LOSS SEVERITY DISTRIBUTION ESTIMATION OF OPERATIONAL RISK USING GAUSSIAN MIXTURE MODEL FOR LOSS DISTRIBUTION APPROACH

***Seli Siti Sholihat*[1]**
***Hendri Murfi*[2]**

[1]*Department of Accounting, Faculty of Economics, Universitas Gunadarma*
[2]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia,*
seli.siti.sholihat@gmail.com
hendri@ui.ac.id

Abstract

Banks must be able to manage all of banking risk; one of them is operational risk. Banks manage operational risk by calculates estimating operational risk which is known as the economic capital (EC). Loss Distribution Approach (LDA) is a popular method to estimate economic capital(EC).This paper propose Gaussian Mixture Model(GMM) for severity distribution estimation of loss distribution approach(LDA). The result on this research is the value at EC of LDA method using GMM is smaller 2 % - 2, 8 % than the value at EC of LDA using existing distribution model.

Keywords: *Loss Distribution Approach, Gaussian Mixture Model, Bayesian Information Criterion, Operational Risk.*

INTRODUCTION

Bank must be able to manage all of banking risk, one of them is operational risk. A common industry definition of Operational Risk is the risk of direct or indirect loss resulting from inadequate or failed internal processes, people or systems, or from external events", Frachot[4]. Bank manage operational risk by calculate estimation of operational risks which is known as the economic capital (EC).

Economic Capital (EC) is the amount of capital that an organization must set aside to offset potential losses. There are three approach to calculate Economic Capital based on Basel Accord II. That are Basic Indicator Approach (BIA), Standardized Approach (SA), dan Advanced Measurement Approach (AMA), (Frachot, 2001). The capital charge using BIA dan SA is calculated by fixed percentage. The capital charge using AMA, bank could calculated EC based on their internal loss data. Internal data is used as an input to compute the probability distribution of loss. The popular approach of AMA is Loss Distribution Approuch(LDA).

Mathematics definition , the total of annual operational Losses :

$$Z(t) = \sum_{i}^{N(t)} X^{(i)}(t) \qquad (1)$$

Where:

N (t) : Random Variable of the number events losses in 1 year.

Distribution of N (t) is called **frecuency Distribution**

$X^{(i)}(t)$: Random Variable of the amount losses for the i-th event.

Distribution of $X^{(i)}(t)$ is called **Severity Distribution**

$Z(t)$: Annual losses, is summarize of the loss $X^{(i)}(t)$ in 1 year.
Distribution of $Z(t)$ is called **Aggregation Distribution**

In LDA method, loss severity distribution (severity distribution) and loss frequency distribution (frequency distribution) must be estimated and then aggregate distribution is formed from both of them. Through LDA method, the value of EC can be gotten from Value at Risk (VaR) in aggregate distribution with the level of confidence reaches 99,9%. Aggregate distribution of the random variable Z can not expressed analytically. So that the numerical approach is needed to determine the distribution. Several well-known numerical method that could be used are the Monte Carlo method, the Fast Fourier Transform, and Panjer Recursion. In the study used the most easily implemented, namely the Monte Carlo method (Shevchenko,2009). That why our research would used its method. One of problems on LDA is severity distribution estimation that used a model on particular distribution cannot describe a data well through. Then severity distribution estimation based on data is used to solved this problem.

RESEARCH METHOD

One of methods that estimate probability distribution function based on data is *Gaussian Mixture Model* (GMM). GMM is parametric method that estimate probability density of random variable. Probability density of GMM is a linear combination of several Gaussian distribution, that is :

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \qquad (2)$$

Where:
$p(x)$: probability of x
K : the number of gaussian distribution that is used
π_k : k-th mixing coefisien, $\sum_k \pi_k = 1$ dan $0 \leq \pi_k \leq 1$.
$\mathcal{N}(x|\boldsymbol{\mu_k}, \Sigma_k)$: Normal /Gaussian Distribution k-th, where k=1,2,…,

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_k) = \frac{1}{\sqrt{(2\pi|\boldsymbol{\Sigma_k}|)}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_k})^{\mathrm{T}} \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu_k})\right\}$$

Each gaussian distribution $\mathcal{N}(x|\mu_k, \Sigma_k)$ is called component of *mixture*, and each componet have different mean $\boldsymbol{\mu_k}$ dan covarian $\boldsymbol{\Sigma_k}$. GMM is formed by parameter $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, dan $\boldsymbol{\Sigma}$, where $\boldsymbol{\pi} = (\boldsymbol{\pi_1}, , \boldsymbol{\pi_2}, \dots, \boldsymbol{\pi_k})$, $\boldsymbol{\mu} = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \dots, \boldsymbol{\mu_k})$ dan $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2} \dots, \boldsymbol{\Sigma_k})$. Parameter π_k is called *mixing* coefisient. Ilustration of GMM show in Bishop, C. M.[1].

The question is "Which is a better K for GMM (K=?)". Number of component in GMM could be selected using model selection. There are two popular model selection that is used, Akaike Information Criterion(AIC) and Bayesian Information Criterion(BIC). Due to the selection model, BIC has proven consistent in estimating the density function of the mixture model, Dempster (1977),. BIC also proved consistent in choosing the number of components in the mixture model (Claeskens and Hjort, 2011). Those are the reason of choosing BIC in this study.

The best model using BIC is taken by giving a score to each model and then choose the model that has the smallest score. Here is the calculation of scores on the model BIC, Claeskens and Hjort[2] :

BIC = -2ln (L (θ)) + dim (θ) ln (n)
where:
L (θ) : the value of the likelihood function model with the estimated parameters θ
n : number of data.

RESULT

The Software of simulations use programming language Python. The simulation in this paper calculate the value of EC using LDA in which severity distribution is estimated by GMM. First step on simulating, we generated data toy for operational risk(Assumed operational risk real data), with those data then we estimate frequency distribution with a Poisson distribution and estimate severity distribution using k-GMM(Selection model for k using BIC). Next, the simulations to be done to generate more data for LDA which appropriate with operasional risk data. The result of simulation on LDA is EC value. Then to see how GMM works on LDA, EC value in which GMM applied compare with EC value in which other distribution model applied.

Data are generate in 3 group of data: 3 years, 5 years, and 10 years. histogram of risk data is below figure 1.

First, estimating frecuency distribution, Frequency of losses per year in operational risk are the values for the random variable N, which is the number of frequency of losses incurred within one year. The distribution of this random variable N can be estimated with a Poisson distribution, this is because the number of frequency of losses incurred in a particular year does not depend on the number of frequencies in other years. Parameters on the Poisson distribution is the mean μ. For data 3 years: $\mu = 165$, for data 5 years: $\mu = 60$, and for 10 years: $\mu = 54$. Frequency distribution that is formed can be seen in the following figure 2.

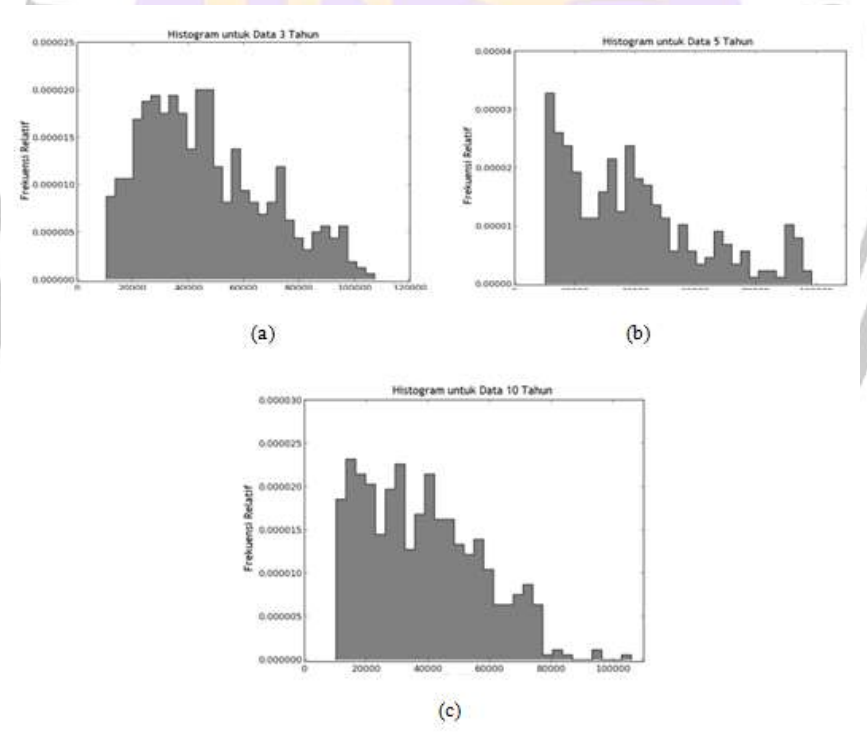

figure 1. (a) Histogram of data risk data of 3 years (b) Histogram of risk data of 5 years and (b) Histogram of risk data of 10 years

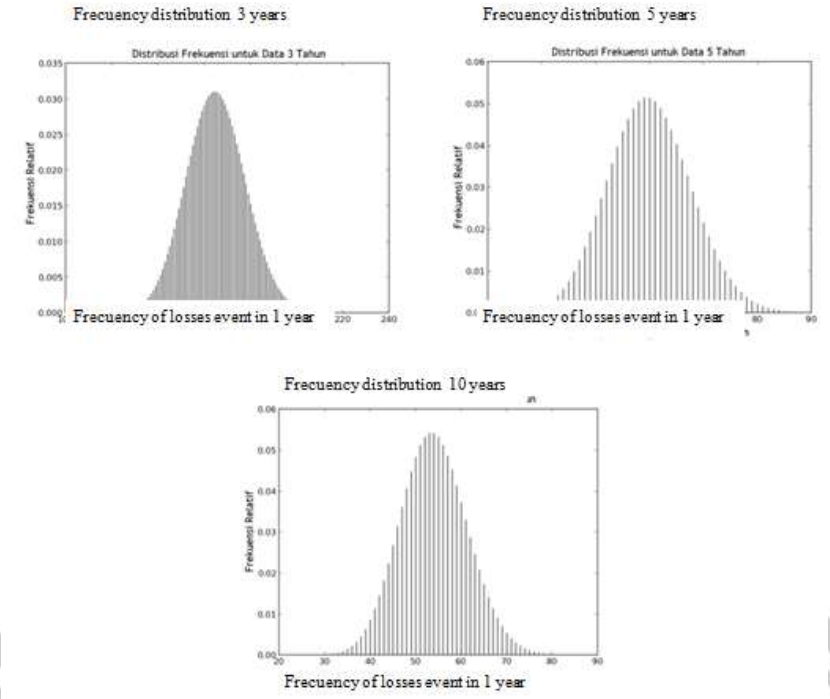

figure 2. Frecuency Distribution

Estimating severity distribution using GMM, GMM is a parametric models, so the thing to do is to determine the parameters-parameters in the GMM. Simulations performed on three groups of data, 3 years, 5 years and 10 years. Component K in GMM for each of data determined in advance by the selection model BIC. BIC methods are iterative methods in determining the optimal model by scoring in each model of different components k, optimal model is a model that has the smallest scoring with K smallest components. At 3 years of data, the method of BIC produce optimal k= 2. At 5 years of data, the method of BIC produce optimal component k= 4. At 5 years of data, the method of BIC produce optimal component K = 3.

Table 1. Parameter-parameter of GMM using data 3 years, 5 years dan 10 years

Data 3 years			
Component	**Coefisien mixing**	**Mean**	**Varians**
1	0.4055	66436.9129	3.23732875e+08
2	0.5945	32211.7052	1.34890412e+08
Data 5 years			
Component	**Coefisien mixing**	**Mean**	**Varians**
1	0.1511	67914.4881	1.13370445e+08
2	0.5348	36794.2922	1.02323533e+08
3	0.2511	14826.8864	1.00002549e+07
4	0.0630	92860.0602	5.95976389e+06
Data 10 years			
Component	**Coefisien mixing**	**Mean**	**Varians**
1	0.4041	54793.2608	2.07430867e+08
2	0.2056	16073.4845	1.20531853e+07
3	0.3903	32175.9726	8.76020500e+07

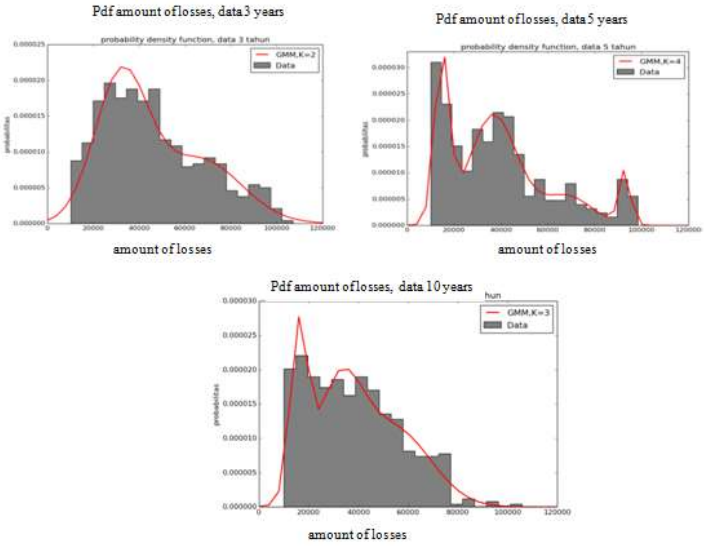

figure 3. Severity Distribution using GMM with K component choosing by BIC

The following is ilustration of GMM for data 10 years with 3 component gaussian where the each parameter on table 1.

$$p(x) = \sum_{k=1}^{3} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

$$= \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \pi_3 \mathcal{N}(x|\mu_3, \Sigma_3)$$

$$p(x) = 0.4041\mathcal{N}(x|54793.2608, 2.07430867\text{e}+08)$$
$$+0.2056\mathcal{N}(x|16073.4845, 1.20531853\text{e}+07)$$
$$+0.3903\mathcal{N}(x|32175.9726, 8.76020500\text{e}+07)$$

The graphic of severity distribution using GMM are on figure 3. The red curve in Figure 3 are curve of GMM for each data. Figure 3 also saw us that the curve is very good in estimating the data, visible from the ridge on the histogram followed properly by the red curve.

How k component on GMM estimate the severity distribution, for k = 1,2,3,4, and k = 10. This estimation was performed on three groups of data, 3 years, 5 years of data, and the data 10 years. This estimating was conducted to visually whether the selection of the best models with BIC able approximating data well and compare it with other GMM models. The following Figure 4, shows the probability density function models GMM for k = 1,2,3,4 and 10. For k = 10, appears to lack of smoothness curve pdf, pdf increasingly tapered curve. Moreover, it appears the estimated GMM with a large k (k = 10) is not too different from the estimated optimal GMM with K obtained by BIC .

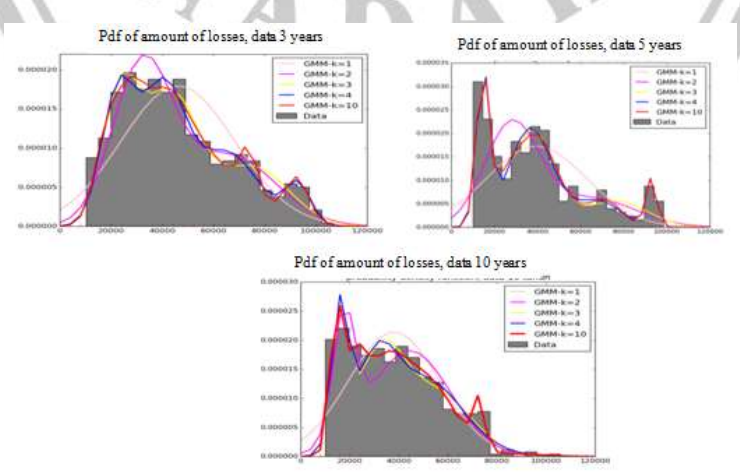

figure 4. Severity Distribution

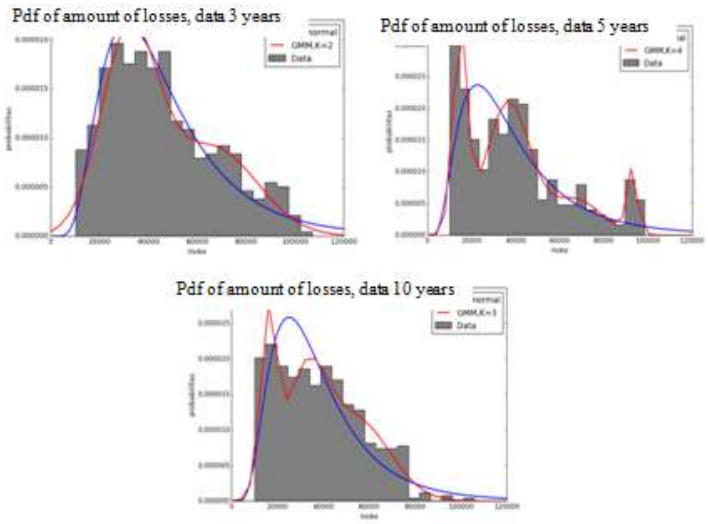

Figure 5. Comparing severity distribution using GMM (red curve) and Log-Normal(blue curve)

Visually to data 3 years, the best model looks to GMM with 3 components (yellow curve), GMM with 2 component is still not good approximation. This is in contrast with the best GMM models produced by the method of BIC which 2 component is an optimal component. Data 5 years, the best model seems to GMM 4 component (blue curve). 10 years of data, the best model seems to GMM 3 component (yellow curve). Data for 5-year and 10 year, the selection of the best models with visualization are same as with the selection of the best model with BIC. How ever best model GMM for each data can be done visually in the case of data one dimension as above. If data have large dimension, it would be difficult to portray the data graphically so the selection of components visually difficult. In addition, the selection of the optimum component in a visual way can not be justified because that are subjective.

Red curve in Figure 5 shows for 3 group of data used in this study, pdf using GMM better in describing the research data because it can estimate the local areas, while pdf model of log-normal can not do it.

The simulation is calculating EC values. As we know that EC obtained from the calculation of VaR on Aggregate distribution (formed from the severity distribution and frequency distribution) with a confidence level of 99.9%. Aggregate distribution calculated numerically using the Monte Carlo method. The purpose of this simulation to determine how much difference the value of EC produced by LDA using GMM and EC produced by LDA using the Log-Normal. The number of samples used were 1.10, 10^2, 10^3, 10^4, 10^5, and 10^6. The simulation was performed 10 times for each sample number. Results of the simulation calculations are presented in table 2.

Tabel 2. EC using GMM and Log-Normal for number of sampel 10^6

Method	**Economic Capital (EC)**		
	Data 3 years	**Data 5 years**	**Data 10 years**
GMM	9.729.364,21	3.557.837,80	3.089.042,94
Log-Normal	9.901.079.80	3.632.659.70	3.178.200.00

Table 2 shows that using GMM on severity distribution of LDA gives a lower EC value than the Log-Normal. EC value by GMM of 3 group of data provide EC value 2% lower than the value of the EC with the Log-Normal.

CONCLUSIONS

The result on this research is estimation of severity distribution through GMM is better than known distribution model in describing the data. The value at EC of LDA method using GMM is smaller 2 % - 2,8 % than the value at EC of LDA using existing distribution model. Then if bank use this method, they could have capital efficiency.

REFERENCES

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer: New York.

Claeskens, G., and Hjort, N.L.(2011). *Model Selection and model Averaging*. Cambrigde University Press.

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, 39(1).*

Frachot, A., Georges, P., and Roncalli, T. (2001). Loss Distribution Approach for Operational Risk. *Working Paper*, Groupe de Recherche Operationnelle: France.

Jimene, E.J., Feria, M.J., and Martin, J.L. (2007). Economic Capital for Operational Risk: Applying Loss Distribution Approach(LDA). *Working Paper*. Consejeria de Innovacions: Andalucia.

McLachlan, G.J., and Krishnan, T. (1997). *The EM Algorithm and its Extensions*. Wiley.

McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Model*. Wiley.

Ming, H.Z., Qian, S.C. (2005). An Approach VaR for capital markets with Gaussian Mixture. *Journal of Applied Mathematics and Computation,* 168, pp. 1079-1085.

Shevchenko, P.V. (2009). Implementing Loss Distribution Approach for Operational Risk. *Applied Stochastic Models in Business and Industry*, vol. 26(3), pp. 277-307.