Bayesian cure-models with a logistic reparameterization, with application to Italian labor market analysis

Modelli Bayesiani per lungo sopravviventi con riparametrizzazione logistica e loro applicazione all'analisi del mercato del lavoro in Italia

Alessio Farcomeni[1], Alessandra Nardi[2], Elena Fabrizi[3]

[1] Dipartimento di Medicina Sperimentale, Sapienza - Università di Roma
e-mail: alessio.farcomeni@uniroma1.it

[2] Dipartimento di Matematica, Università di Roma - Tor Vergata

[3] Dipartimento di Contabilità Nazionale e Analisi dei Processi Sociali, Sapienza - Università di Roma

Riassunto: Per l'analisi della sopravvivenza in quelle popolazioni in cui una frazione di individui potrebbe non sperimentare mai l'evento in esame proponiamo un modello Bayesiano parametrico basato su una mistura a due componenti, di cui una degenere. Sia la distribuzione non degenere che la frazione di lungo sopravviventi possono dipendere da un insieme di variabili esplicative. Vengono affrontati problemi di test di ipotesi e scelta del modello. La proposta é illustrata utilizzando uno studio sul fenomeno della precarietà in Italia. Si dimostra che una percentuale costante nel tempo di persone molto probabilmente rimarrà precaria a vita, in particolare al centro e al sud.

Keywords: bayesian methods, mixture cure-models, model selection, precariousness, variable selection

1. Introduction

Mixture cure models (Boag, 1949; Farewell, 1982) are widely used to model survival data with a cure fraction, i.e., survival data in which a fraction of individuals are expected not to experience the event of interest. They allow the survival function of uncured patients and the cure fraction to depend on covariates. Price and Manatunga (2001) have extended mixture cure models for the use of frailty to capture unobserved heterogeneity and Morbiducci *et al.* (2003) introduce classification approaches and diagnostic tools for cure models. While the use of cure models is widely accepted in the medical literature, there still are open issues and possible extensions to different areas of application. First of all, test theory is not developed in full generality. In particular care is needed when testing for a zero cure fraction since the null hypothesis is on the boundary of the parameter space, and then the usual likelihood ratio test statistic is not distributed like a chi-square. Secondly, common stepwise variable selection methods are not developed even in the simplest cases. Greenhouse and Silliman (1996) tackle the problems, using Schwarz criteria to approximate Bayes factors. They note anyway that Schwarz criteria are more apt for large sample situations. Further, in order to perform variable selection they enumerate the model space, so that their method cannot be considered automatic and it may become unfeasible when the number of covariates is large. Model selection, testing, and variable selection are the focus of this work. We develop strategies in a Bayesian framework for a mixture cure model in which the

cure fraction is reparameterized using a logistic transform. Bayes factors will be well approximated in all settings, even for moderate sample sizes, and variable selection will be performed with an efficient strategy no matter the number of covariates involved. Note that all the techniques we propose are (frequentist) consistent. Furthermore, as a by-product, in the Bayesian framework unobserved heterogeneity is automatically taken into account. This is particularly important when estimating the cure fraction p. The rest of the paper is as follows: in Section 2 we present the cure model, derive the complete likelihood and discuss model fitting. In Section 3 we show how to derive approximate Bayes factors for testing hypotheses on the parameters at stake by extending optimal ratio importance sampling (Chen and Shao, 1997) to the case of general mixture models. In Section 4 we discuss an automatic approach to variable selection. Finally, in Section 5 we show an application to investigate the evolution of precariousness in the Italian labor market. The application is based on microeconomic longitudinal data on work histories.

2. Mixture cure models

Suppose we observe time to the event of interest for n independent subjects, and 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 constant used to include the intercept) and by $S(t|X = x) = \int_t^{+\infty} f(s|X = x)\, ds$ the survival function of T. The general mixture cure-rate model can be easily obtained introducing an unobserved latent indicator Z_i which is equal to 1 when the i-th individual is deemed not to experience the event. In the simpler case of no covariates affecting the latent indicator, a Bernoulli model can be used: $P(Z_i = 1) = p$, where p is the *cure fraction*. Time to event can be modeled parametrically or non-parametrically. While the distribution-free approach can be of interest in different settings, in the mixture case it might become overly adaptive, possibly compressing the cure fraction into one single component. Further, when using a proportional hazards model, the presence of a cure fraction is often translated into a violation of proportionality. This problem can not be overcome with the usual methods and extensions of the semiparametric approaches.

For these reasons in this work we prefer a parametric approach. Among the many possible families of distributions that can be assumed, we recall the Weibull and the Gamma distributions that assume a monotone hazard function and the lognormal and the log-logistic distributions that allow a non monotone, peaked hazard. Although our Bayesian approach can be used in any of these cases, we will focus on the Weibull distribution. The corresponding survival and density functions are:

$$S(t|X = x, Z = 0) = \exp\{-(\mu t)^\gamma\}$$
$$f(t|X = x, Z = 0) = \mu\gamma(\mu t)^{\gamma-1} \exp\{-(\mu t)^\gamma\}$$

In the following we use the short-hand notation of θ for the vector of parameters characterizing $S(t|X = x, Z = 0)$. We have

$$S(t|X = x) = S(t|X = x, Z = 1)p + (1-p)S(t|X = x, Z = 0) = p + (1-p)S(t|X = x, Z = 0),$$

since by assumption $S(t|X = x, Z = 1) = 1$ for any $t \geq 0$.

The complete data likelihood can be then written as

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

where $f(t_i|\theta, X_i, Z_i = 0)$ and $S(t_i|\theta, X_i, Z_i = 0)$ denote the density and the survival function of the Weibull family. Note that the effect of X_i on the risk of failure is modeled by reparametrizing the scale parameter of the Weibull distribution 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 $e^{-\beta' x_i}$. In many cases the same or a different set of covariates, which we denote by $\tilde{X}$, may be assumed to affect the probability of being long survivor. This can be modeled using a logistic transform and setting $P(Z_i = 1|\tilde{X}_i) = (1 + \exp\{-\alpha' \tilde{X}_i\})^{-1}$. In this case one shall only substitute p by its reparameterization in the complete data log-likelihood.

In order to make inference in a Bayesian framework we must specify opportune priors $\pi(\cdot)$ for the parameters at stake. In the presence of prior information this shall be summarized and properly taken into account when choosing the priors. The general approach we propose is to use independent zero-centered normal priors for the α and β regression parameters, and log-normals or inverse gammas for the shape parameter γ.

2.1. Model Fit

The model can be fit adapting an ad-hoc Gibbs sampling scheme for mixture models (Diebolt and Robert, 1994; Robert and Casella, 2000). In practice one shall sample the latent indicators from their posteriors, and then use the complete likelihood conditionally on the latent indicators. The general iteration of the Gibbs sampling scheme we propose for mixture cure models as defined in the previous section is as follows:

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

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

   where θ is 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 α from

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

   Compute $p = \{p_i = (1 + \exp -\alpha' \tilde{X}_i)^{-1}\}$.
3. Sample θ from

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

In order to avoid difficulties linked with setting up Metropolis Hastings, we sample the parameters in θ simultaneously with Adaptive Rejection Metropolis Sampling (Gilks *et al.*, 1995).

Sampling from $\pi(\alpha|Z)$ can be once again performed via an ARMS, even if there are many different common alternative approaches for this standard problem. We finally note that in certain cases one may not wish to let the cure fraction p depend on covariates. Then, a $Beta < \gamma_1, \gamma_2 >$ prior can be assumed for p. The full conditional distribution at Step 2 of the MCMC algorithm would then simply be a Beta with parameters $\gamma_1 + \sum_{i=1}^{n} Z_i$ and $\gamma_2 + \sum_{i=1}^{n}(1 - Z_i)$.

3. Testing and Model Selection

Testing in the Bayesian framework is usually performed through Bayes Factors (BF). Tests 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.

A BF is the ratio between the marginal densities under H_0 and under the alternative hypothesis H_1; that is the normalizing constants of the posterior distributions computed respectively under each model. Roughly, a BF greater than 1 reveals the data provide greater evidence in favor of H_0 than H_1, and the opposite conclusion holds otherwise.

Since the marginals have cumbersome or no closed form expressions in mixture cure models, we must approximate their ratio by an appropriate technique. Approximation of normalizing constants is a very common problem in Bayesian statistics, which here is made harder by the mixture structure in the likelihood.

A procedure especially devised for computing BF in mixture models is in Steele *et al.* (2006), but in practice it requires that the integrals with respect to α, β and γ parameters be solved analytically. Greenhouse and Silliman (1996) develop Schwarz criteria specifically for mixture cure models, but their approximation is satisfactory only for large samples. Farcomeni (2008) extends optimal ratio importance sampling (Chen and Shao, 1997) to general mixture models. In this paper we apply his method to mixture cure-models. The resulting *mixture optimal ratio importance sampling* can be used in general, does not require the posterior distribution to be approximately elliptically contoured, and avoids the problem of label switching. Optimal ratio importance sampling proceeds by first approximating an optimal constant, which we denote by τ, used to tune a proposal distribution from which to draw for importance sampling. The sample from the optimal proposal distribution is used for approximating the BF in the natural way.

The mixture extension simply recognizes that there shall be one optimal constant and one proposal distribution for each configuration of the vector latent variables Z. Since the number of possible configurations of Z is 2^n, which can be very large, also the vectors Z shall be sampled simultaneously with draws from the specific optimal proposal distribution.

The final algorithm runs as follows:

1. Let $\lambda(\phi)$ be an arbitrary density over the union of the support of the two priors (under H_0 and under H_1), where $\phi = \{\theta, \alpha\}$. Let $p_i((\theta, \alpha)|T, Z)$ be the non normalized posterior from the i-th model, $i = 0, 1$.
2. Given a random draw $\phi_1, \ldots, \phi_n$ from $\lambda(\cdot)$, define
$$\tau(Z) = \frac{\sum p_0(\phi_i|T, Z)/\lambda(\phi_i)}{\sum p_1(\phi_i|T, Z)/\lambda(\phi_i)}$$
3. Let the proposal distribution be $\psi(\cdot|T, Z) = |p_0(\cdot|T, Z) - \tau(Z)p_1(\cdot|T, Z)|$.

4. With the approach proposed in Section 2.1 sample $(\phi_1, Z_1), \ldots, (\phi_n, Z_n)$ from $\psi(\cdot|T, Z)$.
5. Define the approximate BF of H_0 against H_1 as

$$B_{01} = \frac{\sum p_0(\phi_i|T, Z_i)/\psi(\phi_i|T, Z_i)}{\sum p_1(\phi_i|T, Z_i)/\psi(\phi_i|T, Z_i)}$$

If H_0 is nested in H_1, as often happens, $\lambda(\phi)$ is simply defined on the space of the largest model. Note that sampling from any of the two posteriors is not needed. Thus, the method can be efficiently implemented before-hand and only the chosen model may be fit. The proposed scheme closely follows usual optimal ratio importance sampling, with the only difference that the non normalized posteriors are substituted with complete posteriors. The optimal tuning parameter τ is now a function of Z, but this does not pose any problem while sampling.

4. Variable Selection

Bayes factors can be used to test hypotheses on regression parameters. Whenever data provide more evidence in favor of $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ we can safely 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 safely conclude that there is no effect of the j-th covariate on the probability of being cured. The number of Bayes factors involved in such a procedure, nevertheless, can be formidable even in small dimensional situations; especially when one wants to test for two-way or even high order interactions, or for the possibility of including transformed continuous variables. There are many automatic approaches to variable selection that can be implemented for mixture cure models. We will adapt here a version of the Stochastic Search Variable Selection (SSVS) of George and McCulloch (1993). When considering the inclusion of interactions or transformations, we will impose hierarchical constraints on the final model. That is, a two-way interaction can be included only if the original variables are included in the model as well. Similarly, we can include the cube of a continuous variable only if both the quadratic and the linear effect are included too, and so on. To do this, we will exploit the approach of Farcomeni (2007), who adapts SSVS to structured variable selection.

Variable selection can be performed by using a two component Gaussian mixture as prior for the regression parameters. The mixture is indicized by a latent indicator variable η_j, which flags the corresponding covariate as in or out of the model. We have:

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

with τ_{1j} slightly larger than τ_{0j}. Whenever $\eta_j = 1$, the usual Gaussian prior is used for β_j and the j-th covariate is included. When $\eta_j = 0$, τ_{0j}^2 is set small enough that the prior mass is concentrated around zero, and the j-th covariate is therefore not included in the model. 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 two original variables are zero. This would result in intractable priors. Farcomeni (2007) 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. Constraint indicators are linked to variable indicators with the following reparameterization:

$$\eta_j = \left(\prod_{j \neq i} \eta_i^{\omega_i(j)} \right).$$

A Bernoulli prior is put on each η_j: $P(\eta_j = 1) = w_j$. Model selection is finally performed by examining the posterior $\pi(\eta_j | T, X)$. Farcomeni (2007) proves that frequentist consistency in model selection is achieved using the median model (Barbieri and Berger, 2004), that is, by inclusion of variables for which $\pi(\eta_j | T, X) > 0.5$. The inclusion of latent indicators η simply adds one step to the Gibbs sampling approach of Section 2.1, in which indicators are sampled from a Bernoulli with the opportune parameter. Let η_{-j} denote the vector of η indicators with the j-th component excluded. It can be proved that

$$P(\eta_j = 1 | \theta, \eta_{-j}, T, X, Z) = \frac{w_j \pi(\beta | \eta_{-j}, \eta_j = 1)}{w_j \pi(\beta | \eta_{-j}, \eta_j = 1) + (1 - w_j) \pi(\beta | \eta_{-j}, \eta_j = 0)},$$

and similarly for the α parameters. Surprisingly enough in fact there is conditional independence between the data and η (George and McCulloch, 1993).

5. Application to WHIP Data

Data come from the Work Histories Italian Panel (WHIP), an employer-employee linked panel database developed by Italian Social Security administrative source (INPS), according to Laboratorio R. Revelli (TO). The reference population is made up of all the individuals who have worked in Italy during the period from January 1985 to December 2003, both Italians and immigrants, aged 16-29 and having first registered experience of work in the private sector; self-employed together with agricultural and public administration workers are excluded. In particular we focus only on new entrants with training on the job (CFL) or apprenticeship contract. The flexible part of the labor force, during the period we are studing, is focused on these contracts, which are fixed term and have contribution rebates. These contribution rebates were introduced in order to give more employment opportunities to underprivileged people out of labor market, and especially to young people without any work experience. Another reason behind these kind of contracts is to stimulate emerging of workers from the shadow economy.

In our analysis we do not include two new kinds of flexible contracts that were introduced at the end of the '90s: temporary agency work and dependent self employment (Co.Co.Co and similar). These are heterogeneous contract forms with respect to the more focused forms that were available also in 1986 and in 1992. We note also that temporary agency contracts are limited for our reference population (less then 1% in 1998). Also, by excluding dependent self employment we are further biasing downwards our estimates about precariusness for new entrances in 1998.

We record the time from the beginning of the first work experience to a removal of all the benefits, if it 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 even be able to get a normal contract.

We underline that benefits may be removed but the job still be a fixed term contract, so that we should even be biasing downwards our estimates of 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.

The measured covariates are: sex, age (minus sixteen), average number of employees of the firm (on log-scale), region (north, center, south), kind of firm (production of goods, services, other). Since the follow-up for the three cohorts are non overlapping, the period is treated as a further covariate.

First, we test the null hypothesis $H_0 : p = 0$ against $H_1 : p > 0$; the BF is around -73 on a log-scale and the likelihood ratio statistic is 9641.48. Data then show overwhelming evidence in favor of the presence of a cure fraction of subjects who will *never* find a stable job. This cure-fraction is estimated as 9.2% (95%HPD: 8.1%-10.2%), when no covariates are included in the model. Secondly, we perform model selection, considering the possibility of hierarchically including two-way interactions. We create a dummy for each level (except one) of each categorical variable. Since we are also interested in equivalence of categories, we let SSVS select the dummies separately. For instance, the dummy for production of goods can be included without the dummy for services, for the variable kind of firm. We note that this enlarges the model space, making variable selection harder.

The median model, together with final estimates of the acceleration factors and logistic regression parameters, is shown in Table 1.

Table 1: *WHIP Data: Median model with parameter estimates, lower and upper 95% credibility intervals and probability of inclusion of covariates.*

	Parameter	$HPD^{low}_{0.95}$	$HPD^{up}_{0.95}$	$P(\eta_j) = 1$
Time component				
γ	1.763	1.722	1.793	
$e^{-\beta_0}$	0.022	0.019	0.024	
third time period	1.214	1.164	1.237	0.765
age	0.935	0.929	0.942	0.794
# employees	0.892	0.870	0.908	0.854
services	0.891	0.870	0.904	0.811
age*services	1.014	1.009	1.021	0.608
p component				
α_0	1.174	0.998	1.291	
age	0.277	0.220	0.317	0.966
region: North	0.314	0.272	0.373	0.875
services	0.344	0.287	0.399	0.899

It is more likely to eventually find a stable job in the communications and other services with respect to people working in the production or other sectors, and with lower waiting times. All the same, elder subjects both have lower time to event and cure-fraction. This could be expected, since people entering the market later are usually more qualified. Larger firms tend to hire faster, and workers in the north are more likely to eventually find a stable job. Despite all the political efforts, there is no change over time period with

respect to the cure fraction (which can be marginally estimated as 9.2% as said) and even an increase over the third period with respect to time to event. SSVS allows us to also evaluate the probability that a variable is important for the model. For instance, we can safely conclude there is no gender effect since its probability of inclusion is estimated as 0.12 for the Weibull and 0.38 for the cure-fraction. As a final analysis, we conducted a small sensitivity study with respect to choice of prior parameters. With such large a sample size there practically is no sensitivity to choice of prior parameters when fitting the model and when doing tests. There instead is some sensitivity when doing model choice, a well known drawback of SSVS (see for instance George and McCulloch, 1997; Farcomeni, 2007).

References

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

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.

Chen M-H, Shao Q-M (1997) On Monte Carlo methods for estimating ratios of normalizing constants. *Annals of Statistics*, 25, 1563–1594.

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. (2007) Bayesian Constrained Variable Selection. Tech. Rep. 21, Dipartimento di Statistica, Probabilità e Statistiche Applicate Sapienza - Università di Roma.

Farcomeni A. (2008) Optimal Ratio Importance Sampling for Mixture Models.

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., 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.

Morbiducci M., Nardi A., Rossi C. (2003) Classification of “cured” individuals in survival analysis: the mixture approach to the diagnostic-prognostic problem. *Computational Statistics and Data Analysis*, 41, 515–529.

Price D.L., Manatunga A.K. (2001) Modelling 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.

Steele R.J, Raftery A.E., Emond M.J. (2006) Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *Journal of Computational and Graphical Statistics*, 15, 712–734.