Importance Sampling and Stratification Techniques for Multivariate Models with Low-Dimensional Structures

by

Yoshihiro Taniguchi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Bruno Rémillard
Professor, Dept. of Decision Sciences, HEC Montreal

Supervisor: Christiane Lemieux
Professor, Dept. of Stats. and Act. Sci.
University of Waterloo

Internal Members: Adam Kolkiewicz
Associate Professor, Dept. of Stats. and Act. Sci.
University of Waterloo
Marius Hofert
Assistant Professor, Dept. of Stats. and Act. Sci.
University of Waterloo

Internal-External Member: Kenneth Vetzal
Associate Professor, School of Accounting and Finance
University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Many problems in finance and risk management involve the computation of quantities related to rare-event analysis. As many financial problems are high-dimensional, the quantities of interest rarely have analytical forms and therefore they must be approximated using numerical methods. Plain Monte Carlo (MC) is a versatile simulation-based numerical technique suitable to high-dimensional problems as its estimation error converges to zero at a rate independent of the dimension of the problem. The weakness of plain MC is the high computational cost it requires to obtain estimates with small variance. This issue is especially severe for rare-event simulation as a very large number, often over millions, of samples are required to obtain an estimate with reasonable precision.

In this thesis, we develop importance sampling (IS) and stratified sampling (SS) schemes for rare-event simulation problems to reduce the variance of the plain MC estimators. The main idea of our approach is to construct effective proposal distributions for IS and partitions of the sample space for SS by exploiting the low-dimensional structures that exist in many financial problems. More specifically, our general approach is to identify a low-dimensional transformation of input variables such that the transformed variables are highly correlated with the output, and then make the rare-event more frequent by twisting the distribution of the transformed variables by using IS and/or SS. In some cases, SS is used instead of IS as SS is shown to give estimators with smaller variance. In other cases, IS and SS are used together to achieve greater variance reduction than when they are used separately. Our proposed methods are applicable to a wide range of problems because they do not assume specific types of problems or distribution of input variables and because their performance does not degrade even in high dimension. Furthermore, our approach serves as a dimension reduction technique, so it enhances the effectiveness of quasi-Monte Carlo sampling methods when combined together.

This thesis considers three types of low-dimensional structures in increasing order of generality and develops IS and SS methods under each structural assumption, along with optimal tuning procedures and sampling algorithms under specific distributions. The assumed low-dimensional structures are as follows: the output takes a large value when at least one of the input variables is large; a single-index model where the output depends on the input variables mainly through some one-dimensional projection; and a multi-index model where the output depends on the input mainly through a set of linear combinations. Our numerical experiments find that many financial problems possess one of the assumed low-dimensional structure. When applied to those problems in simulation studies, our proposed methods often give variance reduction factors of over 1,000 with little additional computational costs compared to plain MC.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my doctoral supervisor, Dr. Christiane Lemieux, who has supported me throughout my thesis work with her encouragement, patient guidance, knowledge and insightful comments. I would also like to thank my thesis committee members Dr. Marius Hofert and Dr. Adam Kolkiewicz for taking time to read my thesis and providing me with suggestions. In addition, I want to thank Dr. Bruno Rémillard and Dr. Ken Vetzal for agreeing to be on my thesis examination committee.

I am grateful to Ethan, David, and Zehao for being wonderful office mates. Many of my research ideas came from talking to them about trivial matters.

My special thanks goes to my best friends Raiyana, Taeho, Andres, Ishraq, and Lichen. I would not have been able to make it through the difficult times without their warm encouragements.

Lastly, I would like to thank my parents for their unconditional love and support throughout my many years in graduate school.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA	analysis of variance
CE	cross entropy
CLT	central limit theorem
CMC	conditional Monte Carlo
CV	control variate
CVIS	control variate with importance sampling
ES	expected shortfall
EV	extreme value
EVIS	extreme value importance sampling
GH	generalized hyperbolic
GHS	Glasserman, Heidelberger and Shahabuddin
GIG	generalized inverse Gaussian
G&L	Glasserman and Li
IG	inverse gamma
IS	importance sampling
ISM	increasing second moment
KLD	KullbackLeibler divergence
LDS	low-discrepancy sequence
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MO	Marshall-Olkin
MVN	multivariate normal
OCIS	optimally calibrated importance sampling
OT	Orthogonal Transformation
PCA	principal component analysis
QMC	quasi-Monte Carlo
RE	relative error
RQMC	randomized quasi-Monte Carlo
SIR	Sampling/Importance Resampling
SIS	stratified importance sampling
SS	stratified sampling
STD	standard
UIS	uniform importance sampling
VaR	Value-at-Risk
VM	variance minimizing
VRF	variance reduction factor
VRT	variance reduction technique

List of Symbols

Φ The distribution function of standard normal distribution.

Φ^{-1} The quantile function of standard normal distribution.

t_ν The distribution function of a t-distribution with ν degrees of freedom.

t_ν^{-1} The quantile function of a t-distribution with ν degrees of freedom.

I_d The $d * d$ identity matrix.

$\overset{D}{\rightarrow}$ Convergence in distribution.

Chapter 1

Introduction

Many problems in finance and risk management involve rare-event analysis. For instance, one may want to compute risk measures such as Value-at-Risk or expected shortfall of a portfolio. As many financial problems are high-dimensional and copula models are often used to model dependence for this type of problems, the quantities that we are interested in rarely have analytical forms and therefore they need to be approximated using some numerical methods. This thesis is concerned with developing estimation techniques for high-dimensional rare-event simulation problems.

Plain Monte Carlo (MC) is a versatile simulation-based numerical technique suitable to high-dimensional problems as its estimation error converges to zero at a rate independent of the dimension of the problem. This is in contrast to more traditional, deterministic numerical methods such as quadrature rules based on tensor products as these methods suffer from the *curse-of-dimensionality*, the phenomenon that the approximation accuracy deteriorates exponentially fast with the dimension of the problem. Unfortunately, plain MC is not a cure-all method. Plain MC is notorious for the high computational cost it requires to obtain estimates with a small error. The estimation error of plain MC converges to zero at the rate of the square root of the number of samples. One needs, for instance, 100 times as many samples to obtain an estimate with one more digit of accuracy, making it computationally demanding to obtain estimates with small errors. This issue is especially severe for rare-event simulation as a very large number, often over millions, of samples are required to obtain an estimate with acceptable precision. This is because one generally needs hundreds to thousands of samples to obtain reasonable estimates, but when the nature of problem is a rare-event, a lot more, frequently hundreds to thousands times more, samples need to be generated to gain the said number of observations of rare event. In plain MC, the size of estimation error is closely connected with the variance of the estimator. The estimation error is likely to be small when the estimator has a small variance. Thus, plain MC is often combined with variance reduction techniques (VRTs) (see [76, Ch. 4] for comprehensive coverage of VRTs) to reduce the variance of the estimator. Plain MC with and without VRTs has been applied to a variety of problems in finance such as security pricing [13, 14, 16, 54, 80] and portfolio risk measurement [15, 43, 44, 73, 77].

Importance sampling (IS) [9, 63] and stratified sampling (SS) [21] are VRTs frequently applied to rare-event simulation. The setup of rare-event simulation that we assume in this thesis is that the goal is to estimate $\mu = \mathrm{E}[\Psi(\boldsymbol{X})]$, where $\boldsymbol{X}$ is a d-dimensional random vector with support $\boldsymbol{\Omega}_{\boldsymbol{X}}$ and pdf $f_{\boldsymbol{X}}(\boldsymbol{x})$, and $\Psi : \mathbb{R}^d \to \mathbb{R}$ is such that $\mathbb{P}(\Psi(\boldsymbol{X}) > 0)$ is small. In IS, instead from the original distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$, samples are generated from a proposal distribution denoted by $g_{\boldsymbol{X}}(\boldsymbol{x})$, a distribution constructed in such a way that it gives heavier likelihood to the rare-event region $\{\boldsymbol{x} \in \Omega_{\boldsymbol{X}} \mid \Psi(\boldsymbol{x}) > 0\}$ than the original distribution does. SS starts by partitioning the domain $\Omega_{\boldsymbol{X}}$ into M disjoint strata $\Omega_{\boldsymbol{X}}^{(1)}, \dots, \Omega_{\boldsymbol{X}}^{(M)}$. SS then separately estimates strata means $m_k = \mathrm{E}[\Psi(\boldsymbol{X}) \,|\, \Omega_{\boldsymbol{X}}^{(k}]$, $k = 1, \dots, M$ and combine them to construct an SS estimate of μ. In a rare-event setting, a small number of strata often cover the entire region of rare event. In such a situation, SS concentrates the sampling effort on the important strata. As the sampling distribution induced by SS can be viewed as a proposal distribution for IS, SS is a way of doing IS in a loose sense. Thus, we focus on the IS side instead of SS in motivating our work. We note that, however, IS and SS can be combined to gain greater variance reduction as done in [38, 39, 40]. Some of the techniques developed in this thesis also combine IS and SS.

If implemented with an effective proposal distribution, IS achieves substantial variance reduction and estimates will be highly precise with a reasonable number of samples. Thus, finding a good proposal distribution is a crucial step in IS. How to construct effective proposal distributions is, unfortunately, problem-specific and no strategy works for all types of problems. This is because the nature of rare-event and what constitutes a good proposal distribution depends on the specific form of Ψ and the distribution of $\boldsymbol{X}$. In the finance and risk management context, the nature of the rare events differs, for instance, when one is dealing with options, equity portfolios, or collateralized debt obligations. Even if we restrict our attention to equity portfolios, the characteristics of rare events change when we assume the Gaussian and t-copula models. Consequently, many existing works assume certain types of financial problems under certain distribution assumptions and they use the structure of the problem to design effective proposal distributions and more broadly the IS schemes.

For pricing path dependent options, Glasserman, Heidelberger, and Shahabuddin (GHS) developed IS and SS methods that shift the mean of the underlying normal variables [38] using IS and then use SS to stratify the sample space along with a certain linear combination of the normal variables. For portfolios consisting of stocks and options, GHS proposed IS and SS methods to efficiently estimate the Value-at-Risk of the portfolio under normal [39] and t-distribution [40] models based on applying exponential twisting [9], a popular IS technique, to alter the distribution of the delta-gamma approximation of the portfolio. GHS then use SS to stratify the sample space along the value of the delta-gamma approximation. To estimate tail probabilities of equity portfolios under generalized hyperbolic [88] marginals with a t-copula assumption, Sak, Hörmann and Leydold [106] propose IS methods that shift the mean of normal variables and change the scale parameter of the chi-square variable. To estimate tail probabilities of credit portfolios under the Gaussian copula models, Glasserman and Li [41] propose IS methods that shift the mean of the normal factor variables and use exponential twisting to alter the default probability of obligors. For t-copula credit portfolio problems, which are equivalent to Gaussian models with a common multiplicative shock variable added to it, Bassamboo et al. [12] propose IS methods that apply exponential twisting to the shock variable and the default probabilities. In the same paper, Bassamboo et al. propose another IS technique where the distribution of the shock variable is altered based on Hazard-Rate Twisting. In the same t-copula setting, Chan and Kroese [20] propose to use conditional Monte Carlo [75, pp. 119-125] to analytically integrate out the shock variable and use IS to change the parameters of the underlying multivariate normal variables.

As the IS techniques we reviewed are designed for specific financial problems under specific distributions, they achieve substantial variance reduction if they are applied to the problems originally designed for. The problem with such specialized techniques is that they may not be applicable to other problems without major modifications, though basic principles may be transferable. For instance, the IS techniques developed for option pricing [38] and credit portfolios [41] both rely on shifting the mean of the normal variables, but they approach the problem of estimating the optimal shift rather differently. Moreover, as copula modelling has become prominent in finance, there are much more choices in the distribution of $\boldsymbol{X}$ other than traditional multivariate normal and t distributions. Thus, designing IS techniques for a specific problem or specific distribution limits the applicability of the designed techniques. The goal of this thesis is to develop IS techniques that can be applied to a wide range of financial problems under flexible distribution assumptions. In particular, we put emphasize on effective IS techniques for copula models.

It has been demonstrated analytically and quantitatively (see [18, 118, 119, 120]) that various high-dimensional financial problems have low-dimensional structures. Building on this observation, we focus on certain low-dimensional structures and develop IS techniques that exploit the assumed low-dimensional structure of the problem. More specifically, our proposed IS techniques transform the problem so that only a few leading variables are very important, and then apply IS only to these most important variables. In some cases, SS is used instead of IS to twist the distribution of the important variables as SS is shown to give estimators with smaller variance. In other cases, IS and SS are used together to form stratified importance sampling (SIS) to achieve greater variance reduction than when they are used separately, following the ideas in [38, 39, 40]. To our knowledge, not many existing IS techniques focus on the low-dimensional structure of the problems. The work closest to ours is [38]. In [38], IS and SS are used to exploit the linear and quadratic part of the payoff function of an option, and we can view such dependence of the payoff function on the input variables as some form of low-dimensional structures. In fact, we show that the techniques developed in [38] can be viewed as a special case of the IS and SS techniques developed in this thesis. The benefit of developing IS based on the low-dimensional structure of the problem is that it encapsulates the exact way in which Ψ depends on $\boldsymbol{X}$. Thus, as long as the financial problems we consider have the assumed low-dimensional structure, we can apply our IS techniques in a very similar way, whether the problem is about option pricing or estimating a tail probability of a portfolio. Moreover, since we apply IS only to a few important variables, our IS techniques are less susceptible to the dimensionality problem of IS discussed in [10, 66, 107]. This means that the performance of our IS techniques do not degrade even if they are applied to high-dimensional problems. While our IS techniques do not assume a specific distribution of $\boldsymbol{X}$, how to samples from the proposal distribution depends on the distribution of $\boldsymbol{X}$. For many distributions, the sampling algorithm can be easily implemented. In this thesis we develop sampling algorithms for Archimedean and generalized hyperbolic skew-t copulas which contain the Gaussian and t-copulas as special cases. We believe that similar sampling algorithms can be developed for different distributions.

Quasi-Monte Carlo (QMC) (see [31, 75, 90]) method is a simulation-based numerical method similar to MC. Instead of drawing samples based on pseudo-random numbers as done in MC, QMC draws samples based on a low-discrepancy sequence, a sequence that produces more uniform sample structure than pseudo-random numbers do. QMC has been applied to various high-dimensional financial problems and gave superior results than plain MC [3, 62, 92, 119]. It is widely accepted that the performance of QMC is largely influenced by the effective dimension of the problem, a concept first introduced by Caflisch, Morokoff, and Owen [18]. More precisely, QMC works significantly better than plain MC if the problem has a low effective dimension (see [18, 118, 119, 120]). Thus, QMC is often combined with dimension reduction techniques, which are techniques aimed at reducing the effective dimension of the problem. Such techniques include Brownian bridge (see [18]), principal component analysis (see [3]), the orthogonal transformation of Wang and Sloan [121], and the linear transformation of Imai and Tan [58]. One notion of effective dimension is truncation dimension (see [120]). Essentially, a problem has a low truncation dimension when only a small number of leading input variables are important. Recalling that our IS techniques transform the problem so that only a few leading variables are very important, our techniques work as a dimension reduction technique. Thus, the synergy between our IS techniques and QMC is of great interest in this thesis. In our simulation studies, we apply our IS techniques with and without QMC and empirically analyze how they work together.

The success of our proposed IS and SS methods depends on whether or not the problem at hand possesses one of the assumed low-dimensional structures. Thus, we investigate a wide variety of financial problems in simulation studies of this thesis: Asian and rainbow Asian option pricing under the Black-Scholes framework, basket option pricing under a t-copula model, computation of the loss probabilities of credit portfolios under the Gaussian and t-copula assumptions, and estimation of Value-at-Risk and expected shortfall of equity portfolios under Archimedean and skew-t copulas. It turns out that all these problems have one of the assumed structures. When applied to those problems in simulation studies, our proposed methods often give variance reduction factors of over 1,000 with little additional computational costs compared to plain MC.

The rest of this thesis is organized follows. In Chapter 2, we give the necessary background for this thesis. In particular, we give a brief introduction to MC and QMC methods, some background on IS and SS, provide some background on copulas, and discuss the properties of IS estimators of Value-at-Risk and expected shortfall [88]. In Chapter 3, we develop IS and SS schemes under the assumption that Ψ takes a large value when at least one component of $\boldsymbol{X}$ is large. Such problems often arise from dependence models in the realm of finance and insurance. Explicit sampling algorithms are presented for the case of Archimedean copulas. The optimal calibration for proposal distribution for IS and the optimal sample allocation for SS are derived by minimizing the variance of the respective estimators. Chapter 4 is the main chapter of this thesis. We develop IS and SS methods for single-index models (see [46], [57], and [95]), where Ψ depends on $\boldsymbol{X}$ mainly through some parametric one-dimensional transformation. In simulation studies, we investigate five problems in finance (two for option pricing, two for credit portfolio, and one for equity portfolio) and find that all the problems considered have the assumed low-dimensional structure. The optimal calibration for proposal distributions for IS and SIS are derived by minimizing the variance of the corresponding estimators. Explicit sampling algorithms for the case of generalized hyperbolic skew-t copulas are presented. The application of the proposed IS methods to credit portfolio problems suggest that the optimally calibrated IS method struggles when multiple tail portabilities need to be estimated in one simulation run. In order to develop IS methods that can better handle multiple estimation for problems with a structure based on the single-index model, we explore the use of the extreme value and uniform distribution in Chapter 5. In Chapter 6, we develop IS methods for a multi-index model where $\Psi(\boldsymbol{X})$ depends on $\boldsymbol{X}$ mainly through a set of $p \leq d$ linear combinations, a structure closely related to the one studied for sufficient dimension reduction (see [4, 22, 24] for an overview of this field fo research). We propose the use of parametric and nonparametric proposal distributions for this setup and develop calibration techniques based on the cross-entropy method [26, 102, 103]. A sampling algorithm for the multivariate normal model is presented.

Some of the very important material in Chapter 3 was originally developed by Arbenz, Cambou and Hofert in the preprint [7]. More specifically, the motivation for IS, the general IS framework, and the derivation of the IS weight function given by Theorem 3.3.1 are due to Arbenz, Cambou and Hofert. These correspond to Section 3.2 – Section 3.3.1 of this thesis. Our contributions were to develop the sampling algorithm for Archimedean copulas, propose the use of SS, the variance analysis of the IS and SS estimators, derive the optimal calibration for IS and SS based on the variance expressions and carry out our simulation study. These correspond to Section 3.3.2 – Section 3.6 of this thesis. The work [7] was never published, instead we decided to join our work and published [8] with them.

Chapter 2

Background

In this chapter, we present the background necessary for the rest of this thesis. When introducing Monte Carlo, quasi-Monte Carlo and variance reduction techniques, we mainly follow the notation in [74]

2.1 Plain Monte Carlo Simulation

Suppose that given a function $\Psi : \mathbb{R}^d \to \mathbb{R}$ and a d-dimensional random vector $\boldsymbol{X}$ whose domain, distribution function, and pdf are $\Omega_{\boldsymbol{X}} \subseteq \mathbb{R}^d$, $F_{\boldsymbol{X}}(\boldsymbol{x})$, and $f_{\boldsymbol{X}}(\boldsymbol{x})$, respectively, we want to evaluate the expectation

$$\mu = \mathrm{E}[\Psi(\boldsymbol{X})] = \int\limits_{\Omega_{\boldsymbol{X}}} \Psi(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{2.1}$$

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ be n independent samples from $f_{\boldsymbol{X}}(\boldsymbol{x})$. Then the plain MC estimator for μ is $\hat{\mu}_{\mathrm{MC},n} = \frac{1}{n} \sum\limits_{i=1}^{n} \Psi(\boldsymbol{X}_i)$. This is an unbiased estimator for μ since

$$\mathrm{E}[\hat{\mu}_{\mathrm{MC},n}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \Psi(\boldsymbol{X}_i)\right] = \mathrm{E}[\Psi(\boldsymbol{X})] = \mu. \tag{2.2}$$

Furthermore, the Strong Law of Large Numbers (see [70, p. 101]) assures that $\hat{\mu}_{\mathrm{MC},n}$ converges to μ as $n \to \infty$ with probability 1.

Although $\hat{\mu}_{\mathrm{MC},n}$ converges to μ as $n \to \infty$, one can draw only a finite number of samples in practice. So, $\hat{\mu}_{\mathrm{MC},n}$ has some approximation error, and it is important to quantify the size of the error. The Central Limit Theorem (CLT) [70, p. 134] allows us to derive a probabilistic error bound in the form of a confidence interval (CI). The CLT states that

$$\sqrt{n}\,\frac{\hat{\mu}_{\mathrm{MC},n} - \mu}{\sigma} \xrightarrow{D} N(0,1), \tag{2.3}$$

where σ denotes the standard deviation of $\Psi(\boldsymbol{X})$. The variance (square of standard deviation) of $\Psi(\boldsymbol{X})$ is given by

$$\sigma^2 = \mathrm{E}[(\Psi(\boldsymbol{X}) - \mu)^2] = \int\limits_{\Omega_{\boldsymbol{X}}} (\Psi(\boldsymbol{x}) - \mu)^2 f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{2.4}$$

Then the approximate $100(1-\alpha)\%$ CI for μ is

$$\left(\hat{\mu}_{\mathrm{MC},n} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},\, \hat{\mu}_{\mathrm{MC},n} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right), \tag{2.5}$$

where z_α denotes the point at which $\mathbb{P}(Z \le z_\alpha) = \alpha$ for $Z \sim N(0,1)$. Since σ is unknown in practice, it is replaced with the sample standard deviation

$$\hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (\Psi(\boldsymbol{X}_i) - \hat{\mu}_{\mathrm{MC},n})^2}. \tag{2.6}$$

The estimation error of a plain MC estimator converges to 0 at the rate $O(\frac{1}{\sqrt{n}})$. Notice that the convergence rate is independent of d, the dimension of the problem. This is one of the main reasons why MC is preferred over traditional deterministic numerical schemes for high-dimensional problems. Deterministic methods based on the tensor product of one-dimensional quadrature rules suffer from what is called the *curse of dimensionality*, the phenomenon that the rate of convergence deteriorates exponentially fast with d.

While MC offers a convergence rate that does not depend on d, the $O(\frac{1}{\sqrt{n}})$ convergence often makes it computationally very expensive to obtain precise results. For instance, one needs 100 times as many evaluations of Ψ to obtain an estimate with one more digit of accuracy. Thus, we often use some kind of technique to improve the accuracy of the estimators. Recalling that the half width of the CI for plain MC estimators is proportional to $\frac{\sigma}{\sqrt{n}}$, there are in general two ways to reduce the size of this quantity: reduce the size of the numerator by using variance reduction techniques or improve the $O(\frac{1}{\sqrt{n}})$ convergence rate using quasi-Monte Carlo. We explain the two VRTs, stratified sampling and importance sampling , used in this thesis in the following section. We give a brief introduction to QMC in Section 2.3.

2.2 Variance Reduction Techniques

VRTs generally transform the problem in such a way that the estimator for the new problem has the same expectation but smaller variance. The first property ensures that the estimator based on variance reduction techniques has the correct expectation, while the latter property means that the estimator has smaller error bounds (in the form of CI). This thesis develops new techniques for SS and IS, so we review the basics of the two methods. Readers are referred to [76] for comprehensive coverage of VRTs, and to [21] and [9] for in-depth coverage of SS and IS, respectively.

2.2.1 Stratified Sampling

The main idea of SS is to partition the domain $\Omega_{\boldsymbol{X}}$ of $\boldsymbol{X}$ into M disjoint strata $\Omega_{\boldsymbol{X}}^{(1)}, \dots, \Omega_{\boldsymbol{X}}^{(M)}$ and estimate the strata means separately. Let $p_k = \mathbb{P}(\boldsymbol{X} \in \Omega_{\boldsymbol{X}}^{(k)})$ and $m_k = \mathrm{E}[\Psi(\boldsymbol{X}) | \boldsymbol{X} \in \Omega_{\boldsymbol{X}}^{(k)}]$. Then we can write

$$\mu = \mathrm{E}[\Psi(\boldsymbol{X})] = \sum_{k=1}^{M} p_k \mathrm{E}[\Psi(\boldsymbol{X}) | \boldsymbol{X} \in \Omega_{\boldsymbol{X}}^{(k)}] = \sum_{k=1}^{M} p_k m_k. \tag{2.7}$$

Suppose we know how to draw samples from each stratum and we draw n_k samples from $\Omega_{\boldsymbol{X}}^{(k)}$. Then the stratified sampling estimator is

$$\hat{\mu}_{\mathrm{SS},n} = \sum_{k=1}^{M} p_k \hat{m}_k, \tag{2.8}$$

where $\hat{m}_k$ is the estimate of strata mean based on n_k samples

$$\hat{m}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\Psi(\boldsymbol{X}_i^{(k)}), \quad \boldsymbol{X}_i^{(k)} \overset{\text{ind.}}{\sim} \boldsymbol{X} \,|\, \Omega_{\boldsymbol{X}}^{(k)}.$$

The variance of the SS estimator is given by

$$\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}) = \sum_{k=1}^{M}\frac{p_k^2}{n_k}v_k^2, \tag{2.9}$$

where $v_k^2 = \mathrm{Var}(\Psi(\boldsymbol{X}) \,|\, \boldsymbol{X} \in \Omega_{\boldsymbol{X}}^{(k)})$ is the stratum variance.

The efficiency of SS depends on how strata $(\Omega_1, \ldots, \Omega_M)$ are designed and how sample allocation $(n_1, \ldots, n_M)$ is chosen. Suppose for a moment that the strata are already constructed. We then want to choose the sample allocation such that $\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{MC},n})$. Proportional allocation gives samples proportionally to the stratum probability, that is, $n_k = np_k$. The variance of the SS estimator under the proportional allocation is

$$\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}^{\mathrm{prop}}) = \frac{1}{n}\sum_{k=1}^{M}p_k v_k^2. \tag{2.10}$$

Let K be a random variable that takes values in $\{1, \ldots, M\}$ with the probabilities $\mathbb{P}(K = k) = p_k$, $k = 1, \ldots, M$. Then we can write

$$\begin{aligned} n\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}^{\mathrm{prop}}) &= \mathrm{E}[\mathrm{Var}(\Psi(\boldsymbol{X}) \,|\, \Omega_{\boldsymbol{X}}^{(K)})] \leq \mathrm{E}[\mathrm{Var}(\Psi(\boldsymbol{X}) \,|\, \Omega_{\boldsymbol{X}}^{(K)})] + \mathrm{Var}(\mathrm{E}[\Psi(\boldsymbol{X}) \,|\, \Omega_{\boldsymbol{X}}^{(K)}]) \\ &= \mathrm{Var}(\Psi(\boldsymbol{X})) = n\mathrm{Var}(\hat{\mu}_{\mathrm{MC},n}). \end{aligned} \tag{2.11}$$

So, the variance of the SS estimator under proportional allocation is always equal to or smaller than the variance of the plain MC estimator, regardless of the choice of the strata, and the equality occurs only if $\mu_1 = \cdots = \mu_M$. The optimal allocation [21, pp. 98-99], often called Neyman allocation, can be found by minimizing (2.9) subject to $n_1 + \cdots + n_M = n$ as

$$n_k = np_k\sigma_k \Big/ \sum_{k=1}^{M}p_k\sigma_k. \tag{2.12}$$

and the variance of the SS estimator under Neyman allocation is

$$\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}^{\mathrm{Ney}}) = \frac{1}{n}\left(\sum_{k=1}^{m}p_k\sigma_k\right)^2, \tag{2.13}$$

and by Jensen's inequality $\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}^{\mathrm{Ney}}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}^{\mathrm{prop}})$, where the equality holds only when $\sigma_1 = \cdots = \sigma_M$.

As for the choice of strata, (2.11) implies that

$$\mathrm{Var}(\hat{\mu}_{\mathrm{MC},n}) - \mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}^{\mathrm{prop}}) = \frac{1}{n}\mathrm{Var}(\mathrm{E}[\Psi(\boldsymbol{X}) \,|\, \Omega_{\boldsymbol{X}}^{(K)}]),$$

so the efficiency gain from SS with proportional allocation depends on the variation among strata means. The ideal stratification is such that $\Psi(\boldsymbol{X})$ behaves homogeneously within each stratum so that $\mathrm{E}[\sigma_K^2]$ is small but behaves heterogeneously among different strata so that $\mathrm{Var}(\mu_K)$ is large. How to form such strata is problem-dependent and we do not discuss it here.

2.2.2 Importance Sampling

IS is a variance reduction technique frequently used in rare-event simulation. The typical setting is that $\Psi(\boldsymbol{X})$ is such that $\mathbb{P}(\Psi(\boldsymbol{X}) > 0)$ is small. Let $A = \{\boldsymbol{x} \in \mathbb{R}^d \mid \Psi(\boldsymbol{X}) f_{\boldsymbol{X}}(\boldsymbol{x}) > 0\}$ denote the rare-event region. Under plain MC, most samples give $\Psi(\boldsymbol{X}) = 0$ so they do not contribute in estimating μ. In IS, one draws samples of $\boldsymbol{X}$ from a proposal distribution, a distribution constructed to oversample A. A proposal distribution is often called an IS proposal distribution and we use the two terms interchangeably. Let $g_{\boldsymbol{X}}(\boldsymbol{x})$ denote the pdf of $\boldsymbol{X}$ under the proposal distribution and we assume that $g_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in A$. The IS estimator is obtained based on the following identity

$$\mathrm{E}_f[\Psi(\boldsymbol{X})] = \int\limits_{\Omega_{\boldsymbol{X}}} \Psi(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = \int\limits_{A} \Psi(\boldsymbol{x}) \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{g_{\boldsymbol{X}}(\boldsymbol{x})} g_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = \mathrm{E}_g\left[\Psi(\boldsymbol{X}) w(\boldsymbol{X})\right],$$

where $w(\boldsymbol{x}) = \frac{dF_{\boldsymbol{X}}(\boldsymbol{x})}{dG_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{g_{\boldsymbol{X}}(\boldsymbol{x})}$ is the IS weight function. The $w(\boldsymbol{x})$ term works as a weight so that the estimator remains unbiased after changing the sampling distribution. The subscript f and g on the expectation operator E indicate that expectations are taken with respect to the original and proposal distribution, respectively. The IS estimator is given by

$$\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n} \sum_{i=1}^{n} \Psi(\boldsymbol{X}_i) w(\boldsymbol{X}_i), \quad \boldsymbol{X}_i \overset{\mathrm{ind.}}{\sim} g_{\boldsymbol{X}}. \tag{2.14}$$

Observe that

$$\begin{aligned}\mathrm{E}_g[\Psi^2(\boldsymbol{X})w^2(\boldsymbol{X})] &= \int_{\{\}g_{\boldsymbol{X}}(\boldsymbol{x})>0\}} \Psi^2(\boldsymbol{x})w^2(\boldsymbol{x})g_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_A \Psi^2(\boldsymbol{x})w(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} = \mathrm{E}_f[\Psi^2(\boldsymbol{X})w(\boldsymbol{X})].\end{aligned}$$

So,

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n}\left(\mathrm{E}_f[\Psi^2(\boldsymbol{X})w(\boldsymbol{X})] - \mu^2\right).$$

The variance of the IS estimator is smaller than or equal to that of the plain MC estimator if and only if

$$\mathrm{E}_f[\Psi^2(\boldsymbol{X})w(\boldsymbol{X})] \leq \mathrm{E}_f[\Psi^2(\boldsymbol{X})].$$

Since $\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n})$ depends entirely on the choice of $g_{\boldsymbol{X}}(\boldsymbol{x})$, finding a "good" proposal distribution is the crucial step in IS. Note that

$$\mathrm{E}_g[\Psi^2(\boldsymbol{X})w^2(\boldsymbol{X})] \geq (\mathrm{E}_g[|\Psi(\boldsymbol{X})|w(\boldsymbol{X})])^2 = (\mathrm{E}_f[|\Psi(\boldsymbol{X})|])^2, \tag{2.15}$$

where the inequality holds as an equality if and only if $\Psi^2(\boldsymbol{x}) \propto w^2(\boldsymbol{x})\ \forall \boldsymbol{x} \in A$ by Jensen's inequality. Since $\mathrm{E}_f[|\Psi(\boldsymbol{X})|]$ is a quantity that does not depend on $g_{\boldsymbol{X}}(\boldsymbol{x})$, we can treat the right hand side of (2.15) as a lower bound for the second moment of the IS estimator. The choice

$$g^*_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{|\Psi(\boldsymbol{x})|f_{\boldsymbol{X}}(\boldsymbol{x})}{\int_{\Omega_{\boldsymbol{X}}}|\Psi(\boldsymbol{x})|f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}}, \quad \forall \boldsymbol{x} \in \Omega_{\boldsymbol{X}} \tag{2.16}$$

satisfies the equality condition for (2.15) thus this is the optimal proposal density of $\boldsymbol{X}$. The optimality of $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ can be also proved using calculus of variation as in Kahn and Marshall [63]. If $\Psi(\boldsymbol{x}) \geq 0$ or $\Psi(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \Omega$, then $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ gives a zero-variance estimator. In practice, such an estimator is not attainable as the normalizing constant is unknown. Nonetheless, the form of $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ implies that a "good" proposal density gives larger weights on the region where the product $|\Psi(\boldsymbol{x})|f_{\boldsymbol{X}}(\boldsymbol{x})$ is large.

2.2.3 Dimensionality effect of Importance Sampling

Designing a good IS distribution is challenging in high dimension as the variance of IS estimators grows without bound as the dimension increases unless the proposal distributions are carefully chosen [10, 66, 107]. This is an important issue as many problems in finance are high-dimensional and the goal of this thesis is to develop IS and SS methods that work well for such problems. We review Au and Beck's [10] argument of how the dimensionality of the problem affects the variance of the IS estimators and discuss an approach that alleviates the dimensionality effect. The IS techniques that we develop in this thesis follows this approach so they remain effective even in high dimension, as long as certain low-dimensional structures exist.

Since $\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \mathrm{Var}_g(\Psi(\boldsymbol{X})w(\boldsymbol{X}))/n$, one wants to choose a proposal density $g_{\boldsymbol{X}}(\boldsymbol{x})$ so that $\mathrm{Var}_g(\Psi(\boldsymbol{X})w(\boldsymbol{X}))$ is small. Generally, a proposal distribution that gives large $\mathrm{Var}_g(w(\boldsymbol{X}))$ also gives large $\mathrm{Var}_g(\Psi(\boldsymbol{X})w(\boldsymbol{X}))$ [10], so it is important that the class of IS distributions considered is such that $\mathrm{Var}_g(w(\boldsymbol{X}))$ remains bounded as $d \to \infty$, if the problem of interest is high-dimensional. However, as the analysis in [10] shows, $\mathrm{Var}_g(w(\boldsymbol{X}))$ could grow exponentially in d unless the class of proposal distributions is carefully selected, making the selection of IS distributions delicate for high-dimensional problems.

Suppose for a moment that $\Omega_{\boldsymbol{X}} = \mathbb{R}^d$ for some large d, say 100. Then, if we use any d-dimensional distribution $g_{\boldsymbol{X}}(\boldsymbol{x})$ such that $g_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$, the IS estimator will be unbiased. As discussed earlier, however, such estimators are likely to have very large variance unless $g_{\boldsymbol{X}}(\boldsymbol{x})$ is chosen appropriately. With a poor choice of $g_{\boldsymbol{X}}(\boldsymbol{x})$, the IS estimator constructed based on a practical number of samples will have a large estimation error, even though the estimator is theoretically unbiased. We refer to such estimators as unreliable estimators. We also refer to the estimates that are clearly far from the true value as unreliable estimates.

If $\Psi(\boldsymbol{X})$ depends on a small subset of variables of $\boldsymbol{X}$, one can apply IS only to those important variables. Then the dimensionality problem will not be as severe as when applying IS to the entire set of variables of $\boldsymbol{X}$. Even if $\Psi(\boldsymbol{X})$ does not have such a structure, it may be possible to transform $\Psi(\boldsymbol{X})$ so that the transformed function has the desired structure. The IS techniques that we develop in this thesis implicitly apply such a transformation and then we apply IS only to the important variables.

2.3 Quasi-Monte Carlo

2.3.1 Motivation of QMC

The goal of QMC is the approximation of integrals (or expectations) of the form

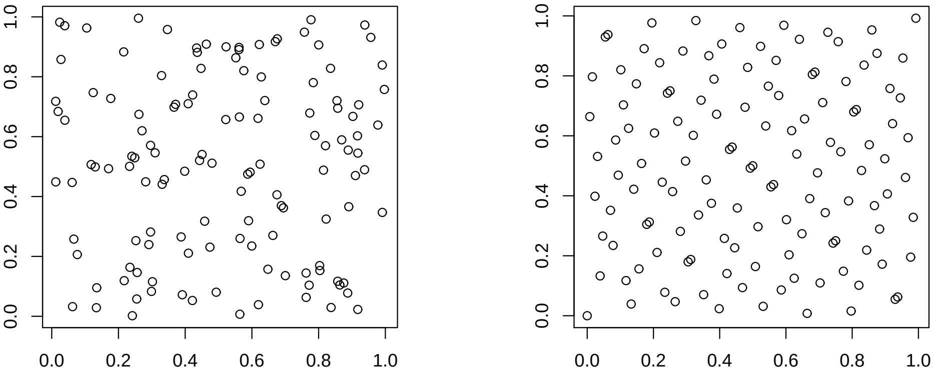
$$\mu = \int_{[0,1)^d} \Psi(\boldsymbol{u}) d\boldsymbol{u} = \mathrm{E}[\Psi(\boldsymbol{U})] \tag{2.17}$$

for $\boldsymbol{U} \sim U[0,1)^d$. It appears restrictive that the domain of integration in (2.17) must be $[0,1)^d$, or equivalently that the expectation assumes that the random vector follows $U[0,1)^d$. However, any integration domain can be transformed to $[0,1)^d$ using a change of variables and any random vector can be generated by transforming $U[0,1)^d$ for large enough d. The formulation (2.17) assumes that such transformations are incorporated in Ψ. The integration domain in (2.17) is assumed to be $[0,1)^d$ instead of $[0,1]^d$ to circumvent possible numerical difficulties. Often time, Ψ is such that $\Psi(\boldsymbol{u}) = \infty$ when any of the component of $\boldsymbol{u}$ is equal to 1. By assuming that the domain is $[0,1)^d$, we avoid accidental evaluation of Ψ at boundaries.

In MC, the samples of $\boldsymbol{U}$ would be generated based on pseudo-random numbers which are designed to mimic the behaviour of samples from $U[0,1)$. The aim of QMC is to improve the $O(1/\sqrt{n})$ convergence rate of MC estimator by replacing pseudo-random numbers with a low-discrepancy sequence (LDS) (see [90]). A LDS offers a more uniform sample structure than pseudo-random numbers do, which leads to a better coverage of the domain. Figure 2.1 compares the plots of 128 samples from two-dimensional pseudo-random numbers (left) and a Sobol sequence (right), a particular construction of LDS. The left figure shows that samples from pseudo-random numbers are not equidistributed; some regions are oversampled and others are undersampled. This poor coverage of the domain partially explains the slow convergence of MC estimators. Intuitively, the estimator is more accurate if the sampled points cover the entire domain more uniformly. QMC takes advantage of such sampling schemes, and can provide faster convergence than MC.

Similarly to the MC estimator, the QMC estimator has the form $\hat{\mu}_{\mathrm{QMC},n} = \frac{1}{n} \sum\limits_{i=1}^{n} \Psi(\boldsymbol{u}_i)$, but the points $\boldsymbol{u}_i$ come from a LDS instead of pseudo-random numbers. The precise definition of LDS and the measure of uniformity will be given in the following section.

Figure 2.1: Plots of 128 samples from two-dimensional pseudo-random numbers and a low-discrepancy sequence



(a) Two dimensional Pseudo-Random Numbers

(b) Two dimensional Sobol sequence

2.3.2 Star-Discrepancy and Error bound of QMC estimate

Again, we follow the notation in [75]. The star discrepancy of a point set P_n is given by

$$D^*(P_n) = \sup_{\boldsymbol{v} \in [0,1)^d} |v_1 \dots v_d - \alpha(P_n, \boldsymbol{v})/n|,$$

where $\alpha(P_n, \boldsymbol{v})$ is the number of points from P_n that are in $\prod_{j=1}^{d} [0, v_j)$. Take a hyper-rectangle H of the form $\prod_{j=1}^{d} [0, v_j)$ and suppose that the volume of H is V. If the point set P_n is truly equidistributed, exactly $V \cdot n$ of all points should lie in H, for all H. In that case, the star-discrepancy of P_n is 0. On the other hand, if all the points lie in a small cluster, the star-discrepancy is close to 1. All sampling schemes are between the two cases. A sequence of points is called a LDS if $D^*(P_n) \in O(n^{-1}(\log n)^d)$. The *Koksma-Hlawka Inequality* [51] relates the star-discrepancy and the error bound of a QMC estimate as

$$|\hat{\mu}_{\text{QMC},n} - \mu| \leq D^*(P_n) V(\Psi), \tag{2.18}$$

provided $V(\Psi)$, *the variation in the sense of Hardy and Krause*, is finite. We can think of (2.18) as the bound on the worst case error of a QMC estimate for functions Ψ with finite *variation* $V(\Psi)$ *in the sense of Hardy and Krause*. Hence, QMC offers approximation error $O(n^{-1}(\log n)^d)$, which is asymptotically smaller than the MC error $O(n^{-\frac{1}{2}})$ for any fixed d, justifying the use of QMC over MC. Notice that, however, this error bound suggests that the accuracy of QMC deteriorates as the dimensionality of a problem increases.

The problem with the error bound (2.18) is that it is virtually impossible to compute since $D^*(P_n)$ and $V(f)$ are both very hard to compute. Even if one is able to compute the error bound, the bound is often too conservative to be useful. Alternatively, one can add randomness to the underlying LDS so that an error bound can be constructed as a CI, much like when one uses MC. This technique is called randomized QMC (RQMC) and is discussed in Section 2.3.4.

2.3.3 Construction of a low-discrepancy sequence

This section introduces two constructions of a LDS. The *van der Corput sequence* is a famous one-dimensional LDS and it forms the basis for many multidimensional constructions. The Sobol sequence [112] can be seen as a type of multidimensional extension of *van der Corput sequence* and this is the QMC point set that we use in this thesis.

van der Corput sequence

Pick some base $b \geq 2$. For a non-negative integer i, compute a sequence $c_l(i)$, $l = 1, 2, \ldots$ through the base b expansion of i as

$$i = \sum_{l=1}^{\infty} c_l(i) b^{l-1}. \tag{2.19}$$

Let ϕ_b denote the *radical inverse function in base b* which is defined as

$$\phi_b(i) = \sum_{l=1}^{\infty} c_l(i) b^{-l}.$$

It is easy to see that $\phi_b(i) \in [0,1)$ and the ith term of *van der Corput sequence* in base b is defined as $\phi_b(i-1)$. The first 10 terms of the sequence with base $b=2$ are 0, 0.5 , 0.25, 0.75, 0.125, 0.635, 0.375, 0.875, 0.0625, 0.5625. Notice that this sequence cover $[0,1)$ more uniformly than a sequence of pseudo random numbers would.

Sobol Sequence

We follow the notation introduced in [75] to describe the construction of the Sobol sequence. The theme of the Sobol sequence is to apply linear transformation to the digits $c_l(i)$ before applying the radical inverse function. For each dimension j, generating Sobol sequence requires a primitive polynomial in $\mathbb{F}_2$, which we denote by $p_j(z)$ and write as

$$p_j(z) = z^{d_j} + a_{d_j,1} z^{d_j-1} + \cdots + a_{j,1},$$

where d_j is the degree of the primitive polynomial and each $a_{j,l}$ is either 0 or 1. We then need d_j *direction numbers* of the form

$$v_{j,r} = \frac{m_{j,r}}{2^r}, \quad r \geq 1$$

where $m_{j,r}$ could be any positive odd integer less than 2^r. Once we have the d_j initial *direction numbers*, we can recursively define the rest as

$$v_{j,k} = a_{j,1} v_{j,r-1} \oplus \cdots \oplus a_{j,d_j} v_{j,r-d_j+1} \oplus v_{j,r-d_j} \oplus (v_{j,r-d_j}/2^{d_j}),$$

where $\oplus$ represents exclusive-or operation. Then $u_{i,j}$, the jth coordinate of the ith point of the Sobol sequence, is given by

$$u_{i,j} = c_1(i) v_{1,j} \oplus c_2(i) v_{1,2} \oplus \cdots ,$$

where the sequence $c_l(i)$, $l = 1, 2, \ldots$ are the coefficients of the base 2 ($b=2$) expansion of i computed as in (2.19).

We emphasize that the quality of the Sobol points heavily depends on the choice of the *direction numbers*. In this thesis, we use the direction numbers from F.Y. Kuo's web page http://web.maths.unsw.edu.au/ fkuo/sobol/, which are originally found by Joe and Kuo using the search algorithm in [60].

2.3.4 Randomized Quasi-Monte Carlo

As mentioned in [75], "randomized quasi-Monte Carlo consists in choosing a deterministic low-discrepancy point set P_n and applying a randomization such that (i) each point $\tilde{\boldsymbol{u}}_i$ in the randomized point set $\tilde{P}_n$ is $U[0,1)^d$ and (ii) the low-discrepancy of P_n is preserved (in some sense) after the randomization". Suppose we already have such randomization scheme and let P_n be a n-point set based on some *low-discrepancy sequence*. Let P_n^l denote a point set from the lth randomization of P_n for $l = 1, \dots m$, that is, we randomize m times.

For $l = 1, \dots, m$, let

$$\hat{\mu}_{l,\text{RQMC}} = \frac{1}{n} \sum_{i=1}^{n} \Psi(\tilde{\boldsymbol{u}}_{l,i}), \tag{2.20}$$

where $\tilde{\boldsymbol{u}}_{l,i}$ denote the ith point in $\tilde{P}_n^l$. Note that $\hat{\mu}_{l,\text{RQMC}}$ is an unbiased estimator of μ since each $\tilde{\boldsymbol{u}}_{l,i} \sim U[0,1)^d$. Randomizing P_n m times, we obtain $\{\hat{\mu}_{1,\text{RQMC}} \dots \hat{\mu}_{m,\text{RQMC}}\}$. Each estimator is an unbiased estimator of μ and they are independent and identically distributed (i.i.d.). This i.i.d. condition allows us to use CLT and derive probabilistic error bounds. Let

$$\hat{\mu}_{\text{RQMC},m} = \frac{1}{m} \sum_{i=1}^{m} \hat{\mu}_{i,\text{RQMC}}.$$

Clearly $\hat{\mu}_{\text{RQMC},m}$ is an unbiased estimator of μ. The approximate $100(1-\alpha)\%$ confidence interval is

$$\{\hat{\mu}_{\text{RQMC},m} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma_{\text{RQMC},m}}{\sqrt{m}}\}.$$

As $\sigma_{\text{RQMC},m}$ is unknown, we replace it with the sample standard deviation

$$\hat{\sigma}_{\text{RQMC},m} = \sqrt{\tfrac{1}{m-1} \textstyle\sum_{i=1}^{m} (\hat{\mu}_{i,\text{RQMC}} - \hat{\mu}_{\text{RQMC},m})^2}.$$

The approximate $100(1-\alpha)\%$ confidence interval becomes

$$\{\hat{\mu}_{\text{RQMC},m} \pm z_{1-\frac{\alpha}{2}} \cdot \tfrac{\hat{\sigma}_{\text{RQMC},m}}{\sqrt{m}}\}.$$

Digital Shift

Digital shift is a simple yet popular randomization technique applicable to Sobol sequence. Suppose we have a QMC point set P_n. Let $\boldsymbol{v} = (v_1, \dots, v_d) \sim U[0,1)^d$ and write the base b expansion of jth component of $\boldsymbol{v}$ as $(v_{j,1}, v_{j,2} \cdots)_b$, that is, we write

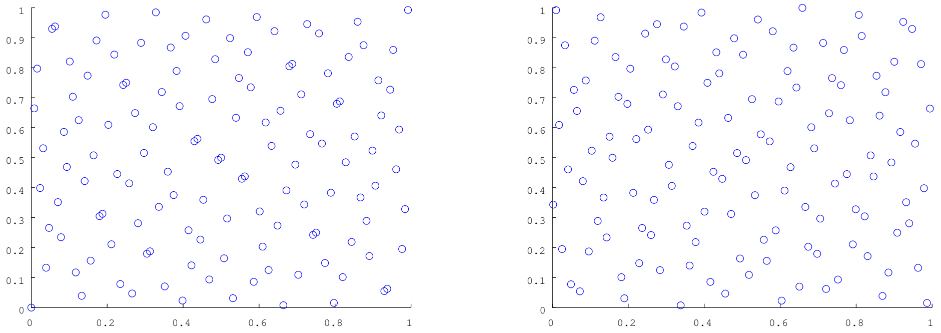
$$v_j = \sum_{l=0}^{\infty} v_{j,l} b^{-l}.$$

Also, let $(u_{i,j,1}, u_{i,j,2}, \cdots)$ represent the base b expansion of jth component of the ith point in P_n. The digitally shifted point set of P_n, which we denote $\tilde{P}_n$, consists of points $\tilde{u}_i$, $i = 1, \dots n$ whose jth component is

$$\tilde{u}_{i,j} = \sum_{l=0}^{\infty} (u_{i,j,l} + v_{j,l}) b^{-l},$$

where the addition is performed in $\mathbb{Z}_b$.

Figure 2.2: Sobol Point Set



(a) Sobol

(b) Digitally Shifted Sobol

Figure 2.2 compares a Sobol point set and a digitally shifted version of it. The shift of $(0.81, 0.91)$ was applied to obtain the point set on the right. As the plot shows, a digital shift transforms a Sobol point set without breaking the uniform structure.

2.3.5 ANOVA Decomposition and Effective Dimension

While the $O(n^{-1}(\log n)^d)$ (worst case) error rate of QMC estimation is asymptotically smaller than the plain MC rate of $O(n^{-\frac{1}{2}})$ for fixed d, the size of n required for $n^{-1}(\log n)^d \leq n^{-\frac{1}{2}}$ to hold can be unrealistically large. For instance, for $d = 10$, n must be at least about 10^{39} for the inequality to hold, as in [75, p. 197]. QMC in this sense appears to suffer from the curse of dimensionality. Nonetheless, QMC has proved to be successful for various high-dimensional problems [3, 62, 92, 119]. A widely accepted explanation of the success of QMC is related to the concept of effective dimension, which was first introduced by Caflisch, Morokoff, and Owen [18]. It has been demonstrated in many examples that QMC works significantly better than plain MC if the problems have low effective dimensions (see for instance [18, 119, 120, 118]). Since the definitions of effective dimension is closely related to the analysis of variance (ANOVA) decomposition [18], we explain ANOVA decomposition first.

ANOVA Decomposition

The ANOVA decomposition of $\Psi(\boldsymbol{u})$ is expressed as

$$\Psi(\boldsymbol{u}) = \sum_{J \subseteq \{1,\ldots,d\}} \Psi_J(\boldsymbol{u}), \tag{2.21}$$

where $\Psi_\emptyset = \mu$ and for a nonempty subset $J \subseteq \{1, \ldots, d\}$, Ψ_J is defined as

$$\Psi_J(\boldsymbol{u}) = \int_{[0,1)^{d-s}} \Psi(\boldsymbol{u}) d\boldsymbol{u}_{-J} - \sum_{K \subset J, K \neq J} \Psi_K(\boldsymbol{u}),$$

where $s = |J|$ and $-J = \{1, \ldots, d\} \setminus J$ is the complement of J. Since

$$\int_{[0,1)^d} \Psi_J(\boldsymbol{u}) \Psi_K(\boldsymbol{u}) d\boldsymbol{u} = 0$$

for all nonempty $J \neq K$, the ANOVA decomposition writes Ψ as a sum of the 2^d orthogonal components.

The significance of ANOVA decomposition is that it decomposes the overall variance $\sigma^2 = \mathrm{Var}(\Psi(\boldsymbol{U})) = \int_{[0,1)^d}(\Psi(\boldsymbol{u}) - \mu)^2 d\boldsymbol{u}$ into the sum of the variance of the ANOVA component as $\sigma^2 = \sum_J \sigma_J^2$, where

$$\sigma_J^2 = \int_{[0,1)^d} \Psi_J^2(\boldsymbol{u}) d\boldsymbol{u}.$$

Sobol's global sensitivity index [113] is defined as $S_J = \frac{\sigma_J^2}{\sigma^2}$ for $J \subseteq \{1, \ldots, d\}$ and it measures the fraction of the variance of Ψ explained by Ψ_J. We can use such indices as the measure of the relative importance of the ANOVA components.

Effective Dimension

Definition 2.3.1. *The effective dimension of Ψ in the truncation sense (truncation dimension) in proportion p is the smallest integer d_{T} such that*

$$\frac{1}{\sigma^2} \sum_{J: J \subseteq \{1,\ldots,d_{\mathrm{T}}\}} \sigma_J^2 \geq p.$$

The effective dimension of Ψ in the superposition sense (superposition dimension) in proportion p is the smallest integer d_{S} such that

$$\frac{1}{\sigma^2} \sum_{J: |J| \leq d_{\mathrm{S}}} \sigma_J^2 \geq p.$$

A truncation dimension of d_{T} indicates that the first d_{T} variables of $\boldsymbol{u}$ explains most of the variation of Ψ. A superposition dimension of d_{T} means that $\Psi(\boldsymbol{u})$ is well approximated by a sum of functions with at most d_{T} variables. This in turn implies that the interaction effects of order larger than d_{T} are not significant.

As noted earlier, the efficiency gain of QMC over MC depends on the effective dimension of the problem. Thus, QMC is often combined with dimension reduction techniques, the techniques aimed at reducing the effective dimension of the problem. Such technique include Brownian bridge (see [18]), principal component analysis (see [3]), the orthogonal transformation of Wang and Sloan [121], and the linear transformation of Imai and Tan [58].

2.4 Copula Models

Since this thesis emphasizes simulation techniques for copula models, we provide a brief introduction to copula in this section. Readers are referred to [88] for a more comprehensive introduction to this topic.

2.4.1 Definitions and Theorems

Definition 2.4.1. *A d-dimensional copula is a distribution function on $[0,1]^d$ with standard uniform marginal distributions.*

The definition of copula above directly gives the following three properties of copula. These properties can be used to define a copula and the two definitions are mathematically equivalent.

Definition 2.4.2. *A d-dimensional mapping $C : [0,1]^d \to [0,1]$ is a copula if*

- $C(u_1, \dots, u_{j-1}, 0, u_{j+1}, \dots u_d) = 0$ *for* $j = 1, \dots, d$
- $C(1 \dots, 1, u_j, 1, 1) = u_j \; j = 1, \dots, d$ *for* $j = 1, \dots, d$
- *For all* $(a_1, \dots, a_d), (b_1, \dots, b_d) \in [0,1]^d$ *with* $a_j \le b_j$

$$\sum_{j_1=1}^{2} \cdots \sum_{j_d=1}^{2} (-1)^{j_1+\dots+j_d} C(u_{1,j_1}, \dots, u_{d,j_d}) \ge 0$$

holds where $u_{j1} = a_j$ *and* $u_{j2} = b_j$ *for all* $j \in \{1, \dots d\}$.

We sometimes refer to a d-dimensional copula as a d-copula. The following theorem due to Sklar [110], states that any multivariate distribution is the composition of a copula and marginals, and is one of the most important theorems in this field.

Theorem 2.4.3 (Sklar 1959)**.** *Let F be a joint distribution function with margins $F_1, \dots F_d$. Then there exists a copula $C : [0,1]^d \to [0,1]^d$ such that, for all $x_1, \dots, x_d \in \bar{\mathbb{R}} = [-\infty, \infty]$,*

$$F(x_1, \dots, x_d) = C(F(x_1), \dots, F(x_d)). \tag{2.22}$$

If the margins are continuous, then C is unique; otherwise C is uniquely determined on Ran $F_1 \times \cdots \times$ Ran F_d, where Ran $F_i = F_i(\bar{\mathbb{R}})$. Conversely, if C is a copula and $F_1, \dots F_d$ are univariate distribution functions, then the function F defined in (2.22) *is a joint distribution with margins $F_1, \dots F_d$.*

The significance of Sklar's Theorem is that it adds a great deal of freedom in modelling joint distributions because the theorem allows us to separate the dependence structure from marginals. With Sklar's Theorem, one can first choose marginal distributions from different parametric families then combine them through a copula of their choice. This two-step procedure is much more flexible than the traditional modelling where full joint distributions have to be specified altogether.

2.4.2 Some well-known copulas

In this section, we introduce some popular copulas in statistical modelling.

Gaussian Copula

The Gaussian copula is the copula underlying a multivariate normal distribution. Suppose $\boldsymbol{X} \sim \mathrm{MVN}(\mathbf{0}, P)$, that is, $\boldsymbol{X}$ follows a multivariate normal (MVN) distribution with the mean vector $\mathbf{0} = (0, \dots, 0)'$ and the correlation matrix P. Then the Gaussian copula C_P^G is defined implicitly as

$$C_P^G(u_1, \dots, u_d) = \Phi_P(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)),$$

where Φ_P denotes the distribution function of a multivariate normal with mean vector $\mathbf{0}$ and the correlation matrix P and Φ^{-1} denotes the quantile function of a univariate standard normal distribution.

Suppose $\boldsymbol{X} \sim \mathrm{MVN}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is the mean vector and Σ is the covariance matrix. If P is the corresponding correlation matrix, the copula of $\boldsymbol{X}$ is C_P^G. The mean vector is irrelevant as it contains no information on the dependence of $\boldsymbol{X}$. The variance of the components of $\boldsymbol{X}$ is also irrelevant for the same reason. The correlation matrix completely captures the dependence of the multivariate normal distribution.

t-copula

Similarly to the Gaussian copula case, the t-copula is the copula underlying the multivariate-t distribution. Suppose $\boldsymbol{X} \sim t_\nu(\mathbf{0}, P)$, that is, $\boldsymbol{X}$ follows the multivariate-t distribution with v degrees of freedom, $\mathbf{0}$ mean vector and correlation matrix P. The t copula $C^t_{v,P}$ is defined as

$$C^t_{\nu,P}(u_1, \dots, u_d) = t_{\nu,P}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)),$$

where $t_{\nu,P}$ denotes the distribution function of a multivariate t distribution with v degrees of freedom, $\mathbf{0}$ mean vector, and the correlation matrix P and t_ν^{-1} denotes the inverse of t_ν, the distribution function of a univariate t-distribution with ν degrees of freedom. Suppose $\boldsymbol{X} \sim t_v(\mu, \Sigma)$. If P is the correlation matrix of Σ, then $C^t_{\nu,P}$ is the copula of X.

Archimedean copula

Unlike the Gaussian and t-copulas where the copulas are implicitly defined, Archimedean copulas are explicitly defined as

$$C(u_1, \dots, u_d) = \psi(\psi^{-1}(u_1) + \dots + \psi^{-1}(u_d)), \quad (u_1, \dots, u_d) \in [0,1]^d, \tag{2.23}$$

where ψ is an Archimedean copula generator, a special univariate function with the following properties;

- $\psi : [0, \infty) \to [0, 1]$ with $\psi(0) = 1$ and $\psi(\infty) = 0$
- $\psi(x)$ is continuous and strictly decreasing on$[0, \psi^{-1}]$.

Satisfying the two conditions above is necessary, but not sufficient for ψ to induce an Archimedean copula. Kimberling's theorem [67] provides the necessary and sufficient condition for a generator to induce an Archimedean copula for any $d \geq 2$.

Theorem 2.4.4 (Kimberling)**.** *Let $\psi : [0, \infty) \to [0, 1]$ be a continuous, strictly decreasing function such that $\psi(0) = 1$ and $\psi(\infty) = 0$. Then ψ induces a copula of any dimension $d \geq 2$ if and only if ψ is completely monotone, that is, $(-1)^k \psi^{(k)}(x) \geq 0$ for $k \geq 1$.*

Bernstein's theorem makes a connection between the notion of complete monotonicity and the Laplace transform of a random variable.

Theorem 2.4.5 (Bernstein)**.** *Let $\psi : [0, \infty) \to [0, 1]$ be a continuous, strictly decreasing function such that $\psi(0) = 1$ and $\psi(\infty) = 0$. Then ψ is completely monotone if and only if ψ is a Laplace transform of a distribution function.*

Combining Kimberling's theorem and Bernstein's theorem, ψ induces an Archimedean copula for any dimension $d \geq 2$ if and only if ψ is a Laplace transform of the distribution function of some positive random variable V, so-called *frailty*. The following algorithm due to Marshall and Olkin [83], to which we refer as Marshall-Olkin algorithm, allows us to efficiently sample from an Archimedean copula C_ψ induced by such generator ψ.

Algorithm 1 Marshall-Olkin Algorithm

```
  Generate $V$ whose Laplace transform is $\psi$
  Generate $E_1, \dots, E_d \overset{\text{ind.}}{\sim} \text{Exp}(1)$
  Let $U_i = \psi\left(\frac{E_i}{V}\right)$ for $i = 1, \dots, d$.
  Return $(U_1, \dots, U_d) \sim C_\psi$.
```

For many popular Archimedean copulas, the frailty random variable V has a known distribution, for instance V is Gamma distributed for Clayton copulas. Table 2.1 lists the information about five popular Archimedean copulas and the corresponding frailty random variables V: see [53, Table 1] for the details concerning Table 2.1.

Table 2.1: Popular Archimedean Copulas

Family	Parameter	$\psi(t)$	V
Ali-Mikhail-Haq	$\theta \in [0, 1)$	$(1-\theta)/(\exp(t)-\theta)$	$Geo(1-\theta)$
Clayton	$\theta \in (0, \infty)$	$(1+t)^{-1/\theta}$	$Gamm(1/\theta, 1)$
Frank	$\theta \in (0, \infty)$	$-\log(1-(1-e^{-\theta})\exp(-t))/\theta$	$Log(1-e^{-\theta})$
Gumbel	$\theta \in [1, \infty)$	$\exp(-t^{1/\theta})$	$S(1/\theta, 1, cos^{\theta}(\pi/(2\theta)), \mathbb{1}\{\theta = 1\}; 1)$
Joe	$\theta \in [1, \infty)$	$1-(1-\exp(-t))^{1/\theta}$	$Sibuya(1/\theta)$

2.5 Risk Measures and Importance Sampling

In this section, we introduce two widely used risk measures, Value-at-Risk (VaR) and expected shortfall (ES), and discuss how to design effective IS distributions to estimate them. We follow the notation in [115] in this section.

2.5.1 VaR and ES

Suppose L is a univariate random variable representing a portfolio loss with distribution function F_L. Let f_L denote the pdf of L if it exists. For $0 \leq \alpha \leq 1$, $100\alpha\%$ Value-at-Risk or VaR_α of L is defined as

$$v = F_L^{-1}(\alpha) = \inf\{x : F_L(x) \geq \alpha\}.$$

That is, VaR_α is the $\alpha\%$ quantile of F_L. Similarly, the $100\alpha\%$ Expected shortfall or ES_α is defined as

$$c = \frac{1}{1-\alpha} \int_\alpha^1 \mathrm{VaR}_u du.$$

Expected shortfall is sometimes called conditional Value-at-Risk. Since $c = \mathrm{E}[L \mid L > v]$ if L has a positive and differential density in the neighbourhood v, we can think of ES as the expected value of a loss given that the loss exceeds the corresponding VaR. In summary, VaR is a quantile of a distribution and ES is a conditional expectation for its tail.

2.5.2 IS estimators of VaR and ES

In risk management, one is usually interested in estimating VaR_α and ES_α for α close to 1. Since this is a rare event simulation, plain MC estimators are not very precise. The idea of applying IS to enhance the precision of the estimates of VaR_α and ES_α is explored in [40, 39] among others. As we apply IS to estimate VaR and ES for copula models in Chapter 3 and Chapter 4, we discuss in this section how to tailor proposal distribution to estimate the said risk measures based on the asymptotic normality results of Sun and Hong [115].

Suppose that the samples $(L_1, \ldots, L_n)$ are generated from a proposal distribution G_L of L. Let $w(x) = \frac{dF_L(x)}{dG_L(x)}$ denote the IS weight function. Define the IS estimate of the empirical distribution of L as

$$\hat{F}_{\mathrm{IS},n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{L_i \le x\} w(L_i).$$

Then the IS estimator of VaR_α and ES_α are given by [115]

$$\hat{v}_{\mathrm{IS},n} = \inf\{x : \hat{F}_{\mathrm{IS},n}(x) > \alpha\} \tag{2.24}$$

and

$$\hat{c}_{\mathrm{IS},n} = \hat{v}_{\mathrm{IS},n} + \frac{1}{n\alpha}\sum_{i=1}^{n}(L_i - \hat{v})^+ w(L_i), \tag{2.25}$$

respectively, where $x^+ = \max\{x, 0\}$. Sun and Hong [115] derive the asymptotic normality of $\hat{v}_{\mathrm{IS},n}$ and $\hat{c}_{\mathrm{IS},n}$ (under some conditions that include the existence of the density of L at v) as

$$\sqrt{n}(\hat{v}_{\mathrm{IS},n} - v) \xrightarrow{D} \frac{\sqrt{\mathrm{Var}_G\left(\mathbb{1}\{L > v\}w(L)\right)}}{f_L(v)} N(0,1) \tag{2.26}$$

$$\sqrt{n}(\hat{c}_{\mathrm{IS},n} - c) \xrightarrow{D} \frac{\sqrt{\mathrm{Var}_G\left((L - v)^+ w(L)\right)}}{\alpha} N(0,1). \tag{2.27}$$

Observe from (2.26) and (2.27) that the asymptotic variance of $\hat{v}_{\mathrm{IS},n}$ and $\hat{c}_{\mathrm{IS},n}$ respectively depends on the proposal distribution through $\mathrm{Var}_g\left(\mathbb{1}\{L > v\}w(L)\right)$ and $\mathrm{Var}_g\left((L - v)^+ w(L)\right)$. If G_L is such that $dF_L(x) < dG_L(x)$ for all $x > v$, then the IS estimator of VaR_α and ES_α have smaller variance than the plain MC estimators would.

Chapter 3

Importance Sampling and Stratification for Copula Models

3.1 Introduction

Many applications in finance and insurance involve the computation of $\mu = \mathrm{E}[\Psi(\boldsymbol{X})]$ where $\boldsymbol{X} = (X_1, \ldots, X_d)'$ is a d-dimensional random vector and $\Psi : \mathbb{R}^d \to \mathbb{R}$ is some function. Since this thesis focuses on rare-event simulation, we assume that $\Psi(\boldsymbol{X})$ takes a non-zero value with small probability. A popular approach in modelling the distribution of $\boldsymbol{X}$ is through the use of copulas. If F is the joint distribution function of $\boldsymbol{X}$ and F_j, $j = 1, \ldots, d$ is the marginal distribution functions of the jth component of $\boldsymbol{X}$, Sklar's Theorem allows use to decompose F as the composition of a copula $C : [0,1]^d \to [0,1]$ and the d marginals as $F(X_1, \ldots, X_d) = C(F_1(X_1), \ldots, F_d(X_d))$. With Sklar's Theorem, one can specify marginals of $\boldsymbol{X}$ first and then choose an appropriate copula. This is in contrast to the traditional approach where the full joint distribution is modelled altogether.

The main contribution of this chapter is the study of IS techniques which we design to be effective for problems where $\Psi(\boldsymbol{X})$ takes a large value when at least one of the components of $\boldsymbol{X}$ is large. Such problems often arise from dependence models in the realm of finance and insurance. We propose a new IS framework which is applicable to all classes of copulas from which sampling is feasible. The main idea of our IS approach is to oversample sets of the form $[0,1]^d\backslash[0,\lambda_k]^d$ for $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M \leq 1$. Explicit sampling algorithms are presented for the case of Archimedean copulas. We show how to construct the optimal IS distribution by analyzing the variance expression of the IS estimator. We further construct an SS estimator based on our general IS setup.

As discussed in Section 2.3, QMC is a simulation based numerical technique much like MC, but it offers a faster convergence for the error rate than MC does by generating samples based on a low-discrepancy sequence. QMC has proven effective for financial security pricing problems, among others, where the underlying model is multivariate normal [3, 13, 62]. Recently, its effectiveness for sampling copula models was studied and demonstrated theoretically and empirically in [19]. Building on that work, we also combine QMC with our proposed IS approach.

The rest of this chapter is organized as follows. Section 3.2 motivates our proposed IS techniques. Section 3.3 introduces a general IS setup for copula models and develops a sampling algorithm for the case of Archimedean copulas. Section 3.4 shows our proposed IS scheme is similar to SS and then develops an SS scheme by building on this connection. A sampling algorithm for SS for the case of Archimedean copulas is also given. Section 3.5 derives the variance expressions for IS and SS estimators. By minimizing such variance expressions, we derive optimal calibration for the proposal distributions, for both IS and SS. Section 3.6 numerically investigates the effectiveness of the proposed IS and SS schemes with and without QMC in simulation studies.

3.2 Motivation and Background

In a copula model, we may write $\mu = \mathrm{E}[\Psi(\boldsymbol{X})] = \mathrm{E}[\Psi_0(\boldsymbol{U})]$, where $\boldsymbol{U} \sim C$, a copula of $\boldsymbol{X}$, and $\Psi_0 : [0,1]^d \to \mathbb{R}$ is given by

$$\Psi_0(u_1,\ldots,u_d) = \Psi(F_1^{-1}(u_1),\ldots,F_d^{-1}(u_d)),$$

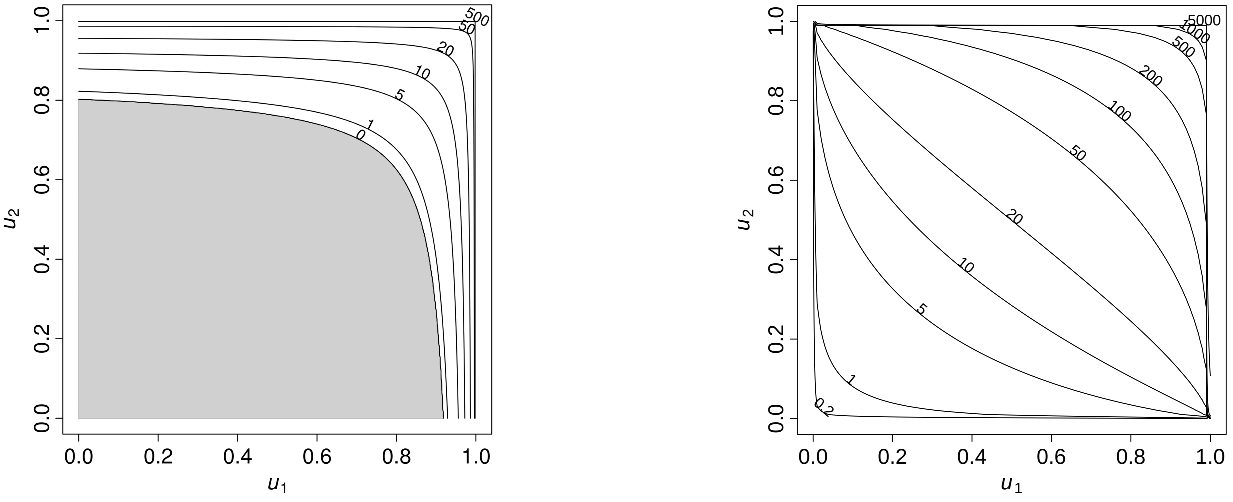
where $F_j^{-1}(p) = \inf\{x \in \mathbb{R} : F_j(x) \geq p\}$ for $j \in \{1,\ldots,d\}$. Sklar's Theorem asserts that C is unique if the d marginals of X are continuously distributed. If C and $F_1,\ldots,F_d$ are known, we can construct a plain MC estimator based on n samples $\{\boldsymbol{U}_1, \dots, \boldsymbol{U}_n\} \overset{\text{ind.}}{\sim} C$ as

$$\hat{\mu}_{\mathrm{MC},n} = \frac{1}{n} \sum_{i=1}^{n} \Psi_0(\boldsymbol{U}_i).$$

In this chapter, we consider the case where Ψ_0 is large only when at least one of its arguments is close to 1, or equivalently, if the maximum component of $\boldsymbol{X}$ is large. This assumption is inspired by several applications in insurance:

- The fair premium of a stop loss cover with deductible D is $\mathrm{E}[\{\sum_{j=1}^{d} X_j - D, 0\}]$. The corresponding functional is $\Psi_0(\boldsymbol{u}) = \max\{\sum_{j=1}^{d} F_j^{-1}(u_j) - D, 0\}$; see the left-hand side of Figure 3.1 (taken from [7, Figure 1]) for a contour plot of Ψ_0 for two Pareto margins.

Figure 3.1: *Left:* Contour lines for the excess function $\Psi_0(u_1, u_2) = \max\{F_1^{-1}(u_1) + F_2^{-1}(u_2) - 10, 0\}$, where the margins are Pareto distributed with $F_1(x) = 1 - (1 + x/4)^{-2}$ and $F_2(x) = 1 - (1 + x/8)^{-2}$. The grey area indicates where Ψ_0 is zero. *Right:* Contour lines for the product function $\Psi_0(u_1, u_2) = F_1^{-1}(u_1)F_2^{-1}(u_2)$, where $X_1 \sim \mathrm{LN}(2, 1)$ and $X_2 \sim \mathrm{LN}(1, 1.5)$.



- Risk measures for an aggregate sum $S = \sum_{j=1}^{d} X_j$, such as value-at-risk, $\mathrm{VaR}_\alpha(S)$, or expected shortfall, $\mathrm{ES}_\alpha(S)$, $\alpha \in (0, 1)$, cannot in general be written as an expectation of type $\mathrm{E}(\Psi(\boldsymbol{X}))$. However, they are functionals of the aggregate distribution function $F_S(x) = \mathbb{P}(S \le x) = \mathrm{E}(\Psi(\boldsymbol{U}; x))$, where $\Psi_0(\boldsymbol{u}; x) = \mathbb{1}_{\{F_1^{-1}(u_1)+\dots+F_d^{-1}(u_d)\le x\}}$. We can therefore write

$$\mathrm{VaR}_\alpha(S) = \inf\{x \in \mathbb{R} : \mathrm{E}(\Psi_0(\boldsymbol{U}; x)) \ge \alpha\}, \quad \mathrm{ES}_\alpha(S) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(S)\, du, \tag{3.1}$$

which depend only on those x for which $\mathrm{E}(\Psi_0(\boldsymbol{U}; x)) \ge \alpha$ holds. This is determined by the tail behaviour of S, which is strongly influenced by the properties of the copula C when at least one component is close to 1. Note that capital allocation methods such as the Euler principle for expected shortfall behave similarly, see [88] and [116], page 260.

3.3 Importance Sampling for Copula Models

3.3.1 Importance Sampling Algorithm

Since we are interested in estimating the quantities related to the tail of $\Psi_0(\boldsymbol{U})$ for $\boldsymbol{U} \sim C$, we use IS to improve the precision of the plain MC estimator. Let G denote the distribution function of $\boldsymbol{U}$ under the IS distribution. The IS estimator has the form

$$\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n}\sum_{i=1}^n \Psi_0(\boldsymbol{U}_i)w(\boldsymbol{U}_i), \quad \boldsymbol{U}_i \overset{\text{ind.}}{\sim} G,$$

where $w(\boldsymbol{u}) = \frac{dC(\boldsymbol{u})}{dG(\boldsymbol{u})}$ is the Radon-Nikodym derivative of C with respect to G.

Let $T = \max\{U_1, \dots, U_d\}$ and $t = \max\{u_1, \dots, u_d\}$. Since we assume that $\Psi_0(\boldsymbol{U})$ is large when at least one component of $\boldsymbol{u}$ is large, the ideal IS distribution places greater weights on the domain of $\boldsymbol{U}$ with large T than the original distribution does. The main idea of our IS scheme is to first draw a discretely distributed threshold random variable $\Lambda \sim F_\Lambda$ which is concentrated on $[0, 1]$ and is defined by $q_k := \mathbb{P}(\Lambda = \lambda_k)$, $k = 0, \dots, M$ and then sample $\boldsymbol{U} \,|\, T > \Lambda$ under the original distribution. This IS scheme is summarized in Algorithm 2.

Depending on the choice of F_Λ, the IS distribution places heavier weights on the region with large T. If for instance $\mathbb{P}(\Lambda = 0) = \mathbb{P}(\Lambda = 0.9) = 0.5$, then greater than 50% of

Algorithm 2 Importance Sampling Estimator

```
for $i = 1, \dots, n$ do
    Draw $\Lambda_i = \lambda_i$ from $F_\Lambda$
    Draw $\boldsymbol{U}_i \sim C$ conditionally on $T > \lambda_i$.
    Compute $w(\boldsymbol{U}_i) = \frac{dC(\boldsymbol{U}_i)}{dG(\boldsymbol{U}_i)}$
end for
return $\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n} \sum_{i=1}^{n} \Psi_0(\boldsymbol{U}_i)\omega(\boldsymbol{U}_i)$.
```

the samples under IS will lie on $[0,1]^d \setminus [0,0.9]^d$ on average. On the other hand, the case $\mathbb{P}(\Lambda = 0) = 1$ yields $G = C$, and then IS becomes MC.

Let $C_\lambda(\boldsymbol{u})$ be the distribution function of $\boldsymbol{U} \,|\, T > \lambda$ under the original distribution. Then the IS distribution function, which we denote by $G(\boldsymbol{u})$, has the mixture representation

$$G(\boldsymbol{u}) = \sum_{k=0}^{M} q_k C_{\lambda_k}(\boldsymbol{u}),$$

where

$$\begin{aligned} C_\lambda(\boldsymbol{u}) &= \mathbb{P}(U_1 \le u_1, \dots, U_d \le u_d \,|\, \max\{U_1, \dots, U_d\} > \lambda) \\ &= \mathbb{P}(U_1 \le u_1, \dots, U_d \le u_d \,|\, \boldsymbol{U} \notin [0,\lambda]^d) \\ &= \frac{C(\boldsymbol{u}) - C\left(\min\{u_1, \lambda\}, \dots, \min\{u_d, \lambda\}\right)}{1 - C(\lambda \mathbf{1})}, \end{aligned} \tag{3.2}$$

Note that the IS weight function is well-defined if C is absolutely continuous with respect to G. In order to guarantee this absolutely continuity for any copula C, we make the following assumption:

Assumption 1. *The random variable* Λ *satisfies* $\mathbb{P}(\Lambda = 0) > 0$.

Since $C_0 = C$, ensuring that $\mathbb{P}(\Lambda = 0) > 0$ is a form of defensive mixture sampling as described in Hesterberg [49]. Then, $w(\boldsymbol{u}) \le \mathbb{P}(\Lambda = 0)^{-1}$ on $\boldsymbol{u} \in [0,1]^d$ under Assumption 1 and the consistency and asymptotic normality of the IS estimator follows. In order to construct an IS estimator as in Algorithm 2, one needs to evaluate the IS weight function. Theorem 3.3.1 derives the expression for $w(\boldsymbol{u})$.

Theorem 3.3.1 ([7, Theorem 4.4, Equation (4.1)], see p. 155 for proof)**.** *The Radon–Nikodym derivative $w(\boldsymbol{u}) = dC(\boldsymbol{u})/dG(\boldsymbol{u})$ is given by*

$$w(\boldsymbol{u}) = \Big(\sum_{k=1}^{M} \frac{\mathbb{1}\{\lambda_k \le \max\{u_1, \dots, u_d\}\}}{1 - C(\lambda_k \mathbf{1})} q_k\Big)^{-1}. \tag{3.3}$$

In order to simplify the notation, let $\widetilde{w} : [0, 1] \to [0, \infty)$ be defined as

$$\widetilde{w}(t) = \Big(\sum_{k=1}^{M} \frac{\mathbb{1}\{\lambda_k \le t\}}{1 - C(\lambda_k \mathbf{1})} q_k\Big)^{-1} \tag{3.4}$$

so that we have $w(\boldsymbol{u}) = \widetilde{w}(\max\{u_1, \dots, u_d\}) = \widetilde{w}(t)$. In order to evaluate $\widetilde{w}$, it is sufficient to calculate (or approximate) $C(\lambda_k \mathbf{1})$ for $k \in \{1, \dots, M\}$. These values must be computed only once and thus this approach is fast and can be easily implemented. In particular, the density of C does not have to be evaluated to calculate w (or $\tilde{w}$). This is an advantage in comparison to most other IS algorithms, for which the existence of the density of C is required.

Remark 3.3.2. *Observe that the IS weight function* (3.3) *depends on* $\boldsymbol{u}$ *only through* t*. That is, the variance of* $w(\boldsymbol{U})$ *under the proposal distribution is a function of the distribution of a univariate random variable* T*. Thus, the weight function of the IS scheme of Algorithm 2 does not suffer from the dimensionality effect discussed in Section 2.2.2. The reason why only the distribution of* T *matters for the variance of* $w(\boldsymbol{U})$ *is due to Step 3 of Algorithm 2. Because we still sample* $\boldsymbol{U} \,|\, T > \Lambda$ *under the original distribution when we apply IS, the density related to* $\boldsymbol{U} \,|\, T > \Lambda$ *appears both in the numerator and the denominator of the weight function and thus they cancel each other out. The only part that remains comes from the distribution of* T*. The conditional sampling step from the original distribution essentially reduces the dimension of the IS weight function to 1. We use this idea of conditional sampling when we develop IS techniques for more general models in Chapter 4 and Chapter 6.*

3.3.2 Sampling Algorithm for Archimedean Copulas

While the IS method from the previous section can be applied to any copula, sampling from C_λ is difficult in general. While it is possible in principle to sample from C_λ, or any multivariate distribution, using a multivariate quantile transform [105], such transform is generally computationally very expensive as it involves evaluations of conditional quantile functions that must be approximated numerically. In this section, we develop an efficient sampling method for C_λ when C is an Archimedean copula.

In light of the MO Algorithm (Algorithm 1), $(U_1, \dots, U_d) \overset{D}{=} \psi(\frac{E_1}{V}, \dots, \frac{E_d}{V})$ where $E_i \overset{\text{ind.}}{\sim}$ Exp(1) and V is the corresponding frailty random variable. Using some algebra, we can write the condition $T > \lambda$ as $E_{(1)} \le \psi^{-1}(\lambda)V$, where $E_{(1)}$ is the first order statistics of $\{E_1, \dots, E_d\}$ which is distributed as Exp(d). In summary, sampling from $\boldsymbol{U} \,|\, T > \lambda$ is equivalent to sampling from $(E_1, \dots, E_d, V) \,|\, E_{(1)} \le \psi^{-1}(\lambda)$. Algorithm 3 summarizes the sampling method for this conditional distribution where we let $\gamma = \psi^{-1}(\lambda)$. Proposition 3.3.3 asserts that samples from Algorithm 3 have the right distribution.

Algorithm 3 Sampling Step of the IS algorithm for Archimedean copulas

```
Require: $0 < \gamma = \psi^{-1}(\lambda) < \infty$.
1: Draw $(E_{(1)}, V) \,|\, (E_{(1)} < \gamma V)$.
2: Draw $(E_1, \dots, E_d) \,|\, E_{(1)}$.
3: Let $U_j = \psi(E_j/V)$ for $j \in \{1, \dots, d\}$.
4: Return $(U_1, \dots, U_d)$.
```

Proposition 3.3.3 (see p. 155 for proof)**.** *Let $E_1, \dots, E_d$ be iid positive random variables and V be a positive random variable independent of the E_j's. Then a sample $(E_1, \dots, E_d, V)$ constructed as in Steps 1–3 of Algorithm 3 has the distribution $(E_1, \dots, E_d, V) \,|\, (E_{(1)} < \gamma V)$.*

While Proposition 3.3.3 holds for general (positive) E_j's and V, we now give detailed explanations of how to do the sampling for Steps 1 and 2 of Algorithm 3, i.e., when $E_j \overset{\text{ind.}}{\sim}$ Exp(1) and V is the frailty random variable.

Step 1: Sample $(E_{(1)}, V) \,|\, (E_{(1)} < \gamma V)$

The objective is to sample from the joint distribution of $(E_{(1)}, V)$ conditioned on the event $(E_{(1)} < \gamma V)$. Let $f_{E_{(1)}}(x)$ denote the density of $E_{(1)}$ and $f_V(v)$ denote the density of V with respect to a reference measure (the Lebesgue measure if V is continuous or the counting measure if V is discrete). Further, let $f_{(E_{(1)},V)|(E_{(1)}<\gamma V)}(x,v)$ be the conditional joint density of $(E_{(1)},V)$ given $E_{(1)}<\gamma V$. Then by independence of $E_{(1)}$ and V

$$f_{(E_{(1)},V)|(E_{(1)}<\gamma V)}(x,v)=\beta f_{E_{(1)}}(x)f_V(v)\mathbb{1}(x<\gamma v), \tag{3.5}$$

where $\beta=1/\mathbb{P}(E_{(1)}<\gamma V)=1/\mathbb{P}(U_{(d)}>\lambda)=1/(1-C(\lambda\mathbf{1}))=1/(1-\psi(d\psi^{-1}(\lambda)))$. We use conditional sampling to sample from this density, that is, we first sample V from the marginal conditional density $f_{V|(E_{(1)}<\gamma V)}$ of (3.5) then draw $E_{(1)}$ from (3.5) given V. Note that

$$f_{V|(E_{(1)}<\gamma V)}(v)=\beta f_V(v)\int_0^{\gamma v} f_{E_{(1)}}(x)\,dx=\beta f_V(v)(1-\exp(-d\gamma v)). \tag{3.6}$$

Unfortunately, the density (3.6) does not belong to a known parametric family for most Archimedean copulas. Nonetheless, there exist efficient numerical algorithms that allow one to sample from a univariate distribution given its probability density function. For instance, the NINIGL Algorithm in [30] achieves this through numerical inversion techniques. Given a density function, the NINIGL Algorithm numerically constructs the inverse CDF function of the density. One can then efficiently draw multiple samples from the density by evaluating the inverse CDF function at samples from $U[0,1)$. Such algorithms could become costly if they had to be applied for several values of Λ. However in our numerical experiments, the threshold random variable Λ only takes a small number of distinct values, such as 10, which is much less than the number of simulations, which is of order 10,000. Hence, for each value of $\Lambda=\lambda$, we sample from (3.6) thousands of times, which makes the overhead required to initialize the sampling algorithms negligible.

After sampling V from (3.6), we want to draw $E_{(1)}$ given V. Let $f_{E_{(1)}|(E_{(1)}<\gamma V,V)}(x\,|\,V)$ denote the conditional density of $E_{(1)}$. Then

$$f_{E_{(1)}|(E_{(1)}<\gamma V)}(x\,|\,V)=\frac{d\exp(-dx)\mathbb{1}(x<\gamma V)}{1-\exp(-d\gamma V)}$$

and we can draw a sample from this density using the inversion technique. In particular, we generate $U\sim \mathrm{U}[0,1)$ and then let $E_{(1)}=-\frac{1}{d}\log(1-U(1-e^{-\gamma dV}))$.

Step 2: Sampling $(E_1,\dots,E_d)\,|\,E_{(1)}$

Suppose we have drawn $E_{(1)}=x_{(1)}$ from Step 1. Let $f(x_1,\dots,x_d)=\exp\left(-\sum_{i=1}^d x_i\right)$

be the joint density of $(E_1, \ldots, E_d)$. Note that each E_j is as likely to be the minimum. Consider the case where E_1 is the minimum. The conditional distribution is

$$f(x_1, \ldots, x_d \,|\, E_1 = E_{(1)},\ E_{(1)} = x_{(1)}) = \frac{e^{-x_{(1)} - \sum_{j=2}^{d} x_j}}{(1/d) d e^{-d x_{(1)}}} = e^{-\sum_{j=2}^{d}(x_j - x_{(1)})} \cdot \mathbb{1}_{\{E_1 = x_{(1)}\}}. \quad (3.7)$$

We can sample from (3.7) by letting $E_j = \mathrm{Exp}(1) + x_{(1)}$ independently for $j \in \{2, \ldots, d\}$.

Since any of the E_j's can be the minimum, we pick the index for the minimum component randomly from 1 to d and sample the rest of the components accordingly. This sampling method works for MC, but may not work very well for QMC. When randomly choosing the index for the minimum component, we potentially destroy the structure of the LDS. So, if we are working with an LDS, the sampling method based on Proposition 3.3.4 below is preferred.

Proposition 3.3.4 (see p. 155 for proof)**.** *Suppose* $E_1, \ldots, E_d$ *are iid* $\mathrm{Exp}(1)$*. Then*

$$\begin{aligned} &\mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \ldots, E_{k-1} = x_{k-1}, E_{(1)} = x) \\ &\qquad = \begin{cases} 1 - \exp\{-(x_k - x)\}, & \text{if } x_j = x \text{ for some } j \in \{1, \ldots, k-1\}, \\ \frac{1}{d-k+1}\mathbb{1}_{\{x_k < x\}} + \frac{d-k}{d-k+1}(1 - \exp\{-(x_k - x)\}), & \text{otherwise.} \end{cases} \end{aligned} \quad (3.8)$$

To sample $E_1, \ldots, E_d$, we let k take the successive values $k \in \{1, \ldots, d\}$ in (3.8) and proceed by inversion.

3.4 Stratified Sampling Alternative to Importance Sampling

Recall from Algorithm 3 that our proposed IS scheme starts with sampling a threshold random variable Λ, and then proceeds by sampling $\boldsymbol{U} \,|\, T > \lambda_k$ under the original distribution. Instead, we can construct an SS estimator based on samples from $\boldsymbol{U} \,|\, \lambda_{k+1} > T \ge \lambda_k$ under the original distribution. Suppose that Λ takes M distinct values as $0 = \lambda_1 < \cdots < \lambda_M < 1$.

Then define M strata as

$$\begin{aligned}\Omega_C^{(k)} &= \{\boldsymbol{u} \in [0, \lambda_{k+1}]^d \mid \lambda_{k+1} > t \geq \lambda_k\}, \quad k = 1, \ldots, M \\ &= \{\boldsymbol{u} \in [0, \lambda_{k+1}]^d \mid \boldsymbol{u} \notin [0, \lambda_k]^d\}, \quad k = 1, \ldots, M. \end{aligned} \tag{3.9}$$

This strata construction stratifies the domain of $\boldsymbol{U}$ along T as $\boldsymbol{U} \in \Omega_C^{(k)}$ if and only if $\lambda_{k+1} > T \geq \lambda_k$. The SS estimator is defined as

$$\hat{\mu}_{\mathrm{SS},n} = \sum_{k=1}^{M} \frac{p_k}{n_k} \sum_{i=1}^{n_k} \Psi_0(\boldsymbol{U}_i^{(k)}), \tag{3.10}$$

where p_k is the stratum probability, n_k is the number of samples allocated to the stratum $\Omega_C^{(k)}$, and $\boldsymbol{U}_i^{(k)} \overset{\text{ind.}}{\sim} \boldsymbol{U} \mid \Omega_C^{(k)}$ under the original distribution. For Archimedean copulas, $p_k = \psi(d\psi^{-1}(\lambda_{k+1})) - \psi(d\psi^{-1}(\lambda_k))$. It is easily shown that sampling from $\boldsymbol{U} \mid \Omega_C^{(k)}$ is equivalent to sampling from

$$(E_1, \ldots, E_d, V) \mid \psi^{-1}(\lambda_{k+1})V < E_{(1)} \leq \psi^{-1}(\lambda_k)V.$$

Define $\lambda_{M+1} = 1$ for convenience and let $\gamma_k = \psi^{-1}(\lambda_k)$ for all $k \in \{1, \ldots, M+1\}$. Algorithm 4 summarizes the procedure to sample from each stratum.

Algorithm 4 Sampling $\boldsymbol{U}_{k,j}$ in SS algorithm for Archimedean copulas

```
Require: $0 < \gamma_{k+1} < \gamma_k < \infty$.
1: Draw $(E_{(1)}, V) \mid (\gamma_{k+1}V < E_{(1)} \leq \gamma_k V)$.
2: Draw $(E_1, \ldots, E_d) \mid E_{(1)}$.
3: Let $U_j = \psi(E_j/V)$ for $j \in \{1, \ldots, d\}$.
4: Return $(U_1, \ldots, U_d)$.
```

In this algorithm, Step 2 is exactly the same as for the IS case (Algorithm 3). For Step 1, we use conditional sampling to draw samples from the joint conditional density of $(E_{(1)}, V) \mid (\gamma_{k+1}V < E_{(1)} \leq \gamma_k V)$. By using an argument similar to the one used for Step 1 of Algorithm 3, one can show that the marginal conditional density of V is

$$f_{V|(E_{(1)}<\gamma V)}(v) = \beta f_V(v)(\exp(-d\gamma_{k+1}v) - \exp(-d\gamma_k v)), \tag{3.11}$$

where $f_V(v)$ is the density of V and $\beta = 1/p_k = 1/\psi[d\psi^{-1}(\lambda_{k+1})) - \psi(d\psi^{-1}(\lambda_k)]$. Conditional on V drawn from (3.11), generate $U \sim \mathrm{U}[0,1)$ and then let

$$E_{(1)} = -\frac{1}{d}\log\left[e^{-\gamma_{k+1}dy} - U(e^{-\gamma_{k+1}dy} - e^{-\gamma_k dy})\right].$$

Then $(E_{(1)}, V)$ follows the desired distribution.

Remark 3.4.1. *We can follow Algorithm 4 to sample from the SS distribution under QMC, if the number of samples to be drawn is fixed. In some cases, however, we want to keep running simulations until some error criterion is met. Since SS requires to have a subset of points allocated to each stratum, combining it with QMC for n not fixed is challenging. This is because when the total sample size is increased by successive increments, it means possibly disjoint subsets of a QMC point set will be used in a given stratum, which is undesirable. Whether or not this allocation over successive increments can be done in a clever way that exploits the uniformity of low-discrepancy sequences is a question we leave for future research.*

3.5 Variance Analysis and Calibration Method

In this section, we analyze the variance of the IS and SS estimators and then propose calibration methods designed to minimize the variance of the respective estimators. We also show that the SS scheme is more flexible when calibrating and it also gives an estimate with a smaller variance than IS does. We define the strata $\Omega_1, \dots, \Omega_M$ as in (3.9) and let $C_k = C(\lambda_k \mathbf{1})$ for $k = 1, \dots, M$. For clarity, the operators $\mathbb{P}_C$, E_C, and Var_C indicate that the probability, expectation and variance are computed under the original distribution C. Similarly, $\mathbb{P}_G$, E_G, and Var_G are for under the IS distribution. The following proposition gives the variance of the IS estimator.

Proposition 3.5.1 (see p. 156 for proof)**.** *Let $\hat{\mu}_{\mathrm{IS},n}$ be the IS estimator given by Algorithm 2. The variance of this estimator is*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n}\left(\sum_{k=1}^{M} p_k \left(\sum_{l=1}^{k} \frac{q_l}{1 - C_k}\right)^{-1} m_k^{(2)} - \mu^2\right), \tag{3.12}$$

where $p_k = \mathbb{P}_C(\boldsymbol{U} \in \Omega_C^{(k)})$, $q_k = \mathbb{P}(\Lambda = \lambda_k)$ and $m_k^{(2)} = \mathrm{E}_C[\Psi_0^2(\boldsymbol{U}) \,|\, \Omega_C^{(k)}]$.

For the optimal calibration, we want to choose q_k's so that (3.12) is minimized. The following proposition gives an analytical expression for the optimal calibration.

Proposition 3.5.2 (see p. 157 for proof). *The set of q_k's that minimize* (3.12) *under the condition $m_1^{(2)} \leq \ldots \leq m_M^{(2)}$ (with $m_0^{(2)} = 0$ for notational convenience), is*

$$q_k^{\text{opt}} = \frac{(1 - C_k)\left(\sqrt{m_k^{(2)}} - \sqrt{m_{k-1}^{(2)}}\right)}{\sum_{k=1}^{M}(1 - C_k)\left(\sqrt{m_k^{(2)}} - \sqrt{m_{k-1}^{(2)}}\right)}, \quad k = 1, \ldots, M. \tag{3.13}$$

Remark 3.5.3. *If the condition $m_1^{(2)} \leq \cdots < m_M^{(2)}$ is not met, some of the q_k^{opt}'s given by* (3.13) *will be negative, which makes the IS scheme infeasible. Note that $q_k^{\text{opt}} < 0$ means that ever having the event $[\Lambda = \lambda_k]$ makes the overall variance greater than when q_k^{opt}. We propose to then remove λ_k from the support of Λ if $q_k^{\text{opt}} < 0$. Accordingly, the strata $\Omega_C^{(k)}$'s will change so the stratum second moments need to be recomputed for the optimal calibration.*

Of course, we do not know the true values of the $m_k^{(2)}$'s in practice, so we have to replace them with estimates. As often done for Neyman allocation, we can first run a pilot study with a small number of simulations and estimate the $m_k^{(2)}$'s. The condition $m_1^{(2)} \leq \ldots \leq m_M^{(2)}$ means that the outer strata must have greater stratum second moments than the inner strata. We refer to this condition as increasing second moment (ISM) condition. Whether this ISM condition is met or not depends on the problem at hand. The assumption that $\Psi_0(\boldsymbol{U})$ is large when T is large and the ISM condition are not incompatible, although there is no guarantee that such $\Psi_0(\boldsymbol{U})$ satisfies the condition. If the ISM condition is satisfied, we can substitute (3.13) into (3.12) to obtain

$$\mathrm{Var}(\hat{\mu}_{\text{IS},n}^{\text{opt}}) = \frac{1}{n}\left(\left(\sum_{k=1}^{M} p_k \sqrt{m_k^{(2)}}\right)^2 - \mu^2\right). \tag{3.14}$$

By Jensen's inequality,

$$\mathrm{Var}(\hat{\mu}^{\mathrm{opt}}_{\mathrm{IS},n}) = \frac{1}{n}\left(\left(\sum_{k=1}^{M} p_k\sqrt{m_k^{(2)}}\right)^2 - \mu^2\right) \le \frac{1}{n}\left(\sum_{k=1}^{M} p_k m_k^{(2)} - \mu^2\right) = \mathrm{Var}(\hat{\mu}_{\mathrm{MC},n}).$$

Equality holds only when $m_k^{(2)}$ is the same for all k. Except for this restrictive case, the IS estimator with the optimal choice of the q_k's always has a smaller variance than the plain MC counterpart. If the ISM condition is not met, there is no analytical form for the optimal q_k's. We can still find the optimal values using widely available convex optimization solvers. If we let $q_1 = 1$ and let $q_k = 0$ for $k = 2, \dots, M$, the proposal distribution becomes the original distribution. That is, IS become plain MC. Hence, if the q_k's are chosen appropriately, the IS estimator cannot do worse than the plain MC estimator. In this sense, the IS estimator is similar to an SS estimator.

Now that we have derived the variance expression and the optimal choice of q_k's for IS estimator, we move on to the SS estimator (3.10). Using simple algebra, one can show

$$\mathrm{Var}(\hat{\mu}_{\mathrm{SS},n}) = \sum_{k=1}^{M} \frac{p_k^2 v_k^2}{n_k}, \tag{3.15}$$

where $v_k^2 = \mathrm{Var}_C(\Psi_0(\boldsymbol{U}) \,|\, \Omega_C^{(k)})$, $k = 1, \dots, M$ are the stratum variances. The optimal choices of the n_k's is given by Neyman allocation [21, pp.98-99]

$$n_k = \frac{n p_k v_k}{\sum\limits_{k=1}^{M} p_k v_k}. \tag{3.16}$$

Unlike the IS estimator, there is no restriction on this optimal allocation. That is, σ_k does not need to increase with k. In this sense, the SS estimator is more flexible.

Since the stratum variances are unknown, we have to replace them with estimates. Investigating the optimal calibration formula for IS (3.13) and SS (3.16), it appears that the estimation error of the strata moments (the $m_k^{(2)}$'s for IS and the v_k^2's for SS) has greater impact on the estimated calibration for IS than for SS. Since q_k for IS depends on $\sqrt{m_k^{(2)}} - \sqrt{m_{k-1}^{(2)}}$, the estimation error comes from both estimating $m_{k-1}^{(2)}$ and $m_k^{(2)}$. On the other hand, for SS, n_k depends on σ_k, so the estimation error comes from estimating v_k^2 alone. Consequently, the approximation is likely to deviate more from the actual optimal calibration for IS than for SS.

The optimal calibration for IS (3.13) and SS (3.16) give the variance minimizing q_k's and n_k's, respectively, for a given set of threshold values $\lambda_1, \dots, \lambda_M$. Another possible optimal calibration is to find the variance minimizing λ_k's for fixed q_k's or n_k's. We do not pursue this approach because finding such λ_k's is difficult, as the variance of the IS and SS estimators are not convex in the λ_k's.

Going back to IS and as discussed in [49], instead of choosing $\Lambda = \lambda_k$ with probability q_k, it is more efficient to stratify Λ. That is, take $n_k = nq_k$ observations with λ_k. Let $\hat{\mu}^{\text{det}}_{\text{IS},n}$ denote such a stratified IS (SIS) estimator. Generally nq_k's will not be integers, so we have to round them. If each n_k is large enough, this rounding effect is negligible. Then we have the following proposition that compares the variance of the three estimators.

Proposition 3.5.4 (see p. 158 for proof)**.** *Suppose we have an IS estimator with* $\mathbb{P}(\Lambda = \lambda_k) = q_k,\ k \in \{1, \dots, M\}$*. If the* $\mu_k = \mathbb{E}_C(\Psi_0(\boldsymbol{U})|\Omega_C^{(k)})$ *are not all equal and* n *is large enough, then there exists some strata sample allocation* $(n_1, \dots, n_M)$ *for the SS estimator such that* $\mathrm{Var}(\hat{\mu}_{\text{SS},n}) \le \mathrm{Var}(\hat{\mu}^{\text{det}}_{\text{IS},n}) \le \mathrm{Var}(\hat{\mu}_{\text{IS},n})$.

This result trivially holds when we use the optimal q_k's (3.13) for stratified and unstratified IS and use the optimal allocation (3.16) for SS. Since the SS estimator is more flexible for calibration and it has a smaller variance than the stratified/unstratified IS estimator, the SS approach is the preferred one if the sampling efforts for (3.6) and (3.11) are not significantly different. Nonetheless, depending on the type of the underlying copula, sampling from the IS distribution could be much easier than sampling from SS distribution.

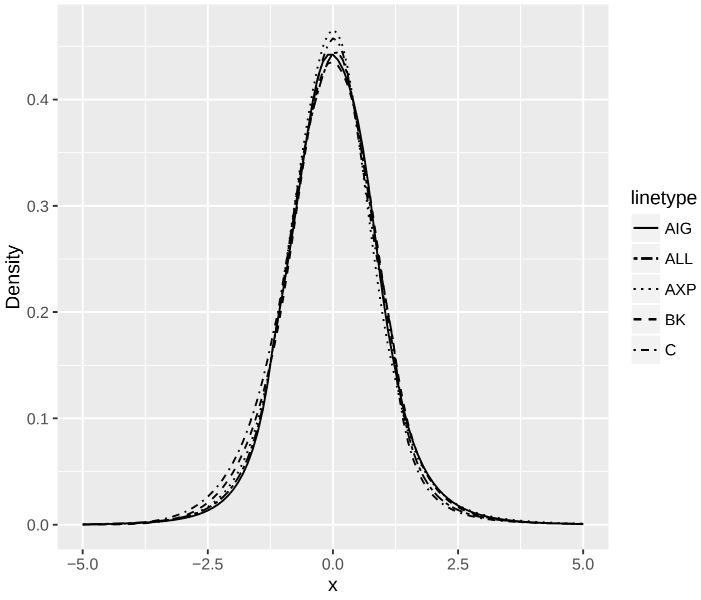
Remark 3.5.5. *The variance minimizing calibration for IS* (3.13) *and SS* (3.16) *assume that the objective is to estimate a standard expectation of the form* $\mu = \mathrm{E}[\Psi_0(\boldsymbol{U})]$*. If the goal is to estimate quantities that cannot be written as a standard expectation, such as* VaR_α *and* ES_α*, we cannot directly apply those calibrations. Fortunately, the asymptotic results* (2.26) *and* (2.27) *allow us to pretend during calibration that the goal is to estimate* $\mathrm{E}[\mathbb{1}\{L > \hat{v}\}]$ *and* $\mathrm{E}[(L - \hat{v})^+]$ *when estimating* VaR_α *and* ES_α*, respectively, where* $\hat{v}$ *is an* *initial estimate of* VaR_α. *Since* $\mathrm{E}[\mathbb{1}\{L > \hat{v}\}]$ *and* $\mathrm{E}[(L-\hat{v})^+]$ *are both expectations, we can use the calibrations of* (3.13) *and* (3.16) *and construct effective proposal distributions for* VaR_α *and* ES_α*, respectively.*

3.6 Numerical examples

In this section, we numerically investigate the efficiency of the IS and SS estimators introduced in this chapter. We consider the valuation of tail-related quantities of a portfolio consisting of stocks from companies in the financial industry listed on the S&P 100. The five stocks in the portfolio are AIG, Allstate Corp., American Express Inc., Bank of New York and Citigroup Inc. Their stock symbols are AIG, ALL, AXP, BK and C, respectively. We assume that the value of the portfolio is 100 and that all the portfolio weights are the same. The data are daily negative log-returns of these five companies from 2010-01-01 to 2016-04-01 (1571 data points). The computations were carried out on a Dell XPS 13 9350, Intel CPU 2.3 GHz on 8 GB RAM. All algorithms are implemented in the R programming environment. We fit GARCH(1,1)-models with t-innovations to each return series to filter out the volatility clustering effect using the R package "rugarch" [34]. The fitted standardized residuals do not exactly follow a t-distribution, so we fit a semi-parametric distribution to the residuals using the R package "spd" [35]. The fitted model uses a kernel density estimate for the centre of the distribution and fits a heavy tailed generalized Pareto distribution to the tails. The use of generalized Pareto distribution to model the GARCH filtered residuals to estimate tail-related risk measures in a univariate setting is studied by McNeil and Frey [87].

Figure 3.2 shows the plot of the density of the semiparametric distributions fitted to the GARCH filtered residuals for the five stocks in the portfolio. As the figure illustrates, the fitted semiparametric densities have slightly different shapes. In particular, the density for ALL and AXP returns have higher peaks and lighter right-tails than the densities for AIG, BK, and C returns do.

Figure 3.2: Comparison of Gumbel and Frank copula.



We let $S = \sum_{j=1}^{d} X_j$ denote the portfolio loss over a one day period with

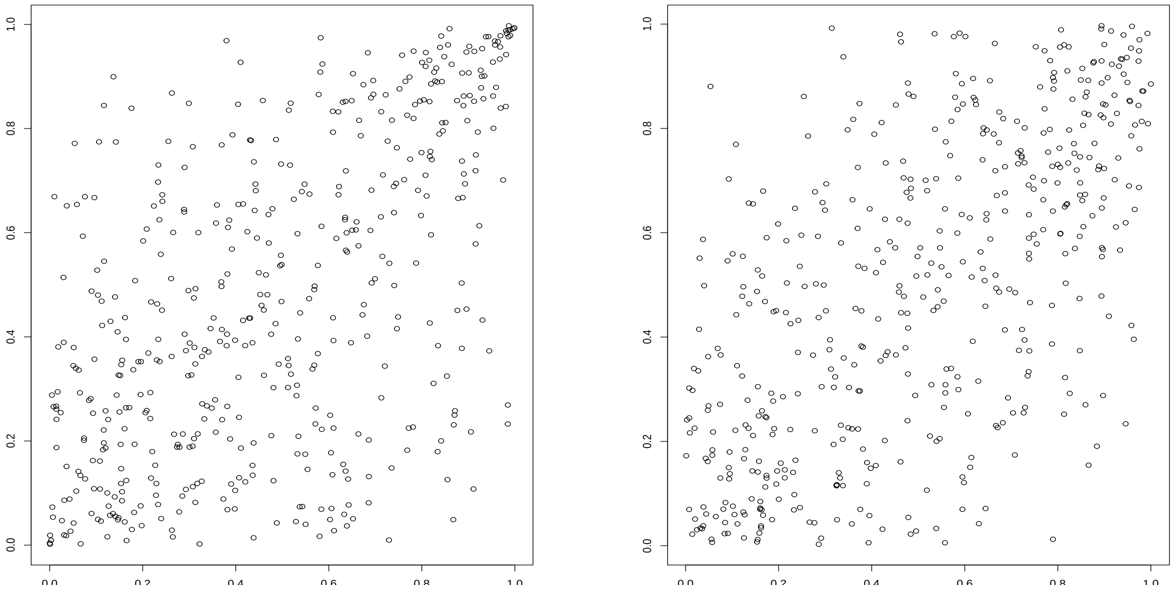
$$X_j = 100\omega_j \left(1 - \sum_{j=1}^{d} \exp(a_j - b_j \tilde{F}_j^{-1}(U_j)) \right),$$

where d is the number of assets, ω_j's are the portfolio weights, a_j's are the means of log-returns, b_j's are the fitted standard deviations from the GARCH(1,1) model, F_j's are the fitted semi-parametric distributions from the R package "spd" [35], and $(U_1, \ldots, U_d)$ follows the fitted copula. We use R package "distr" [104] to sample from (3.6) and (3.11).

Using the R package "copula" [52], we fit the Gumbel, Frank, Clayton and Joe copulas to the standardized residuals based on MLE. The idea of fitting a copula to the residuals of times series models is explored in details by Rémillard [98]. Note that fitting Archimedean copulas implies that we are assuming that the dependence of the standardized residuals is static across time. Dynamic copulas ([61, 93, 100]) relax this assumption and model time-varying dependences. However, we do not purse dynamic copulas in our numerical studies. Among the four Archimedean copulas, the Gumbel copula with $\theta = 1.603$ gives the best fit in terms of log-likelihood, followed by a Frank copula with $\theta = 4.06$. Hence we proceed assuming that the model we consider is well approximated by a Gumbel or a Frank copula.

Figure 3.3 compares 500 independent samples of a two-dimensional Gumbel and Frank copula with $\theta = 1.604$ and $\theta = 4.06$, respectively. As the figure illustrates, the Gumbel copula has a positive upper tail dependence while the Frank copula has no tail dependence. A positive tail dependence means higher chance of multiple components of a sample point being simultaneously large. Intuitively this means that there is a higher chance of large portfolio loss under the Gumbel copula model than under the Frank copula model. Hence, we expect larger VaR and ES under the Gumbel copula model than under the Frank copula model

Figure 3.3: Comparison of Gumbel and Frank copula.



The three functionals we estimate are stop loss $\mathrm{E}(\{L - D\}^+)$ with $D = 3$ for Gumbel and $D = 2$ for Frank, $\mathrm{VaR}_{0.99}$ and $\mathrm{ES}_{0.99}$ of S. To define F_Λ, we use $\lambda_k = 1 - \left(\frac{1}{2}\right)^{k-1}$ for $k \in \{1, \ldots, M\}$, with $M = 10$. When constructing an IS estimator, we stratify Λ regardless of whether we use MC or QMC. When we calibrate the q_k's for IS according to

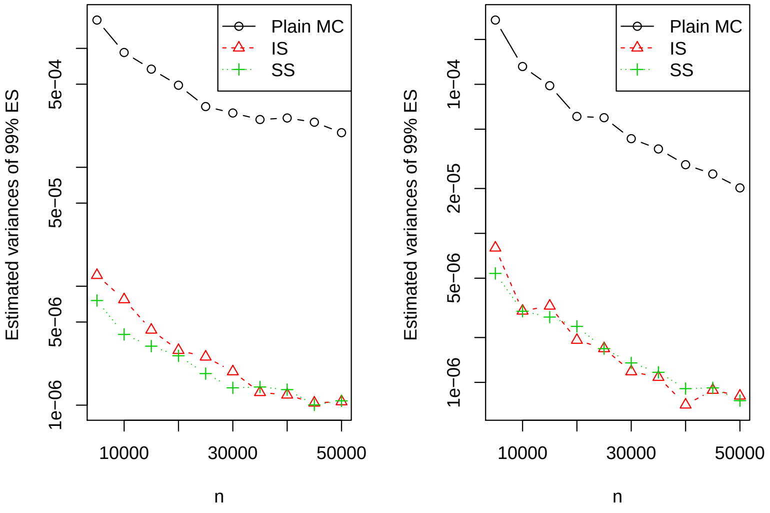
(3.13) and SS according to (3.16), we use ES as our objective function as we expect that the IS distribution that estimates ES well would also estimate the other two quantities well. Since ES is not an expectation, we cannot directly apply the calibrations (3.13) and (3.16). Thus, we use the idea from Remark 3.5.5.

Table 3.1: Estimates and variance reduction factors for the Gumbel and Frank copulas based on $n = 30\,000$.

		Gumbel						Frank					
			MC		QMC				MC		QMC		
Objective function	d	Estimate	IS	SS	Plain	IS	SS	Estimate	IS	SS	Plain	IS	SS
$\mathbb{E}(\max\{S - D, 0\})$	5	0.012	67	168	33	1730	8085	0.011	6.4	11	14	85	161
	20	0.010	49	40	51	1128	3488	0.0034	4.6	4.1	5.7	46	46
$\mathrm{VaR}_{0.99}(S)$	5	3.2	10	26	8.4	39	98	2.4	9.7	9.0	2.6	32	26
	20	3.04	7.9	7.2	5.8	19	28	2.1	4.3	4.8	3.6	16	19
$\mathrm{ES}_{0.99}(S)$	5	4.2	89	175	29	6019	16989	2.8	17	21	7.1	250	373
	20	4.03	49	39	48	1296	4205	2.3	4.6	3.8	4.0	38	36
Run time	5		3.6	3.7	1.8	3.7	3.8		3.6	3.7	1.1	3.6	3.7
	20		2.0	1.9	1.2	1.9	2.0		1.7	1.8	1.1	1.9	2.0

Table 3.1 shows the estimates, variance reduction factors and computational times for the three functionals for five different estimators for Gumbel and Frank copulas, respectively. The estimates shown are based on SS estimators with QMC. Variance reduction factors are defined to be the ratios of the variance of the plain MC estimators over the variance of the estimators with the respective VRTs. As the reciprocal of the variance of a plain MC estimator is proportional to n, the variance reduction factor is the same as the sample size reduction factor. For instance, if IS based on n samples gives an estimator with a variance reduction factor of 2, this equivalently means that IS needs only $n/2$ samples to achieve the same precision as the plain MC estimator. The last row of Table 3.1 shows the increase in computation time compared to plain MC. We see that both IS and SS reduce the variance by large amounts and this is amplified when combined with QMC. Note that SS estimators have variances smaller than the IS estimators do, as suggested by Proposition 3.5.4. For IS and SS estimators with and without QMC, we see that the

Figure 3.4: Estimated variances of plain MC, IS and SIS estimators of $\mathrm{ES}_{0.99}$ for a Gumbel and a Frank copula for different n and for $d = 5$.



(a) Gumbel Copula (b) Frank Copula

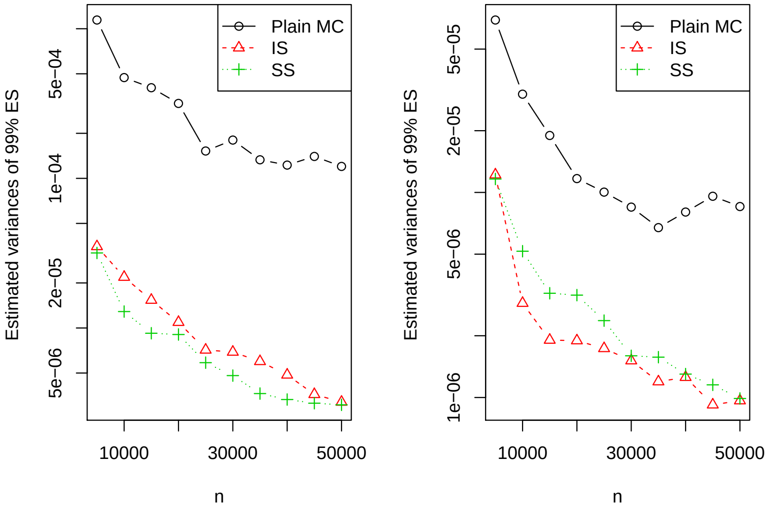
largest variance reduction factors are for ES. This makes sense as we calibrate the q_k's for IS and the n_k's for SS to minimize the variance of the ES estimator.

We also repeat the same experiment but with a portfolio of 20 stocks from large companies in the financial industry traded on NYSE (see Table 3.2 for stocks symbols); the results are displayed under $d = 20$ in Tables 3.1. Overall, the IS and SS schemes introduced in this chapter are effective for the 20-dimensional problem as well.

Table 3.2: Stock symbols for the 20-dimensional model

AIG	ALL	AXP	BAC	BAX	BK	BLK	BRK A	C	CB
COF	GS	JPM	MA	MET	MS	SPG	USB	V	WFC

Figure 3.5: Estimated variances of plain MC, IS and SIS estimators of $\mathrm{ES}_{0.99}$ for a Gumbel and a Frank copula for different n and for $d = 20$.



(a) Gumbel Copula (b) Frank Copula

Chapter 4

Importance Sampling and Stratified Sampling Techniques for Semiparametric Single-Index Models

4.1 Introduction

In Chapter 3, we developed IS, along with SS, techniques for copula models to estimate $\mu = \mathrm{E}[\Psi_0(\boldsymbol{U})]$, where $\boldsymbol{U}$ follows some d-copula and $\Psi_0 : [0,1]^d \to \mathbb{R}$ is some function under the assumption that $\Psi_0(\boldsymbol{U})$ takes a large value only when $\max\{\boldsymbol{U}\} = \max\{U_1, \dots, U_d\}$ is large. The main idea of the IS techniques is to twist the distribution of the maximum component of $\boldsymbol{U}$ by oversampling sets of the form $[0,1]^d \backslash [0, \lambda_k]^d$ for $0 \leq \lambda_1 \cdots \leq \lambda_M < 1$. However, the assumption that $\Psi_0(\boldsymbol{U})$ is large only when $\max\{\boldsymbol{U}\}$ is large may not hold in some applications. In this chapter, we relax this assumption and design IS techniques for problems where the output depends on the input variable mainly thorough some one-dimensional projection. In semiparametric regression, such structure of problems are called single-index models (see [46], [57],[95]), so we refer to our proposed IS technique as single-index IS. We do not specifically assume copula modelling in this chapter, so the problem is to estimate $\mu = \mathrm{E}[\Psi(\boldsymbol{X})]$, where $\Psi : \mathbb{R}^d \to \mathbb{R}$ is some function and $\boldsymbol{X}$ follows some d-dimensional distribution which may or may not be a copula. Since $\max\{\boldsymbol{U}\}$ is a type of one-dimensional projection of $\boldsymbol{U}$, single-index IS generalizes the IS techniques of Chapter 3.

Under a single-index model, $\Psi(\boldsymbol{X})$ is essentially a function of the transformed variable $T = T(\boldsymbol{X})$, where $T : \mathbb{R}^d \to \mathbb{R}$ is some parametric projection function, so we can make the rare-event more frequent by applying IS to T. More specifically, single-index IS draws samples of T from a proposal distribution of T and then draws $\boldsymbol{X} \,|\, T$ under the original distribution. The only conditions that single-index IS requires to work well are that the problem has a strong single-index structure and sampling from $\boldsymbol{X} \,|\, T$ is feasible. As long as the two conditions are met, single-index IS should give large variance reduction. Since the formulation of single-index IS does not assume specific form for Ψ or the distribution of $\boldsymbol{X}$, it is applicable to a wide variety of problems. Moreover, the conditional sampling step of drawing $\boldsymbol{X} \,|\, T$ from the original distribution essentially reduces the dimension of the IS weight function to 1, so single-index IS does not suffer from the dimensionality problem discussed in Section 2.2.3 and works well even in high dimension.

Inspired by the work of GHS [38], we also propose single-index stratified IS (SIS) that combines IS and SS on T in order to achieve further variance reduction. The stratification part of single-index SIS eliminates the variance of $\Psi(\boldsymbol{X})$ captured by the single-index model, which could be as large as 99% in proportion in some problems. In fact, we show that if a drift vector is used as the stratification direction, GHS' IS and stratification techniques [38] are a special case of single-index SIS. Furthermore, single-index IS formulation has a dimension reduction feature, so it enhances the effectiveness of QMC sampling methods if they are used together.

The efficiency of single-index IS comes from exploiting the low-dimensional, namely single-index, structure of the problems at hand. Through literature review, we find that existing IS techniques do not typically take advantage of the possible low-dimensional structure of a given problem. An important application of IS in finance is the estimation of the probability of large losses for a credit portfolio. Glasserman and Li [41] develop IS techniques for Gaussian copula credit portfolio problems based on exponential twisting of default probabilities and mean shifting of the multivariate normal factors. For the same credit portfolio problem, McLeish and Men [86] twist the distribution of the portfolio loss using an extreme value distribution and shift the mean of multivariate normal factors. For

t-copula credit portfolio problems, which are essentially Gaussian models with a common multiplicative shock variable, Bassamboo et al. [12] apply exponential twisting to the shock variable and the default probabilities. In the same paper, Bassamboo et al. propose another IS technique where the distribution of the shock variable is altered based on Hazard-Rate Twisting. In the same t-copula setting, Chan and Kroese [20] use conditional Monte Carlo to analytically integrate out the shock variable and use IS to change the parameters of the underlying multivariate normal variables. None of these methods consider whether or not these credit portfolio problems have a low-dimensional structure. Our simulation studies reveal that credit portfolio problems based on a Gaussian copula have a strong single-index structure and that the t-copula credit problems have a moderate to strong single-index structure depending on the size of the degree of freedom parameter and whether or not the conditional Monte Carlo method proposed in [20] is used. Our proposed single-index IS gives greater variance reduction than Glasserman and Li's IS techniques and when combined with conditional Monte Carlo, it outperforms Chan and Kroese's cross-entropy IS approach. Our simulation studies also show that Asian option pricing problems under the Black-Sholes framework, basket option pricing problems under t-copula models, and the estimation of VaR and ES of equity portfolios based on skew-t copulas also have strong single-index structures and thus single-index IS gives a substantial variance reduction for those problems.

The rest of this chapter is organized as follows. Section 4.2 introduces a single-index model and provides an overview of how single-index IS and SIS achieve variance reduction. Section 4.3 provides the general single-index IS and SIS setup and then derive the variance expressions of the IS and SIS estimators. Based on those expressions, we derive the optimal (variance-minimizing) calibrations for the proposal densities for single-index IS and SIS. The connection between single-index SIS and the IS and stratification techniques in [38] is also shown. Section 4.4 shows that single-index IS reduces the effective dimension of the problem and so it can be seen as a dimension reduction technique. Section 4.5 shows that the stratification part of the SIS scheme is more efficient than control variates in eliminating the variance captured by the single-index model variates. Section 4.6 develops a sampling algorithm for the proposal distribution when $\boldsymbol{X}$ follows a generalized hyperbolic skew-t copula. In Section 4.7, we apply single-index IS and SIS to four problems from finance and numerically evaluate the effectiveness of our proposed methods.

4.2 Semiparametric Single-Index Models and an overview of the single-index (S)IS techniques

In this section, we provide an overview of single-index models and highlight why single-index IS works well for the problems with a strong single-index structure, the structure assumed by single-index model. Readers are refereed to [46], [57], and [95] for more information on single-index models. Let $\Psi : \mathbb{R}^d \to \mathbb{R}$ and $\boldsymbol{X}$ be a d-dimensional random vector whose support, pdf and distribution function are denoted by $\Omega_{\boldsymbol{X}} \subseteq \mathbb{R}^d$, $f_{\boldsymbol{X}}(\boldsymbol{x})$ and $F_{\boldsymbol{X}}(\boldsymbol{x})$, respectively. In rare-event simulation, the goal is often to estimate $\mu = \mathrm{E}[\Psi(\boldsymbol{X})]$ where $\mathbb{P}(\Psi(\boldsymbol{X}) > 0)$ is small. Suppose that Ψ has a single-index structure, that is, there exists some unknown parametric transformation function $T : \mathbb{R}^d \to \mathbb{R}$ such that $\Psi(\boldsymbol{X})$ depends on $\boldsymbol{X}$ mainly through $T = T(\boldsymbol{X})$. Denoting the support, pdf and distribution function of T by Ω_T, $f_T(t)$ and $F_T(t)$, respectively, we have a single-index regression representation

$$\Psi(\boldsymbol{X}) = m(T) + \epsilon_T, \quad \epsilon_T \,|\, T \sim (0, v^2(T)), \tag{4.1}$$

where $m(t) = \mathrm{E}[\Psi(\boldsymbol{X}) \,|\, T = t]$, $v^2(t) = \mathrm{Var}(\Psi(\boldsymbol{X}) \,|\, T = t)$, and ϵ_T is a random error term. Here, $\epsilon \sim (a, b)$ denotes that ϵ follows some distribution with mean a and variance b. This model is called single-index because it assumes that the conditional mean, $\mathrm{E}[\Psi(\boldsymbol{X}) \,|\, \boldsymbol{X}]$, depends on $\boldsymbol{X}$ only through a univariate aggregated information $T = T(\boldsymbol{X})$. The model is semiparametric as it assumes a parametric transformation function $T(\cdot)$ but it does not assume any parametric form for $m(t)$ nor the specific distribution of ϵ_T other than it has a zero mean. By the law of total variance

$$\mathrm{Var}(\Psi(\boldsymbol{X})) = \mathrm{Var}(m(T)) + \mathrm{Var}(\epsilon_T), \tag{4.2}$$

which decomposes the variance of $\Psi(\boldsymbol{X})$ into two pieces: the one captured by the systematic part, $m(T)$, and the other by the random part ϵ_T of the single-index model. The ratio $R^2 = \mathrm{Var}(m(T))/\mathrm{Var}(\Psi(\boldsymbol{X}))$ is the coefficient of determination [72] in regression studies and it measures the fraction of the overall variance explained by the systematic part of the model. In some applications, R^2 is as large as 0.99, implying that $\Psi(\boldsymbol{X})$ is mostly driven by T. We can then apply IS to $\Psi(\boldsymbol{X})$ through changing the distribution of T. In particular, our IS scheme draws T from a proposal distribution of T and then samples $\boldsymbol{X} \,|\, T$ under the original distribution. As our numerical study in Section 4.7 shows, such an IS scheme gives a substantial variance reduction for problems with a single-index structure.

In the stochastic representation (4.1), the form of $T(\cdot)$ is unknown so we must select a specific form of $T(\cdot)$. A popular approach is to assume the parametric form $T(\boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{X}$, for which the model becomes a linear single-index model [57]. We then want to estimate $\boldsymbol{\beta}$, which we call a direction vector, that maximizes the fit of the model. The estimation procedures for such optimal $\boldsymbol{\beta}$ include Ichimura's semiparametric least-squares estimator [57], the average derivative method [114] of Stocker, and the sliced inverse regression [78] of Li. For single-index IS to work, T must satisfy two conditions: the distribution of T is analytical and the conditional sampling of $\boldsymbol{X} \,|\, T$ is feasible. For many distributions of $\boldsymbol{X}$, including the generalized hyperbolic family [88], the two conditions are satisfied under the linear single-index models. We note that all the financial problems considered in the simulation studies have a linear single-index structure.

In this chapter, we also develop a single-index SIS that combines single-index IS and SS on T, following the idea in Glasserman et al. [38]. We give a brief overview of how the SIS scheme accomplishes variance reduction to motivate our work. Recall the variance decomposition $\mathrm{Var}(\Psi(\boldsymbol{X})) = \mathrm{Var}(m(T)) + \mathrm{Var}(\epsilon_T)$. As we will see in Section 4.3.2, the stratification on T essentially stratifies away $\mathrm{Var}(m(T))$, the variance explained by the systematic part of the single-index model. The variance left comes from the random part, ϵ_T, which is potentially less than 1% of the variance of $\Psi(\boldsymbol{X})$, depending on the fit of the model. Noticing that the variance of $\epsilon_T \,|\, T$ depends on the value of T, the IS part of the SIS scheme shifts the distribution of T so that it is proportional to $v(t)f_T(t)$ to minimize the variance contribution from ϵ_T. This form of proposal density has a close resemblance to the Neyman allocation [21, pp. 98-99] in statistical sampling where the optimal allocation is proportional to the product of the stratum probability and the stratum standard deviation.

4.3 Importance Sampling and Stratified Importance Sampling Schemes

4.3.1 Single-index IS and SIS Algorithms

Suppose that the transformation function $T = T(\boldsymbol{X})$ has been selected. We assume that the support of T under the original distribution is an interval $\Omega_T = (t_{\text{inf}}, t_{\text{sup}})$ with possibly $t_{\text{inf}} = -\infty$ and $t_{\text{sup}} = \infty$, but this assumption can be easily generalized. Let $g_T(t)$, and $G_T(t)$ denote the pdf and distribution function of T under the proposal distribution, respectively. Single-index IS draws T from $g_T(t)$ first and then generates $\boldsymbol{X} \mid T$ under the original distribution. Let $t_{\boldsymbol{x}} = T(\boldsymbol{x})$. Since $T(\boldsymbol{x})$ is completely determined by $\boldsymbol{x}$, the conditional density $f_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t)$ of $\boldsymbol{X} \mid T$ under the original distribution is zero if $t \neq t_{\boldsymbol{x}}$. Note that the distribution of $\boldsymbol{X} \mid T$ is identical under the original and the IS distribution by construction, that is, $g_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t_{\boldsymbol{x}}) = f_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t_{\boldsymbol{x}})$. Using this relation, we can write

$$g_{\boldsymbol{X}}(\boldsymbol{x}) = g_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t_{\boldsymbol{x}}) g_T(t_{\boldsymbol{x}}) = f_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t_{\boldsymbol{x}}) g_T(t_{\boldsymbol{x}}).$$

Then the IS weight function becomes

$$w(\boldsymbol{x}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{g_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{f_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t_{\boldsymbol{x}}) f_T(t_{\boldsymbol{x}})}{f_{\boldsymbol{X}|T}(\boldsymbol{x} \mid t_{\boldsymbol{x}}) g_T(t_{\boldsymbol{x}})} = \frac{f_T(t_{\boldsymbol{x}})}{g_T(t_{\boldsymbol{x}})}. \tag{4.3}$$

Thus, the IS weight function is simply the ratio of the original and the IS density of T. As T is univariate regardless of the dimension of $\boldsymbol{X}$, single-index IS is less susceptible to the dimensionality problem discussed in Section 2.2.3. In order to simplify the notation, define $\tilde{w} : \mathbb{R} \to \mathbb{R}$ as $\tilde{w}(t) = \frac{f_T(t)}{g_T(t)}$. For $w(\boldsymbol{x})$ to be well-defined, we need $g_T(t) > 0$ whenever $f_T(t) > 0$. But, we only need $g_T(t) > 0$ whenever $m(t)f_T(t) > 0$ for the IS estimator to be unbiased. Algorithm 5 summarizes this IS scheme.

Remark 4.3.1. *Single-index IS generalizes the IS scheme of Chapter 3 in two ways. Firstly, single-index IS generalizes the form of the transformation function $T(\cdot)$. While single-index IS does not assume any specific form of $T(\cdot)$, the IS of Chapter 3 assumes that $T(\boldsymbol{X}) = \max\{X_1, \ldots, X_d\}$. Secondly, single-index IS generalizes the form of the proposed*

Algorithm 5 Single-index Importance Sampling

```
for $i = 1, \dots, n$ do
    Draw $T_i \sim g_T$
    Draw $\boldsymbol{X}_i \sim f_{\boldsymbol{X}|T}(\boldsymbol{x} \,|\, T_i)$
    Compute $w_i = \tilde{w}(T_i) = f_T(T_i)/g_T(T_i)$.
end for
return $\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n}\sum_{i=i}^{n} \Psi(\mathbf{X}_i) w_i$.
```

density of the transformed variable. The proposal density $g_T(t)$ for the IS of Chapter 3 has a form

$$g_T(t) = \sum_{k=1}^{M} q_k f_{T_h}(t \,|\, T > \lambda_k) = \sum_{k=1}^{M} q_k \frac{f_T(t) I_{\{t>\lambda_k\}}}{1 - F_T(\lambda_k)}, \tag{4.4}$$

where $t_{\mathrm{inf}} = \lambda_1 < \cdots < \lambda_M$, $q_k \geq 0$ and $\sum_{k=1}^{M} q_k = 1$. On the other hand, single-index IS does not impose any restriction on the form of $g_T(t)$, so it is more general.

If T captures a large fraction of the overall variance, that is, if the fit of the single-index model (4.1) is good, we expect that IS on T would give a large variance reduction. In order to achieve further variance reduction, the single-index SIS combines IS and SS on T, inspired by the idea of [38]. The SIS scheme splits the domain of T into n strata of equal probability under G_T and draws one sample of T from each stratum. To do this, let $\lambda_i = {G_T}^{-1}(\frac{i-1}{n})$ for $i = 1, \dots, n+1$, where G_T^{-1} denote the generalized inverse of G_T. Then define the ith stratum as $\Omega_T^{(i)} = [T \in (t_{\mathrm{inf}}, c_{\mathrm{sup}}) \,|\, \lambda_i \leq T < \lambda_{i+1}]$, $i = 1, \dots, n$. By construction, each $\Omega_T^{(i)}$ has probability of $1/n$ under G_T. Algorithm 6 summarizes the SIS scheme.

We note that the combination of IS and SS is not motivated by the same purpose in [38] compared to the single-index SIS of Algorithm 6. In [38], IS are SS are used to remove the variability due to the linear part and the quadratic part, respectively, of $\Psi(\boldsymbol{X})$. In single-index SIS, SS is used to eliminate $\mathrm{Var}(m(T))$, the variance captured by systematic part of the single-index model, and then IS is used to minimize the variance contribution from ϵ_T, the error term in (4.1).

Algorithm 6 Single-index Stratified Importance Sampling Algorithm

```
for $i = 1, \dots, n$ do
    Draw $T_i \sim T \mid \Omega_T^{(i)}$ where $T \sim g_T(t)$
    Draw $\boldsymbol{X}_i \sim f_{\boldsymbol{X}|T}(\boldsymbol{x} \mid T_i)$
    Compute $w_i = f_T(T_i)/g_T(T_i)$.
end for
return $\hat{\mu}_{\text{SIS},n} = \frac{1}{n} \sum_{i=1}^{n} \Psi(\mathbf{X}_i) w_i$.
```

Suppose that $\boldsymbol{X} \sim \text{MVN}(\mathbf{0}, I_d)$. The IS and stratification techniques by GHS [38] shift the mean of $\boldsymbol{X}$ by some drift vector $\mathbf{0} \neq \boldsymbol{\eta} \in \mathbb{R}^d$ so that $\boldsymbol{X} \sim \text{MVN}(\boldsymbol{\eta}, I_d)$ under the IS distribution, and then stratify $\boldsymbol{X}$ along $\boldsymbol{\beta}'\boldsymbol{X}$ for some $\boldsymbol{\beta} \in \mathbb{R}^d$ such that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. In [38], the optimal shift $\boldsymbol{\eta}$ is found by solving some optimization problem and it is argued that setting $\boldsymbol{\beta} = \boldsymbol{\eta}/\sqrt{\boldsymbol{\eta}'\boldsymbol{\eta}}$ often gives a good stratification direction. The following proposition states that the same can be done with single-index IS. This implies that the IS and stratification techniques in [38] that use a normalized drift vector as the stratification direction is a special case of single-index SIS.

Proposition 4.3.2 (see p. 158 for proof)**.** *Suppose that* $\boldsymbol{X} \sim \text{MVN}(\mathbf{0}, I_d)$ *under the original distribution. Fix* $\mathbf{0} \neq \boldsymbol{\eta} \in \mathbb{R}^d$ *and let* $\boldsymbol{\beta} = \boldsymbol{\eta}/\sqrt{\boldsymbol{\eta}'\boldsymbol{\eta}}$*. Consider single-index IS (Algorithm 5) with* $T(\boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{X}$ *and* $T \sim N(\sqrt{\boldsymbol{\eta}'\boldsymbol{\eta}}, 1)$ *under the proposal distribution. Then,* $\boldsymbol{X} \sim \text{MVN}(\boldsymbol{\eta}, I_d)$ *under the IS distribution.*

4.3.2 Variance Analysis and Optimal Calibration for (Stratified) Importance Sampling

In this section, we analyze the variance of single-index IS and SIS estimators defined in Section 4.3.1 and propose calibration methods that minimize the variance of the respective estimators. We first define notation for conditional moments. Recall that $m(t) = \text{E}[\Psi(\boldsymbol{X}) \mid T = t]$ and $v^2(t) = \text{Var}(\Psi(\boldsymbol{X}) \mid T = t)$ and define $m^{(2)}(t) = \text{E}[\Psi^2(\boldsymbol{X}) \mid T = t]$. Note that these conditional moment functions are identical whether $\boldsymbol{X}$ follows the original or the proposal distributions for single-index IS. In what follows, we use the subscript f and

g on expectation, variance, and profitability operators to indicate that they are computed under the original or proposal distribution, respectively.

For a given IS density $g_T(t)$ of T, let $A_t = \{t \in \mathbb{R} \mid g_T(t) > 0\}$ and define the following:

$$\mu_{\text{IS}} = \int_{A_t} m(t) f_T(t) dt,\ \sigma^2_{\text{IS}} = \int_{A_t} m^{(2)}(t) \frac{f_T^2(t)}{g_T(t)} dt - \mu^2_{\text{IS}}, \text{ and } \sigma^2_{\text{SIS}} = \int_{A_t} v^2(t) \frac{f_T^2(t)}{g_T(t)} dt. \quad (4.5)$$

Notice that μ_{IS} depends on $g_T(t)$ through the region A_t of non-zero density of $g_T(t)$. In general, IS and SIS estimators are unbiased only if $g_T(t)$ is such that $\mu_{\text{IS}} = \mu$, but we do not impose this unbiased assumption. The following proposition gives the variance of the IS estimator and the optimal calibration.

Proposition 4.3.3 (see p. 159 for proof)**.** *The mean and the variance of a single-index IS estimator defined as in Algorithm 5 are given by*

$$\mathrm{E}[\hat{\mu}_{\text{IS},n}] = \mu_{\text{IS}}, \quad \mathrm{Var}(\hat{\mu}_{\text{IS},n}) = \sigma^2_{\text{IS}}/n. \quad (4.6)$$

If $\mathrm{E}_g[m^2(T) w^2(T)] < \infty$*, the IS estimator is asymptotically normal as*

$$\sqrt{n}(\hat{\mu}_{\text{IS},n} - \mu_{\text{IS}}) \xrightarrow{d} N(\mu_{\text{IS}}, \sigma^2_{\text{IS}}). \quad (4.7)$$

Suppose that $\Psi(\boldsymbol{x}) \geq 0$ *or* $\Psi(\boldsymbol{x}) \leq 0$ *for all* $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$*. Then the density* $g_T(t)$ *that gives an unbiased IS estimator with the smallest variance is*

$$g_T^{\text{opt}}(t) = \frac{\sqrt{m^{(2)}(t)} f_T(t)}{\int_{t_{\inf}}^{t_{\sup}} \sqrt{m^{(2)}(t)} f_T(t) dt}, \quad t \in (t_{\inf}, t_{\sup}). \quad (4.8)$$

With this choice, the variance of the IS estimator, defined as $\hat{\mu}^{\text{opt}}_{\text{IS},n}$*, is*

$$\mathrm{Var}(\hat{\mu}^{\text{opt}}_{\text{IS},n}) = \frac{1}{n} \left(\left(\int_{t_{\inf}}^{t_{\sup}} \sqrt{m^{(2)}(t)} f_T(t) dt \right)^2 - \mu^2 \right). \quad (4.9)$$

By Jensen's inequality, $\mathrm{Var}(\hat{\mu}^{\text{opt}}_{\text{IS},n}) \leq \mathrm{Var}(\hat{\mu}_{\text{MC},n})$ where the inequality holds as an equality only when $m^{(2)}(t)$ is constant for all $t \in \Omega_T$. The optimal (variance-minimizing) calibration (4.8) requires the knowledge of the conditional second moment function $m^{(2)}(t) =$

$\mathrm{E}_f[\Psi^2(\boldsymbol{X}) \,|\, T = t]$ $\forall t \in \Omega_T$. Using pilot simulations, we can estimate $m^{(2)}(t)$ using nonparametric regression, such as kernel regression [82] or smoothing spline [97]. After approximating $g_T^{\text{opt}}(t)$, we need to draw samples from this density to construct an IS estimator. Numerical inversion techniques such as the NINIGL algorithm of Hörmann, Wolfgang and Leydold [55] is suitable for this purpose as $g_T^{\text{opt}}(t)$ rarely belongs to known parametric family.

The following proposition gives the variance of the SIS estimator and the optimal (variance-minimizing) calibration.

Proposition 4.3.4 (see p. 159 for proof)**.** *The mean and the variance of the SIS estimator defined as in Algorithm 6 are given by*

$$\mathrm{E}[\hat{\mu}_{\text{SIS},n}] = \mu_{\text{IS}}, \quad \mathrm{Var}(\hat{\mu}_{\text{SIS},n}) = \sigma^2_{\text{SIS}}/n + o(1/n). \tag{4.10}$$

where the expression for $\mathrm{Var}(\hat{\mu}_{\text{SIS},n})$ *holds for large enough* n*. If* $\mathrm{E}_g \left|m(T)w(T)\right|^{2+\delta} < \infty$ *for some* $\delta > 0$*, the SIS estimator is asymptotically normal as*

$$\sqrt{n}(\hat{\mu}_{\text{SIS},n} - \mu_{\text{IS}}) \xrightarrow{d} N(\mu_{\text{IS}}, \sigma^2_{\text{SIS}}). \tag{4.11}$$

Suppose that $\Psi(\boldsymbol{x}) \geq 0$ *or* $\Psi(\boldsymbol{x}) \leq 0$ *for all* $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ *and that* $\mathbb{P}_f(v^2(T) = 0,\ m(T) \neq 0) = 0$*. Then the proposal density* $g_T(t)$ *that gives an unbiased SIS estimator with the smallest variance is*

$$g_T^{\text{opt}}(t) = \frac{v(t) f_T(t)}{\int_{t_{\inf}}^{t_{\sup}} v(t) f_T(t) dt}, \quad t \in (t_{\inf}, t_{\sup}). \tag{4.12}$$

With this choice, the variance of the SIS estimator, defined as $\hat{\mu}^{\text{opt}}_{\text{SIS},n}$*, is*

$$\mathrm{Var}(\hat{\mu}^{\text{opt}}_{\text{SIS},n}) = \frac{1}{n} \left(\int_{t_{\inf}}^{t_{\sup}} v(t) f_T(t) dt \right)^2 + o(1/n). \tag{4.13}$$

If $\mathbb{P}_f(v^2(T) = 0,\ m(T) \neq 0) > 0$*, the optimal calibration gives a biased estimator.*

By Jensen's inequality, $\mathrm{Var}(\hat{\mu}^{\text{opt}}_{\text{SIS},n}) \leq \mathrm{Var}(\hat{\mu}_{\text{MC},n})$ where the inequality holds as an equality only when $v^2(t)$ is constant for all $t \in \Omega_T$. The calibration (4.12) requires the knowledge of $v^2(t) = \mathrm{Var}(\Psi(\boldsymbol{X}) \,|\, T = t)$. We can approximate this conditional variance function by fitting a nonparametric regression to the square of the first-order difference of the samples, as proposed by Wang et al. [117]. Unless $m(t) = 0$ for all t, $\mathrm{Var}(\hat{\mu}_{\mathrm{SIS},n}) \leq \mathrm{Var}(\hat{\mu}_{\mathrm{IS},n})$ for the same choice of $g_T(t)$. This in turn implies that $\mathrm{Var}(\hat{\mu}^{\mathrm{opt}}_{\mathrm{SIS},n}) \leq \mathrm{Var}(\hat{\mu}^{\mathrm{opt}}_{\mathrm{IS},n})$. As noted in Proposition 4.3.4, the optimal calibration gives a biased estimator if $\mathbb{P}_f(v^2(T) = 0,\ m(T) \neq 0) > 0$. For many problems, this probability is zero so the optimally calibrated SIS estimator is unbiased. Even if this probability is non-zero so the estimator is biased, it may be possible to debias the estimator, as done for the credit portfolio problem in Section 4.7.3.

Remark 4.3.5. *The calibrations* (4.8) *and* (4.12) *give minimum variance estimators if* $\Psi(\boldsymbol{x}) \geq 0$ *or* $\Psi(\boldsymbol{x}) \leq 0\ \forall\, \boldsymbol{x} \in \Omega$. *This assumption holds for many applications in finance such as when estimating a probability of a certain event, as* $\Psi(\boldsymbol{X})$ *is then an indicator function and when pricing options, as the payoff functions usually take non-negative values. If* $\Psi(\boldsymbol{X})$ *takes both positive and negative values,* $m(t)$ *could be 0 for some values of* t*. We can then improve the optimal calibration by giving zero density over the region where* $m(t) = 0$*. However, since it is generally unknown and hard to estimate for which values of* t *give* $m(t) = 0$*, this improvement may not be implementable. As the objective of this thesis is variance reduction, we call the practice of setting* $g_T(t) = g_T^{\mathrm{opt}}(t)$ *or its approximation as "optimal calibration".*

Remark 4.3.6. *Observe that* $\mathrm{Var}(\hat{\mu}_{\mathrm{SIS},n}) = \sigma^2_{\mathrm{SIS}}/n + o(1/n)$ *does not depend on* $m(t)$*. That is, stratification on* T *asymptotically "stratifies away" the variance captured by the systematic part of the single-index model,* $m(T)$*, so the variance of SIS estimators comes only from the error term,* ϵ_T*, when* n *is large. This in turn means that the stronger the fit of the single-index model is, the greater single-index SIS works. This statement also holds for single-index IS in general. As long as the problem has a strong single-index structure and sampling from* $\boldsymbol{X} \,|\, T$ *is feasible, single-index IS and SIS should give large variance reduction. As those conditions do not assume the specific form for* Ψ *or the distribution of* $\boldsymbol{X}$*, single-index IS and SIS are applicable to a wide range of problems.*

Proposition 4.3.4 asserts the asymptotic normality of an SIS estimator. In order to construct a confidence interval based on this estimator, we must estimate σ^2_{SIS}. Given n samples for an IS estimator, we can estimate σ^2_{IS} defined in (4.5) based on the sample variance as

$$\hat{\sigma}^2_{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} \Psi^2(\boldsymbol{X}_i) w^2(\boldsymbol{X}_i) - \hat{\mu}_{\mathrm{IS},n}. \tag{4.14}$$

The consistency of (4.14) stems from the fact that the IS samples are independently and identically distributed, so the sample variance consistently estimates the population variance. On the other hand, the sample variance of the SIS samples is biased for σ^2_{SIS} as the SIS samples are not identically distributed by construction. In order to construct a consistent estimator for σ^2_{SIS}, we take an approach similar to the one by Wang et al. [117] where the first-order difference of samples is taken to remove the effect of the mean function.

Proposition 4.3.7 (see p. 161 for proof)**.** *Suppose that $G_T(t)$ is the distribution function of T under the proposal distribution and an SIS estimator is constructed as in Algorithm 6 based on n samples. If G_T^{-1}, $\mu(t)$ and $\sigma^2(t)$ are continuously differentiable over the domain of T under the proposal distribution, then*

$$\hat{\sigma}^2_{\mathrm{SIS},n} = \frac{1}{2(n-1)}\sum_{i=1}^{n-1} r_i^2 \tilde{w}^2(T_i) \tag{4.15}$$

is a consistent estimator of σ^2_{SIS}, where $r_i = \Psi(\boldsymbol{X}_{i+1}) - \Psi(\boldsymbol{X}_i)$, $i = 1, \ldots, n-1$.

Proposition 4.3.7 assumes that G_T^{-1} is continuously differentiable which requires that $g_T(t) > 0$ on the support of T under the proposal distribution. This does not hold if there exist intervals where $g_T(t) = 0$. In such a situation, we propose to divide the support of T into disjoint intervals with $g_T(t) > 0$ then apply Proposition 4.3.7 separately to each interval and combine them to obtain $\hat{\sigma}^2_{\mathrm{SIS}}$.

4.4 The Effect of the Indirect Sampling Step on Effective Dimension

Recall that single-index IS draws samples of $\boldsymbol{X}$ indirectly; it generates T first then sample $\boldsymbol{X} \mid T$ under the original distribution. If the problem of interest is not rare-event simulation, IS may not be necessary. Nonetheless, it may be beneficial to take the indirect sampling approach rather than drawing $\boldsymbol{X}$ directly, if samples are drawn using a LDS, that is, QMC is used to estimate μ. This is because, as we show in this section, the indirect sampling step serves as a dimension reduction technique as it transforms the integrand in such a way that its truncation dimension is 1 in proportion $R^2 = \mathrm{Var}(m(T))/\mathrm{Var}(\Psi(\boldsymbol{X}))$, assuming T is sampled using the inversion technique. As discussed in Section 2.3.5, the performance of QMC heavily depends of the effective dimension of the problem, so indirect sampling enhances the effectiveness of QMC if the fit of the single-index model is good.

In order to have a better understanding of the impact of indirect sampling on effective dimension, we first make it explicit how $\boldsymbol{X}$ is sampled. In principle, one can draw samples from any d-dimensional distribution by applying some transformation to a $(d+k)$, $k \geq 0$, dimensional uniform random vector. That is, there exists $\eta : [0,1)^{(d+k)} \to \mathbb{R}^d$ such that $\eta(\mathbf{U}) \sim f_{\boldsymbol{X}}(\boldsymbol{x})$ for $\boldsymbol{U} \sim U[0,1)^{d+k}$. Generally, the transformation function η is not unique and the different choices of η correspond to different sampling methods for $\boldsymbol{X}$.

Suppose $(U_1, \dots, U_{d+k}) \sim U[0,1)^{d+k}$. Let $\eta_1 : [0,1) \to \mathbb{R}$ be defined as $\eta_1(u) = F_T^{-1}(u)$ and $\eta_2 : \mathbb{R} \times [0,1)^{(d+k-1)} \to \mathbb{R}^d$ be a transformation function such that $\eta_2(t, U_2, \dots, U_{d+k}) \sim f_{\boldsymbol{X}|T}(\boldsymbol{x} \,|\, t)$ for $t \in \Omega_T$. Then $\eta : [0,1)^{d+k} \to \mathbb{R}$ defined as

$$\eta(u_1, \dots, u_{d+k}) = \eta_2(\eta_1(u_1), u_2, \dots, u_{d+k})$$

gives samples of $\boldsymbol{X}$ as $\eta(U_1, \dots, U_{d+k}) \sim f_{\boldsymbol{X}}(\boldsymbol{x})$. The problem of estimating the expectation can be then expressed as approximating the integral

$$\mu = \int\limits_{[0,1)^{d+k}} \Psi^*(u_1, \dots, u_{d+k}) d\boldsymbol{u},$$

where $\Psi^* : [0,1)^{d+k} \to \mathbb{R}$ is defined implicitly as $\Psi^*(u_1, \dots, u_{d+k}) = \Psi(\eta(u_1, \dots, u_{d+k}))$.

Note that

$$\begin{aligned}
&\int_{[0,1)^{d+k-1}} \Psi^*(u_1, \dots, u_{d+k}) du_2 \cdots du_{k+d} \\
&= \int_{[0,1)^{d+k-1}} \Psi\left(\eta_2(\eta_1(u_1), u_2, \dots, u_{d+k})\right) du_2 \cdots du_{k+d} \\
&= \mathrm{E}[\Psi(\eta_2(\eta_1(u_1), U_2, \dots, U_{d+k}))], \quad (U_2, \dots, U_{d+k}) \sim U[0,1)^{d+k-1} \\
&= \mathrm{E}[\Psi(\boldsymbol{Y})], \quad \boldsymbol{Y} \sim f_{\boldsymbol{X}|T}(\boldsymbol{X} \mid F_T^{-1}(u_1)) \\
&= \mathrm{E}[\Psi(\boldsymbol{X}) \,|\, T = F_T^{-1}(u_1)] = m(F_T^{-1}(u_1)),
\end{aligned}$$

where the third equality follows as $\eta_2(\eta_1(u_1), U_2, \dots, U_{d+k}) \sim f_{\boldsymbol{X}|T}(\boldsymbol{X} \mid F_T^{-1}(u_1))$ by the construction of η_2. Then recall from Section 2.3.5 that the ANOVA component of Ψ^* for the index set $\{1\}$ is

$$\Psi^*_{\{1\}}(\boldsymbol{u}) = m(F_T^{-1}(u_1)) - \mu,$$

so $\sigma^2_{\{1\}} = \mathrm{Var}(m(T))$ and $\sigma^2_{\{1\}}/\sigma^2 = R^2$. Thus, Ψ^* has a truncation dimension 1 in proportion R^2 and a superposition dimension 1 in proportion greater R^2. If the fit of the single-index model is good, that is, R^2 close to 1, indirect sampling serves as a dimension reduction technique and increase the effectiveness of QMC. The interaction of the indirect sampling and QMC is investigated in detail in the simulation studies in Section 4.7.1.

By construction $\Psi(\boldsymbol{X}) \overset{D}{=} \Psi^*(U_1, \dots, U_{d+k})$ for $(U_1, \dots, U_{d+k}) \sim U[0,1)^d$. Suppose that we apply IS on $\Psi^*(U_1, \dots, U_{d+k})$ by changing only the distribution of U_1. Let $g_{U_1}(u)$ denote the proposal density of U_1. It is easy to check that for any proposal density $g_T(t)$ of T, if we let

$$g_{U_1}(u) = \frac{g_T(F_T^{-1}(u))}{f_T(F_T^{-1}(u))}, \quad u \in [0,1) \tag{4.16}$$

then $\eta_1(U_1) \sim g_T(t)$ when U_1 is sampled from $g_{U_1}(u)$. Thus, the single-index IS scheme that uses $g_T(t)$ as a proposal distribution essentially transforms $\Psi(\boldsymbol{X})$ so that the problem becomes the estimation of $\mu = \mathrm{E}_g[\Psi^*(U_1, \dots, U_{d+k})]$, but using (4.16) as the proposal distribution for U_1, which we recall the variable that accounts for $100R^2\%$ of the variance of $\Psi^*(U_1, \dots, U_{d+k})$. In other words, single-index IS exploits the single-index structure of the problem by transforming it so that the first variable is very important, and then applies IS only to that most important variable.

4.5 Comparison of Stratification and Control Variate

This section shows that the single-index SIS and control variates (CV) with single-index IS (CVIS) estimators asymptotically have the same variance. The connection between the CV and post-stratification is noted by Glynn and Szechtman [42]. Readers are referred to [75, pp. 101-111] for more comprehensive coverage on CV. Suppose that we want to estimate the expectation of $\Psi(\boldsymbol{X})$ and there exists a control variable $C = C(\boldsymbol{X})$ which is correlated with $\Psi(\boldsymbol{X})$ and its exact mean, μ_C, is known. The CV estimator has the form

$$\hat{\mu}_{\mathrm{CV},n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(\boldsymbol{X}_i) + \alpha(\mu_C - C_i), \quad \boldsymbol{X}_i \overset{iid}{\sim} F_{\boldsymbol{X}},$$

where $C_i = C(\boldsymbol{X}_i)$ and α is a constant to be determined. It is easy to check that the CV estimator is unbiased for any value of α. Suppose that $\Psi(\boldsymbol{X})$ and $C(\boldsymbol{X})$ are positively correlated. If $C(\boldsymbol{X}_i) > \mu_C$, the chances are that $\Psi(\boldsymbol{X}_i)$ also exceeds μ, so CV compensates this exceedance by subtracting $\alpha(\mu_C - C_i)$ from $\Psi(\boldsymbol{X}_i)$ for $\alpha > 0$. The same argument holds for the case where $\Psi(\boldsymbol{X})$ and $C(\boldsymbol{X})$ are negatively correlated, other than in that case $\alpha < 0$. In essence, CV uses the correlation between $\Psi(\boldsymbol{X})$ and C to pull the samples of $\Psi(\boldsymbol{X})$ toward its mean.

As in [75, p. 103], the optimal α that minimizes the variance of the CV estimator is

$$\alpha^* = \frac{\mathrm{Cov}(\Psi(\boldsymbol{X}), C)}{\mathrm{Var}(C)}. \tag{4.17}$$

The variance of the CV estimator with α^* is $\mathrm{Var}(\hat{\mu}^{\mathrm{opt}}_{\mathrm{CV},n}) = \frac{1}{n}\mathrm{Var}(\Psi(\boldsymbol{X}))(1-\rho^2_{\Psi(\boldsymbol{X}),C})$, where $\rho_{\Psi(\boldsymbol{X}),C}$ is the correlation coefficient between $\Psi(\boldsymbol{X})$ and C. So,

$$\mathrm{Var}(\hat{\mu}^{\mathrm{opt}}_{\mathrm{CV},n}) = \mathrm{Var}(\hat{\mu}_{\mathrm{MC},n})(1 - \rho^2_{\Psi(\boldsymbol{X}),C}). \tag{4.18}$$

Suppose that we use the conditional mean function of the single-index model as a CV, that is, $C(\boldsymbol{X}) = m(T(\boldsymbol{X}))$ and combine this CV idea with single-index IS, which we refer to as single-index CVIS. Then it is easy to show that $\mu_C = \mu_{\mathrm{IS}}$ defined as (4.5) and $\alpha^* = 1$, so the single-index CVIS estimator is

$$\hat{\mu}_{\mathrm{CVIS},n} = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(T_i)(\Psi(\boldsymbol{X}_i) + \mu_{\mathrm{IS}} - m(T_i)), \quad \boldsymbol{X}_i \overset{iid}{\sim} G_{\boldsymbol{X}},$$

and from (4.10) we have that

$$\begin{aligned} n\mathrm{Var}(\hat{\mu}^{\mathrm{opt}}_{\mathrm{CVIS},n}) &= \mathrm{Var}_g\left(\tilde{w}(T)(\Psi(\boldsymbol{X}) - m(T))\right) = \mathrm{E}_g[\mathrm{Var}(\tilde{w}(T)(\Psi(\boldsymbol{X}) - m(T))\,|\,T] \\ &= \mathrm{E}_g[\tilde{w}^2(T)\sigma^2(T)] = n\mathrm{Var}(\hat{\mu}_{\mathrm{SIS},n}) + o(1). \end{aligned}$$

Thus, single-index SIS and single-index CVIS asymptotically give the same amount of variance reduction. Note that one needs to know $m(t)$ and μ_{IS} to construct the CVIS estimator. If μ_{IS} is known, however, we do not need simulation in the first place. The SIS estimator, on the other hand, requires no knowledge of $m(t)$ nor μ_{IS}, so constructing the SIS estimator is easier than the CVIS estimator, as long as the conditional sampling of $\boldsymbol{X}\,|\,T$ is feasible.

The choice of $C(\boldsymbol{X}) = m(T(\boldsymbol{X}))$ as CV is optimal in the sense that it gives the smallest variance among all CV that depends on $\boldsymbol{X}$ through $T(\boldsymbol{X})$ up to a linear transformation. To see this, observe that for any $\alpha \in \mathbb{R}$, $\tilde{C} : \mathbb{R} \to \mathbb{R}$, and $\mu_C = \mathrm{E}_g[\tilde{C}(T)]$, we have

$$\begin{aligned} &n\mathrm{Var}(\hat{\mu}_{\mathrm{CVIS},n}) = \mathrm{Var}_g\left(w(T)[\Psi(\boldsymbol{X}) + \alpha(\mu_C - \tilde{C}(T))]\right) \\ &= \mathrm{Var}_g\left(\mathrm{E}_g[\tilde{w}(T)(\Psi(\boldsymbol{X}) - \alpha\tilde{C}(T))|T]\right) + \mathrm{E}_g\left[\mathrm{Var}_g(\tilde{w}(T)(\Psi(\boldsymbol{X}) - \alpha\tilde{C}(T))|T)\right] \\ &= \mathrm{Var}_g\left(\tilde{w}(T)(m(T) - \alpha\tilde{C}(T))\right) + \mathrm{E}_g\left[\tilde{w}^2(T)\sigma^2(T)\right] \geq \mathrm{E}_g\left[\tilde{w}^2(T)\sigma^2(T)\right] \end{aligned}$$

and the choice $\tilde{C}(T) = a + b \cdot m(T)$ for any $a, b \in \mathbb{R}$ such that $b \neq 0$ achieves this lower bound for $\alpha^* = 1/b$.

4.6 IS and SIS for the skew-t copula

Single-index (S)IS includes a conditional sampling step of sampling $\boldsymbol{X}\,|\,T$. In this section, we develop the sampling method for this conditional sampling step when $\boldsymbol{X}$ follows the generalized hyperbolic (GH) skew t-copula as in McNeil, Frey, and Embrechts [88]. It takes little effort to generalize this sampling algorithm to the one for a GH copula where GH skew-t copula is a special case.

4.6.1 Formulation and Properties of the skew-t copula

As defined in [56], a random vector $\boldsymbol{X}$ has a d-dimensional GH distribution if it has the following stochastic representation

$$\boldsymbol{X} \overset{D}{=} \boldsymbol{\mu} + W\boldsymbol{\gamma} + \sqrt{W}\boldsymbol{Z}, \tag{4.19}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are the mean and skewness parameters in $\mathbb{R}^d$, respectively, $\boldsymbol{Z} \sim \text{MVN}(\boldsymbol{0}, \Sigma)$, and $W \sim \text{GIG}(\lambda, \chi, \psi)$ is independent of $\boldsymbol{X}$. Here, $W \sim \text{GIG}(\lambda, \chi, \psi)$ means that W follows a generalized inverse Gaussian (GIG) distribution ([88, A.2.5]) with density

$$f_W(w; \lambda, \chi, \psi) = \frac{\chi^{-\lambda}(\sqrt{\chi\psi})^{\lambda}}{2K_{\lambda}(\sqrt{\chi\psi})} w^{\lambda-1} \exp\left(-\frac{1}{2}(\chi w^{-1} + \psi w)\right),$$

where K_λ is a modified Bessel function of the third kind with index λ. See [2] for the details of modified Bessel functions. If $\boldsymbol{\gamma} = \boldsymbol{0}$ and $(\lambda, \chi, \psi) = (\nu/2, \nu, 0)$ for the parameter of W, the distribution of $\boldsymbol{X}$ is the usual symmetric multivariate t distribution with ν degrees of freedom as $\text{GIG}(\nu/2, \nu, 0) \overset{D}{=} \text{IG}(\nu/2, \nu/2)$. Here, $\text{IG}(\alpha, \beta)$ is an inverse gamma (IG) distribution with density

$$f_W(w; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} w^{-\alpha-1} \exp\left(\frac{-\beta}{w}\right).$$

The multivariate normal distribution arises if we further assume $\nu \to \infty$. We have a GH skew-t distribution as in [88] if $W \sim \text{IG}(\nu/2, \nu/2)$ in (4.19). The density of this distribution is derived in p.80 of [88] as

$$f_{st}(\boldsymbol{x}; \nu, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma}) = c\frac{K_{(\nu+d)/2}(\sqrt{(\nu + Q(\boldsymbol{x}))\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}}) \exp((\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}\boldsymbol{\gamma})}{(\sqrt{(\nu + Q(\boldsymbol{x}))\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}})^{-(\nu+d)/2}(1 + (Q(\boldsymbol{x})/\nu))^{(\nu+d)/2}}, \tag{4.20}$$

where $Q(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ and the normalizing constant is

$$c = \frac{2^{1-(v+d)/2}}{\Gamma(v/2)(\pi v)^{d/2}\,|\Sigma|^{1/2}}.$$

The subscript st of f_{st} denotes that it is the density of a skew-t distribution. We denote this distribution and its cumulative distribution function by $t_d(\nu, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ and $F_{st}(\boldsymbol{x}; \nu, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$, respectively.

Since $\boldsymbol{X} \mid W = w \sim N(\boldsymbol{\mu} + w\boldsymbol{\gamma}, w\Sigma)$, the first two moments can be derived as

$$\begin{aligned}\mathrm{E}[\boldsymbol{X}] = \mathrm{E}[\mathrm{E}[\boldsymbol{X} \mid W]] &= \boldsymbol{\mu} + \frac{\nu}{\nu-2}\boldsymbol{\gamma}, \\ \mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}[\mathrm{Var}(\boldsymbol{X} \mid W)] + \mathrm{Var}(\mathrm{E}[\boldsymbol{X} \mid W]) &= \frac{\nu}{\nu-2}\Sigma + \frac{2\nu^2}{(\nu-2)^2(\nu-4)}\boldsymbol{\gamma}\boldsymbol{\gamma}'\end{aligned}$$

as in [28]. The covariance of skew-t distributions is finite only when $\nu > 4$ while it is finite when $\nu > 2$ for symmetric t distributions, so the skewed one imposes a stronger restriction on ν compared to the symmetric counterpart. A useful property of a skew-t distribution is that it is closed under affine transformations. In particular, for any $h \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^d$, we have that $h + \boldsymbol{\theta}'\boldsymbol{X} \sim t_1(\nu, h + \boldsymbol{\theta}'\boldsymbol{\mu}, \boldsymbol{\theta}'\Sigma\boldsymbol{\theta}, \boldsymbol{\theta}'\boldsymbol{\gamma})$. We refer to the copula implied by (4.20) as a GH skew-t copula. In particular, we denote by $C^t_{\nu,P,\boldsymbol{\gamma}}$ the copula of a $t_d(\nu, \mathbf{0}, P, \boldsymbol{\gamma})$ distribution, where P is a correlation matrix. Note that taking $\boldsymbol{\mu} = \mathbf{0}$ as location parameters of a random vector has no effect on their copula. More specifically, for $\boldsymbol{u} \in [0,1)^d$

$$C^t_{\nu,P,\boldsymbol{\gamma}}(\boldsymbol{u}) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_d} f_{st}(\boldsymbol{x}; \nu, \mathbf{0}, P, \boldsymbol{\gamma}) d\boldsymbol{x}, \tag{4.21}$$

where $t_i = {F_{st}}^{-1}(u_j; \nu, 0, 1, \gamma_j)$ for $j = 1, \ldots, d$. The density of this copula, denoted by $c^t_{\nu,P,\boldsymbol{\gamma}}$, is

$$c^t_{\nu,P,\boldsymbol{\gamma}}(\boldsymbol{u}) = \frac{f_{st}(\boldsymbol{x}; \nu, 0, P, \boldsymbol{\gamma})}{\prod_{i=1}^d f_{st}(x_i; \nu, 0, 1, \gamma_i)}, \quad \boldsymbol{u} \in [0,1)^d, \tag{4.22}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $x_j = {F_{st}}^{-1}(u_j; \nu, 0, 1, \gamma_j)$ for $j = 1, \ldots, d$.

The advantage of a skew-t copula over a symmetric one is that the former accommodates asymmetric upper and lower tail dependencies while the latter is limited to the symmetric cases. It is well-accepted that equity returns have greater correlation for downside moves than upside moves (see for example [6] and references therein), supporting the use of a skewed copula for modelling financial returns. Banachewicz and van der Vaart have derived tail coefficients of skew-t copulas in [11]. Let $X_1 = \gamma_1 + \sqrt{W}Z_1$ and $X_2 = \gamma_2 + \sqrt{W}Z_2$ where $W \sim \mathrm{IG}(\nu/2, \nu/2)$ and Z_1 and Z_2 are standard normals with correlation coefficient ρ. The upper tail dependence coefficient λ_u of (X_1, X_2) is given by:

- If $\gamma_1, \gamma_2 > 0$, $\lambda_u = 1$.

- If $\gamma_1 < 0$ or $\gamma_2 < 0$, $\lambda_u = 1$.
- If $\gamma_1 = \gamma_2 = 0$, $\lambda_u = 2t_{\nu+1}(-\sqrt{\nu+1}\sqrt{1-\rho}/\sqrt{1+\rho})$.
- If $\gamma_1 > 0$ and $\gamma_2 = 0$, $\lambda_u = \int\limits_0^1 (1 - \Phi(k_\nu u^{1/\nu}))du$ with $k_\nu = \left(\frac{2^{\nu/2}\Gamma((\nu+1)/2)}{2\sqrt{\pi}}\right)^{1/\nu}$,

As noted in [59], since $f_{\boldsymbol{X}}(x_1, x_2; \gamma_1, \gamma_2, \rho, \nu) = f_{\boldsymbol{X}}(-x_1, -x_2; -\gamma_1, -\gamma_2, \rho, \nu)$, the lower tail dependence coefficient λ_l is equal to the upper tail dependence coefficient with parameters $(-\gamma_1, -\gamma_2, \rho, \nu)$. If a skew-$t$ copula is fitted to bivariate negative daily log-returns of stocks, the skewness parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ are likely to be both positive, thus the fitted copula is comonotonic in the upper-right tail. While this seems restrictive and limits the applicability of GH skew-t copulas, the convergence in the tails are not fast and so this extreme dependence at the limit may not be a problem in actual modelling, as discussed by Joe [59].

4.6.2 Transformation function and sampling algorithm for GH skew-t copulas

Recall from (4.21) that the copula $t_d(\nu, \mathbf{0}, P, \boldsymbol{\gamma})$ of $C^t_{\nu,P,\boldsymbol{\gamma}}$. If $\boldsymbol{X} \sim t_d(\nu, \mathbf{0}, P, \boldsymbol{\gamma})$ then $\boldsymbol{X}$ has the stochastic representation

$$\boldsymbol{X} \overset{D}{=} W\boldsymbol{\gamma} + \sqrt{W}\boldsymbol{Z}, \tag{4.23}$$

where $W \sim \mathrm{IG}(\nu/2, \nu/2)$ and $\boldsymbol{Z} \sim N(\mathbf{0}, P)$, so we can model a problem in terms of $(W, \boldsymbol{Z})$. In order to ensure that the conditional sampling of $\boldsymbol{X} \,|\, T$ is feasible, we use the transformation $T = T(W, \boldsymbol{Z}) = \beta_0 + \beta_1 W + \boldsymbol{\beta}_2'\sqrt{W}\boldsymbol{Z}$ for some constants β_0, $\beta_1 \in \mathbb{R}$ and a vector $\boldsymbol{\beta}_2 \in \mathbb{R}^d$. Then $T \sim t_1(\nu, \beta_0, \sigma^2_{\boldsymbol{\beta}_2}, \beta_1)$ where $\sigma^2_{\boldsymbol{\beta}_2} = \boldsymbol{\beta}_2' P \boldsymbol{\beta}_2$. Algorithm 7 shows the steps for sampling from $\boldsymbol{X} \,|\, T = t$. We explain why this algorithm returns variables with the desired distribution. Using the conditional sampling argument, sampling from $\boldsymbol{X} \,|\, T$ is equivalent to first drawing from $W \,|\, T$, then generating $\boldsymbol{Z} \,|\, \boldsymbol{\beta}_2'\boldsymbol{Z}$. The output variable of the algorithm $\boldsymbol{X} = W\boldsymbol{\gamma} + \sqrt{W}\boldsymbol{Z}$ follows the desired conditional distribution. For the first step, we use the result in [1] that $W \,|\, T = t \sim \mathrm{GIG}\left(-\frac{\nu+1}{2}, \nu + \left(\frac{t-\beta_0}{\sigma_{\boldsymbol{\beta}_2}}\right)^2, \left(\frac{\beta_1}{\sigma_{\boldsymbol{\beta}_2}}\right)^2\right)$. Once we draw W from this GIG distribution, we have $\boldsymbol{\beta}_2'\boldsymbol{Z} = (t - \beta_0 - \gamma W)/\sqrt{W}$ and we denote this quantity by λ. For the last step, we use the result of [48] to get $\boldsymbol{Z} \mid \boldsymbol{\beta}_2'\boldsymbol{Z} = \lambda \sim N\left(\frac{\lambda}{\sigma^2_{\boldsymbol{\beta}_2}}\boldsymbol{\beta}_2, P - P\boldsymbol{\beta}_2\boldsymbol{\beta}_2'P/\sigma^2_{\boldsymbol{\beta}_2}\right)$.

Algorithm 7 Sampling $\boldsymbol{X} \mid \beta_0 + \beta_1 W + \sqrt{W}\boldsymbol{\beta}_2'\boldsymbol{Z} = t$ for skew-t copula

```
Draw $W \sim GIG\left(-\frac{\nu+1}{2}, \nu + \left(\frac{t-\beta_0}{\sigma_{\boldsymbol{\beta}_2}}\right)^2, \left(\frac{\beta_1}{\sigma_{\boldsymbol{\beta}_2}}\right)^2\right)$
Let $\lambda = \frac{t-\beta_0-\beta_1 W}{\sqrt{W}}$
Draw $\boldsymbol{Z} \sim N\left(\frac{\lambda}{\sigma^2_{\boldsymbol{\beta}_2}}\boldsymbol{\beta}_2, P - P\boldsymbol{\beta}_2\boldsymbol{\beta}_2'P/\sigma^2_{\boldsymbol{\beta}_2}\right)$
Return $\boldsymbol{X} = W\boldsymbol{\gamma} + \sqrt{W}\boldsymbol{Z}$.
```

The only non-trivial part of Algorithm 7 is the first step, sampling W from a GIG distribution. Since one of its parameters depends on the conditioning value $T = t$, which changes for each sample, we need GIG generators that support varying parameters. In R, there are several packages such as the "ghyp" package [81] that implement te said generators. They are, however, based on rejection sampling, which do not go well with QMC. QMC requires the quantile function of GIG distributions but they are very computationally expensive to evaluate. To reduce the computational effort, we propose to use the MC step (i.e. rejection sampling) to sample W even if the interest is in constructing QMC estimators. Using MC for some variables and QMC for others may hinder the effectiveness of QMC depending on how important the MC generated variables are. One way to circumvent this problem is to set $\beta_1 = 0$, that is, we remove the linear component in W from $T(W, \boldsymbol{Z})$. In this case, $W \mid T$ is inverse gamma distributed and the quantile function can be evaluated very quickly, even if the parameters vary, so taking this route reduces the overall computational efforts for the IS and SIS schemes. Nonetheless, if the linear part in W is significant, it will reduce the effectiveness of the IS and SIS schemes as the fit of the single-index model will not be as good. We investigate this point numerically in Section 4.7.5.

4.7 Numerical Experiments

In this section, we apply single-index IS and SIS to four problems in finance and numerically investigate their effectiveness. We also examine different aspects of the proposed methods such as its dimension reduction effect in each of the four problems. In Section 4.7.1, we apply IS and SIS to the pricing of arithmetic Asian options under the Black-Scholes framework. We look into the dimension reduction aspect of conditional sampling and also compare stratification to CV. Section 4.7.2 considers the pricing of basket option under a t-copula model. We empirically analyze the finite sample properties of $\widehat{\text{Var}}(\hat{\mu}_{\text{SIS},n})$ defined as in Proposition 4.3.7 as a estimate of $\text{Var}(\hat{\mu}_{\text{SIS},n})$. In Section 4.7.3 and Section 4.7.4, we look at credit portfolio problems under the Gaussian and t-copula assumptions. The important finding from the credit portfolio problems is that the proposal distributions based on optimal calibration for IS and SIS could perform very poorly if multiple quantities are to be estimated in one simulation run. This observation motivates other calibration methods and we explore such calibrations in Chapter 5. In Section 4.7.5, we estimate tail quantities such as VaR and ES of equity portfolios. The focus on this section is the efficiency of the IS and SIS schemes when model deviate significantly from the multivariate normal assumption. In particular, the model considered has marginals with heavy tails and the skewed-t copula as dependence structure. We also compare the performance of two forms of $T(\cdot)$; one with better fit but slower conditional sampling and the other with worse fit but faster sampling.

4.7.1 Arithmetic Asian Option Pricing

For arithmetic Asian option pricing in the Black-Scholes framework, it is widely known that geometric Asian options serve as excellent CV [121]. As the payoff of geometric Asian options has a single-index structure, it is natural to expect that the same holds for the payoff of arithmetic Asian options. Thus, this problem is an ideal candidate for our IS scheme.

Problem Formulation

Suppose that under the risk neutral measure the price of a stock follows a geometric Brownian motion

$$dS_t = rS_t dt + \sigma S_t dW_t, \tag{4.24}$$

where S_t is the price of the stock at time t, r is the risk-free rate, σ is the volatility of the stocks price, and W_t is a Brownian motion. Fix $T \in [0, \infty)$ and $d \in \mathbb{N}$. Let $\Delta t = T/d$ and $t_j = j\Delta t$ for $j = 1, \ldots, d$. The stock price at time t_j has the representation

$$S_{t_j} = S_0 \exp\left((r - \sigma^2/2)t_j + \sigma X_{t_j}\right) \tag{4.25}$$

under the risk-neutral measure, where $X_{t_j} = N(0, t_j)$. Let $\boldsymbol{X} = (X_{t_1}, \ldots, X_{t_d})'$, then by the properties of the Brownian motion $\boldsymbol{X} \sim \mathrm{MVN}(\mathbf{0}, \Sigma)$ where $\Sigma_{k,l} = \Delta t \cdot \min(k, l)$. Let $\boldsymbol{v} = (v_1, \ldots, v_d)$ be a vector of weights such that $\sum_{i=1}^d v_i = 1$. Suppose that the payoff of the option at maturity is $\max(S_a - K, 0)$ where $S_a = \sum_{i=1}^d v_i S_{t_j}$ is the weighted arithmetic average of the stock prices observed at time $t_1 \ldots, t_d$. By risk-neutral pricing (see [37, pp. 27-30]), the price of an arithmetic Asian option with strike K is

$$c_a = \exp(-rT)\mathrm{E}[\max(S_a - K, 0)],$$

where $S_a = \sum_{i=1}^d v_i S_{t_j}$ is the weighted arithmetic average of the stock prices observed at time $t_1 \ldots, t_d$. Since the distribution of S_a is not analytical, there is no closed form solution for c_a. One option is to use MC simulation to estimate c_a.

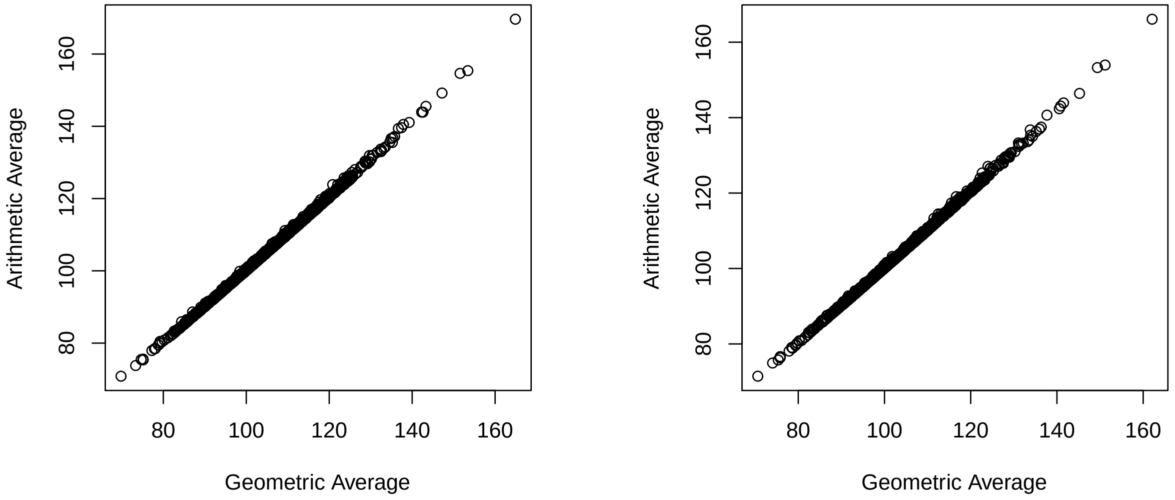
The payoff of a geometric Asian option is highly correlated with its arithmetic counterpart. The price of the geometric option is

$$c_g = \exp(-rT)\mathrm{E}[\max(S_g - K, 0)],$$

where $S_g = \prod_{i=1}^d S_{t_j}^{w_i} = \exp(b + \sigma \boldsymbol{v}'\boldsymbol{X})$ is the weighted geometric average with $b = \log S_0 + (r - \sigma^2/2)\sum_{i=1}^d v_i t_j$. Since the distribution of S_g is $\mathrm{LN}(a, \sigma^2 \boldsymbol{v}'\Sigma\boldsymbol{v})$, a log-normal distribution, the closed-form expression of c_g is easily found, see [121]. Moreover, the payoff $\max(S_g - K, 0)$ depends on $\boldsymbol{X}$ only through $\boldsymbol{v}'\boldsymbol{X}$. Thus, the payoff of geometric Asian options have a perfect linear single-index structure with the direction vector $\boldsymbol{v}$. Given that the payoff of the two types of options are highly correlated, we expect that the arithmetic payoff has a strong linear single-index structure with the same direction vector. This is a rare case where the optimal choice of the direction vector is analytical. The price process (4.24) assumes that the variance of the log-returns is constant over time. We note that the arithmetic payoff has a strong linear single-index structure even under the Heston model (see [50]) where the variance of the log-returns itself follows a stochastic process.

The parameters that we use for our experiments are taken from [121]: $r = 0.1$, $\sigma = 0.2$, $S_0 = 100$, $T = 1$, $d \in \{16, 64\}$, and $K \in \{100, 110\}$. Figure 4.1 shows the scatter plot of the 1,000 realization of (S_g, S_a) for the $d = 16$ and $d = 64$ cases. As the figure shows, there is an almost perfect linear correlation between S_g and S_a, even at their tails, with the correlation coefficient over 0.999 for both $d = 16$ and $d = 64$ cases. This suggests SIS will be effective for this problem.

Figure 4.1: Scatter plots of (S_g, S_a) for $d = 16$ and $d = 64$



Choice of Covariance Decomposition

The payoff of both options is a function of $\boldsymbol{X} \sim \text{MVN}(\mathbf{0}, \Sigma)$ where $\Sigma_{k,l} = \Delta t \cdot \min(k, l)$. Let B be a matrix such that $BB' = \Sigma$. Then $\boldsymbol{X} \overset{D}{=} B\boldsymbol{Z}$ where $\boldsymbol{Z} \sim \text{MVN}(\mathbf{0}, I_d)$, where I_d is the $d * d$ identity matrix. The decomposition of B such that $BB' = \Sigma$ is not unique and different choices of B correspond to different ways of generating the path of the Brownian motion. In MC, the choice of B has no effect on the variance of the estimator because the sampled $\boldsymbol{X}$ has the same distribution regardless of the choice of B. The variance of QMC estimators, on the other hand, often depends on the choice of B primarily because the effective dimension of the problem is influenced by the choice of B.

Cholesky decomposition is the most standard way of constructing B. Thus, we call this STD decomposition. Another popular one is based on the eigenvalue decomposition of Σ. Since this decomposition has a close connection to the principal component analysis (PCA) (see [3]), we call this PCA decomposition. The problem with these two decompositions is that they do not take the nature of the integrand Ψ into account. As long as the covariance matrix C is the same, STD and PCA decompositions always return the same B regardless of how Ψ depends on $\boldsymbol{X}$. Naturally, STD works better than PCA for some problems and vice versa. The work by Sloan and Wang [121] generalizes this idea and states that no fixed decomposition is superior to others. Based on this observation, Wang and Sloan propose Orthogonal Transformation (OT) that takes the integrand into account for finding B. Their main idea is to find a good decomposition for an easy problem and apply it to related problems. In Asian option pricing, geometric option is the easy problem. Recall that S_g depends on $\boldsymbol{X}$ through $\boldsymbol{v}'\boldsymbol{X}$. For a fixed decomposition B, we can write $S_g = h(\boldsymbol{v}'\boldsymbol{X}) = h(\boldsymbol{v}'B\boldsymbol{Z})$ where $h(x) = \exp(a + \sigma x)$. The OT approach constructs B such that $\boldsymbol{v}'B\boldsymbol{Z} = cZ_1$ for some constant $c \in \mathbb{R}$. That is, the geometric average is determined by the first component of $\boldsymbol{Z}$. So if this decomposition is used for geometric option pricing, the problem becomes one-dimensional. The OT approach then uses this B to price the arithmetic option. Since the arithmetic payoff is almost perfectly correlated with the geometric payoff, the first element of $\boldsymbol{Z}$ captures the large majority of the overall variance. Thus, the truncation dimension of the arithmetic option problem with this choice of B will be in proportion over 99.9% for typical sets of model parameters.

The indirect sampling also provides a dimension reduction feature as discussed in Section 4.4. Suppose that we take $T = S_g = h(\boldsymbol{v}'X)$. The indirect sampling step first generates T then samples $\boldsymbol{X} \,|\, T$ under the original distribution. Assuming that T is sampled using the inversion technique, the first input variable determines the geometric average. Recalling that the first input variable also determines the geometric average under OT, the truncation dimension of the arithmetic option problem under single-index IS is 1 in the same proportion as under the OT, which is over 99.9% for typical sets of model parameters. This in turn implies that other variables are fairly irrelevant under OT and indirect sampling. Since OT and indirect sampling draw the most important variable in the same manner, we expect that their performance are comparable for this problem. We test this claim numerically.

Since $\boldsymbol{X} \mid T$ follows a multivariate normal distribution with some covariance matrix Σ_2, we test whether decomposing Σ_2 by STD and PCA has any effect on the variance of the estimator. Two types of weights are considered: Type A is the equal weights $v_j = 1/d$, $j = 1, \ldots, d$ and Type B is the weights of alternating sign $v_j = c(-1)^{j-1}/j$, $j = 1, \ldots, d$, where c is the normalizing constant. The Type B weights are chosen somewhat artificially to illustrate that STD could produce a more QMC friendly decomposition than PCA does, depending on the weights.

Table 4.1 lists the variance reduction factors (VRFs) of QMC estimators with different decompositions over plain MC estimators. I-STD and I-PCA denote indirect sampling with the STD and PCA decomposition of Σ_2, respectively. The table shows that whether we decompose Σ_2 by STD or PCA, it does not have a significant effect on the variance of the QMC estimator when indirect sampling is used. Also, PCA does better than STD for type A weights but the reverse holds for type B. Both OT and indirect sampling work well for either type of weights. As expected, there is no significant difference between OT and indirect sampling in terms of VRFs.

Table 4.1: Variance reduction factors of different decompositions, $n = 2^{15}$, 30 replications

		$K = 100$					$K = 120$				
Type	d	STD	PCA	OT	I-STD	I-PCA	STD	PCA	OT	I-STD	I-PCA
A	16	1.8E+02	6.4E+03	5.9E+03	6.3E+03	6.4E+03	7.1E+01	3.9E+03	3.4E+03	3.7E+03	3.7E+03
A	64	6.7E+01	4.9E+03	5.9E+03	6.2E+03	6.1E+03	2.3E+01	2.9E+03	3.4E+03	3.9E+03	3.5E+03
B	16	1.3E+02	7.8E+01	5.3E+03	5.9E+03	5.3E+03	4.3E+01	6.6E+00	9.2E+02	9.2E+02	8.0E+02
B	64	2.1E+02	3.8E+01	6.6E+03	6.1E+03	5.2E+03	6.0E+00	1.1E+00	2.4E+01	2.4E+01	2.3E+01

Comparison of VRT

We compare the performance of single-index SIS with the IS and stratification techniques of Glasserman, Heidelberger and Shahabuddin [38], which we refer to as GHS IS, and with the CV method that uses geometric option as a control variate with no IS feature. We consider the Asian option with equal weights $v_i = 1/d$, $i = 1, \ldots, d$. Table 4.2 and Table

4.3 list the VRFs with and without QMC for $d \in \{16, 64\}$ and $K \in \{100, 110\}$ for $n = 2^{13}$ and $n = 2^{15}$. The VRFs are computed based on 30 replications. For the plain QMC estimator, OT is used to decompose the covariance matrix Σ.

Table 4.2: Variance reduction factors of different decompositions, $n = 2^{13}$, 30 replications

		MC			QMC			
d	K	CV	GHS IS	SIS	Plain	CV	GHS IS	SIS
16	100	1.12E+03	3.63E+03	2.49E+03	2.04E+03	2.44E+04	2.19E+05	4.93E+05
16	110	5.86E+02	5.17E+03	2.46E+03	1.17E+03	3.21E+03	2.44E+05	7.12E+05
64	100	1.14E+03	5.37E+03	2.25E+03	1.93E+03	3.08E+03	8.45E+04	1.86E+05
64	110	5.82E+02	6.81E+03	2.18E+03	1.06E+03	1.15E+03	5.45E+04	2.00E+05

Table 4.3: Variance reduction factors of different decompositions, $n = 2^{15}$, 30 replications

		MC			QMC			
d	K	CV	GHS IS	SIS	Plain	CV	GHS IS	SIS
16	100	1.14E+03	4.40E+03	2.91E+03	5.87E+03	2.64E+04	6.92E+05	8.91E+05
16	110	5.99E+02	5.11E+03	3.11E+03	3.36E+03	3.92E+03	5.81E+05	9.64E+05
64	100	1.14E+03	3.81E+03	2.66E+03	5.88E+03	9.34E+03	1.91E+05	6.24E+05
64	110	5.78E+02	3.98E+03	2.83E+03	3.37E+03	2.03E+03	1.54E+05	4.09E+05

From the table, we see that GHS IS performs better than SIS for MC but the other way around for QMC. Both techniques give larger variance reduction than CV does, especially when combined with QMC. This is more so when $K = 110$ than when $K = 100$. The reason is that the importance sampling part of SIS and GHS IS shifts the underlying distribution toward the region of non-zero payoff but CV does not have this IS feature. When K is large, there is lower chance of non-zero payoff, so the IS becomes more effective when $K = 110$ than when $K = 100$.

4.7.2 Pricing of Basket Option

Problem Formulation

The experiments in the previous section showed that single-index IS and SIS work well for the Asian option pricing problem where the underlying distribution of the model is multivariate normal. The primary focus of the simulation studies in this section is to analyze how our proposed methods work when the underlying model deviates from multivariate normality. The problem is the pricing of basket options when the log-returns of individual stocks marginally follows a normal distribution but their dependence is described by a t-copula. We chose a t-copula because it is commonly used in financial modelling and it exhibits tail dependence which Gaussian copulas lack. Kole et al. [68] fit Gaussian, t, and Gumbel copulas to the daily returns of indices on stocks, bonds and real estate and find that only t-copula is not rejected with the estimated ν, the degrees of freedom parameter, being 12.1. Note that the Gaussian copula is a special case of a t-copula where $\nu = \infty$.

A basket option is similar to an Asian option in that the payoff of both options depend on the average of stock prices. The two differ in that the average price or return of multiple stocks is used for basket options while the average price of a single stock over time is used for Asian options. Suppose that the price of an asset j, $j = 1, \dots, d$ at time T under the risk-neutral measure is given by

$$S_{j,T} = S_{j,0} \exp\{(r - \varrho_j - \sigma_j^2/2)T + \sigma_j \sqrt{T} Y_j\}, \tag{4.26}$$

where $S_{j,0}$, ϱ_j, σ_j denote the price at time 0, dividend rate and volatility of stock j, respectively, r denotes the risk-free rate, and Y_j's follows a $N(0,1)$ marginally and collectively follow a t-copula. The payoff of the option is a function of the weighted average of the individual stocks returns and the option price can be written as

$$c_b = \exp(-rT)\mathrm{E}\left[\max\left(\sum_{j=1}^{d} v_i \frac{S_{j,T}}{S_{j,0}} - K, 0\right)\right], \tag{4.27}$$

where $\boldsymbol{v} = (v_1, \dots, v_d)'$ is the vector of portfolio weights and K is the strike price.

Under a t-copula model, we have that $\boldsymbol{Y} \overset{D}{=} (\Phi^{-1}(t_\nu(X_1)), \dots, \Phi^{-1}(t_\nu(X_d)))$ for $\boldsymbol{X} = (X_1, \dots, X_d) \sim t(\mathbf{0}, P, \nu)$ where ν is the degrees of freedom parameter, P is an $d * d$

correlation matrix, and t_ν is the distribution function of a t-distribution. For B such that $BB' = P$, $\boldsymbol{Z} \sim \text{MVN}(\boldsymbol{0}, I_d)$ and $W \sim \text{IG}(\nu/2, \nu/2)$, an inverse gamma distribution, we can write $\boldsymbol{X} \overset{D}{=} \sqrt{W} L \boldsymbol{Z}$. We let $T = T(W, \boldsymbol{Z}) = \sqrt{W}\boldsymbol{\beta}'\boldsymbol{Z}$ where $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. We could include a linear term in W for $T(\cdot)$, that is, letting $T = \beta_0 W + \sqrt{W}\boldsymbol{\beta}'\boldsymbol{Z}$, but doing so hardly improves the fit of the single-index model while it makes the conditional sampling more computationally expensive, so we opt not to take this route.

The parameters we use are taken from the numerical example in [89] where the pricing of a basket option on the market indices of G7 nations ($d = 7$) are considered. For the t-copula model, we consider the case with a small degrees of freedom ($\nu = 4$) and one with a large degrees of freedom ($\nu = 10$) to investigate how different values of ν affect the performance of single-index IS and SIS.

Figure 4.2 shows the scatter plot of T against the weighted average of the returns based on 10,000 observations for different distributions of $\boldsymbol{X}$: normal, t_4 and t_{10}. The figure shows that the relationship between T and the average return is generally linear and the fit of the single-index model is fairly good for the normal and t_{10} models but it is worse for the t_4 model especially at both tails. From the plots, we expect that single-index IS and SIS work better for the normal and t_{10} model than the t_4 model.

Finite sample properties of the standard error of the IS and SIS estimators

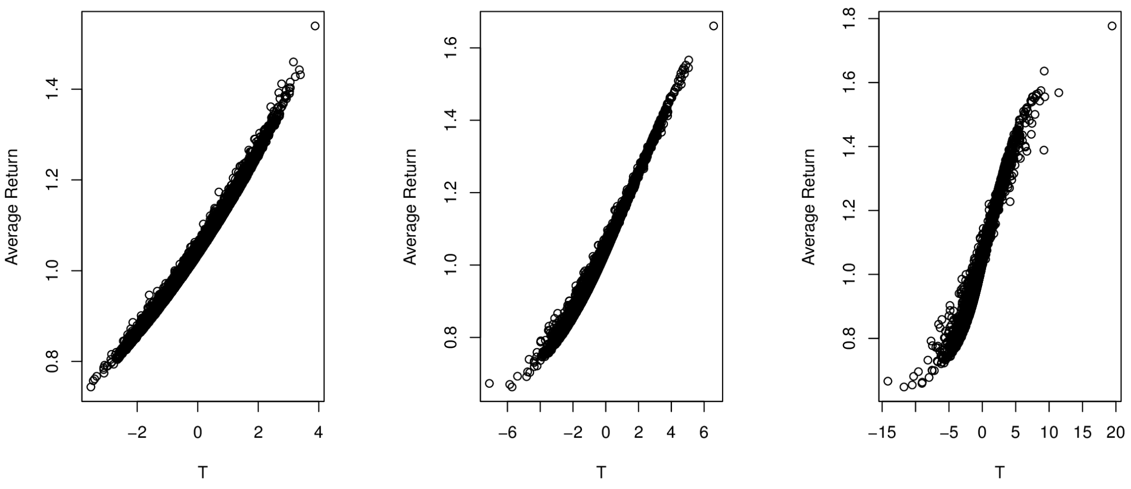
Before analyzing the performance of IS and SIS, we numerically investigate the finite sample properties of $\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n}) := \hat{\sigma}^2_{\text{IS}}/n$ and $\widehat{\text{Var}}(\hat{\mu}_{\text{SIS},n}) := \hat{\sigma}^2_{\text{SIS}}/n$, the squared standard error of $\hat{\mu}_{\text{IS},n}$ and $\hat{\mu}_{\text{SIS},n}$, respectively, where $\hat{\sigma}^2_{\text{IS}}$ and $\hat{\sigma}^2_{\text{SIS}}$ are defined in (4.14) and (4.15), respectively. The approximate $100(1-\alpha)\%$ CI for $\hat{\mu}_{\text{IS},n}$ is constructed as

$$\left(\hat{\mu}_{\text{IS},n} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n})},\, \hat{\mu}_{\text{IS},n} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n})}\right),$$

where z_α denotes the point at which $\mathbb{P}(Z \le z_\alpha) = \alpha$ for $Z \sim N(0,1)$. So, it is important that $\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n})$ is not too far off from $\text{Var}(\hat{\mu}_{\text{IS},n})$ for the CI to be meaningful. The same argument holds for $\widehat{\text{Var}}(\hat{\mu}_{\text{SIS},n})$.

To analyze the finite sample properties of $\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n})$, we repeat the IS procedures 1,000 times under the normal model with $n = 10,000$ samples each. For each replication, we

Figure 4.2: Plot of Transformed variable (T) vs Average Return based on 10,000 observations



(a) normal model

(b) t_{10} model

(c) t_4 model

compute $\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n})$, so we end up with 1000 realization of $\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n})$. We repeat the same procedure for plain MC, SS (SIS without IS), and SIS. For plain MC, $\widehat{\text{Var}}(\hat{\mu}_{\text{MC},n})$ is computed based on the sample variance.

Table 4.4 and Figure 4.3 show the summary statistics and histograms of the 1,000 realized squared standard error for the plain MC, SS, IS, and SIS estimators. The variability of $\widehat{\text{Var}}(\hat{\mu}_{\text{SS},n})$, $\widehat{\text{Var}}(\hat{\mu}_{\text{IS},n})$, and $\widehat{\text{Var}}(\hat{\mu}_{\text{SIS},n})$ appear to be compatible to the variability of $\widehat{\text{Var}}(\hat{\mu}_{\text{MC},n})$. As the variability of the squared standard errors are small, the CIs for IS, SS, and SIS estimators are reliable, at least for this basket option pricing problem.

Table 4.4: Summary statistics of the 1,000 realized squared standard error

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
Plain MC	4.881e-07	5.197e-07	5.265e-07	5.264e-07	5.329e-07	5.632e-07
SS	2.453e-09	2.679e-09	2.743e-09	2.747e-09	2.813e-09	3.130e-09
IS	7.236e-08	7.465e-08	7.538e-08	7.535e-08	7.603e-08	7.855e-08
SIS	1.764e-09	1.956e-09	1.996e-09	1.998e-09	2.040e-09	2.225e-09

Comparisons of VRTs

We investigate the efficiency of single-index IS and SIS estimators with and without QMC for the basket option problem. We also consider the single-index SS estimator which is the single-index SIS estimator without IS part. Table 4.5 shows the variance reduction factors of different methods for $K = 1$ and $K = 1.2$. From the table, we see that the IS and SIS schemes work better when ν is large. Among the VRTs considered, SIS with QMC gives the best results for all three models. Note that QMC with SS performs much better than plain QMC. This numerically confirms that indirect sampling reduces the effective dimension of the problem and enhances the effectiveness of QMC as argued in Section 4.4.

Figure 4.3: Histogram of the 1,000 realized squared standard error

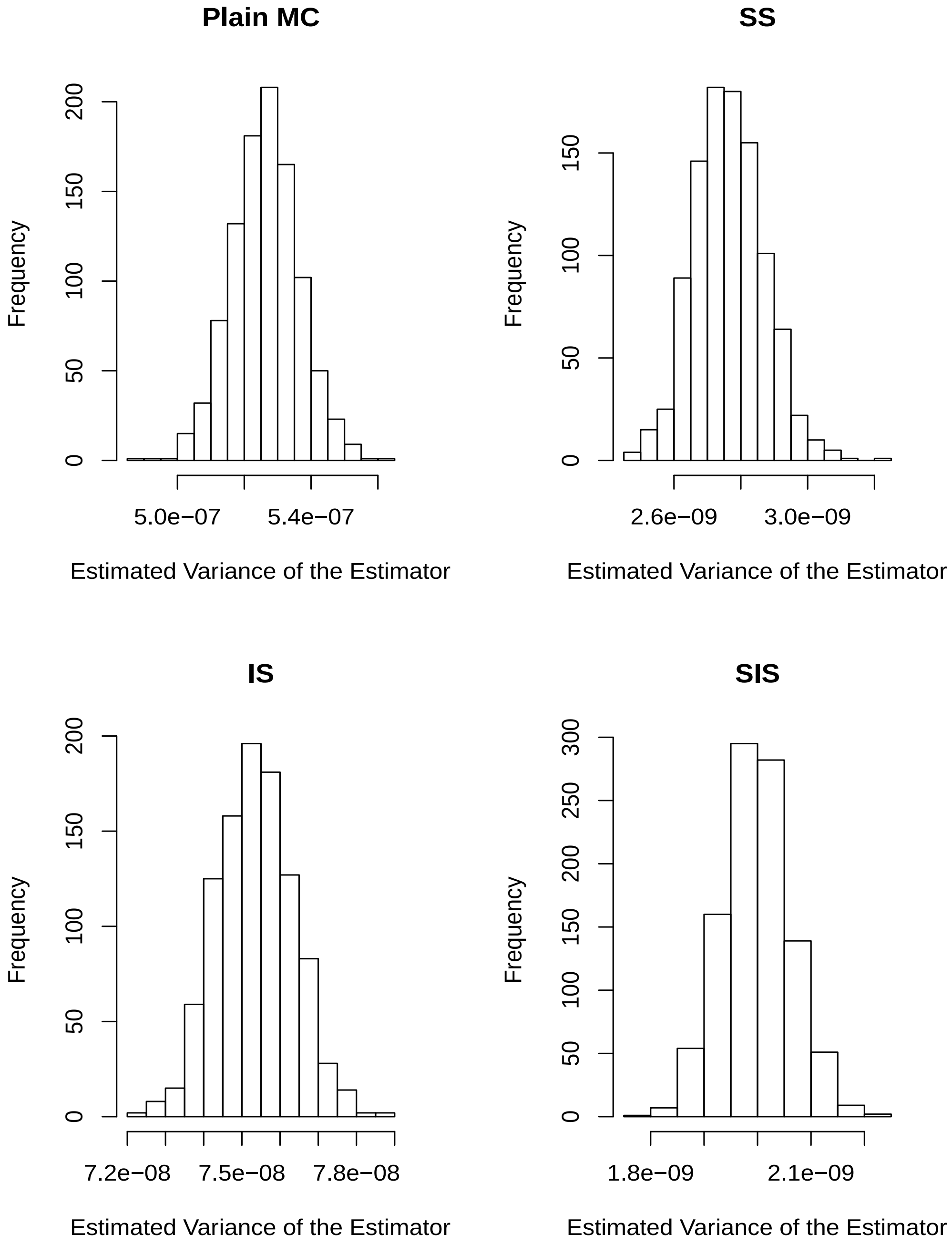


Table 4.5: Variance reduction factors of different VRTs, $n = 2^{14}$, 30 replications

	K=1			K=1.2		
	normal	t_{10}	t_4	normal	t_{10}	t_4
SS	2.26E+02	1.79E+02	5.11E+01	8.64E+01	2.10E+02	2.73E+01
IS	3.38E+01	1.42E+01	8.27E+00	6.92E+01	2.04E+02	1.46E+02
SIS	3.03E+02	1.54E+02	5.31E+01	1.16E+03	2.73E+03	9.69E+01
QMC Plain	7.87E+02	3.77E+02	3.02E+02	7.11E+01	4.52E+01	2.98E+01
QMC SS	2.61E+03	9.19E+02	5.81E+02	4.43E+02	3.66E+02	1.17E+02
QMC IS	3.19E+03	5.67E+02	1.36E+02	2.36E+02	8.97E+02	6.19E+02
QMC SIS	2.15E+04	1.56E+03	2.61E+02	1.21E+04	3.30E+04	9.69E+02

4.7.3 Tail probabilities of a Gaussian Copula Credit Portfolio

In this section, we study the efficiency of the proposed methods for a credit portfolio problem based on a Gaussian copula studied by Glasserman and Li [41], where the goal is to estimate the probability of large losses. We compare single-index IS and SIS to the IS technique of Glasserman and Li, to which we refer as the G&L IS.

Problem Formulation

As in [41], we introduce the following notation:

$$\begin{aligned}
h &= \text{number of obligors to which portfolio is exposed;} \\
Y_k &= \text{default indicator for } k\text{th obligor} \\
&= 1 \text{ if } k\text{th obligor defaults, 0 otherwise;} \\
p_k &= \text{marginal probability that } k\text{th obligor defaults;} \\
c_k &= \text{loss resulting from default of } k\text{th obligor;} \\
L &= c_1 Y_1 + \cdots + c_h Y_h = \text{total loss from defaults.}
\end{aligned}$$

The goal is to estimate $\mathbb{P}(L > l)$ for some large $l \in \mathbb{R}$. Under a Gaussian copula model, the dependence between the default indicators are modelled through multivariate normally distributed latent variables $\boldsymbol{X} = (X_1, \ldots, X_h)$ as

$$Y_k = 1_{\{X_k > x_k\}}, \quad k = 1, \ldots h,$$

where x_k is chosen such that $\mathbb{P}(Y_k = 1) = \mathbb{P}(X_k > x_k) = p_k$. Without loss of generality, assume that each X_k marginally follows a standard normal distribution. Then $x_k = \Phi^{-1}(1-p_k)$. As in [41], assume that each X_k has the following factor structure

$$X_k = a_{k1}Z_1 + \cdots + a_{kd}Z_d + b_k\epsilon_k,$$

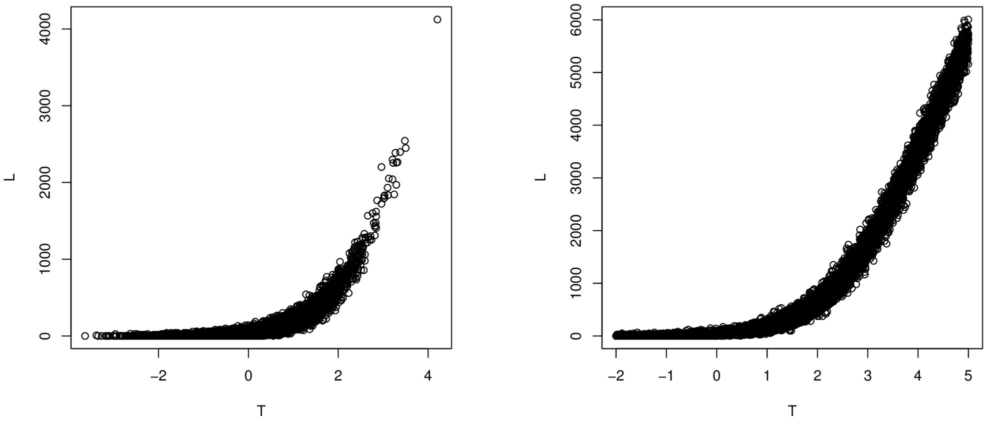
in which

- $Z_1, \ldots, Z_d$ are independent systematic risk factors, each following a $N(0, 1)$ distribution;
- ϵ_k is an idiosyncratic risk associated with kth obligor, independent from $Z_1, \ldots, Z_d$, also $N(0, 1)$ distributed;
- $a_{k1}, \ldots, a_{kd}$ are the factor loadings for the kth obligor, $a_{k1}^2 + \cdots + a_{kd}^2 \leq 1$;
- $b_k = \sqrt{1 - (a_{k1}^2 + \cdots + a_{kd}^2)}$ so that X_k is $N(0, 1)$.

As in [41], we consider a portfolio with $h = 1,000$ obligors in a 10-factor model (i.e. $d = 10$). The marginal default probabilities are $p_k = 0.01 \cdot (1 + \sin(16\pi k/h))$, $k = 1, \ldots, h$ and exposures are $c_k = (\lceil 5k/h \rceil)^2$, $k = 1, \ldots, h$. The marginal default probabilities vary between 0% and 2% and the possible exposures are 1, 4, 9, 16 and 25, with 200 obligors at each level. The factor loadings a_{kj}'s are independently generated from a $U(0, 1/\sqrt{d})$.

Letting $\boldsymbol{Z} = (Z_1, \ldots, Z_d)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_h)'$, we write $L = L(\boldsymbol{Z}, \boldsymbol{\epsilon})$. We investigate whether or not L has a single-index structure. Let $T = \boldsymbol{\beta}'\boldsymbol{Z}$ where $\boldsymbol{\beta} \in \mathbb{R}^d$ such that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$, so $T \sim N(0, 1)$. We estimate $\boldsymbol{\beta}$ using the average derivative method of Stoker [114]. The estimated $\boldsymbol{\beta}$ has almost equal entries close to $\sqrt{1/d}$. This makes intuitive sense as each component of $\boldsymbol{Z}$ is likely to be equally important because the factor loadings are generated randomly. Figure 4.4 shows the scatter plot of (T, L). The left figure is where T is sampled from the original distribution $N(0, 1)$ while the right figure is where T is generated from $U(-2, 5)$ so that more observations from the right tail are sampled. The figure reveals the single-index model fits L well even in the extreme tail, implying single-index IS based on this choice of T will give substantial variance reduction.

Figure 4.4: Plot of Transformed variable (T) vs Portfolio Loss (L) based on 10,000 observations.



(a) T is generated from the original distribution $N(0, 1)$.

(b) T is generated from $U(-2, 5)$ so that more observations from the tail are sampled.

Debiasing SIS estimators

The SIS estimators based on the optimal calibration (4.12) gives a biased estimator for this problem. We describe the procedure to debias the estimators in this section. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_d)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_h)'$. Fix $l > 0$ and let $p_l = \mathbb{P}(L > l)$. Since this is a problem of probability estimation, $\Psi(\boldsymbol{Z}, \boldsymbol{\epsilon}) = \mathbb{1}(L(\boldsymbol{Z}, \boldsymbol{\epsilon}) > l)$. The conditional probability function is $p_l(t) = \mathrm{E}[\Psi(\boldsymbol{Z}, \boldsymbol{\epsilon}) \,|\, T = t] = \mathbb{P}(L > l \,|\, T = t)$ and the conditional variance function is $v^2(t) = \mathrm{Var}(\Psi(\boldsymbol{Z}, \boldsymbol{\epsilon}) \,|\, T = t) = p_l(t)(1 - p_l(t))$. Since $g_T^{\mathrm{opt}}(t) \propto v(t) f_T(t)$, the optimal calibration gives zero density over the region where $v(t) = 0$, or equivalently the region where $p_l(t) = 0$ or $p_l(t) = 1$. In practice, $p_l(t)$ is unknown so we replace it with an estimate $\hat{p}_l(t)$. Similarly, we replace $v^2(t)$ with $\hat{v}^2(t) = \hat{p}_l(t)(1 - \hat{p}_l(t))$. The approximate optimal proposal density of T is then

$$\hat{g}_T^{\mathrm{opt}}(t) \propto \hat{p}_l(t)(1 - \hat{p}_l(t)) f_T(t) \tag{4.28}$$

and this is the proposal density of T that we draw samples from to construct the IS estimator. Noting that $\hat{g}_T^{\mathrm{opt}}(t) > 0$ only if $0 < \hat{p}_l(t) < 1$, we have

$$\mathrm{E}[\hat{\mu}_{\mathrm{SIS},n}] = \int_{\hat{g}_T^{\mathrm{opt}}(t) > 0} p_l(t) \frac{f_T(t)}{g_T(t)} g_T(t) dt = \int_{0 < \hat{p}_l(t) < 1} p_l(t) f_T(t) dt \tag{4.29}$$

$$= \int_{0 < \hat{p}_l(t) < 1} p_l(t) f_T(t) dt = p_l - \mathbb{P}_g(\hat{p}_l(T) = 1), \tag{4.30}$$

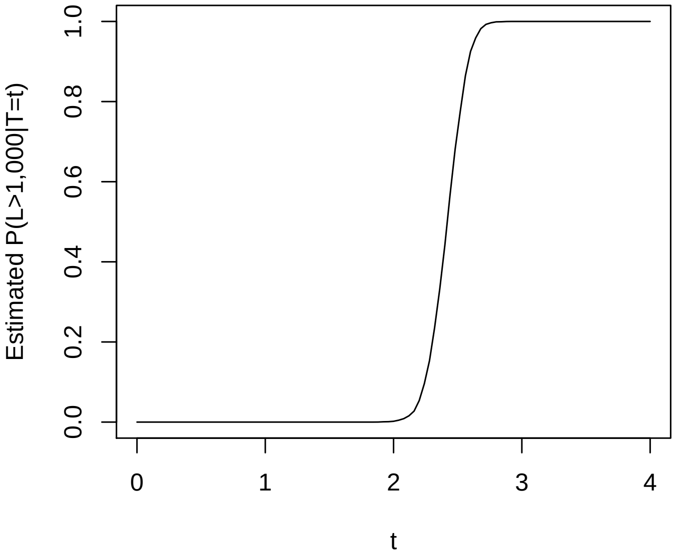
the debiased SIS estimator, $\hat{\mu}_{\mathrm{SIS},n}^{\mathrm{db}}$, is obtained as $\hat{\mu}_{\mathrm{SIS},n}^{\mathrm{db}} = \hat{\mu}_{\mathrm{SIS},n} + \mathbb{P}_g(\hat{p}_l(T) = 1)$. Figure 4.5 shows the plot of $\hat{p}_l(t)$ as a function of t for $l = 1,000$. In further numerical studies, we find that $\hat{p}_l(t)$ has a similar shape for other values of l. As the figure illustrates, $\hat{p}_l(t)$ is a monotone function in t, then there exists $t_l \in \mathbb{R}$ such that $\hat{p}_l(t) = 1$ for all $t > t_l$. Then the debiased estimator becomes

$$\hat{\mu}_{\mathrm{SIS},n}^{\mathrm{db}} = \hat{\mu}_{\mathrm{SIS},n} + \mathbb{P}_g(T > t_l) = \hat{\mu}_{\mathrm{SIS},n} + \Phi(-t_l).$$

Observe that since $\hat{p}_l(t)$ is monotone increasing in t for this problem, the CDF corresponding to $\hat{g}_T^{\mathrm{opt}}(t)$ is

$$G_T(t) = \frac{\int_{-\infty}^{t} \hat{v}(s) f_T(s) ds}{\int_{-\infty}^{\infty} \hat{v}(s) f_T(s) ds} = \frac{\int_{-\infty}^{t} \hat{v}(s) f_T(s) ds}{\int_{-\infty}^{t_l} \hat{v}(s) f_T(s) ds},$$

Figure 4.5: Plot of estimated $\hat{p}_l(t)$ for l=1,000



so $t_l = G_T^{-1}(1)$. Therefore, we can write

$$\hat{\mu}^{\text{db}}_{\text{SIS},n} = \hat{\mu}_{\text{SIS},n} + \mathbb{P}_g(T > t_l) = \hat{\mu}_{\text{SIS},n} + \Phi(-G_T^{-1}(1)).$$

In single-index IS, we generally sample T under the proposal distribution using the inversion technique, which assumes that we have numerically constructed G_T^{-1}, for instance using the NINIGL algorithm [55].

Comparison of Variance Reduction Factors

We now compare single-index IS and SIS to G&L IS by computing the variance reduction factors for estimating $\mathbb{P}(L > l)$ for $l \in \{1,000, 1,500, 2,000, 2,500\}$. All the three methods need to optimize the proposal distribution before running the main simulation. In this comparison, we optimize the proposals at each loss level of l and estimate the corresponding loss probability. Table 4.6 shows the estimated probabilities as reference and the variance reduction factors of the three methods over plain MC. The estimated probabilities are based on SIS with $n = 100,000$ and they are listed to show how rare those events are.

The table shows that single-index SIS gives the greatest variance reduction, followed by in order of single-index IS and G&L IS. For all three methods, VRFs increase as the cutoff value l gets larger.

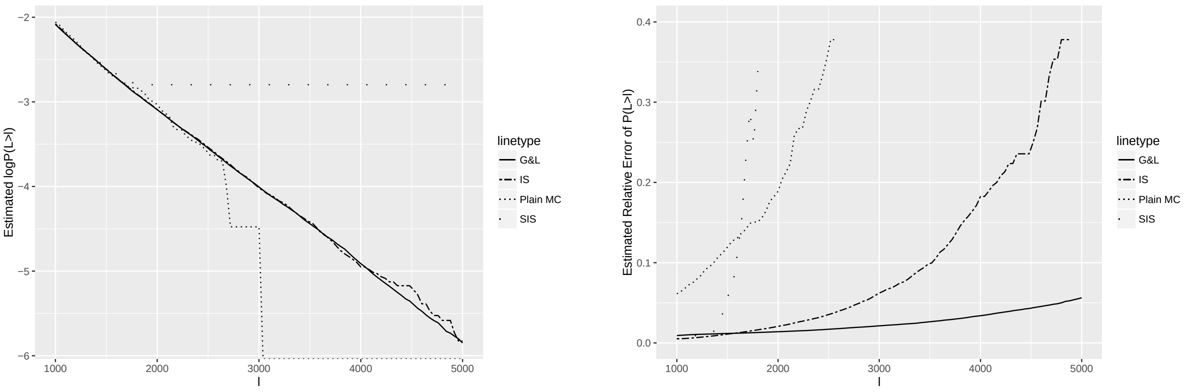
Table 4.6: Variance Reduction Factors based on 30,000 samples

l	1,000	1,500	2,000	2,500
Estimates	8.29E-02	2.43E-02	8.10E-03	2.84E-03
G&L	5.07E+01	1.34E+02	4.38E+02	8.26E+02
IS	1.24E+02	3.97E+02	1.26E+03	2.34E+03
SIS	2.78E+02	9.12E+02	3.00E+03	5.48E+03

For the simulation above, we calibrated the proposal distributions for each value of l. In order to estimate $\mathbb{P}(L > l)$ for $l \in \{1,000, 1,500, 2,000, 2,500\}$ using SIS for instance, we ran the entire simulation four times. Often we want to estimate multiple loss probabilities in one simulation run. To pursue this idea further, next we calibrate the proposal distributions for $l = 1,000$ and estimate $\mathbb{P}(L > l)$ for $l \in [1,000, 5,000]$ in a single simulation. Figure 4.6 shows the estimated probabilities in the base 10 log scale and the estimated relative errors (RE) of the estimators based on $n = 30,000$ samples.

Our experiments based on large n find that $\log \mathbb{P}(L > l)$ is linear in l, consistent with the asymptotic result in [40]. This implies that if the estimated probabilities in log scale deviate from a linear trend, the estimates are inaccurate. From the log-probability plot of Figure 4.6, G&L IS seems to produce estimates with relatively small errors for all values of $l \in [1,000, 5,000]$. The SIS estimator is clearly biased for $l > 1,200$, indicating that the optimally calibrated SIS proposal distribution performs poorly if it is used to estimate multiple loss probabilities. The plain MC and single-index IS appear to struggle estimating the loss probabilities for $l > 2,700$ and $l > 3,500$, respectively. The RE plot of Figure 4.6 agrees with these findings. The plot shows that the RE of the SIS estimator quickly escalates with l and the same holds with plain MC and single-index IS estimators to a lesser extent. Only G&L IS gives an estimator with moderately low RE for all values of $l \in [1,000, 5,000]$.

Figure 4.6: Plot of Estimated Log-Probability and Relative Error based on 30,000 observation. The proposal distributions are calibrated to estimate $\mathbb{P}(L > 1,000)$.



(a) Estimated $\log \mathbb{P}(L > l)$

(b) Estimated Relative Error

Overall, it appears that single-index IS calibrated to estimate p_l for a specific l estimates this probability very well but struggles to estimate probabilities based on different l. This makes sense as the optimal calibrations are constructed to minimize the variance of a given problem and do not take anything else into account. In Chapter 5, we develop calibration methods which balance the performance of multiple estimations.

4.7.4 Tail probabilities of a t-Copula Credit Portfolio

In this section, we apply single-index IS to a credit portfolio problem under a t-copula model. This model can be viewed as being equivalent to the Gaussian copula model studied in Section 4.7.3, but with a multiplicative shock variable added to it. This t-copula model is a special case of the models with extremal dependence studied by Bassamboo et al. [12]. Unlike the Gaussian copula models, the t-copula ones support tail dependence of latent variables, so simultaneous defaults of many obligors are more probable under the t-copula model than its Gaussian copula counterpart.

Problem Formulation

The t-copula model with ν degrees of freedom is the same as the Gaussian copula model except that the latent variables $\boldsymbol{X} = (X_1, \dots, X_d)$ are multivariate-t distributed. That is,

$$X_k = \sqrt{W}(a_{k1}Z_1 + \cdots + a_{kd}Z_d + b_k\epsilon_k), \tag{4.31}$$

where $W \sim \mathrm{IG}(\nu/2, \nu/2)$, $Z_1, \dots, Z_d$, $\epsilon_k \overset{\text{ind.}}{\sim} N(0,1)$. Accordingly, the default threshold for the kth obligor is $x_k = t_\nu^{-1}(1-p_k)$. We use the same set of parameters as in the Gaussian copula model studied in Section 4.7.3. In particular, we consider a portfolio with $h = 1,000$ obligors in a 10-factor model (i.e. $d = 10$). The marginal default probabilities are $p_k = 0.01 \cdot (1 + \sin(16\pi k/h))$, $k = 1, \dots, h$ and exposures are $c_k = (\lceil 5k/h \rceil)^2$, $k = 1, \dots, h$. The marginal default probabilities vary between 0% and 2% and the possible exposures are 1, 4, 9, 16 and 25, with 200 obligors at each level. The factor loadings a_{kj}'s are independently generated from a $U(0, 1/\sqrt{d})$. For the degree of freedom parameter of the t-copula model, we consider $\nu = 12$ and $\nu = 4$.

Let $\boldsymbol{Z} = (Z_1, \dots, Z_d)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_h)$. Chan and Kroese [20] propose a very effective IS technique based on conditional Monte Carlo (CMC). Letting $S_l(\boldsymbol{Z}, \boldsymbol{\epsilon}) = P(L > l \mid \boldsymbol{Z}, \boldsymbol{\epsilon})$ their main idea is to estimate $P(L > l)$ by the sample mean of $S_l(\boldsymbol{Z}, \boldsymbol{\epsilon})$ which can be computed analytically once $(\boldsymbol{Z}, \boldsymbol{\epsilon})$ is sampled. This technique is effective as it analytically integrates out W, the variable which accounts for a large portion of the variance of L. Chan and Kroese further combine this CMC idea and IS on $\boldsymbol{Z}$ and $\boldsymbol{\epsilon}$ to make the event $\{L > l\}$ more frequent based on the cross-entropy method (see [26, 102, 103] and references therein for the details of the cross-entropy method). We refer to Chan and Kroese's method as C&K CMC+IS. The numerical study in [20] demonstrates that C&K CMC+IS achieves substantial variance reduction. We investigate whether or not L and S_l have single-index structures.

Fit of Single-Index models with and without conditional Monte Carlo

We first consider a single-index model without the CMC idea. That is, we investigate whether $L = L(\boldsymbol{Z}, \boldsymbol{\epsilon})$ has a single-index structure. Let Z_W be the quantile-quantile transformed variable of W to a standard normal variable, that is, $Z_W = \Phi^{-1}(F_W(W))$. Let

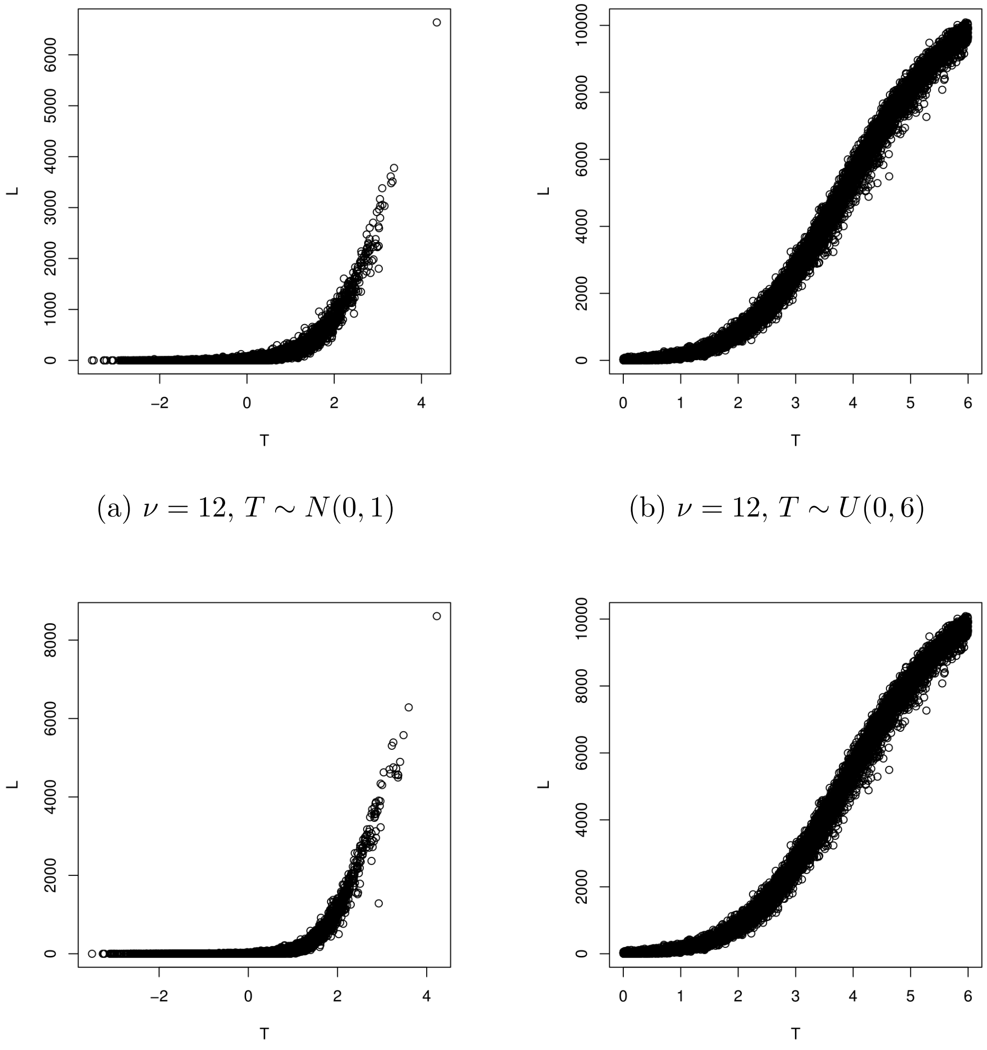
$T_1(W, \boldsymbol{Z}, \boldsymbol{\epsilon}) = \beta_W Z_W + \boldsymbol{\beta}'_L Z$ such that $\beta_W^2 + \boldsymbol{\beta}'_L \boldsymbol{\beta}_L = 1$. Under this constraint, $T_1 \sim N(0,1)$. We estimate the coefficients $(\beta_W, \boldsymbol{\beta}_L)$ that maximize the fit of the single-index model by using the average derivative method of Stoker [114]. Figure 4.7 shows scatter plots of (T_1, L) for $\nu = 12$ and $\nu = 4$ where $T_1 \sim N(0,1)$ o6 $T_1 \sim U(0,6)$. The plots for $T_1 \sim U(0,6)$ show more observations in the right-tail of T_1. The figures show that there is a strong association between T_1 and L but the dependence is stronger when $\nu = 12$ than when $\nu = 4$. When $\nu = 4$, there is a significant variation of L that cannot be captured by the single-index model based on T_1 in the right-tail. This observation holds more generally; the smaller ν is, the worse the fit of the single-index model becomes in the right-tail. Hence, when ν is small, we expect that single-index IS based on T_1 becomes less effective when estimating p_l for large l compared to when ν is large.

We now consider a single-index model with the CMC idea, that is, we examine whether or not $S_l = S_l(\boldsymbol{Z}, \boldsymbol{\epsilon})$ has a single-index structure. Let $T_2 = \boldsymbol{\beta}'_S \boldsymbol{Z}$ with $\boldsymbol{\beta}_S$ such that $\boldsymbol{\beta}'_S \boldsymbol{\beta}_S = 1$. The coefficients $\boldsymbol{\beta}_S$ that maximize the fit of the single-index model are estimated by using the average derivative method [114]. Figure 4.8 shows the scatter plot of (T_2, S_l) for $l = 1,000$ and $l = 2,000$ based on 10,000 observations for degrees of freedom parameters $\nu = 12$ and $\nu = 4$. In order to obtain enough samples in the right-tail, T_2 is drawn from $U[-2, 6]$. From the figure, we see that the fit of T_2 is excellent even in the extreme right-tail for both the $\nu = 4$ and $\nu = 12$ cases. This means that the fit of T_2 is less sensitive to the degree of freedom parameter than T_1 is. This makes sense as the variance due to W, the only variable that depends on ν, is integrated out by the CMC step.

Comparison of Variance Reduction Factors

We compare single-index IS with and without CMC to C&K CMC+IS by computing the variance reduction factors for estimating the probabilities of the form $\mathbb{P}(L > l)$ for $l \in \{1,000, 1,500, 2,000, 2,500\}$. Table 4.7 shows the estimated probabilities as reference and the variance reduction factors of the three IS methods: C&K CMC+IS, single-index IS based on T_2 with CMC, and single-index IS based on T_1 without CMC. For this problem, we calibrate the proposal densities of the three methods for each value of $l \in \{1000, 1500, 2000, 2500\}$. From the table, we see that the IS methods that use CMC

Figure 4.7: Scatter plots of T_1 vs L. The two plots in the first row are for $\nu = 12$ and the second row are for $\nu = 4$. The plots in the first column are for $T_1 \sim N(0,1)$ and the second column for $T_1 \sim U(0,6)$



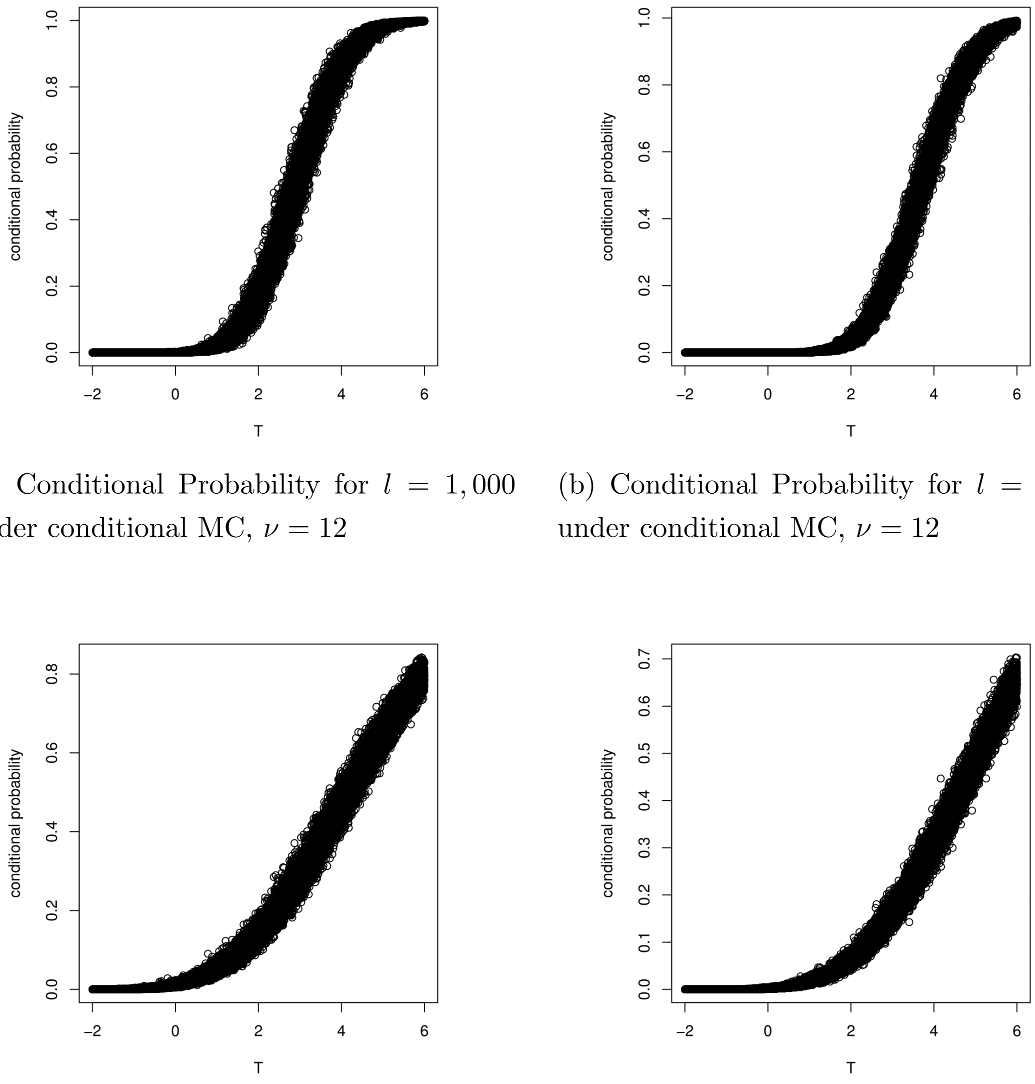
(a) $\nu = 12$, $T \sim N(0,1)$

(b) $\nu = 12$, $T \sim U(0,6)$

(c) $\nu = 4$, $T \sim N(0,1)$

(d) $\nu = 4$, $T \sim U(0,6)$

Figure 4.8: Plot of Transformed variable (T_2) vs Conditional probability (S_l) based on 10,000 observations



(a) Conditional Probability for $l = 1,000$ under conditional MC, $\nu = 12$

(b) Conditional Probability for $l = 2,000$ under conditional MC, $\nu = 12$

(c) Conditional Probability for $l = 1,000$ under conditional MC, $\nu = 4$

(d) Conditional Probability for $l = 2,000$ under conditional MC, $\nu = 4$

(C&K CMC + IS and CMC + IS on T_2) give greater variance reduction than the one that does not use CMC (IS on T_1). Among the two that use CMC, the one based on single-index IS works better than the Chan and Kroese's IS method. Note that single-index IS with CMC gives about 10 times greater variance reduction than the one without CMC. This is consistent with the observation based on Figure 4.7 and Figure 4.8 that S_l has a stronger single-index structure than L does.

Table 4.7: Variance Reduction Factors based on 30,000 samples. "C&K CMC + IS" denotes Chan and Kroese's CMC and IS technique. "CMC+IS on T_2" denotes single-index IS with T_2 combined with CMC. "IS on T_1" denotes single-index IS on T_1 without CMC.

	l	1,000	1,500	2,000	2,500
$\nu = 12$	Estimates	1.68E-02	7.17E-03	3.43E-03	1.74E-03
	C&K CMC + IS	2.18E+02	4.52E+02	8.19E+02	1.56E+03
	CMC + IS on T_2	6.99E+02	1.38E+03	2.57E+03	5.56E+03
	IS on T_1	1.30E+02	2.75E+02	5.55E+02	1.07E+03
$\nu = 4$	Estimates	2.58E-02	1.45E-02	8.81E-03	5.55E-03
	C&K CMC + IS	4.12E+02	6.23E+02	8.47E+02	1.24E+03
	CMC + IS on T_2	9.92E+02	1.89E+03	2.95E+03	4.80E+03
	IS on T_1	1.17E+02	1.74E+02	2.25E+02	3.40E+02

As we did for the Gaussian copula credit portfolio problem in Section 4.7.3, we estimate multiple loss probabilities in one simulation run. The IS schemes based on the CMC idea is not suited to multiple estimation because the CMC becomes very computationally expensive as $\mathbb{P}(L > l \,|\, \boldsymbol{Z}, \boldsymbol{\epsilon})$ must be computed for all samples of $(\boldsymbol{Z}, \boldsymbol{\epsilon})$ for each value of l. Thus, we estimate $l \in [1,500, 7,000]$ using plain MC and single-index IS with T_1. We do not consider single-index SIS as it has already shown to give very unreliable estimates in Section 4.7.3. Figure 4.9 shows the plot of estimated $\log \mathbb{P}(L > l)$ and RE for $\nu = 12$ and $\nu = 4$ cases for plain MC, single-index IS with T_1, and UIS. UIS is a IS technique that we develop specifically for multiple estimation in Chapter 5. As the RE plots of Figure 4.9

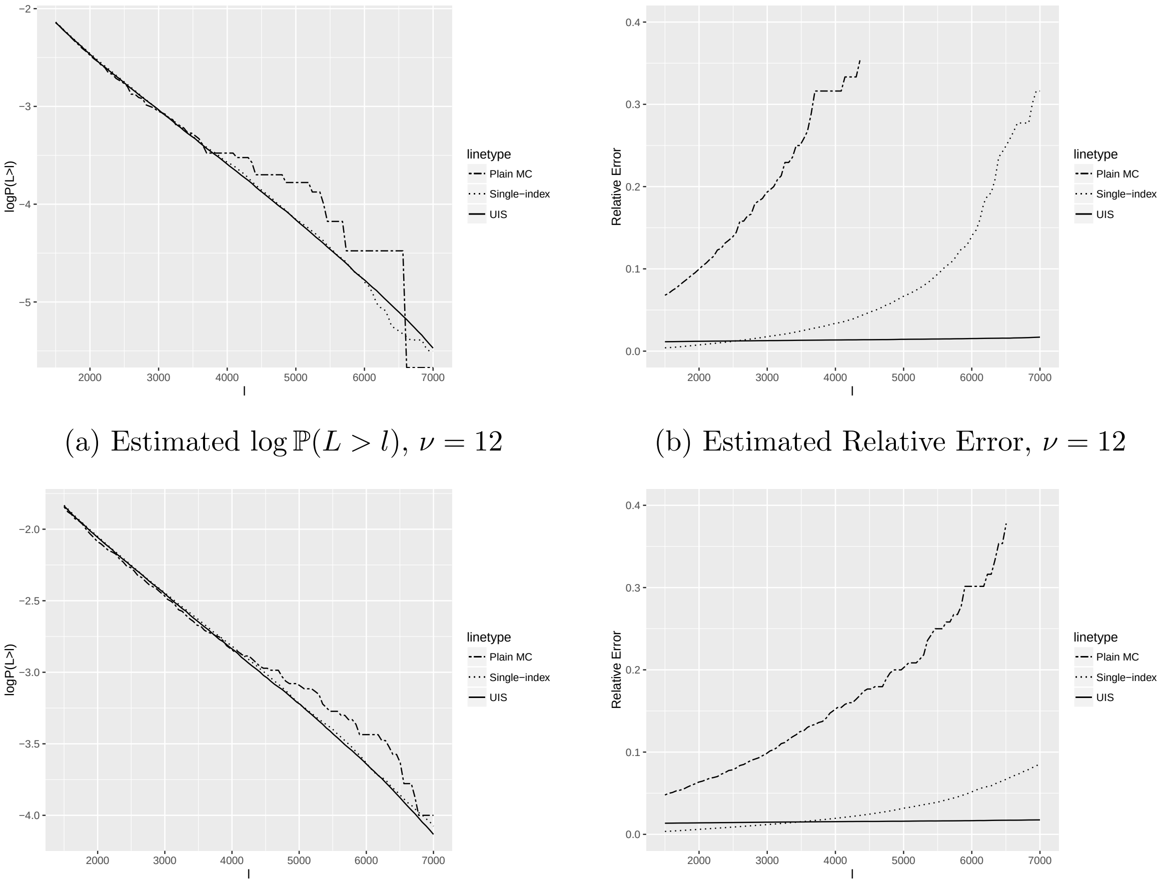
shows, UIS gives estimates of p_l with very low RE for all $l \in [1,500,7,000]$. Thus, we can use the log-probability curve given UIS as the reference for the true log-probabilities. If the estimated probabilities in log scale significantly deviate from the ones based on UIS, the estimates are inaccurate. From the log-probability plot of Figure 4.9, plain MC and single-index IS give unreliable estimates for $l > 4,000$ and $l > 6,000$ when $\nu = 12$ and $l > 4,000$ and $l > 6,500$ when $\nu = 4$, respectively. It appears that plain MC and single-index IS, the estimation quality of p_l deteriorates faster in l when $\nu = 12$ than when $\nu = 4$. From the RE plots of Figure 4.9, the RE of plain MC and single-index IS grows quickly in l, as it was the case for the Gaussian copula model, though this is to a lesser extent when $\nu = 4$ than when $\nu = 12$.

Overall, we have the same conclusion as we had for the Gaussian copula model that single-index IS calibrated to estimate p_l for a specific l estimates this probability very well but struggles to estimate probabilities based on different l. We development calibration methods that handle multiple estimation well in Chapter 5.

4.7.5 Skew-t copula equity portfolio

In this section, we numerically investigate the efficiency of single-index IS and SIS for estimating tail quantities of an equity portfolio under the skew-t copula models. To obtain model parameters, we fit the said copula to daily negative log-returns of the stock of 10 large firms in the financial sector from 2010-01-01 to 2016-04-01 (1571 data points). We fit GARCH(1,1)-models with t-innovations to each return series to filter out the volatility clustering effect using the R package "rugarch" [34]. The fitted standardized residuals do not exactly follow a t-distribution, so we fit a semi-parametric distribution to the residuals using the R package "spd"[35]. We then fit normal, t, and skew-t copulas to the fitted standard residuals. The skew-t copulas are fitted using the procedure in Appendix A. We first examine whether skew t-copulas provide a better fit than t and Gaussian copulas. For the skew t-copula, we consider the one with exchangeable skewness parameter assumption (all skewness parameters are the same) which we call a skew t_{ex} model and the one with different skew parameters which we simply call a skew-t model.

Figure 4.9: Plot of Estimated Probability and Relative Error based on 30,000 observations



(a) Estimated $\log \mathbb{P}(L > l)$, $\nu = 12$

(b) Estimated Relative Error, $\nu = 12$

(c) Estimated $\log \mathbb{P}(L > l)$, $\nu = 4$

(d) Estimated Relative Error, $\nu = 4$

Model Fit

Table 4.8: Fit of the Copulas

	# Param.	Log-like.	AIC	BIC
Gaussian	45	5141.0	-10192.0	-9950.824
symmetric t	46	5478.1	-10864.2	-10617.664
skew t_{ex}	47	5505.7	-10917.4	-10665.505
skew t	56	5517.5	-10923.0	-10622.870

Table 4.9: Estimated Parameters

	$\rho_{1,2}$	ν	γ_1	γ_2	γ_3
Gaussian	0.57				
symmetric t	0.61	10.0			
skew t_{ex}	0.60	10.7	0.275	0.275	0.275
skew t	0.61	10.7	0.387	0.260	0.359

All four copulas considered have a correlation matrix as parameters. The fitted matrix for the four models turn out to be almost identical. Table 4.8 lists the log-likelihood, AIC and BIC of the copula models. Table 4.9 lists the selected parameters of the fitted copulas: $\rho_{1,2}$, the $(1,2)$th element of the correlation matrix, ν, the degrees of freedom parameter, and the first three elements of the skewness parameters, γ. Table 4.10 shows the estimated correlation matrix of the skew-t copula, and then the skewness parameters, $\boldsymbol{\gamma}$, on the last row. Table 4.8 shows that the skew-t copula provides a better fit than the symmetric-t copula, which in turn fits better than the Gaussian copula, in terms of both AIC and BIC. Note that the estimated skewness parameters, $\boldsymbol{\gamma}$, are all positive, suggesting the greater dependence for losses than returns. This aligns with the observation made by Ang and Chen that U.S. stocks have a stronger correlation for downside moves than upside [6]. The skew t_{ex}-copula gives the best fit in terms of BIC while the skew t-copula does the best in terms of AIC. So, there is no clear indication of weather the exchangeable skewness parameters assumption is valid. We use the skew t-copula with varying skewness parameters in our numerical experiments.

Table 4.10: Estimated Correlation Matrix and skewness parameters of skew-t copula

1.00	0.61	0.57	0.64	0.42	0.62	0.60	0.59	0.67	0.61
0.61	1.00	0.58	0.61	0.46	0.64	0.62	0.67	0.63	0.71
0.57	0.58	1.00	0.61	0.47	0.65	0.62	0.61	0.63	0.58
0.64	0.61	0.61	1.00	0.42	0.72	0.62	0.62	0.82	0.59
0.42	0.46	0.47	0.42	1.00	0.47	0.51	0.50	0.46	0.48
0.62	0.64	0.65	0.72	0.47	1.00	0.68	0.66	0.72	0.62
0.60	0.62	0.62	0.62	0.51	0.68	1.00	0.65	0.65	0.61
0.59	0.67	0.61	0.62	0.50	0.66	0.65	1.00	0.64	0.67
0.67	0.63	0.63	0.82	0.46	0.72	0.65	0.64	1.00	0.61
0.61	0.71	0.58	0.59	0.48	0.62	0.61	0.67	0.61	1.00
0.39	0.26	0.36	0.25	0.17	0.29	0.34	0.22	0.23	0.29

Comparison of Variance Reduction Techniques

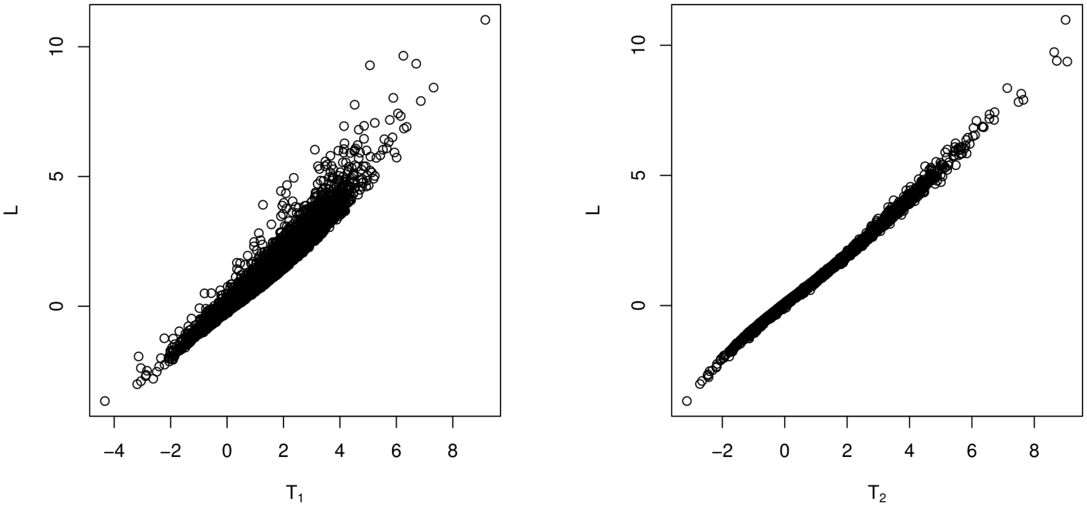
Let L denote the portfolio loss over a one day period,

$$L = 100\left(1 - \sum_{j=1}^{d} v_j \exp\left(a_j - b_j \hat{F}_j^{-1}(U_j)\right)\right),$$

where d is the number of assets, v_j's are the portfolio weights, a_j's are the means of log-returns, b_j's are the fitted conditional standard deviations from the GARCH(1,1) model, $\hat{F}_j$'s are the fitted semi-parametric distributions, and $(U_1, \ldots, U_d)$ follows the fitted copula. In this simulation study, we consider a portfolio with equal weights, that is, $v_j = 1/d$ for $j = 1, \ldots, d$. The transformation functions we use are $T_1(W, Z) = \beta_0 + \sqrt{W}\boldsymbol{\beta}_2'\boldsymbol{Z}$ and $T_2(W, Z) = \beta_0 + \beta_1 W + \sqrt{W}\boldsymbol{\beta}_2'\boldsymbol{Z}$. The difference between the two is that T_2 includes the linear term in W while T_1 does not. The coefficients β_k's are estimated by fitting a linear model. This estimation procedure is valid as the conditional mean $\mathrm{E}[L \mid T_1]$ and $\mathrm{E}[L \mid T_2]$ are linear in T_1 and T_2, respectively, for this problem.

We first examine whether or not the inclusion of the $\beta_1 W$ term improves the fit of the single-index model. Figure 4.10 shows the scatter plot of (T_1, L) and (T_2, L). From the figure we see that the inclusion of the $\beta_1 W$ term significantly improves the fit of the model, which implies that single-index IS and SIS based on T_2 will perform better than the ones based on T_1. However, as discussed in Section 4.6.2, the presence of the $\beta_1 W$ term makes the conditional sampling more computationally expensive. Thus, we compare variance reduction and the computation time relative to plain MC estimators to decide which one is more efficient.

Figure 4.10: Scatter plot of (T_1, L) (left) and (T_1, L) (right) based on 10,000 observations.



We now investigate the efficiency of the IS and SIS schemes. The quantities we are interested in are stop loss $\mathrm{E}[(L-3)^+]$, 99% value-at-risk $\mathrm{VaR}_{0.99}(L)$, and 99% expected shortfall $\mathrm{ES}_{0.99}(L)$. The proposal distribution for IS and SIS are calibrated against $\mathrm{ES}_{0.99}(L)$ as we expect that the proposal that works well for $\mathrm{ES}_{0.99}(L)$ would also work well for $\mathrm{E}[(L-3)^+]$ and $\mathrm{VaR}_{0.99}(L)$.

Table 4.11 shows the estimates of the three objective functions, variance reduction factors of various methods and the ratios of computation time relative to plain MC estimators. The estimates are taken from SIS based on T_2 combined with QMC. From the table, we see that both the IS and SIS schemes improve the estimation precision and this is more so when combined with QMC. Also, the IS and SIS schemes perform better with T_2 than

Table 4.11: Estimates and Variance Reduction Factors: 10-dimensional, $n = 2^{15}$.

		$\mathrm{E}[(L-3)]^+$	$\mathrm{VaR}_{0.99}(L)$	$\mathrm{ES}_{0.99}(L)$	Time
T_1 + MC	IS	4.08E+01	1.45E+01	3.18E+01	1.20
	SIS	3.65E+01	1.65E+01	3.03E+01	1.14
T_1 + QMC	Plain	3.94E+01	9.26E+00	5.35E+01	0.93
	IS	1.18E+02	1.53E+01	7.04E+01	1.24
	SIS	5.94E+01	2.54E+01	4.40E+01	1.20
T_2 + MC	IS	2.36E+02	9.26E+01	2.28E+02	2.15
	SIS	7.63E+03	4.49E+02	7.23E+03	2.16
T_2 + QMC	Plain	3.94E+01	9.26E+00	5.35E+01	0.93
	IS	2.92E+04	1.09E+03	2.66E+04	2.13
	SIS	2.56E+04	6.10E+02	2.09E+04	2.19
	Estimates	5.25E-03	2.68	3.46	

with T_1 as expected from the superior fit of the model with T_2 as shown in Figure 4.10. The computation time does increase when T_2 is used but the reduced variance outweighs the additional computation time.

We repeat the same experiments except the portfolio now consists of 20 stocks (see Table 3.2 for stocks symbols) considered in Section 3.6. Figure 4.11 shows the scatter plot of (T_1, L) and (T_2, L). The plot shows that the fit of the model based on T_1 and T_2 are not much different from the 10-dimensional case. Table 4.12 shows the estimates, variance reduction factors and ratio of computation time relative to plain MC estimators. From the table, we find that the analysis for the 10-dimensional case mostly holds for this 20-dimensional case, except that actual variance reductions are slightly worse this time.

Figure 4.11: Plot of T_1 vs Portfolio Loss (left) and T_2 vs Portfolio Loss (right) based on 10,000 observations, 20 dimensional.

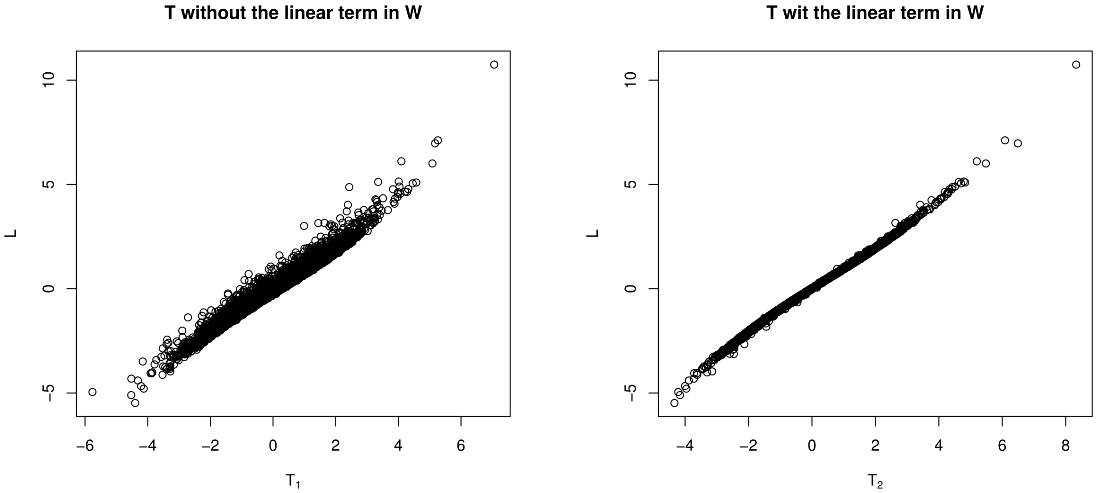


Table 4.12: Estimates and Variance Reduction Factors: 20-dimensional, $n = 2^{15}$.

		$\mathrm{E}[(L-3)]^+$	$\mathrm{VaR}_{0.99}(L)$	$\mathrm{ES}_{0.99}(L)$	Time
T_1 + MC	IS	3.60E+01	1.65E+01	3.39E+01	1.24
	SIS	1.78E+01	2.55E+01	1.79E+01	1.17
T_1 + QMC	Plain	4.68E+01	1.39E+01	4.81E+01	1.01
	IS	1.24E+02	9.70E+01	1.22E+02	1.20
	SIS	1.17E+03	7.26E+01	1.29E+03	1.17
T_2 + MC	IS	4.12E+02	1.15E+02	4.26E+02	1.78
	SIS	1.16E+04	8.84E+02	1.12E+04	1.75
T_2 + QMC	Plain	4.68E+01	1.39E+01	4.81E+01	1.00
	IS	2.01E+04	7.25E+02	1.85E+04	1.80
	SIS	1.41E+04	1.27E+03	1.50E+04	1.80
	Estimates	7.39E-03	2.91	3.73	

Chapter 5

Extreme Value and Uniform Importance Sampling for Multiple Portfolio Loss Probabilities

5.1 Introduction

In Chapter 4, we developed IS and SIS schemes for single-index models and proposed optimal calibration techniques for proposal densities (4.8) and (4.12) that aim to minimize the variance of the resulting IS estimator. The numerical study of credit portfolio problems under a normal and t-copula model in Section 4.7.3 and Section 4.7.4 reveals that such optimally calibrated densities could perform poorly if they are used to estimate multiple loss probabilities in one simulation run. More specifically, when estimating probabilities of the form $\mathbb{P}(L > l)$ for different values of l, the RE of the estimators based on those densities deteriorates relatively quickly as l increases, compared to, for instance, the Glasserman and Li's IS method [41]. In this chapter, we explore the proposal calibrations for single-index IS designed to perform well in estimating multiple loss probabilities. In particular, we propose to use extreme value (EV) distributions and uniform distributions as the parametric families of the transformed variable of the single-index model. We refer to such IS schemes as single-index extreme value importance sampling (EVIS) and uniform importance sampling (UIS), respectively.

The rest of this chapter is organized as follows. In Section 5.2, we give a brief review of single-index models and the IS technique developed for such models in Chapter 4. In Section 5.3, we analyze why the optimally calibrated densities fail for multiple estimation and show that this is partially because those densities have right-tails that decay at the same rate as the originally densities. To make sure that the proposal densities are more heavy-tailed than the original ones, we propose in Section 5.4 to use extreme value (EV) distributions as proposal distributions, following the EVIS idea of McLeish [85] and McLeish and Men [86]. In Section 5.5, we propose another parametric family of IS distributions by observing that only samples from a subset of the domain of the transformed variable is relevant for multiple estimation. More specifically, we propose to use uniform distribution over a truncated domain of the transformed variable. In Section 5.6, we apply single-index EVIS and UIS to the credit portfolio problem under a normal and a t-copula model and numerically evaluate the effectiveness of our proposed IS techniques.

5.2 Single-index Models and Importance Sampling

In this section, we provide a brief review of single-index models and the IS technique for such models developed in Chapter 4 in the context of portfolio loss probability estimation.

Let $\boldsymbol{X}$ be a d-dimensional random vector with a pdf denoted by $f_{\boldsymbol{X}}(\boldsymbol{X})$ and $L : \mathbb{R}^d \to \mathbb{R}$ be a function such that $L = L(\boldsymbol{X})$ represents a portfolio loss. The quantity of interest is $p_l = \mathbb{P}(L(\boldsymbol{X}) > l)$ for a large $l \in \mathbb{R}$ such that p_l is small. Suppose that L has a single-index structure, that is, there exists some unknown parametric transformation function $T : \mathbb{R}^d \to \mathbb{R}$ such that $L(\boldsymbol{X})$ depends on $\boldsymbol{X}$ mainly through $T = T(\boldsymbol{X})$. If we write

$$L(\boldsymbol{X}) = m(T) + \epsilon_T, \quad \epsilon_T \,|\, T \sim (0, v^2(T)), \tag{5.1}$$

where $m(t) = \mathrm{E}[\Psi(\boldsymbol{X}) \,|\, T = t]$, $v^2(t) = \mathrm{Var}(\Psi(\boldsymbol{X}) \,|\, T = t)$, and ϵ_T is a random error term, we say that $L(\boldsymbol{X})$ has a single-index structure if $R^2 = \mathrm{Var}(m(T))/\mathrm{Var}(L)$ is close to 1, say $R^2 > 0.9$.

Let $F_T(t)$ $f_T(t)$, and $f_{\boldsymbol{X}|T}(\boldsymbol{x}\,|\,t)$ denote the distribution function of T, pdf of $\boldsymbol{T}$, and the conditional pdf of $\boldsymbol{X} \mid T = t$, respectively, under the original distribution. We assume that the support of T is an interval $\Omega_T = (t_{\text{inf}}, t_{\text{sup}})$ with possibly $t_{\text{inf}} = -\infty$ and $t_{\text{sup}} = \infty$, but this assumption can be easily generalized. We further let $g_{\boldsymbol{X}}(\boldsymbol{x})$, $g_T(t)$, $F_T(t)$ denote the pdf of $\boldsymbol{X}$, the pdf of T and the distribution function T under the IS distribution. In Chapter 4, we proposed to use an IS distribution of the form

$$g_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}|T}(\boldsymbol{x}\,|\,t_{\boldsymbol{x}})g_T(t_{\boldsymbol{x}}), \tag{5.2}$$

where $t_{\boldsymbol{x}} = T(\boldsymbol{x})$. One can generate samples from (5.2) by drawing T from $g_T(t)$ first and then generating $\boldsymbol{X}\,|\,T$ under the original distribution. Since, $f_{\boldsymbol{X}|T}(\boldsymbol{x}\,|\,t)$ is fixed, we want to choose $g_T(t)$ that minimizes the variance of the resulting IS estimator. Let $p_l(t) = \mathbb{P}(L > l\,|\,T = t)$. Then from (4.8), the optimal proposal density of T for estimating p_l has the form

$$g_{T,l}^{\text{opt}}(t) = c_l\sqrt{p_l(t)}f_T(t), \quad c_l = \left(\int_{\Omega_T} \sqrt{p_l(t)}f_T(t)dt\right)^{-1}, \tag{5.3}$$

and we call the practice of setting $g_T(l) = g_{T,l}^{\text{opt}}(t)$ or its approximation as "optimal calibration". To contrast with single-index EVIS and UIS that we introduce later, we call the IS scheme with optimal calibration as single-index optimally calibrated IS (OCIS). Suppose $g_{T,l}^{\text{opt}}(t)$ has been constructed for $l = l_c$ where $l_c \in \mathbb{R}$ is some calibration point. Since the calibration criterion is to minimize the variance of estimating p_{l_c}, the resulting density is likely to perform poorly if it is used to estimate p_l for a value l that is not very close to l_c. In the following section, we demonstrate this point under a simple linear normal model.

5.3 Analysis of Why Optimal Calibrations Fail in Multiple Loss Probability Estimations

In this section, we show that the proposal density based on optimal calibration does not perform well with multiple estimations, partly because the resulting IS density has a right-tail that decays as fast as that of the original distribution. Based on this finding, we motivate the use of EV distribution and uniform distribution as a proposal distribution.

Consider a linear normal problem. The conclusions drawn for this simple model generalize to more complex models, at least to the normal and t-copula credit portfolio problems studied in Section 5.6. Suppose that

$$L = \alpha T + \epsilon_T, \tag{5.4}$$

where $T \sim N(0,1)$, $\alpha > 0$ and $\epsilon_T \,|\, T \sim N(0,s^2)$ such that $\alpha^2 + s^2 = 1$. Note that $L \sim N(0,1)$ and $\text{Var}(m(T))/\text{Var}(L) = \alpha^2$, so L has a single-index structure when α^2 is close to 1. The model assumptions give analytical expressions for p_l and $p_l(t)$ as

$$p_l = \mathbb{P}(L > l) = \mathbb{P}(N(0,1) > l) = \bar{\Phi}(l) \quad \text{and} \tag{5.5}$$

$$p_l(t) = \mathbb{P}(L > l | T = t) = \mathbb{P}(N(0,s^2) > l - \alpha_1 t) = \bar{\Phi}\left(\frac{l - \alpha_1 t}{s}\right). \tag{5.6}$$

From (5.3), we can write

$$g_{T,l}^{\text{opt}}(t) = c_l \sqrt{\bar{\Phi}\left(\frac{l-\alpha_1 t}{s}\right)} f_T(t), \quad c_l = \left(\int_{t=-\infty}^{\infty} \sqrt{\bar{\Phi}\left(\frac{l-\alpha_1 t}{s}\right)} f_T(t) dt\right)^{-1}. \tag{5.7}$$

Suppose that the model parameters are $\alpha^2 = 0.95$ and $s^2 = 0.05$ and the goal is to estimate $\mathbb{P}(L > l)$ for $l \geq 3$. Figure 5.1 shows the plot of $g_{T,l}^{\text{opt}}(t)$ for $l \in \{3,4,5\}$. The figure shows that the optimally calibrated densities for $l \in \{3,4,5\}$ mostly give samples of T in the range $[l-1, l+1]$, respectively. Suppose that we want to use $g_{T,3}^{\text{opt}}(t)$ to estimate p_l for $l \geq 3$. Note that this density gives very few samples of T in the range $[4,6]$, which is the most important subset of the domain of T when estimating p_5. So IS based on $g_{T,3}^{\text{opt}}(t)$ will give poor estimates of p_l for $l > 5$. Figure 5.2 shows the plot of $p_l(t)$ for $l = 3$. Since $p_3(t) \approx 1$ for $t > 3.5$, it follows that $g_{T,3}^{\text{opt}}(t) \approx c_3 f_T(t)$ for $t > 3.5$. This implies that the optimally calibrated density decays at the same rate as the the original density in the right-tail. Since $c_3 \approx 392$ for this specific problem, the IS with $g_{T,3}^{\text{opt}}(t)$ gives about 400 times more samples in the important region of T for estimating p_5 than the plain MC does. However, since the original density $f_T(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ approaches 0 at a square of an exponential rate in t, the large coefficient such as $c_3 = 392$ becomes irrelevant when estimating p_l for l large enough with $g_{T,3}^{\text{opt}}(t)$.

Figure 5.1: Plot of optimal IS densities $g_{T,l}^{\text{opt}}(t)$ for $l = 3, 4$ and 5

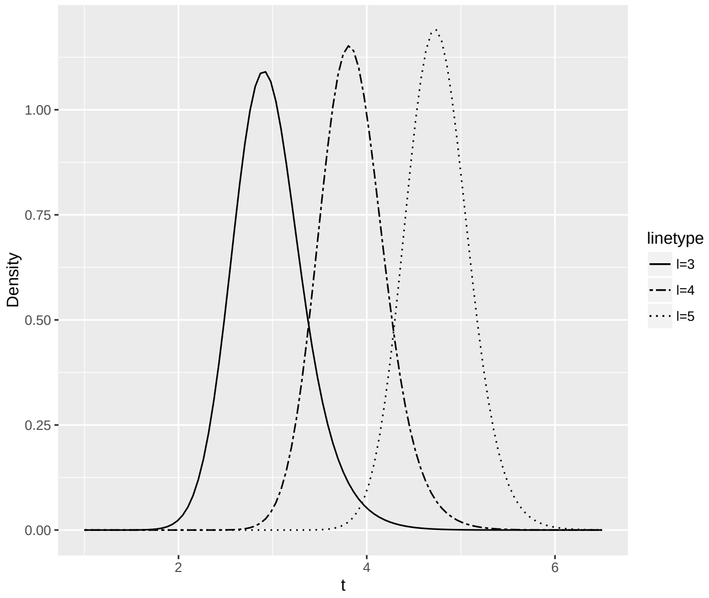
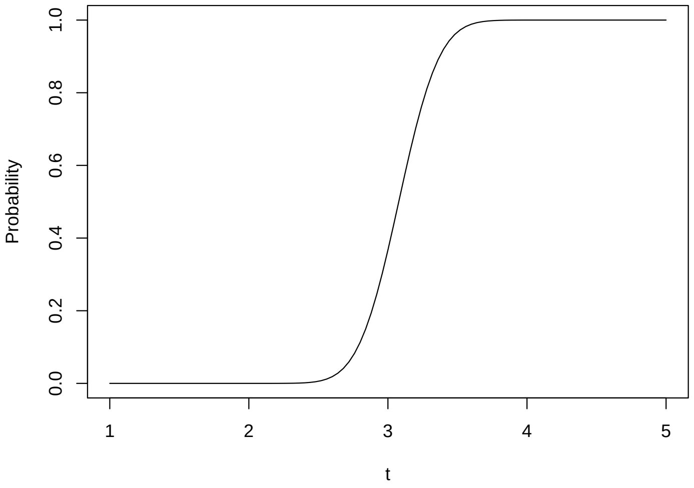


Figure 5.2: Plot of $p_l(t)$ for l=3



If one is more concerned with estimation precision for l close to 3 and less with l further away from 3, we would want a proposal density that gives many samples of T around 3 and moderate number of samples for large T. In order to obtain such samples, we need to make sure that the IS density has a heavier right-tail than the original density does. For this purpose, we propose to use the EV distribution of T as the parametric family of IS distribution, following the ideas of McLeish [85] and McLeish and Men [86]. In Section 5.4, we give the details of EVIS.

If one is equally concerned with the estimation precision of p_l for all l in some finite interval, such as $l \in [3, 5]$, one approach is to construct $g^{\text{opt}}_{T,l}(t)$ for multiple values of l in the interval and construct a mixture of such optimal densities. This approach, however, could be time consuming as we have to construct many $g^{\text{opt}}_{T,l}(t)$ and each requires the approximation of the conditional probability function $p_l(t)$ for many values of l. Instead, we propose to use uniform distribution over a truncated domain of T as a proposal distribution of T and explore this UIS idea in Section 5.5. The numerical studies in Section 5.6 show that UIS gives estimates whose RE remains relatively the same for varying values of l.

5.4 Single-Index Extreme Value Importance Sampling

The main idea of [85] and [86] is that when $d = 1$, the proposal distributions related to the EV distribution of the input variable give estimators with bounded RE. The heavy tailed nature of the EV distributions ensures that the estimation quality does not deteriorate quickly in l in a multiple estimation setting. As we are interested in problems with $d > 2$, single-index EVIS uses an EV distribution as the proposal density of T. Consider a proposal density of the form

$$g_T(t;\theta) = c_\theta e^{-\theta \bar{F}_T(t)} f_T(t)dt, \quad c_\theta = \frac{\theta}{1 - e^{-\theta}}, \tag{5.8}$$

where $\bar{F}_T(t) = 1 - F_T(t)$ and $\theta > 0$ is a parameter to be calibrated. Note that $c_\theta \approx \theta$ when θ is large and the typical value of θ is large enough to make this approximation very accurate. Since $e^{-\theta \bar{F}_T(t)}$ is increasing in t for $\theta > 0$, $g_T(t;\theta)$ has a heavier right-tail than $f_T(t)$ and $g^{\text{opt}}_{T,l}(t)$ do. Given a calibration point l_c, pick a θ that minimizes the variance of the IS estimator for p_{l_c}, that is, find

$$\theta^* = \operatorname*{argmin}_{\theta>0} \frac{1}{c_\theta} \int_{t_{\inf}}^{t_{\sup}} p_{l_c}(t) e^{\theta \bar{F}_T(t)} f_T(t) dt - p_l^2. \tag{5.9}$$

If $p_l(t)$ is known, the integral for (5.9) can be computed numerically for a given θ, so one can find θ^* by running some one-dimensional optimization algorithm. Since $p_l(t)$ is generally unknown, we replace it with a non-parametric approximation of it and proceed with optimization.

Since the proposal distribution of (5.8) does not belong to known parametric families, directly sampling from the IS density requires numerical algorithms. As in [86], we instead draw samples from an EV distribution that approximates this IS density. The advantage of this approach stems from the fact that EV distributions have analytical quantile functions, so the sampling step becomes trivial. Noting that (5.8) is approximately the distribution of the maximum of θ independent copies of $T \sim f_T(t)$ as in [86] when θ takes an integer value, the samples from $g_T(t;\theta)$ follow an EV distribution if normalized appropriately. The proposal distribution based on EV distribution has the following distribution function

$$G_T^{\mathrm{EV}}(t) = \begin{cases} \exp\left(-\left(1+\xi\frac{t-\mu_{\mathrm{ev}}}{\sigma_{\mathrm{ev}}}\right)^{-\frac{1}{\xi_{\mathrm{ev}}}}\right), & \xi_{\mathrm{ev}} \neq 0, \\ \exp\left(-\exp\left(-\frac{t-\mu_{\mathrm{ev}}}{\sigma_{\mathrm{ev}}}\right)\right), & \xi_{\mathrm{ev}} = 0, \end{cases}$$

where μ_{ev}, σ_{ev}, and ξ_{ev} are the location, scale, and shape parameters, respectively. The parameters μ_{ev} and σ_{ev} are calibrated to approximate $g_T(t;\theta^*)$ and ξ is determined from $f_T(t)$. From Theorem 1.18 and Remark 1.19 of Haan et al. [27], we can find these parameters as

$$\begin{aligned} \xi_{\mathrm{ev}} &= \lim_{t\uparrow t_{\sup}} \left(\frac{1-F_T(t)}{f_T(t)}\right)', \\ \mu_{\mathrm{ev}} &= F_T^{-1}\left(1-\frac{1}{\theta^*}\right), \text{ and } \sigma_{\mathrm{ev}} = \frac{1}{\theta^* f_T(\mu_{\mathrm{ev}})}. \end{aligned} \tag{5.10}$$

While (5.10) is applicable to any distribution that attains an EV distribution, we use parameters implicitly defined by

$$\frac{1}{\theta^*} = \frac{\exp(-0.5\mu_{\mathrm{ev}}^2)}{\mu_{\mathrm{ev}}\sqrt{2\pi}} \text{ and } \sigma_{\mathrm{ev}} = \frac{1}{\mu_{\mathrm{ev}}} \tag{5.11}$$

with $\xi_{\mathrm{ev}} = 0$ if $T \sim N(0,1)$, as suggested in Hall [45].

5.5 Single-Index Uniform Importance Sampling

Suppose that we are interested in estimating p_l for $l \in [l_1, l_2]$ for some $l_1, l_2 \in \mathbb{R}$. The main idea of single-index UIS is to truncate the domain of T from $(t_{\text{inf}}, t_{\text{sup}})$ to an interval $[t_1, t_2]$ and then use a uniform distribution over this range as a proposal distribution.

We consider the linear normal problem of (5.4) to illustrate that not all domain of T is relevant and we only need samples from a subset of the domain. Suppose $\alpha^2 = 0.95$ and $s^2 = 0.05$ so that Figure 5.1 shows the optimal proposal densities for p_l for $l \in \{3, 4, 5\}$. If $l_1 = 3$ and $l_2 = 5$, this figure suggests that the optimally calibrated density for any $l \in [3, 5]$ would almost never generated samples of $T < 1.5$ and $T > 6.5$. That is, the relevant domain of T for this specific problem is contained in $[1.5, 6.5]$. Single-index UIS then uses $U[1.5, 6.5]$ as a IS distribution of T.

We now explain how to choose the truncated domain and construct a UIS estimator. Suppose that for small $\epsilon_1, \epsilon_2 > 0$, there exist $t_1, t_2 \in (t_{\text{inf}}, t_{\text{sup}})$ such that $p_l(t) \leq \epsilon_1$ for $t \leq t_1$ and $p_l(t) \geq 1 - \epsilon_2$ for $t \geq t_2$ for all $l \in [l_1, l_2]$. Treating $p_l(t) \approx 0$ for $t < t_1$ and $p_l(t) \approx 1$ for $t > t_2$, we can approximate p_l as

$$\begin{aligned} p_l &= \int_{t_{\text{inf}}}^{t_{\text{sup}}} p_l(t) f_T(t) dt \\ &= \int_{t_{\text{inf}}}^{t_1} p_l(t) f_T(t) dt + \int_{t_1}^{t_2} p_l(t) f_T(t) dt + \int_{t_2}^{t_{\text{sup}}} p_l(t) f_T(t) dt \\ &\approx \int_{t_1}^{t_2} p_l(t) f_T(t) dt + \bar{F}_T(t_2) = p_l^{\text{UIS}}, \end{aligned}$$

and the bias is

$$\begin{aligned} p_l - p_l^{\text{UIS}} &= \int_{t_{\text{inf}}}^{t_1} p_l(t) f_T(t) dt - \int_{t_2}^{t_{\text{sup}}} (1 - p_l(t)) f_T(t) dt \\ &\leq \epsilon_1 F_T(t_1) + \epsilon_2 \bar{F}_T(t_2). \end{aligned} \tag{5.12}$$

Choosing ϵ_1 and ϵ_2 in such a way that the bias (5.12) is much smaller than p_l for $l \in [l_1, l_2]$, we can estimate p_l by constructing an estimator for p_l^{UIS}. This approximation procedure effectively truncates the domain from $(t_{\text{inf}}, t_{\text{sup}})$ to $[t_1, t_2]$. We then use a uniform distribution on $[t_1, t_2]$ as a proposal proposal distribution for T, that is,

$$g_T^{\mathrm{U}}(t) = \frac{1}{t_2 - t_1} \mathbb{1}\left[t \in [t_1, t_2]\right]. \tag{5.13}$$

The UIS estimator based on n samples has the form

$$\hat{p}_{l,n}^{\mathrm{UIS}} = \frac{1}{n}(t_2 - t_1) \sum_{i=1}^{n} \mathbb{1}[L_i > l] \tilde{w}_T(T) + \bar{F}_T(t_2), \quad T_i \overset{\text{ind.}}{\sim} U[t_1, t_2], \quad \boldsymbol{X}_i \sim f_{\boldsymbol{X}|T}(\boldsymbol{x} \,|\, T_i).$$

Generally, we choose $T(\cdot)$ such that $T \sim f_T(t)$ has an analytical distribution so $\bar{F}_T(t_2)$ can be computed with high accuracy.

As we will see in the numerical study of Section 5.6, UIS gives estimators whose RE are less sensitive to l for the normal and t-copula credit portfolio problems, compared to the estimators based on the optimally calibrated IS and EVIS.

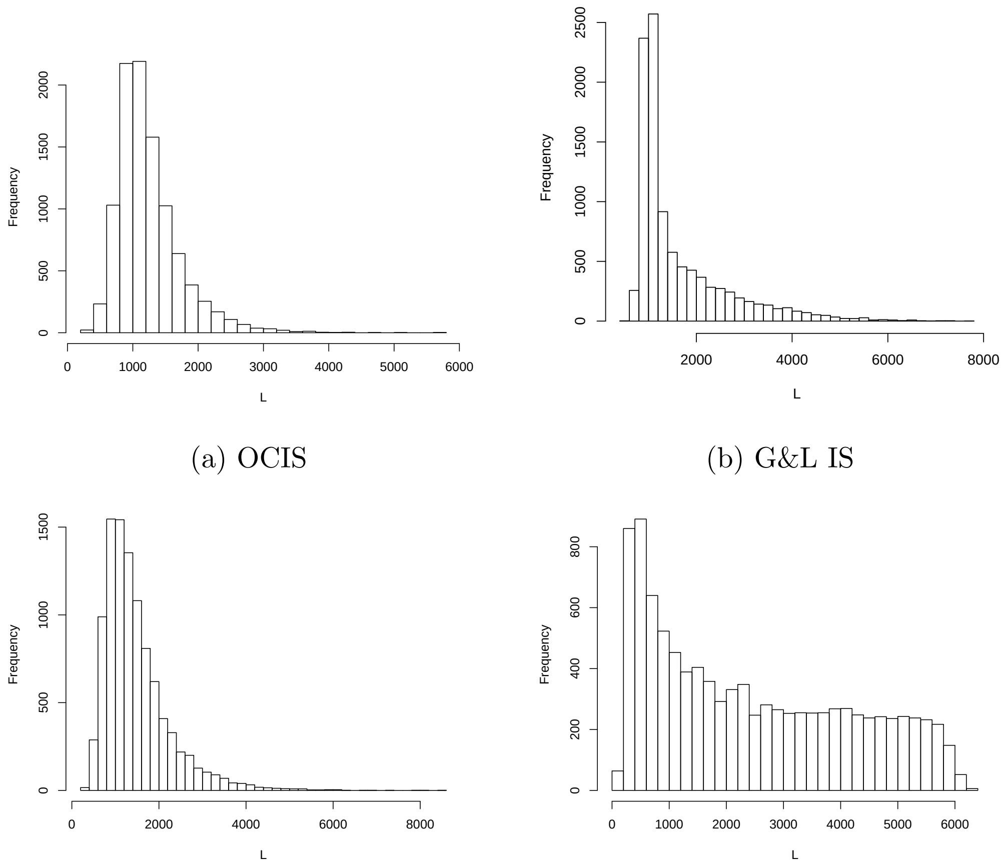
5.6 Numerical Experiments

5.6.1 Gaussian Copula Credit Portfolio Problem

In this section, we study the efficiency of single-index EVIS and UIS for the estimation of large loss probabilities of a credit portfolio under the Gaussian copula model studied by Glasserman and Li [41]. We applied IS and SIS with optimal calibrations to this problem in Section 4.7.3 and found that the optimal calibration is not suitable for multiple loss probabilities estimation. See Section 4.7.3 for the problem formulation and the parameters used for this problem.

The objective is to estimate $p_l = \mathbb{P}(L > l)$ for all $l \in [1,000, 5,000]$ in one simulation run. We apply single-index OCIS, EVIS, and UIS to this problem and compare them to the Glasserman and Li's IS method of [41], to which we refer as G&L IS. For OCIS, EVIS and G&L IS, the proposal distribution is calibrated for $l_c = 1,000$. The reason that calibration is done for $l_c = 1,000$ is that the three IS methods perform poorly in estimating p_l for l smaller than l_c. If for instance we let $l_c = 2,000$, the estimate of p_l for $l \in [1,000, 1500]$ will be very unreliable, which do not meet our objectives. For UIS, the domain of T is truncated to estimate $p_l = \mathbb{P}(L > l)$ for all $l \in [1,000, 5,000]$. Figure 5.3 shows the histogram of 10,000 samples of L based on OCIS, G&L IS, EVIS, and UIS. As expected, EVIS and UIS both give more samples of L in the right-tail than the OCIS does. UIS gives the samples with heaviest tail, followed by in the order of G&L IS, EVIS and optimally calibrated IS.

Figure 5.3: Histogram of Portfolio Loss based on 10,000 observations



(a) OCIS

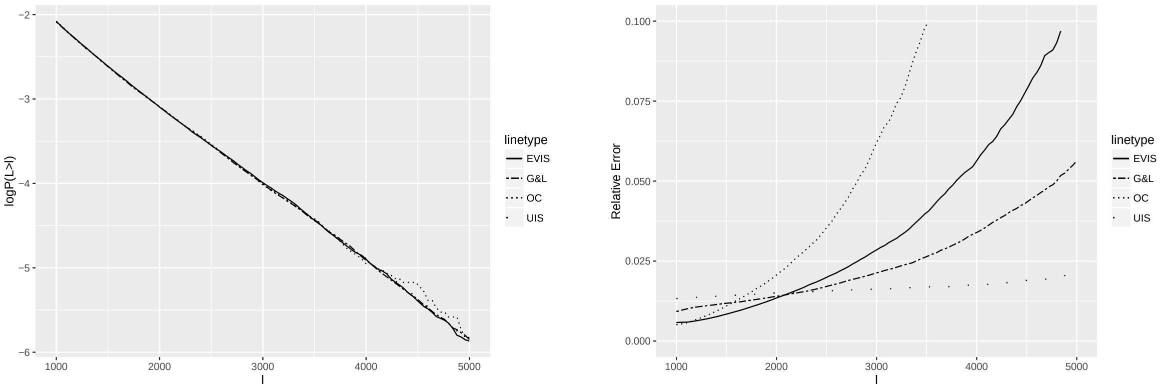
(b) G&L IS

(c) EVIS

(d) UIS

Figure 5.4 shows the estimated probabilities in log-scale and RE of the three methods. The estimated loss probabilities from the three methods are almost identical except when $l > 4,000$. As for RE, no single method dominates other methods for all range of $l \in$

Figure 5.4: Plot of Estimated Probability and Relative Error based on 30,000 observations



(a) Plot of $\log \mathbb{P}(L > l)$

(b) Plot of Relative Error of $\mathbb{P}(L > l)$

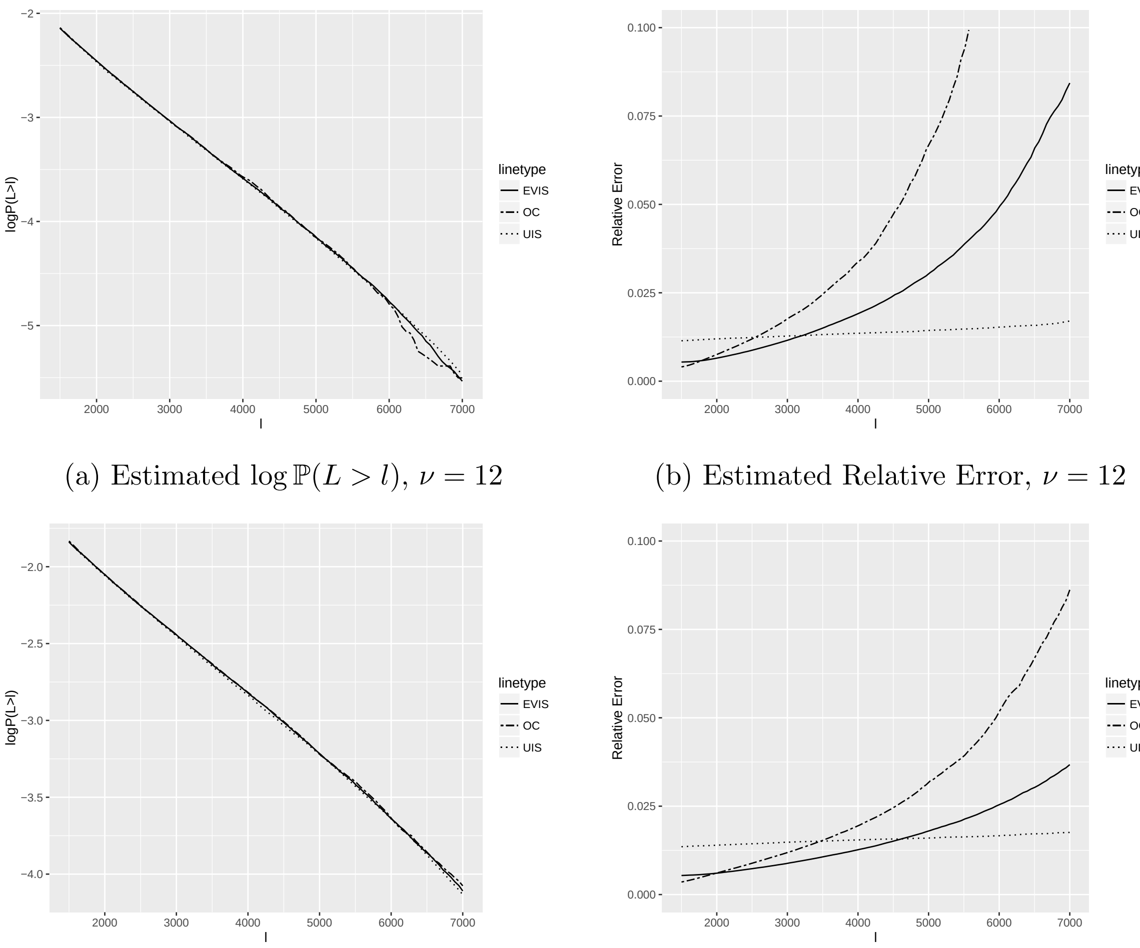
$[1,000, 5,000]$. EVIS and G&L IS provide estimates of p_l with small RE when l is close to l_c and the RE increases as l gets further away from the calibration point. EVIS gives smaller RE than G&L IS does when l is close to l_c, but G&L IS outperforms EVIS when l is much larger than l_c. The RE of the UIS estimates, on the other hand, remains fairly constant across the different values of l, which is desirable if all estimates of p_l, $l \in [1,000, 5,000]$ are equally important. If the estimate of p_l is more important for the value of l closer to $l_c = 1,000$, EVIS and G&L provide better estimates than UIS does.

5.6.2 t-copula Credit Portfolio Problem

In this section, we study the efficiency of single-index EVIS and UIS for the estimation of large loss probabilities of a credit portfolio under the t-copula model that we examined in Section 4.7.4. There, we applied single-index IS based on T_1 to estimate multiple loss probabilities and saw that the optimally calibrated density struggles with multiple estimation. We do not consider the IS schemes based on the conditional Monte Carlo [20] as the CMC becomes very computationally expensive when multiple probabilities are to be estimated. See Section 4.7.3 and Section 4.7.4 for the problem formulation and the parameters used for this problem.

The objective is to estimate $p_l = \mathbb{P}(L > l)$ for all $l \in [1,500, 7,000]$ in one simulation run. Figure 5.5 shows the plot of estimated $\log \mathbb{P}(L > l)$ and RE for $\nu = 12$ and $\nu = 4$ cases for OCIS, EVIS, and UIS based on T_1 (see Section 4.7.4 for how T_1 is defined). The proposal distributions for optimally calibrated IS and EVIS are calibrated for $l_c = 1,500$. For either case of the degree of freedom parameter, optimally calibrated IS gives smallest RE while UIS gives the largest RE for l near $l_c = 1,500$. Similarly to the Gaussian copula case, the RE of UIS estimates remains fairly constant for all $l \in [1,500, 7,000]$. Thus, if one is concerned with the estimation quality of p_l for all $l \in [1,500, 7,000]$, UIS is the preferred approach. The RE of OCIS and EVIS grows with l, but the growth is slower with EVIS. Thus, if the estimate of p_l is more important for the value of l closer to $l_c = 1,500$ and the estimates with moderate size of l, say 5,000, is also of interest, EVIS is the preferred approach.

Figure 5.5: Plot of Estimated Probability and Relative Error based on 30,000 observations



(a) Estimated $\log \mathbb{P}(L > l)$, $\nu = 12$

(b) Estimated Relative Error, $\nu = 12$

(c) Estimated $\log \mathbb{P}(L > l)$, $\nu = 4$

(d) Estimated Relative Error, $\nu = 4$

Chapter 6

Importance Sampling for Multi-Index Model

6.1 Introduction

In Chapter 4 and Chapter 5, we have focused on problems based on single-index models where the output variable, $\Psi(\boldsymbol{X})$, depends on a d-dimensional random vector $\boldsymbol{X}$, mainly through some one-dimensional projection $T = T(\boldsymbol{X})$. The problems that we have examined for simulation studies in those chapters all had the linear single-index structure where $\Psi(\boldsymbol{X})$ depends on $\boldsymbol{X}$ mostly through some linear combination of $\boldsymbol{X}$. Our proposed single-index IS and SIS methods achieved substantial variance reduction for those types of problems. For some applications, however, the single-index model may be too restrictive, even if these applications have some low-dimensional structure. In this chapter, we relax the single-index assumption and develop multi-index IS, the IS techniques for a multi-index model where $\Psi(\boldsymbol{X})$ depends on $\boldsymbol{X}$ mainly through a set of linear combinations of the form $\boldsymbol{\beta}'\boldsymbol{X}$, where $\boldsymbol{\beta} \in \mathbb{R}^{d\times p}$ denotes a direction matrix and p is the number of relevant linear combinations. This multi-index model contains the linear single-index model as a special case when p=1.

Multi-index IS is a two-stage procedure. In the first stage, it estimates the direction matrix $\boldsymbol{\beta}$. The form of $\boldsymbol{\beta}$ may be known analytically or one can use the existing methods that are developed to estimate $\boldsymbol{\beta}$ such as sliced inverse regression of Li [78]. If $\boldsymbol{\beta}$ is correctly specified, one can identify the p-dimensional projection vector $\boldsymbol{T} = \boldsymbol{\beta}'\boldsymbol{X}$ that jointly explains the majority of the variation of $\Psi(\boldsymbol{X})$ under the multi-index model. In the second stage, multi-index IS applies IS on $\Psi(\boldsymbol{X})$ through changing the distribution of $\boldsymbol{T}$ to make the rare event more frequent. In particular, it samples $\boldsymbol{T}$ from some proposal distribution of $\boldsymbol{T}$ and then samples $\boldsymbol{X} \,|\, \boldsymbol{T}$ under the original distribution.

Multi-index have strengths similar to those of single-index IS. Multi-index IS is applicable to many problems as it does not assume specific form of Ψ nor distribution of $\boldsymbol{X}$. The conditional sampling step of sampling $\boldsymbol{X} \,|\, \boldsymbol{T}$ reduces the dimension of the IS weight function of p, so if $p \ll d$, multi-index IS is less susceptible to the dimensionality problem discussed in Section 2.2.3 than the IS methods that change the distribution of $\boldsymbol{X}$ altogether. If p is equal to 2 or 3, the multi-index formulation allows us to use a kernel density as the proposal density.

The focus of this chapter is the second stage of multi-index IS. The emphasis is on the construction and the calibration of effective parametric and nonparametric proposal distributions of $\boldsymbol{T}$. While the accurate estimation of $\boldsymbol{\beta}$ in the first stage is crucial to correctly identify the projected variable $\boldsymbol{T}$, a variety of estimation methods of $\boldsymbol{\beta}$ have been already proposed, including the linear transformation of Imai and Tan [58], sliced inverse regression of Li [78], and principal fitted component model of Cook and Forzani [23]. Since one can use these existing techniques to deal with the estimation of $\boldsymbol{\beta}$, we focus on the design of proposal distributions.

The rest of this chapter is organized as follows. Section 6.2 introduces multi-index models and Section 6.3 presents the general multi-index IS setting. A sampling procedure for $\boldsymbol{X} \,|\, \boldsymbol{T}$ when $\boldsymbol{X}$ follows a multivariate normal distribution is also given. Section 6.4 develops calibration methods for parametric and nonparametric proposal densities. Section 6.5 applies multi-index IS to rainbow Asian option pricing problems.

6.2 Multi-Index Models

Multi-index model assumes that there exists a $d \times p$ direction matrix $\boldsymbol{\beta}$ such that $\Psi(\boldsymbol{X})$ depends on $\boldsymbol{X}$ mainly through $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{X}$, p linear combinations of $\boldsymbol{X}$, as

$$\mathrm{E}[\Psi(\boldsymbol{X})] = m(\boldsymbol{T}), \tag{6.1}$$

where $m(\boldsymbol{t}) = \mathrm{E}[\Psi(\boldsymbol{X}) \,|\, \boldsymbol{T} = \boldsymbol{t}]$ is the conditional mean function. We can equivalently write (6.1) as

$$\Psi(\boldsymbol{X}) = m(\boldsymbol{T}) + \epsilon_{\boldsymbol{T}}, \quad \epsilon_{\boldsymbol{T}} \,|\, \boldsymbol{T} \sim (0, v^2(\boldsymbol{T})), \tag{6.2}$$

where $v^2(\boldsymbol{t}) = \mathrm{Var}(\Psi(\boldsymbol{X}) \,|\, \boldsymbol{T} = \boldsymbol{t})$ is the conditional variance function. Note that in the special case $p = 1$, (6.2) becomes a linear single-index model. As in single-index models, we can decompose the variance of $\Psi(\boldsymbol{Y})$ as the sum of the variance captured by the systematic part of the multi-index model and the variance of the error term as

$$\mathrm{Var}(\Psi(\boldsymbol{X})) = \mathrm{Var}(m(\boldsymbol{T})) + \mathrm{Var}(\epsilon_{\boldsymbol{T}}).$$

Note that a multi-index model perfectly fits any function Ψ if one chooses $p = d$ and $\boldsymbol{\beta}$ to be an identity matrix (or any $d * d$ full rank matrix). Nevertheless, such models provide no insight on the structure of $\Psi(\boldsymbol{X})$. Ultimately we want to learn the low-dimensional representation of the problem, if it exists, and use this information to construct effective proposal distributions Thus, we say that a problem has a multi-index structure if the measure of fit $R^2 = \mathrm{Var}(m(\boldsymbol{T}))/\mathrm{Var}(\Psi(\boldsymbol{X}))$ is close to 1 for $p \ll d$, preferably $p = 2$ or $p = 3$.

The main idea of multi-index IS is to apply IS on $\Psi(\boldsymbol{X})$ through changing the distribution of $\boldsymbol{T}$. To do so, one first needs to identify p, the number of significant linear combinations, and $\boldsymbol{\beta}$, the coefficient matrix for linear combination, to define $\boldsymbol{T}$ and this is the first stage of multi-index IS. In some cases, one may be able to deduce p and the form of $\boldsymbol{\beta}$ by analyzing the structure of the problem. If that is not possible, we propose to use existing methods for estimation. A variety of techniques have been proposed to estimate $\boldsymbol{\beta}$ (and p) in the context of sufficient dimension reduction (see [4, 22] for overview of this field), including sliced inverse regression [78], sliced average variance estimates [25], and principal fitted components [23]. One can also use the linear transformation of Imai and Tan [58] developed in the context of dimension reduction for QMC. Note that conditioning on $\boldsymbol{T} = \boldsymbol{\beta}'\boldsymbol{X}$ is equivalent to conditioning on $(\boldsymbol{\beta}\kappa)'\boldsymbol{X}$ for any full rank $p \times p$ matrix κ, thus $\boldsymbol{\beta}$ is identifiable up to span($\boldsymbol{\beta}$).

6.3 Multi-index Importance Sampling

Suppose that a $\mathbb{R}^{d\times p}$ direction matrix $\boldsymbol{\beta}$ has been estimated and let $\Omega_{\boldsymbol{T}} \subseteq \mathbb{R}^p$ denote the the domain of $\boldsymbol{T}$ under the original distribution. The multi-index IS draws $\boldsymbol{T}$ from a proposal density of $\boldsymbol{T}$ denoted by $g_{\boldsymbol{T}}(\boldsymbol{t})$ and then draws $\boldsymbol{X}\,|\,\boldsymbol{T}$ under the original distribution. Following an argument similar to the one for single-index IS in Section 4.3.1, it is easy to show

$$g_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}\,|\,\boldsymbol{t}_{\boldsymbol{x}})g_{\boldsymbol{T}}(\boldsymbol{t}_{\boldsymbol{x}}) \quad \text{and} \quad w(\boldsymbol{x}) = \frac{f_{\boldsymbol{T}}(\boldsymbol{t}_{\boldsymbol{x}})}{g_{\boldsymbol{T}}(\boldsymbol{t}_{\boldsymbol{x}})}, \tag{6.3}$$

where $\boldsymbol{t}_{\boldsymbol{x}} = \boldsymbol{\beta}'\boldsymbol{x}$. Note that the IS weight function is simply the ratio of the original and the IS density of $\boldsymbol{T}$. In order to simplify the notation, define $\tilde{w} : \mathbb{R}^p \to \mathbb{R}$ as $\tilde{w}(\boldsymbol{t}) = \frac{f_{\boldsymbol{T}}(\boldsymbol{t})}{g_{\boldsymbol{T}}(\boldsymbol{t})}$. For $w(\boldsymbol{x})$ to be well-defined, we need $g_{\boldsymbol{T}}(\boldsymbol{t}) > 0$ whenever $f_{\boldsymbol{T}}(\boldsymbol{t}) > 0$. But, we only need $g_{\boldsymbol{T}}(\boldsymbol{t}) > 0$ whenever $m(\boldsymbol{t})f_{\boldsymbol{T}}(\boldsymbol{t}) > 0$ for the IS estimator to be unbiased. Algorithm 8 summarizes this IS scheme.

Algorithm 8 Multi-index Importance Sampling

```
for $i = 1, \dots, n$ do
    Draw $T_i \sim g_{\boldsymbol{T}}(\boldsymbol{t})$
    Draw $\boldsymbol{Y}_i \sim f_{\boldsymbol{X}|T}(\boldsymbol{x}\,|\,\boldsymbol{T}_i)$
    Compute $w_i = \tilde{w}(\boldsymbol{T}_i) = f_{\boldsymbol{T}}(\boldsymbol{T}_i)/g_{\boldsymbol{T}}(\boldsymbol{T}_i)$.
end for
return $\hat{\mu}_{\text{IS},n} = \frac{1}{n}\sum_{i=i}^{n} \Psi(\mathbf{X}_i)w_i$.
```

We now explain the conditional sampling step for $(\boldsymbol{X}\,|\,\boldsymbol{T} = \boldsymbol{t})$ when $\boldsymbol{X}$ follows a multivariate normal distribution. If $\boldsymbol{X} \sim \text{MVN}(\mu_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$, then $\boldsymbol{T} \sim \text{MVN}(\mu_{\boldsymbol{T}}, \Sigma_{\boldsymbol{T}})$, where

$\mu_{\boldsymbol{T}} = \boldsymbol{\beta}'\mu_{\boldsymbol{X}}$ and $\Sigma_{\boldsymbol{T}} = \boldsymbol{\beta}'\Sigma_{\boldsymbol{X}}\boldsymbol{\beta}$. If $\operatorname{rank}(\Sigma_{\boldsymbol{T}}) = p$, then by [48, Theorem 1]

$$\begin{aligned}
(\boldsymbol{X} \,|\, \boldsymbol{T} = \boldsymbol{t}) &\sim \mathrm{MVN}\left(\mu_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{t}), \Sigma_{\boldsymbol{X}|\boldsymbol{T}}\right), \text{ where} \\
\mu_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{t}) &= \mu_{\boldsymbol{X}} + \Sigma_{\boldsymbol{X}}\boldsymbol{\beta}\,\Sigma_{\boldsymbol{T}}^{-1}(\boldsymbol{t} - \mu_{\boldsymbol{T}}) \text{ and} \\
\Sigma_{\boldsymbol{X}|\boldsymbol{T}} &= \Sigma_{\boldsymbol{X}} - \Sigma_{\boldsymbol{X}}\boldsymbol{\beta}\,\Sigma_{\boldsymbol{T}}^{-1}\boldsymbol{\beta}'\,\Sigma_{\boldsymbol{X}}.
\end{aligned} \tag{6.4}$$

The sampling efficiency of multi-index IS for MVN models comes from the fact that the covariance matrix, $\Sigma_{\boldsymbol{X}|\boldsymbol{T}}$ of (6.4) does not depend on the conditioning value $\boldsymbol{t}$, so this matrix needs to be decomposed only once during the entire IS algorithm. If $\boldsymbol{X} \sim \mathrm{MVN}(\boldsymbol{0}, I_d)$ and $\boldsymbol{\beta}$ is such that $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$, then (6.4) simplifies to

$$(\boldsymbol{X} \,|\, \boldsymbol{T} = \boldsymbol{t}) \sim \mathrm{MVN}\left(\boldsymbol{\beta}\boldsymbol{t}, I_d - \boldsymbol{\beta}\boldsymbol{\beta}'\right). \tag{6.5}$$

In order to sample from (6.5), we need to find a matrix $C \in \mathbb{R}^{d*d}$ such that $CC' = I_d - \boldsymbol{\beta}\boldsymbol{\beta}'$. As in [38], since $(I_d - \boldsymbol{\beta}\boldsymbol{\beta}')(I_d - \boldsymbol{\beta}\boldsymbol{\beta}')' = I_d - \boldsymbol{\beta}\boldsymbol{\beta}'$, we can simply let $C = I_d - \boldsymbol{\beta}\boldsymbol{\beta}'$. Then, for $\boldsymbol{Z} \sim \mathrm{MVN}(0, I_d)$, we have that

$$(\boldsymbol{X} \,|\, \boldsymbol{T} = \boldsymbol{t}) \overset{D}{=} \boldsymbol{\beta}\boldsymbol{t} + (I_d - \boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{Z} = \boldsymbol{\beta}\boldsymbol{t} + \boldsymbol{Z} - \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{Z}). \tag{6.6}$$

The significance of (6.6) is that we do not need to explicitly compute $\boldsymbol{\beta}\boldsymbol{\beta}'$. This saves substantial computational effort when d is large.

6.4 Calibration of Proposal Density for the Multi-Index IS

Recall from (6.3) that the proposal density of $\boldsymbol{X}$ considered by multi-index IS has the form $g_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x} \,|\, \boldsymbol{t})g_{\boldsymbol{T}}(\boldsymbol{t})$. Since $f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x} \,|\, \boldsymbol{t})$ is fixed, we have freedom in choosing $g_{\boldsymbol{T}}(\boldsymbol{t})$, the proposal density of $\boldsymbol{T}$. In this section, we develop calibration methods for parametric and nonparametric proposal densities of $\boldsymbol{T}$.

6.4.1 Variance Minimization and Cross Entropy Method

Let $m^{(2)}(\boldsymbol{t}) = \mathrm{E}\left[\Psi^2(\boldsymbol{X}) \,|\, \boldsymbol{T} = \boldsymbol{t}\right]$, $s(\boldsymbol{t}) = \mathrm{E}\left[|\Psi(\boldsymbol{X})| \,|\, \boldsymbol{T} = \boldsymbol{t}\right]$, and $A_t = \{\boldsymbol{t} \in \mathbb{R}^p \mid g_{\boldsymbol{T}}(\boldsymbol{t}) > 0\}$. Using an argument similar to the variance analysis for the single-index IS (Proposition

4.3.3), it is easy to show that

$$n\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \int_{A_t} m^{(2)}(\boldsymbol{t}) \frac{f_{\boldsymbol{T}}^2(\boldsymbol{t})}{g_{\boldsymbol{T}}(\boldsymbol{t})} d\boldsymbol{t} - \mu^2,$$

assuming that the IS estimator is unbiased. The proposal density $g_{\boldsymbol{T}}(\boldsymbol{t})$ that minimizes the variance of the IS estimator is

$$g_{\boldsymbol{T}}^{\mathrm{VM}}(\boldsymbol{t}) = \frac{\sqrt{m^{(2)}(\boldsymbol{t})} f_{\boldsymbol{T}}(\boldsymbol{t})}{\int_{\mathbb{R}^d} \sqrt{m^{(2)}(\boldsymbol{t})} f_{\boldsymbol{T}}(\boldsymbol{t}) d\boldsymbol{t}}, \quad \boldsymbol{t} \in \mathbb{R}^p. \tag{6.7}$$

The calibration criterion for (6.7) is variance-minimization (VM) as it aims to minimize the variance of the resulting IS estimator. Another popular calibration criterion is the minimization of Kullback-Leibler divergence (KLD) [71] with respect to the theoretically optimal proposal distribution and IS based on such calibration techniques is called the cross-entropy (CE) method (see [26, 102, 103] and references therein). The advantage of the CE method over the VM calibration is that the former often leads to optimization problems that are easier to solve than the latter does. Recall from (2.16) that the theoretically optimal proposal density has the form

$$g_{\boldsymbol{X}}^*(\boldsymbol{x}) = c^* |\Psi(\boldsymbol{x})| f_{\boldsymbol{X}}(\boldsymbol{x}), \quad c^* = \left(\int_{\Omega_{\boldsymbol{X}}} |\Psi(\boldsymbol{x})| f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \right)^{-1}. \tag{6.8}$$

Let $D_{\mathrm{KL}}(\cdot||\cdot)$ denote a KLD operator, that is,

$$D_{\mathrm{KL}}(h(\boldsymbol{x})||r(\boldsymbol{x})) = \mathrm{E}_h \left[\ln \frac{h(\boldsymbol{X})}{r(\boldsymbol{X})} \right] = \int_{\mathbb{R}^d} \ln \left(\frac{h(\boldsymbol{x})}{r(\boldsymbol{x})} \right) h(\boldsymbol{x}) d\boldsymbol{x}$$

for d-dimensional densities $h(\boldsymbol{x})$ and $r(\boldsymbol{x})$. Define

$$g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t}) = c^* s(\boldsymbol{t}) f_{\boldsymbol{T}}(\boldsymbol{t}), \quad \boldsymbol{t} \in \Omega_{\boldsymbol{T}}, \tag{6.9}$$

where the normalizing constant c^* is the same as the one for $g_{\boldsymbol{X}}^*(\boldsymbol{x})$. The following proposition states that when searching for a proposal density of $\boldsymbol{X}$ multi-index IS (6.3), minimizing $D_{\mathrm{KL}}(g_{\boldsymbol{X}}^*(\boldsymbol{x})||g_{\boldsymbol{X}}(\boldsymbol{x}))$ is equivalent to minimizing $D_{\mathrm{KL}}\left(g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})||g_{\boldsymbol{T}}(\boldsymbol{t})\right)$. Trivially, $g_{\boldsymbol{T}}(\boldsymbol{t}) = g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$ is the optimal proposal density of $\boldsymbol{T}$ in the CE method.

Proposition 6.4.1 (see p. 162 for proof)**.** *For a proposal density of* $\boldsymbol{X}$ *defined through* $g_{\boldsymbol{T}}(\boldsymbol{t})$*, that is,* $g_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x} \mid \boldsymbol{t}) g_{\boldsymbol{T}}(\boldsymbol{t})$*,*

$$D_{\mathrm{KL}}\left(g^*_{\boldsymbol{X}}(\boldsymbol{x})||g_{\boldsymbol{X}}(\boldsymbol{x})\right) = D_{\mathrm{KL}}\left(g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})||g_{\boldsymbol{T}}(\boldsymbol{t})\right) + c, \tag{6.10}$$

where $g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$ *is defined as in* (6.9) *and* c *is a constant that does not depend on* $g_{\boldsymbol{T}}(\boldsymbol{t})$*.*

Of course, we do not know the exact shape of $g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$ as it depends on the unknown conditional absolute mean function $s(\boldsymbol{t})$. Nonetheless, Proposition 6.4.1 provides a guideline when searching for a proposal density of T; we want to choose $g_{\boldsymbol{T}}(\boldsymbol{t})$ that gives a heavier weight to the region of Ω_T where the product of $g_{\boldsymbol{T}}(\boldsymbol{t})$ and $s(\boldsymbol{t})$ is large. This observation becomes useful when choosing a parametric form for $g_{\boldsymbol{T}}(\boldsymbol{t})$ in the simulation studies of Section 6.5. The following proposition states how samples from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ and $g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$ are related.

Proposition 6.4.2 (see p. 163 for proof)**.** *Let* $\boldsymbol{X}^* \sim g^*_{\boldsymbol{X}}(\boldsymbol{x})$ *defined as in* (6.8) *and let* $\boldsymbol{T}^* = T(\boldsymbol{X}^*)$*. Then* $\boldsymbol{T}^* \sim g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$ *defined as in* (6.9)*.*

By Proposition 6.4.2, once we have samples from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$, we also have samples from $g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$. Of course, we cannot sample exactly from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ but this observation becomes useful when developing calibration methods for parametric and nonparametric proposal distributions in Section 6.4.2 and Section 6.4.3.

From (6.7) and (6.9), $g^{\mathrm{VM}}_{\boldsymbol{T}}(\boldsymbol{t}) = g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$ if $v^2(\boldsymbol{t}) = 0$ for all $\boldsymbol{t} \in \Omega_{\boldsymbol{T}}$. This occurs when the multi-index model (6.2) gives a perfect fit. Otherwise, $g^{\mathrm{VM}}_{\boldsymbol{T}}(\boldsymbol{t})$ gives the estimator with a smaller variance by construction, but the difference in variance should be small if the fit is near perfect. We were able to use numerical inversion techniques to sample (almost) exactly from the variance-minimizing proposal density for single-index IS (4.8) because T is univariate under single-index model. For multi-index IS, $\boldsymbol{T}$ is multivariate, so sampling exactly from (6.7) and (6.9) is a much more difficult task and one generally needs Markov Chain Monte Carlo (MCMC) [36, 99] for this purpose. As MCMC is computationally intensive, in this chapter we instead use parametric and nonparametric distributions that approximate the optimal densities as proposal distributions.

6.4.2 Parametric IS Density of T

Suppose that we have decided on the parametric form of the proposal density as $g_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is the vector of parameters. Let $g_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) = f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}\,|\,\boldsymbol{t}) g_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta})$ denote the corresponding proposal density of $\boldsymbol{X}$. For simplicity, we assume that the support of $\boldsymbol{T}$ is $\Omega_{\boldsymbol{T}}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Ideally, we want to find the variance-minimizing $\boldsymbol{\theta}^{\mathrm{VM}}$ which satisfies

$$\boldsymbol{\theta}^{\mathrm{VM}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\Omega_{\boldsymbol{X}}} \Psi(\boldsymbol{X}) \frac{f_{\boldsymbol{X}}^2(\boldsymbol{x})}{g_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})} d\boldsymbol{x} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\Omega_{\boldsymbol{T}}} m^{(2)}(\boldsymbol{t}) \frac{f_{\boldsymbol{T}}^2(\boldsymbol{t})}{g_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta})} d\boldsymbol{t}, \tag{6.11}$$

but it is generally difficult to solve this minimization problem. We thus take the CE approach as it leads to optimization problems that are easier to solve. In particular, we search for $\boldsymbol{\theta}^{\mathrm{KL}}$ that minimizes $D_{\mathrm{KL}}\left(g^*_{\boldsymbol{X}}(\boldsymbol{x}) || g_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})\right)$ and find

$$\begin{aligned}
\boldsymbol{\theta}^{\mathrm{KL}} &= \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\Omega_{\boldsymbol{X}}} \ln\left(\frac{g^*_{\boldsymbol{X}}(\boldsymbol{x})}{g_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})}\right) g^*_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\Omega_{\boldsymbol{X}}} \ln\left(g_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})\right) |\Psi(\boldsymbol{x})| f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\Omega_{\boldsymbol{X}}} \ln\left(g_{\boldsymbol{T}}(\boldsymbol{t}_{\boldsymbol{x}};\boldsymbol{\theta})\right) |\Psi(\boldsymbol{x})| f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}[\ln\left(g_{\boldsymbol{T}}(\boldsymbol{T};\boldsymbol{\theta})\right) |\Psi(\boldsymbol{X})|], \quad \boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}) && (6.12) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}[\ln\left(g_{\boldsymbol{T}}(\boldsymbol{T};\boldsymbol{\theta})\right) s(\boldsymbol{T})], \quad \boldsymbol{T} \sim f_{\boldsymbol{T}}(\boldsymbol{t}) && (6.13)
\end{aligned}$$

where $\boldsymbol{t}_{\boldsymbol{x}} = T(\boldsymbol{x})$. It is easier to solve (6.12) than (6.13) as $s(\boldsymbol{T})$ does not need to be known or approximated when solving (6.12). Since analytically solving (6.12) is still difficult, we may solve the stochastic counterpart based on M samples of $\boldsymbol{X}$ from the original distribution

$$\hat{\boldsymbol{\theta}}^{\mathrm{KL}}_M = \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{M} \sum_{i=i}^{M} \ln(g_{\boldsymbol{T}}(\boldsymbol{T}_i;\boldsymbol{\theta})) |\Psi(\boldsymbol{X}_i)|, \quad \boldsymbol{X}_i \sim f_{\boldsymbol{X}}(\boldsymbol{x}). \tag{6.14}$$

If the problem is rare-event simulation, then $\Psi(\boldsymbol{X}) = 0$ for most samples of $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x})$, and therefore $\hat{\boldsymbol{\theta}}^{\mathrm{KL}}_M$ will not be a reliable estimate of $\boldsymbol{\theta}^{\mathrm{KL}}$. We can use IS to improve the quality of the estimation of $\boldsymbol{\theta}^{\mathrm{KL}}$. To distinguish from the proposal distribution $g_{\boldsymbol{X}}(\boldsymbol{x})$ of

$\boldsymbol{X}$ used to construct an IS estimator $\hat{\mu}_{\mathrm{IS},n}$, we refer to the proposal distribution of $\boldsymbol{X}$ used to estimate $\boldsymbol{\theta}^{\mathrm{KL}}$ as a "design distribution" of $\boldsymbol{X}$ and we denote such a distribution of $\boldsymbol{X}$ by $h_{\boldsymbol{X}}(\boldsymbol{x})$. We then refer to $w_h(\boldsymbol{x}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{h_{\boldsymbol{X}}(\boldsymbol{x})}$ as a design weight function. Given M design samples $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M) \overset{\text{ind.}}{\sim} h_{\boldsymbol{X}}(\boldsymbol{x})$ and $\boldsymbol{T}_i = T(\boldsymbol{X}_i)$ for $i = 1, \ldots, M$, one may solve the IS version of (6.14)

$$\hat{\boldsymbol{\theta}}_M^{\mathrm{KL}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \frac{1}{M} \sum_{i=i}^{M} \ln(g_{\boldsymbol{T}}(\boldsymbol{T}_i; \boldsymbol{\theta})) |\Psi(\boldsymbol{X}_i)| w_h(\boldsymbol{X}_i), \quad \boldsymbol{X}_i \sim h_{\boldsymbol{X}}(\boldsymbol{x}). \tag{6.15}$$

If the shape of $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$ is approximately MVN, one may choose $g_{\boldsymbol{T}}(\boldsymbol{t}; \boldsymbol{\theta}) = g_{\boldsymbol{T}}(\boldsymbol{t}; \boldsymbol{\mu}_{\boldsymbol{T}}, \Sigma_{\boldsymbol{T}})$ to be the density of $\mathrm{MVN}(\boldsymbol{\mu}_{\boldsymbol{T}}, \Sigma_{\boldsymbol{T}})$ as the parametric family. Then, the analytical solution to (6.15) is given by [20];

$$\hat{\mu}_{\boldsymbol{T}}^{\mathrm{KL}} = \frac{\sum_{i=1}^{M} |\Psi(\boldsymbol{X}_i)| w_h(\boldsymbol{X}_i) \boldsymbol{T}_i}{\sum_{i=1}^{M} |\Psi(\boldsymbol{X}_i)| w_h(\boldsymbol{X}_i)}, \quad \hat{\Sigma}_{\boldsymbol{T}}^{\mathrm{KL}} = \frac{\sum_{i=1}^{M} |\Psi(\boldsymbol{X}_i)| w_h(\boldsymbol{X}_i) (\boldsymbol{T}_i - \hat{\mu}_{\boldsymbol{T}}^{\mathrm{KL}})(\boldsymbol{T}_i - \hat{\mu}_{\boldsymbol{T}}^{\mathrm{KL}})'}{\sum_{i=i}^{M} |\Psi(\boldsymbol{X}_i)| w_h(\boldsymbol{X}_i)}. \tag{6.16}$$

If the shape of $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$ is heavily skewed or multimodal, an MVN distribution may not be an appropriate parametric family for $g_{\boldsymbol{T}}(\boldsymbol{t}; \boldsymbol{\theta})$. In such a situation, more flexible distributions, such as the generalized hyperbolic (GH) distribution [88] or a mixture of the GH distributions may provide a better fit. The expectation-maximization (EM) algorithm [29] is usually employed to estimate the parameters of the (mixture) GH distributions [17, 88], but this technique is not directly applicable to (6.15). However, if $h_{\boldsymbol{X}}(\boldsymbol{x}) = g_X^*(\boldsymbol{x}) = c^* |\Psi(\boldsymbol{x})| f_{\boldsymbol{X}}(\boldsymbol{x})$ as in (6.15), $(\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_M^*) \overset{\text{ind.}}{\sim} g_{\boldsymbol{X}}^*(\boldsymbol{x})$, and $\boldsymbol{T}_i^* := T(\boldsymbol{X}_i^*)$ for $i = 1, \ldots, M$, then (6.15) becomes

$$\hat{\boldsymbol{\theta}}_M^{\mathrm{KL}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \frac{1}{M} \sum_{i=i}^{M} \ln(g_{\boldsymbol{T}}(\boldsymbol{T}_i^*; \boldsymbol{\theta})) \tag{6.17}$$

and computing such $\hat{\boldsymbol{\theta}}_M^{\mathrm{KL}}$ is similar to calculating maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ based on the samples $(\boldsymbol{T}_1^*, \ldots, \boldsymbol{T}_M^*)$, as pointed out by Akaike [5]. In light of Proposition 6.4.2, this is equivalent to estimating $\boldsymbol{\theta}^{\mathrm{KL}}$ by computing the MLE of $\boldsymbol{\theta}$ based on samples from $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$. If samples from $g_{\boldsymbol{X}}^*(\boldsymbol{x})$ were available, we could use techniques developed to compute MLE, such as the EM algorithm, to obtain $\hat{\boldsymbol{\theta}}_M^{\mathrm{KL}}$. However, in our case sampling exactly from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ is infeasible and we only have samples from some design distribution $h_{\boldsymbol{X}}(\boldsymbol{x})$. We thus propose to use Sampling/Importance Resampling (SIR) [101, 111] to obtain approximate samples from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ based on samples from $h_{\boldsymbol{X}}(\boldsymbol{x})$. Algorithm 9 gives the detailed procedure for the SIR step. Starting with M samples from $h_{\boldsymbol{X}}(\boldsymbol{x})$, Algorithm 9 returns M approximate samples $(\boldsymbol{X}_1^{\#}, \ldots, \boldsymbol{X}_M^{\#})$ from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$. We then let $T_i^{\#} = T(\boldsymbol{X}_i^{\#})$ for $i = 1, \ldots, M$ and compute the MLE of $\boldsymbol{\theta}$ based on the $T_i^{\#}$'s.

Algorithm 9 Sampling/Importance Resampling to sample from $g^*_{\boldsymbol{X}}(\boldsymbol{x}) = c^*|\Psi(\boldsymbol{x})|f_{\boldsymbol{X}}(\boldsymbol{x})$

```
Draw $M$ samples $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M) \overset{\text{ind.}}{\sim} h_{\boldsymbol{X}}(\boldsymbol{x})$;
Compute the weights $w_i = \frac{|\Psi(\boldsymbol{X}_i)|f_{\boldsymbol{X}}(\boldsymbol{X}_i)}{h_{\boldsymbol{X}}(\boldsymbol{X}_i)}$, $i = 1, \ldots, M$;
Normalize the weights $\bar{w}_i = w_i / \sum_{i=1}^{M} w_i$, $i = 1, \ldots, M$;
for $i = 1, \ldots, M$ do
    Draw $r \in \{1, \ldots, M\}$ with probabilities $(\bar{w}_1, \ldots, \bar{w}_M)$ and let $\boldsymbol{X}_i^{\#} = \boldsymbol{X}_r$;
end for
return $(\boldsymbol{X}_1^{\#}, \ldots, \boldsymbol{X}_M^{\#})$.
```

As mentioned in Smith and Gelfand [111], the efficiency of SIR depends on how closely $h_{\boldsymbol{X}}(\boldsymbol{x})$ approximates $g^*_{\boldsymbol{X}}(\boldsymbol{x})$. If $h_{\boldsymbol{X}}(\boldsymbol{x})$ deviates significantly from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$, one needs to start with a very large number of samples from $h_{\boldsymbol{X}}(\boldsymbol{x})$ to obtain approximate samples from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ of acceptable quality.

6.4.3 Nonparametric (Kernel) IS density of T

Construction Kernel IS density Estimator

From Proposition 6.4.1, $g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$ defined in (6.9) is the optimal proposal density of $\boldsymbol{T}$ in the CE method. In this section, we construct a kernel density estimate [46, 109, 108] of $g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$. Let $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ be a p-variate kernel function (see [108, p. 153]), let H denote a (nonsingular) $p \times p$ bandwidth matrix with determinant $\det(H)$, and define $\mathcal{K}_H(\boldsymbol{t}) = \frac{1}{\det(H)}\mathcal{K}\left(H^{-1}\boldsymbol{t}\right)$. See [47, Table 3.1] for a list of common univariate kernel functions and [108, p. 153] for the conditions that $\mathcal{K}$ needs to satisfy to be a proper multivariate kernel.

If we have M independent design samples $(\boldsymbol{T}_1^*, \ldots, \boldsymbol{T}_M^*)$ from $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$, we can construct a nonparametric density estimate of $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$ as

$$g_{\boldsymbol{T}}^{\mathrm{KD}}(\boldsymbol{t}) = \frac{1}{M} \sum_{i=1}^{M} \mathcal{K}_H\left(\boldsymbol{T}_i^* - \boldsymbol{t}\right). \tag{6.18}$$

By Proposition 6.4.2, if we have samples from $g_{\boldsymbol{X}}^*(\boldsymbol{x})$, we can construct the nonparametric proposal density (6.18). Of course, sampling from $g_{\boldsymbol{X}}^*(\boldsymbol{x})$ is infeasible and the estimate (6.18) is not attainable. If we instead have M independent design samples $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M) \overset{\text{ind.}}{\sim} h_{\boldsymbol{X}}(\boldsymbol{x})$, we can compute $T_i = T(\boldsymbol{X}_i)$, $i = 1, \ldots, M$ and construct the IS version of (6.18) as

$$g_{\boldsymbol{T}}^{\mathrm{KD}}(\boldsymbol{t}) = c^{\mathrm{KD}} \sum_{i=1}^{M} \mathcal{K}_H\left(\boldsymbol{T}_i - \boldsymbol{t}\right) \frac{|\Psi(\boldsymbol{X}_i)| f_{\boldsymbol{X}}(\boldsymbol{X}_i)}{h_{\boldsymbol{X}}(\boldsymbol{X}_i)} = \sum_{i=1}^{M} \mathcal{K}_H\left(\boldsymbol{T}_i - \boldsymbol{t}\right) w_i, \tag{6.19}$$

where KD stands for kernel density, $c^{\mathrm{KD}} = \left(\sum_{i=1}^{M} \frac{|\Psi(\boldsymbol{X}_i)| f_{\boldsymbol{X}}(\boldsymbol{X}_i)}{h_{\boldsymbol{X}}(\boldsymbol{X}_i)}\right)^{-1}$ is the normalizing constant, and $w_i = c^{\mathrm{KD}} \frac{|\Psi(\boldsymbol{X}_i)| f_{\boldsymbol{X}}(\boldsymbol{X}_i)}{h_{\boldsymbol{X}}(\boldsymbol{X}_i)}$.

If the kernel $\mathcal{K}(\cdot)$ is a pdf of some p-dimensional random vector, we can interpret (6.19) as a mixture of M distributions centred at $\boldsymbol{T}_i$'s where w_i is the mixture weight of ith component. In this chapter, we use the multivariate normal kernel

$$\mathcal{K}^{\mathrm{MVN}}(\boldsymbol{t}) = (2\pi)^{-\frac{p}{2}} \exp\left(-\frac{1}{2}\boldsymbol{t}'\boldsymbol{t}\right). \tag{6.20}$$

Noting that $\mathcal{K}_H^{\mathrm{MVN}}(\boldsymbol{T} - \boldsymbol{t})$ is the density of $\mathrm{MVN}(\boldsymbol{T}, H'H)$, one can draw a sample from (6.19) based on the MVN kernel (6.20) by first drawing an index $J \in \{1, \ldots, M\}$ according to the probabilities $\{w_1, \ldots, w_M\}$ and then generating a sample from $\mathrm{MVN}(\boldsymbol{T}_J, H'H)$.

Selection of the bandwidth matrix H

The performance of kernel methods, such as density estimation and regression, heavily depends on the choice of the bandwidth parameters. The performance of IS based on the kernel density (6.19) is also sensitive to the selection of bandwidth matrix H. Thus, we discuss our strategy of finding a good bandwidth matrix here.

Suppose that we have M design samples of $(X_1^*, \ldots, X_M^*)$ from $g_{\boldsymbol{X}}^*(\boldsymbol{x})$. We then want to choose H that accurately approximates the density of $T_i^* = T(\boldsymbol{X}_i^*)$, $i = 1, \ldots, M$ in light of Proposition 6.4.2. If the shape of $g_{\boldsymbol{T}}^{\text{KL}}(\boldsymbol{t})$ is well approximated by the density of $\text{MVN}(\eta_{\boldsymbol{T}}^{\text{KL}}, \Sigma_{\boldsymbol{T}}^{\text{KL}})$, Härdle et al. [47, p.73] suggest a generalization of Scott's rule [108, p.152] that sets

$$H = M^{-1/(p+4)}(\Sigma_{\boldsymbol{T}}^{\text{KL}})^{1/2}. \tag{6.21}$$

The estimate $\hat{\Sigma}_{\boldsymbol{T}}^{\text{KL}}$ of $\Sigma_{\boldsymbol{T}}^{\text{KL}}$ can be computed as in (6.16) based on the design samples. Since we essentially have M weighted samples from $g_X^*(\boldsymbol{x})$, we replace M in (6.21) with the effective sample size (see [84],[91, Ch.9]) $M_{\text{e}} = \left(\sum_{i=1}^{M} w_i\right)^2 / \sum_{i=1}^{m} w_i^2$, where w_i is defined in (6.19) and set

$$H = M_e^{-1/(p+4)}(\hat{\Sigma}_{\boldsymbol{T}}^{\text{KL}})^{1/2}. \tag{6.22}$$

If the shape of $g_{\boldsymbol{T}}^{\text{KL}}(\boldsymbol{t})$ is not elliptical, the covariance part of H does not carry much information. In this case, setting the off-diagonal entries of H to 0 may be a better option.

Computational Aspect

The bottleneck of using the nonparametric density (6.19) as a proposal distribution is the high computational cost associated with evaluating the density. If M design samples are used to construct a nonparametric IS distribution and n samples are drawn from this distribution to construct an IS estimator, the nonparametric proposal density must be evaluated n times in order to computes the IS weights, costing $O(Mn)$ operations under naive implementation. Hence, when M and n are moderately large, the computation of the IS weights becomes prohibitively slow, hindering the variance reduction achieved by nonparametric IS meaningless. Thus, techniques that speed up the calculation of kernel density should be employed for IS that uses kernel density estimations, such as the fast Gaussian transformation [32, 122].

6.5 Simulation Study: Pricing an Arithmetic Rainbow Asian Option

In this section, we apply multi-index IS to the pricing of arithmetic rainbow Asian options whose payoffs are essentially the maximum of two arithmetic Asian payoffs. Efficient pricing of those options are investigated by Peng and Peng [94] where the geometric rainbow Asian options are used as CV to reduce the variance of estimating the arithmetic option prices. The success of the CV scheme suggests that the arithmetic payoffs have a multi-index, or more specifically double-index, structure so we can test whether or not our proposed multi-index IS method performs well when the problem indeed has the desired multi-index structure. We compare the efficiency of multi-index IS to that of single-index IS and IS without any dimension reduction features.

6.5.1 Problem Formulation

Suppose that the price of Stock A and Stock B under the risk-neutral measure follow correlated geometric Brownian motions

$$dS_t^A = rS_t^A dt + \sigma_A S_t^A dW_t^A,$$
$$dS_t^B = rS_t^B dt + \sigma_B S_t^B dW_t^B,$$

where S_t^A and S_t^B are the price of Stock A and Stock B at time t, respectively, r is the risk-free rate, σ_A and σ_B are the respective volatility parameters, and W_t^A and W_t^B are correlated Brownian motions with the correlation coefficient ρ such that $dW_t^A \cdot dW_t^B = \rho dt$. Fix $T \in [0, \infty)$ and $d \in \mathbb{N}$. Let $\Delta t = T/d$ and $t_j = j\Delta t$ for $j = 1, \dots, d$. Given (S_0^A, S_0^B), the price of Stock A and Stock B at time t_j, $j = 1, \dots, d$ can be recursively expressed as

$$S_{t_j}^A = S_{t_{j-1}}^A \exp\left((r - \sigma_A^2/2)\Delta t + \sigma_A \sqrt{\Delta t} Z_j^{(1)}\right),$$
$$S_{t_j}^B = S_{t_{j-1}}^B \exp\left((r - \sigma_B^2/2)\Delta t + \sigma_B \sqrt{\Delta t}\left(\rho Z_j^{(1)} + \sqrt{1-\rho^2} Z_j^{(2)}\right)\right),$$

where $Z_1^{(1)}, \dots, Z_d^{(1)}, Z_1^{(2)}, \dots, Z_d^{(2)} \overset{\text{ind.}}{\sim} N(0,1)$. Let $S_a^A = \frac{1}{d}\sum_{j=1}^d S_{t_j}^A$ and $S_a^B = \frac{1}{d}\sum_{j=1}^d S_{t_j}^B$ denote the arithmetic averages of the prices of Stock A and Stock B, respectively, observed at time t_j, $j = 1, \ldots, d$. The payoff of an arithmetic rainbow Asian option with strike price K is $\max(\max(S_a^A, S_a^B) - K, 0)$. Then by risk-neutral pricing (see [37, pp. 27-30]), the price of the option is written as

$$c_a = \exp(-rT)\mathrm{E}[\max(\max(S_a^A, S_a^B) - K, 0)].$$

Let $\boldsymbol{Z} = (Z_1^{(1)}, \ldots, Z_d^{(1)}, Z_1^{(2)}, \ldots, Z_d^{(2)})$, then since S_a^A and S_a^B are functions of $\boldsymbol{Z}$, we can write the payoff function as

$$\Psi(\boldsymbol{Z}) = \max(\max(S_a^A(\boldsymbol{Z}), S_a^B(\boldsymbol{Z})) - K, 0).$$

Note that the dimension of this problem is $2d$ and $\Psi(\boldsymbol{z}) \geq 0$ for all $\boldsymbol{z} \in \mathbb{R}^{2d}$. Since $\mathrm{E}[\max(\max(S_a^A, S_a^B) - K, 0)]$ does not have an analytical form, we use multi-index IS to estimate the price of this option. The parameters we consider are $S_0^A = S_0^B = 100$, $r = 0.05$, $\sigma_A = \sigma_B = 0.3$, $K = 120$, $\rho \in \{0, 0.5, 1\}$, and $d \in \{16, 64, 128\}$. Under these parameters, S_a^A and S_a^B are identically distributed.

6.5.2 Single vs Double-index Model for Rainbow Asian Payoff

Let $S_g^A = (\prod_{j=1}^d S_{t_j}^A)^{1/d}$ and $S_g^B = (\prod_{j=1}^d S_{t_j}^A)^{1/d}$ denote the geometric averages of the prices of Stock A and Stock B, respectively, observed at time t_j, $j = 1, \ldots, d$. The payoff of the geometric rainbow Asian option is $\max(\max(S_g^A, S_g^B) - K, 0)$. For $b = \frac{1+d}{2}(r - \sigma^2/2)\Delta t$, it is easy to show that

$$\begin{aligned} S_g^A &= \left(\prod_{j=1}^d S_{t_j}^A\right)^{1/d} = S_0^A \exp\left(b + \sigma\sqrt{\Delta t}\,\frac{1}{d}\sum_{j=1}^d (d-j+1)Z_j^{(1)}\right), \\ S_g^B &= \left(\prod_{j=1}^d S_{t_j}^A\right)^{1/d} = S_0^A \exp\left(b + \sigma\sqrt{\Delta t}\,\frac{1}{d}\sum_{j=1}^d (d-j+1)\{\rho Z_j^{(1)} + \sqrt{1-\rho^2}Z_j^{(2)}\}\right) \end{aligned}$$

Define $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{2d}$ as $\boldsymbol{\beta}_1 \propto (d, d-1, \ldots, 1, 0, \ldots, 0)'$ and $\boldsymbol{\beta}_2 \propto (0, \ldots, 0, d, d-1, \ldots, 1)'$ such that $\boldsymbol{\beta}_1'\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2'\boldsymbol{\beta}_2 = 1$ so $\boldsymbol{T} = (T_1, T_2)' = (\boldsymbol{\beta}_1^{'}\boldsymbol{Z}, \boldsymbol{\beta}_2^{'}\boldsymbol{Z})' \sim \mathrm{MVN}(\boldsymbol{0}, I_2)$. Notice that the geometric rainbow payoff depends on $\boldsymbol{Z}$ only through $\boldsymbol{T}$, that is, the geometric rainbow

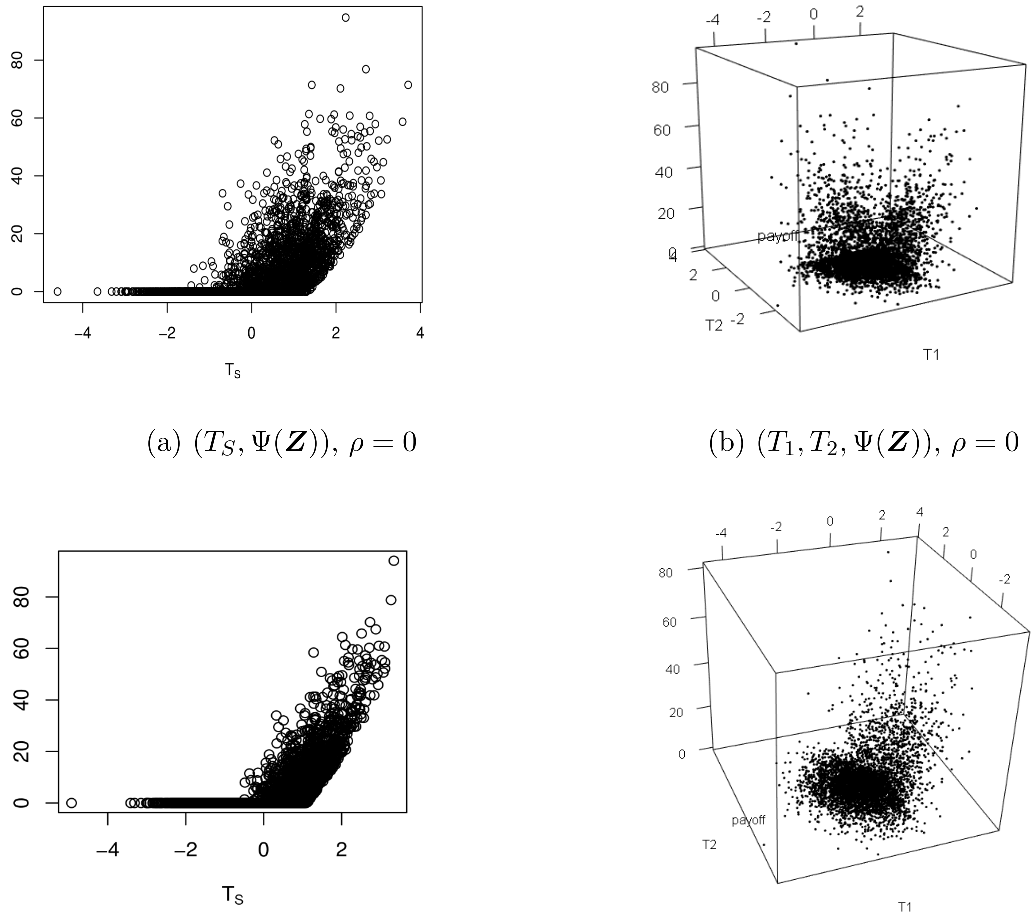
Asian option has a perfect double-index structure. We then expect that the arithmetic counterpart has a strong double-index structure, as implied by the analysis in [94].

The question is whether double-index IS is necessary. If the single-index model fits as well as the double-index one does, one can simply use single-index IS. If $\rho = 1$, we essentially have only one asset and the problem reduces to pricing an arithmetic Asian option on a single asset. As shown in Section 4.7.1, the problem has a nearly perfect single-index structure and single-index IS gives substantial variance reduction. The same argument holds for the case $\rho = -1$, but we mainly consider $\rho > 0$. It is not immediately clear how good the fit of the single-index model will be when $|\rho| \neq 1$, so we numerically compare the fit of the single-index model to that of the double-index one. The transformation variable that we use for single-index IS has the form $T_S = \boldsymbol{\beta}_S^{'}\boldsymbol{Z}$, where $\boldsymbol{\beta}_S^{'}\boldsymbol{\beta}_S = 1$ so that $T_S \sim N(0,1)$. Unlike the double-index case, the form of the direction vector $\boldsymbol{\beta}_S^{'}$ cannot be deduced analytically. Thus, we used the average derivative method of Stoker [114] and found that the optimal $\boldsymbol{\beta}_S$ has the form $\boldsymbol{\beta}_S = \rho\boldsymbol{\beta}_1 + \sqrt{1-\rho^2}\boldsymbol{\beta}_2$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are defined above for multi-index IS.

Figure 6.1 shows the scatter plot of $(T_S, \Psi(\boldsymbol{Z}))$ for the single-index model and $(T_1, T_2, \Psi(\boldsymbol{Z}))$ for the double-index model, for $\rho = 0$ and $\rho = 0.5$ based on 5,000 samples from the original distribution for $d = 16$. The figure shows that the double-index model gives a near perfect fit while single-index model fits very poorly, for both $\rho = 0$ and $\rho = 0.5$. These scatter plots suggest that multi-index IS will perform better than the single-index counterpart if $\rho = 0$ and $\rho = 0.5$. While additional plots are not included, we note that the fit of the single-index and double-index models are fairly constant for different values of d. For instance, the double-index model gives a near perfect fit even when $d = 128$.

We test parametric and nonparametric double-index IS based on the calibration methods developed in Section 6.4.2 and Section 6.4.3. For parametric double-index IS, it is important to choose an appropriate parametric family of the proposal density of $\boldsymbol{T}$. Thus, we use a visual aid to gauge the rough shape of $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$. Figure 6.2 shows the scatter plots of $(T_1, T_2, \Psi(\boldsymbol{X})f_{\boldsymbol{T}}(\boldsymbol{T}))$ for $\rho = 0$ and $\rho = 0.5$ based on 5,000 samples from the original distribution. Since $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t}) \propto \mathrm{E}[\Psi(\boldsymbol{Z}) \,|\, \boldsymbol{T}]f_{\boldsymbol{T}}(\boldsymbol{t})$, the plot shows an unnormalized, unsmoothed version of $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$. From the figure, $\Psi(\boldsymbol{Z})f_{\boldsymbol{T}}(\boldsymbol{T})$ is bimodal so $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t})$ should also be bimodal. Thus, we use a mixture of two bivariate normal distributions as a parametric family.

Figure 6.1: Scatter plot of index variables against payoff based on 5,000 observations



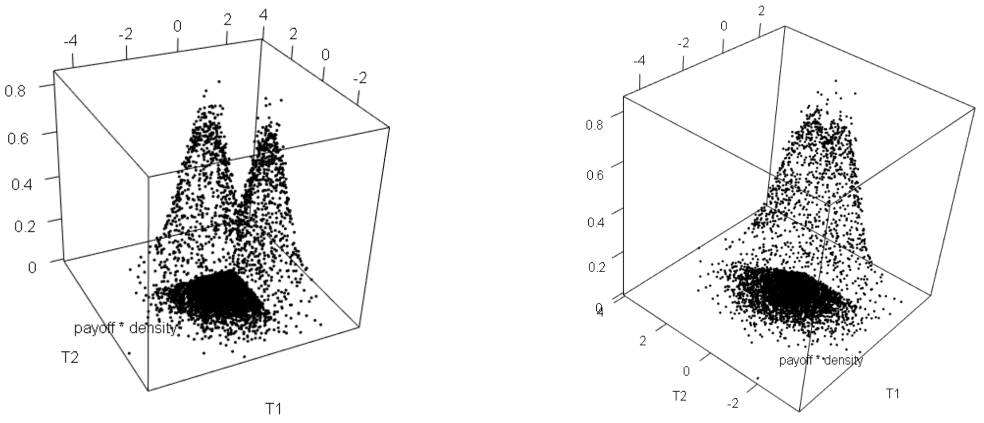
(a) $(T_S, \Psi(\boldsymbol{Z}))$, $\rho = 0$

(b) $(T_1, T_2, \Psi(\boldsymbol{Z}))$, $\rho = 0$

(c) $(T_S, \Psi(\boldsymbol{Z}))$, $\rho = 0.5$

(d) $(T_1, T_2, \Psi(\boldsymbol{Z}))$, $\rho = 0.5$

Figure 6.2: Scatter Plot of index variables against payoff based on 5,000 observations



(a) $(T_S, \Psi(\boldsymbol{Z})f_{\boldsymbol{T}}(\boldsymbol{T}))$, $\rho = 0$

(b) $(T_1, T_2, \Psi(\boldsymbol{Z})f_{\boldsymbol{T}}(\boldsymbol{T})))$, $\rho = 0.5$

6.5.3 Comparison of Single, Multi-index and Full IS

We compare the performance of single-index IS, double-index IS and IS without dimension reduction. Table 6.1 lists the IS techniques considered in this section and we explain each technique here. All IS techniques considered require samples from a design distribution to calibrate the proposal distribution and we use the original distribution as the design distribution, that is, $h_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{\boldsymbol{Z}}(\boldsymbol{z})$ for this simulation study. For single-index IS, referred to as "Single-index", we use the variance-minimizing density (4.8) as the proposal density of T_S. For double-index IS, we consider parametric density of $\boldsymbol{T}$ based on a mixture of two bivariate normal distributions, referred to as "DI-MixBVN", and then a nonparametric proposal density, referred to as "DI-NP", defined by (6.19). For DI-MixBVN, the parameters are calibrated using the EM algorithm as in [96] based on the approximate samples from $g^*_{\boldsymbol{X}}(\boldsymbol{x})$ obtained by applying the SIR method of Algorithm 9 to design samples. To assess the validity of the dimensionality problem of IS discussed in Section 2.2.3, we consider IS schemes that directly applies IS on $\boldsymbol{Z}$ without any dimension reduction technique. Note that directly applying IS on $\boldsymbol{Z}$ is equivalent to the multi-index IS with $p = 2 * d$ and $\boldsymbol{\beta} = I_{2*p}$ so we call such IS schemes as "full IS". For full IS, we consider MVN$(\boldsymbol{\mu}_{\boldsymbol{Z}}, \Sigma_{\boldsymbol{Z}})$, referred to as “Full-MVN1”, MVN($\boldsymbol{\mu_Z}, I_{2*d}$), referred to as “Full-MVN2”, and a nonparametric distribution, referred to as “Full-NP”, defined by (6.19) as proposal distributions of $\boldsymbol{Z}$, where $\boldsymbol{\mu_Z}$ and $\Sigma_{\boldsymbol{Z}}$ are calibrated as in (6.16). Tables 6.2, 6.3, and 6.4 respectively show the estimated price of the option, the variance reduction factors, and the ratio of computation time over plain MC, with and without QMC.

Table 6.1: Summary of Variance Reduction Techniques

Method	Description
Plain	Plain (Q)MC (no IS)
Single-index	Single-index IS with the optimal calibration (4.8).
DI-MixBVN	Multi-index IS with a mixture of two bivariate normal as a proposal for $\boldsymbol{T}$.
DI-NP	Multi-index IS with a noparametric proposal density (6.19) for $\boldsymbol{T}$.
Full-MVN1	IS with MVN($\boldsymbol{\mu_Z}, \Sigma_{\boldsymbol{Z}}$) as the proposal distribution of $\boldsymbol{Z}$, where $\boldsymbol{\mu_Z}, \Sigma_{\boldsymbol{Z}}$ are estimated as (6.16).
Full-MVN2	IS with MVN($\boldsymbol{\mu_Z}, I_{2*d}$) as the proposal distribution of $\boldsymbol{Z}$, where $\boldsymbol{\mu_Z}$ is estimated as (6.16).
Full-NP	IS with a nonparametric proposal density (6.19) for $\boldsymbol{Z}$.

In Table 6.2, the estimates with large estimation errors are underlined. We note that the support of all IS distributions considered are $\mathbb{R}^d$, so they give unbiased estimators in theory. As the estimates are based on a fairly large number of samples ($2^{14} * 30$ samples), estimates with large estimation errors suggest that the corresponding IS methods give unreliable estimates (see Section 2.2.3 for the notion of unreliable estimates and estimators). In general, Full-MVN1 and Full-NP give unreliable estimates unless $d = 16$ and $\rho = 1$. Among the three full IS methods considered, only Full-MVN2 appears to give reliable estimates. This observation is consistent with the result in Au and Beck [10] that altering the covariance matrix of the original distribution for IS often leads to unreliable estimates. The estimates based on single-index and double-index IS all seem reliable.

In Table 6.3, the VRFs corresponding to the estimates given by Full-NP for $d = 64$ and $d = 128$ are underlined. As the estimated prices given by Full-NP are clearly far from the true prices, the seemingly large VRFs do not reflect the actual performance of Full-NP because the mean squared errors are large. Single-index IS gives the greatest variance reduction when $\rho = 1$ but it struggles when $\rho = 0$ and $\rho = 0.5$, consistent with the poor fit of the single-index model for $\rho = 0$ and $\rho = 0.5$, as shown in Figure 6.1. In those two cases, DI-NP and DI-MixBVN achieve the greatest variance reduction for MC and QMC, respectively. It appears that DI-NP does not work well with QMC. Note that Full-MVN1 gives greater variance reduction than Full-MVN2 does when $d = 16$ and $\rho = 1$. This result indicates that having a flexible covariance matrix for the proposal density of $\boldsymbol{T}$ improves the performance of IS compared to when the covariance matrix is constrained to the identity matrix when the dimension of the problem is low. But the advantage disappears and the estimate becomes unreliable in high dimension. Overall, DI-NP and DI-MixBVN seem to be the best methods as they give reliable estimates with small variance.

From Table 6.4, we see that the nonparametric IS, DI-NP and Full-NP, take much more computation time than plain MC and parametric IS, hindering the effectiveness of these methods. As discussed in Section 6.4.3, the bottleneck of the nonparametric IS is the evaluation of the nonparametric proposal densities to compute the IS weights. Hence, the implementation of techniques to speed up the computation of nonparametric proposal densities is necessary to use nonparametric IS in practice. For IS techniques that do not use kernel density, the ones with a dimension reduction feature (Single-index and DI-MixBVN) appear to run faster than the ones without (Full-MVN1, Full-MVN2). When estimation reliability, variance reduction, and computation time are taken into account, DI-MixBVN seems to be the best method.

Table 6.2: The estimated price of the rainbow Asian option: $n = 2^{14}$, 30 replications

	Method	$d = 16$			$d = 64$			$d = 128$		
		$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$
MC	Plain	4.13	3.69	2.16	3.77	3.37	1.96	3.71	3.32	1.93
	Single-index	4.12	3.69	2.15	3.78	3.37	1.97	3.72	3.32	1.94
	DI-MVN	4.13	3.69	2.16	3.77	3.37	1.97	3.71	3.32	1.94
	DI-NP	4.13	3.69	2.16	3.77	3.37	1.97	3.71	3.31	1.94
	Full-MVN1	4.28	3.93	2.15	4.41	4.06	2.63	5.87	8.18	14.45
	Full-MVN2	4.13	3.68	2.16	3.79	3.38	1.97	3.69	3.31	1.94
	Full-NP	4.18	3.60	2.13	0.00	0.00	0.00	0.00	0.00	0.00
QMC	Plain	4.12	3.68	2.16	3.77	3.37	1.97	3.71	3.31	1.94
	Single-index	4.12	3.68	2.15	3.77	3.37	1.97	3.71	3.32	1.94
	DI-MVN	4.13	3.69	2.16	3.77	3.37	1.97	3.71	3.32	1.94
	DI-NP	4.13	3.68	2.16	3.77	3.37	1.97	3.71	3.32	1.94
	Full-MVN1	4.30	3.94	2.16	4.42	4.10	2.84	7.55	14.29	239.6
	Full-MVN2	4.13	3.68	2.16	3.77	3.37	1.97	3.71	3.32	1.94
	Full-NP	4.10	3.57	2.11	0.00	0.00	0.00	0.00	0.00	0.00

Table 6.3: The variance reduction factors of various methods: $n = 2^{14}$, 30 replications

	Method	$d = 16$			$d = 64$			$d = 128$		
		$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$
MC	Single-index	2.0E+00	4.8E+00	3.2E+02	2.4E+00	7.1E+00	5.0E+02	1.9E+00	4.5E+00	5.0E+02
	DI-MVN	1.4E+01	1.7E+01	4.3E+01	4.6E+01	3.1E+01	3.6E+01	2.2E+01	3.3E+01	4.6E+01
	DI-NP	1.9E+01	2.2E+01	6.9E+01	7.0E+01	4.5E+01	4.6E+01	2.5E+01	2.0E+01	9.6E+01
	Full-MVV1	3.1E+00	2.0E+00	3.4E+00	5.0E-02	4.0E-02	1.0E-02	3.4E-04	4.6E-05	6.1E-06
	Full-MVV2	1.6E+00	3.9E+00	1.7E+01	2.7E+00	6.7E+00	1.1E+01	1.7E+00	2.9E+00	1.1E+01
	Full-NP	1.2E-02	1.2E-02	3.2E-02	1.1E+05	9.3E+05	8.9E+06	1.4E+41	4.7E+39	1.4E+39
QMC	Plain	9.2E+00	5.7E+00	2.8E+01	8.0E+00	9.5E+00	5.2E+00	3.4E+00	6.9E+00	4.7E+00
	Single-index	1.8E+01	1.5E+01	9.4E+03	2.3E+01	3.9E+01	1.3E+03	9.4E+00	2.8E+01	2.7E+03
	DI-MVN	3.1E+02	4.4E+02	1.2E+03	1.2E+02	6.4E+01	1.1E+03	2.0E+02	5.4E+02	1.2E+02
	DI-NP	7.0E+01	3.8E+01	2.8E+02	1.3E+02	7.2E+01	8.5E+01	5.6E+01	2.7E+00	6.7E+01
	Full-MVN1	4.7E+00	2.4E+00	5.4E+00	5.6E-02	4.2E-02	5.4E-04	1.2E-04	1.4E-06	1.4E-09
	Full-MVN2	4.0E+00	4.0E+00	6.1E+01	1.6E+01	2.0E+01	2.1E+01	4.3E+00	8.3E+00	3.5E+01
	Full-NP	1.3E-02	1.5E-02	3.6E-02	3.3E+09	2.1E+08	7.8E+09	1.2E+39	8.5E+36	1.1E+39

Table 6.4: The computational time relative to plain MC: $n = 2^{14}$, 30 replications

	Method	$d = 16$			$d = 64$			$d = 128$		
		$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$	$\rho = 0$	$\rho = 0.5$	$\rho = 1$
MC	Single-index	1.2	1.3	1.3	1.8	1.7	1.7	2.1	2.2	2.3
	DI-MVN	1.3	1.3	1.3	1.8	1.7	1.7	2.1	2.2	2.2
	DI-NP	14.0	13.4	8.7	5.5	4.5	3.7	4.1	3.8	3.2
	Full-MVV1	1.4	1.5	1.5	2.6	2.4	2.7	4.2	4.2	4.1
	Full-MVV2	1.3	1.4	1.3	2.0	1.8	2.1	2.9	3.0	3.1
	Full-NP	54	50	31.	28	23	16	23	23	15
QMC	Plain	0.9	1.0	1.0	0.9	0.8	1.1	0.9	1.0	1.0
	Single-index	1.2	1.4	1.2	1.7	1.5	2.0	2.2	2.2	2.3
	DI-MVN	1.2	1.3	1.3	1.6	1.5	1.7	2.2	2.3	2.3
	DI-NP	14.0	13	8.6	5.3	4.5	4.1	4.1	4.0	3.2
	Full-MVV1	1.4	1.4	1.4	2.6	2.3	2.7	3.9	4.0	4.0
	Full-MVV2	1.2	1.3	1.3	2.0	1.8	2.0	2.8	2.9	2.9
	Full-NP	54	50	32	29	22	16	23	22	15

Chapter 7

Concluding Remarks and Future Research Directions

In this thesis, we explored IS techniques that exploit low-dimensional structures commonly observed in high-dimensional financial problems. In particular, we developed IS and SS schemes for three structural assumptions: the output takes a large value when at least one of the input variables is large; a single-index model where the output depends on the input variables mainly through some one-dimensional projection; and a multi-index model where the output depends on the input mainly through a set of linear combinations. We applied the techniques developed in this thesis to a variety of financial problems and our IS and SS schemes achieved substantial variance reduction in many cases.

The two major IS frameworks developed in this thesis are single-index and multi-index. Our treatment of multi-index IS is far from complete and we believe that we can improve multi-index IS in many ways. We discuss three of such ideas that we would like to explore in the future.

The first idea for improvement is the development of multi-index SIS. For the single-index IS developed in Chapter 4, it was fairly straightforward to combine that single-index IS with the SS idea and form single-index SIS. The stratified version of multi-index IS developed in Chapter 6 was never considered in this thesis. Single-index SIS essentially stratifies the domain of $\boldsymbol{X}$ along $T = T(\boldsymbol{X})$. As T is univariate, such stratification makes sure that the sampled T are well structured. The same idea is hard to generalize for multi-index IS as $\boldsymbol{T}$ is now multivariate. If $p = 3$ and we stratify each component of $\boldsymbol{T}$ into 10 strata, it generates 10^3 strata overall. If we look only at one-dimensional projections of $\boldsymbol{T}$, that samples will not be as well structured as the single-index case because we only have 10 strata in each dimension. A more clever approach would be to use a QMC point set when generating $\boldsymbol{T}$. This way, the low-dimensional projection of $\boldsymbol{T}$ will be well structured. For problems involving the pricing of path-dependent European options under the Black-Scholes framework, using a QMC point set to generate T is equivalent to using a QMC point set to stratify certain weighted sums of the path of the Brownian motion. Such idea has been explored by Kolkiewicz [69] and the numerical examples in [69] demonstrate substantial variance reduction when applied to Asian option pricing problems.

The second idea for improvement is modifying the calibration methods for multi-index IS developed in Section 6.4. Recall from Proposition 6.4.1 that the optimal proposal density for $\boldsymbol{T}$ for the CE calibration has the form $g_{\boldsymbol{T}}^{\mathrm{KL}}(\boldsymbol{t}) = c^* s(\boldsymbol{t}) f_{\boldsymbol{T}}(\boldsymbol{t})$. The calibration technique for the parametric proposal density $g_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta})$ developed in Section 6.4.2 searches for $\boldsymbol{\theta}^{\mathrm{KL}}$ such that

$$\begin{aligned}\boldsymbol{\theta}^{\mathrm{KL}} &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} \int_{\Omega_{\boldsymbol{X}}} \ln\left(g_{\boldsymbol{T}}(\boldsymbol{t}_{\boldsymbol{x}};\boldsymbol{\theta})\right) |\Psi(\boldsymbol{x})| f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, \mathrm{E}[\ln\left(g_{\boldsymbol{T}}(\boldsymbol{T};\boldsymbol{\theta})\right) |\Psi(\boldsymbol{X})|], \quad \boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}) \qquad (7.1)\\ &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, \mathrm{E}[\ln\left(g_{\boldsymbol{T}}(\boldsymbol{T};\boldsymbol{\theta})\right) s(\boldsymbol{T})], \quad \boldsymbol{T} \sim f_{\boldsymbol{T}}(\boldsymbol{t}). \qquad (7.2)\end{aligned}$$

We proposed to solve (7.1) instead of (7.2) as $s(\boldsymbol{T})$ does not need to be known or approximated when solving (7.1). If p (the dimension of $\boldsymbol{T}$) is large, it is very hard to nonparametrically approximate the conditional moment function $s(\boldsymbol{T})$. So the procedure that avoids approximating $s(\boldsymbol{T})$ is a better option. If p is 2 or 3, however, it is possible to nonparametrically approximate $s(\boldsymbol{T})$ with reasonable accuracy and solve (7.2) with the approximation. We can use the same idea when calibrating the nonparametric regression. That is, replace $|\Psi(\boldsymbol{X})|$ with $s(\boldsymbol{T})$. Our preliminary work finds that taking this route often gives a more reliable estimate of $\boldsymbol{\theta}^{\mathrm{KL}}$. We would like to investigate this point further in the future.

The third idea for improvement is to compare the techniques developed to estimate the direction matrix $\boldsymbol{\beta}$ for multi-index IS and examine which ones work best with multi-index IS. As mentioned in Section 6.2, Imai and Tan [58] developed an estimation procedure for $\boldsymbol{\beta}$ as a dimension technique for QMC. On the other hand, a variety of estimation procedures for $\boldsymbol{\beta}$ such as sliced inverse regression [78] and sliced average variance estimates [25] have been proposed for sufficient dimension reduction. We did not apply these techniques to estimate $\boldsymbol{\beta}$ for rainbow Asian option pricing in Section 6.5. Rather, we deduced the form of $\boldsymbol{\beta}$ analytically. Since such analytical form for $\boldsymbol{\beta}$ is rarely available, we would like to implement these methods and estimate $\boldsymbol{\beta}$ for financial applications and compare and see which estimation techniques of $\boldsymbol{\beta}$ work best with multi-index IS. Also, it would be interesting to investigate whether the techniques developed for sufficient dimension reduction can be used as dimension reduction techniques for QMC.

References

[1] K. Aas and D.H. Haff. The Generalized Hyperbolic Skew Students t-distribution. *Journal of Financial Econometrics*, 4(2):275–309, 2006.

[2] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Munctions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1965.

[3] P.A. Acworth, M. Broadie, and P. Glasserman. A Comparison of Some Monte Carlo and Quasi Monte Carlo Techniques for Option Pricing. In *Monte Carlo and Quasi-Monte Carlo Methods 1996*, pages 1–18. Springer, 1998.

[4] K.P. Adragni and R.D. Cook. Sufficient Dimension Reduction and Prediction in Regression. *Phil. Trans.: Math., Phys. and Eng. Sci.*, 367(1906):4385–4405, 2009.

[5] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[6] A. Ang and J. Chen. Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63(3):443–494, 2002.

[7] P. Arbenz, M. Cambou, and M. Hofert. An importance sampling algorithm for copula models in insurance. *arXiv preprint arXiv:1403.4291*, 2015.

[8] P. Arbenz, M. Cambou, M. Hofert, C. Lemieux, and Y. Taniguchi. Importance Sampling and Stratification for Copula Models. In *Contemporary Computational Mathematics – a celebration of the 80th birthday of Ian Sloan*. Springer, 2018. to appear.

[9] S. Asmussen and P.W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer-Verlag, 2007.

[10] S.K. Au and J.L. Beck. Important sampling in high dimensions. *Structural Safety*, 25(2):139–163, 2003.

[11] K. Banachewicz and A. Vaart. Tail dependence of skewed grouped t-distributions. *Statistics & Probability Letters*, 78(15):2388–2399, 2008.

[12] A. Bassamboo, S. Juneja, and A. Zeevi. Portfolio Credit Risk with Extremal Dependence: Asymptotic Analysis and Efficient Simulation. *Operations Research*, 56(3):593–606, 2008.

[13] P. Boyle, M. Broadie, and P. Glasserman. Monte Carlo methods for security pricing. *Journal of Economic Dynamics and Control*, 21(8):1267 – 1321, 1997.

[14] P.P. Boyle. Options: A Monte Carlo approach. *Journal of Financial Economics*, 4(3):323–338, 1977.

[15] M. Broadie, Y. Du, and C.C Moallemi. Efficient Risk Estimation via Nested Sequential Simulation. *Management Science*, 57(6):1172–1194, 2011.

[16] M. Broadie and P. Glasserman. Pricing American-style securities using simulation. *J. of Economic Dynamics and Control*, 21(8–9):1323–1352, 1997.

[17] R.P. Browne and P.D. McNicholas. A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2):176–198, 2015.

[18] R.E Caflisch, W. Morokoff, and A. Owen. Valuation of Mortgage Backed Securities Using Brownian Bridges to Reduce Effective Dimension. *Journal of Computational Finance*, 1:27–46, 1997.

[19] M. Cambou, M. Hofert, and C. Lemieux. Quasi-random numbers for copula models. *Statistics and Computing*, 27(5):1307–1329, 2017.

[20] 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.

[21] W. G. Cochran. *Sampling Techniques.* Wiley, 3rd edition, 2005.

[22] R.D. Cook. *Regression Graphics.* Wiley, 1998.

[23] R.D. Cook and L. Forzani. Principal Fitted Components for Dimension Reduction in Regression. *Statistical Science*, 23(4):485–501, 11 2008.

[24] R.D. Cook and L. Forzani. Likelihood-Based Sufficient Dimension Reduction. *J. of the American Statistical Association*, 104(485):197–208, 2009.

[25] R.D. Cook and S. Weisberg. Sliced Inverse Regression for Dimension Reduction: Comment. *J. of the American Statistical Association*, 86(414):328–332, 1991.

[26] P.T. De Boer, D.P. Kroese, S. Mannor, and R.Y. Rubinstein. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, 134(1):19–67, 2005.

[27] L. De Haan and A. Ferreira. *Extreme Value Theory: An Introduction.* Springer Science & Business Media, 2007.

[28] S. Demarta and A.J. McNeil. The t Copula and Related Copulas. *International Statistical Review*, 73(1):111–129, 2005.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[30] G. Derflinger, W. Hörmann, and J. Leydold. Random Variate Generation by Numerical Inversion when Only the Density is Known. *ACM Trans. Model. Comput. Simul.*, 20(4):18:1–18:25, 2010.

[31] J. Dick and F. Pillichshammer. *Digital nets and sequences: Discrepancy Theory and Quasi-Monte Carlo Integration.* Cambridge University Press, 2010.

[32] A. Elgammal, R. Duraiswami, and L.S. Davis. Efficient Kernel Density Estimation Using the Fast Gauss Transform with Applications to Color Modeling and Tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(11):1499–1504, 2003.

[33] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[34] A. Ghalanos. *rugarch: Univariate GARCH models*, 2015. R package version 1.3-6.

[35] A. Ghalanos. *spd: Semi-Parametric Distribution*, 2015. R package version 2.0-1.

[36] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC Press, 1995.

[37] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

[38] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Asymptotically Optimal Importance Sampling and Stratification for Pricing Path-Dependent Options. *Mathematical finance*, 9(2):117–152, 1999.

[39] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Variance Reduction Techniques for Estimating Value-at-Risk. *Management Science*, 46(10):1349–1364, 2000.

[40] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Portfolio Value-at-Risk with Heavy-Tailed Risk Factors. *Mathematical Finance*, 12(3):239–269, 2002.

[41] P. Glasserman and J. Li. Importance Sampling for Portfolio Credit Risk. *Management Science*, 51(11):1643–1656, 2005.

[42] P.W. Glynn and R. Szechtman. Some New Perspectives on the Method of Control Variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 27–49. Springer, 2002.

[43] M. Gordy and S. Juneja. Efficient Simulation for Risk Measurement in Portfolio of CDOs. In *Proceedings of the 2006 Winter Simulation Conference*, pages 749–756. IEEE, 2002.

[44] M.B. Gordy and Sandeep J. Nested Simulation in Portfolio Risk Measurement. *Management Science*, 56(10):1833–1848, 2010.

[45] P. Hall. On the Rate of Convergence of Normal Extremes. *Journal of Applied Probability*, 16(2):433–439, 1979.

[46] W. Hardle, P. Hall, and H. Ichimura. Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, 21(1):157–178, 1993.

[47] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, 2004.

[48] W.A. Harris and T.N. Helvig. Marginal and conditional distributions of singular distributions. *Publications of the Research Institute for Mathematical Sciences, Kyoto University. Ser. A*, 1(2):199–204, 1965.

[49] T. Hesterberg. Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics*, 37(2):185–194, 1995.

[50] S.L. Heston. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6(2):327–343, 1993.

[51] E. Hlawka. Funktionen von beschränkter Variation in der Theorie der Gleichverteilung. *Annali di Matematica Pura ed Applicata*, 54(1):325–333, 1961.

[52] M. Hofert, I. Kojadinovic, M. Maechler, and J. Yan. *copula: Multivariate Dependence with Copulas*, 2016. R package version 0.999-15.

[53] M. Hofert and M. Mächler. Nested Archimedean copulas meet R: The nacopula package. *Journal of Statistical Software*, 39(9):1–20, 2011.

[54] M. Hofert and M. Scherer. CDO pricing with nested Archimedean copulas. *Quantitative Finance*, 11(5):775–787, 2011.

[55] W. Hörmann and J. Leydold. Continuous random variate generation by fast numerical inversion. *ACM Trans. Model. Comput. Simul.*, 13(4):347–362, 2003.

[56] W. Hu and A. Kercheval. Risk Management with Generalized Hyperbolic Distributions. In *Proceedings of the Fourth IASTED International Conference on Financial Engineering and Applications*, pages 19–24, 2007.

[57] H. Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

[58] J. Imai and K.S. Tan. A General Dimension Reduction Technique For Derivative Pricing. *Journal of Computational Finance*, 10:129–155, 2006.

[59] H. Joe and P. Sang. Multivariate models for dependent clusters of variables with conditional independence given aggregation variables. *Computational Statistics and Data Analysis*, 97:114–132, 2016.

[60] S. Joe and F.Y. Kuo. Constructing Sobol Sequences with Better Two-Dimensional Projections. *SIAM J. Sci. Comput.*, 30(5):2635–2654, 2008.

[61] E. Jondeau and M. Rockinger. The Copula-GARCH model of conditional dependencies: An international stock market application. *Journal of International Money and Finance*, 25(5):827–853, 2006.

[62] C. Joy, P.P. Boyle, and K.S. Tan. Quasi-Monte Carlo Methods in Numerical Finance. *Management Science*, 42(6):926–938, 1996.

[63] H. Kahn and A.W. Marshall. Methods of reducing sample size in monte carlo computations. *J. of the Operations Research Society of America*, 1(5):263–278, 1953.

[64] S. Karlin and H.M. Taylor. *A First Course in Stochastic Processes, Second Edition.* Academic Press, 2nd edition, 4 1975.

[65] A.F. Karr. *Probability.* Springer, 1993.

[66] L.S. Katafygiotis and K.M. Zuev. Geometric insight into the challenges of solving high-dimensional reliability problems. *Probabilistic Engineering Mechanics*, 23(2–3):208–218, 2008.

[67] C.H. Kimberling. A Probabilistic Interpretation of Complete Monotonicity. *aequationes mathematicae*, 10(2-3):152–164, 1974.

[68] E. Kole, K. Koedijk, and M. Verbeek. Selecting copulas for risk management. *Journal of Banking & Finance*, 31(8):2405–2423, 2007.

[69] A.W. Kolkiewicz. Efficient Monte Carlo simulation for integral functionals of Brownian motion. *Journal of Complexity*, 30:255–278, 2014.

[70] L. Koralov and Y.G. Sinai. *Theory of Probability and Random Processes.* Springer, 2nd edition, 2007.

[71] S. Kullback. *Information Theory and Statistics.* Dover Publications, 1997.

[72] T.O. Kvalseth. Cautionary Note about R^2. *The American Statistician*, 39(4):279–285, 1985.

[73] H. Lan, B.L. Nelson, and J. Staum. A Confidence Interval Procedure for Expected Shortfall Risk Measurement via Two-Level Simulation. *Operations Research*, 58(5):1481–1490, 2010.

[74] P. L'Ecuyer and C. Lemieux. Variance Reduction via Lattice Rules. *Management Science*, 46(9):1214–1235, 2000.

[75] C. Lemieux. *Monte Carlo and Quasi-Monte-Carlo Sampling.* Springer, 2009.

[76] C. Lemieux and A.B. Owen. Quasi-Regression and the Relative Importance of the ANOVA Components of a Function. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 331–344. Springer, 2002.

[77] V. Lesnevski, B.L. Nelson, and J. Staum. Simulation of Coherent Risk Measures Based on Generalized Scenarios. *Management Science*, 53(11):1756–1769, 2007.

[78] K.-C. Li. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[79] M. Loéve. *Probability Theory.* Van Nostrand, 1963.

[80] F.A. Longstaff and E.S. Schwartz. Valuing American Options by Simulation: A Simple Least-Squares Approach. *The Review of Financial Studies*, 14(1):113–147, 2001.

[81] D. Luethi and W. Breymann. *ghyp: A package on the generalized hyperbolic distribution and its special cases*, 2013. R package version 1.5.6.

[82] Y.P. Mack and B.W. Silverman. Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(3):405–415, 1982.

[83] A.W. Marshall and I. Olkin. Families of Multivariate Distributions. *Journal of the American Statistical Association*, 83(403):834–841, 1988.

[84] L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[85] D.L. McLeish. Bounded Relative Error Importance Sampling and Rare Event Simulation. *ASTIN Bulletin*, 40(1):377–398, 2010.

[86] D.L. Mcleish and Z. Men. Extreme Value Importance Sampling for Rare Event Risk Measurement. *Innovations in Quantitative Risk Management Springer Proceedings in Mathematics and Statistics*, pages 317–335, 2015.

[87] A.J. McNeil and R. Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3-4):271–300, 2000.

[88] A.J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton university press, 2015.

[89] M.A. Milevsky and S.E. Posner. A Closed-Form Approximation for Valuing Basket Options. *The Journal of Derivatives*, 5(4):54–61, 1998.

[90] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods.* SIAM, 1992.

[91] A.B. Owen. *Monte Carlo theory, methods and examples.* 2013.

[92] S.H. Paskov and J.F. Traub. Faster Valuation of Financial Derivatives. *J. of Portfolio Management*, 22:113–120, 1995.

[93] A.J. Patton. Modelling Asymmetric Exchange Rate Dependence. *International Economic Review*, 47(2):527–556, 2006.

[94] B. Peng and F. Peng. Pricing Rainbow Asian Options. *SETP*, 29(11):76–83, 2009.

[95] J.L. Powell, J.H. Stock, and T.M. Stoker. Semiparametric Estimation of Index Coefficients. *Econometrica*, pages 1403–1430, 1989.

[96] R.A. Redner and H.F. Walker. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Rev.*, 26(2):195–239, 1984.

[97] C.H. Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183, 1967.

[98] B. Rémillard. Goodness-of-fit tests for copulas of multivariate time series. *Econometrics*, 5(1):13, 2017.

[99] C. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 2nd edition, 2004.

[100] J.C. Rodriguez. Measuring financial contagion: A Copula approach. *Journal of Empirical Finance*, 14(3):401–423, 2007.

[101] D.B. Rubin. Using the SIR Algorithm to Simulate Posterior Distributions. In *Bayesian Statistics*, pages 395–402. Oxford University Press, 3rd edition, 1988.

[102] R.Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

[103] R.Y. Rubinstein and D.P. Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning.* Springer Science & Business Media, 2013.

[104] P. Ruckdeschel, M. Kohl, T. Stabla, and F. Camphausen. S4 classes for distributions. *R News*, 6(2):2–6, 2006.

[105] L. Rűeschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009.

[106] H. Sak, W. Hörmann, and J. Leydold. Efficient risk simulations for linear asset portfolios in the t-copula model. *European Journal of Operational Research*, 202(3):802–809, 2010.

[107] G.I. Schuëller, H.J. Pradlwarter, and P.S. Koutsourelakis. A critical appraisal of reliability estimation procedures for high dimensions. *Probabilistic Engineering Mechanics*, 19(4):463–474, 2004.

[108] D.W. Scott. *Multivariate Density Estimation.* Wiley, 1992.

[109] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* CRC Press, 1986.

[110] M. Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Université Paris 8, 1959.

[111] A.F. Smith and A.E. Gelfand. Bayesian Statistics Without Tears: A Sampling–Resampling Perspective. *The American Statistician*, 46(2):84–88, 1992.

[112] I.M. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comp. Math. and Math.l Phy.*, 7(4):86–112, 1967.

[113] I.M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, 55(1):271–280, 2001.

[114] T.M. Stoker. Consistent Estimation of Scaled Coefficients. *Econometrica*, 54(6):1461–1481, 1986.

[115] 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.

[116] D. Tasche. Capital Allocation to Business Units and Sub-Portfolios: the Euler Principle. In A. Resti, editor, *Pillar II in the New Basel Accord: The Challenge of Economic Capital*, pages 423–453. Risk Books, London, 2008.

[117] L. Wang, L.D. Brown, T.T. Cai, and M. Levine. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, pages 646–664, 2008.

[118] X. Wang. On the effects of dimension reduction techniques on some high-dimensional problems in finance. *Operations Research*, 54(6):1063–1078, 2006.

[119] X. Wang and K.-T. Fang. The effective dimension and quasi-Monte Carlo integration. *Journal of Complexity*, 19(2):101–124, 2003.

[120] X. Wang and I.H Sloan. Why are high-dimensional finance problems often of low effective dimension? *SIAM Journal on Scientific Computing*, 27(1):159–183, 2005.

[121] X. Wang and I.H. Sloan. Quasi-Monte Carlo Methods in Financial Engineering: An Equivalence Principle and Dimension Reduction. *Operations Research*, 59(1):80–95, 2011.

[122] C. Yang, R. Duraiswami, N.A. Gumerov, and L. Davis. Improved fast gauss transform and efficient kernel density estimation. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 664–671, 2003.

Appendix

A Parameter estimation for GH skew-t copulas

The fitting procedure based on the EM algorithm [29] for the GH skew-t distribution is studied by Aas and Haff [1]. On the other hand, there is currently no work, to our knowledge, that addresses the parameter estimation for the GH skew-t copula. We explore the semi-parametric pseudo-MLE procedure that takes advantage of the EM algorithm in this section.

Suppose we have n samples of independent and identically distributed d-dimensional vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. We write the jth component of ith vector as $X_{i,j}$. Our goal is to find the MLE estimates of the parameters for the GH skew-t copula while the d marginals are estimated non-parametrically by the empirical CDFs. Such estimation procedures are called the semi-parametric pseudo-MLE estimation [33]. The empirical CDF of the jth component of $\boldsymbol{X}$ is

$$\hat{F}_j(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{X_{i,j} \leq x\}.$$

The pseudo-copula samples are constructed as

$$\hat{\boldsymbol{U}}_i = \left(\hat{F}_1(X_{i,1}), \ldots, \hat{F}_d(X_{i,d})\right)', \quad i = 1, \ldots, n.$$

As discussed in [28], the pseudo-observations are dependent even if the original samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent. Nevertheless, the pseudo-MLE method treats the pseudo-copula samples as independent and maximize the copula likelihood. The log-likelihood to maximize is

$$\log L(\nu, \boldsymbol{\gamma}, P; \hat{\boldsymbol{U}}_1, \dots, \hat{\boldsymbol{U}}_d) = \sum_{i=1}^{n} \log c^{t}_{\nu,P,\boldsymbol{\gamma}}(\hat{\boldsymbol{U}}_i), \tag{3}$$

and by (4.22) we can write

$$\log L(\nu, \boldsymbol{\gamma}, P; \hat{\boldsymbol{U}}_1, \dots, \hat{\boldsymbol{U}}_d) = \sum_{j=1}^{n} \log f_{st}(\hat{\boldsymbol{X}}_i; \nu, 0, P, \boldsymbol{\gamma}) - \sum_{j=1}^{n} \sum_{i=1}^{d} \log f_{st}(\hat{X}_{i,j}; \nu, 0, 1, \gamma_j), \tag{4}$$

where $\hat{X}_{i,j} = {F_{st}}^{-1}(\hat{U}_{i,j}; \nu, 0, 1, \gamma_i)$ for $i = 1, \dots, n$ and $j = 1, \dots, d$ and $\hat{\boldsymbol{X}}_i = (\hat{X}_{i,1}, \dots, \hat{X}_{i,d})$. We refer to the $\hat{\boldsymbol{X}}_i$'s as pseudo-skew-t samples. The fact that the marginal transformations $x_{i,j} = {F_{st}}^{-1}(\hat{U}_{i,j}; \nu, 0, 1, \gamma_j)$ depend on the parameters to be estimated makes this log-likelihood hard to maximize. Demarta and McNeil [28] argue that the maximization is not particularly easy in higher dimensions for symmetric t-copulas as we have to maximize over the space of the correlation matrix P. For a skew-t copula, we also need to estimate the skewness parameters $\boldsymbol{\gamma}$, making the maximization harder. The main idea of our approach is that if we fix ν and $\boldsymbol{\gamma}$, we can use the EM algorithm to efficiently estimate the correlation matrix P, turning the $1 + d(d-1)/2 + d$ dimensional optimization problem into a $1 + d$ dimensional problem. This is a large reduction in complexity of the problem, especially in high-dimensional cases.

Observe that the marginal transformations $\hat{X}_{i,j} = {F_{st}}^{-1}(\hat{U}_{i,j}; \nu, 0, 1, \gamma_j)$ do not depend on the correlation matrix P. That is, the second term of (4.22) is constant with respect to P. If we treat ν and $\boldsymbol{\gamma}$ as fixed, we only need to find P that maximizes the first term of (4.22). This is simply a problem of finding the MLE of the correlation matrix under a GH skew-t distribution, and we can use the EM alogrithm for this purpose. Then, we use some non-convex optimization solver to find the MLE of ν and γ. Since the EM algorithm does not require that the estimated covariance matrix be a correlation matrix, we have to scale the estimated covariance matrix to a correlation matrix. By the construction of the pseudo skew-t samples, the d marginals have a unit variance. Thus, the estimated covariance matrix will have its diagonal entries close to 1, so the scaling will not alter the estimated matrix much. That is, we will not lose much by scaling the covariance matrix in terms of log-likelihood.

B Proofs

Proof of Theorem 3.3.1. Due to Leibniz' integral rule, $dG(\boldsymbol{u}) = \int_0^1 dC_\lambda(\boldsymbol{u}) dF_\Lambda(\lambda)$. From the definition of C_λ, we can derive the differential

$$dC_\lambda(\boldsymbol{u}) = \begin{cases} 0, & \boldsymbol{u} \in [0,\lambda]^d, \\ \frac{dC(\boldsymbol{u})}{1-C(\lambda\mathbf{1})}, & \text{otherwise.} \end{cases}$$

Using both identities, we obtain

$$dG(\boldsymbol{u}) = dC(\boldsymbol{u}) \int_0^1 \frac{\mathbb{1}_{\{\lambda \leq \max\{u_1,\dots,u_d\}\}}}{1-C(\lambda\mathbf{1})} \, dF_\Lambda(\lambda),$$

leading to the desired result. □

Proof of Proposition 3.3.3. We sample $(E_1,\dots,E_d,V) \,|\, (E_{(1)} < \gamma V)$ using conditional distribution sampling. That is, we first sample $(E_{(1)},V) \,|\, (E_{(1)} < \gamma V)$, which is the Step 1 of Algorithm 3. Given the $(E_{(1)},V)$ drawn, we then want to sample $(E_1,\dots,E_d) \,|\, (E_{(1)} < \gamma V, E_{(1)}, V)$ which is equivalent to sampling $(E_1,\dots,E_d) \,|\, E_{(1)}$ and this is the Step 2 of the algorithm. □

Proof. Proof of Proposition 3.3.4 First, consider the case where $x_j = x$ for some $j = 1,\dots,k-1$. Without loss of generality assume that $x_1 = x$, i.e., $E_1 = E_{(1)}$. So we want to find $\mathbb{P}(E_k \leq x_k \,|\, E_1 = x_1,\dots,E_{k-1} = x_{k-1}, E_{(1)} = E_1 = x)$. From (3.7), the conditional distribution of E_k is $x + \text{Exp}(1)$. So the above probability equals

$$\mathbb{P}(E_k \leq x_k \,|\, E_1 = x_1,\dots,E_{k-1} = x_{k-1}, E_{(1)} = x) = 1 - e^{-(x_k - x)}. \tag{5}$$

Next, we consider the case $x_j \neq x$ for all $j = 1,\dots,k-1$. This means that $E_j = E_{(1)}$ for some $j = k,\dots,d$. Since all E_j are iid, there is a $\frac{1}{d-k+1}$ probability that $E_k = E_{(1)}$. In such a case $E_k = x$ with probability 1 as we are given $E_{(1)} = x$. Suppose $E_k \neq E_{(1)}$, which occurs with probability of $\frac{d-k}{d-k+1}$. Then we need to find the probability

$$\begin{aligned}
&\mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \dots, E_{k-1} = x_{k-1}, E_{(1)} = x, E_j \ne E_{(1)}, j = 1, \dots k) \\
&= \sum_{j=k+1}^{d} \frac{1}{d-k} \mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \dots, E_{k-1} = x_{k-1}, E_{(1)} = x, E_j = E_{(1)}) \\
&= \mathbb{P}(E_k \le x_k \,|\, E_1 = x_1, \dots, E_{k-1} = x_{k-1}, E_{(1)} = x, E_d = E_{(1)}) = 1 - e^{-(x_k - x)}.
\end{aligned}$$

The last equality again holds by (3.7) and the result follows. □

Proof of Proposition 3.5.1. Recall that the IS estimator with n samples is

$$\hat{\mu}_{\mathrm{IS},n} = \frac{1}{n} \sum_{i=1}^{n} \Psi_0(\boldsymbol{U}_i) w(\boldsymbol{U}_i), \quad \boldsymbol{U}_i \overset{\text{ind.}}{\sim} G \tag{6}$$

and the weight function is

$$w(\boldsymbol{u}) = \Big(\sum_{k=1}^{M} \frac{\mathbb{1}\{\lambda_k \le \max\{u_1, \dots, u_d\}\}}{1 - C(\lambda_k \mathbf{1})} q_k \Big)^{-1}. \tag{7}$$

Note from (7) that $w(\boldsymbol{u})$ is constant over each stratum $\Omega_C^{(k)}$ defined as in (3.9). Thus, for $\boldsymbol{u} \in \Omega_C^{(k)}$, we can define the stratum weight as

$$w_k = \left(\sum_{l=1}^{k} \frac{q_l}{1 - C_l} \right)^{-1}, \quad k = 1, \dots, M. \tag{8}$$

The second moment of $\Psi_0(\boldsymbol{U}) w(\boldsymbol{U})$ under the IS distribution is is

$$\begin{aligned}
\mathrm{E}_G[\Psi_0^2(\boldsymbol{U}) w^2(\boldsymbol{U})] &= \mathrm{E}_G[\Psi_0^2(\boldsymbol{U}) w(\boldsymbol{U})] = \sum_{k=1}^{M} p_k \mathrm{E}_C[\Psi_0^2(\boldsymbol{U}) w(\boldsymbol{U}) \,|\, \Omega_k] \\
&= \sum_{k=1}^{M} p_k \tilde{w}_k \mathrm{E}_C[\Psi_0^2(\boldsymbol{U}) \,|\, \Omega_k] = \sum_{k=1}^{M} p_k w_k \mu_k^{(2)} = \sum_{k=1}^{M} p_k \left(\sum_{l=1}^{k} \frac{1}{1 - C_l} q_l \right)^{-1} \mu_k^{(2)}.
\end{aligned}$$

The third equality holds because $\tilde{w}(t)$ is constant over each stratum. The last equality follows from (8). Then the variance of the importance sampling estimator based on n samples is

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n} \left(\sum_{k=1}^{M} p_k \left(\sum_{l=1}^{k} \frac{1}{1 - C_l} q_l \right)^{-1} \mu_k^{(2)} - \mu^2 \right).$$

□

Proof of Proposition 3.5.2. Since the variance expression (3.12) is convex in the q_k's, the minimization problem can be solved using the Lagrange multiplier method. Nevertheless, we simplify (3.12) so that minimization problem becomes easier. Let $\tilde{p}_k = P(\tilde{\boldsymbol{U}} \in \Omega_k)$, the stratum probability under the proposal distribution. Observe that

$$\begin{aligned}\tilde{p}_k &= \sum_{l=1}^{M} q_l \cdot \mathbb{P}_G(\tilde{\boldsymbol{U}} \in \Omega_k \,|\, \Lambda = \lambda_l) = \sum_{l=1}^{k} q_l \cdot \mathbb{P}_C(\boldsymbol{U} \in \Omega_k \,|\, \Lambda = \lambda_l) \\ &= \sum_{l=1}^{k} q_l \cdot \mathbb{P}_C(\boldsymbol{U} \in \Omega_k \,|\, T > \lambda_l) = \sum_{l=1}^{k} q_l \frac{p_k}{1 - C_l} = p_k \sum_{l=1}^{k} \frac{q_l}{1 - C_l}. \end{aligned} \tag{9}$$

By (8) and (9), we can write $w_k = \frac{p_k}{\tilde{p}_k}$. The stratum weight w_k is the ratio of the stratum probability under the original distribution to the one under the IS distribution. Substituting this expression into (3.12),

$$\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n}) = \frac{1}{n} \left(\sum_{k=1}^{M} \frac{p_k^2}{\tilde{p}_k} \mu_k^{(2)} - \mu^2 \right). \tag{10}$$

Using the Lagrange multiplier method, it is easy to show that the optimal $\tilde{p}_k$ for $k = 1, \ldots, M$ is

$$\tilde{p}_k^{opt} = \frac{p_k \sqrt{\mu_k^{(2)}}}{\sum\limits_{k=1}^{M} p_k \sqrt{\mu_k^{(2)}}}. \tag{11}$$

Note that this optimal choice of $\tilde{p}_k$'s resembles the Neyman allocation, the optimal allocation under stratified sampling. Using the relation $q_k = (1 - C_k)\left(\frac{\tilde{p}_k}{p_k} - \frac{\tilde{p}_{k-1}}{p_{k-1}}\right)$, (with $\frac{\tilde{p}_0}{p_0} = 0$) the optimal q_k has the form

$$q_k^{\mathrm{opt}} \propto (1 - C_k) \left(\sqrt{\mu_k^{(2)}} - \sqrt{\mu_{k-1}^{(2)}} \right), \text{ (with } \mu_0^{(2)} = 0). \tag{12}$$

The assumption that $\mu_1^{(2)} \leq \cdots \leq \mu_M^{(2)}$ ensures that $q_k^{\mathrm{opt}} \geq 0$ for $k = 1, \ldots, M$. □

Proof of Proposition 3.5.4. We have $\hat{\mu}^{\text{det}}_{\text{IS},n} = \frac{1}{n}\sum_{k=1}^{M}\sum_{j=1}^{nq_k} \Psi(\tilde{\boldsymbol{U}}_{ki})w(\tilde{\boldsymbol{U}}_{ki}), \quad \tilde{\boldsymbol{U}}_{ki} \overset{iid}{\sim} \boldsymbol{U}|\Lambda = \lambda_k$. Thus $\operatorname{Var}(\hat{\mu}^{\text{det}}_{\text{IS},n}) = \mathbb{E}\left[\operatorname{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \,|\, \Lambda)\right]/n + O(1/n^2)$ (term due to rounding nq_k). Since $\operatorname{Var}(\hat{\mu}_{\text{IS},n}) = \frac{1}{n}\operatorname{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}})$, we have $\operatorname{Var}(\hat{\mu}^{\text{det}}_{\text{IS},n}) \leq \operatorname{Var}(\hat{\mu}_{\text{IS},n})$ as long as n is large enough for the $O(1/n^2)$ term due to rounding to be smaller than $\operatorname{Var}(\mathbb{E}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}})|\Lambda))/n > 0$. As shown before, $\tilde{p}_k = \mathbb{P}(\tilde{\boldsymbol{U}} \in \Omega_k) = p_k \sum_{l=1}^{k} \frac{q_l}{1-C_l}$. Consider an SS estimator with $n_k = n\tilde{p}_k$. Then $\operatorname{Var}(\hat{\mu}_{\text{SS},n}) = \frac{1}{n}\sum_{k=1}^{M} \frac{p_k^2}{\tilde{p}_k}\sigma_k^2$. Also $\operatorname{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \,|\, \Lambda = \lambda_k) = \operatorname{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \,|\, T > \lambda_k) \geq \mathbb{E}[\operatorname{Var}(\Psi(\tilde{\boldsymbol{U}})w(\tilde{\boldsymbol{U}}) \,|\, T > \lambda_k, T \in \Omega_j)] = \sum_{j=k}^{M} \frac{p_j}{1-C_k} w_j^2 \sigma_j^2$. Then, using (8) and $w_k = p_k/\tilde{p}_k$ we get

$$\begin{aligned}
\operatorname{Var}(\hat{\mu}^{\text{det}}_{\text{IS},n}) &\geq \frac{1}{n}\sum_{k=1}^{M} q_k \sum_{j=k}^{M} \frac{p_j}{1-C_k} w_j^2 \sigma_j^2 = \frac{1}{n}\sum_{k=1}^{M} p_k w_k^2 \sigma_k^2 \sum_{j=1}^{k} \frac{q_j}{1-C_j} \\
&= \frac{1}{n}\sum_{k=1}^{M} p_k w_k \sigma_k^2 = \frac{1}{n}\sum_{k=1}^{M} \frac{p_k^2}{\tilde{p}_k}\sigma_k^2 = \operatorname{Var}(\hat{\mu}_{\text{SS},n}).\square
\end{aligned}$$

□

Proof of Proposition 4.3.2. Since $T \sim N(0,1)$ under the original distribution, we have under the proposal distribution that

$$(\boldsymbol{X} \,|\, T = t) \sim \text{MVN}\left(\boldsymbol{\beta}t, I_d - \boldsymbol{\beta}\boldsymbol{\beta}'\right) \tag{13}$$

by [48, Theorem 1]. For any $\boldsymbol{a} \in \mathbb{R}^d$, we have that

$$\begin{aligned}
&\mathrm{E}_g[\exp(\boldsymbol{a}'\boldsymbol{X})] = \mathrm{E}_g\left[\mathrm{E}_g[\exp(\boldsymbol{a}'\boldsymbol{X}) \,|\, T]\right] = \mathrm{E}_g\left[\exp\left(\boldsymbol{a}'\boldsymbol{\beta} \cdot T + \frac{1}{2}\boldsymbol{a}'\left(I_d - \boldsymbol{\beta}\boldsymbol{\beta}'\right)\boldsymbol{a}\right)\right] \\
&= \mathrm{E}_g\left[\exp\left(\boldsymbol{a}'\boldsymbol{\beta} \cdot T\right)\right] \cdot \exp\left(\frac{1}{2}\boldsymbol{a}'\left(I_d - \boldsymbol{\beta}\boldsymbol{\beta}'\right)\boldsymbol{a}\right) \\
&= \exp\left(\boldsymbol{a}'\boldsymbol{\beta}\sqrt{\boldsymbol{\eta}'\boldsymbol{\eta}} + \frac{1}{2}(\boldsymbol{a}'\boldsymbol{\beta})^2\right) \cdot \exp\left(\frac{1}{2}\boldsymbol{a}'\boldsymbol{a} - \frac{1}{2}(\boldsymbol{a}'\boldsymbol{\beta})^2\right) = \exp\left(\boldsymbol{a}'\boldsymbol{\eta} + \frac{1}{2}\boldsymbol{a}'\boldsymbol{a}\right)
\end{aligned}$$

where the third and fifth equalities follow from the moment generating function of the MVN distribution [37, p. 65]. Thus, $\boldsymbol{X} \sim \text{MVN}(\boldsymbol{\eta}, I_d)$ under the IS distribution by uniqueness of the moment generating function. □

Proof of Proposition 4.3.3. Firstly (4.6) follows from

$$\begin{aligned}\mathrm{E}_g[\hat{\mu}_{\mathrm{IS},n}] &= \mathrm{E}_g[\Psi(\boldsymbol{X})\tilde{w}(T)] = \mathrm{E}_g[\mu(T)\tilde{w}(T)] \\ &= \int_{A_t} m(t)\frac{f_T(t)}{g_T(t)} g_T(t)dt = \int_{A_t} m(t) f_T(t)dt \quad \text{and}\end{aligned}$$

$$\begin{aligned}n\mathrm{Var}_g(\hat{\mu}_{\mathrm{IS},n}) + \mu_{\mathrm{IS}}^2 &= \mathrm{E}_g[\Psi^2(\boldsymbol{X})\tilde{w}^2(T)] \\ &= \int_{A_t} m^{(2)}(t)\frac{f_T^2(t)}{g_T^2(t)} g_T(t)dt = \int_{A_t} m^{(2)}(t)\frac{f_T^2(t)}{g_T(t)}dt.\end{aligned}$$

The asymptotic normality of (4.7) follows from the central limit theorem (CLT) (p.190 of [65]). Now, we want to find the $g_T(t)$ that minimizes $\mathrm{Var}(\hat{\mu}_{\mathrm{IS},n})$ or equivalently $\mathrm{E}_g[m^{(2)}(T)\tilde{w}^2(T)]$ when $\Psi(\boldsymbol{x}) \geq 0$ or $\Psi(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ among the $g_T(t)$ that gives an unbiased estimator. The IS estimator is unbiased when $A_t \supseteq A_t^{ub} = \{t \in \Omega_T \mid m(t)f_T(t) \neq 0\}$. By the assumption on $\Psi(\boldsymbol{x})$, $\sqrt{\mu^{(2)}(t)}f_T(t) = 0$ for $t \notin A_t^{ub}$. Jensen's inequality gives

$$\begin{aligned}\mathrm{E}_g[m^{(2)}(T)w^2(T)] &\geq \left(\mathrm{E}_g\left[\sqrt{m^{(2)}(T)}\tilde{w}(T)\right]\right)^2 \\ &= \left(\int_{A_t} \sqrt{m^{(2)}(t)} f_T(t)dt\right)^2 = \left(\int_{t_{\inf}}^{t_{\sup}} \sqrt{m^{(2)}(t)} f_T(t)dt\right)^2, \qquad (14)\end{aligned}$$

where the last equality follows from the assumption that $g_T(t)$ is such that it gives an unbiased estimator and the inequality holds as equality only if $\sqrt{\mu^{(2)}(t)}\tilde{w}(t)$ is constant in t which occurs when $g_T(t) = g_T^{\mathrm{opt}}(t) \propto \sqrt{\mu^{(2)}(\lambda)}f_T(t)$. Note that the right hand side of the inequality of (14) is a constant independent of $g_T(t)$, so if some $g_T(t)$ achieves this lower bound, it gives an estimator with smallest variance. Since $g_T^{opt}(t)$ achieves the lower bound, $g_T^{\mathrm{opt}}(t)$ is optimal. □

Proof of Proposition 4.3.4. Observe that

$$\begin{aligned}\mathrm{E}[\hat{\mu}_{\mathrm{SIS},n}] &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_g[\Psi(\boldsymbol{X})\tilde{w}(T) \,|\, T \in \Omega_T^{(i)}] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_g[\mathrm{E}_g[\Psi(\boldsymbol{X})\tilde{w}(T) \,|\, T] \,|\, T \in \Omega_T^{(i)}] \\ &= \frac{1}{n}\sum_{i=1}^{n} n \int_{\lambda_i}^{\lambda_{i+1}} m(t)\frac{f_T(t)}{g_T(t)} g_T(t)dt = \int_{A_t} m(t) f_T(t)dt.\end{aligned}$$

The statement that $\mathrm{Var}(\hat{\mu}_{\mathrm{SIS},n}) = \sigma^2_{\mathrm{SIS}}/n + o(1/n)$ is a slight generalization of Lemma 4.1 of [38] in that stratification is combined with IS, but the it can be essentially proved in the same way. Let $\eta_n(t)$ denote the index of the stratum containing t. Then,

$$n\mathrm{Var}(\hat{\mu}_{\mathrm{SIS},n}) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{Var}_g\left(\Psi(\boldsymbol{X})\tilde{w}(T)\,|\,T \in \Omega_T^{(i)}\right) = \mathrm{E}_g\left[\mathrm{Var}_g\left(\Psi(\boldsymbol{X})\tilde{w}(T)\,|\,\eta_n(T)\right)\right].$$

Let $\xi = \mathrm{E}_g[\Psi(\boldsymbol{X})\tilde{w}(T)\,|\,T] = m(T)\tilde{w}(T)$ and define the sequence $\xi_n = \mathrm{E}_g[\xi\,|\,\eta_n(T)]$. Note that the σ-algebra generated by $\eta_n(T)$ forms an increasing family as n increases through a constant multiple of power of two. Observe that $\mathrm{E}_g[|\xi|] < \infty$ and $\sup_n \xi_n < \mathrm{E}_g[\Psi^2(Y)w^2(T)] = \mathrm{E}_g[m^{(2)}w^2(T)] < \infty$. Also, ξ_n is a martingale if n increases through a constant multiple of power of two as it is a Doob's Martingale Process [64, p. 246]. Then using the arguments similar to the proof of Lemma 4.1 of [38], we can show that $\mathrm{Var}_g(\hat{\mu}_{\mathrm{SIS},n}) = \sigma^2_{\mathrm{SIS}}/n + o(1)$.

We then prove the asymptotic normality of the SIS estimator (4.11) by showing that the Lyapunov condition [68, p. 134] holds. Let $m_i = \mathrm{E}_g[\Psi(\boldsymbol{X})\tilde{w}(T)\,|\,T \in \Omega_T^{(i)}]$ and $v_i^2 = \mathrm{Var}_g[\Psi(\boldsymbol{X})\tilde{w}(T)\,|\,T \in \Omega_T^{(i)}]$. It is easily checked that $\frac{1}{n}\sum_{j=1}^n m_j = \mu_{\mathrm{IS}}$ and $\frac{1}{n}\sum_{i=1}^n v_i^2 = \sigma^2_{\mathrm{SIS}} + o(1)$. Now, for all $1 \le i \le n$,

$$\begin{aligned}
&\mathrm{E}_g[|\Psi(\boldsymbol{X}_i)\tilde{w}(T_i) - m_i|^{2+\delta}] \le 2^{2+\delta}\left(\mathrm{E}_g[|\Psi(\boldsymbol{Y_i})\tilde{w}(T_i)|^{2+\delta}] + \mathrm{E}_g[|m_i|^{2+\delta}]\right)\\
&= 2^{2+\delta}\left(\mathrm{E}_g[|\Psi(\boldsymbol{X})\tilde{w}(T)|^{2+\delta}\,|\,T \in \Omega_T^{(i)}] + \mathrm{E}_g\left[\left|\mathrm{E}_g[\Psi(\boldsymbol{X})\tilde{w}(T)\,|\,T \in \Omega_T^{(i)}]\right|^{2+\delta}\right]\right)\\
&\le 2^{2+\delta}\left(\mathrm{E}_g[|\Psi(\boldsymbol{X})\tilde{w}(T)|^{2+\delta}\,|\,T \in \Omega_T^{(i)}] + \mathrm{E}_g[\mathrm{E}_g[|\ \Psi(\boldsymbol{X})\tilde{w}(T)\ |^{2+\delta}\,|\,T \in \Omega_T^{(i)}]]\right)\\
&= 2^{3+\delta}\mathrm{E}_g\left[|\Psi(\boldsymbol{X})\tilde{w}(T)|^{2+\delta}\,|\,T \in \Omega_T^{(i)}\right],
\end{aligned}$$

where the first inequality follows from the c_r-Inequality as in [79, p.155]. The Lyapunov condition is satisfied since

$$\begin{aligned}
&\frac{1}{(\sum_{j=1}^{n}\sigma_j^2)^{1+\delta/2}}\sum_{i=1}^{n}\mathrm{E}_g\left|\Psi(\boldsymbol{X}_i)\tilde{w}(T_i)-m_j\right|^{2+\delta}\\
&\leq \frac{2^{3+\delta}}{(\sum_{j=1}^{n}\sigma_j^2)^{1+\delta/2}}\sum_{i=1}^{n}\mathrm{E}_g[\left|\Psi(\boldsymbol{X})\tilde{w}(T)\right|^{2+\delta} \mid T\in\Omega_j]\\
&= \frac{2^{3+\delta}n}{(n\sigma_{\mathrm{SIS}}^2+o(n))^{1+\delta/2}}\mathrm{E}_g\left|\Psi(\boldsymbol{X})\tilde{w}(T)\right|^{2+\delta}\xrightarrow{n\to\infty}0
\end{aligned}$$

by the assumption that $\mathrm{E}_g\left|\Psi(\boldsymbol{X})\tilde{w}(T)\right|^{2+\delta}<\infty$. Then by Lyapunov Central Theorem [68, p. 134] and Slutsky's Theorem, we have that $\hat{\mu}_{\mathrm{SIS},n}\xrightarrow{D}N(\mu,\sigma_{\mathrm{SIS}}^2/n)$.

We want to find the optimal calibration for the SIS estimator among $g_T(t)$ that gives unbiased estimators when $\Psi(\boldsymbol{x})\geq 0$ or $\Psi(\boldsymbol{x})\leq 0\ \forall\,\boldsymbol{x}\in\Omega_{\boldsymbol{X}}$ and $\mathbb{P}_f(v^2(T)=0,\ m(T)\neq 0)=0$. Under these assumptions, the IS estimator is unbiased when $A_t\supseteq\{t\in\Omega_T\mid v(t)f_T(t)>0\}$. Ignoring the $o(1)$ term, Jensen's inequality gives that

$$\begin{aligned}
n\mathrm{Var}(\hat{\mu}_{\mathrm{SIS},n}) &= \mathrm{E}_g[v^2(T)\tilde{w}^2(T)]\geq\left(\mathrm{E}_g\left[v(T)\tilde{w}(T)\right]\right)^2\\
&=\left(\int_{A_t}v(T)f_T(t)dt\right)^2=\left(\int_{t_{\inf}}^{t_{\sup}}\sigma(T)f_T(t)dt\right)^2,
\end{aligned}\tag{15}$$

where the last equality holds from the unbiased estimator assumption and the inequality holds as equality only if $v(t)\tilde{w}(t)$ is constant in t which occurs when $g_T(t)=g_T^{\mathrm{opt}}(t)\propto v(t)f_T(t)$. By using the argument similar to the one for the proof of Proposition 4.3.3, it is easy to show that $g_T^{\mathrm{opt}}(t)$ givens an unbiased estimator with the smallest variance. Also, it is easy to see that if $\mathbb{P}_f(v^2(T)=0,\ m(T)\neq 0)>0$, the optimal calibration gives a biased estimator. □

Proof of Proposition 4.3.7. Note that by the construction of SIS estimators, the samples of T_i are ordered, that is, $T_1<T_2\cdots<T_n$. Since $T_i\overset{D}{=}G_T^{-1}(\frac{i-1+U_i}{n})$ where $U_i\overset{\text{ind.}}{\sim}U(0,1)$ for $i=1,\ldots,n$,

$$T_{i+1}-T_i=(G_T^{-1})'(\xi_i)\left(\frac{1+U_{i+1}-U_i}{n}\right)=\frac{1}{g_T(G_T^{-1}(\xi_i))}\left(\frac{1+U_{i+1}-U_i}{n}\right)=O(1/n)$$

for some ξ_i between T_{i+1} and T_i, which implies that for any continuously differentiable function a, $a(T_{i+1}) = a(T_i) + O(1/n)$. Then we have

$$\begin{aligned}
r_i^2 &= \left[(m(T_{i+1}) + \epsilon_{T_{i+1}}) - (m(T_i + \epsilon_{T_i})\right]^2 \\
&= (m(T_{i+1}) - m(T_i))^2 + (\epsilon_{T_{i+1}} - \epsilon_{T_i})^2 - 2\left(m(T_{i+1}) - m(T_i)\right)\left(\epsilon_{T_{i+1}} - \epsilon_{T_i}\right) \\
&= (\epsilon_{T_{i+1}} - \epsilon_{T_i})^2 - 2\left(m(T_{i+1}) - m(T_i)\right)\left(\epsilon_{T_{i+1}} - \epsilon_{T_i}\right) + O(1/n^2)
\end{aligned}$$

and so

$$\begin{aligned}
&\mathrm{E}_g[r_i^2 \tilde{w}^2(T_i)] = \mathrm{E}_g[\mathrm{E}_g[r_i^2 \tilde{w}(T_i) \,|\, T_i, T_{i+1}]] = \mathrm{E}_g[\tilde{w}^2(T_i)(v^2(T_i) + v^2(T_{i+1}))] + O(1/n^2) \\
&= 2\mathrm{E}_g[\tilde{w}^2(T_i) v^2(T_i)] + O(1/n),
\end{aligned}$$

which means that

$$\begin{aligned}
&\mathrm{E}_g[\hat{\sigma}^2_{\text{SIS}}] = \frac{1}{2(n-1)} \sum_{j=1}^{n-1} \mathrm{E}_g[r_i^2 \tilde{w}^2(T_i)] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_g[v^2(T)\tilde{w}^2(T) \,|\, T \in \Omega_T^{(i)}] + O(1/n) \\
&= \mathrm{E}_g[v^2(T)\tilde{w}^2(T)] + O(1/n) = \sigma^2_{\text{SIS}} + O(1/n) \to \sigma^2_{\text{SIS}}.
\end{aligned}$$

So, $\hat{\sigma}^2_{\text{SIS}}$ is a consistent estimator of σ^2_{SIS}. □

Proof of Proposition 6.4.1. We first show that the normalizing constant for $g_{\boldsymbol{T}}^{\text{KL}}(\boldsymbol{t})$ is indeed c^* defined as (6.8). Observe that

$$\begin{aligned}
\int_{\Omega_{\boldsymbol{X}}} |\Psi(\boldsymbol{X})| f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} &= \mathrm{E}_f[|\Psi(\boldsymbol{X})|] = \mathrm{E}_f\left[\mathrm{E}_f[|\Psi(\boldsymbol{X}) \,|\, T|]\right] \\
&= \mathrm{E}_f[s(T)] = \int_{\Omega_T} s(\boldsymbol{T}) f_{\boldsymbol{T}}(\boldsymbol{t}) d\boldsymbol{t},
\end{aligned}$$

so the result for the normalizing constant follows. We then have

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(g^*_{\boldsymbol{X}}(\boldsymbol{x})||g_{\boldsymbol{X}}(\boldsymbol{x})\right) = \int_{\Omega_{\boldsymbol{X}}} \ln\left(\frac{g^*_{\boldsymbol{X}}(\boldsymbol{x})}{g_{\boldsymbol{X}}(\boldsymbol{x})}\right) g^*_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} \\
&= -\int_{\Omega_{\boldsymbol{X}}} \ln\left(g_{\boldsymbol{X}}(\boldsymbol{x})\right) g^*_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} + c_1 = -\int_{\Omega_{\boldsymbol{X}}} \ln\left(f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}\,|\,\boldsymbol{t}_{\boldsymbol{x}})g_{\boldsymbol{T}}(\boldsymbol{t}_{\boldsymbol{x}})\right) g^*_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} + c_1 \\
&= -\int_{\Omega_{\boldsymbol{X}}} \ln\left(g_{\boldsymbol{T}}(\boldsymbol{t}_{\boldsymbol{x}})\right) g^*_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} + c_2 = -\mathrm{E}_f\left[\ln\left(g_{\boldsymbol{T}}(\boldsymbol{T})\right)c^*|\Psi(\boldsymbol{X})|\right] + c_2 \\
&= -\mathrm{E}_f\left[\mathrm{E}_f\left[\ln\left(g_{\boldsymbol{T}}(\boldsymbol{T})\right)c^*|\Psi(\boldsymbol{X})|\,|\,\boldsymbol{T}\right]\right] + c_2 = -\mathrm{E}_f\left[\ln\left(g_{\boldsymbol{T}}(\boldsymbol{T})\right)c^* s(\boldsymbol{T})\right] + c_2 \\
&= \int_{\Omega_{\boldsymbol{T}}} -\ln\left(g_{\boldsymbol{T}}(\boldsymbol{t})\right) g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})d\boldsymbol{t} + c_2 = \int_{\Omega_{\boldsymbol{T}}} \ln\left(\frac{g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})}{g_{\boldsymbol{T}}(\boldsymbol{t})}\right) g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})d\boldsymbol{t} + c_3 \\
&= D_{\mathrm{KL}}\left(g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})||g_{\boldsymbol{T}}(\boldsymbol{t})\right) + c_3,
\end{aligned}
$$

where c_1, c_2, and c_3 are given by

$$
\begin{aligned}
c_1 &= \int_{\Omega_{\boldsymbol{X}}} \ln\left(g^*_{\boldsymbol{X}}(\boldsymbol{x})\right) g^*_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}, \quad c_2 - c_1 = -\int_{\Omega_{\boldsymbol{X}}} \ln\left(f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}\,|\,\boldsymbol{t}_{\boldsymbol{x}})\right) g^*_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}, \\
c_3 - c_2 &= \int_{\Omega_{\boldsymbol{T}}} \ln\left(g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})\right) g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})d\boldsymbol{t}
\end{aligned}
$$

so that c_3 is a constant that does not depend on $g_{\boldsymbol{T}}(\boldsymbol{t})$. □

Proof of Proposition 6.4.2. For any $\boldsymbol{r} \in \mathbb{R}^p$,

$$
\begin{aligned}
\mathrm{E}_{g^*_{\boldsymbol{T}}}\left[e^{\boldsymbol{r}'\boldsymbol{T}^*}\right] &= \int_{\Omega_{\boldsymbol{X}}} e^{\boldsymbol{r}'T(\boldsymbol{x})}c^*|\Psi(\boldsymbol{x})|f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} = c^*\mathrm{E}_f[e^{\boldsymbol{r}'T(\boldsymbol{X})}|\Psi(\boldsymbol{X})|] = c^*\mathrm{E}_f[e^{\boldsymbol{r}'\boldsymbol{T}}s(\boldsymbol{T})] \\
&= \int_{\Omega_{\boldsymbol{T}}} e^{\boldsymbol{r}'\boldsymbol{t}}c^*s(\boldsymbol{t})f_{\boldsymbol{T}}(\boldsymbol{t})d\boldsymbol{t} = \int_{\Omega_{\boldsymbol{T}}} e^{\boldsymbol{r}'\boldsymbol{t}}g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})d\boldsymbol{t} = \mathrm{E}_{g^{\mathrm{KL}}_{\boldsymbol{T}}}\left[e^{\boldsymbol{r}'\boldsymbol{T}}\right], \qquad (16)
\end{aligned}
$$

Therefore, $T^* \sim g^{\mathrm{KL}}_{\boldsymbol{T}}(\boldsymbol{t})$ by uniqueness of the moment generating function. □