Computational Statistics and Data Analysis 81 (2015) 52–63

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A new minimal training sample scheme for intrinsic Bayes factors in censored data

Stefano Cabras [a,b,*], Maria Eugenia Castellanos [c], Silvia Perra [d]

[a] Department of Statistics, Universidad Carlos III de Madrid, Spain
[b] Department of Mathematics and Informatics, Università di Cagliari, Italy
[c] Department of Informatics and Statistics, Universidad Rey Juan Carlos, Spain
[d] Department of Social Science and Institutions, Università di Cagliari, Italy

ARTICLE INFO

Article history:
Received 7 January 2014
Received in revised form 10 July 2014
Accepted 22 July 2014
Available online 31 July 2014

Keywords:
Improper priors
Intrinsic Bayes factor
Kaplan–Meier estimator
Model selection
Survival analysis

ABSTRACT

The problem of covariate selection for regression models with right censored data is considered. It is approached from a default Bayesian point of view with Bayes factors (BFs) and in particular with Intrinsic BF (IBF) that depends on the minimal training samples (MTSs). In the presence of censored data, the number of possible MTSs increases, due to the fact that uncensored data, relevant for training the improper prior into a proper posterior, must be combined with censored data. For this purpose, the sequential minimal training sample scheme (SMTS) accounts for such requirements but generally leads to IBF correction factors that do not have an analytical form and thus require numerical approximation. In order to obtain an analytical expression of the correction terms, a different TS scheme is introduced based on the Kaplan–Meier (KM) estimator, termed the KM minimal training sample scheme. This new tool works extremely well in the analyzed simulation setting and also in the applications; it produces results which are similar, if not better, than the IBF calculated using MTSs. The resulting new IBF, being based on analytical expressions, is much faster to compute. Evidence of these results comes from a large simulation study, theoretical arguments, and an application to a real data set.

1. Introduction

The model selection problem for survival regression analysis is studied with particular attention to covariates selection where the response variable Y is censored. From a Bayesian model selection perspective (see e.g. Nott and Leng, 2010; Wagner and Duller, 2012; and Lee et al., 2014) the most common tools are the Bayes factors (BFs), which unfortunately are undefined when default improper priors are employed. The extensive literature on methods that deal with such a situation is mainly focused on the Intrinsic BF (IBF) (Berger and Pericchi, 1996) that, together with the Fractional BF (FBF) (O'Hagan, 1995), have been defined in order to avoid indeterminacy. Under model regularity conditions, also met in this paper, such approximated BFs converge asymptotically to actual BFs as it is possible to show that there is an intrinsic prior leading to the asymptotic IBF (or asymptotic fractional BF). These two BFs are defined upon a minimal training sample (MTS), drawn, in a suitable manner, from the available data. The MTS is used to train the improper prior into a proper posterior, thus making the indetermination of the BF disappear. In particular, both depend on the size of the MTS, while the IBF also depends on the particular MTS that has been drawn.

* Correspondence to: C/ Madrid 126, 28903 Getafe, Spain. Tel.: +34 91 6249314; fax: +34 91 6249849.
E-mail addresses: s.cabras@unica.it, stefano.cabras@uc3m.es (S. Cabras).

http://dx.doi.org/10.1016/j.csda.2014.07.012
0167-9473/

Throughout the paper the IBF is mainly used for censored data. The motivation for exploring model selection using this tool instead of other ones such as FBF, or other possibilities like those discussed, for instance, in Celeux et al. (2012) is that IBFs are posterior model consistent, even if the model dimension grows exactly like the sample size. Posterior model consistency here means that the posterior model probability tends to 1 for the "true" sampling model. The necessary and sufficient conditions for this are discussed in Moreno et al. (2014). Such conditions require the priors for parameters and for models that are the same ones used in this paper.

In survival analysis, because of the presence of censored and uncensored observations, the drawing mechanism of the MTS must take into account these different types of observations because each type leads to a different amount of information to estimate the unknown regression parameters. The problem of exploring the space of the MTSs, which includes censored and uncensored observations, has already been formalized in Assumption 0 of Berger and Pericchi (2004), which states that the hypothetical sampling space of possible MTSs must have probability 1 under each model. Satisfying Assumption 0 is a requirement that leads to the introduction of the sequential minimal training sample scheme (SMTS) (Berger and Pericchi, 2004) which consists in drawing samples sequentially, without replacement, until a specified number of uncensored observations has been obtained. This differs from the ordinary resampling without replacement scheme, denoted here by OMTS, used in calculating the IBF in Berger and Pericchi (1996). The OMTS assigns uniform weights to all observations that train the improper prior into a proper density. However, when dealing with censored data only uncensored observations would lead to a proper trained posterior. In this case the OMTS would provide only unweighted uncensored observations, thus not satisfying Assumption 0 and introducing a bias in the estimation of regression parameters and, hence, in the whole model selection procedure.

A different approach to defining MTSs in the presence of censored data is discussed. This new strategy is very useful when it is possible to obtain closed-form expressions for predictive distributions, which is something that occurs mainly when training samples do not contain censored data. In particular, this is true for some models such as the log-normal one, which is a reference model in survival analysis. Calculating the predictive distribution of Y with only uncensored data, i.e., calculating the likelihood involving only the density function of Y, would significantly reduce the computational efforts because the survival function is usually not available in a closed form expression, as in the case of the log-normal model.

As will be discussed later, the theoretical justification behind the KMMTS is that it assures that Assumption 0 is still asymptotically satisfied, and it generates samples that contain only uncensored observations. This simplifies the calculation of IBF and leads to the possibility of exploring a larger number of models with respect to the IBF defined upon the SMTS. Simplification of calculus consists in substituting an MCMC approach for BF approximation with an exact analytical approximation, and this partially explains, as shown later, the better performance of the IBF defined upon the KMMTS with respect to that upon SMTS.

The rest of the paper is organized as follows: Section 2 describes the IBF and the BIC in the presence of censored data. Section 3 contains definitions of different training sample schemes. Section 4 reconsiders the IBF for right censored data using the KMMTS to be used with the IBF. In Section 5 the proposed technique is applied to the log-normal regression model, providing the corresponding expressions for the IBF. Section 6 compares the BIC and the two versions of the IBFs using a simulation study. Section 7 illustrates an application to a well known data set in survival modeling literature. Some remarks and conclusions can be found in Section 8.

2. The regression model and the objective variable selection

It is useful to first recall the regression model considered although the arguments presented here are of general applicability to the problem of model selection. Later on, objective variable selection is introduced, under a perspective mainly focused on the IBF technique.

Let $(t_i, \delta_i, \boldsymbol{x}_i)$ be the survival time, censored indicator and covariates, respectively, for individual $i = 1, \ldots, n$, where $\delta_i = 0$ if right censored and $\delta_i = 1$ otherwise. Without loss of generality, consider a fixed design matrix $\boldsymbol{X}$ with $p + 1$ columns, including the intercept. Let $y_i = \log(t_i)$ be normally distributed according to the following regression model $\mathcal{M}_k$ with a set of covariates denoted by $\boldsymbol{x}_{k,i}$

$$\mathcal{M}_k : y_i = \boldsymbol{\beta}_k^{\mathsf{T}} \boldsymbol{x}_{k,i} + \sigma_k \epsilon_i, \tag{1}$$

where $k \in \{1, 2, \ldots, K = 2^p\}$ is the model index with the corresponding design matrix $\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{x}_{k,n})^{\mathsf{T}} \in \mathbb{R}^{n \times p_k}$ and model parameter $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k) \in \boldsymbol{\Theta}_k = \mathbb{R}^{p_k} \times \mathbb{R}^+$. Here ϵ_i is assumed to be normally distributed such that the observed times are log-normal. Note that other types of distributions for the time are possible by just fixing that of ϵ_i. For instance, the Weibull model is obtained assuming that ϵ_i is distributed according to a standard Gumbel, or if ϵ_i follows a generalized Gumbel distribution, the observed times would follow a generalized gamma distribution. The proposed approach also fits such models, but the computational advantage is limited, in that when calculating the BF, it is not necessary to integrate the survival function.

The most probable model $\mathcal{M}_k$, given the observed data through BFs and model posterior probabilities (see Kass and Raftery, 1995; Berger, 1999 and Berger and Pericchi, 2001), is selected. In order to calculate BFs a prior distribution $\pi_k(\boldsymbol{\theta}_k)$ needs to be specified separately for each model. This can be complicated because one often initially entertains K models leading to the impossibility of careful subjective prior elicitation. For this purpose, Bayesian model selection is usually done