<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Propriety of posterior distributions arising in categorical and survival models under generalized extreme value distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manuscript ID</strong></td>
<td>SS-11-011R2</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td>10.5705/ss.2011.011</td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
<td>Vivekananda Roy and Dipak K. Dey</td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
<td>Vivekananda Roy</td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
<td><a href="mailto:vroy@iastate.edu">vroy@iastate.edu</a></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
Propriety of posterior distributions arising in categorical and survival models under generalized extreme value distribution

Vivekananda Roy  
Department of Statistics  
Iowa State University

Dipak K. Dey  
Department of Statistics  
University of Connecticut

Abstract

This paper introduces a flexible skewed link function for modeling binary as well as ordinal data with covariates based on the generalized extreme value (GEV) distribution. Extreme value techniques have been widely used in many disciplines relating to risk analysis. However, its application in the binary and ordinal data from a Bayesian context is sparse and its strength as a link function has recently been explored. There are a number of non-regular situations with likelihood method for GEV models where usual asymptotic properties of MLE do not hold. That is why Bayesian methodology is preferred for analyzing GEV models. We also introduce the GEV distribution in reliability and survival models. We show that our proposed model leads to an extremely flexible hazard function. In this paper, first we investigate the property of posterior distributions for binary and ordinal response models under the generalized extreme value link using a uniform prior distribution on the regression parameters. Necessary and sufficient conditions for the propriety of the posterior distribution with the generalized extreme value link function are established. Next we consider similar issues for survival data models, where log survival time has a GEV distribution and propriety of the posterior distribution under uniform prior on the regression coefficients is established. The flexibility of the proposed survival model is illustrated through a dataset involving a lung cancer clinical trial.

Key words and phrases. Complementary log-log link; generalized extreme value distribution; hazard function; improper prior; posterior propriety; skewness.
1 Introduction

The generalized extreme value (GEV) distribution is a family of continuous probability distributions that combines the three families of extreme value distributions namely the Gumbel, Fréchet or the Weibull class of distributions developed in extreme value theory. The term “extreme value” refers to the fact that these distributions can be obtained as the limiting distributions of properly normalized maxima of \( n \) independent and identically distributed random variables. Extreme value analysis finds wide applications in many areas including climatology (Sang and Gelfand, 2009), environmental science (Smith (1989), Sang and Gelfand (2010), Wang, Dey and Banerjee (2010)), financial strategy of risk management (Dahan and Mendelson, 2001) and survival analysis (Mann, Schafer and Singpurwalla (1974), Kim and Ibrahim (2000)). In this article we show the broad applicability of the GEV distribution for analyzing binary, ordinal and survival data.

The most popular model for binary response data is the logistic regression model that is based on the logit link function. Other frequently used link functions are probit and complimentary log-log links. However, these commonly used link functions do not always provide the best fit for a given data set. In particular, if the probability of a given binary response approaches 0 at a different rate than it approaches 1, the use of a symmetric link function such as probit or logit is inappropriate. In this case, if the link function is misspecified, there can be substantial bias in the mean response estimates (Czado and Santner, 1992). One intuitive way of guarding against link misspecification is to embed the symmetric links e.g., probit or logit into a wide parametric class of links. Several authors have introduced such parametric classes for binary response data. For example, Aranda-Ordaz (1981), Guerrero and Johnson (1982), Morgan (1983) and Whitemore (1983) considered different one-parameter families. Stukel (1988) extended these links by proposing a class of generalized logistic models. Stukel’s (1988) models are general and several commonly used link functions, such as the probit and complimentary log-log link models can be approximated by members of this family. However, in the presence of covariates, Stukel’s (1988) models yield improper posterior distributions for many types of non-informative improper priors, including the improper uniform prior for the regression coefficients (Chen, Dey and Shao, 1999). Chen et al. (1999) introduced a class of skewed links that leads to proper posterior distributions for the regression coefficients under standard improper prior. However, Chen et al.’s (1999) model has the limitation that the intercept term is confounded with the skewness parameter. This problem was
overcome in Kim, Chen and Dey (2008) by a class of generalized skewed t-link models, though the constraint on the shape parameter $\delta$ as $0 < \delta \leq 1$ greatly reduces the possible range of skewness provided by this model.

To construct a general and flexible model for binary data that overcomes the constraint of generalized skewed t-link models, Wang and Dey (2010) introduce the GEV distribution as a link function. With a free shape parameter, the GEV distribution provides great flexibility in fitting a wide range of skewness in the response curve. Wang and Dey (2010) illustrated the flexibility of GEV link function through both simulated data sets and a billing data set of the electronic payments system adoption from a Fortune 100 company in 2005.

The misspecification of link function can also occur for ordinal data (Wang and Dey, 2011). Many link functions for ordinal response data proposed in the literature including the probit link, Albert and Chib’s (1993) family of t-links, and Chen and Dey’s (2000) scale mixture of multivariate normal link functions are symmetric and may not be appropriate when the cumulative probabilities are skewed. Wang and Dey (2011) employed the GEV distribution for modeling ordinal response data. They illustrated the flexibility of the GEV model with an application to an ecological survey data about the coverage of Berberis thunbergii in New England. However, the authors did not address the issue of the propriety of the posterior distribution of the regression coefficients and of the cut points under improper uniform priors on the parameters. In this paper, our major contribution is to develop conditions with rigorous proofs for the propriety of the posterior distributions of the associated parameters for binomial as well as ordinal data. We further propose survival models based on the GEV distribution, and provide sufficient conditions for the propriety of the corresponding posterior distributions when an improper uniform prior is used on the regression coefficients.

There is a close connection between categorical and survival data through the link function specification (Banerjee, Chen, Dey and Kim, 2007). Banerjee et al. (2007) proposed a general class of non-proportional hazard models known as generalized odds-rate class of regression models. In a similar spirit, in this paper we develop a class of non-proportional hazard regression models using the GEV distribution. In reliability and survival analysis, the probability distribution of the time-to-failure of an equipment can be characterized by the hazard function (also known as failure rate) $\lambda(t) = f(t)/S(t)$, where $f(t)$ and $S(t)$ are failure density and survival function respectively. Many widely used mod-
els including gamma, Weibull and the truncated normal distribution lead to monotone hazard function. The increasing hazard function is a characteristic of the systems that consistently deteriorate with time, whereas decreasing hazard function is a property of the equipments that consistently improve with time. But, it is long known that in practical situations often the hazard function is not monotone and it is either upside-down shaped or bathtub shaped or a combination of upside-down and bathtub shape (Lieberman (1969), Langlands, Pocock, Kerr and Gore (1979), Bennett (1983)). A popular way of introducing non-monotone hazard function is by considering the mixture distribution models (Barlow and Proschan (1975), Finkelstein (2009)). Mixtures do not always lead to a non-monotone hazard function. For example, it is well known that mixtures of decreasing failure rate always have decreasing failure rate (Barlow and Proschan, 1975). On the other hand mixtures of increasing failure rate can decrease, at least in some time intervals (Gurland and Sethuraman, 1995). Also, it is to be noted that often the mixture modeling might not be desirable since it brings flexibility at the expense of additional parameters, consequently more parameters have to be estimated.

To build an extremely flexible hazard function, we propose the GEV distribution for $\log T$ where $T$ denotes the failure time. We show that by changing the shape parameter of the GEV distribution, we obtain a variety of shape for the hazard function including the upside-down and bathtub shapes. As mentioned before, the GEV distribution, as a special case, includes the Gumbel distribution (also known as type I extreme value distribution). A version of the type I extreme value distribution is a widely used parametric survival model. However, the commonly used version of the extreme value model can be viewed as a parametrization of the Weibull distribution. In particular, if $T$ is assumed to have a Weibull distribution, then $\log T$ has a Gumbel distribution for the minimum extremes (Mann et al. (1974), Kim and Ibrahim (2000)). It is well known that in this case the hazard (failure) rate is some power function of $t$, the time-to-failure and the hazard rate is decreasing (increasing) if the shape parameter of the Weibull distribution is $< 1 (> 1)$ (Mann et al., 1974). However, if $\log T$ has a GEV distribution the modeling framework is much different.

We also consider situations where the distribution of failure time $T$ depends on one or more covariates. In particular, we consider accelerated failure time models, which are linear models for $\log T$. The regression models for $\log T$ arise naturally in accelerated life testing. Life data analysis involves analyzing times-to-failure data in order to quantify the reliability of a product. Unfortunately, for products
with long life time, only a few, if any, items fail during testing (between the design and release of a product) under normal operating conditions. Given this difficulty, the standard method to assess highly reliable product is to test them under extreme operating conditions, referred to as accelerated life testing (Mann et al. (1974), Nelson (1990)). Accelerated failure time or log-location-scale models are also useful in other fields of applications. We introduce accelerated failure time models with GEV as error distribution. We consider a Bayesian analysis of the corresponding model under non-informative priors. Since the Jeffreys prior turns out to be extremely cumbersome in this case, we consider a uniform prior on the regression coefficients. We obtain sufficient conditions for the propriety of the corresponding posterior distribution. We demonstrate the flexibility of the proposed survival model through a lung cancer dataset.

The rest of the paper is organized as follows. Section 2 provides a short introduction to GEV distribution. Section 3 describes the GEV link models for binomial response data and provides necessary and sufficient conditions for propriety of the posterior distributions. Section 4 is devoted to the development of sufficient conditions for posterior propriety under GEV links in ordinal data scenario. Section 5 introduces GEV distribution in reliability and accelerated failure time models. The paper concludes with a discussion in Section 6. The proofs of the theorems have been relegated to appendices.

2 Generalized extreme value distribution

Suppose $Y_1, Y_2, \ldots$ is a sequence of iid random variables and let $M_n = \max\{Y_1, \ldots, Y_n\}$. If the distribution of $Y_i$ is specified then the exact distribution of $M_n$ is known. On the other hand, in the absence of such specification, extreme value theory considers the existence of $\lim_{n \to \infty} P[\frac{(M_n - b_n)}{a_n} \leq y] \equiv F(y)$ for two sequences of real numbers $a_n > 0$ and $b_n$. If $F(y)$ is a non-degenerate distribution, then it belongs to either the Gumbel, Fréchet or the Weibull family of distributions, which can all be represented as members of a single family of GEV distributions with cumulative distribution function as follows:

$$G_{(\mu, \sigma, \xi)}(x) = \begin{cases} 
\exp\left\{-\left\{ 1 + \frac{x - \mu}{\sigma} \right\}_{+}^{\frac{1}{\xi}} \right\} & \text{if } \xi > 0 \text{ or } \xi < 0 \\
\exp(-\exp(-\frac{x - \mu}{\sigma})) & \text{if } \xi = 0,
\end{cases} \quad (1)$$
where $\mu \in \mathbb{R}$ is the location parameter, $\sigma \in \mathbb{R}^+$ is a scale parameter, $\xi \in \mathbb{R}$ is the shape parameter and $x_+ = \max(x, 0)$. The Gumbel, Fréchet and the Weibull family of distributions are obtained from (1) by considering $\xi = 0, \xi > 0$ and $\xi < 0$ respectively. A more detailed discussion on extreme value distributions can be found in Coles (2001) and Smith (1985).

The importance of GEV distribution as a link function arises from the fact that the shape parameter $\xi$ purely controls the tail behavior of the distribution (Wang and Dey, 2010). The Gumbel distribution is the least positively skewed distribution in the GEV class when $\xi$ is non-negative. Wang and Dey (2010) provide a plot of the probability distributions of the GEV family which demonstrates the flexibility of the GEV distribution.

Since the usual definition of skewness $\mu_3 = \left\{ \frac{E(X - \mu)^3}{E(X - \mu)^2} \right\}^{-3/2}$ does not exist for large positive values of $\xi$’s for the GEV model, Wang and Dey (2010) extended Arnold and Groeneveld’s (1995) skewness measure to the GEV distribution in terms of its mode. Wang and Dey (2010) showed that, based on this skewness definition, the GEV distribution is negatively skewed for $\xi < \log 2 - 1$ and positively skewed for $\xi > \log 2 - 1$.

In the next section we consider the GEV link model for binomial response data.

3 Generalized extreme value link for binomial regression models

Suppose $y = (y_1, y_2, \ldots, y_n)$ is a vector of $n$ independent binomial random variables. Also, let $x_i$ be the $k \times 1$ vector of covariates associated with $y_i$, and suppose $X$ denotes the $n \times k$ design matrix with rows $x'_i$. Let $\beta$ be the $k \times 1$ vector of regression coefficients. Assume that $y_i \sim \text{Bin}(n_i, p_i)$ $i = 1, 2, \ldots, n$. Also, we assume that

$$p_i = 1 - G_\xi(-x'_i \beta),$$

where $G_\xi(x)$ represents the cumulative probability at $x$ for the GEV distribution with $\mu = 0, \sigma = 1$, and an unknown shape parameter $\xi$. So the joint pmf of $y$ is given by

$$f(y|\beta, \xi) = \prod_{i=1}^{n} \binom{n_i}{y_i} \left(1 - G_\xi(-x'_i \beta)\right)^{y_i} \left(G_\xi(-x'_i \beta)\right)^{n_i - y_i}.$$

It is possible to estimate the shape parameter $\xi$ in the above GEV link model by the maximum likelihood method. However, there are a number of non-regular situations in using likelihood methods where
the conditions required for the usual asymptotic properties of the maximum likelihood estimator do not hold. Smith (1985) studied the maximum likelihood estimation for the three-parameter GEV distribution in the non-regular cases in detail and obtained that when $\xi < -0.5$ the regularity conditions are not satisfied by the GEV models. This means that the standard asymptotic likelihood results are not automatically applicable. Since Bayesian analysis do not depend on the regularity assumptions required by the asymptotic theory of maximum likelihood, in the unusual situation where $\xi < -0.5$ and the classical theory of maximum likelihood breaks down, Bayesian inference provides a viable alternative. This is one of the reasons for favoring Bayesian methodology for analyzing the GEV link model.

In the next two subsections we are going to consider two different priors on $(\beta, \xi)$, namely the uniform prior and the Jeffreys prior and study the property of the corresponding posterior distributions.

### 3.1 Uniform prior

Here, we consider an improper uniform prior on $\beta$, $\pi(\beta) \propto 1; \beta \in \mathbb{R}^k$ and a proper prior on $\xi$, $\pi(\xi) = 0.5I_{[-1,1]}(\xi)$. So the joint posterior density becomes

$$
\pi(\beta, \xi|y) \propto \prod_{i=1}^{n} \left( \frac{n_i}{y_i} \right) \left( 1 - G_\xi(-x_i'\beta) \right)^{y_i} \left( G_\xi(-x_i'\beta) \right)^{n_i-y_i} \frac{1}{2}I_{[-1,1]}(\xi).
$$

Our main result of this section is Theorem 1, where we provide sufficient conditions for propriety of the posterior density, $\pi(\beta, \xi|y)$. Of course, $\pi(\beta, \xi|y)$ will be proper if and only if

$$
c(y) := \int_{-1}^{1} \int_{\mathbb{R}^k} f(y|\beta, \xi)d\beta d\xi < \infty.
$$

Slightly abusing notation, we denote the pmf of $y_i$ by

$$
f(y_i|\beta, \xi) = \binom{n_i}{y_i} \left( 1 - G_\xi(-x_i'\beta) \right)^{y_i} \left( G_\xi(-x_i'\beta) \right)^{n_i-y_i}.
$$

Notice that

$$
f(y_i|\beta, \xi) \leq \begin{cases} 
G_\xi(-x_i'\beta) & \text{if } y_i = 0 \\
1 - G_\xi(-x_i'\beta) & \text{if } y_i = n_i \\
\binom{n_i}{y_i} \left( 1 - G_\xi(-x_i'\beta) \right) \left( G_\xi(-x_i'\beta) \right)^{n_i-y_i} & \text{if } 1 \leq y_i \leq n_i - 1.
\end{cases}
$$
Let \( N_n = \{1, 2, \ldots, n\} \). We partition \( N_n \) as \( N_n = I_1 \cup I_2 \cup I_3 \), where \( I_1 = \{i \in N_n : y_i = 0\} \), \( I_2 = \{i \in N_n : y_i = n_i\} \) and \( I_3 = \{i \in N_n : 1 \leq y_i \leq n_i - 1\} \) (Chen, Ibrahim and Shao, 2004b). So

\[
\begin{align*}
    f(y|\beta, \xi) &= \prod_{i=1}^{n} f(y_i|\beta, \xi) \\
    &\leq \prod_{i \in I_1} G_\xi(-x_i'\beta) \prod_{i \in I_2} \left(1 - G_\xi(-x_i'\beta)\right) \\
    &\quad \times \prod_{i \in I_3} \left(1 - G_\xi(-x_i'\beta)\right)\left(G_\xi(-x_i'\beta)\right).
\end{align*}
\]

Let \( q = \#(I_3) \) be the cardinality of \( I_3 \). Let the \( q \times k \) matrix with rows \( x_{i}^\prime \), \( i \in I_3 \) be denoted by \( \tilde{X} \) and let \( \tilde{X} \) be the following \((n + q) \times k\) matrix

\[
\begin{pmatrix}
    X \\
    \tilde{X}
\end{pmatrix}.
\]

Define \( \tau_1, \tau_2, \ldots, \tau_{n+q} \) where \( \tau_i = -1 \) if \( i \leq n \) and \( i \in I_1 \cup I_3 \), \( \tau_i = 1 \) if \( i \leq n \) and \( i \in I_2 \) and \( \tau_{n+i} = 1 \) for \( i = 1, 2, \ldots, q \). Let \( X^* \) be the \((n + q) \times k\) matrix with \( i \)th row of \( X^* \) being \( -\tau_i \tilde{x}_{i} \) where \( \tilde{x}_{i} \) is the \( i \)th row of \( \tilde{X} \) and let \( X^*_{\ell,m} \) be the \((m - \ell) \times k\) matrix with rows \( -\tau_i \tilde{x}_{i}, \ell < i \leq m \).

The following theorem states sufficient conditions for the propriety of the posterior density, \( \pi(\beta, \xi|y) \).

**Theorem 1.** Suppose that there exist integers \( p, m_0, \ldots, m_p \) such that \( p > k, 0 = m_0 < \cdots < m_p \leq n + q \), and that \( X^*_{m_{\ell-1}, m_{\ell}} \) is of full rank and positive vectors \( a_1, a_2, \ldots, a_p \) such that \( a_{\ell} X^*_{m_{\ell-1}, m_{\ell}} = 0 \) for \( \ell = 1, 2, \ldots, p \). Then \( c(y) < \infty \).

The proof of the Theorem 1 is given in Appendix A.

Notice that the binary regression models can be obtained as a special case of binomial regression models by taking \( n_i = 1 \) for \( i = 1, 2, \ldots, n \). So in this case \( I_3 = \emptyset \), a null set, \( q = 0 \) and \( X^* \) is an \( n \times k \) matrix with \( i \)th row of \( X^* \) is \( x_i^\prime I_{\{0\}}(y_i) - x_i^\prime I_{\{1\}}(y_i) \). In order to gain intuition behind the conditions in Theorem 1, consider this special case of binary regression models. If \( X \) is of full rank, the existence of a positive vector \( a \) with \( a'X^* = 0 \) implies that there is no point \( \beta_0 \in \mathbb{R}^k \setminus \{0\} \) such that \( x_i^\prime \beta \leq 0 \) for all \( i = 1, 2, \ldots, n \) or \( x_i^\prime \beta \geq 0 \) for all \( i = 1, 2, \ldots, n \) (Roy and Hobert, 2007, page 261). That is, every point in \( \mathbb{R}^k \setminus \{0\} \) lies in the positive side of some of the \( n \) hyperplanes \( x_i^\prime \beta = 0 \) and in the negative side of the rest of the hyperplanes. This condition of the existence of a positive vector \( a \) satisfying \( a'X^* = 0 \) also implies that the data set is overlapped (Albert and Anderson, 1984). (See Albert and Anderson (1984) who beautifully depict this condition through a picture.) Since the GEV
distribution does not have higher order moments (for example, it does not have finite second moment for \( \xi \geq 1/2 \)), we need to impose stronger conditions (mentioned in Theorem 1) on the matrix \( X^* \), than the mere existence of a positive vector \( a \) satisfying \( a'X^* = 0 \) (Chen and Shao, 2000). On the other hand, if we assume that \( \xi < 1/k \), then the GEV distribution has finite \( k \)th moment. In this case, the existence of \( a > 0 \) with \( a'X^* = 0 \) implies that \( c(y) < \infty \). Roy and Hobert (2007) provide a simple way to check the existence of a positive vector \( a \) with \( a'X^* = 0 \) which involves maximizing \( 1'g \) subject to \( g'X^* = 0, (J - I)g \leq 1 \) (element wise) and \( g_i \geq 0 \) for \( i = 1, 2, \ldots, n + q \) where \( 1 \) and \( J \) denote a column vector and the matrix of 1s respectively. This can be easily implemented in many statistical software languages. For example, the “simplex” function in the “boot” library of R (R Development Core Team, 2011) can be used to check the above condition.

The following theorem provides necessary conditions for posterior propriety in the case of binary regression models.

**Theorem 2.** For binary regression models the following two conditions are necessary for \( c(y) < \infty \).

- The design matrix \( X \) is of full rank.
- There exists a positive vector \( a = (a_1, a_2, \ldots, a_n)' \in \mathbb{R}^n \) such that
  \[
a'X^* = 0.
  \]

The proof of the Theorem 2 is given in Appendix B. As mentioned before, if it is assumed that \( \xi < 1/k \) then the above two conditions are, in fact, necessary and sufficient for \( c(y) < \infty \).

In the next section we study the posterior distribution that results when we use the Jeffreys prior on the regression coefficients \( \beta \).

### 3.2 Jeffreys prior

In this section we consider the following prior on \((\beta, \xi)\)

\[
\pi_1(\beta, \xi) = \pi(\beta | \xi)\pi(\xi),
\]

where \( \pi(\beta | \xi) \propto |I(\beta | \xi)|^{1/2} \) with \( I(\beta | \xi) \) being the Fisher information matrix for the Binomial distribution with the GEV link and \( \pi(\xi) = 0.5I_{[-1,1]}(\xi) \). So, the corresponding posterior density is given
by
\[ \pi_1(\beta, \xi | y) \propto f(y | \beta, \xi) | I(\beta | \xi)|^{1/2} I_{[-1,1]}(\xi). \]

The posterior density for the Jeffreys prior is proper, which is stated in the following theorem.

**Theorem 3.** The posterior density \( \pi_1(\beta, \xi | y) \) is proper.

The proof of the Theorem 3 is given in Appendix C.

In the next section we consider the GEV link model for ordinal response data.

## 4 Generalized extreme value link for independent ordinal regression models

Suppose that we have \( n \) observations \( y_1, y_2, \ldots, y_n \), where \( y_i \) takes value in \( \{ j : j = 1, 2, \ldots, J \} \). Note that in Section 3 we used the same notation \( y_i \) to denote a binomial variable. Also, abusing the notation here we use \( y \) to denote \( (y_1, y_2, \ldots, y_n) \). A common way of modeling ordinal data is by considering underlying continuous latent variables \( w_i, i = 1, 2, \ldots, n \) and assuming that we observe

\[ y_i = j \quad \text{if} \quad \gamma_{j-1} < w_i \leq \gamma_j, \]

where \( -\infty = \gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_{J-1} < \gamma_J = \infty \) are cut point parameters that determine the discretization of the data into the \( J \) ordered categories (Albert and Chib, 1993). Here we assume that

\[ w_i = x_i' \beta + \epsilon_i, \]

\( i = 1, 2, \ldots, n \) where \( x_i' \)'s are \( k \)-dimensional vectors of covariates, \( \beta \) is the vector of regression parameters and \( \epsilon_i \sim GEV(\mu = 0, \sigma = 1, \xi) \) (Wang and Dey, 2011). Note that

\[ P(y_i = j) = P(\gamma_{j-1} < w_i \leq \gamma_j) = P(\gamma_{j-1} - x_i' \beta < \epsilon_i \leq \gamma_j - x_i' \beta) = G_\xi(\gamma_j - x_i' \beta) - G_\xi(\gamma_{j-1} - x_i' \beta). \]

So the likelihood function for the above model is given by

\[ L(\beta, \gamma, \xi | y) = \prod_{i=1}^{n} \left[ G_\xi(y_{i} - x_i' \beta) - G_\xi(y_{i-1} - x_i' \beta) \right]. \]
Here we consider a Bayesian analysis of the above model with the following priors on the parameters $\beta$, $\xi$, and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{J-1})$,

$$\Pi(\beta) \propto 1, \quad \beta \in \mathbb{R}^p,$$  
$$\Pi(\xi) = \frac{1}{2} I_{[-1,1]}(\xi),$$  
$$\Pi(\gamma) = 1 I_{[\gamma_1 < \gamma_2 < \cdots < \gamma_{J-1}]}(\gamma).$$

Then the posterior density of $\beta$, $\gamma$, and $\xi$ is given by

$$\Pi(\beta, \gamma, \xi | y) \propto \frac{L(\beta, \gamma, \xi | y) \Pi(\beta) \Pi(\gamma) \Pi(\xi)}{\int_1^1 \cdots \int_{\mathbb{R}^k} \left\{ \prod_{i=1}^n \left[ G_{\xi}(\gamma_{y_i} - x_i^{d} - G_{\xi}(\gamma_{y_i-1} - x_i^{d}) \right] \right\} d\beta d\gamma d\xi < \infty .$$

The main results of this section are Theorem 4 and Theorem 5 where we provide sufficient conditions for the posterior density $\Pi(\beta, \gamma, \xi | y)$ to be proper. Of course, the posterior density, $\Pi(\beta, \gamma, \xi | y)$ is proper if and only if

$$c_1(y) := \int_1^1 \cdots \int_{\mathbb{R}^k} \left\{ \prod_{i=1}^n \left[ G_{\xi}(\gamma_{y_i} - x_i^{d} - G_{\xi}(\gamma_{y_i-1} - x_i^{d}) \right] \right\} d\beta d\gamma d\xi < \infty .$$

In order to state our sufficient conditions, we introduce the following notations: first we partition the set $N_n = \{1, 2, \ldots, n\}$ into $N_n = U \uplus L \uplus M$ where

$$U = \{i \in N_n : y_i = J\}$$  
$$L = \{i \in N_n : y_i = 1\}$$

and

$$M = \{i \in N_n : 1 < y_i < J\}.$$  

Let $X$ be the $n \times k$ design matrix with rows $x_i^d$. Define $x_i^* = (1, x_i^d)'$ for $i = 1, 2, \ldots, n$.

**Theorem 4.** Assume that

(A1) there exists $p > k + J - 1$ such that we can partition $U = \bigcup_{\ell=1}^p U_\ell$, $L = \bigcup_{\ell=1}^p L_\ell$, and $M = \bigcup_{\ell=1}^p M_\ell$, where $U_\ell$, and $L_\ell$ are non-empty for $\ell = 1, \ldots, p$. Define

$$X_{1\ell} = \{x_i^*, i \in U_\ell, -x_j^*, j \in L_\ell \cup M_\ell\}'$$  
$$X_{2\ell} = \{x_j^*, j \in L_\ell, -x_i^*, i \in U_\ell \cup M_\ell\}' , \quad \ell = 1, \ldots, p .$$
Then the posterior is proper if one of the following two conditions is satisfied.

(A2)' $X_{1\ell}$ is of full column rank and there exists $b_{\ell} > 0$ such that $b_{\ell}X_{1\ell} = 0$ for $\ell = 1, \ldots, p$.

(A2)'' $X_{2\ell}$ is of full column rank and there exists $b_{\ell} > 0$ such that $b_{\ell}X_{2\ell} = 0$ for $\ell = 1, \ldots, p$.

The proof of Theorem 4 is given in Appendix D.

Next, we present another set of sufficient conditions for $c_1(y) < \infty$. Following Chen and Shao (1999), we define

$$T_{\ell,1} = \{(i, j) : i \in U, j \in L, x_{i\ell} - x_{j\ell} > 0\}$$

$$T_{\ell,-1} = \{(i, j) : i \in U, j \in L, x_{i\ell} - x_{j\ell} < 0\}.$$ 

For $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$ where $\eta_\ell = \pm 1$, let

$$T(\eta) = \bigcap_{\ell=1}^{k} T_{\ell,\eta_\ell}.$$ 

Suppose there exist $U(\eta) \subset U$ and $L(\eta) \subset L$ such that $U(\eta) \times L(\eta) \subset T(\eta)$. Let

$M^* = \min_{\eta} \min \left( \#U(\eta), \#L(\eta) \right)$, where as before $\#A$ is the cardinality of the set $A$. We now have the following theorem.

**Theorem 5.** If $M^* > k + J - 1$, then $c_1(y) < \infty$.

The proof of Theorem 5 is given in Appendix E.

Note that if either $L$ or $U$ is empty, then obviously there is no information available to estimate $\gamma_1$ or $\gamma_{J-1}$, so we need at least the sets $U$ and $L$ to be non-empty for proper posterior. On the other hand, from the above theorems we see that the posterior density, $\Pi(\beta, \gamma, \xi|y)$ can still be proper even when the set $M$ is empty. Also, $\text{rank}(X) = k$, i.e., the full rank condition of the design matrix is a necessary condition for the posterior density, $\Pi(\beta, \gamma, \xi|y)$ to be proper (Chen and Shao, 1999).

Now we introduce the GEV distribution in the analysis of survival data.

### 5 Generalized extreme value distribution in survival analysis

As we mentioned in Section 1, one can hardly find populations in real life with monotone hazard function. Whereas, many widely used models including gamma, Weibull and the truncated normal distribution lead to monotone hazard function. Mixtures of distributions have been a popular approach to
construct a non-monotone hazard function (See Finkelstein (2009) and the references therein.). But, it might not be possible to create a flexible non-monotone hazard function using mixtures. It is well known that decreasing hazard function (failure rate) is closed under the operation of mixing, that is, mixtures of decreasing failure rate always have decreasing failure rate (Barlow and Proschan, 1975). In the following section we propose the GEV model to obtain an extremely flexible hazard function.

5.1 Shape of the hazard function

Suppose $T$ denote the time-to-failure. In our proposed GEV model we assume that $\log T \sim GEV(\mu = 0, \sigma = 1, \xi)$. As we mentioned in Section 1, a version of the extreme value distribution is a widely used parametric survival model. However, as we mentioned before, the commonly used version of the extreme value model can be viewed as a parametrization of the Weibull distribution and in this case the hazard function is monotone. However, if $\log T$ has a GEV distribution, the hazard function, as we show now, is extremely flexible.

Since we assume that $\log T \sim GEV(\mu = 0, \sigma = 1, \xi)$, the pdf of $T$ is

$$f_\xi(t) = \begin{cases} 
\frac{\exp\left[-(1+\xi \log t)^{-\frac{1}{\xi}}\right]}{t(1+\xi \log t)^{\frac{1}{\xi}+1}} & t > \exp\left(-\frac{1}{\xi}\right) \text{ if } \xi > 0 ; \\
t < \exp\left(-\frac{1}{\xi}\right) & t < \exp\left(-\frac{1}{\xi}\right) \text{ if } \xi < 0 \\
\exp\left(-\frac{1}{T}\right) \cdot \frac{1}{T^2} & 0 < t < \infty \text{ if } \xi = 0 .
\end{cases}$$

So the survival function $S_\xi(t) = P(T \geq t)$ is given by

$$S_\xi(t) = \begin{cases} 
1 - \exp\left(- \left(1 + \xi \log t\right)^{-\frac{1}{\xi}}\right) & \text{if } \xi \neq 0 \\
1 - \exp\left(-\frac{1}{T}\right) & \text{if } \xi = 0 .
\end{cases}$$

Hence, the hazard function $\lambda_\xi(t) = f_\xi(t)/S_\xi(t)$ is given by

$$\lambda_\xi(t) = \begin{cases} 
\frac{1}{t(1+\xi \log t)^{\frac{1}{\xi}+1} \left\{ \exp\left(1+\xi \log t\right)^{-\frac{1}{\xi}}\right\} - 1} & \text{if } \xi \neq 0 \\
\frac{1}{t^2 \left(\exp\left(\frac{1}{T}\right)\right) - 1} & \text{if } \xi = 0 .
\end{cases}$$
Figure 1 shows the plot of the GEV hazard function for five different values of $\xi$, $\xi = 0.3, 0, -0.3, -0.5,$ and $-1.5$. Note that, for $\xi = 0, 0.3$ the hazard function of the GEV distribution is an upside-down function whereas for $\xi = -0.3$ the shape of the hazard function is modified bathtub (it first increases up to a point and then behaves like a bathtub failure rate). Gupta and Warren (2001) showed that certain mixtures of two gamma distributions lead to modified bathtub hazard rate. For $\xi = -1.5$, the GEV hazard function takes the bathtub shape or U-shape, whereas for $\xi = -0.5$, the hazard function is monotone increasing. This shows that a GEV model is extremely flexible in modeling survival data. Another advantage of using a GEV model is that by varying only the shape parameter, $\xi$ we obtain different shapes for the hazard function whereas the mixture models provide flexibility at the expense of many extra parameters. The number of parameters increase with the number of mixing distributions as well as with the unknown change points where the function changes.

Figure 1: Hazard functions of the generalized extreme value distribution for different values of $\xi$. 

We now state a theorem regarding the hazard function for $\xi = 0$, that is, $\lambda_0(t)$. 

...
**Theorem 6.** The hazard function, $\lambda_0(t)$ is an upside down function.

The proof of Theorem 6 is given in Appendix F. In the next section we consider accelerated failure time models with GEV errors.

5.2 Generalized extreme value regression models

In Section 5.1 we demonstrated that the GEV models can be very successfully used for analyzing reliability data. Here we consider GEV as the error distribution in accelerated failure time models. Let $T_i$ denote the failure times and assume that

$$\log T_i \sim GEV(x_i' \beta, \sigma, \xi),$$

for $i = 1, \ldots, n$ where $x_i$’s are the $k$ dimensional covariates, $\beta$ is the vector of regression coefficients and $\sigma$ is the scale parameter. As we mentioned before, a version of the extreme value distribution is widely used in survival data analysis. For example, Kim and Ibrahim (2000) consider the extreme value regression model where they assume that $T_i$ has a Weibull distribution i.e., $\log T_i$ has a type 1 extreme value distribution.

Let $\{(t_i, \nu_i); i = 1, \ldots, n\}$ be the observed data where $t_i, i = 1, \ldots, n$ denotes the observed failure time or right censored time and $\nu_i$ is an indicator variable taking value 1 if $t_i$ is an observed failure time and 0 if $t_i$ is censored. Let $t = (t_1, t_2, \ldots, t_n)$ and $\nu = (\nu_1, \nu_2, \ldots, \nu_n)$. So the likelihood function (assuming right censoring) is given by

$$L(\beta, \sigma, \xi | t, \nu) = \prod_{i=1}^{n} \left\{ \frac{1}{\sigma t_i \left(1 + \frac{y_i - x_i' \beta}{\sigma} \right)^{\frac{1}{\xi} + 1}} \exp \left[ - \left( 1 + \frac{y_i - x_i' \beta}{\sigma} \right)^{-\frac{1}{\xi}} \right] \right\}^{\nu_i} \times \left\{ 1 - \exp \left[ - \left( 1 + \frac{y_i - x_i' \beta}{\sigma} \right)^{-\frac{1}{\xi}} \right] \right\}^{1-\nu_i},$$

(2)

where $y_i = \log t_i$ for $i = 1, \ldots, n$. Note that in Sections 3 and 4 we used $y_i$ to denote different things.

Note that the GEV distribution is irregular as its support depends on the parameters (Smith, 1985). When $\sigma$ and $\xi$ are known, the Jeffreys prior for $\beta$, $\pi(\beta | \sigma, \xi)$ is proportional to the square root of the determinant of the Fisher information matrix, that is, $\pi(\beta | \sigma, \xi) \propto |I(\beta | \sigma, \xi)|^{1/2}$. It can be shown that

$$I(\beta | \sigma, \xi) := E\left( - \frac{\partial^2}{\partial \beta_i \partial \beta_j} \log L(\beta, \sigma, \xi | t, \nu) \bigg| \beta, \sigma, \xi \right)_{1 \leq i, j \leq k} = E(X^T WX | \beta, \sigma, \xi),$$

15
where $X$ is the $n \times k$ covariate matrix, and $W$ is an $n \times n$ diagonal matrix. It turns out that the $i$th diagonal element of $W$ is a very complicated function of $d_i = 1 + \xi(y_i - x_i'\beta)/\sigma$. Therefore, instead of using Jeffreys prior, we use a uniform prior on $\beta$. Kim and Ibrahim (2000) made similar comments regarding Jeffreys prior for their extreme value regression model.

Suppose we consider the following prior on $(\beta, \sigma, \xi)$,

$$
\pi(\beta, \sigma, \xi) \propto \pi(\sigma)\pi(\xi); \beta \in \mathbb{R}^k, \sigma \in \mathbb{R}^+, \xi \in \mathbb{R}^+.
$$

where $\pi(\sigma)$ is a (proper or improper) density on $\mathbb{R}^+$ and $\pi(\xi) = 0.5I_{[-1,1]}(\xi)$. So the posterior density is given by

$$
\pi(\beta, \sigma, \xi | t, \nu) \propto L(\beta, \sigma, \xi | t, \nu)\pi(\sigma)\pi(\xi).
$$

We now state the following theorem regarding the propriety of posterior density $\pi(\beta, \sigma, \xi | t, \nu)$.

**Theorem 7.** Let $\tilde{X}$ be the $n \times k$ matrix with rows $\nu_i x_i$, $i = 1, 2, \ldots, n$. Assume that $r(\tilde{X}) = k$ and

$$
\int_0^\infty \frac{1}{\sigma^{m-k}}\pi(\sigma) d\sigma < \infty,
$$

where $m = \#\{i: \nu_i = 1\}$, is the number of uncensored observations. Then the posterior density, $\pi(\beta, \sigma, \xi | t, \nu)$ in (3) is proper.

The proof of Theorem 7 is given in Appendix G.

**Remark 1.** Kim and Ibrahim (2000) considered conditions for posterior propriety in the special case where $T_i$ is assumed to have a Weibull distribution. One of these conditions was that the likelihood function based on any $n - k$ observations is bounded. Here we establish the propriety of the posterior density in (3) without such stringent restrictions.

### 5.3 An illustrative example

We consider survival data on 40 advanced lung cancer patients presented in (Lawless, 2003, p. 7). The dataset has three covariates: performance status (PS) at diagnosis (a measure of general medical condition on a scale of 10 to 90, with lower numbers indicating poorer conditions), age of the patient at diagnosis in years (age), and the number of months from diagnosis of cancer to entry into the study (diag). Three of the 40 observations are censored. This dataset has been previously analyzed by Kim and Ibrahim (2000) who assumed that the survival time follows a Weibull distribution. The shape parameter of the Weibull distribution was estimated to be 0.949 which implies a monotone decreasing hazard
rate. Figure 2 shows the plots of the estimated baseline hazard function using the nonparametric kernel methods described in Müller and Wang (1994). We used the “muhaz” package in R (R Development Core Team, 2011) to make these plots. The plot in the left panel is obtained using the global bandwidth selection algorithms of Müller and Wang (1994) and the maximum time is taken to be the time at which ten patients remain at risk (default choice in the “muhaz” function). The plot in the right panel is based on the local bandwidth choices as prescribed in Müller and Wang (1994) and the time domain is stretched to the maximum observed survival time (999 days). The plots in Figure 2 suggest that the true hazard rate may be U-shaped or modified bathtub shaped and hence a Weibull model for the survival time T may not be appropriate here. Here we use the GEV accelerated failure time model proposed in Section 5.2 to analyze this data set.

As in Kim and Ibrahim (2000) we consider improper uniform prior on the regression coefficients. Kim and Ibrahim (2000) used gamma prior on the shape parameter of the Weibull distribution which corresponds to using an inverse gamma (IG) prior for the scale parameter of the type 1 extreme value accelerated failure time models. We use IG(1,1) prior on $\sigma$. The posterior estimates reported below are fairly robust with respect to the hyperparameter values of the IG prior. Since $r(\tilde{X}) = 4$, from Theorem 7 we know that the posterior density, $\pi(\beta, \sigma, \xi | t, \nu)$ in (3) is proper. We use Metropolis-Hastings (with normal and truncated normal kernels) within Gibbs sampling algorithm for MCMC sampling. We standardize the covariate values to improve convergence of the MCMC algorithms. The R codes implementing the MCMC sampling scheme is available in the supplementary materials for the paper.
The convergence of all the results was examined by visual trace plots, autocorrelation plots as well as Geweke’s (1992) test statistic and the Gelman-Rubin scale reduction factor (Brooks and Gelman, 1998) based on multiple sequences with widely dispersed starting values. The posterior means and 95% central credible intervals for \( \xi \) and \( \sigma \) are \(-0.34(-0.62, -0.04)\), and \(1.26(0.98, 1.67)\) respectively. The baseline hazard function corresponding to \( \xi = -0.34 \) and \( \sigma = 1.26 \) is modified bathtub shaped (like the hazard function with \( \xi = -0.3 \) in Figure 1). Unlike the Weibull model, the estimates of the parameters of our GEV model successfully capture the non-monotone shape of the baseline hazard function of the cancer survival time. The posterior means and 95% central credible intervals for the intercept parameter and the regression coefficients corresponding to the three variables PS, age, and diag are \(3.72(3.25, 4.16)\), \(1.16(0.74, 1.6)\), \(0.07(-0.38, 0.52)\), and \(0.04(-0.32, 0.42)\) respectively. As noted previously by Lawless (2003), we find that the variable PS is important, whereas the other two variables are not significant.

6 Concluding remarks

In this paper, we consider the GEV link function for analyzing binomial as well as independent ordinal regression model. We have established sufficient conditions for the propriety of the posterior distributions when an improper uniform prior is used for the regression coefficients. This paper also introduces the GEV distribution in reliability and accelerated failure time models. We show that our proposed GEV model leads to an extremely flexible hazard function. We have provided sufficient conditions for the propriety of the posterior distribution that results when a GEV error distribution is used in accelerated failure time models.

Our results can be extended in various challenging directions. Extending our results for multivariate categorical response and discrete choice models are quite challenging in nature (Chen, Dey and Ibrahim (2004a)). For the life testing and survival analysis models further study can be performed for fitting regression models for ordinal response and a proportional hazards model with a frailty distribution. Although we consider the right censored data, the methodology proposed in this paper can be extended to other types of censored data, such as the left censored or interval censored data.

Appendices
A Proof of Theorem 1

Proof of Theorem 1. Define \( u, u_1, \ldots, u_{n+q} \) to be iid random variables with common distribution function \( G_\xi(\cdot) \). Let \( u^* = (\tau_1 u_1, \tau_2 u_2, \ldots, \tau_{n+q} u_{n+q})' \), where \( \tau_i \)'s are as defined in Section 3.1. Then by Fubini’s Theorem,

\[
\begin{align*}
\int_{-1}^{1} \int_{\mathbb{R}^k} f(y|\beta, \xi) d\beta d\xi & \\
& \leq \left[ \prod_{i \in I_3} (n_i) \right] E \left\{ \int_{-1}^{1} \int_{\mathbb{R}^n} I \left( \tau_i u_i \geq -\tau_i x_i' \beta, 1 \leq i \leq n + q \right) d\beta d\xi \right\} \\
& \leq \left[ \prod_{i \in I_3} (n_i) \right] \int_{-1}^{1} \int_{\mathbb{R}^{n+q}} \int_{\mathbb{R}^k} I \left( ||\beta|| \leq c \min_{\ell} \max_{m_{\ell-1} < i \leq m_\ell} |u_i| \right) d\beta dG_\xi d\xi \\
& = c^* \int_{-1}^{1} \int_{\mathbb{R}^{n+q}} \left( \min_{\ell} \max_{m_{\ell-1} < i \leq m_\ell} |u_i| \right)^k dG_\xi d\xi \\
& \leq c^* \int_{-1}^{1} \prod_{\ell=1}^p \left( \sum_{m_{\ell-1} < i \leq m_\ell} E_\xi |u_i|^{k/p} \right) d\xi ,
\end{align*}
\]

(4)

where \( c \) and \( c^* \) are two constants and \( dG_\xi = dG_\xi(u_1) \ldots dG_\xi(u_{n+q}) \).

Note that if we can show that \( E_\xi(|u|^a) \) is a continuous function of \( \xi \) when \( 0 < a < 1 \), then it will immediately follow that the expression in (4) is a finite number. Since \( u \sim G_\xi(\cdot) \) the pdf of \( u \) is given by

\[
g_\xi(u) = \begin{cases} 
\exp \left[ -\{1 + \xi u\}^{-1/\xi} \right] \frac{1}{\{1 + \xi u\}^{1/\xi + 1}} & u > -\frac{1}{\xi} \text{ if } \xi > 0 ; u < -\frac{1}{\xi} \text{ if } \xi < 0 \\
\exp(-\exp(-u)) \exp(-u) & -\infty < u < \infty \text{ if } \xi = 0 .
\end{cases}
\]

For \( \xi \neq 0 \), taking the transformation \( t = (1 + \xi u)^{-1/\xi} \), it follows that

\[
E_\xi(|u|^a) = \int_{0}^{\infty} \left( \frac{1}{\xi} \left( t^{-\xi} - 1 \right) \right)^a e^{-t} dt .
\]

Similarly when \( \xi = 0 \), taking the transformation \( t = e^{-u} \), it follows that

\[
E_\xi(|u|^a) = \int_{0}^{\infty} \left| - \log t \right|^a e^{-t} dt .
\]

For fixed \( t > 0 \) consider the following function of \( \xi \)

\[
h_t(\xi) = \begin{cases} 
\frac{t^{-\xi-1}}{\xi} & \text{if } \xi \neq 0 \\
-\log t & \text{if } \xi = 0 .
\end{cases}
\]

19
It is easy to show that for $\xi \neq 0$, $h_t'(\xi) = \{-(\xi \log t + 1)t^{-\xi} + 1\}/\xi^2$. Since $x - x \log x \leq 1$ for $x > 0$, it follows that $h_t'(\xi) \geq 0$, i.e., $h_t(\xi)$ is a nondecreasing function in $\xi$. So for fixed $t > 0$, we have

$$\frac{-t - 1}{t} \leq h_t(\xi) \leq \frac{1}{t} - 1 \text{ for } -1 \leq \xi \leq 1.$$ 

Since the above inequalities hold for any $t > 0$, if $0 < |\xi| \leq 1$ we have

$$\left| \frac{t^{-\xi} - 1}{\xi} \right| \leq \max \left( \frac{1}{t} + 1, t + 1 \right) \text{ for } t > 0,$$

and

$$|\log t| \leq \max \left( \frac{1}{t} + 1, t + 1 \right) \text{ for } t > 0.$$ 

Define

$$h(t) = \begin{cases} 
(1 + t)^{a} & \text{if } 0 < t < 1 \\
(t + 1)^{a} & \text{if } t \geq 1.
\end{cases}$$

We know from the above inequalities that for $0 < |\xi| \leq 1$

$$\left| \frac{1}{\xi} (t^{-\xi} - 1) \right|^a \leq h(t) \text{ for } t > 0,$$

and

$$|\log t|^a \leq h(t) \text{ for } t > 0.$$ 

Also since $0 < a < 1$

$$\int_{0}^{\infty} h(t)e^{-t}dt < \infty.$$ 

Since for any fixed $t > 0$, $|\frac{1}{\xi} (t^{-\xi} - 1)|^a$ is a continuous function of $\xi$, by dominated convergence theorem it follows that $E|u|^a$ is a continuous function of $\xi$, which completes the proof of Theorem 1.

\[\square\]

B Proof of Theorem 2

The proof of Theorem 2 depends on the following Proposition.

**Proposition 1.** The family of distribution functions $\{G_\xi(\cdot)\}$ is stochastically increasing.
Proof of Proposition 1. We know that a distribution function $G_\xi(\cdot)$ is stochastically larger than a distribution function $G_{\xi'}(\cdot)$ if $G_\xi(x) \leq G_{\xi'}(x)$ for all $x$. In order to prove that $\{G_\xi(\cdot)\}$ is a stochastically increasing family, we show that for given $x$, $G_\xi(x)$ is a strictly decreasing function in $\xi$.

Note that, when $\xi \neq 0$, $G_\xi(x) = \exp[-\{1 + \xi x\}^{-1/\xi}]$. Define $\tilde{G}_x(\xi) := -\log(G_\xi(x)) = \{1 + \xi x\}^{-1/\xi}$, which implies that $\log \tilde{G}_x(\xi) = -\log(1 + \xi x)/\xi$ when $1 + \xi x > 0$. For fixed $x$ we denote $d\tilde{G}_x(\xi)/d\xi$ by $\tilde{G}_x'(\xi)$. So

$$
\frac{\tilde{G}_x'(\xi)}{\tilde{G}_x(\xi)} = -\frac{\xi x - \log(1 + \xi x)}{\xi^2} = \frac{1 - (1 + \xi x) + (1 + \xi x) \log(1 + \xi x)}{\xi^2(1 + \xi x)}
$$

Then since $\tilde{G}_x(\xi) = \{1 + \xi x\}^{-1/\xi} > 0$, from the inequality $y - y \log y \leq 0$ for nonnegative $y$, it follows that for fixed $x$, $\tilde{G}_x'(\xi) > 0$ for $\xi \neq 0$. That is for fixed $x$, $\tilde{G}_x(\xi)$ is an increasing function in $\xi$ on both negative and positive half-lines ($\xi < 0$ or $\xi > 0$). Then by the mean value theorem, it follows that $\tilde{G}_x(\xi)$ is an increasing function in $\xi$ on the entire real line. Hence for fixed $x$, $G_\xi(x)$ is a decreasing function in $\xi$.

Proof of Theorem 2: If $X$ is not a full rank matrix then it is easy to see that $\int_{\mathbb{R}^k} f(y|\beta, \xi)d\beta = \infty$ for all $\xi$. Note that $0 < G_\xi(x) < 1$ for all $x \in [-1/2, 1/2]$ if $\xi \in [-1, 1]$. In particular, fix $\delta \in (0, 1/2)$, then $0 < G_\xi(-\delta) \leq G_\xi(\delta) < 1$ for all $\xi \in [-1/2, 1/2]$. Since $y_i$’s are binary random variables, if there does not exist any positive vector $a \in \mathbb{R}^n$ with $a'X^* = 0$, by doing similar calculations as in the proof of Chen and Shao’s (2000) Theorem 2.2, we have

$$
\int_{-1}^{1} \int_{\mathbb{R}^k} f(y|\beta, \xi)d\beta d\xi \geq \int_{-1/2}^{1/2} \int_{\mathbb{R}^k} f(y|\beta, \xi)d\beta d\xi \\
\geq c_1 \prod_{i:y_i=0} \left\{1 - \sup_{\xi \in [-1/2, 1/2]} G_\xi(\delta) \right\} \prod_{i:y_i=1} \left\{\inf_{\xi \in [-1/2, 1/2]} G_\xi(-\delta) \right\} \int_{s_1 \geq 0, |s_j| \leq \eta, 2 \leq j \leq k} ds \\
= c_1 \{1 - G_{0.5}(\delta)\}^{p_1} \{G_{0.5}(\delta)\}^{n-p_1} \int_{s_1 \geq 0, |s_j| \leq \eta, 2 \leq j \leq k} ds \\
= \infty,
$$

where $c_1$ is a nonzero constant, $\eta > 0$ is chosen such that $k\eta \max_{1 \leq i \leq n} \|x_i\| \leq \delta$, $ds = ds_1 \ldots ds_k$, $p_1 = \#\{i : y_i = 0\}$, and the first equality follows from Proposition 1.
C Proof of Theorem 3

Proof of Theorem 3. Since the likelihood function $f(y|\beta, \xi)$ is bounded, it is enough to show that the prior $\pi_1(\beta, \xi)$ is proper. We know that the Fisher information matrix $I(\beta|\xi)$ can be written as $I(\beta|\xi) = X'\Omega(\beta|\xi)X$ where $\Omega(\beta|\xi)$ is an $n \times n$ diagonal matrix with $i$th diagonal element $\omega_i = n_i v_i \delta_i^2$, and $v_i = v(x'_i \beta) = d^2 b(\theta_i)/d\theta_i^2$, and $\delta_i = \delta(x'_i \beta) = d\theta_i/d\eta_i$ is the so-called “link adjustments” ($\eta_i = x'_i \beta$).

Here we use the standard notation $\gamma$ to denote the canonical parameter for binomial distribution and $b(\theta_i) = \log(1 + e^{\theta_i})$. Then following Ibrahim and Laud (1991) we have

$$\int_{-1}^{1} \int_{\mathbb{R}^k} \pi_1(\beta, \xi) d\beta d\xi \leq \sum_{T} (c(x_{i_1}, x_{i_2}, \ldots, x_{i_k}))^{1/2} \int_{-1}^{1} \prod_{j=1}^{k} n_{ij}^{1/2} v_{ij}^{1/2}\delta_{ij} d\beta d\xi, \quad (5)$$

where $T = \{(i_1, i_2, \ldots, i_k) : 1 \leq i_1 < \cdots < i_k \leq n\}$, $x'_{ij}$ is the $i$th row of $X$, $c(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = |X_s|^2$ and $X_s$ is a $k \times k$ matrix with $j$th column $x_{ij}$. Now without loss of generality we can assume that $X_s$ is non-singular since otherwise $c(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = 0$. Then as in Ibrahim and Laud (1991), considering the transformation $u = X_s \beta$ and letting $r_{ij} = \theta(u_{ij})$, it follows that a non-zero term in the expression on the right hand side of (5) is proportional to

$$\prod_{j=1}^{k} n_{ij}^{1/2} \int_{-\infty}^{1} \int_{-\infty}^{\infty} \left(\frac{d^2 b(r_{ij})}{d r_{ij}^2}\right)^{1/2} dr_{ij} d\xi.$$

Then the proof follows from the fact that $\int_{-\infty}^{1} \int_{-\infty}^{\infty} \left(\frac{d^2 b(r_{ij})}{d r_{ij}^2}\right)^{1/2} dr_{ij} d\xi = 2\pi$. \hfill $\square$

D Proof of Theorem 4

Proof of Theorem 4. Let $r_1, r_2, \ldots, r_n$ be $n$ iid random variables with common distribution $G(\cdot)$. So

$$G(\gamma_{y_i} - x'_i \beta) - G(\gamma_{y_i-1} - x'_i \beta) = \int I(\gamma_{y_i-1} - x'_i \beta < r_i \leq \gamma_{y_i} - x'_i \beta) dG(\gamma_i).$$

So by Fubini’s Theorem,

$$c_1(y) = \int_{-1}^{1} \int_{\mathbb{R}^k} \cdots \int_{\mathbb{R}^k} \left\{ \prod_{i=1}^{n} \left[ G(\gamma_{y_i} - x'_i \beta) - G(\gamma_{y_i-1} - x'_i \beta) \right] \right\} d\beta d\gamma d\xi$$

$$= \int_{-1}^{1} \int_{\mathbb{R}^k} \cdots \int_{\mathbb{R}^k} I\left\{ \gamma_{y_i-1} - x'_i \beta < r_i \leq \gamma_{y_i} - x'_i \beta; 1 \leq i \leq n \right\} d\beta d\gamma dG(\bar{r}) d\xi,$$
where \( dG_\xi(\tilde{r}) = dG_\xi(r_1) \cdots dG_\xi(r_n) \). Let
\[
h(\tilde{r}) = \int_{\gamma_1 < \cdots < \gamma_{J-1}} \cdots \int_{\mathbb{R}^k} I\left\{ \gamma_{y_i-1} - x'_i \beta < r_i < \gamma_{y_i} - x'_i \beta; 1 \leq i \leq n \right\} \, d\beta d\gamma.
\]
Then,
\[
c_1(y) = \int_{-1}^{1} \int_{\mathbb{R}^n} h(\tilde{r}) dG_\xi(\tilde{r}) d\xi.
\]
We will show that
\[
h(\tilde{r}) \leq C \min_{1 \leq i \leq p} \max_{j \in Q_i} |r_j|^{k+J-1},
\]
where \( Q_i = U_i \cup L_i \cup M_i \) and \( C \) is a constant depending on \( X \) and \( y \) only. Then, of course,
\[
c_1(y) \leq C \int_{-1}^{1} \int_{\mathbb{R}^n} \min_{1 \leq i \leq p} \max_{j \in Q_i} |r_j|^{k+J-1} dG_\xi(\tilde{r}) d\xi
\]
\[
\leq C \int_{-1}^{1} \prod_{i=1}^{p} \left( \sum_{j \in Q_i} E_\xi |r_j|^{(k+J-1)/p} \right) d\xi.
\]
Since from the proof of Theorem 1 we know that \( E_\xi |r_j|^{(k+J-1)/p} \) is a continuous function of \( \xi \), it follows that \( c_1(y) < \infty \). Now, we show that (6) holds.

Consider the following transformation \( \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{J-1}) \rightarrow \theta = (\theta_1, \theta_2, \ldots, \theta_{J-1}) \) with \( \theta_1 = \gamma_1, \theta_i = \gamma_i - \gamma_{i-1} \) for \( 2 \leq i \leq J - 1 \). The Jacobian of this transformation can be shown to be 1. Let \( \tilde{\theta} = (\theta_2, \theta_3, \ldots, \theta_{J-1}) \). Thus
\[
h(\tilde{r}) = \int_{(R^+)^{J-2}} \int_{\mathbb{R}^k} \int_{-\infty}^{\infty} \left( \min_{1 \leq i \leq p} \left\{ \min_{j \in U_i \cup M_i} \left[ r_j + x'_i \beta - \sum_{\ell=2}^{y_i-1} \theta_{\ell} \right] \right\} \right) \, d\beta d\tilde{\theta}
\]
\[
\leq \int_{(R^+)^{J-2}} \int_{\mathbb{R}^k} \left( \min_{1 \leq i \leq p} \left\{ \min_{j \in U_i \cup M_i} \left[ r_j + x'_i \beta - \sum_{\ell=2}^{y_i-1} \theta_{\ell} \right] \right\} \right) \, d\beta d\tilde{\theta},
\]
\[
\text{where } f(i) = \min_{j \in U_i \cup M_i} \left( r_j + x'_i \beta - \sum_{\ell=2}^{y_i-1} \theta_{\ell} \right).
\]
and
\[ g(i) = \max_{j \in L_i \cup M_i} \left( r_j + x_j' \beta - \sum_{\ell=2}^{y_i} \theta_\ell \right). \]

Then by doing a similar calculation as in Chen and Shao (1999), we get
\[ f(i) - g(i) = \min_{j \in L_i \cup M_i} \left( -r_j - x_j' \beta + \sum_{\ell=2}^{y_j} \theta_\ell \right) - \max_{j \in U_i \cup M_i} \left( -r_j - x_j' \beta + \sum_{\ell=2}^{y_j} \theta_\ell \right) \]
\[ \leq 2 \max_{j \in Q_i} |r_j| - \tilde{M} \left\{ \max_{j \in U_i \cup M_i} \left( \sum_{\ell=2}^{y_j} \theta_\ell \frac{\theta_\ell}{M} - \sum_{\ell=1}^{k} x_{j} \beta_\ell \right) \right\}, \]

where
\[ \tilde{M} = \max(\beta_\ell, \theta_\ell) > 0. \]

Define
\[ d_i = \inf_{0 \leq a_\ell \leq 1, 2 \leq \ell \leq J-1 -1 \leq b_r \leq 1 \leq r \leq k} \left\{ \max_{j \in U_i \cup M_i} \left( \sum_{\ell=2}^{y_j} a_\ell + \sum_{r=1}^{k} x_{j} b_r \right) - \min_{j \in L_i \cup M_i} \left( \sum_{\ell=2}^{y_j} a_\ell + \sum_{r=1}^{k} x_{j} b_r \right) \right\}, \]

and \[ d = \min_{1 \leq i \leq p} d_i. \] So,
\[ f(i) - g(i) \leq 2 \max_{j \in Q_i} |r_j| - \tilde{M} d. \]

Further from (7) we know that we only need to consider the case when
\[ 0 \leq \min_{1 \leq i \leq p} \left( f(i) - g(i) \right). \]

Thus if \( d > 0 \) then
\[ \tilde{M} \leq \frac{2}{d} \min_{1 \leq i \leq p} \max_{j \in Q_i} |r_j|, \]

which will imply that
\[ h(\tilde{r}) \leq C \min_{1 \leq i \leq p} \max_{j \in Q_i} |r_j|^{k+J-1}, \]
where $C$ is a constant. Thus (6) will be proved if we can show that $d > 0$.

Since $d = \min_{1 \leq i \leq p} d_i$, if we show that $d_i > 0$ for all $i$, it will imply that $d > 0$ and hence we will be done with the proof of Theorem 4. We show that (A2)' implies that $d_i > 0$ for all $i = 1, \ldots, p$. Doing similar calculations as in Chen and Shao (1999), we can show that (A2)' implies that $\forall 0 \leq a_v \leq 1, 2 \leq v \leq J - 1$, and $\forall -1 \leq b_r \leq 1, 1 \leq r \leq k, \Sigma |b_r| > 0$,

$$\min_{j \in L_i \cup M_i} \left( \sum_{v=2}^{y_j} a_v + \sum_{r=1}^{k} x_{jr} b_r \right) \leq \max_{j \in U_i \cup M_i} \left( \sum_{v=2}^{y_j-1} a_v + \sum_{r=1}^{k} x_{jr} b_r \right),$$

and the equality in (8) holds only if

$$\sum_{r=1}^{k} x_{jr} b_r = c$$

for some constant $c$ and for all $j \in Q_i$. That is, the equality in (8) holds only if

$$X_{1i} \left( \frac{-c}{\tilde{b}} \right) = 0$$

where $\tilde{b} = (b_1, \ldots, b_k)'$. But this contradicts the fact that $X_{1i}$ is assumed to be of full column rank. Since $a_v.'s$ and $b_r.'s$ are defined on compact intervals, it follows that $d_i > 0$ for all $i = 1, \ldots, p$, which completes the proof.

\[\Box\]

### E Proof of Theorem 5

**Proof of Theorem 5.** Doing similar calculations as in Chen and Shao (1999), we can show that

$$c_1(y) \leq c_1 \sum_{i=1}^{y_k} \int_{-1}^{1} E \left[ \min_{(i,j) \in T(\eta)} (|r_i| + |r_j|)^{k+J-1} \right] d\xi$$

$$\leq 2^{k+J-1} c_2 \sum_{i=1}^{y_k} \int_{-1}^{1} \prod_{i \in U(\eta)} E\xi \left( |r_i|^{\frac{k+J-1}{\#U(\eta)}} \right) + \prod_{j \in L(\eta)} E\xi \left( |r_j|^{\frac{k+J-1}{\#L(\eta)}} \right) d\xi,$$

where $c_1$ and $c_2$ are two finite constants. Since $M^* > k + J - 1$, from the proof of Theorem 1, it follows that the integrand in (9) is a continuous function of $\xi$ and hence $c_1(y) < \infty$. 

\[\Box\]

### F Proof of Theorem 6

**Proof of Theorem 6.** Since $f_0(t) = \frac{e^{-\frac{1}{t^2}}}{t^2}$, by differentiating we get

$$f_0'(t) = \frac{e^{-\frac{1}{t^2}}(1 - 2t)}{t^4}.$$
So we obtain
\[\eta(t) := -\frac{f_0'(t)}{f_0(t)} = \frac{2t - 1}{t^2},\]
and
\[\eta'(t) = \frac{2(1 - t)}{t^3}.\]

Hence from Glaser (1980) it follows that \(\lambda_0(t)\) is either upside-down or a decreasing function of \(t\). Then the proof follows from the fact that \(\lim_{t \to 0} \lambda_0(t) = 0\).

\[\square\]

G Proof of Theorem 7

Proof of Theorem 7. Recall the likelihood function \(L(\beta, \sigma, \xi|t, \nu)\) in (2) from Section 5.2. Note that if \(\nu_i = 0\), then

\[\left\{1 - \exp\left[-\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{-\frac{1}{\xi}}\right]\right\}^{1-\nu_i} \leq 1. \tag{10}\]

On the other hand when \(\nu_i = 1\), we show that there exists a finite constant \(M\) such that

\[\left\{\frac{1}{\sigma t_i\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{\frac{1}{\xi}+1}} \exp\left[-\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{-\frac{1}{\xi}}\right]\right\}^{\nu_i} \leq \frac{M}{\sigma t_i}. \tag{11}\]

For a fixed \(\xi \geq -1\), let \(f_\xi(v) = v^{\xi+1}e^{-v}, v > 0\). It can be shown that \(f_\xi(v) \leq (\xi + 1)^{\xi+1}e^{-(\xi+1)}\) for all \(v > 0\). Let \(M := \sup_{\xi \in [-1,1]} (\xi + 1)^{\xi+1}e^{-(\xi+1)}\). Then (11) follows since

\[
\frac{1}{\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{\frac{1}{\xi}+1}} \exp\left[-\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{-\frac{1}{\xi}}\right] = \left\{\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{-\frac{1}{\xi}}\right\}^{\xi+1} \exp\left[-\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{-\frac{1}{\xi}}\right].
\]

Since \(\tilde{X}\) is of full rank, there must exist \(k\) linearly independent covariate vectors \(x_{i_1}, \ldots, x_{i_k}\) such that \(\nu_{i_1} = \cdots = \nu_{i_k} = 1\). Without loss of generality, we assume that \(i_1 = 1, \ldots, i_k = k\).

The posterior density \(\pi(\beta, \sigma, \xi)\) in (3) is proper if

\[\int_{-\infty}^{1}\int_{0}^{\infty} \int_{\mathbb{R}^k} L(\beta, \sigma, \xi|t, \nu)\pi(\sigma)\pi(\xi)d\beta d\sigma d\xi < \infty.\]

As before let \(N_k = \{1, 2, \ldots, k\}\). From (10) and (11) we have

\[
\int_{-\infty}^{1}\int_{0}^{\infty} \int_{\mathbb{R}^k} L(\beta, \sigma, \xi|t, \nu)\pi(\sigma)\pi(\xi)d\beta d\sigma d\xi \leq \int_{-\infty}^{1}\int_{0}^{\infty} \int_{\mathbb{R}^k} \left(\prod_{i:i_{\nu_i}=0} \frac{1}{\sigma t_i}\right) \left(\prod_{i,i_{\nu_i}=1,i\notin N_k} \frac{M}{\sigma t_i}\right) \\
\times \left(\prod_{i=1}^{k} \frac{1}{\sigma t_i\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{\frac{1}{\xi}+1}} \exp\left[-\left(1 + \frac{y_i - x_i'\beta}{\sigma}\right)^{-\frac{1}{\xi}}\right]\right) \pi(\sigma)\pi(\xi)d\beta d\sigma d\xi. \tag{12}
\]
Consider the transformation \( w_i = x_i' \beta, i = 1, 2, \ldots, k \). This is a one-to-one, linear transformation. Thus the right hand side of (12) is proportional to
\[
\int_{-1}^{1} \int_{0}^{\infty} \frac{1}{\sigma^{m-k}} \left( \prod_{i=1}^{k} \int_{\mathbb{R}} \frac{1}{\sigma} \left( 1 + \frac{y_i - w_i}{\sigma} \right)^{-1+\frac{1}{2}} \exp \left[ - \left( 1 + \frac{y_i - w_i}{\sigma} \right)^{-1+\frac{1}{2}} \right] dw_i \right) \pi(\sigma) \pi(\xi) d\sigma d\xi
\]
\[
= \int_{-1}^{1} \pi(\xi) d\xi \int_{0}^{\infty} \frac{1}{\sigma^{m-k}} \pi(\sigma) d\sigma < \infty,
\]
completing the proof. \(\square\)

**Acknowledgments**

The authors thank one editor and two anonymous reviewers for helpful comments and valuable suggestions which led to several improvements in the manuscript.

**References**


