Weighted Likelihood Estimator for Operational Risk data:

*Improving the accuracy of capital estimates by robustifying MLE*

Andrea Colombo, Alessandro Lazzarini, Silvia Mongelluzzo

*KPMG Advisory Italy*

Frankfurt 19.03.2015
Disclaimers & Copyright

- The views expressed in this paper are exclusively those of the authors and do not necessarily reflect the viewpoints of KPMG Advisory S.p.A or any other organization.

- The material cannot be copied, modified, distributed or displayed without the authors' explicit written permission.
Objectives of the (applied) research

• The **severity parameters** play a crucial role in the OpRisk capital estimates for AMA banks. Due to the intrinsic features of OpRisk data, characterized by the so-called “isolated-data”, standard estimators like the Maximum Likelihood Estimator (MLE) often suffer for **instability**, with dramatic consequences on reliability of capital estimates.

• Within the operational risk modelling, some **robust alternatives to MLE have already been proposed**. However, their practical use within the OpRisk modelling framework for AMA banks seems to deserve only a marginal role. This is likely the effect of the **concern for introducing additional complexity** - both practical and conceptual - in AMA models, e.g. related to additional tuning parameters to be set, usually on a expert judgement basis.

• We propose the use of a **robust generalization of MLE** for the modelling of operational loss data, based on **Weighted Likelihood Estimator (WLE)**. Beside its robustness, its main properties include:
  - The **bias of the VaR under “real life” scenarios on the dataset is lower** than that based on MLE.
  - It retains some **desirable statistical properties**, including consistency and the fact that it can be constructed from applying MLE to a “transformed” sample.
  - It is **simple both to understand and to implement**.
  - It is **not computationally burdensome**.
Agenda

• Motivation
• Background on estimation procedures
• Weighted Likelihood Estimator
• Properties
• Empirical Analysis
• Further directions of research
• Conclusion
• References
Motivation
Isolated data and the risk of overestimation

- **Well populated sample**: Let us consider a sample of size 200 from a GPD with shape parameter \( \xi = 0.9 \) and scale parameter \( \beta = 1 \), with:
  - 100 completely-random drawn observations
  - 100 observations drawn using a stratified sampling scheme such that there is a draw from each probability interval of wide 1%, i.e. one draw from the interval (0%,1%),...(99%,100%).
- "Perturbation": The sample is partitioned into blocks of 5, 10 and 20 observations each. Parameters estimation (based on MLE) is implemented by deleting only one single block each time, i.e. creating gaps of, respectively, about 2.5%, 5% and 10% between consecutive percentiles.

*The experiment has been repeated 100 times and the curves show the mean of the estimated VaR.

*The biasedness is increasing with the size of the block*, that is to say, with the magnitude of the gap. Furthermore, the magnitude of the upward biases is generally higher than the downward biases, e.g. with block detection of size 20, the peak in the upward biases (+314% at the 70th percentile) is more than 3 times the peak in the downward bias (-99% at the right-extreme percentile).

The results of this experiment show that there is:

- an **upward bias** when removing blocks of observations between the 20th/30th and the 90th percentiles.
- a **downward bias** when removing blocks of observations below the 20th/30th percentile or above the 90th percentile, due to a reduction of the variance (i.e. sample shows a narrower support than population).
- **No bias** when deleting blocks around 20th/30th and 90th percentile.
Background on estimation procedures

OpRisk Modelling context

- The most popular method to model OpRisk data is based on the Loss Distribution Approach (LDA), which relies on the quantification, for each Unit of Measure, of the annual aggregate loss obtained by combining the severity distribution with the frequency distribution over a specific time horizon, e.g. one-year.
- Let $X_i, i \geq 1$ be i.i.d. single losses with (severity) density function $f_X(x; \theta)$ and let $N$ be the (random) number of events.
- The annual aggregate loss, denoted by $Y$, is given by $Y = \sum_{i=1}^{N} X_i$.

MLE

- MLE is at the first go and the most used technique for deriving estimators of OpRisk data.
- The estimate for $\theta$ is obtained by: $\hat{\theta} = \arg\min\left(\sum_{i=1}^{N} \log(f_X(X_i; \theta))\right)$.
- If data are i.i.d. and the model is correctly specified, MLE is consistent, asymptotically normal and asymptotic efficient. Relaxing these regularity conditions, no guarantee exists on the MLE properties.
- In real OpRisk applications, these conditions are not usually satisfied, yielding often to quite sensitive estimates to small variations of the data, especially in small data set characterized by isolated data structure and/or contamination. Upward bias is also a likely issue.

ROBUST ESTIMATORS

- Current regulations allows AMA bank to adopt robust estimator provided that it can demonstrate that its use does not underestimate the risk in the tail (EBA/CP/2014/08, art. 23, and BCBS 196, par. 205).*
- The development of robust estimators within the statistical literature is increasing. Within the operational risk modelling, several articles suggest the use of robust estimators:
  - Optimally-robust procedures based on shrinking neighborhood approach (e.g., Horbenko et al., 2011), such as the Most Bias Robust Estimator (MBRE), Optimal MSE Estimator (OMSE). They aim to minimize either the maximum MSE or the maximum bias in a neighborhood around the model.
  - Optimal B-Robust Estimator - OBRE (e.g., Opdyke and Cavallo, 2012). OBRE is a M-estimator which is B-robust (i.e. IF bounded) and whose estimation procedure relies on a weighted standardized score function.
  - Estimators based on quantile distance (e.g., Ergashev, 2008; Lehérissé and Renaudin, 2013), where the estimates are obtained by minimizing different types of measure of the distance between empirical and theoretical quantiles.

* The normative defines robust estimator a generalization of classical estimator having good statistical properties for a whole neighborhood of the unknown underlying distribution of the data (EBA/CP/2014/08, art. 23, point 7).
Weighted Likelihood Estimator - In a nutshell

The literature offers already several robust and attractive options for OpRisk data, exhibiting also good theoretical properties. So, a natural question arises: since every OpRisk practitioner is aware of the extreme sensitivity of the capital estimates to small changes of the data, why these methods are not usually applied by AMA banks?

In our experience, three are the main reasons:

- the available options are generally more complex (than standard estimators like MLE) both to understand and to implement, often requiring a great computational burden;
- most of them rely on a tuning parameter, whose choice can be somewhat arbitrary;
- “innovations” require time to be internalized within an existing framework.

For limiting these weaknesses, we propose a new and simple robust estimator for modelling OpRisk data, based on the Weighted Likelihood Estimator (WLE) based on Choi et al. (2000).

**IDEA:**

- **MLE approach:** each observation contributes to the objective likelihood function with “just” its own likelihood, meaning that implicitly we are assigning equal weights to all the likelihood contributions.

- **WLE approach:** it assigns to each observation-specific likelihood contribution a different weight, which depends on the likelihood of the corresponding observation, increasing (reducing) those likelihood contributions associated with observations with higher (lower) “likelihood”. In particular, the more (less) an observation is consistent with the assumed model, the higher (lower) is the weight for its corresponding likelihood contribution and - indirectly - for the observation itself.

  - This mechanism is called **tilting** and basically it represents an intensifying mechanism on the likelihood contribution, making bigger the role of the more likely observations and lower that of the less likely observations.
Following the procedure defined in Choi et al. (2000), the WLE is obtained based on the following setting. Let:

- $X_1, \ldots, X_n$ be an i.i.d. sample from the density function $f_X(\cdot; \theta)$.
- $p = (p_1, \ldots, p_n)$ be the vector of weights (depending on the parameters and the data), expressing a multinomial distribution.
- $t = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ be the vector of equal weights, representing an uniform distribution.

The WLE for $\theta$, denoted with $\hat{\theta}$, is obtained by solving the following problem:

$$
\min \left( \sum_{i=1}^{n} p_i \cdot \log(f_X(X_i; \theta)) \right) \quad \text{subject to the constraint that } D_{KL}(p, t) = \delta \text{ and } \sum_{i=1}^{n} p_i = 1
$$

where $D_{KL}(p, t) = \sum_{i=1}^{n} p_i \cdot \log(n \cdot p_i)$ is the Kullback Leibler divergence of $p$ from the uniform distribution $t$.

\textbf{WLE is the estimator which maximizes the weighted likelihood function whose weights have a fixed "distance" with respect to the uniform weights.}

It can be easily proven that, solving the corresponding Lagrangian problem, the functional form of the weights is given by:

$$
p_i \propto f_X(X_i; \theta)^{c(\theta)}
$$

Where $c(\theta)$ is a (scalar) function of the data and the parameters $\theta$. This means that the weight associated with each likelihood contribution is proportional to the density evaluated in the corresponding data point.
Weighted Likelihood Estimator
Insights into the tilting parameter

• The tilting parameter, $\delta$, can always be reparametrized as follows: $\delta = -\log(1-\varepsilon)$, with $\varepsilon$ in $[0,1]$.

• Let us consider the vector assigning equal weights to the first $n-m$ observations and 0 to the last $m$ observations:

$$\tilde{p} = (\tilde{p}_1, ..., \tilde{p}_n) = \left(\frac{n-m\,\text{obs}}{n-m}, ..., \frac{m\,\text{obs}}{n-m}, 0, 0, ..., 0\right)$$

The Kullback Leibler divergence of $\tilde{p}$ from the uniform distribution $t$ is given by:

$$D_{KL}(\tilde{p}, t) = -\log(1 - \varepsilon) \quad \text{with} \quad \varepsilon = \frac{m}{n}$$

It follows that the calibration of tilting parameter, $\delta$, can be based on the calibration of $\varepsilon = \frac{m}{n}$, which can be interpreted as the proportion of outliers in the sample against which the proposed tilting approach is robust, according to the breakdown point concept (Choi et al., 2000).

**Remark:** Since the weights are proportional to the density function (i.e. a continuous function with values always strictly positive on the support), all the likelihood contributions will receive some positive weights. The amount of tilting is split among all the observations, as the weights of outliers / isolated data cannot be set equal to zero, but they will just receive a lower weight than with the MLE approach.

---

We believe that - for categories of Operational Risk (so called Unit of Measure) that mostly affect OpVaR measures – is reasonable assuming (after visual inspection of data) that a data gap exists in the tail and as a consequence there is at least one isolated data, so justifying assumption $m=1$.

However, considered the strong regulatory challenge associated with the use of non standard estimators, a deeper investigation about how to calibrate $m$ is appropriate, at least for benchmarking, since the “ideal” $m$ is likely to vary across different Operational Risk datasets.
Properties (1/2)

Depending on the value of the tilting parameter, $\delta$, the WLE holds different properties:

- If $\delta = 0$, then all the weights are equal and the WLE reduces to MLE, following in a **consistent** and **asymptotically efficient estimator** (under regular conditions).

- If $\delta > 0$, then the estimator is not anymore the optimal estimator (under regular conditions), but it will gain in **robustness**, in the sense that the inclusion of single peculiar data points will not dramatically affect the capital estimates, yet still maintaining good properties.

WLE has the following properties:

**Consistency**

- Setting the tilting parameter $\delta = -\log(1-m/n)$, for each $m$ fixed:

$$\lim_{n \to \infty} \delta = 0 \Rightarrow \text{WLE converges to MLE} \Rightarrow \text{WLE is consistent}$$

**Asymptotic normality**

- Under regularity conditions, WLE is asymptotically normally distributed with asymptotic variance given by the variance of the influence function of WLE divided by $n$. 
Properties (2/2)

Robustness

• Through simulation studies, WLE shows that it is able to stabilize VaR estimation much better than MLE.

In order to evaluate the robustness of WLE, several analysis have been implemented, mainly by adding "extreme" losses to the original dataset and analyzing their impact on the VaR.

The following picture shows results of the same exercise illustrated in slide 3, but using WLE instead of MLE:

*Pattern of the (mean) VaR associated with block-deletion (different size) – WLE*

The pattern of the changes is similar to the one observed using MLE but their magnitudes are smaller.

We can observe that:

• **Dramatically lower maximum upward bias for WLE compared to MLE:** The peak for WLE is +119% vs. +314% of MLE for blocks of 20 observations.

• **Same maximum downward bias for both MLE and WLE:** The downward bias observed when removing blocks of observations in the upper extreme of the support of the distribution is similar to the one observed with MLE.
Empirical analysis - Setting

Data

- We consider a sample of empirical OpRisk data, created by merging different banks and applying resampling scheme. Data have been linearly scaled for non disclosure reason.
- In the following, we present results associated with modelling of Internal Fraud (event type category 1).

Features

- The peculiarity of the Event Type 1 is the joint occurrences - within a relative small sample size - of three big losses into 2013 q2 dataset- which has a visible and remarkable impact on the VaR estimated with MLE.
- The VaR correctly increases also using WLE, but remaining on a range of plausible values.

Set - up

- The focus is on the estimation of the severity distribution based on 5 different distributions (GPD, lognormal, Weibull, Inverse Gauss and Burr distribution).
- The pattern of the VaR is analyzed over time (2012q4, 2013q2, 2013q4) and under different levels of "contaminations" of the 2013q4 dataset:
  - No contamination ("base"): no (further) contamination is included. 2013q4 original data are modelled
  - First Contamination Hypothesis ("Hp 1"): inclusion into 2013q4 data of a new maximum data point equal to the maximum data point (within the original sample) multiplied by 1.5.
  - Second Contamination Hypothesis ("Hp 2"): as Hp1 but including also an extra data point equal to the maximum data point (within the original sample) multiplied by 2.
  - Third Contamination Hypothesis ("Hp 3"): as Hp2 but including also an extra data point equal to the maximum data point (within the original sample) multiplied by 3.
Stability over time analysis

The occurrences of the 3 big new losses in 2013 q2 brings to a considerable increase both on MLE and on WLE:

- With MLE approach, the VaR at 2013q2 reaches peaks of more than 10x Var figure of previous semester (especially when pareto-like distributions (GPD and Burr) are used).
- With WLE the maximum impact is an increase round 100% with the Inverse Gauss distribution.*

Figures on bottom panel provide insight of the relative weighting of observations: more extreme events gets at least nearly 50% of weight. This evidence is useful during model challenge phase in order to assure regulators that WLE is not a fancy trick to simply disregard jumbo losses (as opposite to what some capping or filtering procedure do …)

(*) Figures reported are just meant to compare relative performance of WLE vs. MLE. Variations of model output are typically lower once best fit selection component of the methodology is taken into account
Patterns previously observed for 2013 q2 are similar (although more moderate) to what is observed here for the contamination analysis under the Hp 3, that is to say including 3 new maximum values.*

This similarity is understandable since in the 2013 q2 data the three extremes newly sampled are respectively \(\sim 2\), \(\sim 3.5\) and 4 times the maximum data point at 2012q4.

(*) Figures reported are just meant to compare relative performance of WLE vs. MLE. Variations of model output are typically lower once best fit selection component of the methodology is taken into account.
Using simulated data and different sample size and distributions we also check robustness properties of MLE and WLE by comparison of their influence functions (IF).*

Results shows that IF of VaR measured with WLE (solid line) is generally upper bounded by the IF of MLE (dotted line) for any sample size and distributions

*IF is a standard tool of robustness (infinitesimal) analysis aiming at assessing the impact on estimates of the inclusion of a single extra data point.
Further directions of research

The WLE approach here discussed depends on an interpretable tilting parameter, basically representing the number of “isolated data” which – for the purpose of the calibration – can be interpreted as the number of outlier the estimation procedure is able to cope with, still maintaining “reliable” capital estimates. Although the impact of the tilting parameter $m$ depends on the sample data and the chosen distribution function, it clearly appears how relevant is its choice since it directly quantifies the intensity of the robustification procedure proposed.

How to allow for more flexibility with respect to the “a priori” calibration of a fixed value of $m$?

1. **Outliers detection techniques:**
   **Aim:** to empirically support the definition of the parameter $m$ as the estimated number of outliers within the sample.
   **Tools:** We are currently examining different outliers detection procedures for heavy-tail distributions, such as:
   - *parametric method*, based on the distribution of the ratio between each pair of consecutive order statistics, which depends on the estimate of the tail index. See Schluter and Trede (2008).
   - *non parametric method*, based on the joint use of the Bootstrap Based Outlier Detection Plot (or Bootlier plot), described in Singh and Xie (2003), and the Silverman test.

2. **Inequality constraint on the KL divergence:**
   **Aim:** framing of the optimization problem in a more flexible form, which softens the dependence from the chosen value of $m$ and avoiding to find the exact number of outliers.
   **Tools:** This goal is achievable in principle just formulating a more general statistical problem, substituting the equality constraint with a inequality constraint on the Kullback-Leibler divergence. That is to say, instead of imposing an exact distance between the uniform and the multinomial weights, we can just impose the maximum possible amount of this distance.

Finally, another interesting research area is to find proxies or heuristics rule that while mimicking the rationales underlying the construction of weights are less computationally intensive and as such can be replace MLE even within highly complex and demanding computation schemes, as the one sometimes observed in most advanced AMA models.
Conclusions

The approach we have presented here for Operational Risk is based on weighted likelihood and its statistical derivation relies on the paper on Choi et. al (2000).

It holds several properties. First, it is a robust estimator, behaving in accordance with the key messages defined by Dell’Aquila and Embrechts (2006) related to robust estimators:

- if “extreme” data points are conformed with the majority of the data, then they will receive a relative high density from the parametric model and consequently they will be not down-weighted;
- it provides good statistical properties (high efficiency, low bias) for a whole neighborhood of the assumed model distribution;
- WLE can represent a diagnostic tool to provide an indication on the influential data points, e.g., analyzing the relative magnitude of the estimated weights.

Beside its robustness, this estimator:

- Generalizes the MLE, including the MLE as a special case;
- retains some desirable statistical properties, including consistency and the fact that it can be constructed from applying MLE to a “transformed” sample;
- is simple both to understand and to implement;
- is not computationally burdensome;
- Weights are obtained in an adaptive, data driven way, limiting discretionary choice of the analyst.

The disadvantages include the non-closed form solution for the maximization problem - making the estimation procedure (slightly) longer than the MLE - and the need of calibration of the tilting parameter, which contributes to define the degree of robustness of the procedure. However, even under the point of view of the calibration, we argue the proposed estimator is preferable to most of other robust estimation since its tuning parameter has a simple interpretation and its value is set on a canonical scale that does not depend directly on the sample distribution.
• BCBS 196 - Basel Committee on Banking Supervision (2009), *Observed range of practice in key elements of Advanced Measurement Approaches (AMA).*


### Used acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMA</td>
<td>Advanced Measurement Approach</td>
</tr>
<tr>
<td>IF</td>
<td>Influence Function</td>
</tr>
<tr>
<td>LDA</td>
<td>Loss Distribution Approach</td>
</tr>
<tr>
<td>MBRE</td>
<td>Most Bias Robust Estimator</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimator</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>OMSE</td>
<td>Optimal Mean Square Error</td>
</tr>
<tr>
<td>OBRE</td>
<td>Optimal B-Robust Estimator</td>
</tr>
<tr>
<td>UoM</td>
<td>Unit of Measure</td>
</tr>
<tr>
<td>WLE</td>
<td>Weighted Likelihood Estimator</td>
</tr>
</tbody>
</table>
Annex
The influence function

Let \( G \) be some distribution. What happens when the data doesn't follow the model \( F \) exactly but another, slightly different, "going towards" \( G \) ?

To answer this question we're looking for

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

which is the one-sided directional derivative of \( T \) at \( F \) in the direction of \( G-F \) (\( T \) is the quantity to be estimated, i.e. VaR or severity parameters).

Let's set \( G = \delta_x \), i.e. the probability measure which gives mass 1 to \( x \). The influence function is then defined by

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

It describes the effect of an infinitesimal contamination at the point \( x \) on the estimate we are seeking, standardized by the mass \( \varepsilon \) of the contamination. For a robust estimator, we want a bounded influence function, that is, one which does not go to infinity as \( x \) becomes arbitrarily large. MLE has unbounded Influence function.
Contacts

• Andrea Antonio Colombo
  Senior Manager, KPMG Advisory, Financial Risk Management
  Tel + 39 348 3080010
  Email: andreacolombo@kpmg.it

• Alessandro Lazzarini
  Consultant, KPMG Advisory, Financial Risk Management
  Tel + 39 348 3064851
  Email: alazzarini@kpmg.it

• Silvia Mongelluzzo
  Senior Consultant, KPMG Advisory, Financial Risk Management
  Tel + 39 348 8819739
  Email: smongelluzzo@kpmg.it