Robust Statistics vs. MLE for OpRisk Severity Distribution Parameter Estimation

John Douglas (J.D.) Opdyke*, President
DataMineIt, JDOpdyke@DataMineIt.com

Presented at American Bankers Association Operational Risk Modeling Forum, August 10-11, 2011

Discussant: Bakhodir A. Ergashev, Ph.D.
Lead Financial Economist, Federal Reserve Bank of Richmond

***The views presented herein are the views of the sole author, J.D. Opdyke, and do not necessarily reflect the views of other conference participants or discussants. All derivations, and all calculations and computations, were performed by J.D. Opdyke using SAS®: any errors are my own.**

American Bankers Association

DataMineIt℠

Contents

1. The OpRisk Setting and the Specific Objective

Operational Risk

- **Basel II/III**
  - **Advanced Measurement Approach**
    - **Risk Measurement & Capital Quantification**
      - **Loss Distribution Approach**
        - **Frequency Distribution**
        - **Severity Distribution*** **(arguably the main driver of the aggregate loss distribution)**

Specific Objective:

Select / develop a method to estimate the parameters of the severity distribution based on the following criteria – unbiasedness, (relative) efficiency, and robustness – with an emphasis on how these affect (right) tail-fit.**

* Dependence between the frequency and serverity distributions under some circumstances is addressed later in the presentation.

** Technically, the term "efficient" can refer to an estimator that achieves the Cramér-Rao lower bound. Hereafter in this presentation, the terms "efficient" and "efficiency" are used in a relative sense, as in having a lower mean squared error relative to that of another estimator. See Appendix I.

2. MLE vs. Robust Statistics: Point-Counterpoint

Maximum Likelihood Estimation (MLE):

"MLE does not inappropriately downweight extreme observations as do most/all robust statistics. And focus on extreme observations is the entire point of the OpRisk statistical modeling exercise! Why should we even partially ignore the (right) tail when that is where and how capital requirements are determined?! That's essentially ignoring data – the most important data – just because its hard to model!"

Robust Statistics:

"All statistical models are merely idealized approximations of reality, and OpRisk data clearly violate the fragile, textbook model assumptions required by MLE. Robust Statistics acknowledge and deal with these facts by explicitly and systematically accounting for them, sometimes with weights (and thus, they avoid a bias towards weight=one for every data point). Consequently, under real-world, non-textbook OpRisk loss data, Robust Statistics exhibit less bias, equal or greater efficiency, and far more robustness than does MLE. These characteristics translate into a more reliable, stable estimation approach, regardless of the framework used by robust statistics (i.e. multivariate regression or otherwise) to obtain high quantile estimates of the severity distribution.

…to be revisited

2. MLE vs. Robust Statistics: Point-Counterpoint

- Due to the nature of estimating the far right tail of the OpRisk loss event distribution, some type of parametric statistical estimation is required.
- OpRisk data poses many serious challenges for such a statistical estimation, as described on slides 7-8.
- The validity of MLE, the "classical" approach, relies on assumptions clearly violated by the data.
- The main point of this presentation is to address the issue of whether these violations are material: whether MLE is robust enough to the aforementioned violations, or whether it loses its otherwise good statistical properties in this setting, making it unreliable for OpRisk severity distribution parameter estimation. To determine this, analytic results are derived (simulations are merely confirmatory) borrowing from the toolkit of robust statistics, which are examined as possible alternatives to MLE should the objections against it have merit.

2. MLE vs. Robust Statistics: Point-Counterpoint

Some Specific Questions to be Answered:

- **Does MLE become unusable under relatively modest deviations from i.i.d., especially for the heavy-tailed distributions used in this setting, or are these claims overblown?**
- **Do analytical derivations of the MLE Influence Functions for severity distribution parameters support or contradict such claims? Are they consistent with simulation results? How does (possible) parameter dependence affect these results?**
- **Do these results hold under truncation? How much does truncation and the size of the collection threshold affect both MLE and Robust Statistics parameter estimates?**
- **Are widely used, well established Robust Statistics viable for severity distribution parameter estimation? Are they too inefficient relative to MLE for practical use? Do any implementation constraints (e.g. algorithmic issues) trip them up, especially under difficult-to-fit distributions (say, with infinite mean)?**

3. OpRisk Empirical Challenges

The following characteristics of most Operational Risk loss event data make estimating severity distribution parameters very challenging, and are the source of the MLE vs. Alternatives debate:

1. **Relatively few actual data points on loss events**
2. **Extremely few actual data points on low frequency, high severity losses**
3. **The heavy-tailed nature of most relevant severity distributions**
4. **Heterogeneity, even within well-defined units of measure**
5. **The (left) truncated nature of most loss event data (since smaller losses below a threshold typically are ignored)**
6. **The changing nature, from quarter to quarter, of some of the data already in hand (e.g. financial restatements, dispute resolutions, etc.)**
7. **The real potential for a large quarter of new data to non-trivially change the severity distribution**
8. **The real potential for notable heterogeneity in the form of true, robustly defined statistical outliers (not just extreme events)**
9. **The ultimate need to estimate an extremely high quantile of the severity distribution**

- **Moreover, the *combined* effect of 1-9 increases estimation difficulty far more than the sum of the individual challenges (see Cope et al., 2009).**
- **Bottom line: OpRisk loss data is most certainly not independent and identically distributed ("i.i.d."), which is a presumption of MLE; and even if it was close, from an estimation standpoint the above characteristics greatly magnify the effects of even small departures from i.i.d.**

3. OpRisk Empirical Challenges

The practical consequences of 1-9 above for OpRisk modeling can include:

A. **Unusably large variances on the parameter estimates**

B. **Extreme sensitivity in parameter values to data changes (i.e. financial restatements, dispute resolutions, etc.) and/or new and different quarters of loss data. This would translate into a lack of stability and reliability in capital estimates from quarter to quarter.**

C. **Unreasonable sensitivity of parameter estimates to very large losses**

D. **Unreasonable sensitivity of parameter estimates to very small losses (this counter-intuitive result is documented analytically below)**

E. **Due to any of A-D, unusably large variance on estimated severity distribution (high) quantiles**

F. **Due to any of A-E, unusably large variance on capital estimates**

G. **A theoretical loss distribution that does not sync well with the empirical loss distribution: the quantiles of each simply do not match well. This would not bode well for future estimations from quarter to quarter even if key tail quantiles in the current estimation are reasonably close.**

- **So in the OpRisk setting, when estimating severity distribution parameters, the statistical criteria of unbiasedness, efficiency, and robustness are critical and directly determine the degree to which capital estimates from quarter to quarter are stable, reliable, precise, and robust.**
- **A quantitative definition of statistical "robustness" (more precisely, "B-robustness") is provided in the next several slides, after a brief definition of maximum likelihood estimation (MLE).**

4. Maximum Likelihood Estimation (MLE)

- **Maximum Likelihood Estimation (MLE) is considered a "classical" approach to parameter estimation.**
- **MLE parameter estimates are the values that maximize the likelihood, under the assumed model, of observing the data sample at hand.**
- **When the assumed model is in fact the true generator of the data, and those data are independent and identically distributed ("i.i.d."), MLE estimates are asymptotically unbiased ("consistent"), asymptotically normally distributed, and asymptotically efficient (i.e. they achieve the Cramér-Rao lower bound – see Appendix I).**
- **MLE values are obtained in practice by maximizing the log-likelihood function.**
- **As an example, derivations of MLE estimates of the parameters of the LogNormal distribution are shown below.**

4. Maximum Likelihood Estimation (MLE)

For example, assuming an i.i.d. sample of n observations $x_1, x_2, \cdots x_n$ from the LogNormal distribution

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

- **The likelihood function =** $L\left(\mu,\sigma \mid x\right) = \prod_{i=1}^{n} f\left(x_i \mid \mu,\sigma\right)$
- **The log-likelihood function =** $\hat{l}\left(\theta \mid x_1, x_2, \cdots x_n\right) = \ln\left[L\left(\mu,\sigma \mid x\right)\right] = \sum_{i=1}^{n} \ln\left[f\left(x_i \mid \mu,\sigma\right)\right]$
- **Then** $\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta}\left[\hat{l}\left(\theta \mid x_1, x_2, \cdots x_n\right)\right]$
- **So simply 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 finding $\hat{\sigma}$ such that** $\frac{\partial \hat{l}\left(\theta \mid x\right)}{\partial \sigma} = 0$

4. Maximum Likelihood Estimation (MLE)

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

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

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

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

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

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

4. Maximum Likelihood Estimation (MLE)

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

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

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

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

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

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

4. Maximum Likelihood Estimation (MLE)

- **When the log-likelihood cannot be simplified algebraically, numerical methods often can be used to obtain its maximum. For example, for the parameters of the Generalized Pareto Distribution (GDP), Grimshaw (1993) used a reparameterization to develop a numerical algorithm that obtains MLE estimates.**

5a. Robust Statistics: Background and the IF

- The theory behind Robust Statistics is well developed and has been in use for nearly half a century (see Huber, 1964). Textbooks have institutionalized this sub-field of statistics for the past 30 years (see Huber, 1981, and Hampel et al., 1986).
- Robust Statistics is a general approach to estimation that recognizes all statistical models are merely idealized approximations of reality. Consequently, one of its main objectives is bounding the influence on the estimates of a small to moderate number of data points in the sample that deviate from the assumed statistical model.
- Why? So that in practice, when 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."), estimates generated by robust statistics do not "breakdown" and provide meaningless, or at least notably biased and inaccurate, values: their values remain "robust" to such violations.
- Based on the empirical challenges of modeling OpRisk loss data (which is most certainly not "i.i.d.") satisfying this robustness objective would appear to be central to the OpRisk severity distribution parameter estimation effort: robust statistics may be tailor-made for this problem!
- The tradeoff for obtaining robustness, however, is a loss of efficiency – a larger mean squared error (MSE – see Appendix I) – when the idealized model assumptions are true: if model assumptions are violated, robust statistics can be MORE efficient than MLE.

5a. Robust Statistics: Background and the IF

The Influence Function (IF)

- **Perhaps the most useful analytical tool for assessing whether, and the degree to which, a statistic is "robust" in the sense that it bounds or limits the influence of arbitrary deviations* from the assumed model is the Influence Function (IF), defined below:**

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

where

- F **is the distribution that is the assumed source of the data sample**
- T **is a statistical functional, that is, a statistic defined by the distribution that is the (assumed) source of the data sample. For example, the statistical functional for the mean is** $T\left(F\right) = \int u dF\left(u\right) = \int u f\left(u\right) du$
- x **is a particular point of evaluation, and the points being evaulated are those that deviate from the assumed** F**.**
- δ_x **is the probability measure that puts mass 1 at the point** x**.**

*** The terms "arbitrary deviation" and "contamination" or "statistical contamination" are used synonymously to mean data points that come from a distribution other than that assumed by the statistical model. They are not necessarily related to issues of data quality per se.**

5a. Robust Statistics: Background and the IF

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

- F_ε is simply the distribution that includes some proportion of the data, ε, that is an arbitrary deviation away from the assumed distribution, F. So the Influence Function is simply the difference between the value of the statistical functional INCLUDING this arbitrary deviation in the data, vs. EXCLUDING the arbitrary deviation (the difference is then scaled by ε).
- So the IF is defined by three things: an estimator T, an assumed distribution/model F, and a deviation from this distribution, χ (χ obviously can represent more than one data point as ε is a proportion of the data sample, but it is easier conceptually to view χ as a single data point whereby $\varepsilon = 1/n$: this is, in fact, the Empirical Influence Function (EIF) – see Appendix III).
- Simply put, the IF shows how, in the limit (asymptotically as $\varepsilon \to 0$, so as $n \to \infty$), an estimator's value changes as a function of χ, the value of arbitrary deviations away from the assumed statistical model, F. In other words, the IF is the functional derivative of the estimator with respect to the distribution.

5a. Robust Statistics: Background and the IF

- **IF is a special case of the Gâteaux derivative, but its existence requires even weaker conditions (see Hampel et al., 1986, and Huber, 1977), so its use is valid under a very wide range of application (including the relevant OpRisk severity distributions).**

5a. Robust Statistics: Background and the IF

B-Robustness as Bounded IF

- If IF is bounded as x becomes arbitrarily large/small, the estimator is said to be "B-robust"*; if IF is not bounded and the estimator's values become arbitrarily large as deviations from the model become arbitrarily large/small, the estimator is NOT B-robust.
- The Gross Error Sensitivity (GES) measures the worst case (approximate) influence that an arbitrary deviation can have on the value of an estimator. If GES is finite, an estimator is B-robust; if it is infinite, it is not B-robust.

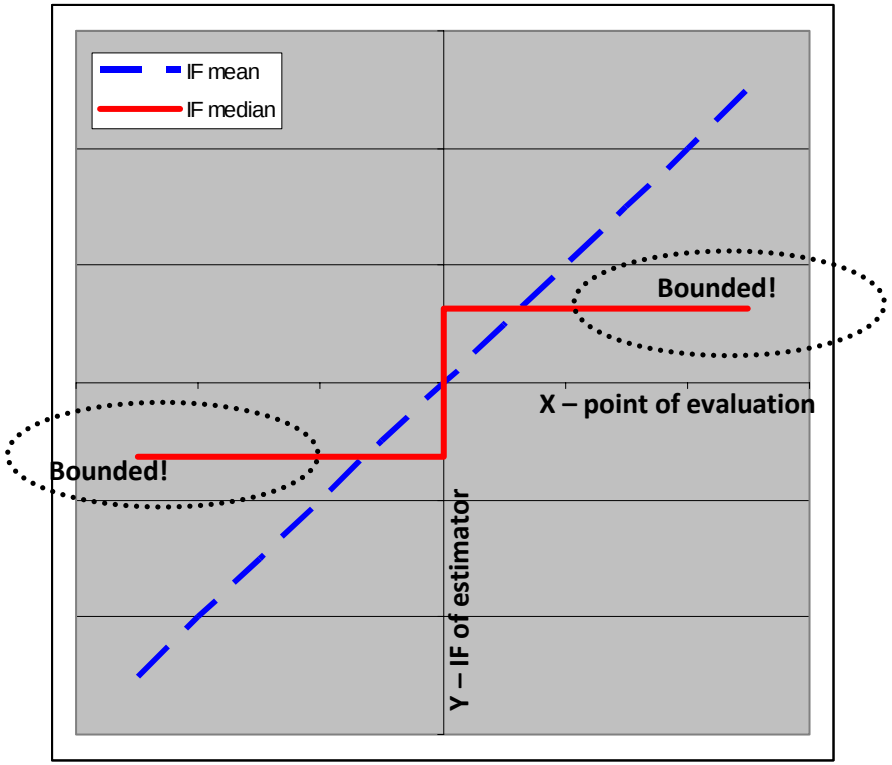
$$GES = \gamma^*(T,F) = \sup_x \left| IF(x;T,F) \right|$$

- A useful example demonstrating the concept of B-robustness is the comparison of the IFs of two common location estimators: the mean and the median. The former is unbounded with an infinite GES, and thus is not B-robust, while the latter is bounded, with a finite GES, and thus is B-robust.

* "B" comes from "bias," because if IF is bounded, the bias of the estimator is bounded.

5a. Robust Statistics: Background and the IF

Graph 1: Influence Functions of the Mean and the Median



- **Because the IF of the mean is <u>unbounded</u>, a single arbitrarily large data point can render the mean meaninglessly large, but that is not true of the median.**
- **The IF of the mean is derived mathematically below (see Hampel et al., 1986, pp.108-109 for a similar derivation for the median, also presented in Appendix II for convenience).**

5a. Robust Statistics: Background and the IF

Derivation of IF of the Mean:

Assuming $F = \Phi$, the standard normal distribution:

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

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

The statistical functional of the mean is defined by $T\left(F\right) = \int u dF\left(u\right) = \int u f\left(u\right) du$, so...

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

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

$$= \lim_{\varepsilon \to 0}\left[\frac{\varepsilon x}{\varepsilon}\right], \text{ because } \int u d\Phi\left(u\right) = 0 \quad \text{so} \quad IF\left(x; T, F\right) = x$$

Or if $F \neq \Phi$ and $\int u dF\left(u\right) \neq 0$, then $IF\left(x \mid T, F\right) = \lim_{\varepsilon \to 0}\left[\frac{-\varepsilon\mu + \varepsilon x}{\varepsilon}\right] = x - \mu$

American Bankers Association

DataMineIt℠

5a. Robust Statistics: Background and the IF

Many important robustness measures are based directly on the IF: brief definitions are presented below, with complete definitions listed in Appendix III.

- **Gross Error Sensitivity (GES):** Measures the worst case (approximate) influence that a small amount of contamination of a fixed size can have on the value of the estimator. If finite, the IF is bounded, and the estimator is "B-robust."
- **Rejection Point:** The point beyond which IF = zero and data points have no effect on the estimate.
- **Empirical Influence Function:** The non-asymptotic, finite-sample influence function.
- **Sensitivity Curves:** The scaled, non-asymptotic, finite-sample influence function (the difference between two empirical functionals, one based on a sample with contamination, one without, multiplied by *n*.)
- **Asymptotic Variance and ARE:** The variance of the estimator, and the ratio of the variances of two estimators.
- **Change-in-Variance Sensitivity:** For M-estimators, the derivative of the asymptotic variance when contaminated, divided by the asymptotic variance. Assesses how sensitive is the estimator to changes in its asymptotic variance due to contamination at *F*. If finite, then estimator is "V-robust," which is stronger than B-robustness.
- **Local Shift Sensitivity:** Assesses how sensitive the estimator is to small changes in the values of the observations; what is the worst effect on an estimator caused by shifting an observation slightly from point x to point y?
- **Breakdown Point:** A measure of global robustness, not local robustness like IF. The percentage of data points that can be contaminated with the estimator still providing useful information, that is, not "breaking down."

5a. Robust Statistics: Background and the IF

- As may now be apparent, the robust statistics approach, and the analytical toolkit on which it relies, can be used to assess the performance of a very wide range of estimators, regardless of how they are classified; it is not limited to a small group of estimators. Hence, it has very wide ranging application and general utility.
- And a major objective of a robust statistics approach, as described above, is to bound the influence function of an estimator so that the estimator remains robust to deviations from the assumed statistical model (distribution). This approach would appear to be tailor-made to tackle many of the empirical challenges resident in OpRisk loss data.

5b. IF Derived: MLE Estimators of Severity Parameters

- **The goal of this section is to derive the IFs of the MLE estimators of the parameters of the relevant severity distributions. For this presentation-format of this paper, these distributions include: Lognormal, Truncated LogNormal, Generalized Pareto Distribution (GPD), and Truncated GPD. I have made similar derivations for additional severity distributions, but include only the above for the sake of brevity. Additional distributions are included in the journal-format version of this paper.**

- **The point is to demonstrate <u>analytically</u> the non-robustness of MLE for the relevant estimations in the OpRisk setting, and hence the utility of IF as a heuristic and applied tool for assessing estimator performance. For example, deriving the IF for the mean (the MLE estimator of the specified model) gave an analytical result above of** $IF\left(x \mid \mu, T\right) = x - \mu$ **We know this is not B-robust because as** x **becomes arbitrarily large, so too does the IF: it is not bounded. Graphs comparing the IFs of these MLE estimators to the corresponding IFs of robust estimators will be shown in Section 7 (technically, the EIFs are compared, but the EIFs converge asymptotically to the IFs, and for the sample sizes used (n=250), the MLE IFs and MLE EIFs are virtually identical).**

5b. IF Derived: MLE Estimators of Severity Parameters

- **Points of Note:**
  - **Derivations of the IFs, MLE or otherwise, must account for dependence between the parameters of the severity distribution: this is something that sometimes has been overlooked in the relevant OpRisk severity modeling literature.**
  - **IFs for the MLE estimators for the (left) truncated* distributions have not been reported in the literature: they are new results.**
  - **OBRE previously has not been applied to truncated data (with one exception that does not use the standard implementation algorithm): so these, too, are new results.**
  - **Truncation induces dependence between the parameters of the severity distribution, if not there already (in which case truncation appears to augment it). This is shown in the formulae and graphs of the IFs, and appears to be the source of the extreme "sensitivity" of MLE estimators of truncated distributions reported in the literature, based on simulations. This is the first paper to present the analytic results under truncation side-by-side with simulation results.**

*** Unless otherwise noted, all truncation herein refers to left truncation, that is, truncation of the lower (left) tail, because data collection thresholds for losses ignore losses below a specified threshold.**

5b. IF Derived: MLE Estimators of Severity Parameters

- **MLEs belong to the class of "M-estimators," so called because they generalize "M"aximum likelihood estimation. Broad classes of estimators have the same form of IF (see Hampel et al. ,1986), so all M-estimators conveniently share the same form of IF.**
- **M-estimators are consistent and asymptotically normal.**
- **M-estimators are defined as any estimator** $T_n = T_n\left(X_1, \cdots, X_n\right)$ **that satisfies**

$$\sum_{i=1}^{n} \rho\left(X_i, T_n\right) = \min_{T_n}! \quad \text{or} \quad \sum_{i=1}^{n} \varphi\left(X_i, T_n\right) = 0 \quad \text{where} \quad \varphi\left(x,\theta\right) = \frac{\partial \rho\left(x,\theta\right)}{\partial \theta}$$

if the derivative of ρ **exists, and** ρ **is defined on** $\wp \times \Theta$ **.**

So for MLE:

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

$$\varphi_\theta\left(x,\theta\right) = \frac{\partial \rho\left(x,\theta\right)}{\partial \theta} = -\frac{\partial f\left(x,\theta\right)}{\partial \theta} \Big/ f\left(x,\theta\right)$$ **(note that this is simply the score function)**

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

5b. IF Derived: MLE Estimators of Severity Parameters

- **And for M-estimators, IF is defined as (assuming a nonzero denominator):**

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

where a and b define the domain of the density (in this setting, typically a = 0 and b = ∞).

So we can write

$$IF_\theta\left(x \mid \theta, T\right) = \frac{-\frac{\partial f\left(y,\theta\right)}{\partial\theta}}{f\left(y,\theta\right)} = \frac{\frac{\partial f\left(y,\theta\right)}{\partial\theta}}{f\left(y,\theta\right)}$$
$$-\int_a^b \frac{-\frac{\partial f^2\left(y,\theta\right)}{\partial\theta^2}\cdot f\left(y,\theta\right) + \left[\frac{\partial f\left(y,\theta\right)}{\partial\theta}\right]^2}{\left[f\left(y,\theta\right)\right]^2} dF\left(y\right) \qquad \int_a^b \frac{\left[\frac{\partial f\left(y,\theta\right)}{\partial\theta}\right]^2 - \frac{\partial f^2\left(y,\theta\right)}{\partial\theta^2}\cdot f\left(y,\theta\right)}{f\left(y,\theta\right)} dy$$

For the (left) truncated densities, $g\left(x,\theta\right) = \frac{f\left(x,\theta\right)}{1 - F\left(H,\theta\right)}$ **where H is the truncation threshold.**

And so the above becomes:

5b. IF Derived: MLE Estimators of Severity Parameters

IF of MLEs for (left) truncated densities:

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

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

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

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

And so the IF is

5b. IF Derived: MLE Estimators of Severity Parameters

IF of MLEs for (left) truncated densities:

$$IF_\theta\left(x;\theta,T\right)=\frac{-\frac{\partial f\left(x;\theta\right)}{\partial\theta}\frac{1}{f\left(x;\theta\right)}-\frac{\partial F\left(H;\theta\right)}{\partial\theta}\frac{1}{1-F\left(H;\theta\right)}}{-\frac{1}{1-F\left(H;\theta\right)}\int_a^b\frac{\left[\frac{\partial f\left(y;\theta\right)}{\partial\theta}\right]^2-\frac{\partial^2 f\left(y;\theta\right)}{\partial\theta^2}\cdot f\left(y;\theta\right)}{f\left(y;\theta\right)}dy+\frac{\left[\frac{\partial F\left(H;\theta\right)}{\partial\theta}\right]^2+\frac{\partial^2 F\left(H;\theta\right)}{\partial\theta^2}\cdot\left[1-F\left(H;\theta\right)\right]}{\left[1-F\left(H;\theta\right)\right]^2}}$$

Note that a and b are now H and (typically) ∞, respectively.

As noted previously, we must account for (possible) dependence between the parameter estimates, and so we must use the matrix form of the IF defined below (see Stefanski & Boos (2002) and D.J. Dupuis (1998)):

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

Where K is either F or G, $A\left(\theta\right)$ is simply the Fisher Information (if the data follow the assumed model), and φ_θ is now vectorized. Parameter dependence exists when the off-diagonal terms are not zero.

5b. IF Derived: MLE Estimators of Severity Parameters

Note that the off-diagonal cross-terms are the second-order partial derivatives:

$$-\int_a^b \frac{\partial \varphi_{\theta_1}}{\partial \theta_2} dG(y) = -\frac{1}{1-F(H;\theta)} \int_a^b \frac{\left[\frac{\partial f(y;\theta)}{\partial \theta_1}\right]\left[\frac{\partial f(y;\theta)}{\partial \theta_2}\right] - \frac{\partial^2 f(y;\theta)}{\partial \theta_1 \partial \theta_2} \cdot f(y;\theta)}{f(y;\theta)} dy + \frac{\left[\frac{\partial F(H;\theta)}{\partial \theta_1}\right]\left[\frac{\partial F(H;\theta)}{\partial \theta_2}\right] + \frac{\partial^2 F(H;\theta)}{\partial \theta_1 \partial \theta_2} \cdot \left[1-F(H;\theta)\right]}{\left[1-F(H;\theta)\right]^2}$$

and

$$-\int_a^b \frac{\partial \varphi_{\theta_1}}{\partial \theta_2} dF(y) = \int_a^b \frac{\left[\frac{\partial f(y,\theta)}{\partial \theta_1}\right]\left[\frac{\partial f(y,\theta)}{\partial \theta_2}\right] - \frac{\partial f^2(y,\theta)}{\partial \theta_1 \partial \theta_2} \cdot f(y,\theta)}{f(y,\theta)} dy$$

With the above defintion, all that needs be done to derive IF for each severity distribution is the calculation of the first and second order derivatives of each density, as well as, for the (left) truncated cases, the first and second order derivatives of the cumulative distribution functions: that is, derive

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

This is done in Appendix IV for the four severity distributions examined herein.

This "plug-n-play" approach makes derivation and use of the IFs corresponding to each severity distribution's parameters considerably more convenient.

5b. IF Derived: MLE Estimators of Severity Parameters

Below, I "plug-n-play" to obtain $A(\theta)$ for the four severity distributions. Note that for the LogNormal, (left) truncation induces parameter dependence, and for the GPD, it augments dependence that was there even before truncation. For the truncated cases and the GPD, after the cells of $A(\theta)$ are obtained, IF is calculated numerically.

From Appendix IV, inserting the derivations of $\frac{\partial f(y;\theta)}{\partial \theta_1}, \frac{\partial f(y;\theta)}{\partial \theta_2}, \frac{\partial^2 f(y;\theta)}{\partial \theta_1 \partial \theta_2}, \frac{\partial^2 f(y;\theta)}{\partial \theta_1^2}$, and $\frac{\partial^2 f(y;\theta)}{\partial \theta_2^2}$

for the LogNormal yields

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

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

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

5b. IF Derived: MLE Estimators of Severity Parameters

Inserting Appendix IV derivations of for the LogNormal yields…

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

(zero off-diagonals indicate no parameter dependence)

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

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

5b. IF Derived: MLE Estimators of Severity Parameters

From Appendix IV, inserting the derivations of

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

for the (left) Truncated LogNormal yields

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

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

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

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

(non-zero cross-terms indicate parameter dependence)

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

5b. IF Derived: MLE Estimators of Severity Parameters

From Appendix IV, inserting the derivations of $\frac{\partial f(y;\theta)}{\partial \theta_1}, \frac{\partial f(y;\theta)}{\partial \theta_2}, \frac{\partial^2 f(y;\theta)}{\partial \theta_1 \partial \theta_2}, \frac{\partial^2 f(y;\theta)}{\partial \theta_1^2}$, and $\frac{\partial^2 f(y;\theta)}{\partial \theta_2^2}$

for the GPD yields

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

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

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

(non-zero cross-terms indicate parameter dependence)

5b. IF Derived: MLE Estimators of Severity Parameters

From Appendix IV, inserting the derivations of

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

for the (left) Truncated GPD yields

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

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

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

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

5b. IF Derived: MLE Estimators of Severity Parameters

From Appendix IV, inserting the derivations of

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

for the (left) Truncated GPD yields

(non-zero cross-terms indicate parameter dependence)

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

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

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

5c. Robust Estimators: OBRE and CvM

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 (1):

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

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

and

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

and *A* and *a* respectively are a dim(θ) x dim(θ) matrix and a dim(θ)-dimensional vector determined by the equations:

$$E\left[\varphi_c^{A,a}\left(x;\theta\right)\cdot\varphi_c^{A,a}\left(x;\theta\right)^T\right] = I$$ **((2) – ensures bounded IF)**

$$E\left[\varphi_c^{A,a}\left(x;\theta\right)\right] = 0$$ **((3) – ensures Fisher consistency)**

$s\left(x;\theta\right)$ **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,** $\sqrt{\dim(\theta)} \le c \le \infty$ **, regulating from very robust to MLE, respectively.**

5c. Robust Estimators: OBRE and CvM

OBRE Defined:

- **The weights make OBRE robust, but it maintains efficiency as close as possible to MLE (subject to its constraints) because it is based on the scores function. Hence, its name: "Optimal" B-Robust Estimator. The constraints – bounded IF and Fisher consistency – are implemented with *A* and *a*, respectively, which can be viewed as Lagrange multipliers. And *c* regulates the robustness-efficiency tradeoff: a lower *c* gives a more robust estimator, and $c = \infty$ is MLE. Bottom line: by minimizing the trace of the asymptotic covariance matrix, OBRE is maximally efficient for a given level of robustness, which is controlled by the analyst with *c*. Many choose *c* to achieve 95% efficiency relative to MLE, but this actual value for *c* depends on the model being implemented.**

- **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 numerically more stable" (see Alaiz and Victori-Feser, 1996). The "standardized" OBRE is used in this study.**

5c. Robust Estimators: OBRE and CvM

OBRE Computed:

To compute OBRE, (1) must be solved under conditions (2) and (3), for a given tuning parameter value *c*, via Newton-Raphson (see 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 $M_k = \int \left[s(x;\theta) - a \right] \cdot \left[s(x;\theta) - a \right]^T \cdot W_c(x,\theta)^k \, dF_\theta(x), \ \text{k=1,2}$

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.**

5c. Robust Estimators: OBRE and CvM

OBRE Computed:

- The idea of the above algorithm is to first compute ***A*** and ***a*** for a given ***θ*** by solving (2) and (3). This is followed by a Newton-Raphson step given these two new matrics, and these steps are iterated until convergence is achieved.

- The above algorithm follows D.J. Dupuis (1998), who 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 <u>STEP 2</u> and M_1 in <u>STEP 3</u>, but Dupuis (1998) cautions that the former calculation of ***a*** requires integration, rather than a weighted average from plugging in the empirical density, or else (1.a) 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 <u>STEP 4</u> rather than just $\Delta\theta > \eta$ as in Alaiz and Victoria-Feser (1996).

- The initial values for ***A*** and ***a*** in <u>STEP 1</u> correspond to the MLE.

5c. Robust Estimators: OBRE and CvM

OBRE Computed:

- 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 5% arbitrary deviations from the assumed model.

- Note that the weights generated and used by OBRE, W_c, can be extremely useful for another important objective of robust statistics – outlier detection. Within the OpRisk setting, this can be especially useful for determining appropriate "units of measure" (uom), the grouping of loss events by some combinations of business unit and event type, each uom with the same (or close) loss distribution. As discussed below, the extreme quantiles that need to be estimated for regulatory capital and economic capital purposes are extremely sensitive to even slight changes in the variability of the parameter estimates. This, along with the a) unavoidable tradeoff between statistical power (sample size) and homogeneity; b) loss-type definitional issues; and c) remaining heterogeneity within units of measure even under ideal conditions, all make defining units of measure an extremely challenging and crucial task; good statistical methods can and should be utilized to successfully execute on this challenge.

5c. Robust Estimators: OBRE and CvM

CvM Defined:

The Cramér von Mises estimator is a "minimum distance" estimator (MDE), yielding the parameter value of the assumed distribution that minimizes its distance from the empirical distribution. Given the CvM statistic $W^2(\theta)$ in its common form,

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

where F_n is the empirical distribution and F_θ is the assumed distribution, the minimum CvM estimator (MCVME) is that value $\hat{\theta}$ of θ, for the given sample, that minimizes $W^2(\theta)$:

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

5c. Robust Estimators: OBRE and CvM

CvM Computed:

The computational formula typically used to calculate the MCVME is:

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

where $X_{(s)}$ **is the ordered (s)'th value of x.**

- **MCVME is an M-class estimator, and as such it is consistent and asymptotically normal.**
- **MDE's are very similar conceptually, and typically differ in how they weight the data points. For example, Anderson-Darling, another MDE, weights the tail more than does CvM. CvM is very widely used, perhaps the most widely used MDE, hence its inclusion.**
- **Before presenting results comparing MLE to OBRE and CvM, I talk briefly about (left) truncation, and reemphasize its analytic and empirical importance in this setting.**

6. Truncation Matters, the Threshold Matters

- The effects of a collection threshold on parameter estimation can be unexpected, even counterintuitive, both in the magnitude of the effect, and its direction.
- Note first that given the size of the economic and regulatory capital estimates generated from severity distribution modeling (into the hundreds of millions and billions of dollars), the size of the thresholds appear tiny, and the % of the non-truncated distributions that fall below the thresholds do not appear shockingly large, either (assuming, of course, that the loss distribution below the threshold is the same as that above it, which is solely a heuristic assumption here).
- However, the effects of (left) truncation on MLE severity distribution parameter estimates can be dramatic, even for low thresholds.
- Not only are the effects dramatic, but arguably very unexpected. The entire shape AND DIRECTION of some of the IFs change as does the threshold, over relatively small changes in the threshold value.
- Note that this is not merely a sensitivity to simulation assumptions, but rather, an analytical result.

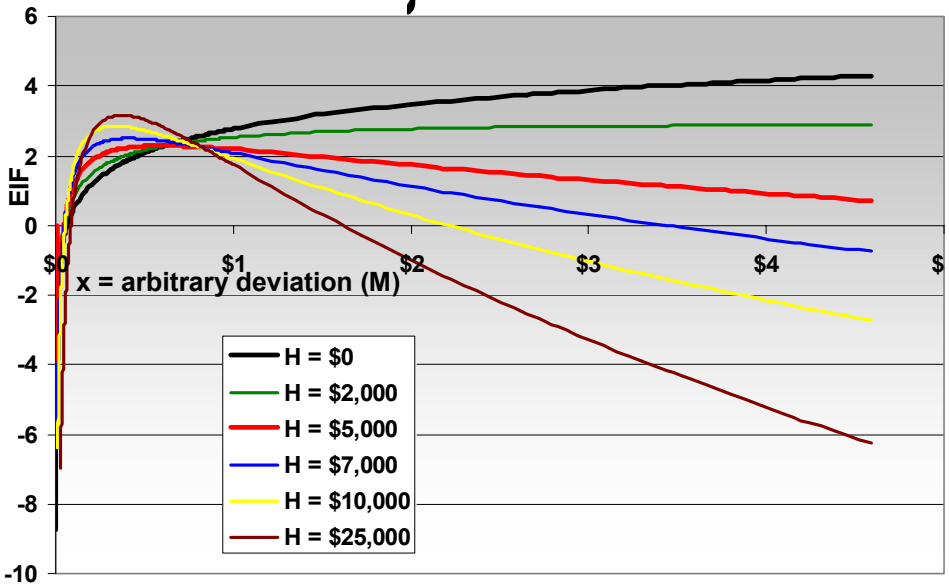
Collection Threshold	LogNormal (μ=10.95, σ=1.75) % Below	GPD (ξ=1.2, β=70,000) % Below
$1,000	1.0%	1.4%
$2,000	2.8%	2.8%
$3,000	4.6%	4.1%
$4,000	6.5%	5.4%
$5,000	8.2%	6.6%
$10,000	16.0%	12.4%
$20,000	27.5%	21.8%
$25,000	31.9%	25.7%

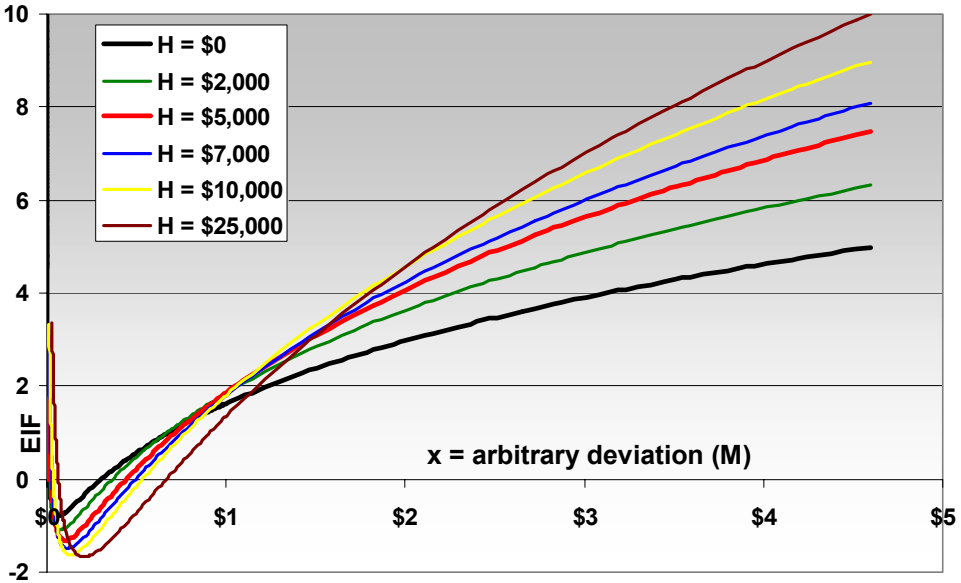
6. Truncation Matters, the Threshold Matters

- **The effects of a collection threshold on parameter estimation can be unexpected, even counterintuitive, both in the magnitude of the effect, and its direction.**

EIF of Truncated LogNormal (μ = 10.95, σ = 1.75) MLE Parameter Estimates:

μ by Size of Threshold σ





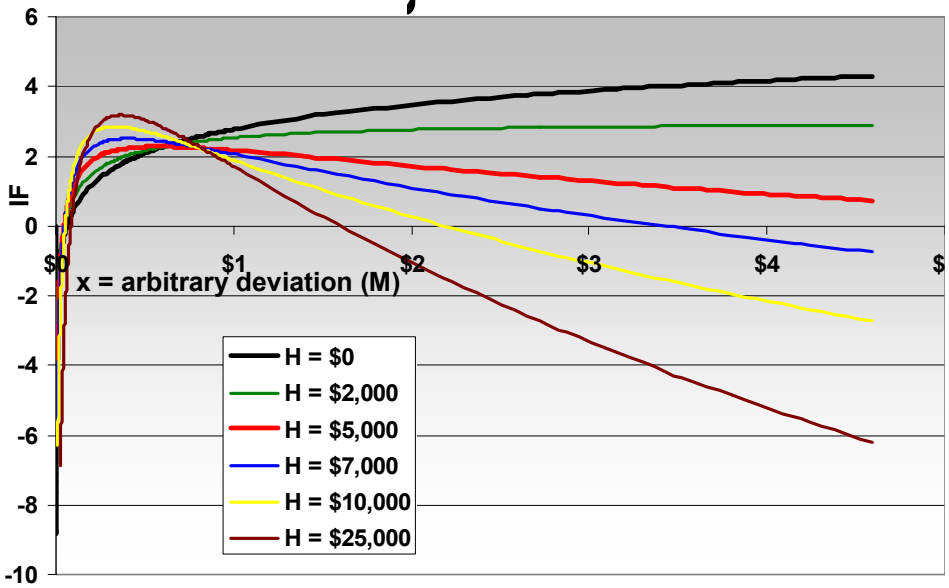
- **Note the NEGATIVE covariance between parameters induced by (left) truncation. Many would call this unexpected, if not counter-intuitive: the location parameter, *μ*, DECREASES under larger and larger arbitrary deviations.**

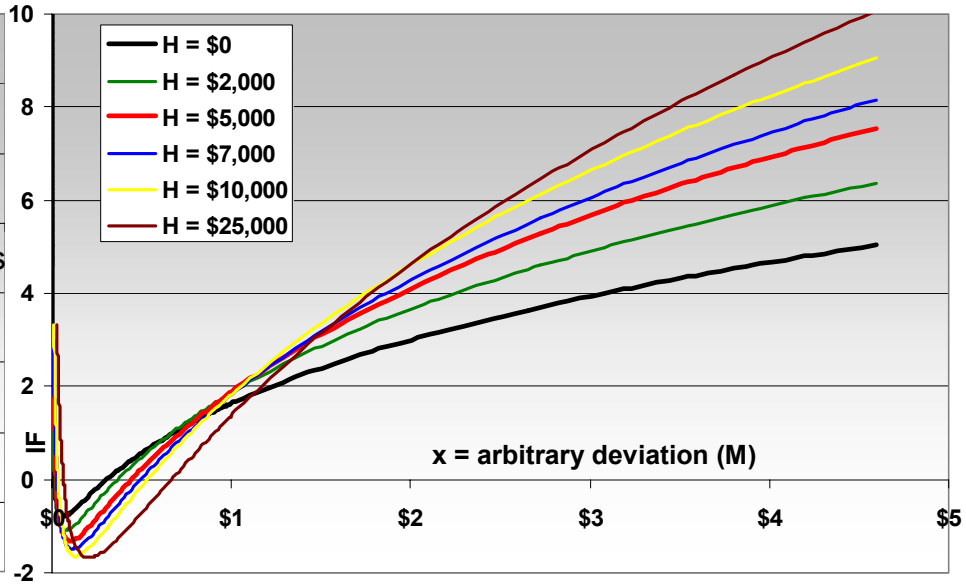
6. Truncation Matters, the Threshold Matters

- **The effects of a collection threshold on parameter estimation can be unexpected, even counterintuitive, both in the magnitude of the effect, and its direction.**

IF of Truncated LogNormal (μ = 10.95, σ = 1.75) MLE Parameter Estimates:

μ by Size of Threshold σ





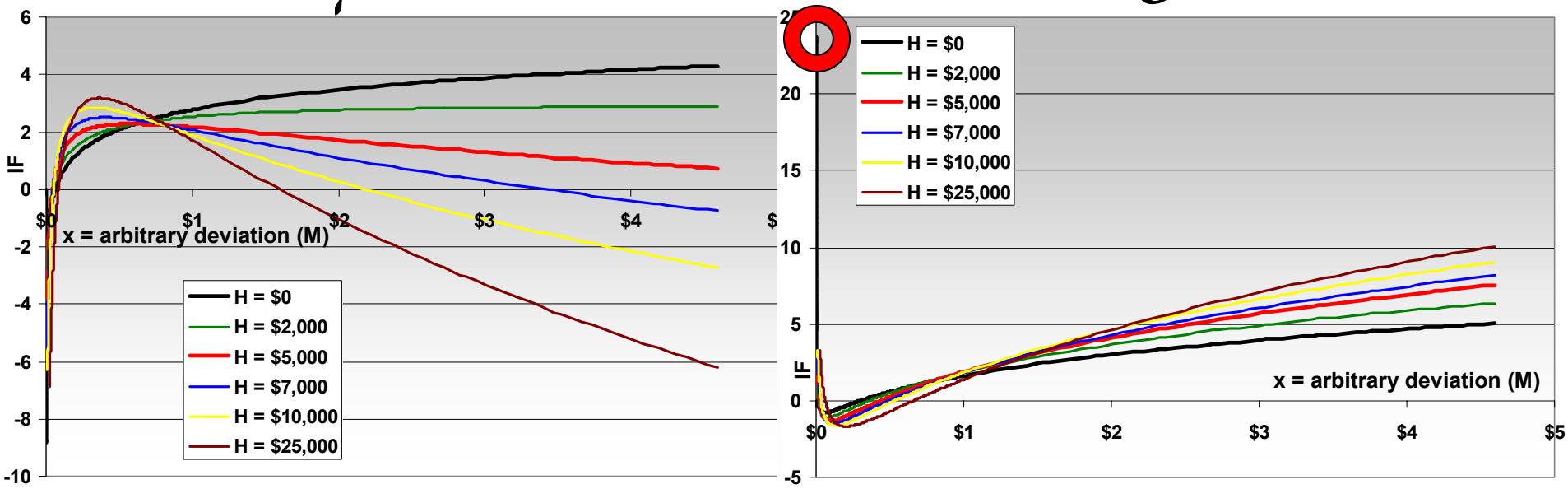
- **Note the NEGATIVE covariance between parameters induced by (left) truncation. Many would call this unexpected, if not counter-intuitive: the location parameter, *μ*, DECREASES under larger and larger arbitrary deviations.**

6. Truncation Matters, the Threshold Matters

- **In an interesting twist, note that (left) truncation actually DECREASES sensitivity of the MLE estimator for σ (and less so for μ) for SMALL deviations (in the left tail) from the assumed model.**

IF of Truncated LogNormal (μ = 10.95, σ = 1.75) MLE Parameter Estimates:

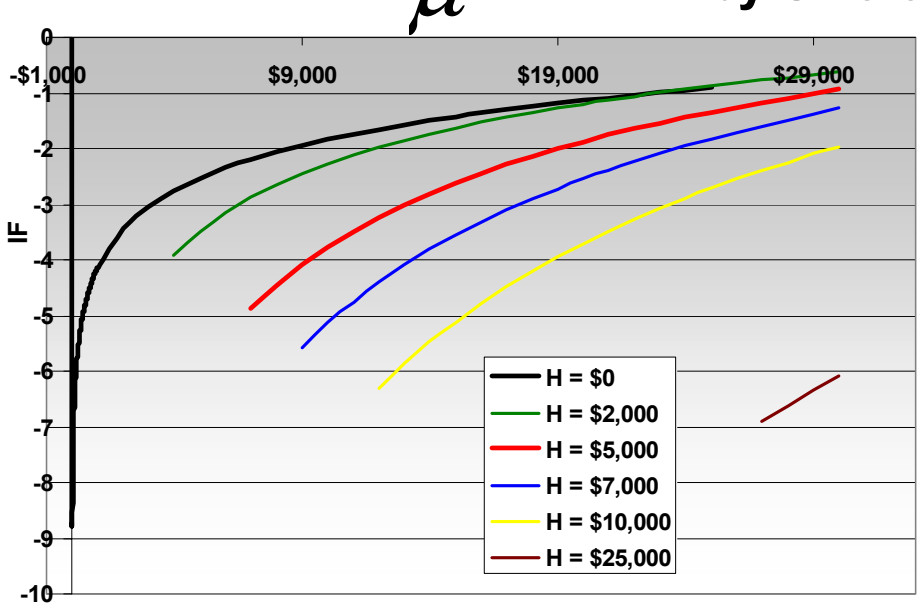
μ by Size of Threshold σ

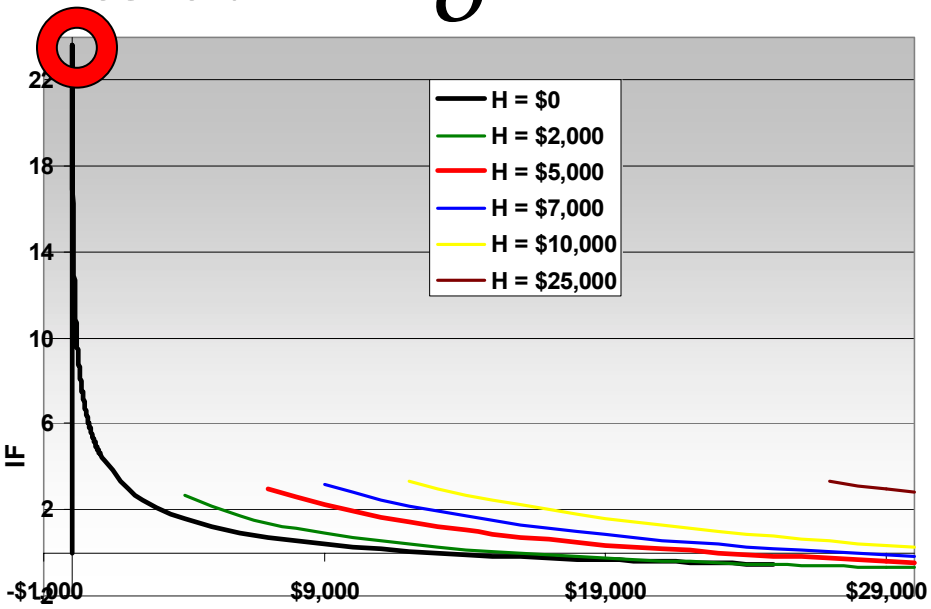


6. Truncation Matters, the Threshold Matters

- **In an interesting twist, note that (left) truncation actually DECREASES sensitivity of the MLE estimator for σ (and less so for μ) for SMALL deviations (in the left tail) from the assumed model.**

IF of Truncated LogNormal (μ = 10.95, σ = 1.75) MLE Parameter Estimates: by Size of Threshold



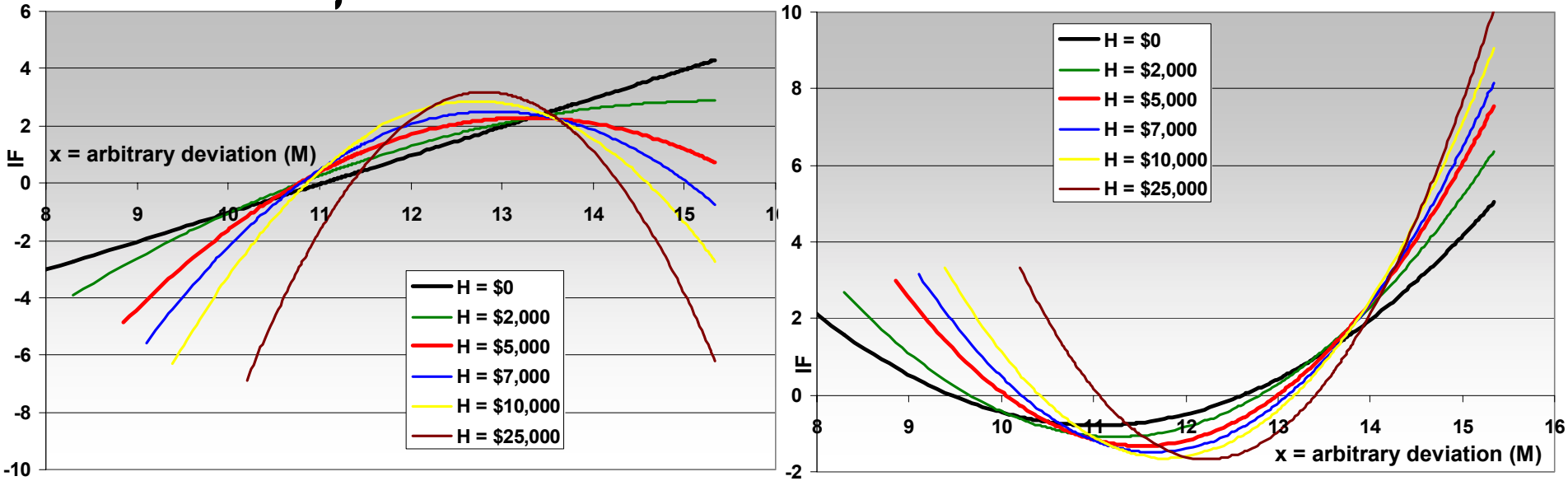


6. Truncation Matters, the Threshold Matters

- **The effects of a collection threshold on parameter estimation can be unexpected, even counterintuitive, both in the magnitude of the effect, and its direction.**

IF of Truncated LogNormal (μ = 10.95, σ = 1.75) MLE Parameter Estimates:

μ **by Size of Threshold** σ



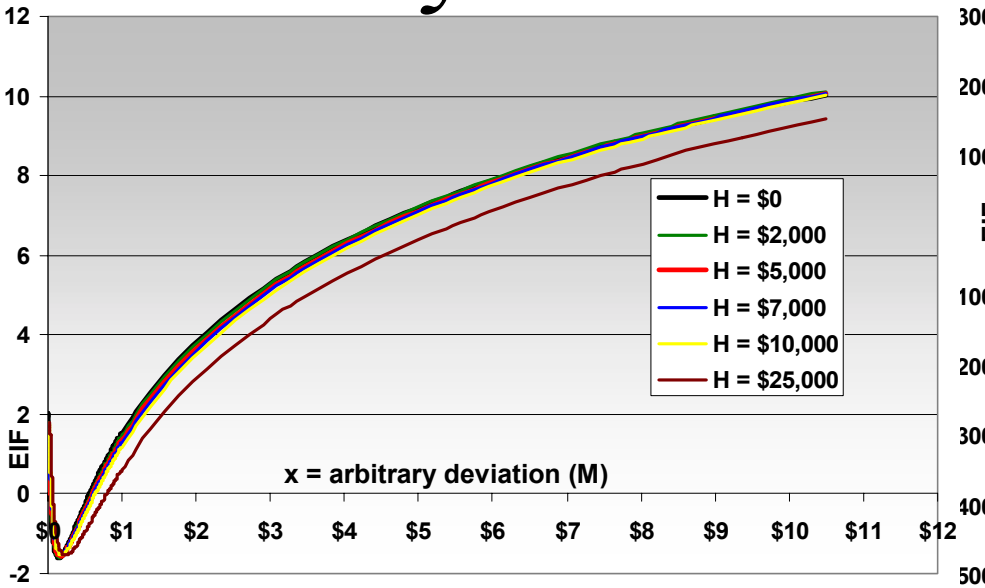
Log Scale

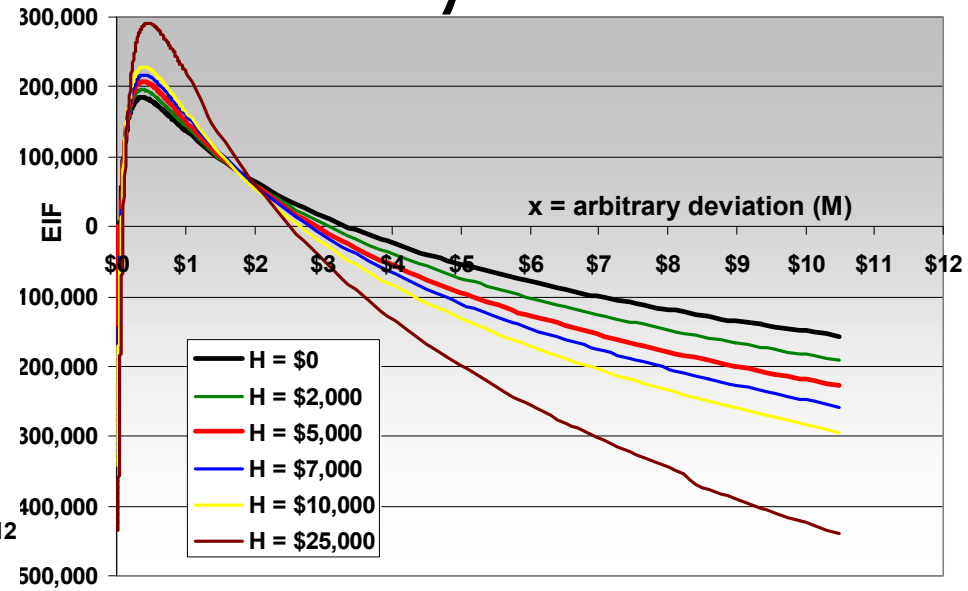
- **Note the log-linear** $IF_{\mu}\left(x;\mu,\sigma;MLE\right)=\ln\left(x\right)-\mu$ **under no truncation is analogous to the** $IF_{\mu}\left(x;\mu,\sigma;MLE\right)=x-\mu$ **obtained earlier under the normal distribution.**

6. Truncation Matters, the Threshold Matters

- **The effects of a collection threshold on parameter estimation can be unexpected, even counterintuitive, both in the magnitude of the effect, and its direction.**

EIF of Truncated GPD (ξ = 1.20, β = 70,000) MLE Parameter Estimates: ξ by Size of Threshold β



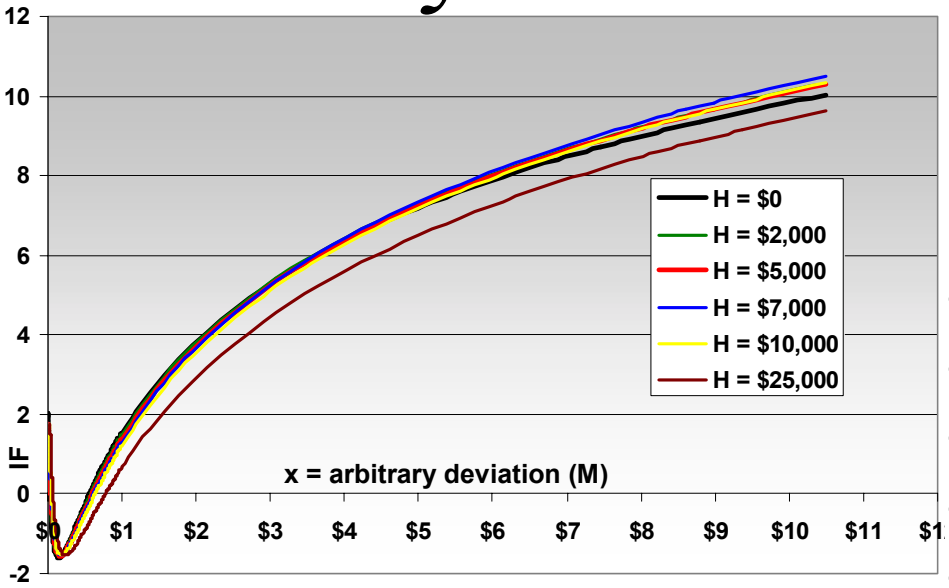


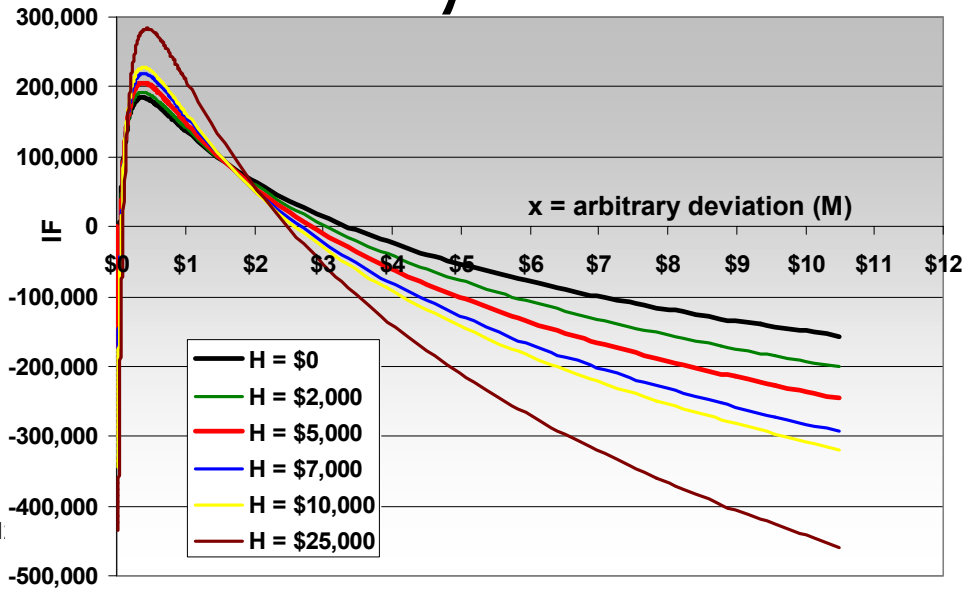
- **β is more strongly affected by (left) truncation, increasing the negative covariance between parameters.**

6. Truncation Matters, the Threshold Matters

- **The effects of a collection threshold on parameter estimation can be unexpected, even counterintuitive, both in the magnitude of the effect, and its direction.**

IF of Truncated GPD (ξ = 1.20, β = 70,000) MLE Parameter Estimates: ξ by Size of Threshold β



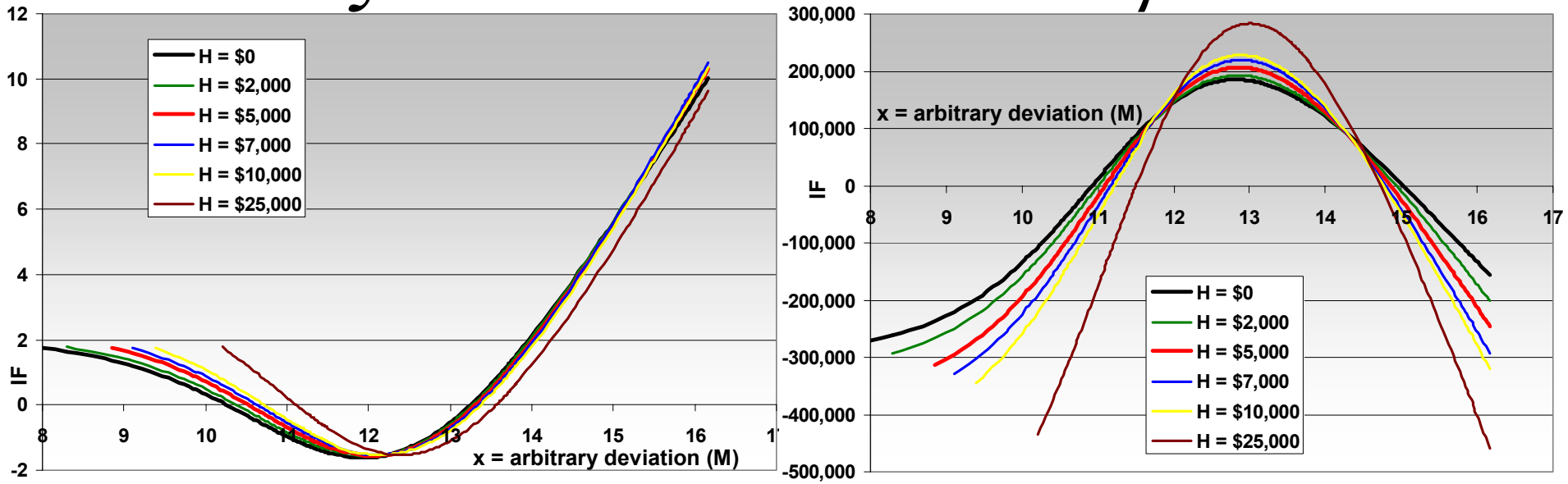


- **β is more strongly affected by (left) truncation, increasing the negative covariance between parameters.**

6. Truncation Matters, the Threshold Matters

- **The effects of a collection threshold on parameter estimation can be unexpected, even counterintuitive, both in the magnitude of the effect, and its direction.**

IF of Truncated GPD (ξ = 1.20, β = 70,000) MLE Parameter Estimates: ξ by Size of Threshold β



- **β is more strongly affected by (left) truncation, increasing the negative covariance between parameters.**

6. Truncation Matters, the Threshold Matters

- **These unexpected, and even counterintuitive results, both in the magnitude of the effect of (left) truncation, and sometimes its direction, are confirmed in the simulations presented below, side-by-side with the analytical IF results. This would appear to explain the extreme sensitivity of MLE estimators under truncation reported in the literature, which has perplexed some.**

7. Results: Simulation Descriptions

The simulations generate MLE parameter estimates vs. OBRE and CvM parameter estimates. The Empirical Influence Functions, which match perfectly with the derived IF formulae presented above (which is actually a very useful QC check), are presented side-by-side with the simulations, which confirm the performance under arbitrary deviations indicated by the IFs.

- Sample Size: n = 250 was chosen as a reasonable size for many units-of-measure. Depending on the bank, some will have larger n, some smaller, but if the results were not useful for this n = 250, then sample size would have been a real issue with these methods going forward, so that is why n = 250 was selected.

- Severity Distributions: the LogNormal and the Generalized Pareto. Both are commonly used in this setting, but they are very distinct distributions, with the latter being more heavy-tailed (see table). The motivation for using GPD obviously was NOT a peaks-over-threshold approach in this study, but rather, the fact that it is in common use in this setting, is heavy-tailed, and is distinct from the LogNormal (i.e. one is not a transformed or limiting version of the other). Results obtained from other distributions will be included in journal-format version of this paper.

Quantile	LogNormal (μ=10.95, σ=1.75)	GPD (ξ=1.2, β=70,000)
50.000%	$56,954	$134,015
75.000%	$185,416	$307,885
90.000%	$536,443	$924,521
95.000%	$1,013,068	$2,123,992
99.000%	$3,338,756	$14,652,671
99.900%	$12,710,088	$232,229,183
99.996%	$56,666,862	$11,052,099,964

7. Results: Simulation Descriptions

- **Truncation and Shifting: The Truncated LogNormal and Truncated GPD, with the relatively low threshold of $5k, also are included, as is the "Shifted" LogNormal (the threshold is subtracted from the losses generated under truncation, the LogNormal is fitted, and then the threshold is added back after estimation).**

- **Parameter values: These were choosen (LogNormal μ = 10.95, σ = 1.75, Truncated LogNormal μ = 9.90, σ = 2.40, and both GPD and Truncated GPD ξ = 1.2, β = 70,000) so as to reflect a) fairly large differences between the Lognormal and the GPD; b) general empirical realities based on OpRisk work I've done (but not proprietary results); c) yet, some "stretching" vis-à-vis fairly large (but still realistic) GPD parameters, to ensure that there were no estimation problems with these methods when they encountered more extreme severity distributions that are more difficult to estimate (actually, because $\xi > 1$, this GPD has infinite mean, a not uncommon occurrence empirically with OpRisk loss data, according to the literature). Obviously, for any given setting, all estimation methods should be tested extensively for parameter value ranges relevant to the specific estimation effort.**

- **OBRE value of *c*: For OBRE, different values for *c*, the tuning parameter, were used with the given parameter values, and values about *c* = 2 most consistently provided what appeared to be a good tradeoff between robustness and efficiency. Again, for any given setting and any given loss dataset, different values of *c* should be tested and evaluated based on the objectives of the specific estimation effort.**

7. Results: Simulation Descriptions

- **OBRE Starting Values: MLE estimates were used as 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 $\eta = 0.01$ usually were used herein, but sometimes $\eta = 0.02$ sufficed and saved computation and time resources; when it did not suffice, it produced unacceptable and unpredictable jumps in the values of the IF. Such behavior is typical of convergence algorithms, so their responsible use requires cognizance of them.**

- **CvM Starting values: A wide range of parameter values were provided for the Gaussian quadrature optimization algorithm. Only one convergence issue was encountered.**

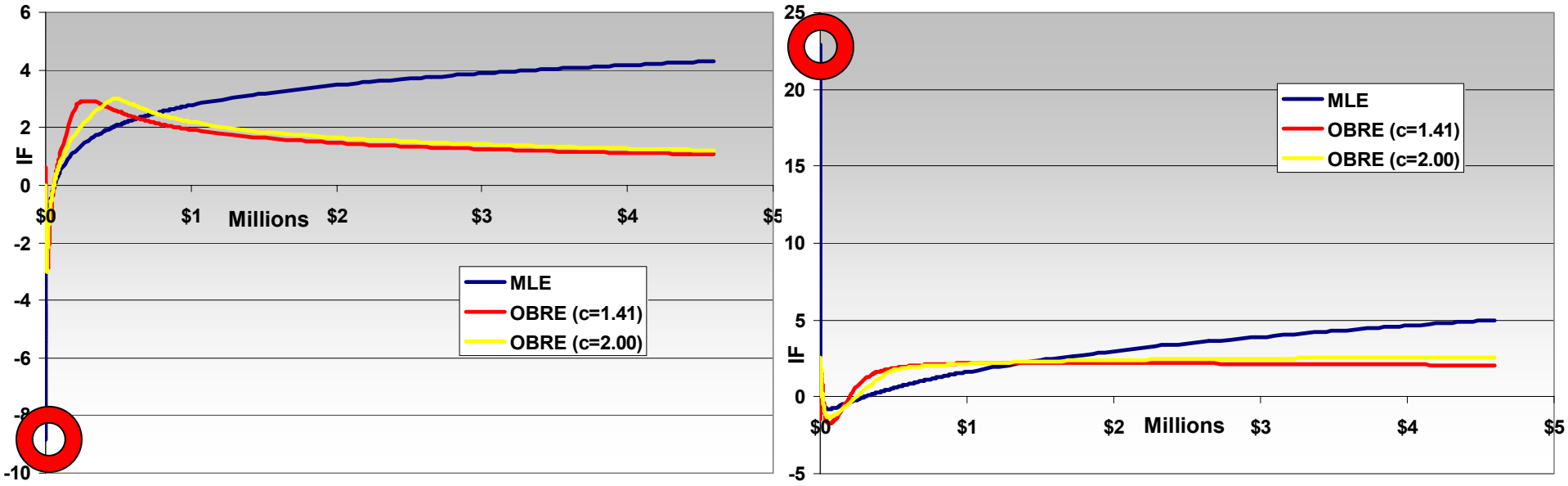
- **Arbitrary Deviations: Observations arbitrarily deviating from the assumed severity distribution, both 2% of all observations and 5% of all observations, are randomly drawn from the top 10%tile (right tail) of the distribution and multiplied by a factor of ten.**

- **Starting points are sometimes noted in the literature as being important for the convergence of these algorithms, although this possibly is due to the relatively small sample sizes (as low as n = 40) being used in some of those studies (see Horbenko, Ruckdeschel, & Bae, 2011). Only one potential convergence issue was noted in this study. Again, the focus here was to compare the robust methods to MLE, and to ensure that they were worth pursuing for application in this setting. An important "next step" for this applied research is a sample size study designed to test "how low can we go."**

7. Results: LogNormal Distribution

- **NOTE: Aribtrary deviations from the assumed model do not have to be large in absolute value to have a large impact on MLE estimates. The IF is a useful tool for spotting such counter-intuitive and important effects.**

EIF of LogNormal (μ = 10.95, σ = 1.75) Parameter Estimates: OBRE v. MLE



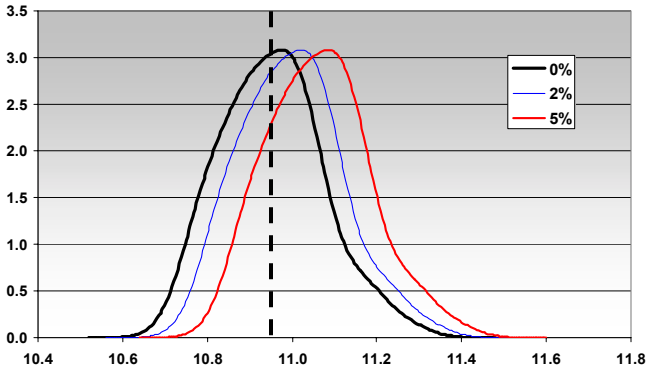
7. Results: LogNormal Distribution (n=250)

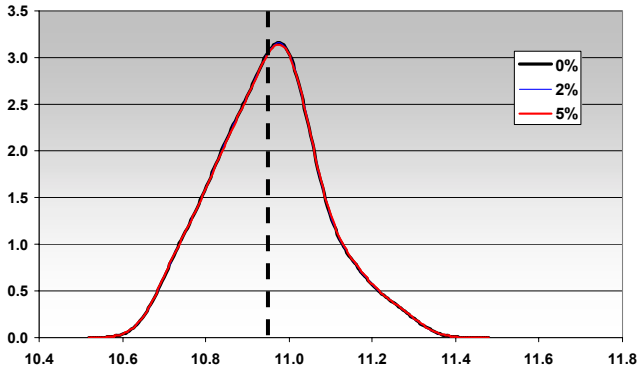
100 Simulations, MLE θ's by % Arbitrary Deviation

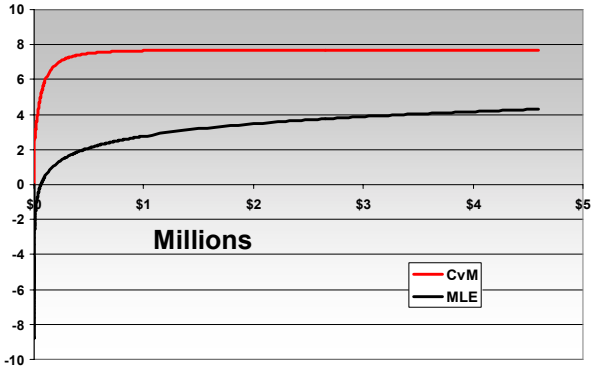
100 Simulations, CvM θ's by % Arbitrary Deviation

(Empirical) Influence Function

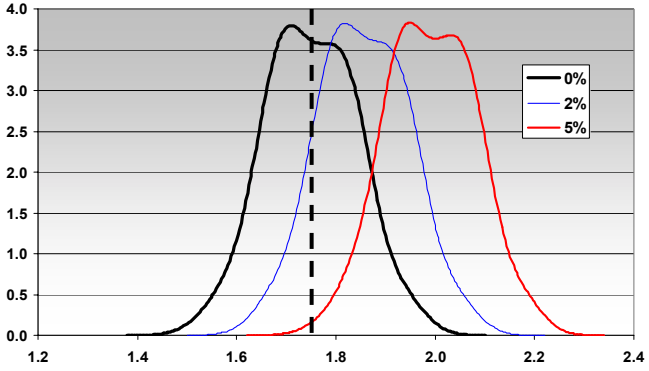
μ = 10.95

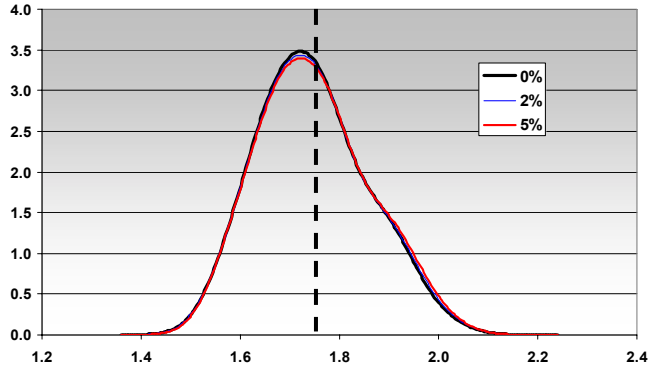


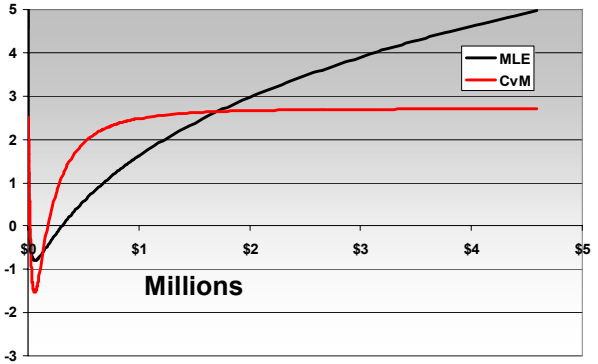




σ = 1.75







American Bankers Association

DataMineIt℠

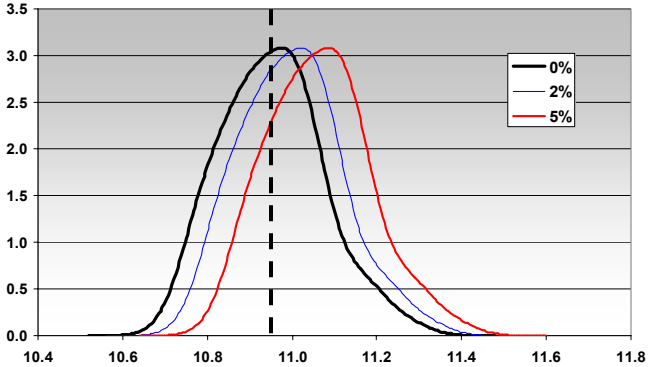
7. Results: LogNormal Distribution (n=250)

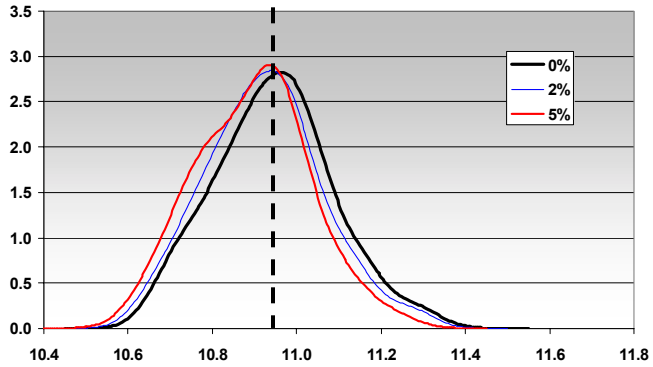
100 Simulations, MLE θ's by % Arbitrary Deviation

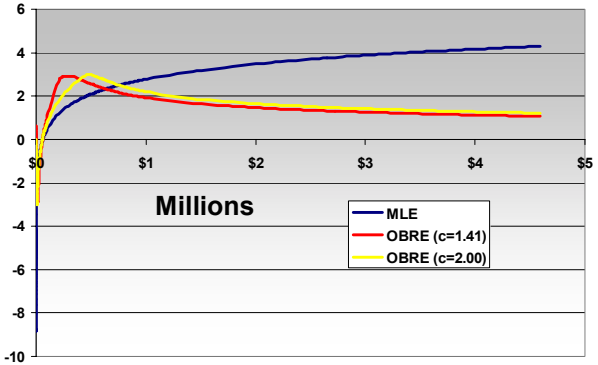
100 Simulations, OBRE (c=2) θ's by % Arbitrary Deviation

(Empirical) Influence Function

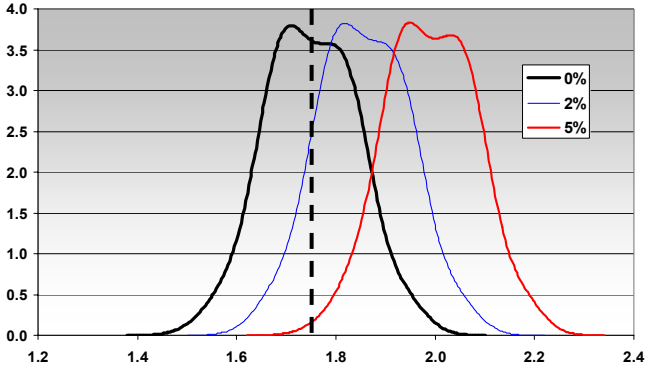
μ = 10.95

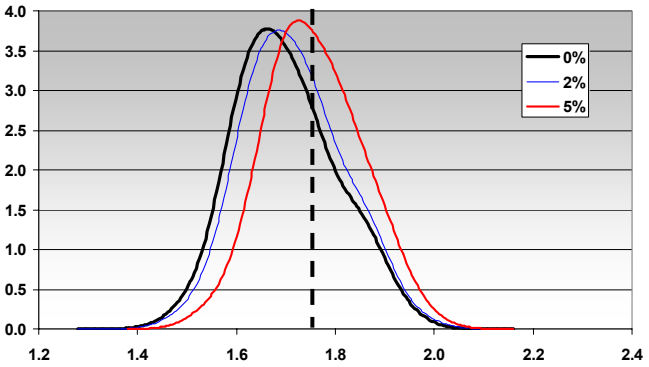


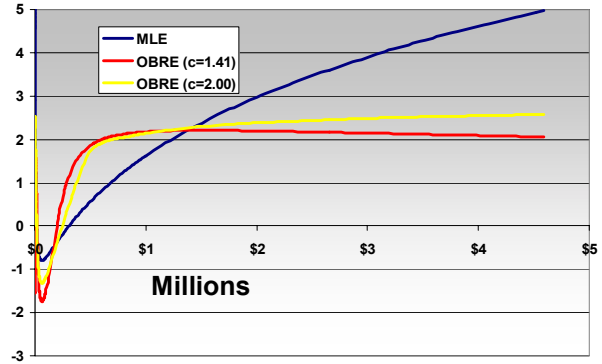




σ = 1.75







American Bankers Association

DataMineIt℠

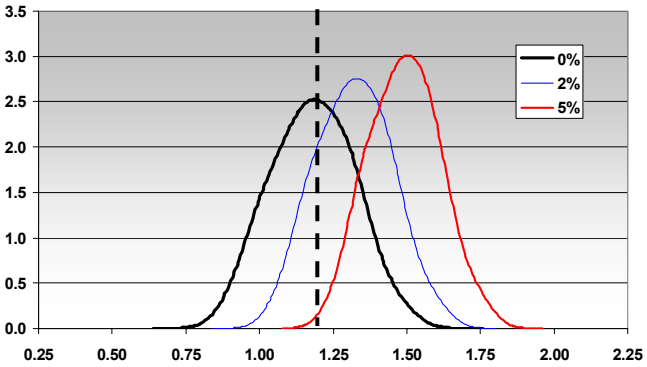
7. Results: Generalized Pareto Distribution (n=250)

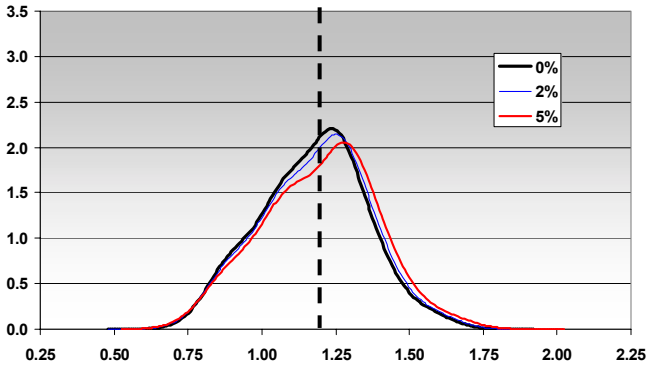
100 Simulations, MLE θ's by % Arbitrary Deviation

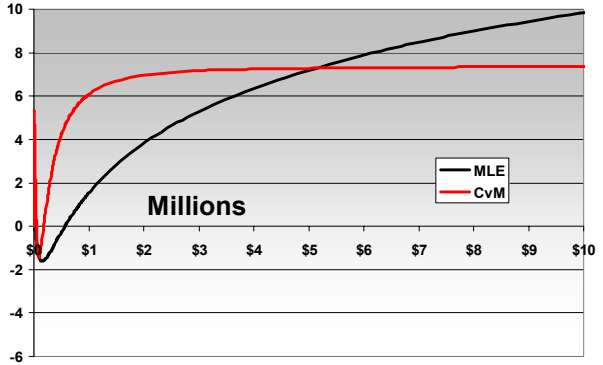
100 Simulations, CvM θ's by % Arbitrary Deviation

(Empirical) Influence Function

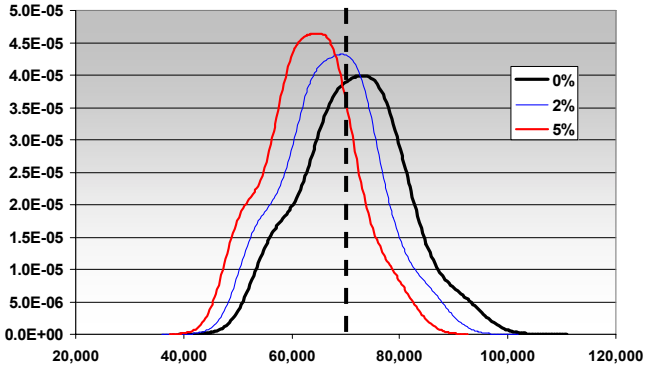
ξ = 1.20

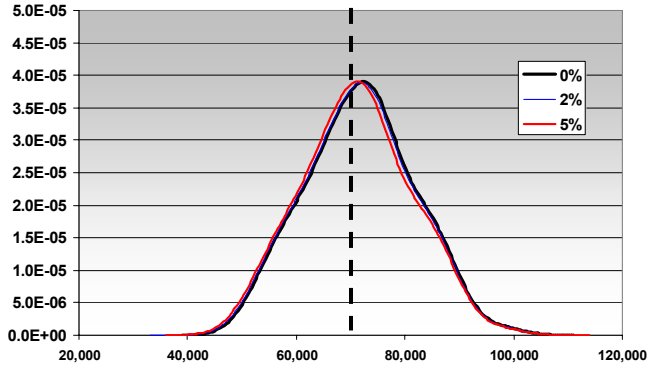


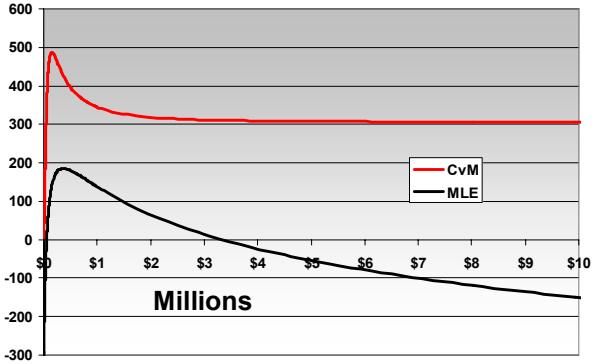




β = 70,000







American Bankers Association

DataMineIt℠

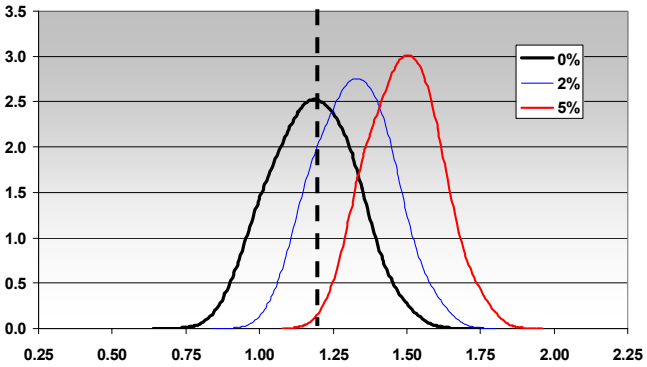
7. Results: Generalized Pareto Distribution (n=250)

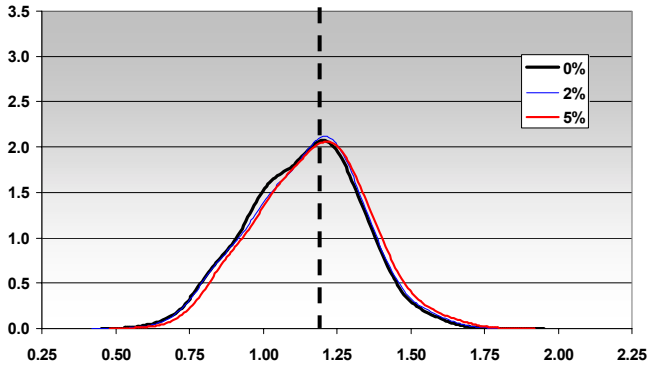
100 Simulations, MLE θ's by % Arbitrary Deviation

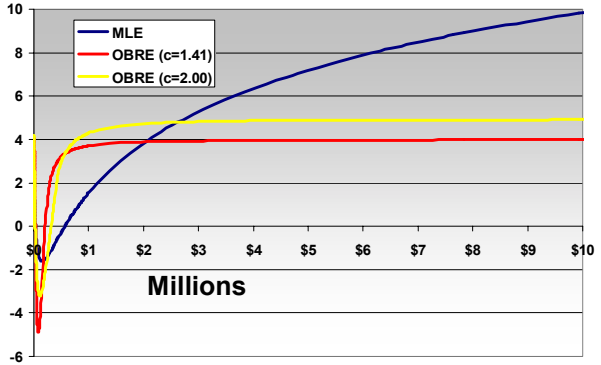
100 Simulations, OBRE (c=2) θ's by % Arbitrary Deviation

(Empirical) Influence Function

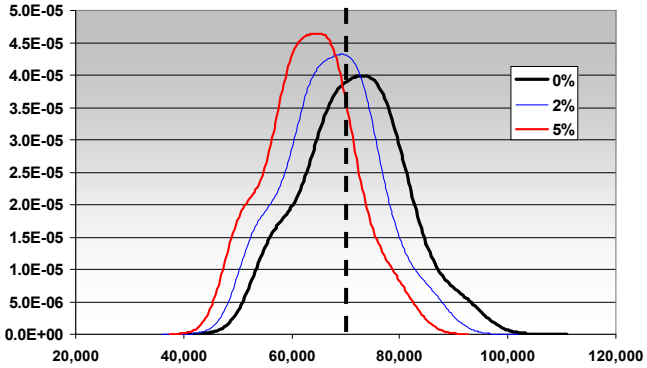
ξ = 1.20

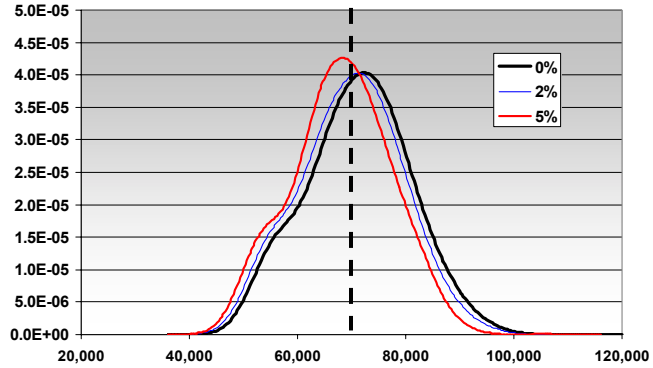


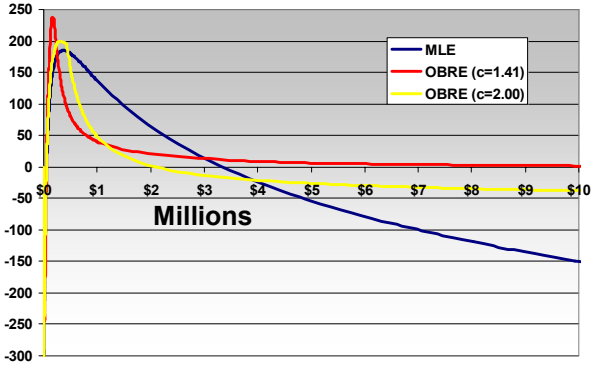




β = 70,000







American Bankers Association

DataMineIt℠

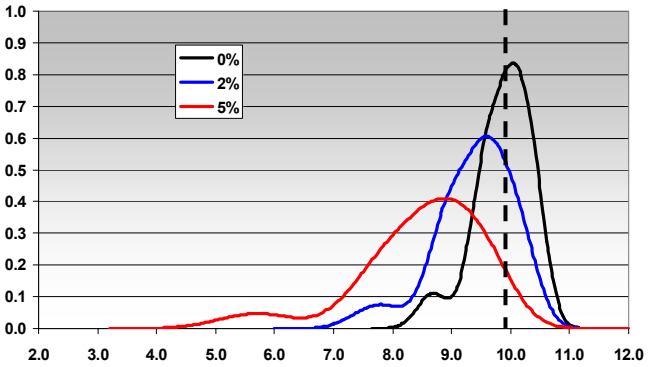
7. Results: Truncated LogNormal (n=250, H=$5k)

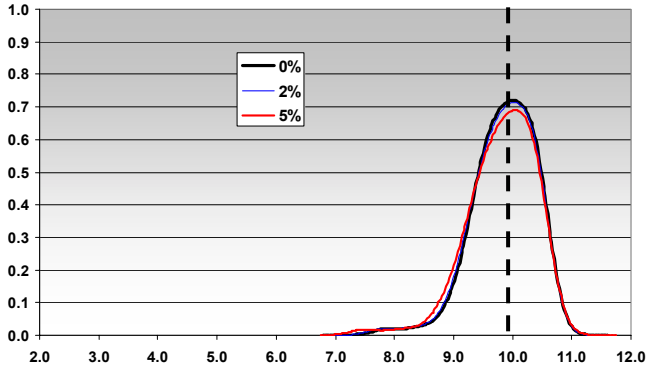
100 Simulations, MLE θ's by % Arbitrary Deviation

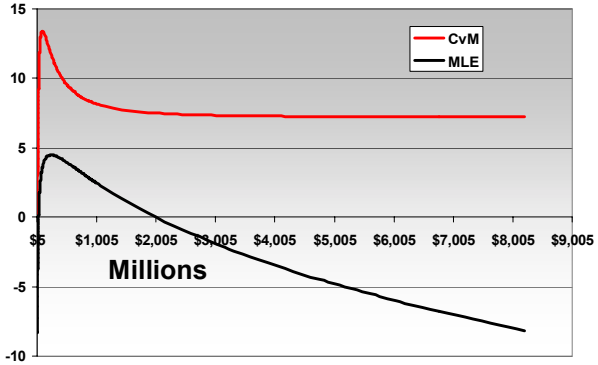
100 Simulations, CvM θ's by % Arbitrary Deviation

(Empirical) Influence Function

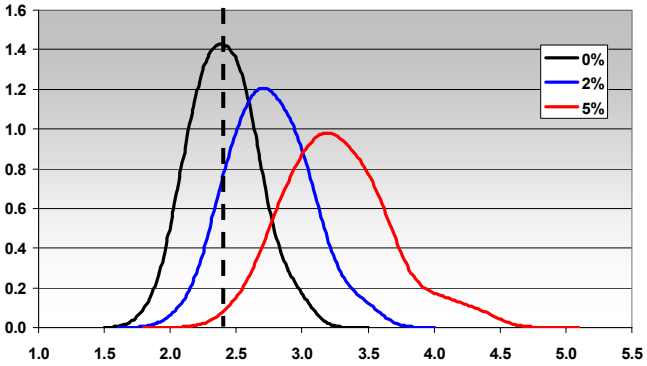
μ = 9.90

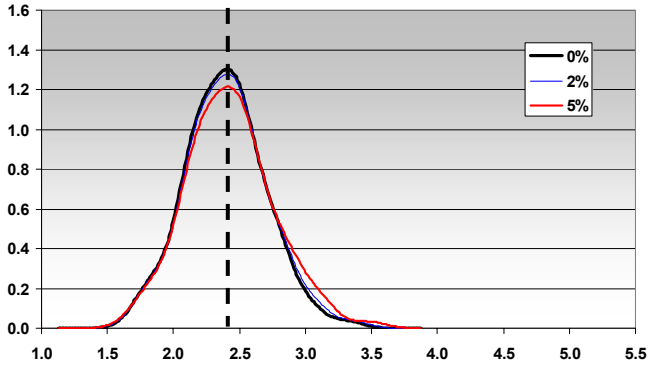


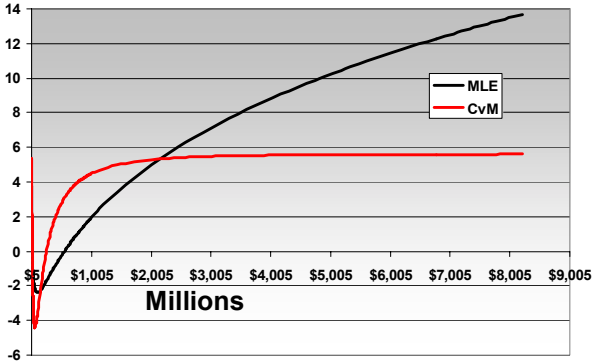




σ = 2.40







American Bankers Association

DataMineIt℠

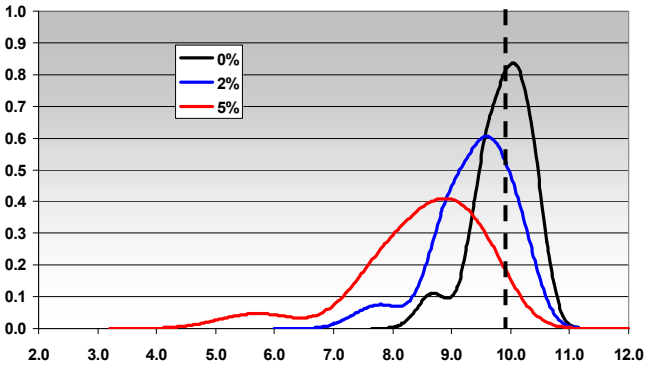
7. Results: Truncated LogNormal (n=250, H=$5k)

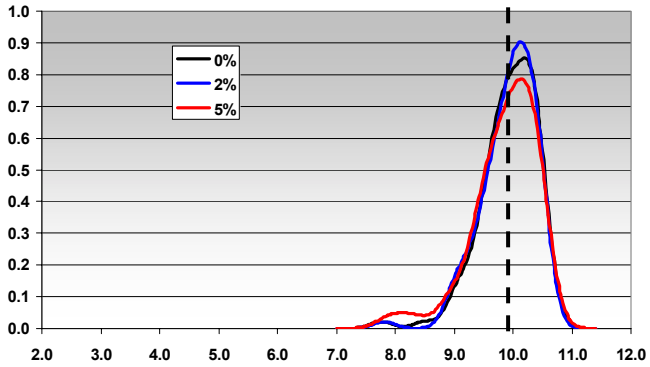
100 Simulations, MLE θ's by % Arbitrary Deviation

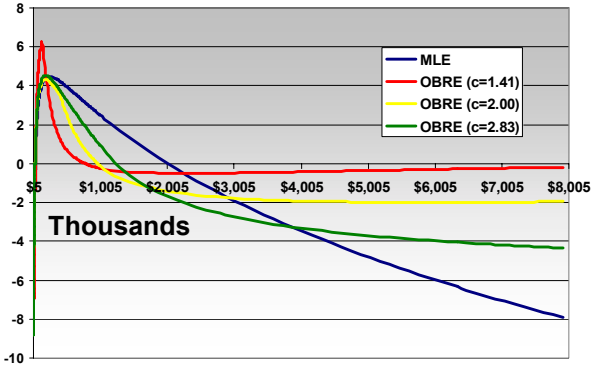
100 Simulations, OBRE (c=2) θ's by % Arbitrary Deviation

(Empirical) Influence Function

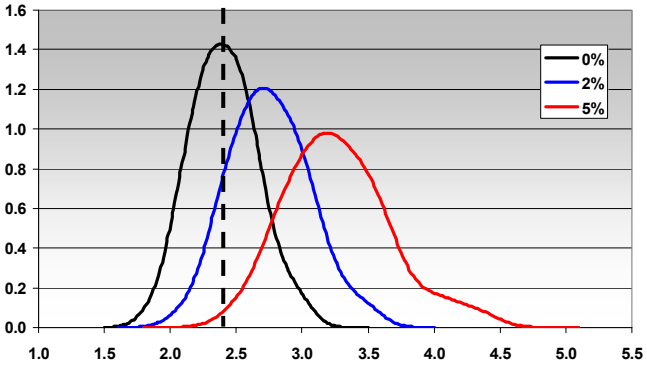
μ = 9.90

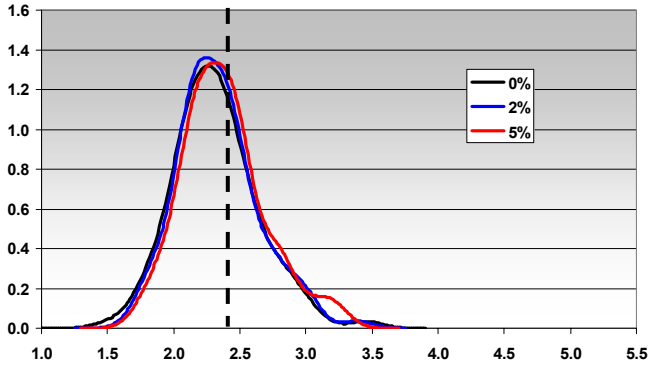


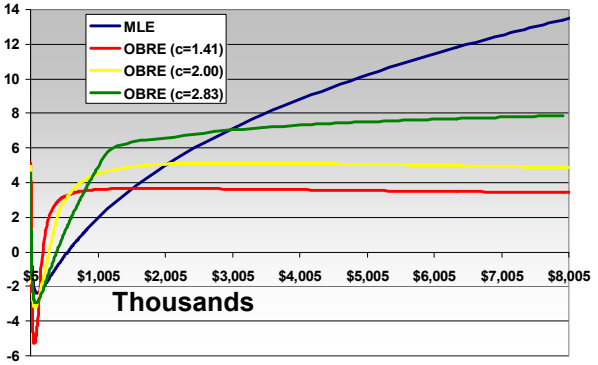




σ = 2.40







American Bankers Association

DataMineIt℠

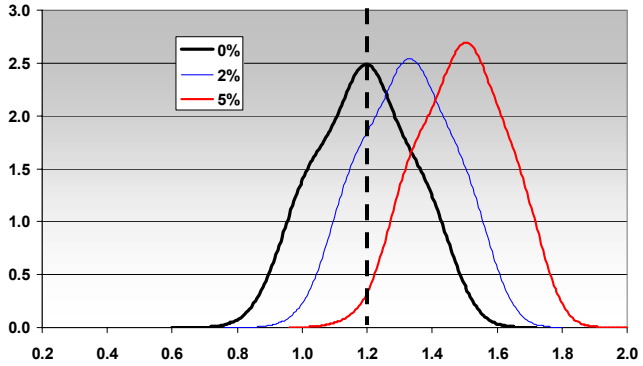
7. Results: Truncated GPD (n=250, H=$5k)

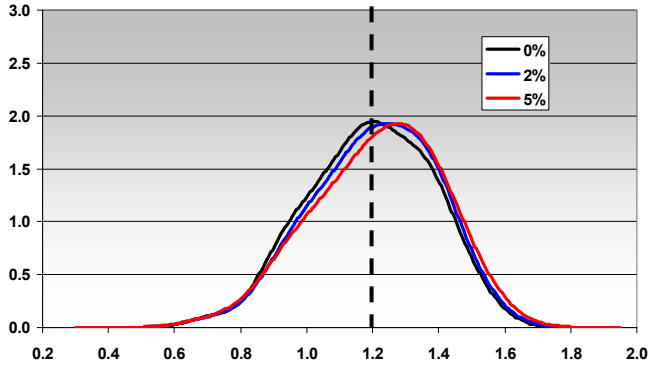
100 Simulations, MLE θ's by % Arbitrary Deviation

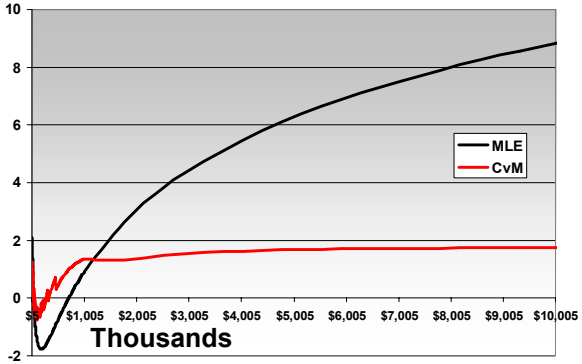
100 Simulations, CvM θ's by % Arbitrary Deviation

(Empirical) Influence Function

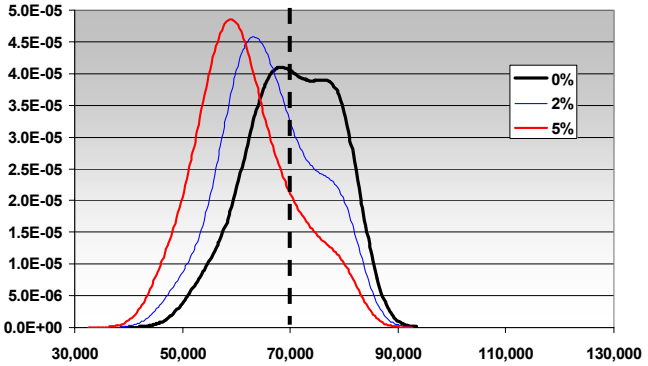
ξ = 1.20

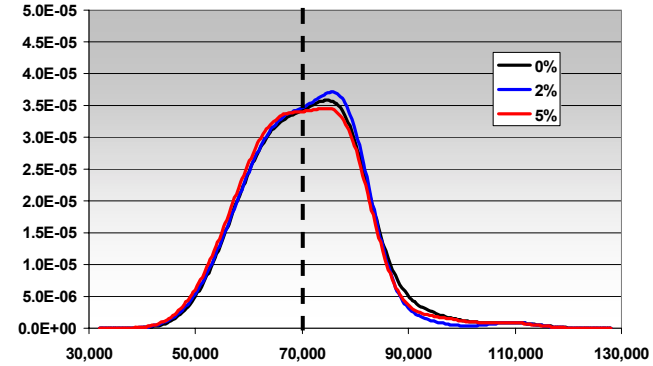


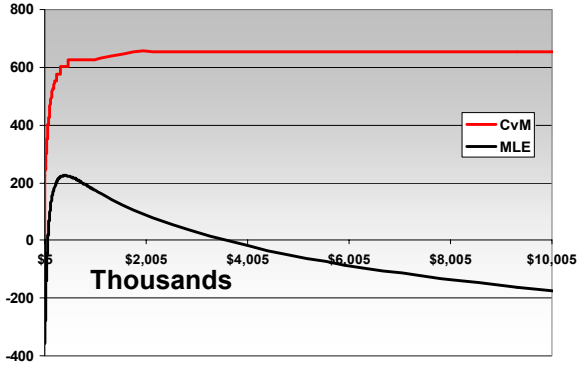




β = 70,000







American Bankers Association

DataMineIt℠

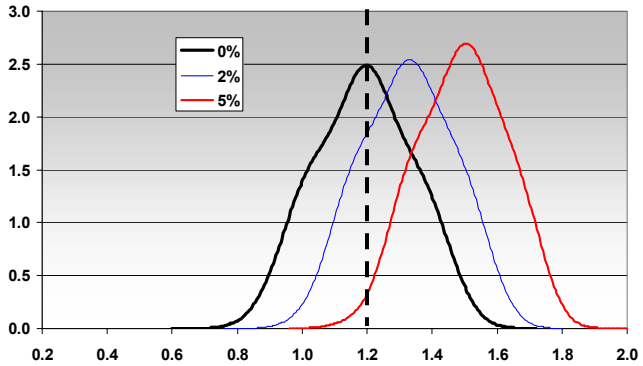
7. Results: Truncated GPD (n=250, H=$5k)

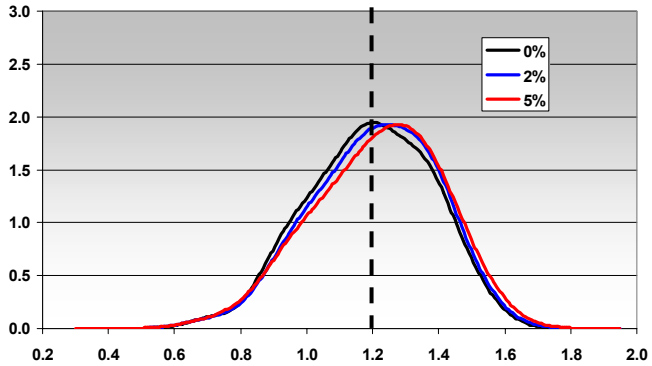
100 Simulations, MLE θ's by % Arbitrary Deviation

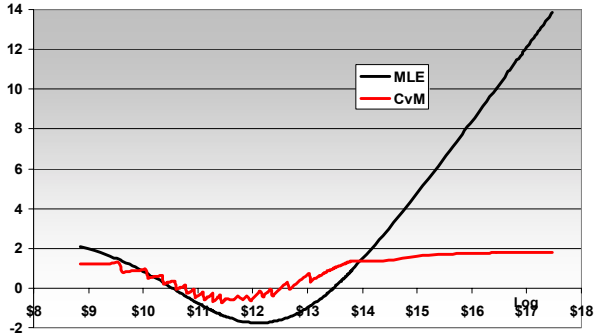
100 Simulations, CvM θ's by % Arbitrary Deviation

(Empirical) Influence Function (Log Scale)

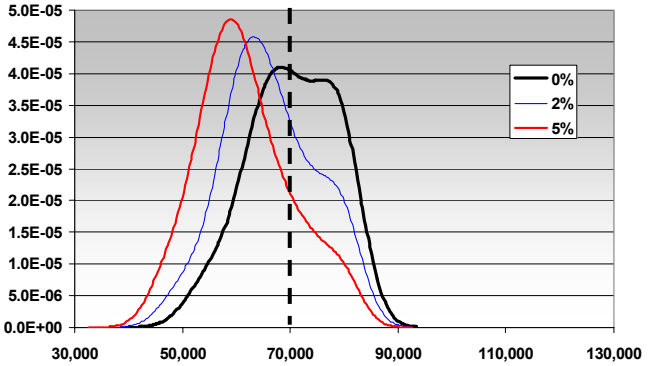
ξ = 1.20

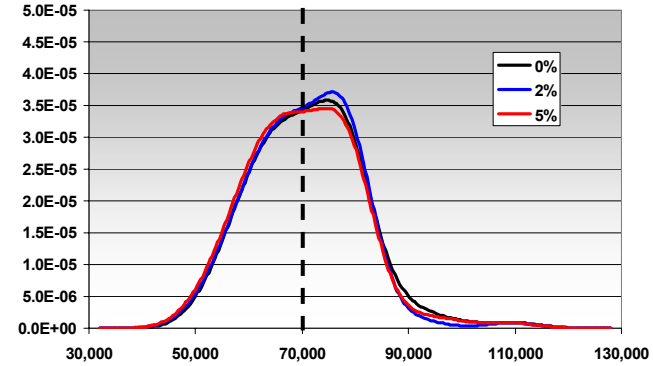


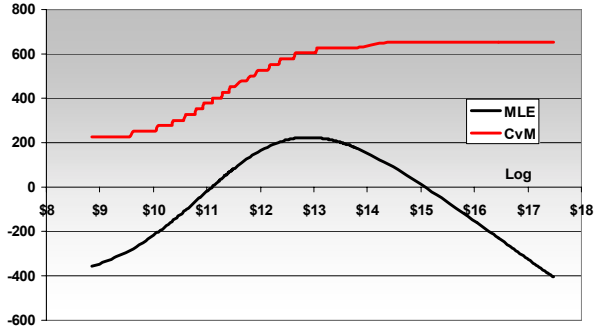




β = 70,000







American Bankers Association

DataMineIt℠

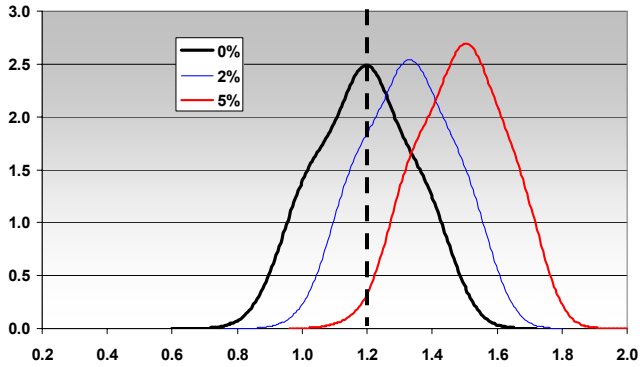
7. Results: Truncated GPD (n=250, H=$5k)

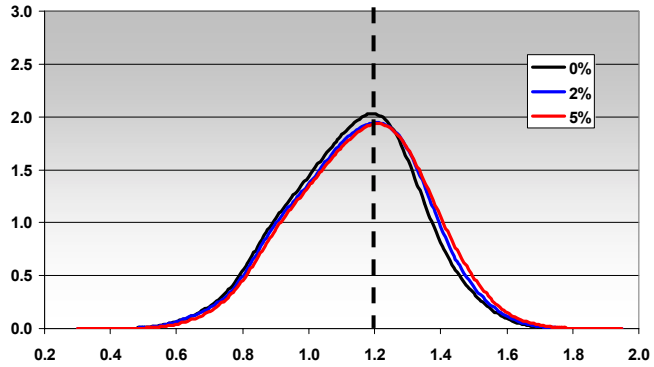
100 Simulations, MLE θ's by % Arbitrary Deviation

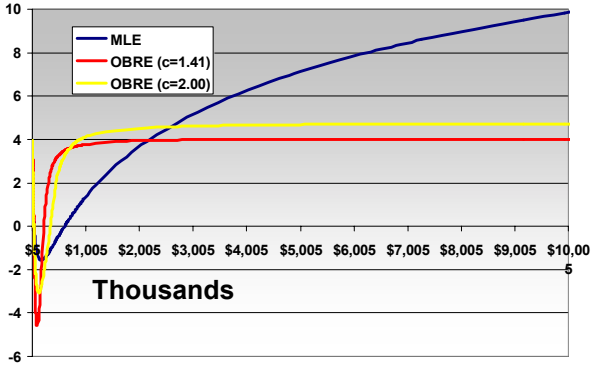
100 Simulations, OBRE (c=2) θ's by % Arbitrary Deviation

(Empirical) Influence Function

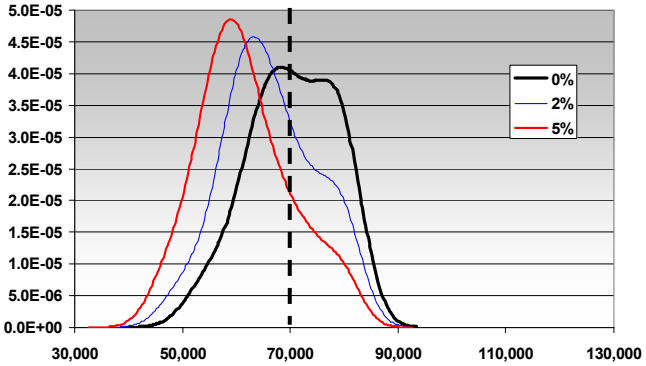
ξ = 1.20

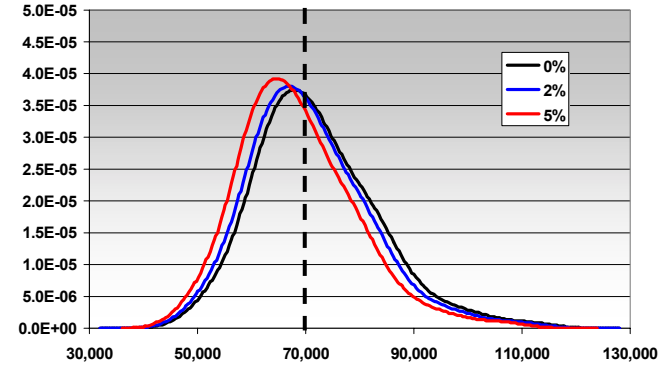


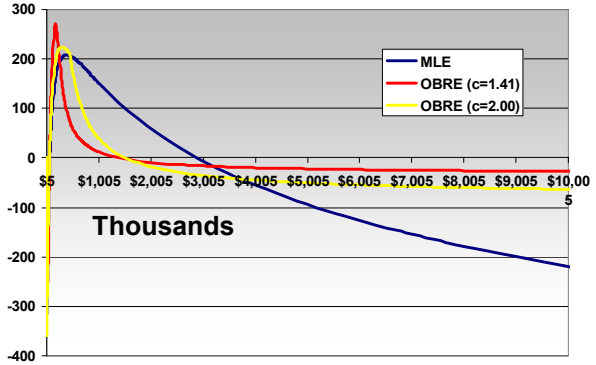




β = 70,000







American Bankers Association

DataMineIt℠

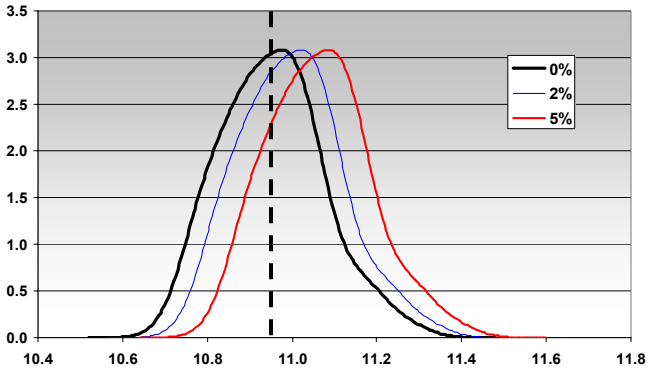
7. Results: LogNormal Distribution (n=250)

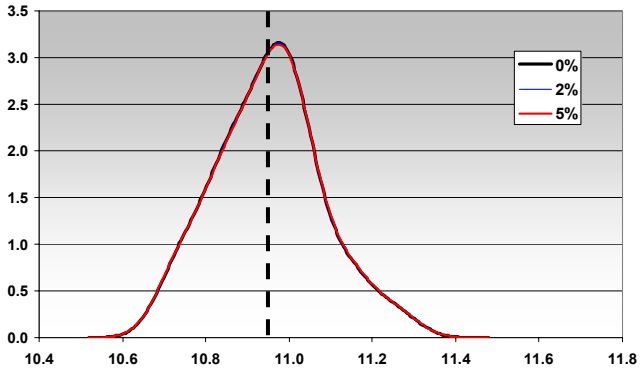
100 Simulations, MLE θ's by % Arbitrary Deviation

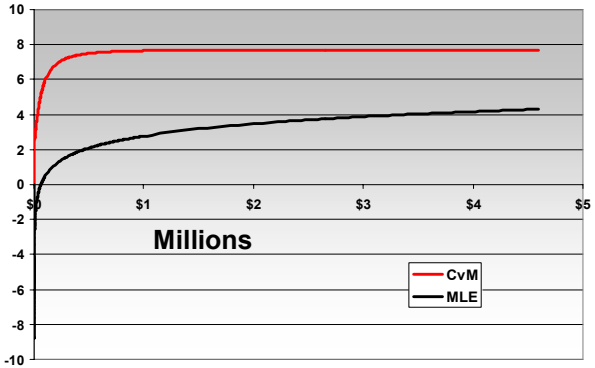
100 Simulations, CvM θ's by % Arbitrary Deviation

(Empirical) Influence Function

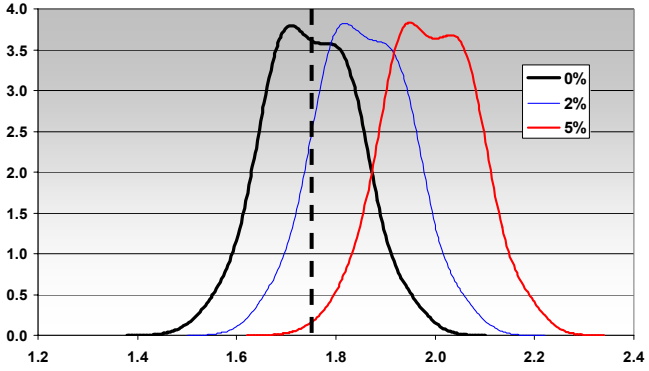
μ = 10.95

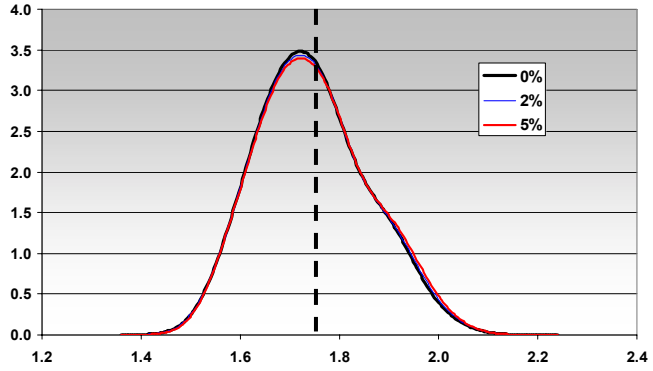


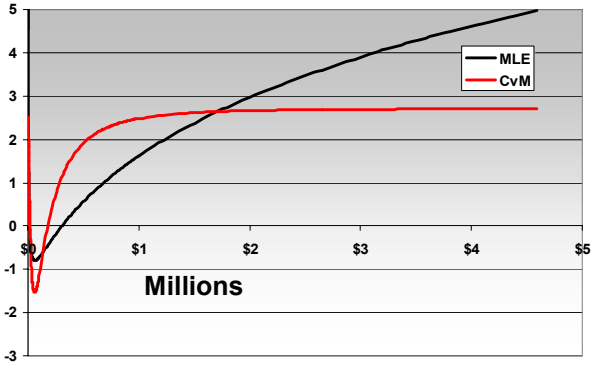




σ = 1.75







American Bankers Association

DataMineIt℠

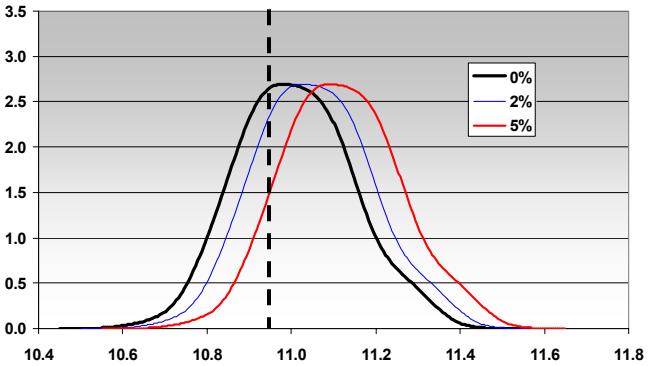
7. Results: Shifted LogNormal (n=250, H=$5k)

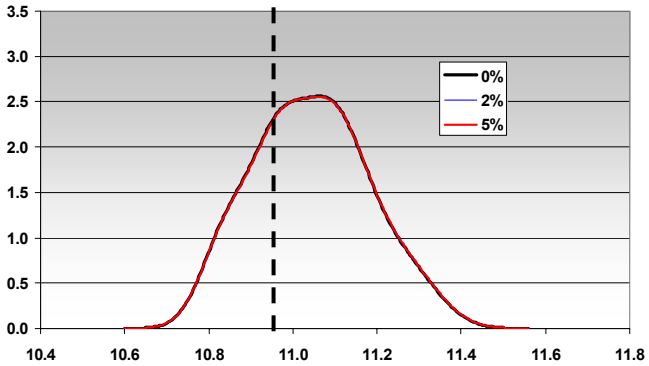
100 Simulations, MLE θ's by % Arbitrary Deviation

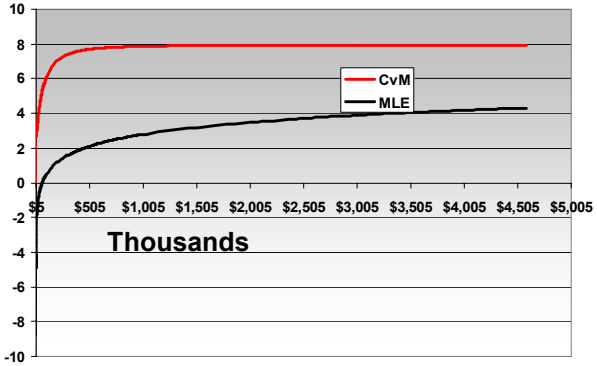
100 Simulations, CvM θ's by % Arbitrary Deviation

(Empirical) Influence Function

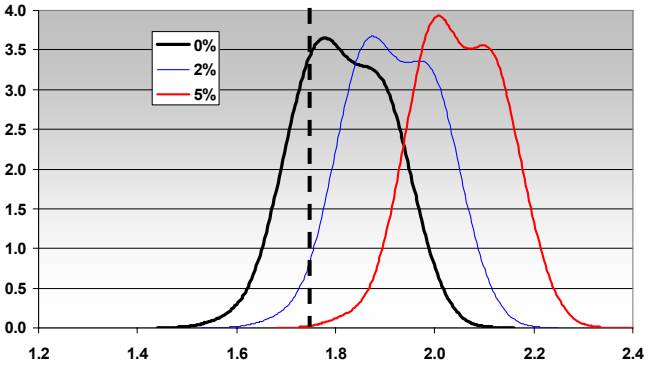
μ = 10.95

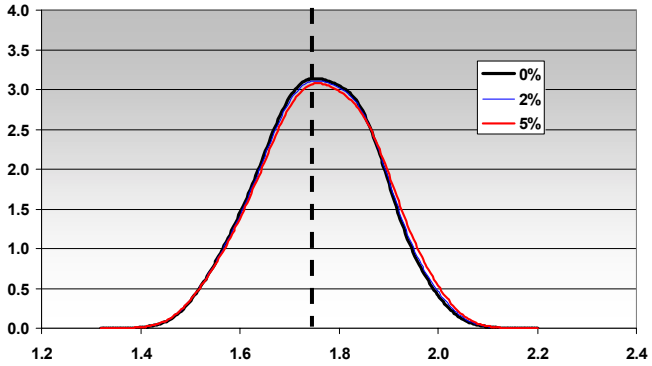


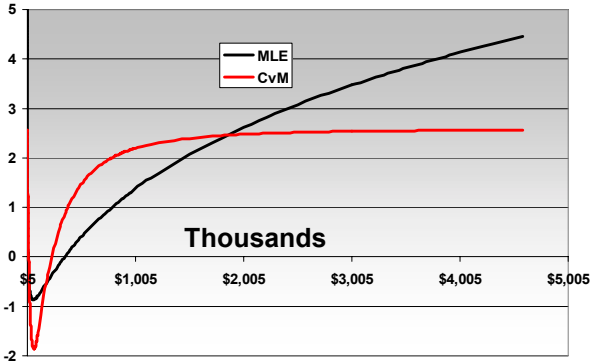




σ = 1.75







American Bankers Association

DataMineIt℠

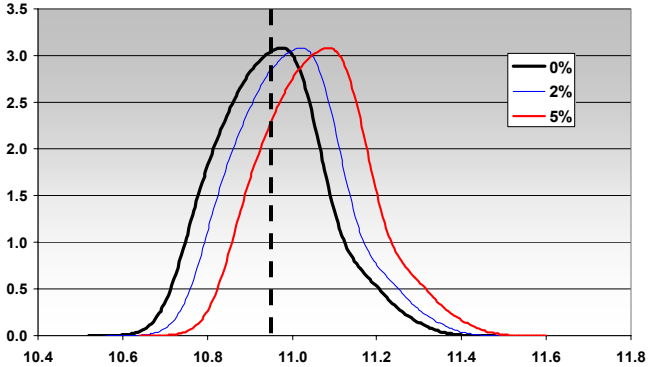
7. Results: LogNormal Distribution (n=250)

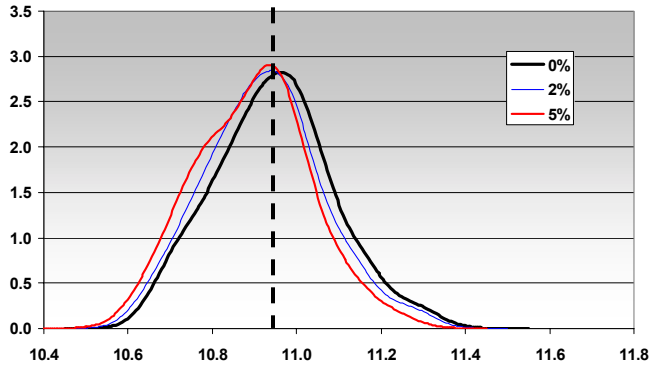
100 Simulations, MLE θ's by % Arbitrary Deviation

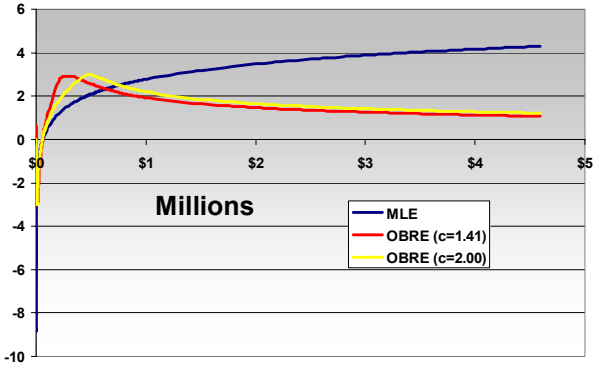
100 Simulations, OBRE (c=2) θ's by % Arbitrary Deviation

(Empirical) Influence Function

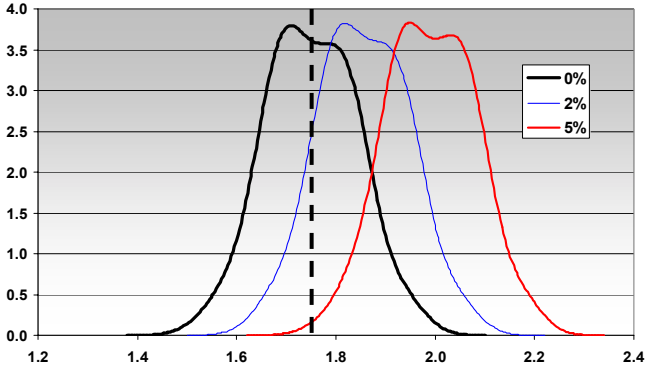
μ = 10.95

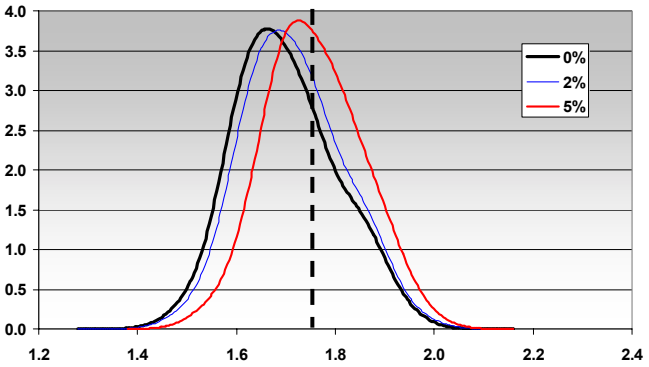


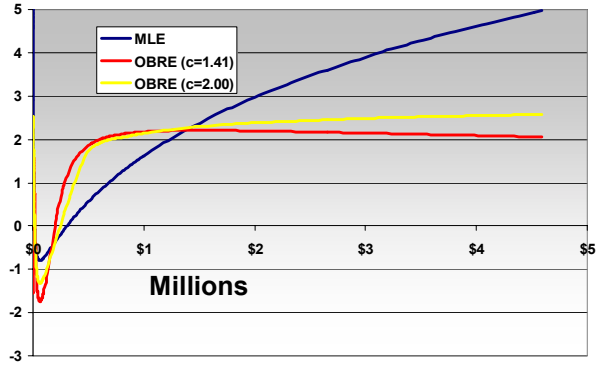




σ = 1.75







American Bankers Association

DataMineIt℠

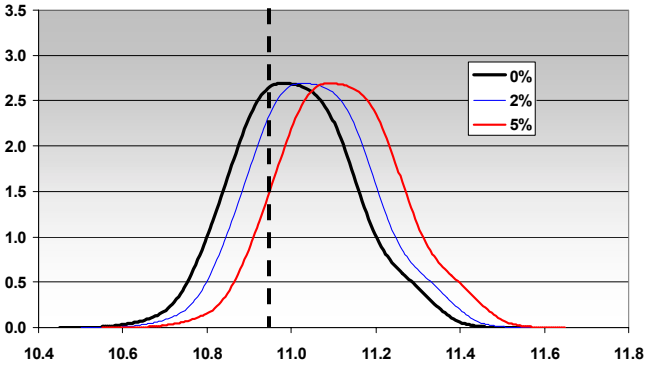
7. Results: Shifted LogNormal (n=250, H=$5k)

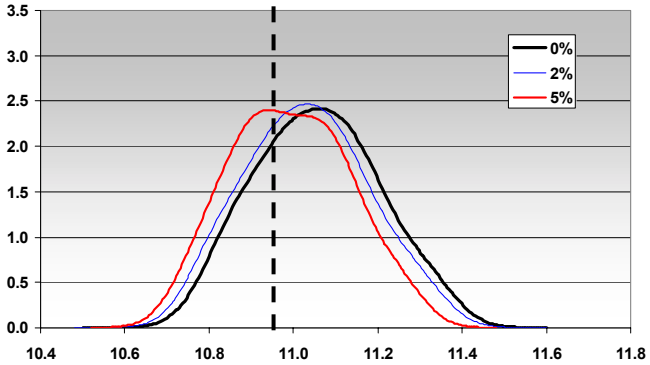
100 Simulations, MLE θ's by % Arbitrary Deviation

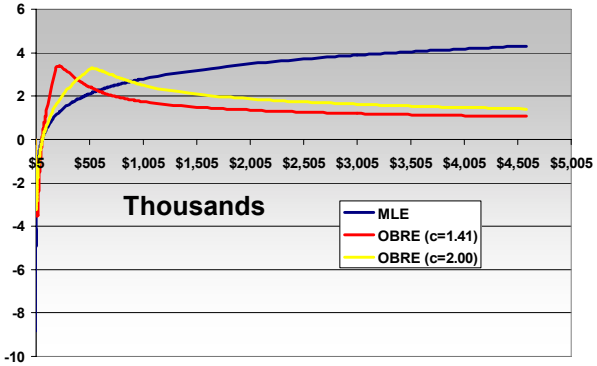
100 Simulations, OBRE (c=2) θ's by % Arbitrary Deviation

(Empirical) Influence Function

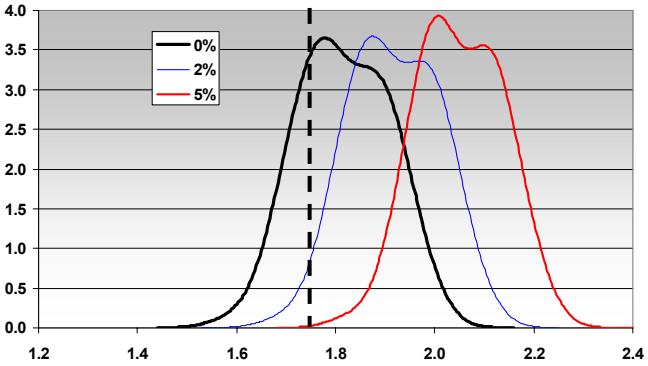
μ = 10.95

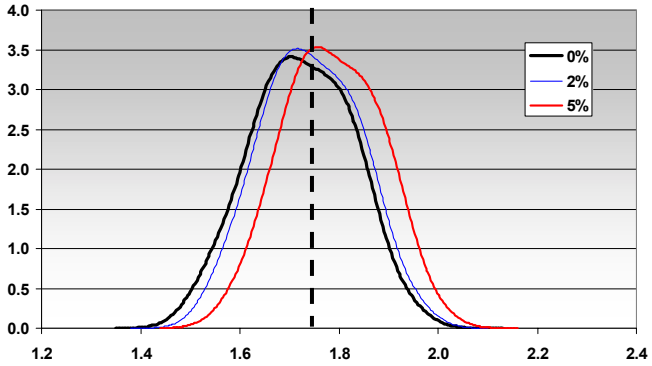


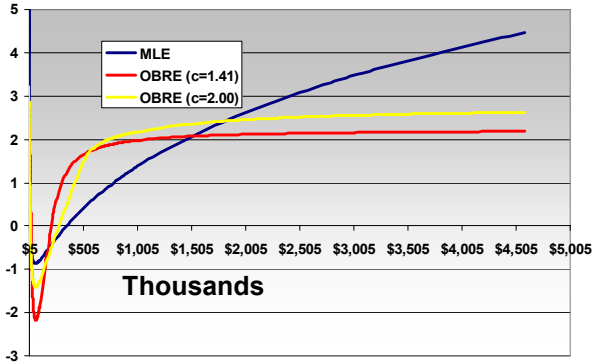




σ = 1.75







American Bankers Association

DataMineIt℠

7. Results: Summary

- **Counter-Intuitive MLE Bias from Small-Sized Deviations:**
  Small aribtrary deviations away from the presumed model (that is, deviations in the left tail) can have very large, disproportionate biasing effects on MLE estimates. This is an analytically derived result of the (LogNormal) IF, not an artifact of sensitivity to simulation assumptions.
- **Notable MLE Bias Under Small % Deviations:**
  MLE estimators are uniformly and strongly biased by even mild deviations (only 2% of all observations) from the assumed severity distribution.
- **Very Strong MLE Bias Under (Left) Truncation for Small % Deviations:**
  MLE bias is especially strong under (left) truncation, both for the LogNormal & GPD.
- **All Analytically Derived IFs Virtually Exactly Match EIFs**
- **(Left) Truncation Induces/Increases Parameter Covariance:**
  For the LogNormal, GPD, Truncated LogNormal, and Truncated GPD distributions, large arbitrary deviation induces positive, negative, negative, and negative covariance between the MLE parameter estimates, respectively. This would appear to be the source of the extreme sensitivity of MLE estimates to truncation often cited in the literature, based on simulations. This is the first study known to this author placing such simulations side-by-side with MLE-IFs derived for truncated distributions.

7. Results: Summary

- **Efficiency – CvM and OBRE Superior to MLE: Under the LogNormal distribution, differences in efficiency amonst the three estimators were negligible. Under the Truncated LogNormal, MLE's greater efficiency under no arbitrary deviation was extremely small, while its INefficiency compared to OBRE and CvM under 2% and 5% deviation was extremely large. For the GPD, not until 5% deviation was inserted did OBRE and CvM become notably more efficient than MLE, which was also true for the Truncated GPD. MLE was notably more efficient under 0% deviation only for β for the Truncated GPD. The winners on the efficiency front clearly were the robust estimators.**
- **Robustness – CvM and OBRE Superior: Unlike MLE, OBRE (c = 2.0) and CvM performed very similarly and very well, exhibiting very good robustness properties with well-bounded EIFs, and virtually none of the bias shown by MLE, even under 5% arbitrary deviations in the (left) truncated models. The one exception to this performance: CvM had difficulty attaining convergence for the Truncated GPD. While this was admittedly the hardest distribution to fit, that also was the point of the exercise, and of using fairly large parameter values in this case. Further scrutiny of this result would be required, with possible focus on starting values and the performance of gaussian quadrature under infinite means.**

7. Results: Summary

- **<u>Shifted LogNormal</u>: Results for the Shifted LogNormal were very similar to those of the LogNormal distribution; for the former compared to the latter, larger MSE and some bias, across estimators, were due to simply fitting the wrong model. But as expected, the robust statistics maintain robustness to arbitrary deviations, whereas MLE is not able to.**

8. CvM and OBRE, Pros and Cons

- **OBRE Advantages**:
  - explicit control over robustness vs. efficiency via the tuning parameter
  - "Optimal" efficiency for given level of robustness obtained (trace of parameter covariance matrix minimized)
  - extremely useful degree-of-deviation weights assigned to each observation
  - generalizable to multivariate regression
- **OBRE Disadvantages**:
  - some coding required, non-trivial implementation / set-up
  - convergence not guaranteed (although no problems in this study)
- **CvM Advantages**:
  - implementation is generally relatively quick and simple
  - statistical performance (unbiasedness, robustness-efficiency tradeoff) appears to be close to "optimal" OBRE performance
  - generalizable to multivariate regression
- **CvM Disadvantages**:
  - possible convergence issues, as noted herein and in the literature (Ergashev, 2008)
  - no explicit control over robustness vs. efficiency (without changing the statistic in non-trivial ways)

9. Potential Limitations of Robust Statistics Generally

- **None of these estimators works perfectly under all conditions. Like all statistical methods, robust estimators have limitations, and their responsible use requires the analyst to remain cognizant of them. These include:**
  - **Starting points – While using MLE starting points did not appear to be an issue in this study for OBRE, this is an issue likely more relevant under more severe sample size constraints. However, this may have been a contributing factor in the one instance that CvM had trouble converging – under a Truncated GPD severity distribution. More robust starting points may be required for smaller sample sizes, or for particularly hard-to-fit data.**
  - **Convergence not guaranteed: For a given set of data, convergence of these algorithms is not guaranteed, although it is worth noting that for heavy-tailed distributions, MLE estimators often, perhaps more often than not, rely on numerical algorithms, too.**
  - **Merely Statistical Estimators: It is important to remember in this OpRisk setting that robust statistics are merely statistical estimators: they do not solve, in and of themselves,**
    - **the inherent difficulties in using distribution parameters to estimate very high quantiles**
    - **data that resist conforming to (mathematically convenient) parametric assumptions**
    - **problems related to initial model selection (where Bayesian methods are strong)**

10. Point-Counterpoint Revisted: Who Wins?

Maximum Likelihood Estimation (MLE):

"MLE does not inappropriately downweight extreme observations as do most/all robust statistics. And focus on extreme observations is the entire point of the OpRisk statistical modeling exercise! Why should we even partially ignore the (right) tail when that is where and how capital requirements are determined?! That's essentially ignoring data – the most important data – just because its hard to model!"

Robust Statistics:

"All statistical models are merely idealized approximations of reality, and OpRisk data clearly violate the fragile, textbook model assumptions required by MLE. Robust Statistics acknowledge and deal with these facts by explicitly and systematically accounting for them, sometimes with weights (and thus, they avoid a bias towards weight=one for every data point). Consequently, under real-world, non-textbook OpRisk loss data, Robust Statistics exhibit less bias, equal or greater efficiency, and far more robustness than does MLE. These characteristics translate into a more reliable, stable estimation approach, regardless of the framework used by robust statistics (i.e. multivariate regression or otherwise) to obtain high quantile estimates of the severity distribution.

10. Point-Counterpoint Revisted: Confirmation

"Estimation of operational risk is badly influenced by the quality of data, as not all external data is relevant, some losses (i.e. 'outliers') may not be captured by the ideal model, and induce bias, and some data may not be reported at all. This can result in systematic over- or under-estimation of operational risk. … robust estimation of the regulatory capital for the operational risk hence provides a useful technique to avoid bias when working with data influenced by outliers and possible deviations from the ideal models." (Horbenko, Ruckdeschel, & Bae, 2010)

"…recent empirical findings suggest that classical methods will frequently fit neither the bulk of the operational loss data nor the outliers well… Classical estimators that assign equal importance to all available data are highly sensitive to outliers and in the presence of just a few extreme losses can produce arbitrarily large estimates of mean, variance and other vital statistics. …On the contrary, robust methods take into account the underlying structure of the data and "separate" the bulk of the data from outlying events, [in – sic] this way avoiding upward bias in the vital statistics and forecasts." (Chernobai & Rachev, 2006)

"Since we can assume that deviation from the model assumptions almost always occurs in finance and insurance data, it is useful to complement the analysis with procedures that are still reliable and reasonably efficient under small deviations from the assumed parametric model and highlight which observations (e.g. outliers) or deviating substructures have most influence on the statistical quantity under observation. Robust statistics achieves this by a set of different statistical frameworks that generalize classical statistical procedures such as maximum likelihood or OLS." (Embrechts & Dell'Aquila, 2006)

11. New Findings, Summary, Recommendations, Next

New Findings:

- **MLE Influence Functions Under Truncation:**
  Influence Functions for MLE severity distribution parameter estimators derived for (left) truncated distributions. This appears to explain the extreme "sensitivity" of MLE estimators of (left) truncated distributions reported in the literature, based on simulations. This is the first paper to present the analytic results side-by-side with confirmatory simulation results.
- **OBRE Under Truncation:**
  OBRE applied to (left) truncated data.* OBRE's performance on these difficult-to-fit distributions, especially GPD with infinite mean, was flawless, and bodes well for its use in this setting.

*** Victoria-Feser & Ronchetti, 1994, compute OBRE using (left) truncated data, but use an EM algorithm rather than the more standard Newton-Raphson algorithm of D.J. Dupuis (1998) to obtain the OBRE estimates.**

11. New Findings, Summary, Recommendations, Next

Summary of Major Findings:

- **MLE loses its good statistical properties under real-world OpRisk loss event data, that is, the moment it steps out of the non-i.i.d. textbook.**
- **MLE exhibits notable bias, and is less efficient and less robust than OBRE and CvM under even modest deviations from the assumed severity distribution(s).**
- **This is especially true under truncated severity distributions, which are prevalent in the OpRisk setting. This is confirmed with analytic derivations of MLE Influence Functions, not simulations alone.**
- **The robust statistics studied herein, namely CvM and OBRE, exhibit good statistical behavior in terms of unbiasedness, efficiency, and robustness under modest deviations from the assumed statistical severity distribution, with and without truncation.**
- **The challenges of OpRisk loss event data appear to be tailor-made for a robust statistics approach, and the results presented herein appear very promising for its application in this setting.**

11. New Findings, Summary, Recommendations, Next

Some Specific Questions to be Answered:

- **Does MLE become unusable under relatively modest deviations from i.i.d., especially for the heavy-tailed distributions used in this setting YES, or are these claims overblown? NO**
- **Do analytical derivations of the MLE Influence Functions for severity distribution parameters support or contradict such claims? NO, THEY SUPPORT THEM Are they consistent with simulation results? YES How does (possible) parameter dependence affect these results? VERY MUCH**
- **Do these results hold under truncation? YES How much does truncation and the size of the collection threshold affect both MLE and Robust Statistics parameter estimates? RESPECTIVELY: VERY BADLY, NOT MUCH/ROBUST**
- **Are widely used, well established Robust Statistics viable for severity distribution parameter estimation? ALL RESULTS INDICATE YES Are they too inefficient relative to MLE for practical use? NO, ACTUALLY BETTER THAN MLE Do any implementation constraints (e.g. algorithmic issues) trip them up, especially under difficult-to-fit distributions (say, with infinite mean)? NO, BUT NEEDS TO BE MONITORED**

11. New Findings, Summary, Recommendations, Next

Recommendations

- **Replace MLE in the OpRisk severity distribution parameter estimation exercise with alternatives robust to modest violations of the assumed models, which by necessity are merely idealized approximations of a non-pristine, somewhat messy data reality. The robust statistics examined herein appear to be very promising candidates for this purpose, and definitely merit further study toward application in this setting. Suggestions are included in "next steps" below.**

Next Steps:

Capital Estimates, Sample Size Study, Variance Reduction via Regression

1. **Capital Estimates:**
   These parameter estimates must be used to obtain high quantiles of the severity distribution for regulatory and economic capital estimation. Initial results via SLA (Single-Loss Approximation, see Böcker & Klüppelberg, 2005), of course using $\xi < 1.0$ for the GPD distributions for which the statistical behavior shown above was identical, indicate notably greater precision of the robust statistics under just 2% deviations, due to bias in the MLE estimates. The robust capital estimates were closer to true required regulatory capital typically by at least several multiples of the true capital value.

11. New Findings, Summary, Recommendations, Next

Next Steps

2. Sample Size Study:

Sizes of units of measure vary, so it is important to know, for any estimation method under consideration, the minimum sample size necessary, under the specific distributional and data circumstances, to provide a parameter and a corresponding quantile estimate with a required level of precision. Some studies have indicated robust results for variants of the estimators examined herein for sample sizes as low as n = 40 (see Horbenko, Ruckdeschel, & Bae, 2010). A thorough "how low can we go" study focusing exclusively on this question under a wide range of scenarios would be a valuable contribution to the possible application of robust statistics in the OpRisk severity modeling setting.

11. New Findings, Summary, Recommendations, Next

Next Steps

3. Variance Reduction – A Multivariate Approach:

As mentioned previously, what may be good estimation statistically, with good relative statistical precision, may not be good enough for reliable, stable capital estimation. The fact that the quantiles that need to be estimated in this setting are so extremely high means that the slightest bias in, or change in variance in, the parameter estimators can dramatically affect the capital estimates in absolute terms. Therefore, variance in parameter estimation needs to be reduced at all costs.

Multivariate Regression may reduce this variance in three ways concurrently by

i) helping to better define units of measure to most efficiently make the tradeoff between homogeneity and statistical power

ii) increasing statistical power by preserving degrees of freedom while simultaneously explaining and accounting for more heterogeneity, and

iii) appropriately handling the problem of time-varying thresholds, which based on the empirical results herein, can have a dramatic effect on parameter estimation

Robust Statistics can and should be applied within a multivariate framework to tackle this difficult challenge of variance reduction.

12. Appendix I

Mean Squared Error: This is the average of the squared devations of sample-based estimate values from the true population value of the parameter being estimated, as shown below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\theta - \hat{\theta}\right)^2 = Variance\left(\theta\right) + \left[\, Bias\left(\theta\right)\right]^2$$

If an estimator is unbiased, bias = 0 and MSE = Variance. "Efficiency" can be defined in slightly different ways, but it is always inversely related to MSE.

The Cramér-Rao Lower Bound: is the inverse of the information matrix – the negative of the expected value of the second-order derivative of the log-likelihood. Because this is the lower bound for the variance of any unbiased estimator, efficiency is usually defined in reference to it, if not in reference to another estimator (in which case it is usually called relative efficiency).

12. Appendix II

For the median, we must use additional results from Hampel et al. (1986) related to L-estimators (of location), which are of the form $T_n(X_1,\ldots,X_n)=\sum_{i=1}^{n} a_i X_{i:n}$, where $X_{1:n},\ldots,X_{n:n}$ is the ordered sample and the a_i are coefficients.

"L" of "L-estimators" comes from "linear" combinations of order statistics. A natural sequence of location estimators is obtained if the weights a_i are generated by $a_i = \int_{[(i-1)/n,\,i/n]} hd\lambda \Big/ \int_{[0,1]} hd\lambda$, where $h:[0,1]\to\Re$ satisfies $\int_{[0,1]} hd\lambda \neq 0$

Under regularity conditions, these estimators are asymptotically normal and the corresponding functional is

$$T(G)=\frac{\int xh\big(G(x)\big)dG(x)}{\int h\big(F(y)\big)dF(y)}$$, which is Fisher consistent with influence function

$$IF(x;T,F)=\frac{\int_{[0,x]} h\big(F(y)\big)d\lambda(y)-\int\left[\int_{[0,t]} h\big(F(y)\big)d\lambda(y)\right]dF(t)}{\int h\big(F(y)\big)dF(y)}$$

where the denominator is nonzero because it equals $\int_{[0,1]} hd\lambda$

Now the median corresponds to $h=\delta_{(1/2)}$, so $\int_{[0,1]} hd\lambda=1$ and $T(G)=\int_{[0,1]} G^{-1}(y)h(y)d\lambda(y)=G^{-1}(1/2)$

So its influence function is $$IF(x;T,F)=\frac{1}{2f\left(F^{-1}(1/2)\right)}\operatorname{sign}\left(x-F^{-1}(1/2)\right)$$

and for standard normal, median=0, $F^{-1}(1/2)=0$, so

$$IF(x;T,F)=\frac{\operatorname{sign}(x-0)}{2f(0)}=\frac{\operatorname{sign}(x)}{2\frac{1}{\sqrt{2\pi}}\exp(-0/2)}=\frac{\operatorname{sign}(x)}{2\frac{1}{\sqrt{2\pi}}\exp(0)}=\frac{\operatorname{sign}(x)}{\frac{2}{\sqrt{2\pi}}}=\operatorname{sign}(x)\sqrt{\frac{\pi}{2}}$$

12. Appendix III

Many important robustness measures are based directly on the IF:

Gross Error Sensitivity (GES) is the supremum being taken over all x where IF exists:

$$\gamma^*(T,F) = \sup_x \left| IF(x;T,F) \right|$$

This measures the worst case (approximate) influence that a small amount of contamination of a fixed size can have on the value of the estimator. If GES is finite, that is, if IF is bounded, the estimator is B-robust ("B" comes from "bias," because GES can be regarded as an upper bound on the (standardized) asymptotic bias of the estimator). Robustifying an estimator typically makes it less efficient, so the conflict between robustness and efficiency is often best solved with Op timal B-robust estimators (OBRE) – estimates which cannot be improved with respect to both GES and asymptotic variance (shown below). So GES is very useful, alongside IF, for comparing two estimators. If, for example, a comparison of the IFs of two estimators leads to ambiguous conclusions, that is, if one estimator's IF has tighter bounds over one range but the other's is tighter over another range, then GES is a useful tool describing which is better under the worst case scenario.

12. Appendix III

Rejection Point: If IF does not exist in some area and is equal to zero, then contamination of points in that area do not have any influence on the estimator at all. The rejection point, then, is defined as

$$\rho^* = \inf\left\{r > 0; IF\left(x; T, F\right) = 0 \text{ when } \left|x\right| > r\right\}$$

Observations farther away than ρ^* are rejected completely, so it is generally desirable if ρ^* is finite. In other words, for estimators with finite rejection point, there will be some point beyond which extreme outlying data points will have no influence on the value of the estimator (because the value of the influence function is zero), and in general, this is a desirable characteristic of an estimator, adding to its robustness against data that deviates notably from the model's assumptions.

12. Appendix III

Empirical Influence Function: The empirical influence function (EIF) naturally corresponds with the IF, and is given by

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

To implement this in practice, EIF is simply a plot of
as a function of *x*, where *x* is the added contamination data point inserted in place of observation . The EIF can be described as an estimation using the original sample, but with only *n* – 1 of the observations, compared to one using the same sample with one additional data point, *x*, the contamination. This also is closely related to the jackknife (the finite sample approximation of the asymptotic variance, treated below, is the jackknife estimator of the variance).

12. Appendix III

Sensitivity Curve: A concept very closely related to the empirical influence function, that is, the non-asymptotic, finite sample IF, is Sensitivity Curves. Analogous to the EIF, these answer the question: how sensitive is the estimator, based on the finite empirical sample at hand, to single-point contaminations at each data point? From Hampel et al. (1986), the sensitivity curve is simply $SC_n(x) = n\left[T_n(x_1,\ldots,x_{n-1},x) - T_{n-1}(x_1,\ldots,x_{n-1})\right]$, which is just a translated and rescaled version of EIF. The functional is applied to two empirical samples (both with one original data point removed): one with a point of contamination, and one without. The difference between the values of the empirical functional, multiplied by *n*, is the sensitivity curve.

Analogously, when the estimator is a functional, then $SC_n(x) = \frac{1}{n}\left[T\left(\left(1-\frac{1}{n}\right)F_{n-1} + \frac{1}{n}\delta_x\right) - T\left(F_{n-1}\right)\right]$, where F_n is the empirical distribution $(x_1,\ldots,x_{n-1})$. In fact, based on the above, $SC_n(x)$ will in many cases converge to $IF(x;T,F)$ asymptotically.

12. Appendix III

Asymptotic Variance and ARE: Based on the IF, an important measure of efficiency is the asymptotic variance, from which the asymptotic relative efficiency (ARE) directly can be calculated. The ARE is simply a measure of the relative size of the variances of two estimators, telling us which is more efficient.

$$V(T,F) = \int IF(x;T,F)^2 \, dF(x)$$

$$ARE_{T,S} = V(S,F) / V(T,F)$$

Understanding the (relative) efficiency of an estimator is especially important within the framework of robust statistics, because some degree of efficiency typically is lost when estimators are made robust. Knowing the extent of efficiency loss is important, because we want estimators that are both robust and efficient, and these are competing criteria by which we need to compare estimators, under different distributions and against each other. Designing estimators to be OBRE (optimally *B*-robust estimators), for example, requires finding estimators that simultaneously can be made no more efficient, and no more robust, and to do this requires knowing how efficient and robust an estimator is.

12. Appendix III

Change-in-Variance Sensitivity: The "change-in-variance" sensitivity assesses how sensitive is the estimator to changes in its asymptotic variance due to contamination at *F*. The denominator of CVS is the asymptotic variance (see section on *M*-estimators above for a definition of ψ), and the numerator is the derivative of the asymptotic variance when contaminated.

$$CVS(\varphi, F) := \sup\left\{\frac{CVF(x;\varphi,F)}{V(\varphi,F)}; x \in \Re \setminus C(\varphi) \cup D(\varphi)\right\}$$ where the change-in-variance function is

$$CVF(x;\varphi,F) = \frac{\partial}{\partial\varepsilon}\left[V\left(\varphi,(1-\varepsilon)F + \varepsilon\left(\frac{1}{2}\delta_x + \frac{1}{2}\delta_{-x}\right)\right)\right]_{\varepsilon=0}$$ for continuous

ψ, for which no delta functions arise. The above is valid for all *M*-estimators. If CVS is finite, *T* is *V*-robust ("V" is for Variance). *V*-robustness is stronger than *B*-robustness: if an estimator is V-robust, it must also be *B*-robust (and if an estimator is not *B*-robust, then it is not *V*-robust). Note that unlike IF, only large positive values for *CVF*, not large negative values, point to nonrobustness.

12. Appendix III

Local Shift Sensitivity: The point of "local shift sensitivity" is to summarize how sensitive the estimator is to small changes in the values of the observations; in other words, how much is the estimator affected by shifting an observation slightly from point x to point y? When the "worst" effect of this "wiggling" is obtained, and it is standardized, the local shift sensitivity is defined as

$$\lambda^* = \sup_{x \neq y} \left| IF\left(y;T,F\right) - IF\left(x;T,F\right) \right| \Big/ \left| y - x \right|$$

This helps to evaluate how sensitive an estimator is to changes in the data, all else equal. And this is relevant in this setting because loss data does change from quarter to quarter, if financials are restated, litigation is settled, etc. So this is an important tool for assessing the robustness of a particular estimator, and can be used in simulation studies to compare the behavior of multiple estimators under such data changes.

12. Appendix III

Breakdown Point: While the IF and its related summary values are related to local robustness, describing the effects of a(n infinitesimal) contamination at point *x*, the "breakdown point" is a measure of global robustness – it describes the global reliability of an estimator by asking, up to what percentage of the data can be contaminated before the estimator stops providing valuable information? The asymptotic contamination breakdown point of the estimate *T* at *F*, denoted ε^*, is the largest $\varepsilon^* \in (0,1)$ such that for $\varepsilon < \varepsilon^*$, $T\left((1-\varepsilon)F + \varepsilon H\right)$ remains bounded as a function of H and also bounded away from the boundary of θ.

Analogously, the finite sample breakdown point ε^* of the estimator T_n at the sample $(x_1,\ldots,x_n)$ is given by

$$\varepsilon_n^*(T_n; x_1,\ldots,x_n) := \frac{1}{n}\max\left\{m; \max_{i_1,\ldots,i_n} \sup_{y_1,\ldots,y_n} \left|T_n(z_1,\ldots,z_n)\right| < \infty\right\},$$

where the sample $(z_1,\ldots,z_n)$ is obtained from the sample $(x_1,\ldots,x_n)$ by replacing the *m* data points $(x_{i_1},\ldots,x_{i_m})$ by arbitrary values $(y_1,\ldots,y_n)$.

The mean, for example, has asymptotic breakdown point and finite sample breakdown point, respectfully, of $\varepsilon^* = 0$ and $\varepsilon^* = 1/n$, because a single observation with value = arbitrarily large (i.e. ∞) renders its values meaningless. In constrast, those of the median are $\varepsilon^* = 0.5$, and $\varepsilon^* = 1/2$ for an even *n* and $\varepsilon^* = (n-1)/2n$ for odd *n*, respectfully, which is far more robust than the mean.

12. Appendix IV

LogNormal Derivatives:

for $0 < x < \infty;\ 0 < \mu < \infty;\ 0 < \sigma < \infty$

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}$$

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

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

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

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

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

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

12. Appendix IV

LogNormal Derivatives (for (left) Truncated case):

Due to Leibniz's Rule, these derivatives can be moved inside these integrals.

$$g(x;\mu,\sigma)=\frac{f(x;\mu,\sigma)}{1-F(H;\mu,\sigma)}$$

$$G(x;\mu,\sigma)=1-\frac{1-F(x;\mu,\sigma)}{1-F(H;\mu,\sigma)}$$

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

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

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

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

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

12. Appendix IV

Generalized Pareto Distribution Derivatives:

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

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

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

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

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

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

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

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

12. Appendix IV

Generalized Pareto Distribution Derivatives (for (left) Truncated Case):

Due to Leibniz's Rule, these derivatives can be moved inside these integrals.

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

$$g(x;\mu,\sigma) = \frac{f(x;\mu,\sigma)}{1 - F(H;\mu,\sigma)}$$

$$G(x;\mu,\sigma) = 1 - \frac{1 - F(x;\mu,\sigma)}{1 - F(H;\mu,\sigma)}$$

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

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

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

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

12. References

- Alaiz, M., and Victoria-Feser, M. (1996), "Modelling Income Distribution in Spain: A Robust Parametric Approach," *TDARP Discussion Paper No. 20*, STICERD, London School of Economics.
- Böcker, K., and Klüppelberg, C. (2005), "Operational VaR: A Closed-Form Approximation," *RISK Magazine*, 12, 90-93.
- Cope, E, G. Mignola, G. Antonini, and R. Ugoccioni, "Challenges and Pitfalls in Measuring Operational Risk from Loss Data," *The Journal of Operational Risk*, Vol. 4, No. 4, 3-27.
- Dupuis, D.J. (1998), "Exceedances Over High Thresholds: A Guide to Threshold Selection," *Extremes*, Vol. 1, No. 3, 251-261.
- Ergashev B., (2008), "Should Risk Managers Rely on the Maximum Likelihood Estimation Method While Quantifying Operational Risk," *The Journal of Operational Risk*, Vol. 3, No. 2, 63-86.
- Grimshaw, S. (1993), "Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution," *Technometrics*, Vol. 35, No. 2, 185-191.
- 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.
- Horbenko, N., Ruckdeschel, P. and Bae, T. (2011), "Robust Estimation of Operational Risk," *The Journal of Operational Risk*, Vol.6, No.2, 3-30.
- Huber, P.J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73-101.
- Huber, P.J. (1977), *Robust Statistical Procedures*, SIAM, Philadelphia.
- Huber, P.J. (1981), *Robust Statistics*, John Wiley and Sons, Inc.
- Stefanski, L., and Boos, D. (2002), *The American Statistician*, Vol. 56, No. 1, pp.29-38.
- Victoria-Feser, M., and Ronchetti, E. (1994), "Robust Methods for Personal-Income Distribution Models," *The Canadian Journal of Statistics*, Vol.22, No.2, pp.247-258.