Biostatistics (2014), **0**, 0, *pp.* 1–9
doi:10.1093/biostatistics/biostats˙supplementary

Supplementary materials for "Bayesian analysis of survival data under generalized extreme value distribution with application in cure rate model"

DOOTI ROY, VIVEKANANDA ROY, DIPAK K. DEY

1. Proofs of Theorems and Lemmas

1.1 *Proof of Theorem 1*

Proof. By summing out the unobserved latent vector $\mathbf{N}$, the complete-data likelihood given in (3.3) reduces to

$$\sum_{\mathbf{N}} L(\beta, \xi|\mathbf{D}) = \prod_{i=1}^{n} (\theta_i f(y_i|\xi))^{\delta_i} \exp\left\{-\theta_i(1 - S(y_i|\xi))\right\}.$$

Notice that if $\delta_i = 0$, $(\theta_i f(y_i|\xi))^{\delta_i} \exp\left\{-\theta_i(1 - S(y_i|\xi))\right\} = \exp\left\{-\theta_i(1 - S(y_i|\xi))\right\} \leqslant 1$. On the other hand, when $\delta_i = 1$, we will show that there exists a constant M such that

$$(\theta_i f(y_i|\xi))^{\delta_i} \exp\left\{-\theta_i(1 - S(y_i|\xi))\right\} \leqslant g_i(\xi)M, \tag{1.1}$$

where $g_i(\xi) = 1/(1 + \xi \log y_i)^{\frac{1}{\xi}+1}$. The left side of (1.1) can be rewritten as

$$\begin{aligned} &\tfrac{f(y_i|\xi)}{1-S(y_i|\xi)} \cdot (\theta_i(1 - S(y_i|\xi)) \exp\left\{-\theta_i(1 - S(y_i|\xi))\right\} \\ &= \tfrac{1}{y_i(1+\xi \log y_i)^{\frac{1}{\xi}+1}} (\theta_i(1 - S(y_i|\xi)) \exp\left\{-\theta_i(1 - S(y_i|\xi))\right\}. \end{aligned} \tag{1.2}$$

Let $g(z) = ze^{-z}$, for $z > 0$, then it can be shown that (see Chen *and others* (1999)) there exists a constant $g_0 > 0$ such that

$$g(z) \leqslant g_0, \forall z > 0. \tag{1.3}$$

Using (1.3), we see that (1.2) is less than or equal to $y_i^{-1} g_0 g_i(\xi)$. Thus taking $M = g_0 \max_{i:\delta_i=1}\{y_i^{-1}\}$, we obtain (1.1).

Since $\mathbf{X}^*$ is of full rank, there must exist k linearly independent row vectors $\mathbf{x}'_{i_1}, \mathbf{x}'_{i_2}, \ldots, \mathbf{x}'_{i_k}$ such that $\delta_{i_1} = \delta_{i_2} = \ldots = \delta_{i_k} = 1$. To prove that the posterior given in (3.4) is proper, we only need to show that

$$\int_s^t \int_{\mathbb{R}^k} \sum_{\mathbf{N}} L(\beta, \xi|D)\pi(\xi) d\beta d\xi < \infty. \tag{1.4}$$

Note that the expression in (1.4) can be written as

$$\begin{aligned}
\frac{1}{(t-s)} \int_s^t \int_{\mathbb{R}^k} \sum_{N} L(\beta, \xi|D) d\beta d\xi &= \frac{1}{(t-s)} \int_s^t \int_{\mathbb{R}^k} \prod_{i=1}^{n} (\theta_i f(y_i|\gamma))^{\delta_i} \exp\left\{-\theta_i (1 - S(y_i|\gamma))\right\} d\beta d\xi \\
&\leqslant \frac{1}{(t-s)} \int_s^t \int_{\mathbb{R}^k} \Big(\prod_{i:\delta_i=0} 1 \Big) \\
&\quad \Big\{ \prod_{j=1}^{k} f(y_{i_j}|\xi)\theta_{i_j} \exp\{-\theta_{i_j}(1 - S(y_{i_j}|\gamma))\} \Big\} \\
&\quad \Big\{ \prod_{i:\delta_i=1, i\neq i_j, j=1,\ldots,k} g_i(\xi) M \Big\} d\beta d\xi \\
&= \frac{1}{(t-s)} \int_s^t \int_{\mathbb{R}^k} M^{d-k} \left[\prod_{i:\delta_i=1, i\neq i_j, j=1,\ldots,k} g_i(\xi) \right] \\
&\quad \prod_{j=1}^{k} f(y_{i_j}|\xi)\theta_{i_j} \exp\{-\theta_{i_j}(1 - S(y_{i_j}|\gamma))\} d\beta d\xi,
\end{aligned} \tag{1.5}$$

where $\theta_{i_j} = \exp(\mathbf{x}'_{i_j}\beta)$ and $\mathbb{R}^k$ denotes the k-dimensional Euclidean space. Now we make the transformation $u_j = x'_{i_j}\beta$ for $j = 1, 2, \ldots, k$. This is a one-to-one linear transformation from β to $\mathbf{u} = (u_1, \ldots, u_k)'$. Thus (1.5) is proportional to

$$\begin{aligned}
&\int_s^t \int_{\mathbb{R}^k} \left[\prod_{i:\delta_i=1, i\neq i_j, j=1,\ldots,k} g_i(\xi) \right] \prod_{j=1}^{k} f(y_{i_j}|\xi) \exp\{u_j - (1 - S(y_{i_j}|\xi)) \exp(u_j)\} d\mathbf{u} d\xi \\
&= \int_s^t \left[\prod_{i:\delta_i=1, i\neq i_j, j=1,\ldots,k} g_i(\xi) \right] \left\{ \prod_{j=1}^{k} f(y_{i_j}|\xi) \int_R \exp\{u_j - (1 - S(y_{i_j}|\xi)) \exp(u_j)\} du_j \right\} d\xi \\
&= \int_s^t \left[\prod_{i:\delta_i=1, i\neq i_j, j=1,\ldots,k} g_i(\xi) \right] \prod_{j=1}^{k} \frac{f(y_{i_j}|\xi)}{1 - S(y_{i_j}|\xi)} d\xi.
\end{aligned} \tag{1.6}$$

The last equality holds, since $\int_R \exp\{u_j - (1 - S(y_{i_j}|\xi))\exp(u_j)\}du_j = 1/(1 - S(y_{i_j}|\xi))$. Notice that

$$\frac{f(y_{i_j}|\xi)}{1 - S(y_{i_j}|\xi)} = \frac{1}{y_{i_j}(1 + \xi \log y_{i_j})^{\frac{1}{\xi}+1}} \leqslant M_1 \cdot g_{i_j}(\xi),$$

where $M_1 = \max_{i:\delta_i=1} y_i^{-1}$. Ignoring the constant, (1.6) is less than or equal to $\int_s^t \prod_{i:\delta_i=1} g_i(\xi)d\xi$. Thus we only need to prove that $g(\xi) := \prod_{i:\delta_i=1} g_i(\xi)$ is bounded in $[s, t]$. We show that for every i with $\delta_i = 1, g_i(\xi)$ is bounded in $[s, t]$. Recall that (3.5) is in force. We consider two cases of this condition separately.

- If $\exp(-1/t) < y_i < \exp(-1/s)$, $-1/\log y_i$ is either less than s or greater than t, $s < 0 < t$. So $1 + \xi \log y_i$ can not be zero for any $\xi \in [s, t]$. Hence $g_i(\xi)$ is bounded in $[s, t]$.

- On the other hand, if $y_i \geqslant \max(\exp(-1/s), e)$, then $1 + \xi \log y_i$ can be zero when $\xi = -1/\log y_i$. But

$$\lim_{\xi \to -\frac{1}{\log y_i}} g_i(\xi) = 0^{\log y_i - 1} = \begin{cases} 1 & if\ y_i = e \\ 0 & o.w. \end{cases}.$$

  Hence $g_i(\xi)$ is bounded on $[s, t]$.

So $g(\xi) = \prod_{i:\delta_i=1} g_i(\xi)$ is bounded in $[s, t]$, thus we have $\int_s^t g(\xi)d\xi < \infty$. □

1.2 *Proof of Theorem 3*

Proof.

By summing out the unobserved latent vector $\mathbf{N}$, the complete-data likelihood given in (3.3) reduces to

$$\sum_{\mathbf{N}} L(\beta, \xi|\mathbf{D}) = \prod_{i=1}^{n} (\theta_i f(y_i|\xi))^{\delta_i} \exp\{-\theta_i(1 - S(y_i|\xi))\}.$$

If $\delta_i = 0$, $(\theta_i f(y_i|\xi))^{\delta_i} \exp\{-\theta_i(1 - S(y_i|\xi))\} = \exp\{-\theta_i(1 - S(y_i|\xi))\} \leqslant 1$. When $\delta_i = 1$, we will show that there exists a constant M such that

$$(\theta_i f(y_i|\xi))^{\delta_i} \exp\{-\theta_i(1 - S(y_i|\xi))\} \leqslant h_i(\xi)M, \tag{1.7}$$

where $h_i(\xi) = 1/(1+\xi \log y_i)$. The left side of (1.7) can be rewritten as

$$\begin{aligned} &\tfrac{f(y_i|\xi)}{1-S(y_i|\xi)} \cdot (\theta_i(1-S(y_i|\xi)) \exp\{-\theta_i(1-S(y_i|\xi))\} \\ &= \tfrac{\frac{1}{y_i}(1+\xi \log y_i)^{\frac{1}{\xi}-1} \exp[-(1+\xi \log y_i)^{\frac{1}{\xi}}]}{1-\exp[-(1+\xi \log y_i)^{\frac{1}{\xi}}]} (\theta_i(1-S(y_i|\xi)) \exp\{-\theta_i(1-S(y_i|\xi))\}, \end{aligned} \tag{1.8}$$

Let $\kappa_1(z) = ze^{-z}$, and $\kappa_2(z) = \frac{ze^{-z}}{1-e^{-z}}$ for $z > 0$. It can be shown that there exists a constant $h_0 > 0$ such that

$$\kappa_i(z) \leqslant h_0, \forall z > 0 \; i = 1, 2. \tag{1.9}$$

Using (1.9), (1.8) is less than or equal to $y_i^{-1} h_0^2 h_i(\xi)$. Thus taking $M = h_0^2 \max_{i:\delta_i=1}\{y_i^{-1}\}$, we obtain (1.7).

Doing similar calculations as in the proof of Theorem 1, we see that the posterior density in (3.4) is proper if $h_i(\xi)$ is bounded in $[s,t]$ for all i with $\delta_i = 1$. Since for all i with $\delta_i = 1$, $\exp(-1/t) < y_i < \exp(-1/s)$, $1+\xi \log y_i$ can not be 0 when $\xi \in [s,t]$. So $h(\xi) := \prod_{i:\delta_i=1} h_i(\xi)$ is bounded in $[s,t]$. Thus $\int_s^t h(\xi) d\xi < \infty$

□

1.3 *Proof of Lemma 1*

Proof.

The cumulative distribution function of T is

$$F(t|\alpha,\lambda) = 1 - \exp\Big\{-\Big(\frac{t}{\lambda}\Big)^{\alpha}\Big\}. \tag{1.10}$$

Thus we have $P(\log T \leqslant y) = P(T \leqslant \exp(y)) = 1 - \exp\{-\frac{\exp(\alpha y)}{\lambda^\alpha}\}$. Assuming $\mu = \log\lambda, \sigma = \frac{1}{\alpha}$, we have $P(\log T \leqslant y) = 1 - \exp\{-\frac{\exp(\alpha y)}{\lambda^\alpha}\} = 1 - \exp(-\exp(\frac{x-\mu}{\sigma}))$. Thus, log T $\sim$ mGEV($\mu = \log(\lambda), \sigma = \frac{1}{\alpha}, \xi = 0$).

□

2. Real Data Application for the MGEV model: Melanoma Cancer Data

We consider a melanoma cancer data set from the National Cancer Institute SEER database Surveillance and End Results (SEER). The data is a population of 1152 subjects who were diagnosed with melanoma cancer in the county of North Bergen of the state of New Jersey between 2000 and 2007 and had been undergoing treatment since diagnosis. All the cases are followed annually and vital status is recorded. All the subjects were diagnosed with only melanoma cancer and were followed up since diagnosis until the end of December, 2007. Subjects who died due to melanoma cancer were considered failed and the rest (those who died due to other causes, dropped, or survived until the end of the study) were considered censored. By the end of 2007, 152 patients had died of melanoma cancer (13.2%) while the remaining were censored. The variable considered in this analysis is: lifetime in months since diagnosis. Subjects with survival time 0 were removed from the data set. The covariates included were: age, gender, stage of the disease (local, other including regional or distant) and marital status (married or divorced/separated/widowed/single). Table 1 lists some basic information for the data set.

We fit both MGEV and Weibull models to the data for better comparisons. We assume the same prior distribution on μ, σ, ξ and β as in Section 5.1. For the hyper parameters, we take $a = 0.5$, $a_\sigma = 0.01$, $b_\sigma = 2$, $\sigma_\beta^2 = 36$ and $\sigma_\mu^2 = 16$. Gibbs sampler with Metropolis-Hastings steps is used as before to get the posterior samples for parameters. For this data set, we have 60,000 MCMC iterations. Convergence was checked using the trace plots, ergodic mean plots and also the autocorrelation plots for all the parameters. And we find that 10,000 iterations are adequate as a burn-in. Further we computed all HPD intervals for all 8 parameters. We fit a Weibull(λ, α) model with covariates as in the previous case. Again for simplicity, we assume that λ, α and β are independent. We assume normal priors for all the β as before. Also, we take $\alpha \sim Gamma(a_\alpha, b_\alpha)$ and $\lambda^2 \sim IG(a_\lambda, b_\lambda)$. We pick $a_\lambda = 0.1$, $b_\lambda = 2$, $a_\alpha = 0.01$, $b_\alpha = 0.01$ and $\sigma_\beta^2 = 49$. We plot the difference of logarithm of CPO at each data point of the two models fitted (MGEV and Weibull)

as before. We also compute LPML and DIC values for each model fit for comparison.

2.1 *Results*

The marginal estimated Kaplan-Meier curve in Figure 1(a) shows the survival curve using MGEV model matches the Kaplan-Meier curves, therefore the proposed model might be a good fit to the data set. The survivval curve is drawn for data without considering the covariates just to provide an initial idea about the appropriateness of the chosen model. Figure 1(a) also shows simulated curve for the fitted Weibull model. There appears a slight benefit of using the MGEV model over the Weibull one. We now consider Bayesian analysis with the covariates included. Tables 2 and 3 show the posterior estimates for all the parameters.

Table 2 shows the estimate of ξ is non-zero, implying there is a positive probability that ξ is non-zero, justifying the need of modeling the data as MGEV which can accommodate an additional shape parameter. The estimates of the β 's from both the fitted models match in sign and with the exception of the intercept, they are also close in values. Figure 1(b) shows the CPO plot of the difference. Most of the points lie above zero (blue dots) giving us an indication that MGEV model is better suited to the data in hand. Table 4 shows the LPML and DIC values of the two model fits. We find that our proposed model MGEV has significantly lower DIC and higher LPML in comparison to Weibull fit.

References

Chen, M.H., Ibrahim, J. and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* **94**, 909–919.

Surveillance, Epidemiology and End Results (SEER) Program, (www.seer.cancer.gov) SEER*Stat. (2013). Total u.s., 1969-2010 counties, national cancer institute, dccps,surveillance research program, surveillance systems branch. *Incidence*

- SEER 17 Regs Research Data, Nov 2011 Sub (1973-2010), Katrina/Rita Population Adjustment, Linked To County Attributes.

Table 1: Summary of the Melanoma Cancer Data in North Bergen.

Survival time(y) (months)	Status(freq)	Age (years)	Gender (freq)	Stage (freq)	Marital status(freq)
Median 37	Censored 1000	Mean 58.84	Male 626	Local 944	Married 742
IQR 47	Death 152	17.00	Female 526	Other 208	Other 410

Table 2: Melanoma Data: Posterior Estimates of the MGEV Model Parameters with covariates included.

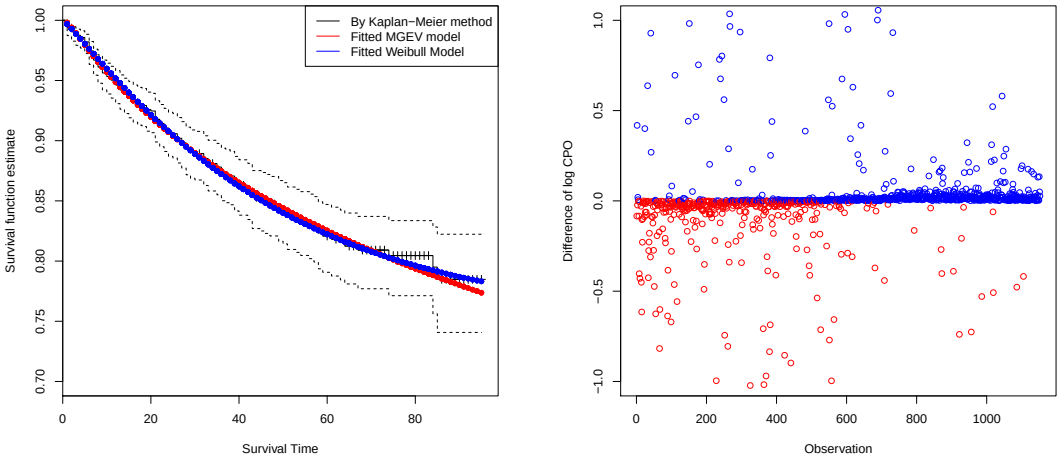
Variable	Posterior mean	Posterior SD	95% HPD interval
μ	13.753	1.106	(12.286, 16.042)
σ	6.579	1.238	(4.486, 8.484)
ξ	0.052	0.141	(-0.196, 0.283)
Intercept	0.077	0.887	(-1.381, 1.253)
Age	0.060	0.008	(0.048, 0.075)
Gender	0.621	0.168	(0.293, 0.942)
Stage	-1.898	0.172	(-2.215, -1.539)
Marital Status	-0.255	0.175	(-0.597, 0.109)

Table 3: Melanoma Data: Posterior Estimates of the Weibull Model Parameters with covariates included.

Variable	Posterior mean	Posterior SD	95% HPD interval
λ	13.289	1.925	(10.17 ,16.16)
α	1.082	0.139	(0.84, 1.32)
Intercept	-4.212	0.715	(-5.25, -2.78)
Age	0.051	0.009	(0.032, 0.065)
Gender	0.594	0.189	(0.240, 0.960)
Stage	-1.841	0.160	(-2.162, -1.524)
Marital Status	-0.220	0.186	(-0.556, 0.171)

Table 4: Model Comparison between Fitted MGEV distribution and Fitted Weibull distribution.

Fitted Model	DIC	LPML
MGEV(μ, σ, ξ)	1764.487	-883.2889
Weibull(α, λ)	1964.006	-983.1095



(a) (b)

Fig. 1:
1(a) Estimated survival curves for the melanoma 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).
1(b) Plot of difference of the log CPO between MGEV and Weibull Model for the melanoma cancer data.