LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTER FOR QUANTITATIVE RISK ANALYSIS
CEQURA

Bakhodir Ergashev, Stefan Mittnik and Evan Sekeris

A Bayesian Approach to Extreme Value Estimation in Operational Risk Modeling

Working Paper Number 10, 2013
Center for Quantitative Risk Analysis (CEQURA)
Department of Statistics
University of Munich

http://www.cequra.uni-muenchen.de

A Bayesian Approach to Extreme Value Estimation in Operational Risk Modeling*

Bakhodir Ergashev
The Federal Reserve Bank of Richmond
Charlotte Office, PO Box 30248, Charlotte, NC 28230, USA
Bakhodir.Ergashev@rich.frb.org

Stefan Mittnik†
Department of Statistics and Center for Quantitative Risk Analysis
Ludwig Maximilians University Munich
Akademiestr. 1/I, 80799 Munich, Germany
finmetrics@stat.uni-muenchen.de

Evan Sekeris
AON
9841 Broken Land Parkway
Columbia, MD 21046, USA
Evangelos.Sekeris@AON.com

October 22, 2013

Abstract

We propose a new approach for estimating operational risk models under the loss distribution approach from historically observed losses. Our method is based on extreme value theory and, being Bayesian in nature, allows us to incorporate other external information about the unknown parameters by use of expert opinions via elicitation or external data sources. This additional information can play a crucial role in reducing the statistical uncertainty about both parameter and capital estimates in situations where observed data are insufficient to accurately estimate the tail behavior of the loss distribution. Challenges of and strategies for formulating suitable priors are discussed. A simulation study demonstrates the performance of the new approach.

Keywords: Generalized Pareto distribution; Model uncertainty; Risk capital; Quantile distance; Simulated annealing

*We would like to thank an anonymous referee for constructive comments. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Banks of Richmond or the Board of the Governors of the Federal Reserve System.

†Corresponding author.

1 Introduction

Extreme value approaches to modeling operational risk severity have gained in popularity in recent years. As a theory that focuses exclusively on understanding and quantifying the behavior of processes at extreme levels, extreme value theory (EVT) seems to be a natural candidate for operational risk modeling. According to EVT, the generalized Pareto distribution (GPD) is well suited to model extreme losses, because, under a broad set of assumptions, it represents the domain of attraction of independent losses beyond a high–level threshold, called the GPD threshold in the sequel. Its popularity in operational risk modeling stems from a general consensus among researchers that historically observed operational losses appear to be heavy tailed (cf. Moscadelli, 2004; and de Fontnouvelle, Rosengren, and Jordan, 2005). Given its successful implementation in modeling similarly heavy–tailed phenomena in other fields and the heavy–tailed losses observed in operational risk, EVT seems to be a natural modeling choice. The adoption of EVT in operational risk modeling has, however, encountered a number of obstacles. The biggest hurdle faced by modelers using EVT in operational risk is the limited sample size. While data scarcity is not unique to EVT and affects operational risk modeling in general, it troubles EVT significantly more, because of its need for sufficiently large tail–event samples. A quick glance at the EVT literature in other fields shows that the technique is commonly applied to large samples and, correspondingly, a large number of tail events (see, for example, Coles and Tawn, 1996).

Data sufficiency is a major but not the only problem faced by practitioners trying to fit a GPD to tail data. Identifying the appropriate GPD threshold is a challenging but crucial task, because a misidentified threshold can greatly impact parameter estimates. The most popular techniques used to determine the threshold, such as the Hill estimator, do not always provide clear answers. Some recommend defining a fixed percentage as tail data (for example, 10% of the largest losses). This strategy benefits from being rule–based, so that the modeler does not have to choose the threshold. However, it may cause tail samples to be contaminated with non-tail observations.

The combination of small samples and issues of threshold identification leads to substantial statistical uncertainty for both parameter and capital estimates. In particular, estimating the shape parameter—i.e., the parameter that plays the dominant role in shaping the tail of the GPD distribution—with sufficient accuracy is extremely difficult with modest samples.[1] Small changes in the estimates of the shape parameter tend to

[1]Ruckdeschel and Horbenko (2013) study robustness properties of several procedures for joint estimation of shape and scale in a GPD model. Robust statistics in this context should provide reliable inference in the presence of moderate deviations from the distributional model assumptions. Our approach is different. We rely on prior assumptions in achieving acceptable levels of stability in parameter

have large and asymmetric impacts on capital estimates. Because of the concave relationship between the shape parameter and the implied capital charge, underestimating the shape parameter will lead to a relatively small underestimation of capital, whereas overestimating the shape parameter can lead to explosive capital charges. Consequently, the resulting capital bias will, on average, be positive and, potentially, large. This is confirmed by research indicating that EVT tends to significantly overestimate operational risk capital—even in reasonably large samples (see, for example, Mignola and Ugoccioni, 2006).[2]

In order to combat the scarcity of severe historical losses, institutions use external data and scenario analysis to complement internal loss data. Both data sources come, however, with their own challenges. On the one hand, external data are usually not directly usable and require scaling; and, because of our poor understanding of the root causes of operational risk, no solid scaling mechanism has yet been developed. On the other hand, scenario analysis data are difficult to incorporate in a quantitative model.

In this paper, we propose a Bayesian estimation method for EVT–based operational risk models which allows us to address both the statistical uncertainty around parameter estimates and the incorporation of alternative sources of information into the modeling process. As a result, the approach we put forth produces compound distributions that are advantageous in operational risk modeling. The proposed method is based on treating the unknown parameters as random variables and deriving suitable estimates, using the Markov chain Monte Carlo (MCMC) methods. By doing so, we avoid maximum likelihood estimation, which can be highly problematic due to the "erratic" behavior of the likelihood function. Moreover, a Bayesian approach allows us to incorporate additional information into the estimation process in form of priors on unknown parameters. This information can be based on expert opinion, scenario analysis, or derived from external data. When the recorded data are insufficient to accurately estimate unknown parameters, priors make it possible to focus on the most plausible region of the parameter space. The use of expert opinions for specifying prior distributions is not necessarily straightforward, as experts may not be familiar with probabilistic descriptions of quantities of interest. However, elicitation techniques have been developed in Bayesian statistics (see, for example, Garthwaite, Kadane and O'Hagan, 2005). As will be discussed, direct and indirect elicitation of expert opinion can be useful strategies for the proposed method.

The paper is organized as follows. Section 2 discusses some important challenges

and capital estimates while being mindful about the potential for significant biases strong priors are capable of generating.

[2]A good overview of the challenges of fitting EVT–based models can be found in Diebold, Schuermann and Stroughair (1997) and Embrechts, Klüppelberg, and Mikosch (1997).

associated with fitting operational risk models with conventional estimation methods. Section 3 introduces the model and our estimation approach and discusses direct and indirect elicitation methods. Section 4 describes the MCMC algorithm employed. In Section 5, we discus the results of a simulation study that assesses how well the proposed method works. Section 6 concludes.

2 Challenges of extreme value estimation

EVT implies that, under certain assumptions, the tail behavior of extreme events closely resembles the GPD with cumulative distribution function (cdf)

$$F_1(x|\tau,\beta,\xi) \;=\; 1-\left(1+\xi\frac{x-\tau}{\beta}\right)^{-\frac{1}{\xi}},\; x>\tau, \tag{2.1}$$

where the unknown parameters $\tau,\ \beta>0$ and ξ are called the GPD threshold (also known as location), scale and shape parameter, respectively. As discussed in the following, several important challenges arise when fitting the GPD.

2.1 Threshold choice

Theoretically, the shape parameter estimate stabilizes as the GPD threshold becomes large. However, in practical applications, the shape tends to decrease when the threshold increases (see, among others, Aue and Kalkbrener, 2006). Therefore, any predetermined value of the threshold becomes questionable. This finding is consistent with the findings of researchers applying the GPD in other areas, such as in studies of rainfalls and floods. Nevertheless, in many applications the threshold value is predetermined, using some preliminary analysis before fitting the extreme value models. Predetermining the threshold may lead to biased estimates, however, because thresholds that produce "plausible" estimates may be selected. For example, one could discard thresholds yielding shape-parameter estimates above one, because, in this case, the mean of the GPD does not exist and is considered to be unrealistic (Tancredi, Anderson and O'Hagan, 2006). Another strategy is to fix the threshold at a predetermined percentage of upper order statistics, as suggested by DuMouchel (1983). This may work well for large sample sizes, but can be quite unfavorable for small sample sizes (Mittnik and Rachev, 1996; Mittnik, Paolella and Rachev, 1998; and Bermudez, Turkman and Turkman, 2002).

Overall conclusion is that the Hill estimator and its alternatives tend to perform well for extremely large samples, but all suffer from small sample bias. The main source of the bias stems from the selection of the appropriate number of tail observations. If one includes too many observations, the variance of the estimate is reduced at the expense of

a bias in the tail estimate. With too few observations the bias declines, but the variance of the estimate becomes overly large. In addition, Hill–type estimators work well in the case of exact Pareto tail behavior, but they may lead to wrong inference for other distributions (Embrechts et al., 1997).

2.2 Statistical uncertainty of estimates

Simulation-based evidence in the literature suggests that there is substantial statistical uncertainty around parameter as well as capital estimates of operational risk models (Mignola and Ugocioni 2006). Statistical uncertainty around estimates is much more substantial for heavy–tailed severity distributions, such as the GPD and the lognormal, relative to light–tailed severity distributions. Standard measures of statistical uncertainty are the bias and the root mean square error (RMSE), given by

$$b(\hat{v}) = \frac{1}{M}\sum_{i=1}^{M}(\hat{v}_i - v_0) \quad \text{and} \quad RMSE(\hat{v}) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{v}_i - v_0)^2}, \tag{2.2}$$

respectively, where M is the number of simulated samples; $\hat{v}_i$ is the estimate of an unknown quantity v corresponding to sample i, with v_0 being the true value. Mignola and Ugocioni (2006) show that, if the number of data points from the GPD severity distribution is 100 or less, the bias around capital estimates could easily exceed true capital by 40% and the RMSE can easily be twice the true capital.[3] Mignola and Ugocioni (2006) calculate bias and RMSE when the shape parameter is 0.8 or less. Clearly, if the true value of the shape parameter exceeds 0.8, the statistical uncertainty is even higher.

Simple simulation exercises demonstrate that the shape of the GPD has a dominant affect on the capital estimates, when the shape parameter is close to or exceeds one. Even slight deviations of this parameter from its true value result in substantial deviations of implied and true capital. Moreover, capital is asymmetrically sensitive to this parameter. Upward deviations of the shape result in much larger capital variations than downward deviations.[4] Yet, shape is the parameter most difficult to estimate. Therefore, the key to success in fitting GDPs is being able to come up with an accurate estimation method, so that capital estimates remain within certain"safety limits."

[3]According to McNeil and Saladin (1997), reasonable safety limits for the two measures are, respectively, 0.1 and 0.6 times true capital.

[4]This asymmetry makes capital a concave function around the shape's true value, which explains why *average* capital estimates are greater than true capital values.

2.3 Existing estimation methods

Due to its favorable properties, such as asymptotic efficiency, maximum likelihood estimation (MLE) is frequently used to fit the GPD. It is well known, however, that MLE does not reach efficiency, even in samples generated from a GPD as large as 500. Moreover, maximum likelihood estimates of GPD quantiles are highly unreliable in small samples with 50 or less observations, particularly when the shape parameter, ξ, is positive (Hosking and Wallis, 1987).

The EVT literature has proposed several alternatives to MLE. One of them, the method of moments, is generally reliable only for $\xi < 0.2$. For $\xi \geq 0.5$, the second and higher moments of the GPD do not exist, so that the method-of-moment estimates do not exist either. Addressing the difficulties with MLE and the methods of moments, Hosking and Wallis (1987) propose the method of probability-weighted moments. This method differs from the methods of moments by using weighted moments with the weights being specific functions of the GPD's cdf. However, the authors study the properties of this estimator only for $\xi < 0.5$. Rootzén and Tajvidi (1997) show that for heavy tailed data with $\xi \geq 0.5$ the method of probability weighted moments leads to seriously biased parameter estimates. They also find that this method systematically and severely underestimates the quantiles of the GPD.

Castillo and Hadi (1997) propose the so–called elemental percentile method to fit the GPD. They assume that the threshold value, τ, is fixed, so that only two parameters need to be estimated. The method is based on the observation that, for any given set of two empirical quantiles, one can find scale and shape as functions of these quantiles. The method consists of two steps. In the first step, one calculates the set of parameter estimates corresponding to all distinct sets of two quantiles. In the second step, one finds the proposed estimates as the median estimates of the parameters coming from the first step. To avoid the difficulties related to dealing with very large sets of quantile pairs, the authors suggest to work with a smaller (for example, randomly selected) subset of pairs. Since the estimates of the elemental percentile method are found through matching a specific set of quantiles of theoretical and empirical distributions, the method is silent about how to fit the rest of the quantiles. Also, in empirical applications, the authors predetermine the threshold value. Our own extensive study of this method reveals that it performs poorly in shape estimation when its true value is around or above one.

Other viable alternatives to MLE include estimation methods that are based on minimizing a certain distance measure between the empirical and the fitted distribution, such as the quantile distance estimation (QDE), or minimizing the Andersen-Darling statistics. An attractive feature of these methods is that, by shifting the focus to accurately

fitting high–level quantiles, they have the potential of extracting useful information contained in extreme data pertaining to the tail behavior of the severity distribution. Also, unlike methods of moments and probability weighted moments, the QDE method does not require the existence of moments. In addition, the estimates of the QDE method are optimal in the sense that the method minimizes a distance between empirical and theoretical quantiles. Our experiments with the QDE method, when applied to fitting the GPD, show that in small samples (with 50 or less tail events) this method slightly outperforms MLE when it comes to fitting scale, shape and capital by median estimates. Surprisingly, QDE outperforms MLE even when there is model uncertainty, i.e., the model is misspecified.[5] Still, both MLE and QDE turn out to induce substantial uncertainties around capital estimates.

Due to the above mentioned difficulties with conventional estimation methods, a branch of the literature explores Bayesian methods for fitting the GPD. Bayesian methods are capable of incorporating additional information in the estimation process, when the sample information is not sufficient to accurately fit a model. The additional information can come in the form of prior knowledge, such as expert opinion, about unknown model parameters. In operational risk applications, external data and scenario losses also become a source of additional information. Among others, Behrens, Lopes, and Gamerman (2004) propose a Bayesian model with GPD tails, in which the prior distribution of the parameters is obtained from experts through elicitation procedures. Their MCMC method works well for shape parameter values below 0.5. Several papers that fit the GPD to the tail of historical operational losses report that, in practice, the shape parameter can even exceed one (see, among many others, de Fontnouvelle, Rosengren, and Jordan, 2005; and Moscadelli, 2004). Therefore, the method of Behrens, Lopes, and Gamerman (2004) may not be well suited to fit operational risk losses.

2.4 Model uncertainty

Practical applications of extreme value estimation in operational risk are associated with significant challenges due to the nature of loss data—tail events are infrequent but highly sever (cf. Chavez–Demoulin, Embrechts, and Neslehova, 2005). In addition, different severe losses could have occurred due to different causes creating heterogeneous sets of tail events. Under these conditions, the strict assumptions of asymptotic theory are not always satisfied. Such a situation occurs when, for example, the true data–generating process is different from a model being fitted. Following the literature, we refer to this situation as model uncertainty. When model uncertainty exists, the estimation bias in

[5]We provide the details of this investigation in Section 2.6.

small samples could be significant, because the sample size is not sufficient for asymptotic properties to work. The simulation exercise of Section 2.6 shows, in particular, that the effect of model uncertainty on capital estimates can be substantial when sample sizes are small.

2.5 Optimization

Samples with rare large losses typically hamper optimization due to irregular likelihood surfaces characterized by both multiple local optima and flat regions. For example, Dutta and Perry (2006) report that the convergence of their estimates is very sensitive to starting values of the parameters and point to the possibility of the likelihood function exhibiting multiple local optima. They also provide evidence that the poor performance of the GPD model may be the result of non–convergence. In such situations, conventional optimization techniques, such as the Newton method, become unreliable as they are based on differential calculus and solutions of first–order conditions. Therefore, in operational risk applications, a conventional optimization techniques might produce a local optimum in the neighborhood of the starting point, provided it converges at all.

To overcome this challenge, we use simulated annealing (Kirkpatrick, Gelatt and Vecchi, 1983) to find the global optimum. This optimization heuristic with probabilistic acceptance criteria is arguably the most popular among several powerful routines that have appeared in recent years as alternatives to classical optimization techniques. It iteratively suggests slight random modifications to the current solution and, by doing so, gradually moves through the search space. A crucial property of simulated annealing is that not only modifications for the better are accepted, but also for the worse, in order to escape local optima. A probabilistic criterion is used to decide whether to accept or reject a suggested worse move. However, this probability declines over time according to some "cooling schedule," allowing the method to converge. The simulated algorithm that we use is described in Ergashev (2008).

An important characteristics of simulated annealing is that it largely avoids non-convergence and, thus, practically always delivers an answer. However, two caveats need to be kept in mind: first, there is no guarantee that simulated annealing always delivers the best possible result, and, second, simulated annealing requires some tuning to make sure that proper inputs are chosen. In practical applications, one could perform tuning, using simulated data with characteristics that are similar to those of the observed sample. This process could be time consuming, however. The new method proposed here does not require high levels of accuracy to obtain the global optimum to find the proposal densities of the MCMC algorithm via simulated annealing. It is sufficient to

run a few iterations of annealing to approximate a proposal density's mode and the curvature at the mode. We discuss details in Section 4.

2.6 A simulation exercise

To emphasize some of the challenges mentioned above, we conduct a simulation exercise in which we compare the performance of MLE and QDE[6] methods in situations with and without model uncertainty. For the case of absence of model uncertainty, we generate loss samples from a GPD and fit a GPD model. In the second case, when model uncertainty is assumed, the fitted model is still the GPD, but the samples are generated from a mixture of two lognormals. Both cases capture a very realistic situation where, due to a small sample size, the asymptotic conditions are not satisfied. To ease the estimation problem, we assume that the location and the frequency parameters, τ and λ, are known. In other words, we apply the MLE and the QDE to estimate only two parameters, namely, scale and shape.

For the simulation exercise, we let each sample contain 50 losses, on average, collected over $n = 5$ years, i.e., an annual loss frequency of $\lambda_0 = 10$. For the case of no model uncertainty, 1,000 samples of losses are generated from the severity GPD with the following true location, scale and shape parameters: $\tau_0 = 10^5$, $\beta_0 = 10^4$, $\xi_0 = 0.9$. The implied true operational risk capital is about 46×10^6. In the case where we allow for model uncertainly, the true data–generating process is a mixture of two truncated lognormal distributions, and we fit 1,000 data samples from the mixture distribution with the following true parameters: $\mu_{10} = 5$, $\sigma_{10} = 2$, $\mu_{20} = 15$, $\sigma_{20} = 1$. Both lognormal distributions are truncated from below at $\tau_0 = 10^5$ to capture only tail losses. Also, we assume that, on average, 9 out of 10 draws come from the first distribution. For this specification, the implied true capital is about 82×10^6.

When applying the QDE, we optimize the equally weighted quantile distance between the theoretical and empirical quantiles of the logarithm of the losses.[7] The main reason for using the different optimization algorithms is to demonstrate the gain in accuracy (relative to the standard estimation technique) that comes with combining QDE with simulated annealing. Also, it is well known that standard optimization techniques, such as the Newton-Raphson, may lead to estimates that are heavily dependent on starting

[6]The quantile distance is defined as a weighted sum of squares of the distances between corresponding empirical and theoretical quantiles. In this paper, we use equal weights and all observed data points from the GPD as the set of empirical quantiles for the definition of the quantile distance, see (4.3). The main reason for using equal weights and the set of all empirical quantiles is to minimize the possibility of affecting final results by fiddling with those choices.

[7]All calculations were carried out in Matlab. To maximizing the likelihood, we use Matlab's built-in standard maximum likelihood function.

Table 1: Accuracy of capital estimates using MLE and QDE to fit a aeneralized Pareto distribution

Each case summarizes the results of fitting 1,000 samples to the GPD. The samples consist of 50 observations on average. In Case 1, the true data–generating process is the GPD; in Case 2, it a the mixture of two lognormals that are truncated from below at the GPD threshold. All estimates are reported as multiples of the true capital numbers.

Case 1: absence of model uncertainty						
Method	Mean	25% quantile	Median	75% quantile	Bias	RMSE
MLE	7.4	0.2	0.6	2.7	6.4	41
QDE	3.6	0.5	1.0	2.4	2.6	14
Case 2: presence of model uncertainty						
Method	Mean	25% quantile	Median	75% quantile	Bias	RMSE
MLE	485	7.1	40	206	484	1,829
QDE	323	9.8	79	281	322	889

values. To properly account for this challenge, we choose the starting points for scale randomly from the lognormal distribution, with parameters 11 and 1, and the starting points for shape randomly from the lognormal distribution, with parameters 0 and 0.1. The latter distribution is truncated from above at 2. The main reason for setting an upper limit of 2 for the starting point of shape is that, to our best knowledge, reported estimates for this parameter never exceed 2.[8]

Table 1 summarizes the results of the above exercise. It reports the mean and median estimates as well as the bias and the root mean square errors (RMSE) of the capital estimates. The table shows that the QDE is more accurate than MLE—both when model uncertainty is absent or present. QDE seems to perform better because in small samples QDE fits, on average, the right tail of loss samples more accurately than MLE. Nevertheless, both methods produce substantial statistical uncertainty around the capital estimates—the biases and RMSE of the capital estimates exceed by far reasonable safety limits set at 10% and 60% of true capital (McNeil and Saladin, 1997).

While estimating Case–2 samples, we encountered the challenge of being unable to reject samples that result in unrealistically large capital estimates. The capital estimates were sometimes greater than the total asset values of the largest U.S. banks. None of the common tests, such as Anderson–Darling, Kolmogorov–Smirnov, Cramer–von Mises, could detect these samples. The results reported for Case 2 in Table 1 are obtained after

[8]We decided to use the above lognormal distributions for the starting points, because these distributions play the role of the priors in the next section when we perform another simulation exercise to evaluate the performance of our proposed method. This way, we are making sure that the results of these two exercises are comparable. We also tried different starting points, such as the uniform distributions $\mathcal{U}(10^3, 10^6)$ and $\mathcal{U}(0.1, 2)$ for scale and shape, respectively. However, our findings did not change qualitatively.

removing a few (3-4 samples out of 1000 for each method) that led to capital estimates greater than 3×10^{12}. The findings of this simulation exercise suggest that the standards tests are not always useful in assessing the goodness–of–fit of operational risk models and that the incorporation of human judgment may be necessary.

In the remainder of the paper, we develop a Bayesian estimation technique, which allows us to incorporate expert judgment into the estimation framework in a transparent and theoretically sound way.

3 Bayesian approach

We propose a novel Bayesian estimation method that is well suited to fit a heavy–tailed GPD with the shape parameter being close to or exceeding one. The method can reduce the uncertainty around the capital estimates through the incorporation of additional information into the estimation process in form of prior assumptions about the unknown parameters. In our setting the severity distribution consists of the body and a GPD tail. We include the body, which is truncated from above at the GPD threshold, to improve the accuracy of the capital estimates. Our estimation method (described in Section 4) treats the GPD threshold as an unknown parameter that needs to be estimated.

In the remainder of the section we present the model and discuss some important features of our estimation method, including the prior elicitation process.

3.1 The EVT model

Losses falling below the GPD threshold are commonly not modeled, because it is believed that there is little to gain from incorporating the information below the threshold, when estimating high quantiles (Hall, 1975). Although this may be true in principle, in operational risk the inclusion of losses from the body of the distribution adds a positive value to operational risk capital. Therefore, rigorous operational risk modeling requires the inclusion of those losses as well. Based on this notion, we model body losses and assume that there is an unobserved truncation point, $\tau > 0$, which separates tail losses form body losses. In other words, the body consists of small losses that occur frequently, with the maximum amount of losses not exceeding τ, while the tail consists of losses that occur infrequently and exceed τ. We assume that the distribution of the body losses is lognormal,[9] which is truncated from above at τ. Since the tail losses are fitted with a GPD, τ is the obvious GPD threshold. Hence, the probability distribution function

[9]Other distributions, such as the gamma, could also be chosen to fit body losses. However, our experience shows that usually the choice of the body distribution does not play a dominant role in determining the amount of capital as long as the GPD threshold estimate is reasonable.

(pdf) of the severity distribution is

$$f(x|\mu,\sigma,\tau,\beta,\xi) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\Big\{-\frac{(\log x-\mu)^2}{2\sigma^2}\Big\}I_{\{0<x\leq\tau\}} + \frac{1-F_2(\tau|\mu,\sigma)}{\beta}\Big(1+\xi\frac{x-\tau}{\beta}\Big)^{-\frac{1}{\xi}-1}I_{\{x>\tau\}}, \quad (3.1)$$

where $F_2(\cdot|\mu,\sigma)$ is the cdf of the lognormal distribution with the parameters μ and σ; and I_A is the indicator function of event A. Although it is theoretically possible that, in operational risk applications, $-\infty<\tau<+\infty$ and $-\infty<\xi<+\infty$, we assume $\tau>0$ and $\xi>0$.[10]

To define the likelihood of the model, let N be the total number of losses in a sample, and $\mathbf{X}=(X_1,\ldots,X_N)$ the set of all losses. Then, the likelihood is given by

$$p(\mathbf{X}|\mu,\sigma,\tau,\beta,\xi) = \prod_{i=1}^{N}\Bigg\{\frac{1}{\sqrt{2\pi}\sigma X_i}\exp\Big\{-\frac{(\log X_i-\mu)^2}{2\sigma^2}\Big\}I_{\{0<X_i\leq\tau\}} + \frac{1-F_2(\tau|\mu,\sigma)}{\beta}\Big(1+\xi\frac{X_i-\tau}{\beta}\Big)^{-\frac{1}{\xi}-1}I_{\{X_i>\tau\}}\Bigg\}. \quad (3.2)$$

3.2 Prior distributions and their elicitation

Bayesian techniques are well suited to incorporate expert opinion about unknown quantities, such as plausible operational losses or the parameters of a model, into the estimation process. To do so, expert knowledge needs to be formulated as a prior distribution of the quantity of interest. Updated knowledge comes in the form of the posterior distribution, which is equivalent to the prior distribution multiplied by the likelihood of an available data sample. Therefore, the posterior distribution combines the information contained in both the prior and the data sample by treating them as independent pieces of information. Usually, and especially when the sample is informative, it is preferred to impose uninformative priors—or not to impose any prior assumptions at all—to let the data drive the results. However, if the data sample is not informative, prior knowledge, provided that it is informative, will tend to prevail in the updated knowledge.

In practical applications, priors can be obtained through elicitation. Elicitation is the process of translating an expert's opinion about some uncertain quantities into probability distributions that reflect the prior knowledge about those quantities. Typically, experts are unfamiliar with the meaning of probabilities. Even if the expert is, it is not easy to assign probability distributions to uncertain quantities. Therefore, elicitation often involves a facilitator, whose role is to assist the expert in formulating his/her

[10]Since operational risk losses are positive, it is natural to assume that $\tau>0$. If $\xi<0$, generalized Pareto random variables are limited from above and cannot exceed $\tau-\beta\xi$.

opinion about a quantity in a meaningful probabilistic form. Typically, elicitation is an iterative process consisting of: summarizing the expert opinion in probabilistic terms, fitting a probability distribution to the summary, and assessing the adequacy of the elicitation with the expert opinion. If the fit is inadequate, then, the facilitator continues to elicit more summaries from the expert until the fitted probability distribution becomes adequate.

If a prior is symmetric, it is easier to elicit that prior based on the mean and standard deviation. In contrast, the elicitation of a skewed prior is easier through the elicitation of expert opinion via quantiles rather than mean or standard deviation. Humans' ability to estimate simple statistical quantities has been examined in psychological research over several decades. Experiments show that, for symmetric distributions, subjects' estimates of the mode, mean, and median tend to be highly accurate. When samples were drawn from highly skewed distributions, subjects' assessment of the median and mode were still reasonably accurate, while their assessments of the mean were, however, biased toward the median (Beach and Swenson, 1966; and Peterson and Miller, 1964). Therefore, when forming skewed priors, a facilitator should elicit expert knowledge based on quantiles, such as the first quartile, median, third quartile, 99% quantile, etc.

As discussed next, elicitation of expert opinions can generally be done directly or indirectly.

Direct elicitation

In direct elicitation, the facilitator requests an expert's opinion about each unknown model parameter. Consider, for example, elicitation of a prior distribution for σ. It is natural to impose an inverse gamma prior on σ^2, a skewed distribution that takes only positive values. To specify this prior, the facilitator needs to obtain the expert's best estimates of the first quartile, median etc. of the distribution of σ^2. Based on the obtained data, the facilitator finds the parameters of the best fitting inverse gamma prior. The elicitation of μ should be easier, since a natural prior is the normal prior, which is symmetric. Therefore, the facilitator should directly elicit the expert's opinion about the mean value of μ and its standard deviation.

The elicitation of a prior for the shape parameter is challenging, because, contrary to μ and σ, the concept of distributional shape is not well understood. Therefore, one may rely on the simplest form of elicitation, namely, asking the expert to specify a range in which the parameter is believed to lie. If this is all the facilitator can elicit from the expert, then, it is natural to assume a uniform distribution over the range. For example, when eliciting an expert's opinion about the shape parameter, this simple form

of elicitation may be preferable, because the implied capital might be sensitive toward the choice of the prior for this parameter. In some cases, it is possible to choose the range for some parameters based on absolute physical limits.

In this study, we use the uniform prior over [0.1, 2] for the shape parameter. The reason for forcing the shape to be positive is that otherwise the GPD tail is truncated from above at a finite value. If the shape is zero, the GPD becomes the exponential distribution, which is, in general, unsuitable as it is a light–tailed distribution. Instead, to be conservative, we assume that the shape parameter is positive and greater than, say, 0.1. The choice of 2 as the upper limit for the shape is motivated by the fact that the resulting capital (assuming that the other parameters are fixed at the levels specified in Section 2.6) is comparable to the market value of largest U.S. banks, which sets a clear physical limit to any imaginable operational loss of any U.S. bank. In addition, the reported estimates of the shape parameter never exceed 2. Therefore, in our view, the uniform prior over [0.1,2] for shape cannot be viewed as an informative prior.

The elicitation of a prior for the GPD threshold is also challenging, because this parameter does not have any direct, intuitive interpretation. Therefore, obtaining expert opinions from, say, line–of–business managers may not be promising. Perhaps the best way of approaching this challenge would be to perform some preliminary analysis using traditional methods, such as the Hill estimator, and then concentrating the prior around the preliminary estimate. Alternatively, one could use indirect elicitation, which is described next.

Indirect elicitation

It should be noted that expressing prior beliefs directly, in term of a model's unknown parameters, is not an easy task. For example, experts may not be able to fully understand the meanings of each parameter of a model. Usually, experts are more comfortable with thinking in terms of the worst losses that could occur once in, say, 5, 10, 50 years or so. To handle this, Coles and Tawn (1996) and Coles and Powell (1996) introduce the idea of elicitating information in terms experts are more familiar with. In the context of finding the priors of the GPD, this can be done as follows. The facilitator explains to the experts that they need to think in terms of tail losses that occur infrequently. The facilitator determines three quantile levels, $0 < p_i < 1$. For example, these could be 80%, 90%, and 98% quantiles, so that $p_1 = 0.80, p_2 = 0.90$, and $p_3 = 0.98$. Then, the facilitator asks the expert to come up with three tail–loss amounts, L_1, L_2 and L_3, of which the i^{th} amount is the worst tail event in a period of n_i years, i.e.,

$$n_i = \frac{1}{1 - p_i}.$$

The three losses specified, say L_1, L_2, and L_3, are linked to the unknown GPD parameters via

$$p_i = \mathsf{exp}\Big\{ -\frac{1}{\lambda}\big[1 - F_1(L_i|\tau,\beta,\xi)\big]\Big\}, \qquad i = 1,2,3, \tag{3.3}$$

where F_1 is the cdf of the GPD defined by (2.1), and λ is an unknown parameter of the Poisson (annual) frequency of losses. The last formula is derived from the well known relationship between the cumulative probability distribution functions of the severity and the maximum loss (see Ergashev, 2012, for further details). One can find the parameters τ, β and ξ by solving the three equations (3.3). By repeating this exercise many times, and possibly with many experts, the facilitator can form priors on the unknown parameters of the model.

4 MCMC sampling

Because the target densities of the parameters are not standard densities, we sample the parameters of the severity distribution using the Metropolis–Hastings (MH) algorithm. The MH algorithm is a general MCMC method to produce samples from a given target density. The target density of a parameter (or set of parameters) is the full conditional density of that parameter (set) conditioned on the rest of the parameters of a model. Bayes theorem implies that this full conditional density is equivalent to the posterior density. By construction, each sample from the MH algorithm constitutes a Markov chain of dependent draws. The algorithm is based on a proposal density that generates a proposal value and a probability of move that is used to determine whether the proposal value should be taken as the next draw from the target density. As proposal density one chooses a standard density (for example, normal, t-, or gamma densities), from which one can easily obtain random values with the help of a random number generator. Each proposal value is accepted as a draw from the target density with a probability of move. If the proposal value is rejected, the last draw in the chain is retained as the next draw. The probability of turning draws from the proposal density to draws from the target density is controlled by the acceptance rate.

4.1 Blocking

The design of the MCMC algorithm should be such that it produces samples that mix well and, thus, quickly converge to the posterior distribution. Perhaps, the first step in designing an efficient MCMC algorithm is sampling the unknown parameters in properly chosen blocks. When a model contains many parameters, sampling highly–correlated parameters in a separate block, using a multiple block–MCMC algorithm, helps to im-

prove mixing. Below we will use this strategy when designing our MCMC algorithm. Specifically, we separate the parameters of the severity distribution into three blocks. In the first block we only sample shape parameter ξ. The second block consists of the parameters τ and β. The third block contains the parameters that essentially form the body distribution, μ and σ. Extensive experimentation indicated that this blocking scheme was the most efficient one among several alternatives explored.

We sample λ in a separate step by the so–called Gibbs sampling algorithm, using a gamma conjugate prior and the Poisson–gamma update. The Gibbs sampler is a particular variant of the MH algorithm, in which the target density is a standard density. Therefore, the sampling can be performed using standard random number generating functions.

4.2 Proposal densities

The next step in designing an efficient MCMC algorithm is making sure that the proposal densities, from which proposal draws are generated in each block, are chosen carefully. It turns out that the straightforward and popular approach to forming proposal densities, the random–walk Metropolis-Hastings (RWMH) algorithm, where the candidate value is drawn according to a random–walk process, is not the best approach for a number of reasons. For one, it is well known that the RWMH algorithm tends to generate draws that are highly serially correlated. Therefore, the chain moves inefficiently through the support of the target distribution. To compensate for this inefficiency, the MCMC algorithm has to be run for many thousands of draws. In addition, the acceptance rates of the draws generated are very low. In our analysis, acceptance rates never exceeded 20%.

For these reasons, we follow the approach of Chib and Greenberg (1995) to sampling proposal draws, in which candidate values are sampled from a tailored multivariate normal or Student t proposal densities. The mode and the curvature of the tailored proposal density is approximated by the mode and the curvature of the target density, with the curvature being computed as the negative of the Hessian at the mode. This strategy has considerable intuitive and theoretical appeal. The proposal density essentially comes from a second–order Taylor series expansion of the log target density. It has been found to be valuable in many applications (see, for example, Chib and Ergashev, 2009, and references therein).

We use the method of tailoring when forming the proposal densities for all three blocks. For the second and third blocks, we find the mode by maximizing the target density as suggested in Chib and Greenberg (1995). For the first block it turns out,

however, that this method of finding the mode works well only when the number of tail events is large enough to contain sufficient information about the tail behavior. If the number of tail events is 50 or less, we prefer using the QDE method to find the mode.

4.3 Some technical considerations

In all three blocks, the optimization in search for the mode is performed using simulated annealing as both the target density and the quantile–distance statistic exhibit numerous local optima. For this purpose, the simulated annealing algorithm does not have to reach convergence. It is enough to find an approximate mode using only a few iterations. We use the following inputs: the initial "temperature" is 1; the "cooling" schedule is such that the new temperature is 0.9 times the previous temperature; the final temperature is 0.7; the new move is proposed as $\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{old} + \frac{1}{50}\mathbf{e}\nu$, where $\boldsymbol{\theta}$ is the vector of unknown parameters of a particular block, $\mathbf{e}$ is the unit vector of the same size as $\boldsymbol{\theta}$ with 1 in a randomly chosen position, and ν is a standard normal random variable (for further details, see Ergashev, 2008). The output of the simulated annealing process (i.e., the final value of $\boldsymbol{\theta}_{new}$) is chosen to be the mode of the multivariate normal proposal density; and the covariance matrix (i.e., curvature) of the proposal density is calculated as the inverse of the negative Hessian of the target density at the mode.

To avoid constrained optimization, we transform all parameters (except for μ), so that the transformed parameters have the real line as their domain:

$$\theta_1 = \log \xi, \quad t = \log \tau, \quad b = \log \beta, \quad s = \log \sigma. \tag{4.1}$$

To show how the MH steps of our algorithm work, we consider the first block that samples θ_1. The target density in this block is

$$f(\theta_1|\mathbf{X}, \mu, \sigma, \tau, \beta) \;\; \propto \;\; \pi(\xi) \prod_{i=1}^{N} \left\{ C_i + C\Big(1 + \xi \frac{X_i - \tau}{\beta}\Big)^{-\frac{1}{\xi}-1} I_{\{X_i > \tau\}} \right\}, \tag{4.2}$$

where $\xi = \exp(\theta_1)$, $\pi(\xi)$ is the priors of ξ, and

$$C_i = \frac{1}{\sqrt{2\pi}\sigma X_i} \exp\Big\{ - \frac{(\log X_i - \mu)^2}{2\sigma^2} \Big\} I_{\{0 < X_i \le \tau\}}, \; i = 1, \cdots, N, \quad C = \frac{1 - F_1(\tau|\mu, \sigma)}{\beta}.$$

It should be noted that quantities, $C_1, ..., C_N$, and C are fixed from the perspective of maximizing the right–hand side of (4.2) with respect to θ_1. If the average number of exceedances over the threshold is greater than 50, we find the mode of the proposal normal density and compute the variance of the proposal density as the inverse of the negative Hessian at the mode.

As pointed out earlier, when searching for the mode, the target density on the right–

hand side of (4.2) becomes practically useless in situations where a the number of exceedances over threshold is about 50 or less. In this case, the likelihood is practically flat around the true value of θ_1. In other words, the likelihood function does not capture any significant information contained in large losses about the true value of this parameter, because the likelihood assigns small weights to those losses. Based on this observation, we decided to find the mode by minimizing the equally weighted distances between the quantiles of the set of all empirical losses exceeding τ and the corresponding quantiles of the GPD. The square of this quantile distance is given by

$$Q^2(\mathbf{Y}|\tau,\beta,\xi) = \sum_{i=1}^{k} \Big(\log Y_i - \log Q_i \Big)^2, \tag{4.3}$$

where k is the number of exceedances over the threshold, τ; $Y_1 \geq, \ldots, \geq Y_k > \tau$ are the empirical quantiles (i.e., the order statistics) of the losses exceeding the threshold; and Q_i is the theoretical quantile corresponding to the quantile-level of

$$p_i = \frac{N-i+1}{N+1}, \qquad i = 1, \ldots, k.$$

The theoretical quantiles are the solutions to

$$p_i = F_2(\tau|\mu,\sigma) + \big\{1 - F_2(\tau|\mu,\sigma)\big\} F_1(Q_i|\tau,\beta,\xi), \qquad i = 1, ..., k.$$

In (4.3), we use the logarithm of the quantiles instead of the quantile, because they turned out to produce more stable results.

To be more specific about how we sample θ_1 in the first block, let us denote the right–hand side of (4.2), i.e., the target density for sampling this parameter, by $g(\theta_1, \mathbf{X})$. Since θ_1 is unconstrained, we use the following univariate normal proposal density $q(\theta_1) = \mathcal{N}(\theta_1|m, v)$, where m is the mode of the target density in terms of the transformed parameter $\theta_1 = \log \xi$; and v is the curvature at the mode. To make sure that the MCMC algorithm explores the posterior base efficiently, the mode and the curvature are recalculated for each MCMC iteration. Let $\theta_1^{(j)}$ denote the value of θ_1 in the j-th iteration of the MCMC algorithm. To draw the next value of θ_1, we draw a proposal value $\theta_1^* \sim q(\theta_1)$ and accept θ_1^* as the next value, $\theta_1^{(j+1)}$, with the probability given by

$$\min\Big\{1, \frac{g(\theta_1^*, \mathbf{X})\pi(\theta_1^*)}{g(\theta_1^{(j)}, \mathbf{X})\pi(\theta_1^{(j)})} \frac{q(\theta_1^{(j)})}{q(\theta_1^*)}\Big\}.$$

If the proposed value is rejected, we take $\theta_1^{(j)}$ as $\theta_1^{(j+1)}$.

We employ similar MH algorithms when sampling $\boldsymbol{\theta}_2 = (t, b)$ and $\boldsymbol{\theta}_3 = (\mu, s)$ in the remaining two blocks. In each of these blocks we find the mode of the bivariate proposal normal density by approximately maximizing the appropriate target density,

using simulated annealing, and compute the covariance matrix of the proposal density as the inverse of the negative Hessian at the mode.

The choice of a gamma–prior for λ allows us to sample the frequency parameter using the Poisson–gamma update (Ergashev, 2009)

$$\mathcal{G}\Big(\lambda\Big|a_\lambda + N, \frac{b_\lambda}{1 + nb_\lambda}\Big). \tag{4.4}$$

In case of no prior assumption, the update is given by

$$\mathcal{G}\Big(\lambda\Big|N, \frac{1}{n}\Big). \tag{4.5}$$

4.4 EVT fitting when all losses are observed

The algorithm for sampling the parameters of the model from their posterior distribution, when there is no data collection threshold and, hence, all losses are observed, can be summarized of follows:

Algorithm 1:

Step 1 Initialize μ, σ, τ, β, and λ; fix j_0 (burn-in) and J (the MCMC sample size); set $j = 1$.

Step 2 While $j \leq J + j_0$:

- **(a)** Sample θ_1 using the MH algorithm from the target density given by (4.2)
- **(b)** Sample $\boldsymbol{\theta_2} = (t, b)$ using the MH algorithm
- **(c)** Sample $\boldsymbol{\theta_3} = (\mu, s)$ using the MH algorithm
- **(d)** Transform (θ_1, t, b, s) to $(\xi, \tau, \beta, \sigma)$ using (4.1)
- **(e)** Sample λ from (4.4) or (4.5)

Step 3 Increment j to $j + 1$ and go to Step 2

Step 4 Discard the draws from the first j_0 iterations and save the subsequent J draws of $\mu,\ \sigma,\ \tau,\ \beta,\ \xi$, and λ.

4.5 EVT fitting when losses below a threshold unobserved

For cost considerations, among other reasons, most institutions do not collect small losses falling below a certain predetermined data collection threshold (BIS, 2008). Fitting operational risk models is challenging when losses falling below a threshold are not

observed, because there is no readily available and sufficient information about the behavior of unobserved losses and their frequency—especially, when the observed sample is small. As a result, profile likelihoods of the parameters of the body distribution become flat and tend to produce meaningless outcomes with poor or substantially biased capital estimates. For example, when fitting the lognormal distribution to the body, a significantly overestimated mean parameter in combination with a significantly underestimated standard deviation parameter, or vice versa, can be obtained. Making matters worse, the optimization algorithm may not even converge, if the likelihood surface exhibits multiple local optima.

The Bayesian approach helps to reduce the severity of such problems in the sense that strong priors on the body parameters can introduce valuable information that turns a flat or irregular likelihood with multiple local optima into a better–behaving posterior, making estimation much easier. One should be aware of the fact that strong priors concentrated far away from the true value can induce a substantial bias. However, our experience suggests that, as long as those strong priors are not unreasonably off the true values, the resulting bias becomes dwarfed relative to the bias caused by inaccuracies about the shape estimate.

Next, we extend the proposed method to account for the existence of a data-collection threshold. For this purpose, we follow Chib (1992) and treat unobserved losses as being censored. Chib (1992) proposes restoring the censored part of the data at each iteration of the MCMC algorithm by sampling that part conditioned on all other parameters of the model. Unfortunately, and different from the setup in Chib (1992), the number of censored losses itself is unknown and changes with every iteration of the MCMC sampler when the other parameters of the model change. Therefore, an extra step in the MCMC algorithm is required to generate the number of censored losses. To discuss further details of how to deal with the censoring issue, we introduce additional notation. We denote by $\overline{\mathbf{X}}$ the set of observed losses (i.e., losses that exceed the data collection threshold) and by $\overline{N}$ the total number of observed losses. These losses, combined with set of the censored losses, denoted by $\underline{\mathbf{X}}$, form the set of all losses, $\mathbf{X} = (\underline{\mathbf{X}}, \overline{\mathbf{X}})$, consisting, altogether, of $N = \underline{N} + \overline{N}$ (censored and observed) losses.

Given μ, σ, τ, and λ, the number of censored losses, $\underline{N}$, is Poisson with the mean parameter $\lambda n F_1(\tau|\mu, \sigma)$. Also, given $\underline{N}, \mu$, and σ, the censored losses are sampled from the following lognormal distribution, truncated from above at the threshold T,

$$f(\underline{\mathbf{X}}|\mu, \sigma) = \mathcal{LN}_{(-\infty,T)}\Big(\underline{\mathbf{X}}|\mu, \sigma\Big). \tag{4.6}$$

Once we simulate censored losses and add them to the set of observed losses, we are

back to a situation compatible with Algorithm 1; and our algorithm for sampling the parameters of the EVT model, when losses below a data collection thresholds are not observed, is summarized s follows:

Algorithm 2:

Step 1 Initialize μ, σ, τ, β, and λ; fix j_0 (burn-in) and J (the MCMC sample size); set $j = 1$.

Step 2 While $j \leq J + j_0$:

(a) Simulate, $\underline{N}$, the number of censored losses from the Poisson distribution with the parameter $\lambda n F_1(\tau|\mu, \sigma)$ and calculate the total number of losses $N = \underline{N} + \overline{N}$

(b) Simulate a random set of $\underline{N}$ censored losses, $\underline{\mathbf{X}}$, from (4.6)

(c) Repeat Step 2 of Algorithm 1

Step 3 Increment j to $j + 1$ and go to Step 2

Step 4 Discard the draws from the first j_0 iterations and save the subsequent J draws of $\mu,\ \sigma,\ \tau,\ \beta,\ \xi$, and λ.

5 A simulation study

To assess whether the proposed method is capable of producing reliable results, we conduct a simulation study. In this study, we chose specifications that are similar to those of the simulation exercise in Section 2.6, to be able to directly compare the performance of the MCMC relative to that of MLE and QDE. Specifically, we consider again two cases: Case 1 covers a situation where there is no model uncertainty; Case 2 allows for model uncertainty.

For Case 1, we generate losses from the same GPD data generating process with the same parameters as in Section 2.6, namely, $\tau_0 = 10^5,\ \beta_0 = 10^4,\ \xi_0 = 0.9$, with the body losses being generated from the lognormal distribution $\mathcal{LN}(5, 3)$ that is truncated from above at τ_0. With $\rho_0 = 0.9$, only 10% of losses are generated from the body, which means that a sample of $n = 5$ years of losses, with true annual frequency $\lambda_0 = 100$, contains, on average, 500 losses of which 50 are tail losses. The implied true capital amount is about 47×10^6. Once the losses have been generated, to be close to a realistic setting, we assume that there is a data collection threshold of 10,000, and we drop all losses falling below the threshold.

In Case 2, the true data generating process for loss severity is the mixture of three lognormal distributions, namely, $\mathcal{LN}(5,2)$, $\mathcal{LN}(12,0.5)$ and $\mathcal{LN}(15,1)$. The first two distributions resemble the body, but all three contribute to the tail losses. The mixture proportions are 0.9, 0.09 and 0.01. This means that for a sample of size 500, on average, 450 losses come from the first distribution, 45 from the second and 5 from the third lognormal distributions, respectively. The implied true capital amount is about 82×10^6. Again, once the losses have been generated, we drop all losses falling below the data collection threshold of 10,000.

In both of these cases, we generate 1,000 loss samples and fit the EVT model to the samples via Algoritm 2.

5.1 Priors

We consider two sets of informative priors on the parameters of the severity distribution. The first set consists of moderately informative flat priors, while the second set consists of informative priors that are concentrated around specific parameter values. When choosing flat priors, we keep two conflicting goals in mind, namely, making priors as little informative as possible, while still minimizing the possibility of sampling from highly unlikely regions of the parameter space. The informative priors chosen are strongly concentrated around the parameter estimates we have observed in practice. Clearly, although these priors serve the purpose of demonstrating the workings of the proposed procedure, they are by no means applicable to the broad range of cases encountered in practice.

The chosen set of flat priors is given by

$$\pi(\mu) = \mathcal{U}(7,8), \quad \pi(\sigma) = \mathcal{U}(2,3),$$

$$\pi(\tau) = \mathcal{U}(\overline{X}_{min}, \overline{X}_{max}), \quad \pi(\beta) = \mathcal{U}(10^3, 10^6), \quad \pi(\xi) = \mathcal{U}(0.1,2),$$

where $\overline{X}_{min}$ and $\overline{X}_{max}$ are the observed minimum and maximum losses. It should be noted that all these priors are moderately informative in the sense that they are uniformly distributed over a broad range of possible values for the parameters.

The concentrated priors are specified by

$$\pi(\mu) = \mathcal{N}(7, 0.01), \quad \pi(\sigma) = \mathcal{IG}(10^4, 3 \times 10^4),$$

$$\pi(\tau) = \mathcal{LN}(11,1), \quad \pi(\beta) = \mathcal{LN}(11,1), \quad \pi(\xi) = \mathcal{LN}(0, 0.1).$$

We choose both sets of priors solely for the purpose of demonstrating how the method works. In practice, one may choose priors based on a preliminary analysis through

elicitation of expert opinion, a study of external data, or some other useful sources of information, such as scenario analysis.[11]

5.2 MCMC diagnostics

To assess the accuracy, we compute, as in Section 2.6, the bias and RMSE for the capital estimates. The smaller the (absolute) values of these quantities, the better the estimation results. When there is no model uncertainty, we summarize the MCMC output in terms of point estimates of the unknown parameters (together with the standard deviations around the estimates), as they can be compared to the true parameter values. We also report the average values of the acceptance rates and the inefficiency factors (Chib, 2001) for each MCMC block. Generally, the higher the acceptance rate the faster the MCMC chain moves. The inefficiency factor indicates the degree of serial dependence in generated draws: the higher the factor, the stronger the serial dependence. Therefore, the larger the inefficiency factors, the larger the necessary size of the MCMC series to obtain a sufficiently accurate summary of the posterior distribution. The inefficiency factor of a particular parameter is measured via

$$1 + 2\sum_{k=1}^{K}\left(1 - \frac{k}{K}\right)\rho(k), \tag{5.1}$$

where $\rho(k)$ is the lag-k autocorrelation of the MCMC draws of that parameter; and K is some large number—in our case, $K = 1,000$.

Some other useful diagnostic tools include monitoring the evolution of sample quantiles and the autocorrelations of the sample output as the sampling proceeds. Instability, unusual trends, or too much stability in the quantiles estimates over time might require a revision of the MCMC algorithm. Also, slowly decaying autocorrelations may indicate problems with the mixing of the MCM chain.

5.3 Estimation results

Table 2 reports the mean, bias, RMSEs, and selected quantiles of the capital estimates both for Case 1 and Case 2. The results clearly demonstrate that our MCMC approach dramatically reduces the bias and the RMSE of the capital estimates compared to the conventional estimation methods (see Table 1). This holds even in the presence of model uncertainty. The main source of this enormous reduction in uncertainty is that properly chosen priors diminish the possibility of obtaining parameter estimates that come from unreasonable regions of the parameter space. When the fitted model matches the data

[11] Note that we do not impose any prior assumption on λ.

Table 2: Accuracy of capital estimates obtained via MCMC method
All estimates are reported as multiples of the true capital numbers.

Case 1: absence of model uncertainty						
Priors	Mean	25% quantile	Median	75% quantile	Bias	RMSE
Flat	2.8	0.3	0.8	1.3	1.8	9.8
Concentrated	1.6	0.6	1.2	1.8	0.7	1.7
Case 2: presence of model uncertainty						
Priors	Mean	25% quantile	Median	75% quantile	Bias	RMSE
Flat	1.2	0.1	0.3	0.8	0.2	13.7
Concentrated	30.4	17.2	25.6	38.3	29.4	35.6

Table 3: Summary of the MCMC parameter estimates of the extreme value model in Case 1 with informative priors
The acceptance rates (A.R.) are in percentages of the total number of draws including burn-in. The inefficiency factor (I.F.) is the sample average of 1,000 inefficiency factors

Para-meter	True Value	Mean	Standard Deviates	Median	Min	Max	A.R.	I.F.
μ	9	6.9	0.02	6.9	6.8	7.0	63	182
σ	2.0	2.4	0.03	2.4	2.3	2.5	63	153
τ	10^5	$9.5 \cdot 10^4$	$2.1 \cdot 10^3$	$1.0 \cdot 10^5$	$9.1 \cdot 10^4$	$1.3 \cdot 10^5$	68	194
β	10^4	$1.2 \cdot 10^4$	$6.8 \cdot 10^3$	$1.1 \cdot 10^4$	$4.6 \cdot 10^3$	$4.5 \cdot 10^4$	68	25
ξ	0.9	0.97	0.07	0.98	0.66	1.11	72	13
λ	100	278	14	278	232	327	100	11

generating model (i.e., absence of model uncertainty), the concentrated priors lead to a lower statistical uncertainty about the capital estimates. In the presence of model uncertainty, however, the choice of concentrated priors may still lead to substantial biases in capital estimates. The results demonstrate how difficult it is to achieve reasonable safety limits around capital estimates, when fitting heavy–tailed samples with insufficient tail losses. None of the results above led to bias and RMSE values that are less than 0.1 and 0.6, respectively.

The seemingly less biased capital estimates for the flat priors under Case 2 (i.e., presence of model uncertainty) are a result of the specifics on how we determine model uncertainty. Specifically, the data generating process is a mixture of three lognormal models, whereas the fitting model has a GDP tail and a lognormal body. Since the tail of a lognormal distribution decays faster, the estimate of the tail parameter is also lower. However, one also should note that the RMSE under the model uncertainty, 13.7, is much higher than that under the absence of model uncertainty, 9.8.

Table 3 presents the average point estimates of the parameters, when applying the

MCMC algorithm described Section 4 and the set of concentrated priors to fit $M = 1,000$ samples generated from the extreme value model under Case 1. For each sample, the MCMC estimate of an unknown parameter is the mean of $J = 10,000$ MCMC draws of that parameter (after a burn-in of $j_0 = 2,000$). Although the estimates of the body parameters are quite off the true values, the effect of this bias on the capital estimate is not dramatic. More importantly, the estimates of the GPD parameters are quite accurate. The average estimate of the annual frequency, λ, is much higher than its true value. This positive bias is a direct result of a negative bias in the mean estimate of μ. The average number of tail losses is still close to 50.

Table 3 also reports the previously mentioned measures of efficacy of the MCMC algorithm: the acceptance rates and the inefficiency factors. The inefficiency factors for β and ξ are quite low, which is very important. The reason for having a high inefficiency factor for τ is that, once the MCMC algorithm figures out where the separation point (between body and tail) lies, it does not much stray away from that point. Therefore, it is natural to obtain a high inefficiency factor for this parameter. All acceptance rates are quite high and much higher than one would expect with the RWMH algorithm.[12]

6 Conclusion

In this paper, we have introduced a new estimation strategy for the extreme value approach to modeling operational risk. The proposed MCMC method gives rise to substantially improved risk capital estimates as compared to conventional estimation methods. Informative priors lower the probability of producing extreme parameter estimates that lead to unrealistically large capital estimates. As a result, the approach leads to a substantial reduction in bias and root mean square error—two crucial measures of statistical uncertainty about capital estimates. However, if priors are concentrated far from regions where true parameter values lie, the statistical uncertainty around capital estimates can still be substantial. Therefore, preference should to be given to "somewhat" informative flat priors—unless there is compelling evidence for the regions of concentration for (some of) the parameters. Such evidence could for example, come from expert opinions, possibly extracted via elicitation, or incorporating information from external data.

Future work should focus on understanding the limits of reasonably informative priors, developing practical procedures for direct and indirect elicitation, and exploring how well the proposed method fits to real operational risk data.

[12] On average, it took less than five minutes to estimate each sample using a computer with Intel Xeon 3GHz processor and 3GB of RAM.

References

Aue, F., and Kalkbrener, M. (2006), "LDA at work: Deutsche Bank's approach to quantifying operational risk," *The Journal of Operational Risk*, 1, 4, 49–93.

Beach, L.R. and Swenson, R.G. (1966), "Intuitive estimation of means," *Psychomic Science*, 5, 161–162.

Behrens, C.N., Lopez, H.F., and Gamerman, D. (2004), "Bayesian analysis of extreme events with threshold estimation," *Statistical Modelling*, 4, 227–244.

BIS (2008), "Results from the 2008 Loss Data Collection Exercise for Operational Risk," The Bank for International Settlements.

Chavez-Demoulin, Embrechts, and Neslehová (2006), "Quantitative models for operational risk: extremes, dependence and aggregation," *Journal of Banking and Finance*, 30,2635–2658.

Chib, S. (1992), "Bayes regression for the Tobit censored regression model," *Journal of Econometrics*, 51, 79–99.

Chib, S. (2001), "Markov chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, eds. Heckman, J. J. and Leamer, E., Amsterdam: North-Holland, vol. 5, pp. 3569–3649.

Chib, S., and Ergashev, B. (2009), "Analysis of multi-factor affine yield curve models," *Journal of the American Statistical Association*, 104(488), pp. 1324–1337.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

Coles, S.G, and Powell, E.A. (1996), "Bayesian Methods in Extreme Value Modelling—A Review and New Develapments", *International Statistical Review*, 64, 114–136.

Coles, S.G. and Tawn, J. A. (1996), "A Bayesian Analysis of Extreme Rainfall Data," *Applied Statistics*, 45, 463–478.

de Fontnouvelle, P., Rosengren, E., and Jordan, J. (2005), "Implications of Alternative Operational Modeling Techniques," NBER Working Paper No. W11103.

Diebold, F.X., Schuermann, T. and Stroughair, J. (1998), "Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management," in *Advances in Computational Finance*, eds. Refens, A.-P.N., Moody, J.D. and Burgess A.N., Amsterdam: Kluwer Academic Publishers, pp. 3–12.

Du Mouchel, W.H. (1983), "Estimation the stable index α in order to measure tail thickness: a critique," *The Anuals of Statistics,* 11, 1019-1031.

Dutta, K., and Perry, J. (2006), "A tail of tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational risk Capital," Working Paper, The Federal Reserve Bank of Boston.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), "Modeling extremal events for insurance and finance," Berlin, Germany:Springer-Verlag.

Ergashev, B. (2008), "Should risk managers rely on maximum likelihood estimation method while quantifying operational risk?," *Journal of Operational Risk*, 3, 2, 63–86.

Ergashev, B. (2009), "Estimating the lognormal-gamma model of operational risk using the MCMC method," *Journal of Operational Risk*, 4,1.

Ergashev, B. (2012), "A Theoretical Framework for Incorporating Scenarios into the Operational Risk Modeling," *Journal of Finacial Services Research,* 41, 145-161.

Garthwaite, P.H., Kadane, J.B., and O'Hagan, A. (2005), "Elicitation," *Journal of the American Statistical Association*, 100, 680–700.

Hosking, J.R.M., and Wallis, J. R. (1987), "Parameter and quantile estimation for the generalized Pareto distribution,", *Technometrics*, 29, 3, 339–349.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. (1983), "Optimization by simulated annealing,", *Science*, 220, 4598, 671–680.

McNeil, A.J., and Saladin, T. (1997), "The peaks over threshods method for estimating high quantiles of loss distributions," in *Proceedings of 28th International ASTIN Colloquium*, (Cairns, Australia). Casualty Actuarial Society, Arlington, Virginia, pp. 23–43.

Mignola, G., and Ugoccioni, R. (2006), "Sources of uncertainty in modeling operational risk losses," *Journal of Operational Risk*, 1, 2, 33-50.

Mittnik, S., Paolella, M.S., and Rachev, S.T. (1998), "A tail estimator for the index of the stable Paretian distribution," *Communications in Statistics: Theory and Methods*, 27, 1239-1262.

Mittnik, S., and Rachev, S.T. (1996), "Tail estimation of the stable index α," *Applied Mathematics Letters*, 9, 53-56.

Moscadelli, M. (2004), "The modelling of operational risk: experiences with the analysis of the data collected by the Basel Committee. Bank of Italy," Working Paper No 517.

Peterson, C.R. and Miller, A. (1964), "Mode, median, and mean as optimal strategies,", *Journal of Experimental Psychology*, 68, 363–367.

Rootzén, H., and Tajvidi, N. (1997), "Extreme value statistics and wind storm losses: a case study,", *Scandinavian Actuarian Journal*, 1, 70-94.

Ruckdeschel, P. and Horbenko, N. (2013), "Optimally robust estimators in generalized Pareto models,", *Statistics*, 47, 762–791.