Risk portfolio modelling: from individual to collective models

Key issues and practical solutions

Nikita AKSENOV
18/07/2012
1 Introduction

The objective of this paper is to provide some practical solutions to ease the transformation of a fully parametrized “individual” portfolio model into a “collective” portfolio model with credit insurance portfolio as an application.

1.1 Risks portfolio modelling as an objective

To understand this objective, we should come back to the definition of risks portfolio modelling. Risks portfolio modelling is the Holy Grail that ultimately all risk-takers from finance to insurance and reinsurance industries would like to possess.

The fact is all players are already doing some form of risks portfolio modelling – in a way it all depends on what is meant really by “modelling” – such as risk monitoring or accumulation control. All this contributes to a better global understanding of risks taken.

Now what we mean by “risk portfolio modelling” is something more complete. For example, we would aim at having the entire distribution of the portfolio losses and/or results. We should also be able to:

- analyze the sensitivity of this distribution to some key parameters;
- decompose the distribution by operational entity or line of business or product etc.

The main problem with portfolio modelling is the number of items to model. Lets take the portfolio of a medium-size insurer with 1 million individual policies. Modelling precisely only one individual policy can be very complex. Ideally all sources of risks should be modeled and then all warranties should be applied. But this is not practical, except for some large industrial risks for which the characteristics are more precisely given. Even this is going to be far from perfection.

Then modelling the entire portfolio can be extremely time-consuming, if even possible. Should the insurer model all risks individually? How should he care of modelling of dependence between risks, if any? This dependence vary by nature of risk source: two risks distant from 10 km are independent from the point of view of fire but may or may not be dependent from the point of view of flood or hurricane...

Therefore, proper portfolio modelling needs to, first, determine the objectives of the modelling and, second, build a model that deals with the right level of granularity so that the model is manageable in term of parameters setting and calculation run-time.

With the above, we can see the main challenges of a portfolio model are:

- to set the right parameters;
- to use the appropriate calculation method to have the model to run with acceptable performance.

This paper will mainly deal with the latter challenge, i.e. how the final result could be calculated, given the parameters which are already set for a fully parameterized individual loss model.
This can be addressed by the use of a “collective model” where risks are not modeled individually (as within an “individual model”) but rather as a whole. For example, if you are interested in knowing what the expected “fire” loss is in a large portfolio, you would certainly not use a Monte-Carlo model of each risk individually and then aggregate the single losses. The use of a closed formula with basic assumptions is going to suffice in this case.

But real life situations are not always that simple. Having a closed formula does not always work. Instead, there are many cases where risks can be modeled individually, dependence between those risks is known or can be assumed. Hence the use of a simulation model (“individual model”) is the answer. Now, if the direct modelling of the result of the portfolio as a whole is possible (“collective model”) this can save time, avoid dealing with unnecessary details and provide numerous additional benefits.

1.2 Motivation

The above description of the portfolio risk modelling suggests multiple ways in which this process could be potentially used in practice.

In the particular case that we came across, the risk portfolio modelling is applied to credit risk reinsurance. A credit insurance portfolio risk was modelled in order to facilitate the analysis of the reinsurance protection strategy and also in order to be able to price this reinsurance protection. Basically, we are interested by such outputs as the insurance portfolio total annual loss distribution, as well as the distribution of the total annual loss burden of the reinsurer (all the relevant details regarding this credit reinsurance contract mechanics will be given in the next chapter which describes the individual loss model applied to a credit risk portfolio).

Different tools have been already developed in order to address this and similar modelling issues. In particular, we have dealt with a simulation platform developed with Matlab software, which was using the fully parameterized individual credit loss model in order to simulate loss occurrences for each individual obligor separately (“obligors” considered here as individual “risk units”). The correlation structure of the portfolio is also parameterized via a so called “frequency correlation matrix” which has as many elements as the total number of “risk units” squared. In fact, this leads quite often to a very large correlation matrix, with several tens or even hundreds of thousands of rows and columns, which could obviously be difficult to handle. In some cases the calculation time is too long to be considered as “efficient”, in other cases the calculation is even not feasible since the computer is unable to handle such a matrix.

The necessity to overcome the above mentioned difficulties and provide more practical methods for credit insurance portfolio modelling was the starting point for the present paper. To what extent could the individual model be replaced by the collective one in our particular context? Since we know that a collective loss model is always an approximation of the individual loss model, then how this approximation could be suitable for the final objective of the whole risk modelling process at stake?
1.3 Novelty of the subject

The computer age gave birth to a large extension of Monte-Carlo technique. Hence the use of closed formulas seems a bit "old fashioned". Using massive computing power apparently tends to reduce the need to use statistics and probability theory in order to reduce the inherent complexity of problems. This is true only for simple problems where calculation run-time is expressed in minutes or hours. Complex problems will always require to be first reduced before being simulated numerically. That is the case of risk portfolio modelling via collective modelling approach.

If

\[ N \] denotes the annual number of losses in the portfolio

and

\[ X \] denotes the individual loss amount given a loss occurs

then

the total portfolio annual loss distribution is a following sum of random variables with a random number of terms:

\[ S = \sum_{i=1}^{N} X_i \]

The calculation of \( S \) variable is straightforward to handle using a simple MonteCarlo simulation scheme if the above mentioned \( N \) and \( X \) variables are independent, which as we will explain below, is the case within an homogeneous risk portfolio composed of independent risks (please note that here we mean both frequency and severity independence).

A collective modelling approach is used quite often in non-life insurance in the particular case when all individual policy frequencies and severities are independent. But the frequency/severity independence for each and every policy (or risk unit) doesn't imply the frequency/severity independence at portfolio scale as a whole.

Also, within such a model even if the portfolio frequency distribution can be shown independent from the portfolio severity distribution, this doesn't imply that the total portfolio loss distribution \( S \) is going to be the same in the collective model and in the individual model.

In this paper we will expand the application scope of the individual and the collective risk models to different cases of risk dependence and also analyze the important implications of the collective model approximation.
The following “classical” actuarial closed formula for the portfolio annual loss (S) variance only holds when the portfolio frequency (N) is independent from the portfolio severity (X) and all X random variables are independent and identically distributed (iid):

$$\text{Var}[S] = E[N] \times \text{Var}[X] + \text{Var}[N] \times (E[X])^2$$

We will also generalize this closed formula for the portfolio annual loss variance calculation for different possible scenarios of risk dependence and heterogeneity.

1.4 Outline

We will first start by the detailed description of the individual loss model, where individual exposures (or policies, or risk units) can be either independent or correlated.

The application example chosen is a credit insurance risk portfolio with a full set of the individual model parameters. We will use one real life portfolio for all the practical applications throughout this document: all the theoretical concepts and principles introduced below are going to be illustrated with this same real life case.

Then we will examine in details the different methods of building of the collective “frequency/severity” model and the related important issues. Since the collective model is going to be an approximation of the individual loss model in most of the cases, we need to know what kind of approximation this could be and how the approximation quality can be improved.

Three different approaches are then suggested and thoroughly examined using the same real life portfolio example throughout:

- Partial collective model approach
- Portfolio stratification approach
- Alternative frequency distribution approach

The above mentioned approaches could be potentially useful when we need to use a collective model approximation.

In the final part, we will generalize the portfolio’s total loss variance formula, which provides us with a versatile tool for instant risk portfolio’s volatility assessment, without any need of numerical simulation.
2  Individual Risk Model: Credit / surety bond insurance portfolio example

Traditionally, a full blown Monte-Carlo simulation of individual obligors (each a risk) is used, an “individual portfolio model”.

In such model, we consider the credit event occurrence for each and every obligor within a portfolio. On the other hand, a collective model (fully discussed later) consists in just modelling of what happens in the portfolio as a whole, without making reference to any obligor in particular.

The individual model that we will use here is a traditional model which uses Gaussian copula defined by the “default” correlation matrix. By “default” we mean an event of “loss occurrence”.

In the individual model simulation process we explicitly generate the copula by Monte Carlo using Choleski decomposition of the default correlation matrix, previously adjusted for “standard Gaussian vector correlation structure” vs. “discrete defaults correlation structure”. (cf - Merton structural credit model)

2.1  General notation for the individual loss model

List of obligors $(C_i)_{i=1}^n = (C_1, ..., C_n)$
I-th obligor index $i$
Exposure (TSI or PML) $m_i$
Probability of default $p_i$
Annual number of losses $N_i$
LGD random variable $X_i$
LGD mean value $\mu_i$
LGD standard deviation $\sigma_i$
Total annual loss occurred to the i-th obligor $Y_i$

Correlation matrix $(\rho_{i,j})_{i=1}^n_{j=1}^{n} = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1n} \\ \vdots & \ddots & \vdots \\ \rho_{n1} & \cdots & \rho_{nn} \end{pmatrix}$

LGD stands for “loss given default”, which is the individual loss amount variable, given a default has occurred.
2.2 Building the individual loss model applied to credit insurance portfolio

2.2.1 Individual model definition:

An individual loss model is a model build and parametrized for each risk separately within a portfolio. We are therefore interested in the loss occurrence process relative to each and every risk, i.e. obligor. If \( C_i \) is the i-th obligor in the portfolio, the losses occurring for this obligor only during a given time horizon (let us fix it to one year for the purpose of this example) are modelled by the following random variables:

\[
Y_i = X_{i,1} + X_{i,2} + \ldots + X_{i,j} + \ldots + X_{i,N_i}
\]

The random variables \( X_{i,j} \) correspond to the j-th individual loss occurring on the i-th obligor during the previously determined time horizon. \( N_i \) is then the random variable which corresponds to the annual total number of losses occurred to the i-th obligor. And the \( Y_i \) variable correspond to the total annual loss occurred on the i-th obligor.

The most important distinctive feature of this model (as opposed to the “collective model” described later) is the fact that the \( Y_i \) loss burden can be modelled and computed as such, relative to one single separate risk: the i-th obligor. Therefore, if we needed to compute afterwards the **global portfolio loss burden**, then we will necessarily need to add together all the relevant \( Y_i \) random variables modeled separately in the following way:

\[
S = \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \sum_{j=1}^{N_i} X_{i,j}
\]

We will assume in our individual model that the maximum possible number of losses for the i-th obligor is one, therefore the \( N_i \) is a Bernoulli variable with a parameter equal to the probability of default (\( p_i \)). If \( N_i \) is Poisson distributed with a parameter \( \lambda_i \), when more than one single loss occurrence is possible during the fixed time horizon, the probability of default can be written as \( p_i = 1 - e^{-\lambda_i} \).
2.2.2 Individual model parameters:

In order to parametrize our individual loss model for the credit risk portfolio the two following elements should be fixed:

1. **Frequency and severity parameters for each i-th obligor.**

These parameters allow us to model the loss burden relative to that particular obligor. This aim can be achieved by using two following random variables:

- $N_i$ total number of losses occurring to the i-th obligor during the fixed time horizon. We have assumed at this stage that it is Bernoulli distributed with $(p_i)$ parameter.
- $X_{i,j}$ the individual amount of the j-th loss occurring to the i-th obligor. We can assume at this stage that independently from their order of occurrence within the loss occurring process, all the $X_{i,j}$ random variables are independent and identically distributed (iid) given a loss occurs. All these variables are then $X_i$ distributed, the latter being any form of continuous or mixed random variable (the particular analytical form for such a variable will be discussed in more detail later).

2. **Default events correlation structure.**

We will assume in this individual model that the only form of dependency within the portfolio is a frequency dependency. This means that all $N_i$ variables can be mutually dependent in some cases, but, on the other hand, the $X_i$ variables are always mutually independent. Such an assumption is quite often made for credit portfolio losses modelling, since the default events in real life situations are often inter-related because of the general economical context which is likely to influence and to trigger many different obligor’s defaults at the same time (a financial and economical crisis, for example). On the other hand, any close relationship between the LGD (Loss Given Default) factors, expressed as a percentage of the exposure for different obligors, is less obvious.

In our case all the default event occurrences within the portfolio during one year period are represented by an n-dimensional Bernoulli vector $(N_i)_{1 \leq i \leq n} = (N_1 \ldots N_n)$.

The random variable which corresponds to the total annual number of losses in the portfolio is then the sum of the elements of the $(N_i)_{1 \leq i \leq n}$ vector, which is:

$$N = \sum_{i=1}^{n} N_i$$

Let us assume here that the complete distribution of the n-dimensional Bernoulli vector $(N_i)_{1 \leq i \leq n}$ is not given in the model parameters, the only parameter available is the correlation matrix $(\rho_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$. This is very often the case in the credit
portfolio modelling. Therefore, there is an infinite number of possible \((N_i)_{1\leq i \leq n}\) vector distributions with given marginals \(N_i\) and with given correlation matrix \((\rho_{i,j})_{1\leq i < j \leq n}\). Then an additional assumption must be made regarding the complete dependency structure of the \((N_i)_{1\leq i \leq n}\) vector if we want to know its complete distribution and to be able to simulate it numerically. This assumption is discussed later, when the simulation scheme used will be described in more detail.

Let us introduce for each obligor \(C_i\) a default time \(\tau_i\) within the fixed one-year period. Then we can now write the probability of default before “\(t\)” as follows:

\[ p_i = u_i(t) = P[\tau_i < t] \]

If we consider again the two following default events: \(1_{\{\tau_i < t\}} = 1_i\) and \(1_{\{\tau_j < t\}} = 1_j\). Then the complete distribution of the bivariate Bernoulli \((N_i)_{1\leq i \leq 2}\) is totally known from the correlation matrix \((\rho_{i,j})_{1\leq i < j \leq 2}\) and the probabilities of default \(p_i\) and \(p_j\).

Indeed, we can easily calculate the following probabilities for the bivariate distribution of \((N_i)_{1\leq i \leq 2}\):

\[ P[\{\tau_i < t\} \cap \{\tau_j < t\}] = COV(1_i, 1_j) + E(1_i)E(1_j) = \rho_{i,j} \sqrt{p_i p_j (1 - p_i)(1 - p_j)} + p_i p_j \]

The above probability obviously corresponds to the \(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\) case of the above Bernoulli vector.

We can also calculate the \(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\) case which is \(P[\{1_i = 1\} \cap \{1_j = 0\}]\) in the following way:

According to the total probability formula we have

\[ P[\{1_i = 1\}] = P[\{1_i = 1\} \cap \{1_j = 1\}] + P[\{1_i = 1\} \cap \{1_j = 0\}] \]

Then \(P[\{1_i = 1\} \cap \{1_j = 0\}] = p_i - \rho_{i,j} \sqrt{p_i p_j (1 - p_i)(1 - p_j)} - p_i p_j \)

The \(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\) case and the \(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\) are eventually obtained by iterating the same total probability formula, using the previously calculated cases.

This bivariate example of the Bernoulli vector \((N_i)_{1\leq i \leq 2}\) is interesting since we can see that starting from 3 variables and more, the complete vector distribution cannot
be known from the individual model parameters, as they have been defined in our example. A further assumption is therefore needed in order to calculate such a distribution. This assumption is going to be done via the Gaussian copula method in our example, which will be described in more details below.

2.2.3 Individual model simulation

Once the parameters of the individual loss model has been defined according to the above description, the next step will be to numerically simulate the individual default events and also their particular loss amounts. What is the purpose of such a simulation?

In some cases an actuary could be interested in discovering a complete distribution of the total annual loss of the portfolio as a whole. This can be achieved using methods like FFT (see “Aggregation of Correlated Risk Portfolios : Models and Algorithms” by Shaun S Wang), Panjer’s recursive algorithm in some cases (see “Recursive Evaluation of a family of Compound Distributions” by Panjer HH), or also Hechman-Meyers method (see “Calculation of Aggregate Loss Distributions from claim severity and claim count distributions” by Hechman PE and Meyers GG), etc.

The calculation of the annual aggregate loss distribution is only one among other required results if the loss modelling is done in order to quantify a reinsurance contract. A reinsurance contract (an excess of loss treaty, for example) is likely to impact the insurance portfolio individual gross losses in a certain way, more or less complex, depending on the particular conditions and terms of the reinsurance contract itself. For example, an actuary could be also interested in the calculation of the probability distribution of such specific random variables as “annual loss ceded to the reinsurer”, “annual aggregate loss retention by the ceding company”, “annual reinstatement premium”, etc. Moreover, some of these distributions need to be aggregated with some others which correspond to different business classes and/or risk types. All this adds a lot of complexity to the distribution calculation in the end.

The above mentioned complexity often makes the methods like FFT and Panjer inappropriate and even sometimes impossible to apply. Therefore, using a Monte Carlo simulation scheme, which allows us to create a numerical sample of the individual model, is often the easiest and the most convenient distribution calculation procedure, since it is adaptable to virtually any situation. In what follows, we will therefore describe in more details the possible MonteCarlo simulation applied to an individual loss model. We will continue to explore our credit portfolio example. But before describing the simulation scheme itself for our individual loss credit model, two different subjects should be introduced and discussed in the first place : Merton structural credit risk model and using gaussian copulas in Monte Carlo simulation of correlated random variables.

1. Merton structural credit risk model and its application to our example.

We have included here a brief description of the Merton credit risk modelling approach, also called structural credit model. A method very close to Merton’s approach will be used here for the numerical simulation of the above individual credit loss model. Also, Merton model offers an explanation for the gaussian copula use in order to define a particular form of the \((N_i)_{i \in \mathbb{N}}\) vector distribution.
The structural approach to credit risk modelling was first initiated by Merton (1974). His method suggests that a default event occurs when the value of assets of a firm falls below a certain threshold (the level of the firm’s liabilities). This clearly means that we use the same approach as the one used in the option pricing model within Black and Scholes framework. This type of credit risk model is currently used by a number of credit insurance market practitioners, like Moody’s KMV model or JP Morgan’s Credit Grade model.

The basic assumption made is the following:

The value of the firm dynamics are modelled using geometrical brownian motion

$$dV_t = V_t(\mu dt + \sigma dW_t)$$

where

$V_t$ is the firm’s value at “t” moment

$\mu$ and $\sigma$ are the constant trend and diffusion parameters of the $V_t$ random process

and $W_t$ is the standard brownian motion process.

In our particular case there will be a simplification of the model, since we are not interested by modelling the default time $\tau_i$ explicitly (and therefore not interested by the $V_t$ dynamics as such), but merely by **modelling the fact that a default event has occurred** or has not occurred during the fixed time horizon, which was fixed at one year so far. So we are working with the Bernoulli vector $(N_i)_{i<j<n}$.

In the individual model parameters we have the correlation matrix $(\rho_{i,j})_{1\leq i \neq j \leq n}$

This is correlation between the default events, i.e. “frequency correlation”.

Let us now consider the occurrence of the two following default events, both modelled by Bernoulli variables:

$$1_{\{\tau_i < t\}} = 1_i \text{ and } 1_{\{\tau_j < t\}} = 1_j$$

Then we know that

$$E[1_{\{\tau_i < t\}}] = P[\tau_i < t] = u_i(t)$$

$$E[1_{\{\tau_j < t\}}] = P[\tau_j < t] = u_j(t)$$
and

\[ \text{VAR}[1_{(t_i \leq \tau_j)}] = u_i(t) - u_i^2(t) \]

\[ \text{VAR}[1_{(t_j \leq \tau_i)}] = u_j(t) - u_j^2(t) \]

Therefore, the linear correlation coefficient \( \rho_{i,j} \) can be defined as follows:

\[
\rho_{i,j} = \frac{\text{COV}(1_{(t_i \leq \tau_j)}, 1_{(t_j \leq \tau_i)})}{\sigma(1_{(t_i \leq \tau_j)}) \times \sigma(1_{(t_j \leq \tau_i)})} = \frac{P(\{t_i \leq \tau\} \cap \{t_j < t\}) - u_i(t) \times u_j(t)}{\sqrt{(u_i(t) - u_i^2(t)) \times (u_j(t) - u_j^2(t))}}
\]

with \( 1 \leq i \leq n, 1 \leq j \leq n \).

Let us assume at this stage that the default event \( 1_{(t \leq \tau)} \) is equivalent to an event of a certain random variable \( R_i \) being below a barrier \( b_i \) (so called “default barrier”):

\[ 1_{(t \leq \tau)} \Leftrightarrow 1_{(R_i < b_i)} \]

We can further assume that \( R_i \) is normally distributed with a probability density function:

\[
f_{R_i}(r) = \frac{1}{\sqrt{2\pi}} e^{-r^2/2}
\]

and the CDF:

\[
F_{R_i}(b_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b_i} e^{-x^2/2} dx = p_i
\]

The default barrier \( b_i = F_{R_i}^{-1}(p_i) = F_{R_i}^{-1}[u_i(t)] \) with \( b_i \in \mathbb{R} \)
is the inverse distribution function of \( R_i \) calculated at the probability of default.

Before going any further into the description of the Monte Carlo simulation scheme of the credit portfolio loss model (defined as an individual loss model), we need to define an important additional parameter: Merton adjusted correlation coefficient.

The basic idea will be to replace the dependency structure of the \( (N_i)_{i \leq 2} \) vector (which cannot be known from the model parameters) by a dependency structure derived from a gaussian vector (hence the gaussian copula). But we know that the
(\(N_i\))_{i \leq 2} vector is a discrete one, containing only “ones” and “zeroes”, whereas a gaussian vector is always a continuous one. The sense and the definition of the correlation coefficient \((\rho_{i,j})_{i \leq 2, j \leq 2}\), between \(1_{\{\tau_i < t\}} = 1_i\) and \(1_{\{\tau_j < t\}} = 1_j\) variables is certainly not the same as the correlation coefficient between the continuous random variables \(R_i\) and \(R_j\).

where \(1_{\{\tau_i < t\}} \Leftrightarrow 1_{\{R_i < b_i\}}\) and \(1_{\{\tau_j < t\}} \Leftrightarrow 1_{\{R_j < b_j\}}\)

And

\[
b_i = F_{R_i}^{-1}[p_i] = F_{R_i}^{-1}[u_i(t)], \quad b_j = F_{R_j}^{-1}[p_j] = F_{R_j}^{-1}[u_j(t)],
\]

\[
f_R (r) = f_{R_j} (r) = \frac{1}{\sqrt{2\pi}} e^{-r^2/2}
\]

Despite the fact that the same event (i.e. default event) can be modelled via two different approaches, the level of correlation between discrete variables \(1_{\{\tau_i < t\}} = 1_i\) and \(1_{\{\tau_j < t\}} = 1_j\) is totally different from the level of correlation required between normally distributed continuous variables \(R_i, R_j\). We need to keep in mind this very important issue while defining the correlation matrix in our simulation, since the Gaussian copula approach will use continuous vectors (gaussian vectors) generated by MonteCarlo in order to eventually simulate discrete events (like a default event).

It is interesting to note that a new “adjusted” correlation matrix for the gaussian vector can be derived from the parameters of the individual loss model that we have at our disposal.

Let us first recall the definition given to the original correlation coefficient between pairs of discrete Bernoulli variables:

\[
\rho_{i,j} = \text{corr}(1_{\{\tau_i < t\}}, 1_{\{\tau_j < t\}}) = \frac{\text{COV}(1_{\{\tau_i < t\}}, 1_{\{\tau_j < t\}})}{\sigma(1_{\{\tau_i < t\}}) \times \sigma(1_{\{\tau_j < t\}})} = \frac{P[\tau_j < t]\{\tau_j < t\} - u_i(t) \times u_j(t)}{\sqrt{(u_i(t) - u_i^2(t)) \times (u_j(t) - u_j^2(t))}}
\]

Note that in the bivariate discrete case the probabilities of the joint distribution \((\bar{N}_i))_{i \leq 2}\) can be extracted from the following bivariate continuous distribution:

\[
P[\{\tau_i < t\} \cap \{\tau_j < t\}] = F_R^{(2)}(b_i(t), b_j(t), \rho_M^M)
\]

Where \(F_R^{(2)}\) is the bivariate gaussian CDF (with standard gaussian marginals), \(b_i(t)\) and \(b_j(t)\) are the points at which the gaussian bivariate CDF is calculated,
\( \rho_{ij}^{M} \) is the “Merton adjusted” coefficient of correlation, used here as an additional parameter in order to define the precise form of \( F_R^{(2)} \).

According to Pugachevsky (Correlations in Multi Credit models, p6) we can define the Merton adjusted coefficient of correlation \( \rho_{ij}^{M} \) as follows:

\( \rho_{ij}^{M} \) is a parameter of a bivariate gaussian CDF \( F_R^{(2)} \) such as

\[
\rho_{i,j} = \frac{F_R^{(2)}[F_R^{-1}(u_i(t)), F_R^{-1}(u_j(t)), \rho_{i,j}^{M}] - u_i(t) \times u_j(t)}{\sqrt{(u_i(t) - u_i^2(t)) \times (u_j(t) - u_j^2(t))}}
\]

Where \( F_R^{-1} \) is the inverse CDF of a standard gaussian variable.

Let us now summarize the suggested method of calculation for \( \rho_{ij}^{M} \):

\[
F_R^{(2)}(b_i(t), b_j(t), \rho_{ij}^{M}) = F_R^{(2)}[F_R^{-1}(u_i(t)), F_R^{-1}(u_j(t)), \rho_{ij}^{M}] = P(\tau_i < t, \tau_j < t)
\]

from the above given definition of the default barriers.

But we also know from our discrete default event model that

\[
P(\tau_i < t, \tau_j < t) = COV(1_i, 1_j) + E(1_i)E(1_j) = \rho_{i,j}\sqrt{u_i(t)u_j(t)(1-u_i(t))(1-u_j(t))} + u_i(t)u_j(t)
\]

Then the following equation can be numerically solved for \( \rho_{ij}^{M} \) from the already known individual model parameters:

\[
F_R^{(2)}[F_R^{-1}(u_i(t)), F_R^{-1}(u_j(t)), \rho_{ij}^{M}] = \rho_{i,j}\sqrt{u_i(t)u_j(t)(1-u_i(t))(1-u_j(t))} + u_i(t)u_j(t)
\]

We also recall that for any gaussian vector of any dimension, its probability distribution is totally defined given the first two moments, i.e. the expected value vector and the variance-covariance matrix. Hence, the knowledge of the default probabilities \( u_1(t) \ldots u_n(t) \), associated to each and every obligor in the portfolio, and also of the variance-covariance matrix derived from \( (\rho_{ij}^{M})_{1 \leq i \leq n, 1 \leq j \leq n} \) is equivalent to the knowledge of the whole n-dimensional gaussian vector distribution \( (R_i)_{1 \leq i \leq n} \) (with standard gaussian marginals) that can be therefore simulated using MonteCarlo technique.
2. **A short introduction to the Copula method and its relevance in the individual loss model.**

The detailed understanding of the Merton adjusted coefficient of correlation (see previous section) is particularly important for the implementation of the gaussian copula simulation scheme, which is described here in more details.

We will start with a very short introduction to the copula approach, relative to the use of it that will be done in the context of the individual loss model MonteCarlo simulation only.

**Definition**

A copula (or n-copula) is a cumulated distribution function of a uniform n-dimensional random vector $U = (U_1, ..., U_n)$ with uniform marginals $U \sim \text{Uniform}(0,1)$:

$$C(u_1, ..., u_n) = P[U_1 \leq u_1, ..., U_n \leq u_n] \text{ with } (u_1, ..., u_n) \in (0,1)$$

The great interest of using a copula method within MonteCarlo simulation scheme is easy to understand from the statement of Sklar theorem (Sklar 1959):

**Theorem (Sklar)**

If $F$ is a n-variate CDF of the vector $X = (X_1, ..., X_n)$ with continuous marginal CDFs $F_1, ..., F_n$,
then a unique copula function $C$ always exists such as

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)) \text{, where } (x_1, ..., x_n) \in \text{Dom}(F_1 \times ... \times F_n)$$

From the above statement we can derive the following important conclusions:

- Since the n-variate CDF function of the vector $X = (X_1, ..., X_n)$ contains all the existing information about the distribution of $X$, then the copula function $C$ contains all the existing information about the dependency structure of $X$.

- The concept of copula isolates the dependency structure from the marginal behaviour within any random vector with continuous marginals. Therefore we can use different well known statistical continuous distributions (lognormal, Pareto, etc) as marginals and a copula to link them together in a certain way.

- The $C$ copula function links the marginal CDFs $F_1, ..., F_n$ to the global vector n-variate CDF $F(x_1, ..., x_n)$
The close link between Sklar theorem and Monte Carlo simulation of a random vector \( X = (X_1, \ldots, X_n) \) can be then understood in the following way:

We recall that the basic principle of Monte Carlo simulation of the occurrence of a one-dimensional random variable \( X \) consists in drawing a random uniform number \( u \) from \( U \sim \text{Uniform}(0,1) \) and then inverting \( u \) by application of the inverse CDF of \( X \):

\[
x = F_X^{-1}(u),
\]

since it can be easily shown that \( U = F_X(X) \) for any \( X \), where \( U \sim \text{Uniform}(0,1) \).

In a similar way, a Monte Carlo simulation of \( n \) variables \( (X_1, \ldots, X_n) \) will need to generate \( n \) uniform variables \( (U_1, \ldots, U_n) \). If the variables \( (X_1, \ldots, X_n) \) are mutually independent, then they can be simulated using independent sample of uniform random variables \( (U_1, \ldots, U_n) \).

If, on the other hand, \( (X_1, \ldots, X_n) \) variables are dependent in some way, then we need to simulate them using a set \( (U_1, \ldots, U_n) \) of dependent uniform variables.

In this particular context, Sklar theorem shows that for a set of given marginal distributions, the dependency structure of the arbitrary vector \( X = (X_1, \ldots, X_n) \) is totally determined by the dependency structure of a certain uniform vector \( U = (U_1, \ldots, U_n) \) with marginals \( U \sim \text{Uniform}(0,1) \).

This gives a key to a very efficient technique of numerical simulation of the sets of dependent random variables. The Bernoulli vector \( (N_i)_{1 \leq i \leq n} = (N_1 \ldots N_n) \) which is used to model default events in our individual loss model can also be simulated in this way. The only relevant question will be that of the choice of the particular copula form, since, at it was already underlined, the complete dependency structure of an \( n \)-dimensional Bernoulli vector is not determined from the individual model parameters (such as probabilities of default and the default event correlation matrix).

3. Using Gaussian Copula for Monte Carlo simulation of the individual loss model

Motivation

Simulation of correlated random variables is most straightforward if these variables have a multivariate normal distribution. However, this does not necessary happen in real life. In this section we will use a normal vector in order to build and mathematically define a normal copula that will be later used to simulate multivariate distributions with arbitrary marginals (i.e. Bernoulli marginals in our particular case). The advantages and the drawbacks of this simulation approach will be discussed below.
Let us assume that \( R = (R_1, \ldots, R_n) \) random vector has a multivariate normal distribution with standard normal marginals \( R_i \sim \text{Normal}(0,1) \) and a positive definite correlation matrix \( \Sigma = (\rho_{ij}^M)_{1 \leq i, j \leq n} \).

Then the random vector \( (R_1, \ldots, R_n) \) has the following joint probability density function:

\[
f(r_1, \ldots, r_n) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left(-\frac{1}{2} r^\top \Sigma^{-1} r\right), \text{ where } r = (r_1, \ldots, r_n)
\]

Let us note \( F^{(n)}_R \) the n-variate CDF of the normal random vector \( R = (R_1, \ldots, R_n) \) and the marginal CDFs:

\[
F_i(r) = \int_{-\infty}^{r} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \text{ for all } i \in (1, n)
\]

**Definition**

According to the above notation for \( F^{(n)}_R \) and for \( F_i = F_i^R \) for all \( i \in (1, n) \), a following multivariate uniform CDF is called the **normal (or gaussian) copula**:

\[
C(u_1, \ldots, u_n) = F^{(n)}_R \left( F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n) \right)
\]

It can be shown that for any set of given marginal arbitrary CDFs \( (F_1, \ldots, F_n) \):

the random variables \( X_1 = F_{X_1}^{-1}(F_1(R_1)), \ldots, X_n = F_{X_n}^{-1}(F_n(R_n)) \)

have a joined CDF \( F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F^{(n)}_R \left[ F_{X_1}^{-1}(F_{X_1}(x_1)), \ldots, F_{X_n}^{-1}(F_{X_n}(x_n)) \right] \)

Although the normal copula does not have a simple analytical expression, it allows a very straightforward simulation algorithm.
According to Wang (Aggregation of Correlated Risk Portfolios, section “Normal copula and Monte Carlo simulation”):

**Gross loss simulation algorithm**

**Step 1**: generate Merton adjusted correlation matrix from the default correlation matrix:

\[
\begin{pmatrix}
\rho_{11} & \ldots & \rho_{1n} \\
\vdots & \ddots & \vdots \\
\rho_{n1} & \ldots & \rho_{nn}
\end{pmatrix}_1 \rightarrow
\begin{pmatrix}
\rho_{11}^M & \ldots & \rho_{1n}^M \\
\vdots & \ddots & \vdots \\
\rho_{n1}^M & \ldots & \rho_{nn}^M
\end{pmatrix}_2
\]

This numerical calculation is performed only once in the process and doesn’t need to be repeated at each Monte Carlo iteration.

Besides, the initially obtained Merton adjusted correlation matrix is not necessarily definite positive, therefore an additional adjustment should be performed on that matrix in some cases.

(a method suggested by R. Rebonato and P. Jackel 1999 is described in more details in the Appendix).

**Step 2**: Choleski decomposition of the Merton adjusted correlation matrix

Find a lower triangular matrix \(A_{n \times n}\) such as:

\[
\begin{pmatrix}
\rho_{11}^M & \ldots & \rho_{1n}^M \\
\vdots & \ddots & \vdots \\
\rho_{n1}^M & \ldots & \rho_{nn}^M
\end{pmatrix}_2 = A \times A^t
\]

The details of the Choleski decomposition algorithm used are given in the Appendix. As for the step 1, this one is performed only once, at the beginning of the simulation process.

**Step 3**: generate a sample of a standard gaussian vector (with independent standard gaussian elements) by Monte Carlo algorithm

\[
Z = (Z_i)_{1 \leq i \leq n} = (Z_1 \ldots Z_n)
\]

**Step 4**: generate a sample of a gaussian vector with standard gaussian elements correlated according to Merton adjusted correlation matrix \((\rho_{i,j}^M)_{1 \leq i, j \leq n}\):

\[
(R_i)_{1 \leq i \leq n} = (R_1 \ldots R_n) = A \times Z
\]

**Step 5**: generate a default event indicator vector

\[
D = (D_i)_{1 \leq i \leq n} = (D_1 \ldots D_n)
\]

with \(D_i\) being equal to either “1” or “0”

We use the default barriers vector \((b_i)_{1 \leq i \leq n} = (b_1 \ldots b_n)\)
where \( b_i = F_{R_i}^{-1}(p_i) \) and \( F_{R_i}(b_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b_i} e^{-x^2/2} dx = p_i \)

These barriers are initially defined and calculated for each and every obligor in the portfolio from the probabilities of default.

At this step we will determine whether an \( i \)-th obligor has defaulted or not:

\[
D_i = \begin{cases} 
1 & \text{if } R_i < b_i \\
0 & \text{otherwise}
\end{cases}
\]

Please note that in Bernoulli case, the event indicator vector elements \( D_i \) will be the same as the elements of the total annual number of losses vector \( N_i \).

**Step 6**: generate a loss amount given default \( X_i \)

At this stage we need to determine an individual loss amount for all obligors for which \( D_i \) values are equal to 1.

We recall from the individual model parameters that each obligor has:
- Exposure (total value insured) \( m_i \)
- Expected LGD (loss given default) \( \mu_i \)
- Standard deviation of LGD \( \sigma_i \)

An additional assumption is made at this stage regarding the particular probability for the LGD distribution. Since the exposure \( m_i \) corresponds to the maximum possible loss amount, then the LGD can be modeled as “damage rate” or “loss rate” \( K_i \), i.e. the ratio between the actual loss amount and the maximum possible loss amount.

Let us suppose that the damage rate is Beta distributed:

\[
K_i = \frac{X_i}{m_i} \sim \text{Beta}(\alpha_i, \beta_i)
\]

where \( \alpha_i \) and \( \beta_i \) are the parameters of the Beta distribution with a domain between 0 and 1

with a probability density function:

\[
f_{K_i}(t) = t^{\alpha_i-1}(1-t)^{\beta_i-1} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}
\]

(where \( \Gamma \) stands for Gamma function)
The above mentioned parameters can be solved using the moments method as follows:

$$\alpha_i = \frac{\mu_i^2 (1 - \mu_i)}{\sigma_i^2} - \mu$$
and

$$\beta_i = (1 - \mu_i) \left( \frac{\mu_i (1 - \mu_i)}{\sigma_i^2} - 1 \right)$$

Then the individual loss amount calculation for the i-th obligor follows:

$$X_i = \begin{cases} 
K_i \times m_i & \text{if } D_i = 1 \\
0 & \text{otherwise}
\end{cases}$$

The total annual loss to the i-th obligor will then be equal to $X_i$, since the frequency variable used in our example is a Bernoulli one (therefore the maximum possible number of losses that can occur to an obligor is 1).

**Step 7**: calculate the total annual loss to the risk portfolio

$$S = \sum_{i=1}^{n} X_i$$

**Reinsurance loss and Net loss simulation algorithm (a very short introduction)**

The above algorithm describes the necessary steps in the gross loss simulation process. The term “gross loss” means portfolio loss before reinsurance. This is the first stage of the analysis. Once the risk portfolio gross losses have been simulated, we can proceed with the simulation of reinsurance losses and the also the net losses. The term “net loss” meaning the portfolio loss after the reinsurance contract conditions have been applied.

The “Excess of Loss” reinsurance contract works in the following way:

The reinsurer will intervene each time the individual loss amount will be higher than one given threshold called “priority” and up to a maximum called “limit”. In our particular case the reinsurance priority and limit will be applied for each and every individual loss occurring from a particular obligor. An annual limit could be also designed, further limiting the maximum amount of the reinsurers intervention per year (in case several “limits” are engaged during the same occurrence year).

If $M$ denotes the reinsurance contract priority (per individual loss)

and

$L$ denotes the reinsurance contract limit (per individual loss)
\( X_{i}^{\text{Reinsured}} \) denotes the reinsured loss from the i-th obligor

\( X_{i}^{\text{Net}} \) denotes the net loss from the i-th obligor

Then

\[
X_{i}^{\text{Reinsured}} = \min(\max(X_{i} - M, 0), L)
\]

\[
X_{i}^{\text{Net}} = X_{i} - X_{i}^{\text{Reinsured}}
\]

And the total annual net loss from the risk portfolio will correspond to

\[
S^{\text{Net}} = \sum_{i=1}^{n} X_{i}^{\text{Net}}
\]

In terms of the algorithm, the above calculations are performed each and every time the \( X_{i} \) variable is simulated during the Gross loss simulation. This approach offers a lot of flexibility for different possible reinsurance calculations (some of which could be potentially very complex), since the numerical sample of gross losses is already available.

### 2.3 Individual loss model applied to credit insurance portfolio: an example

#### 2.3.1 Risk portfolio description

The example that we are going to use here is a real life portfolio with individual credit exposures ranging from 220 KEUR to almost 1 billion EUR maximum exposure per obligor.

The total number of obligors included in this risk portfolio is 2000.

All the individual obligors belong to 4 different classes, each class corresponding to an economic activity sector.

**Probabilities of default**

The individual obligor’s probabilities of default are ranging from 0.03% to 20%, according to the credit quality notation system developed by the insurer. Each credit quality note corresponds to a fixed default probability. The time horizon for the definition of the default probability is one year. In the modelling process described here these probabilities of default are going to be considered as Bernoulli distribution parameters, i.e. we will assume that there is possibility of only a single default, if any, within the one year horizon.
Correlation structure

The correlation structure of the risk portfolio here is only a **frequency correlation**, which in this particular case means that the default events are correlated between them, but the loss amounts given default are independent. In order to define this correlation structure the insurer has defined a square “sector matrix” (size 4x4 in our case). The diagonal elements of this sector matrix correspond to “intra-sector default correlation”, i.e. the coefficient of correlation between default events inside one given economic sector. On the other hand, the non-diagonal elements of the sector matrix correspond to “inter-sector default correlation”, i.e. the coefficient of correlation between default events occurring within different economic sectors.

Since each and every obligor belongs to an economic sector, the above mentioned (4x4) correlation matrix can be expanded to (2000x2000) size in order to obtain the frequency correlation matrix between individual obligors. The correlation coefficients involved represent the correlation between individual default event occurrences, and therefore in our case the correlation matrix will represent the correlation structure of the Bernoulli vector that we will use for the risk portfolio frequency modelling. Like the default probabilities, the correlation coefficients are also defined by the insurer, based on their internal statistical study of defaults in different economic sectors.

**LGD (loss given default) distributions**

The LGD distributions are random variables between 0 and 1 representing the percentage which should be applied to the maximum exposure of an individual obligor in order to obtain the monetary loss amount. Such important elements as possible maximum loss, use factor, recovery rate, etc., well known to all credit insurance practitioners, they are all already taken into account in the parameterization of these LGD distributions. Also, in some particular cases, when all the terms of the individual policies should be taken into account (such as policy deductibles, annual aggregate limits ans deductible, etc.), the LGD distributions could be quite complex.

In our example we will use continuous LGD distributions (Beta form between 0 and 1), although in some real life situations they could be mixed variables (with one continuous part and one discrete part).

Four different types of LGD variables (one per economic sector) were defined by the insurer, the parameters being estimated according to their loss experience during the last 10 years.

**2.3.2 Risk portfolio simulation results**

The simulation results shown below were obtained using our Matlab platform and the individual model algorithm which was described in the previous section.

The total number of Monte Carlo iterations performed in this example is 10 000.
We can observe the high impact of the correlation coefficient on the total annual loss distribution of the risk portfolio:

![Portfolio total annual gross loss CDF](image)

The above graph shows 3 different sets of results:
- without correlation (all frequencies and severities are independent in the individual model);
- 1% correlation (all frequency Bernoulli variables are correlated with a coefficient of correlation equal to 1%, the LGD distributions remain independent);
- 3% correlation (all frequency Bernoulli variables are correlated with a coefficient of correlation equal to 3%, the LGD distributions remain independent);

As we can observe from the above results, the introduction of the frequency correlation in the risk model has important consequences on the total annual loss distribution. The tail tend to become heavier (increased volatility) and also the probability of scenarios with small annual total loss is increased. For example, the probability of zero annual loss (which correspond to a scenario when none of the obligors has defaulted) is very close to 0% within the model without correlation. This same probability jumps to almost 3% within the model including the frequency correlation coefficient = 3%.
Here is the summary of the key statistics for the 3 sets of simulation results:

<table>
<thead>
<tr>
<th>Frequency correlation</th>
<th>Mean annual loss</th>
<th>Std dev of annual loss</th>
<th>100Y annual loss</th>
<th>100Y annual frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>independent</td>
<td>22,762.248</td>
<td>17,450.533</td>
<td>74,119.576</td>
<td>37</td>
</tr>
<tr>
<td>1% correlation</td>
<td>22,612.582</td>
<td>26,294.081</td>
<td>117,448.747</td>
<td>104</td>
</tr>
<tr>
<td>3% correlation</td>
<td>22,301.755</td>
<td>33,586.166</td>
<td>187,577.365</td>
<td>182</td>
</tr>
</tbody>
</table>

We can observe that the mean annual loss stays approximately the same, the slight difference being exclusively due to simulation random error. This is normal since the frequency correlation assumption have no impact of the expected annual loss.

On the other hand, the annual loss volatility is sensitive to different frequency correlation assumptions. We can observe a significant increase in the annual loss standard deviation and in the 100 Year return period annual loss and annual frequency. These last two statistics give an indication regarding the tails of the annual loss and annual frequency distributions.

Since the term “return period” is quite often used in reinsurance specific jargon, we will give here some additional comments clarifying this concept.

The return period corresponding to an annual loss variable $S = s$ is defined as follows:

$$RP(s) = \frac{1}{P(S \geq s)}$$

Which is actually the inverse of the $S$ variable survivor function. Thus, in our case the 100 years return period annual loss will correspond to the 0.99-th percentile of the $S$ variable CDF.

Similarly, the return period corresponding to an annual frequency (number of losses) $N = n$ is defined as follows:

$$RP(n) = \frac{1}{P(N \geq n)}$$

As most of the volatility of the annual loss is driven by different frequency correlation assumptions, we can also observe a significant increase in 100 year return period frequencies. The increase in 100 year return frequencies is more significant than the increase in 100 year return annual losses. This is due the frequency-severity independence assumption made within this individual model.
Building the Frequency/Severity collective model

Using an illustrative example, we will show how credit and surety modelling technique could be adapted to a classical collective “frequency/severity” approach. We will first describe the method itself and then analyze the impact on the results and also the differences compared to the results from the individual model described previously.

In order to build a collective frequency/severity model for a credit portfolio we need to characterize the following random variables:

- Individual loss severity given default occurs in the portfolio \( X \): Severity model
- Total annual number of losses in the portfolio \( N \): Frequency model

3.1 Severity model

Individual loss severity distribution parameters calculation can be done from our initial individual model parameters. For that we need to mention that when a loss occurs somewhere in the portfolio it is a “random draw” from the list of the available exposures:

If \( I \) is an integer valued random variable with probabilities: 
\[
P(I = i) = \frac{p_i}{\sum_{j=1}^{n} p_j}
\]

Then \( I \) random variable indicates which exposure index \((i)\) is affected when a loss has occurred.

Let us now define the \( X \) random variable which in our model, corresponds to the individual loss severity given a loss occurred somewhere within the portfolio.

We will further assume that the \( X \) variable is the the following mixture of random variables:

\[
X = X_i \quad \text{with a probability} \quad P(I = i) \quad \text{, where} \quad \{I = i\} \quad \text{is a complete set of random events (1)}
\]

where \( X_i \) variables correspond to the individual loss severity given a loss has affected the “i-th” exposure within the portfolio.

This definition of the loss severity \( X \) at the portfolio scale as a whole has important implications.

Let us examine these implications and their consequences in the implementation of the collective loss model.
According to the above definition, the individual loss severity CDF given a default occurs can be written as follows:

\[
P\{X \leq x\} = \sum_{i=1}^{n} P\{X_i \leq x \cap I = i\} = \sum_{i=1}^{n} P\{I = i\} \times P\{X_i \leq x / I = i\} = \sum_{i=1}^{n} \left( \frac{p_i}{\sum_{i=1}^{n} p_i} \times P\{X_i \leq x\} \right)
\]

\[
P\{X \leq x\} = \frac{\sum_{i=1}^{n} (p_i \times P\{X_i \leq x\})}{\sum_{i=1}^{n} p_i} \tag{2}
\]

Please note that the probabilities of default here \((p_i)\) are also average annual numbers of losses (loss frequencies) because all defaults are modelled using Bernoulli variables. Besides, the above calculation doesn’t require independent default events, since we use the complete event set property (or total probability formula) which needs incompatible events that represent the total set of all possible events. Indeed, in our case:

For every \(i \neq j \in (1,n)\) the \(P(I = i \cap I = j) = 0\) and \(\sum_{i=1}^{n} P(I = i) = 1\)

Another very important implication here is the fact that in this model we assume that the \(X\) variable is independent from \(N\) variable. In other words, for any possible outcome of \(N\) (annual total number of losses in the portfolio), the conditional distribution of the loss severity \(X\), given a loss occurs, must stay exactly the same:

\[
P(X \leq x / N = n) = P(X \leq x)
\]

In order to understand this important qualification for the above formula (2), we can notice that it is built as a weighted average of the probabilities \(P\{X_i \leq x\}\), where the weights correspond to the probabilities of selecting an “i-th” line in the portfolio given a loss occurs : \(P\{I = i\}\).

Therefore if gives a marginal severity model for the loss severity at the portfolio scale, the term “marginal” meaning here one of the marginal distributions inside the bivariate random vector \((N,X)\).

In order to illustrate the possible cases when this marginal severity model will not be correct for any possible loss severity at the portfolio scale, i.e. the conditional probability \(P(X \leq x / N = n)\) will be different from the marginal distribution \(P(X \leq x)\), let us examine the following two theoretical examples. These examples were simplified in order to ease all the calculations and also for the intuitive understanding of the problem.
3.1.1 Example 1: non-homogeneous portfolio with independent exposures

Our portfolio is composed of 2 very different exposures:
Exposure1 = 100M and Exposure2 = 10M.

In order to simplify further our example, let us fix the same LGD distribution, which will be a constant 100% (in our case $\mu = 100\%$ and $\sigma = 0\%$).
Therefore $X_1 = 100m$ and $X_2 = 10m$

The probabilities of default are different:
$p_1 = 80\%$ and $p_2 = 50\%$

The frequencies of loss events within the risk portfolio are independent: $N_1 \perp N_2$.
$N_1 \sim \text{Binomial}(p_1 ,1)$
$N_2 \sim \text{Binomial}(p_2 ,1)$

We can say that our portfolio is obviously not “homogeneous” since the two exposures are very different, and therefore the two LGD distributions are also very different.

We define the portfolio global frequency distribution as:

$$N = N_1 + N_2$$

and the portfolio global severity distribution as:

$$X = \frac{p_1 X_1 + p_2 X_2}{p_1 + p_2}$$

In order to illustrate the fact that $N$ variable is not independent from $X$ variable we will calculate the conditional expected severity in the individual loss model for every possible scenario of annual number of losses, given a loss occurs: $E(X / N > 0)$.

Scenario 1: $N = 1$

$$E(X / N = 1) = \frac{E(S / N = 1)}{E(N / N = 1)} = E(S / \{N_1 = 1 \cap N_2 = 0\}) + E(S / \{N_1 = 0 \cap N_2 = 1\})$$

$$= E(X_1) \times P(\{N_1 = 1 \cap N_2 = 0\} / \{N = 1\}) + E(X_2) \times P(\{N_1 = 0 \cap N_2 = 1\} / \{N = 1\})$$
The probability $P(N = 1)$ can be calculated in the following way:

\[ P(N = 1) = 1 - P(N = 0) - P(N = 2) = 1 - \left((1 - p_1)(1 - p_2)\right) - p_1p_2 = 0.5 \]

Then

\[ E(X / N = 1) = 100m \times \left((0.8(1 - 0.5) / 0.5) + 10m \times \left((0.5(1 - 0.8) / 0.5\right) = 82m \]

**Scenario 2:** $N = 2$

\[ E(X / N = 2) = \frac{E(S / N = 2)}{E(N / N = 2)} = \frac{1}{2} E(S / \{N_1 = 1 \land N_2 = 1\}) = \frac{1}{2} \times (E(X_1) + E(X_2)) \]

\[ E(X / N = 2) = 0.5 \times (100m + 10m) = 55m \]

The significant difference between the expected loss severity in the first scenario $E(X / N = 1) = 82m$ and in the second scenario $E(X / N = 2) = 55m$ is due to the fact that the exposures are different (hence the non-homogeneous portfolio) and also the probabilities of default are different.

According to the above given definition (1) for the $X$ variable in the collective loss model, the expected loss severity given a default occurs is calculated in the following way:

\[
E(X) = \sum_{i=1}^{n} p_i E(X_i)
\]

\[
= \frac{\sum_{i=1}^{n} p_i E(X_i)}{\sum_{i=1}^{n} p_i}
\]

In our example: $E(X) \approx 65.38m$

Therefore, if the individual loss model is replaced by the collective loss model which assumes independent $X$ and $N$ variables, then the result is likely to be different since within the individual model $X$ and $N$ variables are not independent (as it was shown above).

Obviously, the above $E(X) \approx 65.38m$ calculation is "averaging out" different $E(X / N = n)$ scenarios, being the marginal distribution of the bivariate vector $(X, N)$.

This way of reasoning can be further expanded for a risk portfolio with an arbitrary $n$ exposures. However, the calculation quickly become tedious as the total number of exposures increases, since all the possible combinations should be considered separately. But, in theory, this will allow us to build a regression curve of the $X$
variable on the $N$ variable, by mapping the $E(X / N = n)$ as a function of $(N = n)$.

We can also notice that the correlation structure in the default events of this portfolio doesn’t matter for the above problem: the mentioned exposures can be either correlated or independent, the severity $X$ distribution will not have an identical form given any possible $N = n$ scenario.

Some important comments should be made at this stage.

First of all, once we know that the collective model will necessarily be an approximation of the individual loss model, it is interesting to assess the degree of this approximation. The random variable that we are ultimately interested in is the total annual loss $S$.

We will now calculate the distribution of $S$ according to the individual loss model (denoted as $S_{\text{ind}}$) and according to the collective model approximation (denoted as $S_{\text{coll}}$).

**Individual model calculation**

$$S_{\text{ind}} = S_1 + S_2, \text{ with } S_1 \perp S_2$$

We will use the following convolution formula for the sum of two independent discrete random variables:

$$P(S = s_k) = \sum_j \left( P(S_1 = s_{i_j}) \times P(S_2 = s_k - s_{i_j}) \right)$$

The distribution of $S_1$:

<table>
<thead>
<tr>
<th>0</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>80%</td>
</tr>
</tbody>
</table>

The distribution of $S_2$:

<table>
<thead>
<tr>
<th>0</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>50%</td>
</tr>
</tbody>
</table>

Then, the resulting distribution of $S_{\text{ind}}$ can be calculated as follows:

<table>
<thead>
<tr>
<th>0</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10%</td>
</tr>
<tr>
<td>100</td>
<td>40%</td>
</tr>
<tr>
<td>110</td>
<td>40%</td>
</tr>
</tbody>
</table>
And the resulting expected annual loss:

\[
E(S_{\text{ind}}) = \sum_j (s_j P(S_{\text{ind}} = s_j)) = 85m
\]

**Collective model calculation**

Let us start by calculating the expected annual loss according to the collective loss model assumption:

\[
E(S_{\text{coll}}) = E(N) \times E(X) = 85m
\]

Therefore \(E(S_{\text{coll}}) = E(S_{\text{ind}})\) in this particular case.

But even if the expected annual loss calculation according to the collective loss model assumptions is correct, the distribution of \(S\) will be different. In order to calculate this distribution we will need to calculate the compound distribution as follows:

\[
P(S_{\text{coll}} = s) = \sum_{i=0}^2 \left( P(N = i) \times P(S_{\text{coll}} = s / N = i) \right) = \sum_{i=0}^2 \left( P(N = i) \times P(X^{*i} = s) \right),
\]

where \(X^{*i}\) corresponds to the \(i\)-fold convolution of \(X\) variable.

In general case:

\[
F_{S_{\text{coll}}}(x) = \sum_{i=0}^\infty \left( P(N = i) \times F_{X^{*i}}(x) \right), \quad \text{where } F_{X^{*i}}(x) \text{ is the distribution function of } X \text{ convolved "}i\text{" times with itself.}
\]

The distribution of \(N\):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10%</td>
</tr>
<tr>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td>2</td>
<td>40%</td>
</tr>
</tbody>
</table>

The distribution of \(X\):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8/13</td>
</tr>
<tr>
<td>100</td>
<td>5/13</td>
</tr>
</tbody>
</table>
The distribution of $S_{coll}$ according to each $N = n$ scenario:

<table>
<thead>
<tr>
<th>$s$</th>
<th>$N=0$</th>
<th>$N=1$</th>
<th>$N=2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>10</td>
<td>0%</td>
<td>5/13</td>
<td>0%</td>
</tr>
<tr>
<td>20</td>
<td>0%</td>
<td>0%</td>
<td>4/27</td>
</tr>
<tr>
<td>100</td>
<td>0%</td>
<td>8/13</td>
<td>0%</td>
</tr>
<tr>
<td>110</td>
<td>0%</td>
<td>0%</td>
<td>9/19</td>
</tr>
<tr>
<td>200</td>
<td>0%</td>
<td>0%</td>
<td>25/66</td>
</tr>
</tbody>
</table>

And the resulting final distribution of $S_{coll}$ is the following:

<table>
<thead>
<tr>
<th>$s$</th>
<th>$P(S=s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10%</td>
</tr>
<tr>
<td>10</td>
<td>5/26</td>
</tr>
<tr>
<td>20</td>
<td>1/17</td>
</tr>
<tr>
<td>100</td>
<td>4/13</td>
</tr>
<tr>
<td>110</td>
<td>7/37</td>
</tr>
<tr>
<td>200</td>
<td>5/33</td>
</tr>
</tbody>
</table>

Let us now compare the above calculated $S_{coll}$ and $S_{ind}$ distributions:

![Total annual loss CDF](image)

We can see that for the range $[0,110m]$ $F_{coll} \geq F_{ind}$, and for the range $[110m,200m]$ $F_{coll} \leq F_{ind}$.

Also, we have seen that $E(S_{coll}) = E(S_{ind})$

Then according to the definition of the stochastic dominance of the 2nd order, we have:

$S_{coll} \leq_{2} S_{ind}$
The above notation stands for “the random variable $S_{ind}$ stochastically dominates the variable $S_{coll}$ (at the 2nd order).

This is a very interesting observation, since this conclusion about the stochastic dominance of the individual loss model shows that the collective loss model is more conservative for risk averse insurers. And in this case we didn’t use the Poisson simplification, we directly used the true discrete distribution for $N$, as it results from the individual model parameters.

Here is a very short remainder regarding the concept of stochastic dominance in risk theory:

The stochastic dominance is an instrument of comparison of random variable distributions in terms of risk preference. In other words, stochastic dominance helps to answer the following question: if $X$ represents one risk and $Y$ represents another risk, which one of the two considered risk is preferable for a risk underwriter?

In Deestra, Plantin (Théorie du risque et Réassurance, p15):

→ Definition (Stochastic dominance at the 1st order)

Let $X$ and $Y$ be two arbitrary random variables and $F_X(t), F_Y(t)$ their respective CDFs

If $F_X(t) \geq F_Y(t)$ for all possible values of $t$

Then the $X$ variable is said to stochastically dominate the $Y$ variable at the 1st order:

$X \succeq_1 Y$

This definition is very intuitive since if the probabilities of loss exceedance are higher for $Y$ than for $X$, for any chosen loss threshold, then it is logical to conclude that $X$ represents less risk and therefore preferable to a risk undertaker.

Now, in order to understand the meaning of the stochastic dominance at the 2nd order, we can say that most of the risk undertakers and risk managers accept to diminish their earnings slightly in exchange of the protection they can get against the adverse scenarios in the tail of their total loss distributions. This concept is the key of all insurance and reinsurance economy, since the distribution tails are very often the main drivers of the insurer’s and reinsurer’s results volatility.

→ Definition (Stochastic dominance at the 2nd order)

Let $X$ and $Y$ be two arbitrary random variables and $F_X(t), F_Y(t)$ their respective CDFs
If \( E(X) \leq E(Y) \) and

If there is a number \( c \geq 0 \) such as \( F_X(t) \leq F_Y(t) \) for all \( t \in (0,c) \) and

\[
F_X(t) \geq F_Y(t) \quad \text{for all } t \in [c,+\infty)
\]

Then the \( X \) variable is said to stochastically dominate the \( Y \) variable at the 2nd order:

\[
X \succeq_2 Y
\]

Quite often in real life examples, the annual loss distributions after reinsurance stochastically dominate the same distributions before reinsurance. This is the basic result of the distribution’s tail reduction via reinsurance protection.

If we now get back to our 1st example using two different exposures (10m and 100m), each having a different probabilities of default. We have seen that within the individual model framework the \( X \) variable is not independent from the \( N \) variable. This relation between the \( X \) and the \( N \) variables obviously appears because of the fact that our initial assumption implies different probabilities of default for our exposures. It is interesting to observe that if these default probabilities were the same then the \( X \) and the \( N \) variable would be independent.

Let us take the same assumption for the LGD distributions:

\[
X_1 = 100m \quad \text{and} \quad X_2 = 10m
\]

But change the assumption regarding the probabilities of default:

\[
N_1 \sim \text{Binomial}(80\%,1) \\
N_2 \sim \text{Binomial}(80\%,1)
\]

Then we can easily realize that the conditional distribution of \( X / N = n \) is the same as the marginal distribution of \( X \) :

<table>
<thead>
<tr>
<th>( n )</th>
<th>10</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>50%</td>
<td></td>
</tr>
</tbody>
</table>

Therefore we can say that the \( X \perp Y \) condition can be true, even if the loss portfolio is not homogeneous (i.e. the portfolio exposures are different in size).

But the \( X \perp Y \) condition doesn’t guarantee that \( S_{coll} \) will have the same distribution as \( S_{ind} \) .
As a matter of fact, we can calculate these distributions, using the above mentioned convolution formulas for the collective and for the individual models. The resulting distributions will be different and the stochastic dominance of the individual model can be also shown:

This discrepancy between $S_{coll}$ and $S_{ind}$ distributions can be intuitively understood in the following way. Within the collective loss model each exposure can be affected by a loss more than once, whereas the individual loss model admits only a single loss scenario, if any. This also explains the stochastic dominance of the individual loss model.

### 3.1.2 Example 2: non-homogeneous portfolio with correlated exposures

Our portfolio is composed of 2 very different classes of exposures:
- 100 exposures of class 1 = 100M and
- 100 exposures of class 2 = 10M.

In order to simplify further our example, let us fix the same LGD distribution for all exposures, which will be a constant 100% (in our case $\mu = 100\%$ and $\sigma = 0\%$).

The probabilities of default are all the same: $p_i = 1\%$ for $i \in (1,200)$, which in this case means that the contribution of each class to the expected loss severity is the same.

The frequencies of loss events are correlated for class 1 exposures (100m) and independent for class 2 exposures (10m). The coefficient of correlation for all class 1 frequencies is 5%.
As it was the case in the Example 1, our portfolio is not “homogeneous” since the two exposure classes are very different in size, and therefore the two classes LGD distributions are also very different.

In order to illustrate the fact that $N$ variable is not independent from $X$ variable (defined according to (1)) we will have simulated 10000 iterations of this portfolio and mapped the conditional expected severity for every group of iterations with a given annual number of losses $E(X / N = n)$ in the individual loss model:

Each blue point on the graph corresponds to the simulated $E(X / N = n)$ with $n \in (1,80)$.

As we can observe from the above graph, the scenarios corresponding to a high annual number of losses are more likely to trigger large exposures rather than the small ones because of the correlation involved. Indeed, the relatively high numbers of annual losses are more often due to the correlation between the defaults of large exposures. Therefore, the $X$ distribution is likely to be significantly different for scenarios with rather small total annual number of defaults (losses) from the scenarios with a high total annual number of defaults (losses).
If now we remove the correlation between class 2 large exposures, and re-simulate the same portfolio losses again, then the whole picture will change dramatically:

The regression curve is now totally flat, which shows that in this particular case the marginal X distribution has the same expected value as the simulated ones, for all possible frequency scenarios. The two “outsider” points on the graph are exclusively due to the random error from MonteCarlo simulation process.

Let us now consider the distributions of the total annual loss $S_{coll}$ and $S_{ind}$, according to the collective loss model and individual loss model simulations. In order to perform such a simulation within the collective model framework, we need to know the total annual number of losses distribution. However, the exact parameters cannot be known from the individual model parameters, as it was defined above. Indeed, in order to know the distribution of $N_{coll} = N_1 + N_2 + ... + N_n$ we need to have the knowledge of the distribution of the Bernoulli frequency vector $N = \{N_1, N_2, ..., N_n\}$, which is impossible to derive just from the knowledge of the default probabilities $p_i$ and the correlation matrix $\rho_{ij}$. An additional assumption should be made, this is the subject of the next section.

3.2 Frequency model

The complete characterization of the total annual number of losses distribution is not possible, given the parameters of the individual loss model as described above, since all we know so far is that it is a discrete random variable which corresponds to the sum of correlated Bernoulli variables, each one with different parameters:

$$N = \sum_{i=1}^{n} N_i \quad (3)$$
Given the information about the marginals $N_i$ in the Bernoulli vector and also the correlation matrix $(\rho_{i,j})_{i \leq i, j \leq n}$ it is still impossible to fully characterize such a distribution.

An additional assumption regarding the dependency structure of the Bernoulli vector must be made. The one that our individual model makes is an assumption of the Gaussian vector dependency structure (i.e. normal copula).

Without showing any statistical argument at this stage regarding the adequacy of any particular type of copula in this case, the collective model described in this section implies the use of a different dependency structure than the Gaussian one.

Here is how we suggest to characterize directly $N$ as a single random variable:

**Step 1:** we calculate the average annual number of losses in our portfolio:

$$E[N] = \sum_{i=1}^{n} E[N_i] = \sum_{i=1}^{n} p_i$$  \hspace{1cm} (4)

**Step 2:** we calculate the variance of the annual number of losses in the portfolio:

$$Var[N] = Var[\sum_{i=1}^{n} N_i] = \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(N_i, N_j)$$  \hspace{1cm} (5)

where $N_i$ are the Bernoulli random variables corresponding to the annual number of losses affecting the $i$-th obligor and $Cov(N_i, N_j)$ are the elements of the variance-covariance matrix that can be easily calculated from the correlation matrix $(\rho_{i,j})_{i \leq i, j \leq n}$:

$$Cov(N_i, N_j) = (\rho_{i,j}) \times \sqrt{p_i p_j (1-p_i) (1-p_j)}$$  \hspace{1cm} (6)

**Step 3:** we postulate a Negative Binomial form for $N$

$$N \sim NegBin(E[N], Var[N])$$

At this step our copula assumption is made. Assuming $N$ follows a Negative Binomial distribution implies a copula assumption different from the Gaussian one. The argument for using this particular probability form for $N$ is not a statistical one at this stage. Negative Binomial distribution is one of the most frequently used analytical distributions, and the one that allows a maximum flexibility for any combination of the $1^{st}$ and the $2^{nd}$ moments. Later we will test this assumption against the normal copula used in the individual loss model. The Negative Binomial model is also described in Wang (Aggregation of Correlated risk portfolios, section 3.4.3 “Negative Binomial model”).
3.3 Homogeneous risk portfolio: a definition

If using the above notation we have a frequency/severity model defined for the risk portfolio as a whole with the two key random variables defined as follows:

\[ N = \sum_{i=1}^{n} N_i \] corresponding to the total annual number of losses

and

\[ X \] corresponding to the individual cost of each and every loss, given a loss occurs

Where \( X = X_i \) with a probability \( P(I = i) \) and \( \{ I = i \} \) is a complete set of random events.

Let us now give a definition for a “homogeneous risk portfolio” which will be useful in our particular context:

**Definition**

A risk portfolio is called homogeneous when

- \( X_i \) variables are all equidistributed conditionally to \( N \) variable
- \( X_i \) variables are all independent

From the above theoretical examples we can understand that the key issue regarding the validity of the formula (2) is the “homogeneity” of the risk portfolio. As a matter of fact, if the risk portfolio is perfectly homogeneous, i.e. composed of the exposures all having the same \( X_i \) distributions independent from each other, then the formula (2) holds. We will come back later to this concept of the risk portfolio “homogeneity” and figure out the possible solutions to the cases when this important characteristic is not satisfied.

On the other hand, any correlation between the default events (frequency correlation) will only impact the frequency distribution “\( N \)” and will have no effect on the validity of the severity formula (2) if all the \( X_i \) variables have the same distribution and are all independent.

3.4 Collective loss model applied to credit insurance portfolio: a real life example

Let us take the same real life portfolio example as the one described in the section 2.3.1 above.

The severity discretized CDF is calculated according to the formula (2). This is quite difficult to do manually in Excel, especially if the number of points in the discretized CDF is important. But it can be easily handled via a VBA code in Excel using the parameters of the individual loss model. We just need to determine a sufficiently fine step in the CDF discretization, since afterwards this discretized severity CDF will be used in a MonteCarlo simulation program and the list of discrete points will be interpolated with a
spline function. Thus, we don’t need to postulate any particular form for this severity
distribution, it will be defined as a discretized continuous variable CDF:

\[ x \rightarrow P(X \leq x) \]
calculated for a finite number of selected values of \( x \).

The frequency Negative Binomial distribution parameters are calculated according to the
above defined formulas (4), (5) and (6). This calculation can be tedious if done manually in
Excel, therefore we have designed a VBA code for Excel, performing the total variance
calculation for the Bernouilli vector automatically.

We have first of all tested the results in the case of independent frequencies:

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>18 188 626</td>
<td>18 695 459</td>
<td>-506 832</td>
<td>inf</td>
</tr>
<tr>
<td>60%</td>
<td>21 082 782</td>
<td>21 600 970</td>
<td>-518 188</td>
<td>inf</td>
</tr>
<tr>
<td>70%</td>
<td>24 773 360</td>
<td>25 212 521</td>
<td>-439 161</td>
<td>inf</td>
</tr>
<tr>
<td>80%</td>
<td>29 994 799</td>
<td>30 254 496</td>
<td>-259 697</td>
<td>inf</td>
</tr>
<tr>
<td>90%</td>
<td>39 583 371</td>
<td>39 640 277</td>
<td>-56 906</td>
<td>inf</td>
</tr>
<tr>
<td>95%</td>
<td>49 767 874</td>
<td>49 532 420</td>
<td>235 454</td>
<td>sup</td>
</tr>
<tr>
<td>99%</td>
<td>75 243 443</td>
<td>74 119 576</td>
<td>1 123 867</td>
<td>sup</td>
</tr>
<tr>
<td>99.50%</td>
<td>90 544 056</td>
<td>90 146 243</td>
<td>397 813</td>
<td>sup</td>
</tr>
</tbody>
</table>
We can observe that the two CDFs are very close to each other and we have a good approximation of the distribution tail. The simulated results exhibit the 2\textsuperscript{nd} order stochastic dominance for the individual model but the simulated difference in percentiles is very small.

The following step was to test the results while taking into account of all the frequency correlations according to the parameters of the individual loss model:

![Total annual loss CDF](image)

We can observe an increased discrepancy between the two CDFs, compared to the previous simulation results without correlation between the defaults.

Zooming into the above graph will show multiple crossing points between the CDF simulated with the individual model and the CDF simulated with the collective one, which means that the simulation results as such don’t let us to make any type of comparison between the two CDFs.

Besides, the tail simulated under the collective risk model is underestimated in comparison to the tail simulated under the individual loss model in this case. This means that if we choose the collective loss model in this particular case, then our results are likely to be less conservative from the point of view of “risk adverse” risk underwriter:

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>7 965 284</td>
<td>9 031 261</td>
<td>-</td>
<td>1 065 977</td>
</tr>
<tr>
<td>60%</td>
<td>13 624 407</td>
<td>13 640 396</td>
<td>-</td>
<td>15 989</td>
</tr>
<tr>
<td>70%</td>
<td>22 732 747</td>
<td>20 934 242</td>
<td>1 798 505</td>
<td>sup</td>
</tr>
<tr>
<td>80%</td>
<td>37 388 902</td>
<td>32 945 278</td>
<td>4 443 624</td>
<td>sup</td>
</tr>
<tr>
<td>90%</td>
<td>64 272 440</td>
<td>56 348 981</td>
<td>7 923 459</td>
<td>sup</td>
</tr>
<tr>
<td>95%</td>
<td>92 837 441</td>
<td>86 416 166</td>
<td>6 421 275</td>
<td>sup</td>
</tr>
<tr>
<td>99%</td>
<td>169 127 562</td>
<td>187 577 365</td>
<td>18 449 803</td>
<td>inf</td>
</tr>
<tr>
<td>99.50%</td>
<td>201 912 543</td>
<td>245 960 235</td>
<td>-</td>
<td>44 047 692</td>
</tr>
</tbody>
</table>
3.5 Collective loss model vs Individual loss model: a general case without portfolio’s frequency approximation

In order to better understand the reasons of this difference in the tails of the $S_{\text{coll}}$ and $S_{\text{ind}}$ distributions in the above example, let us first of all explore the stochastic dominance properties of the individual loss model in general case, when the portfolio’s total annual number of losses distribution is exactly reproduced in the collective loss model. Very often in actuarial literature we can see a detailed study of the “classical” Poisson approximation, where basically each Bernoulli frequency distribution $N_i$ is replaced by a Poisson distribution with the same expected value (in Kaas, Von Heerwaarden ans Goovaerts “Between individual and collective model for total claims”). This approximation will always lead to stochastic dominance of the individual loss model at the 2nd order (if, of course all the Bernoulli frequency distributions in the individual loss model are independent). Here we will explore the general case of the collective model where there is no approximation at all in terms of the total annual frequency distribution, i.e. the frequency variable used in the collective model $N = \sum_i N_i$ will exactly represent the sum of Bernoulli variables of different parameters (correlated or independent).

The important initial condition here is the fact that in the collective model there will be no approximation as far as the portfolio’s frequency is concerned:

$$N_{\text{coll}} = N_{\text{ind}} = \sum_i N_i$$

Of course, in real life situation this is unlikely to occur, since in order to reproduce exactly such a distribution in the collective model we need some very tedious calculations within the independence case, and we encounter a sheer impossibility within the correlated case. Therefore, this chapter has a very limited practical utility, rather it helps in the understanding of what we have observed in our real life example.

Let us define the form of the individual frequency distribution $N_i$, conditional to the total annual frequency distribution $N$, within the collective and the individual loss models respectively:

**Collective loss model**

For each and every scenario $N = n$ the annual number of losses for the i-th exposure in the portfolio will follow the Binomial distribution with the probability parameter corresponding to the “chances” of a choice of the i-th exposure given a loss occurs somewhere in the portfolio.

Then, using the total probability formula we can write the $N_i$ variable CDF as a linear combination of the conditional Binomial distribution functions in the following way:

$$P(N_i \leq n_i) = \sum_{j=1}^k \left[ P(N_i^{(N_{\text{coll}} = j)} \leq n_i) \times P(N_{\text{coll}} = j) \right]$$
Individual loss model

Using the same reasoning we can say that for each and every scenario \( N = n \) the annual number of losses for the \( i \)-th exposure in the portfolio will follow the Hypergeometric distribution with the probability parameter corresponding to the “chances” of a choice of the \( i \)-th exposure given a loss occurs somewhere in the portfolio. Compared to the collective model, the \( i \)-th exposure can be affected only once per year by a loss, which corresponds to a sampling scheme without replacement given a fixed \( N = n \) scenario. Therefore, we can write the \( i \) variable CDF as a linear combination of the conditional Hypergeometric distribution functions in the following way:

\[
P(N_i \leq n_i) = \sum_{j=1}^{k} \left[ P\left( \text{Binomial} \left( \frac{p_i}{\sum_{i=1}^n p_i}, j \right) \leq n_i \right) \times P(N_{\text{coll}} = j) \right]
\]

\[
P(N_i \leq n_i) = \sum_{j=1}^{k} \left[ P\left( \text{Hypergeometric} \left( \frac{p_i}{\sum_{i=1}^n p_i}, n, j \right) \leq n_i \right) \times P(N_{\text{ind}} = j) \right]
\]

If \( N_{\text{coll}} = N_{\text{ind}} = \sum_i N_i \) (which is the specified initial qualification),

Then we can see that the hypergeometric sampling scheme (associated with the individual loss model in our example) is going to be less conservative than the binomial sampling scheme since any given exposure can only be affected only once.
4 Improving the quality of the collective risk model approximation

As it was already underlined and shown with a real life example in the case of the credit insurance portfolio, the collective risk model represents an approximation of the individual risk model. The reasons of this approximation are not only related to the portfolio non-homogeneity but also to the approximation made of the portfolio’s frequency distribution:

\[ N_{\text{coll}} \neq N_{\text{ind}} = \sum_i N_i \]

This approximation is a significant source of discrepancy in the resulting total annual loss distributions, especially in case of the frequency correlation:

\[ S_{\text{coll}} \neq S_{\text{ind}} \]

The natural question at this stage is the following: if we use the collective risk model approximation, then are there any ways in which this approximation could be improved?

Also, the interesting point is the approximation quality key drivers: what are the parameters in the individual loss model to which the collective model approximation will be the most sensitive?

In order to answer the above questions, three different approaches will be discussed in the following chapters:

- Partial collective model approximation
- Portfolio stratification
- Alternative frequency distribution hypothesis

4.1 Partial collective model approximation

The idea of the partial collective model approximation is based on a the use of the collective model only for one part of the risk portfolio. The typical choice would be to model a restricted number large exposures individually and the rest of the exposures (so called “attritional exposures”) using the collective risk model. All exposures (or risks) which are below a fixed threshold are considered as “attritional exposures”, and all the exposures which are above this same threshold are considered as “large exposures”.

Since the bulk of the exposures in terms of the number of items is represented by the attritional exposures, the collective model will still play its role in the reduction of the number of simulated variables and the simplification of the correlation issues. Also, the large exposures which are often the major drivers of the annual total loss volatility (and the form of the distribution tails in particular) will still be modeled using the individual model approach. The details of the implementation of this approach, along with the test results from the real life example portfolio will be discussed below.
4.2 Portfolio stratification

To some extend, the portfolio stratification approach corresponds to the previously discussed partial collective model which is further expanded. Instead of separating the portfolio into two segments: “large exposures” and “attritional exposures”, the portfolio could be separated further into a larger number of segments (or strata), using different stratification criteria. All of these segments will be modelled with a separate collective loss model applied for each one and ultimately aggregated together.

The different criteria used for the portfolio stratification could be:

- size of risks (insured valued, possible maximum loss scenarios, etc.),
- level of correlation between risks within the chosen portfolio segment,
- probability of default or any other alternative risk quality measure.

The methods suggested for the aggregation of the attritional and the large exposures are further expanded here in order to aggregate an arbitrary number of portfolio segments. The details of the implementation of this approach, along with the test results from an example portfolio will be discussed below.

4.3 Alternative frequency distribution hypothesis

The important source of discrepancy between the total annual loss distribution calculated via the individual loss model and calculated via the collective loss model is the frequency distribution approximation made within the collective model. In the examples discussed in the previous chapters we have given a detailed description of the portfolio individual frequency vector distribution simulation via a gaussian copula approach. In the collective model approximation we are considering only the distribution of the sum of the elements of the portfolio’s frequency vector. So far, we have assumed the Negative Binomial form for this sum.

The Negative Binomial distribution is the one widely used among other discrete distributions available in non-life insurance field, since it offers a good flexibility in terms of the 1st two moments. Indeed, the Poisson distribution requires the variance being exactly equal to the expected value, and the Binomial distribution requires the variance being strictly inferior to the expected value. Since in the real life situations, where the risks are correlated, the variance of the portfolio’s annual number of losses tends to be significantly higher than the expected value, the choice of the Negative Binomial distribution was the most convenient one.

The Negative Binomial distribution is a member of a large family of distributions, called “Mixed Poisson distributions”. It represents the “Poisson Gamma mixed distribution”, where the Poisson parameter (a parameter of the discrete variable) is a continuous Gamma variable, the so called “structure variable”. Therefore, we can potentially device a new set of alternative Poisson mixtures, where the structure variable is different from Gamma. This will ultimately lead to different collective model approximations.
4.4 Collective model approximation quality measure

In order to be able to compare the collective model approximation obtained via different approaches mentioned above, a certain quantitative measure should be established. The measure that we suggest takes into account the distribution of the total annual loss. An important criteria that we should keep in mind is the fact that the discrepancies in the distribution tails are the most significant for a risk manager (particularly in the reinsurance field). Therefore, we suggest a quality of approximation measure $\omega$ based on the weighted CDFs, the weights being designed in order to give more significance to the distribution’s tails.

Notation specific to this section:

\[ \tilde{F}(s) = p \] denotes the simulated estimate of the CDF of the total annual loss $S$

\[ n \] denotes the total number of Monte Carlo iterations performed

\[ k \] denotes the “standard” number of target values of $p$ used for the measure

Then the numerical estimation of the total annual loss CDF can be written as follows:

\[ \tilde{F}(s) = \tilde{P}(S \leq s) = \frac{\text{card}\{S \leq s\}}{n} = p \]

And the numerical estimation of the total annual loss expected value can be written as follows:

\[ \tilde{E}(S) = \sum_{S} \frac{S}{n} \]

We suggest the following measure based on the comparison of the numerical CDF estimates of the total annual loss calculated via collective and via individual loss models respectively:

\[ \omega = \sum_{i=1}^{k} \left( \frac{\text{abs}(\tilde{F}_{\text{coll}}^{-1}(p_i) - \tilde{F}_{\text{ind}}^{-1}(p_i))}{(1 - p_i)^{-1}} \right) \times \frac{1}{\sum_{i=1}^{k} (1 - p_i)^{-1} \times \tilde{E}_{\text{ind}}(S)} \]

The above sum is done over a standard set of values of $p_i$ in the discretized total annual loss CDFs obtained via Monte Carlo sampling in both collective and individual loss models. The total number of $p_i$ is equal to $k$ and will be standard list for all CDF compared.
5 Partial collective model: separating of the portfolio into “large exposures” and “attritional exposures”

This section will discuss two important issues:

1. If one part of the risk portfolio’s exposures is modelled using the individual loss model and another part using the collective loss model, then how these parts could be put together in order to ultimately model the portfolio’s total annual loss?

2. In some cases one part of the risk portfolio’s annual loss can be modelled as a single random variable. We will study the possible methods for characterization of such a variable. Please note that in this particular case we have a further simplification of the model since only a single random variable is involved.

The useful condition for the feasibility of the previously described frequency/severity method is that the average annual number of losses of the portfolio should not be too big. Indeed, it is very inconvenient to generate 1000 losses (or more) per year on average, especially if the bulk of these losses are attritional (i.e. small individual amounts) and are not likely to trigger a reinsurance contract (for example, an Excess of Loss structure).

The objective is then to sample the minimum number of variables, just necessary for the subsequent financial calculations. Typically, it is useless to sample all losses individually, regardless of their magnitude, if we are only interested by large losses, i.e. losses above a certain fixed threshold. The method allowing to realize this objective within our collective model is the subject of this section.

Definition
Any individual loss that is lower than a fixed threshold (called “large loss threshold”) is considered as an attritional loss.

Motivation
Quite often within reinsurance context, we are interested in modelling of “large” losses on an individual basis, whether for the rest of the losses (“attritional” ones) a modelling done on a cumulated basis as a single annual random variable is generally sufficient. In this case the collective model as a whole could be represented by 3 different random variables instead of two (as discussed previously). An alternative handling could be to continue using the individual loss model for large exposures.

Additional notation:

\( N_L \) corresponding to the total annual number of large losses;

\( X_L \) corresponding to the individual cost of each and every large loss, given a large loss occurs;

\( S_L \) corresponding to the total annual sum of large losses;

\( S_{Li} \) corresponding to the annual large loss from the i-th obligor;
\( S_A \) corresponding the total annual sum of attritional losses.

Then the total annual loss from the portfolio will be calculated as follows:

\[
S = S_L + S_A = \sum_{i=1}^{N_L} X_{L_i} + S_A
\]

In case where all the “large” losses are handled via the individual loss model we have:

\[
S = S_L + S_A = \sum_{i=1}^{n} S_{L_i} + S_A
\]

The objective of this chapter is to suggest a method for a characterization of \( S_A \) as a single random variable, which will make the simulation algorithm (Monte Carlo sampling) a lot easier and more straightforward. Also we need a method for the calculation of the sum of random variables \( S = S_L + S_A \).

This will be also useful for general characterization of the total annual portfolio loss burden as a single random variable, and especially for its 1\textsuperscript{st} and 2\textsuperscript{nd} moments calculation.

The method is based on splitting the modeled portfolio into two following components:

- large exposures (TSI or PML bigger than a fixed threshold)
- attritional exposures (TSI or PML less than a fixed threshold).

There is a slight difference here, compared to the “traditional” separation between large and attritional losses which was described above: in fact, the large exposures can still potentially produce some attritional losses, whereas the attritional exposures will always produce attritional losses by definition.

However, this separation raises 2 following important issues that must be also considered:

1. If we decide to model all annual losses produced by attritional exposures as a single random variable, then how can we characterize it? This question also holds if all losses are modeled as a single annual random variable.

2. Annual losses from attritional exposures are obviously correlated to the annual losses from large exposures in case of default event correlation described earlier in the individual model section. How can we measure and take into account this correlation in order to eventually model all annual losses?
5.1 Modelling of the annual loss from attritional exposures as a single random variable

Let us model the annual loss from attritional exposures as a single random variable. Because of correlations between defaults it is important to note that in certain cases the probability of “zero loss” in the portfolio is not negligible, therefore this variable is a “mixed” random variable, i.e. containing a discrete part and a continuous part. Because of this, the annual loss from attritional exposures cannot be modeled as being strictly continuous or strictly discrete variable.

The characterization of this variable will be done in the four following steps:

**Step1:** estimation of the 1st and the 2nd moments of the annual loss from attritional exposures.

**Step2:** estimation of the 1st and the 2nd moments of the continuous part of the annual loss from attritional exposures.

**Step3:** assuming a particular form (for example lognormal) for the continuous part of the annual loss from attritional exposures.

**Step4:** calculating the complete CDF of the mixed random variable “annual loss from attritional exposures”.

Even if the modelling of the discrete part explicitly seems to complicate the whole procedure, these steps are necessary because of the fact that the probability of the total annual loss equal to zero is often significant and it is therefore impossible to correctly model such a variable as a purely continuous one.

If all the default events are independent the probability of zero loss in the portfolio is calculated as follows:

\[ P(S = 0) = P(N = 0) = \prod_{i=1}^{n} (1 - p_i) \]

Which is quite often very small, especially if \( n \) is a big number. But in case when the default events are highly correlated, this probability often becomes significant and can no longer be ignored.

**Specific notation for this section:**

\( k \) is the number of attritional exposures, among \( n \) exposures in the portfolio

\( S_a \) is a mixed random variable corresponding to the “annual loss from attritional exposures”
$S_L$ is a mixed random variable corresponding to the “annual loss from large exposures”

$S = S_L + S_A$ is the total annual loss from the risk portfolio

$P[S_A = 0] = p$

$P[S_A > 0] = q = 1 - p$

\[
\begin{cases}
0 \quad \text{with} \quad P[S_A = 0] = p \\
Z \quad \text{with} \quad P[S_A > 0] = q
\end{cases}
\]

where $Z$ is a strictly positive continuous random variable which is the continuous part of the $S_A$ mixed random variable.

$N_A$ is the random variable corresponding to the “annual number of losses from attritional exposures”

$X_A$ is the random variable corresponding to the “Loss Given Default (LGD) for attritional exposures”.

5.1.1 Step 1 : Estimation of the 1st and the 2nd moments of the annual loss from attritional exposures.

We can calculate the expected annual loss from attritional exposures in the following way:

\[
E[S_A] = \sum_{i=1}^{k} (m_i \times \mu_i \times p_i) \quad (7)
\]

where

<table>
<thead>
<tr>
<th>Exposure (TSI or PML)</th>
<th>$m_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGD mean value</td>
<td>$\mu_i$</td>
</tr>
<tr>
<td>Probability of default</td>
<td>$p_i$</td>
</tr>
</tbody>
</table>

According to the definition of the attritional exposures the above sum is done for all “$i$” satisfying the following condition : $m_i < \text{threshold}$.

We can also calculate the variance of the annual loss from attritional exposures using the following well known property:

\[
\text{Var}[S_A] = E[N_A] \times \text{Var}[X_A] + \text{Var}[N_A] \times (E[X_A])^2 \quad (8)
\]
The necessary underlying assumption for the above formula is the fact that $N_A$ and $X_A$ variables are independent. Which also requires that all $X_i$ are independent and equally distributed for all “i”.

Let us assume here that the portfolio composed of attritional exposures is a perfectly homogeneous portfolio with independent LGD distributions, then the required assumptions for use of the formula (8) hold.

Besides, note that formula (8) doesn’t require the independence of $N_i$ variables, in our case they can be correlated.

We also need to keep in mind that if the required conditions for the use of this formula are not satisfied, we can always use the generalized variance calculation formula (24) based on the individual loss model, which is going to be explained in more detail in the subsequent sections of this paper.

We use (4) to calculate $E[N_A]$.
We use (5) and (6) to calculate $Var[N_A]$.

We calculate $E[X_A]$ according to the following formula:

$$
E[X_A] = \frac{E[S_A]}{E[N_A]} = \frac{\sum_{i=1}^{k} (m_i \times \mu_i \times p_i)}{\sum_{i=1}^{k} p_i}
$$

(9)

This is based on the fact that $E[S_A] = E[N_A] \times E[X_A]$, since, as it was already mentioned, the part of the risk portfolio composed of attritional exposures is homogeneous with mutually independent severities $X_i$. In other words, if the necessary conditions for the formula (8) hold, then the formula (9) also holds.

We calculate $Var[X_A]$ according to the following formula:

$$
Var[X_A] = E[X_A^2] - (E[X_A])^2 \quad \text{by definition},
$$

where

$$
E[X_A^2] = \frac{\sum_{i=1}^{k} (p_i \times E[X_i^2])}{\sum_{i=1}^{k} p_i}
= \frac{\sum_{i=1}^{k} (p_i \times (E[X_i]^2 + Var[X_i]))}{\sum_{i=1}^{k} p_i}
= \frac{\sum_{i=1}^{k} (p_i \times (\mu_i^2 + \sigma_i^2))}{\sum_{i=1}^{k} p_i}
$$
All these calculations are straightforward from the individual model parameters.

Therefore

\[
\text{Var}[X_A] = \sum_{i=1}^{k} \left( p_i m_i^2 (\mu_i^2 + \sigma_i^2) \right) - \frac{\left( \sum_{i=1}^{k} (m_i \times \mu_i \times p_i) \right)^2}{\sum_{i=1}^{k} p_i^2} \quad (10)
\]

Alternatively, if the required conditions for the use of the formulas (9) and (10) are not satisfied, then we directly use the formula (24) based on the individual loss model (as it was already mentioned above) and therefore we do not need to calculate \( E[X_A] \) and \( \text{Var}[X_A] \) anymore.

At this stage the 1\textsuperscript{st} and the 2\textsuperscript{nd} moments of the variable \( S_A \) are directly known from the individual model parameters. Such a calculation could be also performed for the risk portfolio as a whole (for the expected value and variance of \( S = S_L + S_A \)).

5.1.2 Step 2: Estimation of the 1\textsuperscript{st} and the 2\textsuperscript{nd} moments of the continuous part of the annual loss from attritional exposures.

Our aim here is to be able to calculate the 1\textsuperscript{st} and the 2\textsuperscript{nd} moment for \( Z \) variable, and then via assuming a particular form for it (for example lognormal), be able to model \( S_A \) as a single variable.

In fact, we know already the 1\textsuperscript{st} and the 2\textsuperscript{nd} moment for \( S_A \) and can easily calculate the probabilities \( P[S_A = 0] = p \) and \( P[S_A > 0] = q = 1 - p \).

These probabilities can be obtained from the distribution of the number of losses \( N_A \) supposed to be Negative Binomial as follows:

In fact, if \( N_A \) follows a Negative Binomial distribution with

\[
P[N_A = n] = \frac{\Gamma(\beta + n)}{\Gamma(\beta) \Gamma(n + 1)} \times \alpha^n \beta^n (1 - \alpha)^n \quad \text{with} \quad \alpha \quad \text{and} \quad \beta \quad \text{as parameters},
\]

then \( p = P[S_A = 0] = P[N_A = 0] = \alpha^0 \beta^0 \) and \( q = P[S_A > 0] = 1 - \alpha^0 \beta^0 \).
The above parameters $\alpha$ and $\beta$ are directly obtained from the 1st and the 2nd moments of $N_A$ which are already known in our case:

$$
\alpha = \frac{E[N_A]}{Var[N_A]} \quad \text{and} \quad \beta = \frac{E[N_A]^2}{Var[N_A] - E[N_A]}
$$

The next step is to use the following moment property for mixed variables:

$$
E[S_A] = E[S_A / S_A > 0] \times P[S_A > 0] + E[S_A = 0] \times P[S_A = 0] \quad (11)
$$

Proof:

$$
E[S_A] = E[1_{S_A > 0} S_A] + E[1_{S_A = 0} S_A] = \frac{E[1_{S_A > 0} S_A]}{P[S_A > 0]} \times P[S_A > 0] + \frac{E[1_{S_A = 0} S_A]}{P[S_A = 0]} \times P[S_A = 0] =
$$

$$
= E[S_A / S_A > 0] \times P[S_A > 0] + E[S_A / S_A = 0] \times P[S_A = 0]
$$

In our particular case the following formula for $E[Z]$ and $Var[Z]$ can be obtained using the above moment property:

$$
E[Z] = \frac{E[S_A]}{q} \quad (12)
$$

$$
Var[Z] = \frac{Var[S_A]}{q} - p \left( \frac{E[S_A]}{q} \right)^2 \quad (13)
$$

Proof:

$$
E[S_A] = E[S_A / S_A > 0] \times P[S_A > 0] + E[S_A = 0] \times P[S_A = 0] = E[Z] \times q
$$

Therefore

$$
E[Z] = \frac{E[S_A]}{q}
$$

$$
E[S_A^2] = E[S_A^2 / S_A^2 > 0] \times P[S_A^2 > 0] + E[S_A^2 = 0] \times P[S_A^2 = 0] = q \times E[Z^2] =
$$

$$
= q \times (Var[Z] + E^2[Z]) = qVar[Z] + \frac{1}{q} E^2[S_A]
$$
Therefore
\[
Var[Z] = \frac{E[S_A^2]}{q} - q^{-1}E[Z] = \frac{Var[S_A] + E^2[S_A] - q^{-1}E^2[Z]}{q} = \frac{Var[S_A]}{q} - p\left(\frac{E[S_A]}{q}\right)^2
\]

5.1.3 Step 3: Assuming a particular form (lognormal for example) for the continuous part of the annual loss from attritional exposures

\[Z \sim \text{LogNorm}(E[Z], Var[Z])\]

This allows us to calculate the probabilities in the \(Z\) variable CDF:

\[F_Z(z) = P(Z < z)\]

5.1.4 Step 4: Calculating the CDF of the variable “annual loss from attritional exposures”

We can write the CDF of the \(S_A\) (mixed variable) in the following way using the total probability formula, given the fact that \(S_A\) is a non-negative variable:

for any \(s > 0\)

\[F_{S_A}(s) = F_Z(s) \times q + p \quad (14)\]

\[F_{S_A}(s) = P[S_A < s] = P[S_A < s \cap S_A > 0] + P[S_A < s \cap S_A = 0]\]

\[F_{S_A}(s) = P[S_A < s / S_A > 0] \times P[S_A > 0] + P[S_A < s / S_A = 0] \times P[S_A = 0]\]

\[F_{S_A}(s) = P[Z < s] \times P[S_A > 0] + P[0 < s] \times P[S_A = 0] = F_Z(s) \times q + p\]

This calculation (14) can be done numerically for any list of positive “\(s\)” values. Then the variable \(S_A\) can be directly sampled by MonteCarlo algorithm from the numerical CDF. With large enough number of points in the numerical definition of this CDF and also the appropriate interpolation between these points the simulation will work correctly.

We can also define a mixed variable directly \(S_A\) in a parametric form with \(p\) and \(q\) being the “weights” used for the discrete and the continuous parts respectively.

In the 1st place we need to determine which of the two possible scenarios occurred:
**Scenario 1**: the annual loss is equal to zero. The probability of this scenario is \( P(S_A = 0) = p \)

**Scenario 2**: the annual loss is positive. The probability of this scenario is \( P(S_A > 0) = q = 1 - p \)

We will 1st sample from a discrete Bernoulli variable with a parameter equal to \( p \) in order to determine the particular scenario occurred. The next step will consist in sampling the amount of the annual loss : 

If the scenario 1 is sampled, then \( S_A = 0 \)

If the scenario 2 is sampled, then \( S_A \sim \text{LogNorm}(E[Z], Var[Z]) \) and needs additional sampling from this parametric distribution.

### 5.2 Modelling of the total annual loss using the partial collective model

As we have seen, our method consists in modelling large exposures using the individual loss model and modelling attritional exposures separately using a single random variable or using a classical “frequency/severity” collective model. The resulting problem will then be the way we can put both parts together. Obviously, these two parts of the annual total loss are correlated together but we do not know so far the level of this correlation.

Here is the method we suggest for it:

**Step 1**:

At this step we will estimate the covariance between the annual loss from large exposures and the annual loss from attritional exposures. In order to do this we first need to calculate the variance for the following three random variables :

- variance of the annual loss from large exposures
- variance of the annual loss from attritional exposures
- variance of the total annual loss

This variance calculation is done using the equations (4), (5), (6), (8), (9) and (10).

For large exposures and all exposures it is the same approach as the one used above for attritional exposures only, we just sum over “n-k” terms and “n” terms respectively, instead of summing over “k” terms, as it was done for attritional exposures. Once we know the variance of the total annual loss, the variance of annual loss from large exposures and the variance of annual loss from attritional exposures, we can easily derive the covariance needed here:

\[
\text{Cov}(S_A, S_L) = \frac{1}{2} \times [\text{Var}(S = S_A + S_L) - \text{Var}(S_A) - \text{Var}(S_L)]
\]

where \( S_L \) denotes a random variable “annual loss from large exposures” and \( S = (S_A + S_L) \) denotes a random variable “total annual loss”.

---

54/103
Step 2:

We use normal copula with the required correlation coefficient between the two parts of the portfolio (i.e. large exposures and attritional exposures) in order to calculate the global annual loss distribution by Monte Carlo sampling:

\[
\rho(S_A, S_L) = \frac{\text{Cov}(S_A, S_L)}{\sqrt{\text{Var}(S_A)\text{Var}(S_L)}}
\]

In this case we do not need to make a particular assumption regarding the form of the resulting distribution. All we need is the coefficient of correlation and a normal copula in order to match the total annual loss variance calculated previously.

The sampling method used in this case is the “rank correlation” method which allows to correlate the numerical samples from attritional exposures and from large exposures using the annual loss variable correlation and the gaussian copula. The details of the required algorithm are explained in the Appendix.

5.3 Method application and analysis: real life examples with credit insurance portfolio

We will use the same real life portfolio that is being studied here from the beginning. All the individual frequencies are correlated as it was explained above.

We will start by comparing the numerical CDFs from the simulation of the individual loss model and the partial collective loss model.

Different thresholds are going to be explored for the definition of the partial collective model:

- 50m
- 10m
- 3m

This will show not only the impact of the method on the collective model approximation itself but also the sensitivity of the results regarding the choice of a particular threshold to be used.

5.3.1 Example of the partial collective model built for 50m threshold

The partial collective loss model is defined as follows:

- Exposures bigger than 50m (called “large exposures”) are modeled via individual loss model (based on gaussian copula for frequency correlation).
- Exposures smaller than 50m (called “attritional exposures”) are modelled via collective loss model based on a classical frequency/severity simulation and a Negative Binomial assumption made for the portfolio’s annual frequency distribution form.
Some important comments about this separation of the risk portfolio between “large” and “attritional exposures”.

The total number of large exposures in this example is 24, whereas the total number of attritional exposures is 1976 (the total number of obligors is still 2000). The interesting observation is the fact that even if the large exposure represent a very small part of the total number of obligors (1.2%) they still correspond to a significant part of the portfolio total exposure (32%) This is due to the portfolio’s risk profile where we have some very huge individual exposures that could potentially behave like “risk drivers” for the insurance company having a significant impact on the total annual loss distribution tail. Therefore, the application of the partial collective model in this case could increase the quality of the individual model’s approximation while having a very straightforward handling since the portfolio modelled individually is small in size (number of obligors).

We can observe the similar relation between the CDF estimated using the collective model and the CDF using the individual model. But the discrepancy is decreased via the partial collective model approach. Therefore, the quality of the collective model approximation is enhanced.

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>8 089 917</td>
<td>9 031 261</td>
<td>- 941 345</td>
<td>inf</td>
</tr>
<tr>
<td>60%</td>
<td>13 808 970</td>
<td>13 640 396</td>
<td>168 574</td>
<td>sup</td>
</tr>
<tr>
<td>70%</td>
<td>22 815 350</td>
<td>20 934 242</td>
<td>1 881 108</td>
<td>sup</td>
</tr>
<tr>
<td>80%</td>
<td>36 526 135</td>
<td>32 945 278</td>
<td>3 580 857</td>
<td>sup</td>
</tr>
<tr>
<td>90%</td>
<td>63 389 619</td>
<td>56 348 981</td>
<td>7 040 638</td>
<td>sup</td>
</tr>
<tr>
<td>95%</td>
<td>91 997 846</td>
<td>86 416 166</td>
<td>5 581 680</td>
<td>sup</td>
</tr>
<tr>
<td>99%</td>
<td>175 651 612</td>
<td>187 577 365</td>
<td>- 11 925 753</td>
<td>inf</td>
</tr>
<tr>
<td>99.50%</td>
<td>222 694 548</td>
<td>245 960 235</td>
<td>- 23 265 687</td>
<td>inf</td>
</tr>
</tbody>
</table>
In order to quantify this approximation quality enhancement we will calculate the measure suggested above:

\[ \omega = 0.35 \]

Whereas, for the total collective frequency/severity loss model this same measure is the following:

\[ \omega = 0.42 \]

### 5.3.2 Example of the partial collective model built for 10m threshold

In this example, the partial collective loss model is defined as follows:

- Exposures bigger than 10m (called “large exposures”) are modeled via individual loss model (based on gaussian copula for frequency correlation).

- Exposures smaller than 10m (called “attritional exposures”) are modelled via collective loss model based on a classical frequency/severity simulation and a Negative Binomial assumption made for the portfolio’s annual frequency distribution form.

The total number of large exposures in this example is 184, whereas the total number of attritional exposures is 1816. The large exposures modeled individually represent 9.2% of the total number of obligors and 64% of the total exposure. This again underlines the fact that the risk portfolio studied here is highly concentrated and heterogeneous.

The simulation results using the partial collective model with 10m threshold exhibit further enhancement of the collective model approximation quality:
In order to quantify this approximation quality enhancement we will calculate the measure suggested above:

\[ \omega = 0.19 \]

Whereas, for the total collective frequency/severity loss model this same measure is the following:

\[ \omega = 0.42 \]

### 5.3.3 Example of the partial collective model built for 3m threshold

In this example, the partial collective loss model is defined as follows:

- Exposures bigger than 3m (called “large exposures”) are modeled via individual loss model (based on gaussian copula for frequency correlation)

- Exposures smaller than 3m (called “attritional exposures”) are modelled via collective loss model based on a classical frequency/severity simulation and a Negative Binomial assumption made for the portfolio’s annual frequency distribution form.

The total number of large exposures in this example is 591, whereas the total number of attritional exposures is 1409. The large exposures modeled individually represent 30% of the total number of obligors and 86% of the total exposure.

The simulation results using the partial collective model with 3m threshold exhibit further enhancement of the collective model approximation quality:

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>7 661 695</td>
<td>9 031 261</td>
<td>-1 369 566</td>
<td>inf</td>
</tr>
<tr>
<td>60%</td>
<td>13 501 472</td>
<td>13 640 396</td>
<td>-1 138 924</td>
<td>inf</td>
</tr>
<tr>
<td>70%</td>
<td>21 881 772</td>
<td>20 934 242</td>
<td>947 530</td>
<td>sup</td>
</tr>
<tr>
<td>80%</td>
<td>35 546 280</td>
<td>32 945 278</td>
<td>2 601 002</td>
<td>sup</td>
</tr>
<tr>
<td>90%</td>
<td>62 820 199</td>
<td>56 348 981</td>
<td>6 471 217</td>
<td>sup</td>
</tr>
<tr>
<td>95%</td>
<td>94 114 425</td>
<td>86 416 166</td>
<td>7 698 258</td>
<td>sup</td>
</tr>
<tr>
<td>99%</td>
<td>181 564 364</td>
<td>187 577 365</td>
<td>-6 013 001</td>
<td>inf</td>
</tr>
<tr>
<td>99.50%</td>
<td>232 169 120</td>
<td>245 960 235</td>
<td>-13 791 115</td>
<td>inf</td>
</tr>
</tbody>
</table>

**Total annual loss CDF (simulation results)**
In order to quantify this approximation quality enhancement we will calculate the measure suggested above:

$$\omega = 0.08$$

Whereas, for the total collective frequency/severity loss model this same measure is the following:

$$\omega = 0.42$$

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>8 946 999</td>
<td>9 031 261</td>
<td>-</td>
<td>84 262</td>
</tr>
<tr>
<td>60%</td>
<td>14 133 351</td>
<td>13 640 396</td>
<td>492 955</td>
<td>sup</td>
</tr>
<tr>
<td>70%</td>
<td>21 848 447</td>
<td>20 934 242</td>
<td>914 205</td>
<td>sup</td>
</tr>
<tr>
<td>80%</td>
<td>34 497 319</td>
<td>32 945 278</td>
<td>1 552 041</td>
<td>sup</td>
</tr>
<tr>
<td>90%</td>
<td>59 512 350</td>
<td>56 348 981</td>
<td>3 163 368</td>
<td>sup</td>
</tr>
<tr>
<td>95%</td>
<td>88 995 958</td>
<td>86 416 166</td>
<td>2 579 792</td>
<td>sup</td>
</tr>
<tr>
<td>99%</td>
<td>190 429 582</td>
<td>187 577 365</td>
<td>2 852 216</td>
<td>sup</td>
</tr>
<tr>
<td>99.5%</td>
<td>235 940 771</td>
<td>245 960 235</td>
<td>- 10 019 464</td>
<td>inf</td>
</tr>
</tbody>
</table>
5.3.4 Conclusion for the partial collective model method

The threshold used for the separation of the portfolio into large and attritional exposures is based on the portfolio profile, the lower the threshold the higher is the approximation quality of the collective model (low values of the $\omega$ measure).

In our example the relationship between the threshold and the approximation quality measure $\omega$ is the following:

![Graph showing the relationship between threshold and $\omega$ measure](image)

Therefore, the approximation quality can be potentially improved by modelling only a small part of the global risk portfolio via the individual loss model, particularly for the cases where a small number of largest exposures represent a significant part of the overall portfolio exposure.

5.4 Modelling of the total annual loss as a single random variable: real life example with credit insurance portfolio

As an illustration of the alternative approach, where the annual total loss is modelled as a single random variable, we will calculate the total annual loss distribution using the Lognormal assumption for the continuous part of the total annual loss and compare the results with the individual loss model.

We use the above described method in order to characterize the mixed random variable corresponding to the total annual loss of the risk portfolio:

We assume $N \sim \text{NegBin}(E(N), \text{Var}(N))$
Then, the parameters of the Negative Binomial distribution can be calculated as follows, using the method of moments:

\[ \alpha = \frac{E[N]}{Var[N]} = 0.0192 \]

and

\[ \beta = \frac{E[N]^2}{Var[N] - E[N]} = 0.4936 \]

Then \( P(S = 0) = P(N = 0) = \alpha^\beta = 0.1420 \)

We have already estimated the 1st and the 2nd moments of the random variable “total annual loss”:

\[ E(S) = 23.19m \]
\[ Var(S) = 1.58854E + 15 \text{ (the standard deviation is equal to 39.86m)} \]

This knowledge allows us to calculate the 1st and the 2nd moments of the continuous part of the mixed random variable “total annual loss”:

\[ E[Z] = \frac{E[S]}{P(S > 0)} = 27.03m \]
\[ Var[Z] = \frac{Var[S]}{P(S > 0)} - P(S = 0) \left( \frac{E[S]}{P(S > 0)} \right)^2 = 1.74768E + 15 \]

(the standard deviation is equal to 41.8m)

We assume \( Z \sim \text{LogNorm}(E(Z), Var(Z)) \)

Then, we will perform a two-step MonteCarlo sampling in order to simulate the mixed random variable “total annual loss”:

- Step 1: Sampling from a Bernoulli variable with a parameter equal to \( P(S > 0) \)
- Step 2: If the result of the step 1 sampling is equal to 1, then the distribution of \( S \) is the same as the distribution of \( Z \) and we sample from \( Z \). If the result of the step 1 is equal to zero, then \( S = 0 \) and no further sampling is required.
We obtain the following result:

![Total annual loss CDF](chart.png)

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>11 743 729</td>
<td>9 031 261</td>
<td>2 712 467</td>
<td>sup</td>
</tr>
<tr>
<td>60%</td>
<td>16 539 668</td>
<td>13 640 396</td>
<td>2 899 272</td>
<td>sup</td>
</tr>
<tr>
<td>70%</td>
<td>22 827 842</td>
<td>20 934 242</td>
<td>1 893 600</td>
<td>sup</td>
</tr>
<tr>
<td>80%</td>
<td>33 075 425</td>
<td>32 945 278</td>
<td>130 147</td>
<td>sup</td>
</tr>
<tr>
<td>90%</td>
<td>55 671 614</td>
<td>56 348 981</td>
<td>-</td>
<td>- inf</td>
</tr>
<tr>
<td>95%</td>
<td>84 203 271</td>
<td>86 416 166</td>
<td>- 2 212 895</td>
<td>inf</td>
</tr>
<tr>
<td>99%</td>
<td>185 833 308</td>
<td>187 577 365</td>
<td>- 1 744 058</td>
<td>inf</td>
</tr>
<tr>
<td>99.50%</td>
<td>241 384 889</td>
<td>245 960 235</td>
<td>- 4 575 346</td>
<td>inf</td>
</tr>
</tbody>
</table>

We can observe that the approximation of the total annual loss is better than using the previous methods. The approximation quality measure for this model is significantly lower since the distribution tail is much better approximated:

\[
\omega = 0.15
\]
Even if the approximation is better in this particular case, we need to keep in mind that this method only allows the simulation of the total annual loss. If we need some further details, i.e. individual loss frequencies and amounts, then the “classical” frequency severity approach is required. But even in such a situation the mixed lognormal approximation could be used for the attritional losses modelling in when 3 random variables are used within the collective model:

- Annual number of large losses
- Individual loss amount
- Total annual attritional loss

This technique helps to reduce the number of losses simulated individually in a collective loss model. For example, we model large losses individually and attritional losses as a single random variable in the way described above.
6 Portfolio stratification method

As we have seen from the studied examples, the above described collective frequency-severity model will not work for a portfolio with some type of \( X_i \) dependence or some of dependence between \( X \) variable and \( N \) variable.

In order to solve this problem and still avoid the detailed individual model approach, which basically the central goal of this paper, we will separate the total risk portfolio into distinct parts, each of those being an homogeneous risk portfolio on its own. The number of such parts depends on the risk portfolio profile (distribution of the exposure size), types and variety of the LGD models used and also the default correlation structure. Of course, in some particular cases such a strategy could still be time consuming since the number of the resulting sub-portfolios could be rather big. But in the vast majority of cases, especially when the risk portfolio is highly diversificated, a low number of sub-portfolios could suffice, since we are looking for a reasonable approximation of the detailed individual model by a collective one.

Our portfolio is separated into “\( m \)” homogeneous parts, where for each part we note:

\[
\begin{align*}
N_i & \quad \text{annual number of losses in the } l\text{-th part of the portfolio} \\
X_i & \quad \text{LGD in the } l\text{-th part of the portfolio} \\
S_i & \quad \text{annual loss in the } l\text{-th part of the portfolio}
\end{align*}
\]

With \( l \in (1,m) \)

Besides,

- we assume here that \( N_i \perp X_i \) for each “\( l \)” in our stratification. In the other words this means that each sub-portfolio is homogeneous.
- we also assume that inside each sub-portfolio all individual risk severities are independent : for any \( i \neq j \in (1,n) \) \( X_i \perp X_j \), where “\( n \)” is the total number of risks inside a sub-portfolio.

6.1 Building frequency/severity model for each part

If the above homogeneity assumption is satisfied, then a collective frequency severity model will work for each part of the portfolio taken separately from the others.

We characterize the annual number of losses \( N_i \) using equations (4), (5) and (6):

\[
N_i \sim \text{NegBin}(E[N_i], Var[N_i])
\]
We characterize the LGD distribution for the \( l \)-th part of the portfolio using the following approach:

\[
F_{X_l}(x) = P\{X_l \leq x\} = \sum_{i=1}^{\infty} \left( \frac{p_i}{n} \times P\{X_i \leq x\} \right),
\]

where "\( n \)" is the total number of risks inside the \( l \)-th part of the portfolio.

For each part of the portfolio the frequency and the severity can be sampled by MonteCarlo algorithm as independent variables of the collective model. In the other words, we can consider each part as a separate portfolio with a separately built collective frequency/severity models.

### 6.2 Calculation of the covariance matrix for the annual losses from different parts of the stratified portfolio

Once we get the correct frequency/severity collective model for each and every “homogeneous” part separately, we still need to find out the correlation between losses coming from different parts in order to be able to eventually put these different portfolio parts together.

Let us take \( l = 1 \) and \( l = 2 \) here as an example. The calculation is easily generalized in order to get a \((m \times m)\) covariance matrix.

The 1\textsuperscript{st} step will be to calculate the covariance matrix for the variables “annual number of losses” in each part of the portfolio.

\[
\text{Cov}(N_1, N_2) = \frac{1}{2} \times [\text{Var}(N_1 + N_2) - \text{Var}(N_1) - \text{Var}(N_2)] \quad (15)
\]

where all variances are calculated using equations (5) and (6) from the parameters on the individual loss model.

The 2\textsuperscript{nd} step will be to calculate the covariance matrix for the variables “annual loss” from each part of the portfolio according to the following formula:

\[
\text{Cov}(S_1, S_2) = E(X_1) \times E(X_2) \times \text{Cov}(N_1, N_2) \quad (16)
\]
Proof:

\[
\text{Cov}(S_1, S_2) = \text{Cov} \left( \sum_{i=1}^{N_1} X_{1i}, \sum_{j=1}^{N_2} X_{2j} \right) = E \left[ \sum_{i=1}^{N_1} X_{1i} \times \sum_{j=1}^{N_2} X_{2j} \right] - E \left[ \sum_{i=1}^{N_1} X_{1i} \right] \times E \left[ \sum_{j=1}^{N_2} X_{2j} \right]
\]

\[
E \left[ \sum_{i=1}^{N_1} X_{1i} \times \sum_{j=1}^{N_2} X_{2j} \right] = E \left[ \sum_{i=1}^{N_1} X_{1i} \times \sum_{j=1}^{N_2} X_{2j} / N_1, N_2 \right] = E \left[ E[N_1 N_2 X_1 X_2 / N_1, N_2] \right] =
\]

\[
= E[N_1 N_2 E[X_1 X_2 / N_1, N_2]] = E[N_1 N_2 E[X_1] E[X_2]] = E[X_1] E[X_2] E[N_1 N_2]
\]

Then

\[
\text{Cov}(S_1, S_2) = E[X_1] E[X_2] E[N_1 N_2] - E \left[ \sum_{i=1}^{N_1} X_{1i} \right] \times E \left[ \sum_{j=1}^{N_2} X_{2j} \right] =
\]

\[
= E[X_1] E[X_2] E[N_1 N_2] - E[N_1] E[X_1] E[N_2] E[X_2] = E[X_1] E[X_2] \times \text{Cov}[N_1, N_2]
\]

The above proof needs the following assumptions:

1. \( X_{1i} \) are iid within portfolio part 1
2. \( X_{2j} \) are iid within portfolio part 2
3. \( N_1 \perp X_1 \) and \( N_2 \perp X_2 \)
4. \( X_1 \perp X_2 \)

It can be noticed that the assumptions 1 and 2 are included in the condition number 3. They have been stipulated separately in order to underline the fact that all the LGDs must be equidistributed for the validity of our method of calculation of the distribution of \( X \) variable. Also we would like to underline the fact that all the individual loss severities were assumed independent in our model.

These assumptions are all satisfied in our example since we are considering that both parts of the risk portfolio are homogeneous risk portfolios on their own. Also, there is no severity dependence assumed in the model at this stage. Later we will relax this particular condition.

On the contrary, any type of frequency dependence is allowed in this formula, which corresponds to our case where different default events can be potentially correlated between them. Therefore the random variables \( N_1 \) and \( N_2 \) can be also correlated.

6.3 Simulation of the global portfolio losses

Since we now have the covariance matrix for the variables “annual loss” in our stratified portfolio, we can easily calculate the corresponding correlation matrix using:

\[
\rho(S_1, S_2) = \frac{\text{Cov}(S_1, S_2)}{\sqrt{\text{Var}(S_1) \text{Var}(S_2)}} \quad (17)
\]
Where $\text{Var}(S_1)$ and $\text{Var}(S_2)$ can be obtained from the individual loss model parameters using the equations (4), (5), (6), (8), (9) and (10).

Please note that the formula (8) works for the annual loss variance calculation only if $N_1 \perp X_1$ and $N_2 \perp X_2$, which is assumed to be the case here since our sub-portfolios are all homogeneous.

We are then able to use the above correlation matrix in order to generate the numerical sample for the portfolio as a whole. We use here an approach similar to the one that we have already used for the case of portfolio separation into “attritional exposures” and “large exposures”. The difference is that in this case we can have an arbitrary number of sub-portfolios simulated together. Another difference is that in each sub-portfolio we can have a complete frequency/severity model for large exposures and also an additional random variable corresponding to the annual loss burden from attritional exposures.

As in the case of “attritional exposures” and “large exposures”, we will use the method of “rank correlation”, already mentioned in the previous section and that is discussed in details in the Appendix.

At last, the attritional annual losses, modeled as a single random variable, can be also integrated as an “additional homogeneous part” to the global portfolio, using equations (16) and (17).

We can also calculate the annual loss variance of the portfolio as a whole, using the following formula:

$$\text{Var}(S) = \text{Var}(\sum_{l=1}^{m} S_l) = \sum_{l=1}^{m} \sum_{k=1}^{m} \text{Cov}(S_l, S_k)$$

(18)

since we have already calculated all the above mentioned variances and covariances.

The above (18) formula therefore allows us to calculate the total annual loss variance in cases when the number of losses in the global portfolio is not necessarily independent from the individual loss amount.

### 6.4 Method application and analysis: theoretical portfolio example

As it was already mentioned, different criteria could be used for the portfolio stratification:
- Size of risks
- Level of correlation between risks
- Probabilities of default

The portfolio stratification based on the size of individual risks is closely related to the concept of portfolio homogeneity previously discussed in detail. We have already produced some results for such a stratification in the chapter where the “large” exposures were separated from the “attritional” exposures. Further segments could be created, for example “large”, “mid-sized” and “small” exposures, etc. The more segments we have, the more “homogeneous” they are likely to be. It all depends on the portfolio’s profile and the
The approach explained above gives us the possibility to integrate together an arbitrary number of separate portfolio segments.

Let us first underline an important issue: even a perfectly “homogeneous” portfolio segment could still result in a significant discrepancy between the modeled CDFs from the individual and a collective model.

Let us consider a theoretical risk portfolio with the following parameters:

- 200 homogeneous exposures with a fixed loss given default = 1,000,000
- Each exposure has the same probability of default = 12.5%
- All exposures’ frequencies have the same coefficient of correlation = 50%

We approximate this risk portfolio by using a complete “frequency/severity” collective model based on the Negative Binomial assumption for the portfolio’s frequency distribution and compared this approximation with the individual model simulation results:

![Total annual loss CDF](image)

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>6,000,000</td>
<td>1,000,000</td>
<td>5,000,000</td>
<td>sup</td>
</tr>
<tr>
<td>60%</td>
<td>11,000,000</td>
<td>4,000,000</td>
<td>7,000,000</td>
<td>sup</td>
</tr>
<tr>
<td>70%</td>
<td>21,000,000</td>
<td>13,000,000</td>
<td>8,000,000</td>
<td>sup</td>
</tr>
<tr>
<td>80%</td>
<td>39,000,000</td>
<td>35,000,000</td>
<td>4,000,000</td>
<td>sup</td>
</tr>
<tr>
<td>90%</td>
<td>75,000,000</td>
<td>92,000,000</td>
<td>-17,000,000</td>
<td>inf</td>
</tr>
<tr>
<td>95%</td>
<td>116,000,000</td>
<td>142,000,000</td>
<td>-26,000,000</td>
<td>inf</td>
</tr>
<tr>
<td>99%</td>
<td>221,500,000</td>
<td>193,500,000</td>
<td>28,000,000</td>
<td>sup</td>
</tr>
<tr>
<td>99.5%</td>
<td>270,500,000</td>
<td>198,500,000</td>
<td>72,500,000</td>
<td>sup</td>
</tr>
</tbody>
</table>

We can observe some important discrepancies, particularly in the distribution tails. We can also observe multiple crossing points in the simulated CDFs.

The important conclusion is that even for a perfectly homogeneous portfolio there could be still a significant discrepancy between the total annual loss distributions obtained via the
collective and the individual loss model. Besides, the results are not comparable at the 2\textsuperscript{nd} order (multiple crossing points in CDFs) mainly due to the Negative Binomial assumption made for the portfolio’s frequency distribution.

Despite the above highlighted issue regarding the perfectly homogeneous portfolios, the portfolio stratification could potentially result in better approximation of the individual model by the collective model. We have already seen it when “large” exposures were separated from “attritional” exposures.
7 Alternative frequency distribution hypothesis

In the examples discussed so far, we have made the Negative Binomial assumption for the frequency distribution of the collective risk model. As it was already mentioned, the Negative Binomial distribution represents a “Gamma Poisson mixed variable”, therefore we could expand the possibilities of the collective model approximation by building a different set of discrete random variables while using a “structure variable” different from Gamma variable.

Let us first start with some background on mixed Poisson variables.

7.1 Poisson mixed variables: definition, moment properties and simulation method

In the actuarial litterature the Poisson mixed variables are often used to adress the portfolio’s frequency heterogeneity issues. In our particular case we will use it primarily as a tool for generation of an alternative frequency distribution in a risk portfolio with correlated risks.

Definition and Probability function

Let us consider a random bivariate vector \( (N, \Lambda) \) where \( \Lambda > 0 \),

The conditional distribution \( (N / \Lambda = \lambda) \sim \text{Poisson}(\lambda) \)

and

The marginal CDF of the \( \Lambda \) random variable is denoted as \( H \),

Then the marginal distribution of \( N \) random variable is called “H-mixed Poisson” distribution, where \( \Lambda \) is called “structure variable” and it’s CDF \( H \) is called “structure function”.

The probability function of the H-mixed Poisson distribution can be calculated as follows:

\[
P(N = n) = \int_0^\infty P(N = n / \Lambda = \lambda) dH(\lambda) = \int_0^\infty e^{-\lambda} \frac{\lambda^n}{n!} dH(\lambda) \quad \text{for all} \ n \in N.
\]

In Partrat, Besson “Assurance Non-Vie, Modélisation et Simulation, p228”:

Moment Properties

For a Poisson H-mixed random variable with \( \Lambda \) defined as the structure variable, and assuming the following moments exist, we have:

i) \( E(N) = E(\Lambda) \)

ii) \( Var(N) = E(\Lambda) + Var(\Lambda) \)
Simulation method

Given the fact that for a Poisson mixed variable $N$, the Poisson fixed parameter $\lambda$ becomes a random variable $\Lambda$, there is a natural way to simulate a Poisson mixed variable:

If the structure function of the Poisson mixed variable is defined as follows $H = P(\Lambda < \lambda)$

Then the Poisson mixed variable $N$ can be simulated by Monte Carlo in the following two steps:

**Step 1**: Simulate the $k$-sample of the structure variable $\{\Lambda = \lambda_k\}$

**Step 2**: Simulate the $k$-sample of the Poisson mixed variable $\{N = n_k\}$ by simulating a realisation of the Poisson variable $(N / \Lambda = \lambda_k) \sim \text{Poisson}(\lambda_k)$

A very interesting aspect of this simulation is the fact that we can generate any type of Poisson mixture via this simple method. Even if either the probability function or the CDF of such a mixed variable cannot be written in a closed formula, it still can be easily simulated numerically and thus integrated into the global collective model simulation scheme discussed above.

7.2 Poisson-Gamma mixed variable and alternative mixtures

In the previously discussed examples of the risk portfolio’s frequency approximation we have calculated the 1st and the 2nd moments of portfolio’s frequency distribution, then we have assumed a Negative Binomial form in order to be able to model it and ultimately to be able to simulate the total annual loss numerically.

It can be shown that the Negative Binomial distribution is a Poisson-Gamma mixed distribution:

If $\Lambda \sim \text{Gamma}(r, \frac{p}{q})$ with a density function $f_\Lambda$

Then The Poisson-Gamma mixed variable probability function can be written as follows (according to Partrat, Besson : “Assurance non-vie : Modélisation, Simulation », p205)

$$P(N = n) = \int_0^\infty P(N = n / \Lambda = \lambda)f_\Lambda(\lambda)d\lambda$$

$$= \frac{(p / q)^r}{\Gamma(r)} \int_0^\infty e^{-\lambda} \frac{\lambda^n}{n!} \frac{p\lambda}{q} \lambda^{r-1} d\lambda$$

$$= \frac{(p / q)^r}{\Gamma(r)n!} \int_0^\infty e^{-\left(\frac{1}{q}\frac{n}{q}\right)\lambda} \lambda^{n+r-1}d\lambda$$
\[
\frac{(p/q)^r \Gamma(r + n)}{\Gamma(r + n)} = \frac{\Gamma(r + n)}{\Gamma(r)!} q^n \quad \text{for} \quad n \geq 0
\]

The structure variable \( \Lambda \sim \text{Gamma} \left( r, \frac{p}{q} \right) \) results in the mixed variable \( N \sim \text{NegBin}(r, p) \) according to the above demonstration.

The general idea here will be to replace the Gamma structure variable by a different one. The parameters of this new structure variable can be easily calculated using the Poisson mixed variables moment properties which were mentioned above.

### 7.3 Poisson-Lognormal mixed variable: real life example with credit risk portfolio

We get back to the real life example using the credit risk portfolio. The 1st and the 2nd moment of the total annual number of losses in the risk portfolio was already calculated as follows:

\[
E[N] = \sum_{i=1}^{n} E[N_i] = \sum_{i=1}^{n} p_i = 25.27
\]

\[
Var[N] = Var[\sum_{i=1}^{n} N_i] = \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(N_i, N_j) = 1319.31
\]

Then the respective moments of the structure variable in the Poisson mixture can be calculated as follows:

\[
E[\Lambda] = E[N] = 25.27
\]

\[
Var[\Lambda] = Var[N] - E[N] = 1294.03
\]

The next step is to simulate a numerical sample of the structure variable \( \Lambda \) for which we assume a lognormal form: \( \Lambda \sim \text{LogNorm}(E(\Lambda), Var(\Lambda)) \)

Then for each realisation of the structure variable \( \Lambda = \lambda \) we will simulate another realisation of the Poisson variable with a fixed parameter equal to \( \lambda \).
The resulting mixed distribution will be a discrete Poisson-Lognormal mixed distribution. Here is the comparison between this new Poisson mixture and the previously used Negative Binomial assumption:

We can observe a significant discrepancy in the distribution tails above the 99.5% probability, the Poisson-Lognormal mixed distribution having a heavier tail than the Negative binomial distribution.

This is also confirmed by the following plot:
The next step will be to compare the collective model result using the Poisson-Lognormal assumption for the frequency distribution with the result using the individual loss model:

![Total annual loss CDF](image)

### Total annual loss CDF (simulation results)

<table>
<thead>
<tr>
<th>%ile</th>
<th>Collective model</th>
<th>Individual model</th>
<th>Discrepancy</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>11 727 140</td>
<td>9 031 261</td>
<td>2 695 879</td>
<td>sup</td>
</tr>
<tr>
<td>60%</td>
<td>16 479 928</td>
<td>13 640 396</td>
<td>2 839 532</td>
<td>sup</td>
</tr>
<tr>
<td>70%</td>
<td>23 078 024</td>
<td>20 934 242</td>
<td>2 143 782</td>
<td>sup</td>
</tr>
<tr>
<td>80%</td>
<td>34 051 321</td>
<td>32 945 278</td>
<td>1 106 043</td>
<td>sup</td>
</tr>
<tr>
<td>90%</td>
<td>56 027 804</td>
<td>56 348 981</td>
<td>- 321 177</td>
<td>inf</td>
</tr>
<tr>
<td>95%</td>
<td>81 115 383</td>
<td>86 416 166</td>
<td>- 5 300 783</td>
<td>inf</td>
</tr>
<tr>
<td>99%</td>
<td>163 428 646</td>
<td>187 577 365</td>
<td>- 24 148 719</td>
<td>inf</td>
</tr>
<tr>
<td>99.5%</td>
<td>232 523 358</td>
<td>245 960 235</td>
<td>- 13 436 877</td>
<td>inf</td>
</tr>
</tbody>
</table>

Despite the fact that the extreme distribution tail is better approximated than while using the Negative Binomial assumption for the portfolio’s frequency distribution, the global approximation quality is deteriorated (according to the retained measure):

$$\omega = 0.58$$

(to be compared to $\omega = 0.42$ with the Negative Binomial assumption).
7.4 Poisson-Weibull mixed variable: real life example with credit risk portfolio

We continue studying the same real life example with a different assumption for the form of the portfolio’s frequency distribution.

The 1st two moments of the structure variable in the Poisson mixture was already calculated in the previous section:

\[ E[\Lambda] = E[N] = 25.27 \]
\[ Var[\Lambda] = Var[N] - E[N] = 1294.03 \]

The next step is to simulate a numerical sample of the structure variable for which we assume a Weibull form: \( \Lambda \sim \text{Weibull}(E(\Lambda), Var(\Lambda)) \)

Usually the characterization of a Weibull variable in a simulation software (in this particular case we use “R” software) is done using the following two parameters:

\[ \alpha > 0 \quad \text{and} \quad \beta > 0, \quad \text{such as} \]

The Weibull variable CDF is \( F_{\alpha}(\lambda) = 1 - e^{-\lambda^{\beta}} \)

In order to derive these parameters which are necessary for the Weibull variable simulation we will use the following system of moment equations which is solved numerically for \( \alpha \) and \( \beta \):

\[
\begin{align*}
E(\Lambda) &= \beta \Gamma(1 + 1/\alpha) \\
Var(\Lambda) &= \beta^2 \left( \Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha) \right)
\end{align*}
\]

In our case \( \alpha = 0.7167 \) and \( \beta = 20.41 \)

The next step is to simulate a numerical sample of the structure variable:

\( \Lambda \sim \text{Weibull}(\alpha, \beta) \)

Then for each realisation of the structure variable \( \Lambda = \lambda \), we will simulate another realisation of the Poisson variable with a fixed parameter equal to \( \lambda \).

The resulting mixed distribution will be a discrete Poisson-Weibull mixed distribution. Here is the comparison between this new Poisson mixture and the previously used Negative Binomial assumption:
We can observe a significant discrepancy in the distribution tails above the 99.3% probability, the Poisson-Weibull mixed distribution having a heavier tail than the Negative binomial distribution.

This is also confirmed by the following plot:

The next step will be to compare the collective model result using the Poisson-Weibull assumption for the frequency distribution with the result using the individual loss model:
As opposed to the previous example of the Poisson-Lognormal mixed variable, the fact that the extreme distribution tail is better approximated here than while using the Negative Binomial assumption for the portfolio’s frequency distribution results in better global approximation of the individual model by the collective model (according to the retained measure):

\[ \omega = 0.22 \]

(to be compared to \( \omega = 0.42 \) with the Negative Binomial assumption).
8 Portfolio total annual loss variance calculation generalized formula

In this part of the document we will generalize the use of the closed formula for the portfolio loss variance calculation.

8.1 Motivation

Even if the numerical simulation methods are often used in order to estimate the form of the total annual loss distribution in a risk portfolio, including its moments, the exact knowledge of the 1st and of the 2nd moments of this distribution can be very useful. It is helpful not only for the simulation convergence monitoring (in order to decide whether the number of simulations is sufficient) but also for the collective model calibration itself. Indeed, we can know the exact measure of the total annual loss volatility even before simulating anything, just from the assumed individual loss model parameters.

Also, as we have underlined in the previous sections, the quality of the individual model approximation can be enhanced in some cases. Nevertheless, all the testing and the resulting conclusions that we made there are based on the fact that we could perform the $S$ variable calculation using both collective and individual model. In most of real life situations, we can only perform the collective model calculations. Therefore, the exact knowledge of the variance of the portfolio total loss $S$ from the individual loss model can be also very helpful.

As we have already seen from the above description of the collective frequency/severity model, the following 2 well known formulas can be used for the calculation of the 1st and the 2nd moments of the total annual loss in a risk portfolio.

**Expected annual loss calculation formula** :

$$
E[S] = E\left[ \sum_{i=1}^{n} S_i \right] = \sum_{i=1}^{n} \left( E[N_i] \times E[X_i] \right) = \sum_{i=1}^{n} \left( m_i \times \mu_i \times \lambda_i \right) \tag{19}
$$

where for each “$i$-th” risk in the portfolio

- Individual risk exposure (TSI or PML) $m_i$
- Expected loss severity (% of exposure) $\mu_i$
- Severity standard deviation (% of exposure) $\sigma_i$
- Expected annual number of losses $\lambda_i$
- Individual loss severity given a loss occurs $X_i$
- Total annual number of losses $N_i$
- Total number exposures in the portfolio $n$
Variance of the annual loss calculation formula:

\[ \text{Var}[S] = E[N] \times \text{Var}[X] + \text{Var}[N] \times (E[X])^2 \]  \hspace{1cm} (20)

where

- Individual loss severity given a loss occurs \( X \)
- Total annual number of losses \( N \)

It was previously shown in this paper that the terms in the formula (20) can be calculated from the parameters of the individual loss model using the equations (4), (5), (6), (7) and (10).

The formulas (19) and (20) are very handy and straightforward to calculate from the individual loss model parameters. However, the necessary prerequisites for their use are very different. Let us have a closer look at the conditions that must be satisfied in order to be “authorized” to use these formulas for the moments calculation of the total annual loss.

In fact, in order to be able to correctly apply the variance calculation formula (20) we need the following conditions satisfied:

- \( X \perp N \) in all cases (which also requires \( X_i \) variables are all equidistributed for all “i”)
- \( X_i \) variables are independent for all “i”

Some important comments:

1. The above conditions tell nothing about the frequency \( N_i \) distributions and the correlation between them. In the other words, in this model the loss events can be either independent or correlated. Also the variables \( N_i \) can be either equidistributed for all “i” or not. In all cases relative to the loss frequency the variance formula (20) holds.

2. The condition of the independence \( X \perp N \) is stronger than the condition of independence between \( X_i \) for all “i”. The reason for this is the fact that even if the \( X_i \) are all independent, we can still have the \( X \perp N \) unsatisfied when \( X_i \) are not all identically distributed given a loss occurs. On the other hand, if \( X \perp N \) is satisfied, then it is necessary that \( X_i \) are independent for all “i”.

The 1st moment calculation formula (19) has fewer constraints than the 2nd moment calculation formula. Indeed, we don’t need that \( X \perp N \) in the risk portfolio as a whole, and we don’t need that \( X_i \) are all iid. The expected value of a sum of random variables always equals the sum of the expected values of these variables. The only condition that must be nevertheless satisfied is the following:

- \( X_i \perp N_i \) for all “i”, which means that for each risk (or policy) within the risk portfolio the individual loss severity given a loss occurs must stay independent from the annual number of losses.
This is a much easier requirement to meet in a loss model that the necessary conditions for the use of the total annual loss variance formula. It is always satisfied in the Bernoullian case that was previously described in the credit/surety insurance portfolio. The obvious reason for this is the fact that only one loss is possible in case of loss event occurrence in a given risk (or policy). Therefore, \( X_i \perp N_i \) is always true if \( N_i \sim \text{Bernoulli}(\lambda_i) \).

However, this condition is not always satisfied if more that one loss per year is possible for one given risk in the portfolio. For example if \( N_i \sim \text{Poisson}(\lambda_i) \). This particular case is the only one where the initial formula (19) needs generalization (and this will be specifically discussed later in this document). But in most of the cases the formula (19) will work.

As far as the variance formula (20) is concerned, it needs generalization in every case when the required conditions are not satisfied. That is why we believe it is useful to generalize it by gradually relaxing all the required conditions listed above.

### 8.2 Annual loss variance calculation generalized formula in case of frequency correlation

In this section we will relax the first condition required for the validity of the variance calculation formula (20): \( X \perp N \) condition is not satisfied. This means that, according to our definition, the risk portfolio is not homogeneous.

However, we will still keep the second condition, assuming that \( X_i \) variables are independent for all “\( i \)”.

According to our portfolio stratification method, a non-homogeneous risk portfolio can be transformed into a set of homogeneous sub-portfolios in some cases. Then the total loss variance can be expressed as the sum of all variances and covariances of annual losses from the sub-portfolios.

Please note that such a stratification is always possible, since in most of cases a single policy can be ultimately considered as a smallest “homogeneous portfolio” on its own, if the condition \( X_i \perp N_i \) is satisfied for each separate risk (or policy). If we break the portfolio down to the policy detail (maximum possible granularity for modelling), then the collective model becomes individual model. But still, in such a case the above explained approach to the portfolio loss variance calculation works. Here is the key idea that allows us to calculate the total annual portfolio loss variance, even though the “frequency-severity independence” is not satisfied in the portfolio as a whole. The particular case when \( X_i \perp N_i \) is not satisfied either will be examined later.

We will use the formula used in the previous sections and adapted here to the case of the maximal model granularity, each portfolio “part” representing a single risk (or policy):

\[
\text{Var}(S) = \sum_{i=1}^{n} \text{Var}(S_i) + 2 \sum_{i<j} \text{Cov}(S_i, S_j) \tag{21}
\]
where \( Var(S) \) is calculated as the sum of the elements of the variance-covariance matrix of the annual losses from each and every risk (or policy) in the portfolio.

We use the following equations in order to calculate the variances and the covariances in the above formula (21):

\[
\text{Cov}(S_i, S_j) = E(X_i) \times E(X_j) \times \text{Cov}(N_i, N_j) \quad (22)
\]
\[
\text{Var}[S_i] = E[N_i] \times \text{Var}[X_i] + \text{Var}[N_i] \times (E[X_i])^2 \quad (23)
\]

where

\[
E[N_i] = \lambda_i
\]
\[
\text{Var}[N_i] = \lambda_i(1 - \lambda_i), \text{ if } N_i \sim \text{Bernoulli}(\lambda_i)
\]
\[
\text{Var}[N_i] = \lambda_i \text{ if } N_i \sim \text{Poisson}(\lambda_i)
\]
\[
\text{Cov}(N_i, N_j) = \rho_{i,j} \times \sqrt{\lambda_i \lambda_j (1 - \lambda_i)(1 - \lambda_j)} \text{ if } N_i \sim \text{Bernoulli}(\lambda_i)
\]
\[
\text{Cov}(N_i, N_j) = \rho_{i,j} \times \sqrt{\lambda_i \lambda_j} \text{ if } N_i \sim \text{Poisson}(\lambda_i)
\]

\( E(X_i) \) and \( \text{Var}(X_i) \) are given in the parameters of the individual loss model and depend on the form of severity distribution. In our case:

\[
E(X_i) = m_i \mu_i
\]
\[
\text{Var}(X_i) = m_i^2 \sigma_i^2
\]

Then in general case (for any form of frequency and severity distribution) the total annual loss variance calculation formula in case of frequency correlation will be as follows:

\[
\text{Var}(S) = \sum_{i=1}^{n} \left( E(N_i) \text{Var}(X_i) + \text{Var}(N_i) E(X_i)^2 \right) + 2 \sum_{i<j} \left( E(X_i) E(X_j) \sqrt{\text{Var}(N_i) \text{Var}(N_j)} \times \rho_{i,j} \right) \quad (24)
\]

where \((\rho_{i,j})_{1\leq i,j \leq n}\) are the frequency correlation matrix elements, given in the individual loss model parameters.

In a particular case where \( N_i \sim \text{Poisson}(\lambda_i) \) we have:

\[
\text{Var}(S) = \sum_{i=1}^{n} \left( \lambda_i m_i^2 (\mu_i^2 + \sigma_i^2) \right) + 2 \sum_{i<j} \left( m_i m_j \mu_i \mu_j \sqrt{\lambda_i \lambda_j} \right)
\]
In a particular case where \( N_i \sim \text{Bernoulli}(\lambda_i) \), i.e. only a single loss is possible for each and every policy within the risk portfolio, we have:

\[
\text{Var}(S) = \sum_{i=1}^{n} \left( \lambda_i m_i^2 (\mu_i^2 - \mu_i^2 \lambda_i + \sigma_i^2) \right) + 2 \sum_{i<j} \left( m_i m_j \mu_i \mu_j \rho_{ij} \sqrt{\lambda_i \lambda_j (1-\lambda_i)(1-\lambda_j)} \right)
\]

This method requires the following assumptions:

- \( X_i \) are independent for all “i”
- \( N_i \perp X_i \) for all “i”

It is important to note at this stage that the variance calculation formula (24) relaxes the initial condition \( X \perp N \), which means that it is no longer required for the loss severity in the portfolio as a whole to be independent from the loss frequency. For example, the individual loss policy severities \( X_i \) can have here any arbitrary distribution and not equidistributed for any “i” and the formula (24) still holds.

8.3 Annual loss variance calculation generalized formula in case of both frequency and severity correlation

The above described approach can be further generalized to the cases of both frequency and severity dependent risks.

Let us remind that the “classical” variance calculation formula (20) needs the following conditions to be satisfied:

- \( X \perp N \) in all cases (which also requires \( X_i \) variables are all equidistributed for all “i”)
- \( X_i \) variables are independent for all “i”

In the previous section with the variance calculation generalized formula (24) the first condition \( X \perp N \) was relaxed, while the second condition (\( X_i \) independent for all “i”) was kept.

We can now assume that besides the frequency correlation within the risk portfolio there is also a severity correlation. This means that \( X_i \) random variables are not only arbitrarily distributed for different “i” but also correlated. This is very realistic if we take an example of a fire portfolio hit by a windstorm or any other kind of event likely to impact many risks of different size simultaneously.

But this assumption implies that we have got an additional parameter in the individual loss model: the severity correlation matrix indicating the correlation level between different \( X_i \) random variables.

Let us take the following notation for this matrix in order to distinguish it from the frequency correlation matrix:
Then the frequency correlation matrix will be denoted as follows:

\[
(\rho^X_{ij})_{1\leq i \leq n, 1 \leq j \leq n} = \begin{pmatrix} \rho^X_{11} & \ldots & \rho^X_{in} \\ \ldots & \rho^X_{ij} & \ldots \\ \rho^X_{n1} & \ldots & \rho^X_{nn} \end{pmatrix}
\]

In case of frequency and severity correlation the annual losses from different policies within the risk portfolio are obviously correlated. Therefore, we will use the formula (21), where \( \text{Var}(S) \) is calculated as the sum of the elements of the variance-covariance matrix of the annual losses from each and every risk (or policy) in the portfolio.

The variances in the formula (21) are calculated in the same way as we have done in the previous section, since we continue assuming that for each policy \( X_i \perp N_i \) condition is satisfied.

On the other hand, the covariance calculation for annual losses from different policies becomes more complex, since these covariances of the total annual loss now depend on both frequency and severity covariances.

Here is the formula that we suggest for this covariance calculation:

\[
\text{Cov}(S_i, S_j) = E(X_i)E(X_j)\text{Cov}(N_i, N_j) + E(N_i)E(N_j)\text{Cov}(X_i, X_j) + \text{Cov}(N_i, N_j)\text{Cov}(X_i, X_j) \tag{25}
\]
Proof:

Let us take policy number 1 and policy number 2 for the formula derivation. This can be easily generalized to \( (n \times n) \) covariance matrix. In order to ease the notation, we will use \( i \) and \( j \) indexes in order to note the annual number of losses occurring in the 1st policy and in the 2nd policy. Thus the frequencies can be of any form: Bernoulli, Poisson, etc.

\[
Cov(S_1, S_2) = Cov(\sum_{i=1}^{N_1} X_{1i}, \sum_{j=1}^{N_2} X_{2j}) = E[\sum_{i=1}^{N_1} X_{1i} \times \sum_{j=1}^{N_2} X_{2j}] - E[\sum_{i=1}^{N_1} X_{1i}] \times E[\sum_{j=1}^{N_2} X_{2j}]
\]

\[
E[\sum_{i=1}^{N_1} X_{1i} \times \sum_{j=1}^{N_2} X_{2j}] = E[E[\sum_{i=1}^{N_1} X_{1i} \times \sum_{j=1}^{N_2} X_{2j} / N_1, N_2]] = E[E[N_1 N_2 X_1 X_2 / N_1, N_2]] =
\]

\[
= E[N_1 N_2 E[X_1 X_2 / N_1, N_2]] = E[N_1 N_2 E[X_1 X_2]] = E[X_1 X_2] E[N_1 N_2]
\]

then

\[
Cov(S_1, S_2) = E[X_1 X_2] E[N_1 N_2] - E[\sum_{i=1}^{N_1} X_{1i}] \times E[\sum_{j=1}^{N_2} X_{2j}]
\]

\[
= E[X_1 X_2] E[N_1 N_2] - E[N_1] E[X_1] E[N_2] E[X_2]
\]

\[
= (Cov(X_1 X_2) + E[X_1] E[X_2]) \times (Cov(N_1 N_2) + E[N_1] E[N_2]) - E[N_1] E[X_1] E[N_2] E[X_2]
\]

\[
= E[X_1] E(X_2) Cov(N_1, N_2) + E(N_1) E(N_2) Cov(X_1, X_2) + Cov(N_1, N_2) Cov(X_1, X_2)
\]

The above proof requires the following conditions:

- \( X_1 \perp N_1 \)
- \( X_2 \perp N_2 \)

Then in general case (for any form of frequency and severity distribution) the total annual loss variance calculation formula in case of frequency and severity correlation will be as follows:

\[
Var(S) = \sum_{i=1}^{n} (E(N_i) Var(X_i) + Var(N_i) E(X_i)^2) +
\]

\[
+ 2 \sum_{i<j} \left( E(X_i) E(X_j) \sqrt{Var(N_i) Var(N_j)} \times \rho_{ij}^N \right) +
\]

\[
+ 2 \sum_{i<j} \left( E(N_i) E(N_j) \sqrt{Var(X_i) Var(X_j)} \times \rho_{ij}^X \right) +
\]

\[
+ 2 \sum_{i<j} \left( \sqrt{Var(N_i) Var(N_j) Var(X_i) Var(X_j)} \times \rho_{ij}^N \rho_{ij}^X \right) \quad (26)
\]
In a particular case where \( N_i \sim \text{Poisson}(\lambda_i) \) we have:

\[
\begin{align*}
\text{Var}(S) = & \sum_{i=1}^{n} \left( \lambda_i m_i^2 (\mu_i^2 + \sigma_i^2) \right) + 2\sum_{i<j} \left( m_i m_j \mu_i \mu_j \rho_{ij}^N \sqrt{\lambda_i \lambda_j} \right) + \\
& 2\sum_{i<j} \left( \lambda_i \lambda_j \rho_{ij}^X \sqrt{m_i^2 m_j^2 \sigma_i^2 \sigma_j^2} \right) + 2\sum_{i<j} \left( \rho_{ij}^X \rho_{ij}^N \sqrt{\lambda_i \lambda_j m_i^2 m_j^2 \sigma_i^2 \sigma_j^2} \right)
\end{align*}
\]

In a particular case where \( N_i \sim \text{Bernoulli}(\lambda_i) \), i.e. only a single loss is possible for each and every policy within the risk portfolio, we have:

\[
\begin{align*}
\text{Var}(S) = & \sum_{i=1}^{n} \left( \lambda_i m_i^2 (\mu_i^2 - \mu_i^2 \lambda_i + \sigma_i^2) \right) + 2\sum_{i<j} \left( m_i m_j \mu_i \mu_j \rho_{ij}^N \sqrt{\lambda_i \lambda_j (1-\lambda_i)(1-\lambda_j)} \right) + \\
& 2\sum_{i<j} \left( \lambda_i \lambda_j \rho_{ij}^X \sqrt{m_i^2 m_j^2 \sigma_i^2 \sigma_j^2} \right) + 2\sum_{i<j} \left( \rho_{ij}^X \rho_{ij}^N \sqrt{\lambda_i (1-\lambda_i) \lambda_j (1-\lambda_j) m_i^2 m_j^2 \sigma_i^2 \sigma_j^2} \right)
\end{align*}
\]

The only required condition for the validity of the above formulas is \( X_i \perp N_i \) for all \( i \).

It doesn’t need the \( X_i \) to be either independent or equidistributed. It doesn’t need any further requirement regarding the \( N_i \) distributions: they can also be arbitrarily distributed and correlated. All that is needed is the independence between frequency and severity inside each policy taken separately from the others.

### 8.4 Risk portfolio homogeneity and the annual loss variance calculation

In the previous sections we have discussed various methods that could be potentially used in order to build a collective risk model from the parameters of an individual risk model and that is easy to sample using widely used numerical techniques (like Monte Carlo sampling).

From the above given definition of an “homogeneous” risk portfolio we can easily conclude that in any real life example the studied risk portfolio will certainly not be 100% homogeneous and probably will not satisfy the necessary homogeneity conditions listed above. As a matter of fact, in almost all insurance portfolios the \( X_i \) distributions are not going to be identically distributed for all \( i \) : different risks can of course have different sums insured and thus different LGD distributions. Therefore, a collective risk model is only going to be an approximation of the individual loss model.

Then how “acceptable” this approximation could be?

Of course, it is difficult to answer this question since in the majority of real life situations we have no access to the individual model calculations and the only volatility measure that can be easily compared between the collective and the individual model is the total annual loss variance. As we have seen from the previous section, the portfolio's total annual loss variance can be always directly calculated from the individual model parameters that we have at our disposal.
Here are some propositions that can potentially be helpful in this analysis.

**Notation**

For the following propositions we will slightly simplify the notation in order to ease the formula reading in this section of the document:

- Mean annual number of losses in the “i-th” policy: $E(N_i) = \lambda_i$
- Annual number of losses standard deviation in the “i-th” policy: $\sqrt{Var(N_i)} = \tau_i$
- Mean loss severity in the “i-th” policy: $E(X_i) = \mu_i$
- Loss severity standard deviation in the “i-th” policy: $\sqrt{Var(X_i)} = \sigma_i$

The rest of the notation remains unchanged.

**Proposition 1**

If all individual policy’s annual loss amounts are independent within the risk portfolio $S_i \perp S_j$ for all $i, j \in (1, n)$

and the frequency/severity independence condition is satisfied for each separate policy $X_i \perp N_i$ for all $i \in (1, n)$

and all loss frequencies are poissonian $N_i \sim \text{Poisson}(\lambda_i)$ for all $i \in (1, n)$

Then

The risk portfolio annual loss variance will be the same within the collective model and the individual model framework, even if the individual loss severities $X_i$ are not equidistributed (i.e. the risk portfolio is not homogeneous).

**Proof**

The annual total loss calculation formula $Var[S] = E[N] \times Var[X] + Var[N] \times (E[X])^2$ designed within the collective model framework only works if the independence condition $X \perp N$ is satisfied in all cases, i.e. it only holds in an homogeneous risk portfolios. We will now calculate and compare the portfolio loss variance using the above collective model formula $Var_{coll}[S]$ and also using the exact individual model formula $Var_{indiv}[S]$ which works without any independence restriction.
If the $X_i$ random variables are not equidistributed, then

$$Var_{\text{ind}}[S] = \sum_{i=1}^{n} Var(S_i) = \sum_{i=1}^{n} \left( E(N_i)Var(X_i) + Var(N_i)E(X_i)^2 \right) = \sum_{i=1}^{n} \lambda_i \sigma_i^2 + \sum_{i=1}^{n} \lambda_i \mu_i^2$$

$$Var_{\text{col}}[S] = E(N)Var(X) + Var(N)E(X)^2 = E(N)E(X^2) + Var(N)E(X)^2 - E(N)E(X)^2$$

Since $E(N_i) = Var(N_i) = \lambda_i$

$$Var_{\text{col}}[S] = E(N)E(X^2) = \sum_{i=1}^{n} \lambda_i \times \frac{\sum_{i=1}^{n} \lambda_i E(X_i^2)}{\sum_{i=1}^{n} \lambda_i} = \sum_{i=1}^{n} \lambda_i (V(X_i) + E(X_i)^2) = \sum_{i=1}^{n} \lambda_i \sigma_i^2 + \sum_{i=1}^{n} \lambda_i \mu_i^2$$

Therefore

$$Var_{\text{ind}}[S] = Var_{\text{col}}[S]$$

Therefore, under the discussed conditions, the portfolio annual loss distribution estimated via the collective model approximation will have the same variance as the one estimated via the exact individual model calculation.

**Proposition 2**

If all individual policy’s annual loss amounts are independent within the risk portfolio $S_i \perp S_j$ for all $i, j \in (1, n)$

and the frequency/severity independence condition is satisfied for each separate policy $X_i \perp N_i$ for all $i \in (1, n)$

and all loss frequencies $N_i$ have arbitrary distribution forms for all $i \in (1, n)$

Then

The risk portfolio annual loss variance will be the same within the collective model and the individual model framework if the individual loss severities $X_i$ are all equidistributed (i.e. the risk portfolio is homogeneous).
Proof

If the \(X_i\) random variables are not equidistributed, then

\[
\begin{align*}
\text{Var}_{\text{ind}}[S] &= \sum_{i=1}^{n} \text{Var}(S_i) = \sum_{i=1}^{n} \left( E(N_i) \text{Var}(X_i) + \text{Var}(N_i) E(X_i)^2 \right) \\
&= \sum_{i=1}^{n} \lambda_i \sigma_i^2 + \sum_{i=1}^{n} \tau_i \mu_i^2
\end{align*}
\]

\[
\begin{align*}
\text{Var}_{\text{col}}[S] &= E(N) \text{Var}(X) + \text{Var}(N) E(X)^2 = E(N) E(X^2) + \text{Var}(N) E(X)^2 - E(N) E(X)^2
\end{align*}
\]

\[
\begin{align*}
\text{Var}_{\text{col}}[S] &= \sum_{i=1}^{n} \lambda_i \times \frac{\sum_{i=1}^{n} \lambda_i (\mu_i^2 + \sigma_i^2)}{\sum_{i=1}^{n} \lambda_i} + \sum_{i=1}^{n} \tau_i \times \frac{\left( \sum_{i=1}^{n} \lambda_i \mu_i \right)^2}{\left( \sum_{i=1}^{n} \lambda_i \right)^2} - \sum_{i=1}^{n} \lambda_i \times \frac{\left( \sum_{i=1}^{n} \lambda_i \mu_i \right)^2}{\left( \sum_{i=1}^{n} \lambda_i \right)^2}
\end{align*}
\]

\[
\begin{align*}
\text{Var}_{\text{col}}[S] &= \sum_{i=1}^{n} \lambda_i \mu_i^2 + \sum_{i=1}^{n} \lambda_i \sigma_i^2 + \frac{\left( \sum_{i=1}^{n} \lambda_i \mu_i \right)^2 \times \left( \sum_{i=1}^{n} \tau_i - \sum_{i=1}^{n} \lambda_i \right)}{\left( \sum_{i=1}^{n} \lambda_i \right)^2}
\end{align*}
\]

If now we add the additional condition: all \(X_i\) variables are equidistributed, then this means that for all “\(i\)” we have \(\mu_i = \mu\) and \(\sigma_i = \sigma\).

Then,

\[
\begin{align*}
\text{Var}_{\text{ind}}[S] &= \sigma^2 \sum_{i=1}^{n} \lambda_i + \mu^2 \sum_{i=1}^{n} \tau_i \\
\text{Var}_{\text{col}}[S] &= \mu^2 \sum_{i=1}^{n} \lambda_i + \sigma^2 \sum_{i=1}^{n} \lambda_i + \frac{\mu^2 \left( \sum_{i=1}^{n} \lambda_i \right)^2 \times \left( \sum_{i=1}^{n} \tau_i - \sum_{i=1}^{n} \lambda_i \right)}{\left( \sum_{i=1}^{n} \lambda_i \right)^2} = \\
\text{Var}_{\text{col}}[S] &= \mu^2 \sum_{i=1}^{n} \lambda_i + \sigma^2 \sum_{i=1}^{n} \lambda_i + \mu^2 \sum_{i=1}^{n} \tau_i - \mu^2 \sum_{i=1}^{n} \lambda_i = \sigma^2 \sum_{i=1}^{n} \lambda_i + \mu^2 \sum_{i=1}^{n} \tau_i
\end{align*}
\]
Proposition 3

If all individual policy’s annual frequencies are correlated according to the following correlation matrix:

\[
(\rho_{i,j})_{1 \leq i, j \leq n} = \begin{pmatrix}
\rho_{11} & \cdots & \rho_{1n} \\
\vdots & \ddots & \vdots \\
\rho_{n1} & \cdots & \rho_{nn}
\end{pmatrix},
\]

And all individual policy severities are independent within the risk portfolio:

\[X_i \perp X_j \text{ for all } i, j \in (1, n),\]

and the frequency/severity independence condition is satisfied for each separate policy \[X_i \perp N_i \text{ for all } i \in (1, n)\]

and all loss frequencies \[N_i\] have arbitrary distribution forms for all \[i \in (1, n)\]

Then

The risk portfolio annual loss variance will be the same within the collective model and the individual model framework if the individual loss severities \[X_i\] are all equidistributed (i.e. the risk portfolio is homogeneous).

Proof

If the \(X_i\) random variables are not equidistributed, then

\[
Var_{\text{ind}}[S] = \sum_{i=1}^{n} Var(S_i) + 2 \sum_{i<j} Cov(S_i, S_j)
\]

\[
Var_{\text{ind}}[S] = \sum_{i=1}^{n} \left( E(N_i) Var(X_i) + Var(N_i) E(X_i)^2 \right) + 2 \sum_{i<j} \left( E(X_i) E(X_j) Cov(N_i, N_j) \right)
\]

\[
Var_{\text{ind}}[S] = \sum_{i=1}^{n} \lambda_i \sigma_i^2 + \sum_{i=1}^{n} \tau_i \mu_i^2 + 2 \sum_{i<j} \left( \mu_i \mu_j \rho_{ij} \tau_i \tau_j \right)
\]

\[
Var_{\text{col}}[S] = E(N) Var(X) + Var(N) E(X)^2 = E(N) E(X^2) + Var(N) E(X)^2 - E(N) E(X)^2
\]
\[ V_{\text{coll}}(S) = \sum_{i=1}^{n} \frac{\sum_{i=1}^{n} \lambda_i \left(\mu_i^2 + \sigma_i^2\right)}{\sum_{i=1}^{n} \lambda_i} + \left(\sum_{i=1}^{n} \tau_i + 2 \sum_{i<j} \rho_{ij} \tau_i \tau_j\right) \left(\frac{\left(\sum_{i=1}^{n} \lambda_i \mu_i\right)^2}{\left(\sum_{i=1}^{n} \lambda_i\right)^2} - \frac{n}{\sum_{i=1}^{n} \lambda_i} \left(\frac{\sum_{i=1}^{n} \lambda_i \mu_i}{\sum_{i=1}^{n} \lambda_i}\right)^2 \right) + \left(\sum_{i=1}^{n} \lambda_i \sigma_i^2\right) \left(\sum_{i=1}^{n} \tau_i + 2 \sum_{i<j} \rho_{ij} \tau_i \tau_j\right) \right) \]

\[ V_{\text{col}}(S) = \sum_{i=1}^{n} \lambda_i \mu_i^2 + \sum_{i=1}^{n} \lambda_i \sigma_i^2 \left(\sum_{i=1}^{n} \tau_i - \sum_{i=1}^{n} \lambda_i + 2 \sum_{i<j} \rho_{ij} \tau_i \tau_j\right) \]

If now we add the additional condition: all \( X_i \) variables are equidistributed, then this means that for all \( i \) we have \( \mu_i = \mu \) and \( \sigma_i = \sigma \).

Then,

\[ V_{\text{ind}}(S) = \sigma^2 \sum_{i=1}^{n} \lambda_i + \mu^2 \sum_{i=1}^{n} \tau_i + 2 \mu^2 \sum_{i<j} \rho_{ij} \tau_i \tau_j \]

\[ V_{\text{col}}(S) = \mu^2 \sum_{i=1}^{n} \lambda_i + \sigma^2 \sum_{i=1}^{n} \lambda_i + \mu^2 \sum_{i=1}^{n} \tau_i - \mu^2 \sum_{i=1}^{n} \lambda_i + 2 \mu^2 \sum_{i<j} \rho_{ij} \tau_i \tau_j \]

Since the differences between the \( V_{\text{ind}}(S) \) and \( V_{\text{col}}(S) \) are related to the risk portfolio’s homogeneity, the above results explain why for heterogeneous portfolio we should not expect to get exactly the same total loss variance from the collective model and from the individual model. Other modelling assumptions, such as the choice of the particular form of the total annual frequency distribution, have no impact on the above conclusion, since the 1st and the 2nd moments used for portfolio’s frequency and severity distributions are the same within the collective and the individual models.
9 Conclusion

We can condense our individual credit risk model into a collective frequency/severity model, using the characterization of the random variables $N$ and $X$ according to the above description. This frequency/severity model can be directly sampled as a collective model. In practice, this will significantly reduce the number of sampled random variables per Monte Carlo iteration: instead of sampling tens of thousands (or even millions) of variables per iteration, we will usually sample just tens or hundreds of variables, depending on the $N$ distribution.

However, this model simplification via a single collective model doesn’t necessarily lead to a conservative risk approximation from the “risk adverse” point of view. For risk portfolios with correlated loss frequencies $N_i$, the approximation is essentially due to the assumption made for the portfolio annual frequency $N = \sum N_i$ distribution.

In fact, as we have observed from the real life examples, when the $N$ distribution is approximated in some way (which actually happens in real life situations, where we have used the Negative Binomial approximation), nothing can be said about the comparison of the resulting total annual loss distributions. Indeed, in some cases the total loss distribution tails could be underestimated, in other cases it could be overestimated, since there is no general rule for such a comparison.

This is very different from the “classical” Compound Poisson approximation of the individual model where all risks are independent. In such a case we can usually make a conclusion regarding the stochastic dominance at the 2nd order of the individual loss model, which results in a more conservative risk assessment within the collective risk model.

However, the quality of the collective model approximation could be improved in some cases. Throughout this paper we have explored different possible approaches to this approximation enhancement. The main objective was to reproduce the results of the detailed individual model as closely as possible.

However, we should always bear in mind that the individual model itself is an approximation of the “real” risk and in turn can be sensitive to the particular model’s choice (Gaussian copula used for Monte Carlo simulation, LGD distribution form) and also sensitive to the parameter’s uncertainty (probabilities of default, coefficients of correlation, LGD parameters). But the focus of our exploration here is about how and to what extent this individual risk model could be approximated by a collective one and how the latter could be better defined and parametrized.

As we have seen from the examples in this paper, the collective model is always an approximation of the individual loss model. We have also observed that the degree of such an approximation significantly varies depending on each particular case. The risk portfolios with correlated exposures can be particularly complex, the final result mainly depending on the assumption made for the portfolio’s annual frequency distribution form.

Our first step was to use the Compound Negative Binomial approximation:

$$P(S_{coll} \leq s) = \sum_{i=0}^{\infty} \left( P(\text{NegBin}(E(N),\text{Var}(N)) = i) \times P(X^{\sim i} \leq s) \right),$$
Where

\[ S_{\text{vol}} \] corresponds to the total portfolio’s annual loss calculated from the collective loss model
\[ N \] corresponds to the total portfolio’s annual loss number of losses (loss frequency)
\[ X \] corresponds to the individual loss amount, given a loss occurs (loss severity)
\[ X^* \] corresponds to the i-fold convolution of \( X \).

In the Compound Negative Binomial approximation the \( N \) variable is Negative Binomial distributed with the parameters (expected value and variance) calculated from the individual loss model using the probabilities of default and the frequency correlation matrix. Since the knowledge of the 1st and the 2nd moments of the \( N \) distribution doesn’t allow us to know this distribution completely, an additional assumption is needed. The Negative Binomial assumption if one of these.

Once the results of the Negative Binomial approximation results were compared with the results of the individual loss model, we have explored 3 different approaches to the collective model approximation quality enhancement:

1. **Partial collective model approach**
   The central idea of this approach is to separate the global risk portfolio into two parts: large exposures (higher than a certain threshold) and attritional exposures (lower that a certain threshold). Then the attritional exposures are modeled via a collective loss model, whereas the large exposures continue to be modeled via an individual loss model. The number of items modelled within the individual loss model is therefore significantly reduced. Since most of the exposures that are key drivers of the total portfolio annual loss tail are large exposures, the results of this partial collective model will be significantly closer to the individual model results.

2. **Portfolio stratification**
   The above mentioned partial collective model approach is a particular case of the portfolio stratification. More generally, any kind of the global portfolio split into separate components could be considered as portfolio stratification. The key point will be the choice of the stratification criteria. The explored examples show than the heterogeneity of the risk portfolio itself is not the principal reason of the difference between the individual and the collective models when the risks are correlated. Rather, the portfolio’s annual frequency assumption is the main driver of this discrepancy. Therefore, we concluded that any portfolio stratification into more homogeneous parts (using the size of individual risks as criterion, for example) will not lead to definitive model approximation enhancement. For this approach, as well as for the partial collective model, we have suggested a numerical method for the calculation of the global portfolio annual loss distribution from the annual loss distributions of each and every part of the portfolio.

3. **Alternative portfolio’s annual frequency assumption**
   The initial assumption made for the collective loss model was the Compound Negative Binomial distribution. This assumption implies that the total annual frequency distribution is Negative Binomial. The Negative Binomial distribution is a member of a larger class of distributions called “Poisson mixed distributions”. We have therefore tested other types of
Poisson mixed variables, that still allow us to have a maximum flexibility for the 1st two moments of the resulting mixture. Some of the mixtures, as for example Poisson-Weibull mixed variable could have heavier tails than the Negative Binomial variable (which corresponds to Poisson-Gamma mixture) and also give a better approximation of the individual loss model.

Among the different approximation solutions explored in this paper, the partial collective model and also the Poisson Weibull mixed frequency distribution appeared to be the most efficient.

We have also seen that in cases when the total annual loss can be modeled as a single random variable, the lognormal approximation could give a result very close to the one of the individual loss model.

We think that depending of the final goal (modelling of individual losses or modelling of the total annual loss as a whole) we can use different approaches, which are easily applicable in practice, in order to build a collective risk model in situations where the individual model is extremely difficult if even not possible to simulate.

In many situations a credit insurance company which has built and fully parametrized the individual loss model wants to make an assessment of the total annual loss of their portfolio or any other output relative to their reinsurance program and its impact: total annual loss distribution ceded to reinsurers, impact of the reinsurance structure on the Value at Risk, pricing of reinsurance, etc. We have seen that in some cases such a calculation within the individual loss model remains extremely time consuming or even sometimes not feasible in practice, given the large size of the risk portfolio. In such a cases, the methods explored in this paper will allow a decision maker of risk manager to get a set of practical tools for collective model approximations of different random variables involved.
10 Appendix

10.1 General methodology for creating a valid correlation matrix

According to R. Rebonato and P. Jackel 1999 paper:

Given the right hand side eigensystem \( S \) of the real and symmetric matrix \( C \) and the associated set of eigenvalues \( \lambda_i \):

\[
C \cdot S = \Lambda \cdot S \text{ with } \Lambda = \text{diag}(\lambda_i)
\]

We define the non-zero elements of the diagonal matrix \( \Lambda' \) as follows:

\[
\Lambda': \lambda_i' = \begin{cases} 
\lambda_i & : \lambda_i \geq 0 \\
0 & : \lambda_i < 0
\end{cases}
\]

Next step is to define the non-zero elements of the diagonal scaling matrix \( T \) as follows:

\[
T: t_i = \left[ \sum_m s_{im}^2 \right]^{-1}
\]

Let \( B' := S \sqrt{\Lambda'} \)

And \( B := \sqrt{T} B' = \sqrt{T} S \sqrt{\Lambda'} \)

The result of this operation will be definition of \( \hat{C} \) matrix as follows:

\[
\hat{C} = BB^T
\]
10.2 Choleski decomposition algorithm

According to S.Wang paper “Aggregation of Correlated Risk Portfolios : Models & Algorithms”:

Let \( \Sigma \) be a positive definite correlation matrix of the multivariate normal distribution with standard normal marginals.

\[ \Sigma := \{ \rho_{ij} \} \]

From the correlation matrix \( \Sigma \) we want to construct a lower triangular matrix \( B \) such as:

\[ \Sigma = BB^T \]

The elements of the matrix \( B \) can be calculated from the following Choleski algorithm:

\[
b_{ij} = \frac{\rho_{ij} - \sum_{s=1}^{i-1} b_{is} b_{js}}{\sqrt{1 - \sum_{s=1}^{j-1} b_{js}^2}} \quad \text{for} \quad 1 \leq j \leq i \leq n
\]

For \( i > j \) the denominator of the above equation is equal to \( b_{ji} \)

The elements of \( B \) should be calculated from top to bottom and from left to right.
10.3 Rank correlation method

The rank correlation method allows to numerically simulate a sum of different insurance portfolios or sub-portfolios which are correlated (in case when an insurance portfolio was stratified into different parts or components).

The typical situation for using this method is the following. We have at our disposal the numerically simulated losses from different parts of a risk portfolio. For example, we have performed MonteCarlo simulation for large exposures and attritional exposures separately. In order to obtain a numerical simulation of the risk portfolio as a whole we need to put the two separately simulated components (i.e. large and attritional exposures) together. In order to perform this operation we will use the gaussian copula via the so-called “rank correlation” method.

The necessary steps are the following:

- Step 1
  Simulate each portfolio part $S_i$ (or segment) numerically.
  This can be either a classical “frequency/severity” collective model or all losses modelled as a single random variable.

- Step 2
  Calculate the relevant correlation matrix for the random variables corresponding to the total annual loss from each part of the portfolio:
  \[
  \rho(S_i, S_j) = \frac{\text{Cov}(S_i, S_j)}{\sqrt{\text{Var}(S_i)\text{Var}(S_j)}}
  \]

- Step 3
  Generate a set of uniform random variables samples between 0 and 1 via a gaussian copula according to the detailed explanation of this method given in the chapter on the individual loss model. However, the significant difference with the Bernouilly vector case is the number of dimensions: here the number of dimensions corresponds to the number of portfolio’s parts simulated separately via different collective models, for example we will deal with a 2-dimensional $S_i$ vector in case if large exposures and attritional exposures are modelled separately. The exact size of the uniform samples corresponds to the size of the $S_i$ samples. The copula itself is calibrated numerically.

- Step 4
  Calculate the numerical CDFs for each of the $S_i$ variables.
  The CDF probabilities are going to be considered as “ranks” for the numerical samples of $S_i$.

- Step 5
  Rearrange the samples of $S_i$ variables using the “ranking” and according to the sample of the correlated uniform variables generated via the calibrated gaussian copula. Add
together the samples $S_i$ according to the new arrangement. This last step is performed “simulation-wise”, i.e. sim by sim in the uniform vector numerical sample.

The resulting sample will be the sum $S = \sum_i S_i$, where all $S_i$ are correlated according to $\rho(S_i, S_j)$. 
11 Bibliography

1. AGGREGATION OF CORRELATED RISK PORTFOLIOS: MODELS AND ALGORITHMS  
   Shaun S. Wang, ph.d.

2. POLITICAL RISK REINSURANCE PRICING – A CAPITAL MARKET APPROACH  
   Athula Alwis, Vladimir Kremerman, Yakov Lantsman, Jason Harger and Junning Shi

3. ON THE AGGREGATION OF MARKET AND CREDIT RISKS  
   ISMA papers in Finance, University of Reading UK, Carol Alexander and Jaques Pézier

4. AGGREGATION OF RISKS AND ALLOCATION OF CAPITAL  
   Milliman

5. BETWEEN INDIVIDUAL AND COLLECTIVE MODEL FOR THE TOTAL CLAIMS  
   R. Kaas University of Amsterdam, A. E. Van Heerwaarden University of Amsterdam  
   and M. J. Goovaerts K. U. Leuven and University of Amsterdam

6. COPULES ET DEPENDANCES : APPLICATION PRATIQUE A LA DETERMINATION DU BESOIN EN FONDS

7. PROPRES D’UN ASSUREUR NON VIE  
   Mémoire par David Cadoux et Jean-Marc Loizeau

8. THE CALCULATION OF AGGREGATE LOSS DISTRIBUTIONS FROM CLAIM SEVERITY AND CLAIM COUNT DISTRIBUTIONS  
   Philipe Heckman Glenn G. Meyers

9. THE COMMON SHOCK MODEL FOR CORRELATED INSURANCE LOSSES  
   Glenn G. Meyers

10. CORRELATED DEFAULT RISKS  
    Sanjyv R. Das and Laurence Freed

11. ON DEFAULT CORRELATION : A COPULA FUNCTION APPROACH  
    David X. Li (Risk Metrics Group)

12. AN APPROPRIATE WAY TO SWITCH FROM THE INDIVIDUAL RISK MODEL TO THE COLLECTIVE ONE  
    S. Kuon, M. Radtke and A. Reich The Cologne Re, Cologne

13. MIXING COLLECTIVE RISK MODELS - Leigh J. Halliwell, FCAS, MAAA

14. MAXIMUM LIKELIHOOD ESTIMATE OF DEFAULT CORRELATIONS (RISK MAGAZINE PUBLICATION ON MULTI-FACTOR MERTON MODEL)

15. A NOVEL METHODOLOGY FOR CREDIT PORTFOLIO ANALYSIS : NUMERICAL APPROXIMATION APPROACH  
    Yasushi Takano, Jiro Hashiba

16. RISK AGGREGATION AND ECONOMIC CAPITAL (SAS publication)
15. **CREDIT RISK MODELS II : STRUCTURAL MODELS**  
   Abel Elizalde. CEMFI and UPNA

16. **A PRACTICAL CONCEPT OF TAIL CORRELATION**  
   B. John Manistrea FSA, FCIA, MAAA

17. **THE AGGREGATION AND CORRELATION OF INSURANCE EXPOSURE**  
   Glenn G. Meyers, FCAS, MAAA Fredrick L. Klinker, FCAS, MAAA, and David A. Lalonde, FCAS, MAAA, FCIA

18. **THIRTEEN WAYS TO LOOK AT THE CORRELATION COEFFICIENT**  

19. **CORRELATIONS IN MULTI-CREDIT MODELS.**  
   5th Columbia-JAFEE Conference on Mathematics in Finance, 5-6 April 2002 -Pugachevsky, D.

20. **THE MOST GENERAL METHODOLOGY TO CREATE A VALID CORRELATION MATRIX FOR RISK MANAGEMENT AND OPTION PRICING PURPOSES.**  
   Quantitative Research Center of NatWest Group (1999) - Rebonato R. and Jackel P.

21. **THEORIE DU RISQUES ET REASSURANCE ECONOMICA 2006**  
   G.Deelstra, G.Plantin

22. **ASSURANCE NON-VIE, MODELISATION ET SIMULATION ECONOMICA 2005**  
   C.Partrat, JL Besson
# Table of Contents

1 Introduction ................................................................................................................................... 2  
   1.1 Risks portfolio modelling as an objective .............................................................................. 2  
   1.2 Motivation .............................................................................................................................. 3  
   1.3 Novelty of the subject ............................................................................................................ 4  
   1.4 Outline .................................................................................................................................... 5  

2 Individual Risk Model : Credit / surety bond insurance portfolio example ................................... 6  
   2.1 General notation for the individual loss model ................................................................. 6  
   2.2 Building the individual risk model applied to credit insurance portfolio ......................... 7  
      2.2.1 Individual model definition: ...................................................................................... 7  
      2.2.2 Individual model parameters: ................................................................................... 8  
      2.2.3 Individual model simulation .................................................................................... 10  
   2.3 Individual loss model applied to credit insurance portfolio: an example ......................... 21  
      2.3.1 Risk portfolio description ........................................................................................ 21  
      2.3.2 Risk portfolio simulation results ............................................................................. 22  

3 Building the Frequency/Severity collective model ...................................................................... 25  
   3.1 Severity model ..................................................................................................................... 25  
      3.1.1 Example 1: non-homogeneous portfolio with independent exposures .................... 27  
      3.1.2 Example 2: non-homogeneous portfolio with correlated exposures ....................... 34  
   3.2 Frequency model ................................................................................................................. 36  
   3.3 Homogeneous risk portfolio : a definition ......................................................................... 38  
   3.4 Collective loss model applied to credit insurance portfolio : a real life example ............... 38  
   3.5 Collective loss model vs Individual loss model: a general case without portfolio’s  
      frequency approximation ..................................................................................................... 41  

4 Improving the quality of the collective risk model approximation ............................................. 43  
   4.1 Partial collective model approximation ............................................................................. 43  
   4.2 Portfolio stratification ......................................................................................................... 44  
   4.3 Alternative frequency distribution hypothesis .................................................................... 44  
   4.4 Collective model approximation quality measure ............................................................ 45
5 Partial collective model: separating of the portfolio into “large exposures” and “attritional exposures” ..............................................................................................................................................46

5.1 Modelling of the annual loss from attritional exposures as a single random variable........48

5.1.1 Step 1: Estimation of the 1st and the 2nd moments of the annual loss from attritional exposures. ........................................................................................................................................49

5.1.2 Step 2: Estimation of the 1st and the 2nd moments of the continuous part of the annual loss from attritional exposures ........................................................................................................51

5.1.3 Step 3: Assuming a particular form (lognormal for example) for the continuous part of the annual loss from attritional exposures ..........................................................53

5.1.4 Step 4: Calculating the CDF of the variable “annual loss from attritional exposures” .................................................................................................................................53

5.2 Modelling of the total annual loss using the partial collective model.................................54

5.3 Method application and analysis: real life examples with credit insurance portfolio........55

5.3.1 Example of the partial collective model built for 50m threshold ................................55

5.3.2 Example of the partial collective model built for 10m threshold .................................57

5.3.3 Example of the partial collective model built for 3m threshold ....................................58

5.3.4 Conclusion for the partial collective model method .......................................................60

5.4 Modelling of the total annual loss as a single random variable: real life example with credit insurance portfolio ........................................................................................................60

6 Portfolio stratification method ...............................................................................................64

6.1 Building frequency/severity model for each part ..................................................................64

6.2 Calculation of the covariance matrix for the annual losses from different parts of the stratified portfolio ......................................................................................................................65

6.3 Simulation of the global portfolio losses ..................................................................................66

6.4 Method application and analysis: theoretical portfolio example ..........................................67

7 Alternative frequency distribution hypothesis ........................................................................70

7.1 Poisson mixed variables: definition, moment properties and simulation method ..............70

7.2 Poisson-Gamma mixed variable and alternative mixtures ....................................................71

7.3 Poisson-Lognormal mixed variable: real life example with credit risk portfolio ...............72

7.4 Poisson-Weibull mixed variable: real life example with credit risk portfolio .......................75

8 Portfolio total annual loss variance calculation generalized formula ..................................78

8.1 Motivation ................................................................................................................................78

8.2 Annual loss variance calculation generalized formula in case of frequency correlation ..............................................................80

8.3 Annual loss variance calculation generalized formula in case of both frequency and severity correlation .......................................................................................................................82

8.4 Risk portfolio homogeneity and the annual loss variance calculation .................................85