Methodology and Computing in Applied Probability (2022) 24:3049–3073
https://doi.org/10.1007/s11009-022-09970-1

Single-Index Importance Sampling with Stratification

Erik Hintz[1] · **Marius Hofert**[1] · **Christiane Lemieux**[1] · **Yoshihiro Taniguchi**[2]

Received: 23 November 2021 / Revised: 17 June 2022 / Accepted: 11 July 2022 /
Published online: 21 July 2022

Abstract
In many stochastic problems, the output of interest depends on an input random vector mainly through a single random variable (or index) via an appropriate univariate transformation of the input. We exploit this feature by proposing an importance sampling method that makes rare events more likely by changing the distribution of the chosen index. Further variance reduction is guaranteed by combining this single-index importance sampling approach with stratified sampling. The dimension-reduction effect of single-index importance sampling also enhances the effectiveness of quasi-Monte Carlo methods. The proposed method applies to a wide range of financial or risk management problems. We demonstrate its efficiency for estimating large loss probabilities of a credit portfolio under a normal and t-copula model and show that our method outperforms the current standard for these problems.

Keywords Single-index model · Importance sampling · Stratified sampling · Quasi-Monte Carlo · Loss probabilities

Erik Hintz, Marius Hofert, Christiane Lemieux and Yoshihiro Taniguchi have contributed equally to this work.

✉ Erik Hintz
erik.hintz@uwaterloo.ca

Marius Hofert
marius.hofert@uwaterloo.ca

Christiane Lemieux
clemieux@uwaterloo.ca

Yoshihiro Taniguchi
ytanigucmc@gmail.com

[1] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave W, Waterloo N2L 3G1, Ontario, Canada

[2] Canadian Imperial Bank of Commerce, Toronto, Ontario, Canada

1 Introduction

Many stochastic problems in finance and risk management are high-dimensional with a univariate quantity of interest, say $\mu = \mathbb{E}(\Psi(\boldsymbol{X}))$ for some integrable function $\Psi : \mathbb{R}^d \to \mathbb{R}$ and random vector $\boldsymbol{X} \sim F_{\boldsymbol{X}}$ for some d-dimensional distribution function $F_{\boldsymbol{X}}$. Because μ rarely allows for an analytical expression, the plain Monte Carlo (MC) estimator $\hat{\mu}_n^{\mathrm{MC}} = (1/n)\sum_{i=1}^n \Psi(\boldsymbol{X}_i)$ where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{ind.}}{\sim} F_{\boldsymbol{X}}$ is a popular choice for finding approximate solutions to such problems. Being unbiased and having an estimation error converging to zero at a rate independent of the dimension of the problem makes MC often popular for finding approximate solutions to such problems. The drawback of plain MC is the high computational cost it requires to obtain an estimate with a sufficiently small error. This issue is particularly severe for rare-event simulation, i.e., when $\mathbb{P}(|\Psi(\boldsymbol{X})| > 0)$ is small, as then a typically very large number of samples is required to obtain non-zero observations and therefore an estimator with small variance. As such, plain MC is often combined with variance reduction techniques (VRTs), such as control variates (see, e.g., Lavenberg and Welch (1981)) or stratified sampling (SS) (see, e.g., Cochran (2005)) to make the variance and thus the width of the estimate's confidence interval small.

Importance sampling (IS) is a VRT frequently applied to rare-event analysis in order to improve the reliability of MC estimators; see, e.g., Kahn and Marshall (1953) and Asmussen and Glynn (2007). The main idea of IS is to draw samples from a proposal distribution that puts more mass on the rare-event region of the sample space than the original distribution. As the efficiency of IS depends heavily on the choice of the proposal distribution, finding a good proposal distribution is a crucial step in applying IS. Unfortunately, there is no single best strategy known for finding a good proposal distribution that works in every situation since the nature of the rare event and what constitutes a good proposal distribution depends on the problem at hand; that is, on Ψ and $F_{\boldsymbol{X}}$. As such, much of the existing work on IS in computational finance finds effective proposal distributions by exploiting the structure of specific problems: Glasserman et al. (1999) develop IS methods to price path-dependent options under multivariate normal models; Glasserman et al. (2000, 2002) estimate the Value-at-Risk of a portfolio consisting of stocks and options under a normal and t-distribution; Sak et al. (2010) estimate tail probabilities of equity portfolios under generalized hyperbolic marginals with a t-copula assumption; Glasserman and Li (2005) estimate tail probabilities of credit portfolios under the Gaussian copula, Bassamboo et al. (2008); Chan and Kroese (2010) consider t-copula models. As all these IS techniques are exploiting specific properties of the problem at hand, they can achieve substantial variance reduction but are typically specific techniques not applicable to other problems without major modifications.

The contribution of this work is the development of theory and algorithms to apply IS for a wide range of problems by introducing a conditioning sampling step and optimally twisting the distribution of the conditioning variable. Let $T = T(\boldsymbol{X})$ be some univariate random variable, such as $\boldsymbol{\beta}^\top \boldsymbol{X}$ for some (well chosen) $\boldsymbol{\beta} \in \mathbb{R}^d$, and assume sampling from $\boldsymbol{X} \mid T$ is feasible. Let f be the density of T, $F_{\boldsymbol{X}|T}(\cdot \mid T = t)$ be the distribution of $\boldsymbol{X}$ given $T = t$ and g be a proposal density for T (assumed to have the same support as f), define by

$$\hat{\mu}_n^{\mathrm{SIS}} = (1/n)\sum_{i=1}^n \Psi(\boldsymbol{X}_i) f(T_i)/g(T_i), \quad T_i \overset{\text{ind.}}{\sim} g, \quad \boldsymbol{X}_i \overset{\text{ind.}}{\sim} F_{\boldsymbol{X}|T}(\cdot \mid T = T_i)$$

for $i = 1, \ldots, n$. If T explains much of the variability of the output, so if $R^2 := \mathrm{Var}(\mathbb{E}(\Psi(\boldsymbol{X}) \mid T))/\mathrm{Var}(\Psi(\boldsymbol{X}))$ is large, we can choose g optimally and make the rare event more

likely by changing the distribution of $\boldsymbol{X}$ through changing the distribution of the univariate T. Many high dimensional financial problems are of this nature; see, e.g., Caflisch et al. (1997); Wang and Fang (2003); Wang and Sloan (2005); Wang (2006).

In order to analyze our estimator, we write

$$\Psi(\boldsymbol{X}) = m(T) + \varepsilon_{\boldsymbol{X},T}$$

for some (unknown) transformation $T : \mathbb{R}^d \to \mathbb{R}$, where $m^{(k)}(t) = \mathbb{E}(\Psi(\boldsymbol{X})^k \mid T)$ for $k \in \mathbb{N}$ and $\varepsilon_{\boldsymbol{X},T}$ is a random error so that $\varepsilon_{\boldsymbol{X},T} \mid T$ has mean 0 and variance $v^2(t) = \mathrm{Var}(\Psi(\boldsymbol{X}) \mid T = t)$. We say that $\Psi(\boldsymbol{X})$ has a strong single index structure if R^2 is large (say, $R^2 > 0.9$), and the resulting estimator is referred to as Single Index IS (SIS) estimator. We will show that the optimal proposal distribution for g under SIS is proportional to $(m^{(2)})^{1/2}(t)f(t)$ resulting in an estimator with variance no larger than the plain MC estimator. If the proposal distribution g allows for a simple way to evaluate the quantile function G_T^{-1} of g, we can further reduce the variance by applying equal stratification to the support of T, i.e., instead of sampling $T_1, \dots, T_n \overset{\text{ind.}}{\sim} g$, we can set $T_i = G_T^{-1}(U_i)$ where $U_k \overset{\text{ind.}}{\sim} \mathrm{U}(k/n, (k+1)/n)$ for $k = 0, \dots, n-1$ and $G_T^{\leftarrow}(u) = \inf\{t \in \mathbb{R} : G_T(t) \geq u\}$ is the quantile function of T under g. The resulting method is referred to as stratified SIS (SSIS). We also derive optimal variance expressions in this case and show that (S)SIS gives zero variance when $R^2 = 1$. The derivation of these results along with some more notation and the connection between our methods and the IS and stratification techniques from Arbenz et al. (2018); Glasserman et al. (1999); Neddermeyer (2011) can be found in Sect. 2. There, we also briefly explain how our conditional sampling step reduces the effective dimension of the problem and therefore makes quasi-Monte Carlo (QMC) particularly attractive in our setting; in QMC, pseudo-random numbers (PRNs) are replaced by more homogeneously distributed quasi-random numbers (see, e.g., see Niederreiter (1978); Lemieux (2009); Dick and Pillichshammer (2010)).

Besides the choice of g, the performance of our procedure heavily depends on the choice of the transformation T, which must be chosen such that i) sampling from $\boldsymbol{X} \mid T$ is feasible and ii) T explains a lot of the variability of $\Psi(\boldsymbol{X})$, i.e., R^2 is as close to 1 as possible. The choice of the transformation is clearly not unique. In our numerical examples, we typically assume that T is a linear function of $\boldsymbol{X}$, whose coefficients can be estimated via the average derivative method of Stoker (1986), the sliced inverse regression of Li (1991) or the semiparametric least-squares estimator of Ichimura (1993). We remark that these methods do not require the form of the function $m(t)$ to be known.

As seen earlier, the optimal proposal densities involve a conditional moment function that is not known in practice. We propose to estimate this function using pilot-runs. The resulting point-wise approximation to the optimal density function can then be integrated and inverted numerically using the NINIGL algorithm developed in Hörmann and Leydold (2003). When this is too time-consuming, we suggest finding an approximately optimal g in the same parametric family as f (e.g., a location-scale transform of the original density). We detail this calibration stage, i.e., the process of estimating T, the optimal density and a way to sample from it, in Sect. 3.

In the numerical examples in Sect. 4, we demonstrate that our methods are applicable to a wide range of problems and achieve substantial variance reduction. After investigating a simple linear model example, we consider the problem of tail probability estimation in Gaussian and t-copula credit portfolio problems and show that our methods outperform those of Glasserman and Li (2005) and Chan and Kroese (2010).

As our formulation of (S)SIS does not assume a specific Ψ or $F_{\boldsymbol{X}}$, it is applicable to a wide range of problems and is efficient as long as the problem of interest has a strong

enough single-index structure. It also adapts to the problem through the design of the one-dimensional transformation revealing the single-index structure and through the choice of the proposal distribution. Besides its applicability to a wide range of problems, our proposed method has the following advantages. First, as it applies IS only to the univariate transformation variable, SIS is less susceptible to the dimensionality problem of IS, which is discussed in Au and Beck (2003); Katafygiotis and Zuev (2008); Schüeller et al. (2004). This also simplifies the task of finding an optimal proposal distribution. Second, SIS has a dimension reduction feature, so it enhances the effectiveness of QMC sampling methods. Third, by applying IS to a transformation of the input random vector $\boldsymbol{X}$, our proposal distribution amounts to changing the dependence structure of the problem under study, which can have a significant advantage over methods that only change the marginal distributions.

We conclude this paper in Sect. 5.

2 Variance Analysis and Optimal Calibration for SIS and SSIS

2.1 Notations and Definitions

To fix notation, recall we estimate $\mu = \mathbb{E}(\Psi(\boldsymbol{X}))$ via

$$\hat{\mu}_n^{\mathrm{SIS}} = (1/n) \sum_{i=1}^{n} \Psi(\boldsymbol{X}_i) w(T_i), \quad T_i \overset{\text{ind.}}{\sim} g_T, \quad \boldsymbol{X}_i \overset{\text{ind.}}{\sim} F_{\boldsymbol{X}|T}(\cdot \mid T = T_i),$$

for $i = 1, \ldots, n$, where f_T and g_T denote the original and proposal densities for T with supports $\Omega_f = (t_{\inf}, t_{\sup})$ (with possibly $t_{\inf}, t_{\sup} \in \{\pm\infty\}$) and Ω_g and $w(t) = g_T(t)/f_T(t)$ is the IS weight function.

Furthermore, we model the output $\Psi(\boldsymbol{X})$ as $\Psi(\boldsymbol{X}) = m(T) + \varepsilon_{\boldsymbol{X},T}$, where

$$m^{(k)}(t) = \mathbb{E}(\Psi(\boldsymbol{X})^k \mid T), \quad \mathrm{Var}(\varepsilon_{\boldsymbol{X},T} \mid T) = v^2(t) = \mathrm{Var}(\Psi(\boldsymbol{X}) \mid T).$$

and $\mathbb{E}(\varepsilon_{\boldsymbol{X},T} \mid T) = 0$. We already introduced the coefficient of determination $R^2 = \mathrm{Var}(m(T))/\mathrm{Var}(\Psi(\boldsymbol{X}))$ (see, e.g., Kvalseth (1985)) and said that $\Psi(\boldsymbol{X})$ is a strong single-index model if R^2 is large. This can be true for any model $\Psi(\boldsymbol{X})$, as we allow $\varepsilon_{\boldsymbol{X},T}$ to depend on $\boldsymbol{X}$. However, a pure single index model is a situation where $\varepsilon_{\boldsymbol{X},T} = \varepsilon_T$ only depends on $\boldsymbol{X}$ through T. In that case, it is easy to see that $\mathbb{E}(\Psi(\boldsymbol{X}) \mid T) = m(T)$, so that overall the random variable $\Psi(\boldsymbol{X})$ depends on $\boldsymbol{X}$ only through T. However, we do not impose the assumption of a pure single index model. Readers are referred to Powell et al. (1989), Härdle et al. (1993) and Ichimura (1993) for more information on single-index models.

Based on the representation of $\Psi(\boldsymbol{X})$ and using the law of total variance, we can write

$$\mathrm{Var}(\Psi(\boldsymbol{X})) = \mathrm{Var}(m(T)) + \mathbb{E}(v^2(T)) = \mathrm{Var}(m(T)) + \mathrm{Var}(\varepsilon_{\boldsymbol{X},T}), \tag{1}$$

since $\mathbb{E}(v^2(T)) = \mathrm{Var}(\varepsilon_{\boldsymbol{X},T}) - \mathrm{Var}(\mathbb{E}(\varepsilon_{\boldsymbol{X},T} \mid T)) = \mathrm{Var}(\varepsilon_{\boldsymbol{X},T})$. We see that (1) decomposes the variance of $\Psi(\boldsymbol{X})$ into two pieces: the one of the (random) systematic part, $m(T)$ and the unsystematic error $\varepsilon_{\boldsymbol{X},T}$ of the model. Note that (1) holds irrespective of wether we have a pure single index model or not.

In addition to applying IS on T, we also propose to use stratification on T to further reduce the variance; it will turn out that this essentially "stratifies away" $\mathrm{Var}(m(T))$, the variance of the

systematic part of the model. More precisely, let $\Omega_f = (t_{\inf}, t_{\sup})$ where possibly $t_{\inf} = -\infty$ and $t_{\sup} = \infty$. The SSIS scheme splits Ω_f into n strata of equal probability under g and draws one sample of T from each stratum. This is accomplished by first sampling $U_i \overset{\text{ind.}}{\sim} \mathrm{U}((i-1)/n, i/n)$ and then applying the quantile function to set $T_i = G_T^{\leftarrow}(U_i)$. Our estimator becomes

$$\hat{\mu}_n^{\mathrm{SSIS}} = (1/n)\sum_{i=1}^{n} \Psi(\boldsymbol{X}_i) w(T_i), \quad T_i = G_T^{\leftarrow}(U_i), \quad U_i \sim \mathrm{U}((i-1)/n, i/n),$$

and, as before, $\boldsymbol{X}_i \overset{\text{ind.}}{\sim} F_{\boldsymbol{X}|T}(\cdot \mid T = T_i)$ for $i = 1, \ldots, n$.

For our variance analysis below, it is useful to find an expression for $\mathrm{Var}(\hat{\mu}_n^{\mathrm{MC}})$. Note that the conditional moment functions $m^{(k)}$ do not depend on wether we sample from f_T or g_T. From (1) and the fact that $\mathrm{Var}(m(T)) = \mathbb{E}(m(T)^2) - \mu^2$ as well as $\mathbb{E}(v^2(T)) = \mathbb{E}(m^{(2)}(T)) - \mathbb{E}(m(T))^2$, we find

$$n\mathrm{Var}(\hat{\mu}_n^{\mathrm{MC}}) = \mathrm{Var}(m(T)) + \mathbb{E}(v^2(T)) = \mathbb{E}(m^{(2)}(T)) - \mu^2. \tag{2}$$

As should be clear from the form of our estimators, their bias depends on the support Ω_g of g_T. We define

$$\mu_{\mathrm{SIS}} = \int_{\Omega_g} m(t) f_T(t)\, \mathrm{d}t$$

and

$$\sigma^2_{\mathrm{SIS}} = \int_{\Omega_g} m^{(2)}(t) \frac{f_T^2(t)}{g_T(t)}\, \mathrm{d}t - \mu^2_{\mathrm{SIS}}, \quad \sigma^2_{\mathrm{SSIS}} = \int_{\Omega_g} v^2(t) \frac{f_T^2(t)}{g_T(t)}\, \mathrm{d}t.$$

Notice that μ_{SIS} depends on g_T through the region Ω_g. The SIS and SSIS estimators are unbiased if g_T is such that $g_T(t) > 0$ whenever $m(t)f_T(t) > 0$, which holds, by construction, in all our numerical examples.

2.2 Optimal Densities

We are now able to derive properties of the (S)SIS estimators and derive the optimal (variance-minimizing) proposal distribution of g_T; see the Appendix for the proofs. As the objective of our IS techniques is variance reduction, we call the practice of setting g_T to its optimal density or their approximation as *optimal calibration*, and the resulting methods SIS* and SSIS*.

Proposition 1 (Variance-optimal SIS) *We have* $\mathbb{E}(\hat{\mu}_n^{\mathrm{SIS}}) = \mu_{\mathrm{SIS}}$ *and* $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SIS}}) = \sigma^2_{\mathrm{SIS}}/n$. *If* $\mathbb{E}_g(m^2(T)w^2(T)) < \infty$, *then* $\sqrt{n}(\hat{\mu}_n^{\mathrm{SIS}} - \mu_{\mathrm{SIS}}) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2_{\mathrm{SIS}})$ *as* $n \to \infty$.

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

$$g_T^{\mathrm{opt}}(t) = c^{-1}\sqrt{m^{(2)}(t)} f_T(t), \quad t \in (t_{\inf}, t_{\sup}), \quad c = \int_{t_{\inf}}^{t_{\sup}} \sqrt{m^{(2)}(t)} f_T(t)\, \mathrm{d}t. \tag{3}$$

The variance of the optimal SIS estimator, denoted by $\hat{\mu}_n^{\mathrm{SIS,opt}}$, *is* $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SIS,opt}}) = (c^2 - \mu^2)/n$.

Remark 1

1. Proposition 1 implies that using optimal SIS gives variance no larger than MC. Indeed, by Jensen's inequality, $n\mathrm{Var}(\hat{\mu}_n^{\mathrm{SIS,opt}}) \leq \mathbb{E}(m^{(2)}(T)) - \mu^2$, which is equal to $\mathrm{Var}(\hat{\mu}_n^{\mathrm{MC}})$ using (2). This inequality holds as an equality only when $m^{(2)}(t)$ is constant for all $t \in \Omega_T$.
2. If $R^2 = 1$ (corresponding to the strongest possible single index structure), then $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SIS,opt}}) = 0$: SIS provides a zero-variance estimator if $m^{(2)}(t) = (m(t))^2$ for all t, which is equivalent to having $v^2(t) = 0$ for all t, or equivalently, to having $\mathbb{E}(v^2(T)) = 0$ since $v^2(t) \geq 0$ for all t. This is the same as asking $\mathrm{Var}(m(T))/\mathrm{Var}(\Psi(\boldsymbol{X})) = R^2 = 1$. This is why choosing a function T such that the model is an as good fit as possible is important for the SIS method to achieve significant variance reduction.

The following proposition gives the properties of the SSIS estimator and the optimal (variance-minimizing) proposal distribution of g_T. Its proof is in the Appendix.

Proposition 2 (Variance-optimal SSIS) *It holds that* $\mathbb{E}(\hat{\mu}_n^{\mathrm{SSIS}}) = \mu_{\mathrm{SIS}}$ *and, for large enough* n, $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS}}) = \sigma^2_{\mathrm{SIS}}/n + o(1/n)$. *If* $\mathbb{E}_g\left(|m(T)w(T)|^{2+\delta}\right) < \infty$ *for some* $\delta > 0$, $\hat{\mu}_n^{\mathrm{SSIS}}$ *is asymptotically normal as* $\sqrt{n}(\hat{\mu}_n^{\mathrm{SSIS}} - \mu_{\mathrm{SIS}}) \to \mathrm{N}(0, \sigma^2_{\mathrm{SIS}})$ *for* $n \to \infty$. *Suppose that* $\Psi(\boldsymbol{x}) \geq 0$ *or* $\Psi(\boldsymbol{x}) \leq 0$ *for all* $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ *and that* $\mathbb{P}_f(v^2(T) = 0,\ m(T) \neq 0) = 0$. *The density* g_T *that gives an unbiased SSIS estimator with the smallest variance is*

$$g_T^{\mathrm{opt,s}}(t) = c^{-1} v(t) f_T(t), \quad t \in (t_{\mathrm{inf}}, t_{\mathrm{sup}}), \quad c = \int_{t_{\mathrm{inf}}}^{t_{\mathrm{sup}}} v(t) f_T(t)\, \mathrm{d}t. \tag{4}$$

The variance of the optimal SSIS estimator $\hat{\mu}_n^{\mathrm{SSIS,opt}}$ *is* $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS,opt}}) = c^2/n + o(1/n)$. *If* $\mathbb{P}_f(v^2(T) = 0,\ m(T) \neq 0) > 0$, *then* $\hat{\mu}_n^{\mathrm{SSIS,opt}}$ *is biased.*

Remark 2

1. Proposition 2 implies that using optimal SSIS gives asymptotically a variance no larger than MC. Indeed, Jensen's inequality implies that we have $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS,opt}}) \leq (1/n)\mathbb{E}(v^2(T)) + o(1/n)$ with equality only if $v(t)$ is constant for all $t \in \Omega_T$. From (2) (and ignoring the $o(1/n)$ term), this means $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS,opt}}) \leq \mathrm{Var}(\hat{\mu}_n^{\mathrm{MC}})$, with equality only if $v(t)$ is constant for all $t \in \Omega_T$ and $\mathrm{Var}(m(T)) = 0$, which is unlikely to be the case since $m(T)$ has been chosen specifically such that $R^2 \approx 1$.
2. If $R^2 = 1$ (strongest possible single index structure), then $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS,opt}}) = 0$, since $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS}}) = 0$ iff $m^{(2)}(t) = (m(t))^2$ for all t, or equivalently $v^2(t) = 0$ for all t and thus $\mathbb{E}(v^2(T)) = 0$, which means $R^2 = 1$.
3. Unless $m(t) = 0$, SSIS achieves variance reduction compared to SIS, as $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS}}) \leq \mathrm{Var}(\hat{\mu}_n^{\mathrm{SIS}})$ for the same choice of g_T. This in turn implies that $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS,opt}}) \leq \mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS,opt}})$. The proposal densities g_T^{opt} and $g_T^{\mathrm{opt,s}}$ defined in (3) and (4) give estimators with smallest variance if $\Psi(\boldsymbol{x}) \geq 0$ or $\Psi(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \Omega$, which holds for many applications in finance (e.g., when Ψ is an indicator and thus μ a probability or when Ψ is the payoff of an option). If Ψ takes both positive and negative values, $m(t)$ could be 0 for some values of t. We can then improve the optimal calibration by setting $g_T(t) = 0$ whenever $m(t) = 0$. Since it is

generally unknown and hard to estimate which values of t give $m(t) = 0$, this improvement may not be implementable.

4. The expression for $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS,opt}})$ implies that SSIS* "stratifies away" the variance captured by the systematic part $m(T)$ of the single-index model, so the variance of the SSIS* estimator comes only from the error term $\varepsilon_{\boldsymbol{X},T}$ via $v(t)$. If g_T is not chosen optimally, then $\mathrm{Var}(\hat{\mu}_n^{\mathrm{SSIS}}) = \sigma^2_{\mathrm{SIS}}/n + o(1/n)$ shows that we still make $\mathrm{Var}(m(T))$ vanish by using stratification, but the contribution from $v^2(T)$ might be amplified (compared to how it contributes to the MC estimator's variance) if we do not choose a good proposal density. Irrespective of the choice of g_T it is true that the stronger the fit of the single index model, the better (S)SIS works.
5. These results show that as long as the problem at hand has a strong single-index structure and sampling from T and $\boldsymbol{X} \mid T$ is feasible, SIS and SSIS can be applied and should give large variance reduction. As those conditions do not assume a specific form for Ψ or for the distribution of $\boldsymbol{X}$, SIS and SSIS are applicable to a wide range of problems.

Proposition 2 asserts the asymptotic normality of the SSIS estimator. In order to construct a confidence interval from this estimator, we must estimate σ^2_{SSIS}. We take an approach similar to the one by Wang et al. (2008) where the first-order difference of samples are taken to remove the effect of the mean function. Its proof is in the Appendix.

Proposition 3 (Estimation of σ^2_{SSIS}) *Let G_T be the distribution function corresponding to g_T. If G_T^{-1}, m and v^2 are continuously differentiable over the domain of T under the proposal distribution, then*

$$\hat{\sigma}^2_{\mathrm{SSIS}} = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} r_i^2 w^2(T_i)$$

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

Proposition 3 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 3 separately to each interval and combine them to obtain $\hat{\sigma}^2_{\mathrm{SSIS}}$.

2.3 Connection to Other IS and SS Techniques

In this subsection, we explain some connections of our proposed methods to other IS and SS techniques.

Suppose that $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{0}, I_d)$. A popular strategy for constructing a proposal distribution under the multivariate normal (MVN) model is to shift its mean vector of $\boldsymbol{X}$, that is, letting $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\eta}, I_d)$ under the IS distribution for some $\boldsymbol{0} \neq \boldsymbol{\eta} \in \mathbb{R}^d$. The following proposition states that this type of IS can be achieved within our SIS framework by using $T(\boldsymbol{X}) = \boldsymbol{\theta}^\top \boldsymbol{X}$ where $\boldsymbol{\theta}$ is the normalized version of $\boldsymbol{\eta}$. Based on Proposition 1 and Remark 1, this result thus implies that this popular mean-shifting strategy for MVN models works well if the problem has a strong linear single-index structure based on the specific choice of shift vector $\boldsymbol{\eta}$.

Proposition 4 (SIS in MVN models) *Let* $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{0}, I_d)$ *under the original distribution. Fix* $\boldsymbol{0} \neq \boldsymbol{\beta} \in \mathbb{R}^d$ *with* $\boldsymbol{\beta}^\top \boldsymbol{\beta} = 1$. *Consider SIS with* $T(\boldsymbol{X}) = \boldsymbol{\beta}^\top \boldsymbol{X}$. *If* g_T *is the density of* $\mathrm{N}(c, \sigma^2)$, *then* $\boldsymbol{X} \sim \mathrm{N}_d(c\boldsymbol{\beta}, I_d + (\sigma^2 - 1)\boldsymbol{\beta}\boldsymbol{\beta}^\top)$ *in the IS scheme.*

Proposition 4 implies that $\boldsymbol{X} \sim \mathrm{N}_d(c\boldsymbol{\beta}, I_d)$ if g_T is chosen as the $\mathrm{N}(c, 1)$ density (where we recall that the original distribution f_T is $\mathrm{N}(0, 1)$), so that the previously mentioned mean-shifting strategy is a special case of IS (namely, by merely shifting the mean of T instead of applying SIS*). If $\mathrm{Var}(T) \neq 1$ under g, the dependence structure of the components in $\boldsymbol{X}$ does change in the IS scheme.

The stratification technique proposed in Glasserman et al. (1999) is applied by using the normalized shift vector as the stratification direction and can also be achieved within our SIS framework using the same function T and proposal distribution as in Proposition 4. The combination of IS and SS is not motivated as in Glasserman et al. (1999). In the latter reference, IS and SS are used to remove the variability due to the linear and the quadratic part, respectively, of $\Psi(\boldsymbol{X})$. In SSIS, SS is used to eliminate $\mathrm{Var}(m(T))$, the variance captured by the systematic part of the single-index model, and then IS is used to minimize the variance contribution from $\varepsilon_{X,T}$.

It is easy to see that the NPIS method proposed by Neddermeyer (2011) with $u = 1$ (where u is defined as in Neddermeyer (2011)) is closely connected to SIS with $T(\boldsymbol{X}) = X_1$. It is proposed to choose $g_T(t) = m(t)f_T(t)/\mu$ in Neddermeyer (2011), but by Proposition 1, choosing g_T^{opt} defined in (3) gives an IS estimator with a smaller variance.

SIS also generalizes the IS method in Arbenz et al. (2018) in two ways. First, SIS generalizes the form of the transformation function T, that is, it does not assume any specific form of T, while the IS method in Arbenz et al. (2018) assumes that $T(\boldsymbol{X}) = \max\{F_1(X_1), \dots, F_d(X_d)\}$, where $F_1, \dots, F_d$ are the marginal distribution functions of $\boldsymbol{X}$. Secondly, SIS generalizes the form of the proposed density of the transformed variable, whereas the proposal density g_T for the IS method in Arbenz et al. (2018) has the form

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

for some $M \geq 1$, $t_{\mathrm{inf}} = \lambda_1 < \cdots < \lambda_M$, and $q_1, \dots, q_M \geq 0$ such that $\sum_{k=1}^{M} q_k = 1$.

The single-index structure we exploit to design our SIS and SSIS schemes is strongly related to the idea of conditional MC. In both cases, the goal is to identify a function T of $\boldsymbol{X}$ that explains much of the variability of $\Psi(\boldsymbol{X})$. However, with conditional MC one typically also chooses T so that $m(t) = \mathbb{E}(\Psi(\boldsymbol{X}) \mid T(\boldsymbol{X}) = t)$ is known, and then estimates μ by the sample mean of the $m(T_i)$, $i = 1, \dots, n$. In our case, we do not assume or need this conditional expectation to be known in closed-form. This means we typically do not completely get rid of the $\mathrm{Var}(m(T))$ term in (1), but we aim to reduce it via IS; if SSIS is applied optimally, we actually do make $\mathrm{Var}(m(T))$ vanish.

2.4 Single-Index Importance Sampling and QMC

As mentioned in the introduction, further variance reduction can be achieved by performing the simulation based on quasi-random numbers (QRNs) instead of PRNs. Suppose we are given a sampling algorithm $\phi : [0, 1)^{d+k} \to \mathbb{R}^d$ for some $k \geq 0$ such

that $\phi(\boldsymbol{U}) \sim f_{\boldsymbol{X}}$ for $\boldsymbol{U} \sim \mathrm{U}[0,1)^{d+k}$. For instance, when $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$, then $k = 0$ and $\phi(\boldsymbol{u}) = \boldsymbol{\mu} + C(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d))^\top$, where the matrix C is such that $CC^\top = \Sigma$ and $\Phi(x) = \int_\infty^x (2\pi)^{-0.5} \exp(-t^2/2)\, \mathrm{d}t$ is the distribution function of the standard normal distribution. For a discussion of what the function ϕ is in a more general context, where $\boldsymbol{X}$ has a dependence structure modelled by a copula other than the Gaussian copula, we refer to Cambou et al. (2016). With ϕ at hand, we can write $\hat{\mu}_n^{\mathrm{MC}} = (1/n) \sum_{i=1}^n \Psi(\phi(\boldsymbol{U}_i))$ where $\boldsymbol{U}_i \overset{\text{ind.}}{\sim} \mathrm{U}(0,1)^{d+k}$. With QMC, we replace the $\boldsymbol{U}_i$ with deterministic vectors $\boldsymbol{v}_i \in [0,1)^{d+k}$ that fill the unit hypercube more evenly. A number of constructions for such points have been proposed (see e.g., Lemieux (2009), Ch. 5); we use the Sobol' sequence of Sobol' (1967) for our numerical examples later on. In order to obtain an easy-to-compute error bound, we apply a random digital shift to the $\boldsymbol{v}_i$ to obtain multiple independent and identically distributed realizations of the randomized QMC (RQMC) estimator. Based on the digitally-shifted RQMC estimates, we can compute a probabilistic error bound in the form of a confidence interval.

It is widely accepted that the performance of QMC is largely influenced by the effective dimension of the problem, a concept first introduced in Caflisch et al. (1997). More precisely, QMC works significantly better than plain MC if the problem has a low effective dimension; see also Wang and Fang (2003); Wang and Sloan (2005); Wang (2006). One notion of effective dimension is the truncation dimension; see Wang and Sloan (2005). Essentially, a problem has a low truncation dimension when only a small number of leading input variables are important. Recall that $\boldsymbol{X}$ is sampled indirectly in SIS, that is, T is generated first then $\boldsymbol{X}$ is drawn from $F_{\boldsymbol{X}|T}$. Assuming T is generated using the inversion method and via the first coordinate u_1 of $\boldsymbol{u} \in [0,1)^{k+d}$, the indirect sampling step of SIS transforms the problem in such a way that the first input variable accounts for $R^2 \cdot 100\%$ of the variance of $\Psi(\boldsymbol{X})$, where $R^2 = \mathrm{Var}(m(T))/\mathrm{Var}(\Psi(\boldsymbol{X}))$. That is, the problem has a truncation dimension of 1 in proportion R^2 under SIS. Therefore, if the fit of the single-index model is good, say $R^2 > 0.9$, the indirect sampling step via T serves as a dimension reduction technique and enhances the efficiency of QMC.

3 Calibration Stage in Practice

As mentioned in the introduction, we must estimate the optimal transformation function $T = T(\boldsymbol{X})$ and construct an approximation $\hat{g}_T^{\mathrm{opt}}$ for the optimal density g_T^{opt} before applying (S)SIS. We call the stage in which these two tasks are performed the *calibration stage*. Furthermore, the calibrations in (3) and (4) require the knowledge of the conditional mean function and variance function, respectively. As these are rarely known in practice, they must be estimated in the calibration stage as well.

3.1 Estimating the Optimal Transformation T

In what follows, we assume that T is a linear function of the components in $\boldsymbol{X}$, i.e., $T = \boldsymbol{\beta}^\top \boldsymbol{X}$ for some $\boldsymbol{\beta} \in \mathbb{R}^d$; note that if $\boldsymbol{X}$ is multivariate normal, then T is univariate normal and sampling from $\boldsymbol{X} \mid T$ is straightforward. To find $\boldsymbol{\beta}$ that maximizes R^2, we use the average derivative method of Stoker (1986), which essentially allows us to estimate $\boldsymbol{\beta}$ as if we met the assumptions of a linear regression. That is, we sample independent realizations

$\Psi(\boldsymbol{X}_i) =: \Psi_i$ for $i = 1, \dots, n_1$ (say, $n_1 = 1000$) and compute the sample covariance matrix $\Sigma_{\boldsymbol{X},\boldsymbol{X}}$ as well as the sample cross covariance of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{n_1}$ and $(\Psi_1, \dots, \Psi_{n_1})$, say $\Sigma_{\boldsymbol{X},\Psi}$ to obtain

$$\hat{\boldsymbol{\beta}} = \Sigma_{\boldsymbol{X},\boldsymbol{X}}^{-1} \Sigma_{\boldsymbol{X},\Psi}.$$

In some applications, we may use only a subset of the components in $\boldsymbol{X}$; in later examples, for instance, we only use the systematic risk factors in a credit model to build our transformation T. Sometimes one may even not need to estimate $\boldsymbol{\beta}$, for instance, if it is clear that the d components in $\boldsymbol{X}$ are equally important, one can simply set $\boldsymbol{\beta} = (1/\sqrt{d}, \dots, 1/\sqrt{d})$; see Sect. 4.2 for an example.

3.2 Finding the Optimal Density

The calibration in (3) requires the knowledge of the conditional second moment function $m^{(2)}(t) = \mathbb{E}(\Psi^2(\boldsymbol{X}) \mid T = t)$ for all $t \in \Omega_T$, which, of course, is not known; similarly, the conditional variance function v^2 required for the calibration in (4) is not known either. We now describe how to calibrate (3) in practice; the calibration of (4) can be done similarly.

Our first ingredient is the construction of an estimate of $m^{(2)}(t) = \mathbb{E}(\Psi^2(\boldsymbol{X}) \mid T = t)$ for all $t \in \Omega_T$; we suggest using plain MC for this purpose. To this end, let $t_1 < \cdots < t_M$ be knots at which the function $m^{(2)}$ is to be estimated (e.g., $M = 20$ equally spaced points in the relevant range). Choose some small pilot sample size n_{pilot} (for example, 5% of the total sample size n). For each t_j, sample n_{pilot}-many realizations from $\boldsymbol{X} \mid T = t_j$ and estimate $m^{(2)}(t_j)$ by its empirical equivalent for $j = 1, \dots, M$. Then utilize smoothing splines (see, for example, Reinsch (1967)) and only those t_j associated with a positive estimate to construct an estimate $\hat{m}^{(2)}$ for all $t \in \Omega_T$; for those t where $\hat{m}^{(2)}(t) \leq 0$, one can either leave them as $\hat{m}^{(2)}(t) = 0$ (which may lead to bias as discussed below) or set $\hat{m}^{(2)}$ to be some positive function (e.g., the error function) that resembles the lower tail of ε.

Having constructed an estimate for $m^{(2)}$, we can set $\hat{g}_T^{\text{opt}} \propto \sqrt{\hat{m}^{(2)}(t)} f(t)$ for $t \in \mathbb{R}$. However, $\hat{g}_T^{\text{opt}}$ rarely belongs to known parametric families of distributions that are easily sampled from. One can use numerical techniques such as the NINIGL algorithm to approximate the quantile function of a distribution given its unnormalized density; see Hörmann and Leydold (2003). This approach, however, has three drawbacks: i) sampling from a numerically constructed density is time-consuming and can be prone to numerical problems; ii) the normalizing constant needs to be estimated, and iii) bias can occur when $\hat{g}_T^{\text{opt,s}}$ does not have the same support as $g_T^{\text{opt,s}}$, which in turn happens when $\hat{m}^{(2)}(t) = 0$ even though $m^{(2)}(t) \neq 0$ for some set D with $\int_D f(t)\, \mathrm{d}t > 0$.

The third drawback can be alleviated if we can define $\hat{m}^{(2)}(t)$ to be positive whenever $m^{(2)}(t)$ is (for example, by assuming some lower and upper tail behaviour). Furthermore, recall from Proposition 2 that (4) gives a biased estimator if $\mathbb{P}_f(v^2(T) = 0,\ m(T) \neq 0) > 0$ which in some cases can be debiased. For instance, if $v(t) > 0$ for all $t \in \Omega_t$, but the estimated $\hat{v}(t) = 0$ for $t \geq t_{\max}$ for some $t_{\max} \in \mathbb{R}$ and $m(t) = c$ for some constant c for $t \geq t_{\max}$ (for instance, if μ is a probability, then typically $m(t) = c = 1$ for $t \geq t_{\max}$). If $\hat{\mu}_n^{\text{SSIS}}$ is constructed using $\hat{v}$, we find

$$\mathbb{E}(\hat{\mu}_n^{\mathrm{SSIS}}) = \int_{-\infty}^{t_{\max}} m(t) f_T(t)\,\mathrm{d}t = \mu - \int_{t_{\max}}^{\infty} m(t) f_T(t)\,\mathrm{d}t \approx \mu - c\mathbb{P}_f(T > t_{\max});$$

$\hat{\mu}_n^{\mathrm{SSIS}}$ can therefore be debiased by adding $c\mathbb{P}_f(T > t_{\max})$.

The second drawback can be addressed by using weighted IS (so that the normalizing constant cancels out); (see Lemieux (2009), Sect. 4.5). Alternatively, the normalizing constant can be estimated as follows: Let $\hat{g}_{T,u}^{\mathrm{opt}}(t) = \sqrt{\hat{m}^{(2)}} g(t)$ denote the unnormalized density, and $T_1, \ldots, T_n \overset{\text{ind.}}{\sim} \hat{g}_T^{\mathrm{opt}}$ (obtained, for instance, using the NINIGL algorithm). Now construct an estimate of the density of $T_1, \ldots, T_n$, such as the kernel density estimator, and denote this estimated density by $\hat{h}$; note that $\hat{h}$ is normalized and that each of $\hat{h}(T_i)/\hat{g}_{T,u}^{\mathrm{opt}}(T_i)$ for $i = 1, \ldots, n$ is an estimator for the normalizing constant. As such, we suggest using the sample median of $\{\hat{h}(T_1)/\hat{g}_{T,u}^{\mathrm{opt}}(T_1), \ldots, \hat{h}(T_n)/\hat{g}_{T,u}^{\mathrm{opt}}(T_n)\}$ as an estimator for the normalizing constant.

The first drawback, that is, the construction of an approximation to the quantile function of $\hat{g}_T^{\mathrm{opt}}$ being both slow and potentially prone to numerical problems, is most severe. Below, we propose an alternative method, namely by setting $\hat{g}_T^{\mathrm{opt}}(t) = 1/\sigma f((t-k)/\sigma)$ for carefully chosen $k \in \mathbb{R}$ and $\sigma > 0$. In other words, we suggest using a location-scale transform of the original density as proposal density and will therefore call this method $\mathrm{SIS}^{c,\sigma}$. While this procedure does require estimation of k and σ, it does not suffer from any of the three aforementioned problems: i) if we can sample from f, we can also sample from $f((t-k)/\sigma)/\sigma$; ii) there is no normalizing constant or density to be estimated; iii) f and $f((t-k)/\sigma)/\sigma$ have the same support, so that the resulting estimator is unbiased.

The idea behind using a location-scale transform arises from the observation that in many practical examples (as will be seen later) the optimal density has roughly the same shape as the original density. As such, we try to find k and σ so that $1/\sigma f((t-k)/\sigma)$ is approximately $g_T^{\mathrm{opt}}(t)$. Denote again by $\hat{g}_{T,u}^{\mathrm{opt}}(t) = \sqrt{\hat{m}^{(2)}(t)} f(t)$ the unnormalized, estimated optimal density and assume that the mode of f_T is at zero (otherwise, shift accordingly). Now find $k^* = *argmax_t \hat{g}_{T,u}^{\mathrm{opt}}(t)$ numerically; this makes sure that the theoretical and approximated densities have (roughly) the same mode, thereby both sample from the "important region". Having estimated k^*, the next step is to compute σ such that it minimizes the variance of the resulting estimator. More precisely, given a sample $T_1, \ldots, T_{n_{\mathrm{pilot}}}$ from f, we can estimate the variance of the estimator for a given σ as follows: Set $\tilde{T}_i = k^* + \sigma T_i$ and $w_i = \frac{f(\tilde{T}_i)}{f((\tilde{T}_i - k^*)/\sigma)/\sigma}$ and sample $\boldsymbol{X}_i \mid \tilde{T}_i$ for $i = 1, \ldots, n_{\mathrm{pilot}}$. The second moment of the IS estimator (written as a function of the scale σ) is then

$$V(\sigma) = \sum_{i=1}^{n_{\mathrm{pilot}}} \Psi(\boldsymbol{X}_i) w_i^2, \quad \sigma > 0. \tag{5}$$

We can now solve $\sigma^* = *argmin_{\sigma>0} V(\sigma)$ numerically. Note that due to the nature of a location-scale transform, we only need to sample $T_1, \ldots, T_{n_{\mathrm{pilot}}}$ once. Intuitively, k^* shifts the density to the important region, while σ^* scales it appropriately. If computing $V(\sigma)$ is very time consuming (for example, when the sampling of $\boldsymbol{X} \mid T$ is complicated), one can set $\sigma^* = 1$; the resulting method is then called SIS^{μ} instead of $\mathrm{SIS}^{\mu,\sigma}$.

Algorithm 1 Calibration and estimation stage for estimating μ via $\mathrm{SIS}^{\mu,\sigma}$

Given knots $t_1, \dots, t_{n_{\text{pilot}}}$, a total pilot budget n_{tot} and knot-sample size n_{knot}, target sample size n, estimate μ via:

1. *Estimation of the direction vector.*
   (a) Sample $\boldsymbol{X}_i \overset{\text{ind.}}{\sim} f_{\boldsymbol{X}}$ for $i = 1, \dots, n_{\text{pilot}}$ and compute $\Psi(\boldsymbol{X}_i) =: \Psi_i$, $i = 1, \dots, n_{\text{pilot}}$.
   (b) Compute $\Sigma_{\boldsymbol{X},\boldsymbol{X}}$ and $\Sigma_{\boldsymbol{X},\Psi}$and set $\hat{\boldsymbol{\beta}} = \Sigma_{\boldsymbol{X},\boldsymbol{X}}^{-1} \Sigma_{\boldsymbol{X},\Psi}$.
2. *Estimation of c^* and σ^*.*
   (a) For each $k = 1, \dots, n_{\text{pilot}}$, sample $\boldsymbol{X}_{j,k} \overset{\text{ind.}}{\sim} f_{\boldsymbol{X}|T}(\cdot \mid t_k)$ for $j = 1, \dots, n_{\text{knot}}$.
   (b) Utilize smoothing splines[1] through $(t_k, (1/n_{\text{knot}}) \sum_{j=1}^{n_{\text{knot}}} \Psi(\boldsymbol{X}_{j,k})^2)$, $k = 1, \dots, n_{\text{pilot}}$, to construct an estimate for $\hat{m}^{(2)}(t)$ for $t \in \mathbb{R}$.
   (c) Find $c^* = \operatorname{argmax}_t \sqrt{\hat{m}^{(2)}(t)} f(t)$ numerically.
   (d) Sample $T_1, \dots, T_{n_{\text{pilot}}} \overset{\text{ind.}}{\sim} f$ and find $\sigma^* = \operatorname{argmin}_{\sigma>0} V(\sigma)$ with the function V from (5) numerically.
3. *Estimation of μ.*
   (a) Sample $T_1', \dots, T_n' \overset{\text{ind.}}{\sim} f$, set $T_i = c^* + \sigma^* T_i$ and compute $w_i = \frac{f(T_i)}{f((T_i - c^*)/\sigma^*)/\sigma^*}$ for $i = 1, \dots, n$.
   (b) Sample $\boldsymbol{X}_i \overset{\text{ind.}}{\sim} f_{\boldsymbol{X}|T}(\cdot \mid T_i)$ for $i = 1, \dots, n$.
   (c) Return $\hat{\mu}_n^{\text{SIS}} = (1/n) \sum_{i=1}^n \Psi(\boldsymbol{X}_i) w_i$.

Remark 3

1. Algorithm 1 can be easily adapted to accommodate quasi-random numbers and stratification, as will be discussed in the next section.
2. The effort for the conditional sampling needed in Steps 2a, 2d and 3b is problem specific – for some problems, samples of $\boldsymbol{X} \mid T = t_1$ can be easily transformed to samples from $\boldsymbol{X} \mid T = t_2$ for $t_1 \neq t_2$, making these steps very fast; in some other problems, the conditional sampling is more involved.
3. Our proposed SIS method can also be combined with other VRTs. For instance, in Sect. 4.3, we combine conditional MC (CMC) and SIS to estimate loss probabilities of a credit portfolio whose dependence is governed by a t-copula.

4 Numerical Experiments

In this section, we perform an extensive numerical study to demonstrate the effectiveness of our proposed methods. We start with a simplistic linear model example, in which case calibration of the optimal densities can be done easily. This allows us to investigate the effect of replacing g_T^{opt} by $\hat{g}_T^{\text{opt}}$. In Sect. 4.2, we apply our SIS and SSIS schemes to a credit portfolio problem under the Gaussian copula model studied by Glasserman and Li (2005).

The same financial problem but this time using a more complicated t-copula model is studied in Sect. 4.3. All computations were carried out in R; see R Core Team (2020).

4.1 Linear Model Example

Let $L = \alpha T + \varepsilon_T$ where $T \sim \mathrm{N}(0,1)$, $\varepsilon_T \mid T \sim \mathrm{N}(0, s^2)$ and $\alpha^2 + s^2 = 1$. L has a single index structure when $\alpha^2 \approx 1$ since $R^2 = \mathrm{Var}(m(T))/\mathrm{Var}(L) = \mathrm{Var}(\alpha T) = \alpha^2$.

Assume interest lies in estimating the probability $p_l = \mathbb{P}(L > l) = \bar{\Phi}(l) = \mathbb{E}(\mathbb{1}_{\{L>l\}})$ for some large l; note that we can approximate the true value of p_l efficiently with high precision since $L \sim \mathrm{N}(0,1)$. Furthermore it is easily seen that $p_l(t) = \mathbb{P}(L > l \mid T = t) = \bar{\Phi}((l - \alpha t)/s)$ for $l, t \in \mathbb{R}$. Since the integrand Ψ in this setting is an indicator, we find from Proposition 1 that $g_T^{\mathrm{opt}}(t) \propto \sqrt{p_l(t)} f_T(t)$.

Unlike in this simplistic setting, $p_l(t)$ for $t \in \mathbb{R}$ is unknown in practice as discussed in Sect. 3; thus, this setting serves as an excellent example to also compare whether approximating g_T^{opt} by $\hat{g}_T^{\mathrm{opt}}$ has a significant effect on the accuracy of the estimators. Sampling from the true optimal densities is performed using the R package `Runuran` of Leydold and Hörmann (2020). We consider the methods SIS** (constructed using known $p_l(t)$), SIS* (approximated $p_l(t)$ and NINGL), SIS$^{\mu}$ and SIS$^{\mu,\sigma}$

For the two settings of $\alpha^2 \in \{0.7, 0.99\}$ (corresponding to a weaker and stronger single index structure), we estimate p_l for $l \in \{3, 4, \ldots, 7\}$ using the five aforementioned methods. For each value of l, the optimal density is calibrated separately. In all examples, we use a sample size of $n = 10^6$ and a pilot sample size of 5×10^4. We repeat the experiment 200 times.

Figure 1 displays on the left the optimally calibrated and approximated IS densities. The true optimal density is bell shaped, so it is well approximated by a normal density. It can be confirmed from the plot that in this case, all IS densities seem to cover the important range. The right of Fig. 1 displays a boxplot of run-times needed to estimate p_l; note that the runtime does not depend on α or l. This plot, however, should be interpreted with caution as it highly depends on how the pilot runs are implemented.

Figure 2 displays mean relative errors; recall that we know p_l here. The relative errors for the different methods are similar, though SIS$^{\mu,\sigma}$ seems to give smallest errors. A possible explanation might be that the simplicity of that method (e.g., in terms of the support)

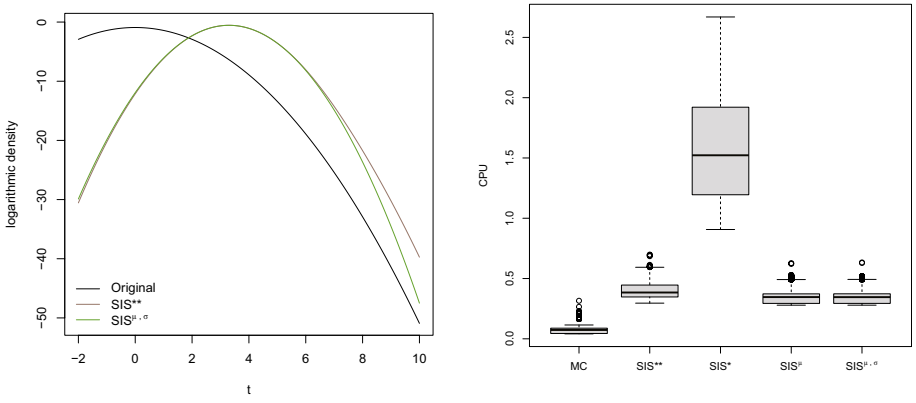

Fig. 1 Left: Calibrated densities for $\alpha = 0.99$, $l = 5$. Right: Run-times for each method including pilot runs

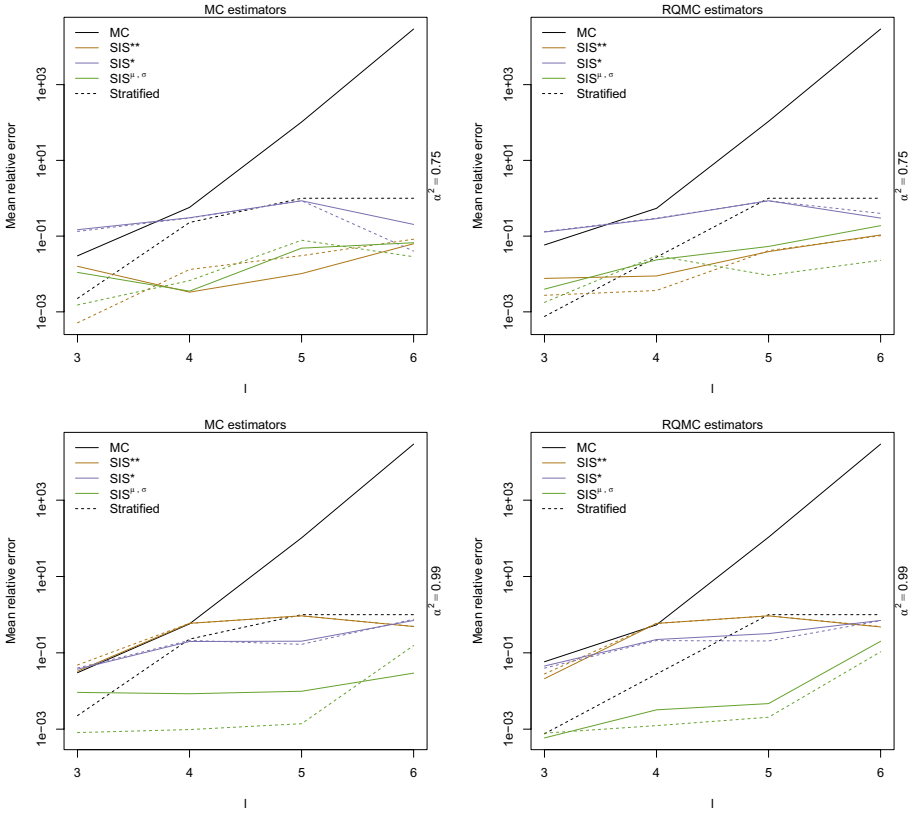

Fig. 2 Mean relative errors for the linear model example when using pseudo-random numbers (left) and quasi-random numbers (right) for $\alpha^2 = 0.75$ (top) and $\alpha^2 = 0.99$ (bottom). Dashed lines indicate the stratified estimators

relative to numerically constructing the optimal density via NINGL might outweigh the benefit of the latter having slightly more theoretical support. Furthermore, note that the IS methods perform much better when $R^2 = \alpha^2$ is larger, i.e., when the single index structure is strong, as expected.

4.2 Loss Distribution of a Credit Portfolio

In this section, we study the effectiveness of the proposed methods for a credit portfolio problem studied in Glasserman and Li (2005), where the goal is to estimate the probability of large portfolio losses under a normal copula model. We compare our proposed methods to the IS technique of Glasserman and Li, to which we refer to as G&L IS.

4.2.1 Problem Formulation

Suppose that Y_k denotes the default indicator of the kth obligor with exposure c_k and a default probability of p_k for $k = 1, \ldots, h$. The incurred loss is then $L = \sum_{k=1}^{h} c_k Y_k$. Let

$$Y_k = \mathbb{1}_{\{X_k > \Phi^{-1}(1-p_k)\}}, \quad X_k = a_{k1}Z_1 + \cdots + a_{kd}Z_d + b_k\varepsilon_k \sim \mathrm{N}(0,1), \quad k = 1, \ldots, h,$$

where

$$(Z_1, \ldots, Z_d) \sim \mathrm{N}_d(\mathbf{0}, I_d), \quad \varepsilon_1, \ldots, \varepsilon_h \overset{\text{ind.}}{\sim} \mathrm{N}(0,1), \quad \sum_{j=1}^{h} a_{kj}^2 \le 1,$$

and $b_k = \sqrt{1 - \sum_{j=1}^{h} a_{kj}^2}$. The a_{kj} represent the kth obligor's factor loadings for the d risky systematic factors; the choice of b_k ensures $X_k \sim \mathrm{N}(0,1)$. Our goal is to estimate $P(L > l)$ for large $l > 0$.

As in Glasserman and Li (2005), we consider a portfolio with $h = 1\,000$ obligors in a 10-factor model (i.e. $d = 10$). The marginal default probabilities and exposures are $p_k = 0.01 \cdot (1 + \sin(16\pi k/h))$ and $c_k = (\lceil 5k/h \rceil)^2$ for $k = 1, \ldots, h$, respectively. 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 $\mathrm{U}(0, 1/\sqrt{d})$. Letting $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_h)^\top$, we write $L = L(\boldsymbol{Z}, \boldsymbol{\varepsilon})$, i.e., the vector $\boldsymbol{X}$ to which we have referred throughout this paper is given by $\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{\varepsilon})$ for this example. We investigate whether or not L has a single-index structure. Let $T = \boldsymbol{\theta}^\top \boldsymbol{Z}$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\boldsymbol{\theta}^\top \boldsymbol{\theta} = 1$, so $T \sim \mathrm{N}(0,1)$. We estimate $\boldsymbol{\theta}$ that maximize the fit by using the average derivative method of Stoker (1986). The estimated $\boldsymbol{\theta}$ 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. The left side of Fig. 3 shows the scatter plot of (T, L). The figure reveals the single-index model fits L well even in the extreme tail, implying SIS based on this choice of T will give substantial variance reduction. The right side of Fig. 3 displays the original density of T, the optimally calibrated SIS* density as well as the estimated function $p_l(t)$. Note that the optimally calibrated density's mode substantially differs from the original one.

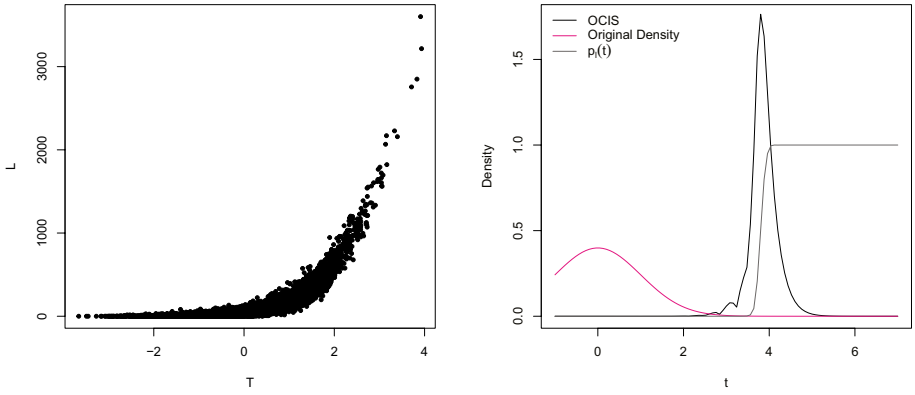

Fig. 3 Plot of Transformed variable (T) vs Portfolio Loss (L) based on 10 000 observations (left) and OCIS density calibrated to $l = 3000$ (right)

4.2.2 Proposed Estimators

The method of Glasserman and Li (2005) consists of a two-step procedure. In a calibration stage, an optimal mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ is found by solving an optimization problem minimizing the variance of the resulting IS estimator. Next, one samples $\boldsymbol{Z} \sim \mathrm{N}_d(\boldsymbol{\mu}, I_d)$ and computes the conditional default probabilities $p_k(\boldsymbol{Z}) = \mathbb{P}(Y_k = 1 \mid \boldsymbol{Z}) = \Phi((\boldsymbol{a}_k^\top \boldsymbol{Z} - x_k)/b_k)$, which enter another optimization problem used to find a number $\theta \in \mathbb{R}$ so that $q_k(\theta, \boldsymbol{Z})$ are variance minimizing default probabilities. Given $\boldsymbol{Z}$, we know that $Y_1, \dots, Y_h$ are independent and can therefore easily sample the loss via $L = \sum_{k=1}^h c_k \mathbb{1}_{\{U_k \le q_k(\theta(\boldsymbol{Z})\}}$ where $(U_1, \dots, U_h) \sim \mathrm{U}(0,1)^h$. Finally, the estimator $\mathbb{1}_{\{L>l\}} \cdot w(\boldsymbol{Z}, L)$ where w denotes the IS weight function is an unbiased estimator.

Our method $\mathrm{SIS}^{\mu,\sigma}$ proceeds as described in Algorithm 1; SIS^{μ} omits sets the scale to unity while the SSIS methods also stratify. Once $\boldsymbol{Z} \mid T$ is sampled, we sample Y_k from $p_k(Z)$ independently. We also include SIS^* and SSIS^*, where the function $p_l(t)$ is estimated as before and the quantile function of the optimal distribution is estimated via the NINiGL algorithm, in our experiments; see also Fig. 3.

4.2.3 Comparison

We compare SIS and SSIS to G&L IS by computing estimates, standard errors and computation times for $l \in \{100, 1000, 2000, 3000, 4000\}$. All methods require a calibration stage. For this comparison, we optimize the proposal distributions at each loss level of l separately and estimate the corresponding loss probability. Table 1 shows the estimated probabilities along with half-widths of estimated confidence intervals (CI) in brackets, Table 2 shows relative error reduction factors. The last column shows the average computational time of each method over all loss levels l. All examples used $n = 5000$ samples and 1000 samples for the calibration.

Table 1 Estimates and CI halfwidths when estimating p_l in the Gaussian Credit Portfolio problem with $h = 1000$ obligors and $d = 10$ factors for various l and methods. The last column displays average run-times

l	100	1000	2000	3000	4000	Avg runtime (sec)
G&L IS	0.28	0.0079	0.00077	9.2e-05	1.1e-05	2.45
	(0.0078)	(0.00036)	(4.1e-05)	(6.3e-06)	(8.8e-07)	
SIS^*	0.28	0.0081	0.00076	9.2e-05	1.1e-05	6.62
	(0.0068)	(0.00021)	(2.1e-05)	(2.4e-06)	(3.5e-07)	
SSIS^*	0.28	0.0082	0.00077	9.5e-05	1.1e-05	12.56
	(0.0046)	(0.00014)	(1.4e-05)	(1.7e-06)	(2.5e-07)	
SIS^{μ}	0.28	0.0077	0.00074	8.6e-05	1e-05	1.41
	(0.0086)	(0.00039)	(4.2e-05)	(5.5e-06)	(6.8e-07)	
SSIS^{μ}	0.28	0.008	0.00075	9.1e-05	1.1e-05	1.45
	(0.0062)	(0.00028)	(2.9e-05)	(4e-06)	(5.1e-07)	
$\mathrm{SIS}^{\mu,\sigma}$	0.28	0.0082	0.00077	9.4e-05	1.1e-05	2.45
	(0.0077)	(0.00034)	(3.3e-05)	(5.2e-06)	(4.6e-07)	
$\mathrm{SSIS}^{\mu,\sigma}$	0.28	0.0081	0.00075	8.9e-05	1.1e-05	2.2
	(0.0059)	(2e-04)	(1.9e-05)	(2.3e-06)	(3e-07)	

Table 2 Relative error reduction factors RE(MC)/RE(RQMC) for the Gaussian credit portfolio with $h = 1000$ obligors and $d = 10$ factors for various l and methods

l	100	1000	2000	3000	4000
G&L IS	1.5	1.5	1.5	1.5	1.6
SIS*	1.3	1.3	1.2	1.7	1.5
SIS$^{\mu,\sigma}$	1.6	1.5	1.9	1.7	1.6
SSIS$^{\mu,\sigma}$	1	1.1	1	1.1	0.9

We see that all our methods lead standard errors smaller than G&L IS, while the estimated CIs for both methods are typically overlapping, supporting the correctness of both approaches. Given the small run-time, unbiasedness and small estimated errors, we can conclude that SSIS$^{\mu,\sigma}$ is the best estimator for this problem. This supports our claim that the optimal density of T can be quickly and accurately approximated by a location scale transform of f_T. Note that SIS* and SSIS* are particularly slow, as it involves numerically approximation the quantile function corresponding to the optimal g_T.

4.3 Tail Probabilities of a *t*-Copula Credit Portfolio

In this section, we apply SIS to a credit portfolio problem under a t-copula model, which is the model studied in Sect. 4.2 with a multiplicative shock variable included. This t-copula model is a special case of the models with extremal dependence studied in Bassamboo et al. (2008). Unlike the Gaussian copula, the t-copula is able to model tail dependence of latent variables, so simultaneous defaults of many obligors are more likely under the t-copula model than under its Gaussian copula counterpart.

4.3.1 Problem Formulation

In the t-copula model, the latent variables $\boldsymbol{X} = (X_1, \dots, X_d)$ are multivariate-t distributed, that is,

$$X_k = \sqrt{W}(a_{k1}Z_1 + \cdots + a_{kd}Z_d + b_k\varepsilon_k), \quad k = 1, \dots, h,$$

where $W \sim \mathrm{IG}(\nu/2, \nu/2)$ is independent of $Z_1, \dots, Z_d, \varepsilon_k \overset{\text{ind.}}{\sim} \mathrm{N}(0,1)$. Accordingly, we define $Y_k = \mathbb{1}_{\{X_k > t_\nu^{-1}(1-p_k)\}}$. We assume the same parameters as in Sect. (4.2.1), except that now we have $h = 50$ obligors, and the two settings for the degrees-of freedom $\nu \in \{5, 12\}$. Let $\boldsymbol{Z} = (Z_1, \dots, Z_d)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_h)$. We consider two transformations. For the first transformation, let $Z_W = \Phi^{-1}(F_W(W))$ and

$$T_1(W, \boldsymbol{Z}, \boldsymbol{\epsilon}) = \beta_W Z_W + \boldsymbol{\beta}_L^\top \boldsymbol{Z},$$

where $\beta_W \in \mathbb{R}$ and $\boldsymbol{\beta}_L \in \mathbb{R}^d$ are such that $\beta_W^1 + \boldsymbol{\beta}_L^\top \boldsymbol{\beta}_L = 1$. Then, $T_1 \sim \mathrm{N}(0,1)$ since $Z_W \sim \mathrm{N}(0,1)$ is independent of $\boldsymbol{Z}$.

Our second transformation relies on the random variable $S_l(\boldsymbol{Z}, \boldsymbol{\varepsilon}) = \mathbb{P}(L > l \mid \boldsymbol{Z}, \boldsymbol{\varepsilon})$ and note that $\mathbb{P}(L > l) = \mathbb{E}(S_l(\boldsymbol{Z}, \boldsymbol{\varepsilon}))$. Based on this and the fact that, given a sample $\boldsymbol{Z}, \boldsymbol{eps}$, the function S_l can be computed analytically, Chan and Kroese (2010) propose to use CMC, i.e.,

estimating $\mathbb{P}(L > l)$ by the sample mean of $S_l(\boldsymbol{Z}_i, \boldsymbol{\epsilon}_i)$ for independent $\boldsymbol{Z}_i, \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$. We propose to use this CMC idea combined with SIS by using the transformation

$$T_2 = \boldsymbol{\beta}_S^\top \boldsymbol{Z}$$

with $\boldsymbol{\beta}_S$ such that $\boldsymbol{\beta}_S^\top \boldsymbol{\beta}_S = 1$, which implies $T_2 \sim \mathrm{N}(0, 1)$.

The second method based on CMC, is very effective as the variable W which accounts for a large portion of the variance of L, is integrated out. Furthermore, Chan and Kroese (2010) additionally employ IS on $(\boldsymbol{Z}, \boldsymbol{\epsilon})$ to make the event $\{L > l\}$ more frequent using the cross-entropy method; see De Boer et al. (2005); Rubinstein (1997); Rubinstein and Kroesse (2013). We refer to Chan and Kroese's method as C&K CMC+IS. The numerical study in Chan and Kroese (2010) demonstrates that C&K CMC+IS achieves substantial variance reduction. We will show in our numerical examples below that combining their CMC idea with our proposed single index IS method gives even greater variance reduction.

4.3.2 Fit of Single-Index Models with and Without Conditional Monte Carlo

We first investigate whether or not L and S_l have single-index structures. As before, the coefficients $\boldsymbol{\beta}$ that maximize the fit of the single-index model are estimated using the average derivative method of Stoker (1986).

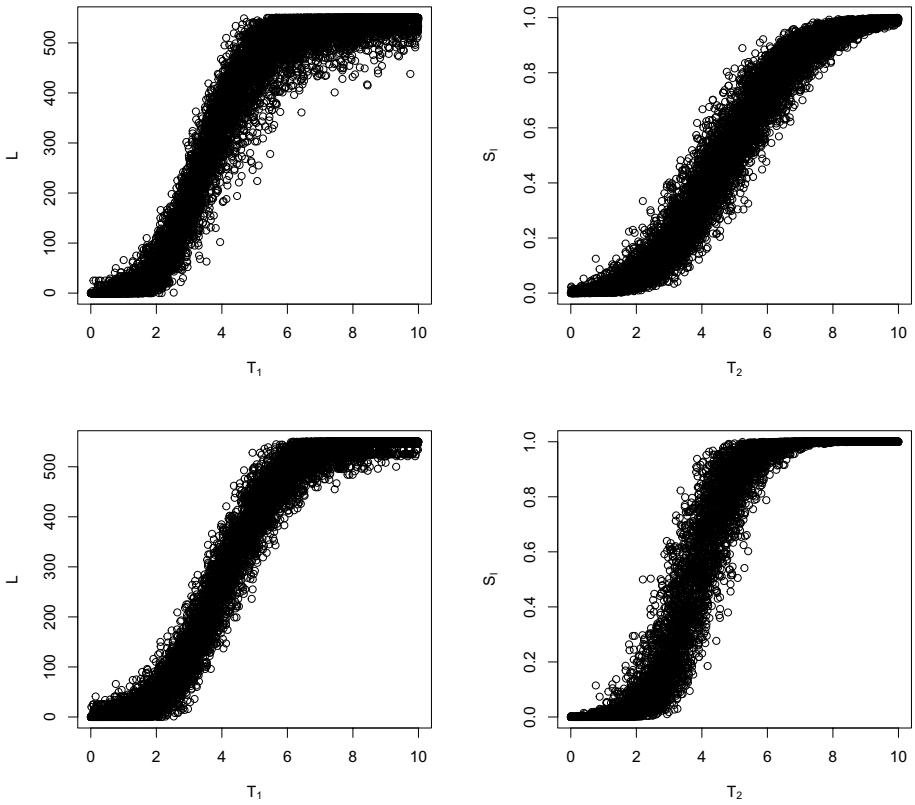

Fig. 4 Scatter plots of L vs T_1 (left) and S_l vs T_2 (right) where $l = 500$ and $\nu = 5$ (top) and $\nu = 12$ (bottom)

Figure 4 shows scatter plots of (T_1, L) and (T_2, S_l) for $\nu = 12$ and $\nu = 5$. 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 (i.e., the stronger the dependence between the $\boldsymbol{X}_i$), the worse the fit of the single-index model becomes in the right-tail. When investigating the fit of (T_2, S_l), recall that the main advantage of CMC is that W is integrated out; the resulting estimators should be less sensitive to the degrees-of-freedom ν, which is the case in the plot. We can see that the fit of T_2 is excellent even in the outer right-tail for all settings of ν and l.

4.3.3 Estimates and Estimated Variances

We compare the original C&K CMC+IS from Chan and Kroese (2010) with SIS with and without CMC. We additionally investigate whether employing RQMC yields a variance reduction. To this end, we estimate p_l for $l = 100$ for various n and methods; see Figs. 5

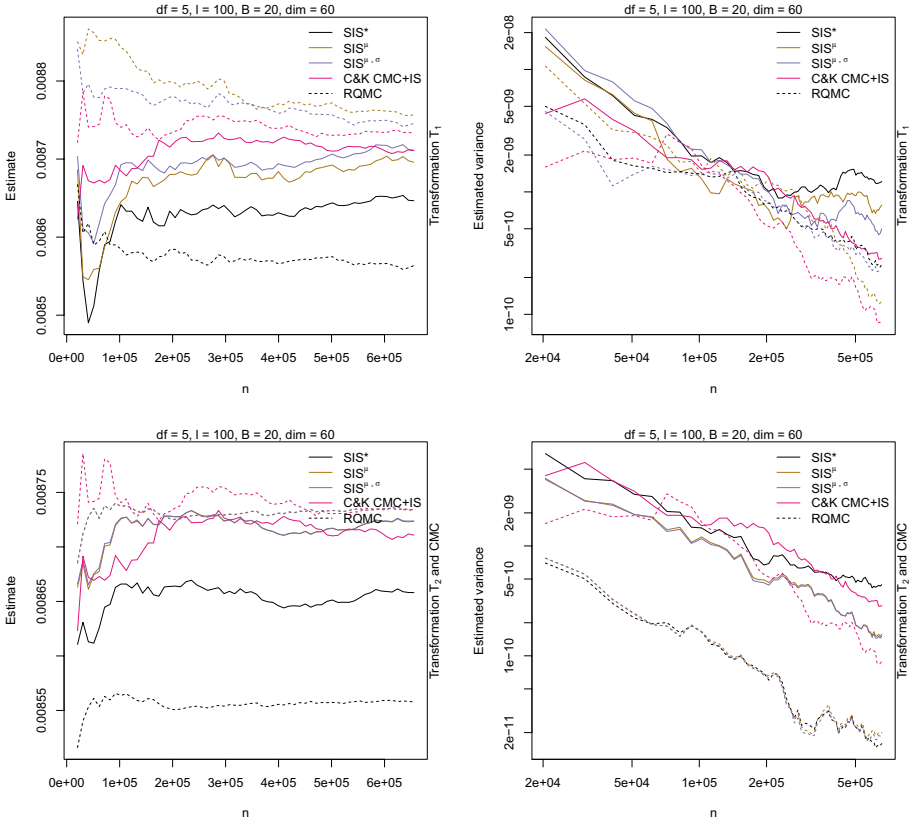

Fig. 5 Estimates (left) and estimated variances (right) for the t-copula credit problem for $\nu = 5$ with transformation T_1 (top) and transformation T_2 (bottom). Dashed lines indicate RQMC estimators

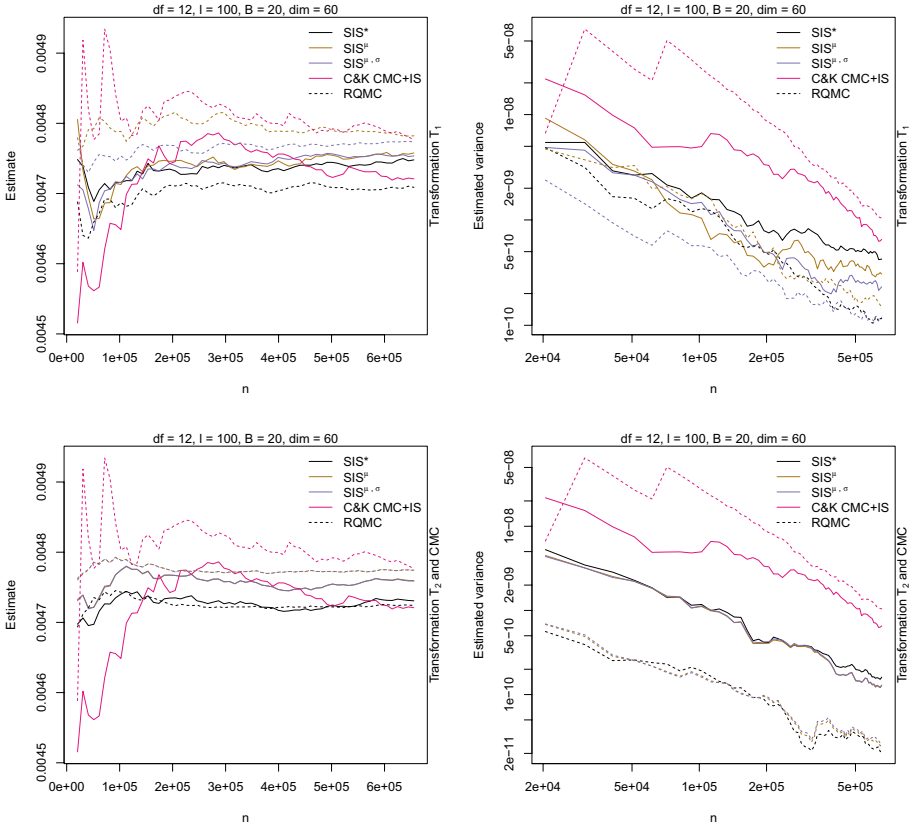

Fig. 6 Estimates (left) and estimated variances (right) for the t-copula credit problem for $\nu = 12$ with transformation T_1 (top) and transformation T_2 (bottom). Dashed lines indicate RQMC estimators

and 6. Variances are estimated as the sample variance of $B = 20$ repetitions; this ensures that the same variance estimator (namely, the sample variance) is used for both methods, rather than using the estimator from Proposition 3 for MC and the sample variance for RQMC.

Note that for fixed ν, the data for C&K CMC+IS are identical independent of which transformation is used, so these lines can be used as reference. As expected, variances with the CMC idea are smaller than without the CMC idea. Note further that all our (S)SIS methods combined with T_1 (which does not integrate out W) give smaller variances than C&K CMC+IS, which does integrate out W.

5 Concluding Remarks

In this paper, we developed importance sampling and stratification techniques that are designed to work well for problems with a single-index structure, i.e., where the response variable depends on input variables mostly through some one-dimensional transformation.

The main theme of our approach is to exploit the low-dimensional structure of a given problem in rare-event simulation by introducing a conditional sampling step on this important transformed random variable and using optimal IS.

We derived expressions for optimal densities of said one-dimensional transformation which achieve minimum variance and discussed boundary cases with zero variance. Furthermore, we demonstrated that our framework includes and generalizes existing mean-shifting techniques. Our theoretical framework and numerical examples suggest substantial variance reduction for problems having strong single-index structures. As the optimal density rarely belongs to a known parametric family, we also give explicit steps to calibrate the proposal distribution.

Our numerical experiments revealed that the proposed methods outperform existing methods that were specifically tailored to the Gaussian and t-copula credit portfolio problem. The success of our method in this framework highlights the flexibility and wide applicability of our approach.

By combining our single-index framework with RQMC methods, we achieve even more precise estimation results, thanks to the dimension reduction feature of our conditional sampling step.

Note that there exist many other low-dimensional structures studied in the literature and they may provide a better fit than single-index models do. For instance, the structure assumed by the sufficient dimension reduction can be seen as a multi-index extension of the linear single-index model; see Cook (1998); Cook and Forzani (2009); Adragni and Cook (2009). We would like to develop importance sampling techniques for problems based on other low-dimensional structures in future research.

Appendix

Proofs

Proof of Proposition 1 The mean and variance follow from

$$\mathbb{E}_g(\hat{\mu}_n^{\mathrm{SIS}}) = E_g(\Psi(\boldsymbol{X})w(T)) = \mathbb{E}_g(m(T)w(T)) = \int_{\Omega_g} m(t)\frac{f_T(t)}{g_T(t)}g_T(t)\,\mathrm{d}t = \mu_{\mathrm{SIS}}$$

and

$$n\mathrm{Var}_g(\hat{\mu}_n^{\mathrm{SIS}}) + \mu_{\mathrm{SIS}}^2 = \mathbb{E}_g\left(\Psi^2(\boldsymbol{X})w(T)\right) = \int_{\Omega_g} m^{(2)}(t)\frac{f_T^2(t)}{g_T^2(t)}\,\mathrm{d}t.$$

Asymptotic normality follows from the central limit theorem. Next, we need to find g_T among all g that give unbiased estimators so that the variance, or equivalently $\mathbb{E}_g(m^{(2)}(T)w(T))$, is minimal when $\Psi(\boldsymbol{x}) \geq 0$ or $\Psi(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \Omega$. Let $\Omega_{\mathrm{ub}} = \{t \in \Omega_f : m(t)f_T(t) \neq 0\}$. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}_g\left(m^{(2)}(T)w^2(T)\right) &\geq \left(\mathbb{E}_g\left(\sqrt{m^{(2)}(T)}w(T)\right)\right)^2 \\ &= \left(\int_{\Omega_g}\sqrt{m^{(2)}(t)}w(t)\,\mathrm{d}t\right)^2 = \left(\int_{\Omega_f}\sqrt{m^{(2)}(t)}w(t)\,\mathrm{d}t\right)^2\end{aligned}$$

The last inequality follows since $\hat{\mu}_n^{\text{SIS}}$ is assumed to be unbiased, i.e., $\Omega_{\text{ub}} \subseteq \Omega_g$ and the fact that $\sqrt{m^{(2)}(t)} f_T(t) = 0$ for $t \notin \Omega_{\text{ub}}$ (as $m(t) = 0$ implies $m^{(2)}(t) = 0$ by the assumption on Ψ). The right hand side of the inequality is a constant independent of the choice of g_T, namely the minimum variance among all SIS estimators. To achieve equality, or equivalently to minimize the variance, set $g_T \propto \sqrt{m^{(2)}(t)} f_T(t)$ for $t \in \Omega_{\text{ub}}$ and the claim follows.

Proof of Proposition 2 Let $\Omega_T^{(i)} = \{t \in (t_{\text{inf}}, t_{\text{sup}}) : \lambda_i \le t < \lambda_{i+1}\}$ where $\lambda_i = G_T^{\leftarrow}((i+1)/n)$ and note that $\mathbb{P}(T \in \Omega_T^{(i)}) = 1/n$ for $i = 1, \ldots, n$. Then

$$\begin{aligned}\mathbb{E}(\hat{\mu}_n^{\text{SSIS}}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_g\Big(\Psi(\boldsymbol{X})w(T) \mid T \in \Omega_T^{(i)}\Big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_g\Big(\mathbb{E}_g(\Psi(\boldsymbol{X})w(T) \mid T) \mid T \in \Omega_T^{(i)}\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{\lambda_i}^{\lambda_{i+1}} m(t)\frac{f_T(t)}{g_T(t)} g_T(t)\,\mathrm{d}t = \mu_{\text{SIS}}\end{aligned}$$

The expression for the variance is a slight generalization of (Glasserman et al. (1999), Lemma 4.1) in that stratification is combined with IS, bit it can be proved similarly. Let $\eta_n(t)$ denote the index i so that $t \in \Omega_T^{(i)}$. Then

$$n\text{Var}(\hat{\mu}_n^{\text{SSIS}}) = \frac{1}{n}\sum_{j=1}^{n} \text{Var}_g\Big(\Psi(\boldsymbol{X})w(T) \mid T \in \Omega_T^{(i)}\Big) = \mathbb{E}_g\Big(\text{Var}_g\big(\Psi(\boldsymbol{X})w(T) \mid \eta_n(T)\big)\Big).$$

Let $\xi = \mathbb{E}_g(\Psi(\boldsymbol{X})w(T) \mid T) = m(T)w(T)$ and define the sequence $\xi_n = \mathbb{E}_g(\xi \mid \eta_n(T))$. Note that the $\sigma-$algebra generated by $\eta_n(T)$ forms an increasing family as n increases through a constant multiple of power two. Observe that $\mathbb{E}_g(|xi|) < \infty$ and $\sup_n \xi_n < \mathbb{E}_g(\Psi(\boldsymbol{X})w^2(T)) = \mathbb{E}_g(m^{(2)}(T)w^2(T)) < \infty$. Also, ξ_n is a martingale if n increases through a constant multiple of powers of two as it is a Doob's martingale; (see Karlin and Taylor (1975), p. 246). Then using the arguments as in (Glasserman et al. (1999), Lemma 4.1), it follows that $\text{Var}_g(\hat{\mu}_n^{\text{SSIS}}) = \sigma_{\text{SIS}}^2/n + o(1)$.

The expression for the optimal density and variance expressions follow as in the proof of Prop. 1 by applying Jensen's inequality. It remains to show that the SSIS estimator is asymptotically normal, which we show by applying the Lyapunov Central Theorem; (see Kole et al. (2007), p. 134). Let $m_i = \mathbb{E}_g(\Psi(\boldsymbol{X})w(T) \mid T \in \Omega_T^{(i)})$ and $v_i^2 = \text{Var}_g(\Psi(\boldsymbol{X})w(T) \mid T \in \Omega_T^{(i)})$. It is easily seen that $(1/n)\sum_{i=1}^n m_i = \mu_{\text{SIS}}$ and $(1/n)\sum_{i=1}^n v_i^2 = \sigma_{\text{SIS}}^2 + o(1)$. For any $i = 1, \ldots, n$, we have

$$\begin{aligned}&\mathbb{E}_g\big(|\Psi(\boldsymbol{X}_i)w(T_i) - m_i|^{2+\delta}\big) \le 2^{2+\delta}\Big(\mathbb{E}_g\big(|\Psi(\boldsymbol{X}_i)w(T_i)|^{2+\delta}\big) + \mathbb{E}_g\big(|m_i|^{2+\delta}\big)\Big) \\ &= 2^{2+\delta}\Big(\mathbb{E}_g\Big(|\Psi(\boldsymbol{X})w(T)|^{2+\delta} \mid T \in \Omega_T^{(i)}\Big) + \mathbb{E}_g\Big(|\mathbb{E}_g(\Psi(\boldsymbol{X})w(T) \mid T \in \Omega_T^{(i)})|^{2+\delta}\Big)\Big) \\ &\le 2^{2+\delta}\Big(\mathbb{E}_g\Big(|\Psi(\boldsymbol{X})w(T)|^{2+\delta} \mid T \in \Omega_T^{(i)}\Big) + \mathbb{E}_g\Big(\mathbb{E}_g(|\Psi(\boldsymbol{X})w(T)|^{2+\delta} \mid T \in \Omega_T^{(i)})\Big)\Big) \\ &= 2^{3+\delta}\mathbb{E}_g\Big(|\Psi(\boldsymbol{X})w(T)|^{2+\delta} \mid T \in \Omega_T^{(i)}\Big),\end{aligned}$$

where the first inequality follows from the c_τ inequality as in (Loeve (1963), p. 155). The Lyapunov condition is satisfied, since

$$\begin{aligned}
&\frac{1}{(\sum_{i=1}^n \sigma_i^2)^{1+\delta/2}} \sum_{i=1}^n \mathbb{E}_g\left(|\Psi(\boldsymbol{X}_i)w(T_i) - m_i|^{2+\delta}\right) \\
&\leq \frac{2^{3+\delta}}{(\sum_{i=1}^n \sigma_i^2)^{1+\delta/2}} \sum_{i=1}^n \mathbb{E}_g\left(|\Psi(\boldsymbol{X}_i)w(T_i)|^{2+\delta} \mid T \in \Omega_T^{(i)}\right) \\
&= \frac{2^{3+\delta} n}{(n\sigma_{\text{SSIS}}^2 + o(n))^{1+\delta/2}} \mathbb{E}_g\left(|\Psi(\boldsymbol{X})w(T)|^{2+\delta}\right) \to 0, \quad n \to \infty,
\end{aligned}$$

by the assumption. The Lyapunov Central Limit Theorem together with Slutsky's Theorem implies $(\hat{\mu}_n^{\text{SSIS}} - \mu_{\text{SIS}})/\sqrt{n} \xrightarrow{\text{d}} \text{N}(0, \sigma_{\text{SSIS}}^2)$.

Proof of Proposition 3 Recall that T_i satisfies $T_i = G_T^{\leftarrow}((i + U_i - 1)/n)$ where $U_i \overset{\text{ind.}}{\sim} \text{U}(0,1)$ for $i = 1, \ldots, n$, and are therefore ordered, i.e., $T_1 < T_2 < \cdots < T_n$. For any $i = 1, \ldots, n$,

$$T_{i+1} - T_i = (G_T^{-1})'(\xi_i)\left(\frac{1 + U_{i+1} - U_i}{n}\right) = \frac{1}{g_T(G_T^{-1}(\xi_i))}\left(\frac{1 + U_{i+1} - U_i}{n}\right) = \mathcal{O}(1/n),$$

for some $\xi_i \in (T_i, T_{i+1})$, which implies that for any continuously differentiable function h, $h(T_{i+1}) = h(T_i) + \mathcal{O}(1/n)$. Then we have

$$\begin{aligned}
r_i^2 &= \left(m(T_{i+1}) + \varepsilon_{T_{i+1}} - m(T_i) - \varepsilon_{T_i}\right)^2 \\
&= \left(m(T_{i+1}) - m(T_i)\right)^2 + \left(\varepsilon_{T_{i+1}} - \varepsilon_{T_i}\right)^2 - 2(m(T_{i+1}) - m(T_i))(\varepsilon_{T_{i+1}} - \varepsilon_{T_i}) \\
&= (\varepsilon_{T_{i+1}} - \varepsilon_{T_i})^2 - 2(m(T_{i+1}) - m(T_i))(\varepsilon_{T_{i+1}} - \varepsilon_{T_i}) + \mathcal{O}(1/n^2),
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{E}_g\left(r_i^2 w^2(T_i)\right) &= \mathbb{E}_g\left(\mathbb{E}_g(r_i^2 w(T_i) \mid T_i, T_{i+1})\right) = \mathbb{E}_g\left(w^2(T_i)(v^2(T_i) + v^2(T_{i+1}))\right) + \mathcal{O}(1/n^2) \\
&= 2\mathbb{E}_g\left(w^2(T_i)v^2(T_i)\right) + \mathcal{O}(1/n),
\end{aligned}$$

which means that

$$\begin{aligned}
\mathbb{E}_g(\hat{\sigma}_{\text{SSIS}}^2) &= \frac{1}{2(n-1)} \sum_{i=1}^n \mathbb{E}_g(r_i^2 w^2(T_i)) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_g\left(v^2(T)w^2(T) \mid T \in \Omega_T^{(i)}\right) + \mathcal{O}(1/n) \\
&= \mathbb{E}_g(v^2(T)w^2(T)) + \mathcal{O}(1/n) = \sigma_{\text{SSIS}}^2 + \mathcal{O}(1/n) \to \sigma_{\text{SSIS}}^2,
\end{aligned}$$

which shows consistency.

Proof of Proposition 4 We use that $(\boldsymbol{X} \mid T = t) \sim \text{N}_d(\boldsymbol{\beta}t, I_d - \boldsymbol{\beta}\boldsymbol{\beta}^\top)$(see Harris and Helvig (1965), Theorem 1) to compute the moment generating function of $\boldsymbol{X}$. For $\boldsymbol{a} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}_g(\mathbb{E}_g(\exp(\boldsymbol{a}^\top \boldsymbol{X})) &= \mathbb{E}_g\left(\mathbb{E}(\exp(\boldsymbol{a}^\top \boldsymbol{X}) \mid T)\right) = \mathbb{E}_g\left(\exp\left(\boldsymbol{a}^\top \boldsymbol{\beta} T + \frac{1}{2}\boldsymbol{a}^\top (I_d - \boldsymbol{\beta}\boldsymbol{\beta}^\top)\boldsymbol{a}\right)\right) \\
&= \mathbb{E}_g(\exp(\boldsymbol{a}^\top \boldsymbol{\beta} T)) \exp\left(\frac{1}{2}\boldsymbol{a}^\top (I_d - \boldsymbol{\beta}\boldsymbol{\beta}^\top)\boldsymbol{a}\right) = \exp\left(c\boldsymbol{a}^\top \boldsymbol{\beta} + \frac{1}{2}(\boldsymbol{a}^\top \boldsymbol{\beta})^2 \sigma^2\right) \times \\
&\times \exp\left(\frac{1}{2}\boldsymbol{a}^\top (I_d - \boldsymbol{\beta}\boldsymbol{\beta}^\top)\boldsymbol{a}\right) = \exp\left(\boldsymbol{a}^\top (c\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{a}^\top (I_d + (\sigma^2 - 1)\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{a}\right).
\end{aligned}$$

By uniqueness of the moment generating function, $\boldsymbol{X} \sim \mathrm{N}_d(c\boldsymbol{\beta}, I_d + (\sigma^2 - 1)\boldsymbol{\beta}\boldsymbol{\beta}^\top)$.

Acknowledgements The second and third author would like to thank NSERC for financial support for this work through Discovery Grant RGPIN-5010-2015 and Grant RGPIN-238959, respectively. We also thank an anonymous referee for their insightful comments which helped improve this paper.

Data Availability All numerical examples presented in this paper can be reproduced with an R script available from the corresponding author upon request.

Declarations

Conflicts of Interest The authors declare no conflicts of interest.

References

Adragni K, Cook R (2009) Sufficient dimension reduction and prediction in regression. Phil Trans Math Phys Eng Sci 367(1906):4385–4405

Arbenz P, Cambou M, Hofert M, Lemieux C, Taniguchi Y (2018) Importance sampling and stratification for copula models. Contemporary Computational Mathematics–a celebration of the 80th birthday of Ian Sloan. Springer

Asmussen S, Glynn P (2007) Stochastic Simulation: Algorithms and Analysis. Springer, Berlin

Au S, Beck J (2003) Important sampling in high dimensions. Struct Saf 25(2):139–163

Bassamboo A, Juneja S, Zeevi A (2008) Portfolio credit risk with extremal dependence: Asymptotic analysis and efficient simulation. Oper Res 56(3):593–606

Caflisch R, Morokoff W, Owen A (1997) Valuation of Mortgage Backed Securities Using Brownian Bridges to Reduce Effective Dimension. Department of Mathematics, University of California, Los Angeles. https://doi.org/10.21314/JCF.1997.005

Cambou M, Hofert M, Lemieux C (2016) Quasi-random numbers for copula models. Stat Comput 27(5):1307–1329. https://doi.org/10.1007/s11222-016-9688-4

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

Cook R (1998) Regression Graphics. Wiley, New York

Cook R, Forzani L (2009) Likelihood-based sufficient dimension reduction. J Am Stat Assoc 104(485):197–208

Cochran W (2005) Sampling Techniques, 3rd edn. Wiley, New York

De Boer P, Kroese D, Mannor S, Rubinstein R (2005) A tutorial on the cross-entropy method. Ann Oper Res 134(1):19–67

Dick J, Pillichshammer F (2010) Digital Nets and Sequences: Discrepancy Theory and quasi-Monte Carlo Integration. Cambridge University Press, Cambridge

Glasserman P, Heidelberger P, Shahabuddin P (1999) Asymptotically optimal importance sampling and stratification for pricing path-dependent options. Math Financ 9(2):117–152

Glasserman P, Heidelberger P, Shahabuddin P (2000) Variance reduction techniques for estimating Value-at-Risk. Manage Sci 46(10):1349–1364

Glasserman P, Heidelberger P, Shahabuddin P (2002) Portfolio value-at-risk with heavy-tailed risk factors. Math Financ 12(3):239–269

Glasserman P, Li J (2005) Importance sampling for portfolio credit risk. Manage Sci 51(11):1643–1656

Härdle W, Hall P, Ichimura H (1993) Optimal smoothing in single-index models. Ann Stat 21(1):157–178

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

Hörmann W, Leydold J (2003) Continuous random variate generation by fast numerical inversion. ACM Trans Model Comput Simul 13(4):347–362

Ichimura H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J Econ 58(1–2):71–120

Kahn H, Marshall A (1953) Methods of reducing sample size in Monte Carlo computations. J Oper Res Soc Am 1(5):263–278

Karlin S, Taylor H (1975) A First Course in Stochastic Processes vol. 1. Gulf Professional Publishing, Houston
Kole E, Koedijk K, Verbeek M (2007) Selecting copulas for risk management. J Bank Financ 31(8):2405–2423
Katafygiotis L, Zuev K (2008) Geometric insight into the challenges of solving high-dimensional reliability problems. Probab Eng Mech 23(2–3):208–218
Kvalseth T (1985) Cautionary note about R^3. Am Stat 39(4):279–285
Lavenberg S, Welch P (1981) A perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo simulations. Manage Sci 27(3):322–335
Lemieux, C.: Monte Carlo and Quasi-Monte Carlo Sampling. Springer, Berlin (2009). https://doi.org/10.1007/978-0-387-78165-5
Leydold J, Hörmann W (2020) Runuran: R Interface to the 'UNU.RAN' Random Variate Generators. R package version 0.30. https://CRAN.R-project.org/package=Runuran
Li K (1991) Sliced inverse regression for dimension reduction. J Am Stat Assoc 86(414):316–327
Loeve M (1963) Probability Theory. Van Nostrand, New York
Niederreiter H (1978) Quasi-Monte Carlo methods and pseudo-random numbers. Bull Am Math Soc 84(6):957–1041
Neddermeyer J (2011) Non-parametric partial importance sampling for financial derivative pricing. Quantitative Finance 11(8):1193–1206
Powell JL, Stock JH, Stoker TM (1989) Semiparametric estimation of index coefficients. Econometrica 1403–1430
R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. http://www.R-project.org
Reinsch C (1967) Numer Math 10(3):177–183
Rubinstein R (1997) Optimization of computer simulation models with rare events. Eur J Oper Res 99(1):89–112
Rubinstein R, Kroesse D (2013) The Cross-entropy Method: a Unified Approach to Combinatorial Optimization. Monte-Carlo Simulation and Machine Learning. Springer, Berlin
Sobol' I (1967) On the distribution of points in a cube and the approximate evaluation of integrals. USSR Comput Math Math Phys 7(4):86–112. https://doi.org/10.1016/0041-5553(67)90144-9
Sak H, Hörmann W, Leydold J (2010) Efficient risk simulations for linear asset portfolios in the *t*-copula model. Eur J Oper Res 202(3):802–809
Stoker T (1986) Consistent estimation of scaled coefficients. Econometrica 54(6):1461–1481
Schüeller G, Pradlwarter H, Koutsourelakis P (2004) A critical appraisal of reliability estimation procedures for high dimensions. Probab Eng Mech 19(4):463–474
Wang X, Fang K (2003) The effective dimension and quasi-Monte Carlo integration. J Complex 19(2):101–124. https://doi.org/10.1016/S0885-064X(03)00003-7
Wang X, Sloan I (2005) Why are high-dimensional finance problems often of low effective dimension? SIAM J Sci Comput 27(1):159–183. https://doi.org/10.1137/S1064827503429429
Wang X (2006) On the effects of dimension reduction techniques on some high-dimensional problems in finance. Oper Res 54(6):1063–1078
Wang L, Brown L, Cai T, Levine M (2008) Effect of mean on variance function estimation in nonparametric regression. Ann Stat 646–664