Improved Probability Distribution Models for Seismic Fragility Assessment

Jianjun Qin
Assistant Professor, Shanghai Institute of Disaster Prevention and Relief, Tongji University, Shanghai, China

Kevin R. Mackie
Associate Professor, Dept. of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, US

Bozidar Stojadinovic
Chair Professor, Institute of Structural Engineering, ETH Zurich, Zurich, Switzerland

ABSTRACT: The paper presents a new formulation for the probabilistic modeling of the variables to reduce the errors between the model and the available data sets and to facilitate seismic fragility assessment. Traditionally, one probability distribution, e.g., lognormal distribution, is adopted to describe one variable throughout the domain of values. However, there are potential errors in the probability content, especially in the high tail. The consequences of such tail sensitivity are particularly detrimental in fragility assessment when the limit states under consideration result in small probabilities of failure. To address the tail sensitivity problem, the data sets of the variables are divided into two parts, i.e., bulk and high tail. Each part is considered separately for the probabilistic modeling and integrated into one continuous distribution for fragility analysis. For illustration purposes, probabilistic seismic assessment of a typical reinforced concrete bridge column is compared using the proposed framework and compared to the results based only on lognormal random variable distributions. Results show that for complementary cumulative distribution function (CCDF), the difference may reach one or more magnitudes; for limit states that exercise the tails of the random variable distributions, the difference in the fragility estimates can increase with the hazard level.

1. INTRODUCTION

Seismic risk assessment of structures is commonly performed to determine probabilities of exceeding limiting values of structural demands, damage states, or losses. In addition, when coupled with a probabilistic seismic hazard assessment at a particular site, mean rates for the structure can be computed and appraised for whether or not the risk is deemed acceptable. One framework for computing limit state probabilities, developed by the Pacific Earthquake Engineering Research (PEER) Center, disaggregates the seismic performance assessment problem into seismic intensity, engineering demand, damage, and decision probabilistic models. Quantification of seismic performance is based on the conditional probability distribution functions used in the demand, damage and decision models. These conditional probability distributions are almost always assumed to be parametric, and defined commonly as lognormal. The cumulative distribution functions (CDF) of each variable (commonly referred to as fragilities when conditioned on the intensity) are therefore fully defined once the parameters of the distribution are known.

The data used to estimate the lognormal distribution parameters in the demand, damage and decision models is obtained from experience during past earthquakes, from laboratory tests, and from simulations using mechanical models, including finite element numerical models. Given that large-magnitude earthquakes are rare, and that structures are designed to have a low probability of failure in design-basis earthquakes, this data are usually clustered in the range corresponding to small and moderate seismic intensity ranges and performance in accordance with design code expectations. Thus, it forms the central part, the bulk of the data set. A comparatively small number of data points, associated with high seismic intensities and performance outside the range of design code expectations, form the high tail of the data set. Given that a single function is used to define the fragility, the bulk of the data set, rather than its tail, govern the selection of the parameters of that function. However, the tail of the data set can be considerably different from the bulk of the data set Caers and Maes (1998). This difference may significantly reduce the reliability of seismic performance evaluation conducted using approaches such as the PEER framework.

Probabilistic extreme value theory has developed strongly in the beginning of twentieth century Kotz and Nadarajah (2000). Even though these theoretical developments have been driven by engineering applications, i.e., attempts to predict the recurrence of extreme events such as floods or earthquakes, there is comparatively little guidance on how to select the type of a tail CDF and how to fit it to the (few) available data points.

A method to improve the CDFs in the demand, damage and decision models used in the PEER framework for seismic performance assessment is proposed in this paper. First, the high tail of the data set is separated from the bulk. Then, these two parts of the data set are described using two different, appropriately selected and fitted, parametric CDFs. These two CDFs are subsequently concatenated into a single consistent distribution. The objective is to minimize the errors between the fragility model and the data set in the entire domain of realizations. To realize the data partition, the boundary of the high tail needs to be identified. A lognormal CDF is used to describe the bulk of the data, but an appropriate distribution to describe the high tail together with the values of the parameters of that distribution, needs to be defined. Once the tail is modeled, the parameters of the lognormal CDF representing the bulk are modified to be consistent with the selected distribution representing the tail of the data.

To formulate an example, a procedure for selection of ground motions, structure types, and damage models similar to that presented in Mackie, et al. (2008) is used to obtain the data sets representing the conditional relation of damage measures (DM) values to engineering demand parameter (EDP) values, and the conditional relation of EDPs to different seismic hazard intensity measures (IMs). However, in this paper, the structure is purposely simplified from an entire bridge to just a single-column bent. The seismic performance of this column bent is evaluated using the fragilities derived using the method proposed in this paper and the evaluation results are compared to those derived using only lognormal fragilities, as presented in Mackie, et al. (2008). This comparison will be used to provide guidance on implementation of the proposed strategy and to examine the effectiveness of proposed strategy to improve the fragility models.

2. METHOD TO IMPROVE THE CDF OF A RANDOM VARIABLE

As introduced above, there are two parts of the data representing the realizations of a random variable, denoted as the bulk and the high tail. The characteristics of these two parts of the data are shown in Table 1. The bulk and the high tail are modeled with different CDFs to minimize the modeling error. Three tail probability distributions described in Kotz and Nadarajah (2000), namely Gumbel maximum, Gumbel minimum, Frechet maximum, and the power law distribution, described in Clauset, et al. (2009), are used. The one that best fits the random variable realization data in the high tail is selected. The bulk is described by a lognormal CDF, whose parameters are adjusted to concatenate the two distributions such that the combined CDF can be used in the PEER seismic performance evaluation framework.

Table 1: Characteristics of the bulk and the high tail of a random variable realization data set.

	Data points	Target seismic events	Engineering consequences
Bulk	Large amounts	Small and moderate events	Low
High tail	Few	Extreme events	High

Denote X as the natural logarithm of a random variable Y ($X = \ln Y$). Let there be N data points representing realizations of X. The data points are denoted by $\boldsymbol{\xi} = \{\xi_1, \xi_2, \cdots, \xi_N\}^T$ $(\xi_1 \le \xi_2 \le \cdots \le \xi_N)$ and their CCDF by $\overline{G}(\xi_i)(i = 1, 2, \cdots, N)$. The bar on *G* differentiates between CCDF and the traditional notation for a CDF. The CCDF is:

$$\overline{G}(\xi_i) = \Pr(x > \xi_i) = \frac{N - i}{N} \qquad (1)$$

Let $\overline{F}_0(x)$ denote the normal CCDF of *X* whose parameters are computed from the entire data set $\boldsymbol{\xi}$. Let $\overline{F}_t(x)$ denote a high tail distribution (Gumbel maximum, Gumbel minimum, Frechet maximum, or power law) fitted using a portion of the data set declared to be in the high tail. Clauset, et al. (2009) presented one approach to obtain the boundary separating the high tail from the bulk of the data, denoted as x_{t_c} here. The idea is to find the data point ξ_i $(i = 1, 2, \cdots, N)$ where the difference between the CCDF of the data $\overline{G}(\xi_i)$ and that of the assumed tail distribution is the smallest. That is:

$$x_{t_c} = \left\{ \xi_i \left| \min_{\xi_i \in \boldsymbol{\xi}} \left| \overline{G}(\xi_i) - \overline{F}_t(\xi_i) \right| \right. \right\} \qquad (2)$$

Comparing absolute values of CCDFs represents a problem: while these values are small, selecting between two adjacent tail boundary candidates may make an order of magnitude difference in the probability estimate for *X*. To avoid such errors, the approach by Clauset, et al. (2009) is modified as follows:

$$x_{t_c} = \left\{ \xi_i \left| \min_{\xi_i \in \boldsymbol{\xi}} \left| \log_{10} \overline{G}(\xi_i) - \log_{10} \overline{F}_t(\xi_i) \right| \right. \right\} \qquad (3)$$

Once the boundary separating the high tail from the bulk of the data is selected as $x_{t_c} = \xi_{N-N_t+1}$, where N_t is the number of data points in the high tail, the parameters of a tail distribution can be obtained by least squares fitting to the data in the high tail. An index of the average residuals in the high tail is proposed to measure the goodness of fit and compare the high tail distribution candidates. It is defined as:

$$R = \frac{\sum_{i=N-N_t+1}^{N} \left| \log_{10}\left(\overline{G}(\xi_i)\right) - \log_{10}\left(\overline{F}_t(\xi_i)\right) \right|}{N_t} \qquad (4)$$

A logarithm is used in this measure to consider the difference of the magnitudes of CCDFs, not their values. The tail distribution with the smallest *R* value is adopted as $\overline{F}_t(x)$.

The CDF representing the bulk of the data, $F_b(x)$, is defined next. This CDF has to satisfy the following two conditions:

(1) $F_b(-\infty) = 0$; and

(2) $F_b(x_{t_c}) = F_t(x_{t_c}) = 1 - \overline{F}_t(x_{t_c})$.

Lind and Hong (1991) proposed a tail entropy approximation approach applied here to modify the normal CDF $F_0(x)$ to obtain $F_b(x)$. The idea is to scale $F_0(x)$ such that:

$$F_b(x) = \frac{F_t(x_{t_c})}{F_0(x_{t_c})} F_0(x) \qquad (5)$$

Finally, the CDF of the random variable *X*(=ln*Y*) is:

$$F(x) = \begin{cases} \dfrac{F_t(x_{t_c})}{F_0(x_{t_c})} F_0(x) & x < x_{t_c} \\ F_t(x) & x \geq x_{t_c} \end{cases} \tag{6}$$

Note that the bulk part might not necessarily follow normal distribution although $F_0(x)$ expresses the CDF of a normal distribution and the rate $F_t(x_{t_c})/F_0(x_{t_c})$ would generally be close to 1, as can be seen in the Example section.

3. EXAMPLE

To illustrate the proposed method for improving CDFs used in seismic performance assessment, a single column similar to the reinforced concrete column designated as Type 1A, previously investigated in Mackie, et al. (2008), is selected. The circular flexural column has diameter 1.3m, clear height 6m with an additional 0.8m to the center of mass of the superstructure, 2% longitudinal reinforcing ratio, and 1% transverse reinforcing ratio. The nominal compression strength of concrete is 45MPa and the nominal yield strength of the reinforcement is 462MPa. The superstructure is represented by a concentrated mass weighing 3000kN. The column is assumed fixed to the foundation at the bottom. The column behaves as a cantilever in both the longitudinal and transverse directions. However, a rigid link is used to represent the portion of the column between the top clear height and the center of mass of the superstructure, potentially creating an inflection point in the column. The initial elastic fundamental vibration periods of the column are 0.54sec in each of the orthogonal lateral directions and 0.042sec in the vertical direction.

A total of 160 ground motions were applied to the OpenSees (http://opensees.berkeley.edu) model of the column to generate a rich numerically simulated data set to evaluate the CDF of an EDP conditioned on the IM. Each ground motion was comprised of three components, the two orthogonal horizontal accelerations and a vertical acceleration component. Rather than the cloud approach taken in Mackie, et al. (2008), all of the ground motions were scaled to predetermined IM levels. The square root-sum of square (SRSS) of the two orthogonal horizontal Peak Ground Velocity (PGV) values is selected as the IM. Correspondingly, the EDP is an SRSS of the obtained horizontal drifts in the two orthogonal directions. A complementary cumulative distribution (CCDF) of the natural logarithm of the EDP (lnEDP) for IM = 51.5 cm/s is shown in Figure 1. This IM intensity corresponds to design-basis seismic hazard. The mean value μ_{lnEDP} = -4.15 and standard deviation σ_{lnEDP} = 0.32. A normal CCDF of lnEDP with these parameters is also shown. Note that the abscissa of the plot is lnEDP, exaggerating the difference in EDP values. Evidently, the lognormal CDF represents the numerically simulated data reasonably well in the range where EDP values are relatively small. However, there is a significant difference between the numerically simulated data and the normal distribution of lnEDP when EDP values are large, in the high tail outlined using a red rectangle. In this example, assuming normal distribution of lnEDP extends into the high tail predicts higher probabilities of exceedance than the numerically simulated data suggests.

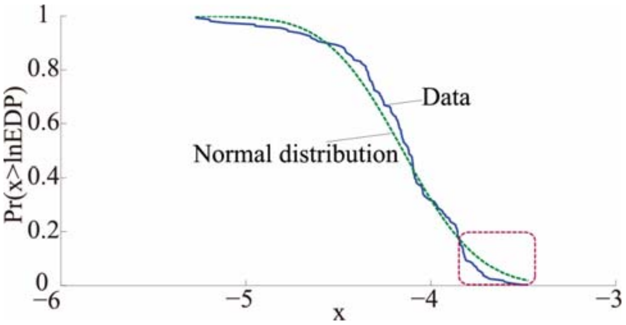

Figure 1: Comparison of the CCDF of numerically simulated lnEDP data for IM = 51.5 cm/s and a normal CCDF of lnEDP.

The boundary between the bulk and the high tail data points is established first, each of the four examined tail CDFs is fitted to the high tail data next, and, last, the lognormal bulk distribution is translated to match the tail distribution at the boundary between them. A comparison of the four considered tail CCDFs is shown in Figure 2 (for the tail region in the rectangle in Figure 1) and in Table 2. In this case, the Gumbel maximum distribution provides the best fit to the high tail data. Note that the power law and the Frechet maximum distribution both have a value of *R* larger than 1, meaning that the average difference between the numerically simulated data and the fitted CCDF is larger than one order of magnitude of the data. The scaling ratio $F_t\left(\ln \mathrm{EDP}_{t_c}\right)/F_0\left(\ln \mathrm{EDP}_{t_c}\right)$ is always close to one, showing that the modification of the lognormal distribution for entire data set to fit the bulk of the data and concatenate to the selected high tail distribution is small. Finally, the power law distribution leaves only four points in the high tail, but generates a very large average residual error.

Table 2: High tail boundary and the parameters of the tail distributions for the lnEDP data set.

Type of distribution	Parameters			
	High tail boundary $\ln \mathrm{EDP}_{t_c}$	High tail boundary point number N_t	Bulk distribution translation ratio $\frac{F_t\left(\ln \mathrm{EDP}_{t_c}\right)}{F_0\left(\ln \mathrm{EDP}_{t_c}\right)}$	Average residual measure R
Power law	-3.61	4	1.02	1.69
Gumbel max	-3.88	38	0.96	0.03
Gumbel min	-3.92	42	0.96	0.76
Frechet max	-3.82	18	1.05	1.25

To generate a CDF for DM conditioned on EDP, the 49 points providing the column drift ratio values for the longitudinal reinforcement bar buckling damage measure for circular spirally reinforced bridge columns are extracted from the data in the PEER column performance database (Berry and Eberhard (2003)). The DM data points are obtained from the concrete columns with different sizes and different material properties compared to the column used in this example. The mean value μ_{lnDM} = -2.86 and standard deviation σ_{lnDM} = 0.44 are computed from this data set.

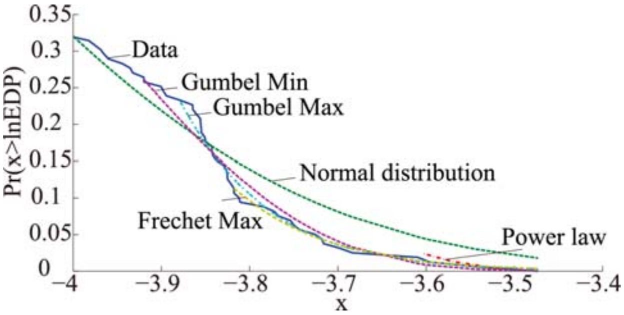

Figure 2: Comparison of the four high tail distributions for lnEDP.

Table 3: Boundary and the parameters of the tail distributions for the data set of lnDM.

Type of distribution	Parameters			
	High tail boundary $\ln \mathrm{DM}_{t_c}$	High tail boundary point number N_t	Bulk distribution translation ratio $\frac{F_t\left(\ln \mathrm{DM}_{t_c}\right)}{F_0\left(\ln \mathrm{DM}_{t_c}\right)}$	Average residual measure R
Power law	-2.26	5	0.98	1.32
Gumbel max	-2.83	20	1.12	0.05
Gumbel min	-2.68	13	1.10	0.68
Frechet max	-2.77	17	1.12	0.72

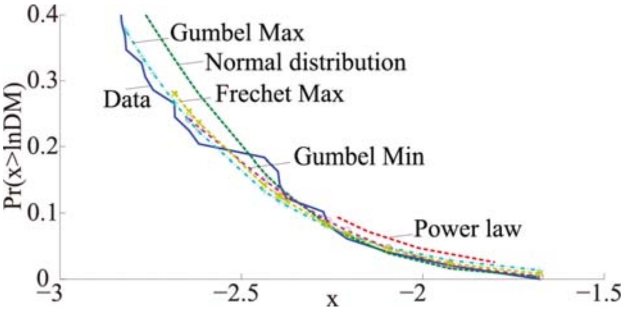

Figure 3: Comparison of the four high tail distributions for lnDM.

The improved CDF for the damage model is computed following the proposed procedure. A comparison of the four considered tail CCDFs is shown in Figure 3 (for the tail region only) and in Table 3. In this case, the Gumbel maximum distribution again provides the best fit in the high tail. The scaling factors required to concatenate the bulk lognormal distribution are somewhat higher than in the lnEDP case, but are still close to 1. Comparing the number of points in the high tail normalized by the number of data points in the data set shows that the tail of lnDM contains (proportionally) more points than the tail of lnEDP.

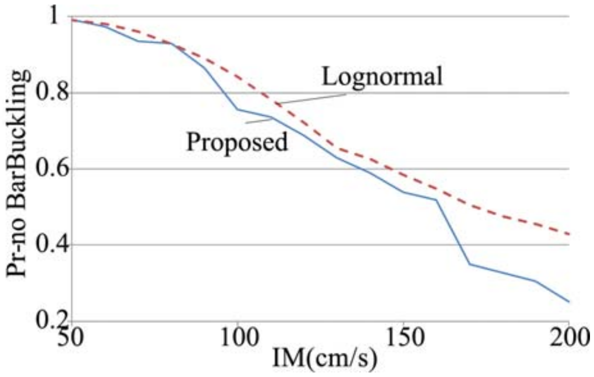

Figure 4: Comparison of the conditional probability of bar buckling computed using the improved and the conventional CDFs for the demand and the damage model.

The improved CDFs for the demand and damage models can now be constructed using Eq. (6). Note that the conventional CDFs for the demand and damage model are lognormal with parameters computed using all data. A comparison of the probability of bar buckling conditioned on the value of IM computed using the improved and the conventional CDFs is made using Monte Carlo simulation. The results are shown in Figure 4. From the figure, it can be seen that the conventional CDFs underestimates the probability of bar buckling and the difference increases with the hazard level.

4. CONCLUSIONS

Traditionally, seismic fragility assessment of structures assumes that the variables relating, for example, intensity, damage and demand follow the lognormal distribution, which is defined by the mean value and the standard deviation of the collected data. This assumption is a potential source of error in the fragility assessment, especially in the tail, the magnitude of which might reach one or more orders of magnitude in the estimation of CCDF. However, the high tail of the data sets of the variables, which represent extreme events and have the potential for broader societal consequences, is generally of interest in risk analysis and decision making.

This paper proposed an improved probability model for seismic fragility assessment that directly accounts for the division of the data into two sets, i.e., bulk and high tail. In the high tail, four tail distributions (Gumbel max, Gumbel min, Frechet max, and power law) are tested to fit the data and the relevant parameters are defined. The best distribution to describe the high tail is identified by the proposed index of the average residuals. Afterwards, the distribution of the bulk part is also modified correspondingly. The two parts of the probabilistic model are concatenated into one continuous distribution. The procedure is illustrated using a single reinforced concrete bridge column, simplified from a larger set of common reinforced concrete bridge models. The proposed probability models are fitted for both the demand models obtained using nonlinear time history analysis, and damage models obtained from databases of experimental column tests. Results show the Gumbel max distribution models the tail well for both the demand and damage models for the realizations data. The errors in the fragility estimates can increase with the hazard level.

In the example, it is assumed that the intensity of the seismic hazard is constant. When the seismic hazard for a site is known, the reliability assessment can be integrated with the hazard to yield mean rates of exceeding the limit states of interest. If the hazard curve focus on the high hazard level, e.g. IM (PGV) is larger than 150cm/s, the integration would make the error remarkable.

5. ACKNOWLEDGEMENTS

The first author greatly acknowledges the support from the National Natural Science Foundation of China [grant number 51408438]; the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State

Education Ministry of the People’s Republic of China; and Shanghai Pujiang Program [grant number 14PJ1408300].

6. REFERENCES

Caers, J., and Maes, M. A. (1998). "Identifying tails, bounds and end-points of random variables." *Structural Safety*, 20(1), 1-23.

Kotz, S., and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*, Imperial College Press.

Mackie, K. R., Wong, J.-M., and Stojadinovic, B. (2008). "Integrated Probabilistic Performance-Based Evaluation of Benchmark Reinforced Concrete Bridges." Pacific Earthquake Engineering Research Center, College of Engineering, University of California, Berkeley, 199.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). "Power-Law Distributions in Empirical Data." *SIAM Review*, 51(4), 661-703.

Lind, N. C., and Hong, H. P. (1991). "Tail Entropy Approximations." *Structural Safety*, 10(4), 297-306.

Berry, M., and Eberhard, M. (2003). "Performance Models for Flexural Damage in Reinforced Concrete Columns." Department of Civil & Environmental Engineering, University of Washington, Berkeley, 162.