18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

An Application of Data Fusion Techniques in Quantitative Operational Risk Management

Sabyasachi Guharay
Systems Engineering & Operations Research
George Mason University
Fairfax, Virginia U.S.A.
sguhara2@masonlive.gmu.edu

KC Chang
Systems Engineering & Operations Research
George Mason University
Fairfax, Virginia U.S.A.
kchang@gmu.edu

Abstract - In this article we show an application of data fusion techniques to the field of quantitative risk management. Specifically, we study a synthetic dataset which represents a typical mid-level financial institution's operational risk loss as defined by the Basel Committee on Banking Supervision (BCBS) report. We compute the economic capital needed for a sample financial institution using a Loss Distribution Approach (LDA) by determining the Value at Risk (VaR) figure along with the correlation measures by using copulas. In addition, we perform computational studies to test the efficacy of using a "universal" statistical distribution function to model the losses and compute the VaR. We find that the Lognormal-Gamma (LNG) distribution is computationally robust in fusing the frequency and severity data when computing the overall VaR.

Keywords: Operational risk, Statistical Distribution fitting, Data Fusion, low-probability events, Value at Risk (VaR), heavy tailed distributions.

1 Introduction

The application of data fusion techniques to various different disciplines in applied sciences and engineering has been a popular research topic recently. In a nutshell, the paradigm of data fusion can be thought of "... the scientific process of integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and technically useful representation" [1]. In the present environment, the tool of "data fusion" has been numerously applied to various engineering fields such as sensor networks; defense and intelligence; aerospace; homeland security; public security; medical technology etc. There has been a somewhat paucity of *direct* application to the field of quantitative risk management. This paper addresses one novel application which serves as an interesting applied problem valuable to practitioners in the field. In the broadest sense of terms, quantification of risk management involves analyzing the events which tend to be *remotely probable* as opposed to focusing only on those which are *reasonably* possible. To better understanding the relevance of this field, we begin by introducing the concept of applying data fusion in the risk framework next. Afterward, we give a brief overview of the risk management framework. We then give a quick overview of the specific risk management framework, namely operational risk. Afterwards, we describe the specific problem of interest studied in this paper and the methodology used. Next we show our results and present discussions. Finally, we narrate our conclusions, current ongoing work and future research directions.

1.1 Data Fusion in Risk Framework

In most scientific and engineering fields, the investigators are interested in studying the behavior of events which are typically occurring (i.e. occur in the "body" of a statistical distribution). In most cases, events which occur rarely are classified as "outliers" and ignored (or even sometimes thrown out). It is in fact a part of human nature as argued by Nobel Laureate economist Daniel Kahneman in Prospect Theory [2] where he shows from psychological experiments that humans view *near-zero* probabilities as identical to zero probability. This mindset is the exact opposite of what is practiced in risk management, specifically operational risk management. The recent 2008 Financial Crisis, showed that the so-called "Black Swan" [3] events can occur and potential devastate the world economy. Thus, it may be "human nature" to ignore or neglect these *low*-probability outlier types of events, but in a risk management context, these events are crucial to be properly modeled and examined. While the mathematics behind low-probability events has been well-studied since the 1940s, applying it in a risk management framework is still considered somewhat of an art partially due to the difficulties that data come from various correlated sources. In the current risk management practice, many simplifications and assumptions are made to the mathematics which makes the risk management decision making process incomplete. The primary reason behind these simplifications is that there are multiple sources of data and the science of integrating them properly is not well understood and practiced. Therefore, we believe that using data fusion in this field is a promising application which has high economic significance. We will next motivate our work further by discussing the basic foundations of the risk management application.

1.2 The Risk Management Framework

Risk management framework has been developed extensively in the past couple of decades mainly used for financial institutions. Most financial institutions for example banks, insurance companies, hedge funds, etc. are regularly exposed to several different types of risks which are easy to observe such as market risk along with credit risk. Market risk can be broadly thought of as changes to the overall/macro financial conditions (such as stock prices, interest rates) which can adversely affect the portfolio value of a financial institution. Credit risk can be broadly thought of the risk from a failing counterparty. These two risks have been extensively studied and there is a good confluence between theory and practice. There is a third, an equally important, branch of risk management which is known as the operational risk management. This is a newer type of risk and is defined as the following: "The risk of loss resulting from inadequate or failed internal processes, people and systems or from external events" [4]. Examples of this can include a rogue trader, hurricane Katrina, credit card fraud, tax non-compliance etc. The losses resulting from this type of risk comes from multiple data sources and types. Thus the application of data fusion principles is apt for this field. To manage the risk, there is a regulatory agency called the Basel Committee for Banking Supervision (BCBS) which regulates and stipulates that financial institutions are *required* to mitigate themselves from this type of risk by holding Economic Capital of an appropriate amount to absorb these losses. In otherwords, financial institutions are required to hold a "rainy day" fund to absorb shocks which result from operational risk. But how much should they hold? If they hold too little, then if a large shock occurs, then the financial institution can get wiped out. But if they hold too much capital, then they are losing out on opportunity costs of making profits. This is one of the fundamental questions. From a mathematical point of view, this concept is described as Value at Risk (VaR). A VaR of V dollars represents that one is X% sure of not losing more than V dollars in time T. So the practitioner sets the time T and probability X *a priori*, and computes V accordingly. One of the goals in operational risk management is to *accurately* compute the VaR value of V when data comes from multiple sources. The other is to compute the *expected* (i.e. average) loss that one can expect.

Using the latest Basel III framework, loss data are officially categorized according to seven Basel defined event types and eight defined business lines [5]. The business lines are the following: (1) Corporate Finance (CF); (2) Sales & Trading (S&T); (3) Retail Banking (RB); (4) Commercial Banking (CB); (5) Payment & Settlement (P&S); (6) Agency Services (AS); (7) Asset Management (AM); and (8) Retail Brokerage (RB) [5]. The seven event types for losses are the following: (1) Internal Fraud (IF); (2) External Fraud (EF); (3) Employee Practices & Workplace Safety (EPWS); (4) Clients, Products, & Business Practice (CPBP); (5) Damages to Physical Assets (DPS); (6) Business Disruption & Systems Failures (BDSF); and (7) Execution, Delivery, & Process Management (EDPM) [5]. After the 2008 financial crisis, the BCBS performed a "Loss Data Collection Exercise for Operational Risk" [5]. In this paper, we study one of these data sets (for an anonymized small financial institution). We use data fusion techniques to model three different business lines and their correlation structure to compute a final VaR figure.

2 Operational Risk Framework

Now that we have introduced the general framework above, we briefly narrate the fundamentals of the modeling of operational risk using the Loss Data Approach (LDA) [6-11]. When modeling operational risk, there are two fundamental components: (1) Frequency of losses; (2) Severity of losses. The simplest explanation is that one is interested in how often losses will occur (frequency), and also how large will the losses be when they occur (severity). Banks and other financial institutions obviously dread the instances where large losses (severity) happen in large occurrences (frequency). This is known as a high probability high impact event. Contrary to the fears of many chief financial officers, these types of event almost never takes place. The reason is that most banks have proper risk management practices which would identify key risk indicators (KRIs) that can prevent/mitigate frequent occurrences of large losses. In otherwords, any good financial institution will have checks in place to ensure that their employees can not regularly steal billions of dollars. So if there is a rogue employee committing theft, it should be a *rare* event, and not a frequent event. Instead, what is more important is the low probability high impact, i.e. rare occurrences of large losses.

According to the guidelines from the BCBS, the aggregated losses from operational risk can be described in a paradigm such as the random sum model [6]. The joint loss process (consisting of frequency and severity) is assumed to follow a stochastic process $\{S_t\}_{t \geq 0}$ expressed as the following:

$$S_t = \sum_{k=0}^{N_t} L_k \ , L_k \overset{iid}{\sim} F_\gamma \qquad (1)$$

The paradigm expressed by the above equation assumes that the severity (i.e. loss magnitudes) are independent and identically distributed (i.i.d.) sequence of $\{L_k\}$. Since the $\{L_k\}$ are i.i.d., one can assume that they come from a cumulative distribution function (CDF), F_γ. This CDF can be statistically characterized as belonging to a parametric family of continuous probability functions. Likewise, the the counting process N_t is assumed to follow a discrete counting process or a probability mass function. The key point here is that in Eq. (1) there is an inherent assumption of independence between severity and frequency distributions. In Figure 1, we graphically illustrate how the frequency and the severity process are traditionally thought as "independent" (silo) processes which come together to

calculate the annualized aggregate loss. The frequency of losses are estimated along with the severity of the losses using two different statistical distributions. Then one can combine these approaches and use Monte Carlo (MC) simulation, to compute the annualized aggregate loss. Once the aggregate loss distribution has been determined, one can estimate the mean (expected) loss and also upper quantiles to get an estimate of the operational risk VaR. Most banks tend to estimate at least a 99.9% (if not higher to 99.99%, which would hold for a 1 in 10,000 year event).

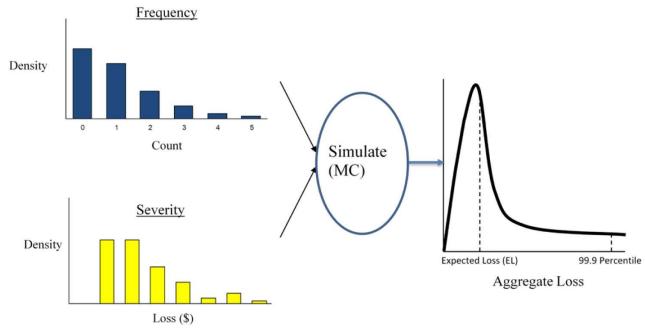

Figure 1. Illustration of computing the VaR

The natural question that arises next is how does one measure the frequency and the severity? In practice, most banks have an internal loss data collection exercise which they calculate for every year. So the operational risk modeler can fit the losses that were collected (L_1, L_2, ..., L_N) to get the severity distribution. Likewise a similar approach can be used to statistically estimate how often the losses are happening to get the frequency distribution. These are thought of as two distinct data sources that need to be "fused" to arrive at a combined estimate.

2.1 Frequency Distributions

There are three main types of distribution which can be used to the model the frequency of losses: (1) Poisson; (2) Binomial; and (3) Negative Binomial distribution. The Poisson distribution has a unique characteristic among the class of statistical distributions in that it's mean (μ) is equal to its standard deviation (σ). Also this distribution is characterized by a single parameter, λ. This distribution is the easiest to model since it involves only fitting a single parameter. The binomial distribution can be fully characterized by two parameters, n (sample size) and p (probability). Similarly, the negative binomial distribution can also be characterized by two parameters, r (number of failures till success) and p (probability). In terms of mean and variance, the binomial distribution is appropriate when $\mu > \sigma$, while the negative binomial distribution is appropriate when $\mu < \sigma$.

In most instances one can tell which frequency distribution to use by simply computing the relationship between sample mean and sample variance. Overall, there is not much difference when using different frequency distributions. Figure 2 shows similarity of the frequency distributions between Poisson, Binomial and Negative-Binomial distributions. It shows that in most cases there is not a great benefit to derive the *ideal* frequency distribution. A notable exception would be if historical loss data collection exercise of a bank shows say $\mu > \sigma$ in all cases (empirically). In this case, one should choose a binomial distribution as a fit for the frequency. Likewise the same would be true if the reverse was observed and then the negative-binomial distribution could be used.

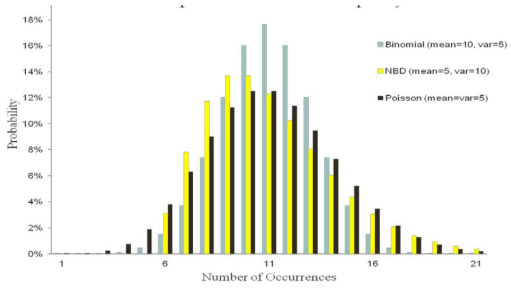

Figure 2. Comparison of different frequency distributions

2.2 Severity Distribution types

Unlike the case of the frequency, there are a plethora of valid statistical distribution that one can use to fit the severity data. We list (for illustrative purposes only) a sample of distributions that one may use: (1) Lognormal (since losses are always non-negative); (2) Burr XII distribution; (3) Generalized Pareto (GPD); (4) Weibull; (5) Pareto; and (6) Lognormal-Gamma (LNG) [7].

Among the distributions, a unique one which we study in this paper is the three parameter Lognormal-Gamma (μ, σ, κ) distribution. The first parameter represents the mean, the second parameter represents the standard deviation, and the third parameter represents the kurtosis (fourth moment). This distribution comes from the statistical property of convolution of distribution functions. Analytically, the CDF for LNG [7] can be expressed as the following:

$$F(x \mid \mu, \sigma, \kappa) = \int_0^\infty \gamma(y|\kappa)\phi(x|\mu, \sigma^2 * y)dy \qquad (2)$$

where $\gamma(y|\ \kappa)$ corresponds to the pdf of the gamma distribution while $\phi(x|\ \mu,\ \sigma^2)$ is the pdf for the normal distribution which is characterized by a population mean μ and population variance σ^2.

Note that there is not a closed form solution for equation (2). Similar to the "error" (Erf) function for the Gaussian distribution cdf, the distribution for the Lognormal-Gamma has to be computed numerically. Thus the problem with this distribution is that one cannot write an analytical expression for the CDF, and thus generating random numbers takes longer since one cannot use the inverse CDF method from simulation. However it is extremely useful for our applications because the Lognormal distribution is a special

case of the Lognormal-Gamma distribution (i.e. when κ = 3). So the strength of this distribution is that one can directly model and interpret "heavy tails" (i.e. those with κ > 3) for any dataset.

Figure 3 illustrates a sample operational risk loss data set for the severity where there exists in almost all cases a loss data collection threshold, T [7]. The reason is that most financial institutions will only keep an inventory of these losses but not the small losses below a threshold T in their own Loss Data Collection exercise that they undertake [5]. That is why in Figure 3, the loss severity histogram is shown starting from $10,000 and moving forward.

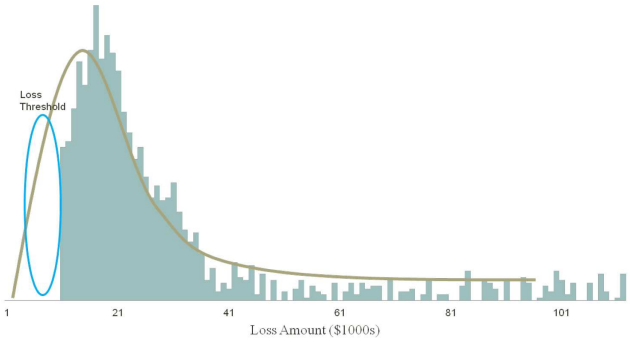

Figure 3. Sample severity loss data

3 Methodology

As mentioned in the section 2, there has been an extensive loss data collection exercise collected by the BCBS in 2009 [5]. Most of these loss data sets are highly proprietary in nature. However, many studies have reported the statistical parameter estimates (severity, frequency, and correlations) for typical financial institutions losses [5-7]. With this in mind and based on the first author's personal experience studying mid-level financial institution's loss data, we generate a synthetic dataset which resembles a mid-level financial institution involving three different business lines along with one event type of Internal Fraud. The three business lines are the following: (1) Corporate Finance; (2) Sales & Trading; and (3) Retail Banking. We first compute the VaR assuming independence between the business lines and then use the methodology of copulas to model correlation among the business lines. In order to do that, we examine if there is a unique and most appropriate severity distribution that can be used for modeling the loss severity. If a universal severity distribution can be found, then this will be useful for fusing the severity and frequency losses when computing the aggregated VaR figure. To this end, we simulate losses from different heavy-tailed severity distributions. We then fit the simulated data to various types of severity distributions and check if one type of severity distribution can perform well universally.

3.1 Fitting the loss data

There are two main statistical techniques to fitting the data: (1) Maximum Likelihood Estimation (MLE); and (2) Minimum Distance Estimation. In this paper, we focus on the MLE method because it is also primarily used by practitioner's in the operational risk field.

The MLE method can be used for a data set of losses L_1, L_2, ..., L_N which come from a distribution F with the parameter set θ. Then the MLE approach requires computing the log-likelihood (LL) function as the following for the density f:

$$LL(\theta|L_1,L_2,...,L_N)=\log\left(\prod_{i=1}^{n} f(L_i|\theta)\right) \quad (3)$$

The MLE approach is to find the value of $\hat{\theta}_{MLE}$ which can maximize the LL function. In almost all cases, this can be computed numerically. As previously mentioned, one of the challenges for operational risk loss data, is that there is a data collection threshold. Therefore, we need to use the corrected MLE approach which accounts for left-censoring of the data [7]. This approach involves computing the new LL function as below with the data collection threshold T:

$$LL_{Truncated}(\theta|T,L_1,L_2,...,L_N)=\log\left(\frac{\left[\prod_{i=1}^{n} f(L_i|\theta)\right]}{[1-F(T|\theta)]^n}\right) \quad (4)$$

One can then maximize the θ vector in Eq. (4), to obtain the correct MLE estimates. The frequency data can be fit by simply using the sample mean as the estimate for the Poisson distribution's parameter.

3.2 Monte Carlo Method for Fusing Severity & Frequency Distributions

Now that the severity and the frequency distribution have been determined, we can calculate via Monte Carlo simulations, the economic capital (EC) for operational risk by integrating the two together. The algorithm is outlined in the following:

1. Determine the Severity Distribution and optimal parameters from censored MLE fits
2. Determine optimal Frequency Distribution parameter
   2.1 Set a simulation number (usually a minimum of 10,000 runs)
3. Set the iteration counter t = 1.
4. Draw a random number of losses from the Frequency Distribution, n
5. Given the number n, draw n losses, L_1, L_2, ..., L_n from the severity distribution.
6. Sum all n of the severity losses to obtain the aggregate value A_t (Aggregate Loss for time t).
7. Set t = t+1, and go to step 4.
8. Iterate till t hits the maximum iteration threshold.

9. $\{A_1, A_2, ..., A_t\}$ is the Aggregate Loss distribution. Empirically compute the mean, and 99.9 percentiles to get expected loss (EL) and VaR.

3.3 Correlation among Business Lines

In many instances, one can treat the severity and frequency data from different business lines as independent. However, for many *smaller* financial institutions, the losses tend to be correlated amongst different lines. Therefore, we need a robust statistical model to account for the correlation.

The standard Pearson's correlation coefficient ρ, is useful if we know *a priori* that the correlations are linear. However, if the dependence across the distribution is not linear, we will have to employ other methodology such as copula [12-13] to model the correlations.

Broadly speaking, copula is a mathematical method for modeling the joint distribution of simultaneous losses. It is used to model the dependence structure of a multivariate distribution (i.e. more than one business line for example) separate from the marginal distribution without having to specify a unified, joint distribution. Mathematically, suppose that the random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ which consists of n random variables, has a multivariate CDF, F_Y with continuous marginal univariate CDFs, $F_{Y_1}, ..., F_{Y_n}$. With the inverse CDF method, one can easily show that $F_{Y_1}(Y_1)$ follows a Uniform[0,1] distribution. Then, the CDF of $\{F_{Y_1}(Y_1), ..., F_{Y_n}(Y_n)\}$, C_Y, is a defined as a copula. We will apply two well-known copulas, Gaussian copula and a *t*-Copula to account for tail dependence between different business lines.

4 Results & Discussion

We begin by showing the characteristics of the data that we analyze from the loss data collection exercise.

4.1 Characteristics of Data set

Figures 4-6 show the scatter plots of the data for each pair of the three business lines. From the figures, it is clear that correlation is present amongst the business lines. Also we notice some potential outliers which we mark in red.

We apply the Gaussian and *t*-Copulas to estimate the correlation across the business lines. We use MATLAB to estimate the correlation structure using the Gaussian and *t*-Copula (including the degrees of freedom (df)) via MLE. The results are shown in Table 1.

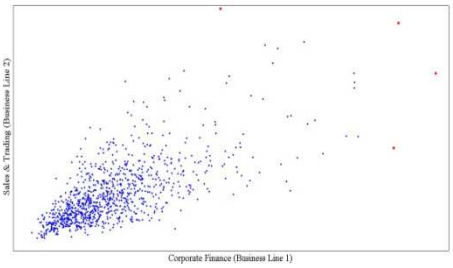

Figure 4. Plot of the loss (severity) across business lines 1 and 2; the red dots indicate potential "outliers"

We next show the plots across Business Line 2 and 3 along with Lines 1 and 3.

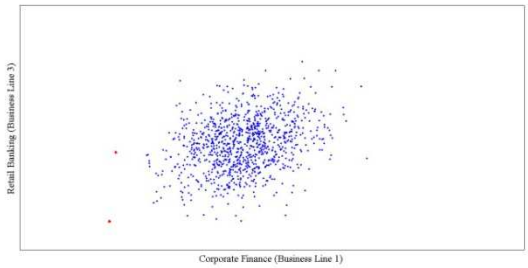

Figure 5. Plot of the loss (severity) across business lines 1 and 3; the red dots indicate potential "outliers"

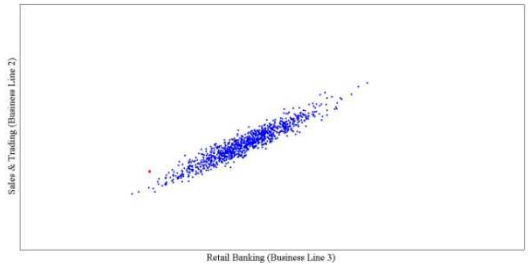

Figure 6. Plot of the loss (severity) across business lines 2 and 3; the red dots indicate potential "outliers"

Correlation	Gaussian	*t*-Copula (df)
Lines 1 & 2	0.04	0.05 (33)
Lines 1 & 3	0.23	0.26 (44)
Lines 2 & 3	0.91	0.89 (55)

Table 1: Copula results for the dataset

4.2 Universal severity distribution for fusing severity and frequency

We need to now determine which severity distribution is most appropriate in fitting the loss data. In Section 2, we mentioned several distributions such as Weibull, Lognormal, Burr etc. Instead of arduously fitting all severity distribution types and then applying statistical goodness of fit tests (such as Chi-squared, Cramér-von Mises, Anderson-Darling, etc.) to identify the best one, we intend

to find a *universal* statistical distribution which can fit most heavy-tailed types of data well.

In order to do so, we conduct extensive computational analysis. We simulated a large dataset (of size 10,000,000) of a heavy tailed distribution modeled by a Lognormal-Gamma distribution with (μ=9,σ=2, κ=5). We then fit it to the following distributions: (1) Weibull, (2) Lognormal, (3) Lognormal-Gamma, (4) GPD, (5) Burr and (6) Pareto. Instead of doing graphical/statistical tests of goodness of fits, we compare the percentile values as shown in Figure 7 below.

Notice how one can get a quick estimate of the fit by just looking at the percentile comparisons. For example at the 99.9%, the true value is around $28 million, and the GPD does an under-estimate of $10 Million, while the Burr does an underestimate of $12 million (if these were losses for example). Notice how the Weibull and Pareto fail completely to fit this heavy-tailed data. This is expected since Weibull is known to be a thin-tailed distribution, and Pareto is a single parameter distribution. Obviously, the Lognormal-Gamma fits itself quite well.

True Severity Distribution		Fitted Severity Distributions					
Percentile	Lognormal-Gamma	GPD	Pareto	Lognormal	Lognormal-Gamma	Burr	Weibull
99.95	67,342,538	42,937,414	-	5,848,528	67,792,930	35,091,196	5,243,867
99.9	28,738,314	19,243,328	-	3,917,913	28,555,005	16,204,984	4,044,601
99.5	3,955,866	2,980,913	-	1,400,242	3,959,597	2,688,693	1,968,048
99	1,660,437	1,332,633	-	850,126	1,662,351	1,236,956	1,344,850
98	691,981	593,784	-	492,813	690,778	566,493	863,546
95	212,919	201,338	-	217,512	212,551	198,482	418,356
90	85,464	86,675	-	105,169	85,316	87,166	204,730
50	8,096	7,977	-	8,103	8,096	8,053	7,857
25	2,731	2,561	-	2,102	2,732	2,505	721
0	0	-	-	-	0	-	-

LNG	
Parameters	Theoretical
Mean	9
Standard Deviation	2
Kurtosis	5

Figure 7. Fitting randomized data; Burr and LNG perform well

The next experiment focuses on the aggregate distribution of losses, which is the primary interest for risk practitioners. Here, we assume a Poisson frequency distribution with a fixed parameter value of λ = 10 (10 losses per annum), and calculate the VaR simulation as shown in Figure 8 below.

It is interesting to note that while the Burr distribution has not performed well in the MLE fit, the Aggregate Loss distribution estimates are very reasonable. The true expected loss was actually around $6 million while the Burr distribution estimated it around $3 million. For the 99% value, the Burr estimated an $17 million value, while the actual value was near $29 million. The Peak-over-Threshold (POT) distributions such as GPD and Pareto completely overestimate the VaR and are not suitable for general practice.

True Severity Distribution		Fitted Severity Distributions					
Percentile	Lognormal-Gamma	GPD	Pareto	Lognormal	Lognormal-Gamma	Burr	Weibull
99.95	67,342,538	42,937,414	-	5,848,528	67,792,930	35,091,196	5,243,867
99.9	28,738,314	19,243,328	-	3,917,913	28,555,005	16,204,984	4,044,601
99.5	3,955,866	2,980,913	-	1,400,242	3,959,597	2,688,693	1,968,048
99	1,660,437	1,332,633	-	850,126	1,662,351	1,236,956	1,344,850
98	691,981	593,784	-	492,813	690,778	566,493	863,546
95	212,919	201,338	-	217,512	212,551	198,482	418,356
90	85,464	86,675	-	105,169	85,316	87,166	204,730
50	8,096	7,977	-	8,103	8,096	8,053	7,857
25	2,731	2,561	-	2,102	2,732	2,505	721
0	0	-	-	-	0	-	-

Simulation	
	λ = 10
n	500,000

Aggregate Loss Distribution		Fitted Aggregate Loss Distributions					
Percentile	Lognormal-Gamma	GPD	Pareto	Lognormal	Lognormal-Gamma	Burr	Weibull
99.95	970,711,142	574,435,196	1.E+110	19,461,640	998,997,550	437,324,025	12,018,353
99.9	422,623,229	272,744,582	1.E+102	14,371,042	441,888,771	200,306,580	10,128,089
99	29,114,170	19,413,124	3.E+74	4,485,347	29,303,601	16,896,568	5,100,455
95	4,351,288	3,404,891	5.E+54	1,865,052	4,376,122	3,112,206	2,793,277
90	1,954,863	1,646,165	2.E+46	1,220,588	1,959,988	1,554,138	2,014,171
75	687,292	639,572	2.E+34	641,118	684,631	626,369	1,140,538
50	282,757	283,867	6.E+23	325,484	283,998	284,422	578,701
25	136,377	141,029	3.E+15	165,024	136,260	142,344	269,970
0	-	-	0.E+00	-	-	-	-
Expected Loss (EL)	6,065,383	9,266,219	5.E+180	598,329	9,294,097	3,087,695	903,043

LNG	
Parameters	Theoretical
Mean	9
Standard Deviation	2
Kurtosis	5

Figure 8. Using fusion of severity and frequency; LNG and Burr perform well when computing the overall VaR.

Figure 9 shows the test results using the GPD as the true distribution. Interestingly as shown in the figure, the GPD fails to fit itself at the $0 threshold. It can only fit itself from a certain positive threshold ($100K in this example). This is not surprising, since GPD comes from the Extreme Value Theory (EVT) class of POT distributions. We also notice from the figure that for the MLE portion only, the Lognormal-Gamma and the Burr does a reasonable job in the fit. Looking at the MLE portion only, the Burr does the best job. For the lower ends of the distribution, like at the 25th percentile, the Burr is showing a value of around $6,641 while the actual value is $6,527. For the higher ends of the tail, the 99.95% actual value is around $141 million while the Burr is showing around $145 million. The Lognormal-Gamma performs the second best under the MLE fits criterion. However, we are primarily interested in the VaR analysis. Therefore, when one moves to the aggregate loss in Figure 9, we observe that the Lognormal-Gamma performs as well as the Burr in fitting this theoretical Aggregate Loss distribution from a GPD severity and Poisson frequency of $\lambda \approx 19$. In reality, the GPD is not commonly used due to its numerical stability issues. However, the figure below shows that even if GPD was the "true" severity distribution, the three parameter Lognormal-Gamma distribution can perform well to estimate the Aggregate Loss. While the three parameter Burr distribution may marginally perform the "best" amongst all distributions, it is not at all intuitive to interpret the meaning of the parameter estimates from a Burr distribution. On the other hand, for each of the three parameters of the Lognormal-Gamma distribution there is a clear intuitive and statistical interpretation, namely, mean, variance and kurtosis. We therefore prefer the LNG over the Burr for overall VaR analysis.

Theoretical		Fitted					
Percentile	Generalized Pareto (GPD)	GPD	Pareto	Lognormal	Lognormal-Gamma	Burr	Weibull
99.95	141,928,759	325,215,655	1,692,405,580	13,243,855	359,861,863	145,378,707	84
99.9	62,839,351	141,739,615	696,458,926	8,961,886	160,626,371	63,295,855	18
99.5	9,091,859	20,649,685	88,621,537	3,286,914	24,392,100	9,170,229	0
99	3,939,528	9,033,654	36,469,545	2,020,786	10,844,545	3,984,361	0
98	1,712,675	3,972,298	15,007,951	1,187,616	4,730,319	1,726,164	0
95	558,754	1,367,342	4,640,600	535,071	1,570,728	564,694	0
90	234,758	631,921	1,909,700	263,486	671,962	237,006	0
50	20,535	146,560	243,001	21,652	83,592	20,813	0
25	6,527	114,802	144,558	5,812	32,846	6,641	0
0	0	100,000	100,000	-	0	0	0

Simulation	λ = 19
n	1,000,000

Theoretical		Fitted					
Percentile	GPD	GPD	Pareto	Lognormal	Lognormal-Gamma	Burr	Weibull
99.95	1,623,600,048	1,872,340,698	1,739,064,861	1,575,066,325	1,610,512,681	1,620,941,942	1,523,879,392
99.9	1,561,823,477	1,688,781,718	1,618,189,717	1,539,488,109	1,554,564,367	1,561,974,340	1,511,772,588
99.5	1,512,704,991	1,537,669,403	1,518,982,040	1,412,571,897	1,509,462,322	1,510,733,088	613,887,772
99	1,036,434,631	1,515,301,991	1,505,604,547	703,530,617	948,294,833	1,033,218,295	354,376,337
98	465,593,179	857,700,137	688,395,096	352,154,853	433,777,113	461,783,585	204,377,468
95	168,569,855	308,147,441	223,833,725	142,348,715	160,154,429	164,335,206	98,727,603
90	81,965,155	148,328,201	100,381,916	73,963,283	78,229,591	79,214,035	57,875,815
50	17,867,542	30,008,964	17,139,090	16,232,686	16,013,287	15,980,502	15,593,591
25	11,314,767	17,854,058	9,730,576	9,610,633	9,478,678	9,460,952	9,554,080
0	1,792,832	603,002	552,935	558,873	237,363	561,911	565,517
Expected Loss (EL)	54,684,921	86,451,318	64,575,399	46,442,791	51,241,954	52,536,969	34,168,060

Generalized Pareto	
Parameters	Theoretical
Scale	19,000
Shape	-1.2
Fit Threshold	100,000
Sample Size	10,000,000

Figure 9. Using fusion of severity and frequency; LNG performs reasonably well when computing the overall VaR.

4.3 Fitting the Loss Data & Computing VaR

From the previous section we found that the Lognormal-Gamma performs well for fitting heavy-tailed distributions. Therefore we apply it for our severity and Poisson for our frequency. We fit across two different thresholds of \$0 and \$100,000 (100^K). The reason is that the data had very few (less than 2% data) between \$0 and 100^K. The results are shown in Figure 10 where the estimated parameters of the Lognormal-Gamma distributions from the three business lines are given.

T=0	Business Line 1	Business Line 2	Business Line 3	T=100K Combined		
Percentile	Lognormal-Gamma	Lognormal-Gamma	Lognormal-Gamma	Lognormal-Gamma	Lognormal-Gamma	Lognormal-Gamma
99.5	79,781	80,307	360,682,883	75,759,526	37,912,861	19,810,792
99	46,796	46,001	174,223,349	52,633,180	21,651,086	11,983,128
98	26,213	25,680	81,526,026	35,224,153	12,223,226	7,142,991
95	11,219	10,840	26,740,876	19,246,995	5,303,192	3,449,836
90	5,349	5,110	10,398,054	11,418,858	2,679,272	1,884,738
50	407	404	441,006	1,888,232	363,892	322,404
25	106	108	86,378	752,063	182,151	173,784
0	0	0	0	100,000	100,000	100,000
Parameters						
Mu (Mean)	6.01	6	13	14.41	11.23	11.54
Sigma (Standard Dev)	2.02	2	2.5	1.43	2.10	1.78
Kurtosis	3.1	3.2	3.3	3.10	3.20	3.30

Figure 10. Fit of the loss severity data using the Lognormal-Gamma (LNG) distribution;

We use the Lognormal-gamma distribution also to measure the heaviness of the tail. We next proceed to fitting the frequency and then using Monte Carlo to compute the VaR. With the copula correlations obtained from Table 1, we conduct the Monte Carlo simulation (using a \$100K threshold) as described in Section 3.2 to estimate the overall VaR by integrating (fusing) severity and frequency of loss events across the three business lines. The results are given in Figure 11. Notice that the frequency we obtained was approximately 2.29 (per annum) for losses above the \$100K threshold.

The dataset in Figures 4-6 show that there are some outlier tail events and this the t-Copula modeling seems to be most suitable. As shown at the bottom of Figure 11, it is interesting to observe that due to the presence of correlations, the VaR t-Copula provides a most conservative economic capital value estimate. The difference is quite large (approximately 50% increase) from the naive independence assumption across business lines. This shows the importance of incorporating copulas when there is evidence of correlations across business lines.

T=100K	Business Line 1	Business Line 2	Business Line 3
Percentile (VaR)	Lognormal-Gamma	Lognormal-Gamma	Lognormal-Gamma
99.95	349,491,530	358,625,560	147,663,614
99.9	260,748,551	219,181,286	100,825,482
99	95,946,014	44,464,445	23,942,258
95	42,961,269	13,235,199	8,244,291
90	28,374,123	7,427,169	5,008,262
EL (Expected Loss)	12,042,764	4,177,325	2,496,125

Simulation Length	500,000
Frequency	2.29
VaR (Normal Copula)	427,880,649
VaR (t-Copula)	520,638,286
VaR (Independent)	355,240,793

Figure 11. Result of the Monte Carlo simulation; the frequency fit is shown here along with the 500,000 simulation runs.

5 Conclusion and Future Research

In this paper, we have studied an application of data fusion techniques to a problem in quantitative risk management. We study a synthetically generated typical mid-level financial institution's operational risk characteristics and computed the VaR value using correlations modeled by copulas. We found the presence of correlations across the Business Lines and the t-Copula estimate was most conservative and appropriate. We also studied data fusion technique of which severity distribution can be *universally* applied a priori. We found strong computational evidence of using the three-parameter Lognormal-Gamma distribution. We found that it can fit many types of heavy-tailed distributions reasonably well.

We are still continuing further study for testing the efficacy of using Lognormal-Gamma distribution as a universal source. Also we will investigate the applicability of using Panjer's algorithm [14-15], a method from actuarial science, along with the Fast Fourier Transform (FFT) from signal processing. The FFT and Panjer methods can only work for specific frequency and severity distributions. We expect to conduct further study with the FFT and Panjer methods to see which can perform the best data fusion among frequency and severity.

References

[1] Lawrence A. Klein, Sensor and data fusion: A tool for information assessment and decision making, SPIE Press, Washington, 2004.

[2] D. Kahneman and A. Tversky, “Prospect Theory: An Analysis of Decision under Risk,” *Econometrica*, vol. 47, no. 2, pp. 263, 1979.

[3] N.N. Taleb, The Black Swan: the impact of the highly improbable, Random House Publishers, USA, 2010.

[4] "Basel II: revised international capital framework," Basel committee, bis.org.

[5] "Results from the 2008 Loss Data Collection Exercise for Operational Risk", Basel Committee on Banking Supervision document.

[6] S.T. Rachev, A. Chaernobai and C. Menn, "Empirical examination of operational loss distributions," *Perspectives on Operations Research*, DUV, pp. 379-401, 2006.

[7] A. Samad-Khan, S. Guharay, B. Franklin, B. Fischtrom, P. Shimpi, M. Scanlon, "A New Approach for Managing Operational Risk: Addressing the Issues Underlying the 2008 Global Financial Crisis," Society of Actuaries, 2010 Research Paper.

[8] B. Ergashev, "Estimating the lognormal-gamma model of operational risk using the Markov Chain Monte Carlo method," *The Journal of Operational Risk*, vol. 4, no. 1, pp. 35, 2009.

[9] K. Dutta and J. Perry, "A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital," Boston Federal Reserve Bank working paper, 2006.

[10] G. Mignola and R. Ugoccioni, "Sources of Uncertainty in modeling Operational Risk Losses," *The Journal of Operational Risk*, vol. 1, no. 2, pp. 35, 2006.

[11] A. Chernobai, and S. Rachev, "Applying robust methods to operational risk modeling," *Journal of Operational Risk*, vol. 1, no. 1, pp. 27-41, 2006.

[12] D. Ruppert, Statistics and Data Analysis for Financial Engineering, Springer-Verlag, New York, 2010.

[13] A. Staudt, "Tail risk, systemic risk and copulas," In *Casualty Actuarial Society E-Forum,* vol. 2, 2010.

[14] D.C.M. Dickson, "A Review of Panjer’s Recursion Formula and Its Application," *British Actuarial Journal*, vol. 1, no. 1, pp. 107-124, 1995.

[15] H.H. Panjer, "Recursive evaluation of a family of compound distributions," *ASTIN Bulletin (International Actuarial Association)*, vol. 12, no. 1, pp. 22–26, 1983.