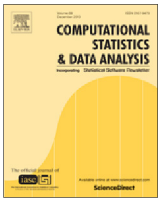
Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda



Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines[☆]

Gressani Oswaldo[a,*], Philippe Lambert[a,b]

[a] *Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Voie du Roman Pays 20, B-1348, Louvain-la-Neuve, Belgium*

[b] *Faculté des Sciences Sociales, Méthodes Quantitatives en Sciences Sociales, Université de Liège, Place des Orateurs 3, B-4000, Liège, Belgium*

ARTICLE INFO

Article history:
Received 27 September 2017
Received in revised form 14 February 2018
Accepted 24 February 2018
Available online xxxx

Keywords:
Promotion time cure model
P-splines
Survival analysis
Laplace approximation
Approximate Bayesian inference

ABSTRACT

Bayesian methods for flexible time-to-event models usually rely on the theory of Markov chain Monte Carlo (MCMC) to sample from posterior distributions and perform statistical inference. These techniques are often plagued by several potential issues such as high posterior correlation between parameters, slow chain convergence and foremost a strong computational cost. A novel methodology is proposed to overcome the inconvenient facets intrinsic to MCMC sampling with the major advantage that posterior distributions of latent variables can rapidly be approximated with a high level of accuracy. This can be achieved by exploiting the synergy between Laplace's method for posterior approximations and P-splines, a flexible tool for nonparametric modeling. The methodology is developed in the class of cure survival models, a useful extension of standard time-to-event models where it is assumed that an unknown proportion of unidentified (cured) units will never experience the monitored event. An attractive feature of this new approach is that point estimators and credible intervals can be straightforwardly constructed even for complex functionals of latent model variables. The properties of the proposed methodology are evaluated using simulations and illustrated on two real datasets. The fast computational speed and accurate results suggest that the combination of P-splines and Laplace approximations can be considered as a serious competitor of MCMC to make inference in semi-parametric models, as illustrated on survival models with a cure fraction.

1. Introduction

There is a growing interest for cure rate models in survival analysis as witnessed by the number of recently published papers on that topic in statistical journals. These models have gained in popularity as they intrinsically account for long-term survivors that will never experience the event of interest even when followed-up for an extended time period. The promotion time (cure) model introduced by Yakovlev et al. (1996) is motivated by cancer tumor kinetics, the biological mechanism underlying the proliferation and growth of carcinogenic cells. In particular, let $N \sim \mathit{Poisson}\big(\phi(\mathbf{x})\big)$ be the number of carcinogenic cells affecting a given subject with mean $\phi(\mathbf{x}) = \exp(\beta_0 + \mathbf{x}^T \boldsymbol{\beta})$. To the ith cell is associated a latent event time $T_i \geq 0$ representing the duration necessary for the cell to grow to a detectable tumor mass. Latent event times $\{T_1, \ldots, T_N\}$ are assumed to be independently and identically distributed with common cumulative distribution function $F(t)$ and the

[☆] R code is available as supplementary material in the electronic version of the paper.

[*] Corresponding author.
E-mail addresses: oswaldo_gressani@hotmail.fr (G. Oswaldo), p.lambert@ulg.ac.be (P. Lambert).

https://doi.org/10.1016/j.csda.2018.02.007
0167-9473/© 2018 Elsevier B.V. All rights reserved.

Please cite this article in press as: Oswaldo G., Lambert P., Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.02.007.

observed survival time is defined as $T = \min\{T_1, \dots, T_N\}$. When a Cox proportional hazards model (Cox, 1972) is used to model the N conditional latent distributions $F(t_i|\mathbf{z}) = 1 - S_0(t_i)^{\exp(\mathbf{z}^T\boldsymbol{\gamma})}$, $i = 1, \dots, N$ one can show that the resulting survival function of T is given by (Tsodikov, 1998; Chen et al., 1999)

$$\begin{aligned} S_p(t|\mathbf{x}, \mathbf{z}) &= \exp\big(-\phi(\mathbf{x})F(t|\mathbf{z})\big) \\ &= \exp\Big(-\exp(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})\big(1 - S_0(t)^{\exp(\mathbf{z}^T\boldsymbol{\gamma})}\big)\Big). \end{aligned} \qquad (1)$$

In this model, a subject is cured when $N = 0$, an event arising with a probability given by $P(N = 0 \,|\mathbf{x}, \mathbf{z}) = \lim_{t\to\infty} S_p(t|\mathbf{x}, \mathbf{z}) = \exp(-\phi(\mathbf{x}))$. Alternative specifications are proposed in the literature to model the distribution of latent event times $F(t_i)$, for example Ibrahim et al. (2001) propose a semi-parametric form for the latent distribution involving a smoothing parameter controlling the degree of parametricity in the right tail of the population survival function, while Zeng et al. (2006) introduce a semi-parametric class of cure models taking into account a subject-specific frailty. Model (1) can be estimated by maximum likelihood methods in a frequentist setting (see Tsodikov, 2002, 2003). From a Bayesian perspective, Yin and Ibrahim (2005) assume a piecewise exponential model for the baseline survival function with a tradeoff between model flexibility and the number of partitions of the time axis. More recently, Bremhorst and Lambert (2016) use a large number of B-splines to specify the baseline hazard and, following Eilers and Marx (1996), counterbalance the flexibility of the model by using a roughness penalty based on finite differences of adjacent B-spline coefficients.

The rather complex structure of the posterior distributions in the latter Bayesian frameworks requires the use of MCMC techniques. For such models, the MCMC toolbox is usually accompanied by a large computational burden and challenging convergence problems under the original parameterization. A crucial component explaining the inefficiency of rejection sampling techniques is a strong posterior correlation appearing firstly among latent variables and secondly between latent variables and hyperparameters of the model, thus having a global impact on convergence speed and autocorrelation. Integrated Nested Laplace Approximations (INLA) is a sampling-free Bayesian methodology recently introduced in the literature that allows to obtain marginal posteriors in the class of latent Gaussian models and has been recognized to be an interesting alternative to standard MCMC methods. In this dimension, Rue et al. (2009) and Martino (2007) are the pioneering references showing how to perform approximate Bayesian inference in latent Gaussian models via Laplace approximations.

While INLA has been shown to work well in a large variety of applications like stochastic volatility models (Martino et al., 2011a), generalized dynamic linear models Ruiz-Cárdenas et al. (2012) and spatio-temporal disease mapping models (Schrödle and Held, 2011), there seems to be little work related to survival analysis or penalized B-spline models. Among the contributions on the subject, we can cite Fong et al. (2010) who combine INLA and O'Sullivan splines in a nonparametric smoothing setting. Martino et al. (2011b) investigate the use of INLA with the R-INLA package (www.r-inla.org) by considering a Cox model where the baseline hazard has a parametric or semi-parametric specification. Also, Jiang et al. (2014) study the effect of environmental radiation on cancer by using a cure fraction mixture survival model with a Weibull distribution for event times.

We investigate how Laplace approximations can be extended and combined with penalized B-splines in the context of a semi-parametric promotion time cure model. Bridging the gap between Laplace's method and regression splines brings a twofold advantage. First, it provides a fast computational approach to approximate posterior distributions and second, the spline dimension allows for a flexible specification of the baseline distribution yielding smooth estimates of survival quantities. Another crucial point is that in contrast to the classic INLA approach which focuses mainly on posterior marginal univariate distributions, our methodology permits to compute reliable approximations to the posterior joint distributions of latent variables including regression parameters, with the implication that set estimators can be derived even for complicated functions of latent variables such as the baseline or conditional population survival functions.

Accordingly, the end user will be endowed with a powerful and rapid tool for making inference in the promotion time cure model. Furthermore, while the code design underlying INLA assumes a one-to-one connection between data points and a subset of the latent field, implying that the dimension of the latter grows with the sample size n, our modeling strategy choice is more efficient as it involves a latent field of a dimension unaffected by the number of observations. Hence, given that the number of B-splines is fixed (to a large value and counterbalanced by a roughness penalty) in the P-spline approach (Eilers and Marx, 2010), the latent field dimension grows only with the number of regressors in the model and not with n.

This manuscript is organized as follows. In Section 2, the Laplace-P-spline promotion time cure model is defined and the gradient and Hessian of the log-likelihood are computed to obtain a Gaussian approximation of the conditional posterior distribution of the latent field. A strategy is proposed to explore the posterior distribution of the hyperparameter vector and the joint posterior of latent field elements are derived. The construction of credible intervals for the baseline and population survival functions is also addressed here. In Section 3, the merits of the proposed methodology will be assessed by extensive simulations with different scenarios regarding the percentages of cured individuals and right censored subjects. Coverage properties of credible intervals will also be considered. In Section 4, we apply the model to two real datasets and Section 5 concludes with a discussion.

2. Laplace-P-spline promotion time model

2.1. Flexible modeling of the baseline hazard

Following Rosenberg (1995), the log-hazard corresponding to the baseline survival function $S_0(t)$ in (1) is specified as a linear combination of cubic B-splines $h_0(t) = \exp\big(\boldsymbol{\theta}^T b(t)\big)$, where $b(\cdot) = (b_1(\cdot), \dots, b_K(\cdot))^T$ is a cubic B-spline basis obtained

Please cite this article in press as: Oswaldo G., Lambert P., Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.02.007.