Robust Risk Aggregation Techniques and Applications

by

Yuyu Chen

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

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

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

External Examiner:	Qihe Tang Professor, School of Risk and Actuarial Studies, University of New South Wales
Supervisor:	Ruodu Wang Professor, Department of Statistics and Actuarial Science, University of Waterloo
Supervisor:	Ken Seng Tan Professor, Nanyang Business School, Nanyang Technological University
Internal Member:	Greg Rice Associate Professor, Dept. of Stats. and Actuarial Science, University of Waterloo
Internal Member:	Alexander Schied Professor, Department of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Tao Chen Associate Professor, Department of Economics, University of Waterloo

Author's Declaration

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

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

Abstract

Risk aggregation, which concerns the statistical behaviors of an aggregation position $S(\mathbf{X})$ associated with a random vector $\mathbf{X} = (X_1, \dots, X_n)$, is an important research topic in risk management, economics, and statistics. The distribution of $S(\mathbf{X})$ is determined by both the marginal behaviors and the joint dependence structure of $\mathbf{X}$. In general, it is challenging to obtain an accurate estimation of the dependence structure of $\mathbf{X}$ compared with the estimation of the marginal distributions. Given the marginal distributions of $\mathbf{X}$, this thesis focuses on studying the aggregation position $S(\mathbf{X})$ with different dependence assumptions in different contexts. We will assume that $\mathbf{X}$ has a specific dependence structure (e.g., independence), or its dependence structure is (partially) unknown. In particular, for the case that the dependence structure is (partially) unknown, we are interested in the worst-case and the best-case scenarios of $S(\mathbf{X})$.

In Chapter 2, we show the surprising inequality that the weighted average of iid ultra heavy-tailed Pareto losses (with infinite mean) is larger than a standalone loss in the sense of first-order stochastic dominance. This result is further generalized to allow for random total number and weights of Pareto losses and for the losses to be triggered by catastrophic events. We discuss several important implications of these results via an equilibrium analysis of a risk exchange market. First, diversification of ultra heavy-tailed Pareto losses leads to increases in portfolio risk, and thus diversification penalty exists. Second, agents with ultra heavy-tailed Pareto losses will not share risks in a market equilibrium. Third, transferring losses from agents bearing Pareto losses to external parties without any losses may arrive at an equilibrium which benefits every party involved.

In Chapter 3, we focus on aggregation sets, which represent model uncertainty due to unknown dependence structure of random vectors. We investigate ordering relations between two aggregation sets for which the sets of marginals are related by two simple operations: distribution mixtures and quantile mixtures. Intuitively, these operations "homogenize" marginal distributions by making them similar. As a general conclusion from our results, more "homogeneous" marginals lead to a larger aggregation set, and thus more severe model uncertainty, although the situation for quantile mixtures is much more complicated than that for distribution mixtures. We proceed to study inequalities on the worst-case values of risk measures in risk aggregation, which represent conservative calculation of regulatory capital. Among other results, we obtain an order relation on VaR under quantile mixture for marginal distributions with monotone densities. Numerical results are presented to visualize the theoretical results. Finally, we provide applications on portfolio diversification under dependence uncertainty and merging p-values in multiple hypothesis testing, and discuss the connection of our results to joint mixability.

In Chapter 4, we study the aggregation of two risks when the marginal distributions are known and the dependence structure is unknown, with the additional constraint that one risk is smaller than or equal to the other. Risk aggregation problems with the order constraint are closely related to the recently introduced notion of the directional lower (DL) coupling. The largest aggregate risk in concave order (thus, the smallest aggregate risk in convex order) is attained by the DL coupling. These results are further generalized to calculate the best-case and worst-case values of tail risk measures. In particular, we obtain analytical formulas for bounds on Value-at-Risk. Our numerical results suggest that the new bounds on risk measures with the extra order constraint can greatly improve those with full dependence uncertainty.

In Chapter 5, we study various methods for combining p-values from multiple hypothesis testing into one p-value, under different dependence assumptions of p-values. We say that a combining method is valid for arbitrary dependence if it does not require any assumption on the dependence structure of the p-values, whereas it is valid for some dependence if it requires some specific, perhaps realistic, but unjustifiable, dependence structures. The trade-off between the validity and efficiency of these methods is studied by analyzing the choices of critical values under different dependence assumptions. We introduce the notions of independence-comonotonicity balance (IC-balance) and the price for validity. In particular, IC-balanced methods always produce an identical critical value for independent and perfectly positively dependent p-values, a specific type of insensitivity to a family of dependence assumptions. We show that among two very general classes of merging methods commonly used in practice, the Cauchy combination method and the Simes method are the only IC-balanced ones. Simulation studies and a real-data analysis are conducted to analyze the size and power of various combining methods in the presence of weak and strong dependence.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisors Dr. Ruodu Wang and Dr. Ken Seng Tan. They have always been inspiring, offering unconditional support, and providing invaluable guidance. Dr. Wang has spent a generous amount of time helping me become a better researcher in many different aspects, no matter how big or small. Dr. Tan has always been supportive, and I am grateful for his encouragement and advice. I would like to thank Dr. Wang and Dr. Tan for taking me on as their student. The opportunity to work in Waterloo means a lot to me.

I would like to thank the rest of the committee members, Dr. Tao Chen, Dr. Greg Rice, Dr. Alexander Schied, and Dr. Qihe Tang for spending their valuable time reading my thesis and providing insightful comments. Special thanks go to Dr. Alexandru Badescu, who encouraged me to pursue a PhD degree and has always been supporting my academic journey.

I would like to thank Dr. Paul Embrechts, Dr. Peng Liu, Dr. Yang Liu, and Liyuan Lin for collaborating and sharing their expertise with me. Discussions with them are always inspiring. This thesis would not have been possible without them. Also, I would like to thank Qiuqi Wang, Wenyuan Li, and Xiyue Han, with whom I had many interesting discussions. My thanks also go to Dr. Zijia Wang, who provided me with much help and guidance since the start of my PhD years.

I would like to thank the faculty and staff in the department, who have created such a supportive environment for all the students. Thanks to the help from Ms. Mary Lou Dufton and Mr. Greg Preston, my life becomes much easier. I am also grateful to Dr. David Landriault, who supported me in taking one of the examinations towards the Associate of the Society of Actuaries.

Last but not least, I would like to thank my partner Meng Yuan for her endless love, unconditional support, and understanding. She accompanied me through all the hard times during the pandemic. Without Meng, none of this would ever have been possible.

Dedication

Dedicated to Yu Wei.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Risk aggregation and risk measures

In quantitative risk management, *risk aggregation* refers to the statistical behaviors of an *aggregation position* $S(\mathbf{X})$ associated with a risk vector $\mathbf{X} = (X_1, \dots, X_n)$ where the random variables $X_1, \dots, X_n$ represent individual losses in a fixed period of time. We will mainly focus on the aggregation position $S(\mathbf{X}) = X_1 + \cdots + X_n$, which is most relevant in quantitative risk management as it can be simply interpreted as the portfolio loss of a financial institution. In this section, we introduce several important topics regarding the aggregation position $S(\mathbf{X})$ not only in the context of quantitative risk management, but also in economics and statistics.

Measuring the risk of a financial portfolio is crucial in both banking and insurance sectors, and it is typically done by calculating the value of a risk measure. A *risk measure*, which maps a loss random variable to a real number, represents the conservative calculation of regulatory capital requirement for financial institutions (see McNeil et al. (2015) and Föllmer and Schied (2016)). Fix an atomless probability space $(\Omega, \mathcal{A}, \mathbb{P})$, let $\mathcal{X}$ be the set of random variables, and let $\mathcal{M}$ be the set of distributions. In this thesis, *law-invariant* risk measures are mappings from $\mathcal{X}$ to $\mathbb{R}$, and we also treat them as mappings from $\mathcal{M}$ to $\mathbb{R}$. We abuse the notation for convenience in the introduction. Two important risk measures in banking and insurance regulatory frameworks are Value-at-Risk (VaR) and Expected Shortfall (ES). For $F \in \mathcal{M}$ and $p \in (0, 1)$, VaR and ES are defined as

$$\mathrm{VaR}_p(F) = \inf\{x \in \mathbb{R} : F(x) \geqslant p\} \quad \text{and} \quad \mathrm{ES}_p(F) = \frac{1}{1-p}\int_p^1 \mathrm{VaR}_u(F)\mathrm{d}u.$$

To calculate a risk measure on the aggregation position $S(\mathbf{X})$, we need to know its distribution, which by Sklar's theorem (see, e.g., Rüschendorf (2013)) is determined by the marginal distributions and the dependence structure (i.e., copula[1]) of $\mathbf{X}$; we refer to Nelsen (2006) for an introduction to copulas. Many researchers have studied the distribution of $S(\mathbf{X})$ with complete information of the marginal distributions and dependence structure of $\mathbf{X}$. Commonly used classes of marginal distributions for financial losses can be found in, e.g., Klugman et al. (2012) and McNeil et al. (2015). Individual risk models in the actuarial literature assume that risks are independent; see Klugman et al. (2012). Other important classes of dependence structures in risk management include copulas derived from elliptical distributions and Archimedean copulas; see McNeil et al. (2015). Arguably, the estimations of univariate distributions are accurate compared to those of dependence structure of risks. In this thesis, we always assume that the marginal distributions of $\mathbf{X}$ are known and study the aggregation position $S(\mathbf{X})$ with different assumptions of dependence structures.

Given the marginal distributions and dependence structure of $\mathbf{X} = (X_1, \dots, X_n)$, one crucial question in risk management is whether a risk measure ρ has the property that

$$\rho(X_1 + \cdots + X_n) \leqslant \rho(X_1) + \cdots + \rho(X_n).$$

Such a property is referred to as *subadditivity*, one of the four axioms of a coherent risk measure (Artzner et al. (1999)). Subadditivity means that diversification of risks leads to a smaller risk assessment than the sum of risk assessments of individual risks, thus *diversification benefit*; see McNeil et al. (2015) for discussions on implications of subadditivity. As for the two regulatory risk measures, ES is subadditive as it is a coherent risk measure, whereas VaR is not subadditive in general. For instance, Ibragimov (2009) showed that VaR_p, $p \in (0.5, 1)$, is non-subadditive for independent risks which follow a convolution of symmetric α-stable distributions with tail index $\alpha \leqslant 1$ (i.e., infinite first moment); see Ibragimov et al. (2009) and Ibragimov et al. (2011) for discussions on diversification of heavy-tailed distributions. In Chapter 2, we discuss the diversification effects of independent Pareto risks without finite mean which can be used to model catastrophic losses, operational losses, and large insurance losses. Several implications of these results in a risk exchange market are presented through an equilibrium model.

In some circumstances, it is very challenging to capture the dependence structure of risks due to limited choices of multivariate models and their statistical inference issues. The failure to capture the dependence structure may lead to a significantly different calculation of risk measures and thus unsound risk management strategies. For instance, using the

[1]Copulas are joint distributions of standard uniform random variables.

Gaussian copula to describe the dependence structure of risks may underestimate the probability of joint large losses as Gaussian copula is asymptotically independent in the tails (e.g., McNeil et al. (2015)). In cases where the dependence structure of risks cannot be estimated accurately, we assume that the dependence structure is (partially) unknown, and the unknown dependence structure is referred to as *dependence uncertainty*; see Bernard et al. (2014), Embrechts et al. (2015) and the references therein.

Under the term *robust risk aggregation*, one studies the aggregation position $S(\mathbf{X})$ with known marginal distributions of $\mathbf{X}$ and dependence uncertainty. Let $\mathbf{F} = (F_1, \dots, F_n)$, where $F_1, \dots, F_n$ are the marginal distributions of $\mathbf{X}$. We define the *aggregation set* (Bernard et al. (2014)) with full dependence uncertainty (i.e., no dependence information is available) as

$$\mathcal{D}_n(\mathbf{F}) = \{\text{Distribution of } X_1 + \cdots + X_n : X_i \sim F_i,\ i = 1, \dots, n\}. \tag{1.1}$$

The aggregation set $\mathcal{D}_n(\mathbf{F})$ fully describes the model uncertainty of $X_1 + \cdots + X_n$ due to the unknown dependence structure of $\mathbf{X}$. However, it is extremely difficult to give an analytical characterization of $\mathcal{D}_n(\mathbf{F})$. The only available analytical results of $\mathcal{D}_n(\mathbf{F})$ are given by Mao et al. (2019) for standard uniform distributions. In Chapter 3, we study the inclusion relations of two aggregation sets whose marginal distributions are related by some simple operations.

Over the aggregation set (1.1), one practically revelant question is what are the largest and the smallest values of a risk measure. For a risk measure ρ, the *worst-case value* of ρ is defined as

$$\overline{\rho}(\mathbf{F}) = \sup\left\{\rho(F) : F \in D_n(\mathbf{F})\right\},$$

and the *best-case value* of ρ is defined as

$$\underline{\rho}(\mathbf{F}) = \inf\left\{\rho(F) : F \in D_n(\mathbf{F})\right\}.$$

We refer to McNeil et al. (2015) for general discussions on the bounds of risk measures. The difference between the worst-case and the best-case values of a risk measure is called *Dependence Uncertainty spread* which is used to measure the model uncertainty of a portfolio; see Embrechts et al. (2015). From the perspective of risk management, the worst-case value of a risk measure is particularly important, as it ensures a sufficient capital requirement for financial institutions. Moreover, the techniques used to find the upper bounds of commonly used risk measures can also be used to find their lower bounds. Therefore, we will mainly focus on the worst-case value of a risk measure in this thesis. For the two important regulatory risk measures, VaR and ES, an explicit result is available for the worst-case value of ES as it is a coherent risk measure (Artzner et al. (1999)) and

additive for comonotonic risks (Dhaene et al. (2006)), whereas an analytic formula for the worst-case value of VaR is not available for general marginal distributions.

We summarize below some results for calculating or approximating the worst-case value of VaR defined as

$$\overline{\mathrm{VaR}}_p(\mathbf{F}) = \sup\left\{\mathrm{VaR}_p(X_1 + \cdots + X_n) : X_i \sim F_i, i = 1, \ldots, n\right\}. \tag{1.2}$$

Early results for (1.2) date back to Makarov (1981) and Rüschendorf (1982), who solved the problem explicitly for the case that $n = 2$. For homogeneous marginal distributions with monotone densities, analytic formulas were obtained by Wang et al. (2013) and Puccetti and Rüschendorf (2013). For heterogeneous marginal distributions with monotone densities, Jakobsons et al. (2016) provided a solution to (1.2) by solving a group of functional equations. More recently, Blanchet et al. (2020) established a new analytic formula which covers its previous results to (1.2). Another direction is to apply ES bounds to approximate (1.2) for large n by using the asymptotic equivalence between VaR and ES in Embrechts et al. (2015). The Rearrangement Algorithm (RA) is available in Puccetti and Rüschendorf (2012) and Embrechts et al. (2013) for numerical calculation. In Chapter 3, we compare the worst-case values of a risk measure over two related aggregation sets by using the inclusion relations of the two aggregation sets and some other techniques. Thus upper bounds for (1.2) can be provided.

Although the worst-case value of a risk measure based on the sole knowledge of marginal distributions ensures sufficient capital requirements for financial institutions, it can be too large to be practically useful. Many efforts have been made to improve the worst-case value of a risk measure by incorporating partial dependence information or adding additional constraints into the problem. For instance, a variance constraint is imposed at the portfolio level by Bernard et al. (2017). A lower bound is placed on the corresponding copula of risks by Puccetti et al. (2016). Puccetti et al. (2017) used independence assumptions among groups of risks and left the dependence structure within each group unknown. Bernard et al. (2017b) considered a partially specified factor model with dependence uncertainty. In Chapter 4, we study the aggregation of two risks with known marginal distributions, unknown dependence structure, and an order constraint that one risk is less than the other almost surely.

The techniques developed in robust risk aggregation are applicable not only in risk management and finance but also in other research areas such as multiple hypothesis testing. One of the main goals in multiple hypothesis testing is to merge multiple p-values into a test statistic and give the corresponding critical value. The critical values can be derived with or without a specific dependence assumption on the p-values. Those

methods, which are valid for arbitrary dependence structures of p-values, are referred to as *VAD* (valid for arbitrary dependence structures) methods. Recently, Vovk and Wang (2020) applied the existing results for (1.2) to develop the VAD averaging methods. On the other hand, methods that are valid for some specific dependence assumption of p-values are referred to as *VSD* (valid for some specific dependence structures) methods. Some dependence assumptions imposed by VSD methods include independence (e.g., Tippett (1931), Pearson (1933), Fisher (1948), and Simes (1986)) and Gaussian copula (e.g., Liu and Xie (2020)). In Chapter 5, the trade-off on validity and efficieny of hypothesis testing between using VAD and VSD methods is discussed.

1.2 Contributions of the thesis

This thesis focuses on developing techniques in risk aggregation and applying these techniques to solve different problems in risk management, economics, and statistics. In the following, we briefly describe the contributions of each chapter.

In Chapter 2, we study the diversification effects of independent Pareto risks, which frequently appear in modeling catastrophic risks, operational risks, and wealth allocations; see Embrechts et al. (1999), McNeil et al. (2015), and Taleb (2020). For two random variables X and Y, we say X is smaller than Y in first-order stochastic dominance, if $\mathbb{P}(X \leqslant x) \geqslant \mathbb{P}(Y \leqslant x)$ for all $x \in \mathbb{R}$. We show that the weighted average of iid ultra heavy-tailed Pareto losses (i.e., the losses have infinite mean) is larger than a standalone ultra heavy-tailed Pareto loss in the sense of first-order stochastic dominance; see Ibragimov (2009) for a relevant result for iid risks which are convolutions of symmetric stable random variables without finite mean. Our result is further generalized to allow for random total number and weights of Pareto losses and for the losses to be triggered by catastrophic events. Special cases of our results are studied in Example 7 of Embrechts et al. (2002) and Embrechts and Puccetti (2010, Figure 5.2).

A direct interpretation of these results is that diversification of ultra heavy-tailed Pareto losses leads to more severe portfolio risk, and thus diversification penalty. We discuss several implications of these results on risk sharing via an equilibrium model. First, agents with ultra heavy-tailed Pareto losses will not share risks in a market equilibrium. Second, transferring losses from agents bearing ultra heavy-tailed Pareto losses to external parties without any losses may arrive at an equilibrium which benefits every party involved. Moreover, we show that if the Pareto losses have finite mean, agents with initial Pareto losses may still prefer diversifying risks. Hence, whether Pareto losses have finite mean plays an

important role in the effects of diversification; see Ibragimov et al. (2009) for discussions on the diversifications of Pareto losses in a different model.

Chapter 3 first studies the inclusion relationship between two aggregation sets $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\mathbf{G})$, where $\mathbf{G}$ is a tuple of distributions obtained by $\mathbf{F}$ through operations of *distribution mixture* or *quantile mixture*. Intuitively, those operations "homogenize" the distribution tuple $\mathbf{F}$. The general message is that the more "homogeneous" the distribution tuple is, the larger its aggregation set is. In particular, $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\mathbf{G})$ if $\mathbf{G}$ is a distribution mixture of $\mathbf{F}$. The set inclusion relationship for quantile mixture does not hold in general but is established for uniform marginal distributions.

A practical relevance of the set inclusions is to compare the worst-case values of risk measures of $\mathbf{F}$ and $\mathbf{G}$, as $D_n(\mathbf{F}) \subset D_n(\mathbf{G})$ implies $\overline{\rho}(\mathbf{F}) \leqslant \overline{\rho}(\mathbf{G})$ for any risk measure ρ. Among other inequalities we obtain for the worst-case values of risk measures, we show that quantile mixture provides an upper bound for (1.2) if the marginal distributions of $\mathbf{F}$ have monotone densities. More specific inequalities are derived for the interesting case that all marginal distributions are Pareto distributions without finite mean. Applications of our results to portfolio diversification and multiple hypothesis testing are discussed. In particular, if VaR is used as the risk measure, diversification under dependence uncertainty may lead to more severe losses.

Chapter 4 studies the aggregation of two risks when the marginal distributions are known, and the dependence structure is unknown. In addition, we impose an order constraint that one risk is smaller than or equal to the other almost surely. Risk aggregation problems with the order constraint are closely related to the recently introduced notion of the directional lower (DL) coupling; see Arnold et al. (2020) and Nutz and Wang (2021). In particular, the largest aggregation position of two ordered risks in concave order[2] (thus, the smallest aggregate risk in convex order) is attained by the DL coupling.

Many commonly used risk measures are consistent with concave (convex) order; see Mao and Wang (2020) for characterizations of risk measures consistent with convex order. Thus, the DL coupling gives the worst-case (best-case) values of those risk measures. These results are further generalized to calculate the best-case and worst-case values of *tail risk measures* (Liu and Wang (2021)), which are not necessarily consistent with concave or convex order. The class of tail risk measures, whose values are solely determined by the tail behavior of the aggregation position, includes VaR, ES, and Range Value-at-Risk (RVaR) (Cont et al. (2010)) as special cases. In particular, we obtain analytical formulas for bounds on VaR, which can also be used to derive the bounds of default probabilities of

[2]A random variable X is said to be smaller than a random variable Y in concave order if $\mathbb{E}[u(X)] \leqslant \mathbb{E}[u(Y)]$ for all concave functions $u : \mathbb{R} \to \mathbb{R}$ provided that the expectations exist.

the aggregation position. Our numerical results show that the new bounds on risk measures with the extra order constraint can greatly improve those solely based on the knowledge of marginal distributions.

In Chapter 5, we apply the techniques for solving (1.2) in robust risk aggregation to study VAD methods of merging p-values, and analyse the trade-off between validity and efficiency for VAD and VSD methods of several test statistics. One key issue of the VSD methods is that it is very difficult to verify the dependence assumptions as only one set of p-values is usually available in practice. If the dependence structure is not correctly specified, VSD methods may not have the correct size. As a comparison, VAD methods can always control the size below the significance level under any dependence structure of p-values, but they may yield less power than the corresponding VSD methods. Therefore, there is always a trade-off between using VAD and VSD methods.

We introduce the notions of independence-comonotonicity balance (IC-balance) and the price for validity. In particular, IC-balanced methods always produce an identical critical value for independent and perfectly positively dependent p-values, thus showing insensitivity to dependence assumptions. We show that, among two very general classes of merging methods, the Cauchy combination (Liu and Xie (2020)) and the Simes method (Simes (1986)) are the only IC-balanced ones. The harmonic averaging (Wilson (2019)) and Cauchy combination methods are asymptotically equivalent in several senses. The price for validity is used to measure the loss of efficiency of the hypothesis test when the dependence assumption is changed from some specific dependence structure to arbitrary dependence structures. The prices for validity of the Simes, the Cauchy combination, and the harmonic averaging methods increase at moderate rates as the number of p-values increases. These theoretical results explain the wide applications of these methods in different statistical procedures.

Chapter 6 concludes the thesis and discusses possible future research problems and topics. The proofs of some theoretical results and technical discussions are put in the appendix of each chapter.

To keep each chapter's content self-contained, important concepts such as risk measures, VaR, and ES will be reintroduced in each chapter, with slightly different conventions. In particular, risk measures are mappings from $\mathcal{X}$ to $\mathbb{R}$ in Chapter 2, and they are mappings from $\mathcal{M}$ to $\mathbb{R}$ in Chapters 3 and 4. Our choice of convention is for the convenience of presentation.

Chapter 2

An unexpected stochastic dominance: Pareto distributions, catastrophes, and risk exchange

2.1 Introduction

Pareto distributions are arguably the most important class of heavy-tailed loss distributions, due to their connection to regularly varying tails, extreme value theory, and power laws in economics and social networks; see, e.g., Embrechts et al. (1997), de Haan and Ferreira (2006) and Gabaix (2009). In quantitative risk management, Pareto distributions are frequently used to model losses from catastrophes such as earthquakes, hurricanes, and wildfires; see, e.g., Embrechts et al. (1999). They are also widely used in economics for wealth distributions (e.g., Taleb (2020)) and modeling the tails of financial asset losses and operational risks (e.g., McNeil et al. (2015)). Andriani and McKelvey (2007) contains over 80 references to diverse fields of applications. By the Pickands-Balkema-de Haan Theorem (Pickands (1975) and Balkema and de Haan (1974)), generalized Pareto distributions are the only possible non-degenerate limiting distributions of the residual life time of random variables exceeding a high level.

Stochastic dominance relations are an important tool in economic decision theory which allow for the analysis of risk preferences for a group of decision makers (Hadar and Russell (1969)). The strongest form of commonly used stochastic dominance relations is first-order stochastic dominance. For two random variables X and Y representing random losses, we say X is smaller than Y in *first-order stochastic dominance*, denoted by $X \leqslant_{\mathrm{st}} Y$, if

$\mathbb{P}(X \leqslant x) \geqslant \mathbb{P}(Y \leqslant x)$ for all $x \in \mathbb{R}$. Write $X \simeq_{\mathrm{st}} Y$ if X and Y have the same distribution. The relation $X \leqslant_{\mathrm{st}} Y$ means that all decision makers with an increasing[1] utility function will prefer the loss X to the loss Y; see Hadar and Russell (1971).

For iid random variables $X_1, \dots, X_n$ following a Pareto distribution with infinite mean and weights $\theta_1, \dots, \theta_n \geqslant 0$ with $\sum_{i=1}^n \theta_i = 1$, our main finding in Theorem 2.1 is the stochastic dominance relation

$$X_1 \leqslant_{\mathrm{st}} \theta_1 X_1 + \cdots + \theta_n X_n, \tag{2.1}$$

and the inequality (2.1) is strict in a natural sense. As far as we are aware, the inequality (2.1) is not known in the literature, even in the case that $\theta_1, \dots, \theta_n$ are equal (i.e., they are $1/n$). It is somewhat surprising that, for infinite mean losses, the inequality (2.1) holds for the strongest form of risk comparison: for every monotone decision maker (with precise definition in Section 2.4), a diversified portfolio of such iid Pareto losses is less preferred to a non-diversified one.

To appreciate the remarkable nature of (2.1), we first remark that for any identically distributed random variables $X_1, \dots, X_n$ with finite mean, regardless of their distribution or dependence structure, for $\theta_1, \dots, \theta_n > 0$ with $\sum_{i=1}^n \theta_n = 1$, (2.1) can only hold if $X_1 = \cdots = X_n$ (almost surely), in which case we have the trivial equality $X_1 = \theta_1 X_1 + \cdots + \theta_n X_n$; see Proposition 2.1. Therefore, the assumption of infinite mean is very important for (2.1) to hold.

Observations similar to (2.1), although with less generality, are made in the literature in different forms. Samuelson (1967) mentioned that having an infinite mean in portfolio diversification may lead to a worse distribution; see also p. 271 in Fama and Miller (1972) and Malinvaud (1972). The inequality (2.1) for $n = 2$ and the Pareto tail parameter $\alpha = 1/2$ (see Section 2.2 for the parametrization) has an explicit formula in Example 7 of Embrechts et al. (2002). A numerical example for $n = 3$ and $\alpha = 1$ is provided by Embrechts and Puccetti (2010, Figure 5.2). A relevant result of Ibragimov (2009) is that for iid random variables $Z_1, \dots, Z_n$ which follow a convolution of symmetric stable distributions without finite mean, $\mathbb{P}(\theta_1 Z_1 + \cdots + \theta_n Z_n \leqslant x) \leqslant \mathbb{P}(Z_1 \leqslant x)$ for $x > 0$ but not for $x < 0$ (and hence first-order stochastic dominance does not hold[2]). The symmetry of distributions is essential for this inequality, and $Z_1, \dots, Z_n$ can take negative values, unlike Pareto losses, which are positive, skewed and more suitable for the modeling of extreme losses.

[1] In this chapter, all terms like "increasing" and "decreasing" are in the non-strict sense.

[2] This relation is closer to second-order stochastic dominance; see Ibragimov and Walden (2007).

In the realm of banking and insurance, Pareto distributions with infinite mean occur as a possible mathematical model after careful statistical analysis in several contexts. For instance, catastrophic losses, operational losses, large insurance losses, and financial returns from technological innovations, are often modelled by Pareto distributions without finite mean; Section 2.1.1 below collects some examples and related literature.

In risk management, the inequality (2.1) yields *superadditivity* of the regulatory risk measure Value-at-Risk (VaR) in banking and insurance sectors; that is, the weighted average of Pareto losses without finite mean gives a larger VaR than that given by an individual Pareto loss. Different from the literature on VaR superadditivity for regularly varying distributions (e.g., Embrechts et al. (2009) and McNeil et al. (2015)), the superadditivity of VaR implied by (2.1) holds for all probability levels, and this not just in some asymptotic sense.

We obtain several generalizations of the inequality (2.1) for other models in Sections 2.2 and 2.3. In particular, Proposition 2.3 in Section 2.2 deals with losses that are Pareto only in the tail region, and Theorem 2.2 in Section 2.3 addresses losses triggered by catastrophic events, a setting where ultra heavy-tailed Pareto losses (hence infinite mean) are relevant.

We discuss in Section 2.4 the implications of (2.1) and related inequalities on the risk management decision of a single agent. It follows from (2.1) that the action of diversification increases the risk of ultra heavy-tailed Pareto losses *uniformly for all risk preferences*, such as VaR, expected utilities, and distortion risk measures, as long as the risk preferences are monotone and well defined. The increase of the portfolio risk is strict, and it provides an important implication in decision making: For an agent who faces iid Pareto losses without finite mean and aims to minimize their risk by choosing a position across these losses, the optimal decision is to take only one of the Pareto losses (i.e., no diversification).

We proceed to study equilibria of a risk exchange market for Pareto losses under a few different settings in Section 2.5. As individual agents do not benefit from diversification in a risk exchange market where iid Pareto losses without finite mean are present, we may expect that agents will not share their losses with each other. Indeed, if each agent in the market is associated with an initial position in one of these Pareto losses, the agents will merely exchange the entire loss positions instead of risk sharing in an equilibrium model (Theorem 2.3 (i)). The situation becomes quite different if the agents with initial losses are allowed to transfer their losses to external parties. If the external agents have a stronger risk tolerance, then it is possible that both the internal and external agents can benefit by transferring losses from the internal to the external agents (Theorem 2.4 (ii)). In Proposition 2.7, we show that agents prefer to share Pareto losses with finite mean among themselves; this is in sharp contrast to the case of Pareto losses without finite mean. The

above results are consistent with the observations made in Ibragimov et al. (2011) based on a different model.

In Section 2.6, numerical and real data examples are presented to illustrate the presence of ultra heavy tails in two real datasets, over which the phenomenon of the inequality (2.1) can be empirically observed, and to study the diversification effects of ultra heavy-tailed Pareto losses with different tail indices. Section 2.7 concludes the chapter. Some background on risk measures is put in Appendix 2.8.1, and proofs of all technical results are put in Appendix 2.8.2.

We fix some notations. Throughout, random variables are defined on an atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denote by $\mathbb{N}$ the set of all positive integers and $\mathbb{R}_+$ the set of non-negative real numbers. For $n \in \mathbb{N}$, let $[n] = \{1, \dots, n\}$. Denote by Δ_n the standard simplex, that is, $\Delta_n = \{(\theta_1, \dots, \theta_n) \in [0,1]^n : \sum_{i=1}^n \theta_i = 1\}$. For $x, y \in \mathbb{R}$, write $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$, and $x_+ = \max\{x, 0\}$. Recall that $X \simeq_{\mathrm{st}} Y$ means equality in distribution. We always assume $n \geqslant 2$.

2.1.1 Infinite-mean Pareto models

The key assumption of this chapter is that Pareto losses have infinite mean, hence are so-called ultra heavy-tailed. Whereas statistical models with some divergent higher moments are ubiquitous throughout the risk management literature, the infinite mean case needs more specific motivation. For power-tail data, a standard approach for the estimation of the underlying tail parameters is the Peaks Over Threshold (POT) methodology from Extreme Value Theory (EVT); see Embrechts et al. (1997). As we will discuss in Proposition 2.3 and Section 2.6.2, our results apply to the case of the generalized Pareto distribution which is the basic model for the POT set-up. Below we discuss some examples from the literature leading to ultra heavy-tailed Pareto models; extra data examples are provided in Section 2.6.2.[3]

In the parameterization used in Section 2.2, a tail parameter $\alpha \leqslant 1$ corresponds to an infinite-mean Pareto model. Ibragimov et al. (2009) used standard seismic theory to show that the tail indices α of earthquake losses lie in the range $[0.6, 1.5]$. Estimated by Rizzo (2009), the tail indices α for some wind catastrophic losses are around 0.7. Hofert and Wüthrich (2012) showed that the tail indices α of losses caused by nuclear power accidents are around $[0.6, 0.7]$; similar observations can be found in Sornette et al. (2013). Based on

[3] These examples show that, at least, we cannot exclude the possibility that infinite-mean models fit these datasets better than finite-mean models.

data collected by the Basel Committee on Banking Supervision, Moscadelli (2004) reported the tail indices α of (over 40000) operational losses in 8 different business lines to lie in the range $[0.7, 1.2]$, with 6 out of the 8 tail indices being less than 1, with 2 out of these 6 significantly less than 1 at a 95% confidence level. For a detailed discussion on the risk management consequences in this case, see Nešlehová et al. (2006). Losses from cyber risks have tail indices $\alpha \in [0.6, 0.7]$; see Eling and Wirfs (2019), Eling and Schnell (2020) and the references therein. In a standard Swiss Solvency Test document (FINMA (2021, p. 110)), most major damage insurance losses are modelled by a Pareto distribution with default parameter α in the range $[1, 2]$, with $\alpha = 1$ attained by some aircraft insurance. As discussed by Beirlant et al. (1999), some fire losses collected by the reinsurance broker AON Re Belgium have tail indices α around 1. Biffis and Chavez (2014) showed that a number of large commercial property losses collected from two Lloyd's syndicates have tail indices α considerably less than 1. Silverberg and Verspagen (2007) concluded that the tail indices α are less than 1 for financial returns from some technological innovations. Besides large financial losses and returns, numbers of deaths in major earthquakes and pandemics modelled by Pareto distributions also have infinite mean; see Clark (2013) and Cirillo and Taleb (2020). Heavy-tailed to ultra heavy-tailed models also occur in the realm of climate change and environmental economics. Weitzman (2009)'s Dismal Theorem discusses the break-down of standard economic thinking like cost-benefit analysis in this context. This led to an interesting discussion with William Nordhaus, a recipient of the 2018 Nobel Memorial Prize in Economic Sciences; see Nordhaus (2009).

The above references exemplify the occurrence of infinite mean models. Our perspective on these examples and discussions is that if these models are the result of some careful statistical analyses, then the practicing modeler has to take a step back and carefully reconsider the risk management consequences. Of course, in practice there are several methods available to avoid such ultra heavy-tailed models, like cutting off the loss distribution model at some specific level, or tapering (concatinating a light-tailed distribution far in the tail of the loss distribution). Our experience shows that in examples like those referred to above, such corrections often come at the cost of a great variability depending on the methodology used. It is in this context that our results add to the existing literature and modeling practice in cases where power-tail data play an important role.

2.2 Diversification of Pareto losses without finite mean

A common parameterization of Pareto distributions is given by, for $\theta, \alpha > 0$,

$$P_{\alpha,\theta}(x) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, \quad x \geqslant \theta.$$

Note that if $X \sim P_{\alpha,1}$, then $\theta X \sim P_{\alpha,\theta}$, and thus θ is a scale parameter. For $X \sim P_{\alpha,1}$, we write $X \sim \text{Pareto}(\alpha)$. Moreover, the mean of Pareto(α) is infinite if and only if the tail parameter $\alpha \in (0, 1]$. We say that the Pareto(α) distribution is *ultra heavy-tailed* if $\alpha \leqslant 1$, and it is *moderately heavy-tailed* if $\alpha > 1$.

Theorem 2.1. *Let $X, X_1, \dots, X_n$ be iid* Pareto(α) *random variables, $\alpha \in (0, 1]$. For $(\theta_1, \dots, \theta_n) \in \Delta_n$, we have*

$$X \leqslant_{\text{st}} \sum_{i=1}^{n} \theta_i X_i. \tag{2.2}$$

Moreover, for $t > 1$, $\mathbb{P}\left(\sum_{i=1}^{n} \theta_i X_i > t\right) > \mathbb{P}\left(X > t\right)$ if $\theta_i > 0$ for at least two $i \in [n]$.

Remark 2.1 (Generalized Pareto distributions)**.** The inequality (2.2) can be stated equivalently for other parameterizations of Pareto distributions without finite mean. For instance, it is often useful to consider generalized Pareto distributions, which provide an approximation for the excess losses beyond some high threshold. The generalized Pareto distribution for $\xi \geqslant 0$ is parametrized by

$$G_{\xi,\beta}(x) = 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}, \quad x \geqslant 0, \tag{2.3}$$

where $\xi \geqslant 0$ ($\xi = 0$ corresponds to an exponential distribution) and $\beta > 0$; see Embrechts et al. (1997). If $\xi \geqslant 1$, then $G_{\xi,\beta}$ does not have finite mean. For $\xi > 0$, a generalized Pareto distribution in (2.3) can be converted to $P_{1/\xi,1}$ through a location-scale transform. Therefore, (2.2) implies that for $\xi \geqslant 1$, $(\beta_1, \dots, \beta_n) \in (0, \infty)^n$ and independent random variables $Y_i \sim G_{\xi,\beta_i}$, $i \in [n]$, we have $Y \leqslant_{\text{st}} \sum_{i=1}^{n} Y_i$, where $Y \sim G_{\xi,\beta}$ with $\beta = \sum_{i=1}^{n} \beta_i$.

We will say that a *diversification penalty* exists if (2.2) holds, which is naturally interpreted as that having exposures in multiple iid ultra heavy-tailed Pareto losses is worse than having just one Pareto loss of the same total exposure. This observation will be generalized to a few other models later.

To better understand the result in Theorem 2.1, we stress that (2.2) cannot be expected if $X_1, \dots, X_n$ have finite mean, regardless of their dependence structure, as summarized in the following proposition.

Proposition 2.1. *For $\theta_1, \dots, \theta_n > 0$ with $\sum_{i=1}^n \theta_n = 1$ and identically distributed random variables $X, X_1, \dots, X_n$ with finite mean and any dependence structure, (2.2) holds if and only if $X_1 = \cdots = X_n$ almost surely.*

Proposition 2.1 implies, in particular, that (2.2) never holds for iid non-degenerate random variables $X, X_1, \dots, X_n$ with finite mean. As such, it seems that Theorem 2.1 yields a clear and elegant methodological distinction between the two modeling environments. Even if $X, X_1, \dots, X_n$ have an infinite mean, we are not aware of any other distributions in the literature for which (2.2) holds other than the ones in this chapter, all built on the basis of Theorem 2.1.

Remark 2.2. The inequality (2.2) also holds for some correlated ultra heavy-tailed Pareto risks. First, the inequality (2.2) simply holds for perfectly positively dependent ultra heavy-tailed Pareto risks (i.e., $X_1 = \cdots = X_n$ almost surely). Therefore, (2.2) remains true if the dependence structure (i.e., copula) of risks $X_1, \dots, X_n$ is a mixture of independence and perfectly positive dependence; see Nelsen (2006) for an introduction to copulas. Besides this specific type of positive dependence structure, the inequality (2.2) may also hold for other dependence structures, but a rigorous analysis is beyond the scope of this chapter.

Remark 2.3. An *ultra heavy-tailed Pareto sum* is a random variable $\sum_{j\in\mathbb{N}} \lambda_j Y_j$ where $Y_j \sim \mathrm{Pareto}(\alpha_j)$, $j \in \mathbb{N}$, are independent, $\alpha_j \in (0,1]$, $\lambda_j \in \mathbb{R}_+$, and $\sum_{j\in\mathbb{N}} \lambda_j < \infty$. The inequality (2.2) in Theorem 2.1 holds also for iid ultra heavy-tailed Pareto sums $X, X_1, \dots, X_n$, and this can be shown by applying Theorem 2.1 to iid copies of each Y_j.

For an equally weighted pool of k iid Pareto losses, it is interesting to see whether enlarging k increases the risk in first-order stochastic dominance, i.e., for iid $\mathrm{Pareto}(\alpha)$ random variables $X_1, \dots, X_\ell$, $\alpha \in (0,1]$, whether it holds that

$$\frac{1}{k}\sum_{i=1}^{k} X_i \leqslant_{\mathrm{st}} \frac{1}{\ell}\sum_{i=1}^{\ell} X_i \quad \text{for } k, \ell \in \mathbb{N} \text{ and } k \leqslant \ell. \tag{2.4}$$

The case of $k = 1$ in (2.4) corresponds to (2.2) with equal weights $\theta_1, \dots, \theta_n$. The inequality (2.4) means that the more we diversify ultra heavy-tailed Pareto losses, the higher the penalty. In the next result, we show this inequality for the case that ℓ is a multiple of k.

Proposition 2.2. *For $m, n \in \mathbb{N}$, let $X_1, \dots, X_{mn}$ be iid $\mathrm{Pareto}(\alpha)$ random variables, $\alpha \in (0,1]$. We have*

$$\frac{1}{m}\sum_{i=1}^{m} X_i \leqslant_{\mathrm{st}} \frac{1}{mn}\sum_{i=1}^{mn} X_i.$$

Based on our numerical results in Section 2.6.1, we conjecture that the inequality (2.4) is true also for the general case that ℓ is not a multiple of k.

Tail Pareto distributions

As reflected by the Pickands-Balkema-de Haan Theorem (see Theorem 3.4.13 (b) in Embrechts et al. (1997)), many losses have a power-like tail, but their distributions may not be power-like over the full support. Therefore, it is practically useful to assume that a random loss has a Pareto distribution only in the tail region; see the examples in Section 2.1.1. For $\alpha > 0$, we say that Y has a Pareto(α) distribution beyond $x \geqslant 1$ if $\mathbb{P}(Y > t) = t^{-\alpha}$ for $t \geqslant x$. Our next result suggests that, under an extra condition, stochastic dominance also holds in the tail region for such distributions.

Proposition 2.3. *Let $Y, Y_1, \dots, Y_n$ be iid random variables distributed as* Pareto(α) *beyond $x \geqslant 1$ and $\alpha \in (0, 1]$. Assume that $Y \geqslant_{\mathrm{st}} X \sim$ Pareto(α). For $(\theta_1, \dots, \theta_n) \in \Delta_n$ and $t \geqslant x$, we have $\mathbb{P}\left(\sum_{i=1}^n \theta_i Y_i > t\right) \geqslant \mathbb{P}\left(Y > t\right)$, and the inequality is strict if $t > 1$ and $\theta_i > 0$ for at least two $i \in [n]$.*

In Proposition 2.3, the assumption $Y \geqslant_{\mathrm{st}} X \sim$ Pareto(α), that is, $\mathbb{P}(Y > t) \leqslant t^{-\alpha}$ for $t \in [1, x]$, is not dispensable. Here we cannot allow the distribution of Y on $[1, x]$ to be arbitrary; the entire distribution is relevant in order to establish the inequality $\mathbb{P}\left(\sum_{i=1}^n \theta_i Y_i > t\right) \geqslant \mathbb{P}\left(Y > t\right)$, even for t in the tail region.

Let $X, X_1, \dots, X_n$ be iid Pareto(α) random variables with $\alpha \in (0, 1]$. As a particular application of Proposition 2.3, it holds that, for any $m \geqslant 1$,

$$X \vee m \leqslant_{\mathrm{st}} \sum_{i=1}^n \theta_i (X_i \vee m). \tag{2.5}$$

This inequality follows by noting that $X \vee m$ has a Pareto distribution beyond m and applying Proposition 2.3 to $t \geqslant m$. A location shift of (2.5) also gives

$$(X - m)_+ \leqslant_{\mathrm{st}} \sum_{i=1}^n \theta_i (X_i - m)_+. \tag{2.6}$$

For (2.5) and (2.6) to hold, it suffices to assume that $X_1, \dots, X_n$ are Pareto(α) beyond m, as their distribution on $(-\infty, m]$ does not matter.

A classic model in insurance

Theorem 2.1 can be easily generalized to include random weights and a random number of risks, which are for instance common in modeling portfolios of insurance losses; see Klugman et al. (2012). Let N be a counting random variable (i.e., it takes values in $\{0, 1, 2, \dots\}$), and W_i and X_i be positive random variables for $i \in \mathbb{N}$. We consider an insurance portfolio where each policy incurs a loss $W_i X_i$ if there is a claim, and N is the total number of claims in a given period of time. If $W_1 = W_2 = \cdots = 1$ and $X_1, X_2, \dots$ are iid, then this model recovers the classic collective risk model. The total loss of a portfolio of insurance policies is given by $\sum_{i=1}^{N} W_i X_i$, and its average loss across claims is $(\sum_{i=1}^{N} W_i X_i)/(\sum_{i=1}^{N} W_i)$ where both terms are 0 if $N = 0$.

Proposition 2.4. *Let $X, X_1, X_2, \dots$ be iid* Pareto(α) *random variables, $\alpha \in (0, 1]$, $W_1, W_2, \dots$ be positive random variables, and N be a counting random variable, such that $X, \{X_i\}_{i\in\mathbb{N}}$, $\{W_i\}_{i\in\mathbb{N}}$, and N are independent. We have*

$$X\mathbb{1}_{\{N\geqslant 1\}} \leqslant_{\mathrm{st}} \frac{\sum_{i=1}^{N} W_i X_i}{\sum_{i=1}^{N} W_i} \quad \textit{and} \quad \sum_{i=1}^{N} W_i X \leqslant_{\mathrm{st}} \sum_{i=1}^{N} W_i X_i. \tag{2.7}$$

If $\mathbb{P}(N \geqslant 2) \neq 0$, then for $t > 1$, $\mathbb{P}(\sum_{i=1}^{N} W_i X_i / \sum_{i=1}^{N} W_i \leqslant t) < \mathbb{P}(X\mathbb{1}_{\{N\geqslant 1\}} \leqslant t)$.

If $W_1 = W_2 = \cdots = 1$ as in the classic collective risk model, then, under the assumptions of Proposition 2.4, we have

$$X_1\mathbb{1}_{\{N\geqslant 1\}} \leqslant_{\mathrm{st}} \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and} \quad NX_1 \leqslant_{\mathrm{st}} \sum_{i=1}^{N} X_i.$$

To interpret the above inequalities, the average of a randomly counted sequence of iid Pareto(α) losses is stochastically larger than one member of the sequence if $\alpha \leqslant 1$. Therefore, building an insurance portfolio for iid ultra heavy-tailed Pareto claims does not reduce the total risk on average. In this setting, it is less risky to insure one large policy than to insure many independent policies of the same type of ultra heavy-tailed Pareto loss and thus the basic principle of insurance does not apply to ultra heavy-tailed Pareto losses.

2.3 A model for catastrophic losses

Catastrophic losses are large losses that usually occur with very small probabilities. It is practical to model an individual catastrophic loss as $X\mathbb{1}_A$, where A is the triggering event

of the loss such that $X\mathbb{1}_A$ is Pareto distributed conditional on A (hence, we can assume that X is Pareto distributed and independent of A). Let $A_1, \dots, A_n$ be the triggering events of independent Pareto losses $X_1, \dots, X_n \sim \text{Pareto}(\alpha)$, $\alpha \in (0, 1]$, such that $A_1, \dots, A_n$ are independent of the loss portfolio $(X_1, \dots, X_n)$. Let $(\theta_1, \dots, \theta_n) \in \mathbb{R}^n_+$ be the exposure vector. The total loss can then be written as $\theta_1 X_1 \mathbb{1}_{A_1} + \cdots + \theta_n X_n \mathbb{1}_{A_n}$. If $A_1 = \cdots = A_n$, meaning that $X_1, \dots, X_n$ represent different losses caused by the same catastrophic event, then, by Theorem 2.1, for $\lambda = \sum_{i=1}^n \theta_i > 0$,

$$X_1 \mathbb{1}_{A_1} \leqslant_{\text{st}} \frac{1}{\lambda} \sum_{i=1}^{n} \theta_i X_i \mathbb{1}_{A_i}. \tag{2.8}$$

Hence, diversification of losses from the same catastrophe *increases* the portfolio risk, and thus there is a diversification penalty. It remains to investigate whether a diversification penalty exists in this model (i.e., (2.8) holds) if $A_1, \dots, A_n$ are different, meaning that $X_1, \dots, X_n$ may represent losses caused by different catastrophic events. Diversification has two competing effects on the loss portfolio: It increases the frequency of having losses and decreases the sizes of the individual losses.

To illustrate the above trade-off, we first look at the diversification of two ultra heavy-tailed Pareto losses. Let X_1, X_2 be iid Pareto(α) random variables, $\alpha \in (0, 1]$, and A_1, A_2 be any events independent of (X_1, X_2). For simplicity, we assume that $(\theta_1, \theta_2) = (1/2, 1/2)$, and $\mathbb{P}(A_1) = \mathbb{P}(A_2)$. We have

$$\begin{aligned}
\frac{1}{2} X_1 \mathbb{1}_{A_1} + \frac{1}{2} X_2 \mathbb{1}_{A_2} &= \frac{1}{2}(X_1 + X_2)\mathbb{1}_{A_1 \cap A_2} + \frac{1}{2} X_1 \mathbb{1}_{A_1 \cap A_2^c} + \frac{1}{2} X_2 \mathbb{1}_{A_1^c \cap A_2} \\
&\simeq_{\text{st}} \frac{1}{2}(X_1 + X_2)\mathbb{1}_{A_1 \cap A_2} + \frac{1}{2} X_1 \mathbb{1}_{(A_1 \cap A_2^c) \cup (A_1^c \cap A_2)} \\
&\geqslant_{\text{st}} X_1 \mathbb{1}_{A_1 \cap A_2} + \frac{1}{2} X_1 \mathbb{1}_{(A_1 \cap A_2^c) \cup (A_1^c \cap A_2)},
\end{aligned}$$

where the second-last equality holds as $A_1 \cap A_2^c$ and $A_1^c \cap A_2$ are mutually exclusive and $X_1 \simeq_{\text{st}} X_2$, and the last inequality uses $\frac{1}{2}(X_1 + X_2)\mathbb{1}_{A_1 \cap A_2} \geqslant_{\text{st}} X_1 \mathbb{1}_{A_1 \cap A_2}$ which follows by combining Theorem 2.1 and Theorem 1.A.14 of Shaked and Shanthikumar (2007). Write

$$X_1 \mathbb{1}_{A_1} = X_1 \mathbb{1}_{A_1 \cap A_2} + X_1 \mathbb{1}_{A_1 \cap A_2^c}.$$

Therefore, whether (2.8) holds in this setting boils down to whether

$$X_1 \mathbb{1}_{A_1 \cap A_2^c} \leqslant_{\text{st}} \frac{1}{2} X_1 \mathbb{1}_{(A_1 \cap A_2^c) \cup (A_1^c \cap A_2)} \tag{2.9}$$

holds. As $\mathbb{P}(A_1)=\mathbb{P}(A_2)$, $\mathbb{P}((A_1\cap A_2^c)\cup(A_1^c\cap A_2))=2\mathbb{P}(A_1\cap A_2^c)$. We write $p=\mathbb{P}(A_1\cap A_2^c)$. We can directly compute, for $t\geqslant 0$,

$$\mathbb{P}(X_1\mathbb{1}_{A_1\cap A_2^c}>t)=p(t^{-\alpha}\wedge 1)\quad\text{and}\quad\mathbb{P}\left(\frac{1}{2}X_1\mathbb{1}_{(A_1\cap A_2^c)\cup(A_1^c\cap A_2)}>t\right)=(2p)((2t)^{-\alpha}\wedge 1).$$

Since $2((2t)^{-\alpha}\wedge 1)=2^{1-\alpha}(t^{-\alpha}\wedge 2^{\alpha})\geqslant(t^{-\alpha}\wedge 1)$, we obtain (2.9). Hence, diversification of two ultra heavy-tailed Pareto losses increases the portfolio risk if the two losses are triggered with the same probability. Theorem 2.2 provides a general result for diversifying any number of ultra heavy-tailed Pareto losses triggered with (possibly) different probabilities. To establish Theorem 2.2, we need the following lemma, which itself has a nice interpretation.

Lemma 2.1. *Let $X\sim\mathrm{Pareto}(\alpha)$, $\alpha\in(0,1]$, and $B_1,\dots,B_n$ be mutually exclusive events independent of X. For $(c_1,\dots,c_n)\in[0,1]^n$, we have*

$$X\mathbb{1}_A\leqslant_{\mathrm{st}}\sum_{i=1}^n c_iX\mathbb{1}_{B_i},$$

where A is an event independent of X satisfying $\mathbb{P}(A)=\sum_{i=1}^n c_i\mathbb{P}(B_i)$.

Lemma 2.1 implies $X\mathbb{1}_A\leqslant_{\mathrm{st}}cX\mathbb{1}_B$, where $\mathbb{P}(A)=c\mathbb{P}(B)$ and $c\in(0,1]$. This implies that if we decrease the size of an ultra heavy-tailed Pareto loss (i.e., multiply X by c) and increase the probability of having the loss (i.e., divide $\mathbb{P}(A)$ by c), the loss becomes larger in first-order stochastic dominance. In general, the stochastic dominance cannot hold if X is a moderately heavy-tailed Pareto loss (i.e., X has a finite mean). For a moderately heavy-tailed Pareto loss X, $\mathbb{E}[cX\mathbb{1}_B]=\mathbb{E}[X\mathbb{1}_A]$. If, in addition, $X\mathbb{1}_A\leqslant_{\mathrm{st}}cX\mathbb{1}_B$ holds, then one has $X\mathbb{1}_A\simeq_{\mathrm{st}}cX\mathbb{1}_B$ (Theorem 1.A.8 of Shaked and Shanthikumar (2007)), which does not hold unless $c=1$. The above observation of ultra heavy-tailed Pareto losses consequently leads to Theorem 2.2.

Theorem 2.2. *Let $X_1,\dots,X_n$ be iid $\mathrm{Pareto}(\alpha)$ random variables, $\alpha\in(0,1]$, and $A_1,\dots,A_n$ be any events independent of $(X_1,\dots,X_n)$. For $(\theta_1,\dots,\theta_n)\in\mathbb{R}_+^n$, we have*

$$\lambda X\mathbb{1}_A\leqslant_{\mathrm{st}}\sum_{i=1}^n\theta_iX_i\mathbb{1}_{A_i},\tag{2.10}$$

where $\lambda\geqslant\sum_{i=1}^n\theta_i$, $X\sim\mathrm{Pareto}(\alpha)$, and A is independent of X satisfying $\lambda\mathbb{P}(A)=\sum_{i=1}^n\theta_i\mathbb{P}(A_i)$.

Remark 2.4. By setting $\mathbb{P}(A_1) = \cdots = \mathbb{P}(A_n) = 1$, $(\theta_1, \dots, \theta_n) \in \Delta_n$ and $\lambda = 1$, Theorem 2.2 recovers the inequality (2.2) in Theorem 2.1. Moreover, a strict inequality

$$\mathbb{P}\left(\sum_{i=1}^{n} \theta_i X_i \mathbb{1}_{A_i} > t\right) > \mathbb{P}\left(\lambda X \mathbb{1}_A > t\right) \tag{2.11}$$

similar to Theorem 2.1 can be expected. A sufficient condition can be obtained using the strict inequality in Theorem 2.1: If there exists $S \subseteq [n]$ with at least two elements such that $\theta_i > 0$ for $i \in S$ and $\mathbb{P}(B_S) > 0$ where $B_S = \left(\bigcap_{i\in S} A_i\right) \cap \left(\bigcap_{i\in S^c} A_i^c\right)$, then (2.11) holds for $t > \sum_{i\in S} \theta_i$.

We discuss a special case of Theorem 2.2, which has practical relevance in risk sharing. Let $\mathbb{P}(A_1) = \cdots = \mathbb{P}(A_n)$, $X = X_1$, $A = A_1$, and $(\theta_1, \dots, \theta_n) \in \Delta_n$. The inequality (2.10) can be rewritten as

$$X_1 \mathbb{1}_{A_1} \leqslant_{\mathrm{st}} \sum_{i=1}^{n} \theta_i X_i \mathbb{1}_{A_i}. \tag{2.12}$$

The left-hand side of (2.12) can be regarded as the loss of an agent who keeps their own risk, and the right-hand side of (2.12) is the loss of an agent who shares risks with other agents. By pooling among ultra heavy-tailed Pareto losses, triggered by (possibly) different catastrophes, agents expect to suffer less loss when their own catastrophic loss occurs. However, every agent in the pool will have a higher frequency of bearing losses. Theorem 2.2 shows that the combined effects of diversification of ultra heavy-tailed Pareto losses lead to a higher probability of default at any capital reserve level, i.e., $\mathbb{P}(\sum_{i=1}^{n} \theta_i X_i \mathbb{1}_{A_i} > t) \geqslant \mathbb{P}(X_1 \mathbb{1}_{A_1} > t)$ for all $t > 0$.

2.4 Risk management decisions of a single agent

As hinted by (2.12) in Section 2.3, in a model of catastrophic losses $(X_1, \dots, X_n)$ and triggering events $(A_1, \dots, A_n)$, an agent who can choose between keeping their own risk or sharing risk with other agents has no incentive to enter the risk sharing pool, because it will increase their total risk. In this section, we make this observation rigorous by formally considering risk preference models.

Some further notation will be useful. Let $\mathcal{X}$ be the set of all random variables, and let $L^1 \subseteq \mathcal{X}$ be the set of random variables with finite mean. For $X \in \mathcal{X}$, denote by F_X the distribution function. Denote by F_X^{-1} the (left) quantile function of X, that is,

$$F_X^{-1}(p) = \inf\{t \in \mathbb{R} : F_X(t) \geqslant p\}, \quad p \in (0, 1].$$

For vectors $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$, their dot product is $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$ and we denote by $\|\mathbf{x}\| = \sum_{i=1}^n |x_i|$.

Measuring the risk of a financial portfolio is a crucial task in both the banking and insurance sectors, and it is typically done by calculating the value of a risk measure which maps the portfolio loss to a real number. A *risk measure* is a functional $\rho : \mathcal{X}_\rho \to \overline{\mathbb{R}} := [-\infty, \infty]$, where the domain $\mathcal{X}_\rho \subseteq \mathcal{X}$ is a set of random variables representing financial losses. We will assume that an agent uses a risk measure ρ for their preference, in the sense that the agent prefers a smaller value of ρ. Our notion of a risk measure is quite broad, and it includes not only risk measures in the sense of Artzner et al. (1999) and Föllmer and Schied (2016) but also decision models such as the expected utility by flipping the sign. However, we need to be clear that most classic expected utility models or convex risk measures (see Appendix 2.8.1) in the literature are not suitable for our setting, because the ultra heavy-tailed Pareto losses do not have a finite mean, and most expected utility functions and convex risk measures will take infinite values when evaluating these losses. Nevertheless, we will soon see that there are still many useful examples of risk measures conforming with our setting.

To interpret our main results, we only need minimal assumptions of monotonicity on ρ, in the following two forms.

(a) Weak monotonicity: $\rho(X) \leqslant \rho(Y)$ for $X, Y \in \mathcal{X}_\rho$ if $X \leqslant_{\mathrm{st}} Y$.

(b) Mild monotonicity: ρ is weakly monotone and $\rho(X) < \rho(Y)$ if $F_X^{-1} < F_Y^{-1}$ on $(0, 1)$.

Each of weak and mild monotonicity implies that $\rho(X) = \rho(Y)$ holds for $X \simeq_{\mathrm{st}} Y$. Common examples of preference models are all mildly monotone; we highlight some examples. First, for an increasing utility function u, the expected utility agent can be represented by a risk measure E_v, namely

$$E_v(X) = \mathbb{E}[v(X)], \qquad X \in \mathcal{X}_{E_v} := \{Y \in \mathcal{X} : \mathbb{E}[|v(Y)|] < \infty\},$$

where $v(x) = -u(-x)$ is also increasing. It is clear that E_v is mildly monotone if v or u is strictly increasing. The next examples are the two widely used regulatory risk measures in insurance and finance, Value-at-Risk (VaR) and Expected Shortfall (ES). For $X \in \mathcal{X}$ and $p \in (0, 1)$, VaR is defined as

$$\mathrm{VaR}_p(X) = F_X^{-1}(p) = \inf\{t \in \mathbb{R} : F_X(t) \geqslant p\}, \tag{2.13}$$

and ES is defined as

$$\mathrm{ES}_p(X) = \frac{1}{1-p} \int_p^1 \mathrm{VaR}_u(X) \mathrm{d}u.$$

For $X \notin L^1$, such as the ultra heavy-tailed Pareto losses, $\mathrm{ES}_p(X)$ can be ∞, whereas $\mathrm{VaR}_p(X)$ is always finite. VaR is mildly monotone on $\mathcal{X}$, whereas ES is mildly monotone only on L^1.

In Theorems 2.1 and 2.2, we have established a diversification penalty for two models, which we will denote by $\mathbf{Y} = (Y_1, \dots, Y_n)$. In both models A and B below, let $X, X_1, \dots, X_n$ be iid Pareto(α) random variables, $\alpha \in (0, 1]$, and $(\theta_1, \dots, \theta_n) \in \Delta_n$.

A. $Y_i = X_i$, $i \in [n]$ and $Y = X$.

B. $Y_i = X_i \mathbb{1}_{A_i}$, $i \in [n]$ and $Y = X\mathbb{1}_A$, where $A_1, \dots, A_n$ are any events independent of $(X_1, ..., X_n)$, and A is independent of X and satisfies $\mathbb{P}(A) = \sum_{i=1}^n \theta_i \mathbb{P}(A_i)$.

From now on, we will assume that $\mathcal{X}_\rho$ contains the random variables in models A and B (this puts some restrictions on v for E_v since $\mathbb{E}[X] = \infty$). The following result on the diversification penalty of ultra heavy-tailed Pareto losses for a monotone agent follows directly from Theorems 2.1 and 2.2.

Proposition 2.5. *For $(\theta_1, \dots, \theta_n) \in \Delta_n$ and a weakly monotone risk measure $\rho : \mathcal{X}_\rho \to \overline{\mathbb{R}}$, for both models A and B, we have*

$$\rho\left(\sum_{i=1}^n \theta_i Y_i\right) \geqslant \rho(Y). \tag{2.14}$$

The inequality in (2.14) is strict for model A if ρ is mildly monotone and $\theta_i > 0$ for at least two $i \in [n]$.

We distinguish strict and non-strict inequalities in (2.14) because a strict inequality has stronger implications on the optimal decision of an agent. As an important consequence of Proposition 2.5, for $p \in (0, 1)$ and $(\theta_1, \dots, \theta_n) \in \Delta^n$, in models A and B,

$$\mathrm{VaR}_p\left(\sum_{i=1}^n \theta_i Y_i\right) \geqslant \mathrm{VaR}_p(Y), \tag{2.15}$$

and if $\theta_i > 0$ for at least two $i \in [n]$, then, in model A,

$$\mathrm{VaR}_p\left(\sum_{i=1}^n \theta_i Y_i\right) > \sum_{i=1}^n \theta_i \mathrm{VaR}_p(Y_i). \tag{2.16}$$

The inequality (2.16) and its non-strict version will be referred to as diversification penalty for VaR_p.

Remark 2.5. Diversification penalty for VaR_p also holds for other models that we consider. For instance, by Proposition 2.3, if $Y, Y_1, \dots, Y_n$ are iid Pareto(α) beyond $x \geqslant 1$ and $Y \geqslant_{\mathrm{st}} X \sim \mathrm{Pareto}(\alpha)$, then inequalities (2.15) and (2.16) hold for $p \geqslant 1 - x^{-\alpha}$.

From now on, we will focus on model A as it allows us to have a simple interpretation of the diversification penalty as in (2.16). Since all commonly used preference models are mildly monotone, Proposition 2.5 suggests that diversification of ultra heavy-tailed Pareto losses is detrimental for the agent.

Proposition 2.5 implies the following optimal decision for an agent in a market where several iid ultra heavy-tailed Pareto losses are present. Suppose that the agent needs to decide on a position $\mathbf{w} \in \mathbb{R}^n_+$ across these losses to minimize the total risk. The agent faces a total loss $\mathbf{w} \cdot \mathbf{Y} - g(\|\mathbf{w}\|)$ where the function g represents a compensation that depends on $\mathbf{w}$ through $\|\mathbf{w}\|$, and $\mathbf{Y}$ is as in model A or B. The agent's optimization problem then becomes

$$\text{to minimize} \quad \rho\left(\mathbf{w} \cdot \mathbf{Y} - g(\|\mathbf{w}\|)\right) \quad \text{subject to } \mathbf{w} \in \mathbb{R}^n_+ \text{ and } \|\mathbf{w}\| = w \text{ with given } w > 0, \tag{2.17}$$

or

$$\text{to minimize} \quad \rho\left(\mathbf{w} \cdot \mathbf{Y} - g(\|\mathbf{w}\|)\right) \quad \text{subject to } \mathbf{w} \in \mathbb{R}^n_+. \tag{2.18}$$

For $i \in [n]$, let $\mathbf{e}_{i,n}$ be the ith column vector of the $n \times n$ identity matrix, and $E_w = \{w\mathbf{e}_{i,n} : i \in [n]\}$ for $w \geqslant 0$, which represents the positions of only taking one loss with exposure w.

Proposition 2.6. *Let $\rho : \mathcal{X}_\rho \to \overline{\mathbb{R}}$ be weakly monotone and $g : \mathbb{R} \to \mathbb{R}$.*

(i) For model A, if ρ is mildly monotone, then the set of minimizers of (2.17) is E_w, and that of of (2.18) is contained in $\bigcup_{w \in \mathbb{R}_+} E_w$.

(ii) For models A and B, if (2.17) has an optimizer, then it has an optimizer in E_w; if (2.18) has an optimizer, then it has an optimizer in $\bigcup_{w \in \mathbb{R}_+} E_w$.

Remarkably, there are almost no restrictions on ρ and g in Proposition 2.6 other than monotonicity of ρ, and hence this result can be applied to many economic decision models.

Remark 2.6. Since ES_p is ∞ for the losses in models A and B, Proposition 2.6 applied to ES gives the trivial statement that every position has infinite risk. The main context of application for Proposition 2.6 should be risk measures which are finite for losses in models A and B, such as VaR, E_v with some sublinear v, and Range Value-at-Risk (RVaR); see Appendix 2.8.1 for the definition of RVaR.

A model of excess-of-loss reinsurance coverage

Next, we assume the agent is an insurance company. In practice, insurers seek reinsurance coverage to transfer their losses. One of the most popular catastrophe reinsurance coverages is the excess-of-loss coverage; see OECD (2018). Therefore, it is interesting to consider heavy-tailed losses bounded at some thresholds. Catastrophe excess-of-loss coverage can be provided on per-loss or aggregate basis. We will see that the result in Proposition 2.5 holds if the excess-of-loss coverage is provided on either per-loss basis with high thresholds or aggregate basis.

We first discuss the case that the excess-of-loss coverage is provided on a per-loss basis, where non-diversification traps may exist for insurers; see Ibragimov et al. (2009). For $X_1, \dots, X_n \sim \mathrm{Pareto}(\alpha)$, $\alpha \in (0,1]$, take $Y_i = X_i \wedge c_i$, where $c_i > 1$ is the threshold, $i = 1, \dots, n$. Note that each Y_i is bounded. Since Y_i has a finite mean, we cannot expect (2.15) or (2.16) to hold for all $p \in (0,1)$. Nevertheless, we will see below that for a given p and large $c_1, \dots, c_n$, (2.16) holds, and thus there exists a diversification penalty for VaR_p.

For $p \in (0,1)$ and $(\theta_1, \dots, \theta_n) \in \Delta_n$, take $c_i \geqslant \mathrm{VaR}_p(\sum_{i=1}^n \theta_i X_i)/\theta_i$ for $i \in [n]$. Given that $X_i \geqslant c_i$ for $i \in [n]$, the distribution of X_i does not contribute to the calculation of $\mathrm{VaR}_p(\sum_{i=1}^n \theta_i X_i)$, and we have $\mathrm{VaR}_p(\sum_{i=1}^n \theta_i Y_i) = \mathrm{VaR}_p(\sum_{i=1}^n \theta_i X_i)$. Therefore, (2.16) holds for this choice of p and $(c_1, \dots, c_n)$. Hence, a diversification penalty for VaR_p exists for a fixed p if the thresholds $c_1, \dots, c_n$ are high enough.

If the excess-of-loss coverage is provided on an aggregate basis, then stochastic dominance holds as $X_1 \wedge c \leqslant_{\mathrm{st}} (\sum_{i=1}^n \theta_i X_i) \wedge c$ where $c > 1$ is the threshold; indeed the inequality is preserved under a monotone transform. Hence, for any weakly monotone risk measure $\rho : \mathcal{X} \to \mathbb{R}$, we have $\rho(X_1 \wedge c) \leqslant \rho((\sum_{i=1}^n \theta_i X_i) \wedge c)$, and a diversification penalty exists for ρ. Unlike the situation of model A in Proposition 2.5, strict inequality may not hold for $\rho = \mathrm{VaR}_p$ because $X_1 \wedge c$ and $(\sum_{i=1}^n \theta_i X_i) \wedge c$ have the same p-quantile c for large p. Nevertheless, for the expected utility preference E_v, we have

$$\mathbb{E}[v(X_1 \wedge c)] < \mathbb{E}[v((\theta_1 X_1 + \cdots + \theta_n X_n) \wedge c)],$$

for $c > 1$ and v strictly increasing on $[1, c]$. This is because E_v is strictly monotone in the sense that for $X \leqslant_{\mathrm{st}} Y$ taking values in $[1, c]$ and $X \not\simeq_{\mathrm{st}} Y$, we have $E_v(X) < E_v(Y)$.

Remark 2.7. If the minimum in the above discussion is replaced by a maximum, then stochastic dominance holds, as discussed in (2.5) and (2.6).

2.5 Equilibrium analysis in a risk exchange economy

2.5.1 The Pareto risk sharing market model

Suppose that there are $n \geqslant 2$ agents in a risk exchange market. Let $\mathbf{X} = (X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are iid Pareto(α) random variables with $\alpha > 0$. The ith agent faces a loss $a_i X_i$, where $a_i > 0$ is the initial exposure. In other words, the initial exposure vector of agent i is $\mathbf{a}^i = a_i \mathbf{e}_{i,n}$, and the corresponding loss can be written as $\mathbf{a}^i \cdot \mathbf{X} = a_i X_i$.

In a risk exchange market, each agent decides whether and how to share the losses with the other agents. For $i \in [n]$, let $p_i \geqslant 0$ be the premium (or compensation) for one unit of loss X_i; that is, if an agent takes $b \geqslant 0$ units of loss X_i, it receives the premium bp_i, which is linear in b. Denote by $\mathbf{p} = (p_1, \dots, p_n) \in \mathbb{R}_+^n$ the (endogenously generated) premium vector. Let $\mathbf{w}^i \in \mathbb{R}_+^n$ be the exposure vector of the ith agent from $\mathbf{X}$ after risk sharing. Then the loss of agent $i \in [n]$ after risk sharing is

$$L_i(\mathbf{w}^i, \mathbf{p}) = \mathbf{w}^i \cdot \mathbf{X} - (\mathbf{w}^i - \mathbf{a}^i) \cdot \mathbf{p}.$$

For each $i \in [n]$, assume that agent i is equipped with a risk measure $\rho_i : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ contains the convex cone generated by $\{\mathbf{X}\} \cup \mathbb{R}^n$. Moreover, there is a cost associated with taking a total risk position $\|\mathbf{w}^i\|$ different from the initial total exposure $\|\mathbf{a}^i\|$. The cost is modelled by $c_i(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|)$, where c_i is a non-negative convex function satisfying $c_i(0) = 0$. Some examples of c_i are $c_i(x) = 0$ (no cost), $c_i(x) = \lambda_i |x|$ (linear cost), $c_i(x) = \lambda_i x^2$ (quadratic cost), and $c_i(x) = \lambda_i x_+$ (cost only for excess risk taking), where $\lambda_i > 0$. We denote by $c_{i-}'(x)$ and $c_{i+}'(x)$ the left and right derivatives of c_i at $x \in \mathbb{R}$, respectively.

The above setting is called a *Pareto risk sharing market*. In this risk sharing market, the goal of each agent is to choose an exposure vector so that their own risk is minimized, i.e., minimizing $\rho_i(L_i(\mathbf{w}^i, \mathbf{p})) + c_i(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|)$ over $\mathbf{w}^i \in \mathbb{R}_+^n$, $i \in [n]$. An *equilibrium* of the market is a tuple $(\mathbf{p}^*, \mathbf{w}^{1*}, \dots, \mathbf{w}^{n*}) \in (\mathbb{R}_+^n)^{n+1}$ if the following two conditions are satisfied.

(a) Individual optimality:

$$\mathbf{w}^{i*} \in \operatorname*{arg\,min}_{\mathbf{w}^i \in \mathbb{R}_+^n} \left\{ \rho_i \left(L_i(\mathbf{w}^i, \mathbf{p}^*) \right) + c_i(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|) \right\}, \quad \text{for each } i \in [n]. \tag{2.19}$$

(b) Market clearance:

$$\sum_{i=1}^n \mathbf{w}^{i*} = \sum_{i=1}^n \mathbf{a}^i. \tag{2.20}$$

In this case, the vector $\mathbf{p}^*$ is an *equilibrium price*, and $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ is an *equilibrium allocation*.

Some of our results rely on a popular class of risk measures, many of which can be applied to ultra heavy-tailed Pareto losses. A *distortion risk measure* is defined as $\rho : \mathcal{X}_\rho \to \mathbb{R}$, via

$$\rho(Y) = \int_{-\infty}^{0} (h(\mathbb{P}(Y > x)) - 1)\mathrm{d}x + \int_{0}^{\infty} h(\mathbb{P}(Y > x))\mathrm{d}x, \tag{2.21}$$

where $h : [0,1] \to [0,1]$, called the *distortion function*, is a nondecreasing function with $h(0) = 0$ and $h(1) = 1$. The distortion risk measure ρ, up to sign change, coincides with the *dual utility* of Yaari (1987) in decision theory. As a class of risk measures, it includes VaR, ES, and RVaR (see Appendix 2.8.1) as special cases, and almost all distortion risk measures are mildly monotone (see Proposition 2.8). We assume that $\mathcal{X}_\rho$ contains the convex cone generated by $\{\mathbf{X}\} \cup \mathbb{R}^n$; this always holds in case ρ is VaR or RVaR, and it holds for ρ being ES if $\alpha > 1$.

2.5.2 No risk exchange for ultra heavy-tailed Pareto losses

As anticipated from Proposition 2.6, each agent's optimal strategy is not to share with the other agents if their risk measure is mildly monotone and the Pareto losses are ultra heavy-tailed. This observation is made rigorous in the following result, where we obtain a necessary condition for all possible equilibria in the market, as well as two different conditions in the case of distortion risk measures. As before, let X be a generic Pareto(α) random variable.

Theorem 2.3. *In a Pareto risk sharing market, suppose that $\alpha \in (0,1]$, and $\rho_1, \dots, \rho_n$ are mildly monotone.*

(i) All equilibria $(\mathbf{p}^*, \mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ (if they exist) satisfy $\mathbf{p}^* = (p, \dots, p)$ for some $p \in \mathbb{R}_+$ and $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ is an n-permutation of $(\mathbf{a}^1, \dots, \mathbf{a}^n)$.

(ii) Suppose that $\rho_1, \dots, \rho_n$ are distortion risk measures on $\mathcal{X}$. If p satisfies

$$c'_{i+}(0) \geqslant p - \rho_i(X) \geqslant c'_{i-}(0) \quad \text{for } i \in [n], \tag{2.22}$$

then the tuple $((p, \dots, p), \mathbf{a}^1, \dots, \mathbf{a}^n)$ is an equilibrium.

(iii) Suppose that $\rho_1, \dots, \rho_n$ are distortion risk measures on $\mathcal{X}$. If $(p, \dots, p)$ is an equilibrium price, then

$$\max_{j\in[n]} c_{i+}'(a_j - a_i) \geqslant p - \rho_i(X) \geqslant \min_{j\in[n]} c_{i-}'(a_j - a_i) \quad \text{for } i \in [n]. \tag{2.23}$$

Theorem 2.3 (i) states that, even if there is some risk exchange in an equilibrium, the agents merely exchange positions entirely instead of sharing a pool. This observation is consistent with Theorem 2.1, which says that diversification among multiple ultra heavy-tailed Pareto losses increases risk in a uniform sense. As there is no diversification in the optimal allocation for each agent, taking any of these iid losses is equivalent for the agent, and the equilibrium price should be identical across losses. Part (ii) suggests that if c_i has a kink at 0, i.e., $c_i'(0+) > 0 > c_i'(0-)$, then p can be an equilibrium price if it is very close to $\rho_i(X)$ in the sense of (2.22). Conversely, in part (iii), if p is an equilibrium price, then it needs to be close to $\rho_i(X)$ for $i \in [n]$ in the sense of (2.23). This observation is quite intuitive because by (i), the agents will not share losses but rather keep one of them in an equilibrium. If the price of taking one unit of the loss is too far away from an agent's assessment of the loss, it may have an incentive to move away, and the equilibrium is broken.

As a general message, the equilibrium price p should be very close to the individual risk assessments, and hence the risk sharing mechanism does not benefit the agents. Indeed, in (ii), the equilibrium allocation is equal to the original exposure, and there is no welfare gain. We will see later in Section 2.5.3 that in the presence of an external market, the picture is drastically different: the agents will benefit from transferring some losses to an external market.

In general, (2.22) and (2.23) are not equivalent, but in the two cases below, they are.

(a) $a_1 = \cdots = a_n$;

(b) $c_1 = \cdots = c_n = 0$.

In either case, both (2.22) and (2.23) are a necessary and sufficient condition for $(p, \dots, p)$ to be an equilibrium price. Hence, the tuple $(\mathbf{p}^*, \mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ is an equilibrium if and only if (2.22) holds and $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ is an n-permutation of $(\mathbf{a}^1, \dots, \mathbf{a}^n)$, which can be checked by Theorem 2.3 (i). In case (a), p cannot be too far away from $\rho_i(X)$ for each $i \in [n]$. In case (b), $p = \rho_1(X) = \cdots = \rho_n(X)$, and an equilibrium can only be achieved when all agents agree on the risk of one unit of the loss and use this assessment for pricing.

Although the agents will not benefit from sharing ultra heavy-tailed Pareto losses, the situation becomes different if these Pareto losses are moderately heavy-tailed, which will be discussed in Section 2.5.4.

Example 2.1 (Equilibrium for VaR agents with no costs)**.** Suppose that $c_i = 0$ for $i \in [n]$. Let $\rho_i = \mathrm{VaR}_q$, $q \in (0,1)$, $i \in [n]$. The tuple $(\mathbf{p}^*, \mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ is an equilibrium where $\mathbf{p}^* = ((1-q)^{-1/\alpha}, \dots, (1-q)^{-1/\alpha})$, and $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ is an n-permutation of $(\mathbf{a}^1, \dots, \mathbf{a}^n)$. For $i \in [n]$, $\rho_i\left(L_i(\mathbf{w}^{i*}, \mathbf{p}^*)\right) = \mathrm{VaR}_q(a_i X) = a_i(1-q)^{-1/\alpha}$.

Remark 2.8. We offer a few further technical remarks on Theorem 2.3.

1. Theorem 2.3 (ii) and (iii) remain valid for all mildly monotone, translation invariant, and positively homogeneous risk measures (see Appendix 2.8.1 for properties of risk measures).

2. If the range of $\mathbf{w}^i = (w_1^i, \dots, w_n^i)$ in (2.19) is constrained to $0 \leqslant w_j^i \leqslant a_j$ for $j \in [n]$, then $((p, \dots, p), \mathbf{a}^1, \dots, \mathbf{a}^n)$ in Theorem 2.3 (ii) is still an equilibrium under the condition (2.22). However, the characterization statement in (i) is no longer guaranteed, which can be seen from the proof of Theorem 2.3 in Appendix 2.8.2. As a result, (iii) cannot be obtained either.

3. The Pareto risk sharing market is closely related to model A in Section 2.4. Since model B has similar properties to model A in Proposition 2.6, we can check that the equilibrium in Theorem 2.3 (ii) still holds if we replace model A by model B, where the triggering events have the same probability of occurrence (i.e., $\mathbb{P}(A_1) = \cdots = \mathbb{P}(A_n)$). However, we cannot guarantee that all equilibria for model B have the form in (i) since holding one of the ultra heavy-tailed Pareto risks may not be the only optimal strategy for agents in model B; see Proposition 2.6.

2.5.3 A market with external risk transfer

In the setting of Section 2.5.2, we have considered a risk exchange within the group of n agents, each of which has an initial loss. Next, we consider an extended market with external agents to which risk can be transferred with compensation from the internal agents.

As we have seen from Theorem 2.3, agents cannot reduce their risks by sharing ultra heavy-tailed losses within the group. As such, they may seek to transfer their risks to other parties external to the group. In this context, the internal agents are risk bearers, and the

external agents are institutional investors without initial position of ultra heavy-tailed Pareto losses.

Consider a Pareto risk sharing market with n internal agents and $m \geqslant 1$ external agents equipped with the same risk measure $\rho_E : \mathcal{X} \to \mathbb{R}$. Let $\mathbf{u}^j \in \mathbb{R}^n_+$ be the exposure vector of the jth external agent after sharing the risks of the internal agents, $j \in [m]$. For the jth external agent, the loss for taking position $\mathbf{u}^j$ is

$$L_E(\mathbf{u}^j, \mathbf{p}) = \mathbf{u}^j \cdot \mathbf{X} - \mathbf{u}^j \cdot \mathbf{p},$$

where $\mathbf{p} = (p_1, \dots, p_n)$ is the premium vector. Like the internal agents, the goal of the external agents is to minimize their risk plus cost. That is, for $j \in [m]$, external agent j minimizes $\rho_E\left(L_E(\mathbf{u}^j, \mathbf{p})\right) + c_E(\|\mathbf{u}^j\|)$, where c_E is a non-negative cost function satisfying $c_E(0) = 0$.

For tractability, we will also make some simplifying assumptions on the internal agents. We assume that the internal agents have the same risk measure ρ_I and the same cost function c_I. Assume that c_I and c_E are strictly convex and continuously differentiable except at 0, and ρ_I and ρ_E are mildly monotone distortion risk measures defined on $\mathcal{X}$. In addition, all internal agents have the same amount $a > 0$ of initial loss exposures, i.e., $a = a_1 = \cdots = a_n$. Finally, we consider the situation where the number of external agents is larger than the number of internal agents by assuming that $m = kn$, where k is a positive integer, possibly large.

An *equilibrium* of this market is a tuple $(\mathbf{p}^*, \mathbf{w}^{1*}, \dots, \mathbf{w}^{n*}, \mathbf{u}^{1*}, \dots, \mathbf{u}^{m*}) \in (\mathbb{R}^n_+)^{n+m+1}$ if the following two conditions are satisfied.

(a) Individual optimality:

$$\mathbf{w}^{i*} \in \operatorname*{arg\,min}_{\mathbf{w}^i \in \mathbb{R}^n_+} \left\{ \rho_I\left(L_i(\mathbf{w}^i, \mathbf{p}^*)\right) + c_I(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|) \right\}, \quad \text{for each } i \in [n]; \tag{2.24}$$

$$\mathbf{u}^{j*} \in \operatorname*{arg\,min}_{\mathbf{u}^j \in \mathbb{R}^n_+} \left\{ \rho_E\left(L_E(\mathbf{u}^j, \mathbf{p}^*)\right) + c_E(\|\mathbf{u}^j\|) \right\}, \quad \text{for each } j \in [m]. \tag{2.25}$$

(b) Market clearance:

$$\sum_{i=1}^{n} \mathbf{w}^{i*} + \sum_{j=1}^{m} \mathbf{u}^{j*} = \sum_{i=1}^{n} \mathbf{a}^i. \tag{2.26}$$

The vector $\mathbf{p}^*$ is an *equilibrium price*, and $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ and $(\mathbf{u}^{1*}, \dots, \mathbf{u}^{m*})$ are *equilibrium allocations* for the internal and external agents, respectively. Before identifying the

equilibria in this market, we first make some simple observations. Let

$$L_E(b) = c_E'(b) + \rho_E(X) \quad \text{and} \quad L_I(b) = c_I'(b) + \rho_I(X), \qquad b \in \mathbb{R}.$$

We will write $L_I^-(0) = c_{I-}'(0) + \rho_I(X)$ and $L_I^+(0) = c_{I+}'(0) + \rho_I(X)$ to emphasize that the left and right derivative of c_I may not coincide at 0; this is particularly relevant in Theorem 2.3 (ii). On the other hand, $L_E(0)$ only has one relevant version since the allowed position is non-negative. Note that both L_E and L_I are continuous except at 0 and strictly increasing.

If an external agent takes only one source of loss (intuitively optimal from Proposition 2.6) among $X_1, \dots, X_n$ (we use the generic variable X for this loss), then $L_E(b)$ is the marginal cost of further increasing their position at bX. As a compensation, this agent will also receive p. Therefore, the external agent has incentives to participate in the risk sharing market if $p > L_E(0)$. If $p \leqslant L_E(0)$, due to the strict convexity of c_E, this agent will not take any risks. On the other hand, if $p \geqslant L_I^-(0)$, which means that it is expensive to transfer the loss externally, then the internal agent has no incentive to transfer. For a small risk exchange to benefit both parties, we need $L_E(0) < p < L_I^-(0)$. This implies, in particular,

$$\rho_E(X) \leqslant L_E(0) < p < L_I^-(0) \leqslant \rho_I(X),$$

which means that the risk is more acceptable to the external agents than to the internal agents, and the price is somewhere between the two risk assessments. The above intuition is helpful to understand the conditions in the following theorem.

Theorem 2.4. *In the Pareto risk sharing market of n internal and $m = kn$ external agents, suppose that $\alpha \in (0, 1]$. Let $\mathcal{E} = (\mathbf{p}, \mathbf{w}^{1*}, \dots, \mathbf{w}^{n*}, \mathbf{u}^{1*}, \dots, \mathbf{u}^{m*})$.*

(i) If $L_E(a/k) < L_I(-a)$, then there is no equilibrium.

(ii) Suppose that $L_E(a/k) \geqslant L_I(-a)$ and $L_E(0) < L_I^-(0)$. Let u^* be the unique solution to

$$L_E(u) = L_I(-ku), \qquad u \in (0, a/k]. \tag{2.27}$$

The tuple $\mathcal{E}$ is an equilibrium if and only if $\mathbf{p} = (p, \dots, p)$, $p = L_E(u^*)$,

$$(\mathbf{u}^{1*}, \dots, \mathbf{u}^{m*}) = u^*(\mathbf{e}_{k_1,n}, \dots, \mathbf{e}_{k_m,n}),$$

and

$$(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*}) = (a - ku^*)(\mathbf{e}_{\ell_1,n}, \dots, \mathbf{e}_{\ell_n,n}),$$

where $k_1, \dots, k_m \in [n]$ *and* $\ell_1, \dots, \ell_n \in [n]$ *such that*

$$u^* \sum_{j=1}^{m} \mathbb{1}_{\{k_j = s\}} + (a - ku^*) \sum_{i=1}^{n} \mathbb{1}_{\{\ell_i = s\}} = a \quad \textit{for each} \quad s \in [n].$$

Moreover, if $u^* < a/(2k)$*, then the tuple* $\mathcal{E}$ *is an equilibrium if and only if* $\mathbf{p} = (p, \dots, p)$*,* $p = L_E(u^*)$*,* $(\mathbf{u}^{1*}, \dots, \mathbf{u}^{m*})$ *is a permutation of* $u^*(\mathbf{e}_{\lceil 1/k \rceil, n}, \dots, \mathbf{e}_{\lceil m/k \rceil, n})$*, and* $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ *is a permutation of* $(a - ku^*)(\mathbf{e}_{1,n}, \dots, \mathbf{e}_{n,n})$*.*

(iii) Suppose that $L_E(0) \geqslant L_I^-(0)$*. The tuple* $\mathcal{E}$ *is an equilibrium if and only if* $\mathbf{p} = (p, \dots, p)$*,* $p \in [L_I^-(0), L_E(0) \wedge L_I^+(0)]$*,* $(\mathbf{u}^{1*}, \dots, \mathbf{u}^{m*}) = (0, \dots, 0)$*, and* $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ *is a permutation of* $a(\mathbf{e}_{1,n}, \dots, \mathbf{e}_{n,n})$*.*

To interpret Theorem 2.4 (i), note that $L_E(a/k) < L_I(-a)$ implies $L_E(u) < L_I(w - a)$ for all $u \in [0, a/k]$ and $w \in [0, a]$. It means that if the price of transferring a unit of risk is in $[L_E(a/k), L_I(-a)]$, the optimal position for each internal agent will be 0, and the external agents will have the incentives to increase their exposures from 0 to more than a/k. In this case, the individual optimality conditions (2.24) and (2.25) and the clearance condition (2.26) cannot be satisfied at the same time. Therefore, there is no equilibrium.

Compared with Theorem 2.3, where no benefits exist from risk sharing among the internal agents, Theorem 2.4 (ii) implies that in the presence of external agents, every party in the market may get better from risk sharing. More specifically, if $L_E(0) < L_I^-(0)$, (i.e., the marginal cost of increasing an external agent's position from 0 is smaller than the marginal benefit of decreasing an internal agent's position from a), there exists an equilibrium price $p \in [L_E(0), L_I^-(0)]$ such that both internal and external agents in the market can improve their objectives. The condition $L_E(0) < L_I^-(0)$ is crucial to such a win-win situation, as a price less than $L_I^-(0)$ will motivate the internal agents to transfer risk, and a price greater than $L_E(0)$ will motivate the external agents to receive risks. As shown by Theorem 2.4 (iii), if $L_E(0) \geqslant L_I^-(0)$, there are no incentives for the internal and external agents to participate in the risk sharing market, and their positions remain the same. Moreover, if $u^* < a/2k$, i.e, the optimal position of each external agent is very small compared with the total position of each loss in the market, the loss X_i for each $i \in [n]$, has to be shared by one internal agent and k external agents in order to achieve an equilibrium.

We make further observations on Theorem 2.4 (ii). From (2.27), it is straightforward to see that if k gets larger (more external agents are in the market), the equilibrium price p gets smaller. Intuitively, as more external agents are willing to take risks, they have to make some compromise on the received compensation to get the amount of risks they

want. The lower price further motivates the internal agents to transfer more risks to the external agents. Indeed, by (2.27), ku^* gets larger as k increases. On the other hand, u^* gets smaller as k increases. In the equilibrium model, each external agent will take less risk if more external agents are in the market. These observations can be seen more clearly in the example below.

Example 2.2 (Quadratic cost)**.** Suppose that the conditions in Theorem 2.4 (ii) are satisfied (this implies $\rho_E(X) < \rho_I(X)$ in particular), $c_I(x) = \lambda_I x^2$, and $c_E(x) = \lambda_E x^2$, $x \in \mathbb{R}$, where $\lambda_I, \lambda_E > 0$. We can compute the equilibrium price

$$p = \frac{k\lambda_I}{k\lambda_I + \lambda_E}\rho_E(X) + \frac{\lambda_E}{k\lambda_I + \lambda_E}\rho_I(X).$$

Therefore, the equilibrium price is a weighted average of $\rho_E(X)$ and $\rho_I(X)$, where the weights depend on k, λ_I, and λ_E. We also have the equilibrium allocations $\mathbf{u}^* = (u, \dots, u)$ and $\mathbf{w}^* = (w, \dots, w)$ where

$$u = \frac{\rho_I(X) - \rho_E(X)}{2(k\lambda_I + \lambda_E)} \quad \text{and} \quad w = \frac{k(\rho_E(X) - \rho_I(X))}{2(k\lambda_I + \lambda_E)} + a.$$

It is clear that p moves in the opposite direction of k. Moreover, if more external agents are in the market, each external agent will take fewer losses, while each internal agent will transfer more losses to the external agents. If λ_I increases, the internal agents will be less motivated to transfer their losses. To compensate for the increased penalty, the price paid by the internal agents will decrease so that they are still willing to share risks to some extent. The interpretation is similar if λ_E changes. Although the increase of different penalties (λ_E or λ_I) have different impacts on the price, the increase of either λ_E or λ_I leads to less incentives for the internal and external agents to participate in the risk sharing market.

2.5.4 Risk exchange for moderately heavy-tailed Pareto losses

In contrast to the settings in Sections 2.5.2 and 2.5.3, we consider moderately heavy-tailed Pareto losses below. The following proposition shows that agents prefer to share moderately heavy-tailed Pareto losses among themselves if they are equipped with ES.

Proposition 2.7. *In the Pareto risk sharing market, suppose that* $\alpha \in (1, \infty)$*, and* $\rho_1 = \cdots = \rho_n = \mathrm{ES}_q$ *for some* $q \in (0, 1)$*. Let*

$$\mathbf{w}^{i*} = \frac{a_i}{\sum_{j=1}^n a_j} \sum_{j=1}^n \mathbf{a}^j \text{ for } i \in [n] \quad \text{and} \quad \mathbf{p}^* = \left(\mathbb{E}\left[X_1|A\right], \dots, \mathbb{E}\left[X_n|A\right]\right),$$

where $A = \{\sum_{i=1}^n a_i X_i \geqslant \mathrm{VaR}_q\left(\sum_{i=1}^n a_i X_i\right)\}$. *Then the tuple* $(\mathbf{p}^*, \mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ *is an equilibrium.*

A sharp contrast is visible between the equilibrium in Theorem 2.3 and that in Proposition 2.7. For $\alpha \in (0, 1]$, the equilibrium price is the same across individual losses, and agents do not share losses at all. For $\alpha \in (1, \infty)$ and ES agents, each individual loss has a different equilibrium price, and agents share all losses proportionally.

We choose the risk measure ES here because it leads to an explicit expression of the equilibrium. Although ES is not finite for ultra heavy-tailed Pareto losses (thus, it does not fit Theorem 2.3), it can be approximated arbitrarily closely by RVaR (e.g., Embrechts et al. (2018)) which fits the condition of Theorem 2.3. By this approximation, we expect a similar situation if ES in Proposition 2.7 is replaced by RVaR, although we do not have an explicit result.

Remark 2.9. Proposition 2.7, in the case of Pareto(α), $\alpha > 1$, works for all convex risk measures (see Appendix 2.8.1). The intuition is that the value of convex risk measures can be reduced by diversification, i.e., $\rho(\lambda X + (1-\lambda)Y) \leqslant \lambda\rho(X) + (1-\lambda)\rho(Y)$ where ρ is a convex risk measure, X and Y are two random variables with finite mean, and $\lambda \in (0, 1)$. Convex risk measures are not suitable for the case of ultra heavy-tailed Pareto risks as they will always be infinite for risks without finite mean (see e.g., Filipović and Svindland (2012)).

2.6 Numerical examples

2.6.1 Diversification effects as n increases

For $\alpha \in (0, 1]$, $p \in (0, 1)$, and iid Pareto(α) random variables $X_1, \dots, X_n$, we compute $\mathrm{VaR}_p(\sum_{i=1}^n X_i/n)$ for $n = 2, \dots, 6$. From Figure 2.1, we observe that $\mathrm{VaR}_p(\sum_{i=1}^n X_i/n)$ increases as n increases. The difference between the curves for different n becomes more pronounced as α becomes smaller, i.e., the tail of the Pareto losses becomes heavier. From these numerical results, we may expect that

$$\frac{1}{k}\sum_{i=1}^{k} X_i \leqslant_{\mathrm{st}} \frac{1}{\ell}\sum_{i=1}^{\ell} X_i,$$

where $k, \ell \in \mathbb{N}$ and $k \leqslant \ell$. We were only able to show the case where ℓ is a multiple of k in Proposition 2.2.

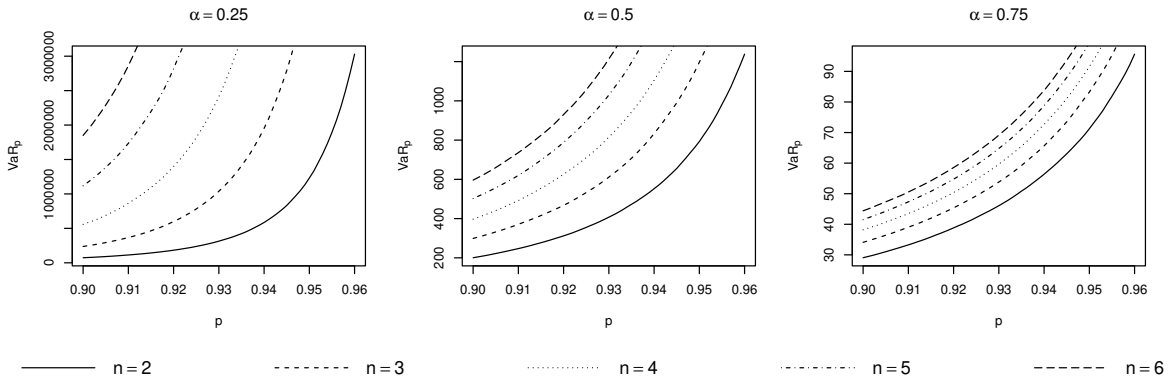


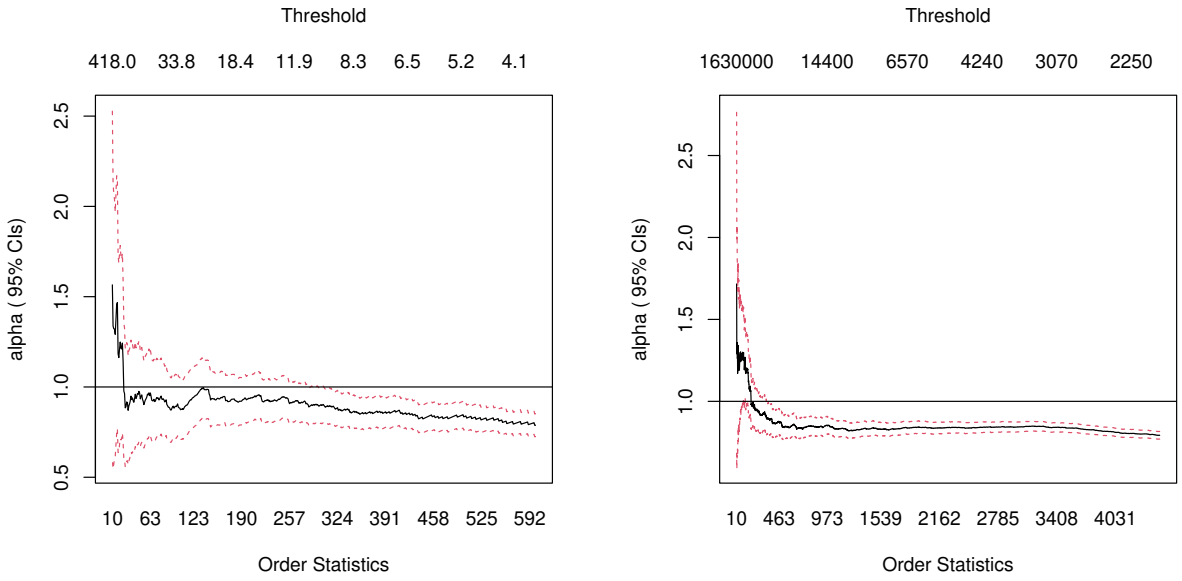
Figure 2.1: $\mathrm{VaR}_p((X_1 + \cdots + X_n)/n)$ for $n = 2, \ldots, 6$ and $p \in (0.9, 0.96)$.

2.6.2 Examples of ultra heavy-tailed Pareto losses

In addition to the many examples mentioned in Section 2.1.1, we provide two further data examples: a first one on marine losses, and a second one on suppression costs of wildfires. Using EVT, we will show that both examples exhibit infinite mean behavior. The marine losses dataset, from the insurance data repository `CASdatasets`[4], was originally collected by a French private insurer and comprises 1,274 marine losses (paid) between January 2003 and June 2006. The wildfire dataset[5] contains 10,915 suppression costs in Alberta, Canada from 1983 to 1995. For the purpose of this section, we only provide the Hill estimates of these two datasets, although a more detailed EVT analysis is available (see McNeil et al. (2015)). The Hill estimates of the tail indices α are presented in Figure 2.2, where the black curves represent the point estimates and the red curves represent the 95% confidence intervals with varying thresholds; see McNeil et al. (2015) for more details on the Hill estimator. As suggested by McNeil et al. (2015), one may roughly chose a threshold around the top 5% order statistics of the data. Following this suggestion, the tail indices α for the marine losses and wildfire suppression costs are estimated as 0.916 and 0.847 with 95% confidence intervals being $(0.674, 1.158)$ and $(0.776, 0.918)$, respectively; thus, these losses/costs have infinite mean if they follow Pareto distributions in their tails regions.

The observations in Figure 2.2 suggest that the two loss datasets may have similar tail parameters. As discussed in Remark 2.1, Theorem 2.1 can be applied to generalized Pareto distributions. If two loss random variables X_1 and X_2 are independent and follow generalized Pareto distributions with the same tail parameter $\alpha = 1/\xi < 1$ (see (2.3)),

[4]Available at `http://cas.uqam.ca/`.
[5]Available at `https://wildfire.alberta.ca/`.



(a) Marine losses

(b) Wildfire suppression costs

Figure 2.2: Hill plots for the marine losses and wildfire suppression costs: For each risk, the Hill estimates are plotted as black curve with the 95% confidence intervals being red curves.

then, for all $p \in (0,1)$,

$$\mathrm{VaR}_p(X_1 + X_2) > \mathrm{VaR}_p(X_1) + \mathrm{VaR}_p(X_2). \tag{2.28}$$

Even if X_1 and X_2 are not Pareto distributed, as long as their tails are Pareto, (2.28) may hold for p relatively large, as suggested by Proposition 2.3 and Remark 2.5.

We will verify (2.28) on our datasets to show how the implication of our main results holds for real data. Since the marine losses data were scaled to mask the actual losses, we renormalize it by multiplying the data by 500 to make it roughly on the same scale as that of the wildfire suppression costs;[6] this normalization does not matter for (2.28) and is made only for better visualization. Let $\widehat{F}_1$ be the empirical distribution of the marine losses (renormalized) and $\widehat{F}_2$ be the empirical distribution of the wildfire suppression costs. Take independent random variables $\widehat{Y}_1 \sim \widehat{F}_1$ and $\widehat{Y}_2 \sim \widehat{F}_2$. Let $\widehat{F}_1 \oplus \widehat{F}_2$ be the distribution

[6]The average marine losses (renormalized) and the average wildfire suppression costs are 12400 and 12899.

with quantile function $p \mapsto \mathrm{VaR}_p(\widehat{Y}_1) + \mathrm{VaR}_p(\widehat{Y}_2)$, i.e., the comonotonic sum, and $\widehat{F}_1 * \widehat{F}_2$ be the distribution of $\widehat{Y}_1 + \widehat{Y}_2$, i.e., the independent sum.

The differences between the distributions $\widehat{F}_1 \oplus \widehat{F}_2$ and $\widehat{F}_1 * \widehat{F}_2$ can be seen in Figure 2.3a. We observe that $\widehat{F}_1 * \widehat{F}_2$ is less than $\widehat{F}_1 \oplus \widehat{F}_2$ over a wide range of loss values. In particular, the relation holds for all losses less than 267,659.5 (marked by the vertical line in Figure 2.3a). Equivalently, we can see from Figure 2.3b that

$$\mathrm{VaR}_p(\widehat{Y}_1 + \widehat{Y}_2) > \mathrm{VaR}_p(\widehat{Y}_1) + \mathrm{VaR}_p(\widehat{Y}_2) \tag{2.29}$$

holds unless p is greater than 0.9847 (marked by the vertical line in Figure 2.3b). Recall that $\widehat{F}_1 * \widehat{F}_2 \leqslant \widehat{F}_1 \oplus \widehat{F}_2$ is equivalent to (2.29) holding for all $p \in (0, 1)$. Since the quantiles are directly computed from data, thus from distributions with bounded supports, for p close enough to 1 it must hold that $\mathrm{VaR}_p(\widehat{Y}_1 + \widehat{Y}_2) \leqslant \mathrm{VaR}_p(\widehat{Y}_1) + \mathrm{VaR}_p(\widehat{Y}_2)$; see the similar observation made in Proposition 2.1. Nevertheless, we observe (2.29) for most values of $p \in (0, 1)$. Note that the observation of (2.29) is entirely empirical and it does not use the fitted models.

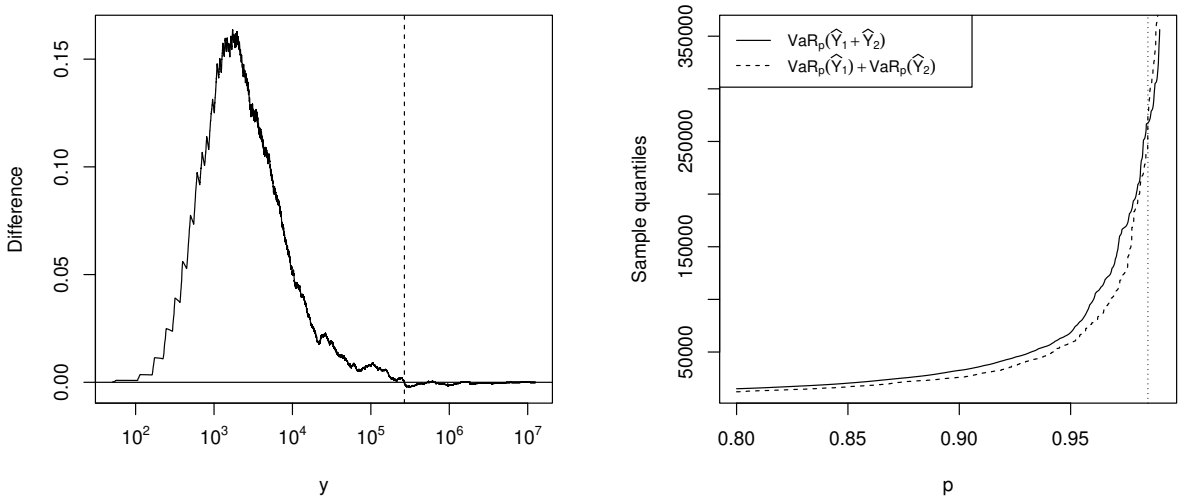
Let F_1 and F_2 be the true distributions (unknown) of the marine losses (renormalized) and wildfire suppression costs, respectively. We are interested in whether the first-order stochastic dominance relation $F_1 * F_2 \leqslant F_1 \oplus F_2$ holds. Since we do not have access to the true distributions, we generate two independent random samples of size 10^4 (roughly equal to the sum of the sizes of the datasets, thus with a similar magnitude of randomness) from the distributions $\widehat{F}_1 \oplus \widehat{F}_2$ and $\widehat{F}_1 * \widehat{F}_2$. We treat these samples as independent random samples from $F_1 \oplus F_2$ and $F_1 * F_2$ and test the hypothesis using Proposition 1 of Barrett and Donald (2003). The p-value of the test is greater than 0.5 and we are not able to reject the hypothesis $F_1 * F_2 \leqslant F_1 \oplus F_2$.

2.6.3 Aggregation of Pareto risks with different parameters

As mentioned above, for independent losses $Y_1, \dots, Y_n$ following generalized Pareto distributions with the same tail parameter $\alpha = 1/\xi < 1$, it holds that

$$\sum_{i=1}^{n} \mathrm{VaR}_p(Y_i) \leqslant \mathrm{VaR}_p\left(\sum_{i=1}^{n} Y_i\right), \text{ usually with strict inequality.} \tag{2.30}$$

Inspired by the results in Section 2.6.2, we are interested in whether (2.30) holds for losses following generalized Pareto distributions with different parameters. To make a first



(a) Differences of the distributions: $\widehat{F}_1 \oplus \widehat{F}_2 - \widehat{F}_1 * \widehat{F}_2$

(b) Sample quantiles for $p \in (0.8, 0.99)$

Figure 2.3: Plots for $\widehat{F}_1 \oplus \widehat{F}_2 - \widehat{F}_1 * \widehat{F}_2$ and sample quantiles

attempt on this problem, we look at the 6 operational losses of different business lines with infinite mean in Table 5 of Moscadelli (2004), where the operational losses are assumed to follow generalized Pareto distributions. Denote by $Y_1, \dots, Y_6$ the operational losses corresponding to these 6 generalized Pareto distributions. The estimated parameters in Moscadelli (2004) for these losses are presented in Table 2.1; they all have infinite mean.

i	1	2	3	4	5	6
ξ_i	1.19	1.17	1.01	1.39	1.23	1.22
β_i	774	254	233	412	107	243

Table 2.1: The estimated parameters ξ_i and β_i, $i \in [6]$.

For the purpose of this numerical example, we assume that $Y_1, \dots, Y_6$ are independent and plot $\sum_{i=1}^6 \mathrm{VaR}_p(Y_i)$ and $\mathrm{VaR}_p(\sum_{i=1}^6 Y_i)$ for $p \in (0.95, 0.99)$ in Figure 2.4. We can see that $\mathrm{VaR}_p(\sum_{i=1}^6 Y_i)$ is larger than $\sum_{i=1}^6 \mathrm{VaR}_p(Y_i)$, and the gap between the two values gets larger as the level p approaches 1. This observation further suggests that, even if the ultra heavy-tailed Pareto losses have different tail parameters, a diversification penalty may still

exist. We conjecture that this is true for any generalized Pareto losses $Y_1, \dots, Y_n$ with shape parameters $\xi_1, \dots, \xi_n \in [1, \infty)$, although we do not have a proof. Similarly, we may expect that $\sum_{i=1}^n \theta_i \mathrm{VaR}_p(X_i) \leqslant \mathrm{VaR}_p(\sum_{i=1}^n \theta_i X_i)$ holds for any Pareto losses $X_1, \dots, X_n$ with tail parameters $\alpha_1, \dots, \alpha_n \in (0, 1]$,

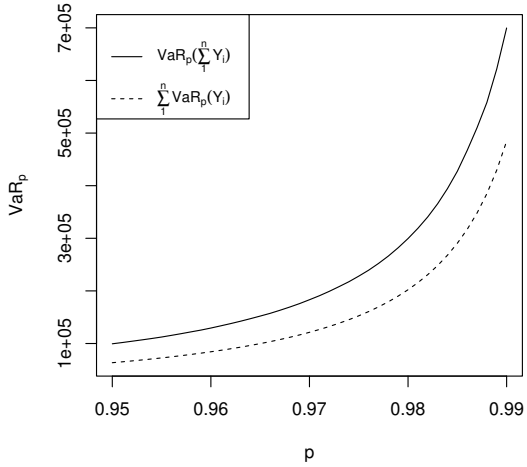


Figure 2.4: Curves of $\mathrm{VaR}_p(\sum_{i=1}^n Y_i)$ and $\sum_{i=1}^n \mathrm{VaR}_p(Y_i)$ for $n = 6$ generalized Pareto losses with parameters in Table 2.1 and $p \in (0.95, 0.99)$.

From a risk management point of view, the message from Sections 2.6.2 and 2.6.3 is clear. If a careful statistical analysis leads to statistical models in the realm of infinite means, then the risk manager at the helm should take a step back and question to what extent classical diversification arguments can be applied. Though we mathematically analyzed the case of equal parameters, we conjecture that these results hold more widely in the heterogeneous case. As a consequence, it is advised to hold on to only one such ultra-heavy tailed risk. Of course, the discussion concerning the practical relevance of infinite mean models remains. When such underlying models are methodologically possible, then one should think carefully about the applicability of standard risk management arguments; this brings us back to Weitzman's Dismal Theorem as discussed towards the end of Section 2.1. From a methodological point of view, we expect that the results from Sections 2.4 and 2.5 carry over to the above heterogeneous setting.

2.7 Concluding remarks

We establish in Theorem 2.1 the inequality that the diversification of iid Pareto losses without finite mean is greater than an individual Pareto loss in the sense of first-order stochastic dominance, which is a very strong dominance relation. The result of stochastic dominance is further generalized to three cases: (i) the losses are Pareto in the tail region (Proposition 2.3); (ii) the number and weights of Pareto losses are random (Proposition 2.4); (iii) the Pareto losses are triggered by catastrophic events (Theorem 2.2). These results provide an important implication in risk management, i.e., the diversification of Pareto losses without finite mean may increase the risk assessment of a portfolio (Proposition 2.6).

The equilibrium of a risk exchange model is analyzed in Theorem 2.3, where agents can take extra Pareto losses with compensations. In particular, if every agent is associated with an initial position of a Pareto loss without finite mean, the agents can merely exchange their entire position with each other. On the other hand, if some external agents are not associated with any initial losses, it is possible that all agents can reduce their risks by transferring the losses from the agents with initial losses to those without initial losses (Theorem 2.4).

2.8 Appendix

2.8.1 Background on risk measures

We collect some common terminology a new result on risk measures. We first present commonly used properties of a risk measure $\rho : \mathcal{X}_\rho \to \mathbb{R}$.

(c) Translation invariance: $\rho(X + c) = \rho(X) + c$ for $c \in \mathbb{R}$.

(d) Positive homogeneity: $\rho(aX) = a\rho(X)$ for $a \geqslant 0$.

(e) Convexity: $\rho(\lambda X + (1-\lambda)Y) \leqslant \lambda\rho(X) + (1-\lambda)\rho(Y)$ for $X, Y \in \mathcal{X}_\rho$ and $\lambda \in [0, 1]$.

A risk measure that satisfies (a) weak monotonicity, (c) translation invariance, and (e) convexity is a *convex risk measure* (Föllmer and Schied, 2002). It is well-known that ES is a convex risk measure. The convexity property means that diversification will not increase the risk of the loss portfolio, i.e., the risk of $\lambda X + (1-\lambda)Y$ is less than or equal to that of

the weighted average of individual losses. However, the canonical space for law-invariant convex risk measures is L^1 (see Filipović and Svindland (2012)) and hence convex risk measures are not useful for losses without finite mean.

For losses without finite mean, such as ultra heavy-tailed Pareto losses, it is natural to consider VaR or Range Value-at-Risk (RVaR), which includes VaR as a limiting case. For $X \in \mathcal{X}$ and $0 \leqslant p < q < 1$, the RVaR is defined as

$$\mathrm{RVaR}_{p,q}(X) = \frac{1}{q-p} \int_p^q \mathrm{VaR}_u(X) \mathrm{d}u.$$

For $p \in (0,1)$, $\lim_{q \downarrow p^+} \mathrm{RVaR}_{p,q}(X) = \mathrm{VaR}_p(X)$. The class of RVaR is proposed by Cont et al. (2010) as robust risk measures; see Embrechts et al. (2018) for its properties and risk sharing results. VaR, ES and RVaR, as well as essential infimum (ess-inf) and essential supremum (ess-sup), belong to distortion risk measures in (2.21). For $X \in \mathcal{X}$, ess-inf and ess-sup are defined as

$$\text{ess-inf}(X) = \sup\{x : F_X(x) = 0\} \quad \text{and} \quad \text{ess-sup}(X) = \inf\{x : F_X(x) = 1\}.$$

The distortion functions of ess-inf and ess-sup are given as $h(t) = \mathbb{1}_{\{t=1\}}$ and $h(t) = \mathbb{1}_{\{0<t\leqslant 1\}}$, $t \in [0,1]$, respectively; see Table 1 of Wang et al. (2020). Distortion risk measures satisfy (a), (c) and (d). Almost all the useful distortion risk measures are mildly monotone, as shown by the following proposition.

Proposition 2.8. *Any distortion risk measure is mildly monotone unless it is a mixture of* ess-sup *and* ess-inf.

Proof. Let ρ_h be a distortion risk measure with distortion function h. Suppose that ρ_h is not mildly monotone. There exist $X, Y \in \mathcal{X}$ satisfying $F_X^{-1}(p) < F_Y^{-1}(p)$ for all $p \in (0,1)$ and $\rho(X) = \rho(Y)$. Suppose that there exist $b \in (0,1)$ such that $h(1-a) < h(1-b)$ for all $a > b$. For $x \in (F_X^{-1}(b), F_Y^{-1}(b))$, we have $F_X(x) \geqslant b > F_Y(x)$; see e.g., Lemma 1 of Guan et al. (2022). Hence, we have $h(1-F_X(x)) \leqslant h(1-b) < h(1-F_Y(x))$ for $x \in (F_X^{-1}(b), F_Y^{-1}(b))$. Since $h(1-F_X(x)) - h(1-F_Y(x)) \leqslant 0$ for all $x \in \mathbb{R}$, by (2.21) we get

$$\rho(X) - \rho(Y) = \int_{-\infty}^{\infty} (h(1-F_X(x)) - h(1-F_Y(x)))\, \mathrm{d}x < 0.$$

This conflicts $\rho(X) = \rho(Y)$. Hence, there is no $b \in (0,1)$ such that $h(1-a) < h(1-b)$ for all $a > b$. Using a similar argument with the left quantiles replaced by right quantiles, we conclude that there is no $b \in (0,1)$ such that $h(1-a) > h(1-b)$ for all $a < b$. Therefore,

for every $b \in (0,1)$, there exists an open interval I_b such that $b \in I_b$ and h is a constant on I_b. For any $\varepsilon > 0$, the interval $[\varepsilon, 1-\varepsilon]$ is compact. There exists a finite collection $\{I_b : b \in B\}$ which covers $[\varepsilon, 1-\varepsilon]$. Since the open intervals in $\{I_b : b \in B\}$ overlap, we know that h is a constant on $[\varepsilon, 1-\varepsilon]$. Sending $\varepsilon \downarrow 0$ yields that h takes a constant value on $(0,1)$, denoted by $\lambda \in [0,1]$. Together with $h(0) = 0$ and $h(1) = 1$, we get that $h(t) = \lambda \mathbb{1}_{\{0<t\leqslant 1\}} + (1-\lambda)\mathbb{1}_{\{t=1\}}$ for $t \in [0,1]$, which is the distortion function of $\rho_h = \lambda\text{ess-inf} + (1-\lambda)\text{ess-sup}$. □

As a consequence, for any set $\mathcal{X}$ containing a random variable unbounded from above and one unbounded from below, such as the L^q-space for $q \in [0,\infty)$, a real-valued distortion risk measure on $\mathcal{X}$ is mildly monotone.

2.8.2 Proofs of all theorems, propositions, and lemmas of Chapter 2

Proof of Theorem 2.1. For $(u_1,\dots,u_n) \in (0,1)^n$ and $\boldsymbol{\theta} = (\theta_1,\dots,\theta_n) \in \Delta_n$, define the generalized weighted average $M_{r,\boldsymbol{\theta}}(u_1,\dots,u_n) = (\theta_1 u_1^r + \cdots + \theta_n u_n^r)^{\frac{1}{r}}$, where $r \in \mathbb{R}$. Note that (2.2) can be equivalently written as

$$M_{r,\boldsymbol{\theta}}(U_1,\dots,U_n) \leqslant_{\mathrm{st}} U, \tag{2.31}$$

where $U, U_1, \dots, U_n$ are iid uniform random variables on $(0,1)$, and $r = -1/\alpha \in (\infty, -1]$. It is well known that $M_{r,\boldsymbol{\theta}} \leqslant M_{s,\boldsymbol{\theta}}$ for $r \leqslant s$; see Theorem 16 of Hardy et al. (1934). Hence, $M_{r,\boldsymbol{\theta}}(U_1,\dots,U_n) \leqslant M_{-1,\boldsymbol{\theta}}(U_1,\dots,U_n)$ for all $r \leqslant -1$. Therefore, for (2.31) to hold for all $r \leqslant -1$, it suffices to show that $M_{-1,\boldsymbol{\theta}}(U_1,\dots,U_n) \leqslant_{\mathrm{st}} U$.

If some of $\theta_1,\dots,\theta_n$ are 0, we can reduce the dimension of the problem. Hence, we will assume $\min_{i\in[n]} \theta_i > 0$ in the proof below. There is nothing to show if only one $\theta_i > 0$ which reduces to dimension 1.

We first show the case of $n = 2$. For a fixed $p \in (0,1)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Delta_2$ where $\min\{\theta_1, \theta_2\} > 0$, let $\delta = \theta_2/(p^{-1} - 1 + \theta_2)$. For $(u_1, u_2) \in (0,1)^2$, if $u_2 \leqslant \delta$, then

$$\theta_1 u_1^{-1} + \theta_2 u_2^{-1} \geqslant \theta_1 + \theta_2 \delta^{-1} = 1 - \theta_2 + p^{-1} - 1 + \theta_2 = p^{-1}.$$

Hence, $M_{-1,\boldsymbol{\theta}}(u_1, u_2) \leqslant p$ if $u_2 \leqslant \delta$. Then, for iid uniform random variables U_1 and U_2 on

$(0,1)$, we have

$$\begin{aligned}
\mathbb{P}\left(M_{-1,\boldsymbol{\theta}}(U_1,U_2)\leqslant p\right) &= \mathbb{P}\left(\theta_1 U_1^{-1}+\theta_2 U_2^{-1}\geqslant p^{-1}\right)\\
&= \mathbb{P}(U_2\leqslant\delta)+\mathbb{P}\left(\theta_1 U_1^{-1}\geqslant p^{-1}-\theta_2 U_2^{-1}, U_2>\delta\right)\\
&\geqslant \mathbb{P}(U_2\leqslant\delta)+\mathbb{P}\left(\theta_1 U_1^{-1}\geqslant p^{-1}-\theta_2, U_2>\delta\right)\\
&= \delta+\theta_1(1-\delta)(p^{-1}-\theta_2)^{-1}\\
&> \delta+\theta_1(1-\delta)p\\
&= \theta_1 p+p\delta(p^{-1}-1+\theta_2)=p.
\end{aligned}$$

Hence, we have shown the case when $n=2$. Next, let $n\geqslant 2$, and $\boldsymbol{\theta}=(\theta_1,\dots,\theta_{n+1})\in\Delta_{n+1}$ where $\min_{i\in[n+1]}\theta_i>0$. Let $U,U_1,\dots,U_{n+1}$ be iid uniform random variables on $(0,1)$. Assume that $U^{-1}\leqslant_{\mathrm{st}}\theta_1/(\sum_{i=1}^n\theta_i)U_1^{-1}+\cdots+\theta_n/(\sum_{i=1}^n\theta_i)U_n^{-1}$. As first-order stochastic dominance is closed under convolutions (e.g., Theorem 1.A.3 (a) of Shaked and Shanthikumar (2007)), we have

$$\theta_1 U_1^{-1}+\cdots+\theta_{n+1}U_{n+1}^{-1}\geqslant_{\mathrm{st}}\left(\sum_{i=1}^n\theta_i\right)U^{-1}+\theta_{n+1}U_{n+1}^{-1}\geqslant_{\mathrm{st}}U^{-1},$$

Thus, $M_{-1,\boldsymbol{\theta}}(U_1,\dots,U_{n+1})\leqslant_{\mathrm{st}}U$. Moreover, for $p\in(0,1)$,

$$\begin{aligned}
\mathbb{P}\left(M_{-1,\boldsymbol{\theta}}(U_1,\dots,U_{n+1})\leqslant p\right) &= \mathbb{P}\left(\theta_1 U_1^{-1}+\cdots+\theta_{n+1}U_{n+1}^{-1}\geqslant p^{-1}\right)\\
&\geqslant \mathbb{P}\left(\left(\sum_{i=1}^n\theta_i\right)U^{-1}+\theta_{n+1}U_{n+1}^{-1}\geqslant p^{-1}\right)>p.
\end{aligned}$$

By induction, we have the desired result. □

Proof of Proposition 2.2. Let $Y_j=(\sum_{i=n(j-1)+1}^{jn}X_i)/n$, $j=1,\dots,m$. By Theorem 2.1, $X_j'\leqslant_{\mathrm{st}}Y_j$ for $j=1,\dots,m$, where $X_1',\dots,X_m'\sim\mathrm{Pareto}(\alpha)$ are independent. Note that $Y_1,\dots,Y_m$ are also independent. As first-order stochastic dominance is closed under convolutions (e.g., Theorem 1.A.3 (a) of Shaked and Shanthikumar (2007)), we obtain

$$X_1+\cdots+X_m\simeq_{\mathrm{st}}X_1'+\cdots+X_m'\leqslant_{\mathrm{st}}Y_1+\cdots+Y_m=\frac{X_1+\cdots+X_{mn}}{n}.$$

Dividing both sides by m yields the desired inequality. □

Proof of Proposition 2.4. By Theorem 2.1 and the law of total expectation, it is easy to verify that, for $n = 2, 3, \ldots$, $\mathbb{P}(\sum_{i=1}^n W_i X_i / \sum_{i=1}^n W_i \leqslant t) < \mathbb{P}(X \leqslant t)$, $t > 1$. As N is independent of $\{W_i X_i\}_{i\in\mathbb{N}}$, for $t > 1$,

$$\begin{aligned}\mathbb{P}\left(\frac{\sum_{i=1}^N W_i X_i}{\sum_{i=1}^N W_i} \leqslant t\right) &= \mathbb{P}(N=0) + \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n W_i} \leqslant t\right) \mathbb{P}(N=n) \\ &\leqslant \mathbb{P}(N=0) + \mathbb{P}(N \geqslant 1)\left(1 - \frac{1}{t^\alpha}\right) = \mathbb{P}\left(X \mathbb{1}_{\{N\geqslant 1\}} \leqslant t\right).\end{aligned}$$

It is obvious that the inequality is strict if $\mathbb{P}(N \geqslant 2) \neq 0$. To show the second inequality in (2.7), note that for each realization of $N = n$ and $(W_1, \ldots, W_N) = (w_1, \ldots, w_n) \in \mathbb{R}^n$, $\sum_{i=1}^n w_i X \leqslant_{\mathrm{st}} \sum_{i=1}^n w_i X_i$ holds by Theorem 2.1. Hence, the second inequality in (2.7) holds. □

Proof of Lemma 2.1. The result is clearly true if $c_1 = \cdots = c_n = 0$. If any components of $(c_1, \ldots, c_n)$ are zero, the problem simply reduces its dimension. Hence, we assume that $(c_1, \ldots, c_n) \in (0, 1]^n$ for the rest of the proof. For $t \geqslant 1 \geqslant \max_{i\in[n]} c_i$,

$$\mathbb{P}\left(\sum_{i=1}^n c_i X \mathbb{1}_{B_i} \leqslant t\right) = \sum_{i=1}^n \left(1 - \frac{c_i^\alpha}{t^\alpha}\right) \mathbb{P}(B_i) + \mathbb{P}\left(\bigcap_{i\in[n]} B_i^c\right).$$

Since $B_1, \ldots, B_n$ are mutually exclusive, $\sum_{i=1}^n \mathbb{P}(B_i) = \mathbb{P}\left(\bigcup_{i\in[n]} B_i\right) = 1 - \mathbb{P}\left(\bigcap_{i\in[n]} B_i^c\right)$. Moreover, as $c_i \in (0, 1]$ and $\alpha \in (0, 1]$, $c_i^\alpha \geqslant c_i$ for $i \in [n]$. Therefore,

$$\mathbb{P}\left(\sum_{i=1}^n c_i X \mathbb{1}_{B_i} \leqslant t\right) = 1 - \sum_{i=1}^n \frac{c_i^\alpha}{t^\alpha} \mathbb{P}(B_i) \leqslant 1 - \frac{1}{t^\alpha} \sum_{i=1}^n c_i \mathbb{P}(B_i) = 1 - \frac{1}{t^\alpha} \mathbb{P}(A) = \mathbb{P}(X \mathbb{1}_A \leqslant t).$$

For $t \in [0, 1)$, $\mathbb{P}\left(\sum_{i=1}^n c_i X \mathbb{1}_{B_i} \leqslant t\right) \leqslant \mathbb{P}\left(\sum_{i=1}^n c_i X \mathbb{1}_{B_i} \leqslant 1\right) \leqslant 1 - \mathbb{P}(A) = \mathbb{P}(X \mathbb{1}_A \leqslant t)$. This yields the desired result. □

Proof of Theorem 2.2. For $S \subseteq [n]$, let $B_S = \left(\bigcap_{i\in S} A_i\right) \cap \left(\bigcap_{i\in S^c} A_i^c\right)$. For $(\theta_1, \ldots, \theta_n) \in \mathbb{R}_+^n$, we write

$$\sum_{i=1}^n \theta_i X_i \mathbb{1}_{A_i} = \sum_{S\subseteq[n]} \mathbb{1}_{B_S} \sum_{i\in S} \theta_i X_i.$$

By Theorem 2.1, $\sum_{i\in S} \theta_i X_i \geqslant_{\mathrm{st}} \sum_{i\in S} \theta_i X$ for any $S \subseteq [n]$. As $A_1, \ldots, A_n$ are independent of $(X_1, ..., X_n)$, by Theorem 1.A.14 of Shaked and Shanthikumar (2007), $\sum_{i\in S} \theta_i X_i \mathbb{1}_{B_S} \geqslant_{\mathrm{st}}$

$\sum_{i\in S}\theta_i X\mathbb{1}_{B_S}$ for any $S\subseteq[n]$. Since B_S and B_R are mutually exclusive for any distinct $S,R\subseteq[n]$, we have

$$\sum_{i=1}^{n}\theta_i X_i\mathbb{1}_{A_i}=\sum_{S\subseteq[n]}\mathbb{1}_{B_S}\sum_{i\in S}\theta_i X_i\geqslant_{\mathrm{st}}\sum_{S\subseteq[n]}\sum_{i\in S}\theta_i X\mathbb{1}_{B_S}.$$

Note that

$$\sum_{S\subseteq[n]}\mathbb{P}(B_S)\sum_{i\in S}\theta_i=\sum_{j=1}^{n}\theta_j\sum_{S\subseteq[n],j\in S}\mathbb{P}(B_S)=\sum_{j=1}^{n}\theta_i\mathbb{P}(A_j)=\lambda\mathbb{P}(A).$$

As $\sum_{i\in S}\theta_i/\lambda\in[0,1]$ for any $S\subseteq[n]$, by Lemma 2.1, $\sum_{S\subseteq[n]}(\sum_{i\in S}\theta_i/\lambda)X\mathbb{1}_{B_S}\geqslant_{\mathrm{st}}X\mathbb{1}_A$. Hence, $\sum_{i=1}^{n}\theta_i X_i\mathbb{1}_{A_i}\geqslant_{\mathrm{st}}\lambda X\mathbb{1}_A$. □

Proof of Proposition 2.6. The proof of (i) follows directly from Theorem 2.1. (ii) follows from Theorem 2.2 by noting that there exists $j\in[n]$ such that $\mathbb{P}(A_j)\leqslant\mathbb{P}(A)$, and hence,

$$wX_j\mathbb{1}_{A_j}\leqslant_{\mathrm{st}}wX\mathbb{1}_A\leqslant_{\mathrm{st}}\sum_{i=1}^{n}w_iX_i\mathbb{1}_{A_i},$$

where X and A are in (2.10) with $\lambda=w$ and $(\theta_1,\dots,\theta_n)=(w_1,\dots,w_n)$. □

Proof of Theorem 2.3. (i) Suppose that $(\mathbf{p}^*,\mathbf{w}^{i*},\dots,\mathbf{w}^{n*})$ forms an equilibrium. We let $p=\max_{j\in[n]}\{p_j\}$ and $S=\arg\max_{j\in[n]}\{p_j\}$. For the ith agent, by writing $w=\|\mathbf{w}^i\|$, using Theorem 2.1 and the fact that ρ_i is mildly monotone, we have for any $\mathbf{w}^i\in[0,1]^n$,

$$\begin{aligned}\rho_i(L_i(\mathbf{w}^i,\mathbf{p}^*))&=\rho_i(\mathbf{w}^i\cdot(\mathbf{X}-\mathbf{p}^*)+\mathbf{a}^i\cdot\mathbf{p}^*)\\&\geqslant\rho_i(\mathbf{w}^i\cdot\mathbf{X}-wp+\mathbf{a}^i\cdot\mathbf{p}^*)\geqslant\rho_i(wX_1-wp+\mathbf{a}^i\cdot\mathbf{p}^*)\end{aligned}$$

and the last inequality is strict if $\mathbf{w}^i$ contains at least two non-zero components by the last statement of Theorem 2.1. Moreover, $c(\|\mathbf{w}^i\|-\|\mathbf{a}^i\|)=c(w-\|\mathbf{a}^i\|)$. Therefore, we know that the optimizer $\mathbf{w}^{i*}=(w_1^{i*},\dots,w_n^{i*})$ to (2.19) has at most one non-zero component w_j^{i*}, and $j\in S$. Hence, $w_k^{i*}=0$ if $k\in[n]\setminus S$ and this holds for each $i\in[n]$. Using $\sum_{i=1}^{n}\mathbf{w}^{i*}=\sum_{i=1}^{n}\mathbf{a}^i$ which have all positive components, we know that $S=[n]$, which further implies that $\mathbf{p}^*=(p,\dots,p)$ for $p\in\mathbb{R}_+$. Next, as each $\mathbf{w}^{i*}$ has only one positive component, $(\mathbf{w}^{i*},\dots,\mathbf{w}^{n*})$ has to be an n-permutation of $(\mathbf{a}^1,\dots,\mathbf{a}^n)$ to satisfy the clearance condition (2.20).

(ii) The clearance condition (2.20) is clearly satisfied. Note that distortion risk measures are translation invariant and positive homogeneous (see Section 2.8.1 for properties of risk measures). Using these two properties and Proposition 2.6, for $i \in [n]$,

$$\begin{aligned}
&\min_{\mathbf{w}^i \in \mathbb{R}^n_+} \left\{ \rho_i \left(L_i(\mathbf{w}^i, \mathbf{p}^*) \right) + c_i(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|) \right\} \\
&= \min_{\mathbf{w}^i \in \mathbb{R}^n_+} \left\{ \rho_i \left(\mathbf{w}^i \cdot \mathbf{X} - (\mathbf{w}^i - \mathbf{a}^i) \cdot \mathbf{p}^* \right) + c_i(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|) \right\} \\
&= \min_{\|\mathbf{w}^i\| \in \mathbb{R}_+} \left\{ (\rho_i \left(\|\mathbf{w}^i\| X \right) - (\|\mathbf{w}^i\| - a_i)p) + c_i(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|) \right\} \\
&= \min_{w \in \mathbb{R}_+} \left\{ w(\rho_i(X) - p) + a_i p + c_i(w - a_i) \right\}. \qquad (2.32)
\end{aligned}$$

Note that $w \mapsto w(\rho_i(X) - p) + c_i(w - a_i)$ is convex and with condition (2.22), its minimum is attained at $w = a_i$. Therefore, $\mathbf{w}^{i*} = \mathbf{a}^{i*}$ is an optimizer to (2.19), which shows the desired statement of equilibrium.

(iii) By (i), $(\mathbf{w}^{1*}, \dots, \mathbf{w}^{n*})$ is an n-permutation of $(\mathbf{a}^1, \dots, \mathbf{a}^n)$. It means that for any $i \in [n]$, there exists $j \in [n]$ such that a_j is the minimizer of (2.32). As c_i is convex, we have

$$c'_{i+}(a_j - a_i) \geqslant p - \rho_i(X) \geqslant c'_{i-}(a_j - a_i), \qquad \text{for each} \qquad i \in [n].$$

Hence, we obtain (2.23). □

Proof of Theorem 2.4. As in Section 2.5.2, an optimal position for either the internal or the external agents is to concentrate on one of the losses X_i, $i \in [n]$. By the same arguments as in Theorem 2.3 (i), the equilibrium price, if it exists, must be of the form $\mathbf{p} = (p, \dots, p)$. For such a given $\mathbf{p}$, using the assumption that ρ_E and ρ_I are mildly monotone and Proposition 2.6, we can rewrite the optimization problems in (2.24) and (2.25) as

$$\min_{\mathbf{u}^j \in \mathbb{R}^n_+} \left\{ \rho_E \left(L_E(\mathbf{u}^j, \mathbf{p}) \right) + c_E(\|\mathbf{u}^j\|) \right\} = \min_{u \in \mathbb{R}_+} \left\{ u \left(\rho_E \left(X \right) - p \right) + c_E(u) \right\}, \qquad (2.33)$$

and

$$\min_{\mathbf{w}^i \in \mathbb{R}^n_+} \left\{ \rho_I \left(L_i(\mathbf{w}^i, \mathbf{p}) \right) + c_I(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|) \right\} = \min_{w \in \mathbb{R}_+} \left\{ w(\rho_I \left(X \right) - p) + ap + c_I(w - a) \right\}, \qquad (2.34)$$

for $j \in [m]$ and $i \in [n]$. Note that the derivative of the function inside the minimum of the right-hand side of (2.33) with respect to u is $L_E(u) - p$, and similarly, $L_I(w - a) - p$ is the derivative of the function inside the minimum of the right-hand side of (2.34). Using strict convexity of c_E and c_I, we get the following facts.

1. The optimizer u to (2.33) has two cases:
   (a) If $L_E(0) \geqslant p$, then $u = 0$.
   (b) If $L_E(0) < p$, then $u > 0$ and $L_E(u) = p$.
2. The optimizer w to (2.34) has four cases:
   (a) If $L_I^+(0) < p$, then $w > a$. This is not possible in an equilibrium.
   (b) If $L_I^+(0) \geqslant p \geqslant L_I^-(0)$, then $w = a$.
   (c) If $L_I^-(0) > p > L_I(-a)$, then $0 < w < a$ and $L_I(w - a) = p$.
   (d) If $L_I(-a) \geqslant p$, then $w = 0$.

From the above analysis, we see that the optimal positions for each different external agent are either all 0 or all positive, and they are identical due to the strict monotonicity of L_E. We can say the same for the internal agents. Suppose that there is an equilibrium. Let u be the external agents' common exposure, and w be the internal agent's exposure. By the clearance condition (2.26) we have $w + ku = a$. If $0 < ku < a$, then from (1.b) and (2.c) above, we have $L_E(u) = L_I(-ku)$. Below we show the three statements.

(i) If $L_E(a/k) < L_I(-a)$, then by strict monotonicity of L_E and L_I, there is no $u \in (0, a/k]$ such that $L_E(u) = L_I(-ku)$. Since u cannot be larger than a/k, if an equilibrium exists, then $u = 0$; but in this case, by (1.a) and (2.b), we have $L_E(0) \geqslant p \geqslant L_{I-}(0)$, which conflicts $L_E(a/k) < L_I(-a)$. Hence, there is no equilibrium.

(ii) In this case, there exists a unique $u^* \in (0, a/k]$ such that $L_E(u^*) = L_I(-ku^*)$. It follows that $u = u^*$ optimizes (2.33) and $w = a - ku^*$ optimizes (2.34). It is straightforward to verify that $\mathcal{E}$ is an equilibrium, and thus the "if" statement holds. To show the "only if" statement, it suffices to notice that $L_E(u) = L_I(-ku) = p$ has to hold, where p is an equilibrium price and u is the optimizer to (2.33), and such u and p are unique. Next, we show the "only if" statement for $u^* < a/2k$. As the optimal position for each external agent is $a - ku^* > a/2$, if more than two of the internal agents take the same loss, then the clearance condition (2.26) is broken. Hence, the internal agents have to take different losses. Moreover, as the optimal position for the internal agents are the same, the loss X_i for each $i \in [n]$, must be shared by one internal and k external agents. The equilibrium is preserved under the permutation of allocations. Thus, we have the "only if " statement for $u^* < a/2k$. The "if" statement is obvious.

(iii) The "if" statement can be verified directly using Theorem 2.3 (ii). Next, we show the "only if" statement. By (2.a), it is clear that the equilibrium price p satisfies $p \leqslant L_I^+(0)$. If $p < L_I^-(0)$, by (1.a), (2.c), and (2.d), the clearance condition (2.26) cannot be satisfied. Thus, $p \geqslant L_I^-(0)$. By a similar argument, we have $p \leqslant L_E(0)$. Hence, we get $p \in [L_I^-(0), L_E(0) \wedge L_I^+(0)]$. From (1.a) and (2.b), we have $u = 0$ and $w = a$ and thus the desired result. □

Proof of Proposition 2.7. The clearance condition (2.20) is clearly satisfied. As ES is translation invariant, it suffices to show that $\mathbf{w}^{i*}$ minimizes $\mathrm{ES}_q(\mathbf{w}^i \cdot \mathbf{X} - \mathbf{w}^i \cdot \mathbf{p}^*) + c_i(\|\mathbf{w}^i\| - \|\mathbf{a}^i\|)$ for $i \in [n]$. Write $r : \mathbf{w} \mapsto \mathrm{ES}_q(\mathbf{w} \cdot \mathbf{X})$ for $\mathbf{w} = (w_1, \dots, w_n) \in [0,1]^n$. By Corollary 4.2 of Tasche (2000),

$$\frac{\partial r}{\partial w_i}(\mathbf{w}) = \mathbb{E}\left[X_i | A_{\mathbf{w}}\right], \quad i \in [n],$$

where $A_{\mathbf{w}} = \{\sum_{i=1}^n w_i X_i \geqslant \mathrm{VaR}_q(\sum_{i=1}^n w_i X_i)\}$. Moreover, using convexity of r, we have (see McNeil et al. (2015, p. 321))

$$r(\mathbf{w}) - \mathbf{w} \cdot \mathbf{p}^* \geqslant \sum_{i=1}^n w_i \frac{\partial r}{\partial w_i}(a_1, \dots, a_n) - \mathbf{w} \cdot \mathbf{p}^* = 0.$$

By Euler's rule (see McNeil et al. (2015, (8.61))), the equality holds if $\mathbf{w} = \lambda(a_1, \dots, a_n)$ for any $\lambda > 0$. By taking $\lambda = a_i / \sum_{j=1}^n a_j$, we get $\|\mathbf{w}\| = a_i = \|\mathbf{a}^i\|$, and hence $c_i(\|\mathbf{w}\| - \|\mathbf{a}^i\|)$ is minimized by $\mathbf{w} = \lambda(a_1, \dots, a_n)$. Therefore, $\mathbf{w}^{i*}$ is an optimizer for each $i \in [n]$. □

Chapter 3

Ordering and Inequalities for Mixtures on Risk Aggregation

3.1 Introduction

Robust risk aggregation has been studied extensively with applications in banking and insurance. A typical problem in this area is to compute the worst-case values of some risk measures for an aggregate loss with unknown dependence structure. Two popular regulatory risk measures used in industry are Value-at-Risk (VaR) and the Expected Shortfall (ES); see McNeil et al. (2015) and the references therein. The worst-case value of ES in risk aggregation is explicit since ES is a coherent risk measure (Artzner et al. (1999)), whereas the worst-case value of VaR in risk aggregation generally does not admit analytical formulas, which is a known challenging problem (see e.g., Embrechts et al. (2013, 2015)). See Cai et al. (2018) on robust risk aggregation for general risk measures, and Eckstein et al. (2020) on computation of robust risk aggregation using neural networks.

The above robust risk aggregation problem involves taking the supremum of a risk measure over an *aggregation set*. Fix an atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{M}$ be the set of cdfs[1] on $\mathbb{R}$. For $F \in \mathcal{M}$, $X \sim F$ means that the cdf of a random variable X is F. Moreover, let $\mathcal{M}_1$ denote the set of cdfs on $\mathbb{R}$ with finite mean. For $\mathbf{F} = (F_1, \dots, F_n) \in \mathcal{M}^n$, the aggregation set (Bernard et al. (2014)) is defined as

$$\mathcal{D}_n(\mathbf{F}) = \{\text{cdf of } X_1 + \cdots + X_n : X_i \sim F_i,\ i = 1, \dots, n\}. \tag{3.1}$$

[1] In this chapter, we treat probability measures on $\mathcal{B}(\mathbb{R})$ and cdfs on $\mathbb{R}$ as equivalent objects.

The obvious interpretation is that $\mathcal{D}_n(\mathbf{F})$ fully describes model uncertainty associated with known marginal distributions $F_1, \dots, F_n$ but unknown dependence structure. The separate modeling of marginals and dependence is a standard practice in quantitative risk modeling, often involving copula techniques; see e.g., McNeil et al. (2015). An analytical characterization of $\mathcal{D}_n(\mathbf{F})$ for a given $\mathbf{F}$ is very difficult and challenging. The only available analytical results are in Mao et al. (2019) for standard uniform marginals.

The main objective of this chapter is to compare model uncertainty of risk aggregation for $\mathbf{F}, \mathbf{G} \in \mathcal{M}^n$ which represent two possible models of marginals. The strongest form of comparison is set inclusion between two aggregation sets $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\mathbf{G})$. It turns out that such a strong relation may be achievable if $\mathbf{F}, \mathbf{G} \in \mathcal{M}^n$ are related by the simple operations of distribution mixtures and quantile mixtures. Distribution mixture produces a tuple whose components are convex combinations of the given distributions and quantile mixture yields a tuple whose components are given by convex combinations of the given quantiles. Both types of operations are common in statistics and risk management, as they correspond to simple operations on the parameters in statistical models or on portfolio construction; see Section 3.7 for an example. Moreover, if $\mathbf{G}$ is obtained from $\mathbf{F}$ via a distribution or quantile mixture, then the mean (assumed to be finite) of any element of $\mathcal{D}_n(\mathbf{G})$ is the same as that of any element of $\mathcal{D}_n(\mathbf{F})$, making the comparison fair. To the best of our knowledge, this chapter is the first systematic study on the order relation between $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\mathbf{G})$ for different $\mathbf{F}$ and $\mathbf{G}$, thus comparing model uncertainty at the level of all possible distributions.

In some cases, a strong comparison via set inclusion is not possible, but we can compare values of a chosen risk measure. For a law-invariant risk measure[2] $\rho : \mathcal{M} \to \mathbb{R}$, we denote by $\overline{\rho}(\mathbf{F})$ the worst-case value of ρ in risk aggregation for $\mathbf{F} \in \mathcal{M}^n$, that is,

$$\overline{\rho}(\mathbf{F}) = \sup\{\rho(F) : F \in \mathcal{D}_n(\mathbf{F})\}.$$

We shall compare $\overline{\rho}(\mathbf{F})$ with $\overline{\rho}(\mathbf{G})$, thus the worst-case values of a risk measure under model uncertainty, which usually represent conservative calculation of regulatory risk capital (e.g., Embrechts et al. (2013)). Certainly, $D_n(\mathbf{F}) \subset D_n(\mathbf{G})$ implies $\overline{\rho}(\mathbf{F}) \leqslant \overline{\rho}(\mathbf{G})$ for all risk measures ρ, implying that the first comparison is stronger than the second one.[3]

Our study brings insights to two relevant problems in risk management. First, suppose that $\mathbf{F}$ and $\mathbf{G}$ are two possible statistical models for the marginal distributions in a risk ag-

[2]We conveniently treat law-invariant risk measures as mappings on $\mathcal{M}$, although it is conventional to treat them as mappings on a space of random variables. The two settings are equivalent for law-invariant risk measures.

[3]In this chapter, the set inclusion "$\subset$" is non-strict; the strict set inclusion is "$\subsetneq$". Similarly, the terms "increasing" and "decreasing" are in the non-strict sense.

gregation setting. Our results allow for a comparison of model uncertainty associated with the two models, regardless of the choice of risk measures. Although a completely unknown dependence structure is sometimes unrealistic, it is commonly agreed that the dependence structure in a risk model is difficult to accurately specify (e.g., Embrechts et al. (2013) and Bernard et al. (2017)). Hence, a comparison of the magnitude of model uncertainty is an important practical issue. On the other hand, the general conclusions remain valid even if the marginal distributions are not completely specific (see the discussion in Section 3.9 on the presence of marginal uncertainty), and thus the assumption of known marginal distributions in our study is not harmful.

Second, our results provide an analytical way to establish inequalities on the worst-case risk measures in the form $\overline{\rho}(\mathbf{F}) \leqslant \overline{\rho}(\mathbf{G})$. Sometimes the worst-case risk measure is difficult to calculate for $\mathbf{F}$, but it may be easier to calculate for $\mathbf{G}$. For instance, formulas on worst-case VaR are available for some homogeneous marginal distributions in Wang et al. (2013) and Puccetti and Rüschendorf (2013), but explicit results on heterogeneous marginal distributions are limited (see Blanchet et al. (2020) for a recent treatment). Therefore, we can use the analytical formula $\overline{\rho}(\mathbf{G})$, if available, as an upper bound on $\overline{\rho}(\mathbf{F})$, and this leads to interesting applications in other fields; see Section 3.7 for applications on portfolio diversification and multiple hypothesis testing and Section 3.8 for a connection to joint mixability.

Our theoretical contributions are briefly summarized below. In Sections 3.2 and 3.3, we analyze general relations on distribution and quantile mixtures. The general message of our results is that the more "homogeneous" the distribution tuple is, the larger its corresponding aggregation set $\mathcal{D}_n$ is. In particular, the set inclusion is established for any tuples connected by distribution mixtures in Theorem 3.1; that is, $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\mathbf{G})$ if $\mathbf{G}$ is a distribution mixture of $\mathbf{F}$. The problem for quantile mixtures is much more challenging. The set inclusion is established for uniform marginals in Proposition 3.2. For other families of distributions, such a general relationship does not hold, as discussed with some examples.

In Section 3.4, we obtain inequalities between the worst-case values of some risk measure ρ in risk aggregation with marginals related by distribution or quantile mixtures. Although quantile mixtures do not satisfy the relationship $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\mathbf{G})$ in general, we can prove an order property between $\overline{\rho}(\mathbf{G})$ and $\overline{\rho}(\mathbf{F})$ for commonly used risk measures. Most remarkably, in Theorem 3.3, we show that under a monotone density assumption, VaR satisfies this order property for a quantile mixture. Section 3.5 is dedicated to the most interesting special case of Pareto risk aggregation, with a special focus on the case of infinite mean.

Numerical results are presented in Section 3.6 to illustrate the obtained results. In

Section 3.7, we provide two applications: portfolio diversification under dependence uncertainty and merging p-values in multiple hypothesis testing. Some further technical discussions on distribution and quantile mixtures are put in Section 3.8. Section 3.9 concludes the chapter by presenting several open mathematical challenges related to quantile mixtures. Some proofs and further properties of Pareto risk aggregation are put in Section 3.10.

3.2 Distribution mixtures

In this section we put our focus on one of the two operations: distribution mixture. The main objective is to establish some ordering relationships on the set $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\mathbf{G})$ where $\mathbf{G}$ is a distribution mixture of $\mathbf{F}$. For greater generality, we investigate a more general f-aggregation set $\mathcal{D}_f(\mathbf{F})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a measurable and symmetric function.[4] Similarly to (3.1), for $\mathbf{F} = (F_1, \dots, F_n) \in \mathcal{M}^n$, the f-aggregation set is defined as

$$\mathcal{D}_f(\mathbf{F}) = \{\text{cdf of } f(X_1, \dots, X_n) : X_i \sim F_i,\ i = 1, \dots, n\}.$$

It is clear that $\mathcal{D}_n$, defined in (3.1), becomes a specific case of D_f if f is a sum function ($f(x_1, \dots, x_n) = \sum_{j=1}^n x_j$). We first present some properties of the f-aggregation set.

Lemma 3.1. *For an n-symmetric function $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{F}, \mathbf{G} \in \mathcal{M}^n$, $\lambda \in [0,1]$ and an n-permutation π, the following hold.*

(i) $\mathcal{D}_f(\mathbf{F}) = \mathcal{D}_f(\pi(\mathbf{F}))$.

(ii) $\lambda\mathcal{D}_f(\mathbf{F}) + (1-\lambda)\mathcal{D}_f(\mathbf{G}) \subset \mathcal{D}_f(\lambda\mathbf{F} + (1-\lambda)\mathbf{G})$. *In particular,*

- *(a)* $\lambda\mathcal{D}_f(\mathbf{F}) + (1-\lambda)\mathcal{D}_f(\mathbf{F}) = \mathcal{D}_f(\mathbf{F})$.
- *(b)* $\mathcal{D}_f(\mathbf{F}) \cap \mathcal{D}_f(\mathbf{G}) \subset \mathcal{D}_f(\lambda\mathbf{F} + (1-\lambda)\mathbf{G})$.

Proof. (i) holds because of the symmetry of f. To prove (ii), for any $H \in \lambda\mathcal{D}_f(\mathbf{F}) + (1-\lambda)\mathcal{D}_f(\mathbf{G})$, there exist $X_1 \sim F_1, \dots, X_n \sim F_n$, $Y_1 \sim G_1, \dots, Y_n \sim G_n$ and an event $A \in \mathcal{F}$ independent of $X_1, \dots, X_n, Y_1, \dots, Y_n$ such that $\mathbb{P}(A) = \lambda$ and

$$(f(X_1, \dots, X_n)\mathbb{1}_A + f(Y_1, \dots, Y_n)\mathbb{1}_{A^c}) \sim H.$$

[4] A function f is symmetric if $f(\mathbf{x}) = f(\pi(\mathbf{x}))$ for any $\mathbf{x} \in \mathbb{R}^n$ and n-permutation π.

We notice that

$$f(X_1,\dots,X_n)\mathbb{1}_A + f(Y_1,\dots,Y_n)\mathbb{1}_{A^c} = f\left(X_1\mathbb{1}_A + Y_1\mathbb{1}_{A^c},\dots,X_n\mathbb{1}_A + Y_n\mathbb{1}_{A^c}\right),$$

and $(X_i\mathbb{1}_A + Y_i\mathbb{1}_{A^c}) \sim \lambda F_i + (1-\lambda)G_i$ for any $i = 1,\dots,n$. Thus we have $H \in \mathcal{D}_f(\lambda\mathbf{F} + (1-\lambda)\mathbf{G})$. This completes the proof of (ii). □

We briefly fix some notation and convention. Let Δ_n be the standard simplex given by $\Delta_n = \{(\lambda_1,\dots,\lambda_n) \in [0,1]^n : \sum_{i=1}^n \lambda_i = 1\}$. Recall that a doubly stochastic matrix is a square matrix of nonnegative real numbers, each of whose rows and columns sums to 1 (i.e. each row or column is in Δ_n). Denote by $\mathcal{Q}_n$ the set of $n \times n$ doubly stochastic matrices. All vectors should be treated as column vectors. For $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_n) \in \Delta_n$ and $\mathbf{F} = (F_1,\dots,F_n) \in \mathcal{M}^n$, their dot product is $\boldsymbol{\lambda}\cdot\mathbf{F} = \sum_{i=1}^n \lambda_i F_i \in \mathcal{M}$. For a matrix $\Lambda = (\boldsymbol{\lambda}_1,\dots,\boldsymbol{\lambda}_n)^\top \in \mathcal{Q}_n$ and $\mathbf{F} \in \mathcal{M}^n$, their product is $\Lambda\mathbf{F} = (\boldsymbol{\lambda}_1\cdot\mathbf{F},\dots,\boldsymbol{\lambda}_n\cdot\mathbf{F}) \in \mathcal{M}^n$.

The vector $\Lambda\mathbf{F}$ is a distribution mixture of $\mathbf{F}$, and we will call it the Λ-*mixture* of $\mathbf{F}$ to emphasize the reliance on Λ. Indeed, $\Lambda\mathbf{F}$ can be seen as a vector of weighted averages of $\mathbf{F}$. In particular, by choosing $\Lambda = (\frac{1}{n})_{n\times n}$ (here $(x)_{n\times n}$ means an $n\times n$ matrix with identical number $x \in \mathbb{R}$), we get the vector $(F,\dots,F)$ where F is the average of components of $\mathbf{F}$. Note that if $\mathbf{F} \in \mathcal{M}_1^n$, then the mean of any element of $\mathcal{D}_n(\mathbf{F})$ is the same as that of $\mathcal{D}_n(\Lambda\mathbf{F})$.

The first result below suggests that the set of aggregation for a tuple of distributions is smaller than that for the weighted averages. The proof is elementary, but the result allows us to observe the important phenomenon that *more homogeneous marginals lead to a larger aggregation set*.

Theorem 3.1. *For an n-symmetric function $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{F} \in \mathcal{M}^n$ and $\Lambda \in \mathcal{Q}_n$, $\mathcal{D}_f(\mathbf{F}) \subset \mathcal{D}_f(\Lambda\mathbf{F})$. In particular, $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\Lambda\mathbf{F})$.*

Proof. Let $\Pi_1,\dots,\Pi_{n!}$ be all different n-permutation matrices, i.e. $\Pi_k\mathbf{F}$ is a permutation of $\mathbf{F}$. By Birkhoff's Theorem (Theorem 2.A.2 of Marshall et al. (2011)), the set $\mathcal{Q}_n$ of doubly stochastic matrices is the convex hull of permutation matrices, that is, for any $\Lambda \in \mathcal{Q}_n$, there exists $(\lambda_1,\dots,\lambda_{n!}) \in \Delta_{n!}$, such that

$$\Lambda = \sum_{k=1}^{n!} \lambda_k \Pi_k.$$

Note that $\mathcal{D}_f(\mathbf{F}) = \mathcal{D}_f(\Pi_k \mathbf{F})$ for $k = 1, \dots, n!$ by Lemma 3.1(i). Further, by Lemma 3.1(ii-b), we have,

$$\mathcal{D}_f(\mathbf{F}) = \bigcap_{k=1}^{n!} \mathcal{D}_f(\Pi_k \mathbf{F}) \subset \mathcal{D}_f\left(\sum_{k=1}^{n!} \lambda_k \Pi_k(\mathbf{F})\right) = \mathcal{D}_f(\Lambda \mathbf{F}).$$

This completes the theorem. □

As the sum aggregation is the most common in financial applications, we will mainly discuss $\mathcal{D}_n$ instead of $\mathcal{D}_f$ in the following context, while keeping in mind that most results on $\mathcal{D}_n$ can be extended naturally to $\mathcal{D}_f$.

Corollary 3.1. *For* $\mathbf{F} = (F_1, \dots, F_n) \in \mathcal{M}^n$ *and* $\Lambda \in \mathcal{Q}_n$, $\mathcal{D}_n(\Lambda \mathbf{F}) \subset \mathcal{D}_n(F, \dots, F)$ *where* $F = \frac{1}{n} \sum_{i=1}^n F_i$.

By taking Λ as the identity in Corollary 3.1, we obtain the set inclusion $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(F, \dots, F)$, which was given in Theorem 3.5 of Bernard et al. (2014) to find the bounds on VaR for heterogeneous marginal distributions.

The doubly stochastic matrices are closely related to majorization order. For $\boldsymbol{\lambda}, \boldsymbol{\gamma} \in \mathbb{R}^n$, we say that $\boldsymbol{\lambda}$ dominates $\boldsymbol{\gamma}$ in *majorization order*, denoted by $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$, if $\sum_{i=1}^n \phi(\gamma_i) \leqslant \sum_{i=1}^n \phi(\lambda_i)$ for all continuous convex functions ϕ. There are several equivalent conditions for this order; see Section 1.A.3 of Marshall et al. (2011). One equivalent condition that is relevant to Theorem 3.1 is that $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$ if and only if there exists $\Lambda \in \mathcal{Q}_n$ such that $\boldsymbol{\gamma} = \Lambda \boldsymbol{\lambda}$. We can similarly define majorization order between $\mathbf{F}, \mathbf{G} \in \mathcal{M}^n$, denoted by $\mathbf{G} \prec \mathbf{F}$, if $\mathbf{G} = \Lambda \mathbf{F}$ for some $\Lambda \in \mathcal{Q}_n$. Then, we have the following corollary.

Corollary 3.2. *For* $\mathbf{F}, \mathbf{G} \in \mathcal{M}^n$, *if* $\mathbf{G} \prec \mathbf{F}$, *then* $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\mathbf{G})$.

Example 3.1 (Bernoulli distributions)**.** We apply Theorem 3.1 to Bernoulli distributions. Let B_p be a Bernoulli cdf with (mean) parameter $p \in [0, 1]$. Note that a mixture of Bernoulli distributions is still Bernoulli, and more precisely, for $\mathbf{p} = (p_1, \dots, p_n) \in [0, 1]^n$ and $\mathbf{q} = (q_1, \dots, q_n) = \Lambda \mathbf{p}$, we have $\Lambda(B_{p_1}, \dots, B_{p_n}) = (B_{q_1}, \dots, B_{q_n})$. Therefore, by Theorem 3.1, for any $\mathbf{p}, \mathbf{q} \in [0, 1]^n$ with $\mathbf{q} \prec \mathbf{p}$, we have $\mathcal{D}_n(B_{p_1}, \dots, B_{p_n}) \subset \mathcal{D}_n(B_{q_1}, \dots, B_{q_n})$. This result will be used later to discuss joint mixability (see Section 3.8) of Bernoulli distributions. For instance, we can set $\mathbf{p} = (0.2, 0.8)$,

$$\Lambda = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}, \text{ and } \mathbf{q} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix} (0.2, 0.8) = (0.65, 0.35).$$

Note that $\Lambda(B_{0.2}, B_{0.8}) = (B_{0.65}, B_{0.35})$. Hence $\mathcal{D}_n(B_{0.2}, B_{0.8}) \subset \mathcal{D}_n(B_{0.65}, B_{0.35})$.

Next, we discuss how Λ-mixtures affect the lower sets with respect to convex order. A distribution $F \in \mathcal{M}_1$ is called smaller than a distribution $G \in \mathcal{M}_1$ in *convex order*, denoted by $F \prec_{\mathrm{cx}} G$, if

$$\int \phi \,\mathrm{d}F \leqslant \int \phi \,\mathrm{d}G \quad \text{for all convex} \quad \phi : \mathbb{R} \to \mathbb{R}, \tag{3.2}$$

provided that both integrals exist (finite or infinite); see Müller and Stoyan (2002) and Shaked and Shanthikumar (2007) for an overview on convex order and the related notion of second-order stochastic dominance. For a given distribution $F \in \mathcal{M}_1$, denote by $\mathcal{C}(F)$ the set of all distributions in $\mathcal{M}_1$ dominated by F in convex order, that is,

$$\mathcal{C}(F) = \{G \in \mathcal{M}_1 : G \prec_{\mathrm{cx}} F\}.$$

For any distributions F and G, we denote by $F \oplus G$ the distribution with quantile function $F^{-1} + G^{-1}$.[5] Moreover, define

$$\mathcal{C}(F_1, \dots, F_n) = \mathcal{C}(F_1 \oplus \cdots \oplus F_n).$$

The following lemmas give a simple link between the sets $\mathcal{D}_n$ and $\mathcal{C}$; see e.g., Lemma 1 of Mao et al. (2019).

Lemma 3.2. *For* $\mathbf{F} \in \mathcal{M}_1^n$, $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{C}(\mathbf{F})$.

Similarly to the set $\mathcal{D}_n(\mathbf{F})$ in Theorem 3.1, $\mathcal{C}(\mathbf{F})$ also satisfies an order with respect to Λ-mixture.

Theorem 3.2. *For* $\mathbf{F} \in \mathcal{M}_1^n$ *and* $\Lambda \in \mathcal{Q}_n$, *we have* $\mathcal{C}(\mathbf{F}) \subset \mathcal{C}(\Lambda\mathbf{F})$.

Proof. Note that $F_1 \oplus \cdots \oplus F_n \in \mathcal{D}_n(\mathbf{F})$ since $F_1 \oplus \cdots \oplus F_n$ corresponds to the sum of comonotonic random variables with respective distributions $F_1, \dots, F_n$. Using Theorem 3.1 and Lemma 3.2, we have $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\Lambda\mathbf{F}) \subset \mathcal{C}(\Lambda\mathbf{F})$. This implies $F_1 \oplus \cdots \oplus F_n \in \mathcal{C}(\Lambda\mathbf{F})$. By definition, $\mathcal{C}(\mathbf{F}) \subset \mathcal{C}(\Lambda\mathbf{F})$. □

[5] In other words, $F \oplus G$ is the distribution of the sum of two comonotonic random variables with respective distributions F and G. Two random variables X and Y are said to be *comonotonic*, if there exists a random variable U and two increasing functions f, g such that $X = f(U)$ and $Y = g(U)$ almost surely. Such U can be chosen as a standard uniform random variable ($U \sim \mathrm{U}[0,1]$), and f and g can be chosen as the inverse distribution functions of X and Y, respectively.

3.3 Quantile mixtures

In Section 3.2, we have seen a set inclusion between $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\mathbf{G})$ where $\mathbf{G}$ is a distribution mixture of $\mathbf{F}$. The general message from Theorem 3.1 is that distribution mixtures enlarge the aggregation sets. As distribution mixture corresponds to the arithmetic average of distribution functions, it would then be of interest to see whether a "harmonic average" of $F_1, \dots, F_n$ would give similar properties. By saying "harmonic average" of $F_1, \dots, F_n$, we mean the distribution F with $F^{-1} = \frac{1}{n}\sum_{i=1}^n F_i^{-1}$, i.e., the average of quantiles. We shall call this type of average as *quantile mixture*.

In many statistical applications, marginal distributions of a multi-dimensional object are modelled in the same location-scale family (such as Gaussian, elliptical, or uniform family). The quantile mixture of such distributions is still in the same family, whereas the distribution mixture is typically no longer in the family. Moreover, a quantile mixture also corresponds to the combination of comonotonic random variables (such as combining an asset price with a call option on it), and hence finds its natural position in finance. As such, it is rather important and practical to consider quantile mixtures.

Remark 3.1. The two types of mixtures are both basic operations on distributions and often lead to qualitatively very different mathematical results. As a famous example in decision theory, the axiom of linearity on distribution mixtures leads to the classic von Neumann-Morgenstern expected utility theory, whereas the axiom of linearity on quantile mixtures leads to the dual utility theory of Yaari (1987).

For a matrix Λ of non-negative elements (not necessarily in $\mathcal{Q}_n$) and $\mathbf{F} \in \mathcal{M}^n$, let $\Lambda \otimes \mathbf{F}$ be a vector of distributions $\mathbf{G}$ such that componentwise, $\mathbf{G}^{-1}$ is equal to $\Lambda\mathbf{F}^{-1}$. If $\Lambda \in \mathcal{Q}_n$, we call $\mathbf{G} = \Lambda \otimes \mathbf{F}$ the Λ-*quantile mixture* of $\mathbf{F}$. If $\mathbf{F} \in \mathcal{M}_1^n$, then the mean of any element of $\mathcal{D}_n(\mathbf{F})$ is the same as that of $\mathcal{D}_n(\Lambda \otimes \mathbf{F})$, similarly to the case of distribution mixture. This suggests that one may compare $\mathcal{D}_n(\mathbf{F})$ with $\mathcal{D}_n(\Lambda \otimes \mathbf{F})$, just like what we did in Section 3.2 for distribution mixture.

The first natural candidates for us to look at are $\mathcal{D}_n(F_1, \dots, F_n)$ and $\mathcal{D}_n(F, \dots, F)$ where $F^{-1} = \frac{1}{n}\sum_{i=1}^n F_i^{-1}$, thus the quantile version of Corollary 3.1. Unfortunately, the sets $\mathcal{D}_n(F_1, \dots, F_n)$ and $\mathcal{D}_n(F, \dots, F)$ are not necessarily comparable, as seen from the following example.

Example 3.2. Take F_1 as a binary uniform distribution (with probability $1/2$ at each point) on $\{0, 1\}$ and F_2 as a binary uniform distribution on $\{0, 3\}$. Clearly, F is a binary uniform distribution on $\{0, 2\}$. $\mathcal{D}_2(F_1, F_2)$ contains distributions supported on $\{0, 1, 3, 4\}$

and $\mathcal{D}_2(F,F)$ contains distributions supported on $\{0,2,4\}$. Therefore, these two sets do not have a relation of set inclusion.

On the other hand, as a trivial example, if $F_2,\dots,F_n$ are point masses (without loss of generality, we assume that they are point masses at 0), then F satisfies $F^{-1}=F_1^{-1}/n$. In this case, $\mathcal{D}_n(F_1,\dots,F_n)=\{F_1\}\subset\mathcal{D}_n(F,\dots,F)$ holds trivially. Therefore, we can expect that the inclusion $\mathcal{D}_n(F_1,\dots,F_n)\subset\mathcal{D}_n(F,\dots,F)$ may hold under some special settings.

Below, we note that both $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\Lambda\otimes\mathbf{F})$ have the same convex-order maximal element. This is in sharp contrast to the case of mixtures in Theorem 3.2. Proposition 3.1 can be verified directly by definition.

Proposition 3.1. *For $\mathbf{F}\in\mathcal{M}_1^n$ and $\Lambda\in\mathcal{Q}_n$, we have $\mathcal{C}(\mathbf{F})=\mathcal{C}(\Lambda\otimes\mathbf{F})$.*

As we see from Example 3.2, $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\Lambda\otimes\mathbf{F})$ are not necessarily comparable. In Mao et al. (2019), a non-trivial result is established for the aggregation of standard uniform distributions, which leads to an interesting observation along this direction.

Proposition 3.2. *Suppose that $F_1,\dots,F_n$ are uniform distributions, $n\geqslant 3$, and $\Lambda=(\frac{1}{n})_{n\times n}$. Then $\mathcal{D}_n(\mathbf{F})\subset\mathcal{D}_n(\Lambda\otimes\mathbf{F})$.*

Proof. Note that the components of $\Lambda\otimes\mathbf{F}$ are uniform distributions with equal length. By Theorem 5 of Mao et al. (2019), we have $\mathcal{D}_n(\Lambda\otimes\mathbf{F})=\mathcal{C}_n(\Lambda\otimes\mathbf{F})$. Using Proposition 3.1, we have $\mathcal{C}_n(\mathbf{F})=\mathcal{C}_n(\Lambda\otimes\mathbf{F})$. Lemma 3.2 further yields $\mathcal{D}_n(\mathbf{F})\subset\mathcal{C}_n(\mathbf{F})$. Putting the above results together, we obtain $\mathcal{D}_n(\mathbf{F})\subset\mathcal{D}_n(\Lambda\otimes\mathbf{F})$. □

It is unclear whether $\mathcal{D}_n(\mathbf{F})\subset\mathcal{D}_n(\Lambda\otimes\mathbf{F})$ under some other conditions, similarly to Proposition 3.2. Note that the set inclusion $\mathcal{D}_n(\mathbf{F})\subset\mathcal{D}_n(\Lambda\otimes\mathbf{F})$ would help us to obtain semi-explicit formulas for bounds on risk measures (such as VaR), since by choosing $\Lambda=(\frac{1}{n})_{n\times n}$, the marginal distributions of $\Lambda\otimes\mathbf{F}$ are the same, and formulas for VaR bounds in e.g., Wang et al. (2013) and Bernard et al. (2014) are applicable; see Section 3.4.

There are several sharp contrasts regarding distribution and quantile mixtures. In addition to the contrast on order relations that we see from Theorem 3.1 and Example 3.2, the two notions also treat location shifts on the marginal distributions very differently. This point will be explained in Section 3.8.1.

3.4 Bounds on the worst-case values of risk measures

This section is dedicated to exploring the inequalities between the worst-cases value of risk measures in risk aggregation with different marginal distribution tuples. Our main results in Sections 3.2 and 3.3 will help to find the inequalities in Proposition 3.5.

3.4.1 Risk measures

We pay a particular attention to the popular regulatory risk measure VaR, which is a quantile functional. For $F \in \mathcal{M}$, for $p \in (0,1)$, define the risk measure $\mathrm{VaR}_p : \mathcal{M} \to \mathbb{R}$ as

$$\mathrm{VaR}_p(F) = F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geqslant p\}.$$

Another popular regulatory risk measure is $\mathrm{ES}_p : \mathcal{M}_1 \to \mathbb{R}$ for $p \in (0,1)$, given by

$$\mathrm{ES}_p(F) = \frac{1}{1-p}\int_p^1 F^{-1}(u)\mathrm{d}u.$$

Given marginals $\mathbf{F}$, the worst-case value of VaR in risk aggregation with unknown dependence structure is then defined as

$$\overline{\mathrm{VaR}}_p(\mathbf{F}) = \sup\{\mathrm{VaR}_p(G) : G \in \mathcal{D}_n(\mathbf{F})\}.$$

In other words, $\overline{\mathrm{VaR}}_p(\mathbf{F})$ is the largest value of VaR_p of the aggregate risk $X_1+\cdots+X_n$ over all possible dependence structures among $X_i \sim F_i,\ i = 1,\dots,n$. Similarly, the worst-case value of ES in risk aggregation is defined as $\overline{\mathrm{ES}}_p(\mathbf{F}) = \sup\{\mathrm{ES}_p(G) : G \in \mathcal{D}_n(\mathbf{F})\}$.

The worst-case value of ES in risk aggregation is easy to calculate since ES is consistent with convex order. On the other hand, worst-case value of VaR in risk aggregation generally does not admit any analytical formula, which is a challenging problem; results under some specific cases are given in Wang et al. (2013), Puccetti and Rüschendorf (2013) and Bernard et al. (2014). To obtain approximations for $\overline{\mathrm{VaR}}_p(\mathbf{F})$, one may use the asymptotic equivalence between VaR and ES in Embrechts et al. (2015) and then directly apply ES bounds, or use a numerical algorithm such as the rearrangement algorithm of Puccetti and Rüschendorf (2012) and Embrechts et al. (2013).

We will discuss a general relationship on risk measures for different aggregation sets. A *risk measure* is a functional $\rho : \mathcal{M}_\rho \to \mathbb{R}$, where $\mathcal{M}_\rho \subset \mathcal{M}$ is the set of distributions of some financial losses. For instance, if ρ is the mean, then $\mathcal{M}_\rho$ is naturally chosen as the set of distributions with finite mean. We denote by $\overline{\rho}(\mathbf{F})$ the worst-case value of ρ in risk aggregation for $\mathbf{F} \in \mathcal{M}^n$, that is, assuming $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{M}_\rho$,

$$\overline{\rho}(\mathbf{F}) = \sup\{\rho(G) : G \in \mathcal{D}_n(\mathbf{F})\}.$$

3.4.2 Inequalities implied by stochastic dominance

Quite obviously, one can compare the worst-case values of some risk measures for two tuples of distributions satisfying some stochastic dominance, which we briefly discuss here.

A distribution $F \in \mathcal{M}$ is smaller than a distribution G in *stochastic order* (also first-order stochastic dominance), denoted by $F \prec_{\mathrm{st}} G$, if $F \geqslant G$. For $\mathbf{F}, \mathbf{G} \in \mathcal{M}^n$, we say that $\mathbf{F}$ is smaller than $\mathbf{G}$ in stochastic order, denoted by $\mathbf{F} \prec_{\mathrm{st}} \mathbf{G}$, if $F_i \prec_{\mathrm{st}} G_i$, $i = 1, \dots, n$. Analogously, for $\mathbf{F}, \mathbf{G} \in \mathcal{M}_1^n$, we say that $\mathbf{F}$ is smaller than $\mathbf{G}$ in convex order, denoted by $\mathbf{F} \prec_{\mathrm{cx}} \mathbf{G}$, if $F_i \prec_{\mathrm{cx}} G_i$, $i = 1, \dots, n$.

We define two relevant common properties of risk measures. A risk measure ρ is *monotone* if $\rho(F) \leqslant \rho(G)$ whenever $F \prec_{\mathrm{st}} G$; it is *consistent with convex order* if $\rho(F) \leqslant \rho(G)$ whenever $F \prec_{\mathrm{cx}} G$. Almost all risk measures used in practice are monotone; ES is consistent with convex order whereas VaR is not. Monetary risk measures (see Föllmer and Schied (2016)) that are consistent with convex order are characterized by Mao and Wang (2020) and they admit an ES-based representation. In particular, all lower semi-continuous convex risk measures, including ES and expectiles (e.g., Ziegel (2016) and Delbaen et al. (2016)), are consistent with convex order; we refer to Föllmer and Schied (2016) for an overview on risk measures.

Now we state in Proposition 3.3 that one can compare the worst-case values of some risk measures for $\mathbf{F}$ and $\mathbf{G}$ if $\mathbf{F}$ is smaller than $\mathbf{G}$ in stochastic order or convex order.

Proposition 3.3. *Let ρ be a risk measure and $\mathbf{F}, \mathbf{G} \in \mathcal{M}^n$ with $\mathcal{D}_n(\mathbf{F}), \mathcal{D}_n(\mathbf{G}) \subset \mathcal{M}_\rho$.*

(i) If ρ is monotone and $\mathbf{F} \prec_{\mathrm{st}} \mathbf{G}$, then $\overline{\rho}(\mathbf{F}) \leqslant \overline{\rho}(\mathbf{G})$.

(ii) If ρ is consistent with convex order and $\mathbf{F} \prec_{\mathrm{cx}} \mathbf{G}$ with $\mathbf{F}, \mathbf{G} \in \mathcal{M}_1^n$, then $\overline{\rho}(\mathbf{F}) \leqslant \overline{\rho}(\mathbf{G})$.

Proof. (i) is straightforward to verify. We next focus on (ii). Since $F_1 \oplus \cdots \oplus F_n$ is the largest distribution in $D_n(\mathbf{F})$ with respect to convex order and ρ is consistent with convex order, we have $\overline{\rho}(\mathbf{F}) = \rho(F_1 \oplus \cdots \oplus F_n)$. Similarly, $\overline{\rho}(\mathbf{G}) = \rho(G_1 \oplus \cdots \oplus G_n)$. Note that $\mathbf{F} \prec_{\mathrm{cx}} \mathbf{G}$ means $F_i \prec_{\mathrm{cx}} G_i$, $i = 1, \dots, n$. For all $p \in (0,1)$, using comonotonic-additivity of ES_p, we have

$$\mathrm{ES}_p(F_1 \oplus \cdots \oplus F_n) = \sum_{i=1}^{n} \mathrm{ES}_p(F_i) \leqslant \sum_{i=1}^{n} \mathrm{ES}_p(G_i) = \mathrm{ES}_p(G_1 \oplus \cdots \oplus G_n),$$

which gives $F_1 \oplus \cdots \oplus F_n \prec_{\mathrm{cx}} G_1 \oplus \cdots \oplus G_n$ (see e.g., Theorem 3.A.5 of Shaked and Shanthikumar (2007)). □

In the following result, we will show that the distribution tuples and their Λ-mixture or Λ-quantile mixture typically do not satisfy stochastic order or convex order, unless the mixture operation is essentially identical ($\Lambda\mathbf{F} = \mathbf{F}$ or $\Lambda \otimes \mathbf{F} = \mathbf{F}$). The proof of Proposition 3.4 is put in Section 3.10.

Proposition 3.4. *Suppose $\Lambda \in \mathcal{Q}_n$. The statements within each of (i)-(iv) are equivalent.*

(i) For $\mathbf{F} \in \mathcal{M}^n$, (a) $\Lambda\mathbf{F} \prec_{\mathrm{st}} \mathbf{F}$; (b) $\mathbf{F} \prec_{\mathrm{st}} \Lambda\mathbf{F}$; (c) $\Lambda\mathbf{F} = \mathbf{F}$.

(ii) For $\mathbf{F} \in \mathcal{M}^n$, (a) $\Lambda \otimes \mathbf{F} \prec_{\mathrm{st}} \mathbf{F}$; (b) $\mathbf{F} \prec_{\mathrm{st}} \Lambda \otimes \mathbf{F}$; (c) $\Lambda \otimes \mathbf{F} = \mathbf{F}$.

(iii) For $\mathbf{F} \in \mathcal{M}_1^n$, (a) $\Lambda \otimes \mathbf{F} \prec_{\mathrm{cx}} \mathbf{F}$; (b) $\mathbf{F} \prec_{\mathrm{cx}} \Lambda \otimes \mathbf{F}$; (c) $\Lambda \otimes \mathbf{F} = \mathbf{F}$.

(iv) For $\mathbf{F} \in \mathcal{M}_1^n$, (a) $\Lambda\mathbf{F} \prec_{\mathrm{cx}} \mathbf{F}$; (b) $\Lambda\mathbf{F} = \mathbf{F}$.

An implication of Proposition 3.4 is that the result on stochastic order in Proposition 3.3 cannot be applied to compare the worst-case values of risk measures for $\mathbf{F}$ and $\Lambda\mathbf{F}$ or $\mathbf{F}$ and $\Lambda \otimes \mathbf{F}$. Nevertheless, this comparison can be conducted by applying our findings in Sections 3.2 and 3.3 and some other techniques. This will be the task in the next subsection.

3.4.3 Inequalities generated by distribution/quantile mixtures

In the following, we will obtain inequalities between the worst-case values of risk measures for $\mathbf{F}$ and $\Lambda\mathbf{F}$ or $\mathbf{F}$ and $\Lambda \otimes \mathbf{F}$. First, we apply Theorem 3.1 and Proposition 3.1 and immediately obtain the following result.

Proposition 3.5. *Let ρ be a risk measure and $\Lambda \in \mathcal{Q}_n$.*

(i) For $\mathbf{F} \in \mathcal{M}^n$ with $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{M}_\rho$ and $\mathcal{D}_n(\Lambda\mathbf{F}) \subset \mathcal{M}_\rho$, we have $\overline{\rho}(\mathbf{F}) \leqslant \overline{\rho}(\Lambda\mathbf{F})$;

(ii) For $\mathbf{F} \in \mathcal{M}_1^n$ with $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{M}_\rho$ and $\mathcal{D}_n(\Lambda \otimes \mathbf{F}) \subset \mathcal{M}_\rho$, if ρ is consistent with convex order, then $\overline{\rho}(\mathbf{F}) = \overline{\rho}(\Lambda \otimes \mathbf{F}) = \rho(F_1 \oplus \cdots \oplus F_n)$.

Note that in Proposition 3.5, the inequality for distribution mixture is valid for all risk measures whereas the equality for quantile mixture is constrained to risk measures consistent with convex order. As ES_p is a special case of risk measures consistent with convex order, we immediately get $\overline{\mathrm{ES}}_p(\mathbf{F}) \leqslant \overline{\mathrm{ES}}_p(\Lambda\mathbf{F})$ and $\overline{\mathrm{ES}}_p(\mathbf{F}) = \overline{\mathrm{ES}}_p(\Lambda \otimes \mathbf{F})$. Since VaR is not consistent with convex order, (ii) of Proposition 3.5 cannot be applied to VaR.

Nevertheless, using a recent result on $\overline{\mathrm{VaR}}$ in Blanchet et al. (2020), we obtain an inequality between $\overline{\mathrm{VaR}}$ for some special marginals and $\overline{\mathrm{VaR}}$ of their corresponding quantile mixture. Denote by $\mathcal{M}_D$ (respectively, $\mathcal{M}_I$) the set of distributions with decreasing (respectively, increasing) densities on their support. Moreover, let $\mathcal{M}_D^n = (\mathcal{M}_D)^n$ and $\mathcal{M}_I^n = (\mathcal{M}_I)^n$.

Theorem 3.3. *For $p \in (0,1)$, $\Lambda \in \mathcal{Q}_n$, and $\mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n$, we have*

$$\overline{\mathrm{VaR}}_p(\mathbf{F}) \leqslant \overline{\mathrm{VaR}}_p(\Lambda \otimes \mathbf{F}).$$

Proof. We start with some preliminaries. Define the upper VaR at level p for a cdf F as

$$\mathrm{VaR}_p^*(F) = \inf\{x \in \mathbb{R} : F(x) > p\}, \quad p \in (0,1).$$

The worst-case value of the upper VaR in risk aggregation is $\overline{\mathrm{VaR}}_p^*(\mathbf{F}) = \sup\{\mathrm{VaR}_p^*(G) : G \in \mathcal{D}_n(\mathbf{F})\}$. For $\mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n$ and $p \in (0,1)$, Lemma 4.5 of Bernard et al. (2014) gives

$$\overline{\mathrm{VaR}}_p^*(\mathbf{F}) = \overline{\mathrm{VaR}}_p(\mathbf{F}).$$

Using Lemma 3.3 in Section 3.10.1 (paraphrased from Theorem 2 of Blanchet et al. (2020)), we have

$$\overline{\mathrm{VaR}}_p(\mathbf{F}) = \inf_{\boldsymbol{\beta} \in \mathbb{B}_n} \sum_{i=1}^n \frac{1}{(1-p)(1-\beta)} \int_{p+(1-p)(\beta-\beta_i)}^{1-(1-p)\beta_i} \mathrm{VaR}_u(F_i)\mathrm{d}u, \tag{3.3}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$, $\beta = \sum_{i=1}^n \beta_i$ and $\mathbb{B}_n = \{\boldsymbol{\beta} \in [0,1)^n : \beta < 1\}$. Note that

$$\Lambda \otimes \mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n \ \text{ if } \ \mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n.$$

Consequently, for $p \in (0,1)$,

$$\begin{aligned}
\overline{\mathrm{VaR}}_p(\Lambda \otimes \mathbf{F}) &= \inf_{\boldsymbol{\beta} \in \mathbb{B}_n} \sum_{i=1}^n \frac{1}{(1-p)(1-\beta)} \int_{p+(1-p)(\beta-\beta_i)}^{1-(1-p)\beta_i} \left(\sum_{j=1}^n \Lambda_{i,j} \mathrm{VaR}_u(F_j) \right) \mathrm{d}u \\
&= \inf_{\boldsymbol{\beta} \in \mathbb{B}_n} \sum_{i=1}^n \sum_{j=1}^n \Lambda_{i,j} M_{i,j}(\boldsymbol{\beta}),
\end{aligned}$$

where the function $M : \mathbb{B}_n \to \mathbb{R}^{n \times n}$, mapping an n-dimensional vector to an $n \times n$ matrix, is given by

$$M_{i,j}(\boldsymbol{\beta}) = \frac{1}{(1-p)(1-\beta)} \int_{p+(1-p)(\beta-\beta_i)}^{1-(1-p)\beta_i} \mathrm{VaR}_u(F_j)\mathrm{d}u, \quad i,j = 1, \dots, n.$$

We can rewrite (3.3) as

$$\overline{\mathrm{VaR}}_p(\mathbf{F}) = \inf_{\boldsymbol{\beta}\in\mathbb{B}_n} \sum_{i=1}^{n} M_{i,i}(\boldsymbol{\beta}).$$

Let $\Pi_1, \dots, \Pi_{n!}$ be all different n-permutation matrices, i.e., $\Pi_k\boldsymbol{\beta}$ is a permutation of $\boldsymbol{\beta}$ for each k. By Birkhoff's Theorem (Theorem 2.A.2 of Marshall et al. (2011)), for $\Lambda \in \mathcal{Q}_n$, there exists $(\lambda_1, \dots, \lambda_{n!}) \in \Delta_{n!}$ such that $\Lambda = \sum_{k=1}^{n!} \lambda_k \Pi_k$. Hence, by writing $\Pi_k\boldsymbol{\beta} = (\beta_1^k, \dots, \beta_n^k)$ for each k, we have

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} \Lambda_{i,j} M_{i,j}(\boldsymbol{\beta}) &= \frac{1}{(1-p)(1-\beta)} \sum_{i=1}^{n} \int_{p+(1-p)(\beta-\beta_i)}^{1-(1-p)\beta_i} \left(\sum_{j=1}^{n} \Lambda_{i,j} \mathrm{VaR}_u(F_j) \right) \mathrm{d}u \\
&= \frac{1}{(1-p)(1-\beta)} \sum_{i=1}^{n} \sum_{k=1}^{n!} \lambda_k \int_{p+(1-p)(\beta-\beta_i^k)}^{1-(1-p)\beta_i^k} \mathrm{VaR}_u(F_i) \mathrm{d}u \\
&= \sum_{k=1}^{n!} \lambda_k \sum_{i=1}^{n} \frac{1}{(1-p)(1-\beta)} \int_{p+(1-p)(\beta-\beta_i^k)}^{1-(1-p)\beta_i^k} \mathrm{VaR}_u(F_i) \mathrm{d}u \\
&= \sum_{k=1}^{n!} \lambda_k \sum_{i=1}^{n} M_{i,i}(\Pi_k\boldsymbol{\beta}).
\end{aligned}$$

Using the above facts, we finally obtain

$$\begin{aligned}
\overline{\mathrm{VaR}}_p(\mathbf{F}) = \inf_{\boldsymbol{\beta}\in\mathbb{B}_n} \sum_{i=1}^{n} M_{i,i}(\boldsymbol{\beta}) &= \sum_{k=1}^{n!} \lambda_k \inf_{\boldsymbol{\beta}\in\mathbb{B}_n} \sum_{i=1}^{n} M_{i,i}(\Pi_k\boldsymbol{\beta}) \\
&\leqslant \inf_{\boldsymbol{\beta}\in\mathbb{B}_n} \sum_{k=1}^{n!} \lambda_k \sum_{i=1}^{n} M_{i,i}(\Pi_k\boldsymbol{\beta}) = \inf_{\boldsymbol{\beta}\in\mathbb{B}_n} \sum_{i=1}^{n}\sum_{j=1}^{n} \Lambda_{i,j} M_{i,j}(\boldsymbol{\beta}) \\
&= \overline{\mathrm{VaR}}_p(\Lambda \otimes \mathbf{F}).
\end{aligned}$$

This completes the proof of the theorem. □

The restriction of marginals to distributions with monotone densities in Theorem 3.3 is because of applying Lemma 3.3. This assumption is common in the literature of VaR bounds (e.g., Wang et al. (2013)). We may expect Theorem 3.3 to hold for more general classes of $\mathbf{F}$; this is supported by the numerical results in Figure 3.4. Moreover, for $\Lambda \in \mathcal{Q}_n$ and and $\mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n$, we may expect $\overline{\rho}(\mathbf{F}) \leqslant \overline{\rho}(\Lambda \otimes \mathbf{F})$ for other risk measures ρ than VaR (Theorem 3.3) and those consistent with convex order (Proposition 3.5). Unfortunately, we are unable to prove the above statements in general. Some related open questions are listed in Section 3.9.

Remark 3.2. The condition $\mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n$ in Theorem 3.3 can be relaxed to that the p-tail distributions of $F_1, \dots, F_n$ are all in $\mathcal{M}_D$ or all in $\mathcal{M}_I$.[6] This should be clear since only the p-tail distributions are involved in the proof of Theorem 3.3. This condition often holds if p is close to 1, and it allows for Theorem 3.3 to be applied to many common distributions in risk management.

Next, we study location-scale distribution families. Let $T_x(F)$ be a shift of $F \in \mathcal{M}$ by adding a constant $x \in \mathbb{R}$ to its location, that is, $T_x(F)$ is the distribution of $X + x$ for $X \sim F$. For $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ and $\mathbf{F} = (F_1, \dots, F_n) \in \mathcal{M}^n$, we use the notation $\mathbf{T}_{\mathbf{x}}(\mathbf{F}) = (T_{x_1}(F_1), \dots, T_{x_n}(F_n))$. Moreover, for $\lambda \geqslant 0$, we denote by F^{λ} the distribution of λX for $X \sim F$ and write $\mathbf{F}^{\boldsymbol{\lambda}} = (F^{\lambda_1}, \dots, F^{\lambda_n})$.

Corollary 3.3. *For $p \in (0,1)$, $F \in \mathcal{M}_D \cup \mathcal{M}_I$, $\boldsymbol{\lambda}, \boldsymbol{\gamma} \in \mathbb{R}_+^n$, and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, if $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$ and $\sum_{i=1}^n x_i \leqslant \sum_{i=1}^n y_i$, then*

$$\overline{\mathrm{VaR}}_p(\mathbf{T}_{\mathbf{x}}(\mathbf{F}^{\boldsymbol{\lambda}})) \leqslant \overline{\mathrm{VaR}}_p(\mathbf{T}_{\mathbf{y}}(\mathbf{F}^{\boldsymbol{\gamma}})). \tag{3.4}$$

Proof. By Section 1.A.3 of Marshall et al. (2011), $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$ if and only if there exists $\Lambda \in \mathcal{Q}_n$ such that $\boldsymbol{\gamma} = \Lambda \boldsymbol{\lambda}$. This implies $\mathbf{F}^{\boldsymbol{\gamma}} = \Lambda \otimes \mathbf{F}^{\boldsymbol{\lambda}}$. By Theorem 3.3, it follows that $\overline{\mathrm{VaR}}_p(\mathbf{F}^{\boldsymbol{\lambda}}) \leqslant \overline{\mathrm{VaR}}_p(\mathbf{F}^{\boldsymbol{\gamma}})$. Moreover, observe that

$$\overline{\mathrm{VaR}}_p(\mathbf{T}_{\mathbf{x}}(\mathbf{F}^{\boldsymbol{\lambda}})) = \overline{\mathrm{VaR}}_p(\mathbf{F}^{\boldsymbol{\lambda}}) + \sum_{i=1}^n x_i \quad \text{and} \quad \overline{\mathrm{VaR}}_p(\mathbf{T}_{\mathbf{y}}(\mathbf{F}^{\boldsymbol{\gamma}})) = \overline{\mathrm{VaR}}_p(\mathbf{F}^{\boldsymbol{\gamma}}) + \sum_{i=1}^n y_i.$$

By the fact that $\sum_{i=1}^n x_i \leqslant \sum_{i=1}^n y_i$, we prove (3.4). □

3.5 Bounds on risk measures for Pareto risk aggregation

In this section we study the worst-case risk measure for a portfolio of Pareto risks, and the risk measure is not necessarily consistent with convex order. Throughout this section, we assume that ρ is a monotone risk measure, such as VaR.

One particular situation of interest for risk aggregation with non-convex risk measures is when the risks in the portfolio do not have a finite mean. Note that for a portfolio

[6] The p-tail distribution of F is the distribution of $F^{-1}(U)$ where U is uniform on $[p,1]$; see e.g., Rockafellar and Uryasev (2002).

without finite mean, any non-constant risk measure that is consistent with convex order (including convex risk measures) will have an infinite value. Therefore, one has to use a non-convex risk measure such as VaR to assess risks in this situation.

Arguably, the most important class of heavy-tailed risk distributions is the class of Pareto distributions due to their regularly varying tails and their prominent appearance in extreme value theory; see e.g., Embrechts et al. (1997). A common parameterization of Pareto distributions is given by, for $\theta, \alpha > 0$,

$$P_{\alpha,\theta}(x) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, \quad x \geqslant \theta.$$

Note that if $X \sim P_{\alpha,1}$, then $\theta X \sim P_{\alpha,\theta}$, and thus θ is a scale parameter. Moreover, the mean of $P_{\alpha,\theta}$ is infinite if and only if $\alpha \in (0,1]$. Limited by the current techniques, we confine ourselves to portfolios of risks with a fixed α and possibly different θ.

For $\alpha > 0$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in (0,\infty)^n$, let $\mathbf{P}_{\alpha,\boldsymbol{\theta}} = (P_{\alpha,\theta_1}, \dots, P_{\alpha,\theta_n})$. We are interested in the worst-case value $\overline{\rho}(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$. We first note some simple properties of the above quantity, which are straightforward to check (a simple proof is put in the Section 3.10).

Proposition 3.6. *Let ρ be a monotone risk measure on $\mathcal{M}$. For $\alpha > 0$ and $\boldsymbol{\theta} \in (0,\infty)^n$,*

(i) $\Lambda \otimes \mathbf{P}_{\alpha,\boldsymbol{\theta}} = \mathbf{P}_{\alpha,\Lambda\boldsymbol{\theta}}$ *for all* $\Lambda \in (0,\infty)^{n\times n}$*;*

(ii) $\overline{\rho}(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ *is decreasing in* α*;*

(iii) $\overline{\rho}(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ *is increasing in each component of* $\boldsymbol{\theta}$*.*

The next result contains an ordering relationship on the aggregation of Pareto risks. In particular, we show that for $\alpha \in (0,1]$, which means the mean of the distribution is infinite, the quantile mixture leads to an even larger worst-case value of risk aggregation than the distribution mixture (this statement is generally not true for $\alpha > 1$; see the figures in Section 3.6). This result is not implied by any comparisons obtained in the previous sections, and it seems to be rather specialized for Pareto distributions, as seen from the proof. It is unclear at the moment whether the result can be generalized to other types of distributions without a finite mean.

Theorem 3.4. *Let ρ be a monotone risk measure on $\mathcal{M}$. For $\alpha \in (0,1]$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in (0,\infty)^n$, and $\Lambda \in \mathcal{Q}_n$, we have $\overline{\rho}(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \overline{\rho}(\Lambda\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \overline{\rho}(\mathbf{P}_{\alpha,\Lambda\boldsymbol{\theta}})$.*

Proof. The first inequality follows directly from Theorem 3.1. Next we focus on the second inequality. Recall that $\Lambda = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n)^\top \in \mathcal{Q}_n$ and let $\boldsymbol{\lambda}_j = (\lambda_{j,1}, \dots, \lambda_{j,n})$ for $j = 1, \dots, n$. For any fixed $j \in \{1, \dots, n\}$, denote the cdf of $(\Lambda \mathbf{P}_{\alpha,\boldsymbol{\theta}})_j$ by F_j, then

$$F_j(x) = \sum_{i=1}^{n} \lambda_{j,i} \left(1 - \left(\frac{\theta_i}{x}\right)^{\alpha}\right)_+, \quad x \in \mathbb{R}.$$

For some fixed $x > 0$ and $\alpha \in (0, 1]$, define $g(t) := 1 - (t/x)^{\alpha}$, $t \geqslant 0$. Note that g is a convex function on $[0, \infty)$. Hence

$$F_j(x) \geqslant \sum_{i=1}^{n} \lambda_{j,i} \left(1 - \left(\frac{\theta_i}{x}\right)^{\alpha}\right) \geqslant 1 - \left(\frac{\sum_{i=1}^{n} \lambda_{j,i}\theta_i}{x}\right)^{\alpha}.$$

This implies

$$F_j(x) \geqslant G_j(x), \ x \geqslant 0,$$

where $G_j = (\mathbf{P}_{\alpha,\Lambda\boldsymbol{\theta}})_j$. As $F_j \leqslant_{\mathrm{st}} G_j$ for $j = 1, \dots, n$ and ρ is monotone, by Proposition 3.3(i), we have the second inequality. □

Next, we combine the results of Theorems 3.3-3.4 and Propositions 3.5-3.6 with a special focus on VaR_p, $p \in (0, 1)$. The proof is straightforward and omitted.

Proposition 3.7. *For $p \in (0, 1)$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in (0, \infty)^n$, and $\Lambda \in \mathcal{Q}_n$,*

(i) If $\alpha \in (0, \infty)$, $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \overline{\mathrm{VaR}}_p(\Lambda \mathbf{P}_{\alpha,\boldsymbol{\theta}})$;

(ii) If $\alpha \in (0, \infty)$, $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\Lambda\boldsymbol{\theta}})$;

(iii) If $\alpha \in (0, 1]$, $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \overline{\mathrm{VaR}}_p(\Lambda \mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\Lambda\boldsymbol{\theta}})$.

Proposition 3.7 is useful for the application in Section 3.7.2 on multiple hypothesis testing, where P^r follows a Pareto distribution for a p-value P and $r < 0$. Some further properties of $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ are put in the Section 3.10.3.

3.6 Numerical illustration

Define a 3×3 doubly stochastic matrix by

$$\Lambda = 0.8 \times I_3 + 0.2 \times \left(\frac{1}{3}\right)_{3\times 3}, \tag{3.5}$$

where I_3 is the 3×3 identity matrix. In this section, we consider a sequence of doubly stochastic matrices $\{\Lambda^k\}_{k\in\mathbb{N}}$ to numerically illustrate the ordering relationships and inequalities obtained throughout this chapter. Note that Λ^k is more "homogeneous" as k grows larger, and $\Lambda^k \to (\frac{1}{3})_{3\times 3}$ as $k \to \infty$. The general messages obtained from the numerical examples are listed as follows.

1. For general marginals, the value of $\overline{\mathrm{VaR}}$ becomes larger after making a distribution mixture (Proposition 3.5(i)); this is shown in all figures.

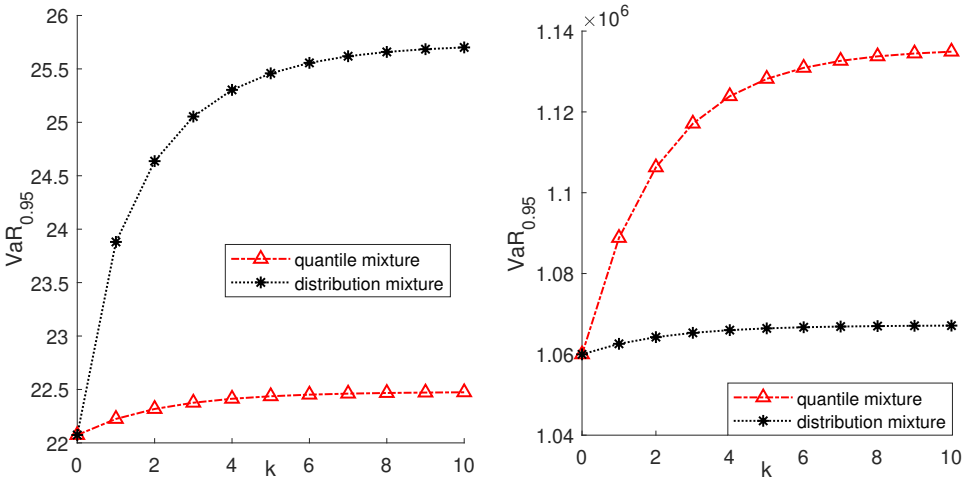
2. For marginals with monotone densities, with a quantile mixture, the value of $\overline{\mathrm{VaR}}$ becomes larger (Theorem 3.3); see Figures 3.1-3.3. Numerical examples in Figure 3.4 indicate that Theorem 3.3 may also hold for marginals with non-monotone densities. Nevertheless, the order does not hold for arbitrary marginals. A counterexample, involving discrete marginals, is provided in Figure 3.5.

3. For Pareto distributions with infinite mean, the value of $\overline{\mathrm{VaR}}$ of the quantile mixture is larger than that of the distribution mixture (Proposition 3.7(iii)); see Figure 3.1(b). This relationship does not hold for Pareto distributions with finite mean; see Figure 3.1(a).

3.6.1 Illustration of theoretical results

In this subsection, we discuss marginals with monotone densities ($\mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n$). We have $\Lambda^k \otimes \mathbf{F}, \Lambda^k\mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n$. According to Lemma 3.3 in Section 3.10.1, we obtain a formula (Equation (3.3)) for $\mathbf{F}$, $\Lambda^k \otimes \mathbf{F}$ and $\Lambda^k\mathbf{F}$ and numerically compute the exact values for $\overline{\mathrm{VaR}}$.

In Figure 3.1, we consider Pareto distributions with finite mean ($\alpha = 3$) and infinite mean ($\alpha = 1/3$), respectively. The ordering relationships in Proposition 3.7(i)-(ii) for Pareto distributions with the same α are visualized as the curves in Figure 3.1 are all increasing in k. In Figure 3.1(b), it turns out that for the case with infinite mean the quantile mixture gives larger value of $\overline{\mathrm{VaR}}$ than that given by the distribution mixture. This coincides with the conclusion in Proposition 3.7(iii). Interestingly, we observe from Figure 3.1(a) that the value of $\overline{\mathrm{VaR}}$ given by distribution mixture is larger than the one with quantile mixture, which is contrary to the case with infinite mean (Figure 3.1(b)). It is an open question whether this conclusion is true for general doubly stochastic matrices Λ and all $\alpha > 1$.



(a) Pareto distribution with finite mean ($\alpha = 3$)

(b) Pareto distribution with finite mean ($\alpha = 1/3$)

Figure 3.1: Quantile mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k \otimes \mathbf{P}_{\alpha,\boldsymbol{\theta}}) = \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\Lambda^k\boldsymbol{\theta}})$; Distribution mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k\mathbf{P}_{\alpha,\boldsymbol{\theta}})$. Setting: $p = 0.95$; $\boldsymbol{\theta} = (1, 2, 3)$, $X_i \sim \mathrm{Pareto}(\alpha, \theta_i)$, $i = 1, 2, 3$; Λ is defined by (3.5); $k = 0, 1, \ldots, 10$.

We next focus on Pareto distributions with different α in Figure 3.2. First observe that the curves of quantile mixture and distribution mixture in Figure 3.2 are both increasing in k, which is consistent with Theorem 3.3 and Proposition 3.5(i). Comparing the two curves, it is shown that value for the distribution mixture in this case is smaller than the one for quantile mixture.

Heterogeneous distribution families with decreasing densities are considered in Figure 3.3. As we can see, the curves are both increasing in Figure 3.3, which coincides with the statements in Theorem 3.3 and Proposition 3.5(i). We can also observe that the value for distribution mixture is smaller than the corresponding one for quantile mixture in Figure 3.3, which is the same as it has been shown in Figure 3.2.

3.6.2 Conjectures for general distributions

Explicit expressions for $\overline{\mathrm{VaR}}_p(\mathbf{F})$ are unavailable for general marginal distributions. Fortunately, we can approximate the value of $\overline{\mathrm{VaR}}_p(\mathbf{F})$ using the rearrangement algorithm (RA) of Embrechts et al. (2013) and get an upper bound on $\overline{\mathrm{VaR}}_p(\mathbf{F})$ using (3.12) in Lemma 3.3.

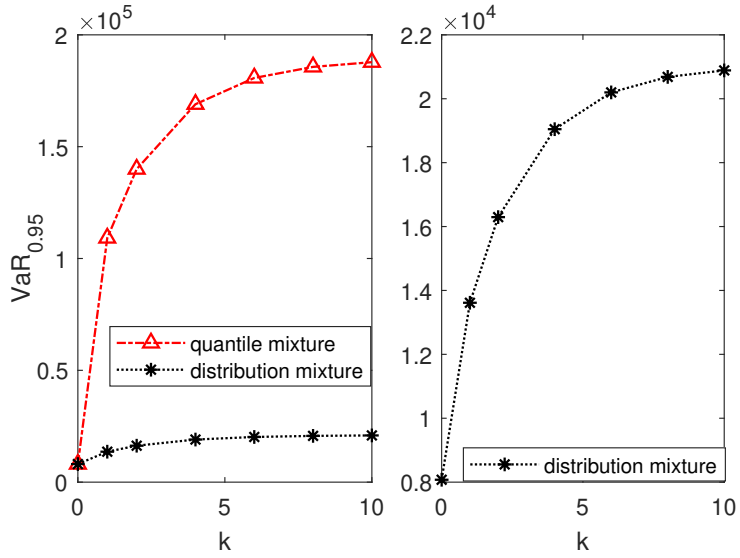


Figure 3.2: Quantile mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k \otimes \mathbf{F})$; Distribution mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k \mathbf{F})$. Setting: $p = 0.95$; $\boldsymbol{\alpha} = (1/3, 4, 5)$, $\boldsymbol{\theta} = (1, 2, 3)$, $X_i \sim \mathrm{Pareto}(\alpha_i, \theta_i)$, $i = 1, 2, 3$; Λ is defined by (3.5); $k = 0, 1, 2, 4, 6, 8, 10$. The right panel zooms in on the range of the distribution mixture.

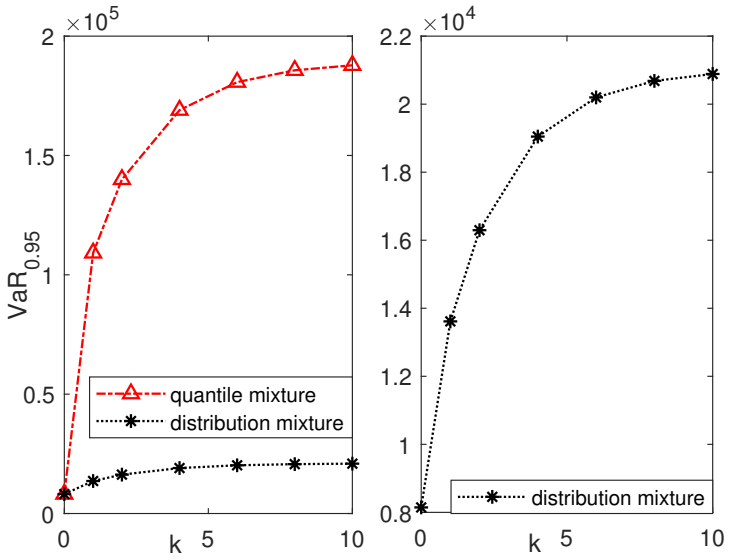
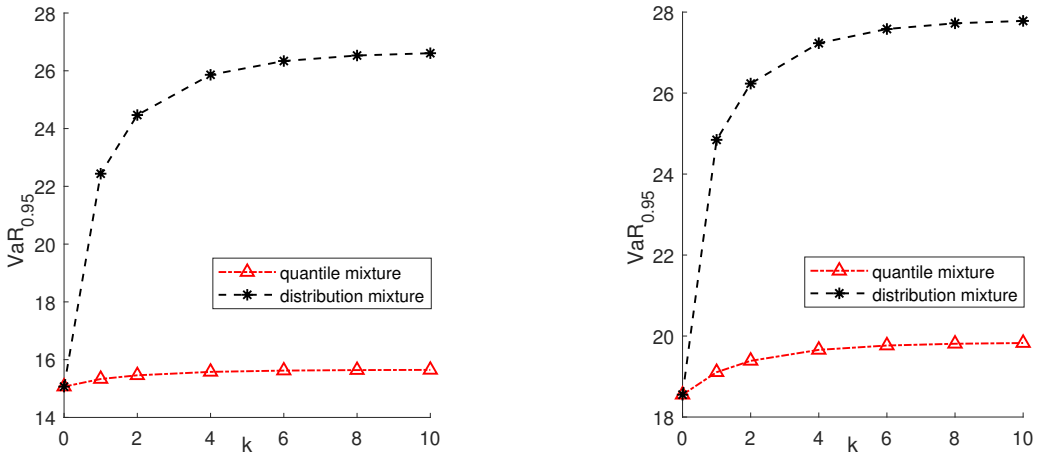


Figure 3.3: Quantile mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k \otimes \mathbf{F})$; Distribution mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k \mathbf{F})$. Setting: $p = 0.95$; $X_1 \sim \mathrm{Pareto}(1/3, 1)$, $X_2 \sim \Gamma(1, 2)$, $X_3 \sim \mathrm{Weibull}(1, 1/2)$; Λ is defined by (3.5); $k = 0, 1, 2, 4, 6, 8, 10$. The right panel zooms in on the range of the distribution mixture.



(a) $X_3 \sim \text{Pareto}(3,1)$ (b) $X_3 \sim \text{LogNormal}(0,1)$

Figure 3.4: Quantile mixture: $\overline{\text{VaR}}_p(\Lambda^k \otimes \mathbf{F})$; Distribution mixture: $\overline{\text{VaR}}_p(\Lambda^k \mathbf{F})$. Setting: $p = 0.95$; $X_1 \sim \Gamma(5,1)$, $X_2 \sim \text{Weibull}(1,5)$, left panel: $X_3 \sim \text{Pareto}(3,1)$, right panel: $X_3 \sim \text{LogNormal}(0,1)$; Λ is defined by (3.5); $k = 0, 1, 2, 4, 6, 8, 10$.

For distributions with non-monotone densities including Gamma and Weibull, the curves of both distribution and quantile mixtures in Figure 3.4 are increasing in k. The result on distribution mixture is consistent with Proposition 3.5(i), and the result on quantile mixture seems to suggest that the conclusion in Theorem 3.3 may be valid for more general distributions with non-monotone densities. This conjectured extension of Theorem 3.3 would hold if (3.3) holds for more general distributions, which is a difficult question.

The above observation is no longer true for discrete distributions. We observe in Figure 3.5 that the curve of the quantile mixture is not increasing at some points (in this example, we have chosen a small $p = 0.01$ for illustration). This shows that the claim in Theorem 3.3 cannot be extended to arbitrary, in particular discrete, distributions.

3.7 Applications

3.7.1 Portfolio diversification with dependence uncertainty

We discuss applications of our results to portfolio diversification in the presence of dependence uncertainty. In this section, we treat risk measures as functionals on the space

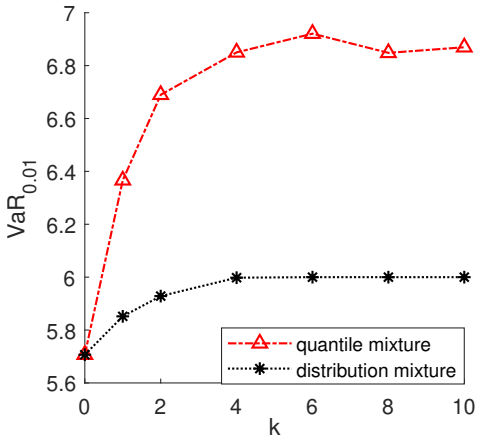


Figure 3.5: Quantile mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k \otimes \mathbf{F})$; Distribution mixture: $\overline{\mathrm{VaR}}_p(\Lambda^k \mathbf{F})$. Setting: $p = 0.01$; $X_1 \sim \mathrm{Binomial}(10, 0.1)$, $X_2 \sim \Gamma(5, 1)$, $X_3 \sim \mathrm{Weibull}(1, 5)$; Λ is defined by (3.5); $k = 0, 1, 2, 4, 6, 8, 10$.

of random variables, that is, for a random variable X and a risk measure ρ, we write $\rho(X) = \rho(F)$ if $X \sim F$.

For tractability, we consider a simple setting where the vector of losses $(X_1, \dots, X_n)$ has identical marginal distributions F. A classic portfolio selection problem is to choose a portfolio position $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n) \in \Delta_n$ to minimize

$$R_\rho(\boldsymbol{\lambda}) := \rho\left(\sum_{i=1}^n \lambda_i X_i\right). \tag{3.6}$$

Alternatively, one may consider an objective which involves both risk and return, such as maximizing the quantity $\mathbb{E}[-\sum_{i=1}^n \lambda_i X_i] - \alpha\rho(\sum_{i=1}^n \lambda_i X_i)$ for some $\alpha > 0$ (e.g., α may arise as a Lagrangian multiplier); in our setting, this problem is equivalent to (3.6) since $\mathbb{E}[\sum_{i=1}^n \lambda_i X_i]$ is constant over $\boldsymbol{\lambda} \in \Delta_n$. Intuitively, for two portfolio positions $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, we can say that $\boldsymbol{\gamma}$ is more diversified than $\boldsymbol{\lambda}$ if $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$, since in this case $\boldsymbol{\gamma}$ can be obtained from averaging components of $\boldsymbol{\lambda}$, i.e., $\boldsymbol{\gamma} = \Lambda\boldsymbol{\lambda}$ for some $\Lambda \in \mathcal{Q}_n$ (see Section 3.2). Due to diversification effect, one may expect, under the assumption that the marginal distributions of $(X_1, \dots, X_n)$ are identical,

$$R_\rho(\boldsymbol{\gamma}) \leqslant R_\rho(\boldsymbol{\lambda}) \quad \text{if } \boldsymbol{\gamma} \text{ is more diversified than } \boldsymbol{\lambda}. \tag{3.7}$$

Note that $(\frac{1}{n}, \dots, \frac{1}{n}) \prec \boldsymbol{\lambda} \prec (1, 0, \dots, 0)$ for any portfolio position $\boldsymbol{\lambda}$, meaning that the

most diversified portfolio is the equally weighted one, and the least diversified portfolio is concentrating on a single source of risk.

To compute the value of $R_\rho(\boldsymbol{\lambda})$ in (3.6) requires a full specification of the joint distribution of $(X_1, \dots, X_n)$. In the presence of dependence uncertainty, we may take a worst-case approach by minimizing

$$\overline{R}_\rho(\boldsymbol{\lambda}) := \sup \left\{ \rho\left(\sum_{i=1}^{n} \lambda_i Y_i \right) : Y_1, \dots, Y_n \sim F \right\}. \tag{3.8}$$

Under the setting of optimizing (3.8), our intuition is that diversification should not yield any benefit, since the portfolio may not have any diversification effect due to unknown dependence; see Wang and Zitikis (2021) for discussions on the absence of diversification effect within the Fundamental Review of the Trading Book from the Basel Committee on Banking Supervision (BCBS (2019)). Hence, one may expect, as the marginal distributions are identical, that

$$\overline{R}_\rho(\boldsymbol{\gamma}) = \overline{R}_\rho(\boldsymbol{\lambda}) \quad \text{even if } \boldsymbol{\gamma} \text{ is more diversified than } \boldsymbol{\lambda} \text{ (in fact, for all } \boldsymbol{\gamma} \text{ and } \boldsymbol{\lambda}). \tag{3.9}$$

A similar observation is made in Proposition 1 of Pflug and Pohl (2018), which says that for a subadditive, comonotonic-additive and positively homogeneous risk measure, diversification under dependence uncertainty does not decrease the aggregate risk. These assumptions on the risk measure are not necessary for our result below.

The next proposition, based on Theorem 3.3 and Proposition 3.5, shows that, under some extra conditions, the two intuitive equations (3.7) and (3.9) hold for risk measures consistent with convex order. For VaR, one arrives at a statement in the reverse direction: the more diversified portfolio has a larger risk under dependence uncertainty.

Proposition 3.8. *Suppose that $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$, $(X_1, \dots, X_n)$ has identical marginal distributions F with finite mean, and ρ is a risk measure.*

(i) If ρ is consistent with convex order and $(X_1, \dots, X_n)$ is exchangeable,[7] *then $R_\rho(\boldsymbol{\gamma}) \leqslant R_\rho(\boldsymbol{\lambda})$.*

(ii) If ρ is consistent with convex order, then $\overline{R}_\rho(\boldsymbol{\gamma}) = \overline{R}_\rho(\boldsymbol{\lambda})$.

(iii) If $\rho = \mathrm{VaR}_p$ for some $p \in (0,1)$ and $F \in \mathcal{M}_D \cup \mathcal{M}_I$, then $\overline{R}_\rho(\boldsymbol{\gamma}) \geqslant \overline{R}_\rho(\boldsymbol{\lambda})$.

[7] A random vector $\mathbf{X}$ is exchangeable if $\mathbf{X}$ is identically distributed as $\pi(\mathbf{X})$ for any permutation π.

Moreover, in (i) and (iii), the inequalities are generally not equalities.

Proof. Write $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$. Take $X \sim F$, and let $\mathbf{F}$ and $\mathbf{G}$ be the tuples of marginal distributions of $(\gamma_1 X, \dots, \gamma_n X)$ and $(\lambda_1 X, \dots, \lambda_n X)$, respectively. Using $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$, there exists $\Lambda \in \mathcal{Q}_n$ such that $\mathbf{F} = \Lambda \otimes \mathbf{G}$. Since $X_1, \dots, X_n \sim F$, $\mathbf{F}$ and $\mathbf{G}$ are also tuples of marginal distributions of $(\gamma_1 X_1, \dots, \gamma_n X_n)$ and $(\lambda_1 X_1, \dots, \lambda_n X_n)$, respectively. Hence, we have

$$\overline{R}_\rho(\boldsymbol{\gamma}) = \overline{\rho}(\mathbf{F}) = \overline{\rho}(\Lambda \otimes \mathbf{G}) \quad \text{and} \quad \overline{R}_\rho(\boldsymbol{\lambda}) = \overline{\rho}(\mathbf{G}). \tag{3.10}$$

(i) As $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$ and $(X_1, \dots, X_n)$ is exchangeable, by Theorem 3.A.35 of Shaked and Shanthikumar (2007), we have $\sum_{i=1}^n \gamma_i X_i \prec_{\mathrm{cx}} \sum_{i=1}^n \lambda_i X_i$. Hence, $R_\rho(\boldsymbol{\gamma}) \leqslant R_\rho(\boldsymbol{\lambda})$. The inequality is strict when, for instance, $\rho = \mathrm{ES}_{0.5}$, $X_1, \dots, X_n$ are iid normal, $\boldsymbol{\gamma} = (\frac{1}{n}, \dots, \frac{1}{n})$, and $\boldsymbol{\lambda} = (1, 0, \dots, 0)$.

(ii) This follows directly from Proposition 3.5(ii) and (3.10).

(iii) The inequality $\overline{R}_\rho(\boldsymbol{\gamma}) \geqslant \overline{R}_\rho(\boldsymbol{\lambda})$ follows directly from Theorem 3.3 and (3.10). The inequality is strict in, for instance, the situation of Figure 3.1(a), where F is a Pareto distribution with $\alpha = 3$. □

We make a few observations from Proposition 3.8. For identical marginal distributions in $\mathcal{M}_D$ or $\mathcal{M}_I$, under dependence uncertainty, VaR yields a bigger risk if the portfolio is more diversified. This may be seen as another disadvantage of VaR, which is well known to be problematic regarding diversification. In contrast, any risk measure consistent with convex order, such as ES, would simply ignore diversification effect in this setting (where diversification benefit is unjustifiable). Moreover, without dependence uncertainty, for an exchangeable vector of losses, a risk measure consistent with convex order rewards diversification, and there is no such general relationship for VaR. For the inequality in Proposition 3.8 (iii), it suffices to require the p-tail distribution of F to be in $\mathcal{M}_D \cup \mathcal{M}_I$; see Remark 3.2.

3.7.2 Merging p-values in hypothesis testing

In this subsection, we apply our results to p-merging methods following the setup of Vovk and Wang (2020). A random variable P is a *p-variable* if $\mathbb{P}(P \leqslant \varepsilon) \leqslant \varepsilon$ for all $\varepsilon \in (0, 1)$, and its realization is called a p-value. In multiple hypothesis testing, one natural problem is to merge individual p-values into one p-value. More specifically, with n

p-variables $P_1, \dots, P_n$, one needs to choose an increasing Borel function $F : [0,1]^n \to [0,\infty)$ as a *merging function* such that $F(P_1, \dots, P_n)$ is a p-variable. F is a *precise merging function* if for each $\varepsilon \in (0,1)$, $\mathbb{P}(F(P_1, \dots, P_n) \leqslant \varepsilon) = \varepsilon$ for some p-variables $P_1, \dots, P_n$.

As explained in Vovk and Wang (2020), an advantage of using averaging methods to combine p-values, compared to classic methods on order statistics, is that we can introduce weights to p-values in an intuitive way. Without imposing any dependence assumption on the individual p-variables, an averaging method uses, for $r \in [-\infty, \infty]$ ($r \in \{-\infty, 0, \infty\}$ are interpreted as limits),

$$F : [0,1]^n \to [0,\infty), \ (p_1, \dots, p_n) \mapsto a_{r,\mathbf{w}}(w_1 p_1^r + \cdots + w_n p_n^r)^{\frac{1}{r}},$$

as the merging function, where $a_{r,\mathbf{w}}$ is a constant multiplier and $\mathbf{w} = (w_1, \dots, w_n) \in \Delta_n$. The constant $a_{r,\mathbf{w}}$ is chosen so that F is a precise merging function, thus the most powerful choice of the constant multiplier. Let $\mathcal{U}$ be the set of uniform random variables distributed on [0,1]. Lemma 1 in Vovk and Wang (2020) gives

$$a_{r,\mathbf{w}} = \begin{cases} -\sup\left\{q_0(-\sum_{i=1}^n w_i P_i^r) \mid P_1, \dots, P_n \in \mathcal{U}\right\}^{-1/r}, & r > 0; \\ \exp\left(\sup\{q_0(\sum_{i=1}^n w_i \log(1/P_i)) \mid P_1, \dots, P_n \in \mathcal{U}\}\right), & r = 0; \\ \sup\left\{q_0\left(\sum_{i=1}^n w_i P_i^r\right) \mid P_1, \dots, P_n \in \mathcal{U}\right\}^{-1/r}, & r < 0, \end{cases}$$

where $q_0 : X \mapsto \inf\{x \in \mathbb{R} : \mathbb{P}(X \leqslant x) > 0\}$ is the essential infimum. Clearly, $a_{r,\mathbf{w}}$ involves calculating $\overline{\mathrm{VaR}}_p(\mathbf{F})$ for Pareto, exponential or Beta distributions, and letting $p \downarrow 0$.

Denote $a_{r,\mathbf{w}}$ by $a_{r,n}$ where $\mathbf{w} = (1/n, \dots, 1/n)$. Analytical results for $a_{r,n}$ has been well studied in Vovk and Wang (2020) whereas results for $a_{r,\mathbf{w}}$ are limited since there are no analytical formulas of $\overline{\mathrm{VaR}}_p(\mathbf{F})$ in general for heterogeneous marginal distributions. Although the rearrangement algorithm of Puccetti and Rüschendorf (2012) and Embrechts et al. (2013) can be used to calculate $a_{r,\mathbf{w}}$ numerically, the calculation burden becomes quite heavy in high-dimensional situation, which is unfortunately very common in multiple hypothesis testing. It turns out that our Theorem 3.3 is helpful to provide a convenient upper bound on $a_{r,\mathbf{w}}$.

Proposition 3.9. *For $r \in \mathbb{R}$, we have $a_{r,\mathbf{w}} \leqslant a_{r,n}$.*

Proof. Note that for $r < 0$, P_i^r, $i = 1, \dots, n$, has a decreasing density, and $(1/n, \dots, 1/n) \prec (w_1, \dots, w_n)$ in majorization order. By letting $p \downarrow 0$ in Proposition 3.7, we have

$$\sup\left\{q_0\left(\sum_{i=1}^n w_i P_i^r\right) \mid P_1, \dots, P_n \in \mathcal{U}\right\} \leqslant \sup\left\{q_0\left(\sum_{i=1}^n \frac{1}{n} P_i^r\right) \mid P_1, \dots, P_n \in \mathcal{U}\right\}.$$

Therefore $a_{r,\mathbf{w}} \leqslant a_{r,n}$ for $r < 0$. If $r \geqslant 0$, the argument can be proved similarly using Corollary 3.3. □

The interpretation of Proposition 3.9 is that, when using a weighted p-merging method, one can safely rely on the same coefficient obtained from a symmetric p-merging method. This is particularly convenient when validity of the test is more important than the quality of an approximation; see Vovk and Wang (2020) for more discussions on such applications.

3.8 Some further technical discussions

3.8.1 Location shifts for distribution and quantile mixtures

In this section we discuss the difference between distribution and quantile mixtures when location shifts are applied. Let $V_x = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_1 + \cdots + x_n = x\}$ for $x \in \mathbb{R}$. For $\mathbf{F} \in \mathcal{M}^n$ and $\mathbf{x} \in V_x$, we have the invariance relation

$$\mathcal{D}_n(\mathbf{T}_\mathbf{x}(\mathbf{F})) = T_x(\mathcal{D}_n(\mathbf{F})). \tag{3.11}$$

The aggregation set of quantile mixture is invariant under location shifts of the marginal distributions, in sharp contrast to the case of distribution mixture. For $\mathbf{F} \in \mathcal{M}^n$ and $\mathbf{x} \in V_x$, it holds that for $\Lambda \in \mathcal{Q}_n$,

$$\mathcal{D}_n(\Lambda \otimes \mathbf{T}_\mathbf{x}(\mathbf{F})) = T_x\left(\mathcal{D}_n(\Lambda \otimes \mathbf{F})\right).$$

That means, $\mathcal{D}_n(\Lambda \otimes \mathbf{T}_\mathbf{x}(\mathbf{F}))$ is the same for all $\mathbf{x} \in V_x$. However, this does not hold for the distribution mixture, that is, generally, $\mathcal{D}_n(\Lambda\mathbf{T}_\mathbf{x}(\mathbf{F}))$ is not the same for $\mathbf{x} \in V_x$, and

$$\mathcal{D}_n(\Lambda\mathbf{T}_\mathbf{x}(\mathbf{F})) \neq T_x\left(\mathcal{D}_n(\Lambda\mathbf{F})\right).$$

In particular, for $x \neq 0$ and $F_1 \neq F_2$,

$$\mathcal{D}_2\left(\frac{1}{2}(T_x(F_1) + F_2), \frac{1}{2}(T_x(F_1) + F_2)\right) \neq \mathcal{D}_2\left(\frac{1}{2}(F_1 + T_x(F_2)), \frac{1}{2}(F_1 + T_x(F_2)))\right).$$

The above example shows that distribution mixture and quantile mixtures treat location shifts differently.

Inspired by the above observation, we slightly generalize Theorem 3.1 by including location shifts. For $\mathbf{F} \in \mathcal{M}^n$, we define the set $\mathcal{A}_n(\mathbf{F})$ of averaging and location shifts of $\mathbf{F}$ as

$$\mathcal{A}_n(\mathbf{F}) = \{\Lambda\mathbf{T}_\mathbf{x}(\mathbf{F}) : \Lambda \in \mathcal{Q}_n,\ \mathbf{x} \in \mathbb{R}^n,\ x_1 + \cdots + x_n = 0\},$$

and denote by $\overline{\mathcal{A}_n(\mathbf{F})}$ the closure of the convex hull of $\mathcal{A}_n(\mathbf{F})$ with respect to weak convergence. It is straightforward to check

$$\overline{\mathcal{A}_n(\mathbf{T}_{\mathbf{y}}(\mathbf{F}))} = \mathbf{T}_{\mathbf{y}}\left(\overline{\mathcal{A}_n(\mathbf{F})}\right), \quad \mathbf{y} = (y, \dots, y) \in \mathbb{R}^n.$$

Proposition 3.10. *For $\mathbf{F} \in \mathcal{M}^n$ and $\mathbf{G} \in \overline{\mathcal{A}_n(\mathbf{F})}$, we have $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\mathbf{G})$.*

Proof. First, by Theorem 3.1 and (3.11), $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\mathbf{G})$ for each $\mathbf{G} \in \mathcal{A}_n(\mathbf{F})$. Denote by $\mathrm{cx}(\mathcal{A}_n(\mathbf{F}))$ the convex hull of $\mathcal{A}_n(\mathbf{F})$. By Lemma 3.1(ii-b), for each $\mathbf{G} \in \mathrm{cx}(\mathcal{A}_n(\mathbf{F}))$, we have $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\mathbf{G})$. Take $\mathbf{G} \in \overline{\mathcal{A}_n(\mathbf{F})}$, and write it as the limit of $\{\mathbf{G}_k\}_{k=1}^{\infty} \subset \mathrm{cx}(\mathcal{A}_n(\mathbf{F}))$. It follows that for any $F \in \mathcal{D}_n(\mathbf{F})$, F is also in $\mathcal{D}_n(\mathbf{G}_k)$. This implies F is also in $\mathcal{D}_n(\mathbf{G})$ by the compactness property in Theorem 2.1(vii-b) of Bernard et al. (2014). □

3.8.2 Connection to joint mixability

Joint mixability (Wang et al. (2013) and Wang and Wang (2016)) is a central concept in the study of risk aggregation with dependence uncertainty, and analytical results are quite limited. In this section, we study the implication of our results on conditions for joint mixability. We denote by δ_x the point mass at $x \in \mathbb{R}$.

Definition 3.1 (Joint mixability)**.** An n-tuple of distributions $\mathbf{F} \in \mathcal{M}^n$ is *jointly mixable* (JM) if $\mathcal{D}_n(\mathbf{F})$ contains a point mass distribution δ_x, where $x \in \mathbb{R}$ is called a *center* of $\mathbf{F}$.

Example 3.1 implies a conclusion on the joint mixability of Bernoulli distributions.

Proposition 3.11. *For $p_1, \dots, p_n \in [0,1]$, $(B_{p_1}, \dots, B_{p_n})$ is jointly mixable if and only if $\sum_{i=1}^n p_i$ is an integer.*

Proof. The "only-if" part is trivial since the sum of Bernoulli random variables takes value in integers. To show the "if" part, let $k = \sum_{i=1}^n p_i$ and $\mathbf{1}_k \in \{0,1\}^n$ be a vector whose first k entries are 1 and the remaining entries are 0. It is clear that $\mathbf{p} \prec \mathbf{1}_k$ (see Section 1.A.3 of Marshall et al. (2011)). Hence, from Example 3.1,

$$\{\delta_k\} = \mathcal{D}_n(\underbrace{B_1, \dots, B_1}_{k}, \underbrace{B_0, \dots, B_0}_{n-k}) \subset \mathcal{D}_n(B_{p_1}, \dots, B_{p_n}).$$

Therefore $(B_{p_1}, \dots, B_{p_n})$ is jointly mixable. □

The set $\overline{\mathcal{A}_n(\mathbf{F})}$ can also be used to obtain joint mixability of some tuples of distributions. In particular, we shall see in the following proposition that $\overline{\mathcal{A}_n(\delta_0,\dots,\delta_0)}$ is the set of all jointly mixable tuples with center 0.

Proposition 3.12. *For $\mathbf{G}\in\mathcal{M}^n$, the following statements are equivalent.*

(i) $\mathbf{G}$ *is jointly mixable.*

(ii) $\mathbf{G}\in\overline{\mathcal{A}_n(\delta_c,\dots,\delta_c)}$ *for some* $c\in\mathbb{R}$.

(iii) $\mathbf{G}\in\overline{\mathcal{A}_n(\mathbf{F})}$ *for some* $\mathbf{F}\in\mathcal{M}^n$ *which is jointly mixable.*

Proof. (ii)⇒(iii) is trivial. (iii)⇒(i): Suppose that $\mathbf{G}\in\overline{\mathcal{A}_n(\mathbf{F})}$ and $\mathbf{F}$ is jointly mixable with center $x\in\mathbb{R}$. By Proposition 3.10, we have $\{\delta_x\}\subset\mathcal{D}_n(\mathbf{F})\subset\mathcal{D}_n(\mathbf{G})$. This shows $\mathbf{G}$ is jointly mixable. Next, we show (i)⇒(ii). Suppose that $\mathbf{G}$ is jointly mixable, and without loss of generality we can assume it has center 0. By definition, there exists a random vector $\mathbf{X}=(X_1,\dots,X_n)$ such that $X_i\sim G_i$ and $X_1+\cdots+X_n=0$. Denote by H the distribution measure of $\mathbf{X}$. For $A\in\mathcal{B}(\mathbb{R})$ and $i=1,2,\dots,n$,

$$G_i(A)=\mathbb{P}(X_i\in A)=\int_{\mathbb{R}^n}\mathbb{P}(X_i\in A|\mathbf{X}=\mathbf{y})H(\mathrm{d}\mathbf{y})=\int_{\mathbb{R}^n}\delta_{y_i}(A)H(\mathrm{d}\mathbf{y}),$$

and as a consequence,

$$\mathbf{G}(A)=(G_1(A),\dots,G_n(A))=\int_{\mathbb{R}^n}(\delta_{y_1}(A),\dots,\delta_{y_n}(A))H(\mathrm{d}\mathbf{y}).$$

Noting that H is supported in $V_0=\{(y_1,\dots,y_n)\in\mathbb{R}^n:y_1+\cdots+y_n=0\}$, we have

$$\mathbf{G}(A)=\int_{V_0}(\delta_{y_1}(A),\dots,\delta_{y_n}(A))H(\mathrm{d}\mathbf{y})=\int_{V_0}\mathbf{T}_{\mathbf{y}}(\delta_0(A),\dots,\delta_0(A))H(\mathrm{d}\mathbf{y}).$$

Hence, we conclude that $\mathbf{G}\in\overline{\mathcal{A}_n(\delta_0,\dots,\delta_0)}$. □

The set $\overline{\mathcal{A}_n(\delta_c,\dots,\delta_c)}$ is quite rich and cannot be analytically characterized. The simple example of uniform distributions might be helpful to understand Proposition 3.12. Suppose that $F_i=\mathrm{U}[0,a_i]$, $a_i>0$, $i=1,\dots,n$, and $\sum_{i=1}^n a_i\geqslant 2\bigvee_{i=1}^n a_i$. By Theorem 3.1 of Wang and Wang (2016), we know that $\mathbf{F}$ is jointly mixable. Then, Proposition 3.12 implies that every tuple in the set $\overline{\mathcal{A}_n(\mathbf{F})}$ is jointly mixable.

It remains an open question whether it is possible to characterize the set $\overline{\mathcal{A}_n(\mathbf{F})}$ for uniform random variables. This would lead to many classes of jointly mixable distributions including those with monotone densities and symmetric densities; see Wang and Wang (2016).

3.9 Concluding remarks

Chapter 3 studies the ordering relationship for aggregation sets where the marginal distributions for different sets are connected by either a distribution mixture or a quantile mixture. For general marginal distributions, the aggregation set becomes larger after making a distribution mixture on the marginal risks, whereas the aggregation sets are not necessarily comparable in general by a quantile mixture on the marginal risks. Nevertheless, we obtain several useful results especially on the comparison of VaR aggregation, which has applications in and outside financial risk management.

Although the marginal distributions are assumed known in our main setting, this assumption is not essential for the interpretation of our results in practical situations. In case both marginal uncertainty and dependence uncertainty are present, our results can be directly applied to obtain ordering relationships, as we explain below. Suppose that $\Lambda \in \mathcal{Q}_n$ and $\mathcal{F} \subset \mathcal{M}^n$ is a set of possible marginal models, representing uncertainty on the marginal distributions. In this case, the set of all possible distributions of aggregate risk is $\bigcup_{\mathbf{F}\in\mathcal{F}} \mathcal{D}_n(\mathbf{F})$, and the worst-case value of a risk measure ρ is $\sup\{\rho(G) : G \in \mathcal{D}_n(\mathbf{F}),\ \mathbf{F} \in \mathcal{F}\} = \sup_{\mathbf{F}\in\mathcal{F}} \overline{\rho}(\mathbf{F})$. Using Theorem 3.1, Proposition 3.5 and Theorem 3.3, we have

$$\bigcup_{\mathbf{F}\in\mathcal{F}} \mathcal{D}_n(\mathbf{F}) \subset \bigcup_{\mathbf{F}\in\mathcal{F}} \mathcal{D}_n(\Lambda\mathbf{F}), \qquad \sup_{\mathbf{F}\in\mathcal{F}} \overline{\rho}(\mathbf{F}) \leqslant \sup_{\mathbf{F}\in\mathcal{F}} \overline{\rho}(\Lambda\mathbf{F}),$$

and, if $\mathcal{F} \subset \mathcal{M}_D^n \cup \mathcal{M}_I^n$,

$$\sup_{\mathbf{F}\in\mathcal{F}} \overline{\mathrm{VaR}}_p(\mathbf{F}) \leqslant \sup_{\mathbf{F}\in\mathcal{F}} \overline{\mathrm{VaR}}_p(\Lambda \otimes \mathbf{F}).$$

Thus, our results on set inclusion and risk measure inequalities remain valid in the presence of marginal uncertainty.

3.10 Appendix

3.10.1 A lemma used in the proof of Theorem 3.3

The following lemma is rephrased from Theorem 2 of Blanchet et al. (2020).

Lemma 3.3. *For $p \in (0,1)$ and any $\mathbf{F} = (F_1, \dots, F_n) \in \mathcal{M}^n$,*

$$\overline{\mathrm{VaR}}_p^*(\mathbf{F}) \leqslant \inf_{\boldsymbol{\beta}\in\mathbb{B}_n} \sum_{i=1}^{n} \frac{1}{(1-p)(1-\beta)} \int_{p+(1-p)(\beta-\beta_i)}^{1-(1-p)\beta_i} \mathrm{VaR}_u(F_i)\mathrm{d}u, \tag{3.12}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$, $\beta = \sum_{i=1}^n \beta_i$ *and* $\mathbb{B}_n = \{\boldsymbol{\beta} \in [0,1)^n : \beta < 1\}$, *and the above inequality is an equality if* $\mathbf{F} \in \mathcal{M}_D^n \cup \mathcal{M}_I^n$.

3.10.2 Proofs of two propositions

Proof of Proposition 3.4. We first focus on (i). We will show (a) $\Leftrightarrow$ (c). (c) $\Rightarrow$ (a) is trivial by the definition of stochastic order. For (a) $\Rightarrow$ (c), note that $\Lambda\mathbf{F} \prec_{\mathrm{st}} \mathbf{F}$ with $\Lambda = (\Lambda_{ij})$ implies

$$\sum_{j=1}^n \Lambda_{ij} F_j(x) \geqslant F_i(x), \ x \in \mathbb{R}, i = 1, \dots, n. \tag{3.13}$$

Adding all the inequalities in (3.13) yields

$$\sum_{i=1}^n \sum_{j=1}^n \Lambda_{ij} F_j(x) \geqslant \sum_{i=1}^n F_i(x), \ x \in \mathbb{R}.$$

Due to the fact that Λ is a doubly stochastic matrix, we have

$$\sum_{i=1}^n \sum_{j=1}^n \Lambda_{ij} F_j(x) = \sum_{i=1}^n F_i(x), \ x \in \mathbb{R}.$$

Hence all the inequalities in (3.13) are essentially equalities. This proves (c). We can analogously show that (b) $\Leftrightarrow$ (c). This establishes the claims in (i). We will omit the proof of (ii) since it is similar to the proof of (i).

We next focus on (iii). Trivially, (c) $\Rightarrow$ (a) and (c) $\Rightarrow$ (b). Next, we will only show (a) $\Rightarrow$ (c) since (b) $\Rightarrow$ (c) is similar. Denote by $\mathbf{G} = (G_1, \dots, G_n) = \Lambda \otimes \mathbf{F}$. Hence

$$G_i^{-1} = \sum_{j=1}^n \Lambda_{ij} F_j^{-1}.$$

By definition, $\Lambda \otimes \mathbf{F} \prec_{\mathrm{cx}} \mathbf{F}$ implies $G_i \prec_{\mathrm{cx}} F_i, i = 1, \dots, n$. It is well known (see e.g., Theorem 3.A.5 of Shaked and Shanthikumar (2007)) that for any two distributions F and G in $\mathcal{M}_1$,

$$F \prec_{\mathrm{cx}} G \quad \Leftrightarrow \quad \mathrm{ES}_p(F) \leqslant \mathrm{ES}_p(G) \text{ for all } p \in (0,1). \tag{3.14}$$

Moreover, by the comonotonic-additivity of ES_p, we have

$$\mathrm{ES}_p(G_i) = \sum_{j=1}^{n} \Lambda_{ij} \mathrm{ES}_p(F_j), \ \ i = 1, \dots, n.$$

Consequently,

$$\mathrm{ES}_p(G_i) = \sum_{j=1}^{n} \Lambda_{ij} \mathrm{ES}_p(F_j) \leqslant \mathrm{ES}_p(F_i), \ \ p \in (0,1), i = 1, \dots, n. \tag{3.15}$$

Noting that Λ is a doubly stochastic matrix, similarly as in the proof of (i), adding all the inequalities in (3.15) leads to

$$\sum_{i=1}^{n} \mathrm{ES}_p(G_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} \mathrm{ES}_p(F_j) = \sum_{i=1}^{n} \mathrm{ES}_p(F_i), \ \ p \in (0,1).$$

This implies that the inequalities in (3.15) are equalities, which means that $\Lambda \otimes \mathbf{F} = \mathbf{F}$ by (3.14). We complete the proof of (iii).

Finally, we consider (iv). (b) $\Rightarrow$ (a) is trivial. We will show (a) $\Rightarrow$ (b). By (3.14), $\Lambda\mathbf{F} \prec_{\mathrm{cx}} \mathbf{F}$ is equivalent to

$$\mathrm{ES}_p(F_i) \geqslant \mathrm{ES}_p\left(\sum_{j=1}^{n} \Lambda_{ij} F_j \right), \quad i = 1, \dots, n. \tag{3.16}$$

Moreover, by the concavity of ES_p on mixtures (e.g., Theorem 3 of Wang et al. (2020)), we have

$$\mathrm{ES}_p\left(\sum_{j=1}^{n} \Lambda_{ij} F_j \right) \geqslant \sum_{j=1}^{n} \Lambda_{ij} \mathrm{ES}_p(F_j).$$

Therefore, we have

$$\mathrm{ES}_p(F_i) \geqslant \mathrm{ES}_p\left(\sum_{j=1}^{n} \Lambda_{ij} F_j \right) \geqslant \sum_{j=1}^{n} \Lambda_{ij} \mathrm{ES}_p(F_j), \quad i = 1, \dots, n. \tag{3.17}$$

Adding the inequalities in (3.17) with noting that Λ is a doubly stochastic matrix yields

$$\sum_{i=1}^{n} \mathrm{ES}_p(F_i) \geqslant \sum_{i=1}^{n} \mathrm{ES}_p\left(\sum_{j=1}^{n} \Lambda_{ij} F_j \right) \geqslant \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} \mathrm{ES}_p(F_j) = \sum_{i=1}^{n} \mathrm{ES}_p(F_i).$$

Hence

$$\sum_{i=1}^{n} \mathrm{ES}_p(F_i) = \sum_{i=1}^{n} \mathrm{ES}_p\left(\sum_{j=1}^{n} \Lambda_{ij} F_j\right),$$

which implies that inequalities in (3.16) are all equalities. We establish the claim by (3.14). □

Proof of Proposition 3.6. (i) Note that $(P_{\alpha,\theta})^{-1} = \theta(P_{\alpha,1})^{-1}$ for $\theta, \alpha > 0$. Hence we prove (i) by showing that

$$(\Lambda \otimes \mathbf{P}_{\alpha,\boldsymbol{\theta}})^{-1} = \Lambda(\mathbf{P}_{\alpha,\boldsymbol{\theta}})^{-1} = \mathbf{P}^{-1}_{\alpha,\Lambda\boldsymbol{\theta}}.$$

(ii) Let $\mathcal{U}$ be the set of uniform random variables on $[0,1]$. By monotonicity of ρ, we have, for $0 < \alpha_1 < \alpha_2$,

$$\begin{aligned}\overline{\rho}(\mathbf{P}_{\alpha_1,\boldsymbol{\theta}}) &= \sup\left\{\rho\left(\theta_1 U_1^{-1/\alpha_1} + \cdots + \theta_n U_n^{-1/\alpha_1}\right) \mid U_1, \dots, U_n \in \mathcal{U}\right\} \\ &\geqslant \sup\left\{\rho\left(\theta_1 U_1^{-1/\alpha_2} + \cdots + \theta_n U_n^{-1/\alpha_2}\right) \mid U_1, \dots, U_n \in \mathcal{U}\right\} = \overline{\rho}(\mathbf{P}_{\alpha_2,\boldsymbol{\theta}}).\end{aligned}$$

This implies that $\overline{\rho}(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ is decreasing in α.

(iii) By monotonicity of ρ, we can establish the claim of (iii) similarly as the proof of (ii). □

3.10.3 Some further properties of the $\overline{\mathrm{VaR}}_p$ for Pareto risks

Properties of $\overline{\rho}(P_{\alpha,\boldsymbol{\theta}})$ in Proposition 3.6 can be strengthened for $\rho = \mathrm{VaR}_p$.

Proposition 3.13. *For $p \in (0,1)$, $\alpha > 0$ and $\boldsymbol{\theta} \in (0,\infty)^n$,*

(i) $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ *is increasing and continuous in p;*

(ii) $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ *is decreasing and continuous in α;*

(iii) $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ *is increasing and continuous in each component of $\boldsymbol{\theta}$;*

(iv) $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ *is homogeneous in $\boldsymbol{\theta}$, that is, for $\lambda > 0$,*

$$\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\lambda\boldsymbol{\theta}}) = \lambda\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}});$$

(v) If $\alpha > 1$, then

$$\frac{\mathbf{1}\cdot\boldsymbol{\theta}}{(1-p)^{1/\alpha}} \leqslant \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \frac{\alpha}{\alpha-1} \times \frac{\mathbf{1}\cdot\boldsymbol{\theta}}{(1-p)^{1/\alpha}}. \tag{3.18}$$

Proof. (i) As the quantile of Pareto distribution is continuous, by Lemma 4.4 and 4.5 of Bernard et al. (2014), $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ is continuous in p on $(0,1)$.

(ii) Let $\mathcal{U}$ be the set of uniform random variables distributed on $(0,1)$. We note that

$$\begin{aligned}\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) &= \sup\left\{\mathrm{VaR}_p\left(\theta_1 U_1^{-1/\alpha} + \cdots + \theta_n U_n^{-1/\alpha}\right) : U_1,\dots,U_n \in \mathcal{U}\right\} \\ &= \sum_{i=1}^{n} \theta_i \sup\left\{\mathrm{VaR}_{1-p}\left(M_{\alpha,\boldsymbol{\theta}}(U_1,\dots,U_n)\right) : U_1,\dots,U_n \in \mathcal{U}\right\}^{-\frac{1}{\alpha}},\end{aligned}$$

where $M_{\alpha,\boldsymbol{\theta}}(u_1,\dots,u_n) = \left(\theta_1 u_1^{-1/\alpha} + \cdots + \theta_n u_n^{-1/\alpha}\right)^{-\alpha} / \left(\sum_{i=1}^n \theta_i\right)^{-\alpha}$, $u_i \in (0,1)$ for $i = 1,\dots,n$. Let $\underline{\boldsymbol{\theta}} = \min(\boldsymbol{\theta}/\left(\sum_{i=1}^n \theta_i\right))$. With the classic averaging inequalities, for $0 < \alpha_1 < \alpha_2$, $M_{\alpha_1,\boldsymbol{\theta}} \leqslant M_{\alpha_2,\boldsymbol{\theta}}$ (Hardy et al. (1934), Theorem 16) and $\underline{\boldsymbol{\theta}}^{\alpha_1} M_{\alpha_1,\boldsymbol{\theta}} \geqslant \underline{\boldsymbol{\theta}}^{\alpha_2} M_{\alpha_2,\boldsymbol{\theta}}$ (Hardy et al. (1934), Theorem 23). We note that $0 < M_{\alpha,\boldsymbol{\theta}} < 1$ and these two inequalities are directly translated to

$$\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha_2,\boldsymbol{\theta}})^{\alpha_2/\alpha_1} \leqslant \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha_1,\boldsymbol{\theta}}) \leqslant \underline{\boldsymbol{\theta}}^{1-\alpha_2/\alpha_1}\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha_2,\boldsymbol{\theta}})^{\alpha_2/\alpha_1}.$$

By letting $\alpha_1 \uparrow \alpha_2$ and $\alpha_2 \downarrow \alpha_1$, we get the continuity of $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$ in $\alpha > 0$.

(iii) Without loss of generality, we assume $\boldsymbol{\theta}_1 = (\theta_1,\dots,\theta_n)$ and $\boldsymbol{\theta}_2 = (\lambda\theta_1,\dots,\theta_n)$, $\lambda > 0$. The monotonicity relative to $\boldsymbol{\theta}$ follows directly from Proposition 3.6. Using the homogeneity of $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}})$, which is proved in (iv), and the monotonicity with respect to $\boldsymbol{\theta}$ if $0 < \lambda < 1$,

$$\lambda\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}_1}) \leqslant \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}_2}) \leqslant \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}_1}),$$

otherwise

$$\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}_1}) \leqslant \overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}_2}) \leqslant \lambda\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}_1}).$$

By letting $\lambda \uparrow 1$ and $\lambda \downarrow 1$, we get the desired result.

(iv) For $\lambda > 0$,

$$\begin{aligned}\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\lambda\boldsymbol{\theta}}) &= \sup\{\mathrm{VaR}_p(G) : G \in \mathcal{D}_n(\mathbf{P}_{\alpha,\lambda\boldsymbol{\theta}})\} \\ &= \sup\left\{\mathrm{VaR}_p\left(G\left(\frac{\cdot}{\lambda}\right)\right) : G \in \mathcal{D}_n(\mathbf{P}_{\alpha,\boldsymbol{\theta}})\right\} \\ &= \lambda \sup\{\mathrm{VaR}_p(G) : G \in \mathcal{D}_n(\mathbf{P}_{\alpha,\boldsymbol{\theta}})\} = \lambda\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}).\end{aligned}$$

(v) For $\alpha > 1$, we haev

$$\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \leqslant \overline{\mathrm{ES}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) = \sum_{i=1}^{n} \mathrm{ES}_p(P_{\alpha,\theta_i}) = \alpha \sum_{i=1}^{n} \theta_i / \left((\alpha-1)(1-p)^{1/\alpha}\right),$$

and $\overline{\mathrm{VaR}}_p(\mathbf{P}_{\alpha,\boldsymbol{\theta}}) \geqslant \sum_{i=1}^{n} \mathrm{VaR}_p(P_{\alpha,\theta_i}) = \sum_{i=1}^{n} \theta_i/(1-p)^{1/\alpha}$. □

Chapter 4

Risk Aggregation under Dependence Uncertainty and an Order Constraint

4.1 Introduction

Quantifying the risk of a portfolio has gained much interest in the literature of finance and actuarial science. To accurately estimate the risk level, the joint distribution of the risks needs to be specified. However, it is challenging to estimate or test the dependence structure of a portfolio. Given known marginal distributions but unspecified dependence structure of risks, one of the most relevant problems is to find the worst-case (the largest possible) and the best-case (the smallest possible) values of a risk measure over all the possible dependence structures; see Embrechts and Puccetti (2006), Bernard et al. (2014) and Embrechts et al. (2013, 2015) for general discussions.

While bounds for risk measures calculated based on the sole knowledge of marginal distributions are generally wide, many attempts have been made to narrow them by incorporating partial dependence information into the problem. For instance, a variance constraint is imposed at the portfolio level by Bernard et al. (2017). A lower bound is placed on the corresponding copula of risks by Puccetti et al. (2016). Puccetti et al. (2017) assumed that certain groups of risks are independent while the dependence structure is unknown within each group. Bernard et al. (2017b) considered a partially specified factor model with dependence uncertainty.

In the literature of isotonic regression, order constraint on the expectations of target variables has been widely used in many practical applications; see Section 1 of Henzi

et al. (2021) for an overview. For two random variables ξ_1 and ξ_2, an isotonic regression problem has the constraint $\mathbb{E}[\xi_1] \leqslant \mathbb{E}[\xi_2]$. In many situations, while the risks ξ_1 and ξ_2 can be affected by a common shock Z (e.g., market risk, pandemic, natural disaster), one can impose a stricter but natural assumption, that is, $\mathbb{E}[\xi_1|Z] \leqslant \mathbb{E}[\xi_2|Z]$. In this chapter, we study the aggregation $S = X + Y$ given known marginal distributions with the order constraint $X \leqslant Y$, which might arise from, for instance, the above setting where $X = \mathbb{E}[\xi_1|Z]$ and $Y = \mathbb{E}[\xi_2|Z]$. In practice, insurance companies can divide the loss of a portfolio into different categories according to the riskiness of the contract, and the order constraint naturally holds in situations where one risk triggers another. For instance, when floods occur, the higher floors of apartments/houses will suffer losses only if there is a huge damage in lower floors. As another example, some cost categories for an insurance company, such as rehabilitation costs, can only occur as a consequence of some severe disease.

Before imposing the order constraint, one should verify that one of the two distributions is stochastically smaller than the other. For real data, this relation can be tested via, e.g., the methods of Barrett and Donald (2003). Statistical inference for distributions ordered stochastically can be carried out through the isotonic distributional regression of Henzi et al. (2021).

Fix an atomless probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let $\mathcal{M}$ be the set of cdfs on $\mathbb{R}$. For $F, G \in \mathcal{M}$ such that F is stochastically smaller than G, define the set

$$\mathcal{F}_2^o(F, G) = \{(X, Y) : X \sim F,\ Y \sim G,\ X \leqslant Y\}.$$

Here and throughout, the inequality $X \leqslant Y$ is understood in the almost sure sense. For a risk measure ρ, we are interested in the worst-case and best-case values of ρ over the set $\mathcal{F}_2^o(F, G)$ denoted by

$$\overline{\rho}(\mathcal{F}_2^o(F, G)) := \sup\{\rho(X + Y) : (X, Y) \in \mathcal{F}_2^o(F, G)\}, \tag{4.1}$$

and

$$\underline{\rho}(\mathcal{F}_2^o(F, G)) := \inf\{\rho(X + Y) : (X, Y) \in \mathcal{F}_2^o(F, G)\}.$$

We mainly deal with the case where ρ is a *tail risk measure* introduced in Liu and Wang (2021). The class of tail risk measures includes some of the most prevalent risk measures such as Value-at-Risk (VaR), Expected Shortfall (ES), and Range Value-at-Risk (RVaR). Generally speaking, the value of a tail risk measure is determined by the risk's upper tail behavior. A key feature for a tail risk measure ρ is that there exists another risk measure ρ^*, called the generator, such that $\rho(X) = \rho^*(X^*)$ where the random variable X^* follows the upper tail distribution of the random variable X.

In an unconstrained problem (i.e., only the marginal distributions of the two risks are known), for a tail risk measure ρ such that ρ^* is consistent with concave order, the worst-case value of ρ is attained by letting the upper tail risks be countermonotonic (i.e., the lower Fréchet-Hoeffding bound). In particular, if ρ is VaR, early results date back to Makarov (1981) and Rüschendorf (1982). Aggregation of more than two risks is much more challenging; see Wang et al. (2013), Puccetti and Rüschendorf (2013), Jakobsons et al. (2016) and Blanchet et al. (2020) for some analytical results. The Rearrangement Algorithm (RA) is developed by Puccetti and Rüschendorf (2012) and Embrechts et al. (2013) for numerical computation.

The problem with the order constraint is more sophisticated. Recently, Arnold et al. (2020) obtained the pointwise lower bound $D_*^{F,G}$ on joint distribution of (X, Y) in $\mathcal{F}_2^o(F, G)$. In mass transportation theory, Nutz and Wang (2021) proposed the *directional optimal transport* between two random variables, and the corresponding joint distribution is also $D_*^{F,G}$. The transport is called directional because of the constraint $Y \geqslant X$; that is, X can only be transported upwards to Y. Since $D_*^{F,G}$ is the smallest distribution function among all joint distributions of $(X, Y) \in \mathcal{F}_2^o(F, G)$, we will call the distribution $D_*^{F,G}$ the *directional lower (DL) coupling*, and $(X^*, Y^*) \sim D_*^{F,G}$ is said to be *DL-coupled*. From the minimality of $D_*^{F,G}$ and results on concordance order of Müller and Scarsini (2000), $X^* + Y^*$ is the largest in concave order among $X + Y$ where $(X, Y) \in \mathcal{F}_2^o(F, G)$.

In general, we are interested in risk measures such as VaR and RVaR, which are not monotone in concave or convex order. Therefore, the DL coupling does not give the maximum or minimum values of these risk measures, and considerable new techniques need to be developed to find bounds on these risk measures. Although VaR is not monotone in convex or concave order, its generator, the essential infimum, is monotone in concave order. As the main contribution of the chapter (Theorem 4.3), we show that for a tail risk measure ρ with a generator ρ^* that is monotone in concave order (such as VaR and RVaR), the solution to the constrained problem (4.1) can be obtained by using the upper tail distributions of risks. Moreover, the worst-case value of ρ with the order constraint is attained by letting the two upper tail risks be DL-coupled. The above assertions on tail risk measures are based on a novel technical result of monotone embedding (Theorem 4.2).

Despite its natural form, the order constraint in this chapter can be quite strong and may not be easy to verify in some applications. Moreover, significant reduction of uncertainty bounds occurs when the two risks have comparable sizes, making the order constraint harder to justify; see Section 4.6. As such, our contributions should be seen as mainly theoretical, and they will serve as fundamental tools for applications emerging in the future.

The rest of the chapter is organized as follows. In Section 4.2, we give a brief review on

comonotonicity, countermonotonicity, and the DL coupling. In Section 4.3, we study the worst-case dependence structures of risk aggregation with the order constraint in concave order. In Section 4.4, the notion of strong stochastic order is introduced. With this notion, we obtain several useful theoretical results. The main technical contributions are contained in Section 4.5, where we obtain worst-case and best-case values of tail risk measures with the order constraint. Analytical results for VaR and probability bounds are obtained. In Section 4.6, numerical studies are conducted to illustrate the impact of the order constraint on the bounds of risk measures. Some concluding remarks and an open question are discussed in Section 4.7.

We conclude this section by providing additional notations and terminologies that will be used throughout this chapter. A cdf F is said to be smaller than a cdf G in stochastic order if $F \geqslant G$, denoted by $F \leqslant_{\mathrm{st}} G$. Throughout, whenever $\mathcal{F}_2^o(F,G)$ appears, F and G are two distributions satisfying $F \leqslant_{\mathrm{st}} G$. A cdf F (or a random variable $X \sim F$) is said to be smaller than a cdf G (or a random variable $Y \sim G$) in concave order if $\mathbb{E}[u(X)] \leqslant \mathbb{E}[u(Y)]$ for all concave functions $u : \mathbb{R} \to \mathbb{R}$ provided that the expectations exist, and we denote this by $X \leqslant_{\mathrm{cv}} Y$ or $F \leqslant_{\mathrm{cv}} G$. Further, F is smaller than G in convex order if $G \leqslant_{\mathrm{cv}} F$, and this is denoted by $F \leqslant_{\mathrm{cx}} G$. The order $X \leqslant_{\mathrm{cx}} Y$ for two random variables X and Y is defined similarly. For more properties of these stochastic orders, we refer to Shaked and Shanthikumar (2007). A law-invariant risk measure ρ is a mapping from $\mathcal{M}$ to $\mathbb{R}$. In addition, we write $\rho(X) = \rho(F)$ for a random variable X with distribution F; thus, ρ can also be interpreted as a mapping from the set of random variables to $\mathbb{R}$. An empty set is denoted by $\emptyset$. By convention, $\inf \emptyset = \infty$ and $\sup \emptyset = -\infty$.

4.2 The directional lower coupling

In this section, we collect some basic results on comonotonicity, countermonotonicity, and the DL coupling, which will be useful for our chapter.

A random vector (X, Y) is said to be *comonotonic* if there exists a random variable U and two increasing functions f and g such that $X = f(U)$ and $Y = g(U)$ almost surely. A random vector (X, Y) is *countermonotonic* if $(X, -Y)$ is comonotonic. We refer to Dhaene et al. (2002, 2006) for a review on comonotonicity and Puccetti and Wang (2015) for negative dependence concepts including countermonotonicity.

Let the random vectors (X^{ct}, Y^{ct}), (X, Y) and (X^c, Y^c) be such that they have the same marginal distributions, (X^{ct}, Y^{ct}) is countermonotonic, and (X^c, Y^c) is comonotonic. It is

well known that

$$X^c + Y^c \leqslant_{\mathrm{cv}} X + Y \leqslant_{\mathrm{cv}} X^{ct} + Y^{ct}; \tag{4.2}$$

see e.g., Rüschendorf (2013, Corollary 3.28).[1] For $F, G \in \mathcal{M}$, let $X^c, X^{ct} \sim F$ and $Y^c, Y^{ct} \sim G$. Note that if $F \leqslant_{\mathrm{st}} G$, then $X^c \leqslant Y^c$, which can be easily checked by choosing U as a uniform random variable over $(0,1)$, and choosing f and g as left quantiles of F and G, respectively. Hence, $\mathcal{F}_2^o(F,G)$ contains comonotonic random vectors. However, (X^{ct}, Y^{ct}) may violate the order constraint unless the essential supremum of F is less than or equal to the essential infimum of G. Therefore, (X^{ct}, Y^{ct}) may not be in $\mathcal{F}_2^o(F,G)$.

To find an alternative for countermonotonicity in $\mathcal{F}_2^o(F,G)$, we need to introduce the DL coupling, whose distribution function is obtained by Arnold et al. (2020). Below, we explain the DL coupling in the context of mass transport following Nutz and Wang (2021), which is motivated by treatment effect analysis and causal inference (e.g., Manski (1997)). A directional coupling of F and G is the joint distribution of a random vector in $\mathcal{F}_2^o(F,G)$ and the DL coupling is the special case of a direction coupling which corresponds to the directional optimal transport of Nutz and Wang (2021). Let $X \sim F$ and $Y \sim G$. Denote by μ_F and μ_G the Borel probability measures generated by F and G, respectively. The directional optimal transport from X to Y can be constructed by considering the common part and the singular parts of μ_F and μ_G separately. We first assume that F and G are continuous distributions. The common part $\mu_F \wedge \mu_G$ is defined as the maximal measure θ such that $\theta \leqslant \mu_F$ and $\theta \leqslant \mu_G$. The singular parts of μ_F and μ_G are defined as $\mu_F' = \mu_F - \mu_F \wedge \mu_G$ and $\mu_G' = \mu_G - \mu_F \wedge \mu_G$. The shaded areas of density plots in Figure 4.1 illustrate the idea of the common and singular parts for two Pareto distributions $F(x) = 1 - 1/x$ for $x \geqslant 1$, and $G(y) = 1 - 2/y$ for $y \geqslant 2$.

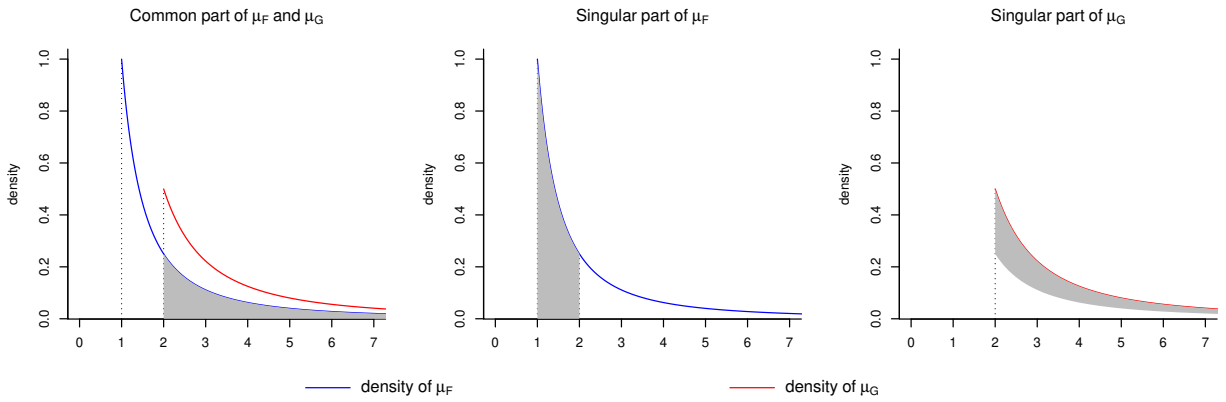
The directional optimal transport between μ_F and μ_G can be described in two pieces. First, the common part of μ_F and μ_G couples identically to each other. The transport from the singular part of μ_F to the singular part of μ_G, denoted by $T^{F,G}$, is defined as

$$T^{F,G}(x) = \inf\left\{z \geqslant x : F(z) - G(z) < F(x) - G(x)\right\}.$$

Corollary 2.4 of Nutz and Wang (2021) gives the following representation of the DL coupling

[1] We choose to work mainly with concave order instead of convex order because a major target of this chapter is to study VaR bounds, and the generator of VaR is increasing in concave-order; see Sections 4.4 and 4.5. Nevertheless, since $\leqslant_{\mathrm{cv}}$ is the same as $\geqslant_{\mathrm{cx}}$, all statements on concave order in this chapter can be equivalently stated using convex order. Convex order is common in the literature of risk management, e.g., Denuit et al. (2005).

Figure 4.1: Common and singular parts of μ_F and μ_G



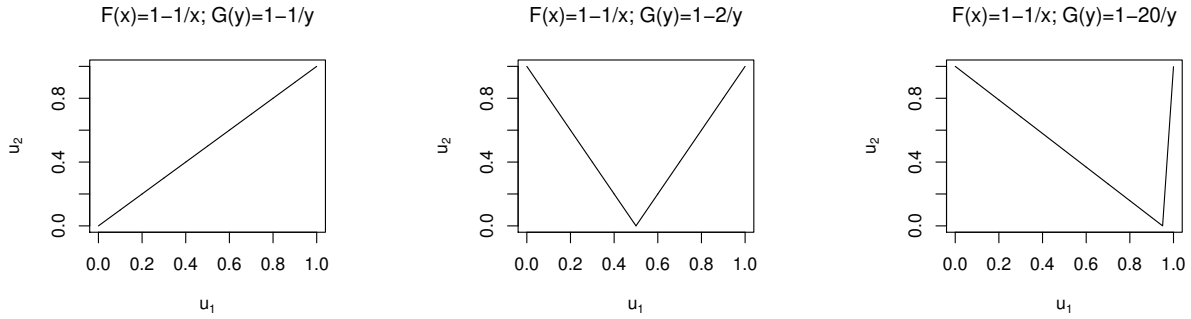
$D_*^{F,G}$, the joint distribution of (X, Y) obtained above,

$$D_*^{F,G}(x,y) = \begin{cases} G(y) & \text{if } y \leqslant x, \\ F(x) - \inf\limits_{z \in [x,y]} \{F(z) - G(z)\} & \text{if } y > x, \end{cases} \tag{4.3}$$

which is also the bivariate distribution function in Theorem 6 of Arnold et al. (2020). For a random vector $(X, Y) \sim D_*^{F,G}$, we say that (X, Y) is *DL-coupled*. Since DL coupling couples the common part of distributions to itself via the identity, it is a maximal coupling which maximizes $\mathbb{P}(X = Y)$ given the marginal distributions of X and Y; see e.g., Thorisson (2000, p. 104-112). DL coupling differs from countermonotonicity in general, and they coincide if the essential supremum of F is less than or equal to the essential infimum of G. In this special case, all couplings between F and G are directional. In Figure 4.2, the support of the copula representing the DL coupling is plotted for three pairs of Pareto distributions. Since F and G are identical in Figure 4.2a, the DL coupling is equivalent to comonotonicity. The DL coupling in Figure 4.2b is a simple combination of comonotonicity and countermonotonicity on the common part and singular parts of μ_F and μ_G, respectively. The DL coupling in Figure 4.2c is more similar to countermonotonicity. We warn the reader that, in general, the DL coupling can be much more complicated than these simple cases for other choices of marginal distributions. Nutz and Wang (2021) showed that the DL coupling is the combination of one comonotonic coupling and countably many countermonotonic couplings, but these countermonotonic couplings may not be between conditional distributions on intervals like in these examples; see Proposition 2.6 and Example 6.3 of Nutz and Wang (2021).

One can also construct DL coupling for non-continuous distributions F and G satisfying

Figure 4.2: Support of the copula of $(U_1, U_2) = (F(X), G(Y))$ where $(X, Y) \sim D_*^{F,G}$



(a) (b) (c)

$F \leqslant_{\mathrm{st}} G$. The idea is first to convert the distributions F and G to continuous distributions F_c and G_c by a monotone transformation. Thereafter, construct DL coupling $D_*^{F_c,G_c}$ and reverse the transformation back to non-continuous case; see Section 5.4 of Nutz and Wang (2021) for more details. As an important fact, DL-coupled (X, Y) always exists if $F \leqslant_{\mathrm{st}} G$, and it has the distribution function (4.3).

Note that DL coupling may render the transport from X to Y randomized. That is, a realization of G may have two pre-images through directional optimal transport. In the language of mass transport theory, the directional optimal transport is Kantorovich-type but not necessarily Monge-type. We use the following example to illustrate how DL coupling introduces such randomness and affects the aggregation of risks.

Example 4.1 (Pareto risks: DL coupling)**.** Suppose that two risks follow Pareto distributions $F(x) = 1 - 1/x$ for $x \geqslant 1$, and $G(y) = 1 - 2/y$ for $y \geqslant 2$. Since $F \leqslant_{\mathrm{st}} G$, we can take $(X, Y) \sim D_*^{F,G}$. The directional optimal transport between the singular parts of μ_F and μ_G is

$$T^{F,G}(x) = \inf\left\{z \geqslant x : \frac{1}{z} < 1 - \frac{1}{x}\right\} = \frac{x}{x-1}, \quad x \in (1, 2].$$

If $x = 1$, $T^{F,G}(x) = \inf \emptyset = \infty$. The directional optimal transport between the common part of μ_F and μ_G is the identical transport. Thus, for a real number $y \geqslant 2$, its pre-image is either y through the identical transport or $y/(y-1)$ through $T^{F,G}$. Let $c^* = (c - \sqrt{c^2 - 4c})/2$

for $c \in [4, \infty]$. By Corollary 2.9 of Nutz and Wang (2021), we have

$$\begin{aligned}\mathbb{P}(X+Y \leqslant c) &= (\mu_F \wedge \mu_G)\left(\left[-\infty, \frac{c}{2}\right]\right) + \mu'_F\left([c^*, 2]\right) \\ &= \left(F\left(\frac{c}{2}\right) - F\left(2\right)\right) + \left(F(2) - F\left(c^*\right)\right) \\ &= \frac{c + \sqrt{c^2 - 4c} - 4}{2c}.\end{aligned}$$

This example will be continued in Examples 4.2, 4.5 and 4.6. In particular, if the order constraint is imposed, the DL coupling leads to the largest essential infimum of $X+Y$, but it does not lead to the largest or the smallest probability $\mathbb{P}(X+Y \leqslant t)$ in general.

4.3 Optimality of the directional lower coupling

In this section, we study the optimal dependence structures of $(X, Y) \in \mathcal{F}_2^o(F, G)$ in the sense of concave order, or equivalently, convex order. As we have seen in (4.2), comonotonicity of (X, Y) yields the smallest $X+Y$ in concave order among all possible dependence structures with given marginal distributions, and hence it also yields the smallest concave order of $X+Y$ for $(X, Y) \in \mathcal{F}_2^o(F, G)$. On the other hand, a simple result from Nutz and Wang (2021) shows that the DL coupling of (X, Y) yields the largest concave order of $X+Y$ over $\mathcal{F}_2^o(F, G)$. The concave ordering bounds are very useful for the calculation of bounds on risk measures.

Lemma 4.1. *For $(X, Y), (X^c, Y^c), (X', Y') \in \mathcal{F}_2^o(F, G)$ such that (X^c, Y^c) is comonotonic and (X', Y') is DL-coupled, we have*

$$X^c + Y^c \leqslant_{\mathrm{cv}} X + Y \leqslant_{\mathrm{cv}} X' + Y'.$$

Proof. The first inequality can be found in, e.g., Theorem 3.5 of Rüschendorf (2013). For the second inequality, by Theorem 2.2 (i) of Nutz and Wang (2021), (X', Y') is the smallest element of $\mathcal{F}_2^o(F, G)$ in concordance order. Equivalently, (X', Y') is the smallest element of $\mathcal{F}_2^o(F, G)$ in supermodular order (e.g., Theorem 2.5 of Müller and Scarsini (2000)). It is well known that the function $(x, y) \mapsto -u(x+y)$ on $\mathbb{R}^2$ for any concave function $u : \mathbb{R} \to \mathbb{R}$ is supermodular. Hence $\mathbb{E}[u(X+Y)] \leqslant \mathbb{E}[u(X'+Y')]$, and therefore $X+Y \leqslant_{\mathrm{cv}} X'+Y'$. □

Next, we use the above concave ordering bounds to obtain bounds on risk measures. We refer to Föllmer and Schied (2016) for an overview on risk measures. For a risk measure $\rho : \mathcal{M} \to \mathbb{R}$, we define three commonly used properties:

(i) A risk measure ρ is *monotone* if $\rho(F) \leqslant \rho(G)$ whenever $F \leqslant_{\mathrm{st}} G$;

(ii) A risk measure ρ is $\leqslant_{\mathrm{cv}}$-*consistent* if $\rho(F) \leqslant \rho(G)$ whenever $F \leqslant_{\mathrm{cv}} G$;

(iii) A risk measure ρ is $\leqslant_{\mathrm{cx}}$-*consistent* if $\rho(F) \leqslant \rho(G)$ whenever $F \leqslant_{\mathrm{cx}} G$.

Many popular risk measures are monotone, such as Value-at-Risk (VaR), Expected Shortfall (ES), and Range-VaR (RVaR). For $F \in \mathcal{M}$ and $q \in (0,1]$, the left VaR denoted by $\mathrm{VaR}_q^L : \mathcal{M} \to \mathbb{R}$ is given by

$$\mathrm{VaR}_q^L(F) = F^{-1}(q) = \inf\{t \in \mathbb{R} : F(t) \geqslant q\}.$$

For $p \in [0,1)$, the right VaR denoted by $\mathrm{VaR}_p^R : \mathcal{M} \to \mathbb{R}$ is given by

$$\mathrm{VaR}_p^R(F) = F^{-1}(p+) = \inf\{t \in \mathbb{R} : F(t) > p\}.$$

For $p = 0$ and $q = 1$, $\mathrm{VaR}_0^R(F)$ and $\mathrm{VaR}_1^L(F)$ correspond to the essential infimum and essential supremum of F which are also denoted by $\text{ess-inf}(F) = F^{-1}(0)$ and $\text{ess-sup}(F) = F^{-1}(1)$, respectively. For $F \in \mathcal{M}$, $\mathrm{ES}_p : \mathcal{M} \to \mathbb{R}$ for $p \in (0,1)$ is defined as

$$\mathrm{ES}_p(F) = \frac{1}{1-p}\int_p^1 \mathrm{VaR}_u^R(F)\mathrm{d}u,$$

and $\mathrm{RVaR}_{p,q} : \mathcal{M} \to \mathbb{R}$ for $0 \leqslant p < q < 1$ is defined as

$$\mathrm{RVaR}_{p,q}(F) = \frac{1}{q-p}\int_p^q \mathrm{VaR}_u^R(F)\mathrm{d}u.$$

The class of RVaR is proposed by Cont et al. (2010) as robust risk measures; see Embrechts et al. (2018) for its properties.

For $p \in (0,1)$, VaR_p^L and VaR_p^R are neither $\leqslant_{\mathrm{cx}}$-consistent nor $\leqslant_{\mathrm{cv}}$-consistent. On the other hand, ES_p and VaR_1^L are $\leqslant_{\mathrm{cx}}$-consistent, and $\mathrm{RVaR}_{0,q}$ and VaR_0^R are $\leqslant_{\mathrm{cv}}$-consistent. Monetary risk measures (see Föllmer and Schied (2016)) that are $\leqslant_{\mathrm{cx}}$-consistent are characterized by Mao and Wang (2020) and they admit an ES-based representation. Using Lemma 4.1, we immediately obtain the following bounds of $\leqslant_{\mathrm{cv}}$-consistent and $\leqslant_{\mathrm{cx}}$-consistent risk measures. Recall that the worst-case and best-case risk measures are defined as, respectively,

$$\overline{\rho}(\mathcal{F}_2^o(F,G)) = \sup\{\rho(X+Y) : (X,Y) \in \mathcal{F}_2^o(F,G)\},$$

and

$$\underline{\rho}(\mathcal{F}_2^o(F,G)) = \inf\{\rho(X+Y) : (X,Y) \in \mathcal{F}_2^o(F,G)\}.$$

Corollary 4.1. *Suppose that* $(X,Y),(X^c,Y^c),(X',Y') \in \mathcal{F}_2^o(F,G)$ *such that* (X^c,Y^c) *is comonotonic and* (X',Y') *is DL-coupled. If* ρ *is* $\leqslant_{\mathrm{cv}}$*-consistent, then*

$$\underline{\rho}(\mathcal{F}_2^o(F,G)) = \rho(X^c+Y^c) \leqslant \rho(X+Y) \leqslant \rho(X'+Y') = \overline{\rho}(\mathcal{F}_2^o(F,G)).$$

If ρ *is* $\leqslant_{\mathrm{cx}}$*-consistent, then*

$$\underline{\rho}(\mathcal{F}_2^o(F,G)) = \rho(X'+Y') \leqslant \rho(X+Y) \leqslant \rho(X^c+Y^c) = \overline{\rho}(\mathcal{F}_2^o(F,G)).$$

As seen from Corollary 4.1, for a $\leqslant_{\mathrm{cx}}$-consistent risk measure, such as a law-invariant convex or coherent risk measure, the extra order constraint does not improve the worst-case risk value obtained under comonotonicity, as in the case without the order constraint.

Since essential infimum and essential supremum are $\leqslant_{\mathrm{cv}}$-consistent and $\leqslant_{\mathrm{cx}}$-consistent respectively, with Corollary 4.1, we give analytical results on the worst-case value of essential infimum and the best-case value of essential supremum in the following theorem. This result will be used to derive the worst-case and best-case values of VaR in Section 4.5.

Theorem 4.1. *For continuous distributions* F*,* G *and* $(X,Y) \sim D_*^{F,G}$*, we have*

$$\overline{\text{ess-inf}}(\mathcal{F}_2^o(F,G)) = \text{ess-inf}(X+Y) = \min\left\{ \inf_{x\in[F^{-1}(0),G^{-1}(0)]} \left\{ T^{F,G}(x)+x \right\}, 2G^{-1}(0) \right\}, \tag{4.4}$$

and

$$\underline{\text{ess-sup}}(\mathcal{F}_2^o(F,G)) = \text{ess-sup}(X+Y) = \max\left\{ \sup_{x\in[F^{-1}(1),G^{-1}(1)]} \left\{ \hat{T}^{F,G}(x)+x \right\}, 2F^{-1}(1) \right\}, \tag{4.5}$$

where $\hat{T}^{F,G}(x) = \sup\{t \leqslant x : F(t)-G(t) < F(x)-G(x)\}$.

Proof. We first prove the statement (4.4) on the worst-case value of the essential infimum. Combining Corollary 4.1 and the fact that essential infimum is $\leqslant_{\mathrm{cv}}$-consistent, we have the first equality in (4.4). To prove the second equality in (4.4), for $(X,Y) \sim D_*^{F,G}$, let

$$t^* := \sup\left\{ t \in \mathbb{R} : \text{ for all } (x,y) \in \mathbb{R}^2 \text{ such that } t = x+y \text{ and } D(x,y) = 0 \right\},$$

where $D = D_*^{F,G}$. It is straightforward to see from the definition that $t^* = \text{ess-inf}(X+Y)$. Therefore, we have

$$F^{-1}(0) + G^{-1}(0) \leqslant \text{ess-inf}(X+Y) \leqslant 2G^{-1}(0). \tag{4.6}$$

If $F^{-1}(0) = G^{-1}(0)$, then clearly $\text{ess-inf}(X+Y) = 2G^{-1}(0)$. For the rest of the proof, we assume $F^{-1}(0) < G^{-1}(0)$.

(i) If $x \leqslant F^{-1}(0)$ or $y \leqslant G^{-1}(0)$, $D(x,y) = 0$.

(ii) If $x > G^{-1}(0)$ and $y > G^{-1}(0)$, it is easy to check $D(x,y) > 0$.

(iii) If $F^{-1}(0) < x \leqslant G^{-1}(0) < y$, we have

$$D(x,y) = \max\left\{ - \inf_{z \in [G^{-1}(0), y]} \{F(z) - G(z) - F(x)\}, 0 \right\}.$$

By definition of $T^{F,G}$, if $F^{-1}(0) < x \leqslant G^{-1}(0) < y \leqslant T^{F,G}(x)$, $D(x,y) = 0$. On the other hand, if $F^{-1}(0) < x \leqslant G^{-1}(0) \leqslant T^{F,G}(x) < y$, $D(x,y) > 0$.

As a result, $D(x,y) = 0$ if and only if one of the following holds: $x \leqslant F^{-1}(0)$, $y \leqslant G^{-1}(0)$ or $F^{-1}(0) < x \leqslant G^{-1}(0) < y \leqslant T^{F,G}(x)$. Let

$$s := \min\left\{ \inf_{x \in [F^{-1}(0), G^{-1}(0)]} \left\{ T^{F,G}(x) + x \right\}, 2G^{-1}(0) \right\}.$$

We will show $t^* = s$. For $x, y \in \mathbb{R}$, suppose that $x + y = s$. If $F^{-1}(0) < x \leqslant G^{-1}(0)$, we have $y \leqslant T^{F,G}(x)$. If $x > G^{-1}(0)$, we have $y < G^{-1}(0)$. That is, for any $x, y \in \mathbb{R}$ such that $x + y = s$, $D(x,y) = 0$. Thus we have $t^* \geqslant s$. For any $g, h \in \mathbb{R}$, suppose that $g + h = t^*$. Then for any $\varepsilon > 0$, we have $D(g, h - \varepsilon) = 0$. Therefore, if $F^{-1}(0) < g \leqslant G^{-1}(0)$, we have $h \leqslant T^{F,G}(g) + \varepsilon$. By letting ε goes to 0, we have $t^* = g + h \leqslant T^{F,G}(g) + g$ for any g in $(F^{-1}(0), G^{-1}(0)]$. As $T(F^{-1}(0)) = \infty$, we have $t^* \leqslant \inf_{x \in [F^{-1}(0), G^{-1}(0)]} \left\{ T^{F,G}(x) + x \right\}$. By (4.6), $t^* \leqslant 2G^{-1}(0)$. Thus we have $t^* \leqslant s$, and the statement (4.4) on the worst-case value of the essential infimum holds.

Next, we show the statement (4.5) on the best-case value of the essential supremum. Let $\hat{F}(t) := 1 - F(-t)$, $\hat{G}(t) := 1 - G(-t)$ for $t \in \mathbb{R}$ and $\hat{T}^{F,G}(x) = \sup\{t \leqslant x : F(t) - G(t) < F(x) - G(x)\}$ for $x \in \mathbb{R}$. $\hat{F}$ and $\hat{G}$ are the distributions of $-X$ and $-Y$. Then we have $-T^{\hat{G},\hat{F}}(x) = \hat{T}^{F,G}(-x)$ for $x \in \mathbb{R}$. Note that

$$\underline{\text{ess-sup}}(\mathcal{F}_2^o(F,G)) = -\sup\{\text{ess-inf}(-X - Y) : (-Y, -X) \in \mathcal{F}_2^o(\hat{G}, \hat{F})\}.$$

Applying (4.4), we get the desired equality. □

Remark 4.1. In the unconstrained case, the worst-case value of the essential infimum and the best-case value of the essential supremum are attained by countermonotonicity, i.e.,

$$\sup\{\text{ess-inf}(X + Y) : X \sim F,\ Y \sim G\} = \inf_{x \in [0,1]} \left\{ F^{-1}(x) + G^{-1}(1 - x) \right\},$$

$$\inf\{\text{ess-sup}(X + Y) : X \sim F,\ Y \sim G\} = \sup_{x \in [0,1]} \left\{ F^{-1}(x) + G^{-1}(1 - x) \right\}.$$

Generally, these bounds are different from the bounds in Theorem 4.1.

Example 4.2 (Pareto risks: Essential infimum)**.** In Example 4.1, we derive the cdf of $X+Y$ for $(X,Y)\sim D_*^{F,G}$ where F and G are two Pareto distributions. We have

$$\mathbb{P}(X+Y\leqslant c)=\frac{c+\sqrt{c^2-4c}-4}{2c},\qquad c\in[4,\infty).$$

By Corollary 4.1, $\overline{\text{ess-inf}}(\mathcal{F}_2^o(F,G))=\text{ess-inf}(X+Y)=4$. Alternatively, we can use the analytical result of Theorem 4.1. As $T^{F,G}(x)=x/(x-1)$ for $x\in(1,2]$ and $T^{F,G}(1)=\infty$, we have $\inf_{x\in[1,2]}\left\{T^{F,G}(x)+x\right\}=4$. Thus, we have $\overline{\text{ess-inf}}(\mathcal{F}_2^o(F,G))=\min\{4,4\}=4$. The unconstrained upper bound for ess-inf is $3+2\sqrt{2}$ which is larger than $\overline{\text{ess-inf}}(\mathcal{F}_2^o(F,G))$. Both the constrained and unconstrained lower bounds of ess-inf are attained when (X,Y) are comonotonic and we have $\underline{\text{ess-inf}}(\mathcal{F}_2^o(F,G))=3$.

4.4 Strong stochastic order and monotone embedding

In this section, we introduce the notion of strong stochastic order, and obtain several theoretical properties. The new notion is crucial for the main results of this chapter in Section 4.5.

For $F,G\in\mathcal{M}$, we say F is smaller than G in *strong stochastic order* if $G(y)-G(x)\geqslant F(y)-F(x)$ for all $y\geqslant x\geqslant G^{-1}(0)$, denoted by $F\leqslant_{\text{ss}}G$. Equivalently, the function $x\mapsto G(x)-F(x)$ is decreasing for $x\geqslant G^{-1}(0)$. Note that the order $\leqslant_{\text{ss}}$ is stronger than $\leqslant_{\text{st}}$, and hence the name. Intuitively, G has more probability in any interval $(x,y]$ than F if $x\geqslant G^{-1}(0)$. If F and G have densities g and f with respect to a dominating measure, then $F\leqslant_{\text{ss}}G$ if and only if $g(x)\geqslant f(x)$ for $x\geqslant G^{-1}(0)$. As far as we know, this notion of stochastic order is new to the literature. We first provide some simple properties of the order $\leqslant_{\text{ss}}$.

Proposition 4.1. *The strong stochastic order satisfies the following properties:*

(i) If $F\leqslant_{\text{ss}}G$ then $F\leqslant_{\text{st}}G$;

(ii) Assuming $F^{-1}(0)=G^{-1}(0)$, $F\leqslant_{\text{ss}}G$ if and only if $F=G$;

(iii) If $G^{-1}(0)=-\infty$, then $F\leqslant_{\text{ss}}G$ means $F=G$;

(iv) The relation $\leqslant_{\text{ss}}$ is a partial order.

Proof. (i) By letting $y \to \infty$, we have $1 - G(x) \geqslant 1 - F(x)$ for all $x \geqslant G^{-1}(0)$. Hence, $G(x) \leqslant F(x)$ for all $x \geqslant G^{-1}(0)$. Moreover, $G(x) = 0$ for $x < G^{-1}(0)$. Hence, $G(x) \leqslant F(x)$ for all $x \in \mathbb{R}$, which gives $F \leqslant_{\mathrm{st}} G$.

(ii) The "$\Leftarrow$" direction is obvious. For the "$\Rightarrow$" direction, by letting $x = F^{-1}(0) = G^{-1}(0)$, we have $F(x) - G(x) = 0$. Thus, for all $y \geqslant x = F^{-1}(0) = G^{-1}(0)$, $F(y) \leqslant G(y)$, which means $G \leqslant_{\mathrm{st}} F$. Together with $F \leqslant_{\mathrm{st}} G$ from (i), we have $F = G$.

(iii) By (i), we have $F^{-1}(0) \leqslant G^{-1}(0) = -\infty$. Thus, $F^{-1}(0) = G^{-1}(0) = -\infty$. Hence, $F = G$ by (ii).

(iv) Reflexivity is obvious and antisymmetry is implied by (i). Suppose that $F \leqslant_{\mathrm{ss}} G$ and $G \leqslant_{\mathrm{ss}} H$. By (i), $G \leqslant_{\mathrm{st}} H$ and $\max\{G^{-1}(0), H^{-1}(0)\} = H^{-1}(0)$. We have $H(y) - H(x) \geqslant F(y) - F(x)$ for all $y \geqslant x \geqslant \max\{G^{-1}(0), H^{-1}(0)\} = H^{-1}(0)$. Transitivity of the order $\leqslant_{\mathrm{ss}}$ is proved. □

Next, we discuss the problem of *monotone embedding*, which is an important issue in the analysis of risk aggregation for tail risk measures in Section 4.5. The problem is formulated as follows. Suppose that $F \leqslant_{\mathrm{st}} F' \leqslant_{\mathrm{st}} G$ and $(X, Y) \in \mathcal{F}_2^o(F, G)$, and the question is whether there exists $X' \sim F'$ such that $X \leqslant X' \leqslant Y$ holds (in the almost sure sense). The existence of such X' is crucial to prove that we can use tail distribution to obtain bounds on tail risk measures (see Theorem 4.3 below). Unfortunately, in general, such X' does not necessarily exist, even if we restrict to DL-coupled (X, Y).

Example 4.3. Let G be the Bernoulli(1/2) distribution. Take $Y \sim G$, let $X = -Y$, and F be the distribution of X. Clearly, (X, Y) is countermonotonic, and hence (X, Y) is DL-coupled. Take another random variable $X' \sim F' = \mathrm{U}[-1, 1]$. It is easy to see that $F \leqslant_{\mathrm{st}} F' \leqslant_{\mathrm{st}} G$. Since $\mathbb{P}(X = Y) = 1/2$ but $\mathbb{P}(X' = Y) = 0$, we know that $X \leqslant X' \leqslant Y$ cannot hold for any $X' \sim F'$.

The next theorem is the most important technical result which allows us to study the DL coupling using the strong stochastic order. The result says that, although $F \leqslant_{\mathrm{st}} F' \leqslant_{\mathrm{st}} G$ is not sufficient for the existence of X' in Example 4.3, assuming the stronger relation $F \leqslant_{\mathrm{ss}} F'$ would suffice.

Theorem 4.2 (Monotone embedding)**.** *Suppose that $F \leqslant_{\mathrm{ss}} F' \leqslant_{\mathrm{st}} G$, and $(X, Y) \sim D_*^{F,G}$. Then there exists $X' \sim F'$ such that $X \leqslant X' \leqslant Y$ almost surely and (X', Y) is DL-coupled.*

Proof. We first consider continuous distributions F, F' and G. Without loss of generality, we assume μ_F and μ_G are not mutually singular. As $(X, Y) \sim D_*^{F,G}$, the common part

$\mu_F \wedge \mu_G$ of μ_F and μ_G are identically coupled. The singular part of μ_F is transported to the singular part of μ_G through $T^{F,G}$. Let P be a joint distribution on $\mathbb{R}^3$ with marginals $P \circ X^{-1} = \mu_F$, $P \circ (X')^{-1} = \mu_{F'}$ and $P \circ Y^{-1} = \mu_G$ such that $(X, Y) \sim D_*^{F,G}$. We will construct P such that (X', Y) is DL-coupled and $X \leqslant X' \leqslant Y$ almost surely.

(i) Let $\theta = \mu_F \wedge \mu_G$ and $\theta' = \mu_{F'} \wedge \mu_G$. As $F \leqslant_{\mathrm{ss}} F'$, $\mu_F(a, b] \leqslant \mu_{F'}(a, b]$ for all $b \geqslant a \geqslant (F')^{-1}(0)$. Thus the common part of $\mu_{F'}$ and μ_G covers the common part of μ_F and μ_G, i.e., $\theta \wedge \theta' = \theta$. Therefore, we can always construct P such that the measure θ of μ_F, $\mu_{F'}$ and μ_G identically couples with each other. By further letting P couple the measure $\theta' - \theta$ of $\mu_{F'}$ and μ_G identically, the common part θ' of $\mu_{F'}$ and μ_G identically couples.

(ii) Next we focus on the directional optimal transports on the singular parts of distributions, i.e., $T^{F,G}$ and $T^{F',G}$. Let P transport the singular part of $\mu_{F'}$ to the singular part of μ_G through $T^{F',G}$. Take x, x' and y satisfying $y = T^{F,G}(x) = T^{F',G}(x')$. Note that we will not consider the sets of x and x' such that $x = y$ or $x' = y$ as we are studying the singular parts of distributions. Thus we have $x < y$ and $x' < y$. A key property of $T^{F,G}$ is that $F(z) - G(z) = F\left(T^{F,G}(z)\right) - G\left(T^{F,G}(z)\right)$ holds for all $z \in \mathbb{R}$; see Lemma 5.2 of Nutz and Wang (2021). With this property and $F \leqslant_{\mathrm{ss}} F'$, we have

$$\begin{aligned} F'(x) - G(x) &= F'(x) - F(x) + F(x) - G(x) \\ &\leqslant F'(y) - F(y) + F(y) - G(y) = F'(x') - G(x'). \end{aligned}$$

Assume that $x' < x$. If $F'(x) - G(x) < F'(x') - G(x')$, as $x' < x < y = T^{F',G}(x')$, by definition of $T^{F',G}$, we have $x = y$ as a contradiction to $x < y$. If $F'(x) - G(x) = F'(x') - G(x')$, as $x' < x < y = T^{F',G}(x')$, x is neither a point of strict increase nor a point of strict decrease of $F' - G$ in the sense of Nutz and Wang (2021). By Proposition 5.1 of Nutz and Wang (2021), the set of points which are neither of strict increase nor of strict decrease is a null set.

By (i) and (ii), we construct P such that $(X', Y) \sim D_*^{F',G}$ and $X \leqslant X' \leqslant Y$ almost surely. Note that $(X, Y) \sim D_*^{F,G}$ and $(X', Y) \sim D_*^{F',G}$ do not necessarily imply $X \leqslant X' \leqslant Y$ almost surely due to the randomness of DL coupling which is illustrated by Example 4.1. Therefore, the construction of P in (i) is necessary. Next, we proceed to complete the proof for non-continuous distributions F, F', and G. As the construction in (i) can also be applied to the common part of non-continuous distributions, we focus on the singular parts of distributions and assume that $\mu_F \wedge \mu_G = 0$ and $\mu_{F'} \wedge \mu_G = 0$. Let

$$j(x) = x + \sum_{y \leqslant x} \left| H(y) - H(y-) + (F(y) - F(y-)) \mathbb{1}_{\{y < (F')^{-1}(0)\}} \right|, \quad x \in \mathbb{R},$$

where $H = F' - G$. The function j is the summation of an identity function, the jumps of H and the jumps of $F(x)$ for $x < (F')^{-1}(0)$. Denote by $j^{-1} : j(\mathbb{R}) \to \mathbb{R}$ the right-continuous inverse function of j. Let

$$J_x = [j(x-), j(x)]$$

be the interval representing the jump of j at x. If there is no jump at x, J_x is a singleton. Next we convert the measure $\mu_{F'}$ to an auxiliary measure $\mu_{F'_c}$ with continuous cdf F'_c. We set $F'_c(z) = F'(j^{-1}(z))$ for $z \in j(\mathbb{R})$. On the complement of $j(\mathbb{R})$, F'_c is defined by linearly interpolating from its values on $j(\mathbb{R})$. In other words, if $\mu_{F'}$ has a jump at x, $\mu_{F'_c}$ is uniformly distributed on the interval J_x with probability $\mu_{F'_c}(J_x) = \mu_{F'}(\{x\})$. The auxiliary measures μ_{F_c} and μ_{G_c} with cdfs F_c and G_c can be constructed similarly from μ_F and μ_G. The transformation implies that G_c is also continuous and $F'_c \leqslant_{\mathrm{st}} G_c$. Note that as $F \leqslant_{\mathrm{ss}} F'$, for $x \geqslant (F')^{-1}(0)$, we have

$$F(x) - F(x-) \leqslant F'(x) - F'(x-).$$

The above inequality implies that for any $x \geqslant (F')^{-1}(0)$, whenever $F(x)$ has a jump, $F'(x)$ must have one. Therefore, the transformation from μ_F to μ_{F_c} reduces all the atoms of μ_F and F_c is continuous. Moreover, as $F \leqslant_{\mathrm{ss}} F'$, the transformation ensures that $F_c \leqslant_{\mathrm{ss}} F'_c$. Consequently, the orders on F, F' and G are preserved after the transformation and we have $F_c \leqslant_{\mathrm{ss}} F'_c \leqslant_{\mathrm{st}} G_c$.

Apply the result for continuous distributions on F_c, F'_c and G_c and convert the transformation back to non-continuous distributions. As all transformations are monotone (see Theorem 5.5 of Nutz and Wang (2021)), the order $X \leqslant X' \leqslant Y$ still holds almost surely for non-continuous distributions F, F' and G. □

In what follows, for any set $A \in \mathcal{A}$ with positive probability, let $H_{X|A}$ be the conditional distribution of X given A. Moreover, $F^{[p,1]}$ is the upper p-tail distribution of F, namely

$$F^{[p,1]}(x) = \frac{(F(x) - p)_+}{1 - p}, \quad x \in \mathbb{R},$$

and $F^{[0,p]}$ is the lower p-tail distribution of F, namely

$$F^{[0,p]}(x) = \frac{F(x) \wedge p}{p}, \quad x \in \mathbb{R}.$$

In other words, $F^{[p,1]}$ is the distribution of $F^{-1}(U)$ where $U \sim \mathrm{U}[p, 1]$, and $F^{[0,p]}$ is the distribution of $F^{-1}(U)$ where $U \sim \mathrm{U}[0, p]$. The next proposition shows that the largest

conditional distribution $H_{X|A}$ for $A \in \mathcal{A}$ with probability $1-p$ in strong stochastic order is the upper p-tail distribution $F^{[p,1]}$ where $X \sim F$. The event A such that $H_{X|A} = F^{[p,1]}$ is called a p-tail event in Wang and Zitikis (2021), which will be formally defined in Section 4.5.

Proposition 4.2. *For $p \in (0,1)$, any set $A \in \mathcal{A}$ of probability $1-p$ and $X \sim F$, $H_{X|A} \leqslant_{\mathrm{ss}} F^{[p,1]}$.*

Proof. For any interval $[x,y]$ with $x \geqslant \left(F^{[p,1]}\right)^{-1}(0)$, we have

$$F^{[p,1]}(y) - F^{[p,1]}(x) = \frac{(F(y)-p)_+ - (F(x)-p)_+}{1-p} = \frac{F(y)-F(x)}{1-p},$$

and

$$H_{X|A}(y) - H_{X|A}(x) = \mathbb{P}(x < X \leqslant y \mid A) = \frac{\mathbb{P}(\{x < X \leqslant y\} \cap A)}{1-p} \leqslant \frac{F(y)-F(x)}{1-p}.$$

Hence, we have $H_{X|A} \leqslant_{\mathrm{ss}} F^{[p,1]}$. □

Combining Theorem 4.2 and Proposition 4.2, we immediately arrive at the following corollary. This corollary will be used to establish the main result on the worst-case value of tail risk measures with the order constraint in Section 4.5.

Corollary 4.2. *Let $A \in \mathcal{A}$ with probability $1-p$ and $p \in (0,1)$. Suppose that $F \leqslant_{\mathrm{st}} G$, $X \sim F$ and $(X_A, Y) \sim D_*^{H_{X|A}, G^{[p,1]}}$. Then there exists $X' \sim F^{[p,1]}$ such that $X_A \leqslant X' \leqslant Y$ almost surely and (X', Y) is DL-coupled.*

Proof. Note that $F^{[p,1]} \leqslant_{\mathrm{st}} G^{[p,1]}$ follows from $F \leqslant_{\mathrm{st}} G$, and $H_{X|A} \leqslant_{\mathrm{ss}} F^{[p,1]}$ follows from Proposition 4.2. Applying Theorem 4.2 with the condition $H_{X|A} \leqslant_{\mathrm{ss}} F^{[p,1]} \leqslant_{\mathrm{st}} G^{[p,1]}$ gives the desired result. □

4.5 Risk measure and probability bounds

4.5.1 Bounds on tail risk measures

Evaluating the "tail risk ", or the behavior of a risk beyond a high level, has become crucial in the regulatory frameworks for banking and insurance. To better understand the

tail risk, Liu and Wang (2021) provided an axiomatic framework of risk measures which can quantify the tail risk. Those risk measures are referred to as tail risk measures. This section is dedicated to studying the worst-case value of tail risk measures with the order constraint.

For $p \in (0,1)$, a risk measure ρ is a *p-tail risk measure* if $\rho(F) = \rho(G)$ for all $F, G \in \mathcal{M}$ such that $F^{[p,1]} = G^{[p,1]}$. In other words, the value of a p-tail risk measure of random variable X is determined by its distribution beyond $F^{-1}(p)$. The class of tail risk measures includes the most important regulatory risk measures VaR and ES, and those popular in the literature, such as RVaR and Gini Shortfall (Furman et al. (2017)).

For a p-tail risk measure ρ, there always exists another risk measure ρ^*, called the generator, such that $\rho(F) = \rho^*\left(F^{[p,1]}\right)$ where $F \in \mathcal{M}$ and $F^{[p,1]}$ is the upper p-tail distribution of F. We call (ρ, ρ^*) a *p-tail pair of risk measures*. The class of $\leqslant_{\mathrm{cv}}$-consistent generators ρ^* includes, for instance,

(i) $\rho^* = \text{ess-inf}$, corresponding to $\rho = \mathrm{VaR}_p^R$;

(ii) $\rho^* = \mathbb{E}$, corresponding to $\rho = \mathrm{ES}_p$;

(iii) $\rho^* : X \mapsto -\mathrm{ES}_t(-X)$, corresponding $\rho = \mathrm{RVaR}_{p,q}$, where $t = (1-q)/(1-p)$ (see Example 5 of Liu and Wang (2021)).

Introduced by Wang and Zitikis (2021), a *p-tail event* of a random variable X is an event $A \in \mathcal{A}$ with $\mathbb{P}(A) = 1 - p \in (0,1)$ such that $X(\omega) \geqslant X(\omega')$ holds for all $\omega \in A$ and $\omega' \in A^c$. It is easy to check that, for $X \sim F$, the upper p-tail distribution of F is the same as the conditional distribution of X on the p-tail event A, i.e., $F^{[p,1]} = H_{X|A}$. Therefore, we can write the p-tail risk measure $\rho(X) = \rho^*(X_A)$ where $X_A \sim H_{X|A}$ and A is a p-tail event of X. Similarly, for risk aggregation $S = X + Y$, we can write the p-tail risk measure $\rho(S) = \rho^*(X_B + Y_B)$ where $X_B \sim H_{X|B}$, $Y_B \sim H_{Y|B}$ and B is a p-tail event of S, but not necessarily a p-tail event of either X or Y.

To investigate the worst-case value of tail risk measures, we use the notion of p-concentration, characterized by Wang and Zitikis (2021). A random vector (X, Y) is *p-concentrated* if X and Y share a common p-tail event of probability $1-p$. Intuitively, for p close to 1, p-concentrated risks will cause simultaneous large losses if the corresponding p-tail event happens.

There is an important connection between p-concentration and the worst-case risk aggregation of a p-tail risk measure. In the unconstrained setting (i.e., without the order

constraint), if ρ is a monotone p-tail risk measure, the worst-case value of ρ can be attained by p-concentrated risks (Theorem 3 of Liu and Wang (2021)). Earlier results of this type for VaR are Theorem 4.6 of Bernard et al. (2014) and Theorem 4 of Bernard et al. (2017). Therefore, in the unconstrained setting, it suffices to look at the tail risk of each marginal distribution when we calculate the worst-case value of ρ. Moreover, if the generator ρ^* of ρ is $\leqslant_{\mathrm{cv}}$-consistent, by (4.2), the worst-case value of ρ is attained when the upper tail risks are countermonotonic.

The following theorem studies the worst-case value of tail risk measures with the order constraint. We show that, if (ρ, ρ^*) is a monotone p-tail pair of risk measures and ρ^* is $\leqslant_{\mathrm{cv}}$-consistent, the worst-case value of ρ with the order constraint can also be attained by p-concentrated risks, and it is attained when the upper tail risks are DL-coupled. This result can be seen as parallel to Liu and Wang (2021, Theorem 3), which does not have the order constraint. However, the proof is quite different, and the strong stochastic order in Section 4.4 through Corollary 4.2 is crucial for this result.

Theorem 4.3. *Suppose that $F \leqslant_{\mathrm{st}} G$, $p \in (0,1)$, (ρ, ρ^*) is a p-tail pair of risk measure, and ρ^* is monotone and $\leqslant_{\mathrm{cv}}$-consistent. We have*

$$\overline{\rho}(\mathcal{F}_2^o(F,G)) = \overline{\rho^*}\left(\mathcal{F}_2^o\left(F^{[p,1]}, G^{[p,1]}\right)\right) = \rho^*(X+Y), \tag{4.7}$$

where $(X,Y) \sim D_*^{F^{[p,1]},G^{[p,1]}}$.

Proof. First, for any $X \sim F^{[p,1]}$ and $Y \sim G^{[p,1]}$, we can always construct $Z \sim F$ and $W \sim G$ such that conditional on a p-tail event of $Z+W$, $Z+W$ has the same law as $X+Y$. This structure can be obtained by using a copula satisfying p-concentration. Hence, we have the "$\geqslant$" direction of the following equality

$$\overline{\rho}(\mathcal{F}_2^o(F,G)) = \overline{\rho^*}(\mathcal{F}_2^o(F^{[p,1]}, G^{[p,1]})). \tag{4.8}$$

Below, we will show the "$\leqslant$" direction of (4.8). We break the proof into several steps.

1. For any $X \sim F$ and $Y \sim G$ such that $X \leqslant Y$ almost surely, let A be a p-tail event of $X+Y$ in the sense of Wang and Zitikis (2021). Hence, $\rho(X+Y) = \rho^*(X_A + Y_A)$ for some $X_A \sim H_{X|A}$ and $Y_A \sim H_{Y|A}$. Note that here we only need to specify the distribution of (X_A, Y_A), which is the conditional distribution of (X,Y) on A.

2. By Propositions 4.1 and 4.2, we have $H_{Y|A} \leqslant_{\mathrm{st}} G^{[p,1]}$. Take $Y' \sim G^{[p,1]}$ satisfying $Y' \geqslant Y_A$. The existence of Y' is guaranteed by, e.g., Theorem 1.A.1 of Shaked and Shanthikumar (2007). By monotonicity of ρ^*, we have $\rho^*(X_A + Y_A) \leqslant \rho^*(X_A + Y')$.

3. Take $\tilde{X}_A \sim H_{X|A}$ and $\tilde{Y} \sim G^{[p,1]}$ such that $(\tilde{X}_A, \tilde{Y})$ is DL-coupled. By $\leqslant_{\mathrm{cv}}$-consistency of ρ^* and Lemma 4.1, we have $\rho^*(X_A + Y') \leqslant \rho^*(\tilde{X}_A + \tilde{Y})$.

4. Using Corollary 4.2, there exists $\tilde{X} \sim F^{[p,1]}$ such that $\tilde{X}_A \leqslant \tilde{X} \leqslant \tilde{Y}$ almost surely. By monotonicity of ρ^*, we have $\rho^*(\tilde{X}_A + \tilde{Y}) \leqslant \rho^*(\tilde{X} + \tilde{Y})$.

We established the chain of inequalities

$$\rho(X+Y) = \rho^*(X_A + Y_A) \leqslant \rho^*(X_A + Y') \leqslant \rho^*(\tilde{X}_A + \tilde{Y}) \leqslant \rho^*(\tilde{X} + \tilde{Y}).$$

where $\tilde{X} \sim F^{[p,1]}$ and $\tilde{Y} \sim G^{[p,1]}$. Therefore, we obtained the "$\leqslant$" direction of the equality in (4.8). The last equality in (4.7) is directly obtained from Lemma 4.1. □

Remark 4.2. For F, $G \in \mathcal{M}$, we look at cases where ρ is one of VaR^R, ES and RVaR.

(i) For $p \in (0,1)$, we have $\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) = \text{ess-inf}(X+Y)$ where $(X,Y) \sim D_*^{F^{[p,1]},G^{[p,1]}}$.

(ii) For $p \in (0,1)$, we have $\overline{\mathrm{ES}}_p(\mathcal{F}_2^o(F,G)) = \mathbb{E}\left[F^{[p,1]}\right] + \mathbb{E}\left[G^{[p,1]}\right] = \mathrm{ES}_p(F) + \mathrm{ES}_p(G)$, which can also be obtained from comonotonic-additivity and subadditivity of ES. Hence the order constraint does not improve the worst-case value of ES. Indeed, the worst-case value of ES in unconstrained case is attained if and only if the two risks are p-concentrated (Theorem 5 of Wang and Zitikis (2021)).

(iii) For $0 \leqslant p < q < 1$, we have $\overline{\mathrm{RVaR}}_{p,q}(\mathcal{F}_2^o(F,G)) = -\mathrm{ES}_t(-X-Y)$, where $(X,Y) \sim D_*^{F^{[p,1]},G^{[p,1]}}$ and $t = (1-q)/(1-p)$.

Similarly, we can derive the best-case value of risk measures. For instance,

$$\underline{\mathrm{RVaR}}_{p,q}(\mathcal{F}_2^o(F,G)) = -\overline{\mathrm{RVaR}}_{1-q,1-p}(\mathcal{F}_2^o(\hat{G},\hat{F}))) = \mathrm{ES}_{p/q}(X+Y),$$

where $\hat{G}$ and $\hat{F}$ are the distributions of $-Y$ and $-X$, respectively, and $(X,Y) \sim D_*^{F^{[0,q]},G^{[0,q]}}$. In Section 4.5.2, we derive analytical results for the best-case and worst-case values of VaR.

In the following example, we calculate the worst-case value of RVaR for two uniformly distributed risks.

Example 4.4. For fixed $p \in (0,1)$ and distributions F, G such that $F \leqslant_{\mathrm{st}} G$, suppose that the upper p-tail distributions are two uniform distributions $F^{[p,1]}(x) = x$ for $x \in [0,1]$ and $G^{[p,1]}(y) = y/b$ for $y \in [0,b]$. It is easy to check that $F^{[p,1]} \leqslant_{\mathrm{st}} G^{[p,1]}$ if and only if $b \geqslant 1$.

We assume that $1 < b < 2$. Let $(X, Y) \sim D_*^{F^{[p,1]}, G^{[p,1]}}$. The directional optimal transport between singular parts of $F^{[p,1]}$ and $G^{[p,1]}$ is

$$T^{F^{[p,1]}, G^{[p,1]}}(x) = \inf\left\{z \geqslant x : 1 - \frac{z}{b} < x - \frac{x}{b}\right\} = b - (b-1)x, \quad x \in [0,1].$$

Then for $c \in [0, b)$ we have $\mathbb{P}(X + Y \leqslant c) = \left(\mu_{F^{[p,1]}} \wedge \mu_{G^{[p,1]}}\right)([\infty, c/2]) = c/2b$. For $c \in [b, 2]$,

$$\mathbb{P}(X+Y \leqslant c) = (\mu_{F^{[p,1]}} \wedge \mu_{G^{[p,1]}})([\infty, c/2]) + \mu'_{F^{[p,1]}}([0, (c-b)/(2-b)]) = \frac{c}{2(2-b)} - \frac{b-1}{2-b}.$$

Therefore, $\mathrm{VaR}_\alpha^R(X+Y) = 2b\alpha$ for $\alpha \in (0, 1/2]$ and $\mathrm{VaR}_\alpha^R(X+Y) = (4-2b)\alpha + 2b - 2$ for $\alpha \in [1/2, 1]$. By Theroem 4.3, we derive the worst-case value of RVaR

$$\overline{\mathrm{RVaR}}_{p,q}(\mathcal{F}_2^o(F,G)) = \begin{cases} ba, & q \in \left(p, \frac{1+p}{2}\right]; \\ \frac{b}{4a} - \frac{1}{4a}(2a-1)(2ba-3b-4a+2), & q \in \left(\frac{1+p}{2}, 1\right), \end{cases}$$

where $a = 1 - (1-q)/(1-p)$.

4.5.2 VaR bounds

The popular risk measure VaR is the most important example of a non-convex risk measure, and it is neither $\leqslant_{\mathrm{cx}}$- nor $\leqslant_{\mathrm{cv}}$-consistent. In this section, we derive analytical solutions for VaR bounds with the order constraint if marginal distributions are continuous. For non-continuous marginal distributions, an algorithm is available in Section 4.6 to approximate the bounds.

Proposition 4.3. *For continuous distributions F and G such that $F \leqslant_{\mathrm{st}} G$ and $p \in (0,1)$, we have*

$$\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) = \min\left\{\inf_{x \in [F^{-1}(p+), G^{-1}(p+)]}\left\{T^{F^{[p,1]}, G^{[p,1]}}(x) + x\right\}, 2G^{-1}(p+)\right\},$$

and

$$\underline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G)) = \max\left\{\sup_{x \in [F^{-1}(p), G^{-1}(p)]}\left\{\hat{T}^{F^{[0,p]}, G^{[0,p]}}(x) + x\right\}, 2F^{-1}(p)\right\},$$

where $\hat{T}^{F^{[0,p]}, G^{[0,p]}}(x) = \sup\left\{t \leqslant x : F^{[0,p]}(t) - G^{[0,p]}(t) < F^{[0,p]}(x) - G^{[0,p]}(x)\right\}$.

Proof. For $p \in (0,1)$, as $\left(\mathrm{VaR}_p^R, \text{ess-inf}\right)$ is a p-tail pair of risk measures and ess-inf is $\leqslant_{\mathrm{cv}}$-consistent, by Theorem 4.3,

$$\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) = \overline{\text{ess-inf}}(\mathcal{F}_2^o(F^{[p,1]}, G^{[p,1]})) = \text{ess-inf}(X+Y),$$

where $(X,Y) \sim D_*^{F^{[p,1]},G^{[p,1]}}$. By Theorem 4.1, we obtain the first result. For the second result, let $X' \sim F$ and $Y' \sim G$. Denote by $\hat{F}$ and $\hat{G}$ the distributions of $-X'$ and $-Y'$. We have

$$\underline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G)) = -\overline{\mathrm{VaR}}_{1-p}^R(\mathcal{F}_2^o(\hat{G},\hat{F})) = \underline{\text{ess-sup}}\left(\mathcal{F}_2^o\left(F^{[0,p]}, G^{[0,p]}\right)\right).$$

Applying Theorem 4.1, we get the desired result. □

Remark 4.3. For $F, G \in \mathcal{M}$ and $p \in (0,1)$, the worst-case value of VaR_p^R and the best-case value of VaR_p^L without the order constraint are attained by letting the upper tail risks and lower tail risks be countermonotonic, respectively, i.e.,

$$\sup\left\{\mathrm{VaR}_p^R(X+Y) : X \sim F,\ Y \sim G\right\} = \inf_{x\in[0,1-p]}\left\{F^{-1}(p+x) + G^{-1}(1-x)\right\},$$

$$\inf\left\{\mathrm{VaR}_p^L(X+Y) : X \sim F,\ Y \sim G\right\} = \sup_{x\in[0,p]}\left\{F^{-1}(x) + G^{-1}(p-x)\right\}.$$

See Makarov (1981) and Rüschendorf (1982).

Example 4.5 (Pareto risks: VaR bounds)**.** Following the marginal assumptions on F and G in Example 4.1, we derive $\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G))$ and $\underline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G))$ by Proposition 4.3. For $p \in (0,1)$, we have

$$F^{[p,1]}(x) = \left(1 - \frac{1}{x(1-p)}\right)\mathbb{1}_{\{x \geqslant 1/(1-p)\}} \quad \text{and} \quad G^{[p,1]}(x) = \left(1 - \frac{2}{x(1-p)}\right)\mathbb{1}_{\{x \geqslant 2/(1-p)\}}.$$

Thus,

$$T^{F^{[p,1]},G^{[p,1]}}(x) = \frac{x}{x(1-p)-1}, \qquad x \in \left(\frac{1}{1-p}, \frac{2}{1-p}\right].$$

And $T^{F^{[p,1]},G^{[p,1]}}(1/(1-p)) = \inf\{\emptyset\} = \infty$. Therefore,

$$\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) = \min\left\{\inf_{x\in[1/(1-p),2/(1-p)]}\left\{\frac{x}{x(1-p)-1} + x\right\}, \frac{4}{(1-p)}\right\} = \frac{4}{1-p}.$$

Similarily, we have $\underline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G)) = 1 + 2/(1-p)$. Those bounds on VaR will be used to calculate probability bounds of $X+Y$ in Example 4.6, where $X \sim F$ and $Y \sim G$.

In the unconstrained problem, Bernard et al. (2014) showed that the worst-case value of VaR_p^R is a continuous function of $p \in (0,1)$ if the marginal distributions are strictly increasing. This continuity result is used to confirm that there is no need to distinguish between VaR^R and VaR^L when we calculate their worst-case values (best-case values). We will see later that the above statement is still true if the order constraint is further imposed. The continuity of the worst-case value of VaR^R with the order constraint is established in Lemma 4.2. The proof of Lemma 4.2 is surprisingly complicated, very different from the case treated by Bernard et al. (2014), and it is put in Section 4.8.

Lemma 4.2. *For strictly increasing continuous distribution functions F and G such that $F \leqslant_{\mathrm{st}} G$, the function $p \mapsto \overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G))$ is continuous on $(0,1)$.*

Using Lemma 4.2, we obtain that the worst-case values (best-case values) of VaR^R and VaR^L with the order constraint are equivalent for strictly increasing continuous distributions.

Proposition 4.4. *Suppose that F and G are strictly increasing continuous distribution functions such that $F \leqslant_{\mathrm{st}} G$. For $p \in (0,1)$, we have*

$$\overline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G)) = \overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) \quad \text{and} \quad \underline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G)) = \underline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)).$$

Proof. For $\varepsilon > 0$, we have $\overline{\mathrm{VaR}}_{p-\varepsilon}^R(\mathcal{F}_2^o(F,G)) \leqslant \overline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G)) \leqslant \overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G))$. By Lemma 4.2, $\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G))$ is a continuous function of $p \in (0,1)$. Letting $\varepsilon \downarrow 0$, we get the desired result for worst-case value of VaR^L and VaR^R. The proof for the best-case value of VaR^L and VaR^R is similar and thus omitted. □

By Proposition 4.4, in practical situations of risk management, there is no need to distinguish between VaR_p^L and VaR_p^R when we calculate their bounds with the order constraint; this observation will be useful in the numerical studies in Section 4.6.

4.5.3 Probability bounds

In risk management and quantitative finance, probability bounds of the aggregate position are also of great interest. In the unconstrained problem (i.e., only marginal distributions are known), the probability bounds on the aggregation of two risks are given by Rüschendorf (1982).

For $F, G \in \mathcal{M}$ such that $F \leqslant_{\mathrm{st}} G$ and $t \in \mathbb{R}$, we are interested in the upper and lower bounds of probability with the order constraint, defined as

$$M^o(t) := \sup\left\{\mathbb{P}(X+Y \leqslant t) : (X,Y) \in \mathcal{F}_2^o(F,G)\right\}$$

and

$$m^o(t) := \inf\left\{\mathbb{P}(X+Y < t) : (X,Y) \in \mathcal{F}_2^o(F,G)\right\}.$$

The above upper and lower bounds of probability can be obtained by inverting the lower and upper bounds of VaR, respectively. In particular, for $p \in (0,1)$, we have

$$\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) = (m^o)^{-1}(p) \quad \text{and} \quad \underline{\mathrm{VaR}}_p^L(\mathcal{F}_2^o(F,G)) = (M^o)^{-1}(p).$$

For continuous marginal distributions, we can invert the analytical solutions in Proposition 4.3 to obtain the probability bounds with the order constraint. While the analytical solutions to VaR bounds does not necessarily lead to an explicit results for probability bounds, a numerical algorithm in Section 4.6 can be used to approximate probability bounds. The following example compares probabilities bounds with and without order constraint for Pareto marginal distributions.

Example 4.6 (Pareto: Probability bounds)**.** Following the assumptions in Examples 4.1 and 4.5, we convert the VaR bounds in Example 4.5 to obtain probability bounds with the order constraint:

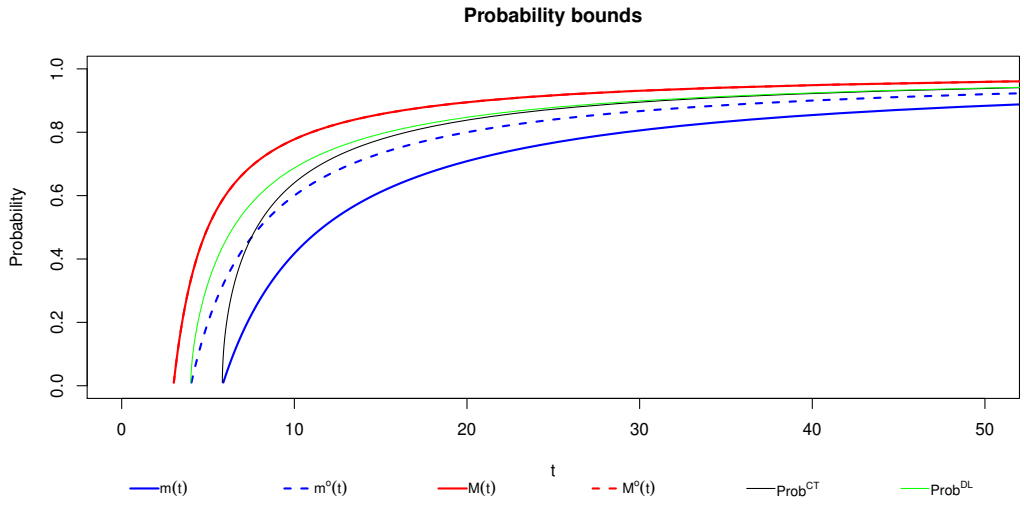
$$M^o(t) = 1 - \frac{4}{t}, \quad t \geqslant 4, \quad \text{and} \quad m^o(t) = 1 - \frac{2}{t-1}, \quad t \geqslant 3.$$

The probability bounds with and without order constraint are plotted in Figure 4.3. The bounds without the order constraint are denoted by M and m. The figure shows that the order constraint improves the lower probability bound a lot while there is no improvement for the upper bound (the difference between M^o and M is invisible). When two risks are countermonotonic or DL-coupled, the corresponding probability (denoted by $\mathrm{Prob}^{\mathrm{CT}}$ and $\mathrm{Prob}^{\mathrm{DL}}$, respectively) lies between the constrained bounds for $t \geqslant 8$.

4.6 Numerical results and a real-data application

In this section, we use numerical examples and a case study to illustrate the impact of the order constraint on VaR bounds (the worst-case and best-case values of VaR^R and VaR^L), and RVaR bounds. For convenience, we do not distinguish between VaR^L and VaR^R when we calculate their bounds (Proposition 4.4). Both VaR^L and VaR^R are referred to as VaR in numerical results. We only illustrate the numerical calculations for VaR bounds. RVaR bounds can be calculated in a similar manner.

Figure 4.3: Probability bounds in Example 4.6



4.6.1 General methodology

Let $F, G \in \mathcal{M}$ be continuous distributions such that $F \leqslant_{\mathrm{st}} G$. As suggested by Theorem 4.3, the best-case and worst-case values of VaR are determined by the lower tail and upper tail distributions of F and G, respectively. To approximately calculate $\overline{\mathrm{VaR}}_p^R$ for $p \in (0,1)$, we first discretize the upper p-tail distributions $F^{[p,1]}$ and $G^{[p,1]}$. Fix an integer n and let

$$x_i = F^{-1}\left(p + \frac{(1-p)(n-i)}{n}\right) \text{ and } y_i = G^{-1}\left(p + \frac{(1-p)(n-i)}{n}\right),$$

for $i = 1, \dots, n$. Define $S_X^{[p,1]} = \{x_1, \dots, x_n\}$ and $S_Y^{[p,1]} = \{y_1, \dots, y_n\}$. If $S_X^{[p,1]}$ and $S_Y^{[p,1]}$ have no identical locations, we use the following algorithm introduced in Nutz and Wang (2021) to approximate the DL coupling between $F^{[p,1]}$ and $G^{[p,1]}$. Let $S_1 = S_Y^{[p,1]}$, we iterate for $k = 1, \dots, n$

(i) $T(x_k) := \min\{y \in S_k : y \geqslant x_k\}$,

(ii) $S_{k+1} = S_k \setminus \{T(x_k)\}$.

Let $s_k = x_k + T(x_k)$, $k = 1, \dots, n$. We use $\min\{s_k : k = 1, \dots, n\}$ as the approximation for $\overline{\mathrm{VaR}}_p^R$. The best-case value of VaR^L can be obtained in a similar manner. The unconstrained bounds of VaR, attained by conditional countermonotonicity, can be numerically

computed by the Rearrangement Algorithm (RA) in Puccetti and Rüschendorf (2012) and Embrechts et al. (2013). Similar procedures can be constructed for discrete distributions.

The difference between the worst-case and best-case values of a risk measure is called the *Dependence Uncertainty spread* (DU-spread) for the risk measure, which is used as a measure of dependence uncertainty (see Embrechts et al. (2015)). We use the DU-spread reduction defined in Puccetti et al. (2017) to measure the improvement on unconstrained VaR bounds due to the order constraint. Denote by L and U the unconstrained best-case and worst-case values of a risk measure ρ. Similarly, denote by L^o and U^o the bounds with the order constraint. The lower and upper reductions of DU-spread are defined as

$$R^L = \frac{L^o - L}{U - L} \quad \text{and} \quad R^U = \frac{U - U^o}{U - L}. \tag{4.9}$$

The DU-spread reduction is defined as the sum of lower and upper DU-spread reductions, which is $R = R^L + R^U \in [0, 1]$.

4.6.2 Numerical examples

Consider distributions F and G_i such that their means are 50 and $50+10i$ and $F \leqslant_{\mathrm{st}} G_i$, $i = 1, 2, 3$. The distributions are specified in uniform and Pareto cases as below.

Table 4.1: Distributions for numerical examples

Uniform	$F(x) = x/100$	$G_1(x) = x/120$	$G_2(x) = x/140$	$G_3(x) = x/160$
Pareto	$F(x) = 1 - (25/x)^2$	$G_1(x) = 1 - (30/x)^2$	$G_2(x) = 1 - (35/x)^2$	$G_3(x) = 1 - (40/x)^2$

For both F and G_i, $i = 1, 2, 3$, being uniform or Pareto distributions, we calculate the improvement (i.e., reduction of DU-spread) on VaR bounds and RVaR bounds. We also present the results of $\mathrm{VaR}_p(X + Y)$ if $X \sim F$ and $Y \sim G_i$ are independent, comonotonic, DL-coupled and countermonotonic, $i = 1, 2, 3$. The results of VaR bounds for uniform and Pareto cases can be found in Figures 4.4 and 4.5, respectively. The results of RVaR bounds can be found in Tables 4.2 and 4.3. We make the following observations.

(i) The DU-spread reductions in all tables and figures show that the improvement due to the order constraint is significant for both VaR and RVaR. The improvement for VaR becomes larger as p increases from 0.9 to 1.

(ii) For all uniform and Pareto cases, as the mean of G_i becomes larger, the improvement becomes smaller. In other words, the more "similar" the distributions F and G_i are, the more improvement is gained from imposing the order constraint.

(iii) The order constraint has an overall larger improvement on the bounds for uniform distributions than those for Pareto distributions. Nevertheless, the improvement on the worst-case value is insignificant for uniform distributions. This is because $\text{ess-inf} G_i^{[p,1]} \geqslant \text{ess-sup} F^{[p,1]}$ for $p \in (0.9, 1)$, and the DL coupling of the upper p-tail distributions is the same as countermonotonicity. While for Pareto distributions, the improvement on the worst-case value is even larger than that on the best-case value.

(iv) For the uniform cases, if the risks are countermonotonic, both VaR and RVaR are close to the unconstrained lower bound. If the risks are DL-coupled, both VaR and RVaR are close to the constrained lower bound for the uniform cases while they lie between the constrained bounds for the Pareto cases. If the risks are comonotonic, both VaR and RVaR lie between the constrained bounds for all cases.

As the observations on VaR and RVaR are similar, we will focus on studying VaR for the rest of the chapter. In previous examples for Pareto distributions, the tail parameter[2] of distributions F and G are fixed (see Table 4.1). Next, we study the improvement of VaR bounds as the tail parameter of the distribution G varies. Let $F(x) = 1 - (25/x)^2$ and $G(x) = 1 - (25/x)^\alpha$, $\alpha \leqslant 2$. VaR bounds are calculated as α increases from 1.3 to 2. Results can be found in Figure 4.6. We observe that the larger α is, the greater the improvement is gained from the order constraint. However, the improvement on the unconstrained lower bound is neligible for small α. As in previous examples for Pareto distributions, comonotonicity and DL coupling produce very close VaR values.

4.6.3 Case study: Health insurance policies

In this case study, we calculate the bounds of VaR with and without order constraint for a health insurance portfolio. Insurance policies can be classified according to certain characteristics of policyholders. For illustration purposes, we use gender to make classifications on health insurance policies (this may not be allowed in certain countries). The aggregate loss of the portfolio can be expressed as $S = X + Y$ where $X \sim F$ and $Y \sim G$ represent the losses caused by females and males, respectively, from a portfolio of 50 males

[2]We use the Pareto(θ, α) distribution parametrized by $F(x) = 1 - (\theta/x)^\alpha$ for $x \geqslant \theta$, where $\theta \in \mathbb{R}$ is the location parameter and $\alpha > 0$ is the tail parameter.

Figure 4.4: Uniform cases: VaR_p bounds, DU reduction and VaR_p of the aggregate risk with different dependence structures are contained in this figure. VaR values with independence, comonotonicity, countermonotonicity and DL coupling are denoted by $\mathrm{VaR}^{\mathrm{Ind}}$, $\mathrm{VaR}^{\mathrm{C}}$, $\mathrm{VaR}^{\mathrm{Co}}$ and $\mathrm{VaR}^{\mathrm{DL}}$, respectively.

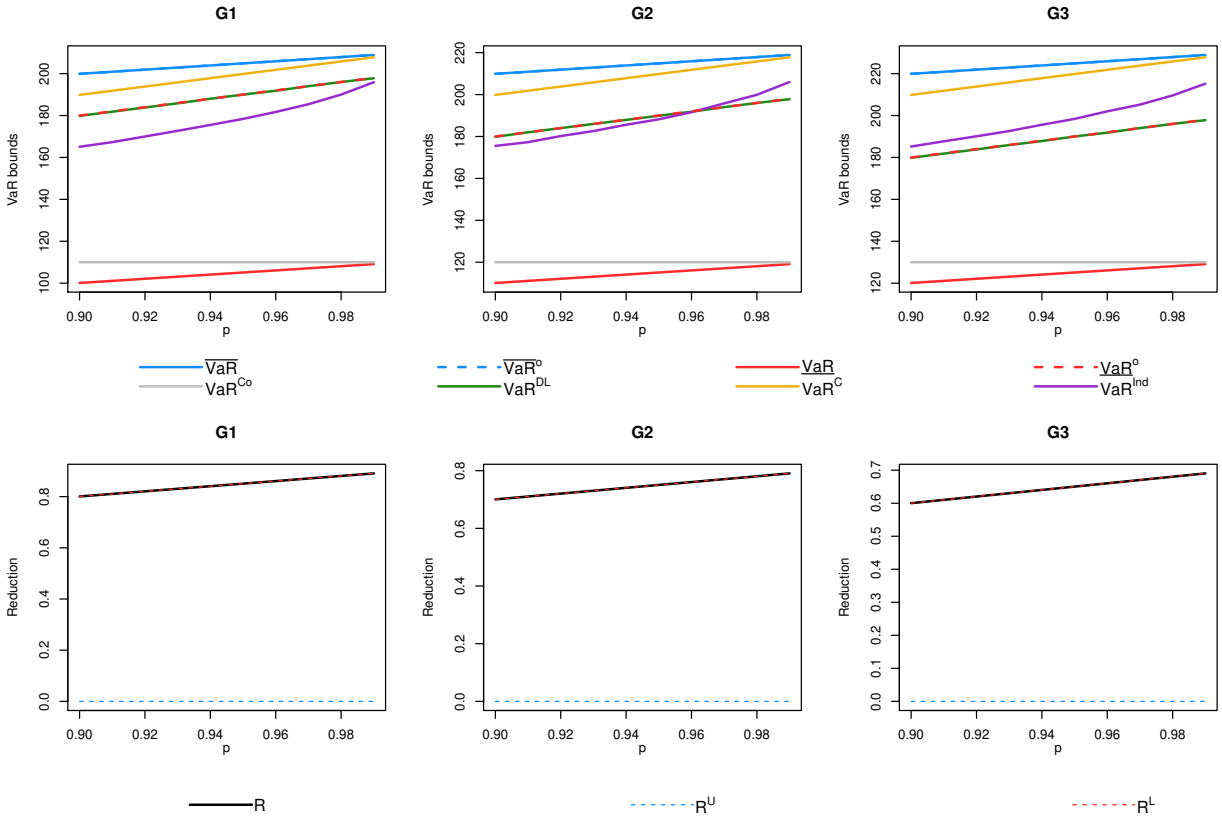


Figure 4.5: Pareto cases: VaR_p bounds, DU reduction and VaR_p of the aggregate risk with different dependence structures are contained in this figure. VaR values with independence, comonotonicity, countermonotonicity and DL coupling are denoted by $\mathrm{VaR}^{\mathrm{Ind}}$, $\mathrm{VaR}^{\mathrm{C}}$, $\mathrm{VaR}^{\mathrm{Co}}$ and $\mathrm{VaR}^{\mathrm{DL}}$, respectively.

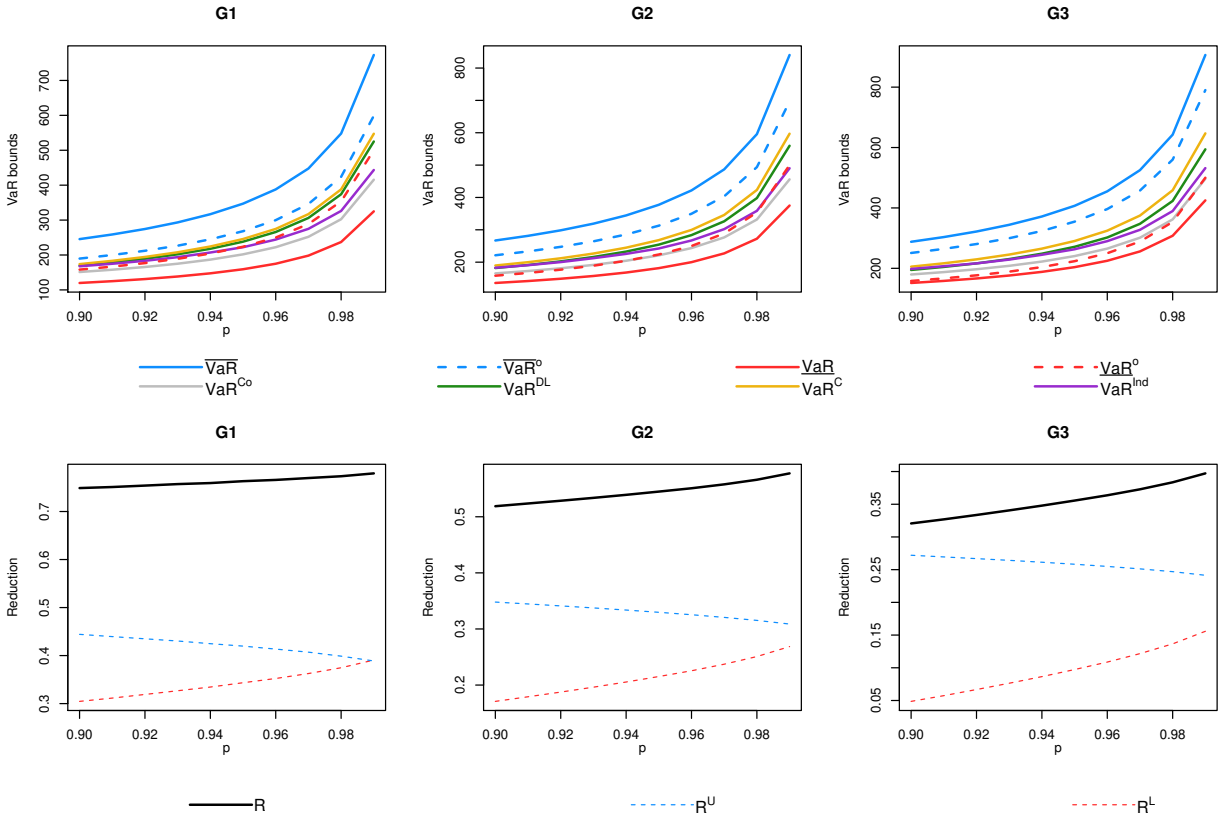


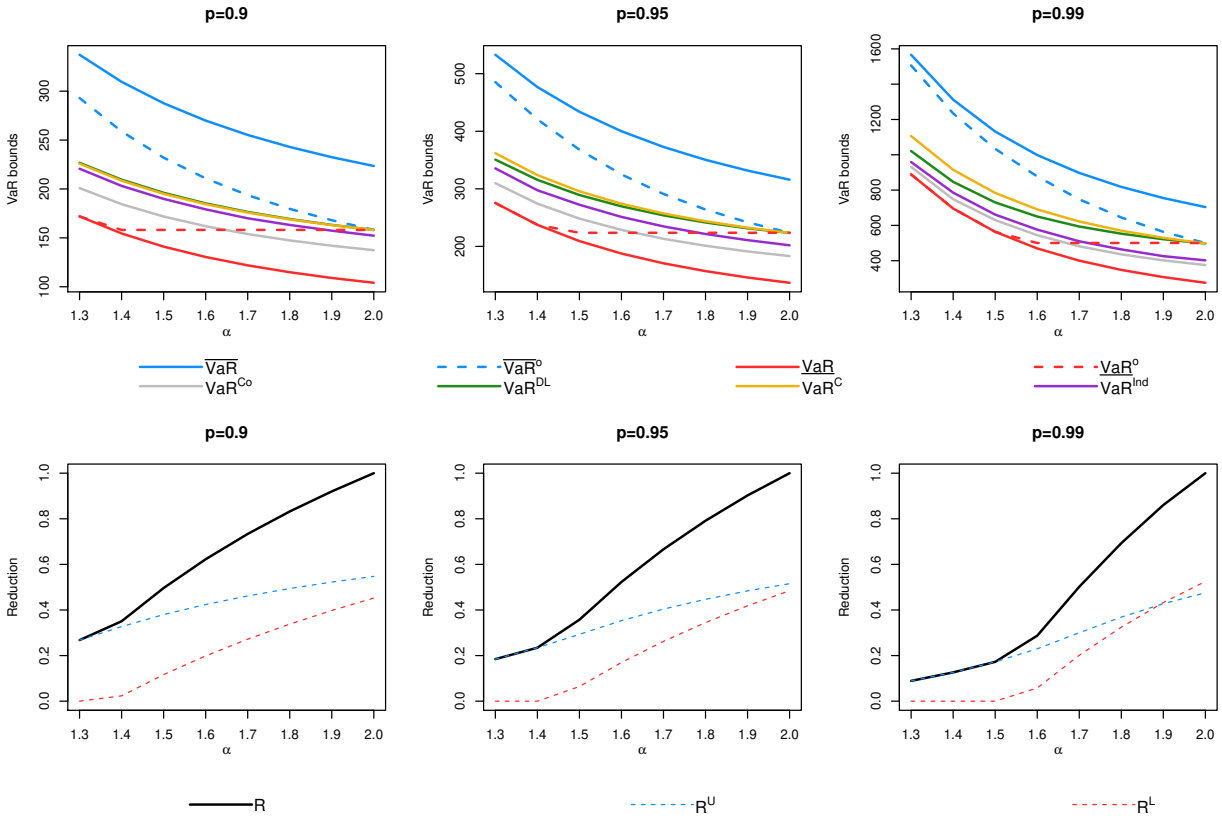
Table 4.2: Uniform cases: $\text{RVaR}_{p,q}$ bounds, DU reduction and $\text{RVaR}_{p,q}$ of the aggregate risk with different dependence structures are contained in this table. Marginal distributions in Case i are uniform distributions F and G_i given in Table 4.1.

	p	75%	90%	95%	99.5%
	q	90%	95%	99.5%	99.9%
Case 1	Constrained bounds	$(165, 182)$	$(185, 200)$	$(195, 205)$	$(200, 209)$
	Unconstrained bounds	$(100, 185)$	$(105, 200)$	$(110, 205)$	$(110, 209)$
	(R^L, R^U, R)	$(0.77, 0.04, 0.81)$	$(0.84, 0, 0.84)$	$(0.89, 0, 0.89)$	$(0.9, 0, 0.9)$
	Independence	151	171	187	203
	Comonotonicity	175	195	204	209
	Countermonotonicity	110	110	110	110
	DL coupling	165	185	194	199
Case 2	Constrained bounds	$(165, 195)$	$(185, 210)$	$(195, 215)$	$(200, 219)$
	Unconstrained bounds	$(110, 195)$	$(115, 210)$	$(120, 215)$	$(120, 219)$
	(R^L, R^U, R)	$(0.65, 0, 0.65)$	$(0.74, 0, 0.74)$	$(0.79, 0, 0.79)$	$(0.8, 0, 0.8)$
	Independence	161	181	197	212
	Comonotonicity	185	205	214	219
	Countermonotonicity	120	120	120	120
	DL coupling	165	185	195	199
Case 3	Constrained bounds	$(165, 205)$	$(185, 220)$	$(195, 225)$	$(200, 229)$
	Unconstrained bounds	$(120, 205)$	$(125, 220)$	$(130, 225)$	$(130, 229)$
	(R^L, R^U, R)	$(0.53, 0, 0.53)$	$(0.63, 0, 0.63)$	$(0.68, 0, 0.68)$	$(0.7, 0, 0.7)$
	Independence	171	192	207	222
	Comonotonicity	195	215	224	229
	Countermonotonicity	130	130	130	130
	DL coupling	165	185	195	199

Table 4.3: Pareto cases: $\mathrm{RVaR}_{p,q}$ bounds, DU reduction and $\mathrm{RVaR}_{p,q}$ of the aggregate risk with different dependence structures are contained in this table. Marginal distributions in Case i are Pareto distributions F and G_i given in Table 4.1.

	p	75%	90%	95%	99.5%
	q	90%	95%	99.5%	99.9%
Case 1	Constrained bounds	$(125, 140)$	$(185, 213)$	$(354, 379)$	$(1012, 1085)$
	Unconstrained bounds	$(103, 164)$	$(140, 254)$	$(262, 409)$	$(679, 1209)$
	(R^L, R^U, R)	$(0.35, 0.39, 0.75)$	$(0.4, 0.36, 0.76)$	$(0.63, 0.21, 0.83)$	$(0.63, 0.23, 0.86)$
	Independence	136	191	316	800
	Comonotonicity	135	204	373	1063
	Countermonotonicity	124	172	292	774
	DL coupling	132	198	360	1017
Case 2	Constrained bounds	$(129, 157)$	$(188, 240)$	$(371, 418)$	$(1042, 1204)$
	Unconstrained bounds	$(114, 178)$	$(156, 276)$	$(289, 446)$	$(755, 1316)$
	(R^L, R^U, R)	$(0.24, 0.33, 0.57)$	$(0.27, 0.3, 0.57)$	$(0.53, 0.18, 0.7)$	$(0.51, 0.2, 0.71)$
	Independence	148	208	345	885
	Comonotonicity	147	222	407	1160
	Countermonotonicity	135	188	320	851
	DL coupling	143	212	383	1075
Case 3	Constrained bounds	$(135, 173)$	$(194, 265)$	$(391, 456)$	$(1083, 1321)$
	Unconstrained bounds	$(125, 192)$	$(174, 298)$	$(317, 482)$	$(838, 1421)$
	(R^L, R^U, R)	$(0.15, 0.29, 0.44)$	$(0.17, 0.27, 0.43)$	$(0.45, 0.16, 0.6)$	$(0.42, 0.17, 0.59)$
	Independence	161	225	375	972
	Comonotonicity	159	241	441	1256
	Countermonotonicity	147	205	349	932
	DL coupling	153	226	407	1144

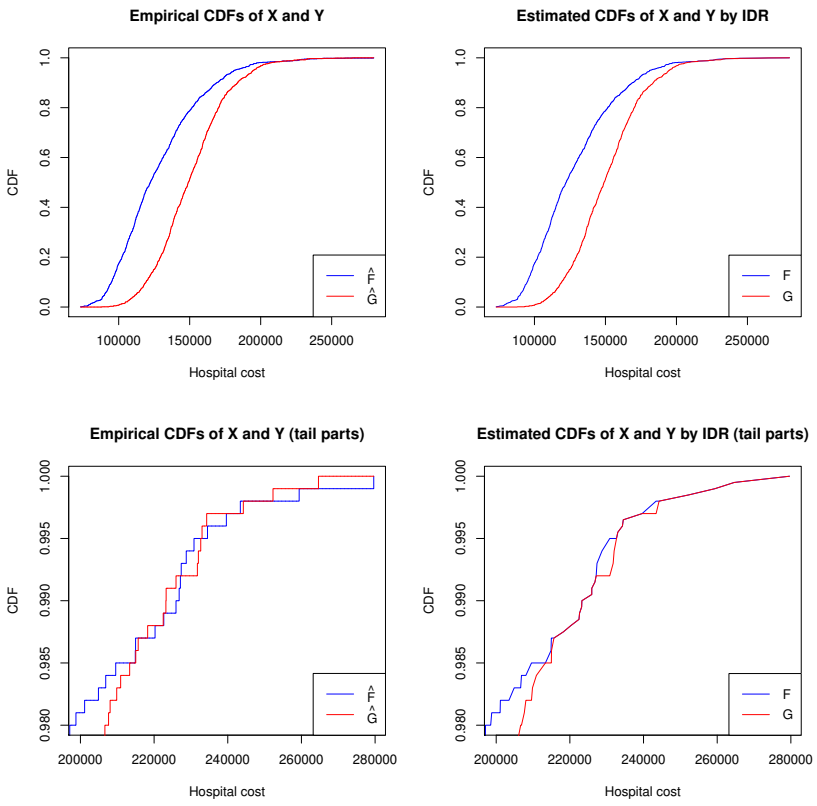
Figure 4.6: This figure contains VaR_p bounds, DU reduction and VaR_p of the aggregated Pareto risks with different dependence structures as α changes, where $F = \mathrm{Pareto}(25, 2)$ and $G = \mathrm{Pareto}(25, \alpha)$. VaR values with independence, comonotonicity, countermonotonicity and DL coupling are denoted by $\mathrm{VaR}^{\mathrm{Ind}}$, $\mathrm{VaR}^{\mathrm{C}}$, $\mathrm{VaR}^{\mathrm{Co}}$ and $\mathrm{VaR}^{\mathrm{DL}}$, respectively.



and 50 females. It is sensible to guess that $F \leqslant_{\mathrm{st}} G$, due to the morbidity differences between males and females; this will be confirmed by our dataset. Moreover, since the losses by males and females are affected by many common factors, the assumption that $X \leqslant Y$ seems also reasonable.

We use the Hospital Costs data of Frees (2009) which were originally from the Nationwide Inpatient Sample of the Healthcare Cost and Utilization Project (NIS-HCUP), to represent the individual losses of the health insurance policies. The data contains 500 observations with 244 males and 256 females. We generate 1000 bootstrapping samples of the total losses caused by 50 males and 50 females, respectively.

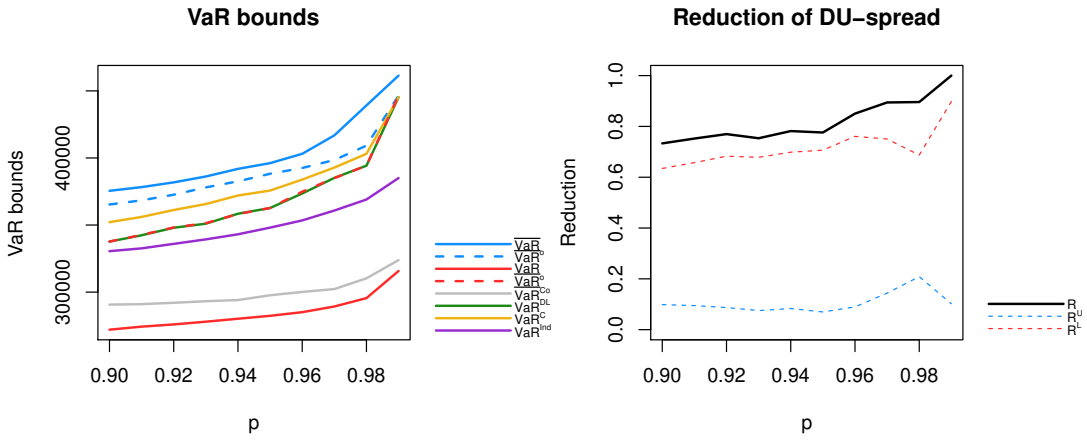
Figure 4.7: Empirical and estimated distributions of X and Y. Top panels: entire region; bottom panels: tail region



The empirical distributions $\hat{F}$ and $\hat{G}$ from the 1000 bootstrapping samples are plotted in the top-left panel of Figure 4.7. Although $\hat{F}$ and $\hat{G}$ do not satisfy $\hat{F} \leqslant_{\text{st}} \hat{G}$, such a violation is almost invisible (see the bottom-left panel of Figure 4.7) and possibly caused by sampling randomness. Indeed, using the Kolmogorov-Smirnov-type test of Barrett and Donald (2003), we cannot reject the hypothesis $F \leqslant_{\text{st}} G$ for the bootstrap data. Hence, $F \leqslant_{\text{st}} G$ is a sensible assumption. The isotonic distributional regression (IDR), introduced by Henzi et al. (2021), is a nonparametric technique to estimate distributions with order restrictions (e.g., stochastic order and hazard rate order). We use IDR to estimate F and G such that the stochastic order holds for the estimated distributions. The estimated distributions are plotted in the top-right panel of Figure 4.7, and they are used to calculate the VaR bounds. However, if $\hat{F} \leqslant_{\text{st}} \hat{G}$ holds already, IDR is not necessary, and we can directly use the empirical distributions.

Using the IDR estimated distributions F and G, VaR bounds and the improvements on the DU-spread in (4.9) are presented in Figure 4.8. VaR values are also presented if risks are independent, comonotonic, DL-coupled and countermonotonic. The extra order constraint greatly improves the unconstrained bounds of VaR, as shown by a DU-spread reduction of more than 69%. In particular, the improvement on the best-case value is greater than that on the worst-case value. The reduction is almost 100% when p is close to 1; that is because the two distributions $F^{[p,1]}$ and $G^{[p,1]}$ are almost identical for such p, making the set $\mathcal{F}_2^o(F^{[p,1]}, G^{[p,1]})$ very small (see Figure 4.7, bottom panels). Moreover, we observe that if the two risks X and Y are countermonotonic, VaR is close to the unconstrained lower bound. If the two risks are DL-coupled or comonotonic, VaR is close to the constrained lower bound.

Figure 4.8: Case study: VaR_p bounds, DU reduction and VaR_p of the aggregate risk with different dependence structures are contained in this figure. VaR values with independence, comonotonicity, countermonotonicity and DL coupling are denoted by $\mathrm{VaR}^{\mathrm{Ind}}$, $\mathrm{VaR}^{\mathrm{C}}$, $\mathrm{VaR}^{\mathrm{Co}}$ and $\mathrm{VaR}^{\mathrm{DL}}$, respectively.



4.7 Concluding remarks

Risk aggregation of two ordered risks in the presence of unknown dependence structure is studied in this chapter. The optimal dependence structures of the aggregate position are discussed in the sense of concave order, which can also be equivalently described via

convex order. The largest (resp. smallest) aggregate position in concave order is attained when the risks are DL-coupled (resp. comonotonic). The concave ordering bounds can be immediately applied to derive the bounds of $\leqslant_{\mathrm{cv}}$-consistent and $\leqslant_{\mathrm{cx}}$-consistent risk measures.

To analyze bounds on tail risk measures such as VaR, we introduce the notion of strong stochastic order and develop several theoretical properties. In particular, if the generator of the tail risk measure is $\leqslant_{\mathrm{cv}}$-consistent, the worst-case value of the tail risk measure with the order constraint can be attained by p-concentrated risks, and it is attained when the upper-tail risks are DL-coupled. With a specific focus on VaR, analytical solutions are derived. Numerical studies show that the extra order constraint on top of the marginal distributions can significantly improve the bounds of risk measures which are solely based on marginal distributions.

There are some limitations of the current setup considered in this chapter. First, the assumption $X \leqslant Y$ for two risks X and Y is arguably quite strong. As we have seen from the numerical results, significant improvement of the constrained bounds over the unconstrained ones requires that the risks are of similar size, which however renders the ordering assumption difficult to satisfy.

4.8 Appendix: Proof of Lemma 4.2

Proof of Lemma 4.2. By Proposition 1 of Embrechts and Hofert (2013), as F and G are strictly increasing and continuous, F^{-1} and G^{-1} are also strictly increasing and continuous. We first show

$$\lim_{\varepsilon \downarrow 0} \overline{\mathrm{VaR}}^R_{p+\varepsilon}(\mathcal{F}_2^o(F,G)) = \overline{\mathrm{VaR}}^R_p(\mathcal{F}_2^o(F,G)). \tag{4.10}$$

By Theorem 4.3, for $\varepsilon \geqslant [0, 1-p)$,

$$\overline{\mathrm{VaR}}^R_{p+\varepsilon}(\mathcal{F}_2^o(F,G)) = \text{ess-inf}(X_\varepsilon + Y_\varepsilon)$$

where $(X_\varepsilon, Y_\varepsilon) \sim D_*^{F^{[p+\varepsilon,1]}, G^{[p+\varepsilon,1]}}$. Since X_0 and Y_0 have continuous distributions, using Corollary 2.5 of Nutz and Wang (2021), we have $(X_\varepsilon, Y_\varepsilon) \to (X_0, Y_0)$ in distribution. Since ess-inf is upper semicontinuous with respect to convergence in distribution, we have

$$\overline{\mathrm{VaR}}^R_p(\mathcal{F}_2^o(F,G)) = \text{ess-inf}(X_0 + Y_0) \geqslant \lim_{\varepsilon \downarrow 0} \text{ess-inf}(X_\varepsilon + Y_\varepsilon) = \lim_{\varepsilon \downarrow 0} \overline{\mathrm{VaR}}^R_{p+\varepsilon}(\mathcal{F}_2^o(F,G)),$$

which implies (4.10). In what follows, we will show

$$\lim_{\varepsilon\downarrow 0}\overline{\mathrm{VaR}}^R_{p-\varepsilon}(\mathcal{F}_2^o(F,G))=\overline{\mathrm{VaR}}^R_p(\mathcal{F}_2^o(F,G)). \tag{4.11}$$

Fix $p\in(0,1)$. If $F^{-1}(p)=G^{-1}(p)$, by Proposition 4.3, $\overline{\mathrm{VaR}}^R_p(\mathcal{F}_2^o(F,G))=F^{-1}(p)+G^{-1}(p)$. For $\varepsilon>0$, by Theorem 4.3 and Corollary 4.1,

$$\overline{\mathrm{VaR}}^R_{p-\varepsilon}(\mathcal{F}_2^o(F,G))=\overline{\text{ess-inf}}\left(\mathcal{F}_2^o(F^{[p-\varepsilon,1]},G^{[p-\varepsilon,1]})\right)\geqslant F^{-1}(p-\varepsilon)+G^{-1}(p-\varepsilon).$$

Thus we have

$$F^{-1}(p-\varepsilon)+G^{-1}(p-\varepsilon)\leqslant\overline{\mathrm{VaR}}^R_{p-\varepsilon}(\mathcal{F}_2^o(F,G))\leqslant\overline{\mathrm{VaR}}^R_p(\mathcal{F}_2^o(F,G))=F^{-1}(p)+G^{-1}(p).$$

As F^{-1} and G^{-1} are continuous, let ε go to 0, we get the desired result.

For the rest of the proof, we assume $F^{-1}(p)<G^{-1}(p)$. We first deal with the case where $F^{-1}(1)<\infty$ and $G^{-1}(1)<\infty$. For $\varepsilon>0$, let

$$\delta(\varepsilon)=\sup_{p\leqslant t\leqslant 1}\left\{\left[F^{-1}(t)-F^{-1}(t-\varepsilon)\right]\vee\left[G^{-1}(t)-G^{-1}(t-\varepsilon)\right]\right\}.$$

As F^{-1}, G^{-1} are continuous and $F^{-1}(1)<\infty$ and $G^{-1}(1)<\infty$, we have $0<\delta(\varepsilon)<\infty$ and $\delta(\varepsilon)\downarrow 0$ as $\varepsilon\downarrow 0$. Furthermore, let

$$h(\varepsilon)=\sup\left\{(F(G^{-1}(p))-p)-(F(z)-G(z)):z\in[G^{-1}(p-\varepsilon),G^{-1}(p)]\right\}.$$

Because $F-G$ is continuous, we have $0\leqslant h(\varepsilon)<\infty$ and $h(\varepsilon)\downarrow 0$ as $\varepsilon\downarrow 0$. As $F^{-1}(p)<G^{-1}(p)$, we can take ε small enough such that $F(G^{-1}(p))-p>h(\varepsilon)$ and $G^{-1}(p)-F^{-1}(p)>\delta(\varepsilon)$. By the definition of $\delta(\varepsilon)$, we also have

$$F^{-1}(p-\varepsilon)<F^{-1}(p)<G^{-1}(p)-\delta(\varepsilon)\leqslant G^{-1}(p-\varepsilon)<G^{-1}(p).$$

Define

$$x_\varepsilon=\inf\left\{x:F(x)-(p-\varepsilon)\geqslant F(G^{-1}(p))-p-h(\varepsilon)\right\}.$$

As F is strictly increasing and continuous, we have $F^{-1}(p-\varepsilon)<x_\varepsilon<G^{-1}(p)$. Furthermore, $x_\varepsilon\uparrow G^{-1}(p)$ as $\varepsilon\downarrow 0$. Let $d(\varepsilon)=G^{-1}(p)-x_\varepsilon$. Thus, $0<d(\varepsilon)<G^{-1}(p)-F^{-1}(p-\varepsilon)$ and $d(\varepsilon)\downarrow 0$ as $\varepsilon\downarrow 0$. Furthermore, for any $x<x_\varepsilon$, we have $F(x)-(p-\varepsilon)<F(G^{-1}(p))-p-h(\varepsilon)$.

From Proposition 4.3, we have

$$\overline{\mathrm{VaR}}^R_{p-\varepsilon}(\mathcal{F}_2^o(F,G))=\min\left\{\inf_{x\in[F^{-1}(p-\varepsilon),G^{-1}(p-\varepsilon)]}\left\{T^{F^{[p-\varepsilon,1]},G^{[p-\varepsilon,1]}}(x)+x\right\},2G^{-1}(p-\varepsilon)\right\}.$$

(i) For any $x \in [G^{-1}(p) - \delta(\varepsilon) \vee d(\varepsilon), G^{-1}(p-\varepsilon)]$, we have

$$\begin{aligned} T^{F^{[p-\varepsilon,1]},G^{[p-\varepsilon,1]}}(x) + x &\geqslant 2x \geqslant 2G^{-1}(p) - 2\delta(\varepsilon) \vee d(\varepsilon) \\ &\geqslant \overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) - 2\delta(\varepsilon) \vee d(\varepsilon). \end{aligned} \tag{4.12}$$

(ii) For any $x \in [F^{-1}(p-\varepsilon), G^{-1}(p) - \delta(\varepsilon) \vee d(\varepsilon))$, let $y = F^{-1}(F(x)+\varepsilon)$. As $F(x)+\varepsilon \geqslant p$, we have

$$y - x = F^{-1}(F(x)+\varepsilon) - F^{-1}(F(x)) \leqslant \delta(\varepsilon), \tag{4.13}$$

and $y \leqslant x + \delta(\varepsilon) \leqslant G^{-1}(p)$. Moreover, we have $y \geqslant F^{-1}(p)$. Therefore, $y \in [F^{-1}(p), G^{-1}(p)]$. By the definition of $h(\varepsilon)$ and $x < x_\varepsilon$, we have for all $z \in [G^{-1}(p-\varepsilon), G^{-1}(p)]$,

$$F(z) - G(z) > F(G^{-1}(p)) - p - h(\varepsilon) > F(x) - (p-\varepsilon).$$

Thus,

$$\begin{aligned} T^{F^{[p-\varepsilon,1]},G^{[p-\varepsilon,1]}}(x) &= \inf\{z > G^{-1}(p-\varepsilon) : F(z) - G(z) < F(x) - (p-\varepsilon)\} \\ &= \inf\{z > G^{-1}(p) : F(z) - G(z) < F(y) - p\} \\ &= T^{F^{[p,1]},G^{[p,1]}}(y). \end{aligned}$$

By (4.13), we have

$$T^{F^{[p-\varepsilon,1]},G^{[p-\varepsilon,1]}}(x)+x = T^{F^{[p,1]},G^{[p,1]}}(y)+y+(x-y) \geqslant T^{F^{[p,1]},G^{[p,1]}}(y)+y-\delta(\varepsilon). \tag{4.14}$$

Combining (4.12), (4.14) and the fact that $G^{-1}(p-\varepsilon) \geqslant G^{-1}(p) - \delta(\varepsilon) \vee d(\varepsilon)$, we conclude that, for $p \in (0,1)$,

$$\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) - 2\delta(\varepsilon) \vee d(\varepsilon) \leqslant \overline{\mathrm{VaR}}_{p-\varepsilon}^R(\mathcal{F}_2^o(F,G)) \leqslant \overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)).$$

Letting $\varepsilon \downarrow 0$, we get (4.11) for the case $F^{-1}(1) < \infty$ and $G^{-1} < \infty$.

If $F^{-1}(1) = \infty$ or $G^{-1}(1) = \infty$, following the proof of Proposition 4 in Blanchet et al. (2020), we have $\overline{\mathrm{VaR}}_p^R(\mathcal{F}_2^o(F,G)) = \overline{\mathrm{VaR}}_p^R\left(\mathcal{F}_2^o\left(F^{[0,m]}, G^{[0,m]}\right)\right)$ for $p \in [0, 2m-1)$ and $1/2 < m < 1$. Intuitively, extremely large values of risks do not contribute to the calculation of the worst-case value of VaR^R. As $\left(F^{[0,m]}\right)^{-1}(1) < \infty$ and $\left(G^{[0,m]}\right)^{-1}(1) < \infty$, $\overline{\mathrm{VaR}}_p^R\left(\mathcal{F}_2^o(F,G)\right)$ is continuous of $p \in (0, 2m-1)$. Letting $m \to 1$, we get (4.11). □

Chapter 5

Trade-off between Validity and Efficiency of Merging p-values under Arbitrary Dependence

5.1 Introduction

In many areas of statistical applications where multiple hypothesis testing is involved, the task of merging several p-values into one naturally arises. Depending on the specific application, these p-values may be from a single hypothesis or multiple hypotheses, in small or large numbers, independent or correlated, and with sparse or dense signals, leading to different considerations when choosing merging procedures.

Let K be a positive integer, and $F : [0,1]^K \to [0,\infty)$ be an increasing Borel function used to combine K p-values, which we shall refer to as a *combining function*. Generally, the combined value may not be a valid p-value itself, and a critical point needs to be specified. Different dependence assumptions on the p-values lead to significantly different critical points, and thus different statistical decisions. The problem of merging p-values has a long history, and early results can be found in Tippett (1931), Pearson (1933) and Fisher (1948) where p-values are assumed to be independent. Based on an idea of Tukey, Donoho and Jin (2004) developed the higher criticism statistics to detect weak and sparse signals effectively using independent p-values. Certainly, these methods do not always produce a valid p-value if the assumption of independence is violated. On the other hand, the independence assumption is often very difficult or impossible to verify in many applications where only one set of p-values is available.

There are, however, some methods that produce valid p-values without any dependence assumption. A classic one is the Bonferroni method by taking the minimum of the p-values times K (we allow combined p-values to be greater than 1 and they can be treated as 1) or equivalently, dividing the critical value by K. Other methods that are valid without assumptions include the ones based on order statistics by Rüger (1978) and Hommel (1983), and the ones based on averaging by Vovk and Wang (2020); details of these merging methods are presented in Section 5.3.

Some other methods work under weak or moderate dependence assumptions, such as the method of Simes (1986), which uses the minimum of $Kp_{(i)}/i$ over $i = 1, \dots, K$, where $p_{(i)}$ is the i-th smallest order statistic of $p_1, \dots, p_K$. The validity of the Simes method is shown under a large class of dependence structures (e.g., Sarkar (1998, 2008); Benjamini and Yekutieli (2001) and Rødland (2006)), although even such dependence assumptions are unlikely to hold in practice (see e.g., Efron (2010, p.51)). Two more recent methods include the Cauchy combination test proposed by Liu and Xie (2020) using the weighted average of Cauchy transformed p-values, and the harmonic mean p-value of Wilson (2019) using the harmonic mean of p-values. Under mild dependence assumptions, these two methods are asymptotically valid as the significance level goes to 0 (see Theorem 5.2).

This chapter is dedicated to a comprehensive and unifying treatment of p-value merging methods under various dependence assumptions. Some methods are valid without any assumption on the interdependence of p-values, and they will be referred to as *VAD methods*. On the other hand, methods that are valid for some specific but realistic dependence assumption (e.g., independence, positive dependence, or joint normality dependence) will be referred to as *VSD methods*. Our main goal is to understand the difference and the trade-off between these methods.

For a fixed combining function F, using a VAD method means choosing a smaller critical value (threshold) for making rejections compared to a VSD method. Thus, the gain of validity comes at the price of a loss of detection power. As it is often difficult to make valid statistical inference on the dependence structure of p-values, our analysis also helps to understand the relative performance of VSD combining methods under the presence of model misspecification. As a byproduct, we obtain several new theoretical results on the popular Simes, harmonic, and Cauchy merging methods.

In the next section, we collect some basic definitions of VAD and VSD merging methods and their corresponding threshold functions. We focus on symmetric merging functions for the tractability in their comparison. In Section 5.3, we introduce two general classes of combining functions, which include all methods mentioned above. Formulas for their VAD and VSD threshold functions are derived, some based on results from robust risk aggrega-

tion, e.g., Wang et al. (2013). In Section 5.4, we introduce independence-comonotonicity balanced (IC-balanced) combining functions, which are indifferent between the two dependence assumptions. We show that the Cauchy combination method and the Simes method are the only IC-balanced ones among two general classes of combining methods, thus highlighting their unique roles. In Section 5.5, we establish strong similarity between the Cauchy combination and the harmonic averaging methods, and obtain an algebraic relationship between the harmonic averaging and the Simes functions. In Section 5.6, the price for validity is introduced to assess the loss of power of VAD methods compared to their VSD versions. Simulation studies and a real data analysis are conducted to analyze the relative performance of these methods. Simulation studies and a real data analysis are presented in Section 5.7 to analyze the relative performance of these methods. Proofs of all technical results are put in Section 5.9.1.

We conclude the section by providing additional notation and terminology that will be adopted in this chapter. All random variables are defined on an atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Random variables $X_1, \dots, X_n$ are comonotonic if there exist increasing functions $f_1, \dots, f_n$ and a random variable Z such that $X_i = f_i(Z)$ for each $i = 1, \dots, n$. For $\alpha \in (0, 1]$, $q_\alpha(X)$ is the left α-quantile of a random variable X, defined as

$$q_\alpha(X) = \inf\{x \in \mathbb{R} \mid \mathbb{P}(X \leqslant x) \geqslant \alpha\}.$$

We also use $F^{-1}(\alpha)$ for $q_\alpha(X)$ if X follows the distribution F. The set $\mathcal{U}$ is the set of all standard uniform random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ (i.e., the set of all measurable functions on $(\Omega, \mathcal{F})$ whose distribution under $\mathbb{P}$ is uniform on $[0, 1]$) and $\mathbb{1}$ is the indicator function. The equality $\overset{\mathrm{d}}{=}$ represents equality in distribution. For given $p_1, \dots, p_K$, the order statistics $p_{(1)}, \dots, p_{(K)}$ are ordered from the smallest to the largest. The equivalence $A_x \sim B_x$ as $x \to x_0$ means that $A_x/B_x \to 1$ as $x \to x_0$. All terms of "increasing" and "decreasing" are in the non-strict sense.

5.2 Merging methods and thresholds

Following the terminology of Vovk and Wang (2020), a *p-variable* is a random variable P such that $\mathbb{P}(P \leqslant \varepsilon) \leqslant \varepsilon$, for all $\varepsilon \in (0, 1)$ (such random variables are called *superuniform* by Ramdas et al. (2019)). Values realized by p-variables are p-values. In the Introduction, p-values are used loosely for p-variables, which should be clear from the context.

Let $P_1, \dots, P_K$ be K p-variables for testing a common hypothesis. A *combining function* is an increasing Borel measurable function $F : [0, 1]^K \to [0, \infty)$ which transforms

$P_1, \dots, P_K$ into a single random variable $F(P_1, \dots, P_K)$. The choice of combining function depends on how one integrates information, and some common options are mentioned in the Introduction. Generally, $F(P_1, \dots, P_K)$ may not be a valid p-variable. For different choices of F and assumptions on $P_1, \dots, P_K$, one needs to assign a critical value $g(\varepsilon)$ so that the hypothesis can be rejected with significance level $\varepsilon \in (0,1)$ if $F(P_1, \dots, P_K) < g(\varepsilon)$. We call g a *threshold (function)* for F and $P_1, \dots, P_K$. Clearly, $g(\varepsilon)$ is increasing in ε. In case g is strictly increasing, which is the most common situation, the above specification of g is equivalent to requiring $g^{-1} \circ F(P_1, \dots, P_K)$ to be a p-variable. To objectively compare various combining methods, one should compare the corresponding values of the function $g^{-1} \circ F$.

In some situations, it might be convenient and practical to assume additional information on dependence structure of p-variables, e.g., independence, comonotonicity (i.e., perfectly positive dependence), and specific copulas. The choice of the threshold g certainly depends on such assumptions. If no assumption is made on the interdependence of the p-variables, the corresponding threshold function is called a *VAD threshold*, otherwise it is a *VSD threshold*. A testing procedure based on a VAD threshold always produces a size less than or equal to the significance level regardless of the dependence structure of the p-variables.

We denote the VAD threshold of a combining function F by a_F. If a merging method is valid for independent (resp. comonotonic) dependence of p-variables, we use b_F (resp. c_F) to denote the corresponding valid threshold function, and we call it the *VI* (resp. *VC*) *threshold*. More precisely, for the equation

$$\mathbb{P}(F(P_1, \dots, P_K) < g(\varepsilon)) \leqslant \varepsilon, \quad \varepsilon \in (0,1), \tag{5.1}$$

a VAD threshold $g = a_F$ satisfies (5.1) for all p-variables $P_1, \dots, P_K$; a VI threshold $g = b_F$ satisfies (5.1) for all independent p-variables $P_1, \dots, P_K$, and a VC threshold $g = c_F$ satisfies (5.1) for all comonotonic p-variables $P_1, \dots, P_K$.

The comonotonicity assumption on the p-variables to combine (actually they are identical if they are uniform on $[0,1]$) is not interesting by itself for statistical practice. Nevertheless, comonotonicity is a benchmark for (extreme) positive dependence, and we analyze c_F for the purpose of comparison; it helps us to understand how valid thresholds for different methods vary as the dependence assumption gradually shifts from independence to extreme positive dependence. This point will be made more clear in Sections 5.4-5.6.

An immediate observation is that the p-variables can be equivalently replaced by uniform random variables on $[0,1]$ as for each p-variable P, we can find $U \in \mathcal{U}$ with $U \leqslant P$; see e.g., Vovk and Wang (2020). Therefore, it suffices to consider p-variables in $\mathcal{U}$. Moreover,

if g satisfies (5.1), then any function that is smaller than g is also valid. Hence, for the sake of power, it is natural to use the largest functions that satisfy (5.1). Putting these considerations together, we formally define the thresholds of interest as follows.

Definition 5.1. The thresholds a_F, b_F and c_F of a combining function F are given by, for $\varepsilon \in (0,1)$,

$$a_F(\varepsilon) = \inf\{q_\varepsilon(F(U_1,\dots,U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\}, \tag{5.2}$$
$$b_F(\varepsilon) = q_\varepsilon(F(V_1,\dots,V_K)), \tag{5.3}$$
$$c_F(\varepsilon) = q_\varepsilon(F(U,\dots,U)), \tag{5.4}$$

where $U, V_1, \dots, V_K$ are independent standard uniform random variables.

In what follows, we focus on the thresholds in Definition 5.1. It is clear that $g = a_F$, b_F or c_F in Definition 5.1 satisfies (5.1) under the respective dependence assumptions.

Remark 5.1. While the objects b_F and c_F in (5.3)-(5.4) can often be explicitly calculated, the object a_F in (5.2) is generally difficult to calculate for a chosen function F due to the infimum taken over all possible dependence structures. Techniques in the field of robust risk aggregation, in particular, results in Wang et al. (2013), Embrechts et al. (2013, 2015) and Wang and Wang (2016), are designed for such calculation, as illustrated by Vovk and Wang (2020). By definition, for any threshold $g(\varepsilon) > a_F(\varepsilon)$, there exists some dependence structure of $(P_1,\dots,P_K)$ such that validity is lost, i.e., (5.1) is violated. Moreover, if the combining function F is continuous, the infimum in (5.2) is attainable; the proof of this statement is similar to that of Lemma 4.2 of Bernard et al. (2014).

5.3 Combining functions

5.3.1 Two general classes of combining functions

We first introduce two general classes of combining functions, the generalized mean class and the order statistics class. Let $p_1,\dots,p_K \in [0,1]$ be the K realized p-values. The first class of combining functions is the generalized mean, that is,

$$M_{\phi,K}(p_1,\dots,p_K) = \phi^{-1}\left(\frac{1}{K}\sum_{i=1}^{K}\phi(p_i)\right),$$

where $\phi : [0,1] \to [-\infty, \infty]$ is a continuous and strictly monotone function and ϕ^{-1} is its inverse on the domain $\phi([0,1])$. Many combining functions used in the statistical literature are included in this class. For example, the Fisher method (Fisher (1948)) corresponds to the geometric mean with $\phi(p) = \log(p)$; the averaging methods of Vovk and Wang (2020) and Wilson (2019) correspond to the functions $\phi(p) = p^r$, and $r \in [-\infty, \infty]$ (including limit cases), and the Cauchy combination method of Liu and Xie (2020) corresponds to $\phi(p) = \tan\left(\pi\left(p - \frac{1}{2}\right)\right)$.

The second class of combining functions is built on order statistics. Let

$$\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K,$$

where $\mathbb{R}_+ = [0, \infty)$. We define the combining function

$$S_{\alpha,K}(p_1, \ldots, p_K) = \min_{i \in \{1,\ldots,K\}} \frac{p_{(i)}}{\alpha_i},$$

where the convention is $p_{(i)}/\alpha = \infty$ if $\alpha = 0$. If $\alpha_1 = 1/K$ and all the other components of α are 0, then using $S_{\alpha,K}$ yields the Bonferroni method based on the minimum of p-values. The VAD method via order statistics of Rüger (1978) uses $S_{\alpha,K}$ by setting $\alpha_i = i/K$ for a fixed $i \in \{1, \ldots, K\}$ and all the other components of α to be 0. On the other hand, if $\alpha_i = i/K$ for each $i = 1, \ldots, K$, then we arrive at the method of Simes (1986); in this case, we will simply denote $S_{\alpha,K}$ by S_K, namely,

$$S_K(p_1, \ldots, p_K) := \min_{i \in \{1,\ldots,K\}} \frac{K p_{(i)}}{i},$$

and S_K will be called the *Simes function*. The method of Hommel (1983) uses $\ell_K S_K$, which is S_K adjusted via the VAD threshold, where

$$\ell_K = \sum_{k=1}^{K} \frac{1}{k}. \tag{5.5}$$

If $\alpha_{i+1} \leqslant \alpha_i$, then the term $p_{(i+1)}/\alpha_{i+1}$ does not contribute to the calculation of the term $S_{\alpha,K}(p_1, \ldots, p_K)$. Hence, we can safely replace α_{i+1} by α_i without changing the function $S_{\alpha,K}$. Thus, we shall assume, without loss of generality, that $\alpha_1 \leqslant \ldots \leqslant \alpha_K$. Admissibility of VAD merging methods in the above two classes are studied by Vovk et al. (2021).

Recall that a function $F : \mathbb{R}_+^K \to \mathbb{R}$ is homogeneous if $F(\lambda \mathbf{x}) = \lambda F(\mathbf{x})$ for all $\lambda > 0$ and $\mathbf{x} \in \mathbb{R}_+^K$. It is clear that the function $S_{\alpha,K}$ is homogeneous, and so are the averaging methods of Vovk and Wang (2020). In such cases, we can show that the VAD threshold a_F is a linear function.

Proposition 5.1. *If the combination function F is homogeneous, then the VAD threshold $a_F(x)$ is a constant times x on $(0,1)$.*

In the subsections below we will discuss several special cases of the above two classes of combining functions, and analyze their corresponding threshold functions. As the first example, we note that the functions a_F, b_F and c_F for the Bonferroni method can be easily verified.

Proposition 5.2. *Let $F(p_1,\dots,p_K) = \min\{p_1,\dots,p_K\}$ for $p_1,\dots,p_K \in [0,1]$. Then $a_F(\varepsilon) = \varepsilon/K$, $b_F(\varepsilon) = 1-(1-\varepsilon)^{1/K}$ and $c_F(\varepsilon) = \varepsilon$ for $\varepsilon \in (0,1)$.*

5.3.2 The averaging methods

The aforementioned averaging methods of Vovk and Wang (2020) use the combining functions given by

$$M_{r,K}(p_1,\dots,p_K) = \left(\frac{p_1^r+\cdots+p_K^r}{K}\right)^{\frac{1}{r}},$$

for $r \in \mathbb{R}\setminus\{0\}$, together with its limit cases

$$\begin{aligned} M_{-\infty,K}(p_1,\dots,p_K) &= \min\{p_1,\dots,p_K\};\\ M_{0,K}(p_1,\dots,p_K) &= \left(\prod_{i=1}^K p_i\right)^{\frac{1}{K}};\\ M_{\infty,K}(p_1,\dots,p_K) &= \max\{p_1,\dots,p_K\}. \end{aligned}$$

Some special cases of the combining functions above are $r=-\infty$ (minimum), $r=-1$ (harmonic mean), $r=0$ (geometric mean), $r=1$ (arithmetic mean) and $r=\infty$ (maximum); the cases $r\in\{-1,0,1\}$ are known as Platonic means. Note that $M_{-\infty,K}$ gives rise to the Bonferroni method, and the geometric mean yields Fisher's method (Fisher (1948)) under the independence assumption. The harmonic mean p-value of Wilson (2019) is a VSD method using the harmonic mean.

Since the mean function $M_{r,K}$ is homogeneous, by Proposition 5.1, the VAD threshold is a linear function $a_F(x) = a_r x$, $x\in(0,1)$ for some $a_r > 0$. The multipliers a_r have been well studied in Vovk and Wang (2020), and here we mainly focus on the cases of Platonic means and the Bonferroni method. It is known that $a_{-\infty} = 1/K$ and $a_1 = 1/2$. For $r=0$ or $r=-1$, the values of a_r and their asymptotic formulas are calculated by Propositions 4 and 6 of Vovk and Wang (2020), summarized below for $K \geqslant 3$.

(i) For $F = M_{0,K}$,

$$a_F(x) = a_0 x = c_K \exp\left(\frac{K-1}{1-Kc_K}\right) \times x, \quad x \in (0,1), \tag{5.6}$$

where c_K is the unique solution to the equation: $\log(1/c - (K-1)) = K - K^2 c$ for $c \in (0, 1/K)$. Moreover, $a_0 \geqslant 1/e$, and $a_0 \to 1/e$ as $K \to \infty$.

(ii) For $F = M_{-1,K}$,

$$a_F(x) = a_{-1} x = \frac{(y_K+1)K}{(y_K+K)^2} \times x, \quad x \in (0,1), \tag{5.7}$$

where y_K is the unique solution to the equation: $y^2 = K((y+1)\log(y+1) - y)$ for $y \in (0, \infty)$. Moreover, $a_{-1} \geqslant (e \log K)^{-1}$, and $a_{-1} \log K \to 1$ as $K \to \infty$.

To determine the VC threshold, it is easy to check that $c_{M_{r,K}}(x) = x$, $x \in (0,1)$ for all $r \in [-\infty, \infty]$, because the generalized mean of identical objects is equal to themselves; this obviously holds for all functions in the family of $M_{\phi,K}$.

Next, we study $b_r := b_{M_{r,K}}$ or its approximate form. For this, we will use stable distributions (e.g., Uchaikin and Zolotarev (2011) and Samorodnitsky (2017)) below. Let F_α be the stable distribution with stability parameter $\alpha \in (0,2)$, skewness parameter $\beta = 1$, scale parameter $\sigma = 1$ and shift parameter $\mu = 0$. The characteristic function of F_α is given by, for $\theta \in \mathbb{R}$,

$$\int \exp(i\theta x) \mathrm{d}F_\alpha(x) = \begin{cases} \exp\left(-|\theta|^\alpha (1 - i\,\mathrm{sgn}(\theta) \tan\frac{\pi\alpha}{2})\right) & \text{if } \alpha \neq 1, \\ \exp\left(-|\theta|(1 + i\frac{2}{\pi}\,\mathrm{sgn}(\theta)\log|\theta|)\right) & \text{if } \alpha = 1, \end{cases}$$

where $\mathrm{sgn}(\cdot)$ is the sign function. For $\alpha \geqslant 2$, let F_α stand for the standard normal distribution.

Proposition 5.3. *Let b_r be the VI threshold of $M_{r,K}$, $r \in \mathbb{R}$.*

(i) If $r < 0$, then for $K \in \mathbb{N}_+$

$$b_r(\varepsilon) \sim K^{-1-1/r}\varepsilon, \quad \text{as } \varepsilon \downarrow 0, \tag{5.8}$$

and for $\varepsilon \in (0,1)$,

$$b_r(\varepsilon) \sim \left(\left(C_\alpha F_\alpha^{-1}(1-\varepsilon) + b_K\right)/K\right)^{\frac{1}{r}}, \quad \text{as } K \to \infty,$$

where $\alpha = -1/r > 0$ and the constants C_α and b_K are given in Table 5.1.

(ii) If $r = 0$, then

$$b_r(\varepsilon) = \exp\left(-\frac{1}{2K} q_{1-\varepsilon}\left(\chi^2_{2K}\right)\right). \tag{5.9}$$

(iii) If $r > 0$, then for $K \in \mathbb{N}_+$,

$$b_r(\varepsilon) = \frac{(\Gamma(1+K/p))^{1/K}\varepsilon^{1/K}}{K^{1/r}\Gamma(1+1/p)}, \qquad \text{if } \varepsilon \leqslant \tfrac{(\Gamma(1+1/p))^K}{\Gamma(1+K/p)},$$

where Γ is the Gamma function. For $\varepsilon \in (0,1)$,

$$b_r(\varepsilon) \sim \left(\frac{\sigma}{\sqrt{K}}\Phi^{-1}(\varepsilon) + \mu\right)^{\frac{1}{r}}, \qquad \text{as } K \to \infty,$$

where $\mu = (r+1)^{-1}$ and $\sigma^2 = r^2(1+2r)^{-1}(1+r)^{-2}$.

Table 5.1: Coefficients C_α and b_K for $r = -1/\alpha < 0$.

$r = -1/\alpha$	C_α	b_K
$-\frac{1}{2} < r < 0$	$\left(K\left(\frac{\alpha}{\alpha-2} - \left(\frac{\alpha}{\alpha-1}\right)^2\right)\right)^{1/2}$	$K\alpha/(\alpha-1)$
$r = -\frac{1}{2}$	$\sqrt{K\log K}$	$K\alpha/(\alpha-1)$
$-1 < r < -\frac{1}{2}$	$K^{1/\alpha}\left(\Gamma(1-\alpha)\cos(\pi\alpha/2)\right)^{1/\alpha}$	$K\alpha/(\alpha-1)$
$r = -1$	$K\pi/2$	$\frac{\pi K^2}{2}\int_1^\infty \sin\left(\frac{2x}{K\pi}\right)\alpha x^{-\alpha-1}\mathrm{d}x$
$r < -1$	$K^{1/\alpha}\left(\Gamma(1-\alpha)\cos(\pi\alpha/2)\right)^{1/\alpha}$	0

5.3.3 The Cauchy combination method

The Cauchy combination method is recently proposed by Liu and Xie (2020) which relies on a special case of the generalized mean via $\phi = \mathcal{C}^{-1}$, where $\mathcal{C}$ is the standard Cauchy cdf, that is,

$$\mathcal{C}(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2},\ x \in \mathbb{R}; \qquad \mathcal{C}^{-1}(p) = \tan\left(\pi\left(p - \frac{1}{2}\right)\right),\ p \in (0,1).$$

We denote this combining function by $M_{\mathcal{C},K}$ (instead of $M_{\mathcal{C}^{-1},K}$ for simplicity), namely,

$$M_{\mathcal{C},K}(p_1,\ldots,p_K) := \mathcal{C}\left(\frac{1}{K}\sum_{i=1}^K \mathcal{C}^{-1}(p_i)\right).$$

It is well known that the arithmetic average of either independent or comonotonic standard Cauchy random variables follows again the standard Cauchy distribution. This feature allows the use of such a combination method to combine p-values under uncertain dependence assumptions. In addition, Liu and Xie (2020) showed that under a bivariate normality assumption of the individual test statistics (i.e., a normal copula), the combined p-value has the same asymptotic behaviour as the one under the assumption of independence (see Theorem 5.2 (ii) below).

Since $\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)$ follows a standard Cauchy distribution if $U_1,\dots,U_K \in \mathcal{U}$ are either independent or comonotonic, we have $b_F(x) = c_F(x) = x$ for all $x \in (0,1)$. This convenient feature will be studied in more details in Section 5.4.

By Definition 5.1, we get, for $F = M_{\mathcal{C},K}$,

$$a_F(\varepsilon) = \mathcal{C}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)\right) \mid U_1,\dots,U_K \in \mathcal{U}\right\}\right). \tag{5.10}$$

The function a_F does not admit an explicit formula, but it can be calculated via results from robust risk aggregation (Corollary 3.7 in Wang et al. (2013)) as in the following proposition.

Proposition 5.4. *For $\varepsilon \in (0,1/2)$, we have*

$$a_F(\varepsilon) = \mathcal{C}\left(-H_\varepsilon(x_K)/K\right), \tag{5.11}$$

where $H_\varepsilon(x) = (K-1)\mathcal{C}^{-1}(1-\varepsilon+(K-1)x) + \mathcal{C}^{-1}(1-x)$, $x \in (0,\varepsilon/K)$, and x_K is the unique solution $x \in (0,\varepsilon/K)$ to the equation

$$K\int_x^{\varepsilon/K} H_\varepsilon(t)\mathrm{d}t = (\varepsilon - Kx)H(x).$$

5.3.4 The Simes method

The method of Simes (1986) uses the Simes function S_K in the order statistics family, given by $S_K(p_1,\dots,p_K) = \min_{i\in\{1,\dots,K\}}\frac{K}{i}p_{(i)}$. For $F = S_K$, the results in Hommel (1983) together with Proposition 5.1 suggest that $a_F(x) = x/\ell_K$ for $x \in (0,1)$. For independent p-variables $P_1,\dots,P_K \in \mathcal{U}$, Simes (1986) obtained

$$\mathbb{P}\left(\min_{i\in\{1,\dots,K\}}\frac{K}{i}P_{(i)} > \varepsilon\right) = 1-\varepsilon, \quad \varepsilon \in (0,1),$$

which gives $b_F(x) = x$ for $x \in (0,1)$. For comonotonic p-variables $P_1, \dots, P_K \in \mathcal{U}$, it is clear that $S_K(P_1, \dots, P_K) = P_{(K)}$, which follows a standard uniform distribution, and hence we again have $c_F(x) = x$ for $x \in (0,1)$. The validity of the Simes function using the VI (VC) threshold (called the Simes inequality) holds under many positive dependence structures; see e.g., Sarkar (1998, 2008).

In the context of testing multiple hypotheses, if p-variables for several hypotheses are independent, the Benjamini-Hochberg procedure for controlling the false discovery rate (FDR) (Benjamini and Hochberg (1995)) also relies on the Simes function (in case all hypotheses are null). Although the Benjamini-Hochberg procedure is valid for many practical models, to control the FDR under arbitrary dependence structure of p-variables, one needs to multiply the p-values by ℓ_K, resulting in the Benjamini-Yekutieli procedure (Benjamini and Yekutieli (2001)). This constant is exactly $x/a_F(x)$, and the function a_F is called a reshaping function by Ramdas et al. (2019) in the FDR context.

5.4 Independence-comonotonicity balance

As we have seen above, the Cauchy function and the Simes function both satisfy $b_F = c_F$, and hence the corresponding merging methods are invariant under independence or comonotonicity assumption, an arguably convenient feature. Inspired by this observation, we introduce the property of *independence-comonotonicity balance* for combining functions in this section. This property distinguishes the Cauchy combination method and the Simes method from their corresponding classes $M_{\phi,K}$ and $S_{\alpha,K}$, respectively.

A combining function is said to be balanced between two different dependence structures of p-variables if the combined random variable under the two dependence assumptions coincide in distribution. Recall that $U, V_1, \dots, V_K$ are independent standard uniform random variables.

Definition 5.2. A combining function $F : [0,1]^K \to [0,\infty)$ is *independence-comonotonicity balanced* (*IC-balanced*) if $F(V_1, \dots, V_K) \overset{\mathrm{d}}{=} F(U, \dots, U)$.

As the VI and VC thresholds are the corresponding quantile functions of $F(P_1, \dots, P_K)$, we immediately conclude that a combining function $F : [0,1]^K \to [0,\infty)$ is IC-balanced if and only if $b_F = c_F$ on $(0,1]$; recall that c_F is the identity for all functions in Section 5.3.

IC-balanced methods have the same threshold $b_F = c_F$ if the dependence structure of

p-variables is a mixture of independence and comonotonicity, i.e., with the copula

$$\lambda \prod_{i=1}^{n} x_i + (1-\lambda) \min_{i=1,\dots,n} x_i, \quad (x_1, \dots, x_n) \in [0,1]^n, \tag{5.12}$$

where $\lambda \in [0,1]$. This is because $\mathbb{P}(F(U_1, \dots, U_K) \leqslant b_F(\varepsilon))$ is linear in the distribution of $(U_1, \dots, U_K)$.

For any combining function F, VI and VC thresholds generally yield more power to the test compared with the corresponding VAD threshold, but the gain of power may come with the invalidity due to model misspecification. If a combining function F is IC-balanced, the validity is preserved under independence, comonotonicity and their mixtures, and we may expect (without mathematical justification) that, to some extent, the size of the test can be controlled properly even if mild model misspecification exists. Therefore, the notion of IC-balance can be interpreted as insensitivity to some specific type of model misspecification (e.g., dependence structure given in (5.12)) for VSD merging methods.

We have already seen in Section 5.3 that the Cauchy combination method and the Simes method are IC-balanced. Below we show that they are the only IC-balanced methods among the two classes of combining functions based on generalized mean and order statistics.

Theorem 5.1. *For a generalized mean function $M_{\phi,K}$ and an order statistics function $S_{\alpha,K}$,*

(i) $M_{\phi,K}$ is IC-balanced for all $K \in \mathbb{N}$ if and only if it is the Cauchy combining function, i.e., $\phi(p)$ is a linear transform of $\tan\left(\pi\left(p - \frac{1}{2}\right)\right)$, $p \in (0,1)$;

(ii) $S_{\alpha,K}$ is IC-balanced if and only if it is a positive constant times the Simes function.

The IC-balance of $M_{\phi,K}$ for some fixed K (instead of all $K \in \mathbb{N}$) does not imply that ϕ is the quantile function of a Cauchy distribution; see the counter-example (Example 5.1) in Section 5.9.1. As a direct consequence of Theorem 5.1, if $S_{\alpha,K}$ is IC-balanced, then $S_{\alpha,k}$ for $k = 2, \dots, K-1$, are also IC-balanced (here we use the first k components of α); a similar statement does not hold in general for the generalized mean functions, also shown by Example 5.1.

Remark 5.2. The property of IC-balance should be seen as a necessary but not sufficient condition for a merging method to be insensitive to dependence between independence and comonotonicity. As shown by Sarkar (1998), the Simes method is valid for positive

regression dependence, which is a large spectrum of dependence structures connecting independence and comonotonicity (larger than (5.12)); on the other hand, the Cauchy combination method using VI threshold is valid under a bivariate Gaussian assumption asymptotically but not precisely (Liu and Xie (2020)); see Theorem 5.2 below and the simulation studies in Section 5.7. Instead of arguing for the practical usefulness of IC-balance, we emphasize it as a necessary condition for insensitivity to dependence. The main aim of Theorem 5.1 is, via this necessary condition, to pin down the unique role of the Simes and the Cauchy combination methods among their respective generalized classes, thus justifying their advantages with respect to dependence.

5.5 Connecting the Simes, the harmonic averaging and the Cauchy combination methods

As we have seen from Theorem 5.1, the Cauchy and Simes combining functions are the only IC-balanced ones among the two classes considered in Section 5.3. Although the harmonic combining function does not satisfy $b_F = c_F$, we observe empirically that the harmonic averaging method and the Cauchy combination method report very similar results in all simulations; see Section 5.7.

In this section, we explore the relationship among the three methods based on S_K, $M_{-1,K}$ and $M_{\mathcal{C},K}$. We first show that the harmonic averaging method is equivalent to the Cauchy combination method asymptotically in a few senses. Second, we show the Simes function S_K and the harmonic averaging function $M_{-1,K}$ are closely connected via $M_{-1,K} \leqslant S_K \leqslant \ell_K M_{-1,K}$, where ℓ_K is given in (5.5). Throughout this section, for fixed $K \in \mathbb{N}$, we write $a_{\mathcal{C}} = a_{M_{\mathcal{C},K}}$, $a_{\mathcal{S}} = a_{S_K}$, $a_{\mathcal{H}} = a_{M_{-1,K}}$ and similarly for $b_{\mathcal{C}}$, $b_{\mathcal{S}}$ and $b_{\mathcal{H}}$.

We will use the following assumption on the p-variables $U_1, \dots, U_K \in \mathcal{U}$.

(G) For each $1 \leqslant i < j \leqslant K$, (U_i, U_j) follows a bivariate Gaussian copula (which can be different for each pair).

The assumption (G) is mild and is imposed by Liu and Xie (2020, Condition C.1). Note that condition (G) includes independence and comonotonicity as special cases. The following theorem confirms the close relationship between the harmonic averaging method and the Cauchy combination method. Recall that the VC thresholds for both methods are the identity function, and thus it suffices to look at VAD and VI thresholds.

Theorem 5.2. *For fixed $K \in \mathbb{N}$, the harmonic averaging and the Cauchy combination methods are asymptotically equivalent in the following senses:*

(i) If $\min_{i\in\{1,\dots,K\}} p_i \downarrow 0$ *and* $\max_{i\in\{1,\dots,K\}} p_i \leqslant c$ *for some fixed* $c \in (0,1)$*, then*

$$\frac{M_{\mathcal{C},K}(p_1,\dots,p_K)}{M_{-1,K}(p_1,\dots,p_K)} \to 1.$$

(ii) For K standard uniform random variables $U_1,\dots,U_K$ satisfying condition (G),

$$\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K) < \varepsilon\right) \sim \mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K) < \varepsilon\right) \sim \varepsilon, \ \text{as } \varepsilon \downarrow 0. \tag{5.13}$$

In particular, $b_{\mathcal{C}}(\varepsilon) \sim b_{\mathcal{H}}(\varepsilon)$ as $\varepsilon \downarrow 0$.

(iii) $a_{\mathcal{C}}(\varepsilon) \sim a_{\mathcal{H}}(\varepsilon)$ as $\varepsilon \downarrow 0$.

(iv) For $r \neq -1$,

$$\frac{M_{\mathcal{C},K}(p_1,\dots,p_K)}{M_{r,K}(p_1,\dots,p_K)} \not\to 1, \ \text{as} \max_{i\in\{1,\dots,K\}} p_i \downarrow 0.$$

Remark 5.3. The statement $\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K) < \varepsilon\right) \sim \varepsilon$ in Theorem 5.2 (ii) is implied by Theorem 1 of Liu and Xie (2020), which gives the same convergence rate for the weighted Cauchy combination method. For the weighted harmonic averaging method, we have a similar result (see (5.26) in Section 5.9.1): For standard uniform random variables $U_1,\dots,U_K$ satisfying condition (G) and any $(w_1,\dots,w_K) \in [0,1]^K$ with $\sum_{i=1}^K w_i = 1$, we have

$$\mathbb{P}\left(\sum_{i=1}^{K} w_i U_i^{-1} > 1/\varepsilon\right) \sim \varepsilon, \ \text{as } \varepsilon \downarrow 0.$$

We omit a discussion on weighted merging methods as the focus of this chapter is comparing symmetric combination functions.

The first statement of Theorem 5.2 means that, if at least one of realized p-values are close to 0, the harmonic averaging and the Cauchy combining functions will produce very close numerical results. This case is likely to happen in high-dimensional situations where the number of p-variables is very large. As the condition (G) for (ii) in Theorem 5.2 is arguably mild, the thresholds of the two methods are similar for a small significance level under a wide range of dependence structures of p-variables (including independence and comonotonicity). Therefore, if the significance level is small, one likely arrives at the same statistical conclusions on the hypothesis testing by using either method. The

third result in Theorem 5.2 illustrates the equivalence between the VAD thresholds of the harmonic averaging and the Cauchy combination methods as the significance level goes to 0. The final result in Theorem 5.2 shows that among all averaging methods, the harmonic averaging method is the only one that is asymptotically equivalent to the Cauchy combination method.

Remark 5.4. We note that the equivalence

$$\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K)<\varepsilon\right)\sim\mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K)<\varepsilon\right)$$

in (5.13) does not always hold under arbitrary dependence structures. Since the Cauchy distribution is symmetric, it is possible that $\mathbb{P}(\mathcal{C}^{-1}(U_1)+\cdots+\mathcal{C}^{-1}(U_K)=0)=1$ for some $U_1,\dots,U_K\in\mathcal{U}$, implying $\mathbb{P}(M_{\mathcal{C},K}(U_1,\dots,U_K)<1/2)=0$. Indeed, Theorem 4.2 of Puccetti et al. (2019) implies that there exist K standard Cauchy random variables whose sum is a constant c, for each $c\in[-K\log(K-1)/\pi, K\log(K-1)/\pi]$. On the other hand, $\mathbb{P}(M_{-1,K}(U_1,\dots,U_K)<\varepsilon)>0$ for all $\varepsilon>0$ and all $U_1,\dots,U_K\in\mathcal{U}$. Thus, $\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\dots,U_K)<\varepsilon\right)\sim\mathbb{P}\left(M_{-1,K}(U_1,\dots,U_K)<\varepsilon\right)$ does not hold.

Remark 5.5. The equivalence in Theorem 5.2 (ii) relies on the p-variables being uniform on $[0,1]$. For p-variables that are stochastically larger than uniform, the behaviour of the Cauchy combination method and that of the harmonic averaging method may diverge; nevertheless, by Theorem 5.2 (i), for a realized vector of p-values with at least one very small component, the two methods would produce similar values.

The next result reveals an intimate relationship between the Simes and the harmonic averaging methods.

Theorem 5.3. *For $p_1,\dots,p_K\in[0,1]$,*

$$M_{-1,K}(p_1,\dots,p_K)\leqslant S_K(p_1,\dots,p_K)\leqslant \ell_K M_{-1,K}(p_1,\dots,p_K).$$

The first inequality holds as an equality if $p_1=\cdots=p_K$. The second inequality holds as an equality if $p_1=p_k/k$ for $k=2,\dots,K$. As a consequence, $a_{\mathcal{S}}/a_{\mathcal{H}}\in[1,\ell_K]$ and $b_{\mathcal{S}}/b_{\mathcal{H}}\in[1,\ell_K]$.

By Proposition 5.3 (i), the VI threshold of the harmonic averaging method satisfies $b_{\mathcal{H}}(\varepsilon)\sim\varepsilon=b_{\mathcal{S}}$ as $\varepsilon\downarrow 0$. Using Theorem 5.3, we further know that $b_{\mathcal{H}}(\varepsilon)<\varepsilon$ (the inequality is strict since $M_{-1,K}<S_K$ has probability 1 for independent p-variables). Therefore, we cannot directly use the asymptotic VI threshold ε of the harmonic averaging method, which needs to be corrected; see Wilson (2019).

To summarize the results in this section, the Cauchy combining function and the harmonic averaging function are very similar in several senses, and the Simes function is more conservative than the harmonic averaging function. Empirically, we see that the Simes function is only slightly more conservative; see Section 5.7.

5.6 Prices for validity

For a given set of realized p-values, the decision to the hypothesis testing for some specific combining function will be determined by the corresponding threshold. The VAD method can always control the size below the significance level; VSD methods may not have the correct size, but they yield more power than the VAD method. Therefore, there is always a trade-off between validity and efficiency, thus a price for validity.

For a combining function F and K standard uniform random variables $U_1, \dots, U_K$ with some specific dependence assumption (e.g., independence, comonotonicity, or condition (G)), let g_F be the VSD threshold, i.e., $g_F(\varepsilon) = q_\varepsilon(F(U_1, \dots, U_K))$. Let a_F be defined as in (5.2). For some fixed $\varepsilon \in (0,1)$, the ratio $g_F(\varepsilon)/a_F(\varepsilon)$ is called the *price for validity* under the corresponding dependence assumption of the p-variables. For instance, $b_F(\varepsilon)/a_F(\varepsilon)$ is the price paid for validity under independence assumption and $c_F(\varepsilon)/a_F(\varepsilon)$ is the corresponding price under the comonotonicity assumption. For a specific application, one may consider the price for validity under other dependence assumptions. The calculation of the price for validity serves for two purposes:

i (Power gain/loss): On the one hand, if additional information on the dependence structure of the p-values is available, the price for validity can be used as a measure for the gain of power from the dependence information. On the other hand, if the dependence information is not available or credible, the price can be used to measure the power loss by switching to the VAD threshold.

ii (Sensitivity to model misspecification): If the dependence structure is ambiguous, VAD thresholds should be used. A small price for validity indicates that a relatively small change of threshold due to the model ambiguity. Hence, the price for validity can be used as a tool to assess the sensitivity of VSD methods to model misspecification.

Remark 5.6. Instead of using the price for validity, a more direct way to assess the trade-off between using VSD and VAD methods is comparing the sizes, i.e., $\mathbb{P}(F(P_1, \dots, P_K) < g_F(\varepsilon))/\mathbb{P}(F(P_1, \dots, P_K) < a_F(\varepsilon))$, where the dependence of p-variables $P_1, \dots, P_K$ corresponds to the VSD method. More precisely, for a fixed $\varepsilon \in (0,1)$, the ratio of sizes is $\varepsilon / g_F^{-1}(a_F(\varepsilon))$, where g_F^{-1} is the (generalized) inverse of g_F. The connection between the price for validity and the ratio of sizes is explained below.

(i) For the Simes and the Cauchy combination methods, the ratios of sizes under independence and comonotonicity are identical to the corresponding price for validity since b_F and c_F are identity functions.

(ii) For the averaging methods, the ratios of sizes under comonotonicity are identical to the price for validity since c_F is identity. The ratios of sizes under independence may be different from $b_F(\varepsilon)/a_F(\varepsilon)$; however, by letting $\delta = a_F(\varepsilon)$, we have ($a_F$ is strictly increasing in all cases we consider)

$$\frac{\varepsilon}{b_F^{-1}(a_F(\varepsilon))} = \frac{a_F^{-1}(\delta)}{b_F^{-1}(\delta)}.$$

This is very similar to $b_F(\varepsilon)/a_F(\varepsilon)$; it is a matter of looking at the ratio of threshold functions or that of their inverses. In fact, if $r < 0$, by Proposition 5.3, we have,

$$\frac{\varepsilon}{b_F^{-1}(a_F(\varepsilon))} \sim \frac{b_F(\varepsilon)}{a_F(\varepsilon)}, \ \varepsilon \downarrow 0,$$

which suggests that the ratio of sizes is almost the same as the price for validity under independence for small significance levels.

We use the Bonferroni method based on the combining function $F = M_{-\infty,K}$ as an example to illustrate the above idea. Using Proposition 5.2 and noting that $K(1-(1-\varepsilon)^{1/K}) \sim \varepsilon$ as $\varepsilon \downarrow 0$, we obtain that the prices for validity of the Bonferroni method satisfy $c_F(\varepsilon)/a_F(\varepsilon) = K$ for $\varepsilon \in (0,1)$ and $b_F(\varepsilon)/a_F(\varepsilon) \to 1$ as $\varepsilon \downarrow 0$. Therefore, for a small ε close to 0, the price for validity under the independence assumption is close to 1 while the price for validity under the comotonicity assumption increases linearly as the number of p-variables increases. This means a model misspecification of independence is not affecting the Bonferroni method much, whereas a model misspecification of comonotonicity greatly affects the statistical conclusion of the Bonferroni method.

Next we numerically calculate the prices for validity under independence and comonotonicity assumptions for various merging methods using results in Section 5.3. We consider the Bonferroni, the harmonic averaging, the geometric averaging, the Cauchy combination, the Simes, and the negative-quartic (using $M_{-4,K}$, a compromise between Bonferroni and harmonic averaging) methods. The (asymptotic) VAD and VI thresholds of these methods are summarized in Table 5.2. The VC threshold is identity for all these methods. The VAD threshold of the negative-quartic method is given by Proposition 5 of Vovk and Wang (2021). Numerical results on the prices for validity are reported in Table 5.3 for $\varepsilon = 0.01$. Although some of the VAD thresholds in Table 5.2 do not have explicit forms,

Table 5.2: Thresholds for K p-variables at significance level $\varepsilon \in (0,1)$.

	Bonferroni	Negative-quartic	Simes	Cauchy	Harmonic	Geometric
$a_F(\varepsilon)$	ε/K	$\frac{3}{4}K^{-\frac{3}{4}}\varepsilon$	ε/ℓ_K	(5.11)	(5.7)	(5.6)
$b_F(\varepsilon)$	$1-(1-\varepsilon)^{1/K}$	(5.8)	ε	ε	(5.8)	(5.9)

Table 5.3: $b_F(\varepsilon)/a_F(\varepsilon)$ and $c_F(\varepsilon)/a_F(\varepsilon)$ for $\varepsilon = 0.01$ and $K \in \{50, 100, 200, 400\}$

	$K=50$		$K=100$		$K=200$		$K=400$	
	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F
Bonferroni	1.005	50.000	1.005	100.000	1.005	200.000	1.005	400.000
Negative-quartic	1.340	25.071	1.340	42.164	1.340	70.911	1.340	119.257
Simes	4.499	4.499	5.187	5.187	5.878	5.878	6.570	6.570
Cauchy	6.625	6.625	7.465	7.465	8.277	8.277	9.058	9.058
Harmonic	6.658	6.625	7.496	7.459	8.314	8.273	9.117	9.072
Geometric	69.903	2.718	78.096	2.718	84.214	2.718	88.694	2.718

the numerical computation is very fast. The results for $\varepsilon = 0.05$ and $\varepsilon = 0.0001$ are similar and reported in Tables 5.5 and 5.6 in Section 5.9.2.

The Bonferroni and the negative-quartic methods pay much lower price under the independence assumption than the comonotonicity assumption, and the geometric averaging method is the absolute opposite. On the other hand, the harmonic averaging, the Simes and the Cauchy combination methods have relatively small prices under both independence and comonotonicity assumptions and their prices increase at moderate rates as K increases, compared to other methods. In particular, the harmonic averaging and the Cauchy combination methods have very similar performance (cf. Theorem 5.2) and their prices are slightly larger than that of the Simes method. If mild model misspecification exists, it may be safer to choose one of the harmonic averaging, the Simes and the Cauchy combination methods and use the corresponding VAD threshold without losing much power. The prices for validity in Table 5.3 can also be interpreted as inflations of sizes by using VSD threshold against VAD threshold except the geometric averaging method (see Remark 5.6).

Next, we show that the prices for validity of the harmonic averaging, the Cauchy combination and the Simes methods behave like $\log K$ for K large enough and ε small enough.

Proposition 5.5. *For $\varepsilon \in (0,1)$, the prices for validity satisfy:*

(i) For the harmonic averaging method, $F = M_{-1,K}$,

$$\lim_{\delta \downarrow 0} \frac{b_F(\delta)}{a_F(\delta)} = \frac{c_F(\varepsilon)}{a_F(\varepsilon)} \sim \log K, \ \ as\ K \to \infty.$$

(ii) For the Cauchy combination method, $F = M_{\mathcal{C},K}$,

$$\lim_{\delta \downarrow 0} \frac{b_F(\delta)}{a_F(\delta)} = \lim_{\delta \downarrow 0} \frac{c_F(\delta)}{a_F(\delta)} \sim \log K, \ \ as\ K \to \infty.$$

(iii) For the Simes method, $F = S_K$,

$$\frac{b_F(\varepsilon)}{a_F(\varepsilon)} = \frac{c_F(\varepsilon)}{a_F(\varepsilon)} \sim \log K, \ \ as\ K \to \infty.$$

Numerical values of the ratios between the price for validity under independence assumption and $\log K$ are reported in Table 5.4; the results for the corresponding ratios under comonotonicity assumption are similar for these methods. The Simes method has the fastest convergence rate among the three methods. The ratios for the harmonic averaging and the Cauchy combination methods converge quite slowly and have similar rates. This fact can also be explained by Theorem 5.3, where we see that the Simes function is generally larger than the harmonic averaging function.

Based on Proposition 5.5, one may be tempted to use $b_F/\log K$ as the corrected critical value under model misspecification; however, for the harmonic averaging and the Cauchy combination methods, the asymptotic rate of $\log K$ can only be expected for very large K (instead, $1.7 \log K$ works for $K \geqslant 100$).

Table 5.4: Numerical values of $\frac{1}{\log(K)} \frac{b_F(\varepsilon)}{a_F(\varepsilon)}$ for the Simes, the Cauchy combination and the harmonic averaging methods.

	ε	$K = 10$	20	50	100	200	500
Simes	0.05	1.272035	1.200955	1.150097	1.126425	1.109415	1.093041
	0.01	1.272035	1.200955	1.150097	1.126425	1.109415	1.093041
Cauchy	0.05	1.979572	1.82826	1.693025	1.620527	1.561670	1.511264
	0.01	1.980144	1.828822	1.693562	1.621011	1.562121	1.504288
Harmonic	0.05	2.026308	1.873762	1.73641	1.661098	1.601539	1.539448
	0.01	1.989255	1.837605	1.701851	1.627702	1.569179	1.508248

5.7 Simulations and a real data example

5.7.1 Simulation studies

We conduct K one-sided z-tests of the null hypothesis: $\mu_i = 0$ against the alternative hypothesis $\mu_i > 0$, $i = 1, \ldots, K$, using the test statistic X_i and the p-value p_i from the ith test, $i = 1, \ldots, K$. The tests are formulated as the following:

$$p_i = \Phi(X_i), \quad X_i = \rho Z + \sqrt{1-\rho^2} Z_i - \mu_i, \quad i = 1, \ldots, K.$$

where Φ is the standard normal distribution function, $Z, Z_1, \ldots, Z_K$ are iid standard normal random variables, $\mu_i \geqslant 0$, $i = 1, \ldots, K$, and ρ is a parameter in $[0, 1]$. Note that for $\rho = 0$, the p-variables are independent, and $\rho = 1$ corresponds to the case where p-variables are comonotonic.

Let $K \in \{50, 200\}$ and set the significance level $\varepsilon = 0.01$. To see how different dependence structures and signals affect the size and the power for various methods using both VAD and VSD thresholds, the rejection probabilities (RPs) are computed over $\rho \in [0, 1]$ under the following four cases:

(i) (no signal) 100% of μ_i's are 0;
(ii) (needle in a haystack) 98% of μ_i's are 0 and 2% of μ_i's are 4;
(iii) (sparse signal) 90% of μ_i's are 0 and 10% of μ_i's are 3;
(iv) (dense signal) 100% of μ_i's are 2.

The RP corresponds to the size under case (i), and it corresponds to the power under (ii), (iii) and (iv). The RP is computed as the ratio between the number of the combined values which are less than the critical threshold and the number of simulations for some $\rho \in [0, 1]$, that is,

$$\text{RP} = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{F_i < g(\varepsilon)\}}}{N},$$

where N is the number of simulations and is equal to 15000 in our study, F_i is the realized value of the combining function for the i-th simulation, $i = 1, \ldots, N$, and $g(\varepsilon)$ is the corresponding critical value. For $\rho \in [0, 1]$, graphs of RPs for different combining methods are drawn using VAD thresholds and VSD thresholds. Some observations from Figures 5.1-5.4 are made below, and those on the averaging methods using $M_{r,K}$ are consistent with the observations in Vovk and Wang (2020).

1. All VAD methods give sizes less than $\varepsilon = 0.01$ as expected. Using VAD thresholds, the Bonferroni, the harmonic averaging, the Cauchy combination and the Simes methods have good powers.

2. The Simes method using thresholds b_F or c_F reports the right size for all values of ρ. Sarkar (1998) showed the validity of the Simes method in the so-called MTP_2 class including multivariate normal distributions with nonnegative correlations (the setting of our simulation).

3. Using thresholds b_F or c_F, the harmonic averaging and Cauchy combination methods perform similarly with sizes possibly larger than 0.01 (see Theorems 5.2 and 5.3).

4. The geometric averaging method using b_F and the Bonferroni and negative-quartic methods using c_F do not yield correct sizes under model misspecification, and the sizes increase rapidly as the misspecification gets bigger.

5. Using b_F or c_F, the harmonic averaging, the Cauchy combination and the Simes methods have good performances on capturing the signals.

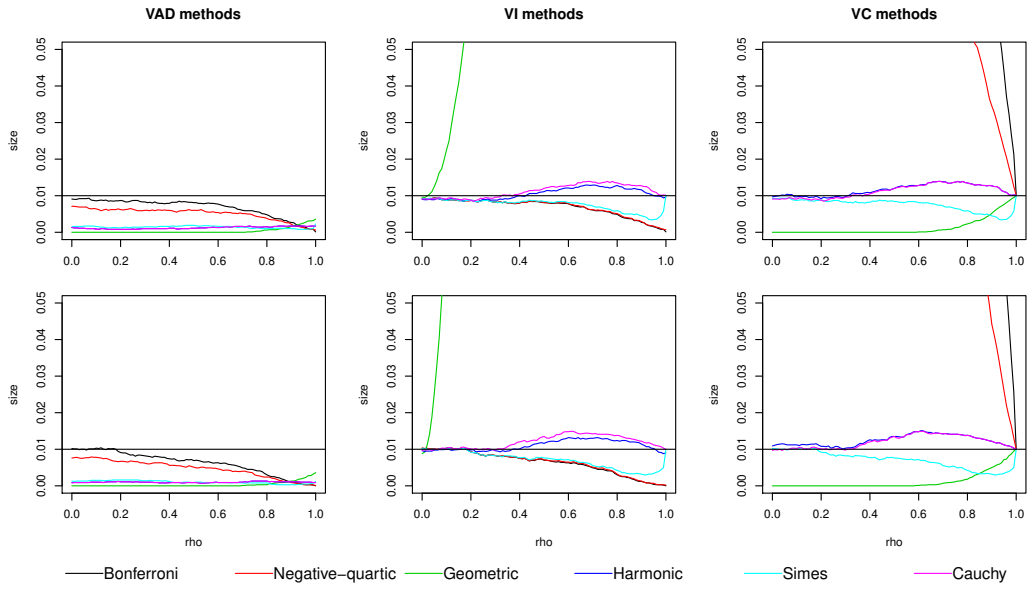


Figure 5.1: Case (i): size (top: $K = 50$, bottom: $K = 200$)

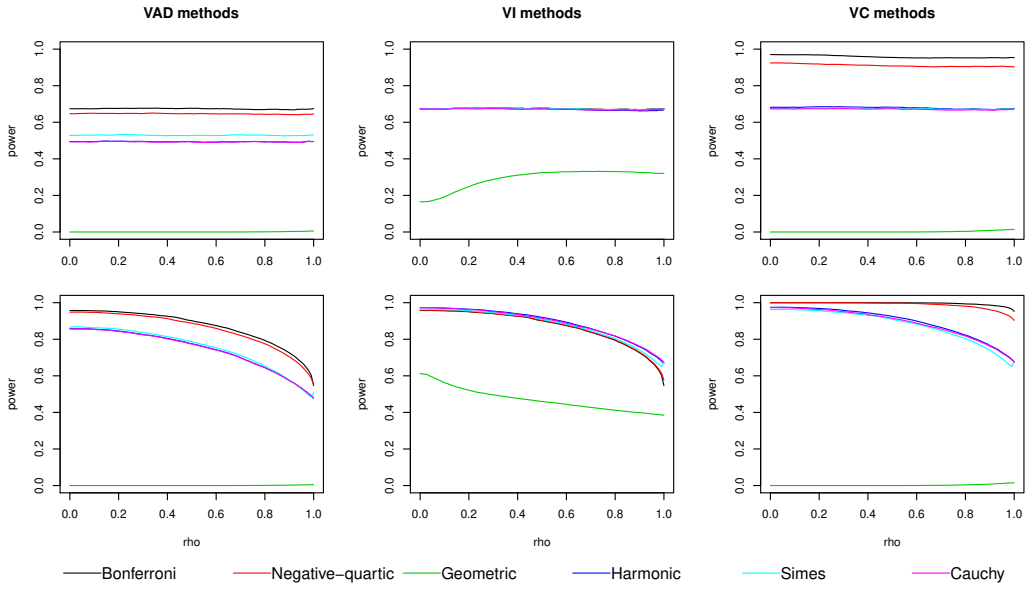


Figure 5.2: Case (ii): needle in a haystack (top: $K = 50$, bottom: $K = 200$)

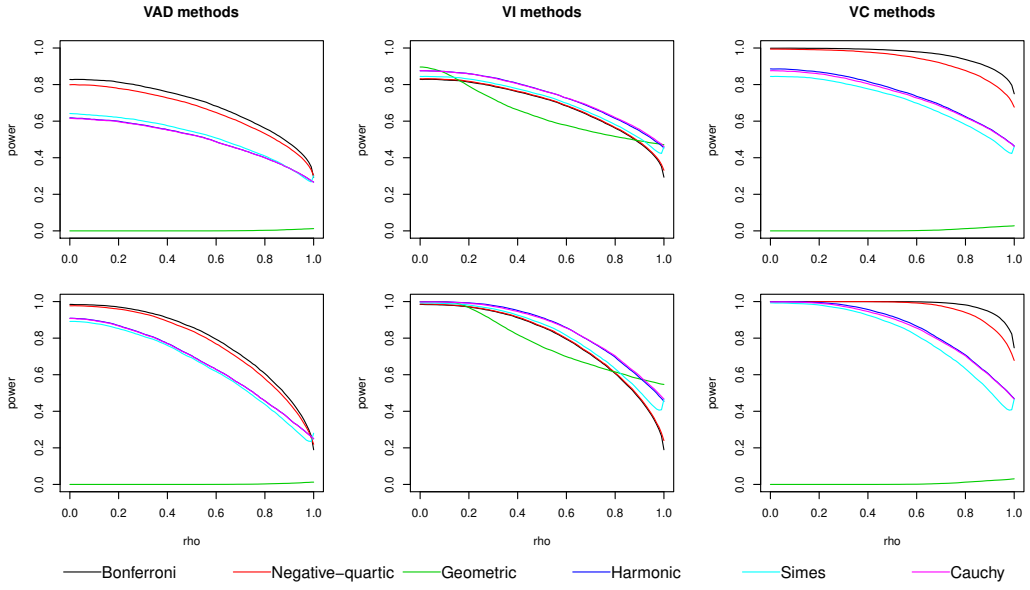


Figure 5.3: Case (iii): sparse signal (top: $K = 50$, bottom: $K = 200$)

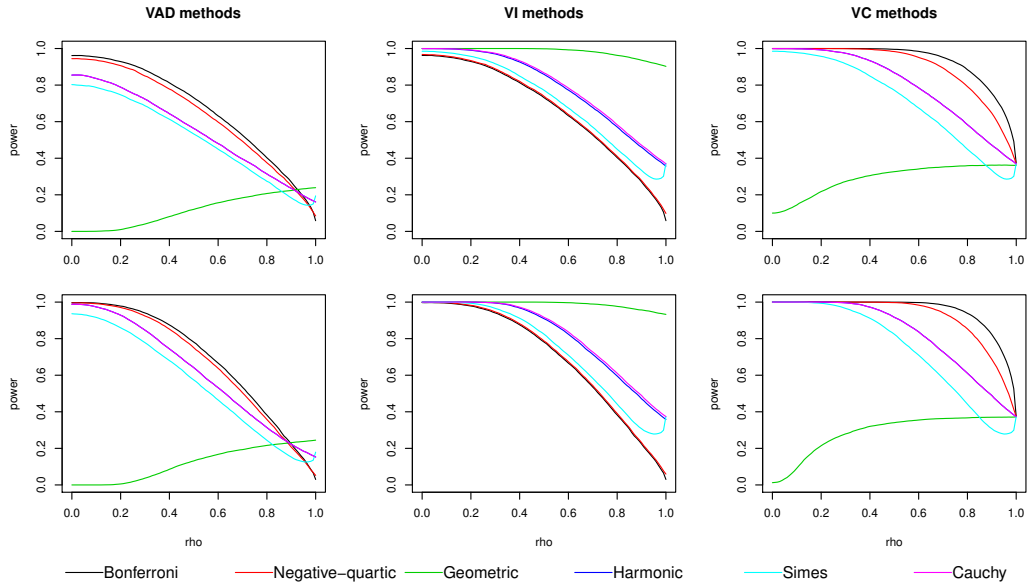


Figure 5.4: Case (iv): dense signal (top: $K = 50$, bottom: $K = 200$)

5.7.2 Real data analysis

We apply several merging methods to a genomewide study to compare their performances. We use the dataset of p-values of Storey and Tibshirani (2003) which contains 3170 p-values computed based on the data from Hedenfalk et al. (2001) for testing whether genes are differentially expressed between BRCA1- and BRCA2-mutation-positive tumors. As mentioned in Section 5.2, $g^{-1} \circ F(P_1, \ldots, P_K)$ is a p-variable if the threshold g is strictly increasing, and it is the quantity we choose to compare combined p-values for different methods.

For each method, we calculate the combined p-value, and remove the smallest p-value from the dataset. Repeat this procedure until the resulting combined p-value loses significance. Using the Bonferroni combining function, this leads to the Bonferroni-Holm (BH) procedure (Holm (1979)); thus we mimic the BH procedure for other methods in a naive manner. The rough interpretation is to report the number of significant discoveries (this procedure generally does not control the family-wise error rate (FWER); to control FWER one needs to use a generalized BH procedure as in Vovk and Wang (2020) or Goeman et al. (2019). This procedure can be seen as a lower confidence bound from a closed testing perspective). For a visual comparison of detection power, the combined p-values against the numbers of removed p-values are plotted in Figure 5.5, where we use both the VAD and the

VI thresholds (comonotonicity is obviously unrealistic here). In the third panel of Figure 5, we present the number of omitted p-values in log-scale for better visualization.

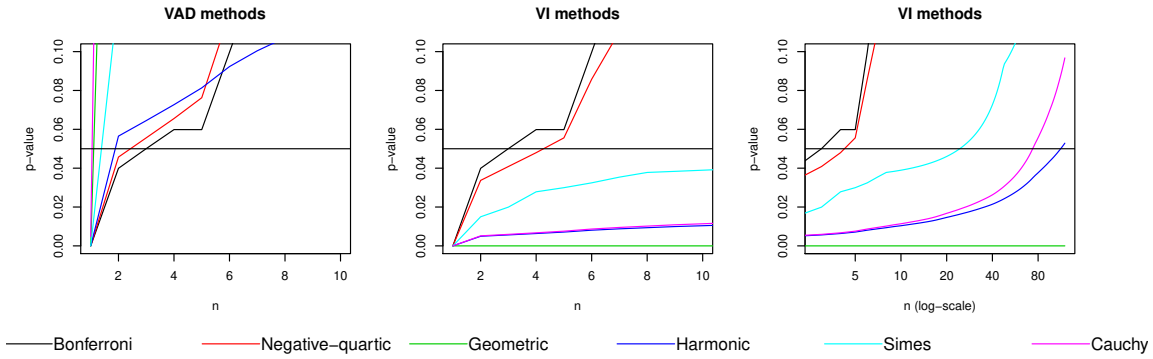


Figure 5.5: Combined p-value after removing n smallest p-values

All VAD methods lose significance at $\varepsilon = 0.05$ after omitting the first or the second smallest p-value (the smallest p-value is 0 and the second smallest is 1.26×10^{-5}). Using thresholds b_F for independence, the Bonferroni and the negative quartic methods behave similarly to their VAD versions (as their price for validity is close to 1). In contrast, the Simes, the Cauchy combination and the harmonic averaging methods lose significance at $\varepsilon = 0.05$ after removing around 20, 70 and 110 p-values respectively. The geometric averaging method (Fisher's) exceeds 0.05 only after removing around 400 p-values. However, this method relies heavily on the independence assumption, which is impossible to verify from just one set of p-values.

5.8 Concluding remarks

We discussed two aspects of merging p-values: the impact of the dependence structure on the critical thresholds and the trade-off between validity and efficiency. The Cauchy combination method and the Simes method are shown to be the only IC-balanced members among the generalized mean class and the order statistics class of combining functions. The harmonic averaging and the Cauchy combination methods are asymptotically equivalent, and the Simes and the harmonic averaging methods have simple algebraic relationship. For the above three methods, the prices for validity under independence (comonotonicity) assumption all behaves like $\log K$ for large K. Moreover, these methods lose moderate

amount of power if VAD thresholds are used, and their performance against model misspecification is better than other methods. This explains the wide applications of these methods in different statistical procedures.

Merging p-values is not only useful for testing a single hypothesis, but also important in testing multiple hypotheses, controlling false discovery rate (Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001)), and exploratory research (Goeman and Solari (2011), Goeman et al. (2019)). In many situations especially involving a large number of hypotheses and tests, dependence information is hardly available. The results in our chapter offer some insights, especially in terms of gain/loss of validity and power, on how the absence of such information influences different statistical procedures of merging p-values.

In many practical applications, p-values arrive sequentially in time, and the existence of the n-th p-variable may depend on previously observed p-values (only promising experiments may be continued); thus the number of experiments to combine is a stopping time. Unfortunately, the current merging method of p-values discussed in this chapter cannot be used to sequentially update p-values with arbitrary stopping rule. To deal with such a situation, one has to rely on anytime-valid methods, typically through the use of a test supermartingale (see Howard et al. (2021) and Ramdas et al. (2020)) or through e-values (see Shafer (2021) and Vovk and Wang (2021)). Moreover, e-values are nicer to combine (e.g., using average and product as in Vovk and Wang (2021)) especially under arbitrary dependence, in contrast to the complicated methods of merging p-values.

R code

An `R` package `pmerge` for various merging methods in this chapter is available at https://github.com/YuyuChen-UW/pmerge.

5.9 Appendix

5.9.1 Proofs of theorems and propositions in Chapter 5

Proof of Proposition 5.1. By definition, we have

$$a_F(\varepsilon) = \inf\{q_\varepsilon(F(U_1, \dots, U_K)) \mid U_1, \dots, U_K \in \mathcal{U}\},\ \varepsilon \in (0, 1).$$

We shall show

$$a_F(\varepsilon) = \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\},\ \varepsilon \in (0,1), \tag{5.14}$$

where $\mathcal{U}_\varepsilon$ denotes the collection of all uniform random variables distributed on $[0,\varepsilon]$. Denote by $S = F(U_1,\dots,U_K)$ and $G_S^{-1}(t) = q_t(S),\ t \in (0,1]$. We can find $U_S \in \mathcal{U}$ such that $G_S^{-1}(U_S) = S$ a.s. (e.g., Lemma A.32 of Föllmer and Schied (2016)). Let $f_i(t) = \mathbb{P}\left(U_i \leqslant t | U_S < \varepsilon\right),\ t \in [0,1]$. Then $f_i(U_i)$ conditionally on $U_S < \varepsilon$ is a uniform random variable on $[0,1]$ and $V_i^\varepsilon := \varepsilon f_i(U_i)$ conditionally on $U_S < \varepsilon$ is a uniform random variable on $[0,\varepsilon]$. We construct the following two random variables:

$$S_1 = S\mathbb{1}_{\{U_S<\varepsilon\}} + d\mathbb{1}_{\{U_S\geqslant\varepsilon\}},\ S_2 = F(V_1^\varepsilon,\dots,V_n^\varepsilon)\mathbb{1}_{\{U_S<\varepsilon\}} + d\mathbb{1}_{\{U_S\geqslant\varepsilon\}}, \tag{5.15}$$

where $d > F(\varepsilon,\dots,\varepsilon)$. Noting the fact that $\varepsilon f_i(t) = \mathbb{P}(U_i \leqslant t, U_S < \varepsilon) \leqslant t,\ t \in [0,1]$ and F is increasing, we have $S_1 \geqslant S_2$. Hence $q_\varepsilon(S_1) \geqslant q_\varepsilon(S_2)$. Moreover, direct calculation shows $q_\varepsilon(S) = q_\varepsilon(S_1)$. Thus $q_\varepsilon(S) \geqslant q_\varepsilon(S_2)$. Let $\hat{V}_1,\dots,\hat{V}_n$ be uniform random variables on $[0,\varepsilon]$ such that $(\hat{V}_1,\dots,\hat{V}_n)$ has the joint distribution identical to the conditional distribution of $(V_1^\varepsilon,\dots,V_n^\varepsilon)$ on $U_S < \varepsilon$. Hence, for $x < d$,

$$\begin{aligned}\mathbb{P}(S_2 \leqslant x) &= \mathbb{P}(F(V_1^\varepsilon,\dots,V_n^\varepsilon) \leqslant x, U_S < \varepsilon)\\ &= \varepsilon\mathbb{P}(F(V_1^\varepsilon,\dots,V_n^\varepsilon) \leqslant x | U_S < \varepsilon)\\ &= \varepsilon\mathbb{P}(F(\hat{V}_1,\dots,\hat{V}_n) \leqslant x).\end{aligned}$$

This implies $q_\varepsilon(S_2) = q_1(F(\hat{V}_1,\dots,\hat{V}_n))$. Thus we have

$$a_F(\varepsilon) \geqslant \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\}.$$

We next show "$\leqslant$" in (5.14). Take $V_1,\dots,V_n \in \mathcal{U}_\varepsilon$ and $U \in \mathcal{U}$ such that U is independent of $V_1,\dots,V_n$. Let $\hat{U}_i = V_i\mathbb{1}_{\{U<\varepsilon\}} + U\mathbb{1}_{\{U\geqslant\varepsilon\}},\ i = 1,2,\dots,n$. It is clear that $\hat{U}_i \in \mathcal{U},\ i = 1,2,\dots,n$ and $F(\hat{U}_1,\dots,\hat{U}_n) = F(V_1,\dots,V_n)\mathbb{1}_{\{U<\varepsilon\}} + F(U,\dots,U)\mathbb{1}_{\{U\geqslant\varepsilon\}}$. Noting that F is increasing, we have $q_1(F(V_1,\dots,V_n)) = q_\varepsilon(F(\hat{U}_1,\dots,\hat{U}_n))$. This implies

$$a_F(\varepsilon) \leqslant \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\}.$$

Therefore, (5.14) holds. By (5.14) and the homogeneity of F we have that for $\varepsilon \in (0,1)$,

$$\begin{aligned}a_F(\varepsilon) &= \inf\{q_1(F(V_1,\dots,V_K)) \mid V_1,\dots,V_K \in \mathcal{U}_\varepsilon\}\\ &= \inf\{q_1(F(\varepsilon U_1,\dots,\varepsilon U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\}\\ &= \varepsilon\inf\{q_1(F(U_1,\dots,U_K)) \mid U_1,\dots,U_K \in \mathcal{U}\}.\end{aligned}$$

This completes the proof. □

Proof of Proposition 5.2. It is well known that the Bonferroni correction yields $a_F(\varepsilon) = \varepsilon/K$. Also, since the average of identical objects is itself, $c_F(\varepsilon) = \varepsilon$ for any averaging method, including the Bonferroni method. For iid standard uniform random variables $V_1, \ldots, V_K$, we have $\mathbb{P}(\min\{V_1, \ldots, V_K\} \leqslant x) = 1 - (1-x)^K$. Therefore, $b_F(\varepsilon) = 1 - (1 - \varepsilon)^{1/K}$ for $\varepsilon \in (0, 1)$. □

Proof of Proposition 5.3. (a) Suppose $r < 0$. We first fix K and find the asymptotic of b_r as $\varepsilon \downarrow 0$ satisfying

$$\mathbb{P}\left(\sum_{i=1}^{K} P_i^r \geqslant K\left(b_r(\varepsilon)\right)^r\right) = \varepsilon.$$

Observe that the random variables P_i^r, $i = 1, \ldots, K$, follow a common Pareto distribution with cdf $\mathbb{P}(P_i^r \leqslant x) = 1 - x^{1/r}$, $x \in (1, \infty)$, $i = 1, \ldots, K$. Note that the tail probability of the sum of iid Pareto random variables is asymptotically the same as that of the maximum of the iid Pareto random variables (e.g., Embrechts et al. (1997), Corollary 1.3.2). Hence

$$\lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}\left(\sum_{i=1}^{K} P_i^r \geqslant K\left(b_r(\varepsilon)\right)^r\right)}{\mathbb{P}\left(\max\{P_1^r, \ldots, P_K^r\} > K\left(b_r(\varepsilon)\right)^r\right)} = \lim_{\varepsilon \downarrow 0} \frac{\varepsilon}{1 - \left(1 - K^{\frac{1}{r}} b_r(\varepsilon)\right)^K} = 1.$$

This implies

$$b_r(\varepsilon) \sim \frac{1 - (1-\varepsilon)^{\frac{1}{K}}}{K^{\frac{1}{r}}} \sim K^{-1-1/r}\varepsilon, \qquad \text{as } \varepsilon \downarrow 0.$$

The case $K \to \infty$ follows directly from the generalized central limit theorem (e.g., Theorem 1.8.1 of Samorodnitsky (2017)).

(b) If $r = 0$, in a similar way, we first have,

$$\mathbb{P}\left(2\sum_{i=1}^{K} \log\frac{1}{P_i} \geqslant 2K \log\frac{1}{b_r(\varepsilon)}\right) = \varepsilon.$$

The random variable $\log\frac{1}{P_i}$, $i = 1, \ldots, K$, follows exponential distribution with parameter 1. Thus $2\sum_{i=1}^{K} \log\frac{1}{P_i}$ follows a chi-square distribution with parameter $2K$. We denote $q_\alpha(\chi^2_\nu)$ the α-quantile of the chi-square distribution with ν degrees of freedom. Hence

$$b_r(\varepsilon) = \exp\left(-\frac{1}{2K} q_{1-\varepsilon}\left(\chi^2_{2K}\right)\right).$$

(c) If $r > 0$, using the result of Wang (2005), we have for $0 \leqslant x \leqslant K^{-r}$,

$$\begin{aligned}\mathbb{P}\left(M_{r,K}(U_1,\ldots,U_K) \leqslant x\right) &= \mathbb{P}\left(\sum_{i=1}^{K} U_i^r \leqslant Kx^r\right) \\ &= \lambda\left\{(x_1,\ldots,x_K) : \sum_{i=1}^{K} x_i^r \leqslant Kx^r,\ x_1,\ldots,x_K \geqslant 0\right\} \\ &= \frac{(\Gamma(1+1/p))^K}{\Gamma(1+K/p)} K^{K/r} x^K,\end{aligned}$$

where λ is the Lebesgue measure. This implies that if $\varepsilon \leqslant \frac{(\Gamma(1+1/p))^K}{\Gamma(1+K/p)}$,

$$b_r(\varepsilon) = \frac{(\Gamma(1+K/p))^{1/K}\varepsilon^{1/K}}{K^{1/r}\Gamma(1+1/p)}. \tag{5.16}$$

The asymptotic behaviour of $b_r(\varepsilon)$ for fixed $\varepsilon \in (0,1)$ as $K \to \infty$ can be obtained by the Central Limit Theorem. Note that the random variables P_i^r, $i = 1,\ldots,K$, follow a common Beta distribution with mean and variance given by, respectively,

$$\mu = (r+1)^{-1}, \text{ and } \sigma^2 = r^2(1+2r)^{-1}(1+r)^{-2}.$$

The Central Limit Theorem gives $(\sum_{i=1}^{K} P_i^r - K\mu)/\sqrt{K}\sigma \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$. Hence

$$b_r(\varepsilon) \sim \left(\frac{\sigma}{\sqrt{K}}\Phi^{-1}(\varepsilon) + \mu\right)^{\frac{1}{r}}, \text{ as } K \to \infty,$$

where Φ^{-1} is the inverse of the standard normal distribution function. □

Proof of Proposition 5.4. By symmetry of the standard Cauchy distribution,

$$\begin{aligned}a_F(\varepsilon) &= \mathcal{C}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)\right) \mid U_1,\ldots,U_K \in \mathcal{U}\right\}\right) \\ &= \mathcal{C}\left(\frac{-1}{K}\sup\left\{q_{1-\varepsilon}\left(\sum_{i=1}^{K}\mathcal{C}^{-1}(U_i)\right) \mid U_1,\ldots,U_K \in \mathcal{U}\right\}\right).\end{aligned}$$

Moreover, $\mathcal{C}^{-1}(U_i)$, $i = 1,\ldots,K$, follow the standard Cauchy distribution with decreasing density on $[\mathcal{C}^{-1}(1-\varepsilon),\infty]$ for $\varepsilon \in (0,1/2)$. The proposition follows directly from applying Corollary 3.7 of Wang et al. (2013). □

Proof of Theorem 5.1. (i) IC-balance of $M_{\phi,K}$ for all $K \in \{2,3,\dots\}$ is equivalent to $\frac{1}{K}\sum_{i=1}^{K}\phi(V_i) \overset{\mathrm{d}}{=} \phi(U)$ for all $K \in \{2,3,\dots\}$, which is further equivalent to the fact that $\phi(U)$ follows a strictly 1-stable distribution. We know that strictly 1-stable distributions are Cauchy distributions (see, e.g., Theorem 14.15 of Sato (1999)). This proves the statement of part (i).

(ii) For the Simes function $S_{\alpha,K} = S_K$, $\alpha_i = i$ for $i \in \{1,\dots,K\}$ and $b_F(x) = c_F(x) = x$ for $x \in [0,1]$. Therefore, $S_{\alpha,K}$ is IC-balanced.

Below we show the opposite direction of the statement. For $n \in \{2,\dots,K\}$, let $V_{(1)},\dots,V_{(n)}$ be the order statistics for n independent standard uniform random variables $V_1,\dots,V_n$. Let $(X_1,\dots,X_{n-1}) = (V_{(1)}/V_{(n)},\dots,V_{(n-1)}/V_{(n)})$ which is identically distributed as the order statistics for $n-1$ independent standard uniform random variables, independent of $V_{(n)}$. Hence, for $x \in (0, 1/\alpha_n)$,

$$\begin{aligned}
&\mathbb{P}\left(S_{\alpha,n}(V_1,\dots,V_n) > x\right) \\
&= \mathbb{P}\left(V_{(1)} > x\alpha_1,\dots,V_{(n-1)} > x\alpha_{n-1}, V_{(n)} > x\alpha_n\right) \\
&= \mathbb{P}\left(X_1 > x\alpha_1/V_{(n)},\dots,X_{n-1} > x\alpha_{n-1}/V_{(n)}, V_{(n)} > x\alpha_1\right) \\
&= \int_{x\alpha_n}^{1} \mathbb{P}\left(X_1 > x\alpha_1/p,\dots,X_{n-1} > x\alpha_{n-1}/p\right) np^{n-1}\mathrm{d}p \\
&= \int_{x\alpha_n}^{1} \mathbb{P}\left(S_{\alpha,n-1}(V_1,\dots,V_{n-1}) > x/p\right) np^{n-1}\mathrm{d}p, \qquad (5.17)
\end{aligned}$$

where for simplicity we use $S_{\alpha,n-1}$ for $S_{(\alpha_1,\dots,\alpha_{n-1}),n-1}$. Note that

$$\mathbb{P}\left(S_{\alpha,1}(V_1) > x\right) = 1 - \alpha_1 x, \quad x \in (0, 1/\alpha_1). \qquad (5.18)$$

Plugging (5.18) in (5.17), we obtain that $\mathbb{P}\left(S_{\alpha,2}(V_1,V_2) > x\right)$ is a polynomial function of x of degree less than or equal to 2. Recursively, using (5.17) we are able to show that the function $\mathbb{P}\left(S_{\alpha,n}(V_1,\dots,V_n) > x\right)$ for $x \in (0,1/\alpha_n)$ is a polynomial of x of degree less than or equal to n for $n = 2,\dots,K$. Hence, there exist K constants $\beta_0,\dots,\beta_{K-1}$ such that

$$\mathbb{P}\left(S_{\alpha,K-1}(V_1,\dots,V_{K-1}) > x\right) = \sum_{i=0}^{K-1}\beta_i x^i, \quad x \in (0,1/\alpha_{K-1}).$$

Moreover, noting that $S_{\alpha,K}$ is IC-balanced, we have

$$\int_{x\alpha_K}^{1} \mathbb{P}\left(S_{\alpha,K-1}(V_1,\dots,V_{K-1}) > x/p\right) Kp^{K-1}\mathrm{d}p = \mathbb{P}\left(S_{\alpha,K}(U,\dots,U) > x\right) = 1 - x\alpha_K,$$

for $x \in (0, 1/\alpha_K)$. Therefore, we have

$$\int_{x\alpha_K}^{1} \left(\sum_{i=0}^{K-1} \beta_i x^i p^{-i} \right) K p^{K-1} \mathrm{d}p = 1 - x\alpha_K,$$

which implies that for $x \in (0, 1/\alpha_K)$,

$$\sum_{i=0}^{K-1} \frac{K\beta_i}{K-i} x^i - \left(\sum_{i=0}^{K-1} \frac{K\beta_i}{K-i} \alpha_K^{K-i} \right) x^K = 1 - x\alpha_K.$$

Solving the above equation, we get $\beta_0 = 1$, $\beta_1 = -\frac{K-1}{K}\alpha_K$ and $\beta_2 = \cdots = \beta_{K-1} = 0$. Consequently,

$$\mathbb{P}\left(S_{\alpha,K-1}(V_1, \ldots, V_{K-1}) > x\right) = 1 - \frac{K-1}{K}\alpha_K x, \quad x \in (0, 1/\alpha_{K-1}).$$

Recursively, using (5.17) we have

$$\mathbb{P}\left(S_{\alpha,n}(V_1, \ldots, V_n) > x\right) = 1 - \frac{n}{K}\alpha_K x, \quad x \in (0, 1/\alpha_n) \tag{5.19}$$

for $n = 1, \ldots, K$, which gives, using (5.18),

$$\alpha_K = K\alpha_1. \tag{5.20}$$

Inserting (5.19) into (5.17), we obtain, for $x \in (0, 1/\alpha_n)$ and $n = 2, \ldots, K$,

$$\begin{aligned} 1 - \frac{n}{K}\alpha_K x &= \int_{x\alpha_n}^{1} \left(1 - \frac{n-1}{K}\alpha_K x p^{-1}\right) n p^{n-1} \mathrm{d}p \\ &= 1 - \frac{n}{K}\alpha_K x + \left(\frac{n}{K}\alpha_K \alpha_n^{n-1} - \alpha_n^n\right) x^n. \end{aligned}$$

Consequently,

$$\alpha_n = \frac{n}{K}\alpha_K, \quad n = 2, \ldots, K,$$

which together with (5.20) implies $\alpha_n = n\alpha_1$, $k = 1, \ldots, K$. This gives the desired statement. □

In the following example, we shall employ several theorems from Sato (1999). To make this chapter more self-contained, we display the useful part of these theorems as below.

Theorem 8.1 in Sato (1999): μ is an infinitely divisible distribution in $\mathbb{R}$ if and only if there exist $d \geqslant 0$, $\gamma \in \mathbb{R}$ and a measure ν on $\mathbb{R}$ satisfying $\nu(\{0\}) = 0$ and $\int_{\mathbb{R}}(|x|^2 \wedge 1)\nu(\mathrm{d}x) < \infty$, such that the characteristic function of μ is

$$\hat{\mu}(z) = \exp\left(-\frac{1}{2}dz^2 + i\gamma z + \int_{\mathbb{R}}(e^{izx} - 1 - izx\mathbb{1}_{[-1,1]}(x))\nu(\mathrm{d}x)\right), \ z \in \mathbb{R}, \tag{5.21}$$

where $\mathbb{1}_{[-1,1]}(\cdot)$ is the indicator function and $i^2 = -1$.

Theorem 27.16 in Sato (1999): Suppose μ satisfies (5.21). If $d = 0$ and ν is discrete with total measure infinite, then μ is a continuous distribution.

Example 5.1 (IC-balanced generalized mean for a finite K)**.** We show that IC-balance of $M_{\phi,K}$ for a finite K does not imply $M_{\phi,K}$ that ϕ is the Cauchy quantile function (up to an affine transform). For this purpose, we construct a continuous distribution μ such that

$$\frac{1}{K}\sum_{i=1}^{K} X_i \overset{\mathrm{d}}{=} X, \tag{5.22}$$

where X and $X_i, i = 1, \ldots, K$ are iid random variables with distribution μ, but μ is not a Cauchy distribution. Define

$$\hat{\mu}(z) = \exp\left(\int_{\mathbb{R}}\left(e^{izx} - 1 - \mathbb{1}_{[-1,1]}(x)\right)\nu(\mathrm{d}x)\right), \ z \in \mathbb{R},$$

where ν is a symmetric measure on $\mathbb{R} \setminus \{0\}$ satisfying

$$\nu(\{K^n\}) = \nu(\{-K^n\}) = K^{-n}, \ n \in \mathbb{Z}, \text{ and } \nu\left(\mathbb{R} \setminus \left(\{0\} \cup \bigcup_{n\in\mathbb{Z}}\{K^n, -K^n\}\right)\right) = 0.$$

It follows from Theorem 8.1 of Sato (1999) that $\hat{\mu}$ is the characterization function of some infinitely divisible distribution μ. Also noting that $\nu(\mathbb{R} \setminus \{0\}) = \infty$, by Theorem 27.16 of Sato (1999) we know that μ is a continuous distribution. By Theorem 14.7 of Sato (1999), $(\hat{\mu}(z))^b = \hat{\mu}(bz), \ z \in \mathbb{R}, b > 0$ holds if and only if

$$T_b\nu(B) = b\nu(B), \text{ and } \int_{1<|x|\leqslant b} x\nu(\mathrm{d}x) = 0,$$

where $T_b\nu(B) = \nu(b^{-1}B)$ for all Borel sets $B \subset \mathbb{R}$. By symmetry of ν, $\int_{1<|x|\leqslant b} x\nu(\mathrm{d}x) = 0$ holds for any $b > 0$. However, $T_b\nu(B) = b\nu(B)$ holds only for $b \in \{K^n, n \in \mathbb{Z}\}$. Consequently, $(\hat{\mu}(z))^b = \hat{\mu}(bz),\ z \in \mathbb{R}$ if and only if $b \in \{K^n, n \in \mathbb{Z}\}$. This implies that μ is not a Cauchy distribution (strictly 1-stable distribution) but (5.22) holds.

Proof of Theorem 5.2. (i) Recall that

$$\mathcal{C}^{-1}(x) = \tan\left(-\frac{\pi}{2} + \pi x\right), \quad x \in (0,1);$$
$$\mathcal{C}(y) = \frac{1}{\pi}\arctan(y) + \frac{1}{2}, \quad y \in \mathbb{R}.$$

Note that $\mathcal{C}^{-1}(x) \sim -1/(\pi x)$ as $x \downarrow 0$ and $\mathcal{C}(y) \sim -1/(\pi y)$ as $y \to -\infty$. For any $\delta_1, \delta_2 \in (0, 1/K)$, there exists $0 < \varepsilon < 1$ and $m < 0$ such that for all $x \in (0, \varepsilon)$ and $y \in (-\infty, m)$,

$$-\frac{(1+\delta_1)}{\pi x} \leqslant \mathcal{C}^{-1}(x) \leqslant -\frac{(1-\delta_1)}{\pi x}; \tag{5.23}$$

$$-\frac{(1-\delta_2)}{\pi y} \leqslant \mathcal{C}(y) \leqslant -\frac{(1+\delta_2)}{\pi y}. \tag{5.24}$$

For $0 < c < 1$, there exists $0 < \varepsilon' < \varepsilon$ such that

$$\sup_{x\in[\varepsilon,c]} \left|\tan\left(-\frac{\pi}{2} + \pi x\right) + \frac{1}{\pi x}\right| \leqslant \frac{\delta_1}{\pi\varepsilon'}. \tag{5.25}$$

Take $(p_1, \ldots, p_K)$ such that $p_{(1)} < \varepsilon'$ and $p_{(K)} \leqslant c < 1$. Let $l = \max\{i = 1, \ldots, K : p_{(i)} < \varepsilon\}$. As a consequence of (5.23), we have

$$-\sum_{i=1}^{l} \frac{(1+\delta_1)}{\pi p_{(i)}} \leqslant \sum_{i=1}^{l} \tan\left(-\frac{\pi}{2} + \pi p_{(i)}\right) \leqslant -\sum_{i=1}^{l} \frac{(1-\delta_1)}{\pi p_{(i)}}.$$

For $j > l$, (5.25) implies

$$\left|\tan\left(-\frac{\pi}{2} + \pi p_{(j)}\right) + \frac{1}{\pi p_{(j)}}\right| \leqslant \frac{\delta_1}{\pi\varepsilon'} \leqslant \frac{\delta_1}{\pi p_{(1)}}.$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{K}\tan\left(-\frac{\pi}{2}+\pi p_i\right) &\leqslant -\sum_{i=1}^{l}\frac{(1-\delta_1)}{\pi p_{(i)}}-\sum_{i=l+1}^{K}\frac{1}{\pi p_{(i)}}+\frac{(K-l)\delta_1}{\pi p_{(1)}}\\ &\leqslant -\sum_{i=1}^{K}\frac{(1-K\delta_1)}{\pi p_{(i)}}\\ &= -\sum_{i=1}^{K}\frac{(1-K\delta_1)}{\pi p_i}.\end{aligned}$$

Similarly, we can show

$$\sum_{i=1}^{K}\tan\left(-\frac{\pi}{2}+\pi p_i\right)\geqslant\sum_{i=1}^{K}-\frac{(1+K\delta_1)}{\pi p_i}.$$

Using (5.24), for any $(p_1,\ldots,p_K)$ satisfying $p_{(1)}<\min(\varepsilon',\frac{K\delta_1-1}{K\pi m})$ and $p_{(K)}\leqslant c<1$,

$$\frac{1-\delta_2}{1+K\delta_1}M_{-1,K}(p_1,\ldots,p_K)\leqslant M_{\mathcal{C},K}(p_1,\ldots,p_K)\leqslant\frac{1+\delta_2}{1-K\delta_1}M_{-1,K}(p_1,\ldots,p_K).$$

We establish the claim by letting $\delta_1,\delta_2\downarrow 0$, and the above inequalities hold as long as $p_{(1)}$ is sufficiently small.

(ii) The statement

$$\mathbb{P}\left(M_{\mathcal{C},K}(U_1,\ldots,U_K)<\varepsilon\right)\sim\varepsilon\quad\text{as }\varepsilon\downarrow 0$$

follows directly from Theorem 1 of Liu and Xie (2020) by noting that standard Cauchy distribution is symmetric at 0. Below we show $\mathbb{P}\left(M_{-1,K}(U_1,\ldots,U_K)<\varepsilon\right)\sim\varepsilon$ as $\varepsilon\downarrow 0$, based on similar techniques as in Theorem 1 of Liu and Xie (2020). Observe that

$$\mathbb{P}\left(M_{-1,K}(U_1,\ldots,U_K)<\varepsilon\right)=\mathbb{P}\left(\frac{1}{K}\sum_{i=1}^{K}U_i^{-1}>1/\varepsilon\right).$$

Condition (G) means that for any $1\leqslant i<j\leqslant K$, $(\Phi^{-1}(U_i),\Phi^{-1}(U_j))$ is a bivariate normal random variable with $\operatorname{cov}(\Phi^{-1}(U_i),\Phi^{-1}(U_j))=\sigma_{ij}$, where Φ is the standard normal distribution function and Φ^{-1} is its inverse. Clearly, $\sigma_{ij}=1$ implies that

$U_i = U_j$ a.s. In this case we can combine them in one and the corresponding coefficient becomes $2/K$. Thus, it suffices to prove the stronger statement

$$\mathbb{P}\left(\sum_{i=1}^{K} w_i U_i^{-1} > 1/\varepsilon\right) \sim \varepsilon, \text{ as } \varepsilon \downarrow 0, \tag{5.26}$$

where $w_i > 0,\ i = 1, \ldots, K$, $\sum_{i=1}^{K} w_i = 1$ and $\sigma_{ij} < 1,\ i, j = 1, \ldots, K$. We choose some positive constant δ_ε depending on ε, such that $\delta_\varepsilon \to 0$ and $\delta_\varepsilon/\varepsilon \to \infty$ as $\varepsilon \downarrow 0$. Denote by $S = \sum_{i=1}^{K} w_i U_i^{-1}$, and define the following events: for $i \in \{1, \ldots, K\}$,

$$A_{i,\varepsilon} = \left\{U_i^{-1} > \frac{1+\delta_\varepsilon}{w_i \varepsilon}\right\}, \quad B_{i,\varepsilon} = \left\{U_i^{-1} \leqslant \frac{1+\delta_\varepsilon}{w_i \varepsilon},\ S > 1/\varepsilon\right\}.$$

Let $A_\varepsilon = \bigcup_{i=1}^{K} A_{i,\varepsilon}$ and $B_\varepsilon = \bigcap_{i=1}^{K} B_{i,\varepsilon}$ and thus we have

$$\mathbb{P}\left(S > 1/\varepsilon\right) = \mathbb{P}(A_\varepsilon) + \mathbb{P}(B_\varepsilon).$$

First we show $\mathbb{P}(B_\varepsilon) = o(\varepsilon)$. Note that $S > 1/\varepsilon$ implies that there exists $i \in \{1, \ldots, K\}$ such that $U_i^{-1} > \frac{1}{w_i K \varepsilon}$. Hence,

$$\begin{aligned}
\mathbb{P}\left(B_\varepsilon\right) &\leqslant \sum_{i=1}^{K} \mathbb{P}\left(\frac{1}{w_i K \varepsilon} < U_i^{-1} \leqslant \frac{1+\delta_\varepsilon}{w_i \varepsilon}, S > 1/\varepsilon\right) \\
&\leqslant \sum_{i=1}^{K} \mathbb{P}\left(\frac{1}{w_i K \varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i \varepsilon}, S > 1/\varepsilon\right) + \sum_{i=1}^{K} \mathbb{P}\left(\frac{1-\delta_\varepsilon}{w_i \varepsilon} < U_i^{-1} \leqslant \frac{1+\delta_\varepsilon}{w_i \varepsilon}\right) \\
&\leqslant \sum_{i=1}^{K} \mathbb{P}\left(\frac{1}{w_i K \varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i \varepsilon}, S > 1/\varepsilon\right) + \sum_{i=1}^{K} w_i \varepsilon \left(\frac{1}{1-\delta_\varepsilon} - \frac{1}{1+\delta_\varepsilon}\right) \\
&=: I_1 + I_2.
\end{aligned}$$

Noting that $\delta_\varepsilon \downarrow 0$ as $\varepsilon \downarrow 0$, we have $I_2 = o(\varepsilon)$. We next focus on I_1. Observe

$$\begin{aligned}
I_1 &\leqslant \sum_{i=1}^{K} \mathbb{P}\left(\frac{1}{w_i K \varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i \varepsilon}, \sum_{j \neq i}^{K} w_j U_j^{-1} > \delta_\varepsilon/\varepsilon\right) \\
&\leqslant \sum_{i=1}^{K} \sum_{j \neq i}^{K} \mathbb{P}\left(\frac{1}{w_i K \varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i \varepsilon}, U_j^{-1} > \frac{\delta_\varepsilon}{w_j K \varepsilon}\right).
\end{aligned}$$

It remains to show for $1 \leqslant i \neq j \leqslant K$,

$$I_{i,j} := \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i\varepsilon}, U_j^{-1} > \frac{\delta_\varepsilon}{w_j K\varepsilon}\right) = o(\varepsilon).$$

Condition (G) implies that there exist $Z_{i,j}$ and $\delta_{i,j}$ such that

$$\Phi^{-1}(U_j) = \sigma_{ij}\Phi^{-1}(U_i) + \delta_{ij}Z_{ij}, \tag{5.27}$$

where Z_{ij} is a standard normal random variable that is independent of U_i and $\sigma_{ij}^2 + \delta_{ij}^2 = 1$. If $\sigma_{ij} = -1$, we have $U_i = 1 - U_j$. This implies that $I_{i,j} = 0$ for $\varepsilon > 0$ sufficiently small. Next, assume $|\sigma_{ij}| < 1$, and write $\gamma_{ij} = \Phi^{-1}(w_i K\varepsilon)$ if $-1 < \sigma_{ij} \leqslant 0$ and $\gamma_{ij} = \Phi^{-1}\left(\frac{w_i\varepsilon}{1-\delta_\varepsilon}\right)$ if $0 < \sigma_{ij} < 1$. We have

$$\begin{aligned}
I_{i,j} &= \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i\varepsilon}, \sigma_{ij}\Phi^{-1}(U_i) + \delta_{ij}Z_{ij} < \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right)\right) \\
&\leqslant \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i\varepsilon}, \delta_{ij}Z_{ij} < \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) - \sigma_{ij}\gamma_{ij}\right) \\
&= \mathbb{P}\left(\frac{1}{w_i K\varepsilon} < U_i^{-1} \leqslant \frac{1-\delta_\varepsilon}{w_i\varepsilon}\right)\mathbb{P}\left(\delta_{ij}Z_{ij} < \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) - \sigma_{ij}\gamma_{ij}\right).
\end{aligned}$$

Note that $\Phi^{-1}(\varepsilon) \sim -\sqrt{-2\ln\varepsilon}$, as $\varepsilon \downarrow 0$, which is a slowly varying function. Taking $\delta_\varepsilon = -1/\log\varepsilon$, we have

$$\Phi^{-1}\left(\frac{w_i\varepsilon}{1-\delta_\varepsilon}\right) \sim \Phi^{-1}(w_i K\varepsilon) \sim \Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) \quad \text{as } \varepsilon \downarrow 0.$$

This implies

$$\Phi^{-1}\left(\frac{w_j K\varepsilon}{\delta_\varepsilon}\right) - \sigma_{ij}\gamma_{ij} \to -\infty, \text{ as } \varepsilon \downarrow 0.$$

Hence $I_{i,j} = o(\varepsilon)$. Consequently, $I_1 = o(\varepsilon)$ and further $\mathbb{P}(B_\varepsilon) = o(\varepsilon)$. Next, we show $\mathbb{P}(A_\varepsilon) \sim \varepsilon$. By the Bonferroni inequality, we have,

$$\sum_{i=1}^{K}\mathbb{P}(A_{i,\varepsilon}) - \sum_{1\leqslant i<j\leqslant K}\mathbb{P}(A_{i,\varepsilon}\cap A_{j,\varepsilon}) \leqslant \mathbb{P}(A_\varepsilon) \leqslant \sum_{i=1}^{K}\mathbb{P}(A_{i,\varepsilon}).$$

Direct calculation gives

$$\sum_{i=1}^{K}\mathbb{P}(A_{i,\varepsilon}) = \sum_{k=1}^{K}\frac{w_i\varepsilon}{1+\delta_\varepsilon} \sim \varepsilon.$$

For any $1 \leqslant i < j \leqslant K$, since the Gaussian copula is tail independent (e.g., Example 7.38 of McNeil et al. (2015)), we have, writing $w = \max\{w_i, w_j\}$,

$$\begin{aligned}\mathbb{P}(A_{i,\varepsilon} \cap A_{j,\varepsilon}) &= \mathbb{P}\left(U_i^{-1} > \frac{1+\delta_\varepsilon}{w_i\varepsilon}, U_j^{-1} > \frac{1+\delta_\varepsilon}{w_j\varepsilon}\right) \\ &\leqslant \mathbb{P}\left(U_i < \frac{w\varepsilon}{1+\delta_\varepsilon}, U_j < \frac{w\varepsilon}{1+\delta_\varepsilon}\right) = o(1)\mathbb{P}\left(U_1 < \frac{w\varepsilon}{1+\delta_\varepsilon}\right) = o(1)\varepsilon.\end{aligned}$$

Hence $\mathbb{P}(A_{i,\varepsilon} \cap A_{j,\varepsilon}) = o(\varepsilon)$. This implies $\mathbb{P}(A_\varepsilon) \sim \varepsilon$, and we establish (5.26).

(iii) By Lemma A.1 of Vovk and Wang (2020), we have

$$a_{\mathcal{H}}(\varepsilon) = \varepsilon \left(\sup\left\{q_0^+\left(\frac{1}{K}\sum_{i=1}^K P_i^{-1}\right) \mid P_1, \dots, P_K \in \mathcal{U}\right\}\right)^{-1}, \quad \varepsilon \in (0,1),$$

where $q_0^+(X) = \sup\{x \in \mathbb{R} \mid \mathbb{P}(X \leqslant x) = 0\}$. Note that for any $\delta > 0$, there exists $0 < \varepsilon_\delta < 1$ such that for all $x \in (0, \varepsilon_\delta)$

$$-\frac{(1+\delta)}{x} < \tan\left(-\frac{\pi}{2} + x\right) < -\frac{(1-\delta)}{x}.$$

For $\delta > 0$, letting $0 < \varepsilon < \varepsilon_\delta/\pi$ and using Theorem 4.6 in Bernard et al. (2014), we have

$$\begin{aligned}&\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^K \mathcal{C}^{-1}(P_i)\right) \mid P_1, \dots, P_K \in \mathcal{U}\right\} \\ &= \inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^K \tan\left(\pi\left(P_i - \frac{1}{2}\right)\right)\right) \mid P_1, \dots, P_K \in \mathcal{U}\right\} \\ &= \inf\left\{q_1\left(\frac{1}{K}\sum_{i=1}^K \tan\left(\pi\left(\varepsilon P_i - \frac{1}{2}\right)\right)\right) \mid P_1, \dots, P_K \in \mathcal{U}\right\} \\ &\leqslant \inf\left\{q_1\left(\frac{1}{K}\sum_{i=1}^K -\frac{1-\delta}{\varepsilon\pi P_i}\right) \mid P_1, \dots, P_K \in \mathcal{U}\right\} \\ &= -\frac{1-\delta}{\varepsilon\pi}\sup\left\{q_0^+\left(\frac{1}{K}\sum_{i=1}^K P_i^{-1}\right) \mid P_1, \dots, P_K \in \mathcal{U}\right\} = -\frac{1-\delta}{a_{\mathcal{H}}(\varepsilon)\pi}.\end{aligned}$$

Similarly, we obtain, for $0 < \varepsilon < \varepsilon_\delta/\pi$,

$$\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(P_i)\right)\right\} \geqslant -\frac{1+\delta}{a_{\mathcal{H}}(\varepsilon)\pi}.$$

Consequently,

$$\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(P_i)\right)\right\} \sim -\frac{1}{a_{\mathcal{H}}(\varepsilon)\pi} \quad \text{as } \varepsilon \downarrow 0.$$

Plugging the above result in the formula for $a_{\mathcal{C}}$ in (5.10), and using $\mathcal{C}(y) \sim -1/(\pi y)$ as $y \to -\infty$, we have, as $\varepsilon \downarrow 0$,

$$\begin{aligned} a_{\mathcal{C}}(\varepsilon) &= \mathcal{C}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(P_i)\right)\right\}\right) \\ &\sim -\frac{1}{\pi}\left(\inf\left\{q_\varepsilon\left(\frac{1}{K}\sum_{i=1}^{K}\mathcal{C}^{-1}(P_i)\right)\right\}\right)^{-1} \sim a_{\mathcal{H}}(\varepsilon). \end{aligned}$$

This completes the proof.

(iv) By (i), it suffices to show that for $r \neq -1$

$$\frac{M_{-1,K}(p_1,\ldots,p_K)}{M_{r,K}(p_1,\ldots,p_K)} \nrightarrow 1, \text{ as } \max_{i\in\{1,\ldots,K\}} p_i \downarrow 0.$$

Take $p_1 = p^2$ and $p_i = x_i p$ with $x_i > 0$ and $p > 0$ for $i = 2,\ldots,K$. By homogeneity of M_r, for $r \leqslant -1$,

$$\frac{M_{-1,K}(p_1,\ldots,p_K)}{M_{r,K}(p_1,\ldots,p_K)} = \frac{M_{-1,K}(p,x_2,\ldots,x_K)}{M_{r,K}(p,x_2,\ldots,x_K)}.$$

Hence

$$\lim_{p\downarrow 0}\frac{M_{-1,K}(p_1,\ldots,p_K)}{M_{r,K}(p_1,\ldots,p_K)} = K^{1/r+1} \neq 1, \quad r < -1.$$

This proves the claim of (iv) for $r < -1$. The case for $r > -1$ can be argued similarly. □

Proof of Theorem 5.3. Take arbitrary $p_1, \ldots, p_K \in (0, 1]$, and let $j \in \{1, \ldots, K\}$ be such that $\min_{k\in\{1,\ldots,K\}} p_{(k)}/k = p_{(j)}/j$. Noting that

$$\sum_{i=1}^{K} \frac{1}{p_i} = \sum_{i=1}^{K} \frac{1}{p_{(i)}}, \text{ and } \frac{p_{(j)}}{j} \leqslant \frac{p_{(i)}}{i}, \ i = 1, \ldots, K,$$

we have

$$\frac{S_K(p_1, \ldots, p_K)}{M_{-1,K}(p_1, \ldots, p_K)} = \frac{1}{j} p_{(j)} \left(\sum_{i=1}^{K} \frac{1}{p_i} \right) = \sum_{i=1}^{K} \frac{1}{j} p_{(j)} \frac{1}{p_{(i)}} \leqslant \sum_{i=1}^{K} \frac{1}{i} p_{(i)} \frac{1}{p_{(i)}} = \sum_{i=1}^{K} \frac{1}{i} = \ell_K.$$

Moreover,

$$\frac{S_K(p_1, \ldots, p_K)}{M_{-1,K}(p_1, \ldots, p_K)} = \frac{1}{j} p_{(j)} \left(\sum_{i=1}^{K} \frac{1}{p_{(i)}} \right) \geqslant \frac{1}{j} p_{(j)} \left(\sum_{i=1}^{j} \frac{1}{p_{(j)}} + \sum_{i=j+1}^{K} \frac{1}{p_{(i)}} \right) \geqslant 1.$$

Therefore, $M_{-1,K} \leqslant S_K \leqslant \ell_K M_{-1,K}$. The two special cases of equalities are straightforward to check. □

Proof of Proposition 5.5. (i) Recall that $a_F(x) = a_F x$ for $x \in (0, 1)$. By (i) of Proposition 5.3, we have $b_F(\delta) \sim \delta$ as $\delta \downarrow 0$. Hence $\lim_{\delta\downarrow 0} b_F(\delta)/a_F(\delta) = 1/a_F$. By Proposition 6 of Vovk and Wang (2020), we have $a_F \sim 1/\log K$, as $K \to \infty$. Consequently,

$$\lim_{\delta\downarrow 0} \frac{b_F(\delta)}{a_F(\delta)} \sim \log K, \text{ as } K \to \infty.$$

Moreover, for the harmonic averaging method, $c_F(\varepsilon) = \varepsilon$. This implies $c_F(\varepsilon)/a_F(\varepsilon) = 1/a_F$. We establish the claim by the fact $a_F \sim 1/\log K$, as $K \to \infty$.

(ii) By Theorem 5.2, we have $a_{\mathcal{C}}(\delta) \sim a_{\mathcal{H}}(\delta)$ and $b_{\mathcal{C}}(\delta) \sim b_{\mathcal{H}}(\delta)$ as $\delta \downarrow 0$, which together with (i) leads to

$$\lim_{\delta\downarrow 0} \frac{b_{\mathcal{C}}(\delta)}{a_{\mathcal{C}}(\delta)} \sim \log K, \text{ as } K \to \infty.$$

The rest of the statement follows by noting that $c_{\mathcal{C}}(\delta) = b_{\mathcal{C}}(\delta)$.

(iii) For the Simes method, recall that $a_F(x) = x/\ell_K$ and $b_F(x) = c_F(x) = x$. The claim follows directly from the fact that $\ell_K = \sum_{k=1}^{K} \frac{1}{k} \sim \log K$, as $K \to \infty$. □

5.9.2 Additional tables

In Tables 5.5 and 5.6 we report numerical results of prices for validity for $\varepsilon = 0.05$ and 0.0001, respectively.

Table 5.5: $b_F(\varepsilon)/a_F(\varepsilon)$ and $c_F(\varepsilon)/a_F(\varepsilon)$ for $\varepsilon = 0.05$ and $K \in \{50, 100, 200, 400\}$

	$K = 50$		$K = 100$		$K = 200$		$K = 400$	
	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F
Bonferroni	1.025	50.000	1.026	100.000	1.026	200.000	1.026	400.000
Negative-quartic	1.367	25.071	1.367	42.164	1.368	70.911	1.368	119.257
Simes	4.499	4.499	5.187	5.187	5.878	5.878	6.570	6.570
Cauchy	6.623	6.623	7.463	7.463	8.274	8.274	9.055	9.055
Harmonic	6.793	6.625	7.650	7.459	8.485	8.273	9.306	9.072
Geometric	15.679	2.718	16.874	2.718	17.755	2.718	18.395	2.718

Table 5.6: $b_F(\varepsilon)/a_F(\varepsilon)$ and $c_F(\varepsilon)/a_F(\varepsilon)$ for $\varepsilon = 0.0001$ and $K \in \{50, 100, 200, 400\}$

	$K = 50$		$K = 100$		$K = 200$		$K = 400$	
	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F	b_F/a_F	c_F/a_F
Bonferroni	1.000	50.000	1.000	100.000	1.000	200.000	1.000	400.000
Negative-quartic	1.333	25.071	1.333	42.164	1.333	70.911	1.333	119.257
Simes	4.499	4.499	5.187	5.187	5.878	5.878	6.570	6.570
Cauchy	6.625	6.625	7.465	7.465	8.274	8.274	9.055	9.055
Harmonic	6.625	6.625	7.459	7.459	8.272	8.272	9.071	9.071
Geometric	5416.222	2.718	6601.414	2.718	7523.231	2.718	8214.151	2.718

Chapter 6

Conclusions and Future works

6.1 Concluding remarks

This thesis studies several problems in risk aggregation under different dependence assumptions. In Chapter 2, we show that the diversification of iid Pareto losses without finite mean is greater than an individual Pareto loss in the sense of first-order stochastic dominance under three different model setups. Several important implications are provided by these results. First, diversification of Pareto losses without finite mean may increase the risk assessment of a portfolio. Second, risk bearers will not share Pareto losses without finite mean in an equilibrium model. Third, transferring Pareto losses without finite mean from risk bearers to external parties may benefit everyone involved in the process.

Chapter 3 studies the ordering relationship for aggregation sets where the marginal distributions are connected by either a distribution mixture or a quantile mixture. We also investigate the ordering relationship for the worst-case value of risk measures on aggregation sets and their mixtures. The general conclusion is that, more "homogeneous" marginal distributions give more severe model uncertainty, thus more dangerous risk aggregation. Applications of our results are discussed in the contexts of portfolio diversification and merging p-values.

In Chapter 4, we study risk aggregation of two ordered risks in the presence of unknown dependence structure. The bounds of $\leqslant_{\mathrm{cv}}$-consistent and $\leqslant_{\mathrm{cx}}$-consistent risk measures are attained by either the DL coupling or comonotonicity. By introducing the notion of strong stochastic order, we analyzed bounds on tail risk measures such as VaR and RVaR, which are neither $\leqslant_{\mathrm{cv}}$-consistent nor $\leqslant_{\mathrm{cx}}$-consistent. In particular, if the generator of the

tail risk measure is $\leqslant_{\mathrm{cv}}$-consistent, the worst-case value of the tail risk measure with the order constraint can be attained by letting the upper-tail risks be DL-coupled. Moreover, analytical formulas for bounds on Value-at-Risk are obtained.

In Chapter 5, we discuss two aspects of merging p-values: the impact of the dependence structure on the critical thresholds and the trade-off between validity and efficiency. Two general classes of merging methods, the generalized mean class and the order statistics class, are studied. We introduce the notion of IC-balance, which serves as a nice property for a merging method to be insensitive to dependence between independence and comonotonicity. Among the two general classes of merging methods, our results show that the Cauchy combination, the Simes, and the harmonic averaging methods keep a good balance on the trade-off between validity and efficiency, and their performances against model misspecification is better than many other commonly used methods.

6.2 Future work and open questions

6.2.1 Diversification effects of Pareto risks

The diversification effects of Pareto risks are investigated in Chapter 2 by numerical studies where two open technical questions arise. The first question is that it is unknown whether

$$\frac{1}{k}\sum_{i=1}^{k} X_i \leqslant_{\mathrm{st}} \frac{1}{\ell}\sum_{i=1}^{\ell} X_i, \tag{6.1}$$

holds for $k, \ell \in \mathbb{N}$ such that $k \leqslant \ell$, where $X_1, \dots, X_l$ are iid Pareto losses without finite mean. The statement is true if ℓ is a multiple of k, as shown in Proposition 2.2. The second question is whether we have

$$\mathrm{VaR}_p\left(\sum_{i=1}^{n} \theta_i X_i\right) \geqslant \sum_{i=1}^{n} \theta_i \mathrm{VaR}_p(X_i). \tag{6.2}$$

for $(\theta_1, \dots, \theta_n) \in \Delta_n$ and independent Pareto losses $X_1, \dots, X_n$ with possibly different tail parameters no larger than 1. Both (6.1) and (6.2) are anticipated to hold from the numerical results in Chapter 2, although a proof seems to be beyond the current techniques.

6.2.2 Open questions related to mixtures of risk aggregation

Many questions on quantile mixtures are still open, and we list four of them below. The first question concerns whether $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\Lambda \otimes \mathbf{F})$ holds for cases other than the uniform distributions in Proposition 3.2. As we have seen from Example 3.2, for $\mathbf{F} \in \mathcal{M}^n$ and $\Lambda \in \mathcal{Q}_n$, $\mathcal{D}_n(\mathbf{F})$ and $\mathcal{D}_n(\Lambda \otimes \mathbf{F})$ are generally not comparable. It remains open whether $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\Lambda \otimes \mathbf{F})$ under some conditions. For instance, Proposition 3.2 requires $n \geqslant 3$ and Λ being a constant times the identity, to use the characterization of $\mathcal{D}_n(\mathbf{F})$ from Mao et al. (2019). It remains unclear whether the same conclusion holds for $n = 2$ or other choices of Λ.

The second question concerns decreasing densities (or increasing densities). A concrete conjecture is presented below, which is inspired by Theorem 3.3. It is unclear how to formulate natural classes of distributions other than $\mathcal{M}_D$ (or $\mathcal{M}_I$) such that similar statements can be expected.

Conjecture 6.1. For $\Lambda \in \mathcal{Q}_n$ and $\mathbf{F} \in \mathcal{M}_D^n$, we have $\mathcal{D}_n(\mathbf{F}) \subset \mathcal{D}_n(\Lambda \otimes \mathbf{F})$. Weaker versions of this conjecture are:

(i) For $F \in \mathcal{M}_D$, and $\boldsymbol{\lambda}, \boldsymbol{\gamma} \in \mathbb{R}_+^n$, if $\boldsymbol{\gamma} \prec \boldsymbol{\lambda}$, then $\mathcal{D}_n(F^{\lambda_1}, \dots, F^{\lambda_n}) \subset \mathcal{D}_n(F^{\gamma_1}, \dots, F^{\gamma_n})$.

(ii) For $F_1, \dots, F_n \in \mathcal{M}_D$, $\mathcal{D}_n(F_1, \dots, F_n) \subset \mathcal{D}_n(F, \dots, F)$ where $F^{-1} = \frac{1}{n} \sum_{i=1}^n F_i^{-1}$.

(iii) For $F \in \mathcal{M}_D$ and $(\lambda_1, \dots, \lambda_n) \in \Delta_n$, $\mathcal{D}_n(F^{n\lambda_1}, \dots, F^{n\lambda_n}) \subset \mathcal{D}_n(F, \dots, F)$.

It is obvious that the main statement in Conjecture 6.1 implies (i) by noting that one can choose Λ such that $\boldsymbol{\gamma} = \Lambda \boldsymbol{\lambda}$ and it implies (ii) by choosing $\Lambda = (\frac{1}{n})_{n \times n}$. Both (i) and (ii) imply (iii). An example is provided below to illustrate the connection of Conjecture 6.1 to joint mixability.

Example 6.1. We make a connection of Conjecture 6.1 to Theorem 3.2 of Wang and Wang (2016), which says that for $F_i \in \mathcal{M}_D$ with essential support $[0, b_i]$, $i = 1, \dots, n$, $\mathcal{D}_n(F_1, \dots, F_n)$ contains a point mass if and only if the *mean-length condition* holds, that is,

$$\sum_{i=1}^n \mu_i \geqslant \max_{i=1,\dots,n} b_i$$

where μ_i is the mean of F_i, $i = 1, \dots, n$. For $\Lambda \in \mathcal{Q}_n$ and $\mathbf{F} \in \mathcal{M}_D^n$, let $(\hat{\mu}_1, \dots, \hat{\mu}_n)$ be the mean vector of $\Lambda \otimes \mathbf{F}$. Note that

$$\sum_{i=1}^n \hat{\mu}_i = \mathbf{1}_n^\top \Lambda \boldsymbol{\mu} = \mathbf{1}_n^\top \boldsymbol{\mu} = \sum_{i=1}^n \mu_i,$$

where $\mathbf{1}_n = (1, \dots, 1) \in \mathbb{R}^n$. On the other hand, each component of $\Lambda \otimes \mathbf{F}$ has a shorter or equal length of support than the maximum length of $\mathbf{F}$. As a consequence, if the mean-length condition holds for $\mathbf{F}$, then it also holds for $\Lambda \otimes \mathbf{F}$. Therefore, if $\mathcal{D}_n(\mathbf{F})$ contains a point mass, then so does $\mathcal{D}_n(\Lambda \otimes \mathbf{F})$; on the contrary, if $\mathcal{D}_n(\Lambda \otimes \mathbf{F})$ contains a point mass, $\mathcal{D}_n(\mathbf{F})$ does not necessarily contains a point mass, since it may have a longer length of the maximum support. This, at least intuitively, suggests that $\mathcal{D}_n(\mathbf{F}) \subsetneq \mathcal{D}_n(\Lambda \otimes \mathbf{F})$ may hold, as in Conjecture 6.1.

The third question is about the order of VaR for quantile mixture. Our numerical results in Figure 3.4 suggest that the VaR relation

$$\overline{\mathrm{VaR}}_p(\mathbf{F}) \leqslant \overline{\mathrm{VaR}}_p(\Lambda \otimes \mathbf{F})$$

holds for more general choices of $\mathbf{F}$ than the ones in Theorem 3.3. We are not sure what general conditions on $\mathbf{F}$ will guarantee this relation to hold.

The last question concerns a cross comparison of distribution and quantile mixtures. As we see from Proposition 3.7,

$$\overline{\mathrm{VaR}}_p(\Lambda\mathbf{F}) \leqslant \overline{\mathrm{VaR}}_p(\Lambda \otimes \mathbf{F})$$

holds for $\mathbf{F}$ being a vector of Pareto distributions with the same shape parameter and infinite mean. We wonder whether the same relationship holds for other distributions without a finite mean. Note that for the case of finite mean, the relationship may be reversed, as illustrated in Figure 3.1; however we do not have a proof for the reverse inequality (assuming finite mean) either. Generally, it is unclear to us whether and in which situation $\mathcal{D}_n(\Lambda\mathbf{F})$ and $\mathcal{D}_n(\Lambda \otimes \mathbf{F})$ are comparable.

6.2.3 Risk aggregation of more than two ordered risks

We have focused on the problem of two ordered random variables in Chapter 4, while a more general problem considering the order constraint among several risks in a large portfolio would also be interesting. Such a constraint is motivated by monotone treatment effect analysis in causal inference (see Manski (1997)). The statistical inference of stochastically ordered distributions can be handled via IDR of Henzi et al. (2021). Let $G_1, \dots, G_n$ be n distributions satisfying $G_1 \leqslant_{\mathrm{st}} \dots \leqslant_{\mathrm{st}} G_n$. Denote by

$$\mathcal{R}_n^o = \{Y_1 + \cdots + Y_n : Y_i \sim G_i,\ i = 1, \dots, n,\ Y_1 \leqslant \dots \leqslant Y_n\}.$$

We are interested in finding the worst-case value of a risk measure ρ over the set $\mathcal{R}_n^o$. If ρ is $\leqslant_{\mathrm{cx}}$-consistent, then the worst-case value is attained by comonotonicity. For ρ that is not $\leqslant_{\mathrm{cx}}$-consistent, such as the interesting case of VaR, the problem is challenging and cannot be solved by the current techniques. Even without the order constraint, only limited analytical results are available for $n \geqslant 3$; see Wang et al. (2013) and Blanchet et al. (2020). We leave the theoretical analysis of this question, as well as the corresponding algorithms, for future work.

References

Andriani, P. and McKelvey, B. (2007). Beyond Gaussian averages: Redirecting international business and management research toward extreme events and power laws. *Journal of International Business Studies*, **38**(7), 1212–1230.

Arnold, S., Molchanov, I. and Ziegel, J. F. (2020). Bivariate distributions with ordered marginals. *Journal of Multivariate Analysis*, **177**, 104585.

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

Balkema, A. and de Haan, L. (1974). Residual life time at great age. *Annals of Probability*, **2**(5), 792–804.

Barrett, G. F. and Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, **71**(1), 71–104.

BCBS (2019). *Minimum Capital Requirements for Market Risk. February 2019.* Basel Committee on Banking Supervision. Basel: Bank for International Settlements. `https://www.bis.org/bcbs/publ/d457.htm`

Beirlant, J., Dierckx, G., Goegebeur, Y. and Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, **2**(2), 177–200.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**(1), 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**(4), 1165–1188.

Bernard, C., Jiang, X. and Wang, R. (2014). Risk aggregation with dependence uncertainty. *Insurance: Mathematics and Economics*, **54**, 93–108.

Bernard, C., Rüschendorf, L. and Vanduffel, S. (2017). Value-at-Risk bounds with variance constraints. *Journal of Risk and Insurance*, **84**(3), 923–959.

Bernard, C., Rüschendorf, L., Vanduffel, S. and Wang, R. (2017). Risk bounds for factor models. *Finance and Stochastics*, **21**(3), 631–659.

Biffis, E. and Chavez, E. (2014). Tail risk in commercial property insurance. *Risks*, **2**(4), 393–410.

Blanchet, J., Lam, H., Liu, Y. and Wang, R. (2020). Convolution bounds on quantile aggregation. arXiv: 2007.09320.

Cai, J., Liu, H. and Wang, R. (2018). Asymptotic equivalence of risk measures under dependence uncertainty. *Mathematical Finance*, **28**(1), 29–49.

Cirillo, P. and Taleb, N. N. (2020). Tail risk of contagious diseases. *Nature Physics*, **16**(6), 606–613.

Clark, D. R. (2013). A note on the upper-truncated Pareto distribution. *Casualty Actuarial Society E-Forum*, Winter, 2013, Volume 1, pp. 1–22.

Cont, R., Deguest, R. and Scandolo, G. (2010). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, **10**(6), 593–606.

de Haan L. and Ferreira A. (2006). *Extreme Value Theory: An Introduction*. Springer.

Delbaen, F., Bellini, F., Bignozzi, V. and Ziegel, J. (2016). Risk measures with convex level sets. *Finance and Stochastics*, **20**(2), 433–453.

Denuit, M., Dhaene, J., Goovaerts, M.J. and Kaas, R. (2005). *Actuarial Theory for Dependent Risks*. Wiley.

Dhaene, J., Vanduffel, S., Goovaerts, M.J., Kaas, R., Tang, Q. and Vynche, D. (2006). Risk measures and comonotonicity: A review. *Stochastic Models*, **22**(4), 573–606.

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R. and Vynche, D. (2002). The concept of comonotonicity in actuarial science and finance: Theory. *Insurance: Mathematics and Economics*, **31**(1), 3–33.

Dhaene, J., Vanduffel, S., Goovaerts, M. J., Kaas, R., Tang, Q. and Vyncke, D. (2006). Risk measures and comonotonicity: a review. *Stochastic Models*, **22**(4), 573–606.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, **32**(3), 962–994.

Eckstein, S., Kupper, M. and Pohl, M. (2020). Robust risk aggregation with neural networks. *Mathematical Finance*, **30**(4), 1229–1272.

Eling, M. and Schnell, W. (2020). Capital requirements for cyber risk and cyber risk insurance: An analysis of Solvency II, the US risk-based capital standards, and the Swiss Solvency Test. *North American Actuarial Journal*, **24**(3), 370–392.

Eling, M. and Wirfs, J. (2019). What are the actual costs of cyber risk events? *European Journal of Operational Research*, **272**(3), 1109–1119.

Embrechts, P. and Puccetti, G. (2010). Risk aggregation. In *Copula Theory and its Applications* (Eds: Jaworski et al.) pp. 111–126. Springer.

Embrechts, P. and Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, **77**(3), 423–432.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.

Embrechts, P., Liu, H. and Wang, R. (2018). Quantile-based risk sharing. *Operations Research*, **66**(4), 936–949.

Embrechts, P., Lambrigger, D. and Wüthrich, M. (2009). Multivariate extremes and the aggregation of dependent risks: examples and counter-examples. *Extremes*, **12**(2), 107–127.

Embrechts, P., McNeil, A. and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In *Risk Management: Value at Risk and Beyond* (Eds: Dempster) pp. 176–223. Cambridge University Press.

Embrechts, P. and Puccetti, G. (2006). Bounds for functions of multivariate risks. *Journal of Multivariate Analysis*, **97**(2), 526–547.

Embrechts, P., Puccetti, G. and Rüschendorf, L. (2013). Model uncertainty and VaR aggregation. *Journal of Banking and Finance*, **37**(8), 2750–2764.

Embrechts, P., Resnick, S. I. and Samorodnitsky, G. (1999). Extreme value theory as a risk management tool. *North American Actuarial Journal*, **3**(2), 30–41.

Embrechts, P., Wang, B. and Wang, R. (2015). Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance and Stochastics*, **19**(4), 763–790.

Efron, B. (2010). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Fama, E. F. and Miller, M. H. (1972). *The Theory of Finance*. Dryden Press.

Filipović, D. and Svindland, G. (2012). The canonical model space for law-invariant convex risk measures is L^1. *Mathematical Finance*, **22**(3), 585–589.

FINMA (2021). Standardmodell Versicherungen (Standard Model Insurance): Technical description for the SST standard model non-life insurance (in German), October 31, 2021, www.finma.ch.

Fisher, R. A. (1948). Combining independent tests of significance. *American Statistician*, **2**(5), 30.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, **6**(4), 429–447.

Föllmer, H. and Schied, A. (2016). *Stochastic Finance. An Introduction in Discrete Time*. Walter de Gruyter, Fourth Edition.

Frees, E. W. (2009). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Furman, E., Wang, R. and Zitikis, R. (2017). Gini-type measures of risk and variability: Gini shortfall, capital allocations, and heavy-tailed risks. *Journal of Banking and Finance*, **83**, 70–84.

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics*, **1**(1), 255–294.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, **26**(4), 584–597.

Goeman, J. J., Meijer, R. J., Krebs, T. J. and Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, **106**(4), 841–856.

Guan, Y., Jiao, Z. and Wang, R. (2022). A reverse Expected Shortfall optimization formula. arXiv:2203.02599.

Hadar, J. and Russell, W. R. (1969). Rules for ordering uncertain prospects. *The American Economic Review*, **59**(1), 25–34.

Hadar, J. and Russell, W. R. (1971). Stochastic dominance and diversification. *Journal of Economic Theory*, **3**(3), 288–305.

Hardy, G. H., Littlewood, J. E. and Pólya, G. (1934). *Inequalities*. Cambridge University Press.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., et al. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **344**(8), 539–548.

Henzi, A., Ziegel, J. F. and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, **83**(5), 963–993.

Hofert, M. and Wüthrich, M. V. (2012). Statistical review of nuclear power accidents. *Asia-Pacific Journal of Risk and Insurance*, **7**(1).

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**(2), 65–70.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, **25**(5), 423–430.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, **49**(2), 1055–1080.

Ibragimov, R. (2009). Portfolio diversification and value at risk under thick-tailedness. *Quantitative Finance*, **9**(5), 565–580.

Ibragimov, R., Jaffee, D. and Walden, J. (2009). Non-diversification traps in markets for catastrophic risk. *Review of Financial Studies*, **22**(3), 959–993.

Ibragimov, R., Jaffee, D. and Walden, J. (2011). Diversification disasters. *Journal of Financial Economics*, **99**(2), 333–348.

Ibragimov, R. and Walden, J. (2007). The limits of diversification when losses may be large. *Journal of Banking and Finance*, **31**(8), 2551–2569.

Jakobsons, E., Han, X. and Wang, R. (2016). General convex order on risk aggregation. *Scandinavian Actuarial Journal*, **2016**(8), 713–740.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (2012). *Loss Models: From Data to Decisions.* 4th Edition. John Wiley & Sons.

Liu, F. and Wang, R. (2021). A theory for measures of tail risk. *Mathematics of Operations Research*, **46**(3), 1109–1128.

Liu, Y. and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, **115**(529), 393–402.

Makarov, G. (1981). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability and its Applications*, **26**(4), 803–806.

Malinvaud, E. (1972). The allocation of individual risks in large markets. *Journal of Economic Theory*, **4**(2), 312–328.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, **65**(6), 1311–1334.

Mao, T., Wang, B. and Wang, R. (2019). Sums of uniform random variables. *Journal of Applied Probability*, **56**(3), 918–936.

Mao, T. and Wang, R. (2020). Risk aversion in regulatory capital calculation. *SIAM Journal on Financial Mathematics*, **11**(1), 169–200.

Marshall, A. W., Olkin, I. and Arnold, B. (2011). *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition.

McNeil, A. J., Frey, R. and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools*. Revised Edition. Princeton University Press.

Moscadelli, M. (2004). The modelling of operational risk: Experience with the analysis of the data collected by the Basel committee. *SSRN*:557214.

Müller, A. and Scarsini, M. (2000). Some remarks on the supermodular order. *Journal of Multivariate Analysis*, **73**(1), 107–119.

Müller, A. and Stoyan, D. (2002). *Comparison Methods for Statistical Models and Risks*. Wiley.

Nešlehová, J., Embrechts, P. and Chavez-Demoulin, V. (2006). Infinite mean models and the LDA for operational risk. *Journal of Operational Risk*, **1**(1), 3–25.

Nelsen, R. (2006). *An Introduction to Copulas*. Second Edition. Springer.

Nordhaus, W. D. (2009). An analysis of the Dismal Theorem, Yale University: Cowles Foundation Discussion Paper 1686.

Nutz, M. and Wang, R. (2021). The directional optimal transport. *Annals of Applied Probability*, **32**(2), 1400-1420.

OECD. (2018). *The Contribution of Reinsurance Markets to Managing Catastrophe Risk*. Available at www.oecd.org/finance/the-contribution-of-reinsurance-markets-to-managing-catastrophe-risk.pdf.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, **25**(3), 379–410.

Pflug, G. C. and Pohl, M. (2018). A review on ambiguity in stochastic portfolio optimization. *Set-Valued and Variational Analysis*, **26**(4), 733–757.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, **3**(1), 119–131.

Puccetti, G., Rigo, P., Wang, B. and Wang, R. (2019). Centers of probability measures without the mean. *Journal of Theoretical Probability*, **32**(3), 1482–1501.

Puccetti, G. and Rüschendorf, L. (2012). Computation of sharp bounds on the distribution of a function of dependent risks. *Journal of Computational and Applied Mathematics*, **236**(7), 1833–1840.

Puccetti, G. and Rüschendorf, L. (2013). Sharp bounds for sums of dependent risks. *Journal of Applied Probability,* **50**(1), 42–53.

Puccetti, G., Rüschendorf, L. and Manko, D. (2016). VaR bounds for joint portfolios with dependence constraints. *Dependence Modeling*, **4**(1), 368–381.

Puccetti, G., Rüschendorf, L., Small, D. and Vanduffel, S. (2017). Reduction of Value-at-Risk bounds via independence and variance information. *Scandinavian Actuarial Journal*, **2017**(3), 245–266.

Puccetti, G. and Wang R. (2015). Extremal dependence concepts. *Statistical Science*, **30**(4), 485–517.

Ramdas, A. K., Barber, R. F., Wainwright, M. J. and Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *Annals of Statistics*, **47**(5), 2790–2821.

Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv:2009.03167.

Rizzo, M. L. (2009). New goodness-of-fit tests for Pareto distributions. *ASTIN Bulletin*, **39**(2), 691–715.

Rødland, E. A. (2006). Simes' procedure is 'valid on average'. *Biometrika*, **93**(3), 742–746.

Rockafellar, R. T. and Uryasev, S. (2002). Conditional Value-at-Risk for general loss distributions. *Journal of Banking and Finance*, **26**(7), 1443–1471.

Rüger, B. (1978). Das maximale signifikanzniveau des tests:"lehneh o ab, wennk untern gegebenen tests zur ablehnung führen". *Metrika*, **25**(1), 171–178.

Rüschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability*, **14**(3), 623–632.

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

Samorodnitsky, G. (2017). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Routledge.

Samuelson, P. A. (1967). General proof that diversification pays. *Journal of Financial and Quantitative Analysis*, **2**(1), 1–13.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Annals of Statistics*, **26**(2), 494–504.

Sarkar, S. K. (2008). On the Simes inequality and its generalization. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (pp. 231–242). Institute of Mathematical Statistics.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, **184**(2), 407–431.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer.

Silverberg, G. and Verspagen, B. (2007). The size distribution of innovations revisited: An application of extreme value statistics to citation and value measures of patent significance. *Journal of Econometrics*, **139**(2), 318–339.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**(3), 751–754.

Sornette, D., Maillart, T. and Kröger, W. (2013). Exploring the limits of safety analysis in complex technological systems. *International Journal of Disaster Risk Reduction*, **6**, 59–66.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9440–9445.

Taleb, N. (2020). *Statistical Consequences of Fat Tails*. STEM Academic Press.

Tasche, D. (2000). Conditional expectation as quantile derivative. arXiv:0104190.

Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. Springer.

Tippett, L.H.C. (1931). *The Methods of Statistics: An Introduction Mainly for Experimentalists*. Williams and Norgate.

Uchaikin, V. V. and Zolotarev, V. M. (2011). *Chance and Stability: Stable Distributions and their Applications*. Walter de Gruyter.

Vovk, V., Wang, B. and Wang, R. (2021). Admissible ways of merging p-values under arbitrary dependence. *Annals of Statistics*, **50**(1), 351-375.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, **107**(4), 791–808.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics*, **49**(3), 1736–1754.

Wang, B. and Wang, R. (2016). Joint mixability. *Mathematics of Operations Research*, **41**(3), 808–826.

Wang, Q., Wang, R. and Wei, Y. (2020). Distortion riskmetrics on general spaces. *ASTIN Bulletin*, **50**(3), 827–851.

Wang, R., Peng, L. and Yang, J. (2013). Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities. *Finance and Stochastics*, **17**(2), 395–417.

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.

Wang, R. and Zitikis, R. (2021). An axiomatic foundation for the Expected Shortfall. *Management Science*, **67**(3), 1413–1429.

Wang, X. (2005). Volumes of generalized unit balls. *Mathematics Magazine*, **78**(5), 390–395.

Weitzman, M. L. (2009). On modeling and interpreting the economics of catastrophic climate change. *Review of Economics and Statistics*, **91**(1), 1-19.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, **116**(4), 1195–1200.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica*, **55**(1), 95–115.

Ziegel, J. (2016). Coherence and elicitability. *Mathematical Finance*, **26**(4), 901–918.