Bayesian analysis of survival data under Generalized Extreme Value distribution with application in cure rate model

Dooti Roy[1], Vivekananda Roy[2] and Dipak Dey[1]

[1]Department of Statistics,University of Connecticut
[2]Department of Statistics and Statistical Laboratory, Iowa State University

December 11, 2014

Table of Contents

Introducing the GEV distribution

Generalized Extreme Value Distribution

GEV arise as limiting distributions for maximums or minimums (extreme values) of a sample of independent, identically distributed random variables, as the sample size increases.

Suppose $Y_1, Y_2, \ldots$ is a sequence of independent and identically distributed random variables each having the distribution function $F(y)$.

Let $M_n = \max\{Y_1, Y_2, \ldots, Y_n\}$ and $m_n = \min\{Y_1, Y_2, \ldots, Y_n\}$. In our paper, we consider limiting distributions of both Maxima (M_n) and Minima (m_n). We name them **MGEV** and **mGEV** from here on.

Fisher-Tippett Theorem: If there exists sequences $a_n > 0$, b_n, and a non degenerate distribution G, so that

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G(x) \tag{1}$$

then G is a GEV distribution.

Fisher-Tippett theorem shows that the only family of distributions which could be considered as the asymptotic limit distribution of the standardized maxima M_n is Generalized Extreme Value (GEV) distribution.

Generalized Extreme Value Distribution

The generalized extreme value (GEV) distribution combines the Gumbel, Frechet and Weibull families also known as type I, II and III extreme value distributions.

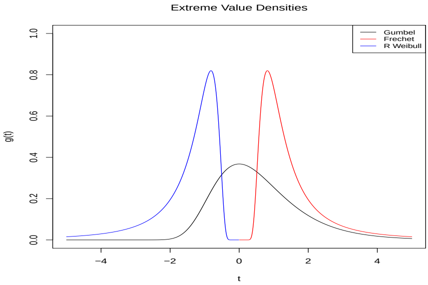


Figure: Frechet (red), Gumbel (black) and Reversed Weibull (blue) distributions.

Cumulative Density Functions of GEV

Cumulative distribution function of the MGEV distribution is given by:

$$G_\xi(x) = \begin{cases} \exp[-(1+\xi\frac{x-\mu}{\sigma})_+^{-\frac{1}{\xi}}] & \text{if } \xi > 0 \text{ or } \xi < 0 \\ \exp[-\exp(-\frac{x-\mu}{\sigma})] & \text{if } \xi = 0, \end{cases}$$

Cumulative distribution function of the mGEV distribution is given by:

$$G_\xi(x) = \begin{cases} 1-\exp[-(1+\xi\frac{x-\mu}{\sigma})_+^{\frac{1}{\xi}}] & \text{if } \xi > 0 \text{ or } \xi < 0 \\ 1-\exp[-\exp(\frac{x-\mu}{\sigma})] & \text{if } \xi = 0, \end{cases}$$

where $\mu \in R$, $\sigma \in R^+$, and $\xi \in R$ are the location, scale, and shape parameters respectively.

Hazard Functions of GEV

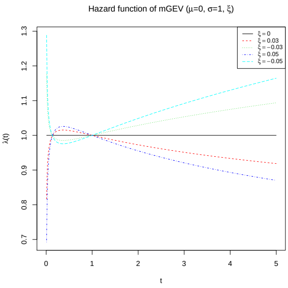
Hazard function of the $MGEV(0,1,\xi)$ is given by:

$$\lambda_M(t|\xi) = \begin{cases} \frac{1}{t(1+\xi\log t)_+^{\frac{1}{\xi}+1}[\exp(1+\xi\log t)_+^{-\frac{1}{\xi}}-1]} & \text{if } \xi \neq 0 \\ \frac{1}{t^2[\exp(\frac{1}{t})-1]} & \text{if } \xi = 0. \end{cases}$$

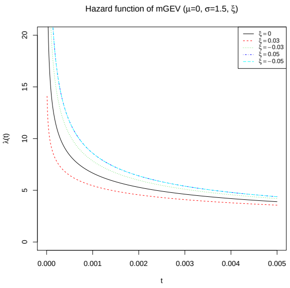
Hazard function of the $mGEV(0,1,\xi)$ is given by:

$$\lambda_m(t|\xi) = \begin{cases} \frac{1}{t}(1+\xi\log t)_+^{\frac{1}{\xi}-1} & \text{if } \xi \neq 0 \\ 1 & \text{if } \xi = 0. \end{cases}$$

Flexibility of Hazard Function Plot for mGEV



(a)

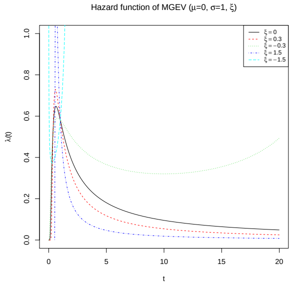


(b)

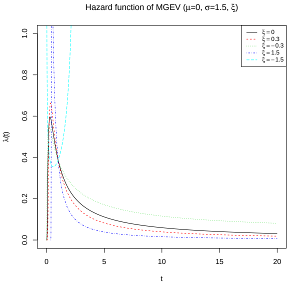
Figure: 1(a) Hazard functions of the generalized extreme value distribution for mGEV($\mu = 0, \sigma = 1, \xi$) for different values of ξ.
1(b) Hazard functions of the generalized extreme value distribution for mGEV($\mu = 0, \sigma = 1.5, \xi$) for different values of ξ .

Note that when $\xi = 0$, this is the hazard from of Weibull$(1, 1)$

Hazard Function Plot of MGEV



(a)



(b)

Figure: 2(a) Hazard functions of the generalized extreme value distribution for MGEV($\mu = 0, \sigma = 1, \xi$) for different values of ξ.
2(b) Hazard functions of the generalized extreme value distribution for MGEV($\mu = 0, \sigma = 1.5, \xi$) for different values of ξ .

Motivation of our work

Typical Characteristic of Survival Data

- Incorporates a cure fraction
- Non monotone hazard functions
- Often is highly skewed
- Has some sort of censoring involved

Past work and our motivation

- Starting from the mixture model introduced by Berkson and Gage (1952) around 60 years ago, several models have been studied in the past. (Farewell (1982), Kuk and Chen (1992), Mallar and Zhou (1996), Peng and Dear (2000), Sy and Taylor (2000) and Roy and Dey (2014))
- In Chen et al. (1999), the authors proposed a proportional hazards model through the cure rate parameter, using Gamma and Weibull distributions, which are quite popular with monotone hazard rates.
- However, the hazard function is often not monotone and is either upside-down shaped or bathtub shaped or a combination of both shapes. These are experienced in relapsed leukemia or lymphoma patients where after initial risk of survival, the patient achieves remission before another relapse.
- We propose both MGEV and mGEV distributions to model log T, where T denotes the failure time/survival time of an individual.

Model setting

- Suppose that we have n subjects
- N_i: unobserved no. of clonogenic cells for the ith subject. Assume N_i's i.i.d. $\mathrm{P}(\theta_i)$, $i = 1, \ldots, n$.
- t_i: failure time for subject i; t_i is right-censored.
- c_i: the censoring time
- observed $y_i = \min(t_i, c_i)$; censoring indicator $\delta_i = I(t_i < c_i)$
- The observed data: $(n, \mathbf{y}, \delta)$; $\mathbf{y} = (y_1, \cdots, y_n)$, $\delta = (\delta_1, \cdots, \delta_n)$.
- $\mathbf{N} = (N_1, \cdots, N_n)$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$.
- $\mathbf{D} = (n, \mathbf{y}, \delta, \mathbf{N})$: The complete data; $\mathbf{N}$ is the unobserved.
- $f(y_i|\xi)$, $S(y_i|\xi)$: density and survival functions of y_i.
- ξ: the shape parameter in the standard MGEV or mGEV distribution

Complete Data Likelihood

The complete data likelihood function of the parameters $(\boldsymbol{\theta}, \xi)$ can be written as:

$$L(\boldsymbol{\theta}, \xi|D) = \prod_{i=1}^{n} S(y_i|\xi)^{N_i-\delta_i}(N_i f(y_i|\xi))^{\delta_i} \times \exp\left\{\sum_{i=1}^{n}(N_i \log(\theta_i) - \log(N_i!) - \theta_i)\right\}. \tag{2}$$

After the covariates are incorporated through θ, the complete-data likelihood of (β, ξ)

$$L(\boldsymbol{\beta}, \xi|\mathbf{D}) = \prod_{i=1}^{n} S(y_i|\xi)^{N_i-\delta_i}(N_i f(y_i|\xi))^{\delta_i} \times \exp\left\{N_i \mathbf{x}_i'\boldsymbol{\beta} - \log(N_i!) - \exp(\mathbf{x}_i'\boldsymbol{\beta})\right\}. \tag{3}$$

Let the prior distribution for (β, ξ) is $\pi(\beta, \xi)$, then the posterior distribution $\pi(\beta, \xi|\mathbf{D}_{obs})$ satisfy this:

$$\pi(\beta, \xi|\mathbf{D}_{obs}) \propto \sum_{\mathbb{N}} L(\beta, \xi|\mathbf{D})\pi(\beta, \xi). \tag{4}$$

Propriety of Posterior Distribution

Conditions for a good Bayes Estimation method

For Bayesian method to be useful and inference to be meaningful, a proper posterior distribution is necessary. Often finding closed form of posterior is a challenge. The second issue is with prior elicitation. There are cases when not much is known about the prior except a range. Finding closed form posterior especially under a wide class of objective priors makes our method readily applicable.

Conditions for MGEV

Let t_i be the failure time for the i'th subject. Let c_i be the censoring time. Then $y_i = \min(t_i, c_i); i = 1, .., n$.
Let $\log y_i \sim \mathrm{MGEV}(\mu = 0, \sigma = 1, \xi)$. We use an diffused uniform prior on β, that is, $\pi(\beta) \propto 1$, and the prior on ξ is $\pi(\xi) = 1/(b-a) I_{[a,b]}(\xi)$,where $a < 0 < b$, are fixed numbers.
We also assume that $\pi(\beta, \xi) = \pi(\beta) \cdot \pi(\xi)$.

Theorem 1

Let $\mathbf{X}^*$ be an $n \times k$ matrix with rows $\delta_i \mathbf{x}_i'$. If the following two conditions hold:

1. *$\mathbf{X}^*$ is of full rank,*
2. *For every i with $\delta_i = 1$,*

$$\exp(-1/b) < y_i < \exp(-1/a) \text{ ; or } y_i \geq \max(\exp(-1/a), e), \quad (5)$$

then the posterior distribution given in (4) is proper.

Corollary to Theorem 1

In the special case when $b = -a = 1$, that is, when $\pi(\xi) = (1/2)I_{[-1,1]}(\xi)$, we have the following corollary.

Corollary 1

Let $\mathbf{X}^*$ *be an* $n \times k$ *matrix with rows* $\delta_i \mathbf{x}_i'$*. If the following two conditions hold:*

1. $\mathbf{X}^*$ *is of full column rank,*
2. *For every* i *with* $\delta_i = 1$, $y_i > 1/e$,

then the posterior distribution given in (4) is proper.

Conditions for MGEV cont.

Now we use the following prior on ξ:
$\pi(\xi) = c\exp(-|\xi|/2), -a < \xi < a, a > 0$, along with the previously used uniform prior on β. Simple calculations show that $c = 4(1 - \exp(-a/2))$. The following theorem provides sufficient conditions for posterior propriety in this case. Our goal is to show that the posterior attains propriety under different reasonable priors.

Theorem 2

Let $\mathbf{X}^*$ be an $n \times k$ matrix with rows $\delta_i \mathbf{x}_i'$. If these two conditions hold:

1. *$\mathbf{X}^*$ is of full column rank,*
2. *For every i with $\delta_i = 1$,*

$$\exp(-1/a) < y_i < \exp(1/a) \ \text{ or } y_i \geq \max(\exp(1/a), e), \qquad (6)$$

then the posterior distribution given in (4) is proper.

Conditions for mGEV

Consider the uniform prior for ξ on (a, b), that is,
$\pi(\xi) = 1/(b-a) I_{[a,b]}(\xi),\ a < 0 < b$, and $\pi(\beta) \propto 1$.

Theorem 3

Let $\mathbf{X}^*$ *be an* $n \times k$ *matrix with rows* $\delta_i \mathbf{x}_i'$. *If the following two conditions hold:*

1. $\mathbf{X}^*$ *is of full column rank,*
2. *For every* i *with* $\delta_i = 1$, $\exp(-1/b) < y_i < \exp(-1/a)$,

then the posterior distribution given in (4) is proper.

Implementation and Real Data Analysis

Comparing with Currently used models

Now that we have shown our model has the desirable properties of flexible hazard and proper posterior, naturally the next question is " How it compares with the current existing models?"

- The Weibull distribution, having exponential and Rayleigh as special sub-models, is a very popular distribution for modeling lifetime data.
- For non-monotone data, distributions like Exponentiated Weibull or Beta Extended Weibull have better performance.
- We conducted a simulation study comparing performance of Weibull, MGEV, mGEV and Exponentiated Weibull models with each other. The simulation was implemented using statistics cluster.

Bayesian Model Selection

For model selection we use:

- Log pseudo marginal likelihood or LPML $= \sum_{i=1}^{n} \log(\widehat{CPO_i})$, where $\widehat{CPO_i}$ is the Monte Carlo approximation of CPO_i defined in Dey et al.(1997). The model with larger LPML provides better fit to the data.
- Deviance information criterion (DIC) proposed by Spiegelhalter (2002). The model with lower DIC is preferred.
- A plot of difference of $log(\widehat{CPO_i})$ from two competing models, for each posterior sample to gauge the superiority of a model.

Comparison of Model Performance

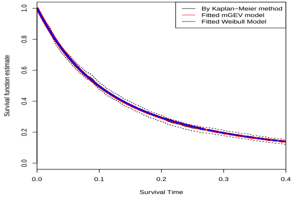
Table: Model Fitting Comparison of mGEV, MGEV, Weibull and Exponentiated Weibull distributions.

		Fitted		
Generated		mGEV	MGEV	Weibull
mGEV	*LPML*	1742.649	1562.961	1301.597
	DIC	-3485.159	-3126.306	-2603.715
MGEV	*LPML*	174.949	695.359	-392.934
	DIC	-348.746	-1390.719	798.289
Weibull	*LPML*	1611.027	1590.493	1595.899
	DIC	-3222.066	-3181.773	-3192.217
Exponentiated Weibull	*LPML*	424.115	476.069	482.26
	DIC	-850.013	-952.832	-964.704

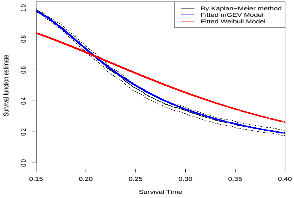
Weibull Distribution as a special case of mGEV

Lemma 1

If $T \sim$ *Weibull*(α, λ)*, then* $\log T \sim$ *mGEV*$(\mu = \log(\lambda), \sigma = 1/\alpha, \xi = 0)$.



(a)



(b)

Figure: 1(a) Estimated Kaplan-Meier Curves for the simulated model Weibull(α=1.03,λ=1) and the fitting models mGEV($\mu = 0, \sigma = 1, \xi$) (the red line) and the Weibull(α,λ=1) (the blue line).
1(b) Estimated Kaplan-Meier Curves for the simulated model mGEV($\mu = 0, \sigma = 1, \xi = 0.5$) and the proposed model Weibull(α,λ=1) (the red line) and the mGEV($\mu = 0, \sigma = 1, \xi$)(the blue line).

Rayleigh and Exponential Distributions

Lemma 2

If $T \sim \textit{Rayleigh}(\lambda)$*, then* $\log T \sim \textit{mGEV}(\mu = \log(\sqrt{2}\lambda), \sigma = \frac{1}{2}, \xi = 0)$.

Lemma 3

If $T \sim \textit{Exponential}(\lambda)$*, then* $\log T \sim \textit{mGEV}(\mu = \log(\lambda), \sigma = 1, \xi = 0)$.

Both *Rayleigh* and *Exponential* distributions are also **special** cases of the mGEV distributions.

Glioblastoma Multiforme (GBM) Data

- Data obtained from National Cancer Institute SEER database.
- Glioblastoma multiforme is one of deadliest form of cancers with an extremely small surviving fraction.
- Medical Research states in adults only 10% patients survive beyond 5 years.
- Smoll, Schaller and Gautschi (2012) discussed the data for the first time, they found a cure fraction of approximately 12% in young adults (typically younger than 40 years).

Glioblastoma Multiforme (GBM) Data

- The data has a patient population of 1725 subjects who are diagnosed with only GBM cancer between the year 1970 and 2004.
- A patient surviving for more than 10 years after diagnosis is considered cured

Table: Summary of the GBM Cancer Data.

Survival time(months)	Status(freq)	Age(years)	Gender(freq)	Radiation(freq)	Marital status
Median 18	Censored 182	Mean 31.4	Male 1053	Had 1453	Married 876
IQR 32	Death 1543	10	Female 672	None 333	Other 897

We consider the mGEV model and perform Bayesian analysis using diffused prior on β and uniform prior on ξ. To facilitate comparison we also fit the Weibull model.

Kaplan-Meier Plot

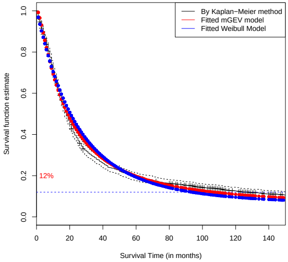


Figure:
3 Estimated survival curves for the GBM data by Kaplan-Meier method(solid line is the estimate, dashed lines are 95% confidence band for the survival function), the fitted Weibull Model(the blue line) and the proposed mGEV model(the red line).

- The Kaplan-Meier curve in Figure 3 shows a clear plateau, so a cure rate model seems to be appropriate.
- From the plot, the empirical cure rate is around 12% which is consistent with the findings in Smoll et al(2012).

Estimation of Parameters

Table: GBM Data: Posterior Estimates of the mGEV Model Parameters with covariates.

Variable	Posterior mean	Posterior SD	95% HPD interval
μ	4.315	0.087	(4.145, 4.485)
σ	1.203	0.052	(1.102, 1.309)
ξ	0.186	0.017	(0.153, 0.216)
Age	0.033	0.002	(0.028, 0.037)
Had Radiation	-0.096	0.062	(-0.218, 0.018)
Marital Status	-0.079	0.056	(-0.188, 0.034)
Gender	0.221	0.050	(0.127, 0.325)

Table: GBM Data: Posterior Estimates of the Weibull Model Parameters with covariates.

Variable	Posterior mean	Posterior SD	95% HPD interval
λ	66.216	0.449	(65.368, 66.870)
α	1.104	0.020	(1.065, 1.143)
Age	0.031	0.003	(0.025, 0.035)
Had Radiation	-0.067	0.074	(-0.193, 0.098)
Marital Status	-0.076	0.055	(-0.173, 0.042)
Gender	0.225	0.057	(0.113, 0.335)

Model Comparison

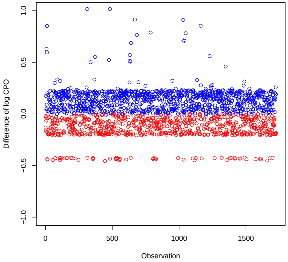


Figure: 4. Plot of difference of the log CPO between mGEV and Weibull Model for the GBM cancer data.

60% of the points lie above zero (blue dots).

Table: Model Comparison between Fitted mGEV distribution and Fitted Weibull distribution.

Fitted Model	DIC	LPML
mGEV(μ, σ, ξ)	14177.23	-7088.581
Weibull(α, λ)	14299.21	-7150.145

Our model performs much better then the Weibull model.

Summary

- We propose modeling the log survival time with censoring as a GEV distribution.
- We show through the hazard and survival plots, that the proposed model achieves a lot of flexibility and thus has an obvious advantage over the commonly used Weibull models.
- We have implemented a new form of survival modeling for right-censored data with a cure fraction using the generalized extreme value distribution.
- We also establish sufficient conditions for the propriety of the posterior distribution when an diffused uniform prior is used for the regression coefficients through cure rates.
- Our model outperforms Weibull models when applied to the GBM data. We also performed similar analysis for several other cancer data sets and each time our model proved better.

Thank you! Questions?