International Journal of Statistics and Probability; Vol. 6, No. 2; March 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Objective Bayesian Snalysis for the Complementary Exponential Geometric Model Applied to Cancer Data

Daniele C. T. Granzotto[1], Vera Tomazella[2] & Francisco Louzada[3]

[1] Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo, 5.790, Maringá-PR, 87020-900, Brazil

[2] Departamento de Estatística, Universidade Federal de São Carlos, Via Washington Luís, km 235, Caixa Postal 676, 13565-905, S o Carlos, SP, Brazil

[3] Instituto de Ciências Matemticas e da Computação, Universidade de São Paulo, Caixa Postal 668, So Carlos, SP, 13566-590, Brazil

Correspondence: Daniele C. T. Granzotto, Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo, 5.790, Maringá-PR, 87020-900, Brazil. E-mail:dctgranzotto@uem.br

Received: December 9, 2016 Accepted: February 15, 2017 Online Published: February 23, 2017

doi:10.5539/ijsp.v6n2p122 URL: https://doi.org/10.5539/ijsp.v6n2p122

Abstract In this paper we provide a reference Bayesian framework to a new two-parameter lifetime distribution with increasing failure rate, the complementary exponential geometric (CEG). To this end, we presented some of the main properties of this model and its characteristics related to the reliability analysis. A simulation study is performed to analyse the frequentist properties of credible intervals from the reference posterior distribution among of the standard error and mean square error (MSE) of estimations. The presented methodology is illustrated by the use of a real data set which presents the study of time until the cure of cervix lesions, that are precursors cancer lesions in the cervix. According to to INCA (Cancer National Institute), cervical cancer stands as the fourth cause of death among women in Brazil. Together with breast cancer, it is one of the most common malignancy affecting women worldwide. For this reason, patients must be carefully evaluated for metastatic disease. These data were collected in the Woman Clinic which is sited in Maringá city (Paraná State, Brazil).

Keywords: CEG distribution, bayesian analysis, objective prior, survival analysis

1. Introdution

Cervical intraepithelial neoplasia (CIN), also known as cervical dysplasia and cervical interstitial neoplasia, is the potentially premalignant transformation and abnormal growth (dysplasia) of squamous cells on the surface of the cervix. Most cases of CIN remain stable, or are eliminated by the host's immune system without intervention. However a small percentage of cases progress to become cervical cancer, usually cervical squamous cell carcinoma (SCC), if left untreated. The major cause of CIN is chronic infection of the cervix with the sexually transmitted human papillomavirus (HPV), especially the high-risk HPV types 16 or 18. Over 100 types of HPV have been identified. About a dozen of these types appear to cause cervical dysplasia and may lead to the development of cervical cancer. The estimated annual incidence in the United States of CIN among women who undergo cervical cancer screening is 4 percent for CIN 1 and 5 percent for CIN 2,3. High grade lesions are typically diagnosed in women 25 to 35 years of age, while invasive cancer is more commonly diagnosed after the age of 40, typically 8 to 13 years after a diagnosis of a high grade lesion. It is estimated that about 500,000 new cases are reported every year, with approximately 230,000 deaths worldwide. In Brazil, the crude incidence rates per 100,000 women, estimated for the year 2012, were 17 for the country and 14 for the Rio Grande do Norte State. The incidence of the disease starts from the age of 20 and the risk gradually increases with age, reaching its peak generally at age 50 to 60, (Limaetal, 2013; Ferenczy, 1982).

In Brazil, the coverage of the cytopathology exam has not still reached the desired indices. It is estimated that approximately 40% of Brazilian women have never undergone the procedure. This is due to several factors, including the difficulty of access to health services, poor knowledge about the Pap smear, and the lack of awareness of the benefits that this exam brings to womens health (for more informations see INCA website). The recognition of the central role of HPV in the etiology of cervical cancer has dramatically changed the vision of how to prevent this cancer. Among the strategies, introduction of HPV testing for primary screening or as an adjunct test and the introduction of HPV vaccines to prevent HPV infection are being evaluated in different settings around the world.

Indeed, these advances are available in many countries that have approved the introduction of the HPV vaccine. However, gaps in knowledge about the causal role of HPV in cervical cancer and the benefits of preventing HPV infection may hamper the successful introduction of the technologies. The greater the womens knowledge of HPV and its role in the development of cervical cancer, the greater will be the adherence to preventive measures.

The presented methodology in this paper is applied to a real study involving patients treated in the Woman Clinic sited in Maringá city, Paraná State, Brazil. The data consist of the time until the cure of CIN (in months) in 363 women followed up from years 2,000 to 2,006. For all patients was observed the maximum time that the initial lesions took to be considered totally cure (after specific diagnostics). We consider a new two-parameters lifetime distribution with increasing failure rate, the complementary exponential geometric distribution proposed by (Louzada, et al., 2011). which is complementary to the exponential geometric model proposed by Adamidis & Loukas (1998). The new distribution arises on a latent complementary risks scenarios, where the lifetime associated with a particular risk is not observable, rather we observe only the maximum lifetime value among all risks. This distribution is based on a generalization of the the exponential distribution, which is a widely used lifetime distribution for modeling many problems in lifetime testing and reliability study.

According to Bernardo (Bernardo, 1979), in the quest for objective posterior distributions, several requirements have emerged which may reasonably be requested as necessary properties of any proposed solution: generality such procedure should be completely general; invariance(Jeffreys, 1946; Datta Ghosh, 1995; Datta Ghosh, 1996); present consistent marginalization; and the properties under repeated sampling of the posterior distribution must be consistent with the model (Neyman Scott, 1948; Lane Sudderth, 1984).

Under these cited assumptions, Bernardo (Bernardo, 1979) introduced the reference analysis which was further developed by Berger and Bernardo (Berger Bernardo, 1992a) and Berger and Bernardo (Berger Bernardo, 1992b). The technique, according to those authors, appears to be the only available method to derive objective posterior distributions which satisfy all these desiderata and even for moderate sample sizes, the information provided by the data should dominate the prior information because of the vague nature of the prior knowledge.

Alternative to this reference prior, the Jeffreys prior is widely used in Bayesian with an important characteristic that it is invariant under injective transformations (Paulino, et al., 2003). Despite the property of invariance, the posterior distribution by using Jeffreys prior may be improper (unlike by using reference prior) and lead to a uninformative distribution. Also, the use of the Jeffreys rule in the multiparametric case is often inadequate. The assumption of a prior independence between parameters of different natures, and the separate use of Jeffreys rule for specification of marginal distributions may give different results than obtained by the Jeffreys principle (Berger, 1985).

So, in this paper the parameter estimation will be considered from the reference Bayesian perspective for the parameters of the the complementary exponential geometric (CEG) distribution (the MLEs of the CEG model in a classical context can be seem in (Louzada, et al., 2011)).

The paper is organized as follows. Section 2, we presented a briefly introduction to a model CEG with some particularities and an overview of reference analysis with emphasis on the case of two parameters. In Section 3, the inference for the CEG model is presented by using the reference prior built. After presenting the model and inference aspects, a simulation study was made and it can be seem in Section 4. Finally, in Sections 5 and 6, respectively, we can see the usefulness of the CEG model in a Bayesian context by studying the time until the cure of precursors cancer cervix lesions and some conclusions.

2. Background

2.1 The CEG Model

Proposed by (Louzada, et al., 2011), the CEG model was formulated to describe lifetime with increasing failure rate. According to the authors, this is complementary to the exponential geometric model proposed by Adamidis and Loukas (Adamidis Loukas, 1998) and arises on a latent complementary risks scenario, in which the lifetime associated with a particular risk is not observable. Instead, we observe only the maximum lifetime value among all risks. For more details on latent risk problem, interested readers can refer to Basu and Klein (Basu Klein, 1982) and Louzada-Neto (Louzada-Neto, 1999).

Let M be a random variable denoting the number of failure causes, $m = 1, 2, \ldots$, and considering M with geometrical distribution of probability given by

$$P(M = m) = \theta(1 - \theta)^{m-1}, \tag{1}$$

where $0 < \theta < 1$ and $M = 1, 2, \ldots$.

Let us also consider t_i, $i = 1, 2, 3, \ldots$, realizations of a random variable denoting the failure time, i.e., the time-to-event due to the ith complementary risk, with T_i having an exponential distribution with probability index λ, given by

$$f(t_i|\lambda) = \lambda \exp\{-\lambda t_i\}. \tag{2}$$

In the latent complementary risk scenario, the number of causes M and the lifetime t_i associated with a particular cause are

not observable (latent variables), and only the maximum lifetime Y among all is usually observed. So, we only observed the random variables given by

$$Y = \max(t_1, t_2, \ldots, t_M). \tag{3}$$

By considering those descriptions, the model CEG was built and the Proposition 2.1 follows.

Proposition. *Let Y be a nonnegative random variable denoting the lifetime of an individual in some population. The random variable y is distributed according to a CEG distribution, with parameters $\lambda \in \Lambda$ and $\theta \in \Theta$, $\Lambda = \{\lambda; \lambda \in (0, +\infty)\}$, $\Theta = \{\theta; \theta \in (0, 1)\}$, if its probability density function (pdf) and cumulative distribution function are given, respectively, by*

$$f(y|\theta, \lambda) = \frac{\lambda\theta e^{-\lambda y}}{\left(e^{-\lambda y}(1-\theta)+\theta\right)^2}, \tag{4}$$

and

$$F(y|\theta, \lambda) = \frac{\theta\left(1 - e^{-\lambda y}\right)}{e^{-\lambda y}(1-\theta)+\theta}. \tag{5}$$

Proof. The build of CEG model and some specific properties can be find in Louzada et al. (Louzada, et al., 2011).

In order to show the behaviour of the probability density and cumulative function are shown, respectively, in left and right panels of the Figure 1.

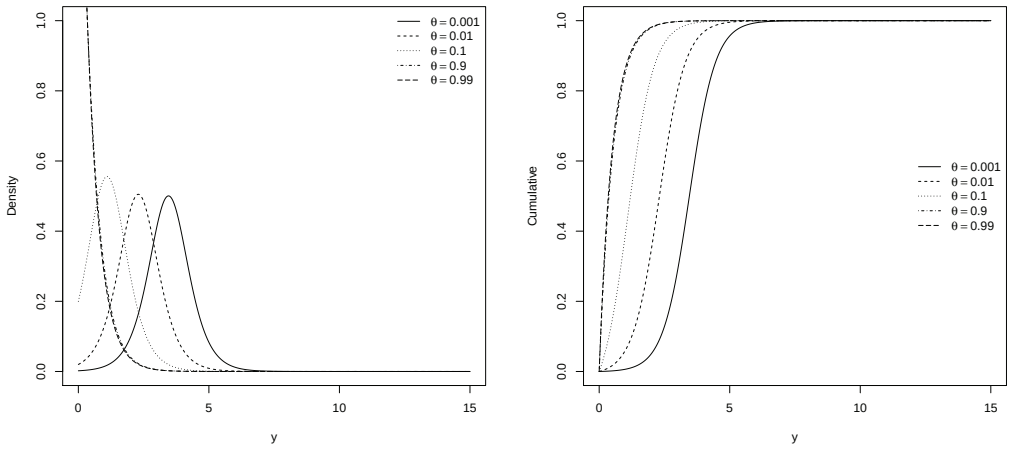

Figure 1. pdf and cdf of the CEG distribution with parameter $\lambda = 2$ fixed.

In reliability analysis, we have interest in two important characteristic of a random variable Y that represents the lifetime, namely, the reliability function $R(y|\theta, \lambda)$, which is the probability of an item not failing prior to some time y, is defined by $R(y|\theta, \lambda) = 1 - F(y|\theta, \lambda)$; and the hazard function, which can be loosely interpreted as the conditional probability of failure, given it has survived to the time y(Lawless, 2003). In what follows, we exhibit the equations of reliability and hazard functions; else, in Figure 2 we show the behaviour of the reliability and hazard functions for some specifics values of parameter θ and $\lambda = 2$ fixed.

Let Y be a nonnegative random variable denoting the lifetime of an individual in some population. The reliability function $R(y|\theta, \lambda)$ is written as

$$R(y|\theta, \lambda) = \frac{e^{-\lambda y}}{e^{-\lambda y}(1-\theta)+\theta}, \tag{6}$$

where $\lambda \in \Lambda$ and $\theta \in \Theta$ are the scale and shape parameters. The hazard, $h(y|\theta, \lambda)$, and cumulative hazard rate, $H(y|\theta, \lambda)$ functions are given by

$$h(y|\theta, \lambda) = \frac{\theta\lambda}{e^{-\lambda y}(1-\theta)+\theta} \tag{7}$$

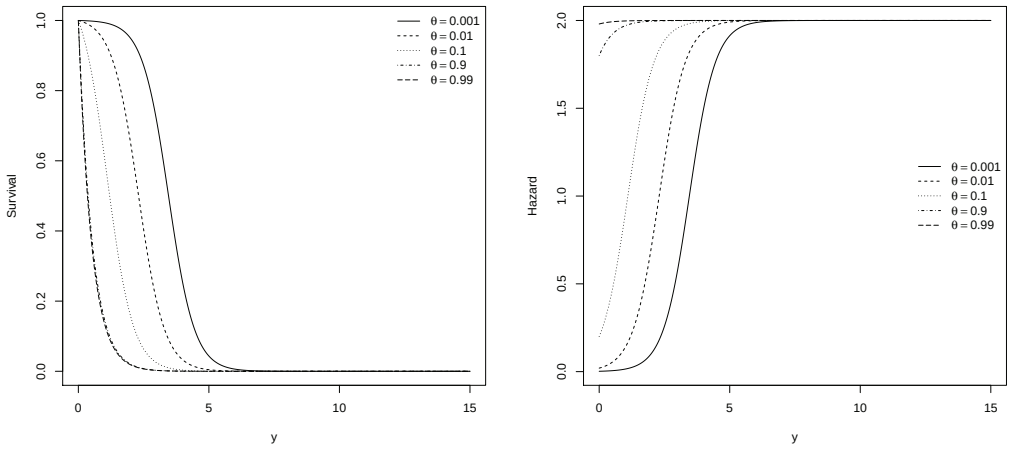

Figure 2. Survival and hazard of the CEG distribution with parameter $\lambda = 2$ fixed.

and

$$H(y|\theta,\lambda) = \ln\left[e^{-\lambda y}(1-\theta)+\theta\right] - \ln\theta - \ln(1-e^{-\lambda y}), \tag{8}$$

$\lambda \in \Lambda$ and $\theta \in \Theta$.

2.2 Reference Analysis

In this section we present the declared objective of reference Bayesian analysis introduced by (Bernardo,1979) and further developed by (Berger Bernardo, 1992a) and(Berger Bernardo, 1992b) is to specify a prior distribution such that, even for moderate sample sizes, the information provided by the data should dominate the prior information because of the "vague" nature of the prior knowledge.

An important feature in the Berger-Bernardo approach to construct a non-informative prior is the different treatment for interest and nuisance parameters. When there are nuisance parameters (typical case in this paper), one must establish an ordered parametrization with the parameter of interest singled out and then follow the procedure below.

Proposition. Let $p(\mathbf{x}|\theta,\lambda)$, $(\theta,\lambda) \in \Theta \times \Lambda(\theta) \subseteq \Re \times \Re$ be a probability model with two real-valued parameters θ and λ, where θ is the quantity of interest. Let $I(\theta,\lambda)$ the corresponding 2x2 Fisher's matrix in terms of θ and λ, and let $V(\theta,\lambda) = I^{-1}(\theta,\lambda)$. Suppose that the joint posterior distribution of (θ,λ) is asymptotically normal with covariance matrix $V(\hat{\theta},\hat{\lambda})$, where $\hat{\theta}$ and $\hat{\lambda}$ are the corresponding consistent estimators of θ and λ. It follows that:

(*i*) the conditional reference prior of λ given θ is

$$\pi(\lambda \mid \theta) \propto [I_{22}(\theta,\lambda)]^{1/2}, \qquad \lambda \in \Lambda(\theta)$$

(*ii*) if $\pi(\lambda|\theta)$ is not proper, a compact approximation $\{\Lambda_i(\theta), \quad i = 1, 2...\}$ to $\Lambda(\theta)$ is required, and the reference prior of λ given θ is

$$\pi_i(\lambda|\theta) = \frac{[I_{22}(\theta,\lambda)]^{1/2}}{\int\limits_{\Lambda_i(\theta)} [I_{22}(\theta,\lambda)]^{1/2}\, d\lambda}, \qquad \lambda \in \Lambda_i(\theta)$$

(*iii*) the sequence of priors can be obtained as

$$\pi_i(\theta) \propto \exp\left\{\int\limits_{\Lambda_i(\theta)} \pi_i(\lambda|\theta) \log\left[v_{11}^{1/2}(\theta,\lambda)\right] d\lambda\right\},$$

where $v_{11}^{1/2}(\theta,\lambda) = I_\theta(\phi,\lambda) = I_{11} - I_{12}I_{22}^{-1}I_{21}$

(*iv*) the reference posterior distribution of θ given data $\{y_1, ..., y_n\}$ is

$$\pi(\theta|y_1,...,y_n) \propto \pi(\theta)\left\{\int_{\Lambda(\theta)}\left\{\prod_{l=1}^{n} p(y_l|\theta,\lambda)\right\}\pi(\lambda|\theta)\,d\lambda\right\}.$$

Proof. See a heuristic justification in Bernardo (2005).

Corollary. If the nuisance parameter space $\Lambda(\theta) = \Lambda$ is independent of θ, and the functions $v_{11}^{-1/2}(\theta,\lambda)$ and $I_{22}^{1/2}(\theta,\lambda)$ factorize in the form

$$\{v_{11}(\theta,\lambda)\}^{-1/2} = f_1(\theta)\,g_1(\lambda), \qquad \{I_{22}(\theta,\lambda)\}^{1/2} = f_2(\theta)\,g_2(\lambda).$$

Then

$$\pi(\theta) \propto f_1(\theta), \pi(\lambda|\theta) \propto g_2(\lambda).$$

Thus, the reference prior relative the ordered parametrization (θ,λ) is given by

$$\pi(\lambda,\theta) = f_1(\theta)\,g_2(\lambda)$$

and there is no need for compact approximation, even if the conditional reference prior $\pi(\lambda|\theta)$ is not proper.

Proof. See proof of Theorem 12 in Bernardo (2005).

3. Reference Prior for the CEG Model

According to Bernardo (2005), let $\mathbf{y}^k = (y_1, \ldots, y_k)$ k-independent replications of the CEG model and consider $h(\theta,\lambda) = 1$ a positive function. Then,

$$q(\theta,\lambda,y_1,\ldots,y_k) \propto L(\theta,\lambda)h(\theta,\lambda)$$

is an asymptotic approximation of posterior distribution. Under certain regularity conditions whereas there is a maximum likelihood estimator $(\hat{\theta}(\mathbf{y^k}), \hat{\lambda}(\mathbf{y^k}))$ we have that the posterior density $q(\theta,\lambda,y_1,\ldots,y_k)$ is approximately normal k-dimensional, i.e,

$$q(\theta,\lambda,y_1,\ldots,y_k) \sim N_k\left((\hat{\theta},\hat{\lambda}), nI(\hat{\theta},\hat{\lambda})\right), \tag{9}$$

where $I(\hat{\theta},\hat{\lambda})$ is the Fisher information matrix.

Considering that the posterior distribution is assymptotically normal, then the reference prior only depends on Fisher information matrix. Here we derive the reference prior considering the approach of one nuisance parameter described above.

Then, the likelihood and log-likelihood functions of (θ,λ) based on the observed sample of size n, $\mathbf{y} = y_1, y_2, \ldots, y_n$, from the CEG distribution (4) are given, respectively, by

$$L(\theta,\lambda|y_1,\ldots,y_n) = \frac{\lambda^n\theta^n \exp(-\lambda\sum_{i=1}^n y_i)}{\prod_{i=1}^n\left[e^{-\lambda y_i}(1-\theta)+\theta\right]^2} \tag{10}$$

and

$$l(\theta,\lambda|y_1,\ldots,y_n) = n\log(\lambda\theta) - \lambda\sum_{i=1}^n y_i - 2\sum_{i=1}^n \log\left[e^{-\lambda y_i}(1-\theta)+\theta\right]. \tag{11}$$

In order to find the Fisher information matrix, the first-order derivates of the log-likelihood function (11) for a single observation are given by

$$\frac{\partial l(\theta,\lambda|y_1,\ldots,y_n)}{\partial\theta} = \frac{1}{\theta} - 2\left[\frac{1-e^{-\lambda y}}{e^{-\lambda y}(1-\theta)+\theta}\right] \tag{12}$$

and

$$\frac{\partial l(\theta,\lambda|y_1,\ldots,y_n)}{\partial\lambda} = \frac{1}{\lambda} - y + 2\left[\frac{(1-\theta)ye^{-\lambda y}}{e^{-\lambda y}(1-\theta)+\theta}\right]. \tag{13}$$

The second-order derivates of the log-likelihood function (11) are given by

$$\begin{aligned}
\frac{\partial^2 l(\theta,\lambda;y)}{\partial\theta^2} &= -\frac{1}{\theta^2}+2\left[\frac{e^{\lambda y}-1}{1+\theta(e^{\lambda y}-1)}\right]^2,\\
\frac{\partial^2 l(\theta,\lambda;y)}{\partial\lambda^2} &= \frac{1+\theta\left(\theta-2+\theta e^{2\lambda y}-2(1-\theta)(1+\lambda^2y^2)e^{\lambda y}\right)}{(1+\theta(e^{\lambda y}-1))^2},\\
\frac{\partial^2 l(\theta,\lambda;y)}{\partial\lambda\partial\theta} &= \frac{2ye^{\lambda y}}{(1+\theta(e^{\lambda y}-1))^2}.
\end{aligned}$$

Finally, the Fisher information matrix for model CEG is given by

$$I(\theta,\lambda)=n\begin{bmatrix} \frac{1}{3\theta^2} & \frac{(\theta-1)\theta-\log(\theta)}{3\lambda\theta(\theta-1)^2} \\ \frac{(\theta-1)\theta-\log(\theta)}{3\lambda\theta(\theta-1)^2} & \frac{-(1+\theta)+2Li_2(\beta)}{3\lambda^2(1-\theta)} \end{bmatrix} \tag{14}$$

where $Li_2(\cdot)$ represents the polylogarithm function given by $Li_2(\beta)=\sum_{k=1}^{\infty}\frac{(-\beta)^k}{k^2}$ and $\beta=\frac{1-\theta}{\theta}$. The inverse of Fisher Information matrix for a single observation, presented in (14), is given by

$$V(\theta,\lambda)=\begin{bmatrix} \frac{3(1-\theta)(1+\theta-2Li_2(\beta))}{\zeta(\theta)} & \\ \frac{3(1-\theta)^2\theta\lambda\left(\theta(\theta-1)-\log\theta\right)}{\zeta(\theta)} & \frac{3(1-\theta)^4\lambda^2}{\zeta(\theta)} \end{bmatrix}, \tag{15}$$

where $\zeta(\theta)=(1-\theta)^2+2(1-\theta)\theta+\log(\theta)^2+2(\theta-1)^3Li_2(\beta)$.

Form Collorary 2.2 and the Equation (14), the join reference prior density of CEG model, with two real-valued parameters θ and λ, $\pi(\theta,\lambda)$, $(\theta,\lambda)\in\Theta\times\Lambda(\theta)\subseteq\Re\times\Re$ is given by

$$\pi(\theta,\lambda)\propto\frac{\pi(\theta)}{\lambda}, \tag{16}$$

where

$$\pi(\theta) \propto \sqrt{\varphi(\theta)}\left[-3(\theta-1)^3\theta^2(1+\theta-2Li_2(\beta))\right]^{-1/2},$$

and

$$\varphi(\theta) = (\theta-1)^2-2(\theta-1)\theta\log(\theta)+\log(\theta)^2+2(\theta-1)^3Li_2(\beta),$$

where $Li_2(\beta)$ is the polylogarithm function defined as $Li_2(\beta)=\sum_{k=1}^{\infty}\frac{(-\beta)^k}{k^2}$ and $\beta=\frac{1-\theta}{\theta}$.

Combining the likelihood function in (11) and the prior (16), the joint posterior distribution for θ,λ is given by,

$$\pi(\theta,\lambda|y_1,\ldots,y_n)\propto\frac{\theta^n\lambda^{n-1}e^{-\lambda\sum_{i=1}^n y_i}}{\prod_{i=1}^n\left[e^{-\lambda y_i}(1-\theta)+\theta\right]^2}\pi(\theta). \tag{17}$$

In Appendix A we show that the join posterior distribution is proper.

4. Simulation Study

Louzada et al. (2011) performed a misspecification simulation study for the CEG model, in order to assess the extent of misspecification errors when testing the exponential geometric distribution against the complementary one in the presence of different sample size and censoring percentage. The authors discovered that it is usually possible to discriminate between the distributions even for small samples in the presence of heavy censoring by using the MLEs in a classical inference context (for sample sizes more than 50).

Aiming to examine the finite sample properties of the MLEs, this section presents the results of a Monte Carlo experiment in a Bayesian context. For that, we use 1,000 Monte Carlo replications in all presented results. The sample sizes n range from 50 to 1,000, generated according to a CEG distribution for each combination of the parameter value $\theta=(0.1;0.3;0.5)$ and $\lambda=2$ fixed. The lifetime times of this model were generated by considering the inverse transformation of the cumulative function (Ross, 2009), as follow.

Proposition. *Let $F(y|\theta,\lambda)$, $y \in (0,+\infty)$, denoted a cumulative distributions function of a CGE model. Define $Y = F^{-1}(U)$, where U has a continuous uniform distribution over the interval $(0,1)$. Then Y is distributed as F, that is, $P(Y \leq x) = F(y), y \in (0,+\infty)$ and the lifetimes are generate as*

$$Y = F^{-1}(U) = -\frac{1}{\lambda}\ln\left(\frac{\theta(1-u)}{\theta(1-u)+u}\right). \tag{18}$$

The samples are subsequently obtained by the Metropolis-Hastings technique through the MCMC implemented in software SAS 9.3 - Statistical Analysis System, by using the procedure MCMC with a single chain of the dimension $50,000$. A burn-in of $10,000$ was adopted in order to eliminate the effect of the initial values, resulting a sample size $40,000$. The convergence of the chain was checked by the criterion proposed by Geweke (1992) for each set of simulated data and an average of the estimates of the parameters and standard deviation (SD), the mean square error and the coverage probability of the 95% posterior credible intervals was obtained by using the reference distribution prior. The coverage probability of the posterior credible interval was proposed, in instead of the confidence interval, once we are interested in study and apply this model in a Bayesian context.

Table 1 show the MLEs by using the reference prior presented in Section 3. We can observed that the variances of the MLEs and their mean square error (MSE) become smaller when the sample size increases. Also, the coverage probability of a 95% two sided credibility intervals, for the model parameters, are observed to be close to the nominal coverage for large sample sizes, though the usually differ from the nominal coverage probability less than 2% for moderate sample sizes (more than 100).

Note that, the simulation was proposed to check the behaviour of the MLEs for finite sample sizes. Although the bias is large and the coverage probability of the credibility intervals for the model parameters is far from the nominal value, for samples with size less than 100, we intend to applie the same inference method in a real study related to the cervical intraepithelial neoplasia with sample size more than 300 patients. When we considere that specific sample size, n=300, we observe that the bias is undercontrol and the coverage probability is really close to the proposed nominal value.

Table 1. Mean of parameters estimated and SD, the mean square error (MSE) and coverage probability for each combination of sample size and generated parameters, by using the reference prior

2*Generated		2*Simulated				Mean Squared Error		Coverage Probability	
n	θ	$\hat{\theta}$	$\hat{\lambda}$	SD($\hat{\theta}$)	SD($\hat{\lambda}$)	MSE$_\theta$	MSE$_\lambda$	CP$_\theta$	CP$_\lambda$
50	0.1	0.325	1.134	0.089	0.402	0.058	0.912	0.660	0.654
	0.3	0.571	1.047	0.154	0.455	0.097	1.115	0.697	0.684
	0.5	0.600	1.575	0.177	0.484	0.041	0.415	0.910	0.872
100	0.1	0.127	1.951	0.047	0.234	0.003	0.057	0.930	0.942
	0.3	0.369	1.928	0.123	0.297	0.020	0.094	0.937	0.939
	0.5	0.544	2.017	0.154	0.307	0.026	0.095	0.968	0.965
150	0.1	0.117	1.972	0.035	0.187	0.001	0.036	0.934	0.935
	0.3	0.344	1.970	0.096	0.232	0.011	0.055	0.933	0.928
	0.5	0.542	2.006	0.135	0.254	0.020	0.065	0.946	0.945
300	0.1	0.108	1.987	0.022	0.132	0.001	0.018	0.935	0.956
	0.3	0.322	1.988	0.063	0.164	0.004	0.027	0.941	0.959
	0.5	0.530	1.996	0.102	0.186	0.011	0.035	0.946	0.955
500	0.1	0.105	1.992	0.017	0.102	0.000	0.010	0.946	0.949
	0.3	0.313	1.993	0.047	0.127	0.002	0.016	0.947	0.952
	0.5	0.519	1.997	0.079	0.145	0.007	0.021	0.949	0.951
1,000	0.1	0.1022	1.997	0.011	0.072	0.000	0.005	0.948	0.950
	0.3	0.306	1.998	0.033	0.090	0.001	0.008	0.952	0.947
	0.5	0.5087	2.000	0.054	0.102	0.003	0.011	0.948	0.946

5. Cervical Intraepithelial Neoplasia Data

Recall the CIN data presented in the introductory section. In this section we illustrate the usefulness of the CEG distribution by using the reference prior on modeling such a real data set.

Firstly, a brief descriptive analysis was made. The minimum observed time was 1 month and the maximum observed one was 47 months, approximately four years. The mean time for the cure is 6.4 months with 5.099 SD. The left panel of the Figure 3 shows the TTT plot (Barlow Campo, 1975), in order to verify the possible shape for the hazard function. If the TTT plot is concave, it indicates increase hazard, which can accommodate by a CEG distribution.

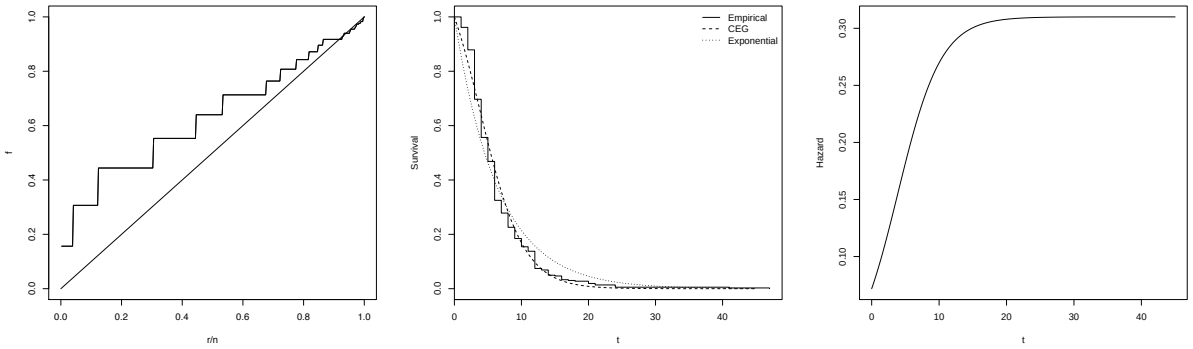

Figure 3. Left panel: TTT plot. Middle panel: estimated survival curves according to the CEG and Exponential distribution fittings over imposed on the Kaplan-Meier empirical curve. Right panel: estimated hazard function according the CEG distribution.

After initial analysis, the CEG model was fitted the data. As the exponential model is a particular case of CEG model when $\theta = 1$, this model was fitted too. The posterior samples (for both models), were generated by the Metropolis-Hastings technique, similar to the simulation study. An single chain of the dimension $100,000$ was considered for each parameter, discarting the first $20,000$ iterations to eliminate the effect of the initial values, and to avoid correlation problems, a lag with size 20 was used, resulting in a final sample $4,000$. Table 2 shows the posterior summaries for the parameters and the 95% credible intervals considering the reference.

Table 2. Posterior model summary of CEG model by considering the reference prior

2*Models	2*Parameters	2*Mean	2*SD	Percentiles			2*IC (95%)	
				25%	50%	75%		
2*CEG	θ	0.232	0.041	0.202	0.228	0.257	0.159	0.318
	λ	0.310	0.023	0.295	0.310	0.310	0.263	0.351
Exponential	λ	0.154	0.008	0.148	0.154	0.160	0.138	0.171

The convergence of the chain was verified by Geweke criterion (Geweke, 1992) which demonstrate that these criteria are satisfied for both models, as follows in Table 3. Also, the acceptance rates were obtained too.

Table 3. Geweke criteria and acceptance rate

Model	Parameter	z	p-value	Acceptance Rate
2*CEG	θ	-0.8442	0.3986	2*0.3289
	λ	0.0037	0.9971	
Exponential	λ	0.3976	0.6909	0.3692

In order to compare these models, the deviance information criterion (DIC), which is a hierarchical modelling generalization of the AIC (Akaike information criterion) and BIC (Bayesian information criterion, also known as the Schwarz criterion), was obtained. This criteria is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation which is our case.

The idea is that models with smaller DIC should be preferred to models with larger DIC which is the case of CEG model with DIC = $2,006.797$ against $2,075.498$ for the exponential model.

In addition to this criterion, it is possible to visually distinguish the model that best fits this data set. The middle panel of the Figure 3 shows the survival curves estimated empirically and by the model CEG and exponential. Clearly, we can see that the estimated survival curve for the model CEG is closer to a empirical curve than the model exponential which suggest that CEG model fits better in this case. Furthermore, in right panel of the Figure 3 we can see the increasing behavior of the estimated hazard curves for models CEG.

Figure 4 shows the marginal posterior density for unknown quantities θ and λ, left and middle panels, respectively. Also, right panel shows the scatterplot to analyze the convergence of the parameters by considering two distinct initial values.

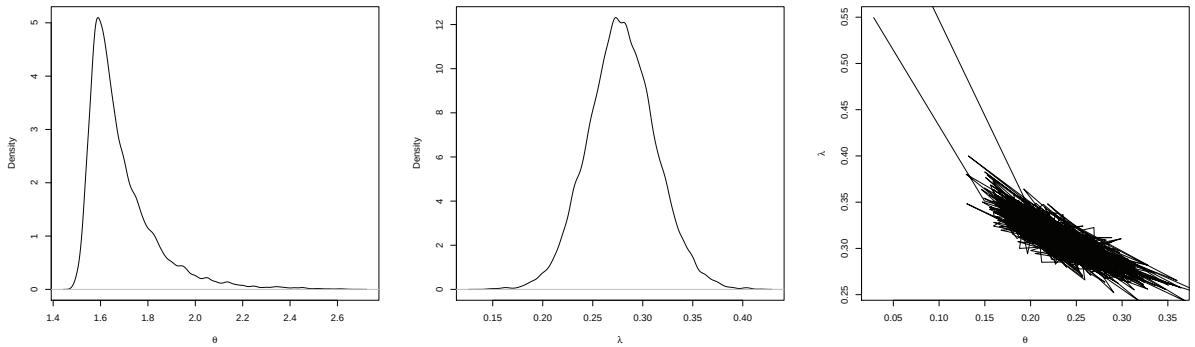

Figure 4. Marginal posteriors densities and scatterplot to analyze the convergence of the parameters by considering two distinct initial values

Considering the CEG distribution fitting, the expected cure time for cervix lesions is $E(T) = -\ln(\hat{\theta})/[\hat{\lambda}(1-\hat{\theta})] = 6.14$ months. The 95% credible interval for expected cure time ranges from 4.79 to 8.30 months. Moreover, the modal value for the time until the cure is given by $M(Y) = \frac{1}{\lambda}\ln\frac{1-\theta}{\theta} = 3.87$ months with a 95% credible interval ranging from 2.17 to 6.32 months.

6. Concluding Remarks

In this paper we presented the reference Bayesian inferential procedure for the CEG distribution. The reference Bayesian analysis is an alternative to the Jeffreys prior. In contrast to the Jeffreys, as we show that the prior is permissible then the reference prior leads to a proper posterior distribution.

After some discussion about the reference prior, we made a simulation study where we can observe the adequacy of the proposed inferential method by using different sample sizes. The simulated results show better frequentist properties for samples sizes of above 100.

In order to illustrate the usefulness and effectiveness of the Bayesian CEG model, one real data set was considered which studies the time to cure of precursors lesions of cervical cancer. The use of complementary model was extremely important in this case because it did not know the times in which each lesions were cured but the final time (in this case the maximum time), in which all had been eliminated. If we chose on an usual model for survival analysis, this important peculiarity has not been observed.

7. Proof that the Posterior Distribution is Proper

The join posterior distribution is given by

$$\pi(\theta,\lambda|y_1,\ldots,y_n) \propto \frac{\theta^n\lambda^{n-1}\exp(-\lambda\sum_{i=1}^n y_i)}{\prod_{i=1}^n\left[\exp(-\lambda y_i)(1-\theta)+\theta\right]^2}\times \left[\frac{(\theta-1)^2-2(\theta-1)\theta\log\theta+(\log\theta)^2+2(\theta-1)^3L_i(\beta)}{3(\theta-1)^3\theta^2[(1+\theta)-2L_i(\beta)]}\right]^{1/2}.$$

Note that

$$\begin{aligned} L_i(\beta) &= \sum_{k=1}^{\infty} \frac{(\theta-1)^k}{k^2} = (\theta-1) + \frac{(\theta-1)^2}{2^2} + \frac{(\theta-1)^3}{3^2} + \frac{(\theta-1)^4}{4^2} + \dots \\ &< +\frac{(\theta-1)^2}{2^2} + +\frac{(\theta-1)^4}{4^2} + \frac{(\theta-1)^6}{6^2} + \dots \\ &< (\theta-1)^2 + (\theta-1)^4 + (\theta-1)^6 + (\theta-1)^8 + \dots \\ &= (\theta-1)^2\left[1 + (\theta-1)^2 + (\theta-1)^4 + (\theta-1)^6 + \dots\right] \\ &= (\theta-1)^2 \frac{1}{1-(\theta-1)^2} \end{aligned} \tag{19}$$

as $\theta \in [0,1]$ than we can conclude that $L_i(\beta) < 0.5$. For the case when $n = 1$

$$\begin{aligned} \pi(\theta,\lambda|y_1) &\propto \frac{\theta\exp(-\lambda y_1)}{[\exp(-\lambda y_1)(1-\theta)+\theta]^2} \times \\ &\quad \left[\frac{(\theta-1)^2 - 2(\theta-1)\theta\log\theta + (\log\theta)^2 + 2(\theta-1)^3 L_i(\beta)}{3(\theta-1)^3\theta^2[(1+\theta) - 2L_i(\beta)]}\right]^{1/2} \\ &\propto \frac{\exp(-\lambda y_1)}{[\exp(-\lambda y_1)(1-\theta)+\theta]^2} \times \\ &\quad \left[\frac{1 - \frac{2\theta\log\theta}{(\theta-1)} + (\frac{\log\theta}{\theta-1})^2 + 2(\theta-1)L_i(\beta)}{3(\theta-1)[(1+\theta) - 2L_i(\beta)]}\right]^{1/2} \end{aligned}$$

but we can note that

$$\left[1 - \frac{2\theta\log\theta}{\theta-1} + \left(\frac{\log\theta}{\theta-1}\right)^2 + 2(\theta-1)L_i(\beta)\right]^{1/2} \leq \left(1 + \left(\frac{\log\theta}{\theta-1}\right)^2\right) \tag{20}$$

and

$$\begin{aligned} \lim_{\theta\to 0}\left(\frac{\log\theta}{\theta-1}\right)^2 &= +\infty \\ \lim_{\theta\to 1}\left(\frac{\log\theta}{\theta-1}\right)^2 &= 1. \end{aligned}$$

Between this and that,

$$\left[1 - \frac{2\theta\log\theta}{\theta-1} + \left(\frac{\log\theta}{\theta-1}\right)^2 + 2(\theta-1)L_i(\beta)\right]^{1/2} \geq 1. \tag{21}$$

Finnaly, we can see that

$$\begin{aligned} \pi(\theta,\lambda|y_1) &< \frac{\exp(-\lambda y_1)}{[\exp(-\lambda y_1)(1-\theta)+\theta]^2}\left[\frac{1}{3(\theta-1)[(1+\theta) - 2L_i(\beta)]}\right]^{1/2} \\ &< \frac{\exp(-\lambda y_1)}{[\exp(-\lambda y_1)(1-\theta)+\theta]^2}\left[\frac{1}{3\theta(1-\theta)}\right]^{1/2} \\ &< \frac{\exp(\lambda y_1)}{\sqrt{(1-\theta)^5\theta}}. \end{aligned}$$

Then

$$\int_{\Theta}\int_{\Lambda}\pi(\theta,\lambda|y_1)d\lambda d\theta < \int_{\Theta}\int_{\Lambda}\frac{\exp(\lambda y_1)}{\sqrt{(1-\theta)^5\theta}} = +\infty$$

i.e., the join posterior distribution is proper for $n = 1$.

Now consider the case where $n > 1$ and take $y_{(n)} = \max\{y_1, \ldots, y_n\}$ and $y_{(1)} = \min\{y_1, \ldots, y_n\}$, then we have

$$\begin{aligned}\pi(\theta,\lambda|y_1,\ldots,y_n) \quad \leq \quad & \frac{\theta^{n-1}\lambda^{n-1}\exp(-\lambda y_{(1)}^n)}{[\exp(-\lambda y_{(n)})(1-\theta)+\theta]^{2n}} \times \\ & \left[\frac{1-\frac{2\theta\log\theta}{(\theta-1)}+(\frac{\log\theta}{\theta-1})^2+2(\theta-1)L_i(\beta)}{3(\theta-1)[(1+\theta)-2L_i(\beta)]}\right]^{1/2}.\end{aligned}$$

By considering (19), (20) and (21)

$$\begin{aligned}\pi(\theta,\lambda|y_1,\ldots,y_n) \quad < \quad & \frac{\theta^{n-1}\lambda^{n-1}\exp(-\lambda y_{(1)}^n)}{[\exp(-\lambda y_{(n)})(1-\theta)+\theta]^{2n}}\left[\frac{1}{3\theta(1-\theta)}\right]^{1/2} \\ < \quad & \frac{\theta^{n-1/2}(1-\theta)^{-1/2}\lambda^{n-1}\exp(-\lambda y_{(1)}^n)}{[\exp(-\lambda y_{(n)})(1-\theta)+\theta]^{2n}}.\end{aligned}$$

Note that

$$\begin{gathered}0<\theta<1 \\ 0<\exp(-\lambda y_{(n)})<1 \\ 0<\exp(-\lambda y_{(n)})(1-\theta)+\theta<1 \\ 0<[\exp(-\lambda y_{(n)})(1-\theta)+\theta]^{2n}<1.\end{gathered}$$

So

$$\frac{\theta^{n-1/2}(1-\theta)^{-1/2}\lambda^{n-1}\exp(-\lambda y_{(1)}^n)}{[\exp(-\lambda y_{(n)})(1-\theta)+\theta]^{2n}} > \theta^{n-1/2}(1-\theta)^{-1/2}\lambda^{n-1}\exp(-\lambda y_{(1)}^n),$$

and

$$\int_{\Theta}\int_{\Lambda}\theta^{n-1/2}(1-\theta)^{-1/2}\lambda^{n-1}\exp(-\lambda y_{(1)}^n)d\lambda d\theta = \frac{\Gamma(n+1/2)\sqrt{\pi}}{y_{(1)}^{n(n+1)}} > 0.$$

But

$$\int_{\Theta}\int_{\Lambda}\pi(\theta,\lambda|y_1,\ldots,y_n)d\lambda d\theta < \int_{\Theta}\int_{\Lambda}\frac{\theta^{n-1/2}(1-\theta)^{-1/2}\lambda^{n-1}\exp(-\lambda y_{(1)}^n)}{[\exp(-\lambda y_{(n)})(1-\theta)+\theta]^{2n}}d\lambda d\theta,$$

$$\int_{\Theta}\int_{\Lambda}\frac{\theta^{n-1/2}(1-\theta)^{-1/2}\lambda^{n-1}\exp(-\lambda y_{(1)}^n)}{[\exp(-\lambda y_{(n)})(1-\theta)+\theta]^{2n}}d\lambda d\theta > \frac{\Gamma(n+1/2)\sqrt{\pi}}{y_{(1)}^{n(n+1)}},$$

and

$$\int_{\Theta}\int_{\Lambda}\pi(\theta,\lambda|y_1,\ldots,y_n)d\lambda d\theta < \frac{\Gamma(n+1/2)\sqrt{\pi}}{y_{(1)}^{n(n+1)}}.$$

Also, we can conclude that the join posterior is proper for $n > 1$ which show us that the posterior distribution is proper for all cases.

References

Adamidis, K., & Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statistics Probability Letters, 39*, 35C42. https://doi.org/10.1016/S0167-7152(98)00012-1

Barlow, R. E., & Campo, R. A. (1975). Total time on test processes and applications to failure data analysis. *Reliability and Fault Tree Analysis,* 451C481.

Basu, A., & Klein, J. (1982). Some recent development in competing risks theory. *In Survival Analysis, Edited by Crowley, J. and Johnson, R. A., Hayward, 1*, 216C229. https://doi.org/10.1214/lnms/1215464851

Berger, J. (1985). *Statistical decision theory and Bayesian analysis.* New York: Springer-Verlag. https://doi.org/10.1007/978-1-4757-4286-2

Berger, J. O., & Bernardo, J. M. (1992a). On the development of reference priors. *Bayesian Statistics, 4*. https://doi.org/10.1016/0378-3758(93)00083-T

Berger, J. O. & Bernardo, J. M. (1992b). Ordered group reference priors with applications to a multinomial problem. *Biometrika, 79*, 25C37.

Bernardo, J. (2005). *Reference analysis, 25*.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B, 41*, 113C147.

Datta, G. S., & Ghosh, M. (1995). Hierarchical Bayes estimators of the error variance in one-way ANOVA models. *J. Statist. Planning and Inference, 45*, 399C411.

Datta, G. S., & Ghosh, M. (1996). On the invariance of noninformative priors. *Ann. Statist., 24*, 141C159.

Ferenczy, A. (1982). *Cervical Intraepithelial Neoplasia,* 156C177. Springer New York, New York, NY.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In Bayesian Statistics 4 (eds. J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), *Oxford: Oxford University Press,* 169C193.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A,* 186, 435C461. https://doi.org/10.1098/rspa.1946.0056

Lane, D. & Sudderth, W. D. (1984). Coherent predictive inference. *Sankhya A, 46,* 166C185.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* John Wiley and Sons, Hoboken.

Lima, E. G., Lima, D. B. S., ..., & Fernandes, J. V. (2013). *Knowledge about HPV and screening of cervical cancer among women from the metropolitan region of Natal, Brazil.* ISRN Obstetrics and Gynecology. https://doi.org/10.1155/2013/930479

Louzada, F., Roman, M., & Gancho, V. G. (2011). The complementary exponential geometric distribution: Model, properties and a comparison with its counterpart. *Computacional Statistics and Data Analysis, 55*, 2516C2524.

Louzada-Neto, F. (1999). Poly-hazard regression models for lifetime data. *Biometrics, 55*, 1121C1125.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1C32. https://doi.org/10.2307/1914288

Paulino, D., Turkman, A., & Murteira, B. (2003). *Estatistica Bayesiana.* Fundaćão Calouste Gulbenkian, Lisboa.

Ross, S. M. (2009). *A First Course in Probability.* Prentice Hall.

Copyrights