Internal data, external data and consortium data for operational risk measurement: How to pool data properly? *

Nicolas Baud, Antoine Frachot and Thierry Roncalli

Groupe de Recherche Opérationnelle, Crédit Lyonnais, France†

This version: June 01, 2002

Abstract

It is widely recognized that calibration on internal data may not suffice for computing an accurate capital charge against operational risk. However, pooling external and internal data lead to unacceptable capital charges as external data are generally skewed toward large losses. In a previous paper, we have developped a statistical methodology to ensure that merging both internal and external data leads to unbiased estimates of the loss distribution. This paper shows that this methodology is applicable in real-life risk management and that it permits to pool internal and external data together in an appropriate way. The paper is organized as follows. We first discuss how external databases are designed and how their design may result in statistical flaws. Then we develop a model for the data generating process which underlies external data. In this model, the bias comes simply from the fact that external data are truncated above a specific threshold while this threshold may be either constant but known, or constant but unknown, or finally stochastic. We describe the rationale behind these three cases and we provide for each of them a methodology to circumvent the related bias. In each case, numerical simulations and practical evidences are given. In the coming weeks, we also plan to release an Excel-based, user-friendly package in order to pool internal and external data while avoiding over-estimation of the capital charge.

*We gratefully thank Maxime Pennequin, Fabienne Bieber, Catherine Duchamp and Nathalie Menkes, all of Operational Risk team at Crédit Lyonnais, for providing us with all the necessary informations and for stimulating discussions. We wish to thank Professor Santiago Carillo Menéndez and the participants at the workshop Seminarios de Matemática Financiera, Instituto MEFF-RiskLab, Madrid for comments and suggestions.

†*Address:* Crédit Lyonnais – GRO, Immeuble ZEUS, 4è étage, 90 quai de Bercy — 75613 Paris Cedex 12 — France; *E-mail:* nicolas.baud@creditlyonnais.fr, antoine.frachot@creditlyonnais.fr, thierry.roncalli@creditlyonnais.fr.

1 Introduction

According to the last proposal [1] by the Basel Committee on Banking Supervision, banks are allowed to use the Advanced Measurement Approaches (AMA) option for the computation of their capital charge covering operational risks (OR). Among these methods, the Loss Distribution Approach (LDA) is probably the most sophisticated one[1]. It is also expected to be the most risk sensitive as long as internal data are used in the calibration process and LDA is thus supposed to be more closely related to the actual riskiness of each bank.

However it is now widely recognized that calibration on internal data only may not suffice to provide accurate capital charge, especially for high severity/low frequency events. In other words, internal data should be supplemented with external data in order to improve the accuracy of capital measurement. This is all the more important that high severity/low frequency events are the risk types which contribute the most to OR capital charge (Baud, Frachot and Roncalli [2002]). Unfortunately mixing internal and external data together is likely to provide unacceptable results as external data are strongly biased toward extreme losses. This bias comes from the fact that external databases only record the highest losses, i.e. the losses which are publicly released. Without any rigorous statistical treatment, the estimated loss distribution is biased toward high losses and the resulting capital requirement is thus dramatically over-estimated.

As a result rigorous statistical treatments are required to make internal and external data comparable and to ensure that merging both databases leads to unbiased estimates. A description of the statistical treatment has been developped in a previous paper [4] (Frachot and Roncalli [2002]), at least from a theoretical point of view. The goal of our paper is to propose a practical, real-life methodology to pool internal and external data together in an appropriate way.

The paper is organized as follows. We first discuss how external databases are designed and how their design may result in statistical flaws. Then we develop a model for the data generating process which underlies external data. In this model, the bias simply comes from the fact that external data are truncated above a specific threshold which may be either constant and known, or constant but unknown, or finally stochastic. We describe the rationale behind these three cases and for each of them we give a statistical method to circumvent the related bias. In each case, numerical simulations and practical evidences are given.

2 Modelling external database bias

2.1 External databases

In this section, we discuss how external databases are built, which is a good starting point for assessing to what extent external databases are biased. Two types of external databases are encountered in practice.

- The first type corresponds to databases which record publicly-released losses. In short these databases are made up of losses that are far too important or emblematic to be concealed away from public eyes. The first version of OpVar® Database pioneered by PwC is a typical example of these first-generation external databases.
- More recent is the development of databases based on a consortium of banks. It works as an agreement among a set of banks which commit to feed a database with their own internal losses, provided that some confidentiality principles are respected. In return banks which are involved in the project are of course allowed to use these data to supplement their own internal data. Gold of BBA (*British Bankers' Association*) is an example of consortium-based data.

[1] see Frachot, Georges and Roncalli [2001] for an extensive presentation of this method.

Remark 1 *The project **ORX** (**O**perational **R**iskdata e**X**change) managed by OpVantage (administrative agent) and PwC Switzerland (custodian) is another example. In this case, "Participants*[2] *deliver specified data that meets quality assurance standards. Custodian anonymise data, clean and scale as required. Administrator consolidates data, performs required analysis, provides reports. Custodian [then] provides standard reports to firms after rescaling or other manipulations. Participating firms [finally] receive back data based upon the business lines and/or locations and/or events for which they provided data" (*Peemöller *[2002]).*

The two types of database differ by the way losses are supposed to be truncated. In the first case, as only publicly-released losses are recorded, the truncation threshold is expected to be much higher than in the consortium-based data. For example, the OpVar® Database declares to record losses greater than USD 1 million while consortium-based data pretend to record all losses greater than USD 25.000 for **ORX** database (or USD 10.000 by 2003 (Peemöller [2002])).

Furthermore public databases, as we name the firt type of external databases, and consortium-based databases differ not only by the stated threshold but also by the level of confidence one can place on it. For example, nothing ensures that the threshold declared by a consortium-based database is the actual threshold as banks are not necessarily able to uncover all losses above this threshold even though they pretend to be so[3]. Rather one may suspect that banks target this threshold although they do not have always the ability to meet this requirement yet. We shall see in the next subsection how the last remark implies a specific statistical modelling in the sense that stated threshold of consortium-based databases should be considered as stochastic.

2.2 Modelling assumptions

For ease of notations, we shall consider one particular business line and one loss type. Internal losses will be denoted by $(\zeta_i)_{i=1,\dots,n}$ where n is the number of recorded internal losses. In the same spirit, $(\zeta_i^\star)_{i=1,\dots,n^\star}$ represent external losses[4]. Let $\mathbf{F}(\zeta;\theta)$ be the parametric loss severity distribution which internal data are assumed to be drawn from. θ is therefore a set of parameters to be estimated. We denote θ_0 the (unknown) true set of parameters.

We do not intend to discuss whether losses are best captured by the set of probability distributions $\mathbf{F}(\zeta;\theta)$. Rather we shall assume that we know the true family of distributions $\mathbf{F}$ and that we only need to uncover (estimate) the true parameter θ_0. Considering for example the lognormal distribution set $\mathcal{LN}(\zeta;\mu,\sigma)$, think of $\theta=(\mu,\sigma)$ as a two-dimensional parameter, i.e. the (theoretical) expected and standard deviation of the logarithm of losses. Finally we shall denote $\hat{\theta}$ an estimator of θ.

Pooling internal and external data together to improve the estimation process makes sense as long as the following assumption holds:

Assumption 1 (Fair Mixing Assumption) *External data are supposed to be drawn from the same distribution $\mathbf{F}(\zeta;\theta_0)$ as internal data except that the recorded (external) data are truncated above a threshold H.*

Under this assumption, external data may be viewed as "implicit internal data", meaning that external and internal data can be pooled together provided external data have been made comparable with internal data. Under this condition, we could supplement internal data with these scaled external data in order to obtain a database with a greater number of observations. Since the accuracy of the estimators increases along with the total number of observations, one expects to estimate the loss distribution more accurately with a pool of both internal and external data.

[2]The first members are (or might be) Deutsche Bank, JP Morgan Chase, ABN-AMRO, Bayerische Landesbank, BNP-Paribas, Commerzbank, Euroclear, Danske Bank, Fortis bank, HypoVereinsbank, ING and Sanpaolo IMI.

[3]The **ORX** project seems more ambitious and proposes reporting control and verification. In particular, the financial institution must demonstrable its capability to collect and to deliver data if it wants to be a member of the **ORX** consortium.

[4]The superscript $\star$ always refers to external data-based concepts.

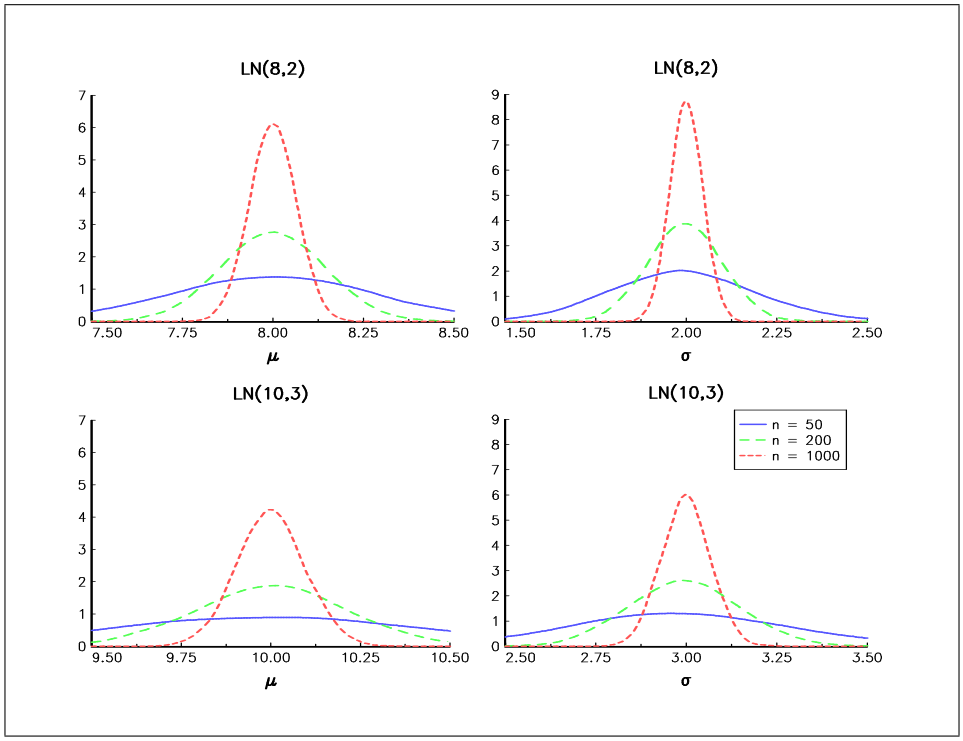

Figure 1: Density of estimators $\hat{\mu}_n$ and $\hat{\sigma}_n$

As a basic illustration, we simulate losses drawn from two lognormal distributions $\mathcal{LN}(8,2)$ and $\mathcal{LN}(10,3)$, and we plot the distribution of the maximum likelihood estimators $\hat{\mu}_n$ and $\hat{\sigma}_n$ (see Figure 1). As expected the accuracy depends positively on n and also on the true parameters. For example when losses are drawn from a lognormal distribution with a higher standard deviation, a greater number of observations is required to achieve similar accuracy as the variance of estimators is larger. Of course the same result holds when computing the quantile of the distribution (see Figures 2 and 3).

Finally we now turn to the truncation process. As said previously, the truncation threshold may be considered either as a constant or as a stochastic variable. The rationale behind this distinction is as follows. The first assumption concerns public database where all losses are recorded, by construction, above a defined threshold while the stochastic threshold case refers to consortium-based database with many different contributors whose internal recording processes differ from one another. Explanations rely on the fact that contributors have different sense of what is worth being released (or what they are able to release) and what should be sent into the external database. As a result a consortium-based database is a collection of data that have been truncated at different thresholds. In short it exactly means that the threshold is itself a random variable.

We shall consider the three following mutually-exclusive asumptions:

Assumption 2 (Known Constant Threshold Assumption) *Threshold H is non-stochastic and its value is known.*

Assumption 3 (Unknown Constant Threshold Assumption) *Threshold H is non-stochastic but its value is unknown.*

Assumption 4 (Stochastic Threshold Assumption) *Threshold H is stochastic.*

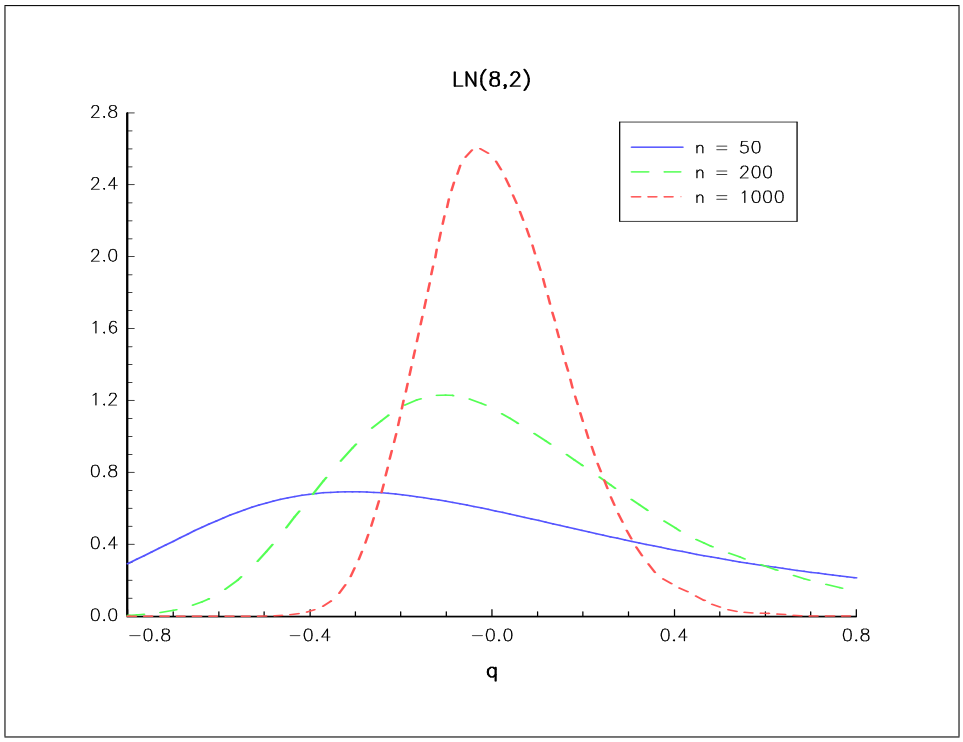

Figure 2: Density of the relative error of the 99.9% quantile when $\zeta \sim \mathcal{LN}(8, 2)$

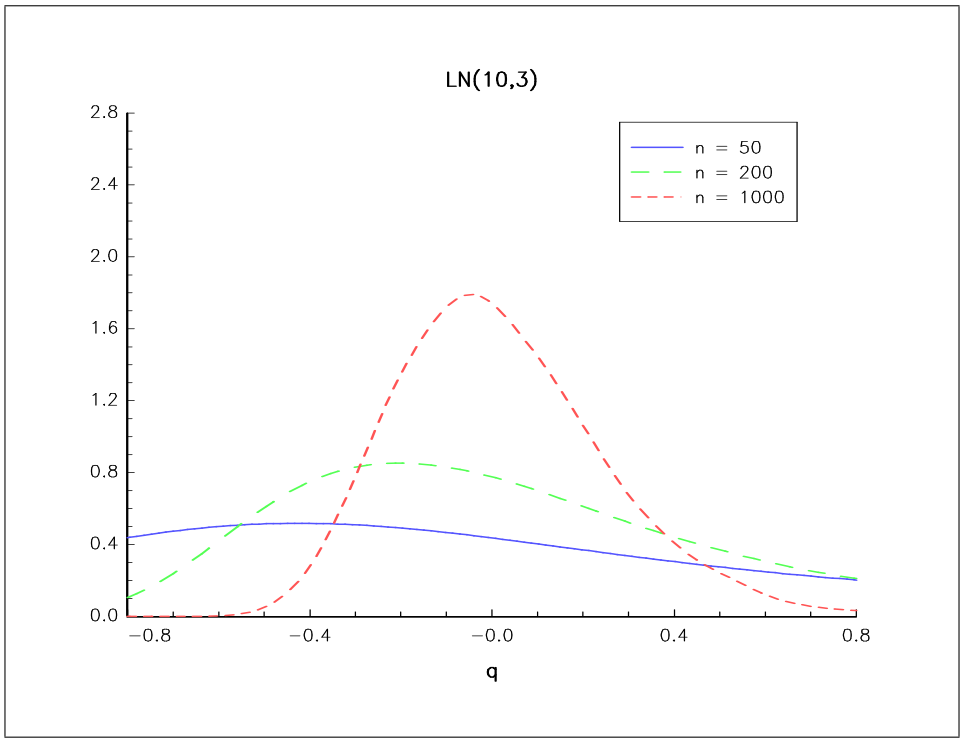

Figure 3: Density of the relative error of the 99.9% quantile when $\zeta \sim \mathcal{LN}(10, 3)$

In the following sections, we explore each assumption and develop the appropriate methodology. Under any assumption, the common and crucial point is that additional parameters have to be estimated along with the specific parameters characterizing the loss distribution. Our methodology is rather simple. It first consists in specifying the actual loss distribution of external data when the threshold assumption is carefully taken into account. Together with the loss distribution of internal data, we develop a standard maximum likelihood scheme illustrated by an implementation on simulated data.

Remark 2 *The use of simulated data instead of real-life data is a requirement demanded for confidentiality reasons but does not, in any case, weaken the scope of our results. Having peformed the same exercice on our real-life data (both internal and external), we are absolutely confident that the essence of both our theoretical and numerical results are fully preserved when using simulated data.*

3 Constant threshold assumptions

3.1 Known constant threshold

External data are supposed to be drawn from the same distribution as internal data except that the recorded data are truncated above a non-random threshold H which is perfectly known. Let f be the density function of an internal loss ζ. Regarding external data, we shall write

$$\zeta_i^\star \sim f^\star(\zeta;\theta) \tag{1}$$

where $f^\star(\zeta;\theta)$ differs from $f(\zeta;\theta)$ because of truncation:

$$f^\star(\zeta;\theta) := \mathbf{1}\{\zeta \geq H\} \cdot \frac{f(\zeta;\theta)}{\int_H^{+\infty} f(x;\theta)\,\mathrm{d}x} = \mathbf{1}\{\zeta \geq H\} \cdot \frac{f(\zeta;\theta)}{1-\mathbf{F}(H;\theta)} \tag{2}$$

where $\mathbf{1}\{\zeta \geq H\}$ stands for a dummy variable equal either to 1 or 0 depending on whether $\zeta \geq H$. Let ℓ be the log-likelihood function. We have

$$\ell(\theta) = \sum_{i=1}^{n} \ln f(\zeta_i;\theta) + \sum_{i=1}^{n^\star} \ln f^\star(\zeta_i^\star;\theta) \tag{3}$$

The maximum likelihood (ML) estimate $\hat{\theta}$ is then the solution of the maximization problem $\hat{\theta} = \arg\max_\theta \ell(\theta)$.

In order to show the influence of the truncation threshold, we perform a Monte Carlo simulation scheme. Two sets of 1000 losses are drawn from distribution $\mathcal{LN}(8,2)$. The first set stands for the internal database while the second is truncated above a threshold H meaning that all losses lower than H are dropped leaving a total of $n^\star \leq 1000$ observed external losses. Repeating the previous scheme as many times as necessary, we obtain the distribution of the maximum-likelihood estimators, that is $\hat{\mu}$ and $\hat{\sigma}$, in the four following situations[5] (see Figures 4 and 5):

1. "Internal": $\hat{\mu}$ and $\hat{\sigma}$ are the ML estimates using only internal data.
2. "External": $\hat{\mu}$ and $\hat{\sigma}$ are the ML estimates using only external data while ignoring truncation.
3. "Mixing": $\hat{\mu}$ and $\hat{\sigma}$ are the ML estimates using internal and external data while ignoring truncation.
4. "Mixing with threshold": $\hat{\mu}$ and $\hat{\sigma}$ are the ML estimates using internal and external data based on the log-likelihood (3).

[5] We use 5000 replications and a gaussian kernel to estimate densities.

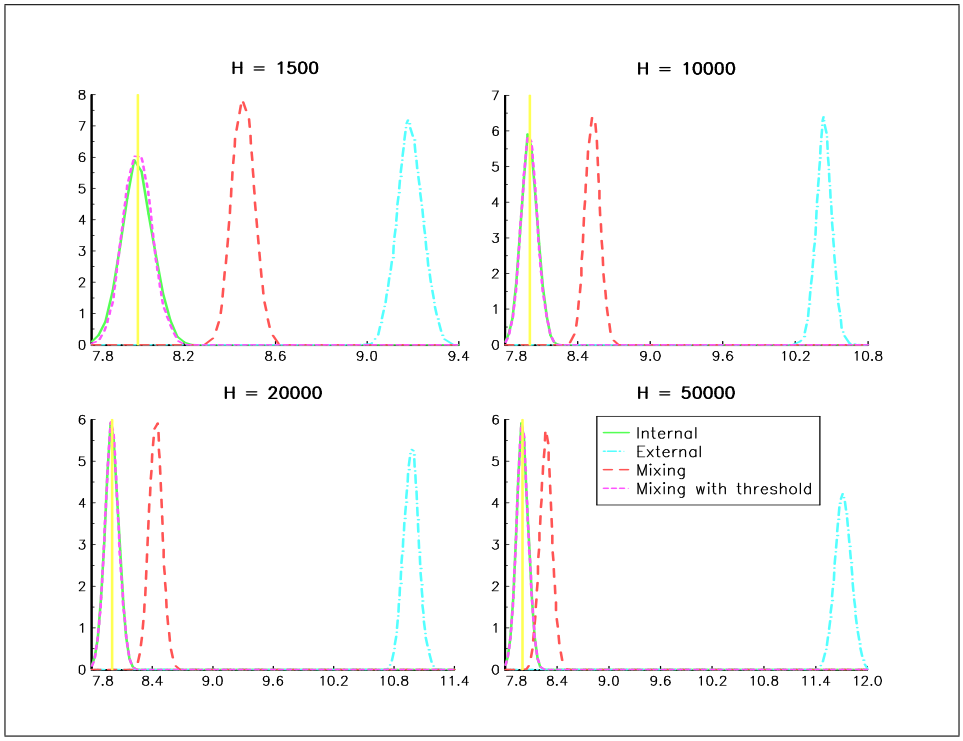

Figure 4: Mixing internal and external data — Density of estimators $\hat{\mu}$ when the threshold H is known

Implementations 1 and 4 are the only ones which are able to provide us with unbiased estimates while implementation 4 is expected to give much more accurate estimates. On the contrary implementations 2 and 3 lead to biased estimates. These theoretical results are entirely confirmed by our numerical simulations as shown on Figures 4 and 5: ignoring truncation results in dramatically spurious estimators. More specifically, under implementations 2 or 3, the expected loss is then over-estimated and subsequent capital charge (based on these estimates) would be much higher than really required.

3.2 Unknown constant threshold

H is no longer assumed to be known. Then it must be considered as an additional parameter to be estimated along with the parameters characterizing the loss distribution. The log-likelihood function is identical to the one given in the previous subsection, except that it is now an explicit function of both θ and H: The program to be maximized is now:

$$\left(\hat{\theta}, \hat{H}\right) = \arg\max_{\theta, H} \ell\left(\theta, H\right) \tag{4}$$

Classical results from maximum likelihood theory still hold meaning that maximizing $\ell\left(\theta, H\right)$ with respect to parameters θ and H is a consistent and efficient procedure for estimating the parameters. Furthermore it is immediately seen that the subsequent estimator for H, denoted by $\hat{H}$, is equal to:

$$\hat{H} = \min_i \zeta_i^\star \tag{5}$$

In practice however we have used a slightly different procedure to account for possible contamination with un-truncated or aberrant data. As a matter of fact, computing $\hat{H}$ as $\min_i \zeta_i^\star$ is generally misleading as external data may contain badly-recorded data. As an example, one external database we had the opportunity to investigate shows some losses which are of an order of magnitude dramatically lower than most other recorded losses. As a consequence $\hat{H} = \min_i \zeta_i^\star$ may significantly underestimate the true threshold and we prefer to interpret this fact as a symptom that external data are

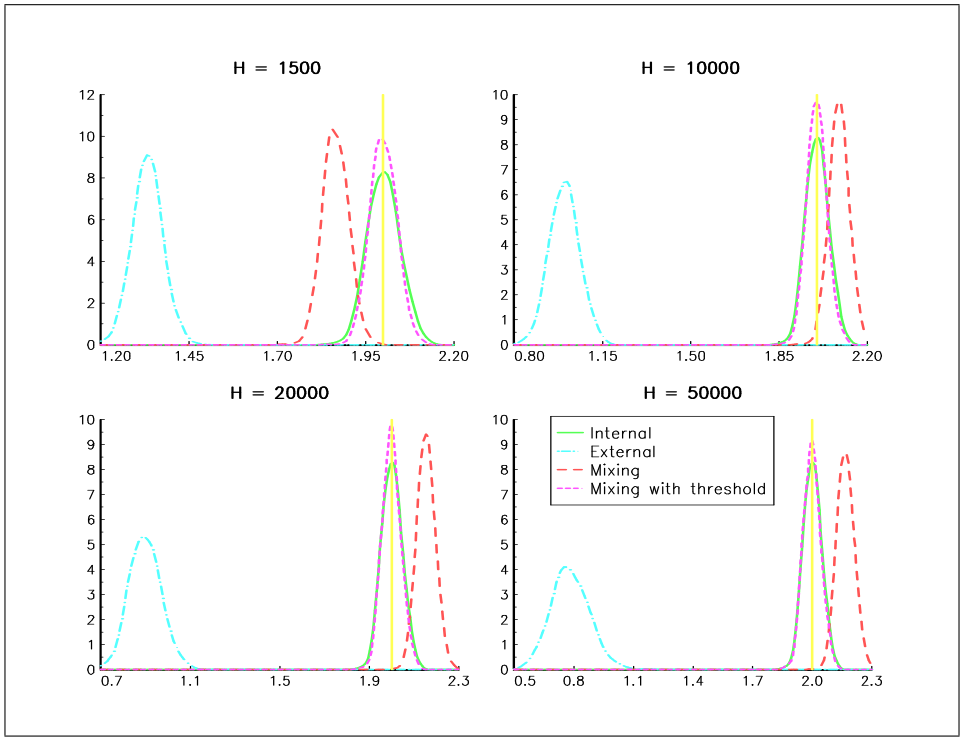

Figure 5: Mixing internal and external data — Density of estimators $\hat{\sigma}$ when the threshold H is known

contaminated by aberrant, non-informative data which should be preferably excluded away from the estimation process. Accordingly, for each threshold H,observations which fall under H are dropped and therefore they do not contribute to the log-likelihood $\ell(\theta, H)$.

Practically our iterative procedure (whose consistency can be rigourously proven) is as follows:

1. Estimate θ for each H ranging from 0 to $+\infty$.
2. Plot the function $H \mapsto \hat{\theta}(H)$ where $\hat{\theta}(H)$ denotes the estimator obtained for each given H.
3. $\hat{H}$ is eventually computed as the threshold beyond which $\hat{\theta}(H)$ remains (approximately) flat.

Figure 6 shows an example of the implementation where we have simulated two sets of log-normally distributed $\mathcal{LN}(8, 2)$ data with a truncation threshold of 1500. Our procedure correctly uncovers both threshold H and parameters μ and σ. Here again it is worth noting how spurious the estimates are when an appropriate correction (for truncation) is not implemented.

Let us consider another example. We assume that the distribution of losses is $\mathcal{LN}(8, 2)$. The size of the internal database is $n = 1000$, whereas the size of the external database is $n^\star = 150$. Let us imagine that only two banks contribute to the external database, one according to a threshold of 10000 and the other one according to a threshold of 50000. Let us suppose that both banks pretend to record losses above 10000 (but only the first one actually does). Then, the estimated loss distribution is biased when our constant-threshold assumption is applied with a (known constant) threshold set to $H = 10000$ (see Figure 7).
This second example may be viewed as the case of a consortium-based database which would be fed by two banks whose respective thresholds are (say) 10000 and 50000. Let us apply our procedure as in real-life, that is as if we were unaware of these heterogeneous thresholds. As expected, our procedure still provides unbiased estimates as soon as we find out the breaking point beyond which

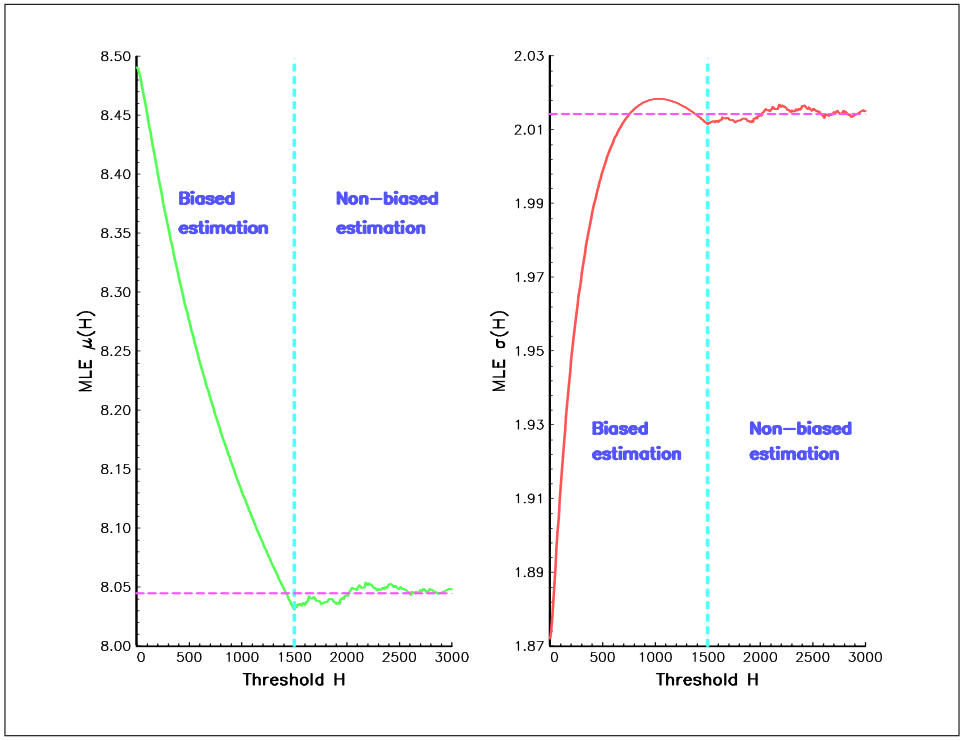

Figure 6: Estimating the unknown threshold

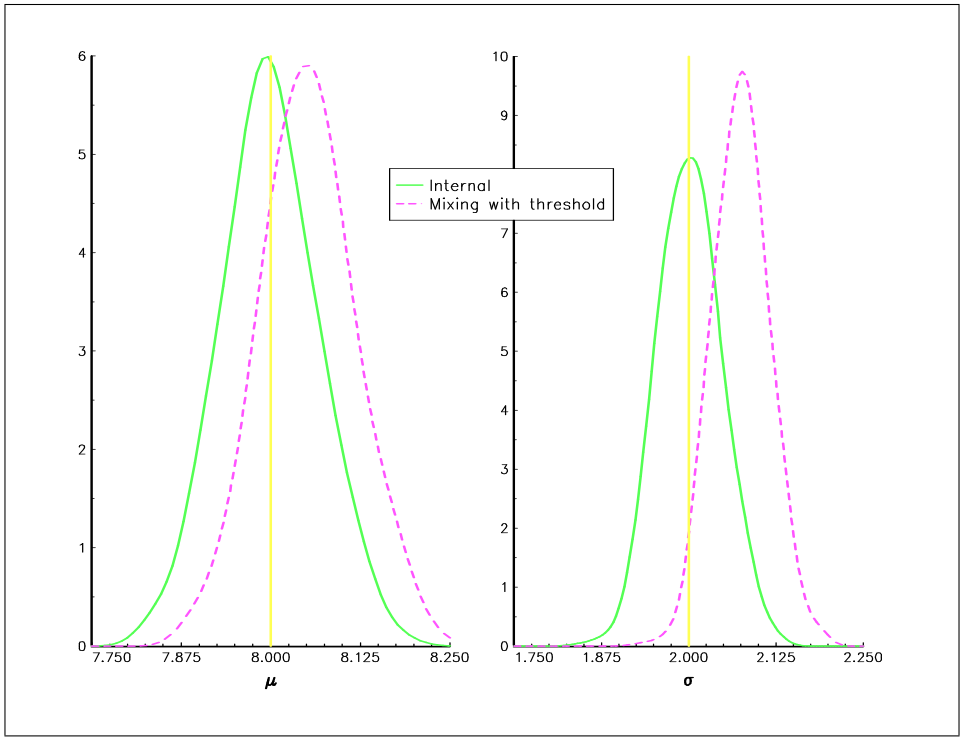

Figure 7: Density of estimators when external data are not exhaustive

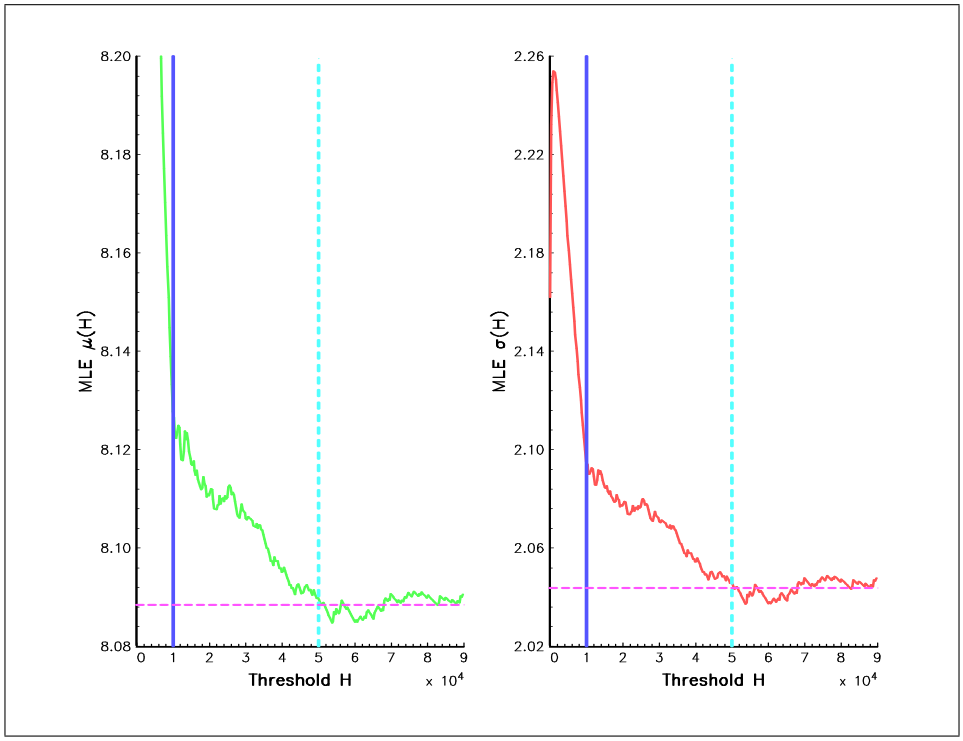

Figure 8: Finding the 'optimal' threshold

the estimates stabilize (see Figure 8). In this second example (which approaches real-life consortium-based databases), the breaking point appears to be the highest threshold adopted by contributing banks. It is rather unsatisfactory because it means that loss parameters are eventually estimated with fewer data than available, i.e. data which stand above the highest threshold. In short, estimations are spoiled by the presence (among contributors) of banks with high threshold, i.e. those banks which are only able to uncover large losses. In order to avoid this loss of information, we go one step further in the following section by considering that the external database results from many contributors whose respective thresholds differ from one another.

4 Stochastic threshold assumption

4.1 The log-likelihood function

Threshold H is no longer assumed to be constant. Instead H follows a non-degenerate probability distribution function:

$$H \sim g\left(h;\delta\right) \tag{6}$$

where δ is a set of parameters characterizing this parametric distribution. As said previously, it can be interpreted as the fact that contributors of the external database truncate the data they supply above contributor-specific thresholds. The density function conditionnally to H being a known constant is the same as before:

$$f^{\star}\left(\zeta;\theta \mid H=h\right) = \mathbf{1}\left\{\zeta \geq h\right\} \cdot \frac{f\left(\zeta;\theta\right)}{\int_{h}^{+\infty} f\left(x;\theta\right)\,\mathrm{d}x} \tag{7}$$

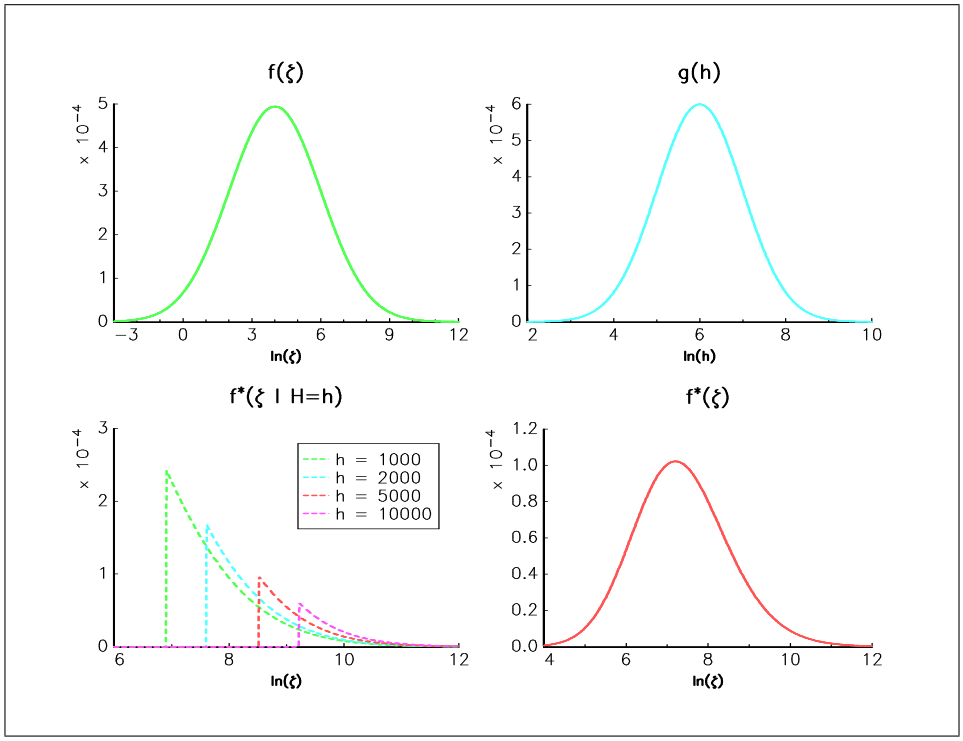

Figure 9: Impact of a stochastic threshold $H \sim \mathcal{LN}(7,1)$ on external data

whereas the unconditional density function is:

$$f^{\star}(\zeta;\theta,\delta) = \int_0^{+\infty} f^{\star}(\zeta;\theta \mid H=h)\, g(h;\delta)\, \mathrm{d}h \tag{8}$$

It is not that easy to guess which distribution is well-suited for modelling the threshold. We left this issue for further discussion in the following subsection but, whichever distribution is considered, its impact on the actual loss distribution of external data is highly significant. As an example Figure 9 gives an illustration of how the distribution of losses is affected when it is compounded by the H-distribution. Here we consider again that loss distribution before truncation is $\mathcal{LN}(8,2)$ while H is assumed to be log-normal $\mathcal{LN}(\mu_H,\sigma_H)$. Figure 10 plots the distribution of external losses $f^{\star}(\zeta;\theta,\delta)$ for various parameters μ_H and σ_H.

Regarding the estimation process, the log-likelihood function is a direct generalization of the previous log-likelihood function considered in the former sections, except that the log-likelihood is now significantly more complex. However, we may 'easily' compute it using numerical algorithms, in particular for managing the integral terms. Accuracy in integral computations is highly recommended. Otherwise, the optimization process may converge with difficulty or may not converge at all. Let us consider an example with $\zeta \sim \mathcal{LN}(8,2)$ and $H \sim \mathcal{LN}(7,1)$. We simulate 1000 external data from this truncation mechanism. For one simulation path, we obtain the following results:

d	32	64	128	256	512
$\hat{\mu}_{\zeta}$	7.308	7.572	7.796	7.796	7.792
$\hat{\sigma}_{\zeta}$	2.213	2.154	2.088	2.087	2.089
$\hat{\mu}_H$	7.059	6.957	6.865	6.866	6.868
$\hat{\sigma}_H$	0.993	0.932	0.904	0.908	0.907

where d denotes the order of the Gauss-Legendre quadrature method. It is worth noting that, for $d \geq 128$, similar results are obtained but large errors remain for lower order.

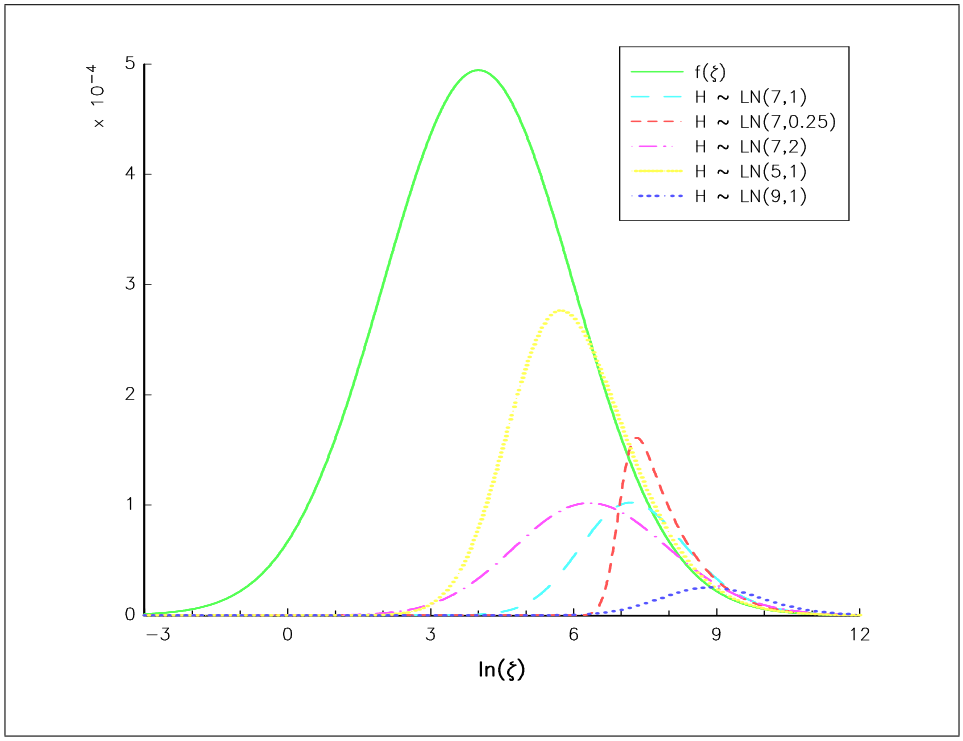

Figure 10: Plot of the density $f^{\star}(\zeta)$ with different distributions for H

Furthermore our experience shows that we may encounter convergence problems when the size of the external database is small. In particular, we may obtain several local maxima. It can be circumvented by using a parametric density function $g(h;\delta)$ with very few parameters (one or two).

Remark 3 *How does a* ***Constant Threshold Assumption*** *behave when applied to data drawn from a* ***Stochastic Threshold*** *distribution? To address this question, we take* $\mathcal{LN}(8,2)$ *as the loss distribution (before truncation) and we assume that* H *is distributed according to a scaled* Beta *distribution* $\mathcal{B}(0.5,1;25000,100000)$ *(see below) with* $n=1000$ *simulated internal data and* $n^{\star}=500$ *external data. We apply the methodology expressed in previous subsections (constant threshold case). Results are reported in Figure 11 where we see that* $\hat{\mu}(H_0)$ *and* $\hat{\sigma}(H_0)$ *are not 'stable' above* $H_0=H_{\_}=25000$ *but stabilize instead for a higher value, which is nevertheless lower than* $H_0=H_{+}=100000$ *(because truncation for these values are less important — we have* $\Pr\{H\geq 60000\}=31.7\%$ *and* $\Pr\{H\geq 75000\}=18.4\%$*). It gives a numerical illustration of the issue raised at the end of the last subsection, i.e. the number of external data above this 'optimal' threshold becomes very small. In our example, only 155 events of the 500 external events correspond to a loss bigger than 75000.* ***The risk when assuming a constant threshold is then that only few losses of the external database are used to estimate the severity loss distribution****. Through stochastic threshold modelling,* ***all*** *external data are used.*

4.2 Remarks about the threshold distribution

The only constraint one has to impose is that H be non negative. Apart from this constraint, one may imagine many different shapes for this distribution, like hump-shaped distributions, bi-modal or multi-modal distributions, etc. depending on how banks are distributed in terms of their ability to uncover internal losses. As an example, one may assume a Beta $\mathcal{B}(\alpha,\beta)$ distribution whose support would be $[H^{-},H^{+}]$. This could be rationalized by considering H^{-} as the stated threshold of the consortium database while H^{+} would be the highest possible value, i.e. corresponding to the "worst"

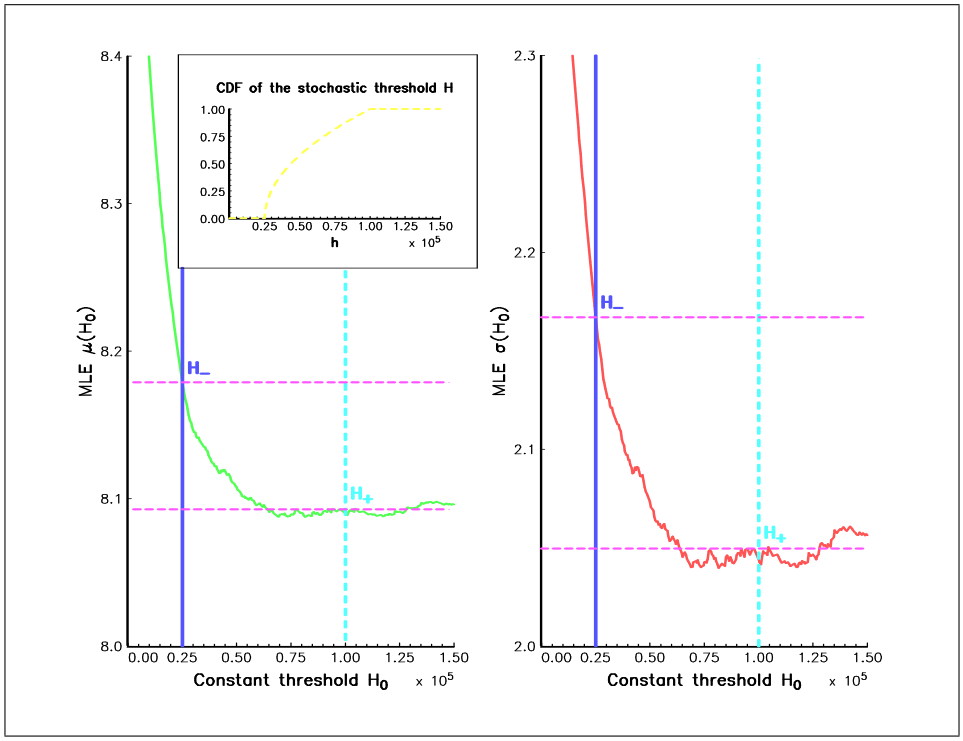

Figure 11: Finding the 'optimal' constant threshold when the threshold is stochastic

contributing bank. The threshold distribution $\mathcal{B}(\alpha, \beta; H^-, H^+)$ would then have the following density:

$$g(h; \alpha, \beta) = \frac{(H^+ - H^-)^{-1}}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \tag{9}$$

with

$$x = \frac{h - H^-}{H^+ - H^-} \wedge 1 \tag{10}$$

Example 1 *For the ORX consortium, the stated threshold is 25000. Suppose that some contributors are unable to report the internal losses lower than* 100000*. In this case, we should logically set* $H^- = 25000$ *and* $H^+ = 100000$*.*

We have reported several corresponding scaled Beta distributions in Figure 12. They implicitly correspond to different threshold mecanisms. For example, $(\alpha = 1, \beta = 1)$ refers to a uniform distribution, meaning that banks' thresholds are uniformly distributed between 25000 and 100000. Alternatively, $(\alpha = 1, \beta = 6)$ is an example of a L shaped distribution.

In practice it is important to have an initial guess of what this distribution looks like. For achieving this, we have tried a 'non-parametric' method where interval $[H^-, H^+]$ is discretized into $K+1$ points uniformly spaced $\{h_k\}$ with:

$$h_k = H^- + \frac{(H^+ - H^-)}{K} k, \qquad k = 0, 1, \ldots, K \tag{11}$$

Denoting p_k the probability $\Pr\{H = h_k\}$, one can consider a discretized version of the log-likelihood function, which in turn becomes a function of $(\theta, p_0, \ldots, p_K)$. The ML estimates of $\theta, p_0, \ldots, p_K$ are then the solution of the contrained optimization problem $\max \ell(\theta, p_0, \ldots, p_K)$ under the constraints $p_k \geq 0$ and $\sum_{k=0}^{K} p_k = 1$.

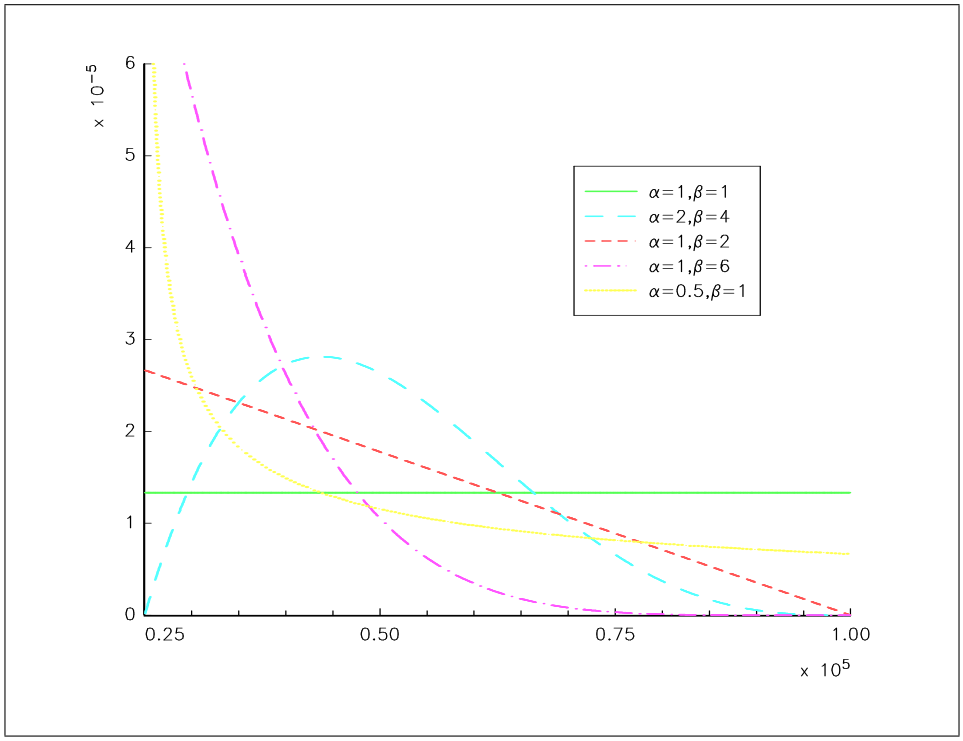

Figure 12: Some scaled Beta distributions

Let us consider the previous example where $\zeta \sim \mathcal{LN}(8,2)$ and $H \sim \mathcal{B}(0.5,1;25000,100000)$. It corresponds to a rather trivial case where the density of H is linearly decreasing to zero. We estimate the parameters $(\mu,\sigma,p_0,p_1,\ldots,p_K)$ for several values of K with 1000 simulated internal data and 500 external data. Results concerning the probabilities p_k are reported in Figure 13 and are linear as expected, provided that the number of discretization points is not too high (because otherwise it implies too many parameters to be estimated), which in turn distorts the estimation process.

5 An example with the Crédit Lyonnais database and the BBA database

We consider the example of loss type 'External Fraud' which is assumed to be log-normally distributed:

$$f(\zeta;\mu,\sigma) = \frac{1}{\zeta\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln\zeta-\mu}{\sigma}\right)^2\right) \tag{12}$$

Under the Unknown Constant Threshold assumption, we have the following density for external data:

$$f^\star(\zeta;\mu,\sigma) = \frac{\mathbf{1}\{\zeta \geq H\}}{1-\Phi\left(\left(\ln H-\mu\right)/\sigma\right)} \cdot \frac{1}{\zeta\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln\zeta-\mu}{\sigma}\right)^2\right) \tag{13}$$

Crédit Lyonnais demands to its business units to report any loss higher than 1500 euro[6] whereas lower losses can be reported in an aggregated form. Capital-at-Risk computations regarding operational risk are made at a 99.9% confidence level. The loss frequency distribution is the distribution of the random number of losses higher than 1500 euro in one year on[7] and the loss severity distribution is the

[6] Crédit Lyonnais has recently lowered its threshold down to 1000 euro.
[7] modelled according to a Poisson distribution.

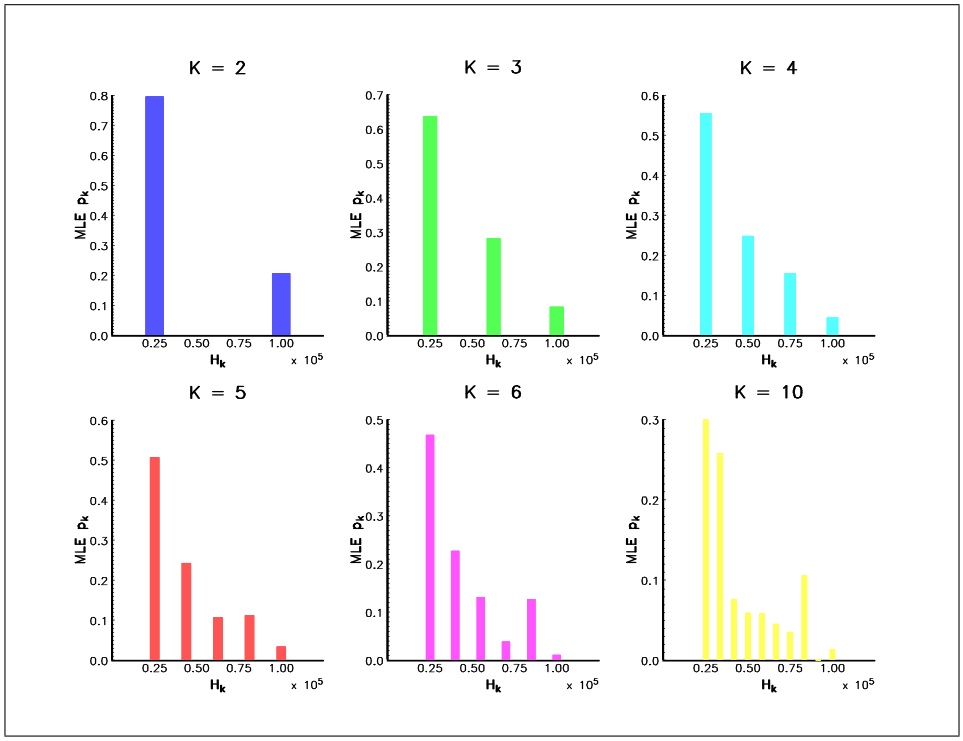

Figure 13: Maximum likelihood estimation of the probabilities p_k

distribution of one loss being higher than 1500 euro. For reasons of robustness, the Capital-at-Risk is computed by loss types.

We pool external frauds of the Crédit Lyonnais database with losses from the BBA database[8]. Figure 14 presents the results[9] when the threshold is supposed to be constant[10]. Let $\hat{\mu}$ and $\hat{\sigma}$ be the values of the parameters when estimated on internal data and $\hat{\mu}(H)$ and $\hat{\sigma}(H)$ when estimated on both internal and external data. We obtain $\hat{\mu} \leq \hat{\mu}\left(\hat{H}\right) \leq \hat{\mu}(0)$ and $\hat{\sigma}\left(\hat{H}\right) \leq \hat{\sigma}(0) \leq \hat{\sigma}$ where $\hat{H}$ corresponds to the 'optimal' threshold. $H = 0$ corresponds to the case where truncation is deliberately ignored, leading to strongly biased estimators: in comparison with the Capital-at-Risk computed on internal data only, the Capital-at-Risk (on both external and internal data) decreases if we use the appropriate threshold correction whereas it increases if we ignore the truncation effect (see Figure 15).

The case of stochastic threshold is considered in Figure 16. It confirms the fact that contributing banks are unable to respect the stated threshold.

[8] We have considered that the loss type External Fraud defined by the Basel Committe on Banking Supervision corresponds to the following loss types in the BBA database:

- External fraud/cheque fraud/forgery
- Fraudulent account opening by client
- Other criminal activity risk
- Robberies (& theft)

Note that the official threshold of the BBA database is US \$50000 for Retail and US \$100000 for Wholesale.

[9] The results presented here are slightly different from those found in [2], as the former have been obtained from a more comprehensive database (from Quarter 3 2000 to Quarter 4 2001 vs Quarter 3 2001 to Quarter 4 2001).

[10] Because oldest events of the Crédit Lyonnais database are often recorded on an aggregated basis, we prefer to estimate the loss severity distribution using the generalized method of moments (GMM), which is easier to implement from a computational point of view.

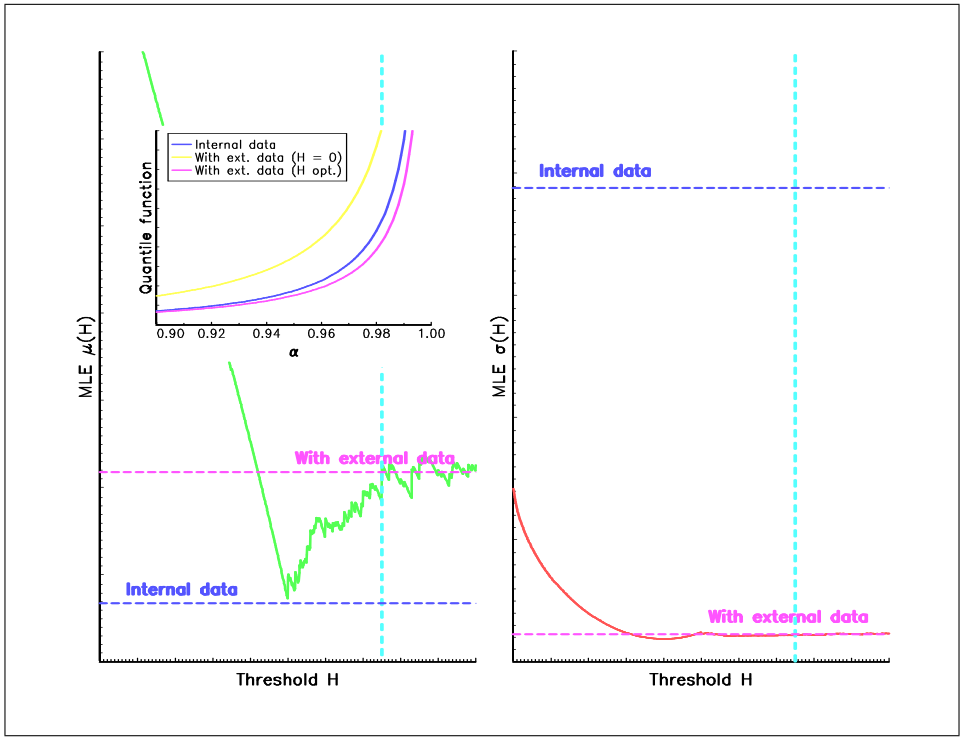

Figure 14: Estimating the 'optimal' threshold for the loss type External Fraud

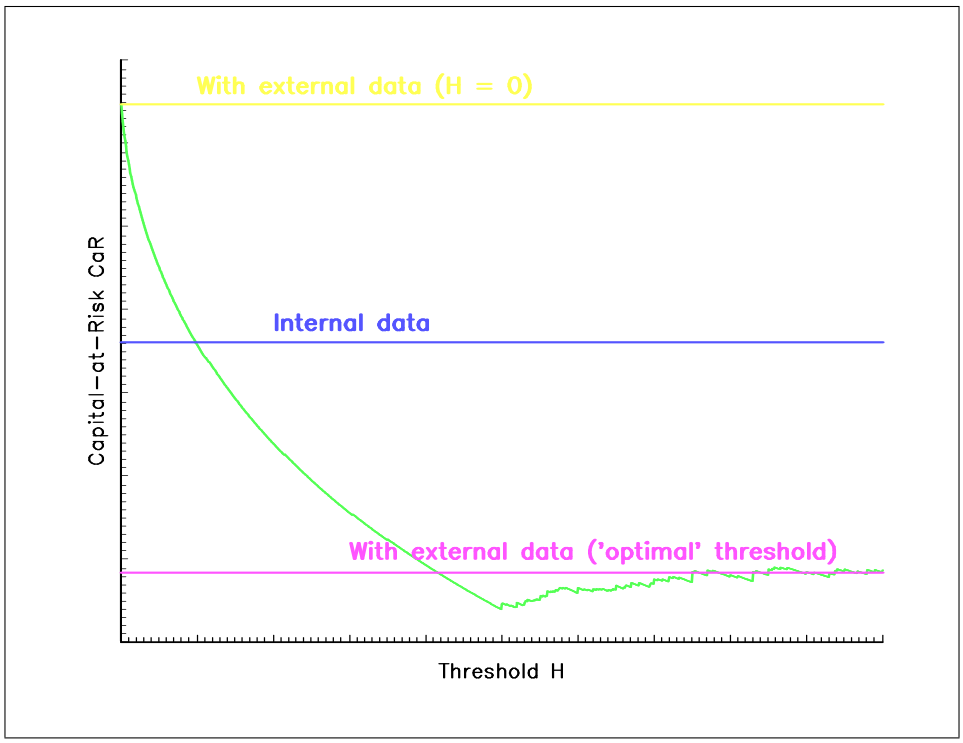

Figure 15: Relationship between the capital-at-risk and the threshold

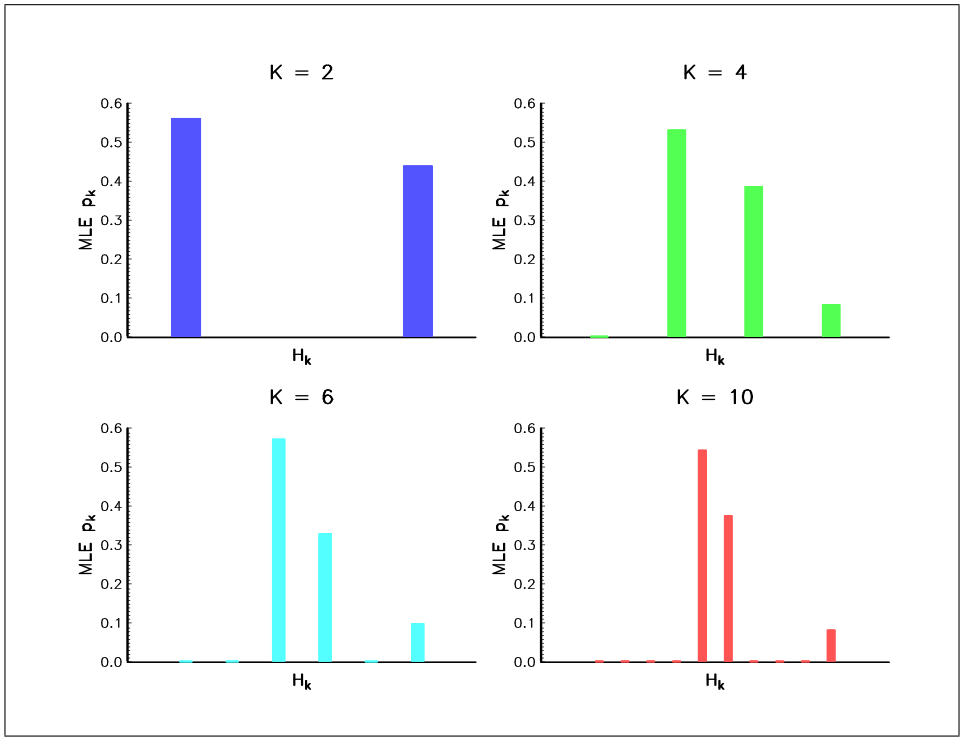

Figure 16: Maximum likelihood estimation of the probabilities p_k for the loss type External Fraud of the BBA database

6 Conclusion

In this paper, we have discussed how to pool internal and external data for measuring operational risk and we have proposed a statistical methodology which can be applied in practice. In the coming weeks, we plan to release a user-friendly routine which implements all the methodologies developped here.

However, it is worth noticing that our methodology is based on the assumption that external data are drawn from the same distribution as internal data except that the recorded data are truncated above a threshold. It means that external data can be viewed as "implicit internal data", provided that our methodology is used. Nevertheless, it is not that obvious that probability distributions for internal and external data are (before truncation) identical. Even though we have not investigated this issue in this paper, our methodology is able to provide a statistical test of the equality of the two distributions. As a matter of fact, it can serve as a reliable indicator of whether internal losses of a specific bank are comparable with losses from other banks. It is then a useful tool to benchmark each bank with respect to the industry.

Unfortunately, if the hypothesis that the two distributions are identical is rejected, then our methodology only provides an 'average' severity loss distribution which must be interpreted with caution since it can no longer be considered as the true bank's loss distribution.We leave the development of an appropriate methodology of this case for further research.

References

[1] Basel Committee on Banking Supervision, Working Paper on the Regulatory Treatment of Operational Risk, september 2001

[2] Baud, N., A. Frachot, and T. Roncalli [2002], An internal model for operational risk computation, Crédit Lyonnais, Groupe de Recherche Opérationnelle, *Slides* of the conference "Seminarios de Matemática Financiera", Instituto MEFF – Risklab, Madrid

[3] Frachot, A., P. Georges and T. Roncalli [2001], Loss Distribution Approach for operational risk, Crédit Lyonnais, Groupe de Recherche Opérationnelle, *Working Paper*

[4] Frachot, A. and T. Roncalli [2002], Mixing internal and external data for managing operational risk, Crédit Lyonnais, Groupe de Recherche Opérationnelle, *Working Paper*

[5] Peemöller, F.A. [2002], Operational risk data pooling, Deutsche Bank AG, *Presentation at CFSforum – Operational Risk*, Frankfurt/Main