 (0.0310)
haschld	-0.0007 (0.0246)	0.2712 (0.4425)	-0.0026 (0.0645)
nchld	0.0053 (0.0162)	-0.0552 (0.3315)	-0.0027 (0.0397)
partnered	-0.0148 (0.0209)	0.3137 (0.2846)	0.0081 (0.0467)
educ	0.0334 (0.0105)	0.0868 (0.1348)	0.0345 (0.0310)
educ.f	-0.0095 (0.0125)	0.1901 (0.1543)	0.0746 (0.0385)
educ.m	-0.0029 (0.0123)	0.2473 (0.1392)	0.0564 (0.0383)
white	-0.0503 (0.0189)	0.0124 (0.1892)	0.0101 (0.0570)
hispanic	-0.0154 (0.0203)	0.0101 (0.2180)	0.0770 (0.0597)
black	-0.0685 (0.0180)	0.0244 (0.1947)	-0.0466 (0.0559)
yr.work1	0.0328 (0.0125)	0.0608 (0.1506)	0.1052 (0.0387)
earn.yr	0.0280 (0.0103)	0.1649 (0.1512)	0.0608 (0.0300)
mosinjob	0.0006 (0.0020)	-0.0196 (0.0261)	-0.0001 (0.0061)
currjob	-0.0120 (0.0118)	0.1350 (0.1334)	-0.0643 (0.0344)
p.inc 3000-6000	0.0476 (0.0140)	0.0042 (0.1849)	-0.0312 (0.0401)
p.inc 6000-9000	0.1016 (0.0236)	0.0989 (0.2455)	-0.0308 (0.0635)
p.inc > 9000	0.0735 (0.0264)	0.0960 (0.3270)	-0.0364 (0.0719)
h.inc 3000-6000	0.0008 (0.0128)	-0.1629 (0.1814)	0.0313 (0.0420)
h.inc 6000-9000	0.0287 (0.0165)	-0.1303 (0.1748)	-0.0658 (0.0496)
h.inc 9000-18000	0.0227 (0.0127)	-0.1577 (0.1599)	-0.0172 (0.0400)
h.inc > 18000	0.0486 (0.0145)	0.1340 (0.1726)	-0.0011 (0.0446)
Treatment	0.0418 (0.0087)	0 (-)	0 (-)

$\hat{\sigma}_{C,EE}$	$\hat{\sigma}_{C,EN,I}$	$\hat{\sigma}_{N,EE}$
0.2272 (0.0040)	0.8095 (0.0401)	0.5311 (0.0095)

	$\hat{\alpha}_{C,EE}$	$\hat{\alpha}_{C,EN}$	$\hat{\alpha}_{C,NN}$	$\hat{\alpha}_{N,EE}$
intercept	1.3542 (0.4793)	0.4194 (1.0224)	2.2326 (0.5715)	0.0837 (0.6226)
age	-0.0085 (0.0242)	-0.0461 (0.0529)	-0.0445 (0.0291)	-0.0052 (0.0316)
female	-0.4160 (0.0874)	-0.2168 (0.1846)	-0.1252 (0.1027)	-0.2023 (0.1136)
evarrst	-0.5418 (0.0896)	-0.3629 (0.1895)	-0.2794 (0.1054)	-0.2695 (0.1193)
haschld	0.1450 (0.1813)	0.0662 (0.6850)	-0.0190 (0.2224)	0.2398 (0.2203)
nchld	-0.2614 (0.1050)	-0.4169 (0.5125)	-0.1939 (0.1295)	-0.1320 (0.1240)
partnered	-0.5915 (0.1490)	-0.7492 (0.4089)	-0.5597 (0.1949)	-0.1003 (0.1778)
educ	0.2179 (0.0990)	0.3858 (0.2036)	-0.0060 (0.1193)	0.5457 (0.1274)
educ.f	-0.1045 (0.1147)	-0.2134 (0.2358)	-0.2530 (0.1376)	-0.1302 (0.1519)
educ.m	0.1299 (0.1197)	0.4602 (0.2225)	0.2554 (0.1380)	0.0359 (0.1577)
white	0.1219 (0.1868)	0.0626 (0.3274)	-0.3716 (0.2163)	0.2403 (0.2440)
hisp	-0.0663 (0.1822)	-0.5223 (0.3637)	-0.2535 (0.2093)	0.1090 (0.2321)
black	-0.2428 (0.1657)	-0.8774 (0.3232)	-0.2178 (0.1893)	-0.3124 (0.2135)
yr.work1	0.0707 (0.1233)	-0.0634 (0.2322)	-0.3418 (0.1452)	-0.0243 (0.1760)
earn.yr	-0.0820 (0.1174)	0.0446 (0.2200)	-0.0495 (0.1471)	0.0317 (0.1401)
mosinjob	0.0622 (0.0225)	0.0168 (0.0426)	0.0231 (0.0278)	0.0508 (0.0278)
currjob	0.1148 (0.1441)	0.3369 (0.2375)	-0.0886 (0.1730)	0.1374 (0.1999)
p.inc 3000-6000	0.0102 (0.1342)	-0.0966 (0.2807)	-0.2194 (0.1673)	0.0043 (0.1707)
p.inc 6000-9000	0.2080 (0.2163)	0.0230 (0.4292)	-0.1215 (0.2646)	0.0503 (0.2712)
p.inc > 9000	0.4249 (0.3382)	0.5623 (0.5403)	0.1551 (0.4070)	0.3852 (0.4058)
h.inc 3000-6000	0.1283 (0.1148)	0.0135 (0.2821)	0.1012 (0.1339)	0.1324 (0.1543)
h.inc 6000-9000	0.0802 (0.1447)	0.5318 (0.2881)	-0.0471 (0.1708)	0.2179 (0.1919)
h.inc 9000-18000	0.1984 (0.1212)	0.3322 (0.2542)	0.2045 (0.1408)	0.3656 (0.1580)
h.inc > 18000	-0.0419 (0.1290)	-0.0059 (0.2802)	-0.0945 (0.1534)	0.1476 (0.1695)

## APPENDIX B

### Baseline characteristics

For each pre-treatment covariate, we provide the (fitted) average in each stratum, for all estimated models; the aim is to illustrate how the baseline characteristics of each individual may affect the compliance behavior and the couple of potential employment status. Tables in the next pages describe the mean covariates values in the  $k$  groups; additionally, a summary of the baseline covariates for compliers and never-takers is returned. The imputed matrix was used; all computations involve design weights. We refer to Appendix A for the description of covariates.

In the NN groups, we generally observe a lower education degree and a prevalence of non-white race, together with a poorer occupational background (as summarized by `yr.work1`, `earn.yr`, `mosinjob`, `currjob`) and a lower personal income. It is difficult to describe the other groups, whose characteristics are generally less marked. For the C.NE group, we can observe a higher education in week 45; for the remaining weeks, due to the small consistency of this group, we prefer to avoid any remarks.

Great differences are observed between compliers and never-takers: with respect to compliers, never-takers are more often females, with a partner, and with one or more children; we can also note that never-takers have a higher mean education. This suggests that the choice of participating to the training program depends on familiar conditions, rather than on personal/household income and job experiences.

Although many remarks could be done, we prefer to not discuss here the issue of *how* each individual chooses to be or not to be a complier and – given the treatment assignment and compliance status – to have a job or to be unemployed. To see how the pre-treatment covariates affect the expected wages, we refer again to Appendix A.



**Model: a.45**  
**(week 45, without monotonicity of truncation)**

	C.EE	C.EN	C.NE	C.NN	Com- pliers	N.EE	N.NN	Never- takers
age	18.8877	19.2098	19.4367	18.1577	18.7098	19.4720	18.8571	19.1160
female	0.4049	0.3109	0.3400	0.4115	0.3843	0.3928	0.4807	0.4437
evarrst	0.2720	0.2764	0.2300	0.2534	0.2550	0.2632	0.3244	0.2986
haschld	0.1686	0.1378	0.1404	0.1539	0.1527	0.1891	0.2512	0.2250
nchld	0.2223	0.1894	0.1904	0.2160	0.2092	0.2620	0.3779	0.3291
partnered	0.0788	0.0507	0.0409	0.0406	0.0505	0.0948	0.0870	0.0903
educ	0.4096	0.4620	0.4825	0.2846	0.3743	0.5306	0.3689	0.4370
educ.f	0.1717	0.2215	0.1995	0.1326	0.1651	0.1934	0.1504	0.1685
educ.m	0.1685	0.2292	0.2122	0.1594	0.1800	0.1978	0.1442	0.1668
white	0.3924	0.3859	0.3421	0.2093	0.2982	0.3810	0.2669	0.3149
hisp	0.1418	0.1468	0.1730	0.1824	0.1674	0.1653	0.1767	0.1719
black	0.4067	0.4018	0.4029	0.5294	0.4608	0.3871	0.4833	0.4428
yr.work1	0.7483	0.7421	0.7686	0.4890	0.6350	0.7698	0.5645	0.6510
earn.yr	0.1514	0.2431	0.2474	-0.2487	0.0009	0.2934	-0.0914	0.0706
mosinjob	4.9203	4.8924	5.3397	2.3175	3.8332	5.4892	3.1649	4.1436
currjob	0.3189	0.3048	0.3213	0.1239	0.2299	0.3343	0.1987	0.2558
p.inc 3000-6000	0.1467	0.1563	0.1486	0.0756	0.1160	0.1862	0.1226	0.1494
p.inc 6000-9000	0.0638	0.0600	0.0564	0.0251	0.0443	0.0728	0.0352	0.0510
p.inc > 9000	0.0344	0.0567	0.0488	0.0140	0.0305	0.0640	0.0190	0.0379
h.inc 3000-6000	0.1813	0.2067	0.1769	0.2152	0.1982	0.1798	0.2276	0.2075
h.inc 6000-9000	0.1065	0.0929	0.1111	0.1126	0.1089	0.1046	0.1072	0.1061
h.inc 9000-18000	0.2942	0.2205	0.2108	0.2503	0.2488	0.2763	0.2100	0.2379
h.inc > 18000	0.2000	0.2730	0.2699	0.1215	0.1871	0.2460	0.1472	0.1888

**Model: b.45**  
**(week 45, with monotonicity of truncation)**

	C.EE	C.EN	C.NN	Com- pliers	N.EE	N.NN	Never- takers
age	19.0415	18.9657	18.5036	18.7085	19.5050	18.7830	19.1201
female	0.3804	0.3192	0.3932	0.3858	0.3878	0.4858	0.4400
evarrst	0.2532	0.3078	0.2516	0.2545	0.2668	0.3289	0.2999
haschld	0.1605	0.1353	0.1514	0.1539	0.1849	0.2550	0.2223
nchld	0.2129	0.1903	0.2117	0.2112	0.2554	0.3845	0.3242
partnered	0.0666	0.0568	0.0420	0.0511	0.0897	0.0879	0.0888
educ	0.4279	0.4423	0.3351	0.3716	0.5400	0.3596	0.4438
educ.f	0.1859	0.1955	0.1513	0.1651	0.1928	0.1474	0.1686
educ.m	0.1864	0.2194	0.1724	0.1792	0.1987	0.1426	0.1688
white	0.3791	0.3749	0.2485	0.2988	0.3727	0.2615	0.3134
hisp	0.1488	0.1471	0.1783	0.1668	0.1709	0.1755	0.1733
black	0.4114	0.4146	0.4931	0.4617	0.3824	0.4914	0.4405
yr.work1	0.7521	0.7187	0.5609	0.6334	0.7762	0.5486	0.6549
earn.yr	0.1860	0.1980	-0.1180	0.0000	0.2959	-0.1224	0.0729
mosinjob	5.0385	4.5638	3.0868	3.8218	5.5361	2.9788	4.1727
currjob	0.3191	0.2927	0.1767	0.2307	0.3314	0.1862	0.2540
p.inc 3000-6000	0.1494	0.1379	0.0934	0.1146	0.1896	0.1207	0.1529
p.inc 6000-9000	0.0607	0.0626	0.0331	0.0439	0.0726	0.0343	0.0522
p.inc > 9000	0.0376	0.0655	0.0230	0.0298	0.0656	0.0173	0.0398
h.inc 3000-6000	0.1826	0.2509	0.2062	0.1999	0.1720	0.2307	0.2033
h.inc 6000-9000	0.1102	0.0636	0.1101	0.1082	0.1065	0.1089	0.1078
h.inc 9000-18000	0.2664	0.2098	0.2392	0.2474	0.2759	0.2114	0.2415
h.inc > 18000	0.2219	0.2832	0.1590	0.1859	0.2506	0.1406	0.1919

**Model: a.135**  
**(week 135, without monotonicity of truncation)**

	C.EE	C.EN	C.NE	C.NN	Com- pliers	N.EE	N.NN	Never- takers
age	18.9747	19.0294	18.8950	18.3669	18.7261	19.2645	18.8120	19.0574
female	0.3607	0.3262	0.7127	0.4169	0.3878	0.3972	0.4741	0.4324
evarrst	0.2515	0.2502	0.3749	0.2584	0.2567	0.2604	0.3305	0.2925
haschld	0.1475	0.1387	0.4223	0.1492	0.1528	0.1969	0.2498	0.2211
nchld	0.2050	0.1759	0.4462	0.2084	0.2086	0.2802	0.3767	0.3244
partnered	0.0489	0.0630	0.0846	0.0382	0.0464	0.0911	0.1052	0.0976
educ	0.4264	0.4243	0.3577	0.3135	0.3781	0.4887	0.3496	0.4250
educ.f	0.1822	0.1588	0.0732	0.1506	0.1649	0.1749	0.1617	0.1689
educ.m	0.1862	0.1930	0.2235	0.1637	0.1782	0.1785	0.1633	0.1715
white	0.3358	0.3946	0.3433	0.2270	0.2961	0.3625	0.2671	0.3188
hisp	0.1589	0.2095	0.1707	0.1647	0.1660	0.1838	0.1641	0.1748
black	0.4365	0.3060	0.2571	0.5418	0.4649	0.3814	0.4965	0.4341
yr.work1	0.6900	0.6945	0.7632	0.5467	0.6325	0.7276	0.5711	0.6560
earn.yr	0.0833	0.1863	0.1854	-0.1358	0.0037	0.1859	-0.0879	0.0606
mosinjob	4.4574	4.7717	5.2723	2.9448	3.8746	4.7865	3.1376	4.0317
currjob	0.2705	0.2816	0.3310	0.1725	0.2321	0.2831	0.2095	0.2494
p.inc 3000-6000	0.1462	0.1184	0.0001	0.0885	0.1169	0.1698	0.1167	0.1455
p.inc 6000-9000	0.0532	0.0503	0.2240	0.0232	0.0439	0.0571	0.0456	0.0518
p.inc > 9000	0.0350	0.0428	0.0655	0.0207	0.0304	0.0511	0.0225	0.0380
h.inc 3000-6000	0.1920	0.1722	0.1857	0.2190	0.2013	0.1884	0.2135	0.1999
h.inc 6000-9000	0.1021	0.1183	0.0517	0.1096	0.1057	0.1157	0.1111	0.1136
h.inc 9000-18000	0.2504	0.2902	0.2922	0.2316	0.2470	0.2676	0.2132	0.2427
h.inc > 18000	0.2193	0.2033	0.2442	0.1389	0.1850	0.2116	0.1721	0.1935

**Model: b.135**  
(week 135, with monotonicity of truncation)

	C.EE	C.EN	C.NN	Com- pliers	N.EE	N.NN	Never- takers
age	18.9723	19.0424	18.3827	18.7249	19.2654	18.8118	19.0603
female	0.3689	0.3125	0.4262	0.3887	0.3972	0.4703	0.4303
evarrst	0.2537	0.2451	0.2624	0.2567	0.2627	0.3284	0.2924
haschld	0.1526	0.1304	0.1579	0.1530	0.1996	0.2463	0.2207
nchld	0.2094	0.1694	0.2163	0.2090	0.2822	0.3734	0.3234
partnered	0.0489	0.0652	0.0398	0.0464	0.0919	0.1048	0.0977
educ	0.4251	0.4270	0.3139	0.3775	0.4886	0.3510	0.4264
educ.f	0.1801	0.1626	0.1476	0.1647	0.1745	0.1632	0.1694
educ.m	0.1869	0.1934	0.1655	0.1783	0.1785	0.1630	0.1715
white	0.3354	0.3959	0.2322	0.2962	0.3630	0.2651	0.3187
hisp	0.1595	0.2084	0.1650	0.1660	0.1837	0.1642	0.1749
black	0.4339	0.3113	0.5296	0.4646	0.3803	0.5008	0.4348
yr.work1	0.6911	0.6907	0.5537	0.6320	0.7293	0.5695	0.6570
earn.yr	0.0821	0.1870	-0.1247	0.0021	0.1912	-0.0896	0.0642
mosinjob	4.4651	4.7402	3.0254	3.8697	4.8046	3.1209	4.0432
currjob	0.2715	0.2773	0.1790	0.2323	0.2833	0.2076	0.2490
p.inc 3000-6000	0.1435	0.1234	0.0849	0.1166	0.1682	0.1195	0.1462
p.inc 6000-9000	0.0558	0.0450	0.0299	0.0438	0.0594	0.0431	0.0520
p.inc > 9000	0.0358	0.0412	0.0220	0.0304	0.0512	0.0220	0.0380
h.inc 3000-6000	0.1917	0.1742	0.2177	0.2014	0.1876	0.2145	0.1997
h.inc 6000-9000	0.1016	0.1195	0.1074	0.1056	0.1149	0.1124	0.1138
h.inc 9000-18000	0.2515	0.2862	0.2340	0.2469	0.2684	0.2121	0.2429
h.inc > 18000	0.2188	0.2020	0.1431	0.1848	0.2134	0.1705	0.1940

**Model: a.208**  
**(week 208, without monotonicity of truncation)**

	C.EE	C.EN	C.NE	C.NN	Com- pliers	N.EE	N.NN	Never- takers
age	18.9152	18.7614	19.4255	18.3956	18.7248	19.2237	18.8570	19.0731
female	0.3604	0.3926	0.4592	0.4214	0.3870	0.4163	0.4643	0.4360
evarrst	0.2376	0.2743	0.2889	0.2752	0.2553	0.2752	0.3285	0.2971
haschld	0.1559	0.1113	0.2046	0.1585	0.1539	0.2117	0.2347	0.2211
nchld	0.2111	0.1419	0.2496	0.2229	0.2100	0.3018	0.3590	0.3253
partnered	0.0518	0.0438	0.1198	0.0404	0.0483	0.0968	0.0923	0.0949
educ	0.4114	0.4264	0.4906	0.3001	0.3742	0.4975	0.3481	0.4362
educ.f	0.1760	0.1842	0.1131	0.1436	0.1639	0.1745	0.1671	0.1715
educ.m	0.1765	0.2334	0.2810	0.1695	0.1809	0.1698	0.1574	0.1647
white	0.3303	0.4398	0.5149	0.2020	0.2973	0.3588	0.2567	0.3169
hisp	0.1666	0.1469	0.1886	0.1648	0.1647	0.1865	0.1670	0.1785
black	0.4388	0.3030	0.0046	0.5625	0.4631	0.3882	0.5079	0.4373
yr.work1	0.7010	0.6755	0.8564	0.5025	0.6304	0.7038	0.6017	0.6619
earn.yr	0.0925	0.1009	0.5166	-0.1787	0.0039	0.1501	-0.0638	0.0623
mosinjob	4.4820	4.3146	6.6762	2.6514	3.8514	4.6557	3.2886	4.0943
currjob	0.2701	0.3058	0.3353	0.1562	0.2335	0.2814	0.1965	0.2465
p.inc 3000-6000	0.1380	0.1233	0.1640	0.0846	0.1180	0.1566	0.1258	0.1440
p.inc 6000-9000	0.0536	0.0560	0.1770	0.0217	0.0447	0.0579	0.0386	0.0500
p.inc > 9000	0.0372	0.0450	0.0750	0.0168	0.0312	0.0463	0.0216	0.0362
h.inc 3000-6000	0.2011	0.1509	0.0944	0.2250	0.2033	0.1831	0.2121	0.1950
h.inc 6000-9000	0.1072	0.1515	0.1635	0.0940	0.1074	0.1123	0.1060	0.1097
h.inc 9000-18000	0.2523	0.2697	0.2748	0.2335	0.2475	0.2696	0.2008	0.2414
h.inc > 18000	0.1970	0.2205	0.2916	0.1466	0.1827	0.2126	0.1806	0.1994

**Model: b.208**  
**(week 208, with monotonicity of truncation)**

	C.EE	C.EN	C.NN	Com- pliers	N.EE	N.NN	Never- takers
age	18.9223	18.7160	18.4401	18.7245	19.2309	18.8453	19.0741
female	0.3627	0.3776	0.4244	0.3871	0.4168	0.4636	0.4358
evarrst	0.2380	0.2749	0.2755	0.2550	0.2769	0.3282	0.2978
haschld	0.1555	0.1081	0.1600	0.1535	0.2142	0.2336	0.2221
nchld	0.2106	0.1383	0.2236	0.2098	0.3039	0.3579	0.3259
partnered	0.0522	0.0431	0.0433	0.0481	0.0987	0.0905	0.0954
educ	0.4121	0.4202	0.3085	0.3737	0.4994	0.3470	0.4374
educ.f	0.1751	0.1834	0.1433	0.1638	0.1743	0.1679	0.1717
educ.m	0.1785	0.2267	0.1745	0.1808	0.1707	0.1567	0.1650
white	0.3338	0.4398	0.2164	0.2979	0.3581	0.2529	0.3153
hisp	0.1663	0.1444	0.1650	0.1641	0.1890	0.1668	0.1800
black	0.4316	0.3188	0.5381	0.4628	0.3855	0.5148	0.4380
yr.work1	0.7032	0.6660	0.5173	0.6303	0.7055	0.5989	0.6622
earn.yr	0.0963	0.0666	-0.1468	0.0024	0.1607	-0.0726	0.0659
mosinjob	4.5085	4.1530	2.8275	3.8475	4.6927	3.2445	4.1041
currjob	0.2707	0.3004	0.1662	0.2337	0.2819	0.1942	0.2462
p.inc 3000-6000	0.1386	0.1220	0.0873	0.1180	0.1567	0.1255	0.1440
p.inc 6000-9000	0.0559	0.0494	0.0270	0.0445	0.0596	0.0374	0.0506
p.inc > 9000	0.0374	0.0404	0.0195	0.0309	0.0478	0.0211	0.0370
h.inc 3000-6000	0.1991	0.1541	0.2195	0.2032	0.1827	0.2134	0.1952
h.inc 6000-9000	0.1087	0.1510	0.0965	0.1075	0.1122	0.1059	0.1096
h.inc 9000-18000	0.2529	0.2711	0.2347	0.2475	0.2696	0.2004	0.2414
h.inc > 18000	0.1980	0.2135	0.1535	0.1825	0.2141	0.1798	0.2002

## References

- Angrist, J.D., Imbens, G.W., Rubin, D.B. (1996). *Identification of causal effects using instrumental variables*. Journal of the American Statistical Association, 91, 444-455.
- Ashenfelter, O. (1978). *Estimating the effect of training programs on earnings*. Review of Economics and Statistics, 60, 47-57.
- Ashenfelter, O., Card, D. (1985). *Using the longitudinal structure of earnings to estimate the effect on training programs*. Review of Economics and Statistics, 67, 648-660.
- Balke, E., Pearl, J. (1997). *Bounds on treatment effects from studies with imperfect compliance*. Journal of the American Statistical Association, 92, 1171-1176.
- Barnard, J., Du, J., Hill, J.L., Rubin, D.B. (1998). *A broader template for analyzing broken randomized experiments*. Sociological Methods and Research, 27, 285-317.
- Bloom, H.S. (1984). *Accounting for no-shows in experimental evaluation designs*. Evaluation Review, 8, 225-246.
- Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others*. New York: Chapman & Hall/CRC.
- Böhning, D., Seidel, W. (2003). *Editorial: recent developments in mixture models*. Computational Statistics & Data Analysis, 41, 349-357, Elsevier.
- Breslow, N.E. (1982). *Clinical trials*. In Encyclopedia of Statistical Sciences, 2, 13-21. Wiley, New York.
- Broniatowski, M., Celeux, G., Diebolt, J. (1983). *Reconnaissance de densités par un algorithme d'apprentissage probabiliste*. In Data Analysis and Informatics, Vol. 3. Amsterdam: North-Holland, 359-374.
- Burghardt, J., McConnell, S., Schochet, P., Johnson, T., Gritz, M., Glazerman, S., Homrighausen, J. (2001). *Does Job Corps work? Summary of the National Job Corps Study*. Document No. PR01-50, Princeton, NJ: Mathematica Policy Research, Inc.
- Burghardt, J., McConnell, S., Schochet, P. (2003). *National Job Corps Study: findings using administrative earnings records data. Final report*. Document No. PR03-92, Princeton, NJ: Mathematica Policy Research, Inc.
- Card, D., Sullivan, D. (1988). *Measuring the effect of subsidized training programs on movements in and out of employment*. Econometrica, Vol. 56, 3, 497-530.
- Casella, B., Berger, R.L. (2002). *Statistical Inference*. Duxbury, USA.
- Cheng, R.C.H., Liu, W.B. (2001). *The Consistency of Estimators in Finite Mixture Models*. Scandinavian Journal of Statistics, Vol. 28, 603-616.
- Dasgupta, A., Raftery, A.E. (1998). *Detecting features in spatial point processes with clutter via model-based clustering*. Journal of the American Statistical Association, 93, 294-302.
- Dempster, A.P., Laird, N., Rubin, D.B. (1977). *Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion)*. Journal of the Royal Statistical Society, Series B 39, 1-38.
- Efron, B., Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- Everitt, B. S., Hand, D. J. (1981). *Finite Mixture Distributions*. London, Chapman & Hall.
- Fisher, R.A. (1925). *The Design of Experiments*. Oliver and Boyd, London, 1st edition.

- Fisher, L., Dixon, D., Herson, J., Frankowski, R., Hearron, M., Peace, K. (1990). *Intention to treat in clinical trials*. In *Statistical Issues in Drug Research and Development* (K. Peace, ed.), 331-350. Dekker, New York.
- Flores, C.A., Flores-Lagunes, A. (2007). *Identification and estimation of causal mechanisms and net effects of a treatment*. University of Miami, Working Papers 0706.
- Flores-Lagunes, A., Gonzalez, A., Neumann, T. (2007). *Estimating the effects of length of exposure to a training program: the case of Job Corps*. IZA Discussion Paper, No. 2846.
- Fraker, T., Maynard, R. (1987). *The adequacy of comparison group designs for evaluations of employment-related programs*. *Journal of Human Resources*, Vol. 22, 2, 194-227.
- Fraley, C., Raftery, A.E. (1998). *How many clusters? Which clustering method? Answers via model-based cluster analysis*. *Computer Journal* 41, 578-588.
- Frangakis, C.E., Rubin, D.B. (1999). *Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes*. *Biometrika*, 86, 365-379.
- Frangakis, C.E., Rubin, D.B. (2002). *Principal stratification in causal inference*. *Biometrics*, 58, 21-29.
- Frangakis, C.E. (2006). *Comment to A. Forcina "Causal effects in the presence of noncompliance: a latent variable interpretation."*. *Metron*, Vol 64, 3, 1-27.
- Fukumizu, K., Akaho, S., Amari, S. (2003). *Critical lines in symmetry of mixture models and its application to component splitting*. *Advances in NIPS*, 15, 865-872.
- Green, P. J. (1995). *Reversible jump markov chain monte carlo computation and bayesian model determination*. *Biometrika*, 82(4), 711-732.
- Heckman, J.J. (1974). *Shadow prices, market wages, and labor supply*. *Econometrica*, 42, 679-694.
- Heckman, J.J. (1979). *Sample selection bias as a specification error*. *Econometrica*, 47, 153-162.
- Heckman, J.J., Hotz, V.J. (1989). *Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training*. *Journal of the American Statistical Association*, 84, 862-880.
- Heckman, J., Robb, R. (1985). *Alternative methods for evaluating the impact of interventions*. In Heckman and Singer (eds.), *Longitudinal analysis of labor market data*, Cambridge, Cambridge University Press.
- Heckman, J.J., Vytlačil, E.J. (1999). *Local instrumental variables and latent variable models for identifying and bounding treatment effects*. *Proceedings of the National Academy of Sciences, USA*, 96, 4730-4734.
- Hirano, K., Imbens, G., Rubin, D.B., Zhou, X. (2000). *Assessing the effect of an influenza vaccine in an encouragement design*. *Biostatistics*, 1, 69-88.
- Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.
- Holland, P. (1986). *Statistics and causal inference*. *Journal of the American Statistical Association*, 81, 945-960.
- Horowitz, J. Manski, C. (2000). *Nonparametric analysis of randomized experiments with missing covariate and outcome data*. *Journal of the American Statistical Association*.



- Imai, K. (2007a). *Sharp bounds on the causal effects in randomized experiments with "truncation-by-death"*. *Statistics & Probability Letters*, 78, 144-149.
- Imai, K. (2007b). *Identification analysis for randomized experiments with noncompliance and truncation by death*. Technical Report, Department of Politics, Princeton University.
- Imbens, G.W., Wooldridge, J.M. (2008). *Recent developments in the econometrics of program evaluation*. Institute for the Study of Labor (IZA), Bonn, Discussion Paper No. 3640.
- Imbens, G.W., Angrist, J. (1994). *Identification and estimation of local average treatment effects*. *Econometrica*, 62, 467-476.
- Imbens, G.W., Rubin, D.B. (1997a). *Bayesian inference for causal effects in randomized experiments with noncompliance*. *The Annals of Statistics*, Vol. 25, 1, 305-327.
- Imbens, G.W., Rubin, D.B. (1997b). *Estimating outcome distributions for compliers in instrumental variables models*. *Review of Economic Studies*, 64, 555-574.
- Imbens, G.W., Rubin, D.B. (2009). *Causal Inference in Statistics and the Medical and Social Sciences*. Cambridge, U.K.: Cambridge University Press (forthcoming).
- Jank, W. (2006). *Ascent EM for fast and global solutions to finite mixtures: an application to curve-clustering of online auctions*. *Computational Statistics & Data Analysis*, 51, 747-761, Elsevier.
- Jo, B. (2002). *Estimation of intervention effects with noncompliance: alternative model specifications*. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4 (Winter, 2002), 385-409.
- Jo, B. (2008). *Bias mechanisms in intention-to-treat analysis with data subject to treatment noncompliance and missing outcomes*. Published on behalf of American Educational Research Association, <http://jeb.sagepub.com/cgi/content/abstract/33/2/158>.
- Karlis, D., Xekalaki, E. (2003). *Choosing initial values for the EM algorithm for finite mixtures*. *Computational Statistics & Data Analysis*, 41, 577-590, Elsevier.
- Lalonde, R.J. (1986). *Evaluating the econometric evaluations of training programs with experimental data*. *American Economic Review*, 76, 604-620.
- Lalonde, R.J. (1995). *The promise of public sector-sponsored training programs*. *The Journal of Economic perspectives*, 9, 149-168.
- Lechner, M., Wunsch, C. (2007). *Are training programs more effective when unemployment is high?* IAB Discussion Paper 200707.
- Lee, D.S. (2008). *Training, wages, and sample selection: estimating sharp bounds on treatment effects*. *Review of Economic Studies*. forthcoming.
- Lee, Y., Ellenberg, J., Hirtz, D., Nelson, K. (1991). *Analysis of clinical trials by treatment actually received: is it really an option?* *Statistics in Medicine*, 10, 1595-1605.
- Lindsay, B.G. (1995). *Mixture models: Theory, Geometry and Applications*. Hayward, California, USA.
- Little, R., Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Little, R., Yau, L. (1998). *Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model*. *Psychological Methods*, 3, 147-159.
- Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

- Mansky, C. (1990). *Nonparametric bounds on treatment effects*. American Economic Review Papers and Proceedings, 80, 319-323.
- Mattei, A., Mealli, F. (2007). *Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination*. Biometrics, 63, 2, 437-446.
- McCullagh, P.A., Nelder, J. (1989). *Generalized Linear Models*. Second edition, London: Chapman & Hall.
- McDonald, P.D.M., Pitcher, T.J. (1979). *Age groups from size-frequency data: a versatile and efficient method of analysing distribution mixtures*. Journal of the Fisheries Research Board of Canada 36, 987-1001.
- McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, USA.
- Mealli, F., Imbens, G.W., Ferro, S., Biggeri, A. (2004). *Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes*. Biostatistics, Vol. 5, No. 2, 207-222.
- Mealli, F., Rubin, D.B. (2002). *Assumptions when analyzing randomized experiments with noncompliance and missing outcomes*. Health Services and Outcomes Research Methodology, Springer Netherlands, Vol. 3, Numbers 3-4, 225-232.
- Meier, P. (1991). *Comment on "Compliance as an explanatory variable in clinical trials" by B. Efron and D. Feldman*. Journal of the American Statistical Association, 86, 19-22.
- Meng, X.L., Rubin, D.B. (1991). *Using the EM to obtain asymptotic variance-covariance matrices: the SEM algorithm*. Journal of the American Statistical Association, 86, 899-909.
- Mercatanti, A. (2004). *Analyzing a randomized experiment with imperfect compliance and ignorable conditions for missing data: theoretical and computational issues*. Computational Statistics and Data Analysis, 46, 493-509.
- Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.
- Neyman, J. (1923). *On the application of probability theory to agricultural experiments: Essay on principles*. Translated in Statistical Science, 5, 465-480, 1990.
- Pernkopf, F., Bouchaffra, D. (2005). *Genetic-based EM algorithm for learning Gaussian Mixture Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8.
- Pilla, R.S., Lindsay, B.G. (2001). *Alternative EM methods for nonparametric finite mixture models*. Biometrika, 88, 535-550.
- R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Robins, J.M., Greenland, S. (1994). *Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial*. Journal of the American Statistical Association, 89, pp 737-749.
- Rosenbaum, P. (1984). *The Consequences of adjustment for a concomitant variable that has been affected by the treatment*. Journal of the Royal Statistical Society, Series A, 147, 656-666.

- Rosenbaum, P.R., Rubin, D.B. (1983). *The central role of the propensity score in observational studies for causal effects*. *Biometrika* 70, 41-55.
- Rosenbaum, P.R., Rubin, D.B. (1984). *Reducing bias in observational studies using subclassification on the propensity score*. *Journal of the American Statistical Association* 79, 516-524.
- Rosenbaum, P.R., Rubin, D.B. (1985). *Constructing a control group using multivariate matched sampling incorporating the propensity score*. *The American Statistician* 39, 33-38.
- Rubin, D.B. (1974). *Estimating causal effects of treatments in randomized and non randomized studies*. *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D.B. (1975). *Bayesian inference for causality: the importance of randomization*. *Proceedings of the Social Statistics Section of the American Statistical Association*, 233-239.
- Rubin, D.B. (1976). *Inference and missing data*. *Biometrika*, 63, 581-592.
- Rubin, D.B. (1977). *Assignment of treatment group on the basis of a covariate*. *Journal of Educational Statistics* 2, 1-26.
- Rubin, D.B. (1978). *Bayesian inference for causal effects*. *Annals of Statistics*, 6, 34-58.
- Rubin, D.B. (1979). *Discussion of "Conditional independence in statistical theory" by A.P. Dawid*. *Journal of the Royal Statistical Society, B* 41, 27-28.
- Rubin, D.B. (1980). *Discussion of "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu*. *Journal of the American Statistical Association*, 75, 591-593.
- Rubin, D.B. (1990). *Formal modes of statistical inference for causal effects*. *Journal of Statistical Planning and Inference*, 25, 279-292.
- Rubin, D.B. (2000). *The utility of counterfactuals for causal Inference – Discussion of "Causal inference without counterfactuals" by A. P. Dawid*. *Journal of the American Statistical Association*, 95, 435-438.
- Rubin, D.B. (2005). *Causal Inference using potential outcomes: design, modeling, decisions*. *Journal of the American Statistical Association*, 100, 322-331.
- Rubin, D.B. (2006). *Causal inference through potential outcomes and principal stratification: application to studies with censoring due to death*. *Statistical Science*, 21, 299-321.
- Schlattmann, P (2003). *Estimating the number of components in a finite mixture model: the special case of homogeneity*. *Computational Statistics & Data Analysis*, 41, 441-451, Elsevier.
- Schochet, P.Z., Bellotti, J., Cao, R.J., Glazerman, S., Grady, A., Gritz, M., McConnell, S., Johnson, T., Burghardt, T. (2003). *National Job Corps Study: data documentation and public use files: Volume I*. Mathematica Policy Research, Inc.
- Scott, W., Szewczyk, W.F. (2001). *From Kernels to Mixtures*. *Technometrics*, Vol. 43, No. 3, Special Tukey Memorial Issue, 323-335.
- Sheiner, L.B.; Rubin, D.B. (1995). *Intention-to-treat analysis and the goals of clinical trials*. *Clinical Pharmacology and Therapy*, 57, 6-10.
- Titterton, D.M., Smith, A.F.M., Makov, U.E. (1985). *Statistical analysis of Finite Mixture Distributions*. New York: Wiley.

- Tohka, J., Krestyannikov, E., Dinov, I.D., Graham, A., Shattuck, D.W., Ruotsalainen, U., Toga, A.W. (2007). *Genetic algorithms for Finite Mixture Model based voxel classification in neuroimaging*. IEEE Transaction in Medical Imaging, Vol. 26, No. 5, May 2007.
- Ueda, N., Nakano, R., Ghahramani, Z., Hinton, G.E. (2000). *SMEM algorithm for mixture models*. Journal of VLSI Signal Processing Systems, Volume 26, 133-140; Kluwer Academic Publishers, Hingham, MA, USA.
- Van Buuren, S., Oudshoorn, C.G.M. (2007). *mice: Multivariate Imputation by Chained Equations*. R package version 1.16. <http://web.inter.nl.net/users/S.van.Buuren/mi/html/mice.htm>.
- van Ours, J. (2004). *The locking-in effect of subsidized jobs*. Journal of Comparative Economics, 32, 37-52.
- Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
- Vlassis, N.A., Likas, A. (2002). *A Greedy EM Algorithm for Gaussian Mixture Learning*. Neural Processing Letters 15(1), 77-87.
- Vose, M.D. (1999). *The Simple Genetic Algorithm: foundations and theory*. MIT Press, Cambridge, MA.
- Wu, C. (1983). *On the convergence properties of the EM algorithm*. The Annals of Statistics 11, 95-103.
- Yang, X., Liu, J. (2002). *Mixture density estimation with group membership functions*. Pattern Recognition Letters, Vol. 23, Issue 5, 501-512; Elsevier Science Inc., New York, USA.
- Zhang, J.L., Rubin, D.B. (2003). *Estimation of causal effects via principal stratification when some outcomes are truncated by "death"*. Journal of Educational and Behavioral Statistics, 28, 353-368.
- Zhang, J.L., Rubin, D.B., Mealli, F. (2008a). *Evaluating the effects of job training programs on wages through principal stratification*. In Modelling and Evaluating Treatment Effects in Econometrics, D.L. Millimet, J.A. Smith, E.J. Vytlačil, eds., Elsevier, 117-145.
- Zhang, J.L., Rubin, D.B., Mealli, F. (2008b). *Likelihood-based analysis of causal effects via principal stratification: new approach to evaluating job-training programs*. Journal of the American Statistical Association, forthcoming.