Monte Carlo Methods for Portfolio Credit Risk

Tim J. Brereton
Dirk P. Kroese

School of Mathematics and Physics
The University of Queensland
Australia

Joshua C. Chan

Research School of Economics
The Australian National University
Australia

1 Introduction

The financial crisis of $2007 - 2009$ began with a major failure in credit markets. The causes of this failure stretch far beyond inadequate mathematical modeling (see Donnelly and Embrechts [2010] and Brigo et al. [2009] for detailed discussions from a mathematical finance perspective). Nevertheless, it is clear that some of the more popular models of credit risk were shown to be flawed. Many of these models were and are popular because they are mathematically tractable, allowing easy computation of various risk measures. More realistic (and complex) models come at a significant computational cost, often requiring *Monte Carlo methods* to estimate quantities of interest.

The purpose of this chapter is to survey the Monte Carlo techniques that are used in portfolio credit risk modeling. We discuss various approaches for modeling the dependencies between individual components of a portfolio and focus on two principal risk measures: Value at Risk (VaR) and Expected Shortfall (ES).

The efficient estimation of the credit risk measures is often computationally expensive, as it involves the estimation of small quantiles. Rare-event simulation techniques such as importance sampling can significantly reduce the computational burden, but the choice of a good importance sampling distribution can be a difficult mathematical problem.

Recent simulation techniques such as the cross-entropy method [Rubinstein and Kroese, 2004] have greatly enhanced the applicability of importance sampling techniques by adaptively choosing the importance sampling distribution, based on samples from the original simulation model.

The remainder of this chapter is organized as follows. In Section 2 we describe the general model framework for credit portfolio loss. Section 3 discusses the crude and importance sampling approaches to estimating risk measures via the Monte Carlo method. Various applications to specific models (including Bernoulli mixture models, factor models, copula models and intensity models) are given in Section 4. Many of these models capture empirical features of credit risk, such as default clustering, that are not captured by the standard Gaussian models. Finally, the Appendix contains the essentials on rare-event simulation and adaptive importance sampling.

2 Modeling Credit Portfolio Losses

Portfolio credit risk is usually evaluated in a *static* setting, whereby the loss of a portfolio is modeled via a single random variable L representing the sum of the losses incurred by the individual

components of the portfolio; that is,

$$L = \text{Loss}_1 + \cdots + \text{Loss}_n \ .$$

If the individual losses are independent, the problem of describing the distribution of L reduces to the problem of describing the marginal distribution of each individual loss. However, in practice the individual losses tend to be dependent on each other. It is therefore important to appropriately model the dependence between the $\{\text{Loss}_i\}$.

Losses can result from changes in credit quality as well as from default. For simplicity we will only consider default events. We write each individual loss as the product of the loss incurred if the individual component defaults and a Bernoulli (that is, indicator) random variable that takes the value 1 when a default occurs and 0 otherwise. Thus, our model is given by

$$L = l_1 D_1 + \cdots + l_n D_n \ , \tag{1}$$

where the $\{l_i\}$ are the magnitudes of individual losses and the $\{D_i\}$ are Bernoulli variables modeling the default events. The $\{l_i\}$ can be random or deterministic. The empirical evidence suggests a strong relation between the magnitudes of losses and the number of defaults. However, many popular credit risk models assume independence between the $\{l_i\}$ and $\{D_i\}$. We will focus on modeling only the default events $\{D_i\}$, though some of the models given below can be modified to incorporate dependence between losses and numbers of defaults.

2.1 Risk Measures

The distribution of L — often called the *loss distribution* and denoted as F_L — is the central object of credit risk modeling. F_L is typically not available in closed form. Instead, certain risk measures are used to describe its key features, particularly its tail behavior. The most widely used risk measure in credit risk is *Value at Risk* (VaR), which describes the quantiles of the loss distribution. For example, the 99% VaR of a portfolio is the value of the loss variable L such that a greater loss would only occur 1% of the time. The VaR for confidence level α is given by

$$v_\alpha = F_L^{-1}(\alpha) \ ,$$

where F_L^{-1} is the generalized inverse of F_L:

$$F_L^{-1}(\alpha) = \inf \{l : F_L(l) \geqslant \alpha\} \ . \tag{2}$$

Common values for α are 0.95, 0.99, 0.995 and 0.999. The use of VaR as a risk measure has been the subject of significant criticism (see Bluhm et al. [2010] and McNeil et al. [2005] for discussions). In particular, it has the counter-intuitive feature that it is not sub-additive: the VaR of two portfolios might be larger than the sum of the VaRs of the individual portfolios. In other words, the VaR of a portfolio is not necessarily reduced through diversification. This led Artzner et al. [1999] to propose a class of *coherent risk measures*, which satisfy certain 'natural' requirements, including sub-additivity. One of the most popular of these is the *Expected Shortfall* (ES), also known as *Conditional Value at Risk* (CVaR). The α expected shortfall is given by

$$\text{c}_\alpha = \mathbb{E}\left[L \,|\, L \geqslant v_\alpha\right] .$$

Expected shortfall is also an example of a spectral risk measure, see Bluhm et al. [2010].

2.2 Modeling Dependency

The key challenge in modeling portfolio credit risk lies in describing the relationship between default events. Defaults do not occur independently of one another, but rather tend to cluster. These default clusters could occur as the result of sector specific conditions, such as a downturn in a particular industry or market, or as a result of broader macroeconomic factors. A major failing of credit models in the financial crisis of 2007 – 2009 was that they failed to adequately model the possibility that a large number of defaults could occur simultaneously. In order to discuss this limitation, we need to introduce a number of different *dependency measures* that describe the relationship between random variables.

The simplest measure of dependency between two random variables X and Y is given by their pairwise linear correlation $\rho(X,Y) = \text{Cov}(X,Y)/\sqrt{\text{Var}(X)\text{Var}(Y)}$. Its multivariate analog is the correlation matrix. The dependency structure of random vector $\mathbf{X}$ is completely specified by its correlation matrix if and only if $\mathbf{X}$ has an *elliptical* distribution; see McNeil et al. [2005]. Important special cases are the multivariate normal and multivariate Student-t distributions.

A drawback of linear correlation (and other correlation measures, such as rank correlation) is that it describes the average joint behavior of random variables. In risk management it is *extremal* events, rather than typical events, that are of primary interest. Two dependency measures that describe extremal behavior are the coefficients of upper and lower tail dependence. Specifically, given two random variables X and Y, with distributions F_X and F_Y, we define the coefficient of upper tail dependence as

$$\lambda_u = \lim_{q \uparrow 1} \mathbb{P}\left(Y > F_Y^{-1}(q) \,|\, X > F_X^{-1}(q)\right),$$

and the coefficient of lower tail dependence as

$$\lambda_l = \lim_{q \downarrow 0} \mathbb{P}\left(Y \leqslant F_Y^{-1}(q) \,|\, X \leqslant F_X^{-1}(q)\right).$$

These measures describe the relationship between variables in the tails of distributions. A joint distribution is said to have upper (lower) tail independence if $\lambda_u = 0$ ($\lambda_d = 0$). Some of the most popular models of credit risk — in particular, the various Gaussian copula models — exhibit tail independence in both tails. This is clearly not a desirable feature in risk models, as empirical evidence tends to indicate that both defaults and risk factors tend to become more correlated in extreme settings. With the exception of the canonical Gaussian models, all of the models described in the following sections possess tail dependence.

3 Estimating Risk Measures via Monte Carlo

For a general loss distribution F_L, analytic calculation of the various risk measures described in the last section is usually impossible. Often the only feasible approach is to estimate these risk measures using Monte Carlo methods. To proceed, we need a method for drawing independent and identically distributed (iid) replicates of the random variable L and a method for estimating

risk measures, given an iid sample $L_1, \ldots, L_N$. The methodology for estimating risk measures is largely model independent, and is the focus of this section.

The Monte Carlo estimation of VaR turns out to be somewhat more difficult than the traditional problem of estimating an expectation. In particular, VaR estimators are non-linear functions of the sample. Many classical Monte Carlo methods cannot be applied to VaR estimation or need to be modified to work well. In addition, it is typically difficult to find confidence intervals for VaR estimators.

3.1 Crude Monte Carlo Estimators

The *Crude Monte Carlo* (CMC) estimator of VaR is the quantile estimator of classical statistics; see van der Vaart [1998] for a discussion of its properties in a statistical context. It replaces the unknown distribution function of L, F_L, in the definition of VaR in (2) with the empirical distribution function $\hat{F}_L$. That is, we estimate VaR using

$$\hat{v}_\alpha = \inf\left\{l : \hat{F}_L(l) \geqslant \alpha\right\},$$

where

$$\hat{F}_L(l) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(L_i \leqslant l) \tag{3}$$

is the empirical distribution function of the iid sample $L_1, \ldots, L_N$. Note that $\hat{F}_L$ is a step function. Consequently, the CMC quantile estimator can be easily obtained by ordering the $\{L_i\}$ as $L_{(1)} \leqslant \cdots \leqslant L_{(N)}$ and finding the $\lceil \alpha N \rceil$th largest value.

Algorithm 3.1 (CMC VaR Estimator)

1. *Generate an iid sample* $L_1, \ldots, L_N$.
2. *Order the sample from smallest to largest as* $L_{(1)} \leqslant \cdots \leqslant L_{(N)}$.
3. *Return* $\hat{v}_\alpha = L_{(\lceil \alpha N \rceil)}$.

The CMC estimator for the ES is more straightforward, as the ES is simply an expectation. The estimator is given by

$$\hat{c}_\alpha = \frac{1}{N(1-\alpha)}\sum_{i=1}^{N} L_i\, \mathbb{I}(L_i \geqslant \hat{v}_\alpha)\ .$$

The variance of the VaR estimator is difficult to evaluate, because the estimator is not an average of iid random variables. However, the following central limit theorems, given with references in Hong and Liu [2011], show that the VaR and ES estimators have asymptotically normal distributions.

Theorem 3.1 (Central Limit Theorems for the CMC VaR and ES Estimators) If $\mathbb{E}L^2 < \infty$ and the density of L, f_L, is positive and continuously differentiable in a neighborhood of v_α, then, as $N \to \infty$

1. $\sqrt{N}\,(\hat{v}_\alpha - v_\alpha) \xrightarrow{D} \frac{\sqrt{\alpha(1-\alpha)}}{f_L(v_\alpha)}\, Z_1$,

2. $\sqrt{N}(\hat{c}_\alpha - c_\alpha) \xrightarrow{D} \frac{\sqrt{\mathrm{Var}(L\mathbb{I}(L>v_\alpha))}}{(1-\alpha)}\, Z_2$,

where Z_1 and Z_2 are standard normal random variables and $\xrightarrow{D}$ denotes convergence in distribution.

3.2 Importance Sampling

The CMC VaR and ES estimators generally require a very large sample size in order to achieve an acceptable level of accuracy. This is because the estimators are focused on the relatively 'rare' event $\{L > v_\alpha\}$. There is a substantial body of theory devoted to efficient Monte Carlo methods for rare events. This theory has mainly been developed in the context of estimating rare-event probabilities of the form $\ell = \mathbb{P}(S(\mathbf{X}) > \gamma)$ for some real-valued function S, threshold γ, and random vector $\mathbf{X}$. Some key concepts and techniques of rare-event simulation are discussed in the Appendix. The following discussion will assume familiarity with these concepts.

The importance sampling approach to quantile estimation was suggested in Glynn [1996]. We replace the CMC estimator of the empirical distribution function with the IS estimator

$$\hat{F}_L^{\mathsf{IS}}(l) = 1 - \frac{1}{N}\sum_{i=1}^{N} W(L_i)\mathbb{I}(L_i > l) \,,$$

where the $\{L_i\}$ are drawn from the IS density g and $W(l) = f_L(l)/g(l)$ is the likelihood ratio. Note that this estimator focuses on the right tail of the distribution — see Glynn [1996] for a motivation. This then leads to the IS VaR estimator

$$\hat{v}_\alpha^{\mathsf{IS}} = \inf\left\{ l : \hat{F}_L^{\mathsf{IS}}(l) \geqslant \alpha \right\}. \tag{4}$$

The corresponding ES estimator is

$$\hat{c}_\alpha^{\mathsf{IS}} = \frac{1}{N(1-\alpha)}\sum_{i=1}^{N} W(L_i) L_i \mathbb{I}\left(L_i \geqslant \hat{v}_\alpha^{\mathsf{IS}}\right) \,, \tag{5}$$

where the L_i are drawn from g. If g is chosen such that draws from the right tail of L happen more frequently, this estimator could provide considerably better performance than the CMC estimator. In practice, the IS VaR estimator is calculated as follows.

Algorithm 3.2 (IS VaR Estimation)

1. *Draw* $L_1, \ldots, L_N$ *from the IS density* g.
2. *Calculate the likelihood ratios* $W(L_1), \ldots, W(L_N)$.
3. *Order the* $\{L_i\}$ *as* $L_{(1)} \leqslant \cdots \leqslant L_{(N)}$.
4. *Find* $N^* = \sup\left\{n : \frac{1}{N}\sum_{i=n}^{N} W(L_{(i)}) \geqslant 1-\alpha\right\}$.
5. *Return* $v_\alpha^{\mathsf{IS}} = L_{(N^*)}$.

So far we have taken g as given. The following central limit theorems, given in Hong and Liu [2011] and Sun and Hong [2010], suggest a good choice of g.

Theorem 3.2 (Central Limit Theorem for the IS VaR Estimator) If L has a positive and differentiable density f_L in a neighborhood of v_α and there exists an $\epsilon > 0$ such that $W(l)$ is bounded for all $l \in (v_\alpha - \epsilon, v_\alpha + \epsilon)$ and $\mathbb{E}_g \mathbb{I}\left(L \geqslant v_\alpha - \epsilon\right)(W(L))^p$ is finite for some $p > 2$, then as $N \to \infty$

1. $\sqrt{N}\left(\hat{v}_\alpha^{\mathsf{IS}} - v_\alpha\right) \xrightarrow{D} \frac{\sqrt{\mathrm{Var}_g(W(L)\mathbb{I}(L \geqslant v_\alpha))}}{f_L(v_\alpha)} Z_1$,
2. $\sqrt{N}(\hat{c}_\alpha^{\mathsf{IS}} - c_\alpha) \xrightarrow{D} \frac{\sqrt{\mathrm{Var}_g(W(L)L\mathbb{I}(L > v_\alpha))}}{(1-\alpha)} Z_2$,

where Z_1 and Z_2 are standard normal random variables and $\xrightarrow{D}$ denotes convergence in distribution.

This suggests that a good choice of g, at least asymptotically, is one that minimizes $\mathrm{Var}_g(W(L)\mathbb{I}(L \geqslant v_\alpha))$. This is equivalent to finding the density g that minimizes the variance of

$$\hat{\ell}^{\mathsf{IS}} = \frac{1}{N}\sum_{i=1}^{N} W(L_i)\mathbb{I}(L_i > v_\alpha) ,$$

where the $\{L_i\}$ are drawn from g. This is the standard IS estimator for

$$\ell = \mathbb{P}(L > v_\alpha) .$$

Of course, the computation of $\hat{\ell}^{\mathsf{IS}}$ involves v_α, which is the unknown quantity we seek to estimate. However, a rough estimate of v_α can often be obtained, either through an approximation or by doing an initial simulation using the CMC VaR estimator. Importance sampling estimators for VaR and ES will often provide very large efficiency gains, even in settings where the initial estimate of v_α is quite inaccurate.

Another complication is that we usually do not know f_L, the density of L. Thus, we cannot apply importance sampling to the $\{L_i\}$ directly. Instead, we seek to represent L as a function S of either a random vector $\mathbf{X}$ with known density $f_{\mathbf{X}}$ or a vector-valued stochastic process $\mathbf{X} = (\mathbf{X}(t), 0 \leqslant t \leqslant T)$, to which we can apply importance sampling.

In practice, the procedure for applying importance sampling is as follows.

Algorithm 3.3 (Importance Sampling Estimation for VaR and ES)
Given a representation $L = S(\mathbf{X})$,

1. *Calculate an initial estimate of v_α, denoted as $\hat{v}_\alpha$.*
2. *Find an appropriate importance sampling density for estimating $\mathbb{P}(L > \hat{v}_\alpha)$.*
3. *Generate $L_1 = S(\mathbf{X}_1), \ldots, L_N = S(\mathbf{X}_N)$ under the IS density and calculate the corresponding likelihood ratios $W(\mathbf{X}_1), \ldots, W(\mathbf{X}_N)$.*
4. *Calculate the VaR estimate as in* (4) *and the ES estimate as in* (5)*.*

3.2.1 Adaptive Importance Sampling

Because credit risk models are generally complicated, it may be difficult (or even impossible) to find *a priori* a good importance sampling density g. Adaptive importance sampling methods aim to avoid difficult theoretical and computational issues by 'learning' a good density from the data. We assume here that f_L, the density of L, is not known and that a representation of the form $L = S(\mathbf{X})$, where $\mathbf{X}$ has density $f_{\mathbf{X}}$, can be used instead. We apply importance sampling to the $\mathbf{X}$. Given a prespecified IS density $g_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$, the idea is to take an initial sample $\mathbf{X}_1, \ldots, \mathbf{X}_M$ and try to learn the optimal parameters using this sample. If the initial sample $\mathbf{X}_1, \ldots, \mathbf{X}_M$ can be sampled directly from the zero-variance density $g^*(\mathbf{x}) = f(\mathbf{x} \,|\, S(\mathbf{x}) > v_\alpha)$, then the parameters can be chosen either to minimize the CE distance to g^*,

$$\hat{\boldsymbol{\theta}}^*_{\mathrm{CE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} \log\left(g_{\boldsymbol{\theta}}(\mathbf{X}_i)\right) ,$$

or to minimize the variance of the estimator

$$\hat{\boldsymbol{\theta}}^*_{\mathrm{VM}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} W_{\boldsymbol{\theta}}(\mathbf{X}_i) .$$

In some settings, g^* is sampled from using Markov Chain Monte Carlo methods (see Kroese et al. [2011] for an introduction). However, because the probability of a loss greater than v_α is not too small, we can often use a more direct acceptance–rejection method here.

Algorithm 3.4 (Sampling Approximately from g^*)

1. *Generate a sample $L_1, \ldots, L_M$.*
2. *Order the sample from smallest to largest as $L_{(1)} \leqslant \cdots \leqslant L_{(M)}$.*
3. *Choose $L_{(\lceil \alpha M \rceil)}, \ldots, L_{(M)}$ as an approximate sample from g^*.*

A very small sample is usually sufficient to find very good CE or VM parameters. The additional computational cost of the trial is generally small compared to the overall costs of the simulation. Indeed, there is hardly any overhead compared with non-adaptive methods for quantile estimation, as such methods use trial runs to find an initial estimate of v_α. A similar adaptive approach is taken in Reitan and Aas [2010]. For an alternative method, where the parameters are updated during the primary sampling phase, see Egloff and Leippold [2010].

4 Specific Models

In this section we discuss four specific classes of credit risk model: Bernoulli mixture models, factor models, copula models, and intensity models. Although each of these models is based on the general framework (1), they use different mathematical structures to model the dependencies between the default variables $\{D_i\}$. As a result, each model requires a different Monte Carlo approach to efficiently estimate the VaR and ES.

4.1 The Bernoulli Mixture Model

Bernoulli mixture models are a fundamental class of credit risk models because many credit risk models can be represented as a mixture model. It is straightforward to apply importance sampling to these models.

In a Bernoulli mixture model, the Bernoulli default variables $D_1, \ldots, D_n$ are conditionally independent given a vector of default probabilities $\mathbf{P} = (P_1, \ldots, P_n)$. It is assumed that these default probabilities are of the form $\mathbf{P}(\mathbf{\Psi})$, where $\mathbf{\Psi}$ is a random vector with a known density $f_{\mathbf{\Psi}}$. Conditional on $\mathbf{P}$, calculating L reduces to calculating a weighted sum of independent light-tailed random variables.

It is quite straightforward to sample from a Bernoulli mixture model.

Algorithm 4.1 (Sampling from a Bernoulli Mixture Model)

1. *Generate a vector of success probabilities* $\mathbf{P} = (P_1, \ldots, P_n)$.
2. *Given* $\mathbf{P}$*, generate* $D_1 \sim \mathsf{Ber}(P_1), \ldots, D_n \sim \mathsf{Ber}(P_n)$.

4.1.1 One-Step Importance Sampling

It is usually not possible to directly apply importance sampling to L, as the distribution of L is often unavailable in closed form. Instead we can apply importance sampling to drawing either $\mathbf{P}$ or the $D_1, \ldots, D_n$ conditional on $\mathbf{P}$. It is simplest to apply importance sampling in the second case. If we assume that $l_1, \ldots, l_n$ are constants, then, conditional on $\mathbf{P}$,

$$L = l_1 D_1 + \cdots + l_n D_n$$

is the sum of independent random variables, with the ith variable taking the value l_i with probability P_i and 0 otherwise. We exponentially twist each of these variables so that the default probability for the ith component is given by

$$\widetilde{P}_i = \frac{P_i \exp(\theta\, l_i)}{P_i \exp(\theta\, l_i) + 1 - P_i}\ .$$

The unique 'asymptotically efficient' choice of θ is the solution to $\kappa_n(\theta^* \,|\, \mathbf{P}) = v_\alpha$, where

$$\kappa_n(\theta \,|\, \mathbf{P}) = \sum_{i=1}^{n} \log\left[P_i \exp(\theta\, l_i) + 1 - P_i\right] \tag{6}$$

is the joint cumulant generating function of the $\{l_i D_i\}$ conditional on $\mathbf{P}$.

Algorithm 4.2 (One-Step Importance Sampling for a Mixture Model)

1. *Generate* $\mathbf{P} = (P_1, \ldots, P_n)$.
2. *Find* θ^**, the solution to* $\kappa_n'(\theta) = v_\alpha$*. (This step usually needs to be done numerically).*
3. *If* $\theta^* < 0$*, set* $\theta^* = 0$.
4. *Calculate* $\widetilde{P}_i = \frac{P_i \exp(\theta^* l_i)}{P_i \exp(\theta^* l_i) + 1 - P_i},\ i = 1, \ldots, n$.
5. *Given* $\widetilde{P}_1, \ldots, \widetilde{P}_n$*, generate* $D_i \sim \mathsf{Ber}\left(\widetilde{P}_i\right), i = 1, \ldots, n$.
6. *Return* $L = l_1 D_1 + \cdots + l_n D_n$ *and the corresponding likelihood ratio*

$$W(L) = \exp\left(\kappa_n(\theta^* \,|\, \mathbf{P}) - \theta^* L\right).$$

Unfortunately, this approach may not give an asymptotically efficient estimator for $\ell = \mathbb{P}(L > v_\alpha)$. This is because $\mathbf{P}$ can play a critical role in driving the dynamics of the rare event. For example, in the context of Gaussian factor models, Glasserman and Li [2005] show that asymptotic efficiency can only be achieved if the correlation between the defaults decreases (at some rate) as $n \to \infty$ and $v_\alpha \to \infty$.

4.1.2 Two-Step Importance Sampling

A potentially more effective importance sampling scheme involves importance sampling in generating $\mathbf{P}$ as well as $D_1, \ldots, D_n$. We can decompose the variance of $\hat{\ell}$ as

$$\mathrm{Var}(\hat{\ell}) = \mathbb{E}\left(\mathrm{Var}\left(\hat{\ell} \,|\, \mathbf{P}\right)\right) + \mathrm{Var}\left(\mathbb{E}\left(\hat{\ell} \,|\, \mathbf{P}\right)\right).$$

The one-step importance sampling procedure detailed above minimizes $\mathrm{Var}(\hat{\ell} \,|\, \mathbf{P})$. Regarding sampling $\mathbf{P}$, we aim to minimize $\mathrm{Var}(\mathbb{E}(\hat{\ell} \,|\, \mathbf{P}))$. This is equivalent to minimizing the variance of $\hat{z}$, the CMC estimator of

$$z = \mathbb{P}\left(L > v_\alpha \,|\, \mathbf{P}(\mathbf{\Psi})\right).$$

The zero-variance density g^* for such a problem is given by

$$g^*_{\mathbf{\Psi}}(\boldsymbol{\psi}) \propto \mathbb{P}(L > v_\alpha \,|\, \mathbf{P}(\boldsymbol{\psi}))f_{\mathbf{\Psi}}(\boldsymbol{\psi}) \,.$$

The normalizing constant is the unknown ℓ, so this is not a practical IS density.

There are two common approaches to finding a good IS density. One approach uses a density $g_{\mathbf{\Psi}}$ whose mean is set equal to the mode of $g^*_{\mathbf{\Psi}}$. This mode is the solution to a generally intractable optimization problem.

Given $g_{\mathbf{\Psi}}$, the two-step importance sampling scheme is summarized as follows.

Algorithm 4.3 (Two-Step Importance Sampling for a Mixture Model)

1. *Draw* $\mathbf{\Psi}$ *from* $g_{\mathbf{\Psi}}$.
2. *Generate* $\mathbf{P} = \mathbf{P}(\mathbf{\Psi})$.
3. *Find* θ^**, the solution to* $\kappa'_n(\theta) = v_\alpha$.
4. *Calculate* $\widetilde{P}_i = \frac{P_i \exp(l_i\theta^*)}{P_i \exp(l_i\theta^*)+1-P_i}, i = 1, \ldots, n$.
5. *Given* $\widetilde{P}_1, \ldots, \widetilde{P}_n$*, generate* $D_i \sim \mathsf{Ber}\left(\widetilde{P}_i\right), i = 1, \ldots, n$.
6. *Return* $L = l_1 D_1 + \cdots + l_n D_n$ *and the corresponding likelihood ratio*

$$W(L) = \frac{f_{\mathbf{\Psi}}(\mathbf{\Psi})}{g_{\mathbf{\Psi}}(\mathbf{\Psi})} \exp\left(\kappa_n(\theta^* \,|\, \mathbf{P}) - \theta^* L\right) .$$

4.1.3 Worked Example: A Bernoulli Mixture Model with Beta Probabilities

We consider a simple Bernoulli mixture model for a portfolio with $n = 1000$ components, with $l_1 = \cdots = l_n = 1$. The default probabilities are all equal, with $P \sim \mathsf{Beta}(0.5, 9)$. We consider three approaches: CMC, CE, and one-step importance sampling. The CE approach finds the outcomes of P corresponding to the highest $N(1-\alpha)$ samples of L. It then computes the MLEs for a Beta distribution numerically. For the IS approach, $\kappa_n(\theta \,|\, P) = \hat{v}_\alpha$ can be solved analytically. However, for this problem, the dynamics of L are largely driven by P. Thus, the IS estimator performs very poorly. Each estimator was used to calculate 100 estimates. The means and standard deviations of these estimators are reported. For IS, the first 10% of the sample was used to calculate a rough estimate of $\hat{v}_\alpha$. For CE, the first 10% of the sample was used to learn the parameters.

Table 1: Estimated VaR and ES for a Bernoulli Mixture Model,

Estimator	$\hat{v}_\alpha$	Std($\hat{v}_\alpha$)	$\hat{c}_\alpha$	Std($\hat{c}_\alpha$)
$\alpha = 0.95$	$N = 10^4$			
CMC	197.5	3.3	270.0	4.3
CE	197.6	1.4	269.9	5.3
IS	197.5	3.2	269.7	4.8
$\alpha = 0.99$	$N = 10^4$			
CMC	316	7.7	382.9	10.0
CE	314.9	3.2	375.6	8.3
IS	316.2	9.3	378.2	9.8
$\alpha = 0.995$	$N = 10^4$			
CMC	363.3	9.9	430.6	10.5
CE	362.6	2.7	421.9	6.6
IS	363.4	9.3	413.0	27.0

4.2 Factor Models

In *factor models*, the ith component defaults when a corresponding random variable X_i crosses a preset threshold ρ_i. That is,

$$D_i = \mathbb{I}\left(X_i > \rho_i\right), \quad i = 1, \ldots, n \ .$$

The variable X_i can sometimes be thought of as corresponding to a default time, as in the Li copula model (see Li [2000]), though this need not be the case. The relationship between the $\{D_i\}$ is imposed by having the $\{X_i\}$ all depend on a vector of common factors, $\mathbf{\Psi}$. A model with one factor is called a *single factor* model; a model with more than one factor is referred to as a *multifactor* model. These factors may correspond to macroeconomic or industry specific factors, though they need not have an economic interpretation. In the simplest case of a *linear factor model*, each X_i is a weighted sum of the factors and another random variable, $\mathcal{E}_i$ which represents the component-specific *idiosyncratic* risk. Conditional on $\mathbf{\Psi}$, factor models are Bernoulli mixture models.

The most popular factor models are based on the normal and Student-t distributions. We focus on three specific factor models.

• In the *Gaussian factor model*, each X_i has the representation

$$X_i = a_{i1}Z_1 + \cdots + a_{im}Z_m + a_i\mathcal{E}_i \ ,$$

where the $\{Z_j\}$ and $\{\mathcal{E}_i\}$ are independent standard normal random variables and the coefficients are chosen such that the marginal distribution of each X_i is standard normal. Here, conditional on $Z_1 = z_1, \ldots, Z_m = z_m$ (thus, $\mathbf{\Psi} = \mathbf{Z}$), the default probability for the ith component is

$$\begin{aligned} P_i &= \mathbb{P}\left(\mathcal{E}_i > \frac{\rho_i - (a_{i1}z_1 + \cdots + a_{im}z_m)}{a_i}\right) \\ &= \Phi\left(\frac{(a_{i1}z_1 + \cdots + a_{im}z_m) - \rho_i}{a_i}\right). \end{aligned}$$

• In the *Student-t factor model*, each X_i is a weighted sum of Student-t random variables. Usually, the Student-t factor model is chosen such that each X_i has the following representation

$$X_i = \sqrt{\frac{r}{V}}\left(a_{i1}Z_1 + \cdots + a_{im}Z_m + a_i\mathcal{E}_i\right),$$

where the $\{Z_j\}$ are standard normals and V has a chi-squared distribution with r degrees of freedom. Here, conditional on $Z_1 = z_1, \ldots, Z_m = z_m$ and $V = v$ (thus, $\mathbf{\Psi} = (\mathbf{Z}, V)$), the default probability is

$$\begin{aligned} P_i &= \mathbb{P}\left(\mathcal{E}_i > \frac{\sqrt{v/r}\rho_i - (a_{i1}z_1 + \cdots + a_{im}z_m)}{a_i}\right) \\ &= \Phi\left(\frac{(a_{i1}z_1 + \cdots + a_{im}z_m) - \sqrt{v/r}\rho_i}{a_i}\right). \end{aligned}$$

• A more general single factor model with heavy tails and tail dependence is introduced in Bassamboo et al. [2008]. It is an extension of the normal mean-variance mixture models described in Frey and McNeil [2001]. Here, each X_i is of the form

$$X_i = \frac{\alpha_i Z + \sqrt{1-\alpha_i^2}\,\mathcal{E}_i}{W},$$

where the $\{\mathcal{E}_i\}$ are iid random variables independent of the random variable Z, and W is a random variable independent of Z and the $\{\mathcal{E}_i\}$, with a density f_W that satisfies

$$f_W(w) = \lambda w^{\nu-1} + o(w^{\nu-1}) \quad \text{as } w \downarrow 0\,. \tag{7}$$

This model includes that single factor Student-t model as a special case, as the chi-squared distribution satisfies (7). Conditional on $Z = z$ and $W = w$ (thus, $\mathbf{\Psi} = (Z, W)$) the default probabilities are

$$P_i = \mathbb{P}\left(\mathcal{E}_i > \frac{w\rho_i - \alpha_i z}{\sqrt{1-\alpha_i^2}}\right).$$

It is usually straightforward to sample from a factor model.

Algorithm 4.4 (Sampling from a Factor Model)

1. *Draw the common factors $\mathbf{\Psi}$ and the idiosyncratic risks $\mathcal{E}_1, \ldots, \mathcal{E}_n$.*
2. *Calculate $X_1, \ldots, X_n$ as per the model.*
3. *Calculate $L = l_1 \mathbb{I}(X_1 > \rho_1) + \cdots + l_n \mathbb{I}(X_n > \rho_n)$.*

4.2.1 Importance Sampling

Factor models are usually Bernoulli mixture models. Thus, importance sampling can be applied as above. It is usually necessary to use a two-step importance sampling scheme, as in Section 4.1.2. The difficulty lies in choosing $g_{\mathbf{\Psi}}$, the IS density for the common factors $\mathbf{\Psi}$.

In the case of Gaussian factor models, where $\Psi = \mathbf{Z}$, Glasserman and Li [2005] use a multivariate normal density $N(\boldsymbol{\mu}, I)$ with the mean vector $\boldsymbol{\mu}$ set equal to the mode of $g^*_{\mathbf{Z}}$. The mode, in turn, can be obtained as the solution to the optimization problem

$$\boldsymbol{\mu}^* = \underset{\mathbf{z}}{\operatorname{argmax}} \, \mathbb{P}\left(L > v_\alpha \,|\, \mathbf{Z} = \mathbf{z}\right) \exp(-\mathbf{z}^\intercal \mathbf{z}/2) \; . \tag{8}$$

Glasserman and Li suggest a number of approximations that simplify this problem. One approach is the *constant approximation*, where L is replaced by $\mathbb{E}\left[L \,|\, \mathbf{Z} = \mathbf{z}\right]$ and $\mathbb{P}(L > v_\alpha \,|\, \mathbf{Z} = \mathbf{z})$ is replaced by $\mathbb{I}\left(\mathbb{E}\left[L \,|\, \mathbf{Z} = \mathbf{z}\right] > v_\alpha\right)$. In this case, (8) becomes

$$\underset{\mathbf{z}}{\operatorname{argmin}} \left\{\mathbf{z}^\intercal \mathbf{z} : \mathbb{E}\left[L \,|\, \mathbf{Z} = \mathbf{z}\right] > v_\alpha\right\} . \tag{9}$$

Another approach is the *tail bound approximation*, which is shown to be asymptotically optimal for the case of a homogeneous single factor portfolio. This approach approximates $\mathbb{P}((L > v_\alpha \,|\, \mathbf{Z} = \mathbf{z})$ by its upper bound, and (8) becomes

$$\underset{\mathbf{z}}{\operatorname{argmax}} \left\{\kappa_n(\theta_{v_\alpha} \,|\, \mathbf{z}) - \theta_{v_\alpha} v_\alpha - \mathbf{z}^\intercal \mathbf{z}/2\right\} ,$$

where $\theta_{v_\alpha} = \theta_{v_\alpha}(\mathbf{z})$ is the solution to $\kappa_n(\theta \,|\, \mathbf{z}) = v_\alpha$ and κ_n is given in (6).

In a multi-factor setting, the problem of finding a good approximation of g^* becomes much more difficult. This is because more than one combination of factors can cause a loss larger than v_α. Glasserman et al. [2008] propose an approach which essentially attempts to partition the rare event $\{L > v_\alpha\}$ into different sub-events; each sub-event corresponds to a particular set of factors taking large values, and they solve (9) for each of these events. This approach is shown to be asymptotically efficient in certain settings. As far as we are aware, this is the only method given in the existing literature that deals adequately with the problem of possibly infinite variance in a multi-factor setting.

In the Student-t factor model setting given above, Kang and Shahabuddin [2005] propose first sampling V, then $Z_1, \ldots, Z_m$. Given V, they proceed as in Glasserman et al. [2008]. They

propose exponentially twisting V by a parameter which is again the solution of a constrained optimization problem. Note that this approach is very computationally expensive, as it requires multiple numerical optimization procedures per sample. Kang and Shahabuddin [2005] suggest using a stratified sampling scheme to minimize this cost.

For the general single-factor model, Bassamboo et al. [2008] introduce two methods. In the first, they propose exponentially twisting W and find a good twisting parameter θ by minimizing the upper bound on the likelihood ratio. This approach gives bounded relative error under some technical conditions. In the second, they apply hazard-rate twisting to $V = 1/W$, see Juneja and Shahabuddin [2006] for a discussion of this method. Again, they choose the twisting parameter to minimize the upper bound on the likelihood ratio. Under some technical conditions, the resulting estimator is shown to be asymptotically efficient.

Another method for applying variance reduction to Student-t factor models is given in Chan and Kroese [2010]. In this approach, VaR can be estimated by calculating the expectations of truncated gamma random variables.

4.2.2 Worked Example: A Gaussian Factor Model

We consider an example suggested in Glasserman and Li [2005]. In this example, the portfolio is of size $n = 1000$, with $l_i = (\lceil 5i/n \rceil)^2$. The barriers are given by $\rho_i = \Phi^{-1}(1 - P_i)$, where $P_i = 0.01 * (1 + \sin(16\pi i/n))$. The $m = 10$ factor loadings, $\{a_{ij}\}$ are drawn uniformly on $(0, 1/\sqrt{m})$.

We calculate the VaR and ES using three different methods: CMC, Glasserman and Li's method, and Cross-Entropy. For Glasserman and Li's algorithm, we only apply importance sampling to the $\{Z_i\}$, as twisting the $\{D_i\}$ does not make a substantial difference in this case, and takes considerably more time. We draw the $\{Z_i\}$ from a $\mathsf{N}(\boldsymbol{\mu}, I)$ distribution, with $\boldsymbol{\mu}$ the solution of (4.2.1) found via numerical root-finding. In the CE approach, we set the means of the $\{Z_i\}$ and the mean of the $\{\mathcal{E}_i\}$ equal to the sample means of the $\{Z_i\}$ and $\{\mathcal{E}_i\}$ corresponding to the $\lfloor N(1-\alpha) \rfloor$ highest values of L.

Table 2 gives the numerical results. The estimators were calculated 100 times each and their means and standard deviations are reported. The Glasserman and Li estimator uses the first 10% of the sample to find an initial estimate of $\hat{v}_\alpha$. The CE estimator uses the first 10% of the sample to learn good parameters. Note that the CE and Glasserman and Li estimators performing better relative to the CMC estimator as α gets larger. Running times are not given here, as they are implementation specific, but we note that the Glasserman Li approach is considerably slower than the CE approach in our implementation.

Table 2: Estimated VaR and ES for a Gaussian factor model.

Estimator	$\hat{v}_\alpha$	Std($\hat{v}_\alpha$)	$\hat{c}_\alpha$	Std($\hat{c}_\alpha$)
$\alpha = 0.95$	$N = 10^4$			
CMC	215	7	488	19
CE	217	3	469	3
GL	216	3	469	3
$\alpha = 0.99$	$N = 10^5$			
CMC	595	31	988	58
CE	600	13	987	12
GL	599	6	987	5
$\alpha = 0.995$	$N = 10^5$			
CMC	833	17	1267	28
CE	837	2	1274	2
GL	837	2	1274	2

4.3 Copula Models

One of the most popular ways of expressing dependency in credit risk models is to use copulas. A copula is simply a multivariate distribution function with uniform marginals:

$$C(u_1, \ldots, u_n) : [0,1]^n \to [0,1] \; .$$

Copulas describe the dependency structure between uniform random variables $U_1, \ldots, U_n$. These can be transformed into random variables $X_1, \ldots, X_n$, with arbitrary distributions $F_1, \ldots, F_n$, by setting $X_1 = F_1^{-1}(U_1), \ldots, X_n = F_n^{-1}(U_n)$. This means that the dependency structure of the $\{X_i\}$ can be modeled separately from their marginal distributions. It can be shown that the dependency structure of any distribution can be defined via a copula (see Nelsen [2006]). Often, the X_i are taken to be default times as, for example, in the Li model, see Li [2000]. However, this need not be the case. If each D_i is of the form $D_i = \mathbb{I}(X_i > \rho_i)$, then the model is said to be a *threshold model*.

We focus on the Gaussian, Student-t and Archimedean copulas, as these are the most popular copulas in credit risk modeling. The Gaussian copula has tail independence. An attractive feature of the other models is that they exhibit tail dependence.

- The *Gaussian copula*, popularized in Li [2000], is of the form

$$C_G(u_1, \ldots, u_n) = \Phi_\Gamma\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\right) \; ,$$

where $\Phi_\Gamma(\cdot)$ is the multivariate normal distribution function with mean vector $\mathbf{0}$ and correlation matrix Γ. The Gaussian factor model, described above, can be interpreted as a Gaussian copula.

- The *Student-t copula* is of the form

$$C_T(u_1, \ldots, u_n) = T_{\nu,\Gamma}\left(T_\nu^{-1}(u_1), \ldots, T_\nu^{-1}(u_n)\right) ,$$

where $T_{\nu,\Gamma}$ is the multivariate Student-t distribution function with ν degrees of freedom, mean vector $\mathbf{0}$, and correlation matrix Γ. The Student-t factor model can be interpreted as a Student-t copula. The Student-T copula has tail dependence in both tails.

- *Archimedean Copulas* are of the form

$$C_\psi(u_1, \ldots, u_n) = \psi^{-1}(\psi(u_1) + \cdots + \psi(u_n)) ,$$

where the *generator* of the copula is a function $\psi : [0,1] \to [0,\infty]$ that satisfies the following conditions:

1. It is strictly decreasing.
2. $\psi(0) = \infty$ and $\psi(1) = 0$.
3. ψ^{-1} is completely monotonic, meaning $(-1)^k \frac{\mathrm{d}^k}{\mathrm{d}u^k}\psi^{-1}(u) \geqslant 0, \quad \forall k \in \mathbb{N}$ and $u \in [0,\infty)$

The class of Archimedean copulas includes the *Gumbel copula*, where $\psi_\eta(u) = (-\log u)^\eta$, and the *Clayton copula*, where $\psi_\eta(u) = u^{-\eta} - 1$. The Gumbel copula has upper tail dependence and the Clayton copula has lower tail dependence.

4.3.1 Sampling from a General Copula

In theory, it is possible to sample from any copula $C(u_1, \ldots, u_n)$. The approach, given in Cherubini et al. [2004], is as follows. Let $C_i(u_1, \ldots, u_i) = C(u_1, \ldots, u_i, 1, \ldots, 1), i = 1, \ldots, n$. The conditional distribution of the copula C_i is

$$\begin{aligned} C_i(u_i \,|\, u_1, \ldots, u_{i-1}) &= \mathbb{P}(U_i \leqslant u_i \,|\, U_1 = u_1, \ldots, U_{i-1} = u_{i-1}) \\ &= \frac{\frac{\partial^{i-1}}{\partial u_1, \ldots, \partial u_{i-1}} C_i(u_1, \ldots, u_i)}{\frac{\partial^{i-1}}{\partial u_1, \ldots, \partial u_{i-1}} C_{i-1}(u_1, \ldots, u_{i-1})} . \end{aligned}$$

We can then decompose $C(u_1, \ldots, u_n)$ as follows

$$C(u_1, \ldots, u_n) = \mathbb{P}(U_1 < u_1)\, C_2(u_2 \,|\, u_1) \cdots C_n(u_n \,|\, u_1, \ldots, u_{n-1}) .$$

Algorithm 4.5 (Sampling from a General Copula)

1. *Draw* U_1 *uniformly on* $(0, 1)$.
2. *Draw* U_i *from the distribution* $C_i(\cdot \mid u_1, \ldots, u_{i-1})$*, for* $i = 2, \ldots, n$.

In general, $C_i(\cdot \mid u_1, \ldots, u_{i-1})$ has to be sampled via the inverse transform method (see Kroese et al. [2011]). This involves drawing a uniform random variable V, and solving $V = C_i(u_i \mid u_1, \ldots, u_{i-1})$ for u_i. This usually needs to be done using a numerical root-finding procedure. In practice, this tends to make sampling from an arbitrary copula too expensive to be feasible.

4.3.2 Sampling from Gaussian and Student-t Copulas

The Gaussian and Student-t copulas are *implicit copulas*. That is, they are copulas implied by the multivariate normal and Student-t distributions. Hence, drawing from these copulas is simply a case of drawing from their respective multivariate distribution. Algorithms for drawing from these distributions are given in Kroese et al. [2011].

Algorithm 4.6 (Sampling from a Gaussian copula)

1. *Draw* $\mathbf{Z} = (Z_1, \ldots, Z_n) \sim \mathsf{N}(0, \Sigma)$.
2. *Return* $U_1 = \Phi(Z_1), \ldots, U_n = \Phi(Z_n)$.

Algorithm 4.7 (Sampling from a Student-t copula)

1. *Draw* $\mathbf{Y}$ *from a multivariate Student-t distribution with* ν *degrees of freedom and correlation matrix* Γ.
2. *Return* $U_1 = T_\nu(Z_1), \ldots, U_n = T_\nu(Z_n)$.

4.3.3 Sampling from Archimedean copulas

Archimedean copulas are particularly easy to sample from. The approach below uses Bernstein's theorem, which states that if ψ satisfies the conditions for an Archimedean generator, then ψ^{-1} is of the form

$$\psi^{-1}(u) = \int_0^\infty \mathrm{e}^{-u\lambda} \, \mathrm{d}F_\Lambda(\lambda) \; .$$

That is, $\psi^{-1}(u)$ is the Laplace transform of some distribution F_Λ. It is easily verified that, if Λ is drawn from F_Λ and $X_1, \ldots, X_n$ are iid and $\mathsf{U}(0, 1)$ distributed, then

$$U_1 = \psi^{-1}\left(\frac{-\log X_1}{\Lambda}\right), \ldots, U_n = \psi^{-1}\left(\frac{-\log X_n}{\Lambda}\right)$$

have the distribution given by the Archimedean copula. Thus, if we know F_Λ, we have the following algorithm for sampling from an Archimedean copula.

Algorithm 4.8 (Sampling from an Archimedean copula)

1. *Draw Λ from the distribution F_Λ.*
2. *Draw iid standard uniform random variables $X_1, \ldots, X_n$.*
3. *Return*
$$U_1 = \psi^{-1}\left(\frac{-\log X_1}{\Lambda}\right), \ldots, U_n = \psi^{-1}\left(\frac{-\log X_n}{\Lambda}\right).$$

Given an arbitrary generator, ψ, F_Λ may not be a known distribution, or one that can be sampled from in a straightforward manner. However, F_Λ is known for both the Gumbel and Clayton copulas. For the Gumbel copula, Λ has a stable distribution $\mathsf{St}(1/\eta, 1, \gamma, \eta)$, where $\gamma = (\cos(\pi\eta/2))^\eta$. In the case of the Clayton copula, Λ is $\mathsf{Gam}(1/\eta, 1)$ distributed.

4.3.4 Importance Sampling

Importance sampling is straightforward for Gaussian and Student-t copula models, as it can be applied directly to the multivariate densities.

In an Archimedean copula model, $U_1, \ldots, U_n$ are independent conditional on Λ. If $D_1, \ldots, D_n$ are generated using a threshold approach, we can represent such a model as a Bernoulli mixture model. This is because,

$$\mathbb{P}(U_i > \rho_i) = \mathbb{P}\left(\psi^{-1}\left(\frac{-\log X_i}{\Lambda}\right) > \rho_i\right) = 1 - \exp\left\{-\Lambda\psi(\rho_i)\right\}. \tag{10}$$

Thus, we can apply importance sampling as in the Bernoulli mixture model case given above.

4.3.5 Worked Example: A Clayton Copula Model

We consider the case where exponentially distributed default times are generated using a Clayton copula. Uniform random variables $U_1, \ldots, U_n$ are drawn from a Clayton copula with parameter $\eta = 1.5$. These are transformed into exponential random variables with parameter $\lambda = 0.1$ by setting

$$X_i = -\frac{\log U_i}{\lambda}.$$

Each D_i is then generated as $\mathbb{I}(X_i < 1)$. VaR and CVaR are both estimated using CMC, CE and one-step importance sampling. In all three cases, the Clayton copula is sampled from via the Laplace transform method detailed above. In the CE case, Λ is sampled from a Gamma distribution with parameters estimated from the elite sample. In the one-step IS case, the importance sampling is applied by twisting the default probabilities $P_1, \ldots, P_n$, which are calculated as in (10). For the CE estimator, the first 10% of the sample is used for learning phase. For the IS estimator, the first 10% of the sample is used as to get a rough estimate of $\hat{v}$. The results are given in the following table. Note that the CE estimator gives significant variance reduction provided that the sample size is large enough to estimate good parameters in the learning phase. The one-step importance sampling estimator performs not much better than CMC, as the value of L is very dependent on the realization of Λ.

Table 3: Estimated VaR and ES for a Clayton Copula model.

Estimator	$\hat{v}_\alpha$	Std($\hat{v}_\alpha$)	$\hat{c}_\alpha$	Std($\hat{c}_\alpha$)
$\alpha = 0.95$	$N = 10^3$			
CMC	72	4.9	89.9	2.4
CE	73	5.2	86.5	9.6
IS	73.5	5.4	86.8	4.8
$\alpha = 0.95$	$N = 10^4$			
CMC	72.7	1.6	88.9	0.8
CE	72.9	0.3	88.7	0.1
IS	72.8	1.5	88.5	0.9
$\alpha = 0.99$	$N = 10^4$			
CMC	97.5	0.6	100.1	0.2
CE	97.6	0.5	99	0.5
IS	97.6	0.6	98.7	0.4

4.4 Intensity Models

In intensity models, the default times of the n components, $\tau_1, \ldots, \tau_n$, are modeled by the arrival times of point processes. Denoting by T the time at which the portfolio is assessed, the Bernoulli default variables are given by $D_1 = \mathbb{I}(\tau_1 < T), \ldots, D_n = \mathbb{I}(\tau_n < T)$. In a *top-down* approach, the defaults are modeled as the arrivals of a single point process. The intensity of this process is given without reference to the portfolio constituents. In a *bottom-up* approach, each component of the portfolio is modeled separately. We will focus on this approach, and refer the reader to Giesecke [2008] for further discussion of modeling approaches. We model each τ_i as corresponding to the arrival time of an indicator process $(N_i(t), t \geqslant 0)$. Such a process has a stochastic intensity $\lambda_i(t), t \geqslant 0$, which is equal to 0 after the first arrival. Intuitively, $\lambda_i(t)$ is the rate at which arrivals occur at time t, conditional on the filtration (that is, the history) of the process up to time t. The default probability for the ith component is given by

$$P_i = 1 - \mathbb{P}(\tau_i < T) = 1 - \mathbb{E}\left[\exp\left\{-\int_0^T \lambda_i(s)\,\mathrm{d}s\right\}\right].$$

Dependency between defaults can be induced by assuming that each intensity λ_i is a function of a common process $(X(t), t \geqslant 0)$ and an idiosyncratic process $(X_i(t), t \geqslant 0)$; for example, $\lambda_i(t) = X(t) + X_i(t)$. A popular modeling choice for the process $(X(t))$ is that it satisfies a

stochastic differential equation with jumps:

$$\mathrm{d}X(t) = \mu(X(t))\,\mathrm{d}t + \sigma(X(t))\,\mathrm{d}B(t) + \Delta J(t)\ , \tag{11}$$

where $(B(t), t \geqslant 0)$ is a standard Brownian motion, $(\Delta J(t), t \geqslant 0)$ is a jump process, and both μ and σ are deterministic functions. The idiosyncratic processes $(X_i(t), t \geqslant 0), i = 1, \ldots, n$ can be modeled in a similar way. If μ and σ are affine functions, then under certain assumptions, the default probabilities $P_1, \ldots, P_n$ can be found by solving a system of ODEs (see Duffie et al. [2003] and Duffie [2005]).

One appeal of intensity models is that they can capture the empirical phenomenon of *contagion*, where defaults tend to happen in clusters. A popular model of contagion is the *generalized Hawke's process*, where the point process $(N(t), t \geqslant 0)$ has a stochastic intensity that satisfies

$$\mathrm{d}\lambda(t) = \kappa(\mu - \lambda(t))\,\mathrm{d}t + \sigma\sqrt{\lambda(t)}\,\mathrm{d}B(t) + \Delta N(t)\ .$$

Point processes in which the intensity depends on the number of arrivals are called *self-exciting*. Intensity models can also capture dependency between credit losses and the default process. A general introduction of using point process models in credit risk is given in Giesecke [2004]. For the relevant background on stochastic differential equations see, for example, Protter [2005].

4.4.1 Sampling from Intensity Models

In practice, though each portfolio component is modeled by a separate point process, we only simulate a single point process. This point process has intensity $\lambda(t) = \sum_{i=1}^{n} \lambda_i(t)$. On the event of a default, the ith component of the portfolio is chosen to default with probability $\lambda_i(t)/\lambda(t)$. The choice of algorithm for simulating from a stochastic intensity model depends on whether the intensity $\lambda(t)$ can be bounded between jumps. If the intensity can be bounded between jumps and it is straightforward to determine $\lambda(t)$ for an arbitrary t, then a thinning method due to Ogata [1981] can be used. At each jump, a piecewise constant process $(\lambda^*(t))$ is identified such that $\lambda(t) < \lambda^*(t)$ almost surely so long as no other jumps occur. A Poisson process with intensity function $\lambda^*(t)$ is simulated, and points are accepted with probability $\lambda(t)/\lambda^*(t)$. This gives the following algorithm.

Algorithm 4.9 (Sampling from a Point Process via Thinning)

1. *Set $i = 0$ and $\tau_0 = 0$;*
2. *Find λ_i^*, the upper bound of $\lambda(t), \tau_i \leqslant t \leqslant T$ given the history of the process up until time τ_i.*
3. *Simulate arrival times $\tilde{\tau}_1, \ldots \tilde{\tau}_n$ for a homogeneous Poisson process with intensity λ^*. Accept each arrival with probability $\lambda_i(\tau)/\lambda_i^*$. Stop after the first arrival time $\tilde{\tau}_i^*$ is accepted.*
4. *Set $\tau_i = \tilde{\tau}_i^* + \tau_{i-1}$.*
5. *Set $i = i + 1$ and repeat from step 2 until $\tau_i > T$.*

There is a general method of sampling from a point process driven by a stochastic intensity. If the compensator $\Lambda(t) \to \infty$ as $t \to \infty$ then $(N(t))$ is a standard Poisson process under the time change defined by $(\Lambda(t))$, with interarrival times given by $\mathsf{Exp}(1)$ random variables (see Giesecke et al. [2011]). The arrival times of the original process can be found by inverting $\Lambda(t)$. That is, given a sequence $Y_1, \ldots, Y_n$ of $\mathsf{Exp}(1)$ random variables representing the interarrival times of the time-changed process, the nth arrival time of the original process, τ_n, can be found by solving,

$$\tau_n = \inf_{t \geqslant 0} \left\{ \int_0^t \lambda(s)\, \mathrm{d}s \geqslant \sum_{i=1}^{n} Y_i \right\}.$$

This suggests the following algorithm.

Algorithm 4.10 (Sampling from a Point Process via a Time Change)

1. *Set* $i = 1$.
2. *Draw* Y_i *from an* $\mathsf{Exp}(1)$ *distribution.*
3. *Return* τ_i*, the time at which* $\Lambda(t)$ *hits* $\sum_{j=1}^{i} Y_j$.
4. *Set* $i = i + 1$ *and repeat from step* 2 *until* $\tau_i > T$.

This method is usually very computationally expensive, as the integral process $\Lambda(t) = \int_0^t \lambda(s)\, \mathrm{d}s, t \geqslant 0$ needs to be approximated on a discrete grid. The conditional distributions of $\Lambda(t)$ may also be unknown, in which case the process may only be approximately sampled at the grid points. An alternative method, that does not require simulating the intensity between jumps is suggested in Giesecke et al. [2011]. However, this method may be difficult or impossible to apply in some settings.

4.4.2 Importance Sampling

Importance sampling can be applied to intensity models in a number of different ways. For example, it can be observed that the events $\{N(t) > \gamma\}$ and $\{\sum_{i=1}^{n} N_i(t) > \gamma\}$ can both be written in the form $\left\{S_{\lceil \gamma \rceil} < T\right\}$, where S_k is the sum of k random variables, representing the first k arrival times. In this setting, exponential twisting can be applied to S_k. Unfortunately, this is often not possible, as the distribution of the S_k is usually either unknown or intractable — see Giesecke and Shkolnik [2011] for a discussion. However, in this setting, standard large deviations techniques can be applied to find good twisting parameters.

Another method is to apply a change of measure to the point process itself. This is the approach taken in Zhang et al. [2009], which considers a generalized Hawke's process. In the approach given in Giesecke and Shkolnik [2011], the change of measure is applied to the intensity processes instead.

If indicator processes are independent of one another conditional on some common factors $\mathbf{X}_t$, then they have a Bernoulli mixture model structure. Thus, the techniques described in Section 4 can be applied. In the particular case where intensities are of the form $\lambda_i(t) = X(t) + X_i(t)$

driven by (11), and the random factors are affine processes, Bassamboo and Jain [2006] propose applying an exponential change of measure to the processes, with a parameter $\boldsymbol{\theta}$ that minimizes the upper bound on the likelihood ratio.

4.5 An Example Point Process Model

In this model, taken from Giesecke and Shkolnik [2011], the individual component intensities are given by

$$\lambda_i(t) = (w_i X_0(t) + X_i(t))(1 - N_i(t)) \ ,$$

where each $X_i(t)$ satisfies the SDE

$$\mathrm{d}X_i(t) = \kappa_i \left(\bar{X}_i(t) - X_i(t)\right) \mathrm{d}t + \sigma_i \sqrt{X_i(t)}\, \mathrm{d}B_i(t) + \delta_i \, \mathrm{d}J_i(t) \ .$$

Here, $J_i(t) = \Delta_1 N_i(t) + \cdots + \Delta_n N_n(t)$ and the $(B_i(t), t \geqslant 0), i = 1, \ldots, n$ are standard Brownian motions. The $\{\kappa_i\}$ are drawn uniformly on $(0.5, 1.5)$. The $\left\{\bar{X}_i\right\}$ are drawn uniformly on $(0.001, 0.051)$ and each σ_i is equal to $\min\left(\sqrt{2\kappa_i \bar{X}_i}, \bar{\sigma}_i\right)$, where the $\{\bar{\sigma}_i\}$ are drawn uniformly on $(0, 0.2)$. Each factor weight w_i is drawn uniformly on $(0, 1)$. The $\{\Delta_i\}$ are drawn uniformly on $(0, 2/n)$ and the $\{\delta_i\}$ are drawn uniformly on $(0, 2)$. We compare the CMC algorithm with one of the two algorithms given in Giesecke and Shkolnik [2011].

In the CMC approach, the process $(N_t, t \geqslant 0)$ is generated using the time-change algorithm (Algorithm 4.10). A single point process is generated with intensity $\lambda(t) = \sum_{i=1}^n \lambda_i(t)$. The intensity processes $\lambda_1(t), \ldots, \lambda_n(t)$ are square-root processes, so they can be simulated exactly on a mesh using non-central chi-squared random variables (see Glasserman [2004]). A mesh of 1000 points is used and the integral $\int_0^t \lambda(s)\, \mathrm{d}s$ is evaluated via the trapezoidal rule. On the event of the kth default, the ith component of the portfolio is selected to default with probability $\lambda_i(\tau_k)/\lambda(\tau_k)$.

The IS algorithm replaces the point process $(N(t))$ with a Poisson process with intensity $\lambda = \hat{v}_\alpha$. The number of defaults, N, is drawn from a Poisson distribution with mean $\hat{v}_\alpha$. The default times $\tau_1, \ldots, \tau_N$ are N ordered uniform random variables on the interval $[0, 1]$. At time T, the Radon–Nikodym derivative for this change of measure is given by

$$M(T) = \exp\left\{\hat{v}_\alpha \tau_N - N(T) \log(\hat{v}_\alpha)\right\} + \sum_{k=1}^{N} \log(\lambda(\tau_k)) - \int_0^{\tau_N} \lambda(s)\, \mathrm{d}s \ .$$

The dynamics of $(\lambda_i(t), t \geqslant 0), i = 1, \ldots, n$ remain unchanged between defaults. A great advantage of this method is a reduction in computational effort, as $\lambda_i(t)$ only needs to be calculated up until the final default time.

The following numerical results are based on a portfolio of size $n = 100$, with each $l_i = 1$. A sample size of $N = 10^3$ was used. The CMC and IS algorithms appear to give different values for c_α. However, for larger sample sizes, the CMC estimates of c_α get closer to the IS estimates. For the importance sampling algorithm, the first 20% of the sample is used to get a rough estimate of $\hat{v}_\alpha$.

Table 4: Estimated VaR and ES for an intensity model.

Estimator	$\hat{v}_\alpha$	Std($\hat{v}_\alpha$)	$\hat{c}_\alpha$	Std($\hat{c}_\alpha$)
$\alpha = 0.95$	$N = 10^3$			
CMC	20	0.0	23.0	0.6
IS	20	0.0	22.6	0.6
$\alpha = 0.99$	$N = 10^3$			
CMC	24.4	0.8	29.5	1.3
IS	24.2	0.4	26.7	0.5
$\alpha = 0.995$	$N = 10^3$			
CMC	26.1	1.0	33.1	1.3
IS	25.8	0.4	27.8	0.7

A Appendix: A Primer on Rare-Event Simulation

The problem of finding good estimators for risk measures such as VaR and ES can, to a large extent, be reduced to the problem of finding good estimators for rare-event probabilities. This is a much better understood problem, and one which has given rise to a large number of effective Monte Carlo techniques. The vast majority of the literature on VaR and ES estimation has focused on a variance reduction method known as *importance sampling* and has used methods from the theory of rare-event simulation to find good classes of importance sampling estimators. These methods can be roughly split into two classes: (1) methods based primarily on Large Deviations asymptotics, and (2) adaptive methods, which 'learn' good estimators. In this appendix, we review the basics of rare-event probability estimation and discuss a number of approaches that work well in the credit risk context. There is an extensive literature on rare-event simulation; we mention, in particular, Bucklew [2004], Rubino and Tuffin [2009], Asmussen and Glynn [2007] and Kroese et al. [2011].

A fundamental problem of rare-event simulation is to estimate $\ell = \mathbb{P}(S(\mathbf{X}) > \gamma)$, when ℓ is very small. Here, S is a real-valued function, $\mathbf{X}$ is a random vector with density f, and γ is a constant. The *Crude Monte Carlo* (CMC) estimator of ℓ is defined as

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(S(\mathbf{X}_i) > \gamma) \ , \tag{12}$$

where the $\{\mathbf{X}_i\}$ are iid draws from f. This estimator performs very well when ℓ is large, but works very badly as $\ell \to 0$. This is because the event of interest $\{S(\mathbf{X}) > \gamma\}$, which is rare by

nature, must happen a large number of times in order to get an accurate estimate. The aim of rare event simulation is to find better estimators in such settings.

A.1 Efficiency

The accuracy of a rare-event estimator is often measured by its *relative error*. This is the normalized standard deviation of the estimator. We can usually think of a rare event estimator as an average of iid replicates of a random variable, which we will label Z. For example, the CMC estimator is an average of iid replicates of $Z = \mathbb{I}(S(\mathbf{X}) > \gamma)$. The relative error is then defined as

$$\mathrm{RE} = \frac{\sqrt{\mathrm{Var}(Z)}}{\ell\sqrt{N}} \ .$$

The relative error of the CMC estimator of ℓ is given by

$$\frac{\sqrt{\ell(1-\ell)}}{\ell\sqrt{N}} \approx \frac{1}{\sqrt{N}\sqrt{\ell}}$$

for small ℓ. This means that a very large sample size is required in order to achieve a low error. For example, estimating a probability of order 10^{-6} to a relative error of 0.01 requires a sample size of approximately 10^{10}. If an estimator is unbiased, its variance is given by

$$\mathrm{Var}(Z) = \mathbb{E}Z^2 - (\mathbb{E}Z)^2 = \mathbb{E}Z^2 - \ell^2 \stackrel{\mathrm{def}}{=} M - \ell^2 \ .$$

This means that the variance of an unbiased estimator is entirely determined by $M = \mathbb{E}Z^2$, the second moment of the random variable Z.

Rare event estimators are often evaluated in terms of their asymptotic performance. To do this, we embed the rare event of interest in a family of increasingly rare events indexed by a rarity parameter γ. For example, we might consider what happens to estimators of $\ell = \mathbb{P}(S(\mathbf{X}) > \gamma)$ as $\gamma \to \infty$. The most common notion of asymptotic efficiency is *logarithmic efficiency*. An estimator is said to be logarithmically or asymptotically efficient if

$$\liminf_{\gamma\to\infty} \frac{|\log M|}{|\log \ell^2|} \geqslant 1 \ .$$

By Jensen's inequality, $M \geqslant \ell^2$. Logarithmic efficiency means that asymptotically the estimator attains this lower bound on a log scale.

A.2 Importance Sampling

Importance sampling is a variance reduction method that is particularly well suited to rare event problems. The idea is to improve upon the efficiency of the CMC estimator by using a different probability measure, under which the rare event is more likely. To do this, we observe that an

expectation with respect to some density f can be rewritten as an expectation with respect to another density g, so long as $f(\mathbf{x}) = 0$ when $g(\mathbf{x}) = 0$. We write

$$\begin{aligned}\mathbb{E}_f \mathbb{I}(S(\mathbf{x}) > \gamma) &= \int \mathbb{I}(S(\mathbf{x}) > \gamma) f(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &= \int \frac{f(\mathbf{x})}{g(\mathbf{x})} \mathbb{I}(S(\mathbf{x}) > \gamma) g(\mathbf{x})\, \mathrm{d}\mathbf{x} = \mathbb{E}_g W(\mathbf{X}) \mathbb{I}(S(\mathbf{X}) > \gamma)\ ,\end{aligned}$$

where $W(\mathbf{x}) = f(\mathbf{x})/g(\mathbf{x})$ is the *likelihood ratio*. This allows us to replace the CMC estimator (12) of ℓ with the *Importance Sampling* (IS) estimator

$$\hat{\ell}_{IS} = \frac{1}{N} \sum_{i=1}^{N} W(\mathbf{X}_i) \mathbb{I}(S(\mathbf{X}_i) > \gamma),$$

where the $\{\mathbf{X}_i\}$ are now drawn from g rather than f. The second moment of the IS estimator is

$$M_{\mathrm{IS}} = \mathbb{E}_g \left(\frac{f(\mathbf{X})}{g(\mathbf{X})} \right)^2 \mathbb{I}(S(\mathbf{X}) > \gamma) = \mathbb{E}_f \frac{f(\mathbf{X})}{g(\mathbf{X})} \mathbb{I}(S(\mathbf{X}) > \gamma) = \mathbb{E} W(\mathbf{X}) \mathbb{I}(S(\mathbf{X}) > \gamma).$$

An importance sampling estimator will have smaller variance than the CMC estimator if $M_{\mathrm{IS}} < \mathbb{E}\hat{\ell}^2$, that is, if

$$\mathbb{E}_f \frac{f(\mathbf{X})}{g(\mathbf{X})} \mathbb{I}\left(S(\mathbf{X}) > \gamma\right) < \mathbb{E}_f \mathbb{I}(S(\mathbf{X}) > \gamma)\ .$$

The optimal IS density is the density that minimizes M_{IS}. It turns out that this density, g^*, actually gives an estimator with zero variance. The zero-variance density is given by

$$g^*(\mathbf{x}) = \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{E}_f \frac{f(\mathbf{X})}{g(\mathbf{X})} \mathbb{I}(S(\mathbf{X}) > \gamma) = \frac{f(\mathbf{x}) \mathbb{I}(S(\mathbf{x}) > \gamma)}{\ell},$$

where $\mathcal{G}$ contains all permissible densities (those such that $g(\mathbf{x}) = 0 \Rightarrow f(\mathbf{x}) = 0$). Unfortunately, the normalizing constant of g^* is ℓ, the estimand, so it is not a practical IS density. However, it provides valuable insight into the structure of good IS densities. In particular, note that,

$$\frac{f(\mathbf{x}) \mathbb{I}(S(\mathbf{x}) > \gamma)}{\ell} = f\left(\mathbf{x} \,|\, S(\mathbf{x}) > \gamma\right)\ .$$

In other words, the optimal IS density, g^* is the original density conditioned on the rare event of interest having occurred. In practice, we usually restrict the IS density g to be a member of a parameterized family of densities $\{g(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. This replaces the infinite-dimensional optimization problem of finding an optimal density with the simpler finite-dimensional problem of finding an optimal vector of parameters $\boldsymbol{\theta}^*$. Even so, it is generally difficult to find a closed-form solution to the *Variance Minimization* (VM) problem

$$\operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_f \frac{f(\mathbf{X})}{g(\mathbf{X}; \boldsymbol{\theta})} \mathbb{I}(S(\mathbf{X}) > \gamma)\ .$$

Instead of solving the VM problem directly, we usually aim to either solve a simpler problem, often using Large Deviations asymptotics, or to 'learn' a good density adaptively.

A.3 The Choice of g

The choice of a good importance sampling density g is highly dependent on the distribution of $\mathbf{X}$ and the properties of the set $\{S(\mathbf{X}) > \gamma\}$. The tail behavior of the $S(\mathbf{X})$ plays an important role in determining the appropriate importance sampling density. A random variable Y is said to be *light-tailed* if $\mathbb{E}e^{\theta Y} < \infty$ for some $\theta > 0$. Light-tailed random variables have tails that decay at least exponentially fast. A random variable that is not light-tailed is said to be *heavy-tailed*. The rare-event behavior of heavy-tailed random variables is considerably different from the behavior of light-tailed random variables. The theory of rare-event simulation for heavy tails is reviewed in Asmussen and Glynn [2007] and Blanchet and Lam [2011].

Sometimes rare events can happen in more than one way. In this case, choosing a g that increases the likelihood of the rare event happening in a certain way may decrease the likelihood of the rare event happening in another way. This means that the likelihood ratio can take extreme values. In the worst case scenarios, this can even lead to estimators with asymptotically infinite variance, as shown in Glasserman and Wang [1997]. In such cases, the appropriate importance sampling density may be a mixture distribution. The use of a mixture distribution may be necessary in some multifactor models, see Glasserman et al. [2007] for a discussion.

In a light-tailed setting, the best importance sampling density is often an *exponentially twisted* density, $f_{\boldsymbol{\theta}}$, derived from the original density f. This density, $f_{\boldsymbol{\theta}}$ is defined as

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\left\{\boldsymbol{\theta}^\intercal \mathbf{x} - \kappa(\boldsymbol{\theta})\right\} f(\mathbf{x}) \,,$$

where

$$\kappa(\boldsymbol{\theta}) = \log \mathbb{E} \exp\left\{\boldsymbol{\theta}^\intercal \mathbf{X}\right\}$$

is the *cumulant generating function* of $\mathbf{X}$. The likelihood ratio of an exponentially twisted density is given by

$$W(\mathbf{x}) = \exp\left\{\kappa(\boldsymbol{\theta}) - \boldsymbol{\theta}^\intercal \mathbf{x}\right\} .$$

Dembo and Zeitouni [2010] and Bucklew [2004] summarize the many attractive properties of likelihood ratios of this form. For example, if there exists an $\boldsymbol{\nu}$ such that

$$\exp\left\{\kappa(\boldsymbol{\theta}) - \boldsymbol{\theta}^\intercal \mathbf{x}\right\} < \exp\left\{\kappa(\boldsymbol{\theta}) - \boldsymbol{\theta}^\intercal \boldsymbol{\nu}\right\}$$

for all $\boldsymbol{\theta}$ and all $\mathbf{x}$ such that $S(\mathbf{x}) > \gamma$, then this is a uniform bound on the likelihood ratio. The parameter $\boldsymbol{\theta}$ can then be chosen to minimize this upper bound, often leading to asymptotically efficient estimators; see, for example, Bucklew [2004].

A.4 Adaptive Importance Sampling

As discussed, the choice of a good importance sampling density is typically model specific and often involves heavy analysis. It is therefore desirable to have an effective way to locate a good importance sampling density in an automatic fashion. In this section we introduce a popular adaptive importance sampling technique for rare-event probability estimation, namely, the Cross Entropy (CE) method. A book-length treatment of the CE method can be found in Rubinstein

and Kroese [2004], and a recent review is given in Kroese [2011]. An improved variant that shows better performance in various high-dimensional settings is recently proposed in Chan and Kroese [2012]. See also Chan, Glynn, and Kroese [2011] for a comparison between the CE and VM methods.

To motivate the CE method, recall that the zero-variance IS density for estimating ℓ is the conditional density given the rare event, i.e.,

$$g^*(\mathbf{x}) = \ell^{-1} f(\mathbf{x}) \mathbb{I}(S(\mathbf{x}) > \gamma).$$

This suggests a practical way to obtain a good importance sampling density. Specifically, if g is chosen to be 'close enough' to g^* so that both behave similarly, the resulting importance sampling estimator should have reasonable accuracy. Therefore, our goal is to locate a convenient density that is, in a well-defined sense, 'close' to g^*.

Now, we formalize this strategy as an optimization problem as follows. Consider the family of density function $\mathcal{G} = \{g(\mathbf{x}; \boldsymbol{\theta})\}$ indexed by the parameter vector $\boldsymbol{\theta}$ within which to obtain the optimal IS density g. One particularly convenient directed divergence measure of densities g_1 and g_2 is the *Kullback–Leibler divergence*, or *cross-entropy distance*:

$$\mathcal{D}(g_1, g_2) = \int g_1(\mathbf{x}) \log \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} \mathrm{d}\mathbf{x} \; .$$

We locate the density g such that $\mathcal{D}(g^*, g)$ is minimized. Since every density in $\mathcal{G}$ can be represented as $g(\cdot; \boldsymbol{\theta})$ for some $\boldsymbol{\theta}$, the problem of obtaining the optimal IS reduces to the following parametric minimization problem:

$$\boldsymbol{\theta}^*_{\text{ce}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{D}(g^*, g(\cdot; \boldsymbol{\theta})) \; .$$

Further, it can be shown that solving the CE minimization problem is equivalent to finding

$$\boldsymbol{\theta}^*_{\text{ce}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E} f(\mathbf{X}) \mathbb{I}(S(\mathbf{X}) \geqslant \gamma) \log g(\mathbf{X}; \boldsymbol{\theta}) \; . \tag{13}$$

The deterministic problem (13) typically does not have an explicit solution. Instead, we can estimate $\boldsymbol{\theta}^*_{\text{ce}}$ by finding

$$\widehat{\boldsymbol{\theta}}^*_{\text{ce}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(S(\mathbf{X}_i) \geqslant \gamma) \log g(\mathbf{X}_i; \boldsymbol{\theta}), \tag{14}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are draws from f. If we are able to draw approximately from g^* — e.g., via Markov Chain Monte Carlo methods — we can instead find

$$\widehat{\boldsymbol{\theta}}^*_{\text{ce}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log g(\mathbf{X}_i; \boldsymbol{\theta}) \; , \tag{15}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are drawn approximately from g^*.

A.5 Importance Sampling for Stochastic Processes

Importance sampling is easily extended to a discrete stochastic process, $\mathbf{X} = \{X_n, n = 0, \ldots, N\}$, as long as the conditional densities $f(x_n \,|\, x_1, \ldots, x_{n-1}), n = 1, 2, \ldots$ are known. A natural importance sampling approach is to simply replace these conditional densities with other conditional densities $g(x_n \,|\, x_1, \ldots, x_{n-1})$, $n = 1, 2, \ldots$. The likelihood ratio is then given by

$$W(\mathbf{x}) = \prod_{n=1}^{N} \frac{f(x_n \,|\, x_1, \ldots, x_{n-1})}{g(x_n \,|\, x_1, \ldots, x_{n-1})} \; .$$

It is less straightforward to apply importance sampling to a continuous-time process, $\mathbf{X} = \{X_t, 0 \leqslant t \leqslant T\}$. The idea is to use the identity

$$\mathbb{E}_{\mathbb{P}} S(\mathbf{X}) = \mathbb{E}_{\mathbb{Q}} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} S(\mathbf{X}) \; ,$$

where $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ is the *Radon-Nikodym derivative*, S is an arbitrary real-valued function and $\mathbb{P}$ and $\mathbb{Q}$ are equivalent measures. This allows us to affect a change of measure similar to that used in discrete setting importance sampling. We note that the stochastic process $\{(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})_t, 0 \leqslant t \leqslant T)$ is a positive martingale. Often, instead of defining $\mathbb{Q}$ explicitly, a positive martingale $\{M_t, 0 \leqslant t \leqslant T\}$ is specified instead. This induces a new measure $\mathbb{Q}$ via Girsanov's theorem. See, for example, Protter [2005] for an in-depth treatment. Examples of specifying a positive martingale and working out the corresponding dynamics of $\mathbb{Q}$ can be found in Bassamboo and Jain [2006], Zhang et al. [2009] and Giesecke and Shkolnik [2011]. A discussion of change of measure for affine jump diffusions, which are of particular importance in credit risk modeling, can be found in Duffie et al. [2000].

References

P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

S. Asmussen and P. W. Glynn. *Stochastic Simulation*. Springer-Verlag, New York, 2007.

A. Bassamboo and S. Jain. Efficient importance sampling for reduced form models in credit risk. In L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, editors, *Proceedings of the 2006 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, inc., 2006.

A. Bassamboo, S. Juneja, and A. Zeevi. Portfolio credit risk with extremal depdendence: Asymptotic analysis and efficient simulation. *Operations Research*, 56(3):593–606, 2008.

J. Blanchet and H. Lam. Rare event simulation techniques. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2011 Winter Simulation Conference*, pages 146–160. Institute of Electrical and Electronics Engineers, inc., 2011.

C. Bluhm, L. Overbeck, and C. Wagner. *Introduction to Credit Risk Modeling: Second Edition*. Chapman & Hall / CRC financial mathematics series, Boca Raton, 2010.

D. Brigo, A. Pallavicini, and R. Torresetti. Credit models and the crisis, or: How I learned to stop worrying and love the CDOs. Working paper, Imperial College, London, 2009. URL `http://ssrn.com/abstract=1529498`.

J. A. Bucklew. *Introduction to Rare Event Simulation*. Springer-Verlag, New York, 2004.

J. C. C. Chan and D. P. Kroese. Efficient estimation of large portfolio loss probabilities in t-copula models. *European Journal of Operational Research*, 205(2):361 – 367, 2010.

J. C. C. Chan and D. P. Kroese. Improved cross-entropy method for estimation. *Statistics and Computing*, 22(5):1031–1040, 2012.

J. C. C. Chan, P. W. Glynn, and D. P. Kroese. A comparison of cross-entropy and variance minimization strategies. *Journal of Applied Probability*, 48A:183–194, 2011.

U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, Chichester, England, 2004.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications: 2nd Edition*. Springer-Verlag, New York, 2010.

C. Donnelly and P. Embrechts. The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. *ASTIN Bulletin*, 40(1):1–33, 2010.

D. Duffie. Credit risk modeling with affine processes. *Journal of Banking and Finance*, 29:2751 – 2802, 2005.

D. Duffie, J. Pan, and K. Singleton. Transform analysis and asset pricing for affine jump diffusions. *Econometrica*, 68(6):1343 – 1376, 2000.

D. Duffie, D. Filipovic, and W. Schachermayer. Affine processes and applications in finance. *The Annals of Applied Probability*, 13(3):984–1053, 2003.

D. Egloff and M. Leippold. Quantile estimation with adaptive importance sampling. *Annals of Statistics*, 37:451–457, 2010.

R. Frey and A. J. McNeil. Modelling depedent defaults. ETH Zentrum, 2001. URL `http://e-collection.ethbib.ethz.ch/show?type=bericht&nr=273`.

K. Giesecke. Credit risk modeling and valuation: An introduction. In D. Shimko, editor, *Credit Risk Models and Management, Vol. 2*. John Wiley & Sons, New York, 2004.

K. Giesecke. Portfolio credit risk: Top down vs. bottom up approaches. In R. Cont, editor, *Frontiers in Quantitative Finance: Credit Risk and Volatility Modeling*. John Wiley & Sons, 2008.

K. Giesecke and A. Shkolnik. Importance sampling for event timing models. Working paper, Stanford, 2011. URL `www.stanford.edu/dept/MSandE/cgi-bin/people/faculty/giesecke/pdfs/is.pdf`.

K. Giesecke, H. Kakavand, and M. Mousavi. Exact simulation of point processes with stochastic intensities. *Operations Research*, 59(5):1233–1245, 2011.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York, 2004.

P. Glasserman and J. Li. Importance sampling for portfolio credit risk. *Management Science*, 51 (11):1643–1656, 2005.

P. Glasserman and Y. Wang. Counterexamples in importance sampling for large deviations probabilities. *Annals of Applied Probability*, 7(3):731 – 746, 1997.

P. Glasserman, W. Kang, and P. Shahabuddin. Large deviations of multifactor portfolio credit risk. *Mathematical Finance*, 17:345–379, 2007.

P. Glasserman, W. Kang, and P. Shahabuddin. Fast simulation of multifactor portfolio credit risk. *Operations Research*, 56(5):1200–1217, 2008.

P. W. Glynn. Importance sampling for Monte Carlo estimation of quantiles. In *Proceedings of the Second International Workshop on Mathematical Methods in Stochastic Simulation and Experimental Design*, pages 180–185, 1996.

L. J. Hong and G. Liu. Monte Carlo estimation of value-at-risk, conditional value-at-risk and their sensitivities. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2011 Winter Simulation Conference*, pages 95–107. Institute of Electrical and Electronics Engineers, inc., 2011.

S. Juneja and P. Shahabuddin. Rare-event simulation techniques: An introduction and recent advances. In S. G. Henderson and B. L. Nelson, editors, *Handbook in Operations Research and Management Science, Vol. 13*. North-Holland, 2006.

W. Kang and P. Shahabuddin. Fast simulation for multifactor portfolio credit risk in the t-copula model. In M. E. Kuhl N. M. Steiger F. B. Armstrong and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 1859–1868. Institute of Electrical and Electronics Engineers, inc., 2005.

D. P. Kroese. The cross-entropy method. In *Wiley Encyclopedia of Operations Research and Management Science*. 2011.

D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. John Wiley & Sons, New York, 2011.

D. X. Li. On default correlation: A copula function approach. *Journal of Fixed Income*, 9(4):43 – 54, 2000.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques, Tools*. Princeton University Press / Princeton Series in Finance, Princeton, 2005.

R. B. Nelsen. *An Introduction to Copulas: Second Edition*. Springer-Verlag, New York, 2006.

Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

P. E. Protter. *Stochastic Integration and Differential Equations: 2nd Edition*. Springer-Verlag, New York, 2005.

T. Reitan and K. Aas. A new robust importance-sampling method for measuring value-at-risk and expected shortfall allocations for credit portfolios. *Journal of Credit Risk*, 6(4):113 – 149, 2010.

G. Rubino and B. Tuffin, editors. *Rare Event Simulation using Monte Carlo Methods*. John Wiley & Sons, New York, 2009.

R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, New York, 2004.

L. Sun and L. J. Hong. Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk. *Operations Research Letters*, 38(4):246–251, 2010.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

X. Zhang, P. Glynn, K. Giesecke, and J. Blanchet. Rare event simulation for a generalized Hawkes process. In M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, editors, *Proceedings of the 2009 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, inc., 2009.