Calculation of Portfolio Loss Distribution Given Default

Yu Zou†‡, Sang-Yeun Shim‡

† PACM and Department of Chemical Engineering, Princeton University, Princeton, NJ 08544
‡The Yield Book, Inc. & Derivative Research, Citigroup Inc., New York, NY 10013

Abstract

Default loss distribution of corporate portfolios plays a crucial role in CDO tranche pricing, tracking error calculation and profit/loss assessment of corporation systems. This work gives an efficient algorithm to calculate the default loss distribution based on Hull-White probability bucketing approach and importance sampling method. The Gaussian copula model is assumed to calculate the conditional default probability for each corporation in a portfolio by the conditional independence framework. The algorithm can significantly reduce the computational time.

1. Introduction

Problem Description

The default loss of a portfolio of \( n \) corporations is defined by

\[
L = \sum_{i=1}^{n} l_i Y_i ,
\]

where \( l_i \) is the loan size of the \( i \)th corporation. \( Y_i \) is the default indicator function which takes value 1 with probability \( q_i \) or 0 with probability \( 1 - q_i \). Here \( q_i \) is the expected default probability of the \( i \)th corporation. In order to calculate the discrete probability of \( L \), one needs to know the joint distribution of all binary \( Y_i \)'s which are not independent. Simply specifying the second-order default correlation \( E(Y_i, Y_j) \) does not suffice this requirement. To determine the joint distribution of all \( Y_i \)'s, the event of \( Y_i \) can be modelled as

\[
\{ Y_i = 1 \} = \{ \Phi(X_i) < q_i \} = \{ X_i < \Phi^{-1}(q_i) \}
\]

where \( \Phi \) is the standard normal cumulative distribution function. The latent variables \( X_i \)'s are assets possessed by each corporation and given by

\[
X_i = \sum_{j=1}^{m} a_{ij} \xi_j + b_i \epsilon_i .
\]

In the above formula, \( \sum_{j=1}^{m} a_{ij}^2 + b_i^2 = 1 \). \( \xi_j \)'s are independent Gaussian random variables and are a set of systematic risk factors that are common to all corporations. \( \epsilon_i \) stands for an idiosyncratic risk factor associated with each corporation. Eqn. (1.2) implies that a corporation will default when its asset \( X_i \) is less than a threshold prescribed by its expected default probability \( q_i \).

By simple mathematical operations we obtain

\[
\{ Y_i = 1 \} = \{ u_i < f(\{ \xi_i \}, q_i) \} ,
\]

where \( u_i \)'s are independent random variables uniformly distributed over \([0,1]\) and
Eqn. (1.5) is a commonly used formula in finance for calculating the credit risk, known as the CreditMetrics (KMV) model (Gordy 2000).

The uniform random variables \( \Phi(X_i) \) in (1.2) constitute a Gaussian copula (Li 2000). By Eqn. (1.2), joint distributions of \( Y_i \)'s can be fully determined relying on joint Gaussian distribution of \( X_i \)'s whose covariance matrix is known if \( a_{ij} \)’s and \( b_i \)'s are given. For example, the probability \( P(Y_i=1, i=1,2,\ldots,n) \) is given by

\[
P(Y_i = 1, i = 1, 2, \ldots, n) = P(X_i < \Phi^{-1}(q_i), i = 1, 2, \ldots, n).
\]

Furthermore, by using Eqn. (1.4) and (1.5), realizations of \( Y_i \)'s can be sampled independently through a conditional independence framework. To do so, independent realizations of the systematic factors \( \xi_j \)'s are sampled first, and a realization of \( Y_i \)'s is sampled according to

\[
P(Y_i = 1 | \{ \xi_j \}) = P(X_i < \Phi^{-1}(q_i) | \{ \xi_j \}) = f(\{ \xi_j \}, q_i) = \Phi((\Phi^{-1}(q_i) - \sum_{j=1}^{m} a_{ij} \xi_j) / b_i).
\]  

The task of this work is to determine the probability distribution of the total loss \( L \). Within the conditional independence framework, the distribution of the conditional loss \( L \) given a realization of \( \xi_j \)'s is calculated first and then integrated over the random domain of \( \xi_j \)'s to obtain the unconditional distribution.

**Comparison of Methods**

Several methods have been proposed in literature to calculate the conditional loss distribution. A Poisson distribution approximation and the FFT were used in Merino 2002. Anderson 2003 proposed a recursive formula to compute the discretized probability of conditional loss. Laurent 2003 also briefly mentioned an FT approach in its discussion on evaluation of default swaps and CDOs. De Prisco 2005 suggested a compound Poisson approximation method. Another recursive approach was later given in Jackson 2007 to improve an algorithm in Hull and White 2004. All these methods cope with corporations whose loan sizes have a common divisor. A different approach that lifts this assumption is the probability bucketing procedure described in Hull and White 2004, which we refer as HW model in this work. This method does not require the loan sizes be located in a common equally-spaced lattice. It can also track the average loss within each bucket. An additional merit of the method is that the loss domain can be truncated at some value beyond which the loss probability may be small enough to be negligible.

Although the above methods can efficiently compute the loss distributions under their respective assumptions, they are not able to effectively calculate the tail distribution. It therefore seems necessary to combine calculation of tail distribution with that in the bulk region. Also based on the conditional independence framework, the saddle-point method (Martin 2001) gives an analytic formula of the tail distribution. This formula, however, is exact only when the true distribution is normal. For distributions far away from normal it may not work well. An alternative numerical procedure, the importance sampling (IS), was recently used to investigate the tail probability of the default loss (Glasserman 2005). This two-step method first applies IS to systematic risk factors and then applies IS to estimate the tail probability conditional on each set of systematic factors. Compared with the saddle-point approach, the IS works properly with wider range of distributions rather than with normal distribution only.

In this work, the probability bucketing procedure and the IS method are combined to calculate the bulk and tail regions of the default loss distribution. Section 2 will describe the theoretical basis of the
algorithm in detail. The third section provides numerical results and convergence analysis. Conclusions are made and future work is discussed in Section 4.

2. Theoretical basis

Determining coefficients in the Gaussian Copula representation

In the conditional independence framework, the coefficients $a_{ij}$ and $b_i$ in (1.3) have to be known a priori in order to generate conditional independent $Y_i$’s. A simple way to calculate the coefficients is to use the asset correlation of corporations, i.e., the correlation between $X_i$’s. Let $X_i$ be a corporation in the $i$th sector of an industry. Assuming the asset correlation of corporations within the same sector is $\rho_1$, the multivariate factor model for corporations in the $i$th sector can be written as

$$X_i = \sqrt{\rho_1} \eta_i + \sqrt{1 - \rho_1} \varepsilon_i ,$$  \hspace{1cm} (2.1)

In the above equation the Gaussian variable $\eta_i$ is the same for all corporations in the sector while the idiosyncratic variable $\varepsilon_i$ is independent between different corporations.

Now suppose the asset correlation of corporations between two distinct sectors, sector $i$ and sector $j$, is $\rho_2$ ($\rho_2 < \rho_1$), simple calculation gives

$$\rho_2 = \rho_1 \eta_i \eta_j ,$$ \hspace{1cm} (2.2)

The covariance matrix $\text{Cov}(\eta, \eta^T)$ can be decomposed by the Singular Value Decomposition. Let

$$\text{Cov}(\eta, \eta^T) = U \Lambda U^T ,$$ \hspace{1cm} (2.3)

where $U$ is orthonormal and $\Lambda$ a diagonal matrix consisting of eigenvalues of $\text{Cov}(\eta, \eta^T)$. The decomposition of $\eta$ follows as

$$\eta = U \Lambda^{1/2} \xi \quad \text{or} \quad \eta_i = \sum_{j=1}^{m} e_{ij} \sqrt{\lambda_j} \xi_j .$$ \hspace{1cm} (2.4)

Here $e_{ij}$ are elements of $U$ and $\lambda_j$ is the $j$th largest eigenvalue of the covariance matrix. $m$ is the number of sectors in the industry.

Therefore, the latent variables can be rewritten as

$$X_i = \sum_{j=1}^{m} \sqrt{\rho_1 \lambda_j} e_{ij} \xi_j + \sqrt{1 - \rho_1} \varepsilon_i .$$ \hspace{1cm} (2.5)

Corporations in sector $i$ share the same $a_{ij}$ and $b_i$

$$a_{ij} = \sqrt{\rho_1 \lambda_j} e_{ij} , \quad b_i = \sqrt{1 - \rho_1} .$$

Calculation of bulk probability based on the Hull-White method

Hull and White 2004 suggested a recursive procedure to calculate the conditional loss distribution given default. This approach updates loss distribution in intervals when the loan size of a corporation is added into the portfolio. Let $0, b_1, b_2, ..., b_M$ be endpoints of intervals $1, 2, ..., M$ in the loss domain. At the beginning of the procedure, probability $1$ is assumed to concentrate at $0$ ($\rho_1 = 1$) and loss probabilities in the rest of intervals are set to $0$ ($\rho_i = 0, i = 2, 3, ..., M$). The conditional average $A_i$ of interval $i$ at this stage is set to the average of the interval’s two endpoints. As the loan size of a corporation is added to the total loss, the loss probability and conditional average in each interval are updated according to the following formulas.
where \( p_i, A_i \) are probability and conditional average of the \( i \)th interval before the updating. \( Q_k = P(Y_k=1) \) and \( l_k \) are respectively the conditional default probability and loan size of the \( k \)th corporation. \( u(i) \) is the interval where \( A_i + l_k \) is located. When \( i = u(i) \) the formulas are

\[
p_i = p_i',
A_i = A_i' + Q_i l_k.
\]  

To ensure that \( p_i \) is not affected before updating on the \( i \)th interval, the above procedure has to be implemented in the descending order of \( i \), i.e., from \( M \) to 1. When \( l_k \) has negative values, the updating should be in the ascending order.

**Importance sampling for tail probability**

This section gives a brief review on the importance sampling method proposed in Glasserman 2005. The conditional tail probability \( P(L>x|\xi) \), after changing its measure, can be written into a new form

\[
P(L>x|\xi) = \hat{E} \left[ I(L>x) \prod_{i=1}^{n} \left( \frac{Q_i}{p_i} \right)^{r_i} \left( \frac{1-Q_i}{1-p_i} \right)^{1-r_i} \right],
\]

where \( p_i \) is the new probability measure associated with the \( i \)th corporation. In this context, \( p_i \) is set to be an exponential twist of the original conditional probability \( Q_i \), i.e.,

\[
p_i = \frac{Q_i e^{\theta_l}}{1+Q_i (e^{\theta_l} - 1)}.
\]

The likelihood ratio in (2.8) can therefore be written as

\[
\prod_{i=1}^{n} \left( \frac{Q_i}{p_i} \right)^{r_i} \left( \frac{1-Q_i}{1-p_i} \right)^{1-r_i} = \exp(-\theta L + \psi(\theta)),
\]

where

\[
\psi(\theta) = \sum_{i=1}^{n} \log(1+Q_i (e^{\theta_l} - 1))
\]

Conditional on \( \xi \), the twisting parameter \( \theta \) is chosen by minimizing the second moment of the estimator in (2.8). It is easy to see that

\[
\hat{E}[I(L>x) e^{-2\theta L + 2\psi(\theta)}] \leq e^{-2\theta x + 2\psi(\theta)}.
\]

Minimizing the second moment directly may be difficult, but it is trivial to minimize its upper bound as shown in (2.11). Since \( \psi \) is strictly convex and passes through the origin, the value of \( \theta \) that minimizes the bound is

\[
\theta_x = \begin{cases} 
\hat{\theta} : \psi'(\hat{\theta}) = x, & x > \psi'(0) \\
0, & x \leq \psi'(0)
\end{cases}
\]

The solution of \( \theta_x \) in the above equation is used to twist the conditional default probability.
Importance sampling can also apply to the systematic factors $\xi$ and find a new measure of $\xi$ that minimizes the variance of the estimator for evaluate $P(L>x)$. Since

$$P(L > x) = \int_{\Omega} P(L > x | \xi) \frac{e^{-\xi^T \xi/2}}{\sqrt{2}} d\xi = \int_{\Omega} \frac{P(L > x | \xi) e^{-\xi^T \xi/2}}{\mu(\xi)} \cdot \mu(\xi) d\xi, \quad (2.13)$$

it is easy to understand that the $\mu$ that zeros the variance of the estimator at the right hand side of (2.13) is

$$\mu(\xi) = P(L > x | \xi) \frac{e^{-\xi^T \xi/2}}{\sqrt{2}} / P(L > x). \quad (2.14)$$

Obtaining an analytical form of $\mu$ in the above equation is difficult because $P(L>x|\xi)$ has to be computed with a formidable time and $P(L>x)$ is what we are solving for. Glasserman 1999 suggested a normal density to serve as a substitute. The new density has the same mode as that in (2.14) which can be determined by maximizing $P(L > x | \xi) e^{-\xi^T \xi/2}$. A more feasible solution for this optimization problem is to maximize its upper bound, i.e., obtain $\xi$ that satisfies

$$\max_{\xi} \left\{ -\theta_x(\xi)x + \psi(\theta_x(\xi), \xi) - \frac{1}{2} \xi^T \xi \right\}.$$ 

The estimator for calculating $P(L>x)$ is therefore

$$I(L > x) e^{-\theta_x(\xi)L + \psi(\theta_x(\xi), \xi)} e^{-\xi^T \xi + \xi^T \xi/2},$$

where $\hat{\xi}$ is the mean (mode) of the new Gaussian measure $\mu$ and $\xi$ is sampled from $\mu$.

The two-step procedure described above can sample the rare events more efficiently with less realizations of $\xi$.

3. Numerical Tests

Several characteristic index portfolios that store a list of corporations with expected default probabilities in the software Yield Book are used to test the method introduced above. The portfolios include high yield index, big credit index and big index. Corporations in these portfolios come from 13 industries consisting of 70 sectors in total. Losses of corporations between different industries are totally independent, which means that their asset correlations are zero. By setting the correlations $\rho_1$ and $\rho_2$ defined above, the coefficients of systematic risk factors are computed through the SVD subroutine dgesvd in lapack. The effective loss domain where loss distributions cannot be negligible is first determined by using Hull-White method with 5000 realizations of systematic factors and 100 intervals. To determine the mean of the new measure $\mu$ in the importance sampling the conjugate gradient method together with line minimization is employed to solve the optimization problem. In order to avoid kinks between the bulk and tail distributions, the cut-off position $x$ for the importance sampling is first set to 60% of the effective domain and then reduced twice by 5% each time. If the ratio of bulk and tail probability at the cut-off point is within a tolerance value of 1.5, then the result of importance sampling is accepted. Otherwise, only the Hull-White result is used to represent the whole distribution. The importance sampling is implemented at only one side of the loss distribution. If the mode of the bulk distribution is in the left (right) half region of the distribution, then the importance sampling will be done at the right (left) side. In both the bulk and tail probability calculations the Sobol quasi Monte-Carlo sequence (Press 2002) is generated to sample the systematic random factors $\xi$ whose dimension is up to 70.
Performance and time
We use 10000 realizations of systematic factors and 400 intervals to implement the calculation. $\rho_1$ and $\rho_2$ are set to 0.5 and 0.3, respectively. Loss distributions modeled by Eqn. (1.1) and (1.2) and calculated from Monte Carlo simulation that directly samples pseudo Gaussian variables $\xi_j$'s, serve as the benchmark for comparison. The number of realizations for the direct MC is $10^7$. The following figures 3.1-3.3 compare the HW-IS result with the benchmark distribution. The comparison shows a good match between the HW-IS and benchmark distribution for the high yield index. The importance sampling fails for the big credit index and big index because the ratio of the bulk and tail probabilities at the cut-off positions is out of the tolerance. Therefore only HW distributions are plotted for these two indices. They agree well with MC distributions in the bulk region. Table 1 compares the CPU time that are required for the computation, which indicates a huge time-saving when the HW-IS method is used.

![Graph showing comparison between HW-IS, MC, and HW for high yield index.](image)

Figure 3.1 – Comparison of portfolio default loss distribution calculated by Hull White probability bucketing procedure with or without importance sampling and direct Monte-Carlo simulation with pseudo random variable sampling. A list of corporations with expected default probability in the high

---

Table 1: Comparison of CPU time required for computation.

<table>
<thead>
<tr>
<th>Index</th>
<th>HW-IS</th>
<th>MC</th>
<th>HW</th>
</tr>
</thead>
<tbody>
<tr>
<td>High yield</td>
<td>1.00E-09</td>
<td>1.00E-08</td>
<td>1.00E-07</td>
</tr>
<tr>
<td>Big credit</td>
<td>1.00E-06</td>
<td>1.00E-05</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>Big index</td>
<td>1.00E-03</td>
<td>1.00E-02</td>
<td>1.00E-01</td>
</tr>
</tbody>
</table>
yield index is used for computation. Top figure: vertical axis in linear scale; Bottom figure: vertical axis in log scale.

Figure 3.2 – Comparison of portfolio default loss distribution calculated by Hull White probability bucketing procedure without importance sampling and direct Monte-Carlo simulation with pseudo random variable sampling. A list of corporations with expected default probability in the big credit index is used for computation. Top figure: vertical axis in linear scale; Bottom figure: vertical axis in log scale.
Figure 3.3 – Comparison of portfolio default loss distribution calculated by Hull White probability bucketing procedure without importance sampling and direct Monte-Carlo simulation with pseudo random variable sampling. A list of corporations with expected default probability in the big index is used for computation. Top figure: vertical axis in linear scale; Bottom figure: vertical axis in log scale.

<table>
<thead>
<tr>
<th></th>
<th>Number of corporations</th>
<th>HW-IS</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>High yield index</td>
<td>636</td>
<td>70 sec</td>
<td>61 min</td>
</tr>
<tr>
<td>Big credit index</td>
<td>757</td>
<td>106 sec</td>
<td>73 min</td>
</tr>
<tr>
<td>Big index</td>
<td>760</td>
<td>112 sec</td>
<td>72 min</td>
</tr>
</tbody>
</table>

Table 1 Comparison of computing time
**Convergence Analysis**

The theoretical unconditional average $\alpha^{\text{theo}}$ and 2nd-order moment of the total loss can be calculated analytically. The measure

$$ e = \sqrt{E \left( \frac{\sum_{i=1}^{N} \alpha(\xi_i, M)}{N} - \alpha^{\text{theo}} \right)^2} $$

(3.1)

is used to calculate the convergence rate of the HW-IS method. In (3.1), $N$ is the number of quasi MC realizations and $M$ is the number of intervals. The average loss conditional on $\xi_i$, $\alpha(\xi_i, M)$, is calculated by assuming a constant probability density in each interval of the effective loss domain. The ensemble average $E(\cdot)$ in (3.1) is performed using 400 realizations.

Figures 3.4-3.6 show the convergence of the measure with respect to $N$ and $M$. The distribution is not sensitive to $M$ when $M$ is beyond 400. However, the effect of $N$ is very significant. It can be seen that for large $M$, $e$ is proportional to $N^\beta$ where $\beta$ can be estimated. Calculation shows that $\beta$ is between -0.6 and -0.7 which is a typical convergence rate for quasi MC simulation (Press 2002).

**Figure 3.4 – Relationship of the measure $e$ with respect to the number of probability intervals $M$ and the number of quasi MC realizations $N$ in the computation of default loss distribution for the high yield index.**
4. Conclusions and Future Work

A method coupling Hull and White probability bucketing procedure and importance sampling is proposed in this work to efficiently compute the default loss distribution of a portfolio of corporations. Comparison with the results from direct Monte-Carlo simulation shows that this method can greatly reduce the computing time for three sample portfolios while retaining desired accuracy. The
convergence rate of this method with respect to the number of quasi MC realizations follows that of the quasi MC simulation.

In addition, the method also employs a multi-factor model to establish connection between the default events of different corporations, which is a simple attempt to model the interdependency of components in a complex financial system. Our future work includes using more realistic and accurate models to characterize interdependency within a complex financial and economic system consisting of multiple interacting components. One interesting approach would be to use agent-based simulation to simulate the dynamics of the default events of these components and their probability conditioned on underlying random system factors. In this manner, not only a more accurate description of the default loss distribution can be obtained but we can also study its dynamical behavior over the time domain.

References