

Zellner-Siow Priors for variable selection with censored data

Stefano Cabras, Mara Eugenia Castellanos

Abstract In this paper we analyze the problem of variable selection in regression analysis when the response variable, Y , follows a lognormal distribution and observations are right censored. The use of conventional priors, namely g -priors and Zellner-Siow priors are analyzed in the context of censored data. Of particular interest is the understanding of the covariance matrix that must be used for specifying such priors in the presence of censored data. A comparison among different options is made through a simulation study.

Abstract *In questo lavoro analizziamo il problema della selezione di regressori quando la variabile risposta, Y , segue una legge log-normale e le osservazioni sono censurate a destra. L'uso di prior convenzionali come le g -prior e la Zellner-Siow prior sono analizzate nel contesto di dati censurati. In particolare è di interesse la determinazione della matrice di covarianza che deve essere utilizzata per specificare queste prior quando si è in presenza di dati censurati. Un confronto, tra le diverse opzioni, è effettuato tramite uno studio di simulazione.*

Key words: Bayes Factor, Cauchy distribution, g -priors, regression survival models

1 Introduction

In this paper we analyze the problem of variable selection in regression analysis when the response variable, Y , follows a parametric distribution (e.g. lognormal)

Stefano Cabras

University of Cagliari (Italy), Department of Mathematics, e-mail: s.cabras@unica.it and University Carlos III de Madrid (Spain), Department of Statistics, e-mail: stefano.cabras@uc3m.es

Mara Eugenia Castellanos

Rey Juan Carlos University (Spain), Department of Statistics and O.R. e-mail: maria.castellanos@urjc.es

and observations are right censored. From a Bayesian perspective the most widely used tools to compare models are the Bayes factors (BFs).

More formally, let $(t_i, \delta_i, \mathbf{x}_i)$ be the survival time, censoring indicator and covariates, respectively for individual $i = 1, \dots, n$, where $\delta_i = 0$ if right censored and 1 if uncensored, and where $\sum_{i=1}^n \delta_i = n_u = n - n_c$ is the number of uncensored observations. Consider $y_i = \log(t_i)$, $\mathbf{y} = (y_1, \dots, y_n)^T$, the vector collecting all log of survival times, and the following regression model $\mathcal{M}_k$ that includes the k -th set of covariates through the design matrix $\tilde{\mathbf{X}}_k \in \mathcal{R}^{n \times p_k}$,

$$\mathcal{M}_k : \mathbf{y} = \mathbf{1}_n \beta_{0,k} + \tilde{\mathbf{X}}_k \boldsymbol{\beta}_k + \sigma_k \boldsymbol{\varepsilon},$$

where k is the model index, $\mathbf{1}_n$ is a vector of ones of length n , $\beta_{0,k}$ represents the intercept for model k , $\boldsymbol{\beta}_k$ is the p_k dimensional vector of regression coefficients, σ_k is a scale parameter and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ is the error term. When no covariates are selected, we have the simplest model explaining $\mathbf{y}$, that we call the “null” model:

$$\mathcal{M}_0 : \mathbf{y} = \mathbf{1}_n \beta_0 + \sigma_0 \boldsymbol{\varepsilon}, \quad (1)$$

So the index model k takes values in $\{0, 1, 2, \dots, K = 2^p - 1\}$. In all these models, if the error term ε_i defining $\boldsymbol{\varepsilon}$, is distributed as a standard normal distribution then survival times would be lognormally distributed, while if they are distributed as a standard Gumbel distribution, times will follow a Weibull one. For purposes of exposition, we consider only normal distributed errors.

The aim is to select the most probable model $\mathcal{M}_k$ given the observed data through BF's and posterior probabilities of each model (see Kass and Raftery (1995); Berger (1999); Berger and Pericchi (2001)). In order to calculate BF's we need to specify a prior distribution $\pi_k(\boldsymbol{\theta}_k)$ where $\boldsymbol{\theta}_k = (\beta_{0,k}, \boldsymbol{\beta}_k, \sigma_k)$, for $k = 0, \dots, K = 2^p - 1$, separately for each model. This can be complicated, because one often initially entertains K models leading to the impossibility of careful subjective prior elicitation. Hence, Bayesian model selection is usually done by means of default methods (Berger and Pericchi, 2001). In literature there exist at least four methods under the label of default Bayesian model selection: the Conventional Prior (CP) approach, Intrinsic and Fractional BF's and the Bayesian Information Criterion (BIC). In this paper we focus on the first method applied to censored data, while more details about Intrinsic, Fractional BF's and the Bayesian Information Criterion (BIC) for censored data can be found in Perra et al. (2013). The problem is that, in general, the use of improper priors for model specific parameters is not allowed in model selection, because improper priors are determined only up to an arbitrary constant. To avoid that, conventional proper priors over model specific parameters, $\boldsymbol{\beta}_k$, are used leading to a procedure named conventional prior approach (Berger and Pericchi, 2001).

Once prior distributions $\pi_k(\boldsymbol{\theta}_k)$ are specified for each model, the predictive density of $\mathbf{y}$, under $\mathcal{M}_k$, is $m_k(\mathbf{y}) = \int_{\boldsymbol{\theta}_k} f_k(\mathbf{y} | \boldsymbol{\theta}_k) \pi_k(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k$, where f_k is the induced density of Y from equation (1).

The BF of $\mathcal{M}_i$ against $\mathcal{M}_j$ is given by

$$B_{ij}(\mathbf{y}) = \frac{m_i(\mathbf{y})}{m_j(\mathbf{y})}, \quad (2)$$

and the posterior probability of $\mathcal{M}_i$ is

$$p(\mathcal{M}_i | \mathbf{y}) = \frac{m_i(\mathbf{y})p(\mathcal{M}_i)}{\sum_{j=1}^K m_j(\mathbf{y})p(\mathcal{M}_j)} = \left(1 + \sum_{j \neq i} \frac{p(\mathcal{M}_j)}{p(\mathcal{M}_i)} B_{ji}(\mathbf{y}) \right)^{-1}, \quad (3)$$

Choices for model prior probabilities in default Bayesian analysis are the uniform, $p(\mathcal{M}_i) = 1/2^p$ (when dimension of models is not large), or the hierarchical uniform prior discussed by Scott and Berger (2010) when the number of covariates is large, in order to account for the multiplicity of comparisons.

2 Conventional priors: g -prior and mixture of g -priors

The conventional prior approach, first used by Jeffreys (1961, Chapter 5), treats the problem of indeterminacy of noninformative priors using: i) noninformative priors only for common orthogonal parameters in all models, so that the arbitrary multiplicative constants for the priors cancel in all BF, and ii) conventional proper priors for model specific parameters. This procedure has been used by many other authors (for instance, Zellner and Siow (1980); Berger and Pericchi (2001); Berger et al. (1998); Bayarri et al. (2012)).

In order to follow the conventional prior approach, we consider the null model defined in (1), and for the rest of $k = 1, \dots, 2^p - 1$, we adopt the following parametrization:

$$\mathcal{M}_k : \mathbf{y} = \mathbf{1}_n \beta_0 + \mathbf{X}_k \boldsymbol{\beta}_k + \sigma_0 \boldsymbol{\epsilon}, \quad (4)$$

where we assume that the columns of $\mathbf{X}_k$ have been centered, so that $\mathbf{1}_n^T \mathbf{X}_k = \mathbf{0}$, in this way the intercept β_0 may be regarded as common to all models. Also σ_0 is considered a common parameter to all 2^p models.

Arguments based on orthogonal parametrizations and invariance to scale and location transformations (Jeffreys, 1961; Liang et al., 2008; Berger et al., 1998) have guided to adopt the following prior under model $\mathcal{M}_k$, $\pi(\boldsymbol{\beta}_k, \beta_0, \sigma_0) = \pi_k(\boldsymbol{\beta}_k | \beta_0, \sigma_0) \pi(\beta_0, \sigma_0) \propto \pi_k(\boldsymbol{\beta}_k | \beta_0, \sigma_0) / \sigma_0$. A common choice of prior for $\boldsymbol{\beta}_k | \beta_0, \sigma_0$ is the normal prior, called g -prior in Zellner (1986).

$$\pi_k^g(\boldsymbol{\beta}_k | \beta_0, \sigma_0) = N_{p_k}(\mathbf{0}, g \sigma_0^2 (\mathbf{X}_k^T \mathbf{X}_k)^{-1}), \quad (5)$$

where g is a hyper parameter to be specified. Common choices are, under the unit information prior criterion (Kass and Wasserman, 1995), to fix g as the amount of information about parameters equal to the amount of information in one observation. For regular parametric families the amount of information is quantified through Fisher information. In the case of the normal regression, without censoring,

this leads to take $g = n$. While in the case of censored data, Volinsky and Raftery (2000) proposed to use $g = n_u$. Along the paper we consider the g -prior using $g = n$ and $g = n_u$, varying the variance matrix for $\boldsymbol{\beta}_k$ according to the information in all the sample without distinguishing between censored and uncensored observations or that provided by only uncensored data.

In order to avoid some paradoxes of g -priors (Liang et al., 2008), and to avoid fixing g to a specific value, the Zellner-Siow prior (ZS) (Zellner and Siow, 1980) is constructed as a mixture of g -priors. The ZS prior can be expressed as a mixture of g -prior with respect to an Inv-Gamma($1/2, n/2$), that is,

$$\pi_k^{ZS}(\boldsymbol{\beta}_k | \beta_0, \sigma_0) = \int N_{p_k}(\boldsymbol{\beta}_k | \mathbf{0}, g\sigma_0^2(\mathbf{X}_k^T \mathbf{X}_k)^{-1}) \pi(g) dg = Ca_{p_k}(\mathbf{0}, n\sigma^2(\mathbf{X}_k^T \mathbf{X}_k)^{-1}) \quad (6)$$

with $\pi(g) = (n/2)^{1/2} / \Gamma(1/2) g^{-3/2} e^{-n/(2g)}$. Again, in the context of censoring, it is interesting to analyze the behavior of the ZS prior, when using n , or n_u in the definition of $\pi(g)$, and the appropriate covariance matrix.

The specification of these priors leads to coherent prior specifications for a pair of hypothesis. But in the context of variable selection in regression models, there is a multiple hypothesis testing problem, that produces many non-nested comparisons. In order to define all possible BFs comparing any pair of hypothesis, we adopt the *encompassing* approach of Zellner and Siow (1980), where each sub-model $\mathcal{M}_i$ is compared with a reference model, for example the null one $\mathcal{M}_0$, $B_{ij} = B_{i0}/B_{j0}$, $i, j = 1, \dots, K, i \neq j$.

3 Application to the lognormal model under censoring

To obtain the above BFs it is necessary to calculate the marginal predictive distributions, for which there is not an analytical form. In this section, we drop the model subscript k , and consider the calculation of the predictive distributions of the log-time for a Normal regression model with s regressors. Under prior defined as in the above section, the posterior kernel is

$$\pi(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) \propto \frac{1}{\sigma} \prod_{i=1}^n \left(\frac{1}{\sigma} f_{\varepsilon} \left(\frac{y_i - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma} \right) \right)^{\delta_i} \left[S_{\varepsilon} \left(\frac{y_i - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma} \right) \right]^{(1-\delta_i)} \pi^*(\boldsymbol{\beta} | \beta_0, \sigma_0),$$

where f_{ε} and S_{ε} are the density and survival functions of ε , and $\pi^*(\boldsymbol{\beta} | \beta_0, \sigma_0)$ denotes π^g or π^{ZS} . $\pi(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X})$ has been approximated using Markov Chain Monte Carlo (MCMC). In particular, we employed a Random Walk Metropolis-Hastings (RW-MH) for parameters $(\boldsymbol{\beta}, \log(\sigma))$ where the proposal is a standard multivariate normal distribution. The marginal predictive distribution of Y , for each model $\mathcal{M}_k$, $m(\mathbf{y} | \mathcal{M}_k)$, has been obtained using the method proposed in Chib and Jeliazkov (2001) (Subsection 2.1).

4 Simulation Study

The conditional distribution of $y_i | \mathbf{x}_i$ has been simulated from the normal regression model and the censoring indicator has been simulated according to a normal distribution, considering 90% of censored data and different sample sizes. The generating model for Y has been chosen by fixing the coefficients of the following regression model:

$$Y_i = \log(T_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma \varepsilon_i \quad i = 1, \dots, n,$$

where $\beta_0 = 0$ and $\sigma = 1$ for all generated samples, while 4 different sets of coefficients are considered: i) $\mathcal{M}_0$ =Null model: $(\beta_1, \beta_2, \beta_3) = (0, 0, 0)$; ii) $\mathcal{M}_1$ =Model with 1 covariate: $(\beta_1, \beta_2, \beta_3) = (2, 0, 0)$; iii) $\mathcal{M}_2$ =Model with 2 covariates: $(\beta_1, \beta_2, \beta_3) = (2, 2, 0)$ and iv) $\mathcal{M}_3$ =Model with 3 covariates: $(\beta_1, \beta_2, \beta_3) = (2, 2, 2)$. Results are summarized in Figure 1. The main differences appear between priors π^g and π^{ZS} ,

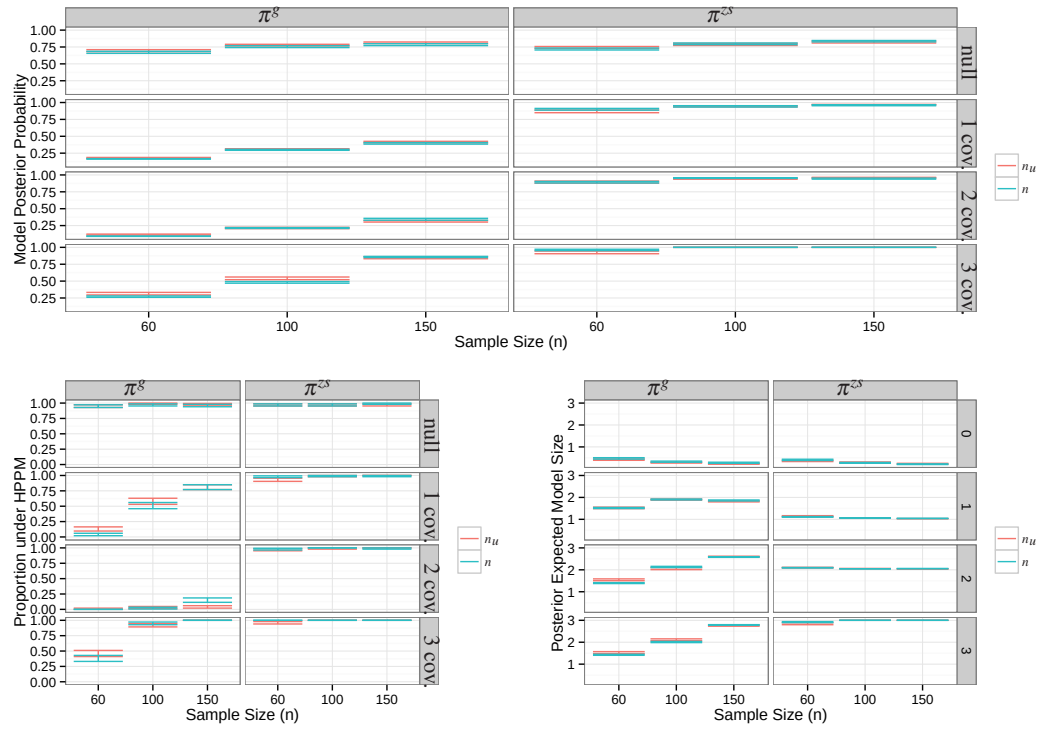


Fig. 1 Results for 100 replications of the simulation study: distribution of the posterior model probability for the generating model (top), the proportion of times that the true model is selected according to the highest posterior probability (bottom left) model criteria along with the posterior expected model size (PEMS) (bottom right).

being π^{ZS} the one that performs better. Secondly, differences between using n_u or n and the appropriate covariance matrix, are not very significative, except for the π^g at smaller sample sizes and complex models.

References

- Bayarri, M., J. Berger, A. Forte, and G. García-Donato (2012). Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics* 40(3), 1550–1577.
- Berger, J. O. (1999). Bayes factors. In S. Kotz, S. Read, and D. Banks (Eds.), *Encyclopedia of Statistical Sciences*, Volume 3, pp. 20–29. New York: Wiley.
- Berger, J. O. and L. R. Pericchi (2001). Objective bayesian methods for model selection: Introduction and comparison. In P. Lahiri (Ed.), *Model Selection*, Volume 38, pp. 135–207. Institute of Mathematical Statistics.
- Berger, J. O., L. R. Pericchi, and J. A. Varshavsky (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 60(3), pp. 307–321.
- Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association* 96(453), 270–281.
- Jeffreys, H. (1961). *The Theory of Probability*. Oxford Univ. Press.
- Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90, 773–795.
- Kass, R. E. and L. Wasserman (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association* 90(431), pp. 928–934.
- Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association* 103(481), pp. 410–423.
- Perra, S., S. Cabras, and M. Castellanos (2013). *Objective Bayesian Variable Selection for Censored Data*. Amazon.
- Scott, J. and J. Berger ((2010)). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* 38(5), 2587–2619.
- Volinsky, C. T. and A. E. Raftery (2000). Bayesian information criterion for censored survival models. *Biometrics* 56(1), pp. 256–262.
- Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In P.K.Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–243. Amsterdam: North-Holland.
- Zellner, A. and A. Siow (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics 1* (Eds. J.M. Bernardo and J.O. Berger and A.P. Dawid and A.F.M. Smith), Volume 31, pp. 585–603. Springer Berlin / Heidelberg. 10.1007/BF02888369.