Joint analysis of occurrence and time to stability after entrance into the Italian labor market: an approach based on a Bayesian cure model with structured stochastic search variable selection

Alessio Farcomeni*

Dipartimento di Malattie Infettive e Sanità Pubblica, Sapienza - Università di Roma

Alessandra Nardi

Dipartimento di Matematica, Università di Roma - Tor Vergata

Elena Fabrizi

Dipartimento di Economia, Sapienza - Università di Roma

Abstract

Precarious employment is a serious social problem, especially in those countries, such as Italy, where there are limited benefits from social security. We investigate this phenomenon by analyzing the beginning point of the career of employees starting off with unstable contracts, for a panel of Italian workers. Our aim is to estimate the probability of obtaining a stable job and to detect factors influencing both this probability and the duration of precariousness. To answer these questions, we use an ad hoc mixture cure rate model in a Bayesian framework.

Key Words: Bayesian methods, mixture cure models, model selection, precariousness, structured variable selection, unobserved heterogeneity

*alessio.farcomeni@uniroma1.it

1 Introduction

In econometric analyses, accelerated failure time models are used to model duration phenomena like employment, life of a firm, etc. (Kiefer, 1988). In analyses of the labor market, duration models are commonly used to evaluate times from an initial state of unemployment to the achievement of a new job (e.g. Lancaster, 1990). Many studies are devoted to the study of working time (Bell and Hart, 2002), transitions from school to work (McVicar and Anyadike-Danes, 2002) and effects of training programmes (Olave and Salvador, 2007). A different aspect is the analysis of transition from a temporary to an open ended contract (for instance Booth *et al.*, 2002; Berton *et al.*, 2008). The main aim in this case is to see whether, among the employees with a fixed term contract, it is possible to identify (and characterize) two groups: one group of subjects in which temporary contracts are stepping stones to more stable open ended contracts, and the other group, trapped in a situation of job precariousness. In Italy and many other countries, an open ended contract is a guarantee for the employee of benefits from social security protection, not only during employment, but also at retirement (Picchio, 2008; Booth *et al.*, 2002; Barbieri and Sestito, 2008).

The kind of temporary contracts we consider are characterized by advantages for the firm, like contribution rebates (i.e., the firm receives funds or tax reductions for hiring) and disadvantages for the employees (i.e., the employee is not supported by an unemployment or sickness benefit). The rights of the worker are transferred to the firm in order to stimulate hiring. A consequence is that in the same firm there could be employees with the same job and different rights: on the one side, those hired with a regular contract (even fixed term), who enjoy benefits; and on the other, those hired with temporary contracts, who do not enjoy benefits. The labor market may be hence possibly segmented by these disparities, with workers (insiders) that are mostly exempted from labor market reforms with respect to the (penalized) outsiders (Barbieri and Sestito, 2008).

We analyse the beginning of the career of employees on entering the labor market with unstable contracts, using longitudinal micro data on work histories. For a panel of Italian workers we set the origin at the time of entering the labor market with a precarious job, and the event of interest is the obtaining of a stable job. Time to event is recorded regardless of the worker working in the same or different firms, and even being temporarily unemployed before the event. As we are only interested in the achievement of stability, a transition to another unstable

job, or even to unemployment can be regarded as an unobserved heterogeneity, which we account for in our model (see below).

The data, definitions of stable and unstable jobs, and all other necessary information will be described in Section 4.

We investigate not only the time necessary to obtain a stable job for new entrants into the labor market and the factors influencing this duration, but also whether a precarious worker will ever hold a stable job. Our answer to the latter question, for our restricted group, is negative since we estimate that about 9% of precarious workers will never be granted a stable contract.

To answer our questions we used an ad hoc mixture cure rate model (Boag, 1949; Farewell, 1982). Cure models are used to model survival data when a positive fraction of individuals is expected not to experience the event of interest. They have been originally proposed to estimate the proportion of patients cured from a disease in medical research but their use has been extended to other areas of research (Hernes, 1972; Bennett *et al.*, 1989). The cure models we propose allow both the survival function of uncured subjects and the cure fraction to depend on covariates, which are possibly different. The model is formulated in the Bayesian framework and, coherently, Bayes factors are used for testing. We use fractional Bayes factors (O'Hagan, 1995), which have good properties (see e.g. O'Hagan, 1997).

Greenhouse and Silliman (1996) suggest the use of Bayes factors also for performing model selection, but their approach requires enumerating the model space, which is time consuming and may not be feasible for even a moderate number of covariates. For this reason, we preferred to adapt a recently developed stochastic search strategy for variable selection to the context of cure models. We used *structured* variable selection, which allows for selecting covariates in groups and hierarchically. We needed to force groups of covariates to enter the model simultaneously because of the presence of many polytomous categorical variables which are reparameterized with dummies. Further, we considered the possibility of including interactions into the model, and we imposed a hierarchical structure (i.e., the interaction cannot enter the model if the original effects are not included) for interpretability reasons. The Bayesian approach for structured automatic variable selection we used has recently been proposed by Farcomeni (2010) for generalized linear models; we extended it to mixture survival models. All the proposed techniques are (frequentist) consistent, in the sense that they asymptotically lead

to the choice of the right model if it is in the collection of possible models, and otherwise of the model closest to the true one with respect to Kullback Leibler divergence otherwise (Farcomeni, 2010).

In our context it is important to take into account possible unobserved heterogeneity, which may bias the estimated parameters. Presumably, some of the workers who do not menage to enter into permanent employment may differ from those who get a permanent job not only along the observables used in estimation but also along other (possibly unobservable) dimensions, e.g., motivation, ability, and work ethics. These are reflected also in the other unmodeled transitions (i.e., to other jobs, to unemployment). The effects of unobserved factors is commonly taken into account by including subject specific random effects. Price and Manatunga (2001) have already extended mixture cure models for the use of frailty to capture unobserved heterogeneity in the maximum likelihood framework. In this work we extend work of Koop (2003) to show that unobserved heterogeneity in mixture cure rate models can be cast in a very simple and straightforward way in the Bayesian framework: we shall specify subject specific intercepts, and model them hierarchically through the addition of a second level of priors.

The contribution of this work is therefore two fold. On one hand, we develop in full the analysis of an original study related to the Italian labor market. On the other hand, we develop in full a Bayesian framework with structured variable selection for cure models. We describe a simple and general Gibbs sampler for approximating the posterior and a systematic approach to maximum likelihood estimation (with the EM algorithm).

The outline of the paper is as follows: in Section 2 we present the cure model, derive the complete likelihood and discuss model fitting. Structured variable selection for mixture cure models is developed in Section 3. In Section 4 we apply the methods for investigating the evolution of precariousness in the Italian labor market.

2 Mixture cure models

In order to model time to a stable job we introduce a parametric cure rate model. Let (T_i, δ_i) denote the observed time and the event indicator for the i-th subject. Denote also by X_i a vector of subject specific covariates with a leading unit term for the intercept, and by $S(t|X = x) = \int_t^{+\infty} f(s|X = x)\, ds$ the survival function

of T.

There are many parametric families of distributions which can be assumed to model failure times (see e.g. Kalbfleisch and Prentice, 1980). and the semiparametric approach that characterises the Cox model could be used as an alternative. However, all these choices are unsatisfactory when there is a fraction of individuals who are not bound to experience the event. These individuals are always censored, with $\delta_i = 0$ for any t, and are the primary interest in our application. Even when not of primary interest, the presence of such a fraction of individuals should be taken into account since it may bias the analyses. When there is a fraction of cured individuals, for instance, the assumption of a constant hazard ratio is trivially violated. Even with non parametric approaches the predicted medians could be biased upwards.

Cure models can be used to overcome these problems. Formally, a fraction p of individuals is assumed to have degenerate survival function $S(t|X = x) = 1$, that is, to have zero risk and never experience the event, while the remaining proportion $1 - p$ of individuals have a failure time distributed according to the assumed model.

To derive the likelihood function we introduce an unobserved latent indicator Z_i which is equal to 1 when the i-th individual is cured, that is, he/she will never obtain an open ended contract. In the simpler case of no covariates affecting the latent indicator, we have $P(Z_i = 1) = p$, where p is the cure fraction. The two components of the survival distribution can then be indexed by Z_i, where $S(t|X = x, Z = 1) = 1$ for any $t \geq 0$ and $S(t|X = x, Z = 0)$ is a proper survival function. We note that in many real applications the assumption $S(t|X = x, Z = 1) = 1$ is just a useful approximation of the survival distribution corresponding to an extremely low hazard. Subjects thought to belong to the cure fraction are often called "long term survivors" for this reason.

In this paper we restricted to parametric assumptions for the failure times associated with $Z = 0$. Extension to semiparametric assumptions is straightforward: also the Cox model can also be used to describe the hazard of non cured individuals (Kuk and Chen, 1992). However, in our experience it might become overly adaptive. The approximate zero risk of the cured individuals may be compressed in the baseline hazard, resulting in a biased estimate of the cure fraction and even in an unstable fit (see also Chen *et al.*, 2002). Note further that the non parametric baseline hazard is compatible with a zero value at any time t, and thus the

interpretation of high/low hazard rates may become ambiguous.

We did not have identifiability issues: any parametric assumption for failure times associated with $Z = 0$ leads to an identifiable model (see e.g. Li *et al.*, 2001, Theorem 3; the theorem directly extends to the situation in which random effects are included).

The effect of the covariates X_i on the risk of failure is modeled by reparametrizing the scale parameter μ according to a logarithmic link function, i.e. $\log(\mu_i) = \beta' x_i$, while the shape parameter γ is assumed to be constant. This assumption corresponds to the classical structure of accelerated failure time models where the effect of covariates is to accelerate or decelerate a baseline survival time T_0 by a factor $e^{-\beta' x_i}$.

The marginal survival function (with respect to Z_i) is given by

$$\begin{aligned} S(t|X=x) &= p\, S(t|X=x, Z=1) + (1-p) S(t|X=x, Z=0) \\ &= p + (1-p) S(t|X=x, Z=0). \end{aligned}$$

Using the short hand notation θ for the vector of parameters characterizing $S(t|X=x, Z=0)$, the corresponding likelihood function can be written as:

$$\begin{aligned} l(\theta, p) &= \prod_{i=1}^{n} (f(t_i|\theta, X_i, Z_i=0)(1-p))^{\delta_i} \\ &\quad (p + S(t_i|\theta, X_i, Z_i=0)(1-p))^{1-\delta_i}. \end{aligned} \tag{1}$$

It is difficult to work directly with the likelihood in mixture models, and in practice, both in the Bayesian and in the maximum likelihood approach, use of the complete likelihood (that is, the likelihood conditional on the latent indicators) is often more convenient. The complete data likelihood can be written as

$$\begin{aligned} L(\theta, p) &= \prod_{i=1}^{n} (f(t_i|\theta, X_i, Z_i=0)(1-p))^{\delta_i} \\ &\quad (p^{z_i} (S(t_i|\theta, X_i, Z_i=0)(1-p))^{1-z_i})^{1-\delta_i} \end{aligned} \tag{2}$$

In many cases the same or a different set of covariates, which we denote by $\tilde{X}$, may affect the probability p of being a long term survivor. In order to let also the distribution of the latent indicator depend on covariates, we can use a logistic link and set

$$P(Z_i = 1|\tilde{X}_i) = \frac{\exp\{\alpha' \tilde{X}_i\}}{1 + \exp\{\alpha' \tilde{X}_i\}} \tag{3}$$

The corresponding complete likelihood is simply obtained by substituting p with its reparameterization (3) in (2).

2.1 Prior Specification

In the Bayesian framework, model is fit after specifying appropriate priors $\pi(\cdot)$ for the parameters. In the most general case we have three sets of parameters:

- A shape parameter γ
- Regression parameters for the time component β.
- Regression parameters for the cure fraction component α.

Available prior information should be summarized and properly taken into account when choosing the priors. The general approach we propose is to make use of independent priors. For the shape parameter γ we use a lognormal distribution. For the regression parameters, in agreement with the literature on Bayesian regression models, we suggest zero centered normal priors with standard deviation σ_β. The vector of standard deviations need not be constant; for instance a larger variance is often used for the intercept. We suggest a similar choice for regression parameters α, as in the literature on Bayesian logistic models. In order to obtain a marginal estimate of the cure fraction, one may not wish to let p depend on covariates; in this case, a Beta distribution can be assumed for p.

Prior sensitivity can be evaluated by repeatedly fitting the model for different prior inputs and comparing posterior summaries, like the posterior means. The choice of priors rarely influences the results significantly when many observations are available. In our example, for instance, posterior means are very close to maximum likelihood estimates for any reasonable choice of prior inputs.

2.2 Unobserved Heterogeneity

While effects of interest are directly adjusted for observed potential confounders, unobserved confounders can still bias the conclusions and affect variability of the posteriors. In econometric studies and many other fields, these problems are usually taken into account including a subject specific intercept into the (regression) models. The subject specific parameter can be expected to capture *individual* effects beyond the observed covariates.

In our framework, a mixture cure rate model with subject specific parameters can still be expressed as $S(t_i|X = x_i) = p_i + (1 - p_i)\exp\{-(\mu_i t)^\gamma\}$, but μ_i and p_i are now reparameterized as:

$$\begin{cases} \log(\mu_i) = \beta_{0i} + \beta' x_i \\ p_i = \frac{\exp\{\alpha_{0i} + \alpha' \tilde{X}_i\}}{1+\exp\{\alpha_{0i} + \alpha' \tilde{X}_i\}} \end{cases} \tag{4}$$

We assume the subject specific intercepts β_{0i} and α_{0i} arise independently from zero centered normal distributions, that is, $\beta_{0i} \sim N(0, \sigma^2_{\alpha_0})$. The variance parameters are modeled through inverse gamma priors. This is the model we will fit to the data.

It is now clear that unobserved heterogeneity can be easily taken into account merely by using a second level of priors: in the Bayesian framework inclusion of a random intercept only involves an additional step of learning on the variance of the random intercepts. We use an inverse gamma prior for $\sigma^2_{\beta_0}$ and for $\sigma^2_{\alpha_0}$. See Koop (2003) and Li and Zheng (2008) for further discussion on the issue of unobserved heterogeneity in Bayesian modeling.

We finally note that we have a random sampling scheme involving random effects. As noted by McCullagh (2008), estimates could thus be biased unless the joint distribution depending only on the list of covariate values on the sampled units has independent components. This is exactly our situation, since subjects appear in the sample only once and intercepts are subject specific.

2.3 Model Fit

In the Bayesian framework, model is fit by computing the update of prior information given by the data, summarized in the posterior distribution for the parameters. Exact calculations on the basis of Bayes theorem are not easy in our case and we must rely on the powerful MCMC tools (Robert and Casella, 2000), by which one can obtain a sequence of simulated values for the parameters which are distributed like a (dependent) random sample from the posterior distribution. The simulated parameters can be used to approximate integral quantities via empirical averages; thus estimating posterior means, medians, standard deviations or even the entire posterior distribution.

The sample from the posterior distribution in our case can be obtained through an ad hoc Gibbs sampling scheme which has been derived for mixture models (Diebolt and Robert, 1994).

The Gibbs sampler for mixture models proceeds by first sampling the latent indicators from their posteriors given the data and the current values for the other parameters, and then sampling the parameters from their full conditionals given the current latent indicators and the data. Since we work conditionally on the latent indicators, for sampling the parameters we make use of the complete likelihood (2).

The general iteration of the Gibbs sampling scheme we propose for mixture cure models is as follows:

1. Sample Z_i, $i = 1, \ldots, n$ from

$$P(Z_i = 1|\delta_i = 0) = \frac{p_i}{p_i + (1 - p_i)S(t_i|X_i, \beta_0, \theta, Z_i = 0)}, \qquad (5)$$

   where θ and β_0 are as in the current iteration of the sampler. Note that in the mixture cure model $P(Z_i = 1|\delta_i = 1) = 0$.

2. Sample $\sigma^2_{\alpha_0}$ from its full conditional, and

$$\pi(\alpha|Z) \propto \pi(\alpha) \prod_{i=1}^{n} \frac{e^{Z_i(\alpha_{0i} + \alpha' \tilde{X}_i)}}{1 + e^{\alpha_{0i} + \alpha' \tilde{X}_i}}.$$

   Compute p_i as in (3).

3. Sample $\sigma^2_{\beta_0}$ from its full conditional, and the remaining parameters θ from

$$\pi(\theta|T, Z) \propto L(\theta, p)\pi(\theta)$$

In the sampling scheme proposed, few full conditionals are not available in closed form. A common solution would be to set up Metropolis Hastings (MH) steps in order to sample from those full conditionals, which are known only up to the normalizing constant. However, there are many difficulties associated with setting up MH. The key to success for MH in fact is a clever candidate transition kernel, which does not seem easily available here. Furthermore, the full conditional distribution at Step 3 is also potentially multimodal, and even if a good candidate transition kernel were available, tuning of MH would be made harder by the presence of the second component in the mixture and volatility in Z. In order to avoid these difficulties, we sample the parameters in θ simultaneously with Adaptive Rejection Metropolis Sampling (ARMS, Gilks *et al.*, 1995).

Sampling from $\pi(\alpha|Z)$ can also be performed via an ARMS, even if there are many different common alternative approaches for this standard problem, which arises when performing Bayesian logistic regression.

If we wish to fit a fixed effects model, we simply set $\sigma^2_{\beta_0} = \sigma^2_{\alpha_0} = 0$. In the case in which the cure fraction is not assumed to depend on random intercepts or covariates and a Beta prior is used for p, a closed form expression for the corresponding full conditional is known and corresponds to a Beta with parameters $\gamma_1 + \sum_{i=1}^n Z_i$ and $\gamma_2 + \sum_{i=1}^n (1 - Z_i)$, where γ_1 and γ_2 are the prior parameters.

We finally note that other more sophisticate methods can be used for approximating the posterior, like slice sampling. The Gibbs sampler is however computationally convenient in this case, and it has proved efficient on our data as we briefly mention below.

2.4 Fractional Bayes Factors for Testing Hypotheses

Bayes Factors (BF) (Kass and Raftery, 1995) can be used for testing hypotheses. The Bayes factor is simply the ratio between the marginal likelihoods of the data under the alternative hypothesis H_1 and under the null hypothesis H_0. It is usually interpreted as a measure of the evidence given by the data in favor of a model compared to a competing one. Roughly, a BF greater than 1 reveals the data provide greater evidence in favor of H_1 than H_0, and the opposite conclusion holds otherwise. Jeffreys (1961) proposed an empirical scale for classifying evidence provided by a BF. For instance, values larger than 3 (smaller than 1/3) are judged as a moderate but clear evidence in favor of the alternative (null) hypothesis.

Null hypotheses of interest include $H_0 : p = 0$, and tests on linear combinations of β and/or α regression parameters, together with comparison of parametric models (for instance, a Weibull versus a Lognormal model). While in the maximum likelihood approach the first and the last test are difficult to perform because the likelihood ratio test statistic may not be asymptotically chi squared, in the Bayesian approach this is not an issue. Further, Bayes factors naturally penalize for model complexity. Bayes factors can be also used to test hypotheses on the regression parameters. Whenever data provide more evidence in favor of $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ we can conclude that the j-th covariate has no effect on time to event for non cured individuals. In the same fashion, whenever data provide more evidence in favor of $H_0 : \alpha_j = 0$ against $H_1 : \alpha_j \neq 0$, we can conclude that there is no effect of the j-th covariate on the probability of being

cured.

Since in mixture cure models the marginal likelihoods have cumbersome or no closed form expressions, a suitable approximation is required. Approximation of normalizing constants is a very common problem in Bayesian statistics, which here is made harder by the mixture structure of the likelihood.

In this paper we use fractional Bayes factors, which are in some sense objective, and convenient from different perspectives (O'Hagan, 1995, 1997). It is well known that using improper priors for performing objective inference would lead to indeterminate values of the Bayes factor. To partially overcome this difficulty, a possibility is given by setting a training fraction b, and using the fractional Bayes factor which is defined as

$$BF_{01}^{b} = \frac{q_0}{q_1}, \tag{6}$$

where

$$q_i = \frac{\int_{\Theta_i} l(\theta, p)\pi(\theta, p)}{\int_{\Theta_i} l(\theta, p)^b \pi(\theta, p)}, \qquad \text{i=0,1}$$

and Θ_0 and Θ_1 denote the parameter spaces respectively under the null and alternative hypotheses.

The training fraction is chosen as $b = \sqrt{n_0/n}$ as suggested by O'Hagan (1997), where n is the sample size and n_0 is the size of the minimal training sample leading to proper inference. In our application we also tried different values of n_0, obtaining approximately the same results. This was to be expected since the sample size is large, and hence any choice of $n_0 << n$ leads to approximately the same b.

Computation of the fractional Bayes factors requires approximating the four integrals in (6). As noted by Gilks (1995) in the discussion of O'Hagan (1995), this can be accomplished with two separate runs of the same MCMC tools used for approximating the full posterior. One can in fact sample from the *fractional* posterior distribution under the null and under the alternative, which is equivalent to the full posterior in which the likelihood is raised to a power of b. Then, q_i is approximated as the average value of the likelihood raised to a power of $1 - b$ computed in the MCMC generated samples. As noted by Gilks (1995), this is expected to work well since the fractional posterior delivers samples in the regions of high likelihood. The strategy does not pose problems in our mixture situation since we can always evaluate the likelihood for any point in the parameter space.

We finally note that due to the large sample size, all alternative forms of Bayes

factors (i.e., regardless of the prior) approximately converge (see, for instance, discussion in O'Hagan, 1995). We could then also use Schwarz criteria (Schwarz, 1978), which are based just on the maximum of the likelihood, to approximate Bayes factors. Greenhouse and Silliman (1996) also use Schwarz criteria in mixture cure models, noting that the approximation to BF is satisfactory in large samples. The Schwartz criterion consists in approximating the logarithm of the Bayes factor with

$$SC = \log(L(\hat{\theta}_{H_1})) - \log(L(\hat{\theta}_{H_0})) - 0.5d\log(n), \tag{7}$$

where d denotes the difference in number of parameters between the two models, and $\log(l(\hat{\theta}_{H_i}))$ is the maximum of the log likelihood under H_i, $i = 0, 1$. It can be proved that the Schwarz criterion approximates the Bayes factor as the number of observations grows, and that the rate of convergence is $O(n^{-1/2})$ whenever the models are nested with a reasonable specification of the priors. When the random intercepts are omitted, the maximum of the likelihoods can be obtained by using an EM algorithm (Dempster *et al.*, 1977). When random intercepts are included in order to take into account unobserved heterogeneity, the EM algorithm shall be modified to an MCEM algorithm (e.g., Wei and Tanner, 1990; Sherman *et al.* (1999)). These algorithms are described in Appendix A.

3 Variable Selection

Variable selection could be performed by comparing each couple of possible models (that is, for each combination of covariates included and excluded from the model), as in Greenhouse and Silliman (1996). However, this procedure involves enumerating the model space, whose dimension increases exponentially with the number of prospective covariates. It is thus unfeasible even in small dimensional situations, especially if one considers interactions and/or transformations. We propose instead an automatic approach for variable selection, adapting a version of Stochastic Search Variable Selection (SSVS) from George and McCulloch (1993).

We consider the possibility of including the original variables and their interactions. In this step, for interpretability reasons, we impose a hierarchical structure on the selected model: an interaction shall not be included in the model if the corresponding main effects are excluded. For categorical variables, we adopt a corner point reparameterization, using a dummy for each category except the baseline one. This implies that grouping constraints are required: the dummies

arising from a single variable should be included or excluded *altoghether*. The context of grouped and hierarchical variable selection is usually referred to as *structured* variable selection.

Structured variable selection for linear and generalized linear models has been introduced in the Bayesian framework by Farcomeni (2010), who provides an opportune extension of the classical SSVS.

The approach involves a slight modification of the priors for the regression parameters α and β. Variable selection can be performed by using a mixture of two centered normals as prior for the regression parameters (excluding the intercepts). The mixture is indicized by a latent indicator variable η_j, which flags the corresponding covariate as in or out of the model. An expression for the prior for the regression parameters of the time component is given by:

$$\pi(\beta_j|\eta_j) \sim \eta_j N(0, \tau_{1j}^2) + (1 - \eta_j)N(0, \tau_{0j}^2),$$

where τ_{1j} is slightly larger than τ_{0j}. The prior variance τ_{0j} is chosen small enough in order to concentrate the second component of the mixture around values operationally close to zero. Whenever $\eta_j = 1$, the usual Gaussian prior is used for β_j and the j-th covariate is included.

The same can be done for the α parameters. Imposing a hierarchical constraint is equivalent to imposing that the η_j corresponding to the interaction is zero whenever any of the ηs corresponding to the original variables are zero. This would result in intractable priors. Farcomeni (2010) proposes a simple solution through a user supplied indicator matrix ω. The user sets an indicator function $\omega_j(i)$ which is 1 if the j-th variable must be included in every model in which i-th is included, and zero if there are no hierarchical constraints between i-th and j-th variable. Further, there are grouping indicators $\phi_k(j)$ which are 1 if the j-th variable belongs to the k-th group, $k = 1, \ldots, g$, and zero otherwise.

Constraint indicators are linked to variable indicators via the following reparameterization:

$$\eta_j = \left(\prod_{k=1}^{g} \nu_k^{\phi_k(j)} \prod_{j \neq i} \eta_i^{\omega_i(j)} \right), \tag{8}$$

where ν_k are second level latent indicator variables which indicize if the k-th group should be included in the model. Finally, a Bernoulli prior is put on each ν_k: $P(\nu_k = 1) = w_k$, where a common choice is $w_k = 0.5$.

The model can be fit as in Section 2.3, with the sole addition of one step in which the indicators ν_k are sampled from their full conditionals, given by a Bernoulli:

$$P(\nu_k = 1|\theta, \nu_{-k}, X, Z) = \frac{w_k \pi(\beta|\nu_{-k}, \nu_k = 1)}{w_k \pi(\beta|\nu_{-k}, \nu_k = 1) + (1 - w_k)\pi(\beta|\nu_{-k}, \nu_k = 0)},$$

where ν_{-k} denotes the vector of ν indicators with the k-th component excluded. After sampling the ν indicators, (8) is applied and the Gibbs sampler continued as in Section 2.3. A similar approach can be used to perform variable selection for the regression parameters for the cure fraction (see Farcomeni (2010) for additional strategies for sampling and inverting (8))

Model selection is based on the posterior for η. Farcomeni (2010) proves that frequentist consistency in model selection is achieved using the median model (Barbieri and Berger, 2004), that is, by inclusion of the j-th variable if the posterior for η_j is above the cut-off 0.5. These results directly extend to the mixture survival case.

4 Application to WHIP Data

4.1 Description of the data

The data source is the Work Histories Italian Panel (WHIP), an employer employee linked panel database developed by the Italian Social Security Administration (INPS), and elaborated by the Laboratorio R. Revelli (Turin, Italy). Refer to `www.laboratoriorevelli.it` for detailed information.

The reference population is made up of all the individuals working in Italy during the period from January 1985 to December 2003, both Italians and immigrants, aged 16-30 and at their first registered work experience in the private sector. The elf employed, together with workers in the agricultural and public administration sectors, are not included in our analyses.

We are not interested in investigating the entire Italian labor market. Among the precarious contracts we focus on new entrants whose first contract is of the "on the job training" (named CFL in Italy) or apprenticeship type. The flexible part of the subordinate labor force, during the period we are studying, is indeed concentrated in these two types of contracts, which are fixed term and have contribution rebates. These contribution rebates were introduced by the Italian government in

order to give more employment opportunities to underprivileged people outside the regular labor market, and especially to young people without any work experience; the main goal was to reduce the duration of the transition from school to work and upgrade young workers' human capital. These kinds of contracts were also intended to stimulate the emergence of workers from the shadow economy.

In our analysis we do not include two new kinds of flexible contracts that were introduced in Italy at the end of the '90s: temporary agency work and collaborators (which are named Co.Co.Co. in Italy). These are heterogeneous contract forms with respect to the more focused forms that were available also in 1986 and in 1992. They are characterized by the absence of a subordinate relationship with the employer (collaborators are formally self employeed while temporary agency workers have a contract with an external agency) and until 2001 these contracts were permitted only for non manual jobs. Temporary agency contracts were anyway less than 1% in 1998 for our reference population (and 0% in 1986 and 1992).

The data are a representative random sample drawn by INPS from the entire population (sampling coefficient is around 1:90). There is also no attrition (missing data): it is compulsory to provide records on employees and firms to the social security administration.

We record the time from the beginning of the first work experience to a removal of all the benefits for the employer, if this happens, in which case the worker has been granted a normal contract and the job is deemed to be stable (even if the contract may still be fixed term). The cure fraction is made up of individuals who will never be able to obtain stable contracts. We deem fixed term contracts as stable because a worker with a fixed term contract but no contribution rebate is equivalent to a worker with open ended contract in terms of labor cost and guaranteed protections. These rights are not equivalently guaranteed in the case of CFL or apprenticeship contract that are also characterised by a lower level of sickness benefits. In this sense the worker without a contribution rebate can be considered as integrated in the labor market. From a different perspective, it can be said that contracts with benefits are different in nature and objectives from contracts without benefits, be they fixed or long term. A more restrictive definition of the event would eventually only provide larger estimates for the cure fraction.

A total of $n = 6648$ subjects are followed for five years in three different, non overlapping, time frames starting in 1986, 1992, 1998. We restrict our analysis to young subjects (16 to 30 years old at the first job, with an average of 21 and

standard deviation of 3.7). We have 5807 (84.3%) events with a median follow up of about two years.

The measured covariates are: sex (60% males), age (minus sixteen), and size of the firm (small, medium, large; according to cut offs of 10 and 50 employees as fixed by UnionCamere, `http://www.unioncamere.it`). We have 42% of the subjects working in small fims, 28% in medium firms and the remaining 30% in large firms. We have also recorded the geographical region (north: 61%, center: 22%, south: 17%), and the firms operational sectors (production of goods: 58%, services: 28%, other: 14%).

Since the follow up periods for the three cohorts are non overlapping, the period is treated as a further covariate. We have 35% of the subjects entering labor market in the first period, 25% in the second, and the remaining 40% in the third.

For the analysis, we create a dummy for each level (except one) of each categorical variable.

4.2 Analysis of Whip Data

We compare different choices for the distribution of the non degenerate component using the fractional Bayes factor. The same approach is used to test the important null hypothesis $H_0 : p = 0$ against $H_1 : p > 0$.

We proceed as follows: we fit six different models with the variable selection method described in Section 3. We consider a generalized extended gamma, a Weibull and a lognormal accelerated failure time model and three cure models: one assuming a generalized extended gamma distribution for the non degenerate component, one assuming a Weibull and one assuming a lognormal distribution.

The extended generalized Gamma model (Kalbfleisch and Prentice, 1980) is a very flexible distribution that includes exponential, Weibull, reciprocal Weibull, lognormal and gamma as special cases; it can model both positively and negatively skewed probability density function for the log times.

When performing model selection, we consider the possibility of hierarchically including two way interactions. We have grouping constraints given by groups of dummies arising from each non binary categorial variable, and hierarchical constraints between each interaction and the generating couple of original variables. Hence the saturated model includes all covariates and all their two way interactions, both on the time and on the cure fraction component for the cure rate

models.

To take into account possible unobserved heterogeneity, we model the prior variance for the subject specific parameters as in (4) with inverse gamma priors.

We let each sampler run for a burn in of 20000 iterations, and then record one every tenth iteration for the following 130000. Diagnostics of convergence give satisfactory results in all cases. We run the sampler from different starting solutions and, after burn in all parallel chains, we approximately obtain the same posterior means.

Further, at the end of the burn in autocorrelations for all parameters vanish after lag 5 (even earlier in many cases).

Table 1 shows the fractional Bayes factors, on the log scale, for comparing the different parametric assumptions considered. We also report Deviance Information Criteria (DIC) in Table 2. DIC were introduced by Spiegelhalter *et al.* (2002), and extended to the case of mixture models by Celeux *et al.* (2006). For the cure models, we use DIC_4 as defined in Celeux *et al.* (2006), which in that paper was found to be more efficient than all other proposals for mixture models.

Table 1: Log fractional Bayes factors for comparing different parametric assumptions: an extended generalized gamma model (M_1), a Weibull (M_2), a lognormal (M_3), an extended generalized gamma cure model (M_{1c}), a Weibull cure model (M_{2c}) and a lognormal cure model (M_{3c}).

	H_0					
H_1	M_1	M_2	M_3	M_{1c}	M_{2c}	M_{3c}
M_1		18	11	-406	-426	-354
M_2	-18		-178	-388	-566	-508
M_3	-11	178		-366	-393	-352
M_{1c}	406	388	366		-29	-33
M_{2c}	426	566	393	29		37
M_{3c}	354	508	352	33	-37	

Both the fractional Bayes factor and DIC suggest using the Weibull cure model. It is not surprising that the more flexible extended generalized gamma model with a cure fraction is discarded in favor of one of its special cases: the prior acts as a penalty so that, when two models fit almost equally well, the simpler is preferred.

Note further that there is an overwhelming evidence in favor of the cure models

Table 2: DIC for comparing different parametric assumptions

Extended Generalized Gamma	49708.67
Weibull	49721.92
LogNormal	49694.45
Extended Geneneralized Gamma cure model	49313.56
Weibull cure model	49282.02
LogNormal cure model	49330.33

confirming the presence of a positive cure fraction of subjects who will *never* find a stable job. The Bayes factor for testing $H_0 : p = 0$ against $H_1 : p > 0$ is estimated as about 150 on the log scale, when we assume Weibull failure times and do not use covariates. The null hypothesis is then rejected and the cure fraction is marginally estimated as 9.4% (95%HPD: 8.1%-10.1%).

In order to verify whether there also was an effect of covariates on the shape of the distribution, conditionally on the chosen Weibull model with the selected covariates, we specified a model for the shape parameter as follows:

$$\log(\gamma) = \beta'_\gamma x_i,$$

We considered the possibility of including any of the available covariates, and two way interactions. We performed variable selection along the same lines as structured variable selection performed for selecting the model on the scale parameter. All approximated probabilities of inclusion were below 0.5, so that we concluded that no covariate should be used to model the shape parameter.

In Table 3 we report the final results for the Weibull cure model: estimates of the acceleration factors for the time component, logistic regression parameters for the cure fraction and posterior probability of inclusion, denoted as $P(\eta_j) = 1$. As stated we included variables for which $P(\eta_j) > 0.5$ (Barbieri and Berger, 2004).

It is understood that variables not included in the model are not of primary importance for the outcome. For instance, we have found that sex does not have a strong effect on occurrence or time to stability while it is well known that there often is a wage advantage of men over women in the labor market. This advantage is in part linked to the life choices, like giving birth to a child, and to career discrimination of women. We can explain the missing effect of gender by the fact that there is no career issue for precarious workers. Moreover, Italian women

beginning their career with an unstable contract often delay choices which would give them a disadvantage with respect to other workers.

Table 3: WHIP Data: Median model with parameter estimates, lower and upper 95% credibility intervals and probability of inclusion of covariates.

	Parameter	$HPD_{0.95}^{low}$	$HPD_{0.95}^{up}$	$P(\eta_j) = 1$
Time component				
γ	1.85	1.81	1.89	
period: second	1.18	1.14	1.23	0.77
period: third	1.32	1.28	1.37	0.77
age	0.94	0.93	0.94	0.81
firm: medium	0.88	0.85	0.91	0.90
firm: large	0.84	0.80	0.88	0.90
sector: services	0.89	0.85	0.95	0.71
sector: other	0.98	0.90	1.09	0.71
sector: services*age	1.02	1.01	1.03	0.64
sector: other*age	1.01	0.99	1.02	0.64
p component				
age	-0.37	-0.40	-0.32	0.99
region: North	-0.63	-0.82	-0.40	0.76
region: South	-0.06	-0.35	0.21	0.76
sector: services	-0.41	-0.66	-0.17	0.89
sector: other	0.42	0.17	0.65	0.89

The parameter estimates suggest a few remarks. Note that by inversion of the posterior credibility intervals we can simply perform tests on the parameters. We have the following comments: finding a stable job is more likely in the services sector (mainly, communications) than in production or other sectors, and with shorter waiting times.

All the same, older subjects have both shorter time to event and cure fraction. We must be careful about the interpretation of the age coefficient. In fact, age is only in part a proxy for the education level, which is not available from the administrative source. People entering the market later are usually more qualified, but of course all subjects above 25 cannot be considered as freshly graduated. To help with interpretation of the age coefficient, we report in Table 4 the distribution

level of education	15-24	25-34
primary school or lower	2.00%	3.85%
secondary school	53.69%	32.40%
high school	42.88%	49.09%
university	1.43%	14.67%

Table 4: Distribution of Italian population by education and age class for ages 15-34. Year 2004. Source: ISTAT (Italian National Statistical Institute).

of the Italian population by education and age class in year 2004, which is taken from the routine survey carried out by the Italian national statistical institute ISTAT (ISTAT, 2005). As can be expected, people younger than 24 years entering the labor market for the first time, with a high probability, have a lower education and this could partly explain their disadvantage with respect to older people.

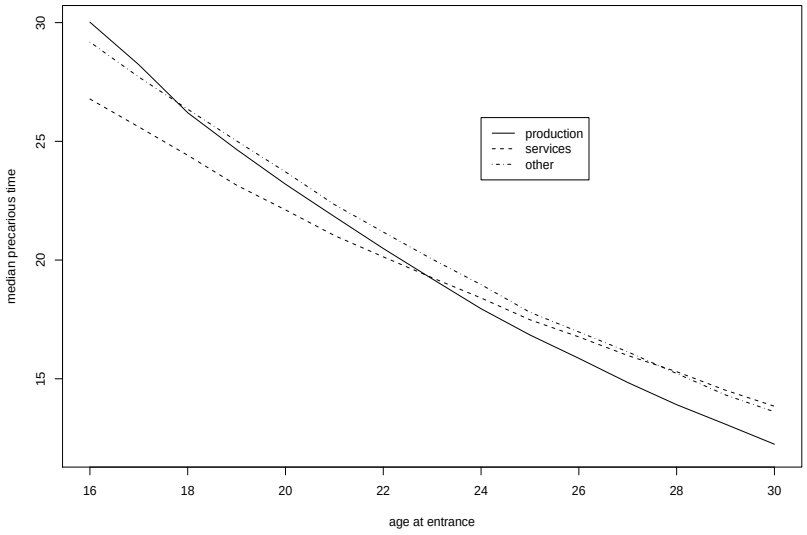
We finally note that exclusion of an important covariate, like education level, is not expected to bias the other estimates since unobserved heterogeneity is taken into account.

Larger firms tend to hire faster, and workers in the north are more likely to eventually find a stable job when compared to workers in the central and southern regions.

Despite all the political efforts, there are no changes over time periods with respect to the cure fraction (which can be marginally estimated to be approximately equal to 9.4%, as discussed above). On the contrary, our results show an increase over the second and the third period of the time to event, which indicates a worsening of the stabilization patterns over time. This is coherent also with the Italian economic conditions during the periods taken into consideration: we could cite many indicators but only mention that the Italian public debt has increased substantially over time, and that the youth unemployment rate has skyrocketed. The latter fact has made many young Italians entering the labor market more prone to accepting less advantageous conditions both from the economic and contractual aspect, and many firms prone to using unstable contracts as tools for reducing risks of firm mortality.

In order to better describe the interaction between sector and age on the time component, in Figure 1 we show the predicted medians (on the y-axis) as a function of age (on the x-axis) and sector, setting all the other covariates at baseline.

Figure 1: Predicted medians for non-cured individuals as a function of age and sector, all the other covariates set at baseline.



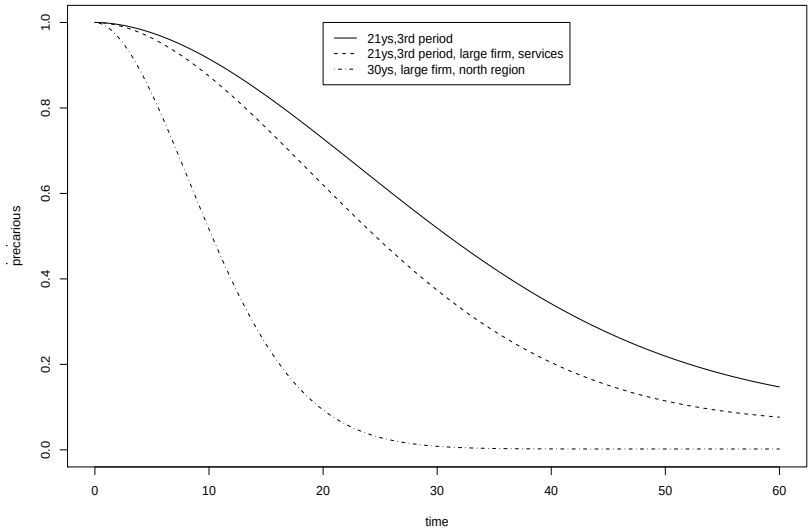
From Figure 1 it can be seen that the youngest new entrants are more penalized in the production and other sectors, while the effect is reversed for new entrants approximately after the twenty fourth year of age. Firms in the production sector can thus be suspected of penalizing less qualified workers but tend to hire qualified workers faster. The reverse seems to hold true for the services sector. This effect is subtle but easily explained, since for instance in services a larger supply of qualified workers, with respect to the production sector, is expected.

To give a further idea of the effect of the covariates, Figure 2 shows the estimated survival functions for different configurations. The covariates which are not mentioned are set at baseline.

It can be seen, for instance, that for older entrants in large production firms, in the northern regions of Italy, the survival functions decrease fast and practically all individuals are expected to eventually reach stability.

Figure 3 shows a histogram of $\Pr(Z_i = 1|\delta_i = 0, X)$, which indicates that for many of the censored individuals the posterior probability of being a long term survivor is estimated as very high. A waiting time of five years is then estimated as long enough for a new entrant to have an high probably of inclusion into the

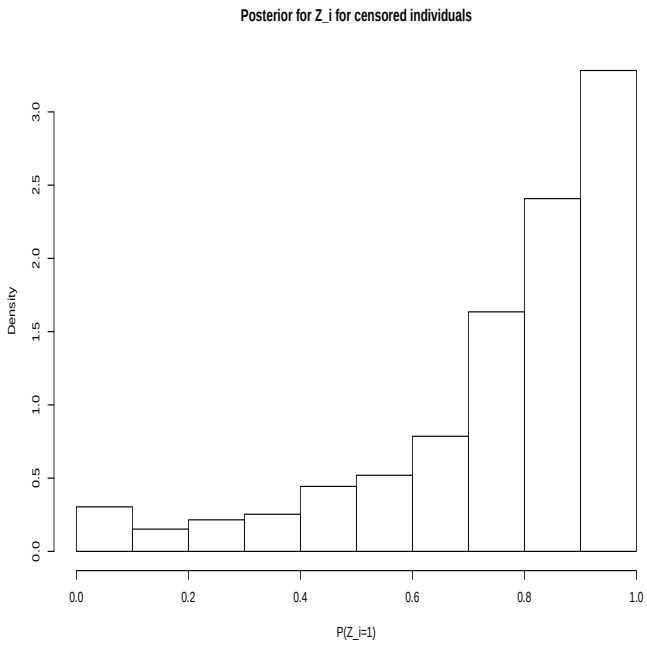
Figure 2: Estimated Survival Functions for Whip Data



cure fraction.

We also conducted a small sensitivity study with respect to the choice of prior parameters. Our simple sensitivity study consisted in choosing a grid of values for the prior parameters and fitting the model for each combination. We checked the posterior means, which were always very close and leading to the same conclusions: with such a large sample size there is practically no sensitivity to the choice of the prior parameters when fitting the model and when carrying out tests. Further, and not surprisingly, posterior means are always very close to maximum likelihood estimates for any reasonable choice of prior inputs. As a final graphical illustration, we report in Figure 4 the posterior distribution for γ obtained with different priors. We use three priors different in shape: a zero centered log normal with standard deviation equal to 2, a gamma distribution with parameters 1.3 and 1.1, which is more peaked and a bit translated to the right, and a gamma distribution with parameters 0.1 and 0.1. It can be seen that the three posterior distributions are very close, and the functional form of the log normal yields a slightly flatter posterior, which is to be preferred in our opinion. We have repeated the same study on the other parameters, obtaining similar results.

Figure 3: Posterior for $Z_i|\delta_i = 0$ for Whip Data



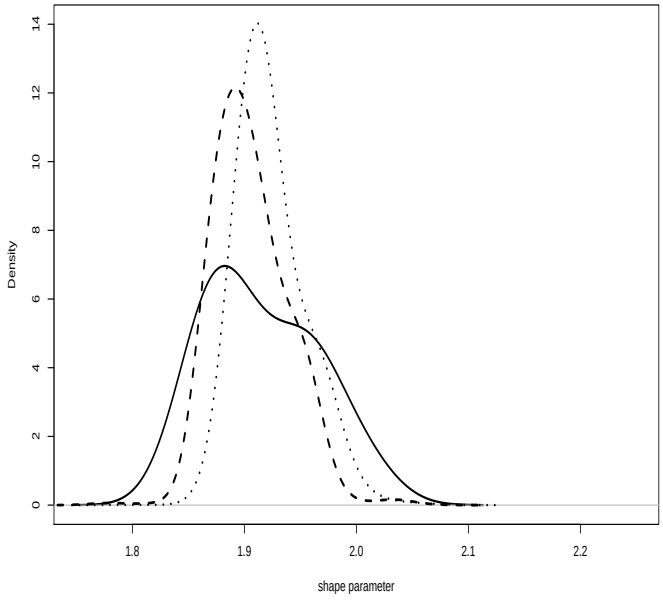
4.3 Conclusions

There have been political efforts in Italy on one hand to increase the number of new entrants in the labor market by incentivating flexible contracts, and on the other hand to decrease the time needed for achieving stability in the labor market (for instance by limiting the applicability of the same flexible contracts).

Our data suggest that, at least for our population, these latter efforts have made no change over time with respect to a fraction of individuals, which can be estimated as 9.4%, who are unlikely to ever reach stability in their working experience. We must be careful in interpreting the estimated cure fraction, considering that this fraction is composed also of discouraged workers who decide to exit the labor market and individuals who start self employment. We remark anyway that mortality of small firms in Italy is very high, so that even individuals who start self employment may often not experience a stable working life.

This fraction of individuals represents a partial failure of a policy based on contracts, like the CFL, which are explicitely intended as a *stepping stone* towards the stability of normal contracts. In this respect a discouraged worker who

Figure 4: Posterior for γ with three different priors: a log normal with parameters 0 and 2 (thick line), a gamma with parameters 1.3 and 1.1 (dashed line), and a gamma with parameters 0.1 and 0.1 (dotted line).



decides to exit the labor market, or even an individual who starts self employment is equivalent to an employed worker who is never granted a stable contract: all of them will never be able to rely on the employment benefits supplied by a stable contract and the social security protection guaranteed to standard workers. Further discussion on these issues can be found in Booth *et al.* (2002), and in Berton *et al.* (2009) for the Italian case.

Ustable contracts were introduced to allow on the job training before stabilization. Some employers may instead have used these contracts for different purposes, as a non negligible fraction of their employees may not reach stabilization. It is acknowledged that some workers may accept a less favourable contract at the beginning of their career if a sort of compensation in terms of chances of stabilization (and an higher wage) is foreseen in a near future. We have formally shown that for almost one out of every ten employees accepting an unstable contract (such as a CFL) this compensation may be neglected. Of course, this fraction is composed also of subjects who decide to start self employment, perhaps using

the CFL for training, subjects whose abilities are not sufficient for entering the labor market and the employer is unable to hire after a trial period. Finally, even for those individuals who are expected to eventually reach stability, from period to period a steady increase in the duration of the precarious part of working life has been detected.

Both features are expected to bring about instability also in the personal life of precarious workers, who must delay marriage, home buying, childbirth, etc.

Acknowledgements: The authors are grateful to three referees for suggestions. A. Farcomeni was supported by EIEF research grant "Advances in non linear panel models with socio economic applications".

References

Barbieri G., Sestito P. (2008) Temporary workers in Italy: who are they and where they end up? *Labour, Review of Labour Economics and Industrial Relations*, 22, 127–166.

Barbieri M.M., Berger J.O. (2004) Optimal Predictive Model Selection. *Annals of Statistics*, 32, 870–897.

Bell D.N.F. and Hart R.A. (2002) Working time in Great Britain, 1975-1994: evidence from the New Earnings survey panel data *Journal fo the Royal Statistical Society (Ser. A)*, 161, 327–348.

Bennett N.G., Bloom D.E. and Craig P.H. (1989) The divergence of Black and White Marriage Patterns. *American Journal of Sociology*, 95, 692–722.

Berton F., Richiardi M. and Sacchi S. (2009) *Flex-insecurity. Perché in Italia la flessibilitá diventa precarietá*, Il Mulino, Bologna.

Berton F., Devicienti F. and Pacelli L. (2008) Temporary jobs: Port of entry, Trap, or just Unobserved Heterogeneity? Working Paper 79 Laboratorio R. Revelli. Moncalieri (TO).

Boag J.W. (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society (Ser. B)*, 11, 15–44.

Booth A. L., Francesconi M., Frank J. (2002) Temporary Jobs: Stepping Stones or Dead Ends? *Economic Journal*, 112, 189–213.

Celeux G., Forbes F., Robert C.P. and Titterington D.M. (2006) Deviance Information Criteria for Missing Data Models *Bayesian Analysis*, 1, 651–674.

Chen K., Jin Z. and Ying Z. (2002) Semiparametric of transformation models with censored data. *Biometrika*, 89, 659–668.

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

Diebolt J., Robert C. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society (Ser. B)*, 56, 363–375.

Farewell V.T. (1982) The use of mixture models for the analysis of survival data with long term survivors. *Biometrics*, 38, 1041–1046.

Farcomeni A. (2010) Bayesian Constrained Variable Selection. *Statistica Sinica*, 20, 1043–1062.

George E.I., McCulloch R.E. (1993) Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88, 881–889.

George E.I., McCulloch R.E. (1997) Approaches for Bayesian Variable Selection. *Statistica Sinica*, 7, 339–373.

Gilks W.R. (1995) Discussion of "Fractional Bayes Factors for model comparison" *Journal of the Royal Statistical Society (Ser. B)*, 57, 99–138.

Gilks W.R., Best, N.G., Tan K.K.C. (1995) Adaptive rejection Metropolis sampling within Gibbs sampling (Corr: 97V46 p541-542 with Neal, R.M.). *Applied Statistics*, 44, 455–472.

Greenhouse J.B., Silliman N.P. (1996) Applications of a mixture survival model with covariates to the analysis of a depression prevention trial. *Statistics in Medicine*, 15, 2077–2094.

Hernes G. (1972) The process of entry into marriage. *American Sociological Review*, 37, 173-182.

ISTAT (2005) *Forze di lavoro media 2004*. Annuario statistico 2005, Rome.

Jeffreys, H. (1961) *Theory of Probability*, Oxford University Press, Oxford.

Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of failure time data* New York; John Wiley.

Kass, R. and Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Kiefer N.M. (1988) Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, 26, 646–679.

Koop G. (2003) *Bayesian Econometrics*. Wiley, Chichester.

Kuk A. Y. C., Chen C. H. (1992) A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79, 531–541.

Lancaster T. (1990) *The Econometric Analysis of Transition Data*. Cambridge, Cambridge University Press.

Li C.-S., Taylor J.M.G. and Sy J.P. (2001) Identifiability of cure models. *Statistics and Probability Letters*, 54, 389–395.

Li T., X. Zheng (2008) Semiparametric Bayesian Inference for Dynamic Tobit Panel Data Models with Unobserved Heterogeneity. *Journal of Applied Econometrics*, 23, 699–728.

McCullagh P. (2008) Sampling bias and logistic models *Journal of the Royal Statistical Sociery (Ser. B)*, 70, 643–677.

McVicar D. and Anyadike-Danes M. (2002) Predicting successful and unsuccessful transitions from school to work by using sequence methods *Journal of the Royal Statistical Society (Ser. A)*, 165, 317–334.

O'Hagan A. (1995) Fractional Bayes Factors for model comparison *Journal of the Royal Statistical Society (Ser. B)*, 57, 99–138.

O'Hagan A. (1997) Properties of intrinsic and fractional Bayes Factors *TEST*, 6, 101–118.

Olave P. and Salvador M. (2007) Semi-parametric Bayesian analysis of the proportional hazard rate model: an application to the effect of training programs on graduate unemployment. *Applied Statistics*, 34, 1185–1205.

Picchio M. (2008) Temporary contracts and transitions to stable jobs in Italy *Labour, Review of Labour Economics and Industrial Relations*, 22 (Special Issue) 147-174.

Price D.L., Manatunga A.K. (2001) Modeling survival data with a cured fraction using frailty models. *Statistics in Medicine*, 20, 1515–1527.

Robert C.P., Casella G. (2000) *Monte Carlo Statistical Methods*. Springer, New York.

Schwarz G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6, 441–464.

Sherman R., Ho Y. and Dalal S. (1999) Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling. *Econometrics Journal*, 2, 248–267.

Spiegelhalter D.J., Best N.G., Carlin B.P. and van der Linde A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Sociery (Ser. B)*, 64, 583–640.

Wei G. and Tanner M. (1990) A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699–704.

A An EM Algorithm for obtaining the Maximum Likelihood

In order to maximize the likelihood, we set up an Expectation-Maximization (EM) algorithm. The EM algorithm is an iterative procedure which alternates the following steps until convergence:

- **E-step**: compute the conditional expected value of the complete data log-likelihood given the current estimate of the parameter vector and the observed data;
- **M-step**: maximize the expected value above with respect to the parameter vector.

We now describe appropriate E-steps and M-steps, according to specific model formulations.

A.1 E-step for a Model without random effects

In our model the expected complete log-likelihood is obtained by replacing the latent variables Z_i in (2) with their conditional expectations given the current parameter estimates. The conditional expectation of Z_i, which we denote with $\tilde{z}_i$, coincides exactly with the right hand side of expression (5) when $\delta_i = 0$, and is equal to 1 when $\delta_i = 1$. We denote the expected complete log-likelihood by $\tilde{l}(\theta)$.

A.2 E-step for a Model with random effects

When there are random intercepts, these must be integrated out. To do so, we set up a Monte Carlo E-Step by jointly drawing a random sample from the full conditional distribution $\pi(\beta_{0i}, \alpha_{0i}|\theta, \alpha, Y)$.

Denote this sample by $(\alpha_{0i}^1, \beta_{0i}^1), \ldots, (\alpha_{0i}^B, \beta_{0i}^B)$. Then, for all i such that $\delta_i = 0$, we have that

$$\tilde{z}_i = \frac{1}{B} \sum_{t=1}^{B} \frac{p_i(\alpha_{0i}^t)}{p_i(\alpha_{0i}^t) + (1 - p_i(\alpha_{0i}^t)) S(t_i | X_i, \beta_0^t, \theta, Z_i = 0)},$$

where we use the notation $p_i(\alpha_{0i}^t)$ to indicate that p_i is recomputed as a function of α_{0i}^t. For all i such that $\delta_i = 1$ we still have that $\tilde{z}_i = 1$.

A.3 M-step for a Model with no Covariates

If there are no covariates involved, the complete log-likelihood can be easily maximized. We can maximize the expected complete likelihood by equating its partial derivatives to zero. We must then solve the following system of equations:

$$\begin{cases} \frac{\partial \tilde{l}(\theta)}{\partial \gamma} = 0 \\ \frac{\partial \tilde{l}(\theta)}{\partial \mu} = 0 \\ \frac{\partial \tilde{l}(\theta)}{\partial p} = \frac{\sum \tilde{z}_j}{p} - \frac{\sum (1 - \tilde{z}_j)}{1-p} = 0. \end{cases}$$

We immediately get the explicit expression $\hat{p} = \frac{\sum \tilde{z}_j}{n}$. If a lognormal model is used for the time component, the system given by the first two equations must be solved numerically, or better the complete log-likelihood can be maximized using an iterative maximization algorithm of Newton-Raphson type.

In case a Weibull model is used for the first component, we obtain the following explicit expressions:

$$\begin{cases} \frac{\partial \tilde{l}(\theta)}{\partial \mu} = \frac{\gamma}{\mu} \sum_j (1 - \tilde{z}_j) \delta_j - \gamma \mu^{\gamma - 1} \sum_j (1 - \tilde{z}_j) t_j^{\gamma} = 0 \\ \frac{\partial \tilde{l}(\theta)}{\partial \gamma} = \frac{\sum_j \delta_j (1 - \tilde{z}_j)}{\gamma} + \sum_j \delta_j (1 - \tilde{z}_j) \log(t_j) - \mu^{\gamma} \sum_j (1 - \tilde{z}_j) t_j^{\gamma} (\log(\mu) + \log(t_j)) = 0 \end{cases}$$

For the first equation, we can then compute the following explicit expression for μ as a function of γ:

$$\mu = \left(\frac{\sum_j (1 - \tilde{z}_j) \delta_j}{\sum_j (1 - \tilde{z}_j) t_j^{\gamma}} \right)^{\frac{1}{\gamma}} \tag{9}$$

Now (9) can be plugged in the second equation, which is nonlinear in γ but can be solved with a simple secant method. The estimated value $\hat{\gamma}$ is finally plugged in (9) to obtain $\hat{\mu}$.

A similar strategy can be used if an exponential distribution is assumed for the time component, since that case corresponds to setting $\gamma = 1$.

A.4 M-step for a Model with covariates

When covariates are involved in the model, for simplicity we can split the complete log-likelihood in two summands, which are conveniently maximized separately: one part for the vector of parameters in the time component, and one for the α parameters: $\tilde{l}(\theta) = \tilde{l}_1(\beta, \gamma) + \tilde{l}_2(\alpha)$, where

$$\begin{aligned} \tilde{l}_1(\beta,\gamma) &= \sum (1-\tilde{z}_j)\log(L_{0j}(\theta)) && (10)\\ \tilde{l}_2(\alpha) &= -\sum (1-\tilde{z}_j)\log(1+\exp\{\alpha'\mathbf{X}_j\}) && (11)\\ &\quad +\sum \tilde{z}_j(\alpha'\mathbf{X}_j - \log(1+\exp\{\alpha'\mathbf{X}_j\})), \end{aligned}$$

where $L_{0j}(\theta)$ denotes the complete likelihood for the j-th observation when $Z_j = 0$. To maximize $\tilde{l}_2(\alpha)$ we can use a standard iterative algorithm of Newton-Raphson type for logit models.

In order to set up a Newton-Raphson algorithm to maximize $\tilde{l}_1(\beta, \gamma)$ we need the first and second derivatives with respect to each parameter. The first derivatives with respect to each parameter are given by

$$\begin{cases} \frac{\partial l_1(\beta,\gamma)}{\partial \gamma} = -\sum (1-\tilde{z}_j) t_j^\gamma \exp(\gamma\beta'\mathbf{X}_j)(\log(t_j)+\beta'\mathbf{X}_j) + \sum (1-\tilde{z}_j)\delta_j(\log(t_j)+\beta'\mathbf{X}_j) \\ \qquad\qquad + \frac{\sum (1-\tilde{z}_j)\delta_j}{\gamma} \\ \frac{\partial l_1(\beta,\gamma)}{\partial \beta_h} = -\gamma \sum (1-\tilde{z}_j) t_j^\gamma \exp(\gamma\beta'\mathbf{X}_j) x_{jh} + \gamma \sum (1-\tilde{z}_j)\delta_j x_{jh} \qquad h = 1,\ldots,d; \end{cases}$$

where d is the length of vector β. The second derivatives with respect to each parameter are given by

$$\begin{cases} \frac{\partial l_1(\beta,\gamma)}{\partial \gamma^2} = -\sum (1-\tilde{z}_j) t_j^\gamma \exp(\gamma\beta'\mathbf{X}_j)(\log(t_j)+\beta'\mathbf{X}_j)^2 - \frac{\sum (1-\tilde{z}_j)\delta_j}{\gamma^2} \\ \frac{\partial l_1(\beta,\gamma)}{\partial \gamma \partial \beta_h} = -\sum (1-\tilde{z}_j) t_j^\gamma \exp(\gamma\beta'\mathbf{X}_j)(\gamma\log(t_j)+\gamma\beta'\mathbf{X}_j+1) x_{jh} + \sum (1-\tilde{z}_j)\delta_j x_{jh} \quad h = 1,\ldots,p; \\ \frac{\partial l_1(\beta,\gamma)}{\partial \beta_h \beta_r} = -\gamma \sum (1-\tilde{z}_j) t_j^\gamma \exp(\gamma\beta'\mathbf{X}_j) x_{jh} x_{jr} \qquad h = 1,\ldots,d; r = 1,\ldots,d. \end{cases}$$

A.5 M-step for a Model with covariates and random effects

If random effects are included, the Monte Carlo M-step is not much different from the M-step just set up. In pratice one can modify (10) and (11) as

$$\begin{aligned}
\tilde{l}_1(\beta,\gamma) &= \frac{1}{B}\sum_t\sum_j (1-\tilde{z}_j)\log(L_{0j}(\theta,\beta_0^t)) \\
\tilde{l}_2(\alpha) &= -\frac{1}{B}\sum_t\sum_j (1-\tilde{z}_j)\log(1+\exp\{\alpha_{0j}^t+\alpha'\mathbf{X}_j\}) \\
&\quad +\frac{1}{B}\sum_t\sum_j \tilde{z}_j(\alpha_{0j}^t+\alpha'\mathbf{X}_j-\log(1+\exp\{\alpha_{0j}^t+\alpha'\mathbf{X}_j\})),
\end{aligned}$$

which lead to similar Newton-Raphson algorithms.