Efficient Simulation of Value at Risk with Heavy-Tailed Risk Factors

Cheng-Der Fuh
Graduate Institute of Statistics, National Central University, Jhong-Li and Institute of Statistical Science, Academia Sinica, Taipei, stcheng@stat.sinica.edu.tw

Inchi Hu
Department of ISOM, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, imichu@ust.hk

Ya-Hui Hsu
Abbott Laboratories, USA, ya-hui.hsu@abbott.com

Ren-Her Wang
Department of Banking and Finance, Tamkang University, Taipei, d92723009@ntu.edu.tw

Simulation of small probabilities has important applications in many disciplines. The probabilities considered in value-at-risk (VaR) are moderately small. However, the variance reduction techniques developed in the literature for VaR computation are based on large deviations methods, which are good for very small probabilities. Modeling heavy-tailed risk factors using multivariate t distributions, we develop a new method for VaR computation. We show that the proposed method minimizes the variance of the importance sampling estimator exactly, while previous methods produce approximations to the exact solution. Thus, the proposed method consistently outperforms existing methods derived from large deviations theory under various settings. The results are confirmed by a simulation study.

Key words: Importance sampling, moderate deviation, multivariate t distribution, quadratic approximation, component VaR.

1. Introduction

We consider the problem of estimating small probabilities using Monte Carlo simulations. Such problems appear in the computation of value-at-risk (VaR) in risk management; cf. ? and ?. The profit-and-loss distribution of a portfolio is of central importance to risk management and VaR is a quantile of this distribution. Since quantiles on the left tail of the profit-and-loss distribution correspond to large losses, it is necessary to estimate the tail probabilities and corresponding quantiles as risk measures.
To measure VaR, we need to determine two quantitative parameters: the length of the time horizon and the level of confidence. For example, the internal model approach of the Basel Committee imposes a 99-percent confidence level over a 10-business-day horizon. The time horizon may be set in terms of hours, days or weeks. Traders, who handle many portfolios, adjust positions immediately based on intraday data. For bank managers, the regulator horizon is two weeks based on the Basel Accord. For investment managers, the time horizon may match the regular reporting period, monthly or quarterly.

With the advancement of structural financial products and derivatives markets, the risk level of a portfolio or a business enterprise can change substantially in a short horizon, say, within a day. Risk assessment has to be done frequently, especially under high financial instability. Therefore, there is a growing need to compute intraday risk measures, i.e. VaR of a portfolio in less than one trading day, using high frequency data. Because VaR is computed frequently and a large amount of computation is involved each time, even a moderate efficiency gain in computation may impact the risk management practice.

Although Monte Carlo simulation is the most powerful method to evaluate portfolio VaR, a major drawback of this method is that it is computationally demanding. There are two reasons. First, the portfolio may consist of many financial instruments and the valuation of an individual instrument may itself require considerable amount of computation. Second, a large number of runs are necessary to obtain accurate estimates corresponding to small tail probabilities of the loss distribution. For any random variable $T$, let $P$ and $Q$ be the probability distributions of $T$ under the probability measures $\mathcal{P}$ and $\mathcal{Q}$, respectively. It is well-known that importance sampling, where one uses observations from an alternative distribution $Q$ to estimate the target distribution $P$, is capable of dramatic efficiency improvement in estimating small probabilities. The goal of this paper is to develop an efficient importance sampling method suitable for VaR computation.

The existing literature on importance sampling for VaR is based on large deviations theory; see e.g. Glasserman, Heidelberger, and Shahabuddin (2000, 2002), which we shall refer to as the GHS method henceforth. The GHS method uses a quadratic approximation to the portfolio loss and
models changes in risk factors by multivariate normal (Glasserman et al., 2000) or multivariate 
*t* (Glasserman et al., 2002) distributions to design the importance sampling procedure. Although 
large deviations theory has been applied successfully in statistical mechanics, quantum mechanics, 
information theory, and risk management, it is most effective for rare events with probabilities in 
the order of $10^{-6}$ or smaller. Since the confidence levels of VaR are usually between 95% and 99%, 
they are events with moderately small probabilities of occurrence. For such events, the mean of 
the optimal alternative distribution is usually in the interior of the event, which differs from the 
dominating point of large deviations theory (Ney, 1983), which is located at the boundary of the 
event. Hence, for VaR computation, the proposed method can be significantly more efficient than 
methods based on large deviations.

Efficient importance sampling has been studied by ?, ?, ?, and Fuh and Hu (2004, 2007). However, 
those papers concern multivariate normal distributions. Because many empirical studies suggest 
that the distributions of risk factors have heavy tails, the aforementioned methods cannot be real-
istically applied to VaR computation. Furthermore, as pointed out by ?, the successful application 
of importance sampling in heavy-tailed settings is a notoriously difficult problem. As the main 
contribution of this paper, we use multivariate *t* distributions to model the heavy-tailed risk fac-
tors and develop an efficient importance sampling method for moderate deviations events, which 
is appropriate for VaR computation.

The rest of this paper is organized as follows. In Section 2, we formulate an optimization problem 
that minimizes the variance of the importance sampling estimator for VaR computation. Section 
3 contains an expression that characterizes the optimal alternative distribution in the sense of 
solving the aforementioned optimization problem. Section 4 describes a recursive algorithm for 
finding the optimal alternative distribution. In addition, we show that the GHS method produces 
an approximation, while the recursive algorithm calculates the exact solution of the optimization 
problem. Section 5 demonstrates that the accuracy of the proposed method is consistently better 
than that of the GHS method through a simulation study. The simulation study also shows that
the proposed method is efficient in terms of computing time. Concluding remarks are given in Section 6. The proofs are deferred to the appendix.

2. Value-at-Risk

As a standard benchmark for market risk disclosure, VaR is the loss in market value over a specified time horizon that will not be exceeded with probability $1 - p$. Hence define VaR as the quantile $l_p$ of the loss $L$ in portfolio value during a holding period of a given time horizon $\Delta t$. To be more specific, we express the portfolio value $V(t, S(t))$ as a function of risk factors and time, where $S(t) = (S_1(t), \ldots, S_m(t))^\top$ comprises the $m$ risk factors to which the portfolio is exposed at time $t$ and $\top$ denotes the transpose of a matrix. The loss of the portfolio over the time interval $[t, t+\Delta t]$ is

$$L = V(t, S(t)) - V(t+\Delta t, S(t+\Delta t)).$$

Therefore VaR, $l_p$, associated with a given probability $p$ and time horizon $\Delta t$, is given by

$$P(L > l_p) = p. \quad (1)$$

Next we shall describe a quadratic approximation to the loss $L$, which follows closely.

2.1. The quadratic approximation

Let $\Delta S = [S(t+\Delta t) - S(t)]$ be the change in $S$ over the corresponding time interval. The delta-gamma methods developed in ?, ?, ?, and ? refine the relationship between risk factors and portfolio value by including quadratic as well as linear terms. The delta-gamma approximation to the change in portfolio value is

$$V(t+\Delta t, S+\Delta S) - V(t, S) \approx \frac{\partial V}{\partial t} \Delta t + \delta^\top \Delta S + \frac{1}{2} \Delta S^\top \Gamma \Delta S,$$

where

$$\delta_i = \frac{\partial V}{\partial S_i}, \quad \Gamma_{ij} = \frac{\partial^2 V}{\partial S_i \partial S_j}, \quad i, j = 1, \ldots, m,$$

and all derivatives are evaluated at the initial point $(t, S)$. Hence we can approximate the loss

$$L \approx a_0 + a^\top \Delta S + \Delta S^\top M \Delta S \equiv a_0 + T,$$
where $a_0 = -\frac{\partial V}{\partial t} \Delta t$ is a scalar, $a = -\delta$ is an $m$-vector and $M = -\Gamma/2$ is a symmetric matrix.

Assume that $\Delta S$ has a multivariate $t$ distribution with degrees of freedom $\nu$ and correlation matrix $\Sigma$. Let $C$ be the square root of the positive definite matrix $\Sigma$ such that $C^T C = \Sigma$. We can transform the multivariate $t$ distributed random variable $\Delta S$ into $X$, which is multivariate $t$ distributed with identity covariance matrix, so that

$$T \overset{d}{=} a^T CX + X^T (C^T MC) X.$$ 

Moreover, $C$ can be chosen so that $C^T MC$ is diagonalized with diagonal elements $\lambda_1, \ldots, \lambda_m$. Let $b^T = a^T C$. We have

$$T \overset{d}{=} b^T X + X^T \Lambda X = \sum_{i=1}^m (b_i X_i + \lambda_i X_i^2),$$

where $X_i = \frac{Z_i}{\sqrt{Y/\nu}}$ is a $t$ distributed random variable, with $Z_i \sim N(0, 1)$ and independent of $Y \sim \chi^2_\nu$.

The issue now is how to estimate a small tail probability of $T$ with the Monte Carlo method so that the variance is as small as possible.

**Remark.** In reality, the portfolio loss would not be exactly quadratic. The quadratic approximation is used to select the alternative distribution in importance sampling. Once the alternative distribution is selected, the actual portfolio loss would be estimated via simulation. The extent of variance reduction will then depend on both the quality of quadratic approximation and the quality of the alternative distribution.

### 2.2. An optimization problem

To estimate the probability of an event $\{T > x\}$, we shall employ the importance sampling method. That is, instead of sampling from the target distribution $P$ of $T$ directly, we sample from an alternative distribution $Q$. The problem is how to choose the alternative distribution $Q$ so that the importance sampling estimator has minimum variance.

The importance sampling estimator for $p = P\{T > x\}$ based on a sample of size $n$ is

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i > x\}} \frac{dP}{dQ}.$$  (3)
where $1_A$ is the indicator function of an event $A$, $T_i$, $i = 1, \ldots, n$ are independent observations from $Q$, and $dP/dQ$ is the Radon-Nikodym derivative assuming $P$ is absolutely continuous with respect to $Q$. Set

$$
v = E_Q[1\{T>x\} \frac{dP}{dQ}]^2 = E[1\{T>x\} \frac{dP}{dQ}],
$$

with the second expectation computed under $P$. Hereafter, the expectation without any qualification is under the target probability measure $P$ unless otherwise stated. Observed that, since the estimator $\hat{p}_n$ is unbiased, the variance of the importance sampling estimator is

$$\text{var}_Q(\hat{p}_n) = n^{-1}(v - p^2).$$

Thus our goal is to solve the optimization problem

$$\min_Q E[1\{T>x\} \frac{dP}{dQ}]. \quad (4)$$

Because the $t$ distribution has a polynomial tail, its moment generating function does not exist. How to embed the optimal alternative distribution $Q$ in a family of distributions so that (4) has a well-behaved solution is a nontrivial task. One key idea of importance sampling for $t$-distributions is to transform the $t$ distributed random variable and we shall consider the following transformation due to (Glasserman et al., 2002)

$$P(T > x) = P\left(\sum_{j=1}^{m} (b_jX_j + \lambda_jX_j^2) - x > 0\right)
= P\left(\frac{Y}{\nu} \left| \sum_{j=1}^{m} (b_jX_j + \lambda_jX_j^2) - x \right| > 0\right)
= P\left(\sum_{j=1}^{m} \sqrt{\frac{Y}{\nu} b_jZ_j + \lambda_jZ_j^2} - x \frac{Y}{\nu} > 0\right).$$

The preceding transformation of the $t$ distributed random variable is event specific. That is, it depends on the event of interest. This transformation converts the underlying sigma-field to a sub sigma-field of a multi-variate normal random variable. The GHS method makes use of this transformation to change the heavy-tailed $t$-distribution into a light-tailed one with an existent
moment generating function. This is necessary for their method, based on large deviations theory, to work.

Our method employs the same transformation. However, there is a critical difference in how the transformed random variable is used in latter developments. Glasserman et al. (2000, 2002), unable to solve the subsequent optimization problem, satisfy themselves with the solution to an approximation version of this problem. Our method, on the other hand, produces a solution to the actual optimization problem, with no approximations needed. The details are given in the next section.

3. The optimal alternative distribution

3.1. A family of alternative distributions

We now develop our method in detail. Let \( Q_\theta \) and \( P \) be the probability measures with distributions \( Q_\theta \) and \( P \), respectively. Following Glasserman et al. (2002), let the likelihood ratio for the family \( \{Q_\theta\} \) of alternative probability measures with respect to the target measure \( P \) be of the form

\[
\frac{dQ_\theta}{dP} = e^{\theta T_x - \psi_x(\theta)},
\]

where

\[
T_x := (T - x)Y/\nu = \sum_{j=1}^{m} \left( \sqrt{\frac{Y}{\nu}} b_j Z_j + \lambda_j Z_{j}^2 \right) - \frac{x}{\nu} Y
\]

and \( \psi_x(\theta) = \log E[\exp(\theta T_x)] \) is the cumulant generating function of \( T_x \) under the target probability measure \( P \). This is equivalent to embedding the alternative measure \( Q \) in a family of probability measures indexed by \( \theta \). The domain of \( \theta \) will be specified after Equation (11). The problem of finding the optimal alternative measure is then reduced to that of identifying the \( \theta \)-value that yields the minimal variance for the importance sampling estimator.

We now proceed to find the probability density under the alternative measure \( Q_\theta \). Even though most of the calculations can be found in Glasserman et al. (2002), we include them here for completeness. The probability density function under \( P \) with respect to Lebesgue measure \( L \) equals

\[
\frac{dP}{dL} = \frac{1}{(2\pi)^{\nu/2}} \exp\left\{-\frac{1}{2} \sum_{j=1}^{m} \frac{z_j^2}{\Gamma(\frac{\nu}{2})2^{\nu/2}} y^{\nu-1} e^{-\frac{y}{2}} \right\}.
\]
That is, $Z_j$, $j = 1, \ldots, m$ are independent standard normally distributed and $Y$ has a chi-square distribution with $\nu$ degrees of freedom. Hence

$$E[e^{\theta T_x} | Y] = \exp\left(-\frac{x\theta}{\nu} + \frac{1}{2\nu} \sum_{j=1}^{m} \frac{\theta^2 b_j^2}{1 - 2\theta \lambda_j}\right) Y \prod_{j=1}^{m} \frac{1}{\sqrt{1 - 2\theta \lambda_j}} = e^{\alpha(\theta) Y} \prod_{j=1}^{m} \frac{1}{\sqrt{1 - 2\theta \lambda_j}},$$

where

$$\alpha(\theta) = -\frac{x\theta}{\nu} + \frac{1}{2\nu} \sum_{j=1}^{m} \frac{\theta^2 b_j^2}{1 - 2\theta \lambda_j}. \quad (8)$$

Integrating $E[e^{\theta T_x} | Y]$ with respect to $Y$, it follows that

$$E[e^{\theta T_x}] = e^{\psi_x(\theta)} = [1 - 2\alpha(\theta)]^{-\frac{\nu}{2}} \prod_{j=1}^{m} \frac{1}{\sqrt{1 - 2\theta \lambda_j}}.$$ 

Thus

$$\psi_x(\theta) = \log E[e^{\theta T_x}] = -\frac{\nu}{2} \log[1 - 2\alpha(\theta)] - \frac{1}{2} \sum_{j=1}^{m} \log(1 - 2\theta \lambda_j). \quad (9)$$

Let $dQ_\theta / d\mathcal{L}$ be the probability density function under $Q_\theta$. From (5), (7) and (9), it follows that the importance sampling is done with an exponential twisting of measure

$$\frac{dQ_\theta}{d\mathcal{L}} \frac{d\mathcal{P}}{dQ_\theta} \frac{d\mathcal{L}}{d\mathcal{P}} = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi \sigma_j}} e^{-\frac{(y_j - \mu_j)^2}{2\sigma_j^2}} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)[2(1 - 2\alpha(\theta))]^{-\frac{\nu}{2}}} e^{-\frac{y}{2[1 - 2\alpha(\theta)]}}, \quad (10)$$

where

$$\mu_j(\theta) = \frac{b_j \theta \sqrt{y_j / \nu}}{1 - 2\theta \lambda_j}, \quad \sigma_j^2(\theta) = \frac{1}{1 - 2\theta \lambda_j}. \quad (11)$$

That is, conditional on $Y = y$, $Z_j$, $j = 1, \ldots, m$, are independent $N(\mu_j(\theta), \sigma_j^2(\theta))$, and $Y$ has gamma distribution with shape parameter $\nu/2$ and scale parameter $2[1 - 2\alpha(\theta)]^{-1}$. To guarantee the existence of $Q_\theta$ under exponential twisting of measures, the constants $\theta$ and $\alpha$ must satisfy $1 - 2\theta \lambda(1) > 0$, $1 - 2\theta \lambda(m) > 0$, and $1 - 2\alpha(\theta) > 0$, where $\lambda(1) = \max \lambda_i$ and $\lambda(m) = \min \lambda_i$.

Define the objective function of the optimization problem (4) as

$$G(\theta) := E[1_{\{T_x > 0\}}] = (1 - 2\alpha(\theta))^{-\frac{\nu}{2}} \prod_{j=1}^{m} \frac{1}{\sqrt{1 - 2\theta \lambda_j}} E[1_{\{T_x > 0\}} e^{-\theta T_x}]. \quad (12)$$
In (12), if \( x \) is so large that the probability \( P\{T > x\} \) is of order \( 10^{-6} \) or smaller, then the event is of large deviations. For events of large deviations, previous authors have shown that the asymptotically optimal alternative distribution is obtained through exponential twisting. That is, \( Q_\theta(dx) = K \exp(\theta x) \mathcal{P}(dx) \), where \( K \) is a normalizing constant and \( \theta \) determines the amount of twisting. The optimal amount of twisting \( \theta \) is such that the mean of the alternative distribution under \( Q_\theta \) measure equals the dominating point located at the boundary of the event concerned; see ?.

Instead of directly minimizing the function \( G \), the GHS method minimizes an upper bound of \( G \). Specifically,

\[
G(\theta) = E[1_{\{T_x > 0\}} \frac{d\mathcal{P}}{dQ_\theta}] \leq E[\exp(\theta T_x)] = e^{\psi_x(\theta)},
\]

and the GHS method minimizes the upper bound \( e^{\psi_x(\theta)} \) by finding the root of

\[
\psi'_x(\theta) := \frac{\partial \psi_x(\theta)}{\partial \theta} = [-x + \sum_{j=1}^{m} \frac{\theta b_j^2 (1 - \theta \lambda_j)}{(1 - 2 \theta \lambda_j)^2} \frac{1}{1 - 2\alpha(\theta)} + \sum_{j=1}^{m} \frac{\lambda_j}{1 - 2\theta \lambda_j}].
\]  

(13)

We claim that GHS method corresponds to the approximation based on the large-deviations theory. To see this, note that the root of (13), \( \theta \), satisfies \( E_\theta(T_x) = 0 \). This is the same as choosing \( \theta \) so that the mean of the alternative distribution under \( Q_\theta \) measure equals the dominating point of the event \( \{T_x > 0\} \), which is 0. This proves our claim.

In sharp contrast to the GHS method, the proposed method minimizes \( G(\cdot) \) instead of \( \psi_x(\cdot) \). In other words, the proposed method yields the exact solution while Glasserman et al. (2000, 2002) gives an approximation to the exact solution. The precise statement is given in Theorem 1. Before stating the theorem, we need to define some quantities that facilitate the presentation of it.

In view of (8) and (11), define

\[
\bar{\alpha}(\theta) = \alpha(-\theta), \quad \bar{\mu}_j(\theta) = \mu_j(-\theta), \quad \bar{\sigma}_j^2(\theta) = \sigma_j^2(-\theta).
\]  

(14)

Let \( \bar{Z}_j = \bar{\sigma}_j(\theta) Z_j + \bar{\mu}_j(\theta) \) and

\[
X_\theta = \sum_{j=1}^{m} \sqrt{\frac{Y}{\nu(1 - 2\bar{\alpha}(\theta))}} b_j \bar{Z}_j + \lambda_j \bar{Z}_j^2 - \frac{xY}{\nu(1 - 2\bar{\alpha}(\theta))}.
\]  

(15)
Theorem 1. Let \( \theta \) be such that \( 1 - 2 \theta \lambda (1) > 0 \), \( 1 - 2 \theta \lambda (m) > 0 \), and \( 1 - 2 \alpha (\theta) > 0 \). Under the quadratic approximation to a portfolio VaR, the optimal alternative distribution \( Q^\theta \) minimizing the variance of the importance sampling estimator has \( \theta \) satisfying

\[
E[X|X > 0] = \psi_x'(\theta).
\]  

Moreover, if the linear approximation is adopted, then

\[
E[X|X > 0] = \frac{\nu(\sum_{j=1}^m \theta b_j^2 - x)}{\nu - \sum_{j=1}^m b_j^2 \theta^2 + 2 \theta x}.
\]  

Remark. The linear approximation case corresponds to \( \lambda_i = 0 \) for \( i = 1, \cdots, m \) in (2).

The proof of Theorem 1 is given in Appendix A.

By (13) and (16), the difference between the GHS method and ours is that the former replaces the left hand side of (16) with zero. If the event \( \{ T_x > 0 \} \) is of very small probability, say, in the order of \( 10^{-6} \), then \( E[X|X > 0] \) is very close to zero and approximating it by zero is justified. However, events in VaR are of moderately small probability and \( E[X|X > 0] \) is not close to zero. Hence approximating by zero can introduce non-negligible errors.

Lemma 1. The optimization problem (4) has a unique solution, which satisfies (16).

Proof. The second derivative of \( G \) equals

\[
\frac{\partial^2 G(\theta)}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2} E[1_{\{ T_x > 0 \}} \frac{dE}{dQ}] = \frac{\partial^2}{\partial \theta^2} E[1_{\{ T_x > 0 \}} e^{-\theta T_x + \psi_x(\theta)}]

= \frac{\partial}{\partial \theta} E[1_{\{ T_x > 0 \}} (-T_x + \psi_x'(\theta))e^{-\theta T_x + \psi_x(\theta)}]

= E[1_{\{ T_x > 0 \}} ((-T_x + \psi_x'(\theta))^2 + \psi_x''(\theta))e^{-\theta T_x + \psi_x(\theta)}].
\]  

Since \( \psi_x(\theta) \) is the cumulant generating function of \( T_x \), its second derivative \( \psi_x''(\theta) > 0 \). It then follows from (18) that \( \frac{\partial^2 G(\theta)}{\partial \theta^2} > 0 \), which implies that \( G \) has a unique minimum.

3.2. Multivariate t function for component VaR

In risk management, it is useful to compute a risk decomposition of the current portfolio. VaRs for individual assets do not serve the purpose because they usually ignore diversification effects. This calls for the concept of Component VaR. Component VaR has been introduced by ? under
the assumption that returns are drawn from a multivariate normal distribution. Let CVaR be the
Component VaR of asset \(i\).

\[
CVaR_i = w_i \beta_i \cdot \text{VaR},
\]

where

\[
\beta_i = \frac{\text{cov}(r_i, r_p)}{\text{var}(r_p)},
\]

\(r_i\) is the \(i\)th-asset return, \(r_p\) is the portfolio return, \(w_i\) is the weight of asset \(i\) in the portfolio, and \(\text{VaR}\) is the portfolio VaR. Note that \(\sum_{i=1}^{m} w_i \beta_i = 1\) and thus \(\text{VaR} = \sum_{i=1}^{m} C\text{Var}_i\) meaning the component VaR forms a decomposition of the portfolio VaR.

Component VaR measures the sensitivity of VaR to the injection of one unit of cash flow in each
dimension of the cash flow space. When presenting VaR in a risk management report, it would be
useful to display what positions constitute big risks and what positions serve as hedges. Component
VaR reveals pockets of risk concentration as well as hedges. Recently, ? and ? provided a new
approach to component VaR. They proposed a probabilistic definition of component VaR and it
has a graphical interpretation that facilitates rigorous analysis. ? relaxes the restrictive normality
assumption and estimates it in a distribution-free setting.

Assuming that returns are from a multivariate \(t\) distribution, consider the losses of all assets
in a portfolio simultaneously. This requires the extension of the proposed method to multivariate
settings. Let \(\tilde{X} = \tilde{Z} / \sqrt{\frac{\nu}{\tilde{Y}}}\) be the multivariate \(t\)-distribution with dimension \(m\), degrees of freedom \(\nu\)
and correlation matrix \(\Sigma\), that is, \(\tilde{Z} \sim N(0, \Sigma)\), \(Y \sim \chi^2(\nu)\) and \(Y, \tilde{Z}\) are independent. Let \(C\) and \(X\) be as in (2) and \(\tilde{X} = CX\). Denote by \(\tilde{b}\) and \(\tilde{\lambda}\) two \(m \times m\) diagonal matrices with diagonal elements
\((b_1, \ldots, b_m)\) and \((\lambda_1, \ldots, \lambda_m)\) respectively. Then the event of interest is

\[
P(\tilde{b} \tilde{X} + \tilde{\lambda} \tilde{X}^2 > \tilde{a}) = P(\tilde{b} \tilde{X} + \tilde{\lambda} \tilde{X}^2 - \tilde{a} > 0) = P(\sqrt{\frac{\tilde{Y}}{\nu}} \tilde{b} \tilde{Z} + \tilde{\lambda} \tilde{Z}^2 - \frac{\tilde{Y}}{\nu} \tilde{a} > 0),
\]

where \(\tilde{a}\) is an \(m \times 1\) vector specifying the loss for each component in the portfolio, and \(\tilde{Z}^2\) and \(\tilde{X}^2\)
are \(m \times 1\) vectors with elements equal to the squares of those in \(\tilde{Z}\) and \(\tilde{X}\), respectively.
Define $\tilde{T}_a = \sqrt{\frac{2}{\nu}} b \tilde{Z} + \lambda \tilde{Z}^2 - \frac{\nu}{\nu} \tilde{a}$. The target density function under $\mathcal{P}$ is given by

$$
\frac{d\mathcal{P}}{d\mathcal{L}} = \frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}Z^T \Sigma^{-1} Z} \frac{1}{\Gamma(\frac{n}{2})2^\frac{n}{2}}Y^\frac{n}{2}e^{-\frac{Y}{2}}.
$$

The likelihood ratio is of the form

$$
\frac{d\mathcal{Q}_\theta}{d\mathcal{P}} = \frac{e^{\theta^T \tilde{T}_a}}{E[e^{\theta^T \tilde{T}_a}]},
$$

where $\theta$ is an $m \times 1$ vector. Straightforward computation of the conditional expectation gives

$$
E[e^{\theta^T \tilde{T}_a} | Y] = \exp\{[-\frac{\theta^T \tilde{a}}{\nu} + \frac{1}{2\nu} \sum_{j=1}^m \frac{\theta_j^2 b_j^2}{1 - 2\theta_j \lambda_j}]Y\} \prod_{i=1}^m \frac{1}{\sqrt{1 - 2\theta_i \lambda_i}},
$$

$$
E[e^{\theta^T \tilde{T}_a}] = [1 - 2(-\frac{\theta^T \tilde{a}}{\nu} + \frac{1}{2\nu} \sum_{j=1}^m \frac{\theta_j^2 b_j^2}{1 - 2\theta_j \lambda_j})]^{\frac{n}{2}} \prod_{i=1}^m \frac{1}{\sqrt{1 - 2\theta_i \lambda_i}},
$$

where $\theta$ is an $m \times 1$ vector. We obtain the likelihood ratio

$$
\frac{d\mathcal{Q}_\theta}{d\mathcal{P}} = \exp\{\theta^T (\sqrt{\frac{2}{\nu}} b \tilde{Z} + \lambda \tilde{Z}^2 - \frac{\nu}{\nu} \tilde{a})\}[1 - 2(-\frac{\theta^T \tilde{a}}{\nu} + \frac{1}{2\nu} \sum_{j=1}^m \frac{\theta_j^2 b_j^2}{1 - 2\theta_j \lambda_j})]^\frac{n}{2} \prod_{i=1}^m \sqrt{1 - 2\theta_i \lambda_i}.
$$

Therefore, the alternative density function under $\mathcal{Q}_\theta$ measure for importance sampling is

$$
\frac{d\mathcal{Q}_\theta}{d\mathcal{L}} = \frac{d\mathcal{Q}_\theta}{d\mathcal{P}} \frac{d\mathcal{P}}{d\mathcal{L}} = (2\pi)^{-\frac{n}{2}} \prod_{i=1}^m \frac{\theta_j^2 b_j^2}{1 - 2\theta_j \lambda_j} \exp\{[-\frac{1}{2} \sum_{i=1}^m [(1 - 2\theta_i \lambda_i)Z_i - b_i \sqrt{\nu Y/\nu}]^2] \} \Gamma(\frac{n}{2})^{\frac{n}{2} - 2} \cdot
$$

$$
[1 - 2(-\frac{\theta^T \tilde{a}}{\nu} + \frac{1}{2\nu} \sum_{j=1}^m \frac{\theta_j^2 b_j^2}{1 - 2\theta_j \lambda_j})]^\frac{n}{2} \prod_{i=1}^m \sqrt{1 - 2\theta_i \lambda_i}.
$$

We specify the optimal alternative distribution for component VaR in Lemma 2, and because the proof is similar to Theorem 1 it is omitted.

**Lemma 2.** For $\theta$ satisfying $\nu + 2\theta^T \tilde{a} - \sum_{j=1}^m \theta_j^2 b_j^2 (1 - 2\theta_j \lambda_j)^{-1} > 0$, under quadratic approximation for component VaR, the optimal alternative distribution has $\theta$ satisfying

$$
E[X_\theta | X_\theta > 0] = [-\tilde{a} + \begin{pmatrix} \theta_1 b_1^2 (1 - \theta_1 \lambda_1) \\ \vdots \\ \theta_m b_m^2 (1 - \theta_m \lambda_m) \\ \end{pmatrix} \frac{1}{1 - 2\tilde{a}^T (\theta)} + \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \\ \end{pmatrix}] = \begin{pmatrix} \tilde{a} \\ \vdots \\ \tilde{a} \end{pmatrix},
$$

where $\tilde{a}(\theta) = \frac{\theta^T \tilde{a}}{\nu} + \frac{1}{2\nu} \sum_{j=1}^m \frac{\theta_j^2 b_j^2}{1 - 2\theta_j \lambda_j}$.

Although Glasserman et al. (2000, 2002) did not provide importance sampling results on component VaR, one can obtain such results from Lemma 2. The procedure is similar to that
described by comments immediately after Theorem 1.

**Remark**: The corresponding results using linear approximation can be obtained by letting all \( \lambda_i = 0 \) on the right hand side of (19).

4. Calculating the optimal alternative distribution

Before employing the importance sampling method, it is first necessary to identify the optimal alternative distribution. Because the optimal \( \theta \) in (16) cannot be computed directly, we use the following recursive algorithm to find the optimal \( \theta \).

The Recursive Algorithm

- **Initialize**: Set \( i = 0 \) and \( \hat{\theta} = \theta^{(0)} \).
- **Integrate**: Compute \( E[X_{\hat{\theta}}(i) | X_{\hat{\theta}}(i) > 0] \) and let it be \( t_i \).
- **Solve**: Find \( \theta \) such that \( \psi_x'(\theta) = t_i \), and let the solution be \( \theta^{(i+1)} \).
- **Terminate**: If \( \frac{\theta^{(i+1)} - \theta^{(i)}}{\theta^{(i)}} \) is small, then output \( \hat{\theta} = \theta^{(i+1)} \) and stop. Otherwise, go to the next step.
- **Iterate**: Return to Integrate and replace \( i \) with \( i + 1 \).

In the preceding algorithm, the initial value \( \theta^{(0)} \) can be chosen as the dominating point of the event \( \{ X_\theta > 0 \} \). Figure ?? shows the solution path of \( \theta^{(i)} \) and \( \theta^* \) is the optimal amount of exponential twisting satisfying (16). When we use the dominating point as the initial value \( \theta^{(0)} \), it can be seen that, after the first iteration, \( \theta^{(1)} \) is in the interior of the set \( \{ T_x > 0 \} \). Furthermore, if \( \theta^{(0)} \) is sufficiently large, the density of \( X_\theta \) decreases rapidly and the solution of (16) is close to \( \theta^{(0)} \). Therefore, fast convergence of the recursive algorithm is to be expected.

Because the exact objective function \( G(\cdot) \) is difficult to optimize directly, both the GHS method and the proposed method replace \( G \) with a quasi objective function and do optimization on the latter. The plots of the quasi objective functions for the GHS method, the proposed method with the number of iterations \( r = 1, 3, 6 \) and true \( G(\cdot) \) are given in Figure 2. We can see that the de facto objective functions of the proposed method approach \( G(\cdot) \) very quickly as the number of iterations of the recursive algorithm increases.
In other words, our method computes the solution to the exact objective function via a sequence of approximations. Each element in the sequence of approximations can be viewed as the solution to a quasi objective function. We show here that the sequence of quasi objective functions converges to the exact objective function very fast. The purpose of this undertaking is to facilitate comparison with the GHS method, which uses only one quasi objective function instead of a sequence of quasi objective functions that converges to the exact one.

Denote the left and right hand sides of (16) by
\[
\tilde{\gamma}(\theta) := E[X_\theta | X_\theta > 0], \quad \gamma(\theta) := \psi_x'(\theta),
\]
respectively. Let \( \theta^* \) be the solution to (16). That is, \( \tilde{\gamma}(\theta^*) = \gamma(\theta^*) \). We have the following theorem. The proof is given in the appendix.

**Theorem 2.** Choose the dominating point of the event \( \{ X_\theta > 0 \} \) as the initial value \( \theta^{(0)} \). Then the recursive algorithm either converges to \( \theta^* \) or alternates in the limit between a pair of values \( \underline{\theta} \neq \bar{\theta} \) satisfying
\[
\gamma(\underline{\theta}) = \tilde{\gamma}(\bar{\theta}) \quad \text{and} \quad \tilde{\gamma}(\underline{\theta}) = \gamma(\bar{\theta}).
\]
Corollary 1. If there does not exist $\theta \neq \bar{\theta}$ such that (??) holds, then the recursive algorithm converges to the solution of (16).

Importance sampling estimate of loss probability

I. For each of the $n$ replications do the following:

1. Generate $Y$ from a gamma distribution with shape parameter $\nu/2$ and scale parameter $2[1 - 2\alpha(\theta)]^{-1}$.

2. Given $Y$, generate independent normals $Z_1, \ldots, Z_m$ with mean and variance given by (11).
3. Set $X = Z/\sqrt{Y/\nu}$, where $Z = (Z_1, \ldots, Z_m)$.

4. Compute $T$ according to (2).

5. Set $T_x = (T - x)Y/\nu$.

6. Multiply the indicator function by the likelihood ratio to get

$$1_{\{T_x > 0\}}e^{-\theta T_x + \psi_x(\theta)},$$ (21)

II. Average (??) over the $n$ independent replications to obtain the estimate.

5. Simulation study

In this section, we report results from four simulation experiments. The first two experiments concern VaR with one and two risk factors, respectively. The third and fourth experiments concern component VaR. Because the main difference between the proposed method and the GHS method is on the choice of alternative distributions in importance sampling, to facilitate comparison, we bypass the valuation step for complicated instruments and focus on the efficiency gain from importance sampling using the proposed method compared to the large deviations method.

To investigate the performance of the importance sampling algorithm described in §4, we conduct four Monte Carlo experiments on the loss probability $P(L > x)$ and report the results in Tables 1-4, respectively. The sample size is $n = 500$ and the number of Monte Carlo replications is $M = 10,000$. Because the recursive algorithm converges very quickly, we set the recursion size to be $r = 6$ in all experiments. We use the recursive algorithm to find the exact solution to the corresponding optimization problem, and thus the computing time is a little in excess of the GHS method. Since we only need to run the recursive algorithm once and use the same optimal $\theta$ in all Monte Carlo replications, the additional computing time is negligible with respect to the total computing time. To be more specific, ? follows previous authors using the product $\sigma^2 \tau$ ($\sigma^2 \tau$ is defined as variance per replication and $\tau$ is defined as expected computing time per replication) as a measure of efficiency. In the case of our algorithm, the extra amount of computing time due to finding the exact solution is only a negligible fraction of the total computing time and thus the computation
time per replication is almost the same. Consequently, it is sufficient to compare the variance per replication to determine which method is more efficient.

### Table 1: Relative Efficiency

<table>
<thead>
<tr>
<th>$x$</th>
<th>1.96</th>
<th>3.93</th>
<th>4.77</th>
<th>6.74</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu = 5$, $b = -1$, $\lambda = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true $P(A)$</td>
<td>5.00E-02</td>
<td>1.00E-02</td>
<td>5.04E-03</td>
<td>1.00E-03</td>
</tr>
<tr>
<td>variance</td>
<td>9.57E-05</td>
<td>1.94E-05</td>
<td>9.98E-06</td>
<td>1.99E-06</td>
</tr>
<tr>
<td>GHS variance</td>
<td>4.99E-02</td>
<td>9.99E-03</td>
<td>5.04E-03</td>
<td>1.00E-03</td>
</tr>
<tr>
<td>variance</td>
<td>2.52E-05</td>
<td>1.49E-06</td>
<td>4.27E-07</td>
<td>2.14E-08</td>
</tr>
<tr>
<td>PSD variance</td>
<td>4.99E-02</td>
<td>9.99E-03</td>
<td>5.04E-03</td>
<td>1.00E-03</td>
</tr>
<tr>
<td>variance</td>
<td>2.00E-05</td>
<td>1.22E-06</td>
<td>3.62E-07</td>
<td>1.87E-08</td>
</tr>
<tr>
<td>eff(PSD,NV)</td>
<td>4.78</td>
<td>15.82</td>
<td>27.56</td>
<td>106.32</td>
</tr>
<tr>
<td>eff(PSD,GHS)</td>
<td>1.26</td>
<td>1.22</td>
<td>1.18</td>
<td>1.14</td>
</tr>
</tbody>
</table>

In Table 1, we compare the relative efficiency of the proposed method (PSD) with respect to the naive method (NV) and the GHS method in estimating the loss probabilities $P\{L > x\}$ with different values of $x$. Here the relative efficiency, $\text{eff}(\text{Method1, Method2})$, is the variance under Method 2 divided by that under Method 1. We can see that the proposed method is significantly more efficient than the naive method and moderately more efficient than the GHS method for all values of $x$. Furthermore, the efficiency gain is larger for smaller probabilities against the naive method and the gain is smaller for smaller probabilities against the GHS method. The results in Table 1 confirm two important facts. First, the importance sampling methods are capable of dramatic improvement over the naive simulation method. Second, among importance sampling methods, large-deviations methods, such as the GHS method, are not as efficient as the the proposed method. Even though the efficiency gain is moderate for the proposed method, it is of practical importance because VaR needs to be computed frequently and each time the amount of computation can be enormous for portfolios with a large number of complicated instruments.

In Table 2, we report the relative efficiency with fifteen risk factors. In addition to the same conclusions as Table 1, we observe that the relative efficiency of importance sampling methods in Table 2 increase compared to those in Table 1 with one risk factor. The proposed method is again consistently better than the GHS method.
Table 2: Relative Efficiency with Fifteen Risk Factors.

The parameters are $\nu = 3$, $b_i = 0.1 + i/100$, and $\lambda_i = 0.05i$, $i = 1, \ldots, 15$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>52.58</th>
<th>161.61</th>
<th>259.3</th>
<th>762.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>true $P(A)$</td>
<td>5.00E-02</td>
<td>1.00E-02</td>
<td>5.00E-03</td>
<td>1.00E-03</td>
</tr>
<tr>
<td>variance</td>
<td>9.23E-05</td>
<td>2.03E-05</td>
<td>9.76E-06</td>
<td>2.01E-06</td>
</tr>
<tr>
<td>GHS</td>
<td>5.01E-02</td>
<td>1.02E-02</td>
<td>5.01E-03</td>
<td>1.00E-03</td>
</tr>
<tr>
<td>variance</td>
<td>5.49E-06</td>
<td>2.44E-07</td>
<td>6.16E-08</td>
<td>2.83E-09</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.175</td>
<td>0.190</td>
<td>0.193</td>
<td>0.196</td>
</tr>
<tr>
<td>PSD</td>
<td>5.00E-02</td>
<td>1.00E-02</td>
<td>5.00E-03</td>
<td>1.00E-03</td>
</tr>
<tr>
<td>variance</td>
<td>4.37E-06</td>
<td>2.02E-07</td>
<td>5.22E-08</td>
<td>2.43E-09</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.224</td>
<td>0.228</td>
<td>0.230</td>
<td>0.231</td>
</tr>
<tr>
<td>eff(PSD, NV)</td>
<td>21.12</td>
<td>100.49</td>
<td>186.97</td>
<td>827.16</td>
</tr>
<tr>
<td>eff(PSD, GHS)</td>
<td>1.26</td>
<td>1.21</td>
<td>1.18</td>
<td>1.16</td>
</tr>
</tbody>
</table>

In Table 3, we compare the relative efficiency $\text{eff}(\text{PSD, NV})$ for component VaR. Since the VaR for each component can be made the same by changing the scales of the corresponding risk factors, without loss of generality, we assume that the component VaRs for each risk factor involved are the same. To demonstrate the effect of the proposed method without complication from many risk factors, we consider only two risk factors here. Let the event of interest be $A = \{\tilde{T}_a > \tilde{d}\}$, where $\tilde{d}$ is the vector in $R^2$ with both components equal $d$. We can adjust $d$ so that the probability of event $A$ equals $p = 0.01, 0.005, 0.001, 0.0005$. The results clearly show that the importance sampling methods are more efficient than naive simulation. The relative efficiency increases as the event probabilities decrease.
Table 3: Component VaR for Two Risk Factors

<table>
<thead>
<tr>
<th></th>
<th>Component VaR</th>
<th>Variance</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV</td>
<td>9.97E-03</td>
<td>9.9E-05</td>
<td>1.99E-06</td>
</tr>
<tr>
<td>PSD</td>
<td>1.00E-02</td>
<td>1.11E-06</td>
<td>6.6E-09</td>
</tr>
</tbody>
</table>

The component VaRs add up to the total VaR, that is, \( CVaR_1 + \cdots + CVaR_n = VaR \), where \( CVaR_i \) is component VaR for \( i \)th asset. Thus, there are two ways to obtain the total VaR: compute directly the portfolio VaR (DT) or compute component VaRs first and add up to total VaR (CP).

In Table 4, under the proposed importance sampling, we report the relative efficiency of computing total VaR directly versus computing component VaR first and then summing up to total VaR.

The event \( B = \{ T > x \} \), corresponding to large losses in a portfolio, cannot be directly decomposed into component VaR. Here we use an event \( B^* = \{ \tilde{T}_a > \tilde{k} \} \) to approximate the event \( B \), where \( \tilde{k} \) is a vector in \( \mathbb{R}^2 \) such that both components equal \( k \) and \( \{ \tilde{T}_a > \tilde{k} \} \) means the vector \( \tilde{T} \) is larger than the vector \( \tilde{k} \) componentwise. In Figure 3, we demonstrate how to find \( B^* \) when there...
are two risk factors and extension to higher dimension is straightforward. The approximation is possible because the probability mass of $\tilde{T}_a$ concentrates around the 45-degree line.

Let $f_\theta(\cdot)$ be the probability density of the random variable $X_\theta$. Then

$$E[X_\theta|X_\theta \in B] = \frac{\int_B x f_\theta(x)dx}{\int_B f_\theta(x)dx} \approx \frac{\int_{B^*} x f_\theta(x)dx}{\int_{B^*} f_\theta(x)dx} = E[X_\theta|X_\theta \in B^*],$$

where $\int_B f_\theta(x)dx = \int_{B^*} f_\theta(x)dx = \alpha$. In other words, we construct the event $B^* = \{\tilde{T}_a > \tilde{k}\}$ by identifying a point $\tilde{k}$ along the 45-degree line so that $P(\tilde{T}_a > \tilde{k}) = \alpha$, Note that $\tilde{k}$ is necessarily in the interior of $B^c$.

The relative efficiency $\text{eff}(\text{CP}, \text{DT})$ is the variance of computing total VaR directly divided by the variance of computing component VaR first. The efficiency gain is more significant in higher dimensions. The situation is the same as the proposed importance sampling under multivariate normal distribution in ?. Therefore, for higher efficiency in simulating VaR, we should add up component VaRs. However, the twisting parameter is multi-dimensional. There is a tradeoff between the efficiency and programming complexity.

Table 4: Component VaR vs. linear combination of VaR.

<table>
<thead>
<tr>
<th>$\nu = 5$, $\tilde{b} = (1,1)'$, $\tilde{a} = (2,2)'$</th>
<th>true &amp; P(A)</th>
<th>1.00E-02</th>
<th>5.00E-03</th>
<th>1.00E-03</th>
<th>5.00E-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>9.99E-03</td>
<td>5.02E-03</td>
<td>9.95E-04</td>
<td>5.03E-04</td>
<td></td>
</tr>
<tr>
<td>variance</td>
<td>1.22E-06</td>
<td>3.90E-07</td>
<td>2.36E-08</td>
<td>6.74E-09</td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td>1.00E-02</td>
<td>5.00E-03</td>
<td>1.00E-03</td>
<td>5.00E-04</td>
<td></td>
</tr>
<tr>
<td>variance</td>
<td>1.07E-06</td>
<td>3.58E-07</td>
<td>2.26E-08</td>
<td>6.64E-09</td>
<td></td>
</tr>
<tr>
<td>eff(CP, DT)</td>
<td>1.14</td>
<td>1.09</td>
<td>1.04</td>
<td>1.01</td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper, we propose an importance sampling algorithm to compute VaR and component VaR of a portfolio. The risk factors are assumed to have heavy-tailed loss distributions, modeled jointly by multivariate $t$-distributions. It is shown that previous importance sampling methods yield approximations to the corresponding optimization problem, while our method produces the exact solution to the problem. Simulation results confirm the theoretical result that our method
always provides greater variance reduction than existing methods based on large deviations theory. Our numerical experiments demonstrate that the gain in variance reduction can be substantial in some cases.

The key features of our method are twofold. First, (16) in Theorem 1 characterizes the optimal alternative distribution for importance sampling. Second, the recursive algorithm in §4.1 facilitates the computation of the optimal solution. The initial value of the recursive algorithm is the alternative distribution used previously by other authors; e.g., Glasserman et al. (2000, 2002). The recursive algorithm then sequentially generates alternative distributions providing greater variance reduction. Due to the nature of the recursive algorithm, the additional programming effort and computing time are negligible.

Our method highlighting the two aforementioned key features is also attainable in other settings. Please see Fuh and Hu (2004) for multivariate normal distributions and Fuh and Hu (2007) for hidden Markov models. Further applications to $K$-distributions, jump diffusion models and Markov switching autoregression models will be published elsewhere. In this paper, we assume that the losses are independent over time. A more challenging project is to model the time dependence using, for example, Markov switching models. The other possible extension is to consider models with multiple degrees of freedom using copula; see e.g., ? and ?.

Appendix A: Proof of Theorem 1

By (6), $T_x = \sum_{j=1}^{m} (\sqrt{Y/\nu}b_j Z_j + \lambda_j Z_j^2) - xY/\nu$, the distribution of $Y \sim \text{Gamma}(\alpha = \nu/2, \beta = 2[1 - 2\alpha(\theta)]^{-1})$, and conditional on $Y = y$, $Z_j \sim N(\mu_j(\theta), \sigma_j^2(\theta))$, where $\mu_j(\theta) = (b_j \sqrt{2}/\sqrt{\nu})/(1 - 2\theta \lambda_j)$ and $\sigma_j^2(\theta) = (1 - 2\theta \lambda_j)^{-1}$. In view of (12) and (5), we have

$$G(\theta) \equiv E[1_{\{T_x > 0\}} \frac{dP}{dQ_\theta}] = E[1_{\{T_x > 0\}} e^{-\theta T_x + \psi_x(\theta)}].$$

To minimize $G(\theta)$, we need to solve the equation $\frac{d}{d\theta} \log G(\theta) = 0$, which is equivalent to

$$-\frac{d}{d\theta} \log E[1_{\{T_x > 0\}} e^{-\theta T_x}] = \frac{d}{d\theta} \psi_x(\theta).$$

(22)
By (13), the right side of (??) equals the right side of (16). We now proceed to show that the left side of (??) equals the left side of (16).

It is easy to see that

\[
-\frac{d}{d\theta} \log E[1_{\{T_z > 0\}} e^{-\theta T_z}] = \frac{E[1_{\{T_z > 0\}} T_x e^{-\theta T_x}]}{E[1_{\{T_z > 0\}} e^{-\theta T_x}]},
\]

and

\[
E[1_{\{T_z > 0\}} e^{-\theta T_z}] = \int_{-\infty}^{\infty} \cdot \int_{-\infty}^{\infty} \int_{0}^{\infty} 1_{\{T_z > 0\}} \exp\left\{ -\sum_{j=1}^{m} \left( \sqrt{\frac{\nu}{\nu}} b_j \theta z_j + \lambda_j \bar{\theta} z_j^2 + \frac{x \theta}{\nu} y \right) \right\} 
\]

\[
\times (2\pi)^{-m/2} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{m} z_j^2 \right\} \left[ \Gamma(\nu/2)2^{\nu/2} \right]^{-1} y^{\nu-1} e^{-\frac{y}{2}} dy dz_1 \cdots dz_m.
\]

In view of (14), the preceding equation can be written as

\[
\int_{-\infty}^{\infty} \cdot \int_{-\infty}^{\infty} \int_{0}^{\infty} 1_{\{T_z > 0\}} \frac{1}{(2\pi)^{m/2}} \exp\left\{ -\sum_{j=1}^{m} \frac{(z_j - \bar{\mu}_j(\theta))^2}{2\sigma_j^2} \right\} 
\]

\[
\times \left[ \Gamma(\nu/2)2^{\nu/2} \right]^{-1} y^{\nu-1} \exp\left\{ -y(1 - 2\bar{\alpha}(\theta))/2 \right\} dy dz_1 \cdots dz_m
\]

\[
= [1 - 2\bar{\alpha}(\theta)]^{-\frac{m}{2}} \prod_{j=1}^{m} \frac{1}{\sqrt{1 + 2\theta \lambda_j}} \int_{-\infty}^{\infty} \cdot \int_{-\infty}^{\infty} \int_{0}^{\infty} 1_{\{T_z > 0\}} \frac{1}{\sqrt{2\pi} \sigma_j} \exp\left\{ -\sum_{j=1}^{m} \frac{(z_j - \bar{\mu}_j(\theta))^2}{2\sigma_j^2} \right\}
\]

\[
\times \left[ \Gamma(\nu/2)2/(1 - 2\bar{\alpha}(\theta)) \right]^{\nu/2} y^{\nu-1} \exp\left\{ -y(1 - 2\bar{\alpha}(\theta))/2 \right\} dy dz_1 \cdots dz_m.
\]

The preceding equation suggests that we consider a transformation of variables,

\[
y' = y(1 - 2\bar{\alpha}(\theta)), \quad z_j' = \frac{z_j - \bar{\mu}_j(\theta)}{\sigma_j},
\]

which has the Jacobian matrix

\[
|J| = \begin{vmatrix} \frac{\partial y}{\partial y} & \frac{\partial y}{\partial z_1} & \cdots & \frac{\partial y}{\partial z_m} \\ \frac{\partial z_1}{\partial y} & \frac{\partial z_1}{\partial z_1} & \cdots & \frac{\partial z_1}{\partial z_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_m}{\partial y} & \frac{\partial z_m}{\partial z_1} & \cdots & \frac{\partial z_m}{\partial z_m} \end{vmatrix} = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ -2\bar{\alpha}(\theta) & 0 & \cdots & 0 \\ 0 & \sqrt{1 + 2\theta \lambda_j} & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{1 + 2\theta \lambda_m} \end{vmatrix} = \frac{1}{1 - 2\bar{\alpha}(\theta)} \prod_{j=1}^{m} \frac{1}{\sqrt{1 + 2\theta \lambda_j}}.
\]

Under the transformation (??), the random variable $T_x$ becomes

\[
X_\theta = \sum_{j=1}^{m} \sqrt{\frac{Y'}{\nu(1 - 2\bar{\alpha}(\theta))}} b_j [\bar{\sigma}_j(\theta)Z'_j + \bar{\mu}_j(\theta)] + \lambda_j [\bar{\sigma}_j(\theta)Z'_j + \bar{\mu}_j(\theta)]^2 - \frac{x Y'}{\nu(1 - 2\bar{\alpha}(\theta))}.
\]
Note that this is consistent with the definition given by (15). Therefore, we have
\[
E[1_{\{T_x > 0\}} e^{-\theta T_x}] = [1 - 2\tilde{\alpha}(\theta)]^{-\frac{\gamma}{2}} \prod_{j=1}^{m} \frac{1}{\sqrt{1 + 2\theta \lambda_j}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 1_{\{X_\theta > 0\}} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-z_j^2/2} \\
\times [1 - 2\tilde{\alpha}(\theta)] \prod_{j=1}^{m} \frac{1}{\Gamma(\nu/2)2^{\nu/2}} y_j^{\nu/2-1} e^{-\nu y_j} \prod_{j=1}^{m} \frac{1}{\sqrt{1 + 2\theta \lambda_j}} dy_j dz_j \\
= [1 - 2\tilde{\alpha}(\theta)]^{-\frac{\gamma}{2}} \prod_{j=1}^{m} \frac{1}{\sqrt{1 + 2\theta \lambda_j}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 1_{\{X_\theta > 0\}} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-z_j^2/2} \\
\times \prod_{j=1}^{m} \frac{1}{\Gamma(\nu/2)2^{\nu/2}} y_j^{\nu/2-1} e^{-\nu y_j} dy_j dz_j = [1 - 2\tilde{\alpha}(\theta)]^{-\frac{\gamma}{2}} \prod_{j=1}^{m} \frac{1}{\sqrt{1 + 2\theta \lambda_j}} E[1_{\{X_\theta > 0\}}].
\]
Following a similar line of arguments, we can show that
\[
E[1_{\{T_x > 0\}} T_x e^{-\theta T_x}] = [1 - 2\tilde{\alpha}(\theta)]^{-\frac{\gamma}{2}} \prod_{j=1}^{m} \frac{1}{\sqrt{1 + 2\theta \lambda_j}} E[1_{\{X_\theta > 0\}} X_\theta],
\]
and thus
\[
\frac{E[1_{\{T_x > 0\}} T_x e^{-\theta T_x}]}{E[1_{\{T_x > 0\}} e^{-\theta T_x}]} = E(X_\theta|X_\theta > 0).
\]
(24)
This establishes (16). To prove (17), observe that the linear case corresponds to letting \(\lambda_i = 0\) for \(i = 1, \ldots, m\) in (2). Applying the preceding observation to (8) and (13), it is easy to see that the right side of (13) reduces to that of (17). This completes the proof of Theorem 1.

Appendix B: Proof of Theorem ??

We need the following lemma

**Lemma 3.** \(\tilde{\gamma}(\cdot)\) is strictly decreasing and \(\gamma(\cdot)\) is strictly increasing.

The lemma is the simple consequence of the following two inequalities.
\[
\frac{d\tilde{\gamma}(\theta)}{d\theta} = \frac{dE[X_\theta|X_\theta > 0]}{d\theta} = -E[1_{\{T_x > 0\}} T_x e^{-\theta T_x}] E[1_{\{T_x > 0\}} e^{-\theta T_x}]^{-1} = -\text{var}(X_\theta) < 0;
\]
\[
\frac{d\gamma(\theta)}{d\theta} = \frac{d^2\psi_x(\theta)}{d\theta^2} = E[T_x^2 e^{\theta T_x}] E[e^{\theta T_x}]^{-1} - E[T_x e^{\theta T_x}] E[e^{\theta T_x}]^{-1} = \text{var}_\theta(T_x) > 0,
\]
where \(\text{var}_\theta\) is the variance under the probability measure \(Q_\theta\) defined by (5). The proof of Lemma ?? is completed and we return to the proof of the theorem.
We claim that the initial value $\theta(0) < \theta^*$. To see why it is true, note that $\gamma(\theta(0)) = 0$ and $\tilde{\gamma}(\theta) > 0$ for all $\theta$ in the domain of $\tilde{\gamma}$. Thus $\tilde{\gamma}(\theta^*) = \gamma(\theta^*) > 0$. Since $\gamma$ is strictly increasing, we have $\theta^* > \theta(0)$ and the claim is established.

Because $\theta(0) < \theta^*$, the recursive algorithm entails

$$\gamma(\theta(1)) = \tilde{\gamma}(\theta(0)) > \tilde{\gamma}(\theta^*) = \gamma(\theta^*).$$

Since $\gamma$ is strictly increasing, we have $\theta(1) > \theta^* > \theta(0)$. Now that $\theta(1) > \theta^*$, the algorithm implies

$$\gamma(\theta(2)) = \tilde{\gamma}(\theta(1)) < \tilde{\gamma}(\theta^*) = \gamma(\theta^*); \quad \gamma(\theta(3)) = \tilde{\gamma}(\theta(2)) > \tilde{\gamma}(\theta^*) = \gamma(\theta^*).$$

It then follows that $\theta(3) > \theta^* > \theta(2)$. Applying the same argument repeatedly, we have

$$\theta(2k+1) > \theta^* > \theta(2k),$$

for $k = 0, 1, 2, \ldots$

Next we show by induction that $\{\theta(2k), \ k = 0, 1, \ldots\}$ is a strictly increasing sequence and $\{\theta(2k+1), \ k = 0, 1, \ldots\}$ is a strictly decreasing sequence.

Note that

$$\gamma(\theta(2)) = \tilde{\gamma}(\theta(1)) > 0 = \gamma(\theta(0)) \Rightarrow \theta(0) < \theta(2),$$

and thus

$$\gamma(\theta(1)) = \tilde{\gamma}(\theta(0)) > \tilde{\gamma}(\theta(2)) = \gamma(\theta(3)) \Rightarrow \theta(1) > \theta(3).$$

Therefore the base case of the induction holds. By induction hypothesis, we have $\theta(2k-2) < \theta(2k)$ and $\theta(2k-1) > \theta(2k+1)$. Hence the algorithm yields

$$\gamma(\theta(2k)) = \tilde{\gamma}(\theta(2k-1)) < \tilde{\gamma}(\theta(2k+1)) = \gamma(\theta(2k+2)),$$

which by the monotonicity of $\gamma$ implies that $\theta(2k+2) > \theta(2k)$. Consequently,

$$\gamma(\theta(2k+1)) = \tilde{\gamma}(\theta(2k)) > \tilde{\gamma}(\theta(2k+2)) = \gamma(\theta(2k+3)),$$

which implies $\theta(2k+1) > \theta(2k+3)$. This completes the induction proof for the monotonicity of $\{\theta(2k)\}$ and $\{\theta(2k+1)\}$. 
Being monotone, the sequences \( \{ \theta^{(2k)} \} \) and \( \{ \theta^{(2k-1)} \} \) converge to limits \( \underline{\theta} \) and \( \bar{\theta} \), respectively. Clearly, \( \gamma(\theta) = \tilde{\gamma}(\bar{\theta}) \) and \( \tilde{\gamma}(\theta) = \gamma(\bar{\theta}) \). If there does not exist a pair of \( \theta \)-values, \( \underline{\theta} \neq \bar{\theta} \) satisfying the two preceding equations, then we must have \( \underline{\theta} = \bar{\theta} = \theta^* \). The proof is completed.

**Acknowledgments**

The research of Cheng-Der Fuh was supported in part by NSC 97-2118-M-008-001-MY3. The research of Inchi Hu was partially supported by Hong Kong Research Grant Council.

**References**


