Estimating Operational Risk Capital: the Challenges of Truncation, the Hazards of MLE, and the Promise of Robust Statistics*

Submitted for publication in the Journal of Operational Risk, January, 2012; Revised, July, 2012

John ("J.D.") Opdyke§
Principal, Bates White LLC, JD.Opdyke@BatesWhite.com

Alexander Cavallo‡
Vice President, Corporate Operational Risk Group, Northern Trust, alc7@ntrs.com

In operational risk measurement, the estimation of severity distribution parameters is the main driver of capital estimates, yet this remains a non-trivial challenge for many reasons. Maximum likelihood estimation (MLE) does not adequately meet this challenge because of its well-documented non-robustness to modest violations of idealized textbook model assumptions, specifically that the data are independent and identically distributed (i.i.d.), which is clearly violated by operational loss data. Yet even under i.i.d. data, capital estimates based on MLE are biased upwards, sometimes dramatically, due to Jensen's inequality. This overstatement of the true risk profile increases as the heaviness of the severity distribution tail increases, so dealing with data collection thresholds by using truncated distributions, which have thicker tails, increases MLE-related capital bias considerably. Truncation also augments correlation between a distribution's parameters, and this exacerbates the non-robustness of MLE. This paper derives influence functions for MLE for a number of severity distributions, both truncated and not, to analytically demonstrate its non-robustness and its sometimes counterintuitive behavior under truncation. Empirical influence functions are then used to compare MLE against robust alternatives such as the Optimally Bias-Robust Estimator (OBRE) and the Cramér-von Mises (CvM) estimator. The ultimate focus, however, is on the economic and regulatory capital estimates generated by these three estimators. The mean adjusted single-loss approximation (SLA) is used to translate these parameter estimates into Value-at-Risk (VaR) based estimates of regulatory and economic capital. The results show that OBRE estimators are very promising alternatives to MLE for use with actual operational loss event

* In 2011, Northern Trust contracted J.D. Opdyke, then President of DataMineIt, to evaluate the use of the robust statistics framework for severity modeling, operational risk measurement, and capital estimation for Basel II risk quantification. The projects included the derivation of the Influence Functions of MLE estimators of the parameters of several loss distributions and a performance assessment of specific robust estimators, including CvM and generalized median estimators. DataMineIt undertook the development, implementation, and analysis of the OBRE estimator, as well as deriving and coding the Influence Functions for truncated severity distribution parameters for this paper, without further compensation from Northern Trust. The views expressed in this paper are solely the views of the authors and do not necessarily reflect the opinions of Bates White LLC or Northern Trust Corporation.
 All analyses were performed by J.D. Opdyke using SAS®.

§ First author, corresponding author. The first author wishes to thank participants in both the August 2011 American Bankers Association-Operational Risk Modeling Forum and the September 2011 Operational Risk eXchange (ORX) Analytics Forum for useful comments on his initial drafts of this paper, and specifically Bakhodir Ergashev, Senior Financial Economist at the Federal Reserve Bank of Richmond, for serving as a discussant at the ABA Forum. The first author also expresses gratitude to Toyo Johnson and Nicole A.J. Opdyke for their support and thoughtful insights.

‡ The second author gratefully acknowledges the assistance and encouragement of colleagues at Northern Trust including Benjamin Rosenthal, Regina Desler, David Humke, Zhenhua Hu, and Shang Xue, and the support of Giovanna, Allison, Natalie, and Nicholas Cavallo.

data, whether truncated or not, when the ultimate goal is to obtain accurate (non-biased) and robust capital estimates.

1. Introduction

Under the Basel II Accord, financial institutions are required to allocate capital for operational risk, defined as the risk of financial loss due to external events or due to inadequate or failed internal processes, people, or systems, including legal risk, but not reputational or strategic risk. Essentially, operational losses are the many different ways that a financial institution may incur a financial loss in the course of business aside from market, credit, or liquidity related exposure.[1] The Basel II Accord describes three potential methods for calculating capital charges for operational risk, with the expectation that internationally active banks and banks with significant operational risk exposure will use The Advanced Measurement Approach (AMA).[2] This is the most empirically sophisticated and risk sensitive of the three and requires the quantification of operational risk exposure at a very high percentile of the enterprise aggregate annual loss distribution.[3] Using the Value-at-Risk (VaR) risk measure, regulatory capital for operational risk is estimated at the 99.9th percentile (which corresponds to the size of total annual loss that would be exceeded no more frequently than once in 1,000 years).[4]

To comply with these regulatory requirements, banks must estimate the enterprise level aggregate annual loss distribution. Because the percentiles required for the capital calculations are so far beyond what can be empirically determined in-sample from the extremely limited amount of extant historical loss data, capital estimation requires out-of-sample extrapolation using a parametric model.[5] The Loss Distribution Approach (LDA) – the most commonly used

[1] See the Basel Committee on Banking Supervision (BCBS), 2006 report entitled "International convergence of capital measurement and capital standards" for additional information about the Basel II Accord and its specific framework for operational risk, including the definition of operational risk and standardized classification schemes for loss events according to business line (Annex 8) and event type (Annex 9).

[2] The other two approaches in the Basel II framework are the Standardized Approach (TSA) and the Basic Indicator Approach (BIA). See BCBS (2006).

[3] National bank regulators typically require internationally active banks and banks with significant operational risk exposure (generally, the largest banks) to use the AMA. Some institutions benchmark their AMA capital estimates against estimates generated from the simpler and less risk sensitive Basic Indicator Approach or The Standardized Approach.

[4] Economic capital is estimated at an even higher percentile of the distribution (usually between the 99.95th to 99.98th percentiles).

[5] The Basel II framework for operational risk was first formally proposed by the Basel Committee on Banking Supervision in June 1999, with substantial revisions released in January 2001 and April 2003, and was finalized in June 2004. In a May 2001 report on Basel II, Daníelson et al. (2001) argue that operational risk cannot be measured reliably due to the lack of comprehensive operational loss data. At this point in time, few financial institutions were systematically collecting operational loss data on all business lines and all operational risk event types. Although vended database products include information on operational losses as far back as the 1970s, this early data cannot be considered systematic. In these databases, information on large operational losses (tail events) is obtained from public records (such as Algo FIRST from Algorithmics, the "PKM" component of Aon's OpBase data, and OpRisk Global Data from SAS) or from insurance brokering activities (Aon's OpBase data). The data compilation process induces data capture bias (because only losses beyond a specific threshold are recorded) and reporting bias (because

method for estimating this extrapolation – decomposes operational risk exposure into its frequency and severity components (that is, distributions of the number and magnitude of losses, respectively), and the aggregate annual loss distribution is estimated as the convolution of these two distributions.

Implementation of the LDA model requires that the underlying historical data share a common severity distribution.[6] Consequently, the Basel II Accord requires the quantification of operational risk for non-overlapping and homogeneous groups of losses ("units of measure") that share a common risk profile. The enterprise level aggregate annual loss distribution is constructed by combining the individual aggregate loss distributions of all the units of measure. Most institutions break this problem into a number of sequential stages and develop specific empirical models for each stage:

1. <u>Unit of measure definition</u>: Partition or segment the historical loss data into non-overlapping and homogeneous "units of measure" that share the same basic statistical properties and patterns.

2. <u>Frequency estimation</u>: Develop empirical models to estimate the annual frequency of losses in each unit of measure.

3. <u>Severity estimation</u>: Develop empirical models to estimate the severity of losses in each unit of measure.

4. <u>Estimation of aggregate annual loss distributions</u>: The distribution of the annual total value of operational losses is computed for each unit of measure based on the convolution of the estimated frequency and severity distributions of each unit of measure.

5. <u>Top-of-house risk aggregation</u>: By default, the Basel II framework presumes perfect dependence of risk across units of measure. For VaR-based risk measures, this amounts to summing the VaR estimates from each unit of measure. An institution may argue that its risks at the level of its units of measure are less than perfectly dependent, in which case there is a reduction in enterprise level capital from risk diversification. This alternative dependence structure is often modeled via correlation matrices or copula models.

Within the AMA framework, obtaining accurate, precise, and robust estimates of operational risk capital that are reasonably stable over time remains a substantial empirical challenge. Among operational risk practitioners, the severity model, as opposed to the frequency model, is generally thought to pose much greater modeling challenges, and to have a much greater impact in the estimation of economic and regulatory capital.[7] Although a variety of estimation techniques can

---

only losses above some nominal threshold (that varies in real terms) become public knowledge or are claimed against an insurance policy). de Fontnouvelle et al. (2003) discuss these biases in analyzing external loss data.

[6] This distribution may take a number of different forms including a single parametric distribution, a finite mixture distribution, or a spliced distribution.

[7] Frachot et al. (2004) demonstrate that the vast majority of variation in capital estimates is due to the variation of the estimated severity parameters, as opposed to the variation of the estimated frequency parameter(s).

be applied to severity modeling, Maximum Likelihood Estimation (MLE) is most widely used by practitioners and most widely accepted by regulatory authorities. The appeal of MLE for estimating the parameters of the severity distribution is the desirable statistical properties of MLE estimators when the MLE modeling assumptions are satisfied. MLE estimators are accurate (asymptotically unbiased), asymptotically normally distributed, and maximally efficient (precise) when loss data is independent and identically distributed ("i.i.d.").[8]

In this paper, we question the usefulness of MLE for the estimation of parameters of loss severity distributions when the objective is to accurately measure the extreme right tail of the aggregated loss distribution; that is, to provide unbiased and robust estimates of regulatory and economic capital. We demonstrate, both analytically and via simulation, that:

a) MLE is not statistically robust for capital estimation ("robustness" is defined below)

b) MLE-based capital estimates are systematically biased upwards, sometimes dramatically so

c) The problems in both a) and b) are exacerbated when using truncated severity distributions to account for data collection thresholds (truncation is the most common and widely accepted approach to dealing with such thresholds)

d) With analytical derivations of the appropriate Influence Functions, we show how both a) and c) explain clearly and definitively some of the previously unexplained and counterintuitive capital estimates that result from using MLE (e.g. MLE-based capital estimates can be extremely sensitive to very minor data changes for very small loss events, and increase or decrease capital in the opposite direction of what might be anticipated based on business logic or risk management intuition)

The analytical results for the above rely upon the Influence Function, a fundamental statistical tool borrowed from the robust statistics literature. This also is the literature to which we turn in our search for a "better" estimator, that is, one that estimates capital more accurately (with less bias), more robustly, and more efficiently (with smaller mean squared error (MSE)).[9] We examine two robust alternatives: the Cramér-von Mises (CvM) estimator, which we found performs poorly in this setting for capital estimation, and the Optimally Bias-Robust Estimator (OBRE). It appears that OBRE may outperform MLE on two of the above three criteria: as a

[8] An observed sample of data points is independent "when no form of dependence or correlation is identifiable across them" (BCBS 2011, fn. 29). An observed sample of data points is identically distributed (homogeneous) when the data are generated by exactly the same data generating process, such as one that follows a parametric probability density function, or "are of the same or similar nature under the operational risk profile" (BCBS 2011, fn. 29). These textbook conditions are mathematical conveniences that rarely occur with actual, real-world data, let alone "messy" operational risk loss event data.

[9] Mean Squared Error is the average of the squared devations of sample-based estimate values from the true population value of the parameter being estimated: $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\theta - \hat{\theta}\right)^2 = Variance\left(\theta\right) + \left[Bias\left(\theta\right)\right]^2$

If an estimator is unbiased, bias = 0 and MSE = Variance.

more robust capital estimator less adversely affected by loss data that do not comport precisely with idealized textbook assumptions; and one that is less biased in that it can at least partially mitigate the systematic overstatement of capital that occurs with MLE. In terms of efficiency, OBRE, as it stands, appears to be comparable to MLE for capital estimation, but later in the paper we address ways in which it may outperform MLE on this criterion as well.

Some of the questions this paper answers follow below:

- Under idealized data conditions (i.i.d. data), how accurate (unbiased) and efficient are capital estimates based on MLE, CvM, and OBRE estimators?
- Under real-world data conditions, that is, data conditions that deviate at least modestly from idealized textbook statistical assumptions (specifically, data that is independently but *not* identically distributed), how accurate (unbiased) and efficient are capital estimates based on MLE, CvM, and OBRE estimators?
- How are these results affected by the treatment of a data collection threshold with a truncated severity distribution?
- How much does the magnitude of the truncation threshold matter?

Again, our answers to the above rely in large part on analytical derivation of the Influence Functions (IFs) for the MLE estimators of parameters of each severity distribution, with well-specified simulations confirming the analytical results. An important contribution of this paper is to clarify that the proper derivation of the IF must permit correlation among the parameters of the distribution, something that sometimes has been overlooked in the operational risk measurement literature. And a new contribution of this work is the derivation of the IF for the MLE estimators of parameters of ***truncated*** distributions commonly used as severity distributions. These IF derivations definitively explain the large and sometimes counterintuitive effects of truncation on both the distributions of the MLE estimators and those of the MLE-based capital estimates, effects that have perplexed some operational risk measurement practitioners and researchers who have simulated and identified it, but have failed to definitively explain it, such as Cope (2011).[10] The correct IF derivations for non-truncated distributions also definitively explain two other disconcerting properties of MLE-based capital estimates: the extreme variability of estimated capital over certain ranges of parameter and data values, and the sometimes counterintuitive and large impact of new loss events, even when small, on capital estimates. Finally, we identify and document the systematic upward bias of MLE-based capital

[10] Based on extensive simulations, Cope (2011) consistently refers to truncation as causing "distortion" and "excess variation" in the maximum likelihood estimators of the severity distribution(s). In contrast, based on definitive analytics, we derive in this paper the Influence Functions (IFs) of MLE estimators of both truncated and non-truncated severity distributions. The MLE estimators do have different variances, but this is simply because they are parameters for different severity distributions (some of which happen to be truncated). These (asymptotic) variances all are analytically defined as $V = \int IF^2 dF$ (see Hampel et al., 1986) for any and all of the commonly used severity distributions in operational risk measurement, whether truncated or not; no simulations are required to obtain these exact values. The sometimes counterintuitive effects of truncation can be seen in the derived IFs which show, formulaically, how and why the relationships (covariances) between severity distribution parameters change under truncation. This is discussed in detail in Section 5.

estimates.[11] This is another new finding in this setting, but it is consistent with the long-known analytical result referred to as Jensen's inequality (see Jensen, 1906). This upward bias of MLE-based capital estimates often is very large and economically material in this setting, yet it has not been noted in the operational risk literature.

In Section 2, we discuss the empirical challenges of estimating a severity distribution and operational risk capital from historical data. We review the classical estimator of severity distribution parameters, that is, MLE, and highlight its underlying assumptions in Section 3. In Section 4, we discuss the robust statistics framework based on the Influence Function, with a focus on how this provides a widely accepted and well established statistical definition of "robustness" (specifically, "B-robustness"). Here we also analytically derive the Influence Functions of MLE estimators of parameters of some of the commonly used medium- to heavy-tailed loss severity distributions. In Section 5 we derive Influence Functions of MLE estimators of the parameters of truncated severity distributions, and closely examine the effects of truncation, and the size of the truncation threshold, on the estimators. We define in Section 6 two robust estimators (CvM and OBRE) and compare the (empirical) Influence Functions of MLE, CvM, and OBRE estimators for different loss severity distributions, with and without truncation. In Section 7, we use simulation studies to evaluate the performance of MLE, CvM, and OBRE estimators for the capital estimation problem. We run simulations for the textbook case that the data perfectly conforms to idealized modeling assumptions (i.i.d. data), as well as the more realistic case that the data generating process includes some sources of heterogeneity that are not captured by the model. In Section 8 we summarize and discuss implications of the results for future work, and in Section 9 we conclude.

2. Empirical Challenges to Severity Distribution Parameter Estimation

The very nature of operational loss data makes estimating severity parameters challenging.

- Data paucity: As mentioned above, systematic collection of operational loss data by banks is a relatively recent development, and as a result, most institutions have a limited amount of historical loss data. Sample sizes for the critical low frequency-high severity loss events are even smaller.[12]
- Parameter instability: Parameter estimates for heavy-tailed loss distributions typically are quite unstable and can be extremely sensitive to individual data points, both small and large (depending on the estimator used). Extreme sensitivity of parameter estimates and

[11] All (asymptotically) unbiased estimators of the severity parameters, not just MLE, will generate upwardly biased capital estimates in this setting, but the magnitude of this bias, and its materiality, varies notably across estimators based on three criteria discussed in Section 7.3.

[12] Pooling data from multiple financial institutions in the 2002 Loss Data Collection Exercise, Moscadelli (2004) estimates GPD distributions on as few as 42 data points. Chapelle et al. (2008) estimate GPD distributions with sample sizes of only 30 to 50 losses, and other parametric distributions with sample sizes of approximately 200, 700, and 3,000 losses. The smaller sample sizes are in stark contrast to important publications in the literature, both seminal (see Embrechts et al., 1997) and directly related to operational risk VaR estimation (see Embrechts et al., 2003) which make a very strong case, via "Hill Horror Plots" and similar analyses, for the need for sample sizes much larger to even begin to approach stability in parameter estimates.

capital estimates to large losses is well documented, but the comparable sensitivity of parameter estimates to small losses has been largely overlooked.[13] This is true in spite of the apparently counterintuitive, and potentially large, impact of small losses on operational risk capital estimates, which is well documented herein and in Opdyke and Cavallo (2012).

- Heterogeneity: To achieve a reasonable sample size, even the most well-defined unit of measure typically must group together losses with different event characteristics, such as losses arising from different business processes, losses incurred in very different economic or market conditions, losses arising in different geographies, etc.[14] As a result, it is unlikely that the critical MLE assumptions of i.i.d. data are credibly satisfied.
- Left truncation: Most institutions collect information on operational losses above a specific data collection threshold, and a common way to deal with this – assuming that the data follow a truncated distribution – complicates and makes less stable the parameter estimation.[15]
- Data instability or revision: Some institutions include provisions or reserves in the analytical data set: over time, the severity of these losses may be adjusted upward or downward to reflect additional information about the event (litigation-related events of notable durations are common examples).[16] Also, due to the inherent application of judgment in interpreting and applying classification schemes, individual loss events can even be reclassified into other business lines or event types. This is not an uncommon occurrence.
- Outliers: Unambiguous, robustly defined statistical outliers may be present in the data, particularly if external and internal data are pooled together. This would have more dramatic implications for statistical estimators than modest heterogeneity.

From an empirical modeling perspective, these certainly are very challenging data conditions that at the very least lead the applied researcher to question the credibility of the fundamental

[13] Cope (2011) documents the substantial sensitivity of MLE parameter estimates to large losses using a mixture approach to induce misspecification in the right tail. The analysis does not examine the ultimate impact to operational risk capital, nor does it examine misspecification in the left tail. Other than Opdyke and Cavallo (2012), we are not aware of any papers that examine the sensitivity of MLE severity distribution parameter estimates and capital estimates to potential misspecification in the left tail of the distribution, as is done herein.

[14] Heterogeneity of operational loss data has been flagged as a major problem by a number of authors. Daníelson et al. (2001) state "the loss intensity process will be very complicated, depending on numerous economic and business related variables" (p. 13). For example, Cope and Labbi (2008) and Cope (2010) make use of country level characteristics and bank gross income to build location-scale models that define (more) homogeneous units of measure, which, of course, would be more heterogeneous if such factors were ignored in unit-of-measure definitions, as they typically are.

[15] Cope (2011) provides a detailed analysis of the lognormal distribution that demonstrates via simulations how truncation changes the shape of the likelihood surface, creating instability and imprecision in estimated parameters. He does not, however, analytically derive its influence function as is done herein, and this actually mathematically defines the likelihood surface. Ergashev (2008) documents how truncation may induce dependence between the parameters of the frequency and severity distributions, although Chernobai et al. (2007) describe the presumption of independence here as one of the three major assumptions behind the use of truncated severity distributions in practice.

[16] Recent AMA-related guidance states that banks must have a process for updating legal event exposure after it is financially recognized on the general ledger until the final settlement amount is established. See BCBS (2011).

assumption of MLE: that losses within a unit of measure are independent and identically distributed. Lack of knowledge of the "true" data generating process(es),[17] and the undeniable heterogeneity of losses by observable event characteristics both strongly suggest that real world operational loss data is not only not independent and identically distributed, but probably not even close to this idealized standard. The relevant question then becomes, "Does the deviation of actual operational loss data from idealized i.i.d. assumptions adversely affect MLE parameter estimation and capital estimation in a material way? Or as an empirical matter, is any adverse effect relatively minor?" Given actual loss data, and no knowledge of the "true" data generating process(es), this only can be answered indirectly with the converse question: "Given simulated data that we know deviates from i.i.d. conditions only modestly, how severe are the effects on MLE parameter estimates and the corresponding capital estimates?" This is addressed, both analytically and via simulation, later in the paper.

3. Maximum Likelihood Estimation

Maximum Likelihood Estimation is considered a "classical" approach to parameter estimation from a frequentist perspective. In the Basel II framework, MLE is the typical choice for estimating the parameters of severity distributions in operational risk.[18] To maintain its desirable statistical properties (described below), MLE requires the following assumptions:

- A1) Independence: Individual loss severities are statistically independent from one another
- A2) Homogeneity: Loss severities are identically distributed within a unit of measure (perfect homogeneity)
- A3) Correctly Specified Model: The probability model of the severity distribution is correctly specified

Under these restrictive and idealized textbook assumptions, MLE is known to be asymptotically unbiased ("consistent"), asymptotically normal, and asymptotically efficient.[19]

Given an i.i.d. sample of losses $\left(x_1, x_2, \cdots x_n\right)$ and knowledge of the "true" family of the probability density function $f\left(x \mid \theta\right)$ (that is, knowledge of the pdf, but not the parameter values), the MLE parameter estimates are the values of $\hat{\theta}_{MLE}$ that maximize the likelihood function

[17] Frachot and Roncalli (2002) state that "Mixing internal and external severity data is an almost impossible task because no one knows which data generating process external severity data are drawn from" (p. 4). Perfect knowledge of the data generating process is as implausible for internal data as it is for external data.

[18] The recent AMA guidance from the Basel Committee acknowledges the recent application of robust statistics in operational risk, but refers to Maximum Likelihood Estimation and Probability Weighted Moments as "classical" methods. See BCBS (2011) ¶ 205.

[19] The term "efficient" here is used in the absolute sense, indicating an estimator that achieves the Cramér-Rao lower bound – the inverse of the information matrix, or the negative of the expected value of the second-order derivative of the log-likelihood function. This is the minimal variance achievable by an estimator. See Greene (2007) for more details. The term "efficient" also can be used in a relative sense, when one estimator is more efficient – that is, all else equal, achieves a smaller variance (usually under unbiasedness) – than another.

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{1}$$

so

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max} \left[L(\theta \mid x_1, x_2, \cdots x_n) \right] \tag{2}$$

The same numeric result is obtained with greater computational convenience by maximizing the log-likelihood function

$$\hat{l}(\theta \mid x_1, x_2, \cdots x_n) = \ln\left[L(\theta \mid x) \right] = \sum_{i=1}^{n} \ln\left[f(x_i \mid \theta) \right] \tag{3}$$

so

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max} \left[\hat{l}(\theta \mid x_1, x_2, \cdots x_n) \right] \tag{4}$$

MLE is amongst the class of M-Class estimators, so called because they generalize "M"aximum Likelihood Estimation. M-Class estimators include a wide range of statistical models for which the optimal values of the parameters are determined by computing sums of sample quantities. The general form of the estimator (5) is extremely flexible and can accommodate a wide range of objective functions, including the MLE approach and various robust estimators.

$$\hat{\theta}_{M} = \underset{\theta \in \Theta}{\arg\min} \sum_{i=1}^{n} \rho(x_i, \theta) \tag{5}$$

Assuming the regularity conditions commonly assumed when using MLE,[20] all M-Class estimators are asymptotically normal and consistent (asymptotically unbiased),[21] which are very useful properties for statistical inference. Specifically, the MLE estimator is an M-class estimator with $\rho(x,\theta) = -\ln\left[f(x,\theta) \right]$, so

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\min} \sum_{i=1}^{n} \rho(x_i, \theta) = \underset{\theta \in \Theta}{\arg\min} \sum_{i=1}^{n} -\ln f(x_i, \theta) \tag{6}$$

As a straightforward example of the above, we show the derivation of the MLE estimators for the parameters of the LogNormal distribution in Appendix 1.

An important, but to date completely overlooked complication of using MLE to estimate severity parameters for operational risk capital estimation is the fact that even under textbook i.i.d. data conditions, the expected value of estimated capital (which is essentially an estimate of a very high quantile of the severity distribution) will be biased upwards as a consequence of Jensen's

[20] Greene (2007) is one of many statistics textbooks that discuss the standard regularity conditions required for the application of MLE.

[21] The regularity conditions needed for the consistency and asymptotic normality of M-Class estimators are discussed in many textbooks on robust statistics such as Huber and Ronchetti (2009).

inequality.[22] Note that this is true even though the MLE estimates themselves are (asymptotically) unbiased. We discuss this in detail in Sections 6 and 7, showing that this overstatement of capital often can be quite substantial and economically material in this setting. We also show that one robust estimator (OBRE) may be able to mitigate this bias substantially.

Bias in capital estimation aside, an objective assessment of real world operational risk data must acknowledge that each one of the key assumptions required for MLE estimators to retain their desirable statistical properties (consistency, asymptotic efficiency, and asymptotic normality) is unlikely to hold in practice.

- Independence (A1): If the severity of operational losses has both deterministic and stochastic components, then operational losses by definition fail to be independent due to common determinants. For example, systematic differences in loss severity may be explained by event characteristics such as geography, legal system, client segment, time period effects, etc.[23]

- Homogeneity (A2): Because a unit of measure typically pools internal loss events from multiple business processes that undoubtedly have different data generating processes, achieving perfect homogeneity, as required by MLE, is virtually impossible. However, such pooling typically is required to overcome severe data paucity: without a large enough sample of losses, estimation with any level of precision simply becomes impossible. And even for a single, well described business process with a large number of internal loss events, the loss distribution may change over time due to changes in firm size, business strategy, enhancements to risk management practices, the introduction of new products, or any number of other internal or external factors. So even under the most hopeful assumptions, loss data at the unit of measure level cannot achieve perfect homogeneity. Pooling of internal and external loss data for severity modeling further augments heterogeneity. Each institution's operational risk profile is unique and is moderated by its specific characteristics – the specific combination of products and service offerings, technology, policies, internal controls, culture, risk appetite, scale of operation, governance, and other factors. Pooling such data, no matter how it is scaled or adjusted, simply cannot result in homogenous units of measure.

---

[22] In its most general form, Jensen's inequality states that a convex function of the mean of a random variable is less than or equal to the mean of the function after it is applied to the random variable (the opposite is true for concave functions): if β is a random variable and $g()$ is a strictly convex function, then $g(E[\beta]) < E[g(\beta)]$. This applies to quantile estimation of all commonly used severity distributions as follows: the random variable β is the parameter estimates of the severity distribution and the convex function $g()$ is the inverse cumulative distribution function of the severity distribution. For commonly used severity distributions, the cumulative distribution function is strictly concave, and therefore, the inverse cumulative distribution function is strictly convex. See Kennedy (2008) for an intuitive and thorough discussion of Jensen's inequality). The magnitude of this upward bias in MLE-based capital estimates is sometimes very large and is discussed in more detail in Sections 6 and 7.

[23] Cope et al. (2011) find systematic variation in loss severity by region, country characteristics, and certain macroeconomic variables. And Wei (2006) finds a statistically significant relationship between bank size (measured by assets) and loss severity.

- Correctly Specified Model (A3): MLE has desirable asymptotic statistical properties only when the correct form of the loss distribution is "known." In practice, however, operational risk analysts do not have such knowledge and they must rely on goodness-of-fit tests and other model diagnostics to determine not only parameter values, but also which parametric distribution is the "true," or closest to the "true," data generating process. Such tests themselves are statistical estimators which always will be characterized by sampling error, so the best the analyst can hope for is a large enough sample to increase the precision of these estimates and the certainty that they are very close to representing the "true" data generating process. Note, however, that that larger the sample of losses for a given unit of measure, the less homogeneous will be the sample. This is the classic and unavoidable tradeoff between homogeneity and statistical power: satisfying mathematically convenient statistical modeling assumptions of i.i.d. data (homogeneity) versus having a sample size that is large enough to attain the required, or at least reasonable, level of statistical power and precision. MLE, however, must have both to retain its useful statistical properties.

Any violation of the assumptions required for MLE can result in parameter estimates that are no longer asymptotically unbiased and asymptotically efficient (see Greene, 2007). Violations of MLE assumption 2 (identically distributed) and assumption 3 (correctly specified) can be particularly problematic for statistical inference and result in parameter estimates that are unstable, easily influenced by individual data values, and potentially unusable for business decision making. For example, the mean is an MLE estimator of central tendency that can be rendered practically unusable by a single arbitrarily extreme data value, in the sense that such a data point completely dominates the information contained in the rest of the data sample. A "robust" estimator, on the other hand, such as the median, is one that would not be rendered meaningless or unusable by such a data point, or few such data points. Since data heterogeneity is endemic to operational risk loss data, a formal definition of robustness may be very useful for assessing and/or deriving estimators in this setting, and just such a definition can be borrowed from the robust statistics framework using the Influence Function, as shown below.

4. The Robust Statistics Framework and the Influence Function

The theory behind Robust Statistics is very well developed and has been in use for over half a century. Seminal theoretical contributions include Tukey (1960), Huber (1964), and Hampel (1968). Textbooks have institutionalized this sub-field of statistics for the past 30 years; classic textbooks include Huber (1981) and Hampel et al. (1986). More recent textbooks like Huber and Ronchetti (2009) capture the many theoretical advances made possible by the tremendous expansion of computing power in the last 20 years.

Robust statistics have been used widely in many different applications, including the analysis of extreme values arising from both natural phenomena and financial outcomes (a detailed summary table is available from the authors upon request). And many of the distributions relied upon in these cases have been exactly those used as severity distributions in operational risk, so the applied methodology is not new – only its application to operational risk is relatively new. Some recent applications of robust statistics to operational risk severity estimation include

Opdyke and Cavallo (2012), Ruckdeschel and Horbenko (2010), and Horbenko, Ruckdeschel, and Bae (2011). Older publications include Chernobai and Rachev (2006) and Dell'Aquila and Embrechts (2006). However, only Opdyke and Cavallo (2012) have used robust statistics with truncated severity distributions in this setting, both for parameter estimation and to generate capital estimates as is done herein. Before presenting our findings on these fronts, we first discuss the robust statistics framework and one of its most useful analytical tools in this setting: the Influence Function.

4.1. The Robust Statistics Framework

Robust Statistics is a general approach to estimation that explicitly recognizes and accounts for the fact that all statistical models are by necessity idealized and simplified approximations of complex realities. Consequently, one of the main objectives of robust statistics is to bound the influence on parameter estimates of a small to moderate number of data points in the sample which happen to deviate from the assumed statistical model. The justification for this objective is simple pragmatism: actual data samples generated by real-world processes do not exactly follow mathematically convenient textbook assumptions (e.g. all data points are not perfectly i.i.d., and rarely exactly follow parametric distributions). Robust statistics uses estimation techniques that do not "breakdown" under these conditions (i.e. they do not provide unusable, or at least notably biased and inaccurate, parameter estimates). Instead, parameter estimates from robust estimators are "robust" to such violations. The tradeoff for obtaining robustness, however, is a loss of efficiency relative to MLE, that is, a larger mean squared error (MSE), but this is only when the idealized model assumptions are true. When the i.i.d. assumptions are violated, robust statistics can be even *more* efficient than MLE, because as noted above, MLE's desirable statistical properties, including that of maximal efficiency, do not hold under non-i.i.d. data conditions.

The very nature of operational risk loss data and the regulatory requirements to incorporate both internal and external data in some fashion raise serious doubts about the plausibility of the assumption of identically distributed loss data, even within seemingly well-defined units of measure. Below we show that modest deviations from i.i.d. data can render severity modeling and capital estimation with MLE biased and unstable. The Influence Function is a critical tool from the robust statistics framework that allows researchers to analytically determine and describe the sensitivity of parameter estimates, and the capital estimates based on them, to specific deviations from the assumed statistical model.

4.2. The Influence Function

At the core of the robust statistics framework is the pragmatic recognition that the assumed statistical model may not perfectly describe the data generating process(es) of all of the data. The Influence Function is an analytical tool that, for a particular estimator, describes how parameter estimates of the assumed severity distribution are affected if some portion of the data follows another unspecified distribution at a particular severity value x. The Influence Function will have different properties or shapes depending on each of its arguments below:

- Assumed severity distribution: $F\left(y,\theta\right)=F\left(\ \right)$
- Specified estimator (the statistical functional[24]): $T\left(F\left(y,\theta\right)\right)=T\left(F\right)$
- Location of arbitrary deviation: x
- Fraction of data deviating: ε

Given the above, we present the Influence Function below as an analytic formula for assessing this impact of an infinitesimal deviation from the assumed distribution.

$$IF\left(x\,|\,T,F\right)=\lim_{\varepsilon\to 0}\left[\frac{T\left\{\left(1-\varepsilon\right)F+\varepsilon\delta_x\right\}-T\left(F\right)}{\varepsilon}\right]=\lim_{\varepsilon\to 0}\left[\frac{T\left(F_\varepsilon\right)-T\left(F\right)}{\varepsilon}\right] \tag{7}$$

where δ_x is the cumulative distribution function of the Dirac delta function D_x , a probability measure that puts mass 1 at the point x

$$D_x\left(y\right)=\begin{Bmatrix}1 & \text{if } y=x\\ 0 & \text{otherwise}\end{Bmatrix} \quad\text{and}\quad \delta_x\left(y\right)=\begin{Bmatrix}1 & \text{if } y\geq x\\ 0 & \text{otherwise}\end{Bmatrix} \tag{8}$$

and $T\left\{\left(1-\varepsilon\right)F+\varepsilon\delta_x\right\}=T\left(F_\varepsilon\right)$ is simply the estimator evaluated with contamination.[25] So the Influence Function is simply the difference between the values of the estimator with and without contamination, normalized by the amount of contamination, ε. As such the Influence Function (IF) describes the marginal impact that arbitrarily distributed data at the specific value of x has on the estimated values of the model parameters. Stated yet another way, the Influence Function is the functional derivative of the estimator (the statistical functional) with respect to the assumed distribution. An extensive and detailed description of the Influence Function and related concepts is in Hampel et al. (1986).

The generalizability and power of this framework lies in the fact that it allows us to compare the exact asymptotic behavior of alternative estimators when faced with less-than-ideal, non-textbook data, regardless of the nature of the deviating data. By systematically varying the location of the arbitrary deviation (x), the impact of data contamination from any distribution can be assessed, without having to rely on arguably subjective or inconclusive simulations. Simply put, when one needs an answer to the question, "How does the parameter estimate change when the data sample changes at data point = x?" the Influence Function is the definitive, exact, asymptotic answer.[26]

---

[24] A statistical functional is an estimator expressed as a function of the assumed distribution of the model.

[25] The term "statistical contamination" does not indicate a problem with data quality per se, but instead reflects the realistic possibility (probability) that most of the data follows the assumed distribution, but some fraction of the data comes from a different distribution (this portion is called "contaminated"). In the remainder of this paper, we use the more neutral term "arbitrary deviation" synonymously with "statistical contamination."

[26] The conditions required for the existence of the Influence Function are detailed in Hampel et al. (1986), and Huber (1977). The Influence Function is a special case of a Gâteaux derivative and it requires even weaker

When using actual data points and the empirical cumulative distribution function is used as *F*( ), so $F(\ ) = \hat{F}(\ )$, what we have is the Empirical Influence Function (EIF), defined below.

$$EIF\left(x;T,\hat{F}\right) = \lim_{\varepsilon \to 0}\left[\frac{T\left\{\left(1-\varepsilon\right)\hat{F} + \varepsilon\delta_x\right\} - T\left(\hat{F}\right)}{\varepsilon}\right] = \lim_{\varepsilon \to 0}\left[\frac{T\left(\hat{F}_\varepsilon\right) - T\left(\hat{F}\right)}{\varepsilon}\right] \quad (9)$$

This is how the Influence Function often is used in practice, with actual data, and typically $\varepsilon = 1/n$ for evaluating the effect of a single data point, with various values of *x* used to plot the EIF function. Even with relatively small sample sizes, EIF often very nearly identically matches IF when both are graphed simultaneously, making EIF a good practical tool for validating IF derivations or for approximating IF for other reasons (we show examples of this in practice later in the paper).

4.3. B-Robustness

The behavior of the Influence Function over the domain of the severity distribution *F* ( ) is used to define a specific type of statistical robustness, "B-Robustness."[27] If the Influence Function is bounded (meaning, not heading towards ±∞) over the domain of *F* ( ), then the estimator is said to be "B-Robust" for the distribution *F* ( ); otherwise, it is not "B-Robust" for the distribution *F*(). If the Influence Function is unbounded, then an arbitrary deviation can result in meaningless or unusable parameter estimates in practice, as when parameter estimates become arbitrarily large or small (i.e. heading toward ±∞). As shown below, the mean is an example of an estimator of central tendency that is not B-robust: one arbitrarily extreme data point approaching ±∞ causes the mean to approach ±∞, which is practically meaningless and completely masks all other information that might be gleaned about the central tendency of the rest of the data sample.

A comparison of the influence functions of the mean and median estimators is useful for illustrating the concept of B-Robustness. Assume the data follows a standard normal distribution, $F = \Phi$. The statistical functional for the mean is $T(F) = \int y dF(y) = \int y f(y) dy$, so to derive the IF of the mean, we have:

---

conditions for existence than a Gâteaux derivative, as noted in Hampel et al. (1986) and Huber (1977). Since very weak differentiability (smoothness) conditions are satisfied by all commonly used severity distributions, IF is widely applicable in the operational risk setting.

[27] "B-robustness" is so named because the concept of defining robustness by bounding the Influence Function originally was associated with limiting the "B"ias of an estimator: if an estimator's Influence Function is bounded, so too must be its bias.

$$IF(x;T,F) = \lim_{\varepsilon \to 0}\left[\frac{T(F_\varepsilon) - T(F)}{\varepsilon}\right] = \lim_{\varepsilon \to 0}\left[\frac{T\{(1-\varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon}\right] =$$

$$= \lim_{\varepsilon \to 0}\left[\frac{\int yd\{(1-\varepsilon)\Phi + \varepsilon\delta_x\}(y) - \int yd\Phi(y)}{\varepsilon}\right] = \lim_{\varepsilon \to 0}\left[\frac{(1-\varepsilon)\int yd\Phi(y) + \varepsilon\int yd\delta_x(y) - \int yd\Phi(y)}{\varepsilon}\right] =$$

$$= \lim_{\varepsilon \to 0}\left[\frac{\varepsilon x}{\varepsilon}\right] \text{ because } \int ud\Phi(u) = 0 \text{, so}$$

$$IF(x;T,F) = x \qquad (10)$$

From the mathematical derivation above, it is evident that the Influence Function for the mean of a standard normal random variable is unbounded. As the point of arbitrary deviation (x) increases to $+\infty$, so does the Influence Function, and as a result, the mean becomes arbitrarily large and meaningless. Similarly, as the point of deviation decreases to $-\infty$, the Influence Function does as well, and the mean becomes arbitrarily small (but large in absolute value).

Figure 1: Influence Functions of the Mean and Median

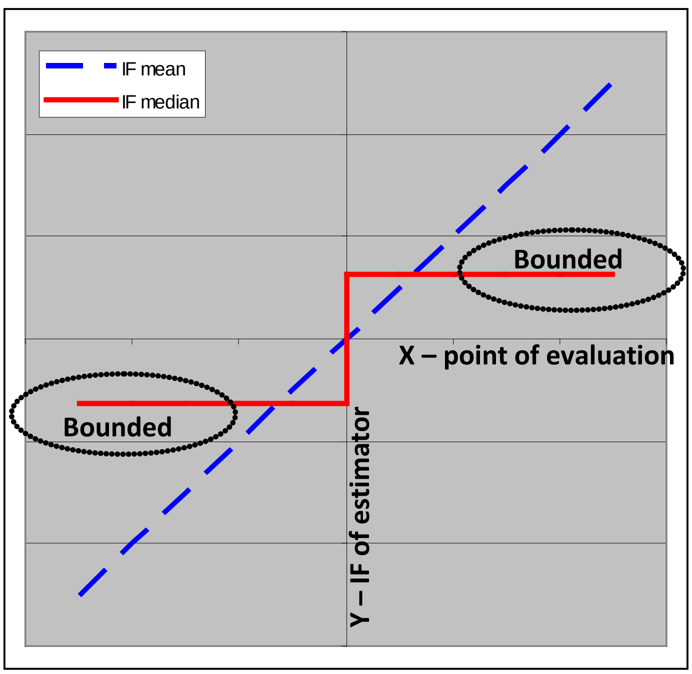


Figure 1 displays the Influence Functions of the mean and the median of a standard normal distribution.[28] Consistent with the mathematical derivation above, the Influence Function for the

[28] See Hampel et al. (1986) for a derivation of the influence function of the median.

mean has a positive slope of 1 and increases without bound in both directions. In contrast, the Influence Function for the median is bounded and never tends toward ±∞. And as expected, the non-robustness and B-robustness of the mean and the median, respectively, hold even when the *F*( ) is not the standard normal distribution, as shown below for the mean

$$IF\left(x;T,F\right)=\lim_{\varepsilon\to 0}\left[\frac{\left(1-\varepsilon\right)\int ydF\left(y\right)+\varepsilon\int yd\delta_x\left(y\right)-\int ydF\left(y\right)}{\varepsilon}\right]=\lim_{\varepsilon\to 0}\left[\frac{\varepsilon x-\varepsilon\mu}{\varepsilon}\right]=x-\mu$$

where μ is the mean of *F*( ), so

$$IF\left(x;T,F\right)=x-\mu \tag{11}$$

As the above begins to show, the Influence Function can be used to assess the performance of a very wide range of estimators regardless of how they are classified (e.g. M-class, L-class, R-class, etc.).[29] Within the context of operational risk, the influence function can be used to:

- compare the robustness of an estimator, and its corresponding capital estimates, across different severity distributions
- compare the robustness of completely different estimators using the same severity distribution
- help identify "deviant" subgroups of losses within the unit of measure for possible alternative treatment
- assess the exact impact of a new loss event of any specified size on capital estimates
- assess the exact impact of revisions of loss severity for existing loss events on capital estimates

A detailed and extensive discussion of such practical, applied uses of the Influence Function in this setting is provided in Opdyke and Cavallo (2012).

In the next section, we present the formula for the Influence Function for M-class estimators generally and MLE specifically. We use this to derive and present the MLE Influence Functions of parameters of specific severity distributions widely used in operational risk.

4.4. The Influence Function for MLE: General Form

As mentioned above, MLE is an M-class estimator (a class of estimators that are generalizations of "M"aximum Likelihood Estimation). M-class estimators generally are defined as any estimator

[29] See Hampel et al. (1986) for detailed definitions and descriptions of broad classes of statistical estimators.

$T_n = T_n\left(X_1, \cdots, X_n\right)$ whose optimized objective function satisfies $\sum_{i=1}^{n} \varphi\left(X_i, T_n\right) = 0$, or equivalently, $\sum_{i=1}^{n} \rho\left(X_i, T_n\right) = \min_{T_n}!$, where $\varphi(x,\theta) = \dfrac{\partial \rho(x,\theta)}{\partial \theta}$ is the derivative of ρ which is defined over $\wp \times \Theta$, the sample space and parameter space, respectively.

For MLE specifically,

$$\rho(x,\theta) = -\ln\left[f\left(x,\theta\right)\right] \tag{12}$$

is the objective function,

$$\varphi_\theta\left(x,\theta\right) = \frac{\partial \rho\left(x,\theta\right)}{\partial \theta} = -\frac{\partial f\left(x,\theta\right)}{\partial \theta} \Big/ f\left(x,\theta\right) \tag{13}$$

is the negative of the score function and

$$\varphi_\theta'\left(x,\theta\right) = \frac{\partial \varphi_\theta\left(x,\theta\right)}{\partial \theta} = \frac{\partial^2 \rho\left(x,\theta\right)}{\partial \theta^2} = \frac{-\dfrac{\partial^2 f\left(x,\theta\right)}{\partial \theta^2} \cdot f\left(x,\theta\right) + \left[\dfrac{\partial f\left(x,\theta\right)}{\partial \theta}\right]^2}{\left[f\left(x,\theta\right)\right]^2} \tag{14}$$

is its derivative (in matrix form, this is the Hessian of the objective function). Hampel et al. (1986) show that Influence Functions for broad classes of estimators conveniently have the same general form, and for M-class estimators, the Influence Function is

$$IF_\theta\left(x \mid \theta, T\right) = \frac{\varphi_\theta\left(x,\theta\right)}{-\displaystyle\int_a^b \varphi_\theta'\left(y,\theta\right) dF\left(y\right)} \tag{15}$$

This is simply the negative of the score function normalized by its variance (the negative of the expected value of the second order derivative), where *a* and *b* define the endpoints of support of the distribution *F*( ). Given the above, the IF for MLE can be written as

$$IF_\theta\left(x \mid \theta, T\right) = \frac{\dfrac{\;\dfrac{\partial f\left(x,\theta\right)}{\partial \theta}\;}{f\left(x,\theta\right)}}{\displaystyle\int_a^b \frac{\left[\dfrac{\partial f\left(y,\theta\right)}{\partial \theta}\right]^2 - \dfrac{\partial^2 f\left(y,\theta\right)}{\partial \theta^2} \cdot f\left(y,\theta\right)}{f\left(y,\theta\right)} dy} \tag{16}$$

However, for distributions with more than one parameter, the MLE Influence Function must take into account the possibility of correlation between parameters. The consequent matrix form of equation (16) above is:

$$IF_{\theta}\left(x;\theta,T\right)=A\left(\theta\right)^{-1}\varphi_{\theta}=\begin{bmatrix} -\int_a^b \frac{\partial \varphi_{\theta_1}}{\partial \theta_1}dF\left(y\right) & -\int_a^b \frac{\partial \varphi_{\theta_1}}{\partial \theta_2}dF\left(y\right) \\ -\int_a^b \frac{\partial \varphi_{\theta_2}}{\partial \theta_1}dF\left(y\right) & -\int_a^b \frac{\partial \varphi_{\theta_2}}{\partial \theta_2}dF\left(y\right) \end{bmatrix}^{-1}\begin{bmatrix}\varphi_{\theta_1}\\ \varphi_{\theta_2}\end{bmatrix} \tag{17}$$

where θ and φ are now vectorized and $A(\theta)$ is simply the Fisher information matrix (see Stefanski and Boos (2002) and Dupuis (1998)). Note that the parameters are correlated when the off-diagonal cross-partial derivative terms are non-zero.

For the one-parameter case, "B-robustness" of the MLE estimator can be determined simply by determining whether the score function is bounded, as long as it is monotonic over the relevant domain (see Huber, 1981). But for the multiple-parameter case, possible correlation of parameters mandates that the cross-partial derivative terms also be taken into account for each parameter. It is important to emphasize that the MLE Influence Function is an analytically determined function given an assumed distribution and parameter values. With it, no simulation is required to assess the behavior of MLE parameter estimators because this is a definitive, analytic result.[30]

4.5. The Influence Function for MLE: LogNormal, LogGamma, and GPD Distributions

With the general form of the MLE Influence Function derived above, for both the single- and multiple-parameter cases, we can now simply "plug-n-play" for any severity distribution. All that needs to be done is the derivation of the first and second order derivatives of each density,

$$\frac{\partial f\left(y;\theta\right)}{\partial\theta_1},\frac{\partial f\left(y;\theta\right)}{\partial\theta_2},\frac{\partial^2 f\left(y;\theta\right)}{\partial\theta_1\partial\theta_2},\frac{\partial^2 f\left(y;\theta\right)}{\partial\theta_1^2},\text{ and }\frac{\partial^2 f\left(y;\theta\right)}{\partial\theta_2^2}$$

which are then plugged into (17) above to obtain the IFs. We provide two examples below for two commonly used severity distributions – the LogNormal and the LogGamma. The former is medium- to heavy-tailed and its parameters are uncorrelated, while the latter is very heavy-tailed and its parameters are correlated.

The usual parameterization of the LogNormal distribution is

[30] Note, however, that some IFs do require numeric integration, as shown in Sections 4.5 and 5.

$$f\left(x;\mu,\sigma\right)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^{2}} \text{ and } F\left(x;\mu,\sigma\right)=\frac{1}{2}\left[1+erf\left(\frac{\ln(x)-\mu}{\sqrt{2\sigma^{2}}}\right)\right]$$

for $0<x<\infty$ and $0<\sigma$ (18)

The first and second order derivatives are shown below:

$$\frac{\partial}{\partial\mu}f\left(x;\mu,\sigma\right)=\left[\frac{\ln(x)-\mu}{\sigma^{2}}\right]f\left(x;\mu,\sigma\right) \quad (19)$$

$$\frac{\partial}{\partial\sigma}f\left(x;\mu,\sigma\right)=\left[\frac{\left(\ln(x)-\mu\right)^{2}}{\sigma^{3}}-\frac{1}{\sigma}\right]f\left(x;\mu,\sigma\right) \quad (20)$$

$$\frac{\partial^{2}}{\partial\mu^{2}}f\left(x;\mu,\sigma\right)=\left[\frac{\left(\ln(x)-\mu\right)^{2}}{\sigma^{4}}-\frac{1}{\sigma^{2}}\right]f\left(x;\mu,\sigma\right) \quad (21)$$

$$\frac{\partial^{2}}{\partial\sigma^{2}}f\left(x;\mu,\sigma\right)=\left(\left[\frac{1}{\sigma^{2}}-\frac{3\left(\ln(x)-\mu\right)^{2}}{\sigma^{4}}\right]+\left[\frac{\left(\ln(x)-\mu\right)^{2}}{\sigma^{3}}-\frac{1}{\sigma}\right]^{2}\right)f\left(x;\mu,\sigma\right) \quad (22)$$

$$\frac{\partial}{\partial\mu\partial\sigma}f\left(x;\mu,\sigma\right)=\left[\frac{\ln(x)-\mu}{\sigma^{2}}\right]\left[\frac{\left(\ln(x)-\mu\right)^{2}}{\sigma^{3}}-\frac{3}{\sigma}\right]f\left(x;\mu,\sigma\right) \quad (23)$$

Now we simply insert the above into formula (17) to obtain the Influence Functions:

$$\varphi_{\theta}=\begin{bmatrix}\varphi_{\mu}\\ \varphi_{\sigma}\end{bmatrix}=\begin{bmatrix}\partial\rho(x,\theta)/\partial\mu\\ \partial\rho(x,\theta)/\partial\sigma\end{bmatrix}=\begin{bmatrix}-\frac{\partial f(x,\theta)}{\partial\mu}\Big/f(x,\theta)\\ -\frac{\partial f(x,\theta)}{\partial\sigma}\Big/f(x,\theta)\end{bmatrix}=\begin{bmatrix}\frac{\mu-\ln(x)}{\sigma^{2}}\\ \frac{1}{\sigma}-\frac{\left(\ln(x)-\mu\right)^{2}}{\sigma^{3}}\end{bmatrix} \quad (24)$$

$$-\int_0^\infty \frac{\partial\varphi_\mu}{\partial\mu}dF(y) = -\int_0^\infty\left[\frac{\ln(y)-\mu}{\sigma^2}\right]^2 - \left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^4}-\frac{1}{\sigma^2}\right]f(y)dy = -\int_0^\infty \frac{1}{\sigma^2}f(y)dy = -\frac{1}{\sigma^2} \quad (25)$$

$$-\int_0^\infty \frac{\partial\varphi_\sigma}{\partial\sigma}dF(y) = -\int_0^\infty\left(\frac{3\left(\ln(y)-\mu\right)^2}{\sigma^4}-\frac{1}{\sigma^2}\right)f(y)dy = \frac{-3}{\sigma^4}\int_0^\infty\left(\ln(y)-\mu\right)^2 f(y)dy + \frac{1}{\sigma^2} = \frac{-3\sigma^2}{\sigma^4}+\frac{1}{\sigma^2} = -\frac{2}{\sigma^2}$$

(26)

$$-\int_0^\infty \frac{\partial\varphi_\mu}{\partial\sigma}dF(y) = -\int_0^\infty \frac{\partial\varphi_\sigma}{\partial\mu}dF(y) = \int_0^\infty\left(\left[\frac{\ln(y)-\mu}{\sigma^2}\right]\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}\right]-\left[\frac{\ln(y)-\mu}{\sigma^2}\right]\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}\right]\right)f(y)dy = 0$$

(27)

Note that the cross-partial derivatives are zero, indicating that the parameters are uncorrelated.

$$IF_\theta(x;\theta,T) = A(\theta)^{-1}\varphi_\theta = \begin{bmatrix} -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_1}dF(y) & -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_2}dF(y) \\ -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_1}dF(y) & -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_2}dF(y)\end{bmatrix}^{-1}\begin{bmatrix}\varphi_{\theta_1}\\ \varphi_{\theta_2}\end{bmatrix} =$$

$$= \begin{bmatrix} -1/\sigma^2 & 0 \\ 0 & -2/\sigma^2\end{bmatrix}^{-1}\begin{bmatrix}\frac{\mu-\ln(x)}{\sigma^2} \\ \frac{1}{\sigma}-\frac{\left(\ln(x)-\mu\right)^2}{\sigma^3}\end{bmatrix} = \begin{bmatrix} -\sigma^2 & 0 \\ 0 & -\sigma^2/2\end{bmatrix}\begin{bmatrix}\frac{\mu-\ln(x)}{\sigma^2} \\ \frac{1}{\sigma}-\frac{\left(\ln(x)-\mu\right)^2}{\sigma^3}\end{bmatrix} = \begin{bmatrix}\ln(x)-\mu \\ \frac{\left(\ln(x)-\mu\right)^2-\sigma^2}{2\sigma}\end{bmatrix}$$

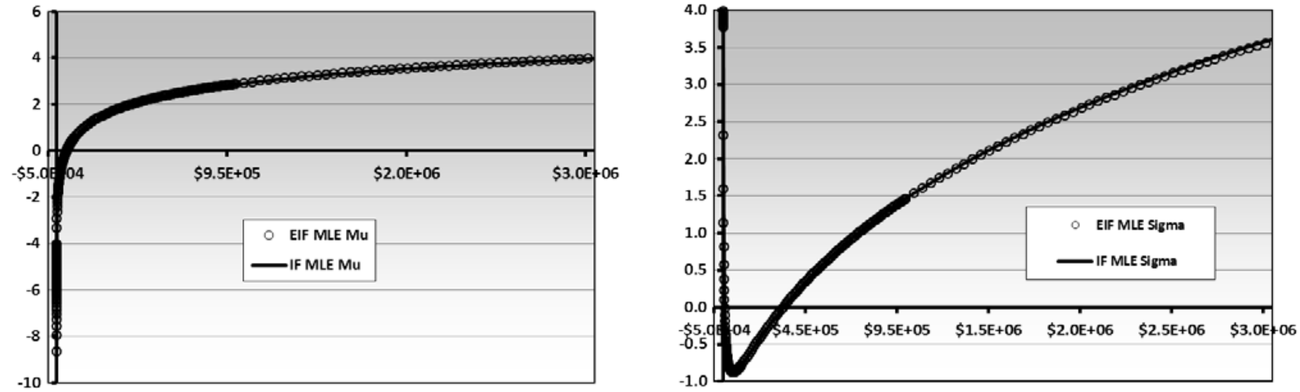
(28)

This is a well known result. And because there is no parameter correlation, we easily can see that the MLE estimators for neither μ nor σ are B-robust: the ln(*x*) terms in both IFs are unbounded, with only constant terms in the rest of each IF, so both diverge toward ±∞. This can be seen graphically in Figure 2 below, which also incorporates the EIFs of each of the parameters (this demonstrates that samples do not have to be large (here, only *n* = 250) for the EIF to converge very well to the IF)[31]:

[31] As mentioned above, graphically comparing EIF to IF is a simple and direct way to validate a correct analytic derivation and implementation of IF, which sometimes can be quite mathematically involved. Recall that EIF is simply the difference between the estimator value with and without “contamination” at data point x, graphed over the domain of all relevant x; the difference is then scaled by the amount of contamination (for only one data point, this scaling is simply dividing by ε = 1/n). With only n = 250 loss events, we see in Figures 2, 3, and 4 below that EIF matches IF virtually identically, providing some assurance that our *analytically* derived IFs are correct.

Figure 2: IF and EIF of MLE Estimators of LogNormal Parameters (μ = 10.95, σ = 1.75)

IF_μ and EIF_μ IF_σ and EIF_σ



Note that the y axes in Figure 2 are scaled for visual clarity: as the IFs analytically show, as $x \to 0^+$, $IF_\mu \to -\infty$ and $IF_\sigma \to +\infty$. This point highlights a very important finding not often noted, even for well-known distributions, when examining the IFs of MLE estimators: due to the logged terms in (many of) the IFs, the MLE parameter values become arbitrarily large or small under contamination by data points with ***very small values***! This is somewhat counterintuitive, and practitioners typically are concerned with parameter sensitivity caused by large-valued contamination, so this is often overlooked, especially in the operational risk setting where focus is on large losses. But for MLE estimators of severity distribution parameters, small valued losses can be just as devastating and destabilizing to the estimation process as large losses, as the IFs clearly show. So aside from not being B-robust, MLE estimators also can be very unstable from quarter to quarter since only a handful of new small losses in this setting, which are far more common than "low frequency, high severity" losses, can dramatically affect and change their value from quarter to quarter. Consequently, a few small losses can dramatically affect estimates of required capital from quarter to quarter. This is a consistent analytical result across many severity distributions, including truncated distributions, and will be discussed further below.

The above process is the same for the LogGamma distribution, whose probability density and cumulative distribution functions are shown below:

$$f(x;a,b) = \frac{b^a \left(\log(x)\right)^{(a-1)}}{\Gamma(a)\, x^{b+1}} \quad \text{and} \quad F(x;a,b) = \int_{\ln(0^+)}^{\ln(x)} \frac{b^a \left(\log(y)\right)^{(a-1)}}{\Gamma(a)\, y^{b+1}} dy$$

for $0 < x < \infty;\ 0 < a;\ 0 < b$ (29)

where $\Gamma(a)$ is the complete gamma function.

The first and second order derivatives (with domain $x > 1$, due to ln(ln(x)) terms) are shown below:[32]

$$\frac{\partial}{\partial a} f\left(x;a,b\right) = \left[\ln\left(b\right) + \ln\left(\ln\left(x\right)\right) - digamma(a)\right] f\left(x;a,b\right) \quad (30)$$

$$\frac{\partial}{\partial b} f\left(x;a,b\right) = \left[\frac{a}{b} - \ln\left(x\right)\right] f\left(x;a,b\right) \quad (31)$$

$$\frac{\partial^2}{\partial a^2} f\left(x;a,b\right) = \left(\left[\ln\left(b\right) + \ln\left(\ln\left(x\right)\right) - digamma(a)\right]^2 - trigamma\left(a\right)\right) \cdot f\left(x;a,b\right) \quad (32)$$

$$\frac{\partial^2}{\partial b^2} f\left(x;a,b\right) = \left[\frac{a\left(a-1\right)}{b^2} - \frac{2a\left(\ln\left(x\right)\right)}{b} + \left(\ln\left(x\right)\right)^2\right] \cdot f\left(x;a,b\right) \quad (33)$$

$$\frac{\partial}{\partial a \partial b} f\left(x;a,b\right) = \left(\frac{1}{b} + \left[\ln\left(b\right) + \ln\left(\ln\left(x\right)\right) - digamma(a)\right] \times \left[\frac{a}{b} - \ln\left(x\right)\right]\right) f\left(x;a,b\right) \quad (34)$$

Again, we simply insert the above into formula (22) to obtain the Influence Functions for both parameters:

$$\varphi_\theta = \begin{bmatrix} \varphi_a \\ \varphi_b \end{bmatrix} = \begin{bmatrix} \partial\rho\left(x,\theta\right)/\partial a \\ \partial\rho\left(x,\theta\right)/\partial b \end{bmatrix} = \begin{bmatrix} -\dfrac{\partial f\left(x,\theta\right)}{\partial a} \Big/ f\left(x,\theta\right) \\ -\dfrac{\partial f\left(x,\theta\right)}{\partial b} \Big/ f\left(x,\theta\right) \end{bmatrix} = \begin{bmatrix} -\left[\ln\left(b\right) + \ln\left(\ln\left(y\right)\right) - digamma(a)\right] \\ -\left[\dfrac{a}{b} - \ln\left(y\right)\right] \end{bmatrix} \quad (35)$$

$$-\int_{1^+}^{\infty} \frac{\partial \varphi_a}{\partial a} dF\left(y\right) = -\int_{1^+}^{\infty} \frac{\partial\left(-\ln\left(b\right) - \ln\left(\ln\left(y\right)\right) + digamma\left(a\right)\right)}{\partial a} f\left(y\right)dy = -\int_{1^+}^{\infty} trigamma(a) f\left(y\right)dy = -trigamma(a) \quad (36)$$

$$-\int_{1^+}^{\infty} \frac{\partial \varphi_b}{\partial b} dF\left(y\right) = -\int_{1^+}^{\infty} \frac{\partial\left(-\frac{a}{b} + \ln\left(y\right)\right)}{\partial b} f\left(y\right)dy = -\int_{1^+}^{\infty} \frac{a}{b^2} f\left(y\right)dy = -\frac{a}{b^2} \quad (37)$$

$$-\int_{1^+}^{\infty} \frac{\partial \varphi_a}{\partial b} dF\left(y\right) = -\int_{1^+}^{\infty} \frac{\partial \varphi_b}{\partial a} dF\left(y\right) = -\int_{1^+}^{\infty} \frac{\partial\left(-\ln\left(b\right) - \ln\left(\ln\left(y\right)\right) + digamma\left(a\right)\right)}{\partial b} f\left(y\right)dy =$$

[32] The digamma and trigamma functions are the first and second order logarithmic derivatives of the complete gamma function: *digamma*(z) = *d*/*dz* ln [Γ(z)] and *trigamma*(z) = d^2/dz^2 ln [Γ(z)].

$$= -\int_{1^+}^{\infty} \frac{\partial\left(-\frac{a}{b} + \ln(y)\right)}{\partial a} f(y)dy = -\int_{1^+}^{\infty} -\frac{1}{b}dy = \frac{1}{b} \tag{38}$$

Note that the cross-partial derivatives are *not* zero, indicating that the parameters are correlated.

$$IF_\theta(x;\theta,T) = A(\theta)^{-1}\varphi_\theta = \begin{bmatrix} -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_1}dF(y) & -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_2}dF(y) \\ -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_1}dF(y) & -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_2}dF(y) \end{bmatrix}^{-1} \begin{bmatrix} \varphi_{\theta_1} \\ \varphi_{\theta_2} \end{bmatrix} =$$

$$= \begin{bmatrix} -trigamma(a) & 1/b \\ 1/b & -a/b^2 \end{bmatrix}^{-1} \begin{bmatrix} -\ln(b) - \ln(\ln(x)) + digamma(a) \\ -\frac{a}{b} + \ln(x) \end{bmatrix} =$$

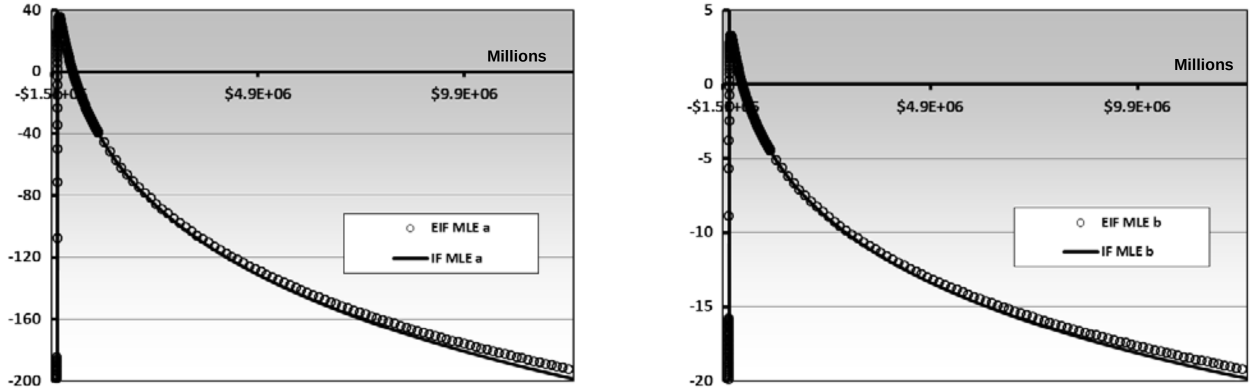
$$= \frac{1}{(a/b^2)\cdot trigamma(a) - 1/b^2} \begin{bmatrix} -a/b^2 & -1/b \\ -1/b & -trigamma(a) \end{bmatrix} \begin{bmatrix} -\ln(b) - \ln(\ln(x)) + digamma(a) \\ -\frac{a}{b} + \ln(x) \end{bmatrix} =$$

$$= \begin{bmatrix} \dfrac{\frac{a}{b^2}\left[\ln(b) + \ln(\ln(x)) - digamma(a)\right] - \frac{1}{b}\left[\ln(x) - \frac{a}{b}\right]}{trigamma(a)\left(\frac{a}{b^2}\right) - \frac{1}{b^2}} \\ \dfrac{\frac{1}{b}\left[\ln(b) + \ln(\ln(x)) - digamma(a)\right] - trigamma(a)\left[\ln(x) - \frac{a}{b}\right]}{trigamma(a)\left(\frac{a}{b^2}\right) - \frac{1}{b^2}} \end{bmatrix} \tag{39}$$

(Note that the limits of integration in (39), *a* and *b*, are not the parameters of the LogGamma – they just coincidentally share the same letters). Even though there is correlation between the parameters of the LogGamma distribution, we still can easily see that the –ln(*x*) terms in both IF numerators dominate the ln(ln(*x*)) terms so both *a* and *b* diverge to –∞ as $x \to +\infty$; and ln(ln(*x*)) – ln(*x*), which inflects at exp(1), becomes a large negative number as $x \to 1^+$, so both *a* and *b* also diverge to –∞ as $x \to 1^+$. So neither MLE estimator is B-robust. This result can be seen graphically in Figure 3 below (note that y axes are cut off for visual clarity: as $x \to 1^+$, IF → –∞ so the LogGamma, too, exhibits extreme sensitivity to small left tail losses like the LogNormal; also note the very different scales of the two y-axes)

Figure 3: IF and EIF of MLE Estimators of LogGamma Parameters (a = 35.5, b = 3.25)

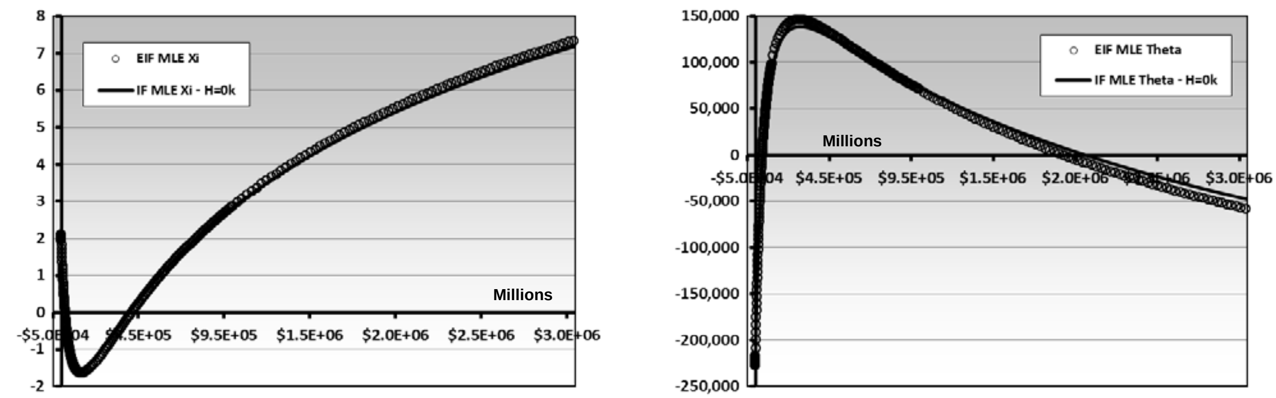
IF_a and EIF_a IF_b and EIF_b



A complete derivation of the Influence Functions for the Generalized Pareto Distribution is presented in Appendix 2, and the corresponding IF graphs are shown below in Figure 4.

Figure 4: IF and EIF of MLE Estimators of GPD Parameters (ξ = 0.875, β = 57,500)

IF_ξ and EIF_ξ IF_β and EIF_β



Working through these examples demonstrates clearly how one uses the IF for different severity distributions in practice, as well as how one can obtain IFs for more complicated severity distributions, such as truncated severity distributions, as shown in the next section.

5. Truncated Severity Distributions

Most banks record losses only above a certain threshold H (typically H = \$5,000, \$10,000, or €20,000 for the case of some external consortium data), so data on smaller losses generally are not available. The most widely accepted and utilized method to account for this when estimating severity distribution parameters is to assume that losses below the threshold follow the same parametric loss distribution, f ( ), as those above it, whereby the severity distribution becomes g ( ), a (left) truncated distribution, with pdf and cdf below.

$$g\left(x,\theta\right)=\frac{f\left(x,\theta\right)}{1-F\left(H,\theta\right)} \quad \text{and} \quad G\left(x;\theta\right)=1-\frac{1-F\left(x;\theta\right)}{1-F\left(H;\theta\right)} \tag{40}$$

Under truncation, the terms of the Influence Function for the MLE estimator now become

$$\rho\left(x;\theta\right)=-\ln\left(g\left(x;\theta\right)\right)=-\ln\left(\frac{f\left(x;\theta\right)}{1-F\left(H;\theta\right)}\right)=-\ln\left(f\left(x;\theta\right)\right)+\ln\left(1-F\left(H;\theta\right)\right) \tag{41}$$

$$\varphi_{\theta}\left(x,H;\theta\right)=\frac{\partial\rho\left(x;\theta\right)}{\partial\theta}=-\frac{\frac{\partial f\left(x;\theta\right)}{\partial\theta}}{f\left(x;\theta\right)}-\frac{\frac{\partial F\left(H;\theta\right)}{\partial\theta}}{1-F\left(H;\theta\right)} \tag{42}$$

and

$$\varphi_{\theta}'\left(x,H;\theta\right)=\frac{\partial\varphi_{\theta}\left(x,H;\theta\right)}{\partial\theta}=\frac{\partial^{2}\rho\left(x;\theta\right)}{\partial\theta^{2}}=$$

$$=\frac{-\frac{\partial^{2}f\left(x;\theta\right)}{\partial\theta^{2}}\cdot f\left(x;\theta\right)+\left[\frac{\partial f\left(x;\theta\right)}{\partial\theta}\right]^{2}}{\left[f\left(x;\theta\right)\right]^{2}}+\frac{-\frac{\partial^{2}F\left(H;\theta\right)}{\partial\theta^{2}}\cdot\left[1-F\left(H;\theta\right)\right]-\left[\frac{\partial F\left(H;\theta\right)}{\partial\theta}\right]^{2}}{\left[1-F\left(H;\theta\right)\right]^{2}} \tag{43}$$

leading to the general form of the Influence Function for one parameter as

$$IF_{\theta}\left(x;\theta,T\right)=$$

$$=\frac{-\frac{\frac{\partial f\left(x;\theta\right)}{\partial\theta}}{f\left(x;\theta\right)}-\frac{\frac{\partial F\left(H;\theta\right)}{\partial\theta}}{1-F\left(H;\theta\right)}}{-\frac{1}{1-F\left(H;\theta\right)}\int_{a}^{b}\frac{\left[\frac{\partial f\left(y;\theta\right)}{\partial\theta}\right]^{2}-\frac{\partial^{2}f\left(y;\theta\right)}{\partial\theta^{2}}\cdot f\left(y;\theta\right)}{f\left(y;\theta\right)}dy+\frac{\left[\frac{\partial F\left(H;\theta\right)}{\partial\theta}\right]^{2}+\frac{\partial^{2}F\left(H;\theta\right)}{\partial\theta^{2}}\cdot\left[1-F\left(H;\theta\right)\right]}{\left[1-F\left(H;\theta\right)\right]^{2}}} \tag{44}$$

where *a* and *b* define the endpoints of support of *G*( ), which are *H* and, typically, ∞, respectively. The structure of the multi-parameter version of the IF does not change except that the differential, of course, corresponds with the cdf of the truncated severity distribution, *G*().

$$IF_{\theta}\left(x;\theta,T\right)=A\left(\theta\right)^{-1}\varphi_{\theta}=\begin{bmatrix}-\int_{a}^{b}\frac{\partial\varphi_{\theta_1}}{\partial\theta_1}dG\left(y\right) & -\int_{a}^{b}\frac{\partial\varphi_{\theta_1}}{\partial\theta_2}dG\left(y\right)\\ -\int_{a}^{b}\frac{\partial\varphi_{\theta_2}}{\partial\theta_1}dG\left(y\right) & -\int_{a}^{b}\frac{\partial\varphi_{\theta_2}}{\partial\theta_2}dG\left(y\right)\end{bmatrix}^{-1}\begin{bmatrix}\varphi_{\theta_1}\\ \varphi_{\theta_2}\end{bmatrix} \tag{45}$$

So comparing (17) and (45), we can see that the numerator of the Influence Function of a truncated distribution simply is a shift of the score function for the non-truncated distribution, and the magnitude of the shift depends only on the threshold H and the parameter values θ, but *not* on the location of the arbitrary deviation x. The denominator of the Influence Function for a truncated distribution differs substantially from that of its non-truncated distribution. The expected value of the Hessian is computed over the truncated domain (H, ∞), multiplied by a truncation constant, and added to an additional constant term based on the cdf evaluated at the threshold and its first and second order derivatives. As is the case for the φ function, the constant terms depend only on the threshold H and the parameter values θ, but *not* on the location of the arbitrary deviation x. These changes in the Fisher information matrix (relative to the non-truncated case) fundamentally alter the correlation structure of the parameters of the distribution, introducing dependence, or magnifying it if already present before truncation.

However, nothing about the "plug-n-play" approach changes under truncation except that derivatives of the cumulative distribution function, in addition to those of the probability density function, must be derived for each severity distribution. So in addition to

$$\frac{\partial f(y;\theta)}{\partial\theta_1},\frac{\partial f(y;\theta)}{\partial\theta_2},\frac{\partial^2 f(y;\theta)}{\partial\theta_1\partial\theta_2},\frac{\partial^2 f(y;\theta)}{\partial\theta_1^2},\text{ and }\frac{\partial^2 f(y;\theta)}{\partial\theta_2^2}$$

we also need to derive

$$\frac{\partial F(H;\theta)}{\partial\theta_1},\frac{\partial F(H;\theta)}{\partial\theta_2},\frac{\partial^2 F(H;\theta)}{\partial\theta_1\partial\theta_2},\frac{\partial^2 F(H;\theta)}{\partial\theta_1^2},\text{ and }\frac{\partial^2 F(H;\theta)}{\partial\theta_2^2}$$

to solve (45) above. To work through an example, we take as the truncated severity distribution of the truncated LogNormal, the cdf partial derivatives for which are shown below. Conveniently, due to Leibniz's Rule, the derivatives can be moved inside the cdf integrals. [33]

$$\frac{\partial F(H;\mu,\sigma)}{\partial\mu}=\frac{\partial}{\partial\mu}\int_0^H f(y;\mu,\sigma)dy=\int_0^H\frac{\partial}{\partial\mu}f(y;\mu,\sigma)dy=\int_0^H\left[\frac{\ln(y)-\mu}{\sigma^2}\right]f(y;\mu,\sigma)dy \tag{46}$$

$$\frac{\partial F(H;\mu,\sigma)}{\partial\sigma}=\frac{\partial}{\partial\sigma}\int_0^H f(y;\mu,\sigma)dy=\int_0^H\frac{\partial}{\partial\sigma}f(y;\mu,\sigma)dy=\int_0^H\left[\frac{(\ln(y)-\mu)^2}{\sigma^3}-\frac{1}{\sigma}\right]f(y;\mu,\sigma)dy \tag{47}$$

[33] See Flanders (1973).

$$\frac{\partial^2 F(H;\mu,\sigma)}{\partial\mu^2}=\frac{\partial^2}{\partial\mu^2}\int_0^H f(y;\mu,\sigma)dy=\int_0^H\frac{\partial^2}{\partial\mu^2}f(y;\mu,\sigma)dy=\int_0^H\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^4}-\frac{1}{\sigma^2}\right]f(y;\mu,\sigma)dy \quad (48)$$

$$\frac{\partial^2 F(H;\mu,\sigma)}{\partial\sigma^2}=\frac{\partial^2}{\partial\sigma^2}\int_0^H f(y;\mu,\sigma)dy=\int_0^H\frac{\partial^2}{\partial\sigma^2}f(y;\mu,\sigma)dy=$$
$$=\int_0^H\left[\frac{1}{\sigma^2}-\frac{3\left(\ln(y)-\mu\right)^2}{\sigma^4}\right]+\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}\right]^2 f(y;\mu,\sigma)dy \quad (49)$$

$$\frac{\partial F(H;\mu,\sigma)}{\partial\mu\partial\sigma}=\frac{\partial}{\partial\mu\partial\sigma}\int_0^H f(y;\mu,\sigma)dy=\int_0^H\frac{\partial}{\partial\mu\partial\sigma}f(y;\mu,\sigma)dy=$$
$$=\int_0^H\left[\frac{\ln(y)-\mu}{\sigma^2}\right]\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{3}{\sigma}\right]f(y;\mu,\sigma)dy \quad (50)$$

In which case, we have for φ

$$\varphi_\theta=\begin{bmatrix}\varphi_\mu\\ \varphi_\sigma\end{bmatrix}=\begin{bmatrix}\partial\rho(x,\theta)/\partial\mu\\ \partial\rho(x,\theta)/\partial\sigma\end{bmatrix}=\begin{bmatrix}-\frac{\partial f(x,\theta)}{\partial\mu}\Big/f(x,\theta)\\ -\frac{\partial f(x,\theta)}{\partial\sigma}\Big/f(x,\theta)\end{bmatrix}=\begin{bmatrix}-\left[\frac{\ln(x)-\mu}{\sigma^2}\right]-\frac{\int_0^H\left[\frac{\ln(y)-\mu}{\sigma^2}\right]f(y;\mu,\sigma)dy}{1-F(H;\mu,\sigma)}\\ -\left[\frac{\left(\ln(x)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}\right]-\frac{\int_0^H\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}\right]f(y;\mu,\sigma)dy}{1-F(H;\mu,\sigma)}\end{bmatrix} \quad (51)$$

and the cells of the Fisher Information $A(\theta)$ are

$$-\int_H^\infty\frac{\partial\varphi_\mu}{\partial\mu}dG(y)=-\frac{1}{\sigma^2}+\frac{\left[\int_0^H\frac{\ln(y)-\mu}{\sigma^2}f(y)dy\right]^2+\int_0^H\frac{\left(\ln(y)-\mu\right)^2}{\sigma^4}-\frac{1}{\sigma^2}f(y)dy\cdot\left[1-F(H;\mu,\sigma)\right]}{\left[1-F(H;\mu,\sigma)\right]^2} \quad (52)$$

$$-\int_H^{\infty}\frac{\partial\varphi_\sigma}{\partial\sigma}dG(y)=-\frac{1}{\left[1-F(H;\mu,\sigma)\right]}\cdot\int_H^{\infty}\frac{3\left(\ln(y)-\mu\right)^2}{\sigma^4}f(y)dy+\frac{1}{\sigma^2}+\frac{\left[\int_0^H\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}f(y)dy\right]^2}{\left[1-F(H;\mu,\sigma)\right]^2}+$$

$$+\frac{\int_0^H\left[\frac{1}{\sigma^2}-\frac{3\left(\ln(y)-\mu\right)^2}{\sigma^4}\right]+\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}\right]^2 f(y)dy\cdot\left[1-F(H;\mu,\sigma)\right]}{\left[1-F(H;\mu,\sigma)\right]^2} \quad (53)$$

$$-\int_H^{\infty}\frac{\partial\varphi_\mu}{\partial\sigma}dG(y)=-\int_0^{\infty}\frac{\partial\varphi_\sigma}{\partial\mu}dF(y)=-\frac{1}{\left[1-F(H;\mu,\sigma)\right]}\cdot\int_H^{\infty}\frac{-2\left(\ln(y)-\mu\right)}{\sigma^3}f(y)dy+$$

$$+\frac{\left[\int_0^H\frac{\ln(y)-\mu}{\sigma^2}f(y)dy\right]\times\left[\int_0^H\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}f(y)dy\right]}{\left[1-F(H;\mu,\sigma)\right]^2}$$

$$+\frac{\left(\int_0^H\frac{-2\left(\ln(y)-\mu\right)}{\sigma^3}f(y)dy+\int_0^H\left[\frac{\ln(y)-\mu}{\sigma^2}\right]\cdot\left[\frac{\left(\ln(y)-\mu\right)^2}{\sigma^3}-\frac{1}{\sigma}\right]f(y)dy\right)\cdot\left[1-F(H;\mu,\sigma)\right]}{\left[1-F(H;\mu,\sigma)\right]^2} \quad (54)$$

And just as with non-truncated distributions, (51)-(54) are inserted into (45) to obtain the IF.

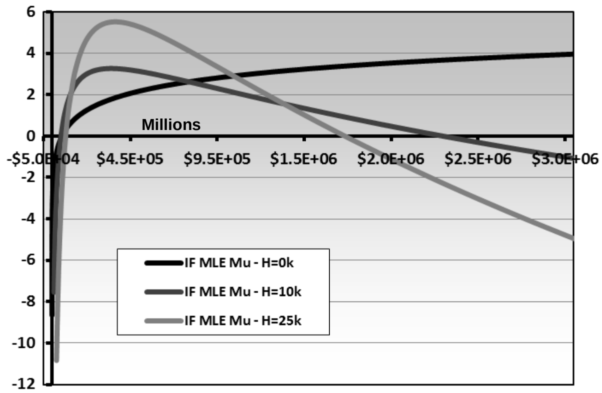
Obviously, the integration for the truncated severity distribution above requires that IF be solved numerically. While inspection of both φ and the graphs of IF below (Figure 5) leaves little doubt that the MLE estimators of the truncated LogNormal are not B-robust, we leave a formal proof for another paper. Instead, we focus on a more notable and unexpected, if not counterintuitive finding. Note, first, that the off-diagonal cells of $A(\theta)^{-1}$ are non-zero, indicating that truncation has induced parameter correlation where none existed before for the non-truncated LogNormal. Secondly, the entire shape *and direction* of IF_μ has changed. Consequently, the direction of the relationship between the two parameters has changed: previously large values of x would increase both μ and σ, but now large values of x move μ and σ in opposite directions. Arbitrarily large contamination now leads to arbitrarily small μ, but arbitrarily large σ. So, quite counterintuitively, larger and larger loss values will actually *decrease* the value of what many consider the location parameter, μ, of the truncated LogNormal, while its scale parameter, σ, increases (this is easily confirmed in straightforward simulations).[34] Finally, as seen in Figure 5, the *size* of the threshold has a dramatic effect on the above findings, with larger thresholds magnifying the parameter correlation, as well as the sensitivity of the parameters to arbitrary deviations from the presumed model, as seen in the more extreme values in the IFs for large x.

[34] In fact, exp($\boldsymbol{\mu}$) is the scale parameter of the LogNormal.

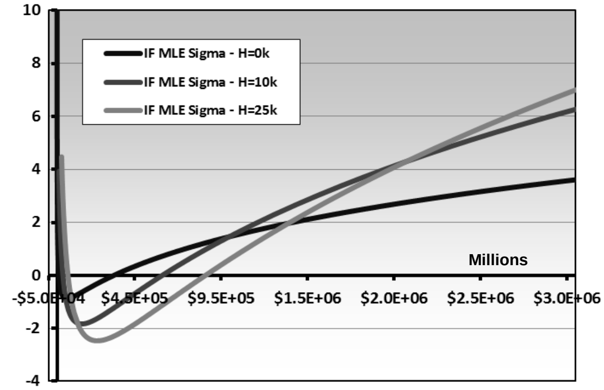
Figure 5:

IF of MLE Estimators of Truncated LogNormal Parameters (μ = 10.95, σ = 1.75) by Threshold, *H*

IF$_\mu$



IF$_\sigma$

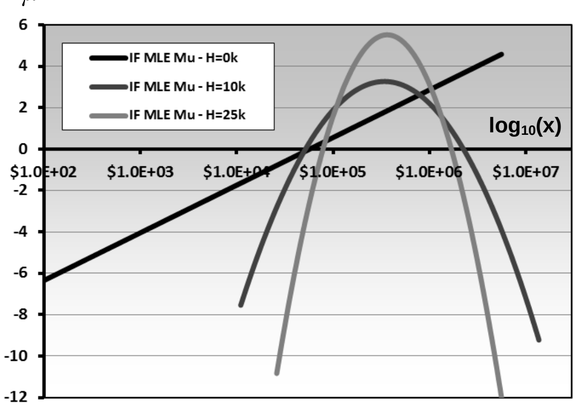


Some of these findings regarding the sensitivity of MLE estimators of truncated distributions, and some of their counterintuitive results, have been reported previously based on authors' simulations (for example, see Cope (2011). In addition, some practitioners have encountered MLE parameter estimates for the truncated LogNormal distribution with $\mu < 0$ and very large values of σ. Given the numerical challenges of solving the estimating equations for MLE under truncation, such parameter estimates are often attributed to convergence problems for the optimization routine. This explanation in fact is incorrect because here, for the first time, are analytically derived Influence Functions that conclusively show how exactly this counterintuitive parameter behavior can be caused by just a few large data values (see IF$_\mu$ in Figure 5 under *H* = \$25,000). Such examples show just how useful the IF can be as it explains, definitively, the cause of the simulation results that have perplexed other authors. They also show how relatively unstable and sensitive MLE estimators can be to what many would consider to be a fairly innocuous change in the statistical model – accounting for collection thresholds by assuming truncation, and truncation at relatively low thresholds at that. This effect is even more apparent in Figure 5a below, which shows Figure 5 in Log Scale.

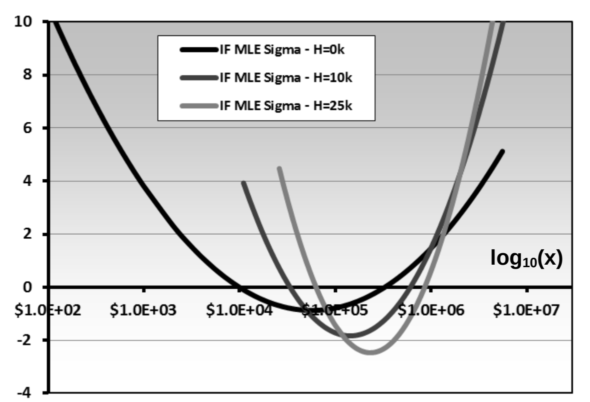
Figure 5a:

IF of MLE Estimators of Truncated LogNormal Parameters (μ = 10.95, σ = 1.75) by *H*

IF$_\mu$



IF$_\sigma$



And just as with the non-truncated severity distributions, the EIFs for the Truncated LogNormal virtually exactly match the corresponding IFs, as expected.[35]

The above derivations are completed for the truncated LogGamma distribution and the truncated Generalized Pareto Distribution in Appendix 3, and the corresponding graphs are shown below in Figures 6, 6a, 7, and 7a, respectively. We note from these Figures that for the LogNormal and LogGamma severity distributions, truncation mitigates the extreme sensitivity of MLE estimators to small, left-tail losses, but it does not come close to eliminating it altogether. Table 1 shows this most clearly: new losses do not have to be extremely close to the data collection thresholds (points of truncation) to generate strongly disproportionate effects on capital requirements.

For example, even when a single, new loss is $4,000 away from a $25,000 threshold, under a Truncated LogNormal (with $\mu \approx 10.95$, $\sigma \approx 1.75$, H = $25,000) regulatory capital increases by over $2million, and economic capital increases by $3.6million. The same numbers for a Truncated LogGamma ($a \approx 35.5$, $b \approx 3.25$) are even more dramatic: $19.2million and $57.4million, respectively. And for the GPD ($\xi \approx 0.875$, $\beta \approx 57{,}500$), truncation does not have the same mitigating effect, but this is arguably because the extreme sensitivity to small losses was not as pronounced to begin with. Whether truncated or not, GPD sensitivity to small losses is grossly disproportionate to the size of the loss: regulatory capital increases range from about $20 to $40million when a new loss is within a few thousand dollars of the lower threshold, and economic capital ranges from just under $70million to over $130million. Can an increase in economic capital of $133.7million ever be justified by a single, new loss of $27,000? According to the proper use of MLE under LDA, it can.

These dramatic results in Table 1 demonstrate why, when using MLE estimators for most if not all widely used severity distributions, large swings in capital estimates can result from just a handful of new losses near the collection threshold in any given quarter. The bank experiences no new "low frequency, high severity" losses, and its risk profile typically will not change in material ways from quarter to quarter, but capital requirements easily can change by 20%, 30%, or 40% due to these small losses. This is a completely counter-intuitive, but mathematically inescapable result attributable entirely to MLE estimators. It is not the much ballyhooed "low frequency, high severity" losses that are the main cause of quarter-to-quarter instability in MLE-based capital estimates, but rather, the much more common small losses closer to the data collection thresholds. This is just one of the drawbacks of MLE estimators that robust statistics are designed to avoid.

[35] EIFs for all truncated severity distributions very closely track the corresponding IFs and are available from the authors upon request.

Figure 6:
IF of MLE Estimators of Truncated LogGamma Parameters (a = 35.5, b = 3.25) by Threshold, H

IF_a IF_b

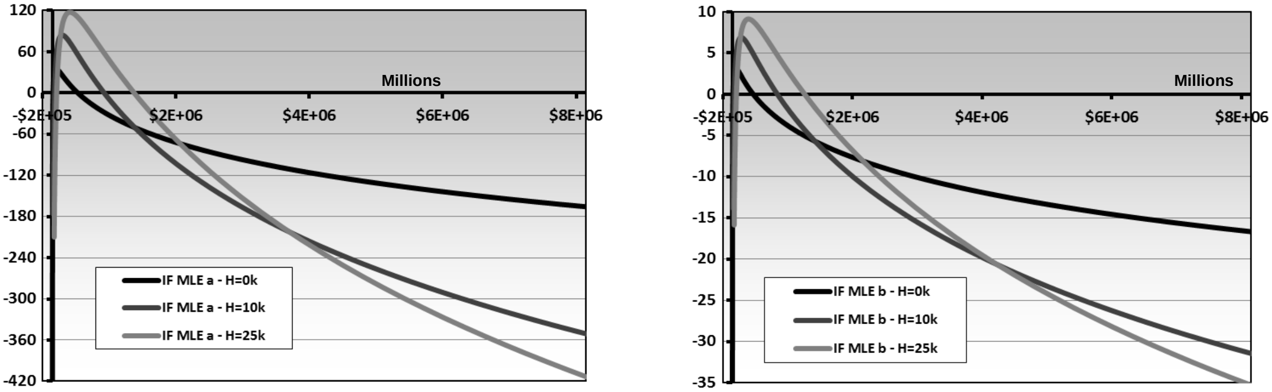


Figure 6a (Figure 6 Log Scale):
IF of MLE Estimators of Truncated LogGamma Parameters (a = 35.5, b = 3.25) by Threshold, H

IF_a IF_b

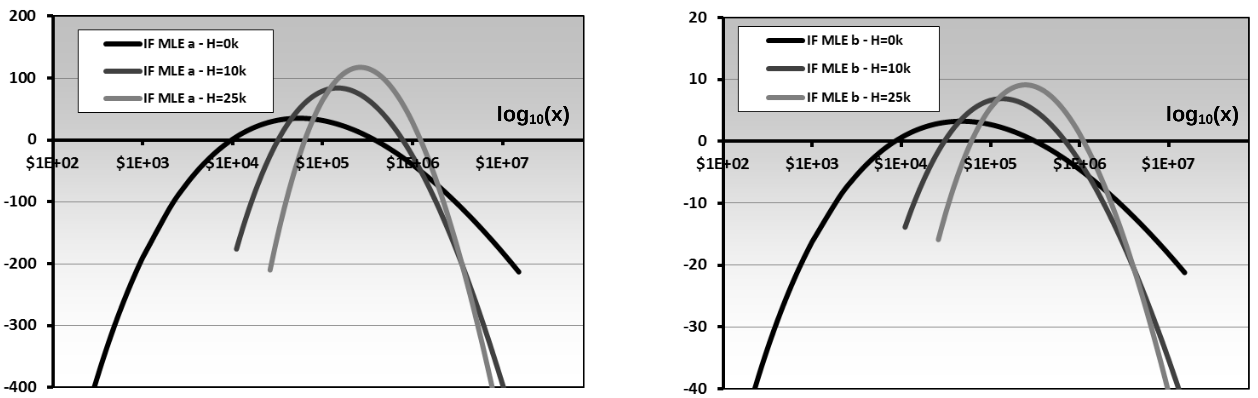


Figure 7:
IF of MLE Estimators of Truncated GPD Parameters (ξ = 0.875, β = 57,500) by Threshold, H

IF_ξ IF_β

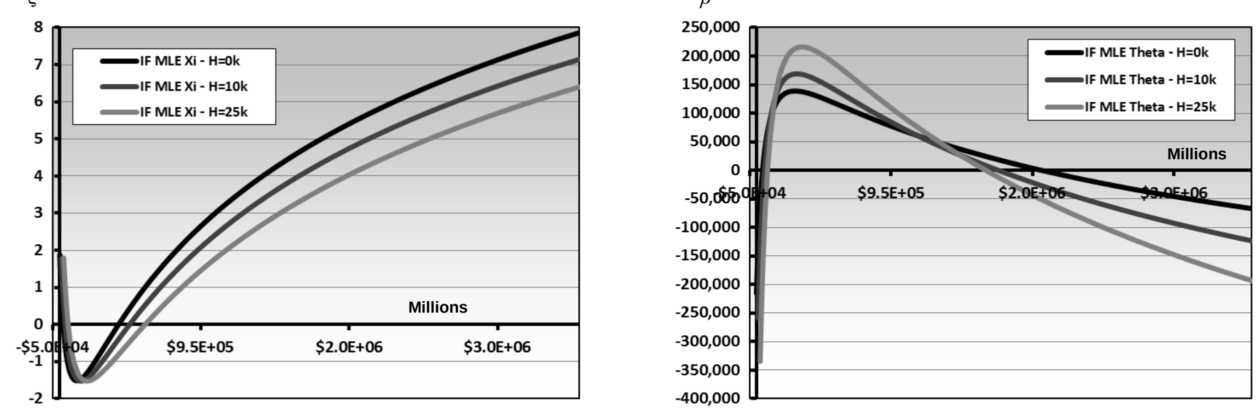


Figure 7a (Figure 7 Log Scale):
IF of MLE Estimators of Truncated GPD Parameters (ξ = 0.875, β = 57,500) by Threshold, H

IF_{ξ} IF_{β}

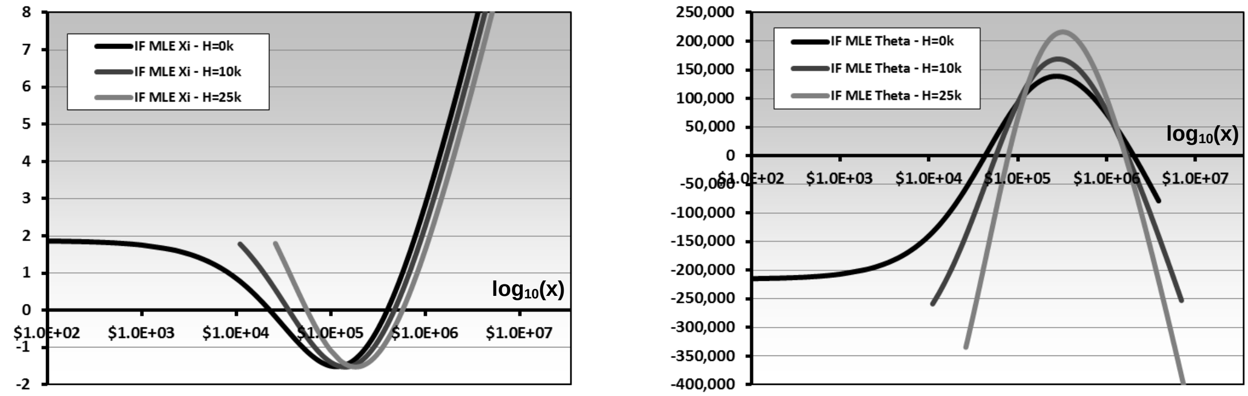


Table 1: Change in MLE-based Capital Due to One New Loss by Severity Distribution (n = 250, λ = 25)

Severity Dist.	Threshold H	Parameter Names	Parm 1	Parm 2	Change in Capital ($mill) H + $10 loss RC	H + $10 loss EC	H + $2k loss RC	H + $2k loss EC	H + $4k loss RC	H + $4k loss EC
LogN	$0	μ, σ	10.953	1.749	$19.0	$33.3	$1.3	$2.4	$0.4	$0.8
LogN	$10,000	μ, σ	10.954	1.750	$2.6	$4.2	$2.0	$3.6	$1.5	$2.4
LogN	$25,000	μ, σ	10.917	1.749	$2.6	$4.8	$2.3	$4.2	$2.0	$3.6
LogG	$0	α, β	35.484	3.252	$590.9	$1,469.8	$14.1	$34.1	$3.6	$9.2
LogG	$10,000	α, β	35.513	3.263	$24.1	$62.2	$18.0	$43.1	$13.2	$33.5
LogG	$25,000	α, β	35.410	3.252	$26.4	$67.0	$22.8	$57.4	$19.2	$57.4
GPD	$0	ξ, β	0.8713	57,584	$27.9	$92.2	$24.0	$79.5	$20.4	$67.8
GPD	$10,000	ξ, β	0.8825	57,484	$31.2	$95.6	$26.4	$95.5	$24.0	$76.4
GPD	$25,000	ξ, β	0.8798	57,340	$38.4	$133.8	$36.0	$133.7	$31.2	$95.5

6. Robust Statistics

Given the empirical challenges of operational risk loss data discussed above (data paucity, non-i.i.d. data/heterogeneity, data instability/revision, unambiguously extreme outliers, etc.), we turn to robust statistics not only to establish the non-robustness of MLE estimators in this setting, but also to find better estimators that can surmount these challenges. Two candidates widely used for decades are the Cramér-von Mises (CvM) estimator, a minimum distance estimator (MDE), and the Optimally Bias-Robust Estimator (OBRE). Both are B-robust, and as M-class estimators, both are asymptotically normal and consistent (asymptotically unbiased). The price they pay for their robustness is less efficiency than MLE, but this is only true under perfectly i.i.d. data; under real world data conditions, they can, in fact, be more efficient than MLE. We define both below.

6.1. The CvM Estimator

CvM Defined:

The Cramér-von Mises (CvM) estimator yields the parameter values of the assumed distribution that minimize its distance from the empirical distribution (see Duchesne et al., 1997). Given the CvM statistic in its common form,

$$W^2(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left[F_n(x_i) - F_\theta(x_i) \right]^2 \tag{55}$$

where F_n is the empirical distribution and F_θ is the assumed distribution, the minimum CvM estimator is that value $\hat{\theta}$ of θ, for the given sample, that minimizes:

$$\hat{\theta}_{MCVME} = \arg\min_{\theta} \left\{ n \cdot \int \left[F_n(x) - F_\theta(x) \right]^2 dF_\theta(x) \right\} \tag{56}$$

CvM Computed:

The computational formula typically used to calculate the minimum CvM is:

$$W^2(\theta) = \frac{1}{12n} \cdot \sum_{s=1}^{n} \left[F_\theta\left(x_{(s)}\right) - \frac{2s-1}{2n} \right]^2 \tag{57}$$

where $x_{(s)}$ is the s'th ordered value of *x* (see Duchesne et al., 1997).

Other MDE's are very similar conceptually, and typically differ in how they weight the data points. For example, Anderson-Darling, another MDE, weights data points in the tail more than does CvM. CvM is a very commonly used MDE, hence its inclusion in this study.[36]

6.2. The OBRE Estimator

OBRE Defined:

The Optimally Bias-Robust Estimator (OBRE) is provided for a given sample of data as the value $\hat{\theta}$ of θ that solves (63):

[36] See Parr and Schucany (1980) for a discussion of minimum distance estimation using a variety of empirical distribution function tests.

$$\sum_{i=1}^{n} \varphi_c^{A,a}\left(x_i;\theta\right) = 0 \tag{58}$$

where $\varphi_c^{A,a}\left(x;\theta\right) = A\left(\theta\right)\cdot\left[s\left(x;\theta\right) - a\left(\theta\right)\right]\cdot W_c\left(x;\theta\right)$ (59)

and

$$W_c\left(x;\theta\right) = \min\left\{1;\frac{c}{\left\|A\left(\theta\right)\cdot\left[s\left(x;\theta\right)-a\left(\theta\right)\right]\right\|}\right\} \tag{60}$$

and $A(\theta)$ is a dim(θ) x dim(θ) matrix and $a(\theta)$ is a dim(θ) vector determined by the equations

$$E\left[\varphi_c^{A,a}\left(x;\theta\right)\cdot\varphi_c^{A,a}\left(x;\theta\right)^T\right] = I \tag{61}$$

and

$$E\left[\varphi_c^{A,a}\left(x;\theta\right)\right] = 0 \tag{62}$$

$A(\theta)$ and $a(\theta)$ can be viewed as Lagrange multipliers which ensure the constraints of B-robustness and Fisher consistency, respectively.[37] $s\left(x;\theta\right)$ is simply the score function, $s\left(x;\theta\right) = \left[\partial f\left(x;\theta\right)/\partial\theta\right] / f\left(x;\theta\right)$, so OBRE is defined in terms of a weighted standardized scores function, where $W_c\left(x;\theta\right)$ are the weights. c is a tuning parameter chosen by the researcher, $\sqrt{\dim\left(\theta\right)} \leq c \leq \infty$, regulating from very robust to MLE, respectively.

So OBRE actually is just a constrained MLE. And because OBRE is based on the score function, it maintains efficiency as close as possible to MLE, subject to its constraints. Hence, its name: "Optimal" B-Robust Estimator. OBRE achieves maximal possible efficiency, defined for the "standardized" OBRE as the smallest possible trace of its asymptotic covariance matrix, for a given level or robustness as determined by *c*. Many researchers choose *c* to achieve 95% efficiency relative to MLE, but this actual value for *c* depends on the model being implemented and the data sample under analysis.

Several versions of the OBRE exist with minor variations on exactly how they bound the IF. The OBRE defined above is the so-called "standardized" OBRE "which has proved to be

[37] A statistical function $T(\ )$ is Fisher consistent if, when applied to the "true" probability distribution, F_θ, it recovers the "true" population parameter values, θ: $T(F_\theta) = \theta$; or equivalently, if the limiting value of the sample analog of the statistical functional, $T(F_n)$ converges to the "true" population values, θ. See Huber and Ronchetti (2009) for further details.

numerically more stable" according to Alaiz and Victoria-Feser (1996). The "standardized" OBRE is used in this study.

OBRE Computed:

To compute OBRE, (63) must be solved under constraints (66) and (67), for a given tuning parameter value *c*, via a convergence algorithm (typically of the Newton-Raphson form). The specific algorithm used in this study follows D. J. Dupuis (1998):

STEP 1: Decide on a precision threshold, η, and initial value for θ, and initial values $a = 0$ and $A = \sqrt{\left[J(\theta)^{-1}\right]^T}$ where $J(\theta) = \int s(x;\theta)\cdot s(x;\theta)^T \, dF_\theta(x)$ is the Fisher Information.

STEP 2: Solve for a and A in the following equations:

$$A^T A = M_2^{-1} \quad \text{and} \quad a = \int s(x,\theta) W_c(x,\theta) dF_\theta(x) \Big/ \int W_c(x,\theta) dF_\theta(x)$$

where $M_k = \int \left[s(x;\theta) - a\right] \cdot \left[s(x;\theta) - a\right]^T \cdot W_c(x,\theta)^k \, dF_\theta(x)$, k=1,2

which gives the "current values" of θ, a, and A used to solve the given equations.

STEP 3: Now compute M_1 and $\Delta\theta = M_1^{-1} \cdot \left\{ \frac{1}{n} \cdot \sum_{i=0}^{n} \left[s(x_i;\theta) - a\right] \cdot W_c(x_i,\theta) \right\}$

STEP 4: If $\max_j \left| \frac{\Delta\theta_j}{\theta_j} \right| > \eta \;\; (j = 1,2)$ then $\theta \rightarrow \theta + \Delta\theta$ and return to STEP 2, otherwise stop.

The idea of the above algorithm is to first compute A and a for a given θ by solving the Step 2 equations. This is followed by a Newton-Raphson step given these two new matrices, and these steps are iterated until convergence is achieved. Dupuis (1998) cautions on two points of implementation in an earlier paper by Alaiz and Victoria-Feser (1996):

- Alaiz and Victoria-Feser (1996) state that integration can be avoided in the calculation of a in STEP 2 and M_1 in STEP 3, but Dupuis (1998) cautions that the former calculation of a requires integration, rather than a weighted average based on plugging in the empirical density, or else (64) will be satisfied by all estimates.

- Also, perhaps mainly as a point of clarification, Dupuis (1998) clearly specifies $\max_j \left| \frac{\Delta\theta_j}{\theta_j} \right| > \eta \;\; (j = 1,2)$ in STEP 4 rather than just $\Delta\theta > \eta$ as in Alaiz and Victoria-Feser (1996).

The initial values for *A* and *a* in STEP 1 correspond to the MLE. The algorithm converges if initial values for θ are reasonably close to the ultimate solution. Initial values can be MLE, or a more robust estimate from another estimator, or even an OBRE estimate obtained with *c* = large and initial values as MLE, which would then be used as a starting point to obtain a second and final OBRE estimate with *c* = smaller. In this study, MLE estimates were used as initial values, and no convergence problems were encountered, even when the loss dataset contained 6% arbitrary deviations from the assumed model.

It should be noted that OBRE weights contain very valuable information: they are indicators of the degree to which a particular data point (within the context of the data sample at hand) deviates from the assumed statistical model. As such they can be used for outlier detection, unit-of-measure construction, and possibly in the parameter estimation process itself. For the latter, they are arguably superior to "trimming" (observation deletion) based on sample quantiles, maximum/minimum *k* observations, absolute deviations, or other relatively arbitrary and inflexible metrics.

6.3. Influence Functions of Robust Estimators versus MLE

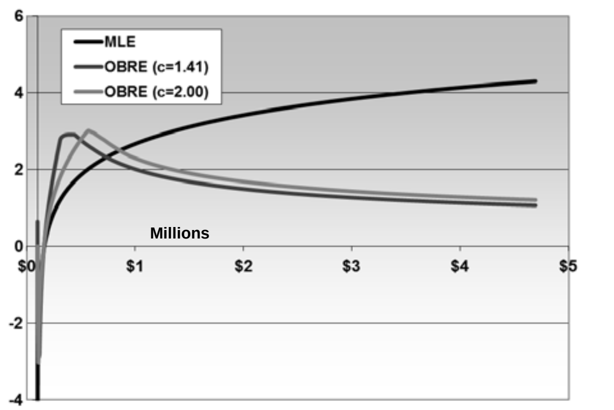
6.3.1 Non-truncated Severity Distributions

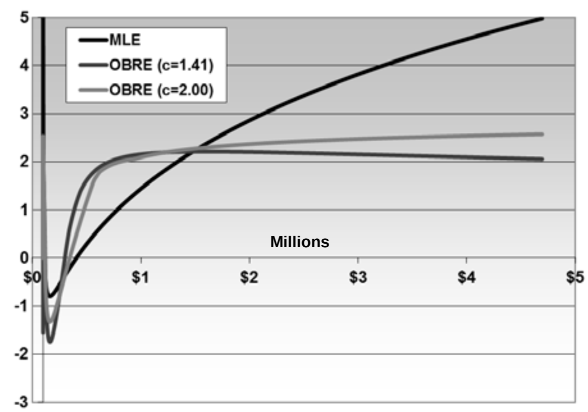
The B-robustness of CvM and OBRE can be seen when their EIFs are compared to those of the MLE estimators for the parameters of the examined severity distributions. While the EIFs of the MLE estimators clearly are unbounded, those of OBRE and CvM clearly are bounded. And importantly, note, too, that the extreme sensitivity exhibited by MLE estimators to *small* arbitrary data deviations, not just large deviations, is mitigated by the boundedness of the robust estimators. This is highlighted in Figures 9, and 12, although it is true for the EIFs represented on all the graphs, even when the y-axis is not scaled for visual clarity.

Figure 8:

EIFs of LogNormal (n = 250, μ = 11, σ = 2) Parameter Estimates: OBRE vs. MLE

EIF$_\mu$ EIF$_\sigma$



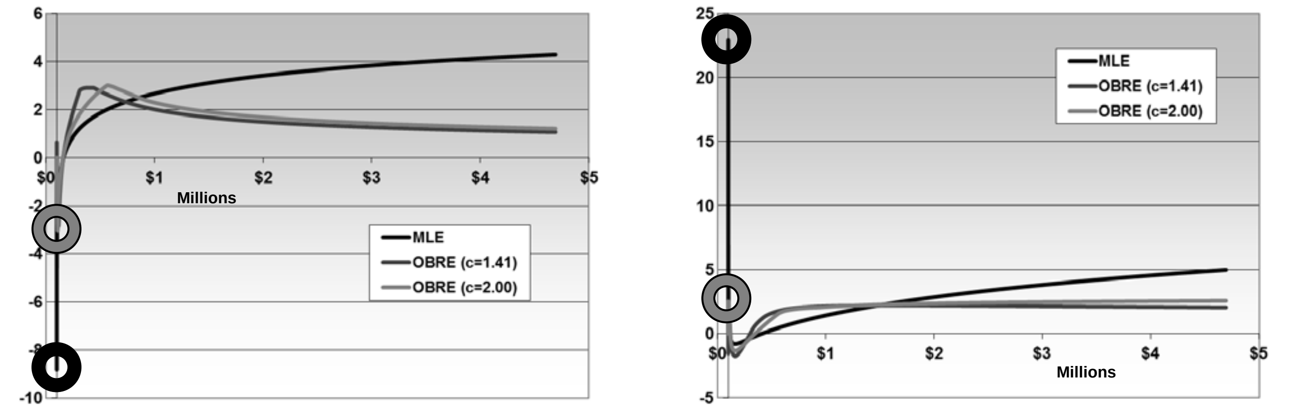


For the LogNormal, as seen previously the EIFs for the MLE estimators of μ and σ are unbounded, heading toward $-\infty$ and $+\infty$, respectively, as x approaches 0, and both heading toward $+\infty$ as x approaches $+\infty$. In contrast, the EIFs for the OBRE estimators are bounded over the entire domain of possible losses $(0, \infty)$. Note that as c increases, the OBRE estimators are less robust, which can be seen in the EIF for c = 2.0 deviating further from the x-axis in the right tail.

Figure 9 (Figure 8 with larger y-axis range):

EIFs of LogNormal (n = 250, μ = 11, σ = 2) Parameter Estimates: OBRE vs. MLE

EIF$_\mu$ EIF$_\sigma$

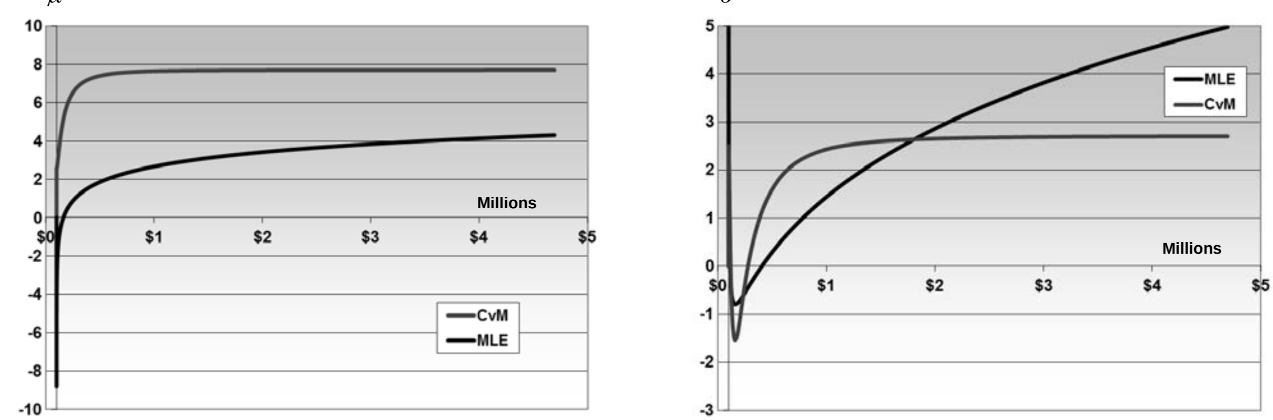


Note in Figure 10 that CvM, like OBRE, is B-robust; EIFs for MLE diverge to $\pm\infty$, but EIFs for CvM do not over the entire domain of possible losses $(0, \infty)$.

Figure 10:

EIFs of LogNormal (n = 250, μ = 11, σ = 2) Parameter Estimates: CvM vs. MLE

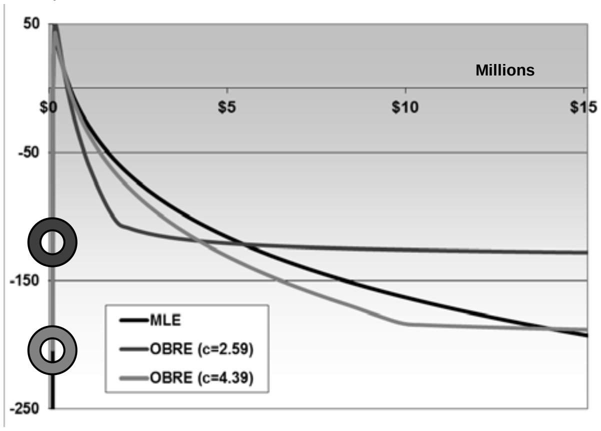
EIF$_\mu$ EIF$_\sigma$



For the LogGamma distribution, the EIFs for the MLE estimators of a and b are unbounded, heading toward $-\infty$ as x approaches both $-\infty$ and $+\infty$. In contrast, the EIFs for the OBRE estimators are bounded as x approaches both $-\infty$ and $+\infty$. Note that as c increases, the OBRE estimators are less robust, which can be seen with the EIF for c = 4.39 deviating further from the x-axis in the right tail than does that for c = 2.59.

Figure 11:
EIF of LogGamma (n = 250, a = 35.5, b = 3.25) Parameter Estimates: OBRE vs. MLE

EIF$_a$



EIF$_b$

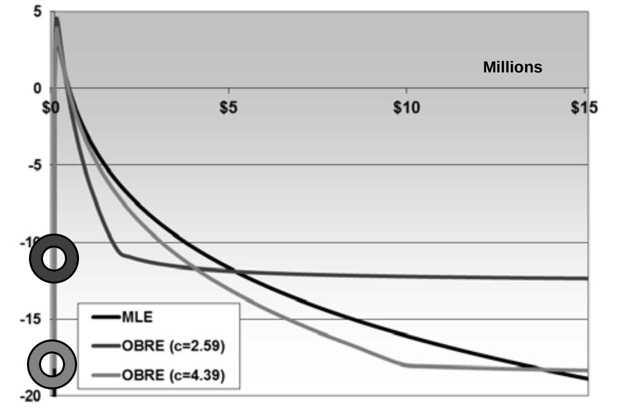
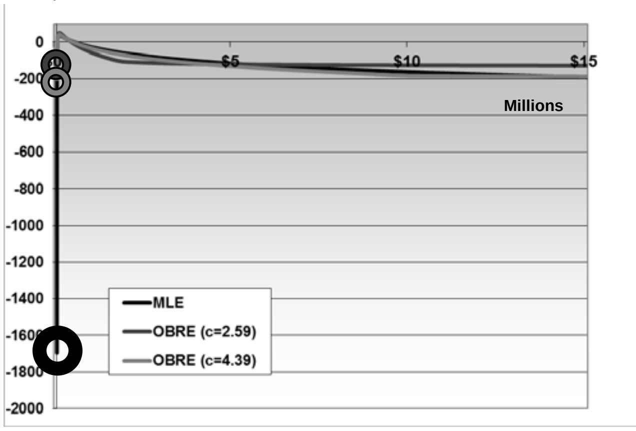


Figure 12 (Figure 15 with larger y-axis range):
EIF of LogGamma (n = 250, a = 35.5, b = 3.25) Parameter Estimates: OBRE vs. MLE

EIF$_a$



EIF$_b$

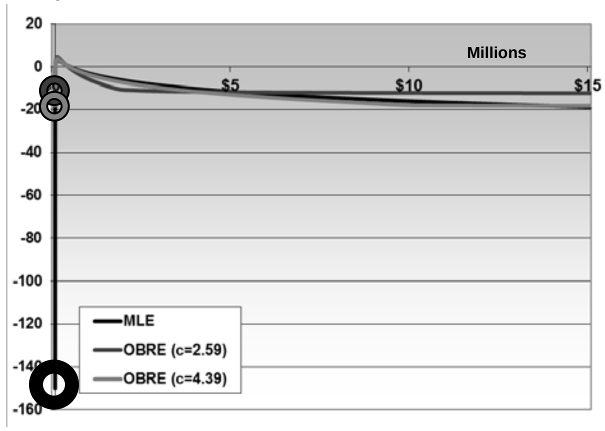
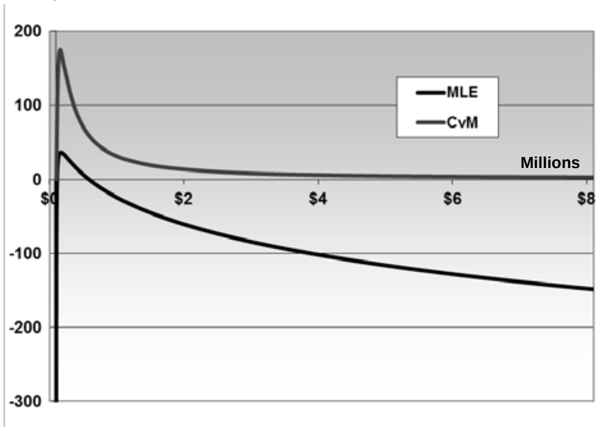
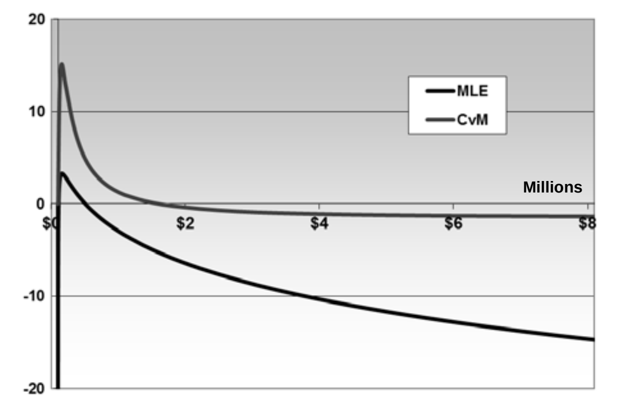


Figure 13
EIF of LogGamma (n = 250, a = 35.5, b = 3.25) Parameter Estimates: CvM vs. MLE

EIF$_a$



EIF$_b$



Note in Figure 13 that CvM, like OBRE, is B-robust; IFs for MLE diverge to ±∞, but IFs for CvM do not over the entire domain of possible losses (0, ∞).

6.3.2 Truncated Severity Distributions

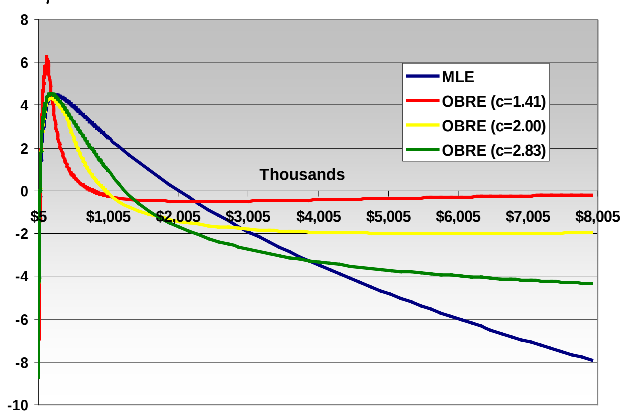
As the analytical derivation of the Influence Function made clear, truncation substantially changes the sensitivity of parameter estimates to arbitrary deviations from the assumed model. The B-robustness of CvM and OBRE can be seen below when their EIFs are compared to those of the MLE estimators for the parameters of the examined severity distributions. While the EIFs of the MLE estimators clearly are unbounded, those of OBRE and CvM clearly are bounded. And importantly, under truncation, just as under non-truncated severity distributions (above), the extreme sensitivity of MLE estimators to *small* arbitrary data deviations, not just large deviations, is not shared by the robust estimators (this is true, of course, even when y-axes are scaled and large MLE EIF values are cut-off for visual clarity).

Figure 14:

EIFs of Truncated LogNormal (n = 250, μ = 11, σ = 2, H = \$5,000) Parameter Estimates:

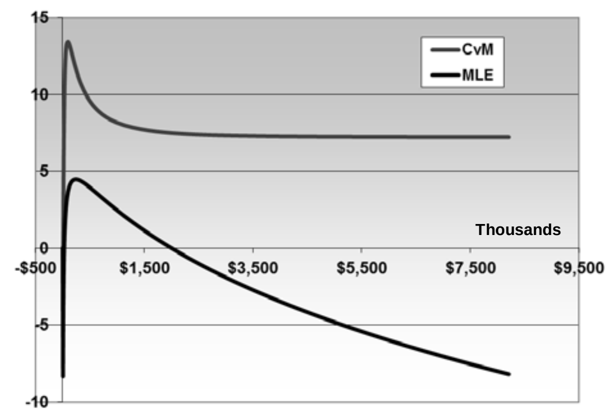
OBRE vs. MLE

EIF_{μ}

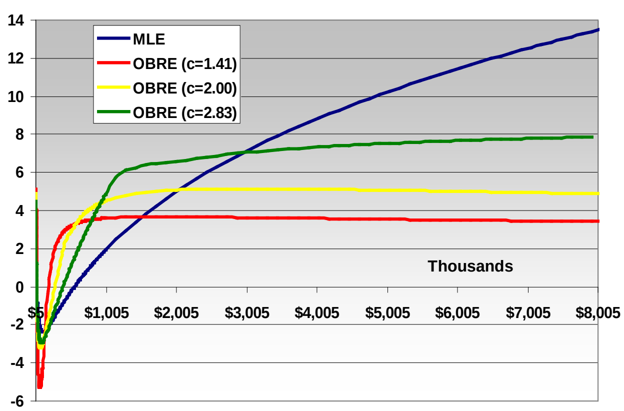


CvM vs. MLE

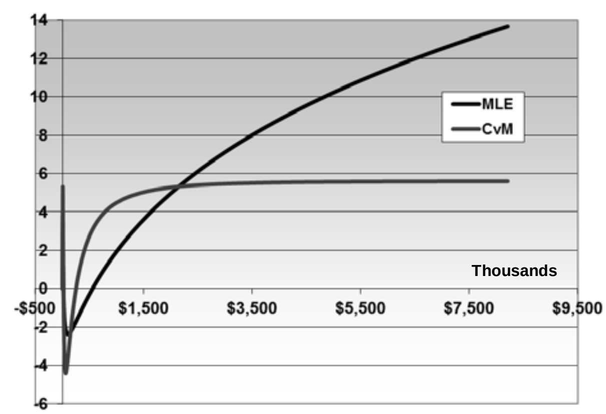
EIF_{μ}



EIF_{σ}



EIF_{σ}

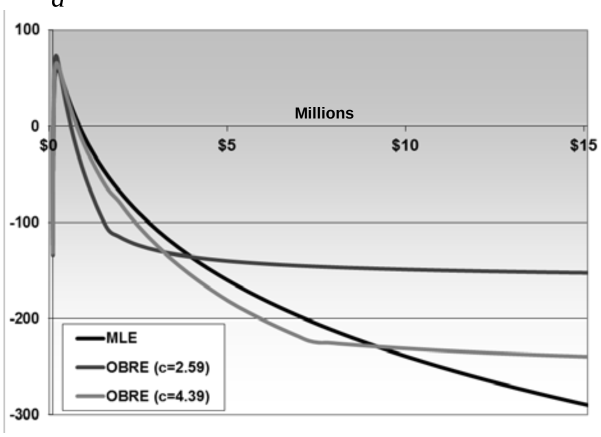


Under truncation, the EIFs for the MLE estimators of μ and σ are non-robust: while they are somewhat less sensitive to new, small, left-tail losses (as seen in Table 1 above), they are much more sensitive to large losses, diverging more quickly to $-\infty$ and $+\infty$, respectively, as x approaches $+\infty$. In contrast, the EIFs for the OBRE estimators exhibit a range of robustness for different values of c, as evidenced by the differing locations on the y-axis where the IF becomes nearly flat for large x. This demonstrates the flexibility of the estimator, via values of its tuning parameter, to provide different levels of robustness in the face of loss samples and severity distributions that can vary notably across different units of measure. CvM, too, is B-robust.

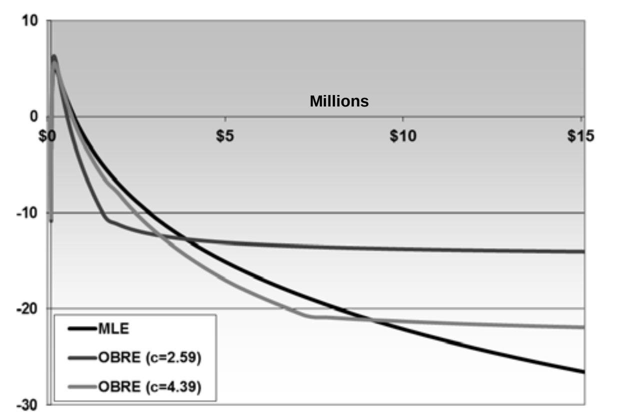
Figure 15:
EIFs of Truncated LogGamma (n=250, a = 35.5, b = 3.25, H = \$5,000) Parameter Estimates

OBRE vs. MLE

EIF$_a$

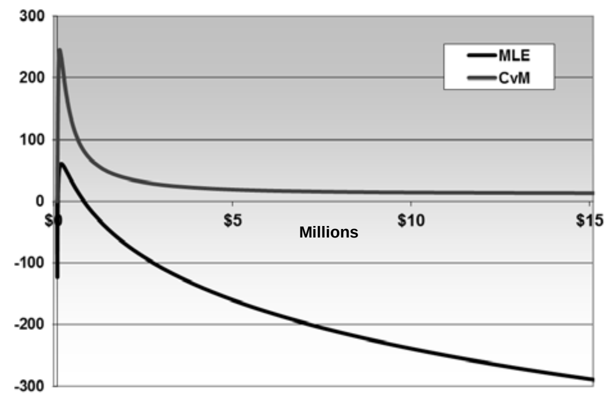


EIF$_b$

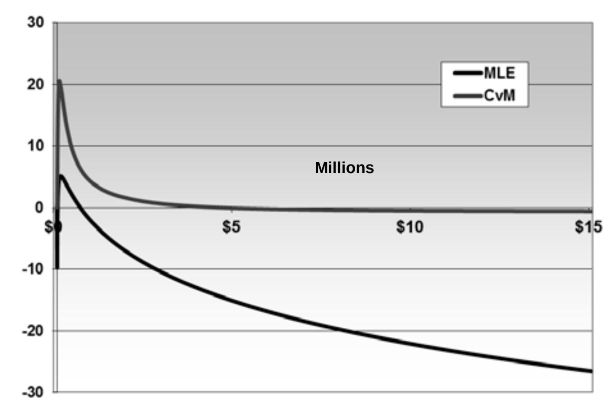


CvM vs. MLE

EIF$_a$



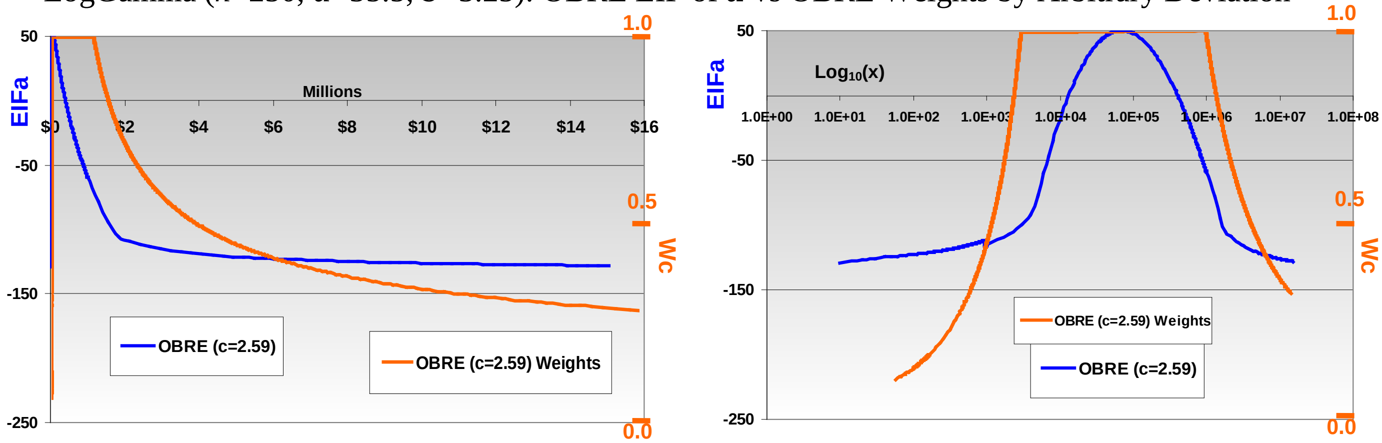
EIF$_b$



6.4. The Information Contained in OBRE Weights

Because of their potential utility for many uses in the operational risk measurement space, including, possibly, in the severity distribution parameter estimation process itself, we present below graphs of the data point weights generated by the OBRE algorithm over its corresponding IF – here, the LogGamma IF_a.

Figure 16:
LogGamma (*n*=250, *a*=35.5, *b*=3.25): OBRE EIF of ***a*** vs OBRE Weights by Arbitrary Deviation



The OBRE EIF begins to flatten out, that is, it diverges from the EIF of an unconstrained MLE, exactly where it should: where the OBRE weights deviate from the value of 1.0. In this specific case, this happens roughly at the fifth and ninety-fifth quantiles of the severity distribution [38].

7. Capital Estimation

As the convolution of the frequency and severity distributions, the annual aggregate loss distribution, for which we must obtain a 99.9^{th}%tile VaR as the capital estiamte, has no general closed form solution; therefore, large scale Monte Carlo simulations are the gold standard for obtaining the "true" capital requirement for a given set of frequency and severity distribution parameters. However, a number of less computationally intensive methods exist, the most convenient of which is the mean adjusted Single Loss Approximation (SLA) of Degen (2010).[39] Given a desired level of statistical confidence (α), an estimate of forward-looking annual loss

[38] GPD and truncated GPD were not examined in Section 6 because time did not permit their inclusion in the capital simulation study in Section 7. Preliminary results for Section 7 indicate findings consistent with those presented, although the simulations differed in one respect: when random samples produced estimates of $\xi \geq 1$, (68a) rather than (68) were, and must be, used for capital calculations..

[39] Degen's (2010) fomula is supported by analytic derivations, whereas that of Böcker and Sprittulla (2006), which is commonly used, is based on empirical observation (although both are very similar: that latter has a second term of $(\lambda - 1)\mu$ rather than $\lambda\mu$).

frequency (λ), an assumed severity distribution ($F(\ )$), and values for the parameters of the severity distribution (θ), capital requirements are approximately given by:[40]

$$C_{\alpha} \approx F^{-1}\left(1-\frac{1-\alpha}{\lambda}\right)+\lambda\mu \tag{63}$$

where
α = single year loss quantile (0.999 for regulatory capital; 0.9997 for economic capital)
λ is the average number of losses occurring within one year (the frequency estimate)
μ is the mean of the estimated severity distribution

From (68) we can see that the VaR of the aggregate loss distribution is essentially just a high quantile of the severity distribution on a single loss (the first term) with a mean adjustment that typically is very small (the second term) relative to the first term (in the remainder of the paper, our use of the "SLA" refers to the mean adjusted SLA). And for the case of severity distributions with infinite mean, say, a GPD severity with $\xi \geq 1$, Degen derives an SLA approximation that is not dependent upon the mean of the distribution:

$$C_{\alpha} \approx F^{-1}\left(1-\frac{1-\alpha}{\lambda}\right)-\left(1-\alpha\right)F^{-1}\left(1-\frac{1-\alpha}{\lambda}\right)\cdot\left(\frac{c_{\xi}}{1-1/\xi}\right) \tag{63a}$$

where $c_{\xi}=\left(1-\xi\right)\dfrac{\Gamma^{2}\left(1-1/\xi\right)}{2\Gamma\left(1-2/\xi\right)}$ if $1<\xi<\infty$ and $c_{\xi}=1$ if $\xi=1$

But in either case, the important point here is that the estimate of the high quantile of the severity distribution is what really matters in estimating capital. In other words, the statistical properties of this high quantile estimate (accuracy, robustness, and precision), NOT those of the parameter estimates per se, should be the ultimate arbiter of what makes a good estimator for quantifying operational risk exposure. Those who focus only on the performance of the estimated parameters while ignoring the performance of the corresponding capital estimates do so at their peril, because they most certainly are not the same thing, and those who blithely assume so may miss very important issues that adversely affect capital estimation. Unfortunately, we believe that is at least part of the reason that the operational risk measurement literature has, to date, missed the fundamental issue of MLE-related capital bias. The source of this bias is Jensen's inequality, which is an analytical result proven over a century ago.

[40] Capital estimates based on the mean VaR of three million fully simulated aggregate loss distributions, with both frequency and severity parameters simulated (which is the gold standard here) were compared to SLA approximations to verify the accuracy of the latter. Across the 16 "true" approximated SLA capital numbers in Table 3, the mean absolute deviation from the corresponding fully simulated numbers was only 1.2%, and the standard deviation of the absolute deviations was only 1.2%.

We employ a simulation study to evaluate the performance of capital estimates based on severity parameters estimated with MLE, CvM, and OBRE. The results of the simulation study demonstrate that Jensen's inequality can have a very large impact on estimated capital.

7.1. Experimental Design for Simulation Study

With a comprehensive simulation study, we compare SLA-based capital estimates using three estimators to model the severity distribution: MLE, CvM, and OBRE. The experimental design covers four of the severity distributions examined closely above – the LogNormal, Truncated LogNormal, LogGamma, and Truncated LogGamma. We vary the underlying data characteristics to capture the potential effects of heterogeneity, considering in turn idealized data conditions (i.i.d. losses), and different types of data contamination (left tail only, right tail only, and both tails).

- N = # of simulations = 500
- n = # of losses per unit of measure (for a single unit of measure) = 250
  (Many units of measure will be smaller, and many will be larger, so 250 was chosen as a realistic "typical" size)
- H = data collection threshold for truncation = $5,000
  (Even this relatively small collection threshold illustrates the sometimes dramatic effects of truncation on capital estimation)
- Confidence level for Basel II regulatory capital: α = 99.9%
  (Corresponding to the confidence level for regulatory capital requirements)
- Loss frequency λ = 25
  (This value, which is what typically is estimated by the frequency model, was arbitrarily selected to illustrate the capital impact. For larger banks, the average annual number of losses may be far larger, thus increasing capital requirements dramatically, ceteris paribus).
- Parameter Values:
  - For both LogNormal and Truncated LogNormal, μ = 11, σ = 2
  - For both LogGamma and Truncated LogGamma, a = 35.5, b = 3.25

    (These parameter values were selected to represent some of the loss distributions and capital estimates encountered by the authors from analysis of actual operational loss data. In addition, these parameter values demonstrate the fairly large differences between the LogNormal and the LogGamma, especially in the tails, as shown in Table 2 below. Yet they yield similar means of about $442,000 and $467,000, respectively. Obviously, the capital impacts only can be assessed using parameter values and sample sizes relevant to a particular institution's unit of measure.)

Table 2: Quantiles by Loss Severity Distributions

%Tile	LogNormal (μ=11, σ=2)	LogGamma (a=35.5, b=3.25)
50.0000%	$59,874	$50,045
75.0000%	$230,724	$179,422
90.0000%	$776,928	$614,477
95.0000%	$1,606,723	$1,333,228
99.0000%	$6,278,840	$6,162,960
99.9000%	$28,932,168	$38,778,432
99.9700%	$57,266,640	$92,087,922
99.9960%	$159,698,811	$355,104,952
99.9988%	$279,358,818	$760,642,911

- Arbitrary Deviations/Contamination: Mixture distributions are used to test the robustness of the estimators to deviations from i.i.d. data.
    - Three scenarios are studied:
        - 6% Left tail contamination – contamination for "small" losses
        - 6% Right tail contamination – contamination for "large" losses
        - 3% Left tail + 3% Right tail contamination
    - Data contamination: The deviating data has a mean that deviates, in both directions, by just under $350,000 from the respective base distributions (which is sizeable given the base loss distributions have means of about $450,000)
        - LogNormal contamination:
          Left tail contamination: LogNormal (μ = 9.5, σ = 2)
          Right tail contamination: LogNormal (μ = 11.576, σ = 2)
        - LogGamma contamination
          Left tail contamination: LogGamma (a = 31.8, b = 3.25)
          Right tail contamination: LogGamma (a = 37, b = 3.25)
- OBRE Tuning Parameter: For OBRE, different values for c, the tuning parameter, were used with the given parameter values, and those which provided the most obviously appropriate tradeoff between accuracy and precision of the corresponding SLA capital estimates were used. Algorithms that may allow for purely data-driven estimation of c, which is key to its usage in practice, are discussed below.
- OBRE Starting Values: MLE estimates were used as the starting point for the OBRE algorithm, and for this study, no convergence problems were encountered. That said, values of η, c, n, and the distribution parameters all are very interrelated, and like any convergence algorithm, must be carefully monitored. For example, values of η = 0.01 were sufficient for LogNormal parameter estimation, but for LogGamma estimation, η = 0.005 and even η = 0.0001 were sometimes required due to its longer tail and the need for greater precision. Such variation under different data conditions is typical of convergence algorithms, so their responsible use requires an awareness of these issues. While starting values are sometimes noted in the literature as being important for the

convergence of OBRE algorithms, this emphasis may be due to the relatively small sample sizes (as low as n = 40) being used in some of those studies (see Horbenko, et al., 2011).

- CvM Starting values: A wide range of parameter values were provided for the Gaussian quadrature optimization algorithm used by the statistical software utilized here (SAS®). No convergence issues were encountered with the LogNormal and Truncated LogNormal distributions, but that was not the case in fully a third of the LogGamma and Truncated LogGamma distributions (whether or not contaminated) where second-order optimality conditions were violated (the simulated capital results were similar whether or not these instances were discarded and replaced).

7.2. Simulation Results

Table 3 summarizes the potential bias mitigation that OBRE provides relative to MLE.

Table 3: Summary of SLA Estimates of Regulatory Capital (#sims=500):
MLE, CvM, and OBRE (Optimal Tuning Parameter and Weight Exclusion for OBRE)

	0% Deviation	**3% Each Tail**	**6% Left Tail**	**6% Right Tail**
LogNormal				
True SLA Capital	$170,317,921	$173,118,560	$165,323,008	$180,654,136
Deviation of **OBRE** Mean	$671,849	$3,996,815	$8,387,953	-$3,033,449
Deviation of **MLE** Mean	$7,504,017	$11,745,639	$15,748,335	$5,806,548
Deviation of **CvM** Mean	$14,893,442	$14,554,328	$17,167,913	$8,703,128
Truncated LogNormal				
True SLA Capital	$180,486,144	$183,180,240	$175,278,136	$190,682,320
Deviation of **OBRE** Mean	$225,670	$8,732,300	$12,744,475	$5,867,546
Deviation of **MLE** Mean	$20,985,417	$24,473,149	$28,282,561	$24,238,437
Deviation of **CvM** Mean	$25,357,240	$32,479,789	$37,850,365	$31,725,240
LogGamma				
True SLA Capital	$366,309,627	$370,407,112	$353,009,568	$387,304,656
Deviation of **OBRE** Mean	-$5,326,671	$13,270,864	$21,020,814	-$2,168,419
Deviation of **MLE** Mean	$48,715,951	$60,143,554	$67,193,035	$47,375,062
Deviation of **CvM** Mean	$70,389,855	$90,109,056	$96,544,056	$77,895,220
Truncated LogGamma				
True SLA Capital	$388,391,019	$392,310,056	$374,657,472	$409,562,640
Deviation of **OBRE** Mean	$18,617,463	$6,390,621	$15,298,931	$1,331,382
Deviation of **MLE** Mean	$81,838,600	$78,081,913	$88,430,354	$69,997,575
Deviation of **CvM** Mean	$136,214,444	$126,769,342	$134,760,825	$114,930,957

Complete simulation results providing the characteristics of the distributions of each estimator's capital estimates are shown in Tables 3a-d, which are the backup to the summary Table 3.

OBRE means generally are much more centered on the true capital than those of MLE, which exhibit notable upward bias that increases dramatically for economic capital (where $\alpha = 0.9997$ instead of $\alpha = 0.999$) as shown in the next Section. CvM exhibits notable bias as well. It appears that OBRE's tuning parameter provides some measure of mitigation of the effects of Jensen's inequality, without which CvM clearly is too blunt a tool for capital estimation (i.e. high quantile estimation).

In terms of robustness, we see from Tables 3a-d that OBRE means under deviations from i.i.d. data are much closer to true capital requirements than are those of MLE or CvM, but at least part of this effect is likely due to Jensen's inequality. It is not clear that simply subtracting the MLE-related bias under no deviations (i.i.d. data) allows for the disentanglement of these two effects, but what is clear is that OBRE is far closer to the true required capital under all four data scenarios. Approaches to mitigate bias specifically due to Jensen's inequality are discussed in the next section.

In terms of precision, OBRE and MLE are generally comparable, and CvM's performance was markedly inferior across the board. The RMSE of OBRE typically was slightly larger than that of MLE, although it sometimes was smaller (see Truncated LogGamma, Table 3d).[41] In terms of the percentage of estimates within 50% of the true value, OBRE and MLE are comparable. Ways in which OBRE's precision might be increased notably via factor models are discussed in following sections.

It seems that while both OBRE and CvM are B-robust, when it comes to capital estimation the tuning parameter of OBRE gives it a dramatic advantage, perhaps providing the key adaptability that CvM lacks for estimation across very different severity distributions. As noted above in the definition of OBRE, the tuning parameter is selected by the researcher: it is not (currently) a parameter whose value is determined by the data sample at hand. In the simulations above, tuning parameter value selection was handled by "backing into" near optimal values knowing the true SLA value ex ante. Informal robustness tests were then conducted ex post to provide an initial assessment as to the feasibility of developing a fully data-driven process for estimating, rather than choosing, the tuning parameter. This only would be viable if the value of c ultimately chosen is robust to initial parameter misspecification. Preliminary tests indicate that it is.

For the LogNormal and Truncated LogNormal, the values of c chosen for Tables 3a-3b were subsequently tested on data samples generated from distributions with both parameters a full standard deviation away from the original parameters in the same direction (for two independent normally distributed parameters, the probability of this occurring is very small: $p < 0.03$). This created distributions with means **–**\$115K / +\$160K smaller/larger, respectively, for the LogNormal, and **–**\$185 / +\$311K smaller/larger for the Truncated LogNormal. In both cases, the original value of c was chosen as the best c, indicating robustness to initial parameter misspecification. For these two distributions, these results bode well for the development of a robust, fully data-driven estimation process for selecting the value of c, the tuning parameter, as has been done for other robust estimators with tuning parameters (see for example Hong and

[41] The Root Mean Square Error (RMSE) simply is the square root of the Mean Square Error (MSE).

Table 3a: LogNormal SLA Estimates of Regulatory Capital (#sims=500):
MLE, CvM, OBRE (Optimal Tuning Parameter)

		0% Deviation	6% Deviation	6% Deviation	6% Deviation
			Both Tails (3% Each)	Left Tail	Right Tail
	True SLA Capital	$170,317,921	$173,118,560	$165,323,008	$180,654,136
OBRE*	Mean	$170,989,770	$177,115,375	$173,710,961	$177,620,687
MLE	Mean	$177,821,938	$184,864,199	$181,071,343	$186,460,684
CvM	Mean	$185,211,363	$187,672,888	$182,490,921	$189,357,264
	OBRE Closer v. MLE	**$6,832,168**	**$7,748,825**	**$7,360,382**	**$2,773,099**
OBRE*	Mean %Difference from True	0.4%	2.3%	5.1%	-1.7%
MLE	Mean %Difference from True	4.4%	6.8%	9.5%	3.2%
CvM	Mean %Difference from True	8.7%	8.4%	10.4%	4.8%
OBRE*	% within +/- 50%	80.0%	84.0%	82.0%	86.0%
MLE	% within +/- 50%	80.0%	83.0%	80.0%	87.0%
CvM	% within +/- 50%	74.0%	70.0%	75.0%	77.0%
OBRE*	RMSE	$79,571,542	$76,325,792	$70,325,414	$73,332,644
MLE	RMSE	$79,516,780	$68,157,312	$66,129,189	$66,662,079
CvM	RMSE	$102,185,795	$93,890,094	$85,334,514	$88,837,541

*NOTE: $c = 2^{(11/8)} \approx 2.59$

Table 3b: Truncated LogNormal SLA Estimates of Regulatory Capital (#sims=500):
MLE, CvM, OBRE (Optimal Tuning Parameter)

		0% Deviation	6% Deviation	6% Deviation	6% Deviation
			Both Tails (3% Each)	Left Tail	Right Tail
	True SLA Capital	$180,486,144	$183,180,240	$175,278,136	$190,682,320
OBRE*	Mean	$180,711,814	$191,912,540	$188,022,611	$196,549,866
MLE	Mean	$201,471,561	$207,653,389	$203,560,697	$214,920,757
CvM	Mean	$205,843,384	$215,660,029	$213,128,501	$222,407,560
	OBRE Closer v. MLE	**$20,759,747**	**$15,740,849**	**$15,538,087**	**$18,370,891**
OBRE*	Mean %Difference from True	0.1%	4.8%	7.3%	3.1%
MLE	Mean %Difference from True	11.6%	13.4%	16.1%	12.7%
CvM	Mean %Difference from True	14.0%	17.7%	21.6%	16.6%
OBRE*	% within +/- 50%	72.0%	70.0%	71.0%	76.0%
MLE	% within +/- 50%	71.0%	73.0%	72.0%	74.0%
CvM	% within +/- 50%	59.0%	64.0%	60.0%	64.0%
OBRE*	RMSE	$133,209,674	$110,730,346	$116,252,565	$129,840,945
MLE	RMSE	$140,551,905	$109,436,060	$111,794,444	$118,952,011
CvM	RMSE	$318,935,475	$148,248,416	$151,366,218	$176,135,956

*NOTE: $c = 2^{(9/8)} \approx 2.18$

Table 3c: LogGamma SLA Estimates of Regulatory Capital (#sims=500):
MLE, CvM, OBRE (Optimal Tuning Parameter and Weight Cut-off)

		0% Deviation	6% Deviation Both Tails (3% Each)	6% Deviation Left Tail	6% Deviation Right Tail
	True SLA Capital	$366,309,627	$370,407,112	$353,009,568	$387,304,656
OBRE*	Mean	$360,982,956	$383,677,976	$374,030,382	$385,136,237
MLE	Mean	$415,025,578	$430,550,666	$420,202,603	$434,679,718
CvM	Mean	$436,699,482	$460,516,168	$449,553,624	$465,199,876
	OBRE Closer v. MLE	**$43,389,280**	**$46,872,690**	**$46,172,221**	**$45,206,643**
OBRE*	Mean %Difference from True	-1.5%	3.6%	6.0%	-0.6%
MLE	Mean %Difference from True	13.3%	16.2%	19.0%	12.2%
CvM	Mean %Difference from True	19.2%	24.3%	27.3%	20.1%
OBRE*	% within +/- 50%	59.0%	71.0%	72.0%	76.0%
MLE	% within +/- 50%	63.0%	75.0%	70.0%	78.0%
CvM	% within +/- 50%	54.0%	62.0%	59.0%	63.0%
OBRE*	RMSE	$222,205,047	$258,303,584	$252,743,932	$252,990,317
MLE	RMSE	$271,095,454	$243,734,467	$233,682,773	$244,208,780
CvM	RMSE	$331,448,466	$332,027,462	$310,337,566	$332,386,275

*NOTE: $c = 2^{(19/8)} \approx 5.187$, and weight inclusion criterion = $W \geq 0.85$

Table 3d: Truncated LogGamma SLA Estimates of Regulatory Capital (#sims=500):
MLE, CvM, OBRE (Optimal Tuning Parameter and Weight Cut-off)

		0% Deviation	6% Deviation Both Tails (3% Each)	6% Deviation Left Tail	6% Deviation Right Tail
	True SLA Capital	$388,391,019	$392,310,056	$374,657,472	$409,562,640
OBRE*	Mean	$407,008,482	$398,700,677	$389,956,403	$410,894,022
MLE	Mean	$470,229,619	$470,391,969	$463,087,826	$479,560,215
CvM	Mean	$524,605,463	$519,079,398	$509,418,297	$524,493,597
	OBRE Closer v. MLE	**$63,221,137**	**$71,691,291**	**$73,131,423**	**$68,666,193**
OBRE*	Mean %Difference from True	4.8%	1.6%	4.1%	0.3%
MLE	Mean %Difference from True	21.1%	19.9%	23.6%	17.1%
CvM	Mean %Difference from True	35.1%	32.3%	36.0%	28.1%
OBRE*	% within +/- 50%	56.0%	60.0%	66.0%	67.0%
MLE	% within +/- 50%	63.0%	67.0%	66.0%	76.0%
CvM	% within +/- 50%	50.0%	51.0%	54.0%	62.0%
OBRE*	RMSE	$273,966,583	$237,477,157	$237,181,395	$272,922,481
MLE	RMSE	$360,712,711	$237,737,636	$270,317,853	$311,345,233
CvM	RMSE	$584,908,158	$393,702,817	$341,259,642	$440,805,965

*NOTE: $c = 2^{(19/8)} \approx 5.187$, and weight inclusion criterion = $W \geq 0.85$

However, for the LogGamma and Truncated LogGamma, selecting the "optimal" tuning parameter was not enough to mitigate the effects of Jensen's inequality: something else was needed. The information-laden OBRE weights are a natural place to turn to attempt to generate Kim (2001), Warwick (2004), Warwick and Jones (2004), and Wu et al., (2010)). And initial results from research currently under way are confirming this promise. SLA capital estimates with greater accuracy. One possibility is to first obtain OBRE weights on the data points, and then reestimate OBRE excluding observations with weights below a certain value, i.e. excluding those observations that deviate dramatically from the assumed distribution. Recall that weight values will change from sample to sample, because they are based on deviations from the presumed distribution (which is different for each sample), not on an arbitrary absolute value, or an arbitrary trimming requirement. Some samples will exclude no observations based on this criterion, and others will exclude several.

To obtain the SLA capital estimates for the LogGamma and Truncated LogGamma in Tables 3, 3c and 3d, a process similar to that used with the tuning parameter was followed: the optimal weight-exclusion value was found through simulation tests ex ante, and then tested for robustness to initial parameter misspecification ex post (this typically required the exclusion of only one or two data points in a given sample; the actual weight values for the inclusion rules are noted in Tables 3a-3d). The results of this ex post testing showed less robustness to parameter misspecification than did the selection of the tuning parameter (shifting parameter values by only half a standard deviation, in the same direction, began to yield values of exclusion weight thresholds different from the "optimal" values). This indicates that, for these distributions, developing a fully data-driven procedure for estimating data exclusion weight criteria appears more challenging than doing the same for the tuning parameter alone. This research is currently under way.

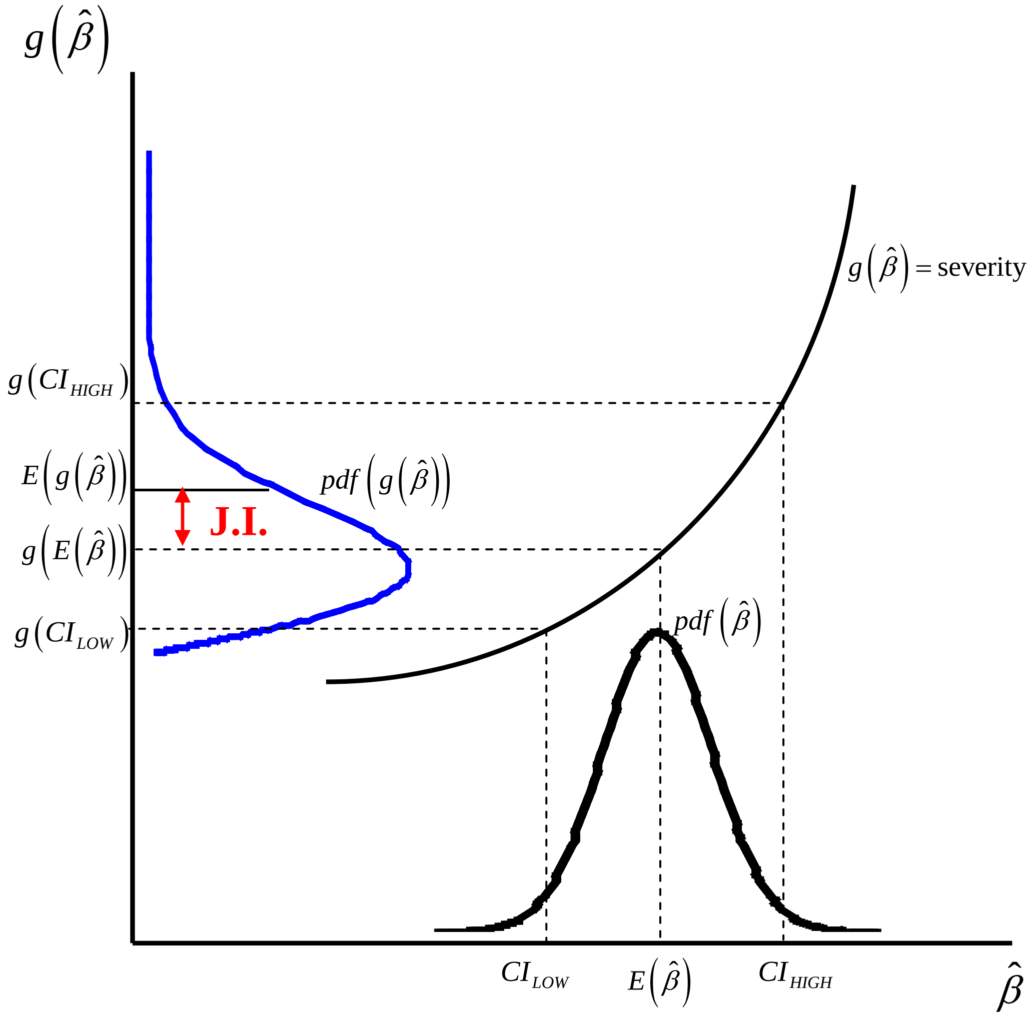
The capital estimates shown for OBRE in Table 3 require selection of the tuning parameter, as well as, for very heavy-tailed distributions, judicious use of OBRE's weights. Because neither of these selections is as yet fully data-driven, this should be considered research-in-progress, but the initial results on tuning parameter selection appear very promising and will be further developed.[42] For the very heavy-tailed severity distributions, in addition to tuning parameter selection, use of either OBRE's weights, or one of the proposed adjustments for Jensen's inequality discussed below would be required to mitigate the bias caused by Jensen's inequality.

7.3. Discussion: MLE Systematically Overstates Capital Requirements, Sometimes Dramatically

In addition to possible bias and inefficiency due to nonrobustness in the face of non-i.i.d. loss data, MLE-based capital estimates are biased even under i.i.d. data due to Jensen's inequality. As mentioned above, Jensen's inequality states that a (strictly) convex function of a mean is less than the mean of the function after its (convex) transformation (and the opposite is true for concave functions). This is illustrated in Figure 17 below (from Kennedy, 1992, p.37).

[42] Initial results that utilize OBRE's asymptotic variance to increase computational efficiency are very promising.

Figure 17: Graphical Display of Jensen's Inequality with Convex Function (right-skewed cdf)



This applies to quantile estimation of all commonly used severity distributions: if β is a random variable (here, our severity distribution parameter estimates) and $g(\ )$ is a (strictly) convex function (here, the inverse of our severity distribution cumulative distribution function), then $g(\mathrm{E}[\beta]) < \mathrm{E}[g(\beta)]$, and our quantile estimate (capital estimate) is biased upwards. The size of the MLE-based capital bias due to Jensen's inequality is a function of how convex is the function being applied to the data: the more convex is the function, the larger the bias. In this setting, the convexity of the right-skewed, heavy-tailed cumulative distribution function (cdf) of the severity distribution is a function of three things: a) the size of the quantile (the larger the quantile, the greater the convexity and the bias); b) for a given quantile, the thickness of the tail for a given severity distribution as determined by its parameter values (the heavier the tail, the greater the convexity and the bias); and c) the size of the variance and (finite sample) skewness of the parameter estimate (which are functions of the size of the sample of loss data for that unit of measure: the larger the sample, the smaller the convexity and the bias because the smaller the variance (and, for asymptotically normally distributed parameter estimates, as here, the smaller any finite sample skewness)).

Figure 18: Bias in Capital Estimates Due to Jensen's Inequality
by Severity Distribution by Quantile by Unit of Measure Sample Size

Truncated LogNormal ($\mu = 11$, $\sigma = 2$) Truncated LogGamma ($a = 35.5$, $b = 3.25$)

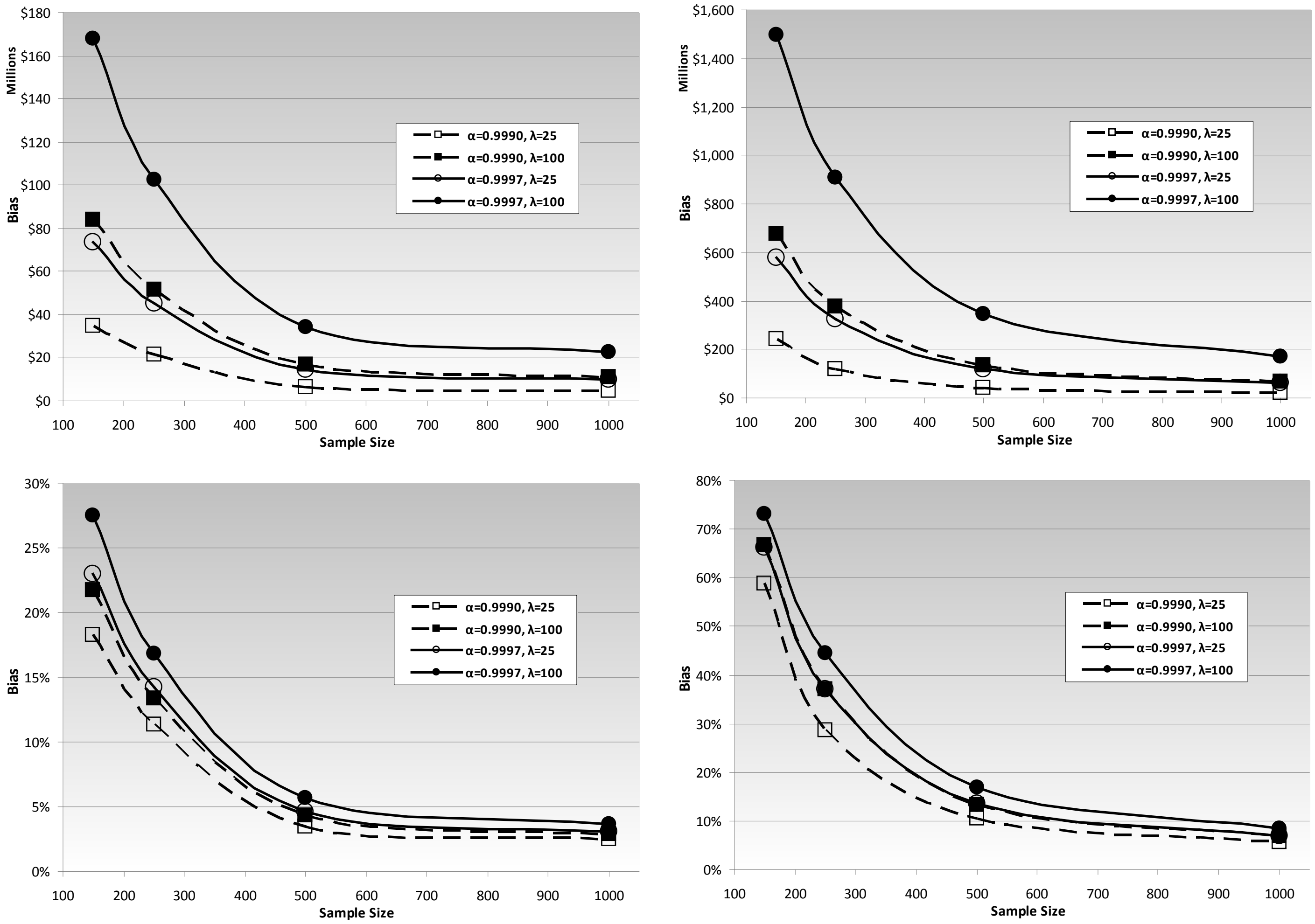


Figure 18 shows the magnitude, in both absolute and percentage terms, of the MLE-based capital bias due to Jensen's inequality by the severity distribution used; by the number of losses in the unit of measure sample; and by the size of the quantile (for regulatory capital, α = 99.9%tile and for economic capital, α = 99.97%tile with, conservatively and arbitrarily, λ= 25; the typically larger values of λ for most banks will yield much larger quantiles, and thus, much larger bias due to Jensen's inequality). A summary table is provided in Table 4 below, and complete backup tables to Table 4 and Figure 18 are available from the authors upon request.[43] On the one hand,

[43] These results represent calculations based on very extreme quantiles, and as such numerical precision issues related to the particular statistical software platform and the specific estimation algorithms used can, and as a general rule do, affect the results. These numbers all were generated using SAS®, within which multiple parameter estimation procedures (Proc Severity and Proc NLMixed) produced nearly identical results. When these calculations were reproduced and validated using R as a secondary check, "true" capital calculations were virtually identical, but for the means of the estimated capital, noticeable, though still relatively modest deviations occurred. Across all 192 estimated means in the backup tables (available from the authors upon request), the mean of the absolute deviations was less than 2%, the median of the absolute deviations was 1.6%, the standard deviation of the

the larger the sample size, the smaller is the bias, so for very large units of measure this effect probably is not material. On the other hand, for most units of measure which do not have large samples of loss data, the vast majority of the severity distributions in use, including the truncated LogNormal and truncated LogGamma, have tails that easily are thick enough to cause very sizeable bias in the capital estimates. In absolute terms, inflated capital estimates due to this bias can be hundreds of millions of dollars higher than the true capital required. And of course, the largest effect on the size of the bias is the size of the quantile: bias for economic capital is dramatically larger than that of regulatory capital, with the former sometimes over a billion dollars for large banks with large λ. And recall that this number refers to only one unit of measure out of what typically are dozens estimated for any given bank.

So why has this well established analytical result (proved in 1906; see Jensen, 1906) not been identified as relevant to this estimation problem until now? [44] We believe there are several possible reasons:

1. When examining the issue of the statistical estimation of severity distribution parameters, many operational risk researchers have focused exclusively on the statistical properties of the parameters without regard to the final and most important step: using those parameter estimates to estimate capital requirements.

2. Of those researchers who have examined capital distributions, most have done so using a relatively low regulatory capital quantile of the aggregate loss distribution (see Mignola & Ugoccioni (2006), who use a quantile based on very few loss events (e.g. λ = 10) which, all else equal, yields a much lower quantile on the aggregate loss distribution). However, for most severity distributions, higher quantiles are needed to achieve sufficient convexity in the cdf to actually notice the capital estimate bias due to Jensen's inequality. The magnitude of these quantiles, however, are well within the range of what must be used by banks in practice (e.g. λ = 25 to 100 used in this study corresponds to a very reasonable range of sample sizes for typical units of measure).

absolute deviations was 1.6%, and the interquartile range of the absolute deviations was 1.7%. These SAS® vs. R deviations were very similar to those produced merely by altering random number generation seeds, the corresponding numbers for one run of which were 1.2%, 0.7%, 1.1%, and 1.3%, respectively.

[44] We have been unable to find any reference to Jensen's inequality, directly or indirectly vis-à-vis mention of bias in MLE-based capital estimates, in the operational risk quantification and measurement literature.

Table 4: Summary Table of MLE-based Capital Estimate Bias due to Jensen's Inequality:
Bias based on the Mean of SLA Capital Estimates (#sims = 1,000)

Sample Size	Capital Requirement: Regulatory = 0.9990 Economic = 0.9997 VaR Quantile, alpha	Lambda	LogNormal (Mu=11, Sig=2)	Truncated LogNormal H = 10k (Mu=11, Sig=2)	LogGamma (a=35.5, b=3.25)	Truncated LogGamma H = 10k (a=35.5, b=3.25)
	VaR Quantile, alpha	Lambda	**True Required Capital**			
	0.999	25	$170,316,732	$189,116,361	$366,314,579	$415,653,661
	0.9997	100	$555,258,263	$610,623,698	$1,816,149,861	$2,045,751,095
	VaR Quantile, alpha	Lambda	**MLE Overshoot Bias a la Jensen's Inequality**			
150	0.999	25	$14,136,149	$34,469,523	$72,059,870	$243,889,760
250	0.999	25	$5,852,122	$21,479,420	$41,724,493	$118,740,721
500	0.999	25	$1,838,248	$6,629,785	$15,997,643	$43,640,327
1,000	0.999	25	$1,054,769	$4,650,714	$6,379,407	$22,913,648
150	0.9997	100	$61,421,882	$167,879,982	$542,712,191	$1,494,602,815
250	0.9997	100	$26,447,945	$102,541,530	$310,519,592	$908,017,304
500	0.9997	100	$9,171,088	$34,425,161	$120,957,673	$343,467,656
1,000	0.9997	100	$5,231,911	$22,379,201	$49,879,949	$170,870,431
	VaR Quantile, alpha	Lambda	**% Overshoot Bias a la Jensen's Inequality**			
150	0.999	25	8.3%	18.2%	19.7%	58.7%
250	0.999	25	3.4%	11.4%	11.4%	28.6%
500	0.999	25	1.1%	3.5%	4.4%	10.5%
1,000	0.999	25	0.6%	2.5%	1.7%	5.5%
150	0.9997	100	11.1%	27.5%	29.9%	73.1%
250	0.9997	100	4.8%	16.8%	17.1%	44.4%
500	0.9997	100	1.7%	5.6%	6.7%	16.8%
1,000	0.9997	100	0.9%	3.7%	2.7%	8.4%

3. Generating capital distributions forces practitioners to face the unpleasant reality that the variance of the capital distribution, regardless of the estimator used, is typically so large for many units of measure that the 95% confidence interval on the mean capital estimate approaches zero for the lower limit, and can be multiples of the mean capital estimate for the upper limit.[45] The tremendous width of such confidence intervals on estimated capital highlight the importance of the third statistical criterion previously mentioned – efficiency (precision), in addition to accuracy (unbiasedness) and robustness, *of the capital distribution*, and not of the distribution of the parameter estimates per se.

[45] Cope et al. (2009) discuss regulatory capital estimates using the Böcker and Klüppelberg (2005) single loss approximation for the case of a LogNormal(10, 2) severity distribution with 10 and 100 losses per year. With 10 losses per year, the capital estimate is 37.4 million with a 95% confidence interval that is 100 million wide. Assuming 100 losses per year, the capital estimate is 111.5 million with a 95% confidence interval that is 355 million wide. Note that this single loss approximation does not include a mean adjustment term.

Alas, if researchers are not focusing on the most important distribution – the capital distribution – when assessing the utility of an estimator, they will not only fail to solve the capital efficiency problem, but also remain completely unaware of the material upward bias on capital that is induced by Jensen's inequality. As an empirical matter, both can be very large issues posing serious problems for capital estimation, and yet the former has been virtually completely ignored to date, while the latter has gone undetected in both the operational risk measurement literature and in its commercial application. Addressing both of these issues must be an important line of additional research if capital estimates in this setting, regardless of the estimator used, are to have any real utility in shaping the strategic choices and business decisions faced by a financial institution.

7.4. Practical Solutions to the Systematic Overstatement of Operational Risk Capital with MLE Severity Modeling

There are several approaches that may be able to address the material upward bias in operational risk capital estimates that is induced by Jensen's inequality. The operational risk practitioner who must regularly generate estimates of regulatory and economic capital may consider the following possibilities:

1. Qualitative adjustment: Reduce the magnitude of other upwards adjustments to capital estimates (based on management judgment or other qualitative post-modeling adjustments) to account for the built-in conservatism of capital estimates resulting from MLE-based severity models. Justifying the ad hoc nature of this approach may be challenging, and it leaves unaddressed the problem of the inefficiency of capital estimates, let alone their non-robusteness.

2. Quantitative adjustment: Apply bias adjustment methodologies from other empirical models to the capital estimation problem. Analytic adjustments are described by a number of authors (see Kennedy, 1981, 1983; Goldberger, 1968; and Miller, 1984). Duan (1983) describes a simulation based approach to bias adjustment for the log transformation. We are not aware of applications of such methods to the problem of bias adjustment in operational risk capital estimation, but research on this application is currently underway by the authors of this paper.

3. Robust estimation: Aside from the OBRE estimator discussed here, we are unaware of other methods that simultaneously address the systematic upward bias related to Jensen's inequality, at least in part, while remaining reasonably efficient and reasonably robust when confronted with non-textbook, non-i.i.d. data, *and* are applicable to truncated severity distributions. The initial results of this study indicate that OBRE as it stands has the potential to accomplish this on severity distributions that are at least medium- to heavy-tailed. However, for very heavy-tailed distributions, the impact of Jensen's inequality in the context of operational risk capital estimation appears to be such that even OBRE requires an explicit, additional adjustment to mitigate it. The form of that adjustment, used simultaneously with OBRE to appropriately achieve robustness, also is

the subject of ongoing research by these authors, and may include one of the above-mentioned methods, or an innovative use of OBRE weights as was done on a preliminary basis in this study.

8. Summary and Future Directions

We have shown analytically, via derivation of the appropriate Influence Functions, that the MLE estimator of severity distribution parameters is i) non-robust; and ii) extremely sensitive not only to large losses, but also to small losses as it exhibits extreme asymptotic and counter-intuitive behavior in the left tail. Dealing with a data collection threshold via truncation only partially mitigates extreme sensitivity to small losses and exacerbates its non-robustness to large losses, as shown by the derived Influence Functions for truncated distributions: truncation either induces or augments correlation of the parameters of the severity distribution (depending on whether it existed before truncation), speeding up their divergence to $-\infty$ or $+\infty$ as the size of new losses increase. The Influence Functions conclusively show that the size of the threshold strongly influences the magnitude of these effects. Finally, the same derivations also definitively explain the sometimes counterintuitive and perplexing behavior of MLE estimators under truncation that has been reported in the simulation results of other authors (for example, see Cope (2011)).

However, these parameter estimates only are used in this setting for the purpose of estimating a high quantile of the severity distribution, which is essentially the capital estimate. So while all of these aforementioned effects are important, and do affect the capital distribution, the distributions of the severity parameter estimates themselves are not the endgame: the capital distribution is, and if the latter is ignored because all the research focus is placed on the properties of the parameter estimates, researchers and practitioners will miss very important issues limiting the current statistical framework's ability to provide good capital estimates. One of these is upward bias in capital estimates due to Jensen's inequality, and another is the extreme inefficiency of the capital distribution, regardless of the severity distribution estimator being used.

In addition to exposing the inability of MLE to provide a capital estimate that is accurate, robust, and efficient, the second objective of this study has been to begin to develop an estimator that possesses the first two of these three characteristics. OBRE appears to be a promising candidate for this, but as it stands it is no more efficient than MLE vis-à-vis the capital distribution. We believe that additional data is required to achieve variance reduction, and thus, efficiency and statistical power, without sacrificing robustness or accuracy. One approach that may be able to accomplish this is OBRE regression. Such a multivariate approach could utilize the additional loss characteristics, at least on internal losses, that currently are going unused in this estimation exercise as covariates in a (robust) regression. The OBRE methodology described above is fully generalizable to the multivariate setting (for example, see Cantoni and Ronchetti (2006), Carroll and Ruppert (1988), Farcomeni (2010), Kunsch, et al. (1989), and Victoria-Feser (2002)), so there is no reason why researchers could not turn to this framework to utilize the additional information on internal loss data as regression covariates. We present a very simplified and stylized example below for purposes of discussion.

Suppose that two units of measure for the Clients, Products, and Business Processes (CPBP) event type have been determined. Assume that they differ primarily in the size of their losses, but not in their severity distribution family: both are roughly LogNormal, but the one with larger losses has a larger μ, while the one with smaller losses has a smaller μ, and assume that both units of measure have approximately the same σ. To make the example concrete, assume that one unit of measure consists of losses in North America, and the other unit of measure includes losses incurred outside North America.[46] These two units of measure could be combined into a single unit of measure, thus doubling degrees of freedom and dramatically increasing estimation precision, if an OBRE regression with a region dummy variable was used. The μ would be allowed to vary by the value of the dummy covariate, but the σ would be restricted: we therefore increase estimation precision without sacrificing the homogeneity of the unit of measure, and neither do we sacrifice the accuracy or robustness of the parameter estimate or the capital estimate.

So the overall goal of this line of research would be to develop an estimation method that provides superior performance on all three characteristics of the capital distribution: more accuracy, more robustness, and more efficiency. We began to address the first two in this study, and OBRE appears to be the most promising candidate moving forward. Further research on this front is focused on finalizing a method of estimating rather than choosing OBRE's tuning parameter and, under very heavy-tailed severity distributions, the most appropriate method for mitigating the biasing effects of Jensen's inequality. Removing any subjectivity from the determination of the value of the tuning parameter obviously is key to OBRE's utilization in this setting. Once this is accomplished, we then hope to implement these methods with an OBRE regression to obtain better efficiency in the capital distribution as well.

9. Conclusion

The final determinants of an estimator's utility for the purpose of estimating the parameters of the severity distribution should be the characteristics of the operational risk estimated capital distribution: its accuracy (unbiasedness), robustness, and precision (efficiency). Characteristics of the distribution of the estimator itself are important, and certainly impact the characteristics of the capital distribution, but they are not the ultimate objective. The operational risk measurement literature to date has failed to focus on the characteristics of capital estimates, and as a consequence, it has missed very fundamental, long-proven analytical results that dramatically affect and limit the ability of the current statistical framework to provide "good" capital estimates. One such example identified and demonstrated herein is the material upward bias caused by Jensen's inequality, which can be extremely large and translate into large amounts of capital that banks are unnecessarily "leaving on the table" if capital estimates are computed with MLE-based severity parameter estimates. This ultimately can distort important business decisions by understating risk adjusted rates of return (RAROC). We believe that part of the reason this has been missed, in fact, is related directly to the failure of researchers and

[46] Cope (2010) documents a loss scaling effect in which CPBP losses are systematically larger in countries that permit class action litigation.

practitioners to focus on the capital distribution. The near exclusive focus on the properties of the estimated severity parameters has left another paramount issue essentially ignored: the extreme inefficiency of the distribution of capital estimates, regardless of the estimator in use. As a consequence, methods that may be able to at least partially address this inefficiency, such as multivariate approaches like OBRE regression, also have been ignored.

The goal of this study was twofold:

1. To definitively document the limitations of MLE for estimating operational risk capital by showing, via both analytic derivation and simulation, that under the actual data conditions encountered in real-world operational losses, MLE severity modeling is neither an accurate, nor a robust, nor a precise estimator of capital requirements; and

2. To begin to develop an alternative estimation approach that is superior in terms of accuracy and robustness.

We believe the OBRE approach shows great promise here, although as described above additional work must be done to develop a fully data-driven estimation process utilizing it. We believe a promising approach that can simultaneously address the need for greater efficiency in the estimation of the capital distribution is an OBRE regression, but we have left this, too, for future study.

Appendix 1. Derivation of MLE Estimators of LogNormal Parameters

Assuming an i.i.d. sample of *n* loss observations $x_1, x_2, \cdots x_n$ from the LogNormal distribution

$$f\left(x;\mu,\sigma\right)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2} \text{ and } F\left(x;\mu,\sigma\right)=\frac{1}{2}\left[1+erf\left(\frac{\ln(x)-\mu}{\sqrt{2\sigma^2}}\right)\right]$$

for $0<x<\infty$ and $0<\sigma<\infty$, (A1)

we maximize the objective function $\hat{l}\left(\theta \mid x\right)=\sum_{i=1}^{n}\ln\left[f\left(x_i \mid \mu,\sigma\right)\right]$ by finding $\hat{\mu}$ such that

$\frac{\partial \hat{l}\left(\theta \mid x\right)}{\partial \mu}=0$ and by finding $\hat{\sigma}$ such that $\frac{\partial \hat{l}\left(\theta \mid x\right)}{\partial \sigma}=0$

$$\hat{l}\left(\theta \mid x\right)=\sum_{i=1}^{n}\ln\left[f\left(x_i \mid \mu,\sigma\right)\right] \quad \text{(A2)}$$

$$=\sum_{i=1}^{n}\ln(1)-\ln\left(\sqrt{2\pi}\sigma x_i\right)-\frac{1}{2}\left(\frac{\ln(x_i)-\mu}{\sigma}\right)^2$$

$$=\sum_{i=1}^{n}-\ln\left(\sqrt{2\pi}\right)-\ln(\sigma)-\ln(x_i)-\frac{\left[\ln(x_i)-\mu\right]^2}{2\sigma^2}$$

$$0=\frac{\partial \hat{l}\left(\theta \mid x\right)}{\partial \mu}=\sum_{i=1}^{n}\frac{\partial}{\partial \mu}\left(-\frac{\left[\ln(x_i)-\mu\right]^2}{2\sigma^2}\right)=\sum_{i=1}^{n}\frac{2\left[\ln(x_i)-\mu\right]}{2\sigma^2}=\sum_{i=1}^{n}\frac{\left[\ln(x_i)-\mu\right]}{\sigma^2}$$

$$0=-\frac{n\mu}{\sigma^2}+\frac{1}{\sigma^2}\sum_{i=1}^{n}\ln(x_i)$$

$$n\mu=\sum_{i=1}^{n}\ln(x_i) \quad \text{so} \quad \hat{\mu}_{MLE}=\frac{\sum_{i=1}^{n}\ln(x_i)}{n}$$

and

$$0=\frac{\partial \hat{l}\left(\theta \mid x\right)}{\partial \sigma}=\sum_{i=1}^{n}\frac{\partial}{\partial \sigma}\left(-\ln(\sigma)-\frac{\left[\ln(x_i)-\mu\right]^2}{2\sigma^2}\right)=\sum_{i=1}^{n}-\frac{1}{\sigma}-\frac{(-2)\left[\ln(x_i)-\mu\right]^2}{2\sigma^3}$$

$$= \sum_{i=1}^{n} \frac{\left[\ln\left(x_i\right) - \mu\right]^2}{\sigma^3} - \frac{n}{\sigma}; \quad \frac{n}{\sigma} = \frac{1}{\sigma^3} \sum_{i=1}^{n} \left[\ln\left(x_i\right) - \mu\right]^2; \quad n\sigma^2 = \sum_{i=1}^{n} \left[\ln\left(x_i\right) - \mu\right]^2$$

$$\hat{\sigma}_{MLE}^2 = \frac{\sum_{i=1}^{n} \left[\ln\left(x_i\right) - \hat{\mu}_{MLE}\right]^2}{n}$$, which is asymptotically unbiased.

When the log-likelihood cannot be simplified to obtain closed-form algebraic solutions as above, numerical methods often can be used to obtain its maximum. For example, for the parameters of the Generalized Pareto Distribution, Grimshaw (1993) used a reparameterization to develop a numerical algorithm that obtains MLE estimates. Similarly, for the LogGamma distribution, Bowman and Shenton (1983, 1988) provide numerical methods to obtain MLE parameter estimates. For heavy-tailed severity distributions used in this setting (and generally), the use of numerical methods to obtain MLE estimates of distributional parameters is the rule rather than the exception, so MLE proponents cannot use this as an objection to other methods of estimation that rely on numerical algorithms for their implementation (like OBRE).

Appendix 2. Influence Function Derivations for MLE Estimators of GPD Parameters

The two-parameter Generalized Pareto Distribution is defined as

$$f\left(x;\varepsilon,\beta\right) = \frac{1}{\beta}\left[1 + \varepsilon\frac{x}{\beta}\right]^{\left[-\frac{1}{\varepsilon}-1\right]} \quad \text{and} \quad F\left(x;\varepsilon,\beta\right) = 1 - \left[1 + \varepsilon\frac{x}{\beta}\right]^{\left[-\frac{1}{\varepsilon}\right]}$$

assuming $\varepsilon \geq 0$, for $0 \leq x < \infty;\ 0 < \beta < \infty$ (A3)

Its first and second order partial derivatives are below.

$$\frac{\partial}{\partial\beta} f\left(x;\beta,\varepsilon\right) = -\frac{1}{\beta}\left[\frac{\beta - x}{\beta + \varepsilon x}\right] f\left(x;\beta,\varepsilon\right) \tag{A4}$$

$$\frac{\partial}{\partial\varepsilon} f\left(x;\beta,\varepsilon\right) = \left[\left(\frac{-x\left(1+\varepsilon\right)}{\beta\varepsilon + \varepsilon^2 x}\right) + \frac{\ln\left(1 + \frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right] f\left(x;\beta,\varepsilon\right) \tag{A5}$$

$$\frac{\partial^2}{\partial\beta^2} f\left(x;\beta,\varepsilon\right) = \left(\left[\frac{1}{\beta^2} - \frac{x\left(1+\varepsilon\right)\left(2\beta + \varepsilon x\right)}{\left(\beta^2 + \beta\varepsilon x\right)^2}\right] + \frac{1}{\beta^2}\left[\frac{\beta - x}{\beta + \varepsilon x}\right]^2\right) f\left(x;\beta,\varepsilon\right) \tag{A6}$$

$$\frac{\partial^2}{\partial\varepsilon^2} f(x;\beta,\varepsilon) = \left(\left[\frac{x\beta + 2\varepsilon x^2 + \varepsilon^2 x^2}{\left(\beta\varepsilon + \varepsilon^2 x\right)^2} + \frac{x}{(\beta+\varepsilon x)\varepsilon^2} - \frac{2\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^3} \right] + \left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon + \varepsilon^2 x} \right) + \frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2} \right]^2 \right) f(x;\beta,\varepsilon) \tag{A7}$$

$$\frac{\partial}{\partial\varepsilon\partial\beta} f(x;\beta,\varepsilon) = \left(\left[-\frac{1}{\beta}\left(\frac{\beta - x}{\beta+\varepsilon x} \right) \right] \left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2 x} \right) + \frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2} \right] + \left[\frac{\varepsilon x(1+\varepsilon)}{\left(\beta\varepsilon+\varepsilon^2 x\right)^2} - \frac{x}{\beta\varepsilon(\beta+\varepsilon x)} \right] \right) f(x;\beta,\varepsilon) \tag{A8}$$

So for $IF_\theta(x;\theta,T) = A(\theta)^{-1}\varphi_\theta = \begin{bmatrix} -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_1} dF(y) & -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_2} dF(y) \\ -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_1} dF(y) & -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_2} dF(y) \end{bmatrix}^{-1} \begin{bmatrix} \varphi_{\theta_1} \\ \varphi_{\theta_2} \end{bmatrix}$, the terms of φ are

$$\varphi_\theta = \begin{bmatrix} \varphi_\beta \\ \varphi_\xi \end{bmatrix} = \begin{bmatrix} \partial\rho(x,\theta)/\partial\beta \\ \partial\rho(x,\theta)/\partial\xi \end{bmatrix} = \begin{bmatrix} -\frac{\partial f(x,\theta)}{\partial\beta} \Big/ f(x,\theta) \\ -\frac{\partial f(x,\theta)}{\partial\xi} \Big/ f(x,\theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{\beta}\left[\frac{\beta-x}{\beta+\varepsilon x} \right] \\ -\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2 x} \right) + \frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2} \right] \end{bmatrix} \tag{A9}$$

and the terms of the Fisher information $A(\theta)$ are

$$-\int_0^\infty \frac{\partial\varphi_\varepsilon}{\partial\varepsilon} dF(x) = -\int_0^\infty \left[\frac{x\beta + 2\varepsilon x^2 + \varepsilon^2 x^2}{\left(\beta\varepsilon+\varepsilon^2 x\right)^2} + \frac{x}{(\beta+\varepsilon x)\varepsilon^2} - \frac{2\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^3} \right] f(x)dx \tag{A10}$$

$$-\int_0^\infty \frac{\partial\varphi_\beta}{\partial\beta} dF(x) = -\int_0^\infty \left[\frac{1}{\beta^2} - \frac{x(1+\varepsilon)(2\beta+\varepsilon x)}{\left(\beta^2+\beta\varepsilon x\right)^2} \right] f(x)dx \tag{A11}$$

$$-\int_0^\infty \frac{\partial\varphi_\varepsilon}{\partial\beta} dF(x) = -\int_0^\infty \frac{\partial\varphi_\beta}{\partial\varepsilon} dF(x) = -\int_0^\infty \left[\frac{x}{\beta\varepsilon(\beta+\varepsilon x)} - \frac{\varepsilon x(1+\varepsilon)}{\left(\beta\varepsilon+\varepsilon^2 x\right)^2} \right] f(x)dx \tag{A12}$$

Note that parameter correlation is indicated by the non-zero off-diagonal, cross-partial derivative terms. The above IF must be solved numerically, and as shown above in Section 4.5, it is validated by a matching EIF (n = 250). However, note that Smith (1987),[47] for the GPD specifically, was able to conveniently simplify the Fisher Information to yield

$$A(\theta)^{-1} = (1+\xi)\begin{bmatrix} 1+\xi & -\beta \\ -\beta & 2\beta^2 \end{bmatrix} \qquad \text{(A13)}$$

(Ruckdeschel and Horbenko (2010) later re-present this result in the operational risk setting). This gives the exact same result, as shown in the graphs in Section 4.5 above, as the numerical implementation of (A12), thus providing further independent validation of the more general framework presented herein (which, of course, can be used with *all* commonly used severity distributions).

Appendix 3. Influence Function Derivations for MLE Estimators of Truncated LogGamma and Truncated GPD Parameters

The first and second order partial derivatives of the cdf of the Truncated LogGamma are listed below (those of the pdf are listed in Section 4.5).

$$\frac{\partial F(H;a,b)}{\partial a} = \int_{1^+}^{H} \left[\ln(b) + \ln(\ln(y)) - digamma(a) \right] f(y;a,b)\,dy \qquad \text{(A14)}$$

$$\frac{\partial F(H;a,b)}{\partial b} = \int_{1^+}^{H} \left[\frac{a}{b} - \ln(y) \right] f(y;a,b)\,dy \qquad \text{(A15)}$$

$$\frac{\partial^2 F(H;a,b)}{\partial a^2} = \int_{1^+}^{H} \left(\left[\ln(b) + \ln(\ln(y)) - digamma(a) \right]^2 - trigamma(a) \right) \cdot f(y;a,b)\,dy \qquad \text{(A16)}$$

$$\frac{\partial^2 F(H;a,b)}{\partial b^2} = \int_{1^+}^{H} \left[\frac{a(a-1)}{b^2} - \frac{2a(\ln(y))}{b} + (\ln(y))^2 \right] \cdot f(y;a,b) \cdot dy \qquad \text{(A17)}$$

$$\frac{\partial F(H;a,b)}{\partial a \partial b} = \int_{1^+}^{H} \left(\frac{1}{b} + \left[\ln(b) + \ln(\ln(y)) - digamma(a) \right] \times \left[\frac{a}{b} - \ln(y) \right] \right) f(y;a,b)\,dy \qquad \text{(A18)}$$

---

47 Smith (1987) was the earliest example of this result that we were able to find in the literature.

So for $IF_\theta(x;\theta,T) = A(\theta)^{-1}\varphi_\theta = \begin{bmatrix} -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_1} dG(y) & -\int_a^b \frac{\partial\varphi_{\theta_1}}{\partial\theta_2} dG(y) \\ -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_1} dG(y) & -\int_a^b \frac{\partial\varphi_{\theta_2}}{\partial\theta_2} dG(y) \end{bmatrix}^{-1} \begin{bmatrix} \varphi_{\theta_1} \\ \varphi_{\theta_2} \end{bmatrix}$, the terms of φ are

$$\varphi_\theta = \begin{bmatrix} \varphi_a \\ \varphi_b \end{bmatrix} = \begin{bmatrix} \partial\rho(x,\theta)/\partial a \\ \partial\rho(x,\theta)/\partial b \end{bmatrix} = \begin{bmatrix} -\frac{\partial f(x,\theta)}{\partial a} \Big/ f(x,\theta) \\ -\frac{\partial f(x,\theta)}{\partial b} \Big/ f(x,\theta) \end{bmatrix} =$$

$$= \begin{bmatrix} -\left[\ln(b) + \ln(\ln(y)) - digamma(a)\right] - \dfrac{\int_1^H \left[\ln(b) + \ln(\ln(y)) - digamma(a)\right] f(y;a,b)dy}{1-F(H;\mu,\sigma)} \\ -\left[\frac{a}{b} - \ln(y)\right] - \dfrac{\int_1^H \left[\frac{a}{b} - \ln(y)\right] f(y;a,b)dy}{1-F(H;\mu,\sigma)} \end{bmatrix} \quad \text{(A19)}$$

(Note that the limits of integration in (A19), *a* and *b*, are not the parameters of the LogGamma – they just coincidentally share the same letters).

and the terms of the Fisher information $A(\theta)$ are

$$-\int_H^\infty \frac{\partial\varphi_a}{\partial a} dG(x) = -trigamma(a) + \frac{\left[\int_{1^+}^H \ln(b) + \ln(\ln(x)) - digamma(a) f(x)dx\right]^2}{\left[1-F(H;a,b)\right]^2} + $$
$$+ \frac{\left[1-F(H;a,b)\right] \cdot \int_{1^+}^H \left[\ln(b) + \ln(\ln(x)) - digamma(a)\right]^2 - trigamma(a) f(x)dx}{\left[1-F(H;a,b)\right]^2}$$

(A20)

$$-\int_H^\infty \frac{\partial\varphi_b}{\partial b} dG(x) = -\frac{a}{b^2} + \frac{\left[\int_{1^+}^H \left(\frac{a}{b} - \ln(y)\right) f(x)dx\right]^2 + \left[1-F(H;a,b)\right] \cdot \int_{1^+}^H \frac{a(a-1)}{b^2} - \frac{2a\ln(y)}{b} + \left[\ln(y)\right]^2 f(x)dx}{\left[1-F(H;a,b)\right]^2}$$

(A21)

$$-\int_{H}^{\infty}\frac{\partial\varphi_a}{\partial b}dG(x)=-\int_{H}^{\infty}\frac{\partial\varphi_b}{\partial a}dG(x)=\frac{1}{b}+\frac{\left[1-F(H;a,b)\right]\cdot\frac{1}{b}\cdot F(H;a,b)}{\left[1-F(H;a,b)\right]^2}+$$

$$+\frac{\left[1-F(H;a,b)\right]\cdot\int_{1^+}^{H}\left[\ln(b)+\ln(\ln(x))-digamma(a)\right]\cdot\left[\frac{a}{b}-\ln(x)\right]f(x)dx}{\left[1-F(H;a,b)\right]^2}$$

$$+\frac{\int_{1^+}^{H}\ln(b)+\ln(\ln(x))-digamma(a)\,f(x)dx\cdot\int_{1^+}^{H}\left(\frac{a}{b}-\ln(x)\right)f(x)dx}{\left[1-F(H;a,b)\right]^2}$$

(A22)

The above IF must be solved numerically. Note that parameter correlation is indicated by non-zero off-diagonal terms.

The first and second order partial derivatives of the cdf of the Truncated GPD are listed below (those of the pdf are listed in Appendix 2).

$$\frac{\partial F(H;\beta,\varepsilon)}{\partial\beta}=\int_{0}^{H}-\frac{1}{\beta}\left[\frac{\beta-x}{\beta+\varepsilon x}\right]f(x;\beta,\varepsilon)dx \quad \text{(A23)}$$

$$\frac{\partial F(H;\beta,\varepsilon)}{\partial\varepsilon}=\int_{0}^{H}\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]f(x;\beta,\varepsilon)dx \quad \text{(A24)}$$

$$\frac{\partial^2 F(H;\beta,\varepsilon)}{\partial\beta^2}=\int_{0}^{H}\left(\left[\frac{1}{\beta^2}-\frac{x(1+\varepsilon)(2\beta+\varepsilon x)}{\left(\beta^2+\beta\varepsilon x\right)^2}\right]+\frac{1}{\beta^2}\left[\frac{\beta-x}{\beta+\varepsilon x}\right]^2\right)f(x;\beta,\varepsilon)dx \quad \text{(A25)}$$

$$\frac{\partial^2 F(H;\beta,\varepsilon)}{\partial\varepsilon^2}=\int_{0}^{H}\left(\left[\frac{x\beta+2\varepsilon x^2+\varepsilon^2x^2}{\left(\beta\varepsilon+\varepsilon^2x\right)^2}+\frac{x}{(\beta+\varepsilon x)\varepsilon^2}-\frac{2\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^3}\right]+\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]^2\right)f(x;\beta,\varepsilon)dx$$

(A26)

$$\frac{\partial F(H;\beta,\varepsilon)}{\partial\varepsilon\partial\beta}=\int_0^H\left(\left[-\frac{1}{\beta}\left(\frac{\beta-x}{\beta+\varepsilon x}\right)\right]\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]+\left[\frac{\varepsilon x(1+\varepsilon)}{\left(\beta\varepsilon+\varepsilon^2x\right)^2}-\frac{x}{\beta\varepsilon(\beta+\varepsilon x)}\right]\right)f(x;\beta,\varepsilon)dx \quad \text{(A27)}$$

So for $IF_\theta(x;\theta,T)=A(\theta)^{-1}\varphi_\theta=\begin{bmatrix}-\int_a^b\frac{\partial\varphi_{\theta_1}}{\partial\theta_1}dG(y) & -\int_a^b\frac{\partial\varphi_{\theta_1}}{\partial\theta_2}dG(y)\\ -\int_a^b\frac{\partial\varphi_{\theta_2}}{\partial\theta_1}dG(y) & -\int_a^b\frac{\partial\varphi_{\theta_2}}{\partial\theta_2}dG(y)\end{bmatrix}^{-1}\begin{bmatrix}\varphi_{\theta_1}\\ \varphi_{\theta_2}\end{bmatrix}$, terms of φ are

$$\varphi_\theta=\begin{bmatrix}\varphi_\beta\\ \varphi_\xi\end{bmatrix}=\begin{bmatrix}\partial\rho(x,\theta)/\partial\beta\\ \partial\rho(x,\theta)/\partial\xi\end{bmatrix}=\begin{bmatrix}-\frac{\partial f(x,\theta)}{\partial\beta}\Big/f(x,\theta)\\ -\frac{\partial f(x,\theta)}{\partial\xi}\Big/f(x,\theta)\end{bmatrix}=$$

$$=\begin{bmatrix}-\left[-\frac{1}{\beta}\left[\frac{\beta-x}{\beta+\varepsilon x}\right]\right]-\frac{\int_0^H-\frac{1}{\beta}\left[\frac{\beta-x}{\beta+\varepsilon x}\right]f(x;\beta,\varepsilon)dx}{1-F(H;\mu,\sigma)}\\ -\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]-\frac{\int_0^H\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]f(x;\beta,\varepsilon)dx}{1-F(H;\mu,\sigma)}\end{bmatrix} \quad \text{(A28)}$$

and the terms of the Fisher information $A(\theta)$ are

$$-\int_0^\infty\frac{\partial\varphi_\varepsilon}{\partial\varepsilon}dG(x)=-\frac{1}{\left[1-F(H;\beta,\varepsilon)\right]}\cdot\int_H^\infty\left[\frac{x\beta+2\varepsilon x^2+\varepsilon^2x^2}{\left(\beta\varepsilon+\varepsilon^2x\right)^2}+\frac{x}{(\beta+\varepsilon x)\varepsilon^2}-\frac{2\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^3}\right]f(x)dx$$

$$+\frac{\left(\int_0^H\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]f(x;\beta,\varepsilon)dx\right)^2}{\left[1-F(H;\beta,\varepsilon)\right]^2}+$$

$$+\frac{\left[1-F(H;\beta,\varepsilon)\right]\cdot\int_0^H\left(\left[\frac{x\beta+2\varepsilon x^2+\varepsilon^2x^2}{\left(\beta\varepsilon+\varepsilon^2x\right)^2}+\frac{x}{(\beta+\varepsilon x)\varepsilon^2}-\frac{2\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^3}\right]+\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]^2\right)f(x;\beta,\varepsilon)dx}{\left[1-F(H;\beta,\varepsilon)\right]^2}$$

(A29)

$$-\int_0^\infty\frac{\partial\varphi_\beta}{\partial\beta}dG(x)=-\frac{1}{\left[1-F(H;\beta,\varepsilon)\right]}\cdot\int_H^\infty\left[\frac{1}{\beta^2}-\frac{x(1+\varepsilon)(2\beta+\varepsilon x)}{\left(\beta^2+\beta\varepsilon x\right)^2}\right]f(x)dx$$

$$+\frac{\left(\int_0^H-\frac{1}{\beta}\left[\frac{\beta-x}{\beta+\varepsilon x}\right]f(x;\beta,\varepsilon)dx\right)^2+\left[1-F(H;\beta,\varepsilon)\right]\cdot\int_0^H\left(\left[\frac{1}{\beta^2}-\frac{x(1+\varepsilon)(2\beta+\varepsilon x)}{\left(\beta^2+\beta\varepsilon x\right)^2}\right]+\frac{1}{\beta^2}\left[\frac{\beta-x}{\beta+\varepsilon x}\right]^2\right)f(x;\beta,\varepsilon)dx}{\left[1-F(H;\beta,\varepsilon)\right]^2}$$

(A30)

$$-\int_0^\infty\frac{\partial\varphi_\varepsilon}{\partial\beta}dG(x)=-\int_0^\infty\frac{\partial\varphi_\beta}{\partial\varepsilon}dG(x)=-\frac{1}{\left[1-F(H;\beta,\varepsilon)\right]}\cdot\int_H^\infty\left[\frac{x}{\beta\varepsilon(\beta+\varepsilon x)}-\frac{\varepsilon x(1+\varepsilon)}{\left(\beta\varepsilon+\varepsilon^2x\right)^2}\right]f(x)dx$$

$$+\frac{\left(\int_0^H\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]f(x;\beta,\varepsilon)dx\right)\times\left(\int_0^H-\frac{1}{\beta}\left[\frac{\beta-x}{\beta+\varepsilon x}\right]f(x;\beta,\varepsilon)dx\right)}{\left[1-F(H;\beta,\varepsilon)\right]^2}$$

$$+\frac{\left[1-F(H;\beta,\varepsilon)\right]\cdot\int_0^H\left(\left[\frac{x\beta+2\varepsilon x^2+\varepsilon^2x^2}{\left(\beta\varepsilon+\varepsilon^2x\right)^2}+\frac{x}{(\beta+\varepsilon x)\varepsilon^2}-\frac{2\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^3}\right]+\left[\left(\frac{-x(1+\varepsilon)}{\beta\varepsilon+\varepsilon^2x}\right)+\frac{\ln\left(1+\frac{\varepsilon x}{\beta}\right)}{\varepsilon^2}\right]^2\right)f(x;\beta,\varepsilon)dx}{\left[1-F(H;\beta,\varepsilon)\right]^2}$$

(A31)

The above IF must be solved numerically. Note that parameter correlation is indicated by non-zero off-diagonal terms.

References

Alaiz, P., and Victoria-Feser, M. (1996), "Modelling Income Distribution in Spain: A Robust Parametric Approach," DARP Discussion Paper, No. 20 (PDF), STICERD, LSE. Available online at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1094765.

Basel Committee on Banking Supervision. (2006), "International convergence of capital measurement and capital standards," Bank of International Settlements. Available online at http://www.bis.org/publ/bcbs128.htm.

Basel Committee on Banking Supervision. (2009), "Observed range of practice in key elements of Advanced Measurement Approaches (AMA)," Bank of International Settlements. Available online at http://www.bis.org/publbcbs160.htm.

Basel Committee on Banking Supervision. (2011), "Operational Risk – Supervisory Guidelines for the Advanced Measurement Approaches," Bank of International Settlements. Available online at http://www.bis.org/publ/bcbs196.htm.

Böcker, K., and Klüppelberg, C. (2005), "Operational VaR: A Closed-Form Approximation," *RISK Magazine*, 12, 90-93.

Böcker, K., and Sprittulla, J. (2006), "Operational VaR: meaningful means." *RISK Magazine*, 19 (12), 96–98.

Bowman, K.O., and Shenton, L.R. (1983), "Maximum Likelihood Estimators for the Gamma Distribution Revisited," *Communications in Statistics-Simulation and Computation*, 12, 697-710.

Bowman, K.O., and Shenton, L.R. (1988), Properties of Estimators for the Gamma Distribution, Marcel Dekker, Inc., New York, New York.

Brazauskas, V. and Kleefeld, A. (2009). "Robust and efficient fitting of the generalized Pareto distribution with actuarial applications in view". *Insurance: Mathematics and Economics*, 45 (3), 424-435.

Brazauskas, V., and Serfling, R. (2003). "Favorable estimators for fitting Pareto models: a study using goodness-of-fit measures with actual data," *ASTIN Bulletin*, 33 (2), 365–381.

Brazzale, A., and Ventura, L. (2002). "Hampel estimation for extreme value distributions," Proceedings of the XLI Scientific Meeting of the SIS, Milan, 5-7 June 2002, 613-616.

Brazauskas, V., Jones, B., and Zitikis, R. (2009). "Robust fitting of claim severity distributions and the method of trimmed moments". *Journal of Statistical Planning and Inference*, 139 (6), 2028--2043.

Cantoni, E., and Ronchetti, E. (2006). "A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures," *Journal of Health Economics*, 25, 198-213.

Carroll, R.J., and Ruppert, D. (1988), Transformation and Weighting in Regression, Chapman Hall, New York, New York.

Chapelle, A., Crama, Y., Hübner, G., and Peters, J. (2008), "Practical methods for measuring and managing operational risk in the financial sector: A clinical study," *Journal of Banking & Finance*, 32 (6), 1049-1061.

Chernobai, A., and Rachev, S. (2006), "Applying robust methods to operational risk modelling." *The Journal of Operational Risk*, 1 (1), 27–41.

Chernobai, A., Rachev, S., and Fabozzi, F. (2007), Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis, John Wiley and Sons, Hoboken, New Jersey.

Cope, E. (2010), "Modeling operational loss severity distributions from consortium data," *The Journal of Operational Risk*, 5 (4), 35-64.

Cope, E. (2011), "Penalized likelihood estimators for truncated data," *Journal of Statistical Planning and Inference*, 141 (1), 345-358.

Cope, E., and Labbi, A. (2008), "Operational risk scaling by exposure indicators: evidence from the ORX database," *The Journal of Operational Risk*, 3 (4), 25–46.

Cope, E., Mignola, G., Antonini, G., Ugoccioni, R. (2009), "Challenges and pitfalls in measuring operational risk from loss data," *The Journal of Operational Risk*, 4 (4), 3–27.

Cope, E., Piche, M., and Walter, J. (2012), "Macroenvironmental determinants of operational loss severity," *Journal of Banking and Finance*, Available online at http://dx.doi.org/10.1016/j.jbankfin.2011.11.022.

Daníelson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O. and Shin, H.S. (2001), "An Academic Response to Basel II," LSE Financial Markets Group, Special Paper No 130. Available online at http://www.bis.org/bcbs/ca/fmg.pdf.

de Fontnouvelle, P., DeJesus-Rueff, V., Jordan, J., and Rosengren, E. (2003), "Capital and Risk: New Evidence on Implications of Large Operational Losses," Working Paper 03-5, Federal Reserve Bank of Boston. Available online at http://www.bos.frb.org/economic/wp/wp2003/wp035.htm.

de Fontnouvelle, P., Jordan, J., and Rosengren, E. (2006), "Implications of Alternative Operational Risk Modelling Techniques," in M. Carey and R. Stulz, eds., The Risks of Financial Institutions, NBER/University of Chicago Press.

Dell' Aquila, R., and Embrechts, P. (2006), "Extremes and Robustness: A Contradiction?" *Financial Markets and Portfolio Management*, Vol. 20, 103-118.

Duan, N., (1983), "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association*, 78, 605-610..

Duchesne, T., Rioux, J., and Luong, A. (1997), "Minimum Cramér-von Mises Estimators and Their Influence Function," *Actuarial Research Clearing House*, Vol. 1, 349-361.

Dupuis, D. J. (1998), "Exceedances Over High Thresholds: A Guide to Threshold Selection," *Extremes*, 1 (3), 251-261.

Dupuis, D. J., and Field, C. A. (1998). "Robust estimation of extremes," *Canadian Journal of Statistics*, 26 (2), 199–215.

Dupuis, D.J., and Mills, J. (1988), "Robust estimation of the Birnbaum-Saunders distribution," *IEEE Transactions on Reliability*, 47 (1), 88-95.

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

Farcomeni, A., (2010), "An Overview of Robust Methods in Medical Research," *Statistical Methods in Medical Research*, October 25, 2010.

Flanders, H., (1973), "Differentiation Under the Integral Sign," *The American Mathematical Monthly*, 80(6), 615-627.

Frachot, A., and Roncalli, T. (2002), "Mixing internal and external data for managing operational risk," Working Paper, Groupe de Recherche Opérationelle, Crédit Lyonnais. Available online at http://www.thierry-roncalli.com/download/mixing-riskop.pdf.

Frachot, A., Moudoulaud, O., Roncalli, T. (2004), "Loss Distribution Approach in practice," In Ong, M. (Ed.), The Basel Handbook: A Guide for Financial Practitioners, Risk Books, London.

Goldberger, A.S., (1968), "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica*, 35, 464-472.

Greene, W. (2007), Econometric Analysis, 6th Edition. Prentice Hall, Upper Saddle River, NJ.

Grimshaw, S. (1993), "Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution," *Technometrics*, 35 (2), 185-191.

Hampel, F.R. (1968), "Contributions to the Theory of Robust Estimation," Ph.D. thesis, University of California, Berkeley.

Hampel, F.R., E. Ronchetti, P. Rousseeuw, and W. Stahel. (1986), Robust Statistics: The Approach Based on Influence Functions, John Wiley and Sons, New York.

Hong, C. and Kim, Y. (2001), "Automatic Selection of the Tuning Parameter in the Minimum Density Power Divergence Estimation," *Journal of the Korean Statistical Society*, 30 (3), 453-465.

Horbenko, N., Ruckdeschel, P. and Bae, T. (2011), "Robust Estimation of Operational Risk," *The Journal of Operational Risk*, 6 (2), 3-30.

Huber, P.J. (1981), Robust Statistics, John Wiley & Sons, New York.

Huber, P.J. (1977), *Robust Statistical Procedures*, Regional Conference Series in Applied Mathematics, 27, SIAM, Philadelphia, PA.

Huber, P.J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73-101.

Huber, P.J., and Ronchetti, E. (2009), Robust Statistics, 2nd Edition, John Wiley & Sons, New York.

Jensen, J. L. W. V. (1906), "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, 30 (1), 175–193.

Juárez, S. and Schucany, W. (2004). "Robust and efficient estimation of the generalized Pareto distribution," *Extremes*, 7 (3), 237–251.

Kennedy, P. (1992), A Guide to Econometrics, 3rd Edition, Wiley-Blackwell, Hoboken, NJ.

Kennedy, P. (2008), A Guide to Econometrics, 4th Edition, Wiley-Blackwell, Hoboken, NJ.

Kennedy, P., (1981), "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations," *American Economic Review*, 71, 802.

Kennedy, P. (1983), "Logarithmic Dependent Variables and Prediction Bias," *Oxford Bulletin of Economics and Statistics*, 45, 389-392.

Kunsch, H.R., Stefanski, L.A., and Carroll, R.J., (1989) "Conditionally Unbiased Bounded Influence Estimation in General Regression Models with Applications to Generalized Linear Models," *Journal of the American Statistical Association*, 84, 460-466.

Goldberger, A.S., (1968), "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica*, 35, 464-472.

Kunsch, H.R., Stefanski, L.A., and Carroll, R.J., (1989) "Conditionally Unbiased Bounded Influence Estimation in General Regression Models with Applications to Generalized Linear Models," *Journal of the American Statistical Association*, 84, 460-466.

Marazzi, A.; Paccaud, F,; Ruffieux, C.; Beguin, C., (1998), "Fitting the distributions of length of stay by parametric models," *Medical Care*, 36 (6), 915-927.

Mignola, G., and Ugoccioni, R. (2006), "Effect of a data collection threshold in the loss distribution approach," *The Journal of Operational Risk*, 4 (1), 35–47.

Miller, D.M., (1984), "Reducing Transformation Bias in Curve Fitting," *The American Statistician*, 38, 124-126.

Moscadelli, M. (2004), “The modelling of operational risk: experience with the analysis of the data collected by the Basel Committee,” Working Paper TD 517, Bank of Italy, Economic Research Department.

Opdyke, J.D., and Cavallo, A. (2012), “Operational Risk Capital Estimation and Planning: Exact Sensitivity Analyses and Business Decision Making Using the Influence Function,” forthcoming in Operational Risk: New Frontiers Explored, Davis, E., ed., Risk Books, London.

Parr, W., and Schucany, W. (1980), “Minimum Distance and Robust Estimation,” *Journal of the American Statistical Association*, 75 (371), 616-624.

Peng, L., and Welsh, A. (2001) “Robust estimation of the generalized Pareto distribution,” *Extremes*, 4 (1), 53–65.

Ruckdeschel, P., and Horbenko, N. (2010), “Robustness properties of estimators in generalized Pareto models,” Technical Report ITWM 182. Available online at http://www.itwm.fraunhofer.de/presse/berichte-des-itwm.html,

Smith, J. (1987), “Estimating the Upper Tail of Flood Frequency Distributions,” *Water Resources Research*, Vol.23, No.8, 1657-1666.

Stefanski, L., and Boos, D. (2002), “The Calculus of M-Estimation,” *The American Statistician*, 56 (1), 29-38.

Tukey, J.W. (1960), “A Survey of Sampling from Contaminated Distributions,” in I. Olkin, ed., Contributions to Probability and Statistics, Stanford University Press, Stanford, CA., 448-485.

Victoria-Feser, M. (2002), “Robust Inference with Binary Data,” *Psychometrica*, 67(1), 21-32.

Warwick, J. (2004), “A Data-based Method for Selecting Tuning Parameters in Minimum Distance Estimators,” *Computational Statistics & Data Analysis*, 48 (3), 571-585.

Warwick, J., and Jones, M. (2004), “Choosing a Robustness Tuning Parameter,” *Journal of Statistical Computation and Simulation*, 75 (7), 581-588.

Wu, Y., Zhang, H., and Liu, Y. (2010), “Robust Model-Free Multiclass Probability Estimation,” *Journal of the American Statistical Association*, 105 (489).

Wei, R. (2007). “Quantification of operational losses using firm-specific information and external database,”. *The Journal of Operational Risk*, 1 (4), 3–34.

[US Final Rule (2007)] Federal Register, “Risk-Based Capital Standards: Advanced Capital Adequacy Framework—Basel II; Final Rule,” *Federal Register*, 72 (2), 69288-69445.