Applied Probability Trust (20 December 2004)

A NOTE ON THE ESTIMATION OF THE FREQUENCY AND SEVERITY DISTRIBUTION OF OPERATIONAL LOSSES

A. CHERNOBAI,* *University of California, Santa Barbara*

C. MENN ** AND

S. TRÜCK,** *Universität Karlsruhe*

S. T. RACHEV,* ** *Universität Karlsruhe and University of California, Santa Barbara*

Abstract

The Basel II Capital Accord requires banks to determine the capital charge to account for operational losses. Compound Poisson process with Lognormal losses is suggested for this purpose. The paper examines the impact of possibly censored and/or truncated data on the estimation of loss distributions. A procedure on consistent estimation of the severity and frequency distributions based on incomplete data samples is presented. It is also demonstrated that ignoring the peculiarities of available data samples leads to inaccurate Value-at-Risk estimates that govern the operational risk capital charge.

Keywords: Operational Risk; Censored and Truncated Data; EM-Algorithm; Loss Distribution

Email address: annac@pstat.ucsb.edu, menn@statistik.uka.de, trueck@statistik.uka.de, rachev@statistik.uka.de

* Postal address: Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA.

** Postal address: Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, Geb. 20.12., D-76128 Karlsruhe, Germany.

1. Introduction

The exposure of banks and other financial institutions to various types of risk has accelerated as a consequence of financial globalization, deregulation and development of new products, instruments and services. A large proportion of financial risks that is attributed to neither market nor credit risk is known as *operational risk*. Examples include losses due to unauthorized trading, fraud, human errors on the Orange County (USA, 1994), Barings (Singapore, 1995), Daiwa (Japan, 1995) or the risks of natural disasters, significant computer failures, terrorist attacks. Under the 2001 Basel II Capital Accord (cf. (1)), banks are required to determine the capital charge to account for unexpected operational losses. Banks currently allocate roughly 60% of their regulatory capital to credit risks, 15% to market risks and 25% to operational risk (cf. (13)) (Cruz (cf. (6)) suggests alternative figures, 50%, 15% and 35%, respectively).

The Bank of International Settlements (BIS) described five methodologies for quantifying the capital charge, to be adopted by each institution depending on its size and risk exposure structure by the end of 2006. Under the Loss Distribution Approach (hereforth, LDA) banks compute separately the loss severity and frequency distribution functions for each business line and risk type pair, over a one year period. The total capital charge is then determined by the sum of the one-year Value-at-Risk (hereforth, VaR) measures with the confidence level $(1-\alpha)$, α such as $\alpha = 1\% - 5\%$, across all combinations, based on the compounded losses. The BIS suggest using the compound Poisson process with Lognormal loss amounts (2).

The correct estimation of the operational loss frequency and severity distributions is the key to determining an accurate aggregated one-year operational capital charge. However, the presence of minimum collection thresholds imposes a problem to data recording. In banks' internal databases losses are recorded starting from roughly 10,000 USD, and the external databases generally collect losses exceeding 1,000,000 USD (cf. (3)). We refer to the recorded data as thus left-truncated and incomplete.

Some evidence suggests that the possibility of losses lying under the specified threshold level is often omitted from the estimation of the loss distribution. Fitting unconditional distribution to the observed (incomplete) losses and ignoring the missing data, in our belief, would lead to inaccurate estimates of the parameters of both

severity and frequency distributions. The magnitude of the effect is dependent on the threshold level and the underlying distribution. In light of BIS discussion of the treatment of operational losses, this paper focuses on the Lognormal case, but the methodology can by all means be extended to more general cases. The resulting VaR measure would tend to be under- or overestimated, depending on the threshold level, underlying distribution, the VaR confidence level, and the steps undertaken to account for the missing data.

The aim of this paper is two-fold. The first purpose is to present the general methodology for the treatment of missing and censored data. Missing data is classified into two types: randomly missing data and non-randomly missing data. This paper treats the second case. Secondly, we analyze the bias of the parameters of the Lognormal severity and Poisson frequency distributions, and the limiting properties of the subsequent impact on the aggregated losses and hence the operational capital charge.

The paper is organized as follows. Section 2 explains the operational loss data problem. Section 3 discusses the general methodology of treating data with non-randomly missing observations of two possible types: the number of missing data points known and unknown. The Expectation-Maximization algorithm (cf. (8)) is often an efficient technique to estimate the parameters of complete data set's distribution based on the incomplete data set, for a class of distributions belonging to the exponential families. Section 4 analyzes possible adjustment procedures to the severity and frequency distributions which account for the missing data, and examines the effects on the VaR measure for the case when the missing data is ignored in the estimation. It is demonstrated that ignoring the missing data leads to misleading VaR estimates. Section 5 concludes and states final remarks.

Expectation-Maximization algorithm has been used in a variety of applications such as probability density mixture models, hidden Markov models, cluster analysis, factor analysis, survival analysis. References include (14), (15), (17), (7), among many others.

2. Problem Description

The following model is based on the well-known setup from the insurance theory. As suggested by the Basel Committee, we assume that the stochastic process $(S_t)_{t\geq 0}$ describing the cumulative operational losses faced by a company A over the time interval $[0,t]$ is modeled by a compound Poisson process of the form:

$$S_t = \sum_{k=0}^{N_t} L_k, \quad L_k \overset{\text{iid}}{\sim} Q_\gamma, \tag{2.1}$$

where N_t denotes a homogenous Poisson process with intensity $\lambda > 0$. The loss distribution Q_γ is assumed to belong to a parametric family $\mathcal{Q} = \{Q_\gamma \,|\, \gamma \in \Gamma\}$ of continuous probability distributions. Depending on the distribution, γ is a parameter vector or a scalar. For simplicity, we would refer to it as a parameter throughout the paper. We assume that the family $\mathcal{Q}$ is sufficiently well behaved so that the parameter γ can be estimated consistently by maximum likelihood. To avoid the possibility of negative losses we restrict the distribution to be concentrated on the positive half line, i.e. for all γ we have $Q_\gamma(\mathbb{R}_{>0}) = 1$. The distribution function (d.f.) of Q_γ is denoted as F_γ and its density as f_γ. The d.f. of the compound Poisson process is given by:

$$P(S_t \leq s) = \begin{cases} \sum_{n=1}^{\infty} P(N_t = n)\ F_\gamma^{n*}(s) & s > 0 \\ \\ P(N_t = 0) & s = 0 \end{cases} \tag{2.2}$$

where F_γ^{n*} denotes the n-fold convolution with itself.

Representation (2.1) assumes independence between frequency and severity distributions. The process N_t uniquely governs the frequency of the loss events, and the distribution Q_γ controls the loss severity. In practice, model (2.1) can be used to determine the required capital charge imposed by regulators. The capital charge is measured as the $(1-\alpha)$ quantile of the cumulative loss distribution over a one year period, the so-called Value-at-Risk (VaR). $\text{VaR}_{\triangle t, 1-\alpha}$ for the tolerated risk level α and the time interval of length $\triangle t$ is defined as the solution of the following equation:

$$P(S_{t+\triangle t} - S_t > \text{VaR}_{\triangle t, 1-\alpha}) = \alpha \tag{2.3}$$

In practice, generally $\alpha = 1\%-5\%$ and $\triangle t$ is a one year period. It is possible to measure VaR, at least theoretically, given the estimated values for the unknown parameters λ and γ.

Given a sample $\mathbf{x} = (x_1, x_2 \ldots, x_n)$ containing all losses which have occurred during some time interval $[T_1, T_2]$, under the imposed assumptions on the structure of $\mathcal{Q}$, the task of estimating λ and γ can be performed with the maximum likelihood estimation (MLE) principle:

$$\hat{\lambda} = \hat{\lambda}_{\mathrm{MLE}}(x) = \frac{n}{T_2 - T_1} \quad \text{and} \quad \hat{\gamma} = \hat{\gamma}_{\mathrm{MLE}}(x) = \arg\max_{\gamma} \sum_{i=1}^{n} \log f_\gamma(x_i) \tag{2.4}$$

In reality it often occurs that not all losses over a certain time interval are recorded accurately, for various reasons. Moreover, the estimation becomes more delicate if some of the losses which have occurred in $[T_1, T_2]$ have been recorded only partially, or not recorded at all. In the framework of operational risk, losses of magnitudes below a certain threshold are not recorded in the databases (cf. (3)):

1. *External databases:* Operational losses are recorded starting from approximately 1,000,000 USD;

2. *Internal databases:* Operational losses are recorded starting from approximately 10,000 USD.

The problem of operational loss data recording and analysis is thus complicated by the presence of *non-randomly missing data* on the left side of the loss distribution. The question addressed in the subsequent analysis is whether ignoring the missing data has a significant impact on the estimation of the frequency parameter λ and the severity parameter γ. From the statistical viewpoint, with non-randomly missing data, it is clear that all estimates will be biased if the missing data is not accounted for. However in practical applications a possible rationale for ignoring the missing data would be that the major part of losses is in excess of the threshold, and losses smaller than 10,000 USD can not have a significant impact on the operational VaR that is determined by the *upper* quantiles of the loss distribution.

In the following section we describe a general sample design covering all cases which seem to be relevant for practical purposes. The presented sample design for censored and truncated data was first presented by Dempster et al. in (cf. (8)).

3. Estimation of complete-data distributions

3.1. Sample design

We assume that losses are generated by the process given in equation (2.1). Given a time interval $[T_1, T_2]$ - the sample window - the collected data which is available for estimating λ and γ is assumed to be incomplete in the following sense. We are given two non-negative pre-specified thresholds u_1 and u_2 with $0 \leq u_1 \leq u_2$ defining a partition on $\mathbb{R}_{\geq 0}$ through the events $A_1 = [0, u_1]$, $A_2 = (u_1, u_2]$ and $A_3 = (u_2, \infty)$. The interpretation of these events for our sample design is the following.

A_1 : If a random outcome of the loss distribution belongs to A_1 then it will not enter the data sample at all. Neither the frequency nor the severity of losses below u_1 are recorded (truncated data).

A_2 : Losses in A_2 are counted but the values of the losses are not recorded, i.e. the sample contains only the information on the number of losses with values between u_1 and u_2 (censored data).

A_3 : Realizations in A_3 are fully reported, i.e. both the frequency and the loss amount are specified.

We now introduce some new notations. The observed sample will be of the form $\mathbf{x} = (n_2, n_3, \mathbf{z_3})$, where n_2 denotes the number of observations in A_2, n_3 the number of observations in A_3 and $\mathbf{z_3} = (z_{31}, \ldots, z_{3n_3})$ the concrete observations in A_3. The corresponding sample space will be denoted as $\mathcal{X}$. Following the notations of Dempster et al. (cf. (8)), the observed sample $\mathbf{x}$ can be interpreted as the image of a mapping T from the complete sample space $\tilde{\mathcal{X}}$ into $\mathcal{X}$. The complete sample $\tilde{\mathbf{x}}$ has the form

$$\tilde{\mathbf{x}} = (\mathbf{n}, \mathbf{z}) = (n_1, n_2, n_3, \mathbf{z_1}, \mathbf{z_2}, \mathbf{z_3}) \in \tilde{\mathcal{X}} \tag{3.1}$$

and the mapping T is given by:

$$T : \begin{cases} \tilde{\mathcal{X}} & \to & \mathcal{X} \\ \tilde{\mathbf{x}} = (\mathbf{n}, \mathbf{z}) & \mapsto & \mathbf{x} = (n_2, n_3, \mathbf{z_3}) \end{cases} \tag{3.2}$$

Given that the total number of observations in the complete sample is unknown, one possible *joint* density on $\mathcal{X}$ (with respect to the product of counting and Lebesgue

measures) which is consistent with the model specification in equation (2.1), is given by the following expression:

$$g_{\lambda,\gamma}(\mathbf{x}) = \frac{(\triangle T\tilde{\lambda})^{n_2+n_3}}{(n_2+n_3)!}e^{-\triangle T\tilde{\lambda}}\binom{n_2+n_3}{n_2}\left(\frac{q_2}{1-q_1}\right)^{n_2}\left(\frac{q_3}{1-q_1}\right)^{n_3}\prod_{i=1}^{n_3}\frac{f_\gamma(z_{3i})}{q_3} \quad (3.3)$$

where $q_j := Q_\gamma(A_j)$ denotes the probability for a random realization of Q_γ to fall into set A_j, $j = 1, 2, 3$ such that $q_1 + q_2 + q_3 = 1$, $\tilde{\lambda} := (1 - q_1)\lambda$ and $\triangle T := T_2 - T_1$ is the length of the sample window. In the representation (3.3), the Poisson process $\tilde{N}_t$ that counts only the losses of magnitudes greater than u_1 is interpreted as a *thinning* of the original process N_t, with a new intensity $\tilde{\lambda} = (1 - q_1)\lambda$.

It is easy to see that the maximization of the corresponding log-likelihood function can be divided into two separate maximization problems, each depending on only one parameter:

$$\hat{\lambda}_{\text{MLE}} = \arg\max_{\lambda} \log g_{\lambda,\hat{\gamma}_{\text{MLE}}}(n_2, n_3, z_3) = \frac{n_2+n_3}{\triangle T(1 - Q_{\hat{\gamma}_{\text{MLE}}}(A_1))} \quad (3.4)$$

$$\hat{\gamma}_{\text{MLE}} = \arg\max_{\gamma} \log\left(\left(\frac{q_2}{1-q_1}\right)^{n_2}\cdot\left(\frac{q_3}{1-q_1}\right)^{n_3}\cdot\prod_{i=1}^{n_3}\frac{f_\gamma(z_{3i})}{q_3}\right) \quad (3.5)$$

$$(3.6)$$

3.2. Expectation-Maximization algorithm

Calculations of the two parts (3.5) and (3.4) separately will simplify the calculation considerably. However, the maximization of the first expression (3.5) might be tedious. In the cases where no closed-form expression for the MLE-estimate of γ is available we cannot further simplify the problem. Equation (3.5) must be solved numerically. But in the cases where a closed-form expression for both the unconditional MLE-estimate of γ as well as the conditional expectation $E_{\tilde{\gamma}}(\log f_\gamma(z) \,|\, z \in A_j)$, $j = 1, 2, 3$ for given $\tilde{\gamma}$ is available (for distributions such as Gaussian, Lognormal, Exponential, 1-parameter Pareto) we can achieve a significant reduction of the computational effort by applying the *Expectation-Maximization* algorithm (EM-algorithm). The algorithm, examined by Dempster et al. in (cf. (8)), is exactly designed for the situation of likelihood estimation with incomplete data.

The procedure of the EM-algorithm is the following: given an initial guess for the unknown parameter γ the missing data values in the log-likelihood function are replaced by their expected values. This leads to the guess for the complete log-likelihood function which is further maximized given the closed-form expression for the MLE. The solution is then used as the initial guess in the next iteration of the algorithm. The EM-algorithm is a two-step procedure, consisting of the E-step, and the M-step:

Start: Choose initial values $(\lambda^{(0)}, \gamma^{(0)})$ for the unknown parameters (λ, γ).

E-step: Calculate the *expected log-likelihood function*

$$E_{(\lambda^{(0)},\gamma^{(0)})}\left(\log \tilde{g}_{\lambda,\gamma}(\tilde{\mathbf{x}}) \,|\, \mathbf{x}\right)$$

of the complete sample $\tilde{\mathbf{x}}$ given the incomplete sample $\mathbf{x}$ and the guess $(\lambda^{(0)}, \gamma^{(0)})$.

M-step: Maximize the expression for the expected log-likelihood function with respect to the unknown parameters

$$(\lambda^{(1)}, \gamma^{(1)}) := \arg\max_{\lambda,\gamma} E_{(\lambda^{(0)},\gamma^{(0)})}\left(\log \tilde{g}_{\lambda,\gamma}(\tilde{\mathbf{x}}) \,|\, \mathbf{x}\right)$$

to obtain new values for λ and γ.

Iteration: Repeat the E-step and the M-step with the new values for λ and γ, until convergence is achieved.

In order to make use of the described EM-algorithm we need a specification of the corresponding complete sample. Therefore, we will provide in a next step an expression for the density $\tilde{g}$ of the complete sample $\tilde{\mathbf{x}}$ on $\tilde{\mathcal{X}}$, assuming at first that the full information is available.

$$\begin{aligned}\tilde{g}_{\lambda,\gamma}(\tilde{\mathbf{x}}) &= \frac{(\triangle T\lambda)^{n_1+n_2+n_3}}{(n_1+n_2+n_3)!}e^{-\triangle T\lambda}\cdot\frac{(n_1+n_2+n_3)!}{n_1!n_2!n_3!}\cdot q_1^{n_1}q_2^{n_2}q_3^{n_3}\prod_{j=1}^{3}\prod_{i=1}^{n_j}\frac{f_\gamma(z_{ji})}{q_j}\\ &= \frac{(\triangle T\lambda)^{n_1+n_2+n_3}}{n_1!n_2!n_3!}e^{-\triangle T\lambda}\cdot\prod_{j=1}^{3}\prod_{i=1}^{n_j}f_\gamma(z_{ji}) \qquad (3.7)\end{aligned}$$

The idea of representation (3.7) can be summarized as follows. Given the Poisson distributed total number of losses $n_1+n_2+n_3$ we can treat the numbers (n_1, n_2, n_3) as the outcomes of a Multinomial random variable with corresponding probabilities

(q_1, q_2, q_3). The specific observations $z_j i$, $i = 1, ..., n$ in A_j are realizations of the corresponding conditional distribution with density $f_\gamma / q_j \cdot \mathbf{1}_{A_j}$. Let us consider the corresponding log-likelihood function $l_{\lambda,\gamma}(\tilde{\mathbf{x}}) := \log \tilde{g}_{\lambda,\gamma}(\tilde{\mathbf{x}})$ associated with equation (3.7):

$$l_{\lambda,\gamma}(\tilde{\mathbf{x}}) = (n_1 + n_2 + n_3) \log(\triangle T \lambda) - \triangle T \lambda - \sum_{j=1}^{3} \log n_j! + \sum_{j=1}^{3} \sum_{i=1}^{n_j} \log f_\gamma(z_{ji}) \quad (3.8)$$

In the E-step, we will treat the occurring magnitudes separately. With the known values of n_2, n_3 and $\mathbf{z_3}$, we first estimate the expected values of $q_j^{(0)}$ and $n_1^{(0)}$, given the initial guess values of $\gamma^{(0)}$ and $\lambda^{(0)}$, in the following steps. Given $\gamma^{(0)}$, the expected values $q_j^{(0)}$ of q_j, $j = 1, 2, 3$ are known as well. Then, given $(\lambda^{(0)}, \gamma^{(0)})$, the expected value of n_1 is given by $n_1^{(0)} = q_1^{(0)} \lambda^{(0)} \triangle T$. The expression $\sum_{j=1}^{3} \log n_j!$ can be left out from the maximization problem as it contains neither λ nor γ. With these notations we hence obtain in the E-step:

$$\begin{aligned} E_{(\lambda^{(0)},\gamma^{(0)})}(l_{\lambda,\gamma}(\tilde{\mathbf{x}}) \,|\, \mathbf{x}) &= E_{(\lambda^{(0)},\gamma^{(0)})}\left(\log \tilde{g}_{\lambda,\gamma}(\tilde{\mathbf{x}}) \,|\, \mathbf{x}\right) \\ &= (n_1^{(0)} + n_2 + n_3) \log(\triangle T \lambda) - \triangle T \lambda + \ldots \\ &\quad \ldots + n_1^{(0)} \cdot E_{\gamma^{(0)}}(\log f_\gamma(z) \,|\, z \in A_1) + \ldots \\ &\quad \ldots + n_2 \cdot E_{\gamma^{(0)}}(\log f_\gamma(z) \,|\, z \in A_2) + \ldots \\ &\quad \ldots + \sum_{i=1}^{n_3} \log f_\gamma(z_{3i}) \end{aligned}$$

By assumption, we have a closed-form expression for the conditional expectations of the log-likelihood function. The M-step can now be treated separately for λ and γ. We obtain:

$$\lambda^{(1)} := \frac{n_1^{(0)} + n_2 + n_3}{\triangle T} \quad (3.9)$$

$$\begin{aligned} \gamma^{(1)} := \arg\max_\gamma \Bigg(& n_1^{(0)} \cdot E_{\gamma^{(0)}}(\log f_\gamma(z) \,|\, z \in A_1) + \ldots \\ & \ldots + n_2 \cdot E_{\gamma^{(0)}}(\log f_\gamma(z) \,|\, z \in A_2) + \sum_{i=1}^{n_3} \log f_\gamma(z_{3i}) \Bigg) \end{aligned} \quad (3.10)$$

By assumption, the solution to the latter maximization problem (3.10) can be expressed as a function of the observed sample $\mathbf{x}$ and the conditional expected value of

the log-likelihood function. Repetition of the steps leads to a sequence $(\lambda^{(k)}, \gamma^{(k)})_{k\in\mathbb{N}_0}$ that converges to the desired MLE-estimates $(\hat{\lambda}_{\text{MLE}}, \hat{\gamma}_{\text{MLE}})$.

In this section we have presented an efficient methodology for calculating consistent parameter estimates. A consistent estimation of the unknown parameters under the missing data specification can be conducted in one of the two following ways:

(a) If the closed-form expression for the MLE-estimators do not exist in terms of the sufficient statistics, then the conditional likelihood function is maximized numerically, using equations (3.5), (3.4);

(b) If the closed-form expression for the MLE-estimators exist in terms of the sufficient statistics, then the EM-algorithm is an efficient procedure for parameter estimation, using equations (3.9), (3.10).

The next section examines the effect of ignoring and accounting for the missing and/or censored data on the VaR measure under alternative adjustment procedures.

4. Value at risk and implications of wrong data specifications

4.1. Aggregated operational loss process

In the previous section it was demonstrated how the distributional parameters of the complete-data set can be estimated given the observed data with censored and/or missing data. Having estimated the unknown parameters of the frequency and severity distributions, we are now able to estimate the VaR measure that governs the required capital charge, for each business line and event type combination. We fix a tolerated risk level α and a time horizon of length $\triangle t$ - the Basel Committee suggests to use $\triangle t = 1$ year (cf. (2)), and α can be taken as $\alpha = 1\% - 5\%$. The $\text{VaR}_{\triangle t, 1-\alpha}$ is then given by equation (2.3). It equals the capital charge which must be maintained in order to protect against potential operational losses in $\triangle t$ from now with a probability of $(1-\alpha)$. Generally no closed-form expression for the cumulative loss distribution is available. The Lévy-Chintschin-formula leads to the following expression for the characteristic function:

$$\varphi_{S_t}(u) := Ee^{iuS_t} = \exp(t\int_{\mathbb{R}}(e^{ius} - 1)\lambda Q_\gamma(ds)) \tag{4.1}$$

The upper quantiles have to be determined numerically through approximations such as the recursive Panjer-Euler scheme, FFT-inversion from the characteristic function or simulation. For the special case of a sub-exponential loss distributions Q, such as Lognormal, Pareto and the heavy-tailed Weibull, we have the following approximation for large values of s (cf. (11), (5)):

$$P(S_t > s) \sim EN_t \cdot Q(s, \infty), \quad s \to \infty \tag{4.2}$$

whereas for light tailed loss distributions the Cramer-Lundberg estimate can be used to get an initial guess for the desired quantile. References include (4), (12) on approximation methods, (10) on the heavy-tailed case, (9) and (16) on the Panje-Euler scheme. We will focus our exposition on the case where the loss distribution family $\mathcal{Q}$ equals the class of Lognormal distributions, i.e. we suppose $Q_\gamma = \mathcal{LN}(\mu, \sigma^2)$. This specification equals the one advised by the Basel Committee (cf. (2)). Given a complete sample $\mathbf{x} = (x_1, ..., x_n)$ of n loss amounts, the Lognormal distribution fulfils the necessary requirements of the EM-algorithm as the following closed form expressions for the MLE-estimates of μ and σ^2 are available for the general case

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \log x_i \tag{4.3}$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \log^2 x_i - \hat{\mu}^2_{\text{MLE}} \tag{4.4}$$

Therefore, this leads to the required conditional expectations that can be explicitly obtained, from our earlier discussion.

4.2. Impact of density misspecification on the capital charge

The operational loss problem with data collection includes the case when the data below a specified threshold is not recorded. Thus, we suppose that we have $u_1 = u_2 =: u$, i.e. our observed sample $\mathbf{z} = (z_1, \ldots, z_m)$ of loss amounts that follow a Lognormal distribution over the time interval $[T_1, T_2]$ consists of a truncated sample and contains no censored data. In this case, we see several possible approaches banks may undertake for the parameter estimation and subsequently the VaR determination:

1. Determine the MLE-estimates $\hat{\lambda}_{\text{MLE}}$, $\hat{\mu}_{\text{MLE}}$ and $\hat{\sigma}^2_{\text{MLE}}$ for the unknown parameters λ, μ and σ^2 with the presented EM-algorithm and determine the

asymptotic VaR estimate using equation (4.2):

$$\widehat{\mathrm{VaR}}_{\triangle t,1-\alpha} = \exp(\hat{\mu}_{\mathrm{MLE}} + \hat{\sigma}_{\mathrm{MLE}}\Phi^{-1}(1 - \frac{\alpha}{\hat{\lambda}_{\mathrm{MLE}}\triangle t})) \tag{4.5}$$

For simulations, draw losses from the unconditional distribution using the complete-data estimated parameters via the EM-algorithm, and use the complete-data frequency parameter.

2. Estimate the frequency parameter λ by the observed frequency $\hat{\lambda} = m/(T_2 - T_1)$ and estimate the complete-data conditional distribution using the EM-algorithm. Simulating losses from the conditional distribution $Q_{\hat{\gamma}}(\cdot|\cdot > u)$ with the estimated observed frequency will lead to asymptotically correct results 4.5. This follows from (4.2).

3. (Naive approach) Use the observed frequency $\hat{\lambda} = m/(T_2 - T_1)$ and fit the unconditional distribution to the truncated data. This oversimplified and misspecified approach will lead to biased estimates for the parameters λ, μ and σ^2 of the loss distribution. The bias can be expressed analytically:

$$\begin{aligned} E\hat{\lambda}_{\mathrm{observed}} &= \lambda \cdot Q_\gamma(Z > u) = \lambda + \mathrm{bias}(\hat{\lambda}_{\mathrm{observed}}) \\ &= \lambda \cdot \left(1 - \Phi\left(\frac{\log u - \mu}{\sigma}\right)\right) \end{aligned} \tag{4.6}$$

$$\begin{aligned} E\hat{\mu}_{\mathrm{observed}} &= E\left(\frac{1}{m}\sum \log Z_i \,|\, Z_i > u\right) = \mu + \mathrm{bias}(\hat{\mu}_{\mathrm{observed}}) \\ &= \mu + \sigma \cdot \frac{\varphi\left(\frac{\log u - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log u - \mu}{\sigma}\right)} \end{aligned} \tag{4.7}$$

$$\begin{aligned} E\hat{\sigma}^2_{\mathrm{observed}} &= E\left(\frac{1}{m}\sum \log^2 Z_i - \hat{\mu}^2_{\mathrm{observed}} \,|\, Z_i > u\right) = \sigma^2 + \mathrm{bias}(\hat{\sigma}^2_{\mathrm{observed}}) \\ &= \sigma^2 \left(1 + \frac{\log u - \mu}{\sigma} \cdot \frac{\varphi\left(\frac{\log u - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log u - \mu}{\sigma}\right)} - \left(\frac{\varphi\left(\frac{\log u - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log u - \mu}{\sigma}\right)}\right)^2\right) \end{aligned} \tag{4.8}$$

where φ and Φ denote the density and distribution function of the standard Normal law.

In (4.6) above, $\mathrm{bias}(\lambda) < 0$ always. Since the bias of the location parameter μ is always positive, then the observed μ is always overstated. The effect on the VaR

estimates according to equation (4.2) depend on the values of u, μ and σ^2. For practical purposes, it is reasonable to expect that $\log u < \mu$ (the threshold level is relatively low), and so the bias of the scale parameter is negative, and hence σ^2 is underestimated. The error of the VaR estimate $\widehat{\text{VaR}}$ given by equation (4.5) by replacing the estimates for μ and σ^2 with their expected values given by equations (4.7) and (4.8). We obtain the following approximation:

$$\widehat{\text{VaR}}_{\triangle t,1-\alpha} \approx \exp\left(\mu + \text{bias}(\mu) + (\sigma + \text{bias}(\sigma))\Phi^{-1}(1 - \frac{\alpha}{(\lambda + \text{bias}(\lambda))\triangle t})\right)$$

which can be demonstrated by the following simple example:

$q = 5\%$		$\mu = 14.5$	$\mu = 15.5$	$\mu = 16.5$
		$\sigma^2 = 0.1732$	$\sigma^2 = 1.0488$	$\sigma^2 = 2.6636$
	VaR $[_{10^{10}}]$	0.00093	0.02430	0.63368
	$\widehat{\text{VaR}}$ $[_{10^{10}}]$	0.00083	0.01832	0.40403
	abs.$[_{10^{10}}]$	0.00010	0.00598	0.22965
	rel.(%)	10.84	24.60	36.24
$q = 10\%$		$\mu = 14.5$	$\mu = 15.5$	$\mu = 16.5$
		$\sigma^2 = 0.2853$	$\sigma^2 = 1.7277$	$\sigma^2 = 4.3878$
	VaR $[_{10^{10}}]$	0.00146	0.07154	3.54110
	$\widehat{\text{VaR}}$ $[_{10^{10}}]$	0.00116	0.04183	1.50560
	abs.$[_{10^{10}}]$	0.0003	0.0297	2.0355
	rel.(%)	19.59	41.53	57.48

TABLE 1: Estimates of absolute and relative errors for VaR.

Table 1 presents the absolute and relative VaR error results for two examples: $q := Q_\gamma(Z < u) = 5\%$, and 10%, for which the values of σ^2 are chosen to keep the fractions constant. We chose $u = 1,000,000$ (based on external databases' minimum threshold), $\alpha = 1\%$, $n = 100$, $\triangle t = 1$ year. The parameters μ and σ^2 are based on realistic values for external operational loss data. The absolute values of the errors from the true VaR (original figures are negative values) were estimated using abs. $= \widehat{\text{VaR}}_{1\text{yr},99\%} - \text{VaR}_{1\text{yr},99\%}$, rel. $= \frac{\widehat{\text{VaR}}_{1\text{yr},99\%} - \text{VaR}_{1\text{yr},99\%}}{\text{VaR}_{1\text{yr},99\%}} \times 100\%$. Higher true σ^2 values (or, alternatively, lower true μ values) result in more significant underestimation of the VaR figures. Moreover, for a higher fraction of missing data (q) the errors

are more dramatic. Thus, if the missing data is completely ignored, the resulting operational capital charge would be understated.

5. Conclusions and final remarks

This paper has illustrated that ignoring the missing (or truncated) data and fitting unconditional distribution to the available data set leads to biased estimates. The bias depends on the threshold level at which the data is truncated. The paper has focused on a particular example of operational losses, that are subject to minimum collection thresholds in the internal and external banks' databases. The presented estimation procedure can be useful for practitioners, and may be equally applied to modeling other financial losses such as credit losses, insurance claims, etc.

The paper has in particular demonstrated the use of the Expectation-Maximization algorithm in estimating the true parameters of Lognormal loss distribution as suggested by the Basel Committee. Given the assumption that operational losses follow the Lognormal law, the paper estimates the 'information loss' as the fraction of the data that has not been recorded. The model can be of course generalized to other loss distributions. We treat the more general framework and conduct an extensive empirical analysis with the operational loss data in a forthcoming paper.

An inevitable outcome of misspecified distributions is misleading (underestimated) capital charge estimate governed by the VaR measure. The paper provides the estimates for the impact of misparametrization on the asymptotic behavior of VaR. VaR figures are accurate only provided that the conditional distribution is fit to the incomplete recorded data set at hand. A more coherent measure of risk, the Expected Tail Loss, would be also false under the wrong parametrization. These and other issues will be addressed in our further papers.

Acknowledgements

Rachev gratefully acknowledges research support by grants from Division of Mathematical, Life and Physical Sciences, College of Letters and Science, University of California, Santa Barbara, the German Research Foundation (DFG) and the German Academic Exchange Service (DAAD).

References

1 BIS (2001). Consultative document: operational risk. `www.bis.org`.

2 BIS (2001). Working paper on the regulatory treatment of operational risk. `www.bis.org`.

3 BIS (2003). The 2002 loss data collection exercise for operational risk: summary of the data collected. `www.bis.org`.

4 Buchwalder, M. ., Chevallier, E. . and Klüppelberg, C. . (1993). Approximation methods for the total claimsize distributions - an algorithmic and graphical presentation. *Mitteilungen SVVM* 187–227.

5 Cline, D. B. H. . (1987). Convolutions of distributions with exponential and subexponential tails. *Journal of Australian Mathematical Society Series* **A 43,** 347–365.

6 Cruz, M. G. . (2002). *Modeling, Measuring and Hedging Operational Risk*. John Wiley & Sons, New York, Chichester.

7 DeCanio, S. J. . and Watkins, W. E. . (1998). Investment in energy efficiency: Do the characteristics of firms matter? *The Review of Economics and Statistics* **80,** 95–107.

8 Dempster, A. P. ., Laird, N. M. . and Rubin, D. B. . (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* **39,** 1–38.

9 Dickson, D. C. M. . (1995). A review on panjer's recursion formula and its applications. *British Actuarial Journal* **1,** 107–124.

10 Embrechts, P., Grübel, R. and Pitts, S. M. . (1993). Some applications of the fast fourier transform algorithm in insurance mathematics. *Statistica Neerlandica* **47,** 59–75.

11 Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin.

12 Hipp, C. and Michel, R. (1990). *Riskotheorie: Stochastische Modelle und Statistische Methoden*. Verlag Versicherungswirtschaft e.V., Karlsruhe.

13 Jorion, P. (2000). *Value-at-Risk: the New Benchmark for Managing Financial Risk* second ed. McGraw-Hill, New York.

14 McLachlan, G. and Krishnan, T. (1997). *The* EM *Algorithm and Extensions*. Wiley Series in Probability and Statistics, John Wiley & Sons.

15 Meng, X.-L. . and van D̃yk, D. (1997). The em algorithm - an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B (Methodological)* **59,** 511–567.

16 Panjer, H. H. . and Willmot, G. (1992). *Insurance Risk Models*. Society of Actuaries, Schaumburg, Illinois.

17 Wulfsohn, M. S. . and Tsiatis, A. A. . (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53,** 330–339.