Application of a generalized lognormal distribution to engineering data fitting

J. Martín & C.J. Pérez
Departamento de Matemáticas, Universidad de Extremadura, Cáceres, Spain

ABSTRACT: The lognormal distribution is commonly used to model certain types of data that arise in several fields of engineering as, for example, different types of lifetime data or coefficients of wear and friction. However, a generalized form of the lognormal distribution can be used to provide better fits for many types of experimental or observational data. In this paper, a Bayesian analysis of a generalized form of the lognormal distribution is developed. Bayesian inference offers the possibility of taking expert opinions into account. This makes this approach appealing in practical problems concerning many fields of knowledge, including reliability of technical systems. The full Bayesian analysis includes a Gibbs sampling algorithm to obtain the samples from the posterior distribution of the parameters of interest. Empirical proofs over a wide range of engineering data sets have shown that the generalized lognormal distribution can outperform the lognormal one in this Bayesian context.

Keywords: Bayesian analysis, Generalized normal distribution, Engineering data, Lognormal distribution, Markov chain Monte Carlo methods.

1 INTRODUCTION

The lognormal distribution is commonly used to model certain types of data that arise in several fields of engineering as, for example, different types of lifetime data (see, e.g., Meeker and Escobar (1998)) or coefficients of wear and friction (see, e.g., Steele (2008)). Particular properties of the lognormal random variable (as the non-negativeness and the skewness) and of the lognormal hazard function (which increases initially and then decreases) make lognormal distribution a suitable fit for some engineering data sets. However, a generalized lognormal distribution can be used to provide better fits for many types of experimental or observational data.

If a random variable $X$ has a lognormal distribution, the random variable $Y = \log X$ is normally distributed. This allows the more well-known analysis techniques for the normal distribution to be applied to the lognormal data through transformation. The previous relationship suggests a possible generalization of the lognormal distribution by using a similar transformation for the generalized normal distribution. The generalized normal distribution (GN for short) is a generalization of the normal distribution that also has the $t$ and Laplace distributions as particular cases (see Nadarajah (2005)). In fact, it is a re-parametrization of the exponential power (EP) distribution. The first formulation of the EP distribution could be attributed to Subbotin (1923). Since then, several different parameterizations can be found in the literature (see, for example, Box and Tiao (1973), Gómez et al., (1998), and Mineo and Ruggieri (2005)). This family provides distributions with both heavier and lighter tails compared to the normal ones. The GN distributions allow the modeling of kurtosis providing, in general, a more flexible fitting to experimental data than the normal distributions. The main reason why the GN or EP distributions have not been used as often as desirable has been purely computational, i.e., because most standard statistical software did not contain procedures using GN distributions. Currently, the GN distribution is considered as a factible alternative to the normal one as a general distribution for random errors.

In this paper, a generalized lognormal distribution (logGN for short) is analyzed from a Bayesian viewpoint. If a random variable $X$ has a logGN distribution, the random variable $Y = \log X$ is distributed as a GN. The logGN distribution has the lognormal one as a particular case. Bayesian inference offers the possibility of taking expert opinions into account. This makes this approach appealing in practical problems concerning many fields of knowledge, including reliability of technical systems. The Bayesian approach is also interesting when no prior information is obtained,
in this case a noninformative prior distribution is used. The full Bayesian analysis includes a Gibbs sampling algorithm to obtain the samples from the posterior distribution of the parameters of interest. Then, the predictive distribution can be easily obtained. Empirical proofs over a wide range of engineering data sets have shown that the generalized lognormal distribution can outperform the lognormal one in this Bayesian context.

The outline of this work is as follows. In Section 2, the logGN distribution is described. Section 3 presents the Bayesian analysis with both noninformative and informative prior distributions. An example with friction data illustrates the application of the proposed approach in Section 4. Finally, Section 5 presents the main conclusions.

2 THE GENERALIZED LOGNORMAL MODEL

If a random variable $X$ has a logGN distribution, the random variable $Y = \log X$ is distributed as a GN. Therefore, the probability density function of a logGN distribution with parameters $\mu, \sigma$, and $s$ is given by:

$$f(x) = \frac{s}{2x\sigma \Gamma\left(\frac{1}{s}\right)} \exp\left(-\left|\frac{\log x - \mu}{\sigma}\right|^s\right),$$

with $x > 0$, $-\infty < \mu < +\infty$, $\sigma > 0$ and $s \geq 1$. Note that $\Gamma$ denotes the gamma function.

This distribution has the lognormal distribution as a particular case by taking $s = 2$ and changing $\sigma$ to $\sqrt{2}\sigma$. The log-Laplace distribution is recovered when $s = 1$. Figure 1 shows the probability density functions for some values of $s$ with $\mu = 0$ and $\sigma = 1$.

The capacity of a distribution to provide an accurate fit to data depends on its shape. The shape can be defined by the third and fourth moments and they represent the asymmetry and flatness coefficients of a given distribution. The logGN distributions allow the modeling of kurtosis providing, in general, a more flexible fit to experimental data than the lognormal distribution.

Random variates from the logGN distribution can be generated from random variates of the GN distribution via exponentiation. Since the GN distribution is a reparameterization of the EP distribution, the techniques for random generation of these distributions can be used for the GN distribution (see, for example, Devroye (1986), Johnson (1987) and Barabesi (1993)).

Walker and Gutiérrez-Peña (1999) suggested a mixture representation for the EP distribution that is adapted here to be valid for the logGN distribution. The following result will be used in the next section to determine the full conditional distributions necessary to apply the Gibbs sampling method. The proof is immediate.

**Proposition 1** Let $X$ and $U$ be two random variables such that $f(x|u) = \frac{1}{2\sigma u^{1/s}} I[\exp(\mu - \sigma u^{1/s}) < x < \exp(\mu + \sigma u^{1/s})]$ and $f(u) = \text{Gamma}(\text{shape} = 1 + 1/s, \text{scale} = 1)$, then $X \sim \log \text{GN}(\mu, \sigma, s)$.

This result can also be used to generate random variates from a logGN($\mu, \sigma, s$). Generating from $U$ is standard, and generating from $X|U$ is obvious through the inverse transformation method. Then, the algorithm to generate random values is given by the following steps:

1. Generate $W \sim \text{Gamma}(1 + 1/s, 1)$
2. Generate $V \sim \text{Uniform}(-1, 1)$
3. Set $X = \exp\{\sigma W^{1/s}V + \mu\}$

The next section presents a Bayesian analysis for the logGN distribution.

3 BAYESIAN ANALYSIS

Bayesian analyses with both noninformative and informative prior distributions are addressed in this section.

3.1 Noninformative case

Following the suggestions in Box and Tiao (1973) and Portela and Gómez-Villegas (2004) for the EP distribution, independence between parameters is considered.

Jeffreys’ noninformative prior distributions are considered here. Jeffreys’ choice for the noninformative density is $\pi(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher
information for $\theta$ (see, e.g., Box and Tiao (1973) or Gelman et al. (2004)). This prior distribution is noninformative in the sense that it maximizes the entropy. In order to obtain the expressions for the noninformative prior distributions of the parameters, the calculation of the Fisher information matrix is required. The Fisher information matrix for the logGN distribution is given by:

$$I(\mu, \sigma, s) = \begin{pmatrix} \frac{(s-1)\Gamma(1-1/s)}{\sigma^{1-1/s}} & 0 & 0 \\ 0 & \frac{s^{1+2/s}}{\sigma^2} & -\frac{A}{\sigma^2 s} \\ 0 & \frac{A}{\sigma^2 s} & (1+1/s)\psi'(1+1/s)+A^2-1 \end{pmatrix},$$

where $\psi$ is the digamma function and $A = \log(s) + \psi(1+1/s)$.

Noninformative prior distributions are derived for the parameters, i.e.,:

$$\pi(\mu) \propto 1,$$
$$\pi(\sigma) \propto \frac{1}{\sigma},$$
$$\pi(s) \propto \sqrt{(1 + 1/s)\psi'(1+1/s) + A^2-1},$$

with $-\infty < \mu < +\infty, \sigma > 0$ and $s \geq 1$.

Since the expression for $\pi(s)$ is very involved, a simple and similar distribution is used, i.e., $\pi(s) \propto 1/s$, see Figure 2.

A Markov Chain Monte Carlo (MCMC) method is applied to generate samples from the posterior distribution. Specifically, a Gibbs sampling algorithm is derived. The mixture representation given in Proposition 1 is used here to obtain the likelihood. Then, the likelihood of a sample $x = (x_1, x_2, \ldots, x_n)$, given the vector of mixing parameters $u = (u_1, u_2, \ldots, u_n)$, is:

$$L(\mu, \sigma, s, u|x) = \frac{1}{(2\sigma)^n} \prod_{i=1}^{n} \frac{1}{x_i u_i^{1/s}} I[e^{\mu - \sigma u_i^{1/s}} < x_i < e^{\mu + \sigma u_i^{1/s}}].$$

Therefore, the posterior distribution is given by:

$$f(\mu, \sigma, s, u|x) \propto \frac{s^{n-1}}{\sigma^{n+1} \Gamma_n(1/s)} \prod_{i=1}^{n} \frac{e^{-u_i}}{x_i} I[e^{\mu - \sigma u_i^{1/s}} < x_i < e^{\mu + \sigma u_i^{1/s}}].$$

The full conditional distributions are derived:

$$f(\mu|\sigma, s, u, x) \propto \frac{1}{\sigma^{n+1}} e^{-\mu \log(x_i)}$$
$$f(\sigma|\mu, s, u, x) \propto \max_i \left\{ \frac{\log(x_i) - \mu}{u_i^{1/s}} \right\}$$
$$f(s|\mu, \sigma, u, x) \propto s^{n-1} \Gamma_n(1/s),$$
$$f(u_i|\mu, \sigma, s, x) \propto e^{-u_i} u_i^{s} \left( \frac{\log(x_i)}{\sigma} \right)^{s}, \quad i = 1, 2, \ldots, n.$$

Random variates from these densities can be generated by using standard methods. Note that the densities given in (1), (2) and (4) are uniform, Pareto and exponential, respectively, and they are generated by using the inverse transformation method. The density given

![Figure 2. Comparison of functions $\sqrt{T(s)}$ and $2/(3s)$](image-url)
in (3) is non-standard, but it can also be easily generated by using the rejection method (see, e.g., Devroye (1986)).

Iterative generations from the above conditional distributions produce a posterior sample of \((\mu, \sigma, s)\).

3.2 Informative case

In many situations, the data analyst is interested in including relevant initial information in the inference process. The choice of the prior distribution must be carefully determined to allow the inclusion of this information. Since the prior distribution choice depends on the problem in hand, there are multiple references related to this topic in the literature (see, e.g., DeGroot (1970), Berger (1985), Ibrahim et al. (2001), and Akman and Huwang (2001). Kadane and Wolfson (1998) present an interesting review on elicitation of expert opinion. O’Hagan (1998) considers the elicitation of engineers’ prior beliefs. Gutiérrez-Pulido et al. (2005) present a comprehensive methodology to specify prior distributions for commonly used models in reliability.

The following prior distributions have been proposed because they can accommodate many possible shapes for the kind of parameters involved in the logGN distribution. Besides, they allow to make efficient posterior calculations and recover the noninformative distribution for each parameter. The proposed prior distributions are given by:

\[
\pi(\mu) \propto p(\mu), \quad -\infty < \mu < +\infty, \quad (5)
\]

\[
\pi(\sigma) \propto \sigma^{-(a_0+1)} e^{-b_0/\sigma}, \quad \sigma > 0 \quad (6)
\]

\[
\pi(s) \propto s^{-(c_0+1)} e^{-d_0/s}, \quad s \geq 1, \quad (7)
\]

where (5) is any distribution in its support, (6) is an inverse-gamma distribution and (7) is a truncated inverse gamma distribution. Note that the noninformative prior distributions are recovered when \(p(\mu)\) is constant and \(a_0 = b_0 = c_0 = d_0 = 0\). This fact is specially interesting because it allows to use a noninformative prior distribution for one or two parameters and informative prior distributions for the remaining parameters.

Analogously to the noninformative case, the posterior distribution is derived:

\[
f(\mu, \sigma, s, \mathbf{u} | \mathbf{x}) \propto \frac{p(\mu) \sigma^{n-c_0-1} e^{-d_0/s}}{\sigma^{a_0+n+1} \Gamma(d_0) \Gamma((1/s))} \times \prod_{i=1}^{n} \frac{e^{-\mu_i}}{x_i} I[\mu - \sigma u_i^{1/s} < x_i < \mu + \sigma u_i^{1/s}].
\]

The full conditional distributions are:

\[
f(\mu | \sigma, s, \mathbf{u}, \mathbf{x}) \propto p(\mu), \quad \max_i \{\log(x_i) - \sigma u_i^{1/s}\} < \mu < \min_i \{\log(x_i) + \sigma u_i^{1/s}\} \quad (8)
\]

\[
f(\sigma | \mu, s, \mathbf{u}, \mathbf{x}) \propto \sigma^{-(a_0+n+1)} e^{-b_0/\sigma}, \quad \sigma > \max_i \left\{ \frac{|\mu - \log(x_i)|}{u_i^{1/s}} \right\} \quad (9)
\]

\[
f(s | \mu, \sigma, \mathbf{u}, \mathbf{x}) \propto s^{-(c_0+1)} e^{-d_0/s} \frac{1}{\Gamma((1/s))}, \quad \max_i \{1, a_i\} < s < \min_{i \in S^+} a_i \quad (10)
\]

\[
f(u_i | \mu, \sigma, s, \mathbf{x}) \propto e^{-u_i}, \quad u_i > \left( \frac{|\log(x_i) - \mu|}{\sigma} \right)^{i}, \quad i = 1, 2, \ldots, n. \quad (11)
\]

where \(S^- \), \(S^+ \) and \(a_i, \ i = 1, 2, \ldots, n\), are defined as in the previous subsection.

In this case, generating from these truncated densities is also easy to perform. Generating from (8) depends on the chosen density. Note that the density given in (9) is a left truncated inverse-gamma. Random variates from this distribution are obtained by taking the reciprocal of variates from a truncated gamma distribution. The density for the conditional distribution given in (10) is again non-standard. Similarly to the noninformative case, a rejection method is implemented. Finally, note that (11) is the same as (4).

4 APPLICATION TO FRICTION DATA

In basic tribology, dimensional issues and the central limit theorem are used to argue that the distribution of the coefficients of friction and wear are typically lognormal. Moreover, empirical evidences from many data sources support this argument (see, e.g., Wallbridge and Dowson (1987) and Steele (2008)). Steele (2008) recommends that engineers, without evidence to suggest differently, allocate a lognormal distribution to the coefficients of friction and wear.

The use of the proposed approach is illustrated with a data set (presented in Nica (1969)) containing the coefficients of friction of clean steel in a vacuum. The size of this data set is 23. These data are used as historical data to extract knowledge on the predictive distribution of the friction coefficient. The most usual case in engineering studies is to have some prior information about the process on which they are trying
to make inferences. This corresponds to the informative case. Then, the historical information and the prior information provided by the engineer are embedded in the posterior distribution.

The following step is to choose the hyperparameter values for the prior distributions of the parameters of interest $\mu$, $\sigma$, and $s$. There are many ways to elicit these values. One possibility is to specify the values according to previous direct knowledge on the parameters (see, e.g., Berger (1985) and Akman and Huwang (2001)). Another one consists in using partial information elicited by the expert. In this case, there are many criteria to obtain the hyperparameters values as, for example, maximum entropy and maximum posterior risk (see Savchuk and Martz (1994)). A third possibility considered here is to use expert information on the expected data and not on the parameters. This is easier for engineers who are not familiarized with parameters but have an approximate knowledge of the process. Finally, it is remarkable that noninformative prior distributions can be used for any of the parameters and informative prior distributions for the remaining ones.

In this application, the hyperparameters are obtained by using a method similar to the one presented by Gutiérrez-Pulido et al. (2005). In this case the expert is asked to provide occurrence intervals for some usual quantities as the mode, median and third quartile. The expert considered that these quantities should be in the following intervals: $[L_{\text{Mo}}, U_{\text{Mo}}] = [0.935, 0.955]$, $[L_{\text{Me}}, U_{\text{Me}}] = [0.95, 0.96]$, and $[L_{\text{Q3}}, U_{\text{Q3}}] = [0.97, 0.985]$. By using the informative prior distributions presented in subsection 3.2, with $\mu \sim N(\mu_0, \sigma_0)$, and following the development in Gutiérrez-Pulido et al. (2005), the hyperparameters obtained are $\mu_0 = -0.0460$, $\sigma_0 = 0.0017$, $a_0 = 80.8737$, $b_0 = 4.0744$, $c_0 = 60.2888$, and $d_0 = 220.4130$.

After it has been considered that the chain convergence has been achieved, a sample of size 10,000 for the parameters of the posterior distribution is generated. The 95% Highest Density Regions (HDR) for $\mu$ and $\sigma$ are shown in Figure 3. Note that the HDR for $s$ does not contain the value $s = 2$, that recovers lognormality.

In order to make a performance comparison, a similar procedure is implemented to obtain the hyperparameters in the lognormal case. Then, a posterior sample is generated by using the lognormal distribution instead of the logGN distribution. Here the 95% HDR for $\mu$ and $\sigma$ are $(-0.0494, -0.0427)$ and $(0.0923, 0.1416)$. The posterior predictive distributions are presented in Figure 4.

The comparison between both performances is presented based on the generated posterior predictive distributions. The criterion used to compare them is based on the use of the logarithmic score as a utility function in a statistical decision framework. This was proposed by Bernardo (1979) and used, for example, by Walker and Gutiérrez-Peña (1999) in a similar context. The expected utilities for the lognormal and logGN models can be estimated as:

$$
\bar{U}_0 = \frac{1}{n} \sum_{i=1}^{n} \log(p_0(x_i))
$$

$$
\bar{U}_1 = \frac{1}{n} \sum_{i=1}^{n} \log(p_1(x_i)),
$$

Figure 3. 95% HDR for $\mu$, $\sigma$, and $s$.

Figure 4. Posterior predictive distributions for friction data. Solid line: logGN, and dashed line: lognormal.
where $p_0$ and $p_1$ are the posterior predictive distributions of the lognormal and logGN models, respectively. The estimated values are $\bar{U}_0 = 1.4118$ and $\bar{U}_1 = 1.4373$, so the logGN model performs better than the lognormal one. The same happens for the noninformative model ($\bar{U}_0 = 1.5533$ and $\bar{U}_1 = 1.5881$).

5 CONCLUSION

The generalized form of the lognormal distribution, presented and analyzed from a Bayesian viewpoint, offers the possibility of taking expert opinions into account. The proposed approach represents a viable alternative to analyze data that are supposed to follow a lognormal distribution and provides flexible fits to many types of experimental or observational data. The technical development is based on a mixture representation that allows to perform inferences via Gibbs sampling. It is remarkable that the logGN family provides very flexible distributions that can empirically fit many types of experimental or observational data obtained from engineering studies.

ACKNOWLEDGEMENTS

This research has been partially supported by Ministerio de Educación y Ciencia, Spain (Project TSI2007-66706-C04-02).

REFERENCES


