A Revisit of the Optimal Excess-of-Loss Contract

Ernest Aboagye* Vali Asimit† Tsz Chai Fung*‡ Liang Peng* Qiuqi Wang*

September 22, 2024

Abstract

It is well-known that Excess-of-Loss reinsurance has more marketability than Stop-Loss reinsurance, though Stop-Loss reinsurance is the most prominent setting discussed in the optimal (re)insurance design literature. We point out that optimal reinsurance policy under Stop-Loss leads to a zero insolvency probability, which motivates our paper. We remedy this peculiar property of the optimal Stop-Loss reinsurance contract by investigating the optimal Excess-of-Loss reinsurance contract instead. We also provide estimators for the optimal Excess-of-Loss and Stop-Loss contracts and investigate their statistical properties under many premium principle assumptions and various risk preferences, which, according to our knowledge, have never been investigated in the literature. Simulated data and real-life data are used to illustrate our main theoretical findings.

Keywords and phrases: Risk analysis, Optimal Insurance, Nonparametric Estimation.

1 Introduction

1.1 Literature Review

Risk transfer is an effective risk management exercise and consists of transferring liabilities from one or multiple risk holders (known as *insurance buyer(s)*) to another or multiple insurance

*J. Mack Robinson College of Business, Georgia State University, USA. Email addresses: eaboagye1@gsu.edu (Ernest Aboagye), tfung@gsu.edu (Tsz Chai Fung), lpeng@gsu.edu (Liang Peng), qwang30@gsu.edu (Qiuqi Wang).

†Bayes Business School, City, University of London, UK. Email address: asimit@city.ac.uk

‡Corresponding author.

carriers (known as *insurance seller(s)*). Finding the optimal contact between (amongst) two (or more than two) parties has received a huge amount of attention in the literature of actuarial science and operations research. A simple Google Scholar search on September 18, 2024 with the keywords "optimal insurance" and "risk transfer" resulted in 2,910,000 and 5,950,000, respectively, research outputs. This is not surprising since the optimality of such risk management exercise goes beyond understanding insurance liabilities. This paper aims to contribute to the problem of optimal insurance contract of insurance liabilities, which has a very specific trait that is not shared with other sector-specific liabilities (e.g., financial liabilities) in the sense that the insurance liabilities do not have a liquid market so that their value is market-based valuation. *Cost-of-Capital* (CoC) approach is a practical methodology for evaluating insurance liabilities, which are based on the cost of meeting the local capital requirements to hold such liabilities in that territory. In other words, CoC is a regulatory-based methodology that is used within the insurance sector.

The optimal risk transfer problem is often understood in the optimal insurance literature as how an insurer and reinsurer would share the aggregate liability between the two insurance players so that the risk position of the insurer is optimized; the optimization from the reinsurer's point of view is also possible. One may view the problem from both the insurer's and reinsurer's point of view, a case in which the analysis becomes a Pareto optimal insurance contract problem that is a long-standing strand of research established in economic theory with ramifications in insurance and risk literature, but also in the wider operations research field; an insurance perspective could be found in Ruschendorf (2013) and references therein. Other equilibrium concepts are possible; for example, Boonen and Ghossoub (2023) investigate the Bowley equilibrium with risk sharing and optimal reinsurance formulations and focus on the common traits of Bowley optimality and Pareto efficiency under fairly general preferences. Bespoke conditions could be imposed on the optimal (re)insurance contract besides the usual absence of moral hazard; one interesting setting is the so-called Vajda condition that is discussed in Boonen and Jiang (2022).

Depending on the risk preferences, the optimal reinsurance literature is quite rich; e.g., Cai et al. (2008) and Cai and Tan (2007) consider *Value-at-Risk* (VaR) and *Expected Shortfall* (ES) buyer's preferences, while quantile risk and expectile preferences are investigated in Asimit, Badescu and Verdonck (2013), and Cai and Weng (2016), respectively; Balbás, Balbás and Heras (2009) investigates some general risk preferences. In particular, optimal reinsurance problems

under distortion risk measures are well-studied in actuarial science and operations research, where the optimal indemnities are generally piecewise linear; see e.g., Assa (2015), Asimit and Boonen (2018), Assa, Sharifi and Lyons (2021), etc. In light of this, we focus our study on deductible reinsurance contracts, which are natural and common in the literature; see e.g., the early work of Arrow (1963) and Raviv (1979) for insurance and the recent work of Klages-Mundt and Minca (2020) and Cai, Liu and Yin (2024) for operations research.

Optimal reinsurance problems have also been studied under other aspects. The optimal contract from the buyer's point of view in the presence of the seller's counterparty default risk is discussed in Chen (2024), Chi and Tan (2021), Cai, Lemieux, and Liu (2014), Asimit, Badescu and Cheung (2013), and Bernard and Ludkovski (2012). Regulatory considerations are discussed, for example, in Asimit, Chi, and Hu (2015) and Bernard and Tian (2009). Robust formulations are investigated, for example, in Asimit, Hu and Xie (2019), Asimit *et al.* (2017), Balbás, Balbás and Heras (2011), Boonen and Jiang (2024) and Gollier (2014), while Cai, Li and Mao (2023) and Pesenti, Wang and Wang (2024) provide a theoretical perspective to robust decision-making when preferences are ordered by distortion risk measures which are considered in our paper and many other papers in the optimal (re)insurance literature. Non-standard settings are considered in the literature; e.g., Bäuerle and Glauner (2018) investigate the optimal transfer in an insurance network from an economic point of view, while Asimit *et al.* (2016) studies Solvency II capital efficiency through risk transfers within an insurance group.

The optimal insurance problem under expected utility settings is often defined without making any assumption regarding the seller's *premium principle*. When risk preferences are ordered by risk measures, then premium principle assumptions are required. Kaluszka (2001) studies the mean-variance premium principle, Asimit, Badescu and Verdonck (2013) investigate quantile risk premium principles, Tan *et al.* (2020) discuss mean-CVaR premium principle, and Chi and Tan (2013) consider general premium principles, though many other papers rely on certain premium principle assumptions that are specific to the buyer's risk preferences.

1.2 Background and Problem Definition

Throughout this paper, the insurance field is represented by $(\Omega, \mathcal{F}, \mathbb{P})$, an atomless probability space endowed with $L^0 := L^0(\Omega, \mathcal{F}, \mathbb{P})$, the set of all non-negative real-valued random variables on this probability space. Let L^q, $q \in [0, \infty)$, be the set of random variables with finite q^{th} moment, and L^∞ be the set of bounded random variables. A risk measure φ is a function

that maps an element of L^0 to a (extended) real number, i.e. $\varphi : L^0 \to \overline{\Re}$. We recall below some properties for a generic risk measure and generic random variable Y – with *cumulative distribution function* (cdf) F_Y, *survival distribution function* $\bar{F}_Y$, and *generalized left-continuous inverse* $F_Y^-(s) := \inf_{x\in\Re}\left\{F_Y(x) \geq s\right\}$ – representing the future loss of a financial asset or insurance liability.

Convexity: $\varphi(aY_1 + (1-a)Y_2) \leq a\varphi(Y_1) + (1-a)\varphi(Y_2)$ for any $Y_1, Y_2 \in L^0$ and $a \in [0,1]$;

Homogeneous of order $\tau > 0$*:* $\varphi(cY) = c^\tau \varphi(Y)$ for any $Y \in L^0$ and $c \geq 0$;

Shift invariance: $\varphi(Y + c) = \varphi(Y)$ for any $Y \in L^0$ and $c \in \Re$;

Translation invariance: $\varphi(Y + c) = \varphi(Y) + c$ for any $Y \in L^0$ and $c \in \Re$.

These properties are well-known in the literature, and an extensive introduction to risk measures can be found in Föllmer and Schied (2011). Two well-known risk measures are *Value-at-Risk* (VaR) and *Expected Shortfall* (ES), defined as

$$\mathrm{VaR}_p(Y) = F_Y^-(p) \quad \text{and} \quad \mathrm{ES}_p(Y) = \min_{t\in\Re}\left\{t + \frac{1}{1-p}\mathbb{E}(Y-t)_+\right\},$$

where $(\cdot)_+ = \max(\cdot, 0)$ and $p \in (0,1)$ is the risk level. It is evident that the two risk measures are homogeneous of order 1 and translation invariant, and ES is convex.

We are now ready to provide the mathematical formulation of the problem of interest. Suppose that an insurer has insured a large number of policies with independent and identically distributed non-negative losses X_i for $1 \leq i \leq N$ with cdf $F_{X_1}(x)$.

We consider now that the reinsurance premium is calculated by the expected value principle. Thus, the total cost for this portfolio of policies after buying *Excess-of-Loss (EoL)* reinsurance becomes

$$T(d, N, \rho) = \sum_{i=1}^{N}(X_i \wedge d) + (1+\rho)\mathbb{E}\left(\sum_{i=1}^{N}(X_i - d)_+\right), \tag{1}$$

where $\rho > 0$ is the loading factor and $X_i \wedge d = \min(X_i, d)$. A practical question is to find the optimal retention d for $T(d, N, \rho)$ by minimizing the buyer's risk when its perception of risk is modeled by some given risk measures such as VaR and ES.

To better appreciate our study, we first point out an issue with the *Stop-loss (SL)* optimal reinsurance (SL is EoL with $N = 1$) in Cai and Tan (2007), where the total cost $T(d, 1, \rho)$ is stud-

ied; that is, the optimal retention is found via minimizing $\mathrm{VaR}_p\big(T(d,1,\rho)\big)$ or $\mathrm{ES}_p\big(T(d,1,\rho)\big)$, which leads to the following optimal retention

$$d^* = F^-_{X_1}\left(1-\frac{1}{1+\rho}\right) \text{ when } 1-p<(1+\rho)^{-1}. \tag{2}$$

Hence, the optimal retention d^* can be estimated nonparametrically using empirical quantile estimation. Unfortunately, for $1-p<(1+\rho)^{-1}$, we have

$$\mathbb{P}\Big(T(d^*,1,\rho)>\mathrm{VaR}_p\big(T(d^*,1,\rho)\big)\Big)=\mathbb{P}(X_1\wedge d^*>d^*)=0,$$

implying no high risk to the buyer, which is mathematically explained by the truncated buyer's liability $X_1\wedge d$. The same issue remains if one replaces $X_1\wedge d$ by $(\sum_{i=1}^N X_i)\wedge d$, i.e., considering the SL for the total loss instead of one loss in Cai and Tan (2007). Furthermore, the SL optimal retention d^* in (2) is not an explicit function of the risk level p. This is also counter-intuitive as the SL optimal retention may remain constant while p, which implies the insurance company's level of risk aversion, increases.

However, when the number of policies is large enough, $\sum_{i=1}^N(X_i\wedge d)$ will not have such a truncation issue to cause a severely distorted risk level for optimal retention, and thus, the optimal EoL (with $N>1$) retention would not share the same counter-intuitive property as SL (when $N=1$). But, the difficulty in studying the case of $N>1$ is how to minimize a risk measure of $T(d,N,\rho)$ because the exact distribution of $T(d,N,\rho)$ for a given loss distribution of X_i is extremely complicated. Also, it is impossible to estimate the risk measure nonparametrically as we do not have copies of $T(d,N,\rho)$. This motivates us to consider approximately optimal retention by using a normal distribution to approximate the distribution of $\sum_{i=1}^N(X_i\wedge d)$ for a large N.

Before outlining our main contributions, we would like to further differentiate the EoL and SL contracts, which are compared in this paper. Note that EoL has more marketability than SL, as the latter is prohibitively expensive to buyers since the deductible is applied to the annual aggregate loss and not to the individual claims (as for EoL). Other negative traits of SL are not shared with EoL. For example, the loss development of an insurance claim is the process of a claim from reporting until the claim is fully settled, which takes a significant amount of time for many lines of business such as personal accident insurance, medical malpractice insurance, workers compensation, liability claims, etc.; the lag is even larger for long-tail lines of coverage

where arbitrage or court proceedings are more likely to occur. Long lags are big impediments to activating SL contracts since the deductible is applied to the aggregate loss, which is known when all claims from that year are fully settled and that may require many years; this is not the case to EoL where each claim is shared between the buyer and seller.

The main contributions of this paper are two-fold. *First*, we point out that optimal reinsurance policy under SL – one of the most prominent settings discussed in the literature – leads to zero insolvency probability for VaR-based regulatory environments as is the case for EU and UK insurance companies where capital requirements are designed on the 1/200 event basis over a one-year time horizon. This peculiar property of the optimal SL reinsurance contract is the main motivation of our paper, and we show that a remedy is possible if one investigates the optimal EoL reinsurance contract instead. *Second*, we propose an approximately optimal retention, provide a nonparametric estimator for it, and derive its statistical properties under many premium principle assumptions and various risk preferences, which, according to our knowledge, have never been investigated in the literature.

The paper is organized as follows: EoL risk model is considered under the VaR_p risk measure in Section 2 for various premium principles, which are further generalized in Section 3 when the risk preferences are ordered by distortion risk measures. Some simulation studies are provided in Section 4, while real data analysis is employed in Section 5. Additional simulation results and discussions on using a higher-order approximation are provided in the online supplementary file.

2 Approximately optimal retention for VaR

In this section, we consider the total cost of $T(d, N, \rho)$ under the VaR_p risk measure. Later, we will generalize the result to distortion risk measures in Section 3. Throughout, we use $A_N = O(B_N), o(B_N), O_p(B_N)$, and $o_p(B_N)$ to denote A_N/B_N is bounded, goes to zero, bounded in probability, and goes to zero in probability, respectively, as $N \to \infty$. We also use $\Phi(x)$ and $\Phi^-(x)$ to denote a standard normal distribution and its quantile function, respectively.

Because VaR is translation invariant, we have

$$\mathrm{VaR}_p\big(T(d, N, \rho)\big) = \mathrm{VaR}_p\left(\sum_{i=1}^{N}(X_i \wedge d)\right) + (1+\rho)N\mathbb{E}\{(X_1 - d)_+\}. \tag{3}$$

Define

$$\begin{cases} \mu_1(d) = \mathbb{E}(X_1 \wedge d) = \int_0^d \bar{F}_{X_1}(x)\,dx, \\ \mu_2(d) = \mathbb{E}(X_1^2 \wedge d^2) = 2\int_0^d \bar{F}_{X_1}(x)x\,dx, \\ \nu_1(d) = \mathbb{E}\{(X_1 - d)_+\} = \int_d^\infty \bar{F}_{X_1}(x)\,dx. \end{cases}$$

For large N, it follows from the Central Limit Theorem that

$$\mathrm{VaR}_p\left(\sum_{i=1}^N (X_i \wedge d)\right) = N\mu_1(d) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\Phi^-(p) + o(\sqrt{N}), \tag{4}$$

Therefore, instead of minimizing $\mathrm{VaR}_p\big(T(d, N, \rho)\big)$ to obtain the optimal retention d, an intuitive idea is to ignore the $o(\sqrt{N})$ term in (4) to approximate the right hand side of (3), i.e., we propose to minimize

$$G_{N,\rho}(d) := N\mathbb{E}(X_1) + N\rho\nu_1(d) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\Phi^-(p), \tag{5}$$

whose solution is called an *approximately optimal retention*. The above approximation method needs to be carefully justified because the $o(\sqrt{N})$ term in (4) may depend on d. Hence, we have the following result for a bound on the $o(\sqrt{N})$ term that is independent of d.

Proposition 1. *For any $0 < d_1 < d_2 \le \infty$ such that $\mu_2(d_1) - \mu_1^2(d_1) > 0$ and $\mathbb{E}(|X_i \wedge d_2|^3) < \infty$, we have*

$$\sup_{d_1 \le d \le d_2} \big|\mathrm{VaR}_p\big(T(d, N, \rho)\big) - G_{N,\rho}(d)\big| \le C \tag{6}$$

for some constant $C > 0$ depending on $\mu_2(d_1) - \mu_1^2(d_1)$ and $\mathbb{E}(X_i^3 \wedge d_2^3)$ but independent of d and $F_{X_1}(x)$.

Proof. Define

$$S_N(d) = \frac{\sum_{i=1}^N (X_i \wedge d) - N\mu_1(d)}{\sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}}.$$

It follows from the Berry-Esseen Theorem that

$$\sup_x |\mathbb{P}(S_N(d) \le x) - \Phi(x)| \le \frac{C_1\mathbb{E}(X_i^3 \wedge d^3)}{(\sqrt{\mu_2(d) - \mu_1^2(d)})^3\sqrt{N}}$$

for some constant $C_1 > 0$ independent of d and F_{X_1}. Because

$$\max_{d_1 \le d \le d_2} \mathbb{E}(X_i^3 \wedge d^3) \le \mathbb{E}(X_i^3 \wedge d_2^3) \text{ and } \min_{d_1 \le d \le d_2} \{\mu_2(d) - \mu_1^2(d)\} \ge \mu_2(d_1) - \mu_1^2(d_1),$$

we have

$$\sup_x |\mathbb{P}(S_N(d) \leq x) - \Phi(x)| \leq C_2 N^{-1/2} \text{ uniformly in } d \in [d_1, d_2] \tag{7}$$

for some constant $C_2 > 0$ depending on $\mu_2(d_1) - \mu_1^2(d_1)$ and $\mathbb{E}(X_i^3 \wedge d_2^3)$ but independent of d and $F_{X_1}(x)$. Hence,

$$\mathbb{P}(S_N(d) \leq \Phi^-(p + C_2 N^{-1/2})) \geq p \text{ and } \mathbb{P}(S_N(d) \leq \Phi^-(p - C_2 N^{-1/2})) \leq p$$

uniformly in $d \in [d_1, d_2]$, implying that

$$\Phi^-(p - C_2 N^{-1/2}) \leq \text{VaR}_p(S_N(d)) \leq \Phi^-(p + C_2 N^{-1/2}) \text{ uniformly in } d \in [d_1, d_2],$$

i.e.,

$$\sup_{d_1 \leq d \leq d_2} |\text{VaR}_p(S_N(d)) - \Phi^-(p)| \leq C_3 N^{-1/2}$$

for some constant $C_3 > 0$ depending on $\mu_2(d_1) - \mu_1^2(d_1)$ and $\mathbb{E}(X_i^3 \wedge d_2^3)$ but independent of d and $F_{X_1}(x)$. Hence, the proposition follows by noting that

$$\text{VaR}_p(\sum_{i=1}^{N}(X_i \wedge d)) = N\mu_1(d) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\text{VaR}_p(S_N(d)).$$

□

Hence, the $o(\sqrt{N})$ term indeed has an order of $O(1)$ uniformly. The above proposition yields the following results, which ensure that both the actual and approximate optimal retention, along with the actual and approximate minimum VaR of the total costs, are asymptotically equivalent. These findings validate the application of the approximation method in (5).

Proposition 2. *Suppose that for sufficiently large N, there exists an approximately optimal retention d_N^* solving $G'_{N,\rho}(d) = 0$, such that $d_N^* \to d^* \in (0, \infty)$, where d^* solves $G'(d) = 0$ and the function $G(d)$ satisfies $\sup_{d_1 \leq d \leq d_2} |G_{N,\rho}(d) - G(d)| = o(\sqrt{N})$ with $d_1 < d^* < d_2$. Here, ρ may depend on d and N. Then, under conditions of Proposition 1, there exists an N_0 and a local minimizer d_N° of $VaR_p(T(d, N, \rho))$ for each $N > N_0$ such that: (i) $d_N^\circ \to d^*$ as $N \to \infty$; and (ii) $\lim_{N\to\infty} |\text{VaR}_p(T(d_N^\circ, N, \rho)) - N\mathbb{E}(X_1)|/\sqrt{N} = \lim_{N\to\infty} |G_{N,\rho}(d_N^*) - N\mathbb{E}(X_1)|/\sqrt{N}$.*

Proof. From Proposition 1 and the conditions above, we have $\sup_{d_1 \leq d \leq d_2} |\text{VaR}_p\big(T(d, N, \rho)\big) - G(d)| = o(\sqrt{N})$. The results follow from an application of Theorem 5.7 of Van der Vaart

(2000). □

Remark 1. *To enhance accuracy in (5), particularly when N is small, one might consider using Edgeworth expansion, which incorporates higher moments to approximate the distribution of $\sum_{i=1}^{N}(X_i \wedge d)$. Mathematical details and numerical results for this approach are provided in Section 2 of the supplementary materials for those interested. However, due to the complexity of the mathematical expressions involved, deriving the theoretical properties of this advanced approximation method is excessively complicated, and we leave it as a topic for future research.*

In the following subsections, we examine the optimal retention under different specifications of the loading factor ρ: constant loading factor (Section 2.1), decreasing loading factor (Section 2.2), standard deviation principle (Section 2.3), and Sharpe ratio principle (Section 2.4).

2.1 Constant Loading Factor

We now solve (5) with a constant loading factor $\rho > 0$. In this case, the optimal retention turns out to be a solution to

$$H_{N,\rho}(d) := \{d - \mu_1(d)\}^2 - \left(\frac{\sqrt{N}\rho}{\Phi^-(p)}\right)^2 \{\mu_2(d) - \mu_1^2(d)\} = 0. \tag{8}$$

The next result stated as Theorem 1 shows that (8) admits a unique solution under some very mild regularity conditions. Recall that we allow $F_{X_1}(0) > 0$ in Theorem 1, which means that the event of having no claim is not a null set.

Theorem 1. *Assume $\mathbb{E}(X_1) < \infty$, $F_{X_1}(\cdot)$ has the support $[0,\infty)$ (i.e., $F_{X_1}(0) > 0$) or $(0,\infty)$, and is continuous on $(0,\infty)$. When the support is $[0,\infty)$, we further assume $F_{X_1}(0) < \frac{N\rho^2}{N\rho^2+(\Phi^-(p))^2}$, which is always true when N is large enough. Then, there exists a unique approximately optimal retention $d^*_{N,\rho} \in (0,\infty)$ such that*

$$d^*_{N,\rho} = \operatorname*{argmin}_{d>0} G_{N,\rho}(d) \quad \textit{and} \quad H_{N,\rho}(d^*_{N,\rho}) = 0.$$

Proof. Note that $\mu_1'(d) = \bar{F}_{X_1}(d)$, $\mu_2'(d) = 2d\bar{F}_{X_1}(d)$, and $\nu_1'(d) = -\bar{F}_{X_1}(d)$, and in turn,

$$G'_{N,\rho}(d) \;=\; -N\rho\bar{F}_{X_1}(d) + \sqrt{N}\bar{F}_{X_1}(d)\frac{d-\mu_1(d)}{\sqrt{\mu_2(d)-\mu_1^2(d)}}\Phi^-(p) \tag{9}$$

$$= \frac{\sqrt{N}\bar{F}_{X_1}(d)\Phi^-(p)}{\left(\frac{d-\mu_1(d)}{\sqrt{\mu_2(d)-\mu_1^2(d)}}+\frac{\sqrt{N}\rho}{\Phi^-(p)}\right)\{\mu_2(d)-\mu_1^2(d)\}}H_{N,\rho}(d).$$

Hence, solving $G'_{N,\rho}(d)=0$ for $d\in(0,\infty)$ is equivalent to solving $H_{N,\rho}(d)=0$ for $d\in(0,\infty)$. That is, we only need to show that there is a unique solution for $H_{N,\rho}(d)=0$.

Since $d-\mu_1(d)>0$ and $\mu_2(d)-\mu_1^2(d)>0$ for all $d>0$, and

$$H'_{N,\rho}(d)=2\{d-\mu_1(d)\}\left(F_{X_1}(d)\left(1+\left(\frac{\sqrt{N}\rho}{\Phi^-(p)}\right)^2\right)-\left(\frac{\sqrt{N}\rho}{\Phi^-(p)}\right)^2\right),$$

we conclude that $H'_{N,\rho}(d)$ is negative, zero, and positive for $0<d<d_1$, $d_1\le d\le d_2$, and $d>d_2$, respectively, where $F_{X_1}(d)=\frac{N\rho^2}{N\rho^2+(\Phi^-(p))^2}$ happens and only happens on $d\in[d_1,d_2]$, which is ensured by the conditions that the right endpoint of $F_{X_1}(x)$ is infinity, $F_{X_1}(x)$ is continuous on $(0,\infty)$, and $F_{X_1}(0)<\frac{N\rho^2}{N\rho^2+(\Phi^-(p))^2}$. That is,

$$H_{N,\rho}(d) \text{ is strictly decreasing on } (0,d_1),\ \text{constant on } [d_1,d_2],\ \text{and strictly increasing on } (d_2,\infty), \tag{10}$$

Note that

$$\lim_{d\to\infty}\frac{d^2}{\mu_2(d)}=\begin{cases}\lim_{d\to\infty}\frac{2d}{2d\bar{F}_{X_1}(d)}=\infty & \text{if } \mu_2(\infty)=\infty,\\ \infty & \text{if } \mu_2(\infty)<\infty.\end{cases}$$

Thus, $\lim_{d\to\infty}H_{N,\rho}(d)/d^2=1$ and $\lim_{d\to\infty}H_{N,\rho}(d)=\infty$. The latter, (9) and (10), and the fact that $\lim_{d\to0}H_{N,\rho}(d)=0$ conclude that $H_{N,\rho}(d)=0$ has a unique solution on (d_2,∞). The proof is now complete. □

Note that $d^*_{N,\rho}$ diverges to infinity as $N\to\infty$, which is not surprising since a constant ρ for any N implies that the seller does not include the diversification effect in its premium calculation, case in which the seller would not be incentivized to participate in such reinsurance contract. The divergence of $d^*_{N,\rho}$ also violates the conditions of Proposition 2, indicating that the approximately optimal retention can still differ significantly from the actual optimal retention, even when N is large under the constant loading factor rule. This issue will be demonstrated in the numerical example in Section 4.1. To mitigate this, a more practical approach would be to adjust or reduce the loading factor as N increases, ensuring a more realistic reinsurance premium. This will be further explored in the following subsections.

2.2 Decreasing Loading Factor

To estimate $d^*_{N,\rho}$ and study its asymptotic properties, we consider instead a bounded approximated optimal retention by assuming $\rho = \rho_N$ such that

$$\lim_{N\to\infty} \rho_N \sqrt{N} = \delta \in (0, \infty). \tag{11}$$

It follows from Theorem 1 that d^*_{N,ρ_N} is the unique solution to $H_{N,\rho_N}(d) = 0$. Using (11), we know that $\lim_{N\to\infty} d^*_{N,\rho_N}$ exists and is the unique solution to

$$\{d - \mu_1(d)\}^2 - (\frac{\delta}{\Phi^-(p)})^2\{\mu_2(d) - \mu_1^2(d)\} = 0.$$

Hence, write $G(d) = N\mathbb{E}(X_1) + \sqrt{N}\delta\nu_1(d) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\Phi^-(p)$, and the conditions of Proposition 2 holds because $\sup_{d_1\le d\le d_2} |G_{N,\rho_N}(d) - G(d)| = |\rho_N\sqrt{N} - \delta| \sup_{d_1\le d\le d_2} \sqrt{N}\nu_1(d) \le |\rho_N\sqrt{N} - \delta|\sqrt{N}\mathbb{E}(X_1) = o(\sqrt{N})$. This justifies the minimization of $G_{N,\rho_N}(d)$ instead of $\mathrm{VaR}_p(T(d, N, \rho_N))$.

To estimate d^*_{N,ρ_N} nonparametrically, we solve the following equation

$$\hat{H}_{N,\rho_N}(d) := \{d - \hat{\mu}_1(d)\}^2 - \left(\frac{\sqrt{N}\rho_N}{\Phi^-(p)}\right)^2 \{\hat{\mu}_2(d) - \hat{\mu}_1^2(d)\} = 0, \tag{12}$$

where

$$\hat{\mu}_1(d) = \frac{1}{N}\sum_{i=1}^N (X_i \wedge d) \text{ and } \hat{\mu}_2(d) = \frac{1}{N}\sum_{i=1}^N (X_i^2 \wedge d^2). \tag{13}$$

Let $\hat{d}^*_{N,\rho_N}$ denote this solution, which is an estimator for d^*_{N,ρ_N}. Let $\Sigma(d)$ denote the covariance matrix of $\boldsymbol{Z}_i(d)$, where $\boldsymbol{Z}_i(d) = (X_i \wedge d, X_i^2 \wedge d^2)^\tau$, and define

$$\hat{\mu}_1^*(d) = \frac{1}{N}\sum_{i=1}^N I(X_i > d) \text{ and } \hat{\mu}_2^*(d) = \frac{2d}{N}\sum_{i=1}^N I(X_i > d), \tag{14}$$

which estimate the first-order derivatives, $\mu_1'(d)$ and $\mu_2'(d)$, respectively. The asymptotic properties of $\hat{d}^*_{N,\rho_N}$ are given in Theorem 2.

Theorem 2. *Under conditions of Theorem 1 and (11), we have*

$$\frac{\sqrt{N}\{\hat{d}^*_{N,\rho_N} - d^*_{N,\rho_N}\}}{\hat{c}_0^{-1}\sqrt{(\hat{c}_1, \hat{c}_2)\hat{\Sigma}_0(\hat{c}_1, \hat{c}_2)^\tau}} \xrightarrow{d} N(0, 1),$$

where

$$\hat{c}_0 = 2\{\hat{d}^*_{N,\rho_N} - \hat{\mu}_1(\hat{d}^*_{N,\rho_N})\}\{1 - \hat{\mu}^*_1(\hat{d}^*_{N,\rho_N})\} - \left(\frac{\rho_N\sqrt{N}}{\Phi^-(p)}\right)^2 \{\hat{\mu}^*_2(\hat{d}^*_{N,\rho_N}) - 2\hat{\mu}_1(\hat{d}^*_{N,\rho_N})\hat{\mu}^*_1(\hat{d}^*_{N,\rho_N})\},$$

$$\hat{c}_1 = 2\{\hat{d}^*_{N,\rho_N} - \hat{\mu}_1(\hat{d}^*_{N,\rho_N})\} - \left(\frac{\rho_N\sqrt{N}}{\Phi^-(p)}\right)^2 2\hat{\mu}_1(\hat{d}^*_{N,\rho_N}),\ \hat{c}_2 = \left(\frac{\rho_N\sqrt{N}}{\Phi^-(p)}\right)^2,$$

$$\hat{\Sigma}_0 = \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{Z}_i(\hat{d}^*_{N,\rho_N}) - \frac{1}{N}\sum_{i'=1}^{N}\boldsymbol{Z}_{i'}(\hat{d}^*_{N,\rho_N})\right]\left[\boldsymbol{Z}_i(\hat{d}^*_{N,\rho_N}) - \frac{1}{N}\sum_{i'=1}^{N}\boldsymbol{Z}_{i'}(\hat{d}^*_{N,\rho_N})\right]^\tau.$$

Proof. For simplicity, the proof uses d^* and $\hat{d}^*$ for d^*_{N,ρ_N} and $\hat{d}^*_{N,\rho_N}$, respectively. Then, the central limit theorem implies

$$\sqrt{N}\begin{pmatrix}\hat{\mu}_1(d^*) - \mu_1(d^*)\\ \hat{\mu}_2(d^*) - \mu_2(d^*)\end{pmatrix} \xrightarrow{\mathrm{d}} N(\mathbf{0}, \Sigma_0) \tag{15}$$

when

$$\Sigma(d^*) \to \Sigma_0 \text{ as } N \to \infty. \tag{16}$$

It follows from (15) that

$$\begin{aligned}\hat{\mu}_1(\hat{d}^*) - \mu_1(d^*) &= \hat{\mu}_1(\hat{d}^*) - \mu_1(\hat{d}^*) + \mu_1(\hat{d}^*) - \mu_1(d^*)\\ &= \{\hat{\mu}_1(d^*) - \mu_1(d^*)\} + \mu'_1(d^*)\{\hat{d}^* - d^*\} + o_p(|\hat{d}^* - d^*|),\end{aligned}$$

$$\hat{\mu}_2(\hat{d}^*) - \mu_2(d^*) = \{\hat{\mu}_2(d^*) - \mu_2(d^*)\} + \mu'_2(d^*)\{\hat{d}^* - d^*\} + o_p(|\hat{d}^* - d^*|),$$

$$\begin{aligned}&\{\hat{d}^* - \hat{\mu}_1(\hat{d}^*)\}^2 - \{d^* - \mu_1(d^*)\}^2\\ =\ & 2\{d^* - \mu_1(d^*)\}\{1 - \mu'_1(d^*)\}(\hat{d}^* - d^*)\\ & -2\{d^* - \mu_1(d^*)\}\{\hat{\mu}_1(d^*) - \mu_1(d^*))\} + o_p(|\hat{d}^* - d^*|),\end{aligned}$$

$$\begin{aligned}&\{\hat{\mu}_2(\hat{d}^*) - \hat{\mu}^2_1(\hat{d}^*)\} - \{\mu_2(d^*) - \mu^2_1(d^*)\}\\ =\ & \{\hat{\mu}_2(d^*) - \mu_2(d^*)\} - 2\mu_1(d^*)\{\hat{\mu}_1(d^*) - \mu_1(d^*)\}\\ & +\{\mu'_2(d^*) - 2\mu_1(d^*)\mu'_1(d^*)\}\{\hat{d}^* - d^*\} + o_p(|\hat{d}^* - d^*|),\end{aligned}$$

implying that

$$\begin{aligned}0 &= \hat{H}_{N,\rho_N}(\hat{d}^*) - H_{N,\rho_N}(d^*)\\ &= 2\{d^* - \mu_1(d^*)\}\{1-\mu_1'(d^*)\}\{\hat{d}^* - d^*\} - 2\{d^* - \mu_1(d^*)\}\{\hat{\mu}_1(d^*) - \mu_1(d^*)\}\\ &\quad - \left(\tfrac{\delta}{\Phi^-(p)}\right)^2 \Big\{\{\hat{\mu}_2(d^*) - \mu_2(d^*)\} - 2\mu_1(d^*)\{\hat{\mu}_1(d^*) - \mu_1(d^*)\}\\ &\quad + \{\mu_2'(d^*) - 2\mu_1(d^*)\mu_1'(d^*)\}\{\hat{d}^* - d^*\}\Big\} + o_p(1/\sqrt{N}) + o_p(|\hat{d}^* - d^*|),\end{aligned}$$

i.e.,

$$c_0\{\hat{d}^* - d^*\} = c_1\{\hat{\mu}_1(d^*) - \mu_1(d^*)\} + c_2\{\hat{\mu}_2(d^*) - \mu_2(d^*)\} + o_p(1/\sqrt{N}) + o_p(|\hat{d}^* - d^*|), \quad (17)$$

where

$$c_0 = 2\{d^* - \mu_1(d^*)\}\{1-\mu_1'(d^*)\} - \left(\frac{\delta}{\Phi^-(p)}\right)^2 \{\mu_2'(d^*) - 2\mu_1(d^*)\mu_1'(d^*)\},$$

$$c_1 = 2\{d^* - \mu_1(d^*)\} - \left(\frac{\delta}{\Phi^-(p)}\right)^2 2\mu_1(d^*), \text{ and } c_2 = \left(\frac{\delta}{\Phi^-(p)}\right)^2.$$

Note that the terms $o_p(1/\sqrt{N})$ and $o_p(|\hat{d}^* - d^*|)$ in (17) can be neglected. This is because, as seen from (15), both $c_1\{\hat{\mu}_1(d^*) - \mu_1(d^*)\}$ and $c_2\{\hat{\mu}_2(d^*) - \mu_2(d^*)\}$ are of order $O_p(1/\sqrt{N})$, and $o_p(|\hat{d}^* - d^*|)$ is clearly dominated by $\hat{d}^* - d^* = O_p(|\hat{d}^* - d^*|)$. Hence,

$$\sqrt{N}\{\hat{d}^* - d^*\} \xrightarrow{d} N\left(0, \frac{1}{c_0^2}(c_1, c_2)\Sigma_0(c_1, c_2)^\tau\right),$$

which implies our main result since $\hat{c}_0, \hat{c}_1, \hat{c}_2, \hat{\Sigma}_0$ are consistent estimators of c_0, c_1, c_2, Σ_0, respectively. The proof is now complete. □

2.3 Standard Deviation Premium Principle

We extend the analysis in Section 2.2 by assuming a decreasing loading factor and the standard deviation principle. Specifically, a particular choice of ρ is assumed in (5) as

$$\rho = \rho_0 \mathrm{SD}\left(\frac{1}{N}\sum_{i=1}^N (X_i - d)_+\right) = \rho_0 N^{-1/2}\sqrt{\nu_2(d) - \nu_1^2(d)}, \quad (18)$$

which depends on both N and d and satisfies (11), where

$$\nu_2(d) = \mathbb{E}\{(X_i - d)_+^2\} = 2\int_d^\infty \bar{F}_{X_1}(x)(x-d)\,dx \text{ satisfying } \nu_2'(d) = -2\nu_1(d).$$

Hence, the total cost for the insurer becomes

$$\tilde{T}(d) = \sum_{i=1}^{N}(X_i \wedge d) + N\nu_1(d) + \rho_0\sqrt{N}\nu_1(d)\sqrt{\nu_2(d) - \nu_1^2(d)},$$

and the optimal retention should minimize

$$\begin{aligned}\mathrm{VaR}_p(\tilde{T}(d)) \quad &= N\mu_1(d) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\Phi^-(p) + o(\sqrt{N}) \\ &\quad + N\nu_1(d) + \rho_0\sqrt{N}\nu_1(d)\sqrt{\nu_2(d) - \nu_1^2(d)}.\end{aligned}$$

Once again, we seek for d that minimizes (19) below as we ignore the $o(\sqrt{N})$ terms:

$$\tilde{G}(d) = N\mathbb{E}(X_1) + \sqrt{N}\{\Phi^-(p)\sqrt{\mu_2(d) - \mu_1^2(d)} + \rho_0\nu_1(d)\sqrt{\nu_2(d) - \nu_1^2(d)}\}. \tag{19}$$

The existence of the approximately optimal retention is shown in Theorem 3 below.

Theorem 3. *Assume $F_{X_1}(x)$ has the support $[0,\infty)$ (i.e., $F_{X_1}(0) > 0$) or $(0,\infty)$, is continuous on $(0,\infty)$, and*

$$\lim_{t\to\infty}\frac{\bar{F}_{X_1}(tx)}{\bar{F}_{X_1}(t)} = x^{-\alpha} \text{ for all } x > 0 \text{ and some } \alpha > 2. \tag{20}$$

If $F_{X_1}(0) > 0$, we further assume that

$$\Phi^-(p)\sqrt{\bar{F}_{X_1}(0)F_{X_1}(0)} < \rho_0\bar{F}_{X_1}(0)\sqrt{\mathbb{E}(X_1^2) - \{\mathbb{E}(X_1)\}^2} + \rho_0\frac{F_{X_1}(0)\{\mathbb{E}(X_1)\}^2}{\sqrt{\mathbb{E}(X_1^2) - \{\mathbb{E}(X_1)\}^2}}. \tag{21}$$

Then, there exists at least one solution of $\tilde{G}'(d) = 0$, and an approximately optimal retention $d^* \in (0,\infty)$ is its smallest solution, which is a local minimum of $\tilde{G}(d)$.

Proof. Because

$$\frac{\tilde{G}'(d)}{\sqrt{N}} = \Phi^-(p)\frac{\bar{F}_{X_1}(d)\{d - \mu_1(d)\}}{\sqrt{\mu_2(d) - \mu_1^2(d)}} - \rho_0\bar{F}_{X_1}(d)\sqrt{\nu_2(d) - \nu_1^2(d)} - \rho_0\frac{F_{X_1}(d)\nu_1^2(d)}{\sqrt{\nu_2(d) - \nu_1^2(d)}}$$

and

$$\lim_{d\to 0}\frac{\{d - \mu_1(d)\}^2}{\mu_2(d) - \mu_1^2(d)} = \lim_{d\to 0}\frac{2\{d - \mu_1(d)\}F_{X_1}(d)}{2\bar{F}_{X_1}(d)\{d - \mu_1(d)\}} = \frac{F_{X_1}(0)}{\bar{F}_{X_1}(0)}, \tag{22}$$

it follows from (21) in Theorem 3 that

$$\begin{aligned}\lim_{d\to 0}\frac{\tilde{G}'(d)}{\sqrt{N}} &= \Phi^{-}(p)\sqrt{\bar{F}_{X_1}(0)F_{X_1}(0)} - \rho_0\bar{F}_{X_1}(0)\sqrt{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2} \\ &\quad -\rho_0\frac{F_{X_1}(0)\{\mathbb{E}(X_1)\}^2}{\sqrt{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2}} \\ &< 0.\end{aligned} \tag{23}$$

By (20), we have

$$\lim_{d\to\infty}\frac{\nu_1(d)}{d\bar{F}_{X_1}(d)} = \frac{1}{\alpha-1} \text{ and } \lim_{d\to\infty}\frac{\nu_2(d)}{d^2\bar{F}_{X_1}(d)} = \frac{2}{(\alpha-1)(\alpha-2)}, \tag{24}$$

implying that

$$\lim_{d\to\infty}\frac{\tilde{G}'(d)}{\sqrt{N}\bar{F}_{X_1}(d)d} = \frac{\Phi^{-}(p)}{\sqrt{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2}} > 0. \tag{25}$$

Hence, it follows from (23) and (25) that there exists at least one solution of $\tilde{G}'(d) = 0$ for $d \in (0,\infty)$, and let d^* be the smallest solution. Then, there exists $d_1 > d^*$ such that $\tilde{G}'(d) < 0$ for $d \in (0, d^*)$, $\tilde{G}'(d) \geq 0$ for $d \in (d^*, d_1)$, and $\tilde{G}'(d_1) > 0$, implying that d^* is a local minimum of $\tilde{G}(d)$ for $d \in (0,\infty)$. The proof is now complete. □

Since the solution d^* in Theorem 3 is independent of N, it is clear that the conditions of Proposition 2 hold, justifying our approximation strategy. To estimate the optimal retention nonparametrically, we minimize the following function for d:

$$\hat{\tilde{G}}(d) = \Phi^{-}(p)\sqrt{\hat{\mu}_2(d)-\hat{\mu}_1^2(d)} + \rho_0\hat{\nu}_1(d)\sqrt{\hat{\nu}_2(d)-\hat{\nu}_1^2(d)}, \tag{26}$$

where $\hat{\mu}_1(d)$ and $\hat{\mu}_2(d)$ are given by (13),

$$\hat{\nu}_1(d) = \frac{1}{N}\sum_{i=1}^{N}(X_i-d)_+, \text{ and } \hat{\nu}_2(d) = \frac{1}{N}\sum_{i=1}^{N}(X_i-d)_+^2. \tag{27}$$

Denote $\tilde{d}^*_{N,\rho_0}$ and $\hat{\tilde{d}}^*_{N,\rho_0}$ as the smallest solution of $\tilde{G}'(d) = 0$ and $\hat{\tilde{G}}'(d) = 0$ with $\tilde{G}(d)$ and $\hat{\tilde{G}}(d)$ given in (19) and (26), respectively. Put $\tilde{\boldsymbol{Z}}_i(d) = (I(X_i > d), X_i\wedge d, X_i^2\wedge d^2, (X_i-d)_+, (X_i-d)_+^2)^\tau$ with A^τ denoting the transpose of vector or matrix A, and let $\tilde{\Sigma}(d)$ be the covariance matrix of $\tilde{\boldsymbol{Z}}_i(d)$. Define

$$\hat{\nu}_1^*(d) = -\frac{1}{N}\sum_{i=1}^{N}I(X_i > d) \text{ and } \hat{\nu}_2^*(d) = -2\hat{\nu}_1(d) \tag{28}$$

to estimate $\nu_1'(d)$ and $\nu_2'(d)$ on top of $\hat{\mu}_1^*(d)$ and $\hat{\mu}_2^*(d)$ which are defined in (14). We further denote $\hat{\bar{F}}_{X_1}(d) = \sum_{i=1}^N I(X_i > d)/N$ as the empirical survival function of X_1 and $\hat{f}_{X_1}(d)$ as any consistent estimator of the density function for X_1, e.g., a kernel density estimation. The asymptotic properties of the approximately optimal retention are provided in Theorem 4.

Theorem 4. *Under conditions of Theorem 3 and (18), and that X_1 has a density function f_{X_1}, we have*

$$\frac{\sqrt{N}\{\hat{\tilde{d}}^*_{N,\rho_0} - \tilde{d}^*_{N,\rho_0}\}}{\hat{b}_0^{-1}\sqrt{\hat{\boldsymbol{b}}\hat{\tilde{\Sigma}}_0\hat{\boldsymbol{b}}^\tau}} \xrightarrow{d} N(0,1),$$

with $\hat{\boldsymbol{b}} := (\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4, \hat{b}_5)$, where

$$\begin{aligned}
\hat{b}_0 &= \frac{\Phi^-(p)\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})}{\sqrt{\hat{\mu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\mu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})}} - \hat{b}_1\hat{f}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0}) + \hat{b}_2\hat{\mu}_1^*(\hat{\tilde{d}}^*_{N,\rho_0}) + \hat{b}_3\hat{\mu}_2^*(\hat{\tilde{d}}^*_{N,\rho_0}) \\
&\qquad + \hat{b}_4\hat{\nu}_1^*(\hat{\tilde{d}}^*_{N,\rho_0}) + \hat{b}_5\hat{\nu}_2^*(\hat{\tilde{d}}^*_{N,\rho_0}), \\
\hat{b}_1 &= \Phi^-(p)\frac{\hat{\tilde{d}}^*_{N,\rho_0} - \hat{\mu}_1(\hat{\tilde{d}}^*_{N,\rho_0})}{\sqrt{\hat{\mu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\mu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})}} - \rho_0\frac{\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - 2\hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})}{\sqrt{\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})}}, \\
\hat{b}_2 &= \Phi^-(p)\left[-\frac{\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})}{\sqrt{\hat{\mu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\mu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})}} + \frac{\hat{\mu}_1(\hat{\tilde{d}}^*_{N,\rho_0})\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})[\hat{\tilde{d}}^*_{N,\rho_0} - \hat{\mu}_1(\hat{\tilde{d}}^*_{N,\rho_0})]}{(\hat{\mu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\mu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0}))^{3/2}}\right], \\
\hat{b}_3 &= -\Phi^-(p)\frac{\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})[\hat{\tilde{d}}^*_{N,\rho_0} - \hat{\mu}_1(\hat{\tilde{d}}^*_{N,\rho_0})]}{2(\hat{\mu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\mu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0}))^{3/2}}, \\
\hat{b}_4 &= -2\rho_0\hat{\nu}_1(\hat{\tilde{d}}^*_{N,\rho_0})\left[\frac{1 - 2\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})}{\sqrt{\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})}}\right. \\
&\qquad \left. + \frac{\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0})\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0}) + \hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})[1 - 2\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})]}{2(\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0}))^{3/2}}\right], \\
\hat{b}_5 &= \rho_0\left[\frac{\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})}{\sqrt{\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})}} - \frac{\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0})\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0}) + \hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0})[1 - 2\hat{\bar{F}}_{X_1}(\hat{\tilde{d}}^*_{N,\rho_0})]}{2(\hat{\nu}_2(\hat{\tilde{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\tilde{d}}^*_{N,\rho_0}))^{3/2}}\right], \\
\hat{\tilde{\Sigma}}_0 &= \frac{1}{N}\sum_{i=1}^N\left[\tilde{\boldsymbol{Z}}_i(\hat{\tilde{d}}^*_{N,\rho_0}) - \frac{1}{N}\sum_{i'=1}^N\tilde{\boldsymbol{Z}}_{i'}(\hat{\tilde{d}}^*_{N,\rho_0})\right]\left[\tilde{\boldsymbol{Z}}_i(\hat{\tilde{d}}^*_{N,\rho_0}) - \frac{1}{N}\sum_{i'=1}^N\tilde{\boldsymbol{Z}}_{i'}(\hat{\tilde{d}}^*_{N,\rho_0})\right]^\tau.
\end{aligned}$$

Proof. For notational convenience, we write d^* and $\hat{d}^*$ for $\tilde{d}^*_{N,\rho_0}$ and $\hat{\tilde{d}}^*_{N,\rho_0}$, respectively. Then,

the central limit theorem implies

$$\sqrt{N}\begin{pmatrix} \hat{\bar{F}}_{X_1}(d^*) - \bar{F}_{X_1}(d^*) \\ \hat{\mu}_1(d^*) - \mu_1(d^*) \\ \hat{\mu}_2(d^*) - \mu_2(d^*) \\ \hat{\nu}_1(d^*) - \nu_1(d^*) \\ \hat{\nu}_2(d^*) - \nu_2(d^*) \end{pmatrix} \xrightarrow{\mathrm{d}} N(\mathbf{0}, \tilde{\Sigma}_0) \tag{29}$$

when $\tilde{\Sigma}(d^*) \to \tilde{\Sigma}_0$ as $N \to \infty$. Expansion of $\hat{\tilde{G}}'(\hat{d}^*) - \tilde{G}'(d^*)$ yields

$$\begin{aligned} 0 &= \hat{\tilde{G}}'(\hat{d}^*) - \tilde{G}'(d^*) \qquad (30) \\ &= \frac{\Phi^-(p)\bar{F}_{X_1}(d^*)}{\sqrt{\mu_2(d^*) - \mu_1^2(d^*)}}\left[\hat{d}^* - d^*\right] + b_1\left[\hat{\bar{F}}_{X_1}(\hat{d}^*) - \bar{F}_{X_1}(d^*)\right] \\ &\quad + b_2\left[\hat{\mu}_1(\hat{d}^*) - \mu_1(d^*)\right] + b_3\left[\hat{\mu}_2(\hat{d}^*) - \mu_2(d^*)\right] \\ &\quad + b_4\left[\hat{\nu}_1(\hat{d}^*) - \nu_1(d^*)\right] + b_5\left[\hat{\nu}_2(\hat{d}^*) - \nu_2(d^*)\right] + o_p(1/\sqrt{N}) + o_p(|\hat{d}^* - d^*|), \end{aligned}$$

where

$$\begin{aligned} b_1 &= \Phi^-(p)\frac{\tilde{d}^*_{N,\rho_0} - \mu_1(d^*)}{\sqrt{\mu_2(d^*) - \mu_1^2(d^*)}} - \rho_0\frac{\nu_2(d^*) - 2\nu_1^2(d^*)}{\sqrt{\nu_2(d^*) - \nu_1^2(d^*)}}, \\ b_2 &= \Phi^-(p)\left[-\frac{\bar{F}_{X_1}(d^*)}{\sqrt{\mu_2(d^*) - \mu_1^2(d^*)}} + \frac{\mu_1(d^*)\bar{F}_{X_1}(d^*)[\tilde{d}^*_{N,\rho_0} - \mu_1(d^*)]}{(\mu_2(d^*) - \mu_1^2(d^*))^{3/2}}\right], \\ b_3 &= -\Phi^-(p)\frac{\bar{F}_{X_1}(d^*)[\tilde{d}^*_{N,\rho_0} - \mu_1(d^*)]}{2(\mu_2(d^*) - \mu_1^2(d^*))^{3/2}}, \\ b_4 &= -2\rho_0\nu_1(d^*)\left[\frac{1 - 2\bar{F}_{X_1}(d^*)}{\sqrt{\nu_2(d^*) - \nu_1^2(d^*)}} + \frac{\nu_2(d^*)\bar{F}_{X_1}(d^*) + \nu_1^2(d^*)[1 - 2\bar{F}_{X_1}(d^*)]}{2(\nu_2(d^*) - \nu_1^2(d^*))^{3/2}}\right], \\ b_5 &= \rho_0\left[\frac{\bar{F}_{X_1}(d^*)}{\sqrt{\nu_2(d^*) - \nu_1^2(d^*)}} - \frac{\nu_2(d^*)\bar{F}_{X_1}(d^*) + \nu_1^2(d^*)[1 - 2\bar{F}_{X_1}(d^*)]}{2(\nu_2(d^*) - \nu_1^2(d^*))^{3/2}}\right]. \end{aligned}$$

We also have

$$\begin{pmatrix} \hat{\bar{F}}_{X_1}(\hat{d}^*) - \bar{F}_{X_1}(d^*) \\ \hat{\mu}_1(\hat{d}^*) - \mu_1(d^*) \\ \hat{\mu}_2(\hat{d}^*) - \mu_2(d^*) \\ \hat{\nu}_1(\hat{d}^*) - \nu_1(d^*) \\ \hat{\nu}_2(\hat{d}^*) - \nu_2(d^*) \end{pmatrix} = \begin{pmatrix} \hat{\bar{F}}_{X_1}(d^*) - \bar{F}_{X_1}(d^*) \\ \hat{\mu}_1(d^*) - \mu_1(d^*) \\ \hat{\mu}_2(d^*) - \mu_2(d^*) \\ \hat{\nu}_1(d^*) - \nu_1(d^*) \\ \hat{\nu}_2(d^*) - \nu_2(d^*) \end{pmatrix} + \begin{pmatrix} -f_{X_1}(d^*) \\ \mu_1'(d^*) \\ \mu_2'(d^*) \\ \nu_1'(d^*) \\ \nu_2'(d^*) \end{pmatrix}\left(\hat{d}^* - d^*\right) + o_p(|\hat{d}^* - d^*|).$$

Multiply $\boldsymbol{b}$ to both sides of the above equation and use (30), we have

$$
\begin{aligned}
-\frac{\Phi^{-}(p)\bar{F}_{X_1}(d^*)}{\sqrt{\mu_2(d^*)-\mu_1^2(d^*)}}(\hat{d}^*-d^*) \quad &= \boldsymbol{b}\begin{pmatrix} \hat{\bar{F}}_{X_1}(d^*)-\bar{F}_{X_1}(d^*) \\ \hat{\mu}_1(d^*)-\mu_1(d^*) \\ \hat{\mu}_2(d^*)-\mu_2(d^*) \\ \hat{\nu}_1(d^*)-\nu_1(d^*) \\ \hat{\nu}_2(d^*)-\nu_2(d^*) \end{pmatrix} + \boldsymbol{b}\begin{pmatrix} -f_{X_1}(d^*) \\ \mu_1'(d^*) \\ \mu_2'(d^*) \\ \nu_1'(d^*) \\ \nu_2'(d^*) \end{pmatrix}\left(\hat{d}^*-d^*\right) \\
&\quad +o_p(1/\sqrt{N})+o_p(|\hat{d}^*-d^*|).
\end{aligned} \tag{31}
$$

Similar to the proof of Theorem 2, the terms $o_p(1/\sqrt{N})$ and $o_p(|\hat{d}^*-d^*|)$ in (31) can be neglected because the first term on the right-hand side of (31) has an order of $O_p(1/\sqrt{N})$ given by (29). Hence, it follows from (29) that

$$
\sqrt{N}(\hat{d}^*-d^*) = -\sqrt{N}b_0^{-1}\boldsymbol{b}\begin{pmatrix} \hat{\bar{F}}_{X_1}(d^*)-\bar{F}_{X_1}(d^*) \\ \hat{\mu}_1(d^*)-\mu_1(d^*) \\ \hat{\mu}_2(d^*)-\mu_2(d^*) \\ \hat{\nu}_1(d^*)-\nu_1(d^*) \\ \hat{\nu}_2(d^*)-\nu_2(d^*) \end{pmatrix} + o_p(1) \xrightarrow{\mathrm{d}} N(\mathbf{0}, b_0^{-2}\boldsymbol{b}\tilde{\Sigma}_0\boldsymbol{b}^\tau), \tag{32}
$$

where $\boldsymbol{b}=(b_1,b_2,b_3,b_4,b_5)$ and

$$
\begin{aligned}
b_0 = \Phi^{-}(p)\bar{F}_{X_1}(d^*)/\sqrt{\mu_2(d^*)-\mu_1^2(d^*)} - b_1 f_{X_1}(d^*) + b_2\mu_1'(d^*) + b_3\mu_2'(d^*) \\
+b_4\nu_1'(d^*) + b_5\nu_2'(d^*).
\end{aligned}
$$

Hence, the theorem follows as $\hat{b}_0$, $\hat{\boldsymbol{b}}$ and $\hat{\tilde{\Sigma}}_0$ are consistent estimators of b_0, $\boldsymbol{b}$ and $\tilde{\Sigma}_0$, respectively. The proof is now complete. □

Remark 2. *While Theorem 4 above only applies specifically to the smallest solution, as it is guaranteed to lead to a local minimum (Theorem 3), it can be extended to accommodate other possible solutions in case of non-uniqueness. For instance, if we define both $\tilde{d}^*_{N,\rho_0}$ and $\hat{\tilde{d}}^*_{N,\rho_0}$ as the largest solutions, Theorem 4 would still hold. The key requirement is that both $\tilde{d}^*_{N,\rho_0}$ and $\hat{\tilde{d}}^*_{N,\rho_0}$ must be selected using the same criterion. If, for example, $\tilde{d}^*_{N,\rho_0}$ is chosen as the smallest solution while $\hat{\tilde{d}}^*_{N,\rho_0}$ is the largest, the resulting asymptotic normality may no longer apply.*

2.4 Sharpe Ratio Premium Principle

We now recast the results in Section 2.3 by assuming the Sharpe Ratio premium principle

$$\mathbb{E}\left(\sum_{i=1}^{N}(X_i-d)_+\right)+\rho_0\frac{\mathbb{E}(\sum_{i=1}^{N}(X_i-d)_+)}{\mathrm{SD}(\sum_{i=1}^{N}(X_i-d)_+)}=N\nu_1(d)+\rho_0\sqrt{N}\frac{\nu_1(d)}{\sqrt{\nu_2(d)-\nu_1^2(d)}}, \tag{33}$$

leading to the total cost for the insurer as

$$\tilde{T}(d)=\sum_{i=1}^{N}(X_i\wedge d)+N\nu_1(d)+\rho_0\sqrt{N}\frac{\nu_1(d)}{\sqrt{\nu_2(d)-\nu_1^2(d)}}.$$

In this case, the loading factor becomes

$$\rho=\frac{\rho_0}{\mathrm{SD}(\sum_{i=1}^{N}(X_i-d)_+)}=\frac{\rho_0}{\sqrt{N}\sqrt{\nu_2(d)-\nu_1^2(d)}},$$

which is decreasing in N and satisfies (11). As before, the optimal retention minimizes

$$\begin{aligned}\mathrm{VaR}_p(\tilde{T}(d)) &= N\mu_1(d)+\sqrt{N}\sqrt{\mu_2(d)-\mu_1^2(d)}\Phi^-(p)+o(\sqrt{N})\\ &\quad +N\nu_1(d)+\rho_0\sqrt{N}\tfrac{\nu_1(d)}{\sqrt{\nu_2(d)-\nu_1^2(d)}},\end{aligned}$$

and we seek for d that minimizes (34) below by ignoring the $o(\sqrt{N})$ terms:

$$\bar{G}(d)=N\mathbb{E}(X_1)+\sqrt{N}\left\{\Phi^-(p)\sqrt{\mu_2(d)-\mu_1^2(d)}+\rho_0\frac{\nu_1(d)}{\sqrt{\nu_2(d)-\nu_1^2(d)}}\right\}. \tag{34}$$

The existence of the approximately optimal retention is shown in Theorem 5 below.

Theorem 5. *Assume $F_{X_1}(x)$ has the support $[0,\infty)$ (i.e., $F_{X_1}(0)>0$) or $(0,\infty)$, is continuous on $(0,\infty)$, and*

$$\lim_{t\to\infty}\frac{\bar{F}_{X_1}(tx)}{\bar{F}_{X_1}(t)}=x^{-\alpha}\ \textit{for all } x>0 \textit{ and some } \alpha\in(2,4). \tag{35}$$

If $F_{X_1}(0)>0$, we further assume that

$$\Phi^-(p)\sqrt{\bar{F}_{X_1}(0)F_{X_1}(0)}<\rho_0\frac{\bar{F}_{X_1}(0)}{\sqrt{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2}}-\rho_0\frac{F_{X_1}(0)\{\mathbb{E}(X_1)\}^2}{\{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2\}^{3/2}}. \tag{36}$$

Then, there exists at least one solution of $\bar{G}'(d)=0$, and an approximately optimal retention $d^*\in(0,\infty)$ is its smallest solution, which is a local minimum of $\bar{G}(d)$.

Proof. Because

$$\frac{\bar{G}'(d)}{\sqrt{N}} = \Phi^-(p)\frac{\bar{F}_{X_1}(d)\{d-\mu_1(d)\}}{\sqrt{\mu_2(d)-\mu_1^2(d)}} - \rho_0\frac{\bar{F}_{X_1}(d)}{\sqrt{\nu_2(d)-\nu_1^2(d)}} + \rho_0\frac{F_{X_1}(d)\nu_1^2(d)}{\{\nu_2(d)-\nu_1^2(d)\}^{3/2}},$$

it follows from (20) and (36) that

$$\begin{aligned}\lim_{d\to 0}\frac{\tilde{G}'(d)}{\sqrt{N}} &= \Phi^-(p)\sqrt{\bar{F}_{X_1}(0)F_{X_1}(0)} - \rho_0\frac{\bar{F}_{X_1}(0)}{\sqrt{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2}} \\ &\quad + \rho_0\frac{F_{X_1}(0)\{\mathbb{E}(X_1)\}^2}{\{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2\}^{3/2}} \\ &< 0.\end{aligned} \tag{37}$$

By (35) with $\alpha < 4$ and (24),

$$\lim_{d\to\infty}\frac{\bar{G}'(d)}{\sqrt{N}\bar{F}_{X_1}(d)d} = \frac{\Phi^-(p)}{\sqrt{\mathbb{E}(X_1^2)-\{\mathbb{E}(X_1)\}^2}} > 0. \tag{38}$$

Hence, it follows from (37) and (38) that there exists at least one solution of $\bar{G}'(d) = 0$ for $d \in (0,\infty)$, and let d^* be the smallest solution. Then, there exists $d_1 > d^*$ such that $\bar{G}'(d) < 0$ for $d \in (0, d^*)$, $\bar{G}'(d) \geq 0$ for $d \in (d^*, d_1)$, and $\bar{G}'(d_1) > 0$, implying that d^* is a local minimum of $\bar{G}(d)$ for $d \in (0,\infty)$. The proof is now complete. □

Again, d^* in Theorem 3 is independent of N, verifying the conditions of Proposition 2. To estimate the optimal retention nonparametrically, we minimize the following function for d:

$$\hat{\bar{G}}(d) = \Phi^-(p)\sqrt{\hat{\mu}_2(d)-\hat{\mu}_1^2(d)} + \rho_0\frac{\hat{\nu}_1(d)}{\sqrt{\hat{\nu}_2(d)-\hat{\nu}_1^2(d)}}. \tag{39}$$

Denote $\bar{d}^*_{N,\rho_0}$ and $\hat{\bar{d}}^*_{N,\rho_0}$ as the smallest solution of $\bar{G}'(d) = 0$ and $\hat{\bar{G}}'(d) = 0$ with $\bar{G}(d)$ and $\hat{\bar{G}}(d)$ given in (34) and (39), respectively. The asymptotic properties of the approximately optimal retention are provided in Theorem 6.

Theorem 6. *Under conditions of Theorem 5 and (33), and that X_1 has a density function f_{X_1}, we have*

$$\frac{\sqrt{N}\{\hat{\bar{d}}^*_{N,\rho_0} - \bar{d}^*_{N,\rho_0}\}}{\hat{a}_0^{-1}\sqrt{\hat{\boldsymbol{a}}\hat{\bar{\Sigma}}_0\hat{\boldsymbol{a}}^\tau}} \xrightarrow{d} N(0,1),$$

with $\hat{\boldsymbol{a}} := (\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5)$, where $\hat{a}_2$, $\hat{a}_3$ and $\hat{\bar{\Sigma}}_0$ are identical to $\hat{b}_2$, $\hat{b}_3$ and $\hat{\bar{\Sigma}}_0$, respectively, in

Theorem 4 though $\hat{\bar{d}}^*_{N,\rho_0}$ *is replaced by* $\hat{\bar{d}}^*_{N,\rho_0}$, *and*

$$\begin{aligned}
\hat{a}_0 &= \frac{\Phi^-(p)\hat{\bar{F}}_{X_1}(\hat{\bar{d}}^*_{N,\rho_0})}{\sqrt{\hat{\mu}_2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\mu}_1^2(\hat{\bar{d}}^*_{N,\rho_0})}} - \hat{a}_1\hat{f}_{X_1}(\hat{\bar{d}}^*_{N,\rho_0}) + \hat{a}_2\hat{\mu}_1^*(\hat{\bar{d}}^*_{N,\rho_0}) + \hat{a}_3\hat{\mu}_2^*(\hat{\bar{d}}^*_{N,\rho_0}) \\
&\qquad + \hat{a}_4\hat{\nu}_1^*(\hat{\bar{d}}^*_{N,\rho_0}) + \hat{a}_5\hat{\nu}_2^*(\hat{\bar{d}}^*_{N,\rho_0}), \\
\hat{a}_1 &= \Phi^-(p)\frac{\hat{\bar{d}}^*_{N,\rho_0} - \hat{\mu}_1(\hat{\bar{d}}^*_{N,\rho_0})}{\sqrt{\hat{\mu}_2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\mu}_1^2(\hat{\bar{d}}^*_{N,\rho_0})}} - \rho_0\frac{\hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0})}{(\hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\bar{d}}^*_{N,\rho_0}))^{3/2}}, \\
\hat{a}_4 &= \rho_0\hat{\nu}_1(\hat{\bar{d}}^*_{N,\rho_0})\left[\frac{2}{(\hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\bar{d}}^*_{N,\rho_0}))^{3/2}} + \frac{3[\hat{\nu}_1^2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0})\hat{\bar{F}}_{X_1}(\hat{\bar{d}}^*_{N,\rho_0})]}{(\hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\bar{d}}^*_{N,\rho_0}))^{5/2}}\right], \\
\hat{a}_5 &= -\rho_0\left[\frac{\hat{\bar{F}}_{X_1}(\hat{\bar{d}}^*_{N,\rho_0})}{(\hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\bar{d}}^*_{N,\rho_0}))^{3/2}} + \frac{3[\hat{\nu}_1^2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0})\hat{\bar{F}}_{X_1}(\hat{\bar{d}}^*_{N,\rho_0})]}{2(\hat{\nu}_2(\hat{\bar{d}}^*_{N,\rho_0}) - \hat{\nu}_1^2(\hat{\bar{d}}^*_{N,\rho_0}))^{5/2}}\right].
\end{aligned}$$

Proof. Since the proof is similar to that for Theorem 4, we omit the details. □

Remark 3. *Another choice of the reinsurance premium beyond the Sharpe Ratio may also seem natural. That is, we could use the Standard Deviation to determine the reinsurance premium as follows:*

$$\mathbb{E}\left(\sum_{i=1}^{N}(X_i - d)_+\right) + \rho_0 SD\left(\sum_{i=1}^{N}(X_i - d)_+\right) = N\nu_1(d) + \rho_0\sqrt{N}\sqrt{\nu_2(d) - \nu_1^2(d)}.$$

However, the resulting objective function has a positive derivative at $d = 0$, *which often leads to a trivial approximately optimal retention being either zero or infinity. Therefore, we do not discuss this setting in the paper as the optimal retention is trivial.*

3 Generalization to distortion risk measures

We show now that our results in Section 2 can be naturally extended to optimal reinsurance problems under general distortion risk measures. A large class of quantile-based risk measures is the distorted class, for which Definition 1 is needed.

Definition 1. *A distortion function is a non-decreasing function* $h : [0,1] \to [0,1]$ *such that* $h(0) = h(0+) = 0$ *and* $h(1) = h(1-) = 1$.

Yarri's dual theory of choice under risk – e.g., see Yaari (1987) – postulates that the risk preferences of a nonrisk neutral decision maker could be modeled by an expectation concerning

a reweighed or distorted probability measure, where the distortion function is as in Definition 1. We are ready to define a *distortion risk measure*, which is given as Definition 2.

Definition 2. *Let Y be a non-negative random variable and h be a distortion function. The Choquet integral*

$$\varphi_h(Y) := \int_0^\infty h \circ \bar{F}_Y(x)\, dx = \int_0^\infty \big(1 - \tilde{h} \circ F_Y(x)\big)\, dx \tag{40}$$

is called a distortion risk measure, where $\tilde{h}(\cdot) = 1 - h(1-\cdot)$ on $[0,1]$.

Note that $\tilde{h}$ is a distortion function since h is a distortion function. Further, $\rho_h(Y)$ is an expectation with respect to a reweighed probability measure, namely $\tilde{h} \circ F_Y$; that is, $\rho_h(Y) = \int_0^\infty x\ d\tilde{h} \circ F_Y(x)$ is a Lebesgue-Stieljes integral. It is not difficult to see that VaR and ES are distortion risk measures with distortion functions $h_{\mathrm{VaR}_p}(s) := I_{\{p \le s \le 1\}}$ and $h_{\mathrm{ES}_p}(s) := \min\left(\frac{s}{1-p}, 1\right)$, respectively for all $s \in [0,1]$. Other examples are i) *Dual-power* with $h_{DP}(s) := 1 - (1-s)^\beta, \beta \ge 1$, ii) *Gini* with $h_G(s) := (1+\beta)s - \beta s^2, 0 \le \beta \le 1$, iii) *Proportional hazard transform (PHT)* with $h_{PHT}(s) := s^{1-\beta}, 0 \le \beta < 1$, and iv) *Wang transform* with $h_{WT}(s) := \Phi\big(\Phi^-(s) + \beta\big), \beta \ge 0$.

The results in Section 2 could be generalized to the class of distortion risk measures through Lemma 1 after verifying some robustness conditions. Note that the case in which the risk preferences are ordered by ES is a special case of our main results in this section.

Lemma 1. *Let Z follow the standard normal distribution* $\mathrm{N}(0,1)$*. We have*

$$\varphi_h\big(T(d,N,\rho)\big) = N(\mathbb{E}(X_1) + \rho\nu_1(d)) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\varphi_h(Z) + o(\sqrt{N}).$$

Proof. It follows from the *Central Limit Theorem* that

$$F^-_{T(d,N,\rho)}(p) = N(\mathbb{E}(X_1) + \rho\nu_1(d)) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\Phi^-(\alpha) + o(\sqrt{N}), \quad p \in (0,1).$$

As $0 \le \sum_{i=1}^n (X_i \wedge d) \le Nd$ and Z is integrable, we have $\{Z\} \cup \{T(d,N,\rho) : N = 1, 2 \dots\}$ is h-uniformly integrable.* By Theorem 4 of Wang et al. (2020), translation invariance and

*For a distortion function h, a set of random variables $\mathcal{X}$ is called *h-uniformly integrable* if

$$\lim_{k \downarrow 0} \sup_{S \in \mathcal{X}} \int_0^k |F_S^-(1-t)|\, dh(t) = 0 \quad \text{and} \quad \lim_{k \uparrow 1} \sup_{S \in \mathcal{X}} \int_k^1 |F_S^-(1-t)|\, dh(t) = 0.$$

homogeneity of φ_h of order 1,

$$\varphi_h\big(T(d,N,\rho)\big) = N\big(\mathbb{E}(X_1)+\rho\nu_1(d)\big) + \sqrt{N}\sqrt{\mu_2(d)-\mu_1^2(d)}\varphi_h(Z) + o(\sqrt{N}). \qquad \square$$

Therefore, instead of minimizing $\varphi_h\big(T(d,N,\rho)\big)$ to define the optimal retention d, we seek an approximately optimal retention d by minimizing

$$G_\varphi(d) := N\mathbb{E}(X_1) + N\rho\nu_1(d) + \sqrt{N}\sqrt{\mu_2(d)-\mu_1^2(d)}\varphi_h(Z). \tag{41}$$

Similar to Theorem 1, we can show the unique solution after replacing $\Phi^-(p)$ in Theorem 1 with $\varphi_h(Z)$. Further, we can estimate this unique approximately optimal retention and derive its asymptotic normal limit. Specifically, we have a generalized result below following a proof similar to that of Theorem 1, which is given as Theorem 7 that needs the following notation.

$$H_\varphi(d) := \{d-\mu_1(d)\}^2 - \left(\frac{\sqrt{N}\rho}{\varphi_h(Z)}\right)^2 \{\mu_2(d)-\mu_1^2(d)\} = 0.$$

Theorem 7. *Assume $\mathbb{E}(X_1) < \infty$, $F_{X_1}(x)$ has the support $[0,\infty)$ (i.e., $F_{X_1}(0) > 0$) or $(0,\infty)$, and is continuous on $(0,\infty)$. When the support is $[0,\infty)$, we further assume $F_{X_1}(0) < \frac{N\rho^2}{N\rho^2+(\varphi_h(Z))^2}$, which is always true when N is large enough. Then, there exists a unique approximately optimal retention $d^*_{\varphi,N} \in (0,\infty)$ such that*

$$d^*_{\varphi,N} = \operatorname*{argmin}_{d>0} G_\varphi(d) \quad \textit{and} \quad H_\varphi(d^*_{\varphi,N}) = 0.$$

In light of Lemma 1, we can also extend Theorems 2-4 in a similar sense to Theorem 1 by changing $\Phi^-(p)$ to $\varphi_h(Z)$ for which no other adjustments are needed.

4 Simulation studies

We conduct two simulation studies in this section. Section 4.1 assesses the validity of substituting the actual VaR of total cost $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ as specified in (3) with the normal-approximated VaR $G_{N,\rho}(d)$ outlined in (5); this assessment is conducted under the four loading factor rules delineated in Sections 2.1–2.4. Section 4.3 empirically examines the statistical properties of the optimal retention estimators introduced in Theorems 2, 4, and 6.

4.1 Examining validity of approximately optimal retention

We generate samples of X_i from a Pareto (type-II) distribution with a probability density function given by $f_X(x) = (\alpha/\lambda)(1+x/\lambda)^{-(\alpha+1)}$, where the shape parameter $\alpha = 9$ and the scale parameter $\lambda = 8$, such that $\mathbb{E}[X_i] = \alpha/(\lambda-1) = 1$. We set the risk level at $p = 0.75$ and consider sample sizes of $N = 10$, 25, and 100. In this analysis, we examine all four loading factor rules by setting $\rho = 0.3$ for the constant loading factor, choosing $\delta = 0.5$ for the decreasing loading factor, and $\rho_0 = 0.5$ for the remaining two rules.

We compute the true VaR of the total cost $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ using (3), where $\mathbb{E}\{(X_1-d)_+\}$ and $\mathrm{VaR}_p(\sum_{i=1}^N (X_i \wedge d))$ are approximated through $B = 50000$ simulated samples. For example, we have $\mathbb{E}\{\sum_{i=1}^N (X_i \wedge d)\} \approx (1/B)\sum_{j=1}^B \sum_{i=1}^N (X_{ij} \wedge d)$ with X_{ij} iid sampled from a Pareto distribution for $i = 1,\ldots,N$ and $j = 1,\ldots,B$. We also compute the normal-approximated VaR of total cost $G_{N,\rho}(d)$ in (5) by numerical integration.

Figure 1 plot $G_{N,\rho}(d)$ (red curves) and $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ (black curves) as functions of d for $N = 10$ and 100 under the four loading factor rules. We note that $G_{N,\rho}(d)$ approximates closely to $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ under all loading factor rules especially for $N = 100$. For all loading factor rules except for the constant loading factor, we observe an optimal retention $d^* \in [0,1]$ that minimizes the VaR of the total cost. Conversely, for the constant loading factor, both $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ and $G_{N,\rho}(d)$ become flat as d increases, especially when N is large. Hence, it is difficult to identify d^* by solely observing the plots.

To further compare the actual and approximately optimal retention levels d^*, we plot $\partial \mathrm{VaR}_p\big(T(d,N,\rho)\big)/\partial d$ and $G'_{N,\rho}(d)$ (computed using numerical method) as functions of d in Figure 2, representing the first derivative of the VaR of the total cost with respect to d, for $N = 10, 100$ under the four loading factor rules. The actual and approximately optimal retentions are determined by solving $\partial \mathrm{VaR}_p\big(T(d,N,\rho)\big)/\partial d = 0$ and $G'_{N,\rho}(d) = 0$, respectively. In general, $G'_{N,\rho}(d)$ closely approximates $\partial \mathrm{VaR}_p\big(T(d,N,\rho)\big)/\partial d$ when the sample size is sufficiently large. However, for small sample size, $\partial \mathrm{VaR}_p\big(T(d,N,\rho)\big)/\partial d$ tends to be more volatile and deviates significantly from $G'_{N,\rho}(d)$ when $d < 0.5$. For example, in the case where $N = 10$ under the Sharpe ratio principle, the black curves frequently cross the horizontal grey line, indicating multiple local minima for $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$. On the other hand, $G'_{N,\rho}(d)$ remains a smooth function of d in all scenarios, with each red curve crossing the horizontal gray line only once, indicating a unique solution in this Pareto simulation setting. This highlights the computational advantage of approximating the optimal retention rather than computing an exact one. For the

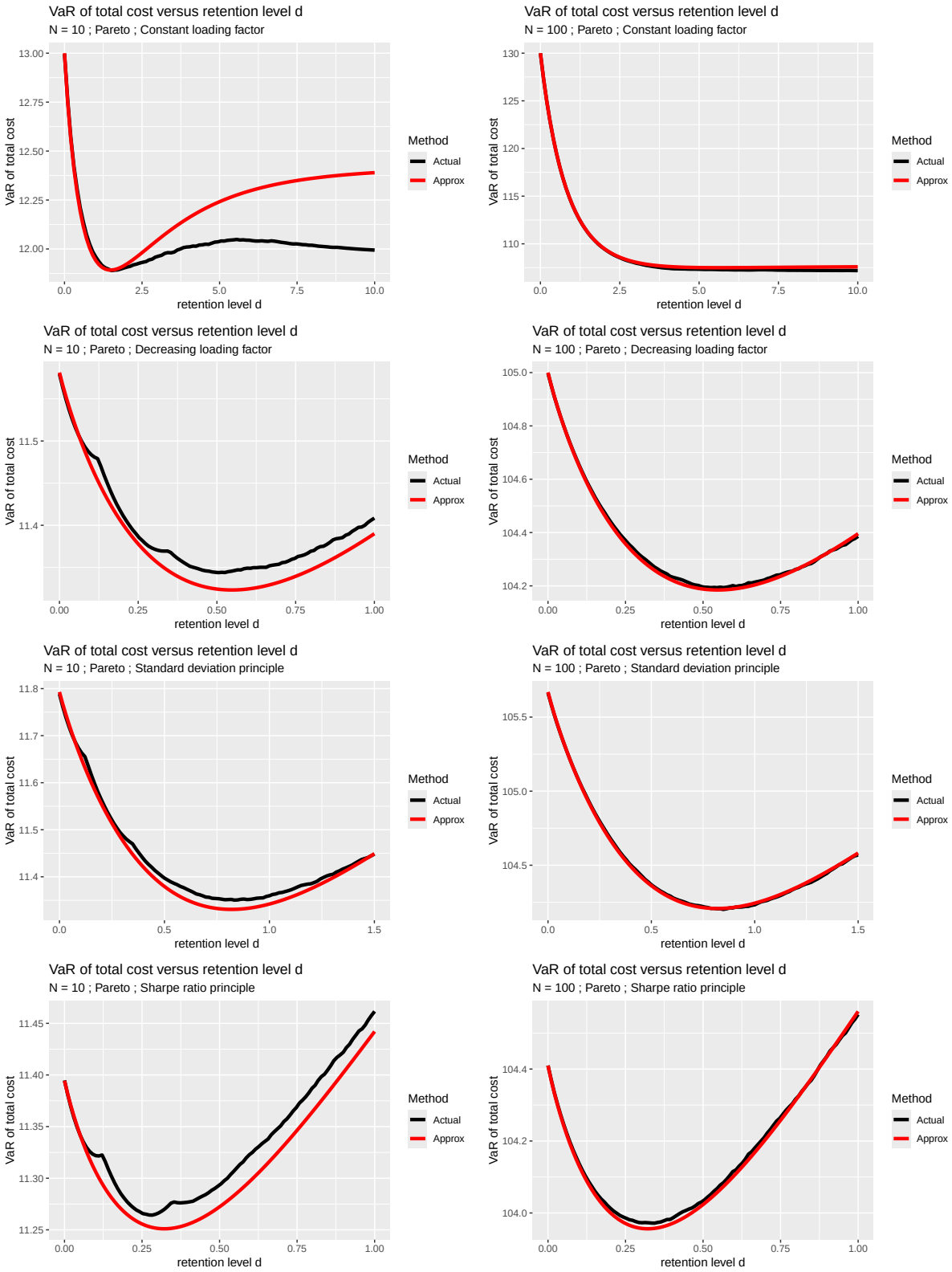

Figure 1: $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ (black curves) and $G_{N,\rho}(d)$ (red curves) versus d for $N = 10, 100$ under various loading factor rules when X_i follows Pareto distribution.

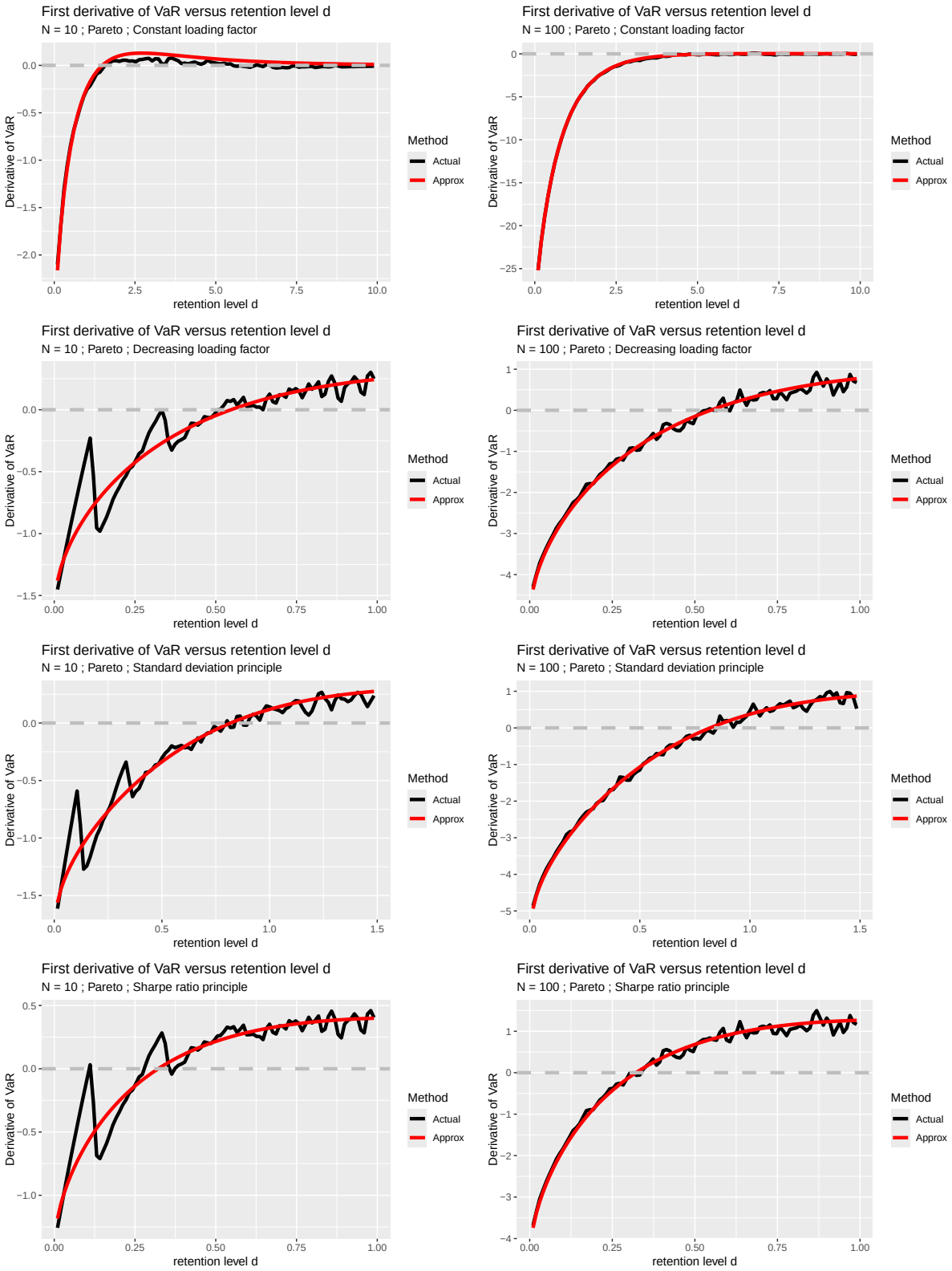

Figure 2: $\partial \mathrm{VaR}_p\big(T(d,N,\rho)\big)/\partial d$ (black curves) and $G'_{N,\rho}(d)$ (red curves) versus d for $N = 10, 100$ under various loading factor rules when X_i follows Pareto distribution.

constant loading factor rule, we observe that both $\partial \mathrm{VaR}_p\big(T(d,N,\rho)\big)/\partial d$ and $G'_{N,\rho}(d)$ approach zero as d increases, especially when $N = 100$, consistent with the observations in Figure 1 where the optimal retention is visually difficult to distinguish.

We then numerically calculate the actual and approximately optimal retentions d^* and $d^*_{N,\rho}$, and the results are displayed in Table 1. For all loading factor rules other than the constant loading factor, the actual optimal retention d^*, which minimizes the true VaR of total cost $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$, yields similar values as the approximately optimal retention $d^*_{N,\rho}$, which minimizes the approximated VaR of total cost $G_{N,\rho}(d)$. Also, the relative difference $|(d^*_{N,\rho}-d^*)/d^*_{N,\rho}|$ generally reduces as N increases. Furthermore, d^* does not change substantially as N changes. As a result, the approximate optimal retention approach for VaR works well under these three loading factor rules.

Table 1: Actual optimal retention d^*, approximately optimal retention $d^*_{N,\rho}$, and the relative difference between d^* and $d^*_{N,\rho}$ (in %) across various loading factor rules and N.

Loading factor rule	N	Actual	Approx.	Diff. (%)
Constant loading factor	10	1.8549	1.4856	-19.91
Constant loading factor	25	3.4442	2.6838	-22.08
Constant loading factor	100	7.1241	5.6581	-20.58
Decreasing loading factor	10	0.5034	0.5472	8.70
Decreasing loading factor	25	0.5835	0.5472	-6.21
Decreasing loading factor	100	0.5472	0.5472	0.02
Standard deviation principle	10	0.7847	0.8189	4.36
Standard deviation principle	25	0.8187	0.8189	0.03
Standard deviation principle	100	0.8499	0.8189	-3.64
Sharpe ratio principle	10	0.2797	0.3218	15.06
Sharpe ratio principle	25	0.3149	0.3218	2.18
Sharpe ratio principle	100	0.3203	0.3218	0.45

For the constant loading factor, the optimal retention numerically exists even if the curves in the top panels of Figure 1 are flat and the curves in the top panels of Figure 2 approach to zero. However, there are noticeable discrepancies between d^* and $d^*_{N,\rho}$ for a given N, especially when N is large. Under this loading factor rule, the proposed approximation method with an order of $o(\sqrt{N})$, as given in (4), lacks sufficient accuracy in determining the optimal retention. To address this, one can employ the Edgeworth expansion to enhance the precision of the approximation. Section 2 in the supplementary materials provides detailed mathematical formulations based on the Edgeworth approximation method, along with the corresponding numerical results for optimal retentions. Overall, by using higher-order approximation methods, the discrepancies between the actual and approximated optimal retentions can be significantly reduced. This

suggests that when using a constant loading factor, a higher-order Edgeworth expansion is necessary to achieve an accurate approximation for optimal retention, especially as N increases.

4.2 Unraveling non-uniqueness of optimal retentions

In the previous numerical study (Section 4.1), we observed a unique approximately optimal retention $d^*_{N,\rho}$ under all loading factor rules when the X_i samples were generated from a Pareto distribution. This uniqueness of the approximately optimal retention is commonly observed across most generating distributions for X_i. However, Theorems 1, 3, and 5 guarantee uniqueness only under the constant or decreasing loading factor rules, but not under the standard deviation or Sharpe ratio premium principles. In this study, we demonstrate the potential for non-unique solutions under the standard deviation and Sharpe ratio premium principles by carefully selecting the generating distribution for X_i.

The X_i samples are now generated from a three-component finite mixture of log-normal Pareto distributions, with the probability density function $f_X(x) = 0.4 \cdot \tilde{\phi}(x; \mu = -1.6, \sigma = 0.1) + 0.4 \cdot \tilde{\phi}(x; \mu = 0, \sigma = 0.01) + 0.2 \cdot f_X(x; \alpha = 51, \lambda = 150)$, where $\tilde{\phi}(x; \mu, \sigma)$ represents the log-normal density with mean and standard deviation parameters μ and σ, while $f_X(x; \alpha, \lambda)$ is a Pareto density with shape and scale parameters α and λ. We set $p = 0.75$, $N = 100$, and $\rho_0 = 0.34$ (for the standard deviation principle) or $\rho_0 = 1.07$ (for the Sharpe ratio principle). Figure 3 presents the actual (black) and approximated (red) VaRs plotted against d under both loading factor rules. The plots are zoomed in for clarity. It is evident that two local minima exist for both the actual and approximated VaRs, demonstrating the non-uniqueness of solutions in this specific setting.

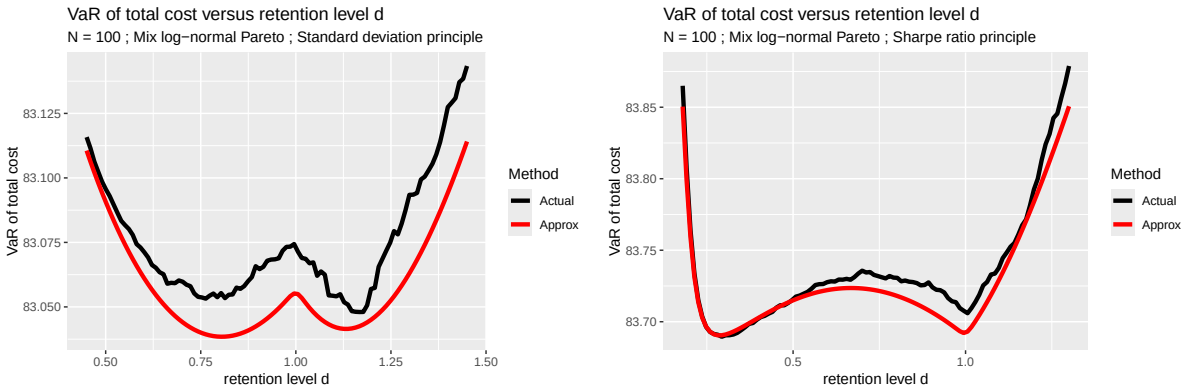

Figure 3: $\mathrm{VaR}_p\big(T(d, N, \rho)\big)$ (black curves) and $G_{N,\rho}(d)$ (red curves) versus d under various loading factor rules when X_i follows finite mixture distribution.

4.3 Verifying statistical properties of nonparametric approach

We adopt the same simulation setup as in Section 4.1 except assessing larger sample sizes of $N = 500$, 2000, and 10000. The computational complexity associated with such large sample sizes is given by the fact the direct estimation of the true VaR of the total cost $\mathrm{VaR}_p\big(T(d, N, \rho)\big)$ is unfeasible without resorting to the normal approximation technique. In each simulation run, we compute the nonparametric estimation of the approximately optimal retention, which is $\hat{d}^*_{N,\rho_N}$ under the decreasing loading factor, $\hat{\tilde{d}}_{N,\rho_N}$ under the standard deviation principle, or $\hat{\bar{d}}_{N,\rho_N}$ under the Sharpe ratio principle. To obtain an M-vector of estimated optimal retentions, which is $\{\hat{d}^{*(m)}_{N,\rho_N}\}_{m=1,\ldots,M}$ for each of the three loading factor rules, we repeat the simulation runs $M = 5000$ times. Additionally, through numerical integration, we compute the "true" approximately optimal retention, denoted as d^*_{N,ρ_N} under the decreasing loading factor, $\tilde{d}_{N,\rho_N}$ under the standard deviation principle, or $\bar{d}_{N,\rho_N}$ under the Sharpe ratio principle. We calculate the sample mean of $\{\hat{d}^{*(m)}_{N,\rho_N}\}_{m=1,\ldots,M}$ and compare it with the true approximately optimal retention to evaluate the bias of the proposed nonparametric estimation approach.

Table 2: *Columns 2–4*: True approximately optimal retention, the sample mean of the nonparametrically estimated approximately optimal retentions, and their relative difference; *Columns 5–7*: Theoretical and empirical standard error of the estimated optimal retention, and their relative difference.

	Mean optimal retention			Std. Error optimal retention		
	True	Estimated	Bias (%)	Theoretical	Empirical	Diff. (%)
Decreasing loading factor						
$N = 500$	0.5472	0.5478	0.10	0.0392	0.0392	-0.05
$N = 2000$	0.5472	0.5478	0.11	0.0196	0.0201	2.57
$N = 10000$	0.5472	0.5474	0.04	0.0088	0.0087	-0.24
Standard deviation principle						
$N = 500$	0.8189	0.8185	-0.05	0.1115	0.1199	7.54
$N = 2000$	0.8189	0.8187	-0.03	0.0577	0.0594	2.98
$N = 10000$	0.8189	0.8191	0.02	0.0263	0.0262	-0.22
Sharpe ratio principle						
$N = 500$	0.3218	0.3259	1.29	0.0442	0.0468	5.85
$N = 2000$	0.3218	0.3233	0.46	0.0229	0.0235	2.64
$N = 10000$	0.3218	0.3220	0.07	0.0104	0.0105	1.73

Utilizing the results from Theorems 2, 4 and 6, we compute the theoretical standard error of the optimal retention estimators, represented as, for example, $N^{-1/2}\hat{c}_0^{-1}\sqrt{(\hat{c}_1, \hat{c}_2)\hat{\Sigma}_0(\hat{c}_1, \hat{c}_2)^\tau}$ under the decreasing loading factor, and compare it with the sample standard deviation of $\{\hat{d}^{*(m)}_{N,\rho_N}\}_{m=1,\ldots,M}$ to assess the validity of the theoretical results. Note that for the computation of standard errors under Theorems 4 and 6, we utilize the kernel density estimator $\hat{f}_{X_1}(d)$ with

a Gaussian kernel function and a bandwidth of 0.1. While alternative kernel functions and bandwidths are possible, we have observed that they exert negligible influence on the computed standard errors. Hence, we do not delve into further details regarding these alternatives. Table 2 summarizes the findings across various N and loading factor rules. Our observations indicate minimal estimation biases of the nonparametrically estimated optimal retention in all scenarios, and in turn, we empirically confirm the consistency of the proposed nonparametric estimators. Moreover, the empirical standard deviations of our estimators closely align with the theoretical standard errors across all cases, which provides empirical validation to the asymptotic properties outlined in Theorems 2, 4 and 6. Additionally, as N becomes large, we note a decline in the relative bias of the estimated optimal retention, as well as the relative difference between the theoretical and empirical standard errors, consistent with the asymptotic theories.

5 Real Data analysis

We analyze the `frecomfire` dataset, which consists of 9,613 commercial fire losses located in France, spanning from 1982 to 1996. This dataset is publicly accessible via the `R` package `CASdatasets`. The left panel of Figure 4 displays the empirical density of claim severities, with each claim expressed in million euros (at the 2007 value). The distribution of claim sizes exhibits significant right-skewness, as evidenced by several extreme losses indicated by arrows. In the right panel of Figure 4, the Lorenz curve illustrates the cumulative share of claim amounts against the cumulative normalized rank of claims. A substantial deviation of the Lorenz curve from the equality line indicates considerable disparities between large and small claims. The pronounced gap between the Lorenz curve and the equality line reflects the wide dispersion of claim amounts. Notably, the median, mean, and maximum loss amounts are 0.7633, 1.9811, and 315.54, respectively, with the 20 largest loss comprising more than 10% of the total loss. The heavy-tailed nature of the claim distribution, coupled with several exceptionally large losses, highlights the importance for insurance companies to transfer individual losses, rather than aggregate liabilities, to reinsurers. This motivates the analysis of EoL reinsurance, rather than SL reinsurance, as explored, for instance, by Cai and Tan (2007).

Our primary objective is to investigate the variations in nonparametric estimates of the approximately optimal retention across various effective loading factors ρ and risk levels p under three loading factor rules: decreasing loading factor, standard deviation principle, and Sharpe

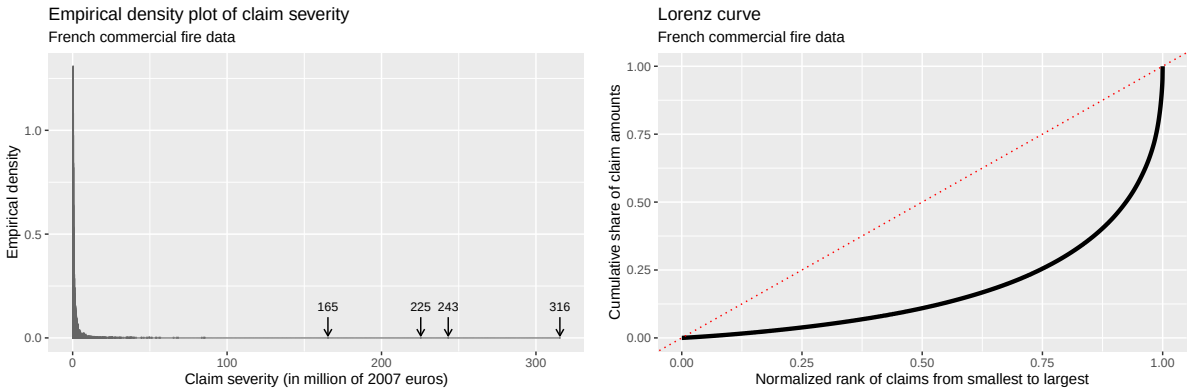

Figure 4: *Left panel*: Empirical density plot of claim severity; *Right panel*: Lorenz curve (thick solid curve) of claim severity with the equality line (dotted 45-degree line).

ratio principle. It is important to highlight that we assess the effective loading factors ρ by using expressions such as (18) under the standard deviation principle, rather than relying on the nominal loading factors like ρ_0 in (18) so that we ensure equitable comparisons among the three loading factor rules.

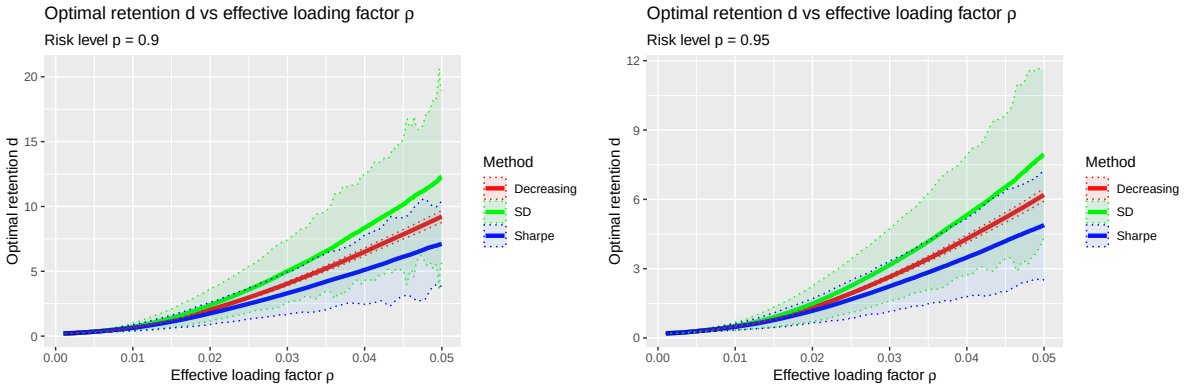

Figure 5: Optimal retention versus effective loading factor (solid curves) with fixed $p = 0.9$ (left panel) or $p = 0.95$ (right panel). The 95% confidence intervals are displayed as shaded areas.

Figure 5 displays the optimal retention as a function of the effective loading factor ρ, with fixed values of $p = 0.9$ (left panel) or $p = 0.95$ (right panel) under each of the three loading factor rules, accompanied by corresponding 95% confidence intervals determined based on Theorems 2, 4 and 6. Across all loading factor rules, it is evident that the optimal retention level increases with ρ for any fixed p. This observation is intuitive, as a higher ρ implies a greater cost for risk transfer, thereby incentivizing insurers to retain losses up to a higher level. Furthermore, it is observed that the standard deviation loading factor principle yields the highest optimal retention for any fixed p and ρ, followed by the decreasing loading factor and, finally, the Sharpe

ratio principle. This trend can be rationalized by considering that the standard deviation of the excess loss $(X_1 - d)_+$ decreases as d increases. Consequently, the effective loading factor under the standard deviation principle diminishes with increasing d, encouraging insurers to select a higher retention level to mitigate reinsurance costs. Additionally, the confidence bands under the standard deviation and Sharpe ratio principles are notably wider than those under the decreasing loading factor. This discrepancy arises because the loading factor ρ under either principle, which is contingent on the second moment of the excess loss, may be heavily influenced by extreme losses, leading to increased standard errors. Conversely, the normal-approximated VaR of the total cost relies solely on the excess loss up to its first moment under the decreasing loading factor, resulting in decreased sensitivity of the estimated optimal retention to extreme losses.

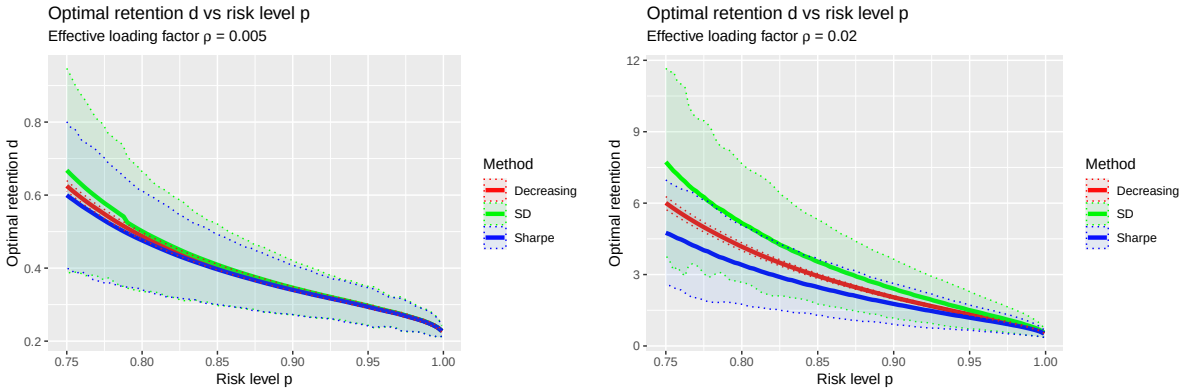

Figure 6: Optimal retention versus risk level p (solid curves) with fixed $\rho = 0.005$ (left panel) or $\rho = 0.02$ (right panel). The 95% confidence intervals are displayed as shaded areas.

Figure 6 illustrates the optimal retention versus the risk level p, with fixed values of $\rho = 0.005$ (left panel) or $\rho = 0.02$ (right panel) under the three loading factor rules, accompanied by 95% confidence intervals. Across all loading factor rules, it is observed that the optimal retention decreases as p increases. This outcome is logical, as a higher risk level p signifies insurers' greater aversion to risk, thereby reducing their inclination to retain extreme losses by opting for a smaller retention level. Notably, our proposed method addresses the counterintuitive finding of Cai and Tan (2007) that the optimal retention remains unchanged as p varies.

References

Arrow, K.J. (1963). Uncertainty and the welfare economics of medical care. *American Eco-*

nomic Review, 53(5), 941–973.

Asimit, A.V., Badescu, A.M. and Cheung, K.C. (2013). Optimal reinsurance in the presence of counterparty default risk. *Insurance: Mathematics and Economics, 53(3),690–697.*

Asimit, A.V., Badescu, A.M., Haberman, S. and Kim, E.-.S. (2016). Efficient risk allocation within a non-life insurance group under Solvency II Regime. *Insurance: Mathematics and Economics, 66, 69–76.*

Asimit, A.V., Badescu, A.M. and Verdonck, T. (2013). Optimal risk transfer under quantile-based risk measures. *I*nsurance: Mathematics and Economics, 53(1), 252–265.

Asimit, A.V., Bignozzi, V., Cheung, K.C., Hu, J. and Kim, E.-.S. (2017). Robust and Pareto optimality of insurance contracts. *European Journal of Operational Research, 262(2), 720–732.*

Asimit, A.V. and Boonen, T.J. (2018). Insurance with multiple insurers: A game-theoretic approach. *European Journal of Operational Research, 267(2), 778–790.*

Asimit, A.V., Chi, Y. and Hu, J. (2015). Optimal non-life reinsurance under Solvency II Regime. *Insurance: Mathematics and Economics, 65, 227–237.*

Asimit, A.V., Hu, J. and Xie, Y. (2019). Optimal robust insurance with a finite uncertainty set. *Insurance: Mathematics and Economics, 87, 67–81.*

Assa, H. (2015). On optimal reinsurance policy with distortion risk measures and premiums. *Insurance: Mathematics and Economics, 61, 70–75.*

Assa, H., Sharifi, H. and Lyons, A. (2021). An examination of the role of price insurance products in stimulating investment in agriculture supply chains for sustained productivity. *European Journal of Operational Research, 288(3), 918–934.*

Balbás, A., Balbás, B. and Heras, A. (2009). Optimal reinsurance with general risk measures. *Insurance: Mathematics and Economics 44, 374–384.*

Balbás, A., Balbás, B. and Heras, A. (2011). Stable solutions for optimal reinsurance problems involving risk measures. *European Journal of Operational Research, 214(3), 796–804.*

Bäuerle, N., and Glauner, A. (2018). Optimal risk allocation in reinsurance networks. *Insurance: Mathematics and Economics, 82, 37-47.*

Bernard, C. and Ludkovski, M. (2012). Impact of counterparty risk on the reinsurance market. *North American Actuarial Journal 16(1), 87–111.*

Bernard, C., and Tian, W. (2009). Optimal reinsurance arrangements under tail risk measures. *Journal of Risk and Insurance 76, 709–725.*

Boonen, T. J., and Ghossoub, M. (2023). Bowley vs. Pareto optima in reinsurance contracting. *European Journal of Operational Research, 307(1), 382-391.*

Boonen, T. J. and Jiang, W. (2022). A marginal indemnity function approach to optimal reinsurance under the Vajda condition. *European Journal of Operational Research, 303(2), 928–944.*

Boonen, T. J. and Jiang, W. (2024). Robust insurance design with distortion risk measures. *European Journal of Operational Research, Forthcoming.*

Cai, J., Lemieux, C., and Liu F. (2014) Optimal reinsurance with regulatory initial capital and default risk. *Insurance: Mathematics and Economics 57, 13–24.*

Cai, J., Li, J. Y. M. and Mao, T. (2023). Distributionally robust optimization under distorted expectations. *Operations Research, Forthcoming.*

Cai, J., Liu, F. and Yin, M. (2024). Worst-case risk measures of stop-loss and limited loss random variables under distribution uncertainty with applications to robust reinsurance. *European Journal of Operational Research, 318(1), 310–326.*

Cai, J. and Tan, K.S. (2007). Optimal retention for a stop-loss reinsurance under the VaR and CTE risk measures. *Astin Bulletin 37, 93–112.*

Cai, J., Tan, K. S., Weng, C. and Zhang, Y. (2008). Optimal reinsurance under VaR and CTE risk measures. *Insurance: Mathematics and Economics 43, 185–196.*

Cai, J., and Weng C. (2016). Optimal reinsurance with expectile. *Scandinavian Actuarial Journal (7), 624–645.*

Chen, Y. (2024). Optimal insurance with counterparty and additive background risk. *ASTIN Bulletin: The Journal of the IAA, 1-22.*

Chi, Y. and Tan, K. S. (2021). Optimal incentive-compatible insurance with background risk. *ASTIN Bulletin: The Journal of the IAA, 51(2), 661–688.*

Chi, Y. and Tan, K. S. (2013). Optimal reinsurance with general premium principles. *Insurance: Mathematics and Economics 52, 180–189.*

Denneberg, D. (1994a). Non-additive measure and integral. *Kluwer Academic Publishers, Dordrecht.*

Denneberg, D. (1994b). Conditioning (updating) non-additive measures. *Annals of Operations Research, 52, 21–42.*

Embrechts, P. and Hofert, M. (2013). A note on generalized inverses, *Mathematical Methods of Operations Research, 77, 423–432.*

Föllmer, H. and Schied, A. (2011). Stochastic Finance: An Introduction in Discrete Time, *Third ed., Walter de Gruyter.*

Gollier, C. (2014). Optimal insurance design of ambiguous risks. *Economic Theory 57, 555–576.*

Kaluszka, M. (2001). Optimal reinsurance under mean-variance premium principles. *Insurance: Mathematics and Economics 28, 61–67.*

Klages-Mundt, A. and Minca, A. (2020). Cascading losses in reinsurance networks. *Management Science, 66(9), 4246–4268.*

Pesenti, S., Wang, Q. and Wang, R. (2024). Optimizing distortion riskmetrics with distributional uncertainty. *Mathematical Programming, Forthcoming.*

Raviv, A. (1979). The design of an optimal insurance policy. *American Economic Review, 69(1), 84–96.*

Rüschendorf, L. (2013). *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios. Springer.*

Tan, K.S., Wei, P., Wei, W. and Zhuang, S.C. (2020). Optimal dynamic reinsurance policies under a generalized Denneberg's absolute deviation principle. *European Journal of Operational Research, 282(1), 345–362.*

Van der Vaart, A.W. (2000). Asymptotic statistics. Vol. 3. *Cambridge university press*

Wang, R., Wei, Y. and Willmot, G. E. (2020). Characterization, robustness and aggregation of signed Choquet integrals. *Mathematics of Operations Research, 45(3), 993–1015.*

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society, 95–115.*

Supplementary materials for "A Revisit of the Optimal Excess-of-Loss Contract"

Ernest Aboagye, Vali Asimit, Tsz Chai Fung, Liang Peng, Qiuqi Wang

September 22, 2024

1 Numerical study: Comparing EoL and SL approaches

In this section, we substantiate the assertions presented in Section 1.2 of the manuscript regarding the comparison between the EoL and SL approaches. This is accomplished through a numerical investigation with a small N supported by theoretical reasoning. We simulate X_i from a Pareto (type-II) distribution with pdf $f_X(x) = (\alpha/\lambda)(1 + x/\lambda)^{-(\alpha+1)}$ with shape parameter $\alpha = 9$ and scale parameter $\lambda = 8$, such that the mean is $E[X_i] = \alpha/(\lambda - 1) = 1$. We choose $p = 0.75$ for the risk level, $\rho = 0.2$ for the loading factor, and consider $N = 2, 3, 5, 10$. We first compute the true VaR of the total cost $\mathrm{VaR}_p\big(T(d, N, \rho)\big)$ using (3), where $\mathrm{VaR}_p(\sum_{i=1}^N (X_i \wedge d))$ and $\mathbb{E}\{(X_1 - d)_+\}$ are approximated, respectively, by the empirical p-VaR and expectation of $\sum_{i=1}^N (X_i \wedge d)$ and $(X_1 - d)_+$ from $B = 50,000$ simulated samples. For example, we compute $\mathbb{E}\{\sum_{i=1}^N (X_i \wedge d)\} \approx (1/B) \sum_{j=1}^B \sum_{i=1}^N (X_{ij} \wedge d)$ with X_{ij} iid sampled from the Pareto distribution for $i = 1, \ldots, N$ and $j = 1, \ldots, B$.

Figure 1 plots $\mathrm{VaR}_p\big(T(d, N, \rho)\big)$ as a function of d for various N. We note that $\mathrm{VaR}_p\big(T(d, N, \rho)\big)$ is piecewise differentiable except for N turning points, which are illustrated by the gray dotted vertical lines added in Figure 1. In addition, if we denote the i-th turning point as $\tilde{d}_N(i)$ for $i = 1, \ldots, N-1$, we observe that $\tilde{d}_N(i) = \{d : P(\sum_{i=1}^N (X_i \wedge d) \geq (N - i + 1)d) = 1 - p\}$; indeed, the density function of $\sum_{i=1}^N (X_i \wedge d)$ exhibits jump points at integer multiples of d, causing the turning points.

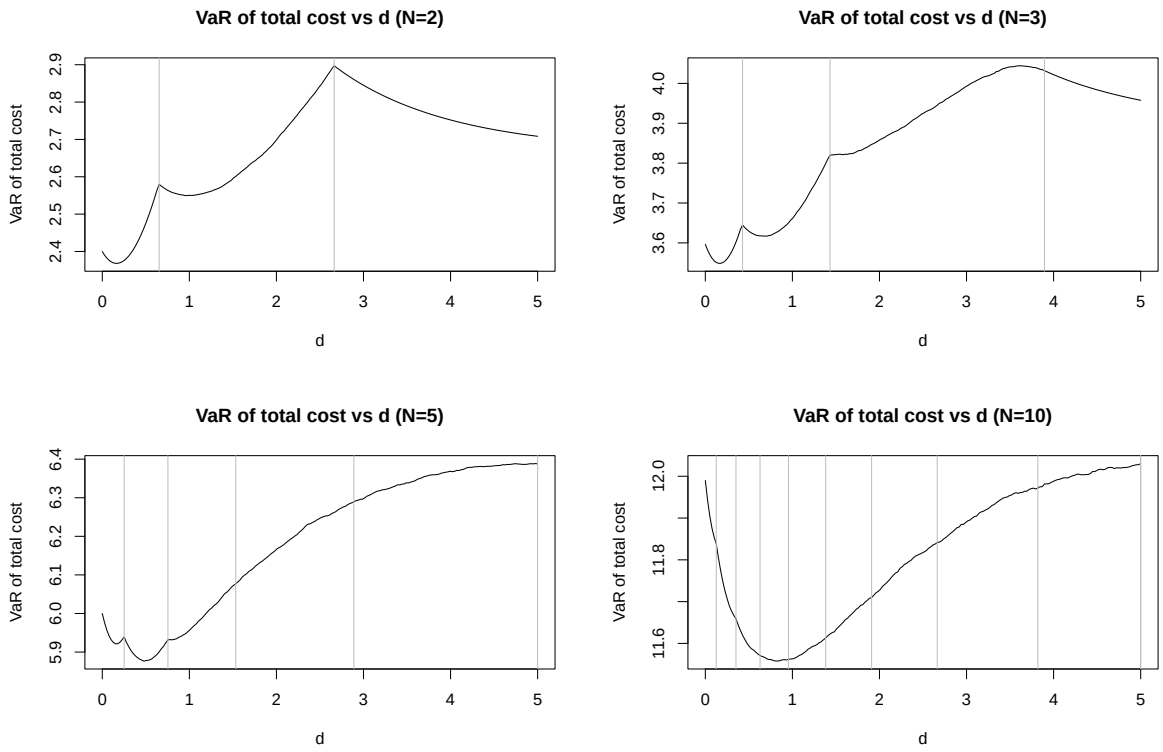

Figure 1: $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ versus $d \in [0,5]$ for various N. Gray vertical lines represent the turning points of the curves.

We then numerically calculate d^*, i.e., minimize $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ over d, and compute the probability $P(T(d^*,N,\rho) > \mathrm{VaR}_p(T(d^*,N,\rho)))$ for various N. An interesting series of observations is further noted.

First, the calculated value of d^* is 0.1633, 0.1643, 0.5025, 0.8779, respectively, for $N = 2,3,5,10$. This coincides with the analysis in Theorem 1 that the optimal retention increases with N if ρ is fixed.

Second, the calculated probability $P(T(d^*,N,\rho) > \mathrm{VaR}_p(T(d^*,N,\rho)))$ is 0, 0, 0.25 and 0.25, respectively, for $N = 2,3,5,10$. As N is large enough, the probability will be exactly 0.25, and otherwise, the probability will be exactly zero.

Third, we observe from the figure that $P(T(d^*,N,\rho) > \mathrm{VaR}_p(T(d^*,N,\rho))) = 0$ if $d^* \leq \tilde{d}_N(1)$ and $P(T(d^*,N,\rho) > \mathrm{VaR}_p(T(d^*,N,\rho))) = 0.25$ if $d^* > \tilde{d}_N(1)$. Indeed, one

can justify it theoretically as follows. From the definition of $\tilde{d}_N(i)$, if $d^* \leq \tilde{d}_N(1)$, we have

$$1-p = P(\sum_{i=1}^{N} X_i \wedge \tilde{d}_N(i) \geq N\tilde{d}_N(i)) = P(X_i \geq \tilde{d}_N(i))^N$$
$$\leq P(X_i \geq d^*)^N = P(\sum_{i=1}^{N} X_i \wedge d^* \geq Nd^*).$$

Since $P(\sum_{i=1}^N X_i \wedge d^* \geq Nd^*) \geq 1-p$, $\sum_{i=1}^N X_i \wedge d^*$ is upper bounded by Nd^* and hence $P(\sum_{i=1}^N X_i \wedge d^* > Nd^*) = 0$, we have $\mathrm{VaR}_p(\sum_{i=1}^N X_i \wedge d^*) = Nd^*$ and hence $P(\sum_{i=1}^N X_i \wedge d^* > \mathrm{VaR}_p(\sum_{i=1}^N X_i \wedge d^*)) = 0$. If $d^* > \tilde{d}_N(1)$, we have $\mathrm{VaR}_p(\sum_{i=1}^N X_i \wedge d^*) < Nd^*$, and the distribution function of $\sum_{i=1}^N X_i \wedge d^*$ is continuous on $(0, Nd^*)$. Hence, $P(\sum_{i=1}^N X_i \wedge d^* > \mathrm{VaR}_p(\sum_{i=1}^N X_i \wedge d^*)) = 1-p$ by the basic definition of quantile. Therefore, we conclude that the VaR of the total cost with the optimal retention is appropriate if and only if the optimal retention is above the first turning point.

Fourth, one can also show that $P(T(d^*, N, \rho) > \mathrm{VaR}_p(T(d^*, N, \rho))) = 1-p$ if and only if $P(X_i \geq d^*) \leq (1-p)^{1/N}$. With a larger N, this condition is more likely to hold. With $N = 1$, i.e., the SL approach following Cai and Tan (2007), we have $d^* = F^-_{X_1}(1-1/(1+\rho))$ given $1-p < (1+\rho)^{-1}$, and hence $P(X_i \geq d^*) = (1+\rho)^{-1} > (1-p)$, meaning that the condition never holds.

Overall, while $\mathbb{P}\left(T(d^*, 1, \rho) > \mathrm{VaR}_p(T(d^*, 1, \rho))\right) = 0$ under the SL approach, we empirically and theoretically show that $\mathbb{P}\left(T(d^*, N, \rho) > \mathrm{VaR}_p(T(d^*, N, \rho))\right) = 1-p$, the correct level, for sufficiently large N under the EoL approach. Hence, the EoL optimal retention would not inherit the same counter-intuitive property as the SL optimal retention.

2 Approximating VaR of total cost by Edgeworth expansion

To address the issue of insufficient accuracy outlined by Section 4.1 under the constant loading factor rule, we employ Edgeworth expansion to improve the approximation precision from (4) in the manuscript. Suppose that $Z_1, \ldots, Z_N$ are iid random variables

with zero mean, unit variance, and $E[Z_1^4] < \infty$. Then, standard Edgeworth expansion yields

$$P\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} Z_i \leq x\right) = \Phi(x) - \frac{1}{\sqrt{N}}p_1(x)\phi(x) + \frac{1}{N}p_2(x)\phi(x) + o(N^{-1}), \tag{1}$$

where $p_1(x) = -\kappa_3 H_2(x)/6$, $p_2(x) = -(\kappa_4 H_3(x)/24 + \kappa_3^2 H_5(x)/72)$. Here, $\kappa_3 = E[(Z_1 - E[Z_1])^3]$ and $\kappa_4 = E[(Z_1 - E[Z_1])^4] - 3E[(Z_1 - E[Z_1])^2]^2$ are the moment quantities, and $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$ and $H_5(x) = x^5 - 10x^3 + 15x$ are the Hermite polynomials. From (1) above and (3) in the manuscript, one can apply the Cornish-Fisher expansion and write

$$\begin{aligned}\mathrm{VaR}_p\left(T(d,N,\rho)\right) = \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\left[\Phi^-(p) + \frac{1}{\sqrt{N}}\tilde{p}_1(p;d) + \frac{1}{N}\tilde{p}_2(p;d)\right] \\ + N\mathbb{E}[X_1] + N\rho\nu_1(d) + o\left(\frac{1}{\sqrt{N}}\right),\end{aligned} \tag{2}$$

where $\tilde{p}_1(p;d) = -\frac{\tilde{\kappa}_3(d)}{6}H_2(p)$ and

$$\tilde{p}_2(p;d) = \frac{\tilde{\kappa}_4(d)}{24}H_3(p) + \frac{\tilde{\kappa}_3(d)^2}{72}\left(H_5(p) + 2H_2'(p)H_2(p) - pH_2(p)^2\right)$$

with

$$\tilde{\kappa}_3(d) = \frac{\mathbb{E}[(X_1 \wedge d - \mathbb{E}[X_1 \wedge d])^3]}{\mathbb{E}[(X_1 \wedge d - \mathbb{E}[X_1 \wedge d])^2]^{3/2}}, \qquad \tilde{\kappa}_4(d) = \frac{\mathbb{E}[(X_1 \wedge d - \mathbb{E}[X_1 \wedge d])^4]}{\mathbb{E}[(X_1 \wedge d - \mathbb{E}[X_1 \wedge d])^2]^2} - 3.$$

We alternatively propose to compute the approximate optimal retention by minimizing

$$G_{N,\rho}^{(2)}(d) = N\mathbb{E}[X_1] + N\rho\nu_1(d) + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\left[\Phi^-(p) + \frac{1}{\sqrt{N}}\tilde{p}_1(p;d),\right] \tag{3}$$

or

$$\begin{aligned}G_{N,\rho}^{(3)}(d) = N\mathbb{E}[X_1] + N\rho\nu_1(d) \\ + \sqrt{N}\sqrt{\mu_2(d) - \mu_1^2(d)}\left[\Phi^-(p) + \frac{1}{\sqrt{N}}\tilde{p}_1(p;d) + \frac{1}{N}\tilde{p}_2(p;d)\right],\end{aligned} \tag{4}$$

where the error terms underlying $G^{(2)}_{N,\rho}$ and $G^{(3)}_{N,\rho}$ are, respectively, $o(1)$ and $o(1/\sqrt{N})$. Recall that the approximation order of $G_{N,\rho}(d)$ in (5) of the manuscript is $o(\sqrt{N})$.

We now numerically evaluate how the Edgeworth approximation technique improves the accuracy of the VaR approximation under the constant loading factor rule, as discussed in the simulation study in Section 4.1 of the manuscript. The left panels of Figure 2 plot $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ (black curves), $G_{N,\rho}(d)$ (red curves), $G^{(2)}_{N,\rho}(d)$ (green curves), and $G^{(3)}_{N,\rho}(d)$ (blue curves) as functions of d for $N = 10$ and $N = 100$ under the constant loading factor rule. The right panels of Figure 2 illustrate the differences between the approximated and actual VaRs for various approximation methods. It is evident that the approximation errors of VaR, particularly for the higher-order Edgeworth expansions (blue curves), are significantly reduced in comparison to the normal approximations. Additionally, we compute the optimal retention that minimizes the approximated VaRs $G^{(2)}_{N,\rho}(d)$ (with an approximation order of $o(1)$) and $G^{(3)}_{N,\rho}(d)$ (with an approximation order of $o(1/\sqrt{N})$), and the results are summarized in Table 1. This table expands upon Table 1 in the manuscript to include the results of higher-order approximation methods. The table demonstrates that the discrepancies between d^* and $d^*_{N,\rho}$ are significantly reduced when using higher-order approximation techniques. In summary, employing a constant loading factor requires higher-order Edgeworth expansions for accurately approximating the optimal retention, especially for larger N.

Figure 2: *Left panels*: $\mathrm{VaR}_p\big(T(d,N,\rho)\big)$ (black curves), $G_{N,\rho}(d)$ (red curves), $G^{(2)}_{N,\rho}(d)$ (green curves) and $G^{(3)}_{N,\rho}(d)$ (blue curves) for $N = 10, 100$ under the constant loading factor. *Right panels*: $[G_{N,\rho}(d) - \mathrm{VaR}_p\big(T(d,N,\rho)\big)]$ (red curves), $[G^{(2)}_{N,\rho}(d) - \mathrm{VaR}_p\big(T(d,N,\rho)\big)]$ (green curves) and $[G^{(3)}_{N,\rho}(d) - \mathrm{VaR}_p\big(T(d,N,\rho)\big)]$ (blue curves) versus d.

Table 1: Actual optimal retention d^*, approximately optimal retention $d^*_{N,\rho}$, and the relative difference between d^* and $d^*_{N,\rho}$ (in %) across various loading factor rules, N, and approximation orders.

Loading factor rule	N	Approx. order	Actual	Approx.	Diff. (%)
Constant loading factor	10	$o(\sqrt{N})$	1.8549	1.4856	-19.91
Constant loading factor	10	$o(1)$	1.8549	1.6276	-12.25
Constant loading factor	10	$o(1/\sqrt{N})$	1.8549	1.5921	-14.17
Constant loading factor	25	$o(\sqrt{N})$	3.4442	2.6838	-22.08
Constant loading factor	25	$o(1)$	3.4442	2.9634	-13.96
Constant loading factor	25	$o(1/\sqrt{N})$	3.4442	2.9969	-12.99
Constant loading factor	100	$o(\sqrt{N})$	7.1241	5.6581	-20.58
Constant loading factor	100	$o(1)$	7.1241	6.3361	-11.06
Constant loading factor	100	$o(1/\sqrt{N})$	7.1241	6.6660	-6.43
Decreasing loading factor	10	$o(\sqrt{N})$	0.5034	0.5472	8.70
Decreasing loading factor	25	$o(\sqrt{N})$	0.5835	0.5472	-6.21
Decreasing loading factor	100	$o(\sqrt{N})$	0.5472	0.5472	0.02
Standard deviation principle	10	$o(\sqrt{N})$	0.7847	0.8189	4.36
Standard deviation principle	25	$o(\sqrt{N})$	0.8187	0.8189	0.03
Standard deviation principle	100	$o(\sqrt{N})$	0.8499	0.8189	-3.64
Sharpe ratio principle	10	$o(\sqrt{N})$	0.2797	0.3218	15.06
Sharpe ratio principle	25	$o(\sqrt{N})$	0.3149	0.3218	2.18
Sharpe ratio principle	100	$o(\sqrt{N})$	0.3203	0.3218	0.45