CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Cyber risk modelling using copulas

Master's thesis

Author: Bc. Michal Spišiak

Study program: Economics and Finance

Supervisor: prof. PhDr. Petr Teplý, Ph.D.

Year of defense: 2020

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, July 31, 2020

Michal Spišiak

Abstract

Cyber risk or data breach risk can be estimated similarly as other types of operational risk. First we identify problems of cyber risk models in existing literature. A large dataset consisting of 5,713 loss events enables us to apply extreme value theory. We adopt goodness of fit tests adjusted for distribution functions with estimated parameters. These tests are often overlooked in the literature even though they are essential for correct results. We model aggregate losses in three different industries separately and then we combine them using a copula. A t-test reveals that potential one-year global losses due to data breach risk are larger than the GDP of the Czech Republic. Moreover, one-year global cyber risk measured with a 99% CVaR amounts to 2.5% of the global GDP. Unlike others we compare risk measures with other quantities which allows wider audience to understand the magnitude of the cyber risk. An estimate of global data breach risk is a useful indicator not only for insurers, but also for any organization processing sensitive data.

Keywords cyber risk, operational risk, data breach, extreme value theory, copula, value at risk, conditional value at risk

Title Cyber risk modelling using copulas

Abstrakt

Kybernetické riziko nebo riziko úniku dat lze odhadnout podobně jako ostatní typy operačního rizika. Nejprve identifikujeme problémy modelů kybernetického rizika v současné literatuře. Rozsáhlý datový soubor obsahující 5 713 pozorování nám umožňuje aplikovat teorii extrémních hodnot. Používáme testy dobré shody přizpůsobené distribučním funkcím s odhadnutými parametry. Tyto testy jsou v literatuře často přehlíženy, přestože jsou nezbytné pro správné výsledky. Ztráty modelujeme samostatně ve třech různých odvětvích a pak je zkombinujeme pomocí kopule. Prostřednictvím t-testu zjišťujeme, že potenciální roční celosvětové ztráty v důsledku rizika úniku dat jsou větší než HDP České republiky. Navíc roční kybernetické riziko měřené s 99% CVaR dosahuje 2,5 % světového HDP. Na rozdíl od ostatních porovnáváme míry rizika s jinými hodnotami, což umožňuje pochopit závažnost kybernetického rizika i širšímu

publiku. Odhad globálního rizika úniku dat je užitečným ukazatelem nejen pro pojišťovny, ale také pro jakoukoli organizaci zpracovávající citlivá data.

Klíčová slova kybernetické riziko, operační riziko, únik dat, teorie extrémních hodnot, kopule, hodnota v riziku, podmíněná hodnota v riziku

Název práce Modelování kybernetického rizika pomocí kopula funkcí

Acknowledgments

I would like to express my deepest gratitude to my supervisor prof. PhDr. Petr Teplý, Ph.D. for inspiration, guidance and valuable comments.

I would also like to thank prof. Ing. Michal Mejstřík, CSc., doc. PhDr. Tomáš Havránek, Ph.D. and doc. PhDr. Zuzana Havránková, Ph.D. for their advice.

Last but not least, I owe much to my parents who have been supporting me in my study efforts by all means.

Bibliographic Record

Spišiak, Michal: *Cyber risk modelling using copulas.* Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2020, pages 87. Advisor: prof. PhDr. Petr Teplý, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

EDF empirical distribution function

EVT extreme value theory

POT peaks over threshold

GPD generalized Pareto distribution

VaR value at risk

CVaR conditional value at risk

Master's Thesis Proposal

Author	Bc. Michal Spišiak
Supervisor	prof. PhDr. Petr Teplý, Ph.D.
Proposed topic	Cyber risk modelling using copulas

Motivation The risk that a cyber attack disrupts a part of the economy is clearly rising while at the same time the level of cyber risk research is globally far behind what is needed. Cyber risk research is currently concentrated around consulting firms and insurance companies which share their findings only exceptionally. The supply of publicly available data is insufficient because the organisations mentioned above profit from such low supply by selling their overpriced products.

The procedure of operational risk assessment is described for example in (Lebovič, 2012). A more advanced method combining extreme value theory and copulas is applied in the work of (Abbate, Farkas, & Gourier, 2009). However, cyber risk modelling is more challenging than traditional operational risk modelling what is also confirmed by (Biener, Eling, & Wirfs, 2015). Even though (Lloyd's, 2017) do not fully disclose their methodology, their cyber risk estimates acquired with a scenario analysis are a substantial contribution to the research. Similar impact has a report by (Verizon, 2018) which is an annual publication with descriptive statistics related to data breach risk.

The aim of this thesis is to provide fully reproducible research while using the best publicly available data. This thesis will discuss cyber risk, its modelling and its impact on organisations around the world. It will build upon the model used to measure cyber risk using copulas introduced by (Herath & Herath, 2011) and (Shah, 2016). Particular attention will be paid to data breach risk which is a major and relatively easily measurable component of cyber risk. The importance of data breach risk is highlighted also by (Verizon, 2018).

Hypotheses

Hypothesis #1: Frequencies of losses caused by data breaches follow a Poisson distribution.

Hypothesis #2: Severities of losses caused by data breaches follow a log-normal distribution.

Hypothesis #3: A Gaussian copula describes dependencies between aggregate losses caused by data breaches in different industries.

Hypothesis #4: The possible total worldwide cost of data breaches per one year is smaller than the nominal GDP of the Czech Republic.

Methodology The first two hypotheses will be tested with Kolmogorov-Smirnov and Anderson-Darling tests. The third hypothesis will be tested with copulas goodness of fit tests such as those described in (Genest, Rémillard, & Beaudoin, 2009).

We will consider more different copulas, and we will select the most suitable one for risk measures calculation. The actuarial model used for risk measures calculation will broadly follow the one described by (Chernobai, Rachev, & Fabozzi, 2007). Nonetheless, many enhancements and optimisations introduced for instance by (Clemente & Romano, 2004) will be incorporated into the model.

The last hypothesis will be tested by calculating confidence intervals of means of risk measures using Monte Carlo simulation. Rejecting the last hypothesis will allow us to claim that the possible total worldwide cost of data breaches per one year is either lower or higher than the nominal GDP of the selected country. The parametric model used to test these hypotheses will be calibrated using publicly available data from Breach Level Index database published by Gemalto.

Expected Contribution Modelling cyber risk is generally more challenging than modelling other types of operational risk because of the constantly changing nature of cyber risk resulting in the insufficient availability of suitable data. Operational risk is a widely researched area; however cyber risk still lacks sufficient attention. Some consulting firms already offer services that assess cyber risks of interested organisations; however, their models are kept private; thus public evaluation of their validity is impossible. This thesis will offer comprehensive and fully transparent research of cyber risk. It will use carefully selected and unique data which have been overlooked by many researchers in this field. This publication will extend the research of (Abbate, Farkas, & Gourier, 2009) to the field of cyber risk. Moreover, it will make public some methods that might have been used in (Lloyd's, 2017) and that have not been published. Results can be to some degree directly used by financial institutions and other organisations to assess their exposure to cyber risk. However, it is more expected that the model and the methodology used in this thesis will serve as a prototype for other cyber risk researches. This thesis will be especially valuable to actuaries specialising in cyber risk.

Outline

1. Introduction
2. Theoretical background
   (a) Definition of cyber risk
   (b) Trends in cyber risk assessment
   (c) Current cyber threats
   (d) GDPR and the data breach risk
3. Literature review
4. Methodology
5. Empirical analysis
6. Discussion of results
7. Results
8. Conclusion

The theoretical part will begin with a definition of cyber risk as a part of the operational risk. It will be followed by literature review concentrated on trends in the cyber risk assessment and current cyber threats. Next, there will be a short analysis of the impact of GDPR on the data breach risk. An overview of basic actuarial models, risk measures, extreme value theory and copulas will conclude this part.

The empirical part will include a specification of a particular aggregate loss model where a copula explains dependencies among severities of losses in different industries. This model will be estimated using real-world data. Resulting risk measures such as value at risk or conditional value at risk will be presented with scenario analysis.

Core bibliography

Abbate, D., Farkas, W., & Gourier, E. (2009). Operational Risk Quantification using Extreme Value Theory and Copulas: From Theory to Practice.

Biener, C., Eling, M., & Wirfs, J. (2015). Insurability of Cyber Risk: An Empirical Analysis. Geneva Papers on Risk and Insurance, 40(1), 32.

Clemente, A., & Romano, C. (2004). A copula-Extreme Value Theory approach for modelling operational risk. In Operational Risl Modelling and Analysis: Theory and Practice.

Genest, C., Rémillard, B., & Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and Economics, 44(2), 199-213.

Herath, H., & Herath, T. (2011). Copula Based Actuarial Model for Pricing Cyber-Insurance Policies.

Chernobai, A., Rachev, S., & Fabozzi, F. (2007). Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis.

Lebovič, M. (2012). The use of coherent risk measures in operational risk modeling.

Lloyd's. (2017). Counting the cost. Retrieved from https://www.lloyds.com/news-and-risk-insight/risk-reports/library/technology/countingthecost

Shah, A. (2016). Pricing and Risk Mitigation Analysis of a Cyber Liability Insurance using Gaussian, t and Gumbel Copulas – A Case for Cyber Risk Index. CANADIAN ECONOMICS ASSOCIATION.

Verizon. (2018). 2018 Data Breach Investigations Report. Retrieved from https://www.verizonenterprise.com/verizon-insights-lab/dbir/

Author

Supervisor

Chapter 1

Introduction

Cyber attacks have become a part of our daily lives. Lloyd's and Cyence (2017) discovered that a mass vulnerability attack can cost as much as USD 9.7 billion or USD 28.7 billion if it is an extreme event. Everybody knows that we cannot avoid them. The best alternative is to be prepared. If we want to prepare, we have to know how large the threat is. This is the reason why we need to calculate the size of cyber risk. IBM Security and Ponemon Institute (2019) estimate average total cost of a data breach at USD 3.92 million. Moreover, 71 percent of data breaches are financially motivated and 25 of data breaches are motivated by espionage (Verizon 2019). Data breaches constitute particularly dangerous part of cyber risk. This is the reason why we concentrate our investigation on data breach risk.

The objective of this thesis is to model data breach risk with respect to all specifics of this type of risk. Cyber risk or data breach risk is a type of operational risk. It is well known that severity distribution of operational losses has heavy right tail. Therefore we have to use extreme value theory in order to properly take care of this property. Also there can be a dependence structure between losses in different industries. Therefore we have to verify if this dependence structure cannot be described with a copula. We have to establish a model. We use an actuarial model suitable for modelling operational risk. It consists of estimating parameters of distributions from pre-selected distribution families. We consider different distribution families for loss frequency and loss severity. For instance discreet probability distributions are more suitable for loss frequency while continuous distributions are more appropriate for loss severity. Second part of the model involves combining loss frequency and severity distributions into an aggregate loss distribution. We extend this model with

extreme value theory and copula. Then we calculate risk measures and we obtain an estimate of the data breach risk.

Based on literature review we test the following hypotheses:

> Hypothesis #1: Frequencies of losses caused by data breaches follow a Poisson distribution.
>
> Hypothesis #2: Severities of losses caused by data breaches follow a log-normal distribution.
>
> Hypothesis #3: A Gaussian copula describes dependencies between aggregate losses caused by data breaches in different industries.
>
> Hypothesis #4: The possible total worldwide cost of data breaches per one year is less than the nominal GDP of the Czech Republic.

In order to test these hypotheses we propose a statistical model described in the rest of this thesis.

The thesis is structured as follows: Chapter 2 gives theoretical background. Chapter 3 provides literature review. Chapter 4 gives detailed information about methodology. Chapter 5 describes our dataset. Chapter 6 contains results and their interpretation. Chapter 7 describes contribution of this thesis. Chapter 8 summarizes our findings.

Chapter 2

Theoretical background

2.1 Definition of cyber risk

"Operational risk is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk. Legal risk includes, but is not limited to, exposure to fines, penalties, or punitive damages resulting from supervisory actions, as well as private settlements" (Basel Committee on Banking Supervision, Bank for International Settlements 2019).

One part of operational risk is cyber risk. "Operational cyber security risks are defined as operational risks to information and technology assets that have consequences affecting the confidentiality, availability, or integrity of information or information systems" (Cebula, Popeck, and Young 2014).

We will concentrate only on one part of the cyber risk and it is the data breach risk. Nonetheless, it is one of the largest parts of cyber risk. "A data breach is defined as an event in which an individual's name and a medical record and/or a financial record or debit card is potentially put at risk, either in electronic or paper format" (IBM Security and Ponemon Institute 2019). Other definition by NortonLifeLock (2020) says that a data breach is a security incident in which information is accessed without authorization. Verizon (2019) distinguishes between an incident and a breach. Incident is "a security event that compromises the integrity, confidentiality or availability of an information asset." Breach is "an incident that results in the confirmed disclosure, not just potential exposure, of data to an unauthorized party."

2.2 Trends in cyber risk assessment

Cyber attacks come in many forms. Fewer organizations surveyed by Department for Digital, Culture, Media & Sport, Her Majesty's Government (2019) reported a data breach in 2019 than a year ago while more data breaches happened on average to those organizations which experienced at least one data breach. The attackers themselves choose which types of attacks are the most prevalent. They prefer attacks that yield the highest income. By causing a disruption they do not gain much reward. If they decide to disable certain service it might be because they have other than financial interests such as self-realization. They can cause a disruption using a denial of service attack or a distributed denial of service attack if they use a network of compromised computers called botnet to conduct the attack. Verizon (2019) reports that distributed denial of service attacks make the majority of all reported security incidents. In any of these two types of denial of service the attackers normally cannot gain access to information stored in the attacked system or cause any damage to it. This is also confirmed by Verizon (2019). Information systems running in clouds are usually able to recover by themselves without a human interaction. If we assume that attackers make rational decision regarding the use of their skills we can expect to see a shift from other types of cyber attacks into cyber attacks with an intention of a data breach.

IBM Security and Ponemon Institute (2019) find that the reason for 51 percent of data breaches in 2019 was malicious attack. This is 21 percentage points more than in 2014. This finding can have several reasons. Either the employees are generally more aware about data protection possibly due to GDPR. Or more attackers started to operate due to existence of positive profit margin. Also attacker focusing on other types of cyber attacks might have switched to cyber attacks with an intention of a data breach. Data breach itself does not constitute a cyber attack. Data breach is a result of a cyber attack during which an attacker obtains sensitive information and decides to release them either for free or for a compensation. Various malware can be used to perform a data breach depending on what information is targeted, how it is protected and to what extent social engineering can be used. The number of data breaches utilising social engineering rose by 18 percentage points between 2013 and 2018 (Verizon 2019).

In 2019 it took longer to identify and contain a data breach resulting from a malicious attack than from other sources (IBM Security and Ponemon Institute

2019). These types of data breaches cost 27 percent more than data breaches due to human error and 37 percent more than data breaches due to system glitches (IBM Security and Ponemon Institute 2019). When an attacker creates a data breach it is usually with an intention of making some profit. Whereas the other two reasons do not involve intentional profit driven behaviour. It can be discussed, if the 27 percent difference is not too little for such a substantial difference in the type of the data breach. Also human error resulting data breaches can be of two types. They are either intentional or not. Given the punishments for intentional data breaches caused by employees, this category should not be disturbing. Unintentional human error data breaches might be difficult to discover, but nobody tries to sell the breached records what would rise the cost. Also preventive measures can be applied. They can include mandatory password changes, etc. Verizon (2019) describes an attack which starts with a malicious outsider and at a later stage through the use of social engineering results to a data breach caused by a human error. It happens as follows. The attacker compromises an email account of one employee and enters a conversation between several people. By forwarding emails to the right targets the attacker forces other employee to make a fraudulent financial transaction.

On average the cost of data breach for large organizations in 2019 was USD 204 per employee while the same figure for small organizations was USD 3,533 per employee (IBM Security and Ponemon Institute 2019). Smaller organizations, which profit from economies of scale to a smaller extent than larger organizations, have to face higher cost of data breaches. This disproportionality can block new entrants into the industry. Or can make entry more risky because a random event such as a data breach can cause a business loss to an immature company. According to Verizon (2019) 43 percent of data breaches were targeted against small business.

IBM Security and Ponemon Institute (2019) found that involvement of third parties and complexity of infrastructure are major reasons for high data breach costs. The communication between an organization and a partner naturally creates an extra burden which causes a rise in the data breach cost. Data breach liquidation is often about information and time. When information is hidden inside the infrastructure of the third party then it influences the ability to provide an adequate response to the data breach in a timely manner. The additional time it takes to exchange information with partner can be used by the attacker to further misuse acquired data. Potentially legal fees for a dispute with the third party might come into play. Cloud platforms operated by third parties

became the main storage platform for sensitive data (Thales 2019). There are two reasons why attackers target software as a service. First, the investigation of a data breach is more complicated which gives them some extra time to hide their malicious activities and make use of the data. Second, organizations prefer to save their data into such services and therefore the concentration of possible victims is higher.

Presence of a dedicated incident response team drops the cost of a data breach on average by USD 1.23 million per data breach (IBM Security and Ponemon Institute 2019). The question is not whether a data breach happens, but when it happens. For an organization of a certain size it is legitimate to expect that a data breach happens in the near future. By spending now for the incident response team the organization can save money when the data breach happens. The cost saving can be such large that it covers the cost of having this team during calm periods. Nevertheless even then they can run simulations and prepare for the inevitable. Department for Digital, Culture, Media & Sport, Her Majesty's Government (2019) suggest that organizations should take more steps to protect themselves against cyber risk. For instance they should plan an incident response.

IBM Security and Ponemon Institute (2019) reports 95 percent higher average costs of data breach at organizations without automated systems for data breach response than at organizations with such tools installed. The aftermath of a data breach constitutes a period of an emergency for multiple departments. Using automation for other tasks than data breach response can also be vital in such an event because at least some operations of the data breach response can be automated.

Data breaches in the United States cost on average more than twice as much as globally (IBM Security and Ponemon Institute 2019). Healthcare is an industry where data breaches cost the most. An average data breach in healthcare costs 60 percent more than an average data breach in all industries combined (IBM Security and Ponemon Institute 2019). Health data are together with financial data among the most sensitive. Out of all data breach victims 15 percent are from healthcare industry and 10 percent are from financial industry (Verizon 2019). An attacker might have several different reasons for trying to perform a data breach. First, the reason might be to make a fraudulent financial transaction in order to steal money. Second, the reason can be to blackmail either the user whose data are breached or the organization which suffers the data breach. Third, the reason can be to gain information either

for own needs or to sell it. Depending on the breached data they can be replaced without any further loss. For example if the breached data are account credentials and the accounts do not contain any sensitive information then in the worst case scenario the attacker or its client might use the service tied to the account without paying for it. If the accounts contain sensitive information it depends on how quickly the attacker can extract the information from the accounts before being discovered. If the breached data directly constitute the sensitive information then the breached organization has substantially limited options. We can call the data breach in this case a data theft. One solution can be to accept the situation, apologize and provide a compensation if possible. Healthcare data breaches usually create a situation in which the organization cannot do anything else to prevent the harm because the harmful potential is already included in the breached data. When this is combined with the high sensitivity of the breached data it makes sense that the data breaches in healthcare have a high cost.

Data breaches have various consequences for organizations. Among the most common are loss of customer trust and lost business (IBM Security and Ponemon Institute 2019). Even if an organization makes sure that accounts with breached credentials are secured and no information is leaked the trust of the public always suffers. All users always have to understand that data which they provide to any organization can be breached. Nonetheless, some organization are more successful at protecting their users' data than other. Loss of business can happen when the organization does not have enough capital to cover the cost of the data breach. For this purpose it is necessary to model data breach risk and calculate risk measures. Even if a company knows how much cyber risk its operation involves, it might not be able to raise enough capital to cover this risk. This might be the reason why lost business is so common consequence of a data breach. Huge data breaches are rare and in general it might be an acceptable solution to let the few unlucky companies go bankrupt. Nonetheless knowing how much capital is needed is still a useful idea because at least it can guide the mangers towards the goal of sufficient capital to protect against losses due to data breaches.

The danger of a data breach does not lie only in its immediate consequences. On average one-third of data breach related expenses are incurred more than one year after the data breach itself (IBM Security and Ponemon Institute 2019). Negligence to indemnify the victims of the data breach immediately is unlikely to reduce the cost of the data breach. If an organization compensates victims for

the data breach instantly, it might not only be able to preserve its reputation, but also save on legal fees. It is not because the organization would not be able to determine if a legal case is worth defending, it is because some of their customers are loyal are they are willing to accept certain amount of mistakes. Nowadays this factor is even more intensified because many companies are building communities of customers both online and offline. Irritating certain key members of the community can lead to a herd effect which displeases practically the whole community. Such situation could lead to unprecedented losses. Nonetheless, there are exceptions represented by organizations which do not pay particular attention to their public relations. It might be for example because they produce a commodity. Managers should always examine each data breach of their organization on a case by case basis. One-quarter of data breaches were conducted by an act of espionage (Verizon 2019). State or state financed attackers represent a specific type of threat. Their attacks usually do not influence consumers directly. Their interests are often propaganda and secret information. On the other hand they can influence people indirectly through manipulations of elections or promoting extremism.

There are differences between industries in a timing of the costs after the data breach. In highly regulated industries like healthcare and finance 53 percent and 32 percent of costs related to a data breach appear in the first and the second years after the data breach, respectively. For the whole sample the figures are 67 percent and 22 percent, respectively (IBM Security and Ponemon Institute 2019). This suggests that in highly regulated industries the organizations cannot just hide the data breaches. They must have procedures for dealing with them and properly follow these procedures. Another reason might be that the highly regulated industries also happen to be the industries where customers care much more about what happens with their data. The average cost of a data breach is USD 3.34 million if it is contained in less than 200 days and USD 4.56 million if it is contained in more than 200 days. Thomson Reuters (2019)

In 2019 it took on average 279 days to identify and contain a data breach. This is an increase from 266 in 2018 (IBM Security and Ponemon Institute 2019). Verizon (2019) reports that 56 percent of data breaches were discover after months or later. The faster the organization identifies the data breach the lower the final cost might be. When an attacker acquires data the time goes against the organization. The more time it takes to discover the data breach, the longer the attacker can use the data for malicious purposes. The attacker might decide to wait some time before using the acquired data in order to hide

the exact date and time of the attack. Nonetheless as the time passes the data start becoming irrelevant. Therefore the attacker balances between these two factors when deciding when to take what kind of action with the breached data. The time between the first interaction of the attacker with the targeted system and acquisition of data is generally measured in minutes while the time between the incident and its discover is generally measured in months (Verizon 2019).

2.3 Current cyber threats

More than 90 percent of identified malware comes from email attachments (Thomson Reuters 2019). It is worrying on one hand, that such a simple and naive method can be used to overcome robust security measures recently implemented by most organizations. Moreover, it is always easier to block malware than to fight with it. If it is let inside the information systems by an employee then the most powerful weapon against attackers is disabled. On the other hand it means that an easy solution is available. While many news articles, projects and influencers have been informing about the dangers of opening email attachments, they might not have been targeting those who need this information the most. Many organizations would benefit from improved guidance regarding cyber risk. Department for Digital, Culture, Media & Sport, Her Majesty's Government (2019) Nonetheless, they are not actively seeking this information and they expect that regulators provide the information to them. Only 16 percent of businesses and 11 percent of charities have set up formal cyber security procedures. A proper response to cyber threats involves action from all participants in the economy and in all industries.

Using encryption in an organization is the largest factor for lowering data breach cost (IBM Security and Ponemon Institute 2019). Other factors include threat intelligence sharing and integrating security protection into software development process. Using encryption should go without saying because it instantly disables a substantial amount of attacks. It is relatively easy and cheap to implement. Since the penetration of encryption use is relatively high, we can assume that organizations, which experienced high data breach costs and did not use encryption, also neglected other factors of cyber security protection. The more organizations widely use encryption, the less profitable it is to develop malware targeting this type of vulnerability. Since the use of encryption is so widespread and because it is such a critical component of software, the encryption libraries are well maintained. Therefore it would be natural to see a

growth in the attacks targeting specific vulnerabilities of encryption protocols. Using encryption by an organization has the potential to decrease the cost of a data breach by USD 360,000 on average. Thales (2019) found in their survey that only a half of organizations currently use encryption in some way while almost all are planing to start using it in the next 12 months. And only 30 percent of organizations use file encryption in most of their applications (IBM Security and Ponemon Institute 2019).

Sharing security information is a reasonable protection against cyber attacks. Attackers might design an attack in such a way that it does not target a specific organization but specific software. Depending on how well that particular software is maintained, this approach can be more efficient for the attacker. However, sharing threat intelligence is a powerful tool against this attack. Nevertheless, it always depends on the purpose for which the information obtained from other sources is used. Of equal significance is how much willingness other organizations using the same software product have to share their data. Embedding security into software should be considered a necessity. However given the pace of some industries, it sounds impossible to give integrating security into development process a high priority. If embedding security into the software by the same team which makes the software decreases the speed of software production to such extent that competitiveness of the company is threatened then it makes sense considering if it is worth it. Integrating security protection into newly written software still does not solve issues with legacy software. It is always a difficult decision whether to keep using old software and patch it as often as possible or to create new software with security in mind but potentially introducing new security vulnerabilities. Department for Digital, Culture, Media & Sport, Her Majesty's Government (2019) reports that in 2019 more organizations improved their cyber security than a year ago.

Chapter 3

Literature review

Existing literature related to cyber risk research can be divided into these categories:

1. studies on cyber risk,
2. studies on general operational risk,
3. theoretical studies and monographies, and
4. reports on cost of data breaches and reports on cyber risk.

3.1 Studies on cyber risk

Studies focused primarily on cyber risk are rare and not all of them use copulas. On the other hand, there are many studies on general operational risk modelling which can also be used in the context of cyber risk. The core studies on cyber risk include Eling and Wirfs (2019), Mukhopadhyay et al. (2013), Biener, Eling, and Wirfs (2014) and H. Herath and T. Herath (2011).

Eling and Wirfs (2019) study cyber risk with an actuarial model. They use extreme value theory, but no copulas. Mukhopadhyay et al. (2013) do a meta study on the feasibility of cyber risk insurance. Biener, Eling, and Wirfs (2014) analyse cyber risk from the point of view of cyber risk insurance. H. Herath and T. Herath (2011) calculate cyber risk premiums with the use of copulas. Shah (2016) propose a cyber risk index that would facilitate trade with cyber risk. Shah (2016) use copulas to model dependence.

Eling and Wirfs (2019) criticise other research papers on cyber risk for being limited to data breach risk. At the same time Eling and Jung (2018),

by assessing purely data breach risk, do exactly what Eling and Wirfs (2019) criticise.

Aldasoro et al. (2020) use a comprehensive dataset of losses due to cyber risk to investigate which variables have effect on the cost of cyber attacks. They have enough evidence to say that both a development of technological skills by staff and a higher reliance on cloud technology reduces the cost. They also recommend an optimal level of investment into information technology security for each sector and compare it with actual spending.

3.2 Studies on general operational risk

Main studies on operational risk in general include Di Clemente and Romano (2004), Valle, Fantazzini, and Giudici (2008), Abbate, Gourier, and Farkas (2009) and Chavez-Demoulin, Embrechts, and Hofert (2016).

Di Clemente and Romano (2004) represents a typical research paper on operational risk. The use of copulas makes it exceptional. It starts with an overview of extreme value theory and copula theory. The part dealing with extreme value theory closely follows Embrechts, Klüppelberg, and Mikosch (1997) as most papers on operational risk do. A small disadvantage might be slightly vague explanation of how aggregate loss distribution is calculated. Valle, Fantazzini, and Giudici (2008) are similar to Di Clemente and Romano (2004). They discovere that the choice of margins has larger impact on risk measures than the choice of copula. Abbate, Gourier, and Farkas (2009) show that operational risk models can have high sensitivity to the parameters of the generalized Pareto distribution. Moreover they claim that full dependence structure produces higher value at risk than when the dependence structure is modelled with copulas. Di Clemente and Romano (2004) claim the opposite that under copula dependence structure the risk measures are on average 10 percent lower than under full dependence.

Both Chavez-Demoulin, Embrechts, and Hofert (2016) and Chavez-Demoulin, Embrechts, and Nešlehová (2006) build an operational risk model with application of peaks over threshold method and they explain the application of extreme value theory in a great detail. They do not use copulas. Chapelle et al. (2008) explain a gap between a robust mathematical models on one hand and simplified pragmatic approaches on the other hand. Gençay, Selçuk, and Ulugülyağci (2003) introduces an approach combining extreme value theory, value at risk and GARCH model. Jarrow (2008) criticises the operational risk

models based on distribution of aggregate losses. Y. Wang, Li, and Zhu (2017) improves traditional model for operational risk with piecewise loss frequency and severity distribution functions.

Carrillo-Menéndez and Suárez (2012) analyse pitfalls of operational risk modelling. They point out insufficiency of relevant data and they question the suitability of extreme value theory. Han, W. Wang, and J. Wang (2015), Lu (2011), Yao, Wen, and Luan (2013) and Xu et al. (2019) analyse capital requirements for operational risk in Chinese banks. Pan et al. (2019) combine copula theory with Bayesian approach. Even though Bhatti and Do (2019) do not focus on operational risk, they build a robust model with copula.

3.3 Theoretical studies and monographies

The core theoretical studies include Ghosh and Resnick (2010), Darling (1957), Braun (1980), Anderson and Darling (1954), Rockafellar and Stanislav Uryasev (2000) and Artzner et al. (1999).

Using mean excess plots correctly is challenging. Ghosh and Resnick (2010) start with definitions and then they give practical advice for using mean excess plot in empirical research. They also warn against common mistakes. Watson (1958) explains the use of Pearson's chi-squared test for testing goodness of fit. Braun (1980) introduces a novel approach for goodness of fit tests which allows using distribution functions with estimated parameters. Anderson and Darling (1954) introduces the well-known Anderson-Darling goodness of fit test. Darling (1957) provide description of Kolmogorov-Smirnov and Cramér-von Mises goodness of fit tests. Anderson and Darling (1952) provide further description of goodness of fit tests. Stephens (1974) give formulas for calculating goodness of fit test statistics. Rockafellar and Stanislav Uryasev (2000) introduce conditional value at risk and advocate about its advantages. Stan Uryasev (2010) in detail explains properties of conditional value at risk. Artzner et al. (1999) introduce coherent risk measures. Yamai and Yoshiba (2002a) and Yamai and Yoshiba (2002b) describe differences between risk measures and advantages of coherent risk measures.

Several monographies related to operational risk, actuarial models, extreme value theory and copula theory are also available. Chernobai, Rachev, and Fabozzi (2007) provide simple instructions for modelling operational risk. It includes an overview of applicable frequency and severity loss distribution families, extreme value theory and risk measures. Chernobai, Rachev, and

Fabozzi (2007) often do not go into details, but there are always references to research paper. In the end of each chapter there is always a review of empirical research papers related to the topic of the chapter. Embrechts, Klüppelberg, and Mikosch (1997) is one of the first books on extreme value theory. The book contains a large amount of various figures which complement otherwise very technical content. Only relatively small part of the book is concerned with peaks over threshold method. It is still more than enough for the purpose of application of this method in our model.

Hofert et al. (2018) present definition of copula and related terms along with plenty of motivational examples and figures. Furthermore they guide the reader through the process of copula parameter estimation, copula family selection, goodness of fit testing and graphical representation. Hofert et al. (2018) has two advantages. First, R code is provided for each example which simplifies practical application of concepts presented in the book. Second, they start with simple ideas and gradually extend them. Especially useful is their attention to detail. They cover a broad range of topics related to application of copulas in empirical research. Nelsen (2006) is more technical than Hofert et al. (2018) and therefore not intended as an introduction into copula theory. On the other hand, this might be useful in certain situations when simple explanations in other resources are too simple to explain complicated ideas. Klugman, Panjer, and Willmont (2012) provide introduction into actuarial models. The book is intended for a broad audience. Mejstřík, Pečená, and Teplý (2015) explain many concepts in finance and banking. Few pages are related to value at risk and risk assessment in general. D'Agostino and Stephens (1986) is a traditional book on goodness of fit tests. They provide a comprehensive comparison of tests in regards to their power in various situations.

3.4 Reports on cost of data breaches and reports on cyber risk

IBM Security and Ponemon Institute (2019) and Verizon (2020) are two most comprehensive reports on data breach costs. We use the cost from IBM Security and Ponemon Institute (2019) to test our hypothesis regarding the size of data breach risk. IBM Security and Ponemon Institute (2019) are more serious and organized while Verizon (2020) tries to put more weight on a few surprising findings. Verizon (2018) and Verizon (2019) are older versions of Verizon (2020).

Reports on cyber risk include Lloyd's and Cyence (2017), Department for Digital, Culture, Media & Sport, Her Majesty's Government (2019) and Thales (2019). Lloyd's and Cyence (2017) focus on cyber attacks from the perspective of cyber insurance. The purpose of their publication was to help insurers quantify cyber risk. They create two scenarios and calculate possible losses. The scenarios are cloud service provider hack and mass vulnerability attack. The first scenario can result into a loss between USD 4.6 and 53.1 billion. The amount for the second scenario is between USD 9.7 and 28.7 billion.

Department for Digital, Culture, Media & Sport, Her Majesty's Government (2019) represents a survey of organizations in the United Kingdom about their experience with cyber attacks. Thales (2019) is a summary of a survey which targeted executive officers. Presented statistics are related to data security. Gaidosch et al. (2019) and Bouveret (2018) provide recommendations from a perspective of a financial regulator in regards to cyber attack prevention.

Chapter 4

Methodology

4.1 Frequency and severity distributions

In order to measure the data breach risk we first build a model, second we calibrate the model with data of past loss events and finally we calculate risk measures. We use a parametric approach for modelling loss frequency and loss severity distributions. For loss frequency we consider Poisson and negative binomial distributions. For loss severity we consider normal, exponential, log-normal, Weibull and Cauchy distributions.

For random variable X following Poisson distribution with parameter $\lambda \geq 0$ we have

$$\mathrm{P}\left(X = x\right) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ...$$

For random variable X following negative binomial distribution with parameters $n > 0$ and $p \in (0, 1]$ we have

$$\mathrm{P}\left(X = x\right) = \frac{\Gamma\left(x + n\right)}{\Gamma\left(n\right) x!} p^n \left(1 - p\right)^x, \quad x = 0, 1, 2, ...$$

Normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ has density function

$$\mathrm{f}\left(x\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Figure 4.1 shows a plot of density function of normal distribution with two different choices of parameters.

Exponential distribution with parameter $\lambda > 0$ has density function

$$\mathrm{f}\left(x\right) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Figure 4.1: Density function of normal distribution with two different choices of parameters

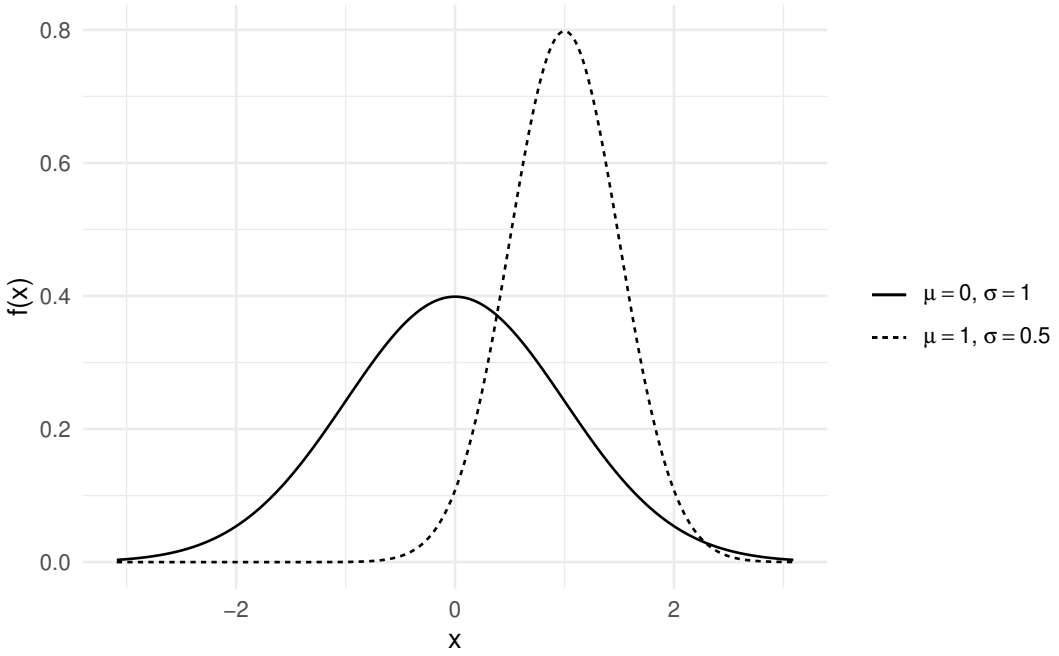
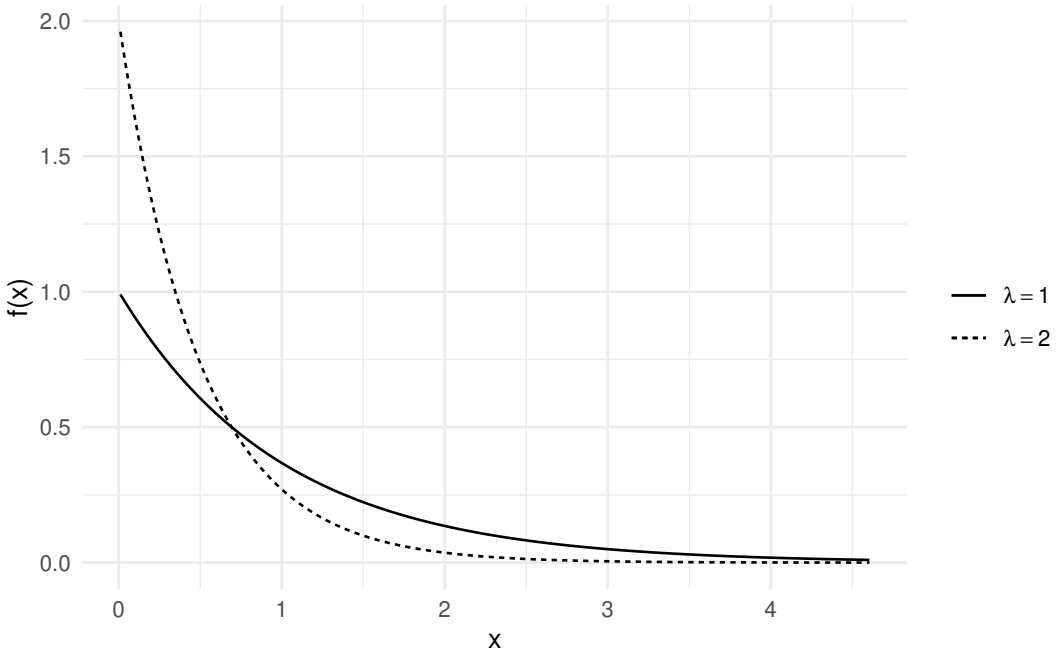


Figure 4.2 shows a plot of density function of exponential distribution with two different choices of parameter.

Figure 4.2: Density function of exponential distribution with two different choices of parameter



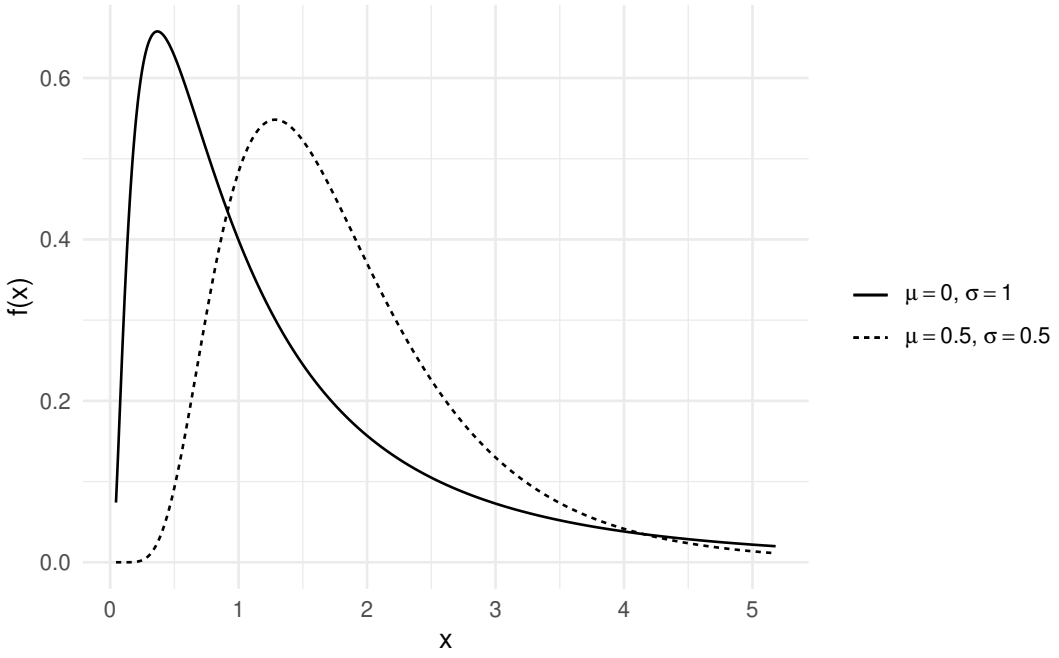
Log-normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ has density

function

$$\mathrm{f}(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Figure 4.3 shows a plot of density function of log-normal distribution with two different choices of parameters.

Figure 4.3: Density function of log-normal distribution with two different choices of parameters



Weibull distribution with shape parameter $a > 0$ and scale parameter $\sigma > 0$ has density function

$$\mathrm{f}(x) = \frac{a}{\sigma}\left(\frac{x}{\sigma}\right)^{a-1} \exp\left(-\left(\frac{x}{\sigma}\right)^{a}\right), \quad x > 0.$$

Cauchy distribution with location parameter $l > 0$ and scale parameter $s > 0$ has density function

$$\mathrm{f}(x) = \frac{1}{\pi s}\left(1 + \left(\frac{x-l}{s}\right)^2\right)^{-1}, \quad x \in \mathbb{R}.$$

Figure A.1 and Figure A.2 in Appendix show plots of density functions of Weibull and Cauchy distributions with two different choices of parameters.

It is a stylized fact that severity distributions of operational loss data have heavy tails. Normal and exponential distributions are light tailed. Log-normal

distribution is moderately heavy tailed. Weibull and Cauchy distributions are heavy tailed (Chernobai, Rachev, and Fabozzi 2007).

4.2 Goodness of fit tests

Suppose $X = X_1, \ldots, X_n$ is a random sample from a distribution with a distribution function $\mathrm{F_X}$. Then

$$X_{(1)} < X_{(2)} < \cdots < X_{(n-1)} < X_{(n)}$$

is ordered random sample and

$$\mathrm{F_n}(u) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\{X_i \leq u\}$$

is empirical distribution function.

Suppose F is a fully specified distribution function. We can use a goodness of fit test for the following null and alternative hypotheses:

$$\begin{aligned} &\mathrm{H_0}\colon\ \mathrm{F}_X(x) = \mathrm{F}(x)\ \ \forall x \in \mathbb{R}, \\ &\mathrm{H_1}\colon \exists x \in \mathbb{R} : \mathrm{F}_X(x) \neq \mathrm{F}(x). \end{aligned}$$

We use Kolmogorov-Smirnov, Anderson-Darling and Cramér-von Mises goodness of fit tests to determine which severity and frequency distributions provide the best fit to the data.

A statistic measuring the difference between $\mathrm{F_n}(x)$ and $\mathrm{F}(x)$ is called empirical distribution function (EDF) statistic. We work with three tests based on different EDF statistics. The null hypothesis is always rejected when the test statistic exceeds corresponding critical value. First, the Kolmogorov-Smirnov test described for example by Darling (1957) has test statistic

$$\mathrm{D} = \sqrt{n} \sup_{x \in \mathbb{R}} \mid \mathrm{F_n}(x) - \mathrm{F}(x) \mid = \sqrt{n} \max\left(\mathrm{D}^+, \mathrm{D}^-\right)$$

where

$$\begin{aligned} \mathrm{D}^+ &= \sup_{x \in \mathbb{R}} \left(\mathrm{F_n}(x) - \mathrm{F}(x)\right), \\ \mathrm{D}^- &= \sup_{x \in \mathbb{R}} \left(\mathrm{F}(x) - \mathrm{F_n}(x)\right). \end{aligned}$$

D^+ is the largest difference between $F_n(x)$ and $F(x)$ while D^- is the largest difference between $F(x)$ and $F_n(x)$ (Chernobai, Rachev, and Fabozzi 2007).

Second, the Anderson-Darling test invented by Anderson and Darling (1952) has test statistic

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x).$$

Anderson and Darling (1954) argues that the Anderson-Darling test is "sensitive to discrepancies at the tails of the distribution rather than near the median."

Finally, the Cramér-von Mises test described for instance by Darling (1957) has test statistic

$$W^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x).$$

D'Agostino and Stephens (1986) recommend using the Anderson-Darling test instead of the Cramér-von Mises test because the former is more powerful when F departs from F_X in the tails. Nonetheless, if the distribution tail is modelled separately this can also be a disadvantage. For instance this can happen if the extreme value theory is applied to the severity distribution. In general, the Kolmogorov-Smirnov test is often the least powerful among all three considered goodness of fit tests (D'Agostino and Stephens 1986).

Stephens (1974) provides formulas for calculation of the test statistics from random sample:

$$\begin{aligned}
D &= \sqrt{n} \max\left(\max_{1 \le i \le n} \left(\frac{i}{n} - F\left(X_{(i)}\right) \right), \max_{1 \le i \le n} \left(F\left(X_{(i)}\right) - \frac{i-1}{n} \right) \right), \\
A^2 &= -n - \frac{1}{n} \sum_{i=1}^{n} (2i - i) \left(\log\left(F\left(X_{(i)}\right)\right) + \log\left(1 - F\left(X_{(n+1-i)}\right)\right) \right), \\
W^2 &= \frac{1}{12n} + \sum_{i=1}^{n} \left(F\left(X_{(i)}\right) - \frac{2i-1}{2n} \right)^2.
\end{aligned}$$

If the distribution function F has parameters estimated from a random sample then these tests might have lower power to reject the null hypothesis than if the distribution function was fully specified. In other words, if F has estimated parameters, the sampling distribution of the test statistics are different from those presented above. Let $F_{G(X)}$ be the distribution function with estimated parameters which depends on the random sample. For instance

the test statistic of the Kolmogorov-Smirnov test becomes

$$\mathrm{D} = \sqrt{n} \sup_{x \in \mathbb{R}} \mid \mathrm{F_n}\left(x\right) - \mathrm{F_{G(X)}}\left(x\right) \mid ,$$

which makes finding a closed form formula for the p-value practically impossible.

Goodness of fit tests represent a specific statistical method. Consider using a test for testing the same null hypothesis as the t-test with the same assumptions as the t-test, but with lower power than the t-test. Such test would require more evidence against the null hypothesis in order to reject it than the t-test. It might not be beneficial to use such test, but it would certainly not be a problem from a methodological point of view. This is not true for goodness of fit tests. When we use the t-test we usually draw a useful conclusion after we reject the null hypothesis. With goodness of fit tests we draw a conclusion that a particular distribution is suitable for a particular random sample when we do not reject the null hypothesis. Therefore using a goodness of fit test with lower power than conventional tests can lead to misleading conclusions.

Braun (1980) proposes a modification of goodness of fit tests which solves the previously described problem. The suggested method works with distribution functions with estimated parameters. The idea is to randomly divide data into a large number of groups of equal size. The test statistic is calculated for each group separately. If there are m groups and we aim for significance level α then the test statistic in each group is compared with $(1-\alpha)/m$ quantile of sampling distribution of test statistic. The composite null hypothesis is rejected if the null hypothesis in any group is rejected. We always use this method with Anderson-Darling and Cramér-von Mises tests.

Other solution is to use the Pearson's chi-squared test studied by Watson (1958) which does not suffer from this disadvantage. It is based on a proposition that observed frequencies of some events should follow theoretical frequencies. Pearson's chi-squared test usually performs worse than any of the three previously discussed tests based on the empirical distribution function (D'Agostino and Stephens 1986). The reason is that this test needs to group data into bins. It constitutes a loss of information which results into a drop in the test power. Other disadvantage of grouping data into bins is that there is not any rule for choosing the number and length of bins. Situation might be easier for discreet probability distributions where the bins can be equivalent with the values that a random variable following this distribution attains. Nonetheless, for continuous distribution the choice of bins is not natural at all. Thus the

test result substantially depends on the researcher's judgement. It opens an opportunity for the researcher to manipulate the results. Therefore we do not use this test.

4.3 Extreme value theory

We can use extreme value theory (EVT) to describes high severity losses which happen with low frequency. "EVT is considered as a useful set of tools for analysing rare events" (Chavez-Demoulin, Embrechts, and Nešlehová 2006). We use an approach of extreme value theory called peaks over threshold (POT) to model the right tail of loss severity distribution. This approach is based on exceedances of high thresholds (Embrechts, Klüppelberg, and Mikosch 1997).

Let F be a distribution function. Then $\overline{\mathrm{F}}(x) = 1 - \mathrm{F}(x)$ for $x \geq 0$ is the tail of the distribution function F.

The following steps loosely follow Embrechts, Klüppelberg, and Mikosch (1997). Suppose $X_1, \ldots X_n$ are independent identically distributed random variables following a distribution with distribution function F. Suppose $u > 0$ is a high threshold. Then

$$N_u = \sum_{i=1}^{n} \mathrm{I}\{X_i > u\}$$

is the number exceedances. Let $Y_1, \ldots, Y_{N_u}$ be the corresponding excesses. The excess distribution function of X is

$$\mathrm{F}_u(x) = \mathrm{P}(X - u \leq y \,|\, X > u) = \mathrm{P}(Y \leq y \,|\, X > u), \quad y \geq 0.$$

Using the definition of conditional probability we obtain

$$\overline{\mathrm{F}}(u + y) = \overline{\mathrm{F}}(u)\,\overline{\mathrm{F}}_u(y).$$

The distribution function of the generalized Pareto distribution (GPD) $\mathrm{G}_{\xi,\beta}$ with parameters $\xi \in \mathbb{R}$ and $\beta > 0$ has the following tail:

$$\overline{\mathrm{G}}_{\xi,\beta}(x) = \begin{cases} \left(1 + \frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}} & \xi \neq 0, \\ e^{-\frac{x}{\beta}} & \xi = 0, \end{cases} \qquad x \in \mathrm{D}(\xi, \beta)$$

where

$$\mathrm{D}(\xi,\beta) = \begin{cases} [0,\infty) & \xi \geq 0, \\ \left[0, -\frac{\beta}{\xi}\right) & \xi < 0. \end{cases}$$

We have for $\overline{\mathrm{F}}_u(y)$:[1]

$$\lim_{u\to\infty} \sup_{0<x<\infty} |\,\overline{\mathrm{F}}_u(x) - \overline{\mathrm{G}}_{\xi,\beta(u)}(x)\,| = 0.$$

Therefore we can use an approximation

$$\overline{\mathrm{F}}_u(y) \approx \overline{\mathrm{G}}_{\xi,\beta(u)}(y).$$

A natural estimator for $\overline{\mathrm{F}}(u)$ is an empirical distribution function

$$\widehat{\overline{\mathrm{F}}(u)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{X_i > u\} = \frac{N_u}{n}$$

and for $\overline{\mathrm{F}}_u(y)$ it is

$$\widehat{\overline{\mathrm{F}}_u(y)} = \overline{\mathrm{G}}_{\hat{\xi},\hat{\beta}}(y)$$

where $\hat{\xi}$ and $\hat{\beta}$ are parameters estimated from fitting GPD to excesses $Y_1, \ldots, Y_{N_u}$. In other words, we use only data above the threshold u to fit the generalized Pareto distribution. Finally, combining the previous two results we get an estimator for $\overline{\mathrm{F}}(u+y)$ for $y > 0$:

$$\widehat{\overline{\mathrm{F}}(u+y)} = \frac{N_u}{n}\left(1 + \frac{\hat{\xi}y}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}.$$

Substituting x for $u+y$ in the previous equation and solving for x yields the following estimator for quantile $\mathrm{q}_t(p)$:

$$\widehat{\mathrm{q}_t(p)} = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{n}{N_u}(1-p)\right)^{-\hat{\xi}} - 1\right).$$

Suppose that body and tail of loss severity distribution are modelled with different distribution functions. Di Clemente and Romano (2004) combine body and tail distribution functions into a piecewise function. We use a similar approach for combining body and tail quantile functions. Combining body

[1] For details see Theorem 3.4.13(b) in Embrechts, Klüppelberg, and Mikosch (1997).

quantile function q_b and tail quantile function $\widehat{\mathrm{q}_t}$ gives quantile function

$$\mathrm{q}(p) = \begin{cases} \mathrm{q}_b(p) & p \leq 1 - \frac{N_u}{n}, \\ \widehat{\mathrm{q}_t(p)} & p > 1 - \frac{N_u}{n}. \end{cases}$$

We can use this quantile function when generating random numbers from the loss severity distribution.

The choice of the high threshold u is arbitrary. There is no universal approach. Nonetheless we can use a method based on mean excess function (Embrechts, Klüppelberg, and Mikosch 1997). Suppose X is a random variable following generalized Pareto distribution $\mathrm{G}_{\xi,\beta}$. The mean excess function of $\mathrm{G}_{\xi,\beta}$ is

$$\mathrm{e}(u) = \mathrm{E}(X - u | X > u) = \frac{\beta + \xi u}{1 - \xi}, \quad u \in \mathrm{D}(\xi, \beta).$$

For $\xi \in (0, 1)$ and $\beta + \xi u > 0$ the mean excess plot is linear and upward sloping. "The heavier the tail of the loss distribution (i.e., the closer ξ is to 1), the steeper the plot."(Chernobai, Rachev, and Fabozzi 2007) The empirical mean excess function is

$$\mathrm{e}_n(u) = \frac{1}{N_u} \sum_{i=1}^{n} \mathrm{I}\{X_i > u\}(X_i - u), \quad u > 0.$$

A plot of the mean excess function is called mean excess plot. Ghosh and Resnick (2010) suggest to choose such value of u that that the plot of $\mathrm{e}_n(x)$ is roughly linear for $x \geq u$.

4.4 Risk measures

We use a simple actuarial model detailed by Klugman, Panjer, and Willmont (2012) to measure the amount of global data breach risk. First we define a model for estimating aggregate losses in one industry. Later we extend this model so that it captures the dependence structure between aggregate losses in different industries. This does not affect how risk measures are calculated. What follows is a definition of the former model.

Let N be a random variable representing the number of loss events during one week, and let $X_1, \ldots, X_N$ be random variables representing loss amounts. Then

$$S = \sum_{i=1}^{N} X_i$$

is a random variable representing the total or aggregate losses during one week. This model has the following assumptions (Klugman, Panjer, and Willmont 2012):

1. conditional on $N = n$, $X_1, ..., X_n$ are independent identically distributed random variables,

2. conditional on $N = n$, the multivariate probability distribution of $X_1, ..., X_n$ does not depend on n and

3. the probability distribution of N does not depend on any of $X_1, X_2, ...$

Simply put, we assume that loss events happen independently. We also assume that the probability that a loss event happens and the amount of that loss do not depend on each other. Random variable N follows the loss frequency distribution. Random variables $X_1, ..., X_N$ follow the loss severity distribution. And random variable S follows an aggregate loss distribution.

This model enables us to model loss frequency and loss severity distributions separately. We use a parametric approach for these distributions. Which means that for each considered distribution family we estimate distribution parameters from the data. Then with the help of goodness of fit tests and experience we choose the most suitable distribution families for loss frequency and loss severity distributions. These two distributions are then aggregated into an aggregate loss distribution with the previously described model. Figure 4.4 illustrates this principle. It tries to explain the essence of the model.

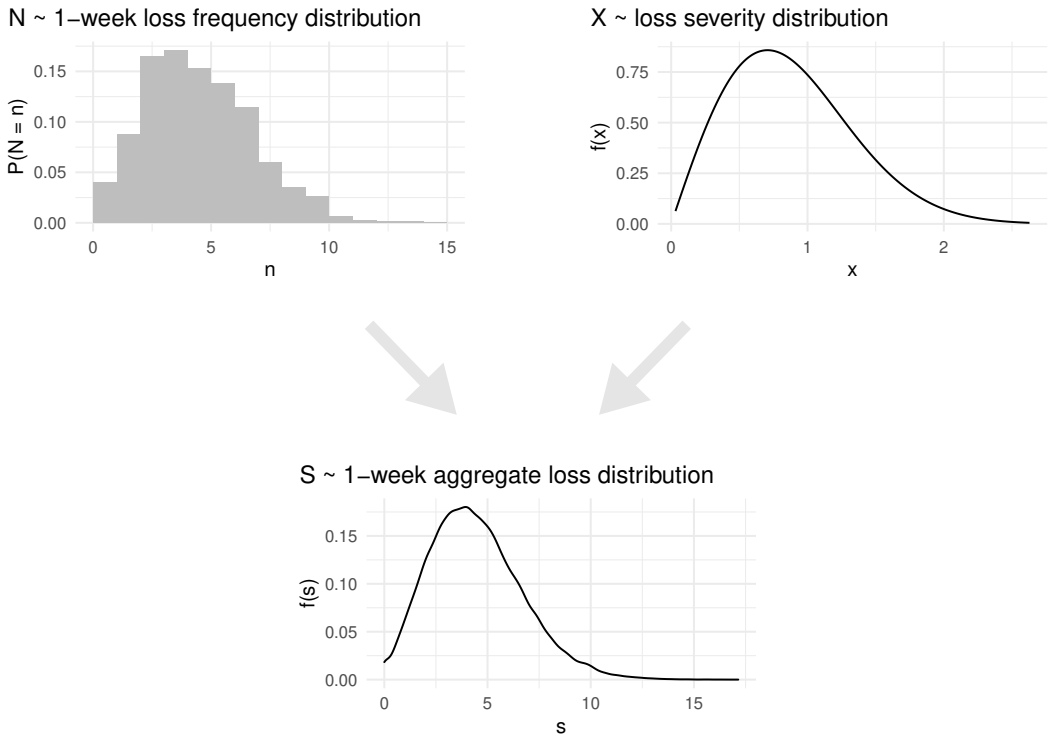
We calculate risk measures of random variable S which follows the distribution of aggregate losses. We consider two risk measures. They are value at risk (VaR) and conditional value at risk (CVaR). "Value at risk (VaR) is the maximum expected portfolio (asset) depreciation at a specified confidence level over a specified holding period" (Mejstřík, Pečená, and Teplý 2015). For instance, a one-year 99% VaR is such a quantity that a random variable, which is observed once a year, exceeds the quantity exactly once in every 100 years.

If random variable S has distribution function F and $\alpha \in [0, 1]$ then $\mathrm{VaR}_\alpha(S)$, i.e. VaR of S at confidence level $(100 \cdot \alpha)\%$ is the $(1 - \alpha)$ quantile of F:

$$\mathrm{VaR}_\alpha(S) = \mathrm{F}^{-1}(1 - \alpha).$$

Rockafellar and Stanislav Uryasev (2000) and Artzner et al. (1999) criticised VaR because it is not a coherent risk measure. Rockafellar and Stanislav Uryasev

Figure 4.4: The principle of aggregation of loss frequency and loss severity distributions. N follows Poisson distribution with parameter $\lambda = 5$ and X follows Weibull distribution with shape parameter 2 and scale parameter 1. Aggregate loss distribution is obtained from loss frequency and loss severity distributions with Monte Carlo simulation. For illustration purposes only, aggregate loss distribution is smoothed with Gaussian kernel density.



Source: Author based on Chernobai, Rachev, and Fabozzi (2007).

(2000) introduced CVaR which is defined by

$$\mathrm{CVaR}_{\alpha}(S) = \mathrm{E}\left(S \,|\, S \geq \mathrm{VaR}_{\alpha}(S)\right).$$

$\mathrm{CVaR}_{\alpha}(S)$ is CVaR of S at confidence level $(100 \cdot \alpha)\%$, $\alpha \in [0, 1]$. Yamai and Yoshiba (2002a): "Expected shortfall measures how much one can lose on average in states beyond the VaR level."

Rockafellar and Stanislav Uryasev (2000) and Stan Uryasev (2010) defined

VaR and CVaR more flexibly as

$$\mathrm{VaR}_\alpha(S) = \min(s \mid \mathrm{F}(s) \geq \alpha),$$
$$\mathrm{CVaR}_\alpha(S) = \int_{-\infty}^{\infty} s \, d\,\mathrm{F}^\alpha(s)$$

where

$$\mathrm{F}^\alpha(s) = \begin{cases} 0 & s < \mathrm{VaR}_\alpha(S), \\ \frac{\mathrm{F}(s)-\alpha}{1-\alpha} & s \geq \mathrm{VaR}_\alpha(S). \end{cases}$$

Artzner et al. (1999) introduced a coherent risk measure which describes properties that a regulator can reasonably require a risk measure to have. Suppose $L = \{X_1, X_2, ...\}$ is a set of all losses and $\rho(X)$ is a risk measure. Then ρ is a coherent risk measure if it satisfies the following four axioms:

1. **Translation invariance.** For $X \in L$ and $\alpha \in \mathbb{R}$ we have $\rho(X+\alpha) = \rho(X) - \alpha$. This means that by reducing the value of a risky asset by α we reduce the risk measure ρ by α.

2. **Sub-additivity.** For $X_1, X_2 \in L$ we have $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$. This means that "a merger does not create extra risk" (Artzner et al. 1999). In other words, the opposite effect of diversification does not exist. It is not possible to split certain risk into two separate entities and reduce the total risk. Such risk must be either the same or greater. Diversification still decreases the risk measure though.

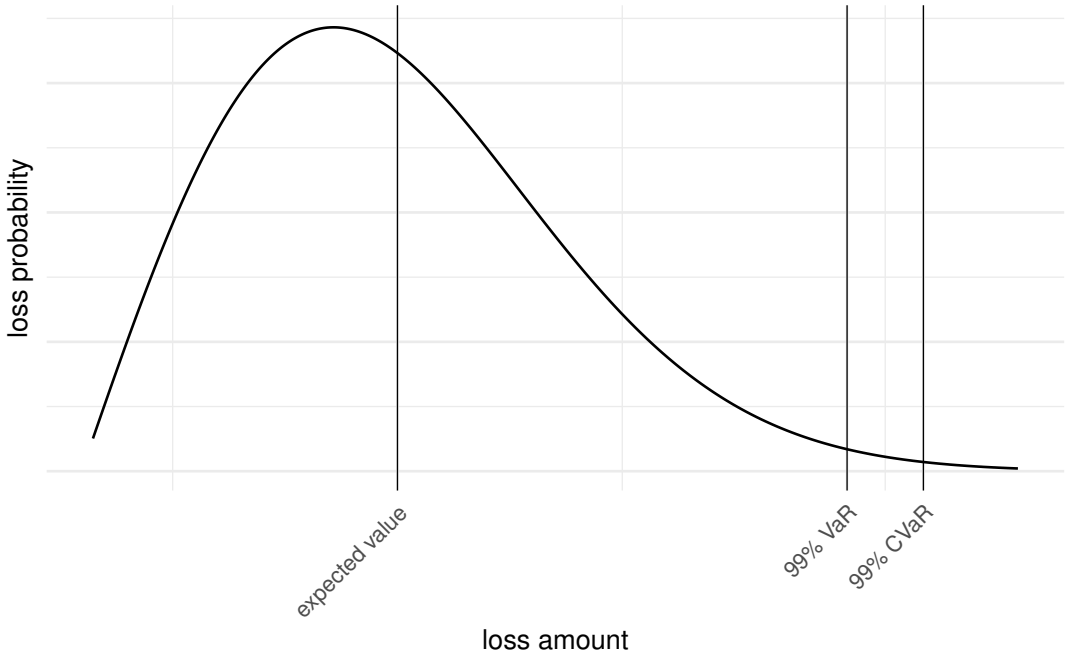
3. **Positive homogeneity.** For $\lambda \geq 0$ and $X \in L$ we have $\rho(\lambda X) = \lambda\rho(X)$. This means that a larger amount of the same type of risk increases the risk measure proportionally.

4. **Monotonicity.** For $X_1, X_2 \in L$ such that $X_1 \leq X_2$ we have $\rho(X_1) \leq \rho(X_2)$. This means that a greater possible loss must result into greater value of risk measure than a smaller possible loss does.

If the distribution of the aggregate losses S is not normal then VaR does not satisfy the axiom of sub-additivity and it is not a coherent risk measure. Contrarily, CVaR is always a coherent risk measure (Yamai and Yoshiba 2002a). Therefore in most cases CVaR should be preferred over VaR. On the other hand, simplicity of VaR might outweigh its deficiencies. Especially when the properties of coherent risk measures are not relevant for the discussed type

of risk. Yamai and Yoshiba (2002b): "One should not conclude that VaR is inappropriate only because it is not sub-additive, since sub-additivity itself may be irrelevant for risk managers." Figure 4.5 demonstrates the difference between VaR and CVaR.

Figure 4.5: Aggregate loss distribution and risk measures. One-week aggregate loss distribution is a probability distribution of the sum of future losses happening over a period of one week. Risk measures quantify the associated risk. Losses below and above expected value are called expected and unexpected losses, respectively. 99% value at risk (VaR) is 0.99 quantile of the aggregate loss distribution. 99% conditional value at risk (CVaR) is expected value of loss amounts exceeding 0.99 quantile of the aggregate loss distribution.



Source: Author based on Chernobai, Rachev, and Fabozzi (2007).

We should note that the following three parameters should be specified together with risk measure values: confidence level, forecast horizon, and unit of measurement (Chernobai, Rachev, and Fabozzi 2007). In our case the unit of measurement is number of breached records. Nonetheless, there are several studies which estimated the monetary cost of one breached record. For instance IBM Security and Ponemon Institute (2019) discovered in their 2019 study that one breached record results on average into USD 150 cost. Using this

information we can convert our estimates of risk measures measured in numbers of breached records into estimates of risk measures measured in USD.

The advantage of VaR is that it allows directly comparing various types of risk (Chernobai, Rachev, and Fabozzi 2007). We can use the following formula to convert between one-week and one-year VaR at the same confidence level:

$$\mathrm{VaR}_{\mathrm{year}} = \sqrt{52}\,\mathrm{VaR}_{\mathrm{week}}\,.$$

"Disclosure of quantitative measures of market risk, such as value at risk is enlightening only when accompanied by a thorough discussion of how the risk measures were calculated how they relate to actual performance" (Greenspan 1996). We should always remember that VaR and CVaR are just numbers which have no meaning unless we know exactly what these numbers represent and under what assumptions we calculated them.

We can improve the trustworthiness of VaR estimates by VaR backtesting. It involves testing adequacy of VaR in regards to its confidence level on historical data. One way how to do this is by using Kupiec's proportion of failures test studied by Chernobai, Rachev, and Fabozzi (2007). Let T be the total number of days in the VaR backtesting period. In our case it is the number of days in 6 years. Let N be the number of days when the total daily loss is greater than one-day VaR. N is also called a number of exceedances or violations. N is a sum of Bernoulli trials and therefore it has binomial distribution:

$$\mathrm{P}\left(N = x\right) = \binom{T}{x} p^x \left(1-p\right)^{T-x},\ x = 0, 1, 2, \ldots, T.$$

Under the null hypothesis, the empirical probability of exceedance N/T is equal to the population probability p from the previous equation. A likelihood-ratio test can be used to calculate confidence intervals for the numbers of exceedances N consistent with the null hypothesis. The confidence intervals for a period of 6 years are $(8, 24)$ and $(60, 94)$ for VaR at confidence levels 99% and 95%, respectively. This test is designed for one-day VaR, therefore we have to convert our one-week VaR estimates into one-day VaR estimates.

We use Monte Carlo simulations to obtain a distribution of weekly aggregate losses from which we can calculate risk measures. The algorithm goes as follows:

1. Generate $N = 100,000$ random numbers $n_1, \ldots, n_N$ from a frequency distribution.

2. For each $i = 1, \ldots, N$ generate n_i random numbers from a severity distribution and sum them up.

3. The resulting N numbers from the previous step constitute weekly aggregate losses.

The process of obtaining aggregate loss distribution can be described either as a calculation or as an estimation. It is a calculation because we perfectly understand the data generating process. It is a question if the underlying loss and severity distributions are estimated correctly, but let us leave that aside for now. The loss and severity distributions are specified with parametric distribution families with fixed parameters. By generating large number of random numbers from this aggregate loss distribution we obtain such a smooth estimate of the distribution function that we can say that we calculate the distribution function. It is also an estimation because we do not have a formula for the aggregate loss distribution and via an empirical distribution function we estimated the underlying distribution function of aggregate losses. Under this approach we consider the random numbers generated from the aggregate loss distribution to be observations.

4.5 Copulas

We model data breach risk in three industries. There are three options how we can do this. First, we can have one model for all industries combined and estimate risk measure from weekly aggregate losses of all industries combined. The second and third options require having a separate model for aggregate losses for each industry.

Second, we calculate risk measures from weekly aggregate losses for each industry individually. We sum these risk measures and we obtain one final risk measure. This option is called perfect dependence.

Third, we find dependence structure between weekly aggregate losses. Using this dependence structure we combine models for individual industries into one model of weekly aggregate losses. Then we calculate risk measure from these losses. We try all three options and compare them.

Dependence between two random vectors can be expressed in terms of a linear correlation coefficient. Two alternatives of the linear correlation coefficient are Spearman's rho and Kendall's tau. They are called rank correlation coefficients and unlike the linear correlation coefficient they always exist (Hofert et al. 2018).

Let (X_1, X_2) and (Y_1, Y_2) be independent random vectors with continuous marginal distribution functions F_1 and F_2. Then Spearman's rho is defined by

$$\rho_s\left(X_1, X_2\right) = \mathrm{Cor}\left(\mathrm{F}_1\left(X_1\right), \mathrm{F}_2\left(X_2\right)\right)$$

and Kendall's tau is defined by

$$\tau(X_1, X_2) = \mathrm{E}\left(\mathrm{sign}\left(\left(X_1 - Y_1\right)\left(X_2 - Y_2\right)\right)\right).$$

Hofert et al. (2018) argue that "a copula is a multivariate distribution function with standard uniform univariate margins, that is, $\mathrm{U}(0,1)$ margins."

Sklar's theorem is the fundamental building block of copula theory. It is necessary for decomposing multivariate distribution into copula and margins and also for generating random numbers from multivariate distributions described by a copula and margins.

Nelsen (2006) presented Sklar's theorem as follows. Let F be a d-dimensional continuous multivariate distribution function with marginal distribution functions $\mathrm{F}_1, \ldots, \mathrm{F}_\mathrm{d}$. Then there exists a d-dimensional copula C such that

$$\mathrm{F}\left(x\right) = \mathrm{C}\left(\mathrm{F}_1\left(x_1\right), \ldots, \mathrm{F}_\mathrm{d}\left(x_d\right)\right), \quad x = \left(x_1, \ldots, x_d\right) \in \mathbb{R}^d$$

and

$$\mathrm{C}\left(u\right) = \mathrm{F}\left(\mathrm{F}_1^{\text{-}1}\left(u_1\right), \ldots, \mathrm{F}_\mathrm{d}^{\text{-}1}\left(u_d\right)\right), \quad u = \left(u_1, \ldots, u_d\right) \in [0,1]^d.$$

The dependence structure between industries can be described with a copula. We consider an independence copula and two elliptical copulas: normal and t. We also consider four Archimedean copulas: Clayton, Frank, Gumbel-Hougaard and Joe.

Independence copula is defined by

$$\Pi\left(u\right) = \prod_{i=1}^{d} u_i, \quad u = \left(u_1, \ldots, u_d\right) \in [0,1]^d.$$

In other words, it is a multivariate distribution function of a random vector with standard uniform univariate margins.

Let P be a correlation matrix. Let Φ_P and Φ^{-1} be distribution function of $\mathrm{N}_\mathrm{d}\left(0, P\right)$ and quantile function of $\mathrm{N}\left(0,1\right)$, respectively. Then Nelsen (2006)

defines normal copula $\mathrm{C}_{\mathrm{P}}^{\mathrm{n}}$ by

$$\begin{aligned}\mathrm{C}_{\mathrm{P}}^{\mathrm{n}}(u) &= \Phi_P\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right) \\ &= \int_{-\infty}^{\Phi^{-1}(u_d)} \ldots \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{\exp\left(-\frac{1}{2}x' P^{-1} x\right)}{(2\pi)^{\frac{d}{2}}\sqrt{\det P}}\, dx_1 \ldots dx_d.\end{aligned}$$

In other words, we obtain normal copula by applying Sklar's Theorem to distribution function of multivariate normal distribution and quantile function of univariate normal distribution.

The t copula is obtained similarly. Let $\mathrm{t}_{P,\nu}$ be a multivariate t distribution with correlation matrix P and ν degrees of freedom. Let t_{ν}^{-1} be a quantile function of univariate Student t distribution with ν degrees of freedom. Then Nelsen (2006) defines t copula $\mathrm{C}_{P,\nu}^{t}$ by

$$\begin{aligned}\mathrm{C}_{P,\nu}^{t}(u) &= \mathrm{t}_{P,\nu}\left(\mathrm{t}_{\nu}^{-1}(u_1), \ldots, \mathrm{t}_{\nu}^{-1}(u_d)\right) \\ &= \int_{-\infty}^{\mathrm{t}_{\nu}^{-1}(u_d)} \ldots \int_{-\infty}^{\mathrm{t}_{\nu}^{-1}(u_1)} \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{d}{2}}\sqrt{\det P}} \left(1 + \frac{x' P^{-1} x}{\nu}\right)^{-\frac{\nu+d}{2}} dx_1 \ldots dx_d.\end{aligned}$$

Figure 4.6 and Figure 4.7 contain two panels. Figure 4.6 is for normal copula and Figure 4.7 is for t copula. First panel displays a wireframe plot of copula density function with such a value of the parameter ρ that Kendall's tau of observations from this copula is $\tau = 0.5$. Second panel depicts a sample of size $n = 1000$ from this copula.

For normal and t copula we use the following symmetric correlation matrix

$$P = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}, \quad \text{where } \rho \in \mathbb{R}.$$

Nelsen (2006) defines an Archimedean copula by

$$\mathrm{C}(u) = \psi\left(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)\right), \quad u = (u_1, \ldots, u_d) \in [0,1]^d$$

where ψ is a function defined as follows. For Clayton copula

$$\psi(t) = (1+t)^{-\frac{1}{\theta}}, \quad \theta \in (0, \infty).$$

Figure 4.6: Wireframe plot of density function of normal copula with a Kendall's tau $\tau = 0.5$ and a sample of size $n = 1000$ from the same copula

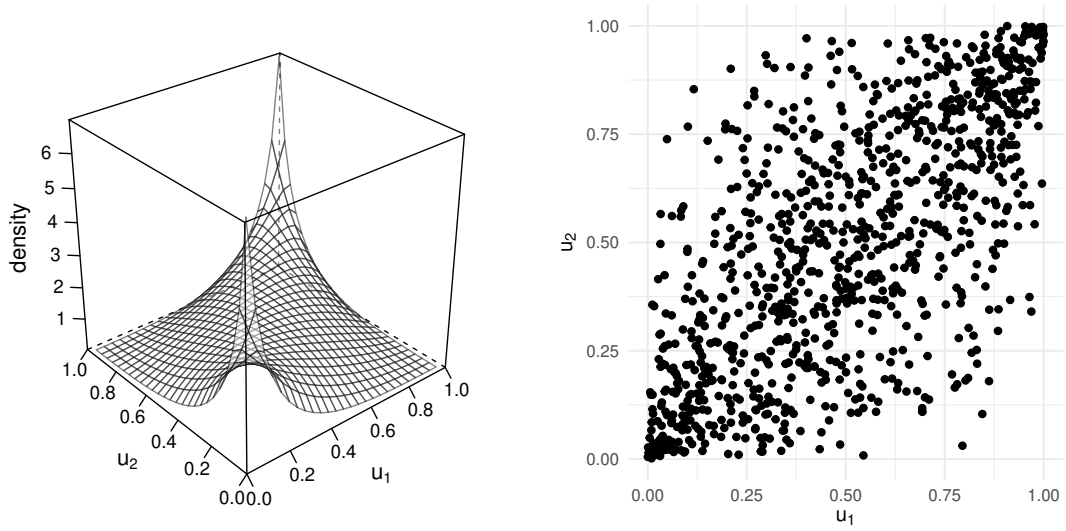
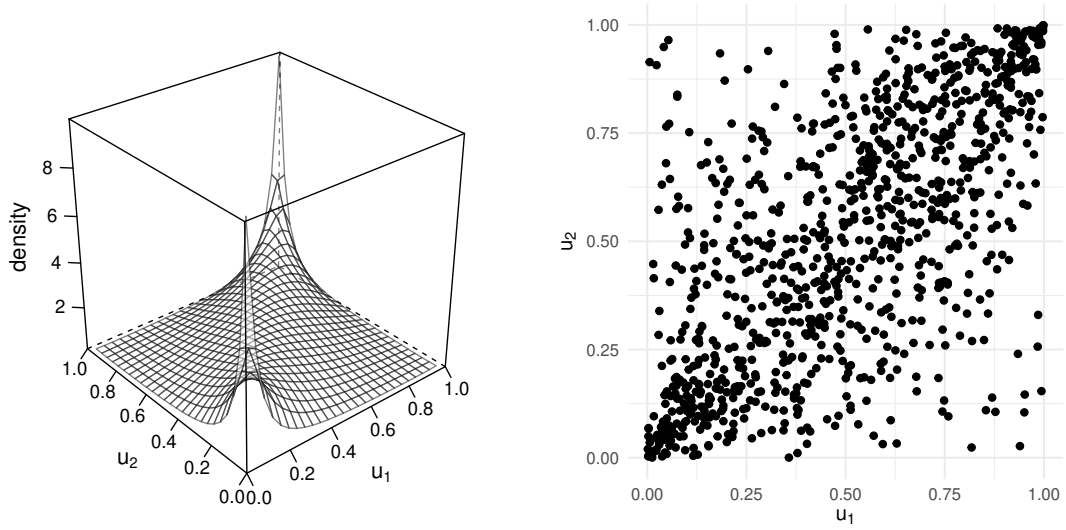


Figure 4.7: Wireframe plot of density function of t copula with a Kendall's tau $\tau = 0.5$ and a sample of size $n = 1000$ from the same copula



For Frank copula

$$\psi(t) = -\frac{1}{\theta}\log\left(1-\left(1-e^{-\theta}\right)e^{-t}\right), \quad \theta \in (0,\infty).$$

For Gumbel-Hougaard copula

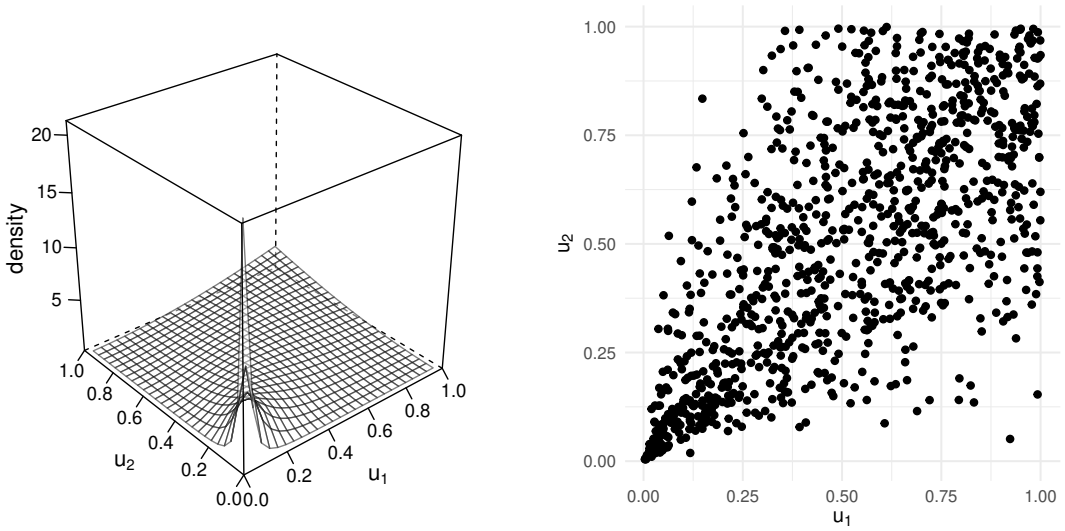
$$\psi(t) = \exp\left(-t^{\frac{1}{\theta}}\right), \quad \theta \in [1,\infty).$$

And for Joe copula

$$\psi(t) = 1 - (1 - e^{-t})^{\frac{1}{\theta}}, \quad \theta \in [1, \infty).$$

Figure 4.8 contains two panels. First panel displays a wireframe plot of Clayton copula density function and second panel displays a sample from this copula. Figure A.3, Figure A.4 and Figure A.5 in Appendix provide similar representations for Frank, Gumbel-Hougaard and Joe copulas, respectively.

Figure 4.8: Wireframe plot of density function of Clayton copula with a Kendall's tau $\tau = 0.5$ and a sample of size $n = 1000$ from the same copula



Suppose $X_1, \ldots, X_n$ are d-dimensional random vectors with marginal distribution functions $\mathrm{F}_1, \ldots, \mathrm{F}_\mathrm{d}$. These marginal distribution functions can be estimated with (Hofert et al. 2018)

$$\mathrm{F}_{\mathrm{n,j}}(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathrm{I}\{X_{ij} \leq x\}, \quad x \in \mathbb{R}.$$

We can then use these empirical distribution functions to create a sample of pseudo-observations

$$U_{i,n} = \left(\mathrm{F}_{\mathrm{n},1}(X_{i1}), \ldots, \mathrm{F}_{\mathrm{n,d}}(X_{id})\right), \quad i \in \{1, \ldots, n\}.$$

Finally, we can use these pseudo-observations to construct a maximum

pseudo-likelihood estimator for estimating copula parameters

$$\theta_n = \arg\sup_{\theta \in \Theta} \sum_{i=1}^{n} \log \mathrm{c}_\theta \left(U_{i,n} \right)$$

where c_θ is copula density function.

We use a copula goodness of fit test proposed by Genest, Rémillard, and Beaudoin (2009) based on the following Cramér-von Mises test statistic

$$\mathrm{S_n^{gof}} = \sum_{i=1}^{n} \left(\mathrm{C_n} \left(U_{i,n} \right) - \mathrm{C}_{\theta_n} \left(U_{i,n} \right) \right)^2$$

where θ_n is an estimator of a parameter vector θ computed from pseudo-observations $U_{1,n}, \ldots, U_{n,n}$.

Hofert et al. (2018) describes leave-one-out cross validation copula information criterion which can be used to compare copula families in a similar fashion as Akaike information criterion can be used for univariate models.

We estimate copula parameters from weekly aggregate loss data. We use the following algorithm for combining models for individual industries into one distribution of weekly aggregate losses.

1. Generate $N = 100,000$ random vectors $n_1, \ldots, n_N$ from a copula.
2. For each $i = 1, \ldots, N$ and for each industry $j \in \{1, 2, 3\}$ select $n_{i,j}$-th quantile from empirical quantile function of weekly aggregate losses for j-th industry and sum together all 3 obtained quantiles.
3. The resulting N numbers from the previous step constitute weekly aggregate losses.

4.6 Comparing risk measures with other quantities

We do not have to stop with our research after calculating risk measures. 99% CVaR is a property of the aggregate loss distribution. In our case the aggregate loss distribution is unknown. We can only obtain an empirical aggregate loss distribution function with a Monte Carlo simulation. This empirical distribution function is an observation of an unknown population distribution function. We can calculate an empirical 99% CVaR from this empirical aggregate loss distribution function. This empirical 99% CVaR has a population counterpart which is also unknown, but we know that it exists. This population 99% CVaR is

a random variable following some distribution. The observed empirical aggregate distribution function depends on random values generated from the frequency and severity distributions. Simply put, there are two ways how we can look at the empirical aggregate loss distribution function. First, it can be an observation of the population aggregate loss distribution function. Second, it can be a function that is so close to the population aggregate loss distribution function that for practical purposes we assume that it is the population aggregate loss distribution function. We use a large number of scenarios in the Monte Carlo simulation in order to obtain an empirical aggregate loss distribution function which is reasonably close to its population counterpart. Nonetheless, we cannot remove the noise from the empirical aggregate loss distribution introduced by generating random numbers from frequency and severity distributions. We can only decrease its proportion by increasing the number of scenarios.

The empirical 99% CVaR calculated from an empirical aggregate loss distribution function resulting from one Monte Carlo simulation is one observation of the population 99% CVaR. In other words, one execution of the previous algorithm produces one observation of the 99% CVaR. The algorithm can be executed 100 times to create a random sample of 99% CVaR observations. Consequently we can use a t-test to test a hypothesis about the mean of the 99% CVaR.

It would be natural to compare the 99% CVaR with a quantity that a broad audience can easily imagine. Such quantity can be for instance the GDP of the Czech Republic. Both quantities can be measured in USD, therefore we can directly compare them. We chose 99% CVaR because it is a coherent risk measure and the confidence level seems reasonable in terms of average human lifespan. Nevertheless, a similar test could be conducted with other risk measures. We still have to make one adjustment. The 99% CVaR calculated from the model proposed earlier has a forecast horizon of one week. Since we are comparing this risk measure with one-year GDP, we have to convert the calculated one-week 99% CVaR to a one-year 99% CVaR. The null and alternative hypotheses of the proposed t-test are following:

H_0: Mean of one-year 99% CVaR is equal or smaller than the GDP of the Czech Republic in 2019.

H_1: Mean of one-year 99% CVaR is greater than the GDP of the Czech Republic in 2019.

Chapter 5

Data description

We downloaded data used for calibration of the model from a website called Breach Level Index (Gemalto 2019). The main purpose of this website was to "tracks publicly disclosed breaches." One of its sections was Data Breach Database and it contained a dataset with recorded data breaches. Apart from this dataset the website used to offer an online assessment tool which allowed its users to calculate a data breach risk score of their organization. Additionally the website used to inform about news in the field of corporate information systems protection.

Each loss event represents one data breach. The dataset contains loss events recorded between years 2013 and 2018. The dataset has several variables. We are interested in three of them. They are date of the data breach, number of breached records and industry in which the data breach happened. We group the industries into three groups in order to reduce the dimensionality of the model. The three industries are technology, services and government. We try to put similar industries in one group. We also try to make the groups balanced in terms of number of observations.

Table 5.1 depicts summary statistics of the dataset. Losses are reported in number of breached records. Even though the maximum loss in services industry is more then twice as large as the maximum losses in technology and government industries, the medians are close to each other.

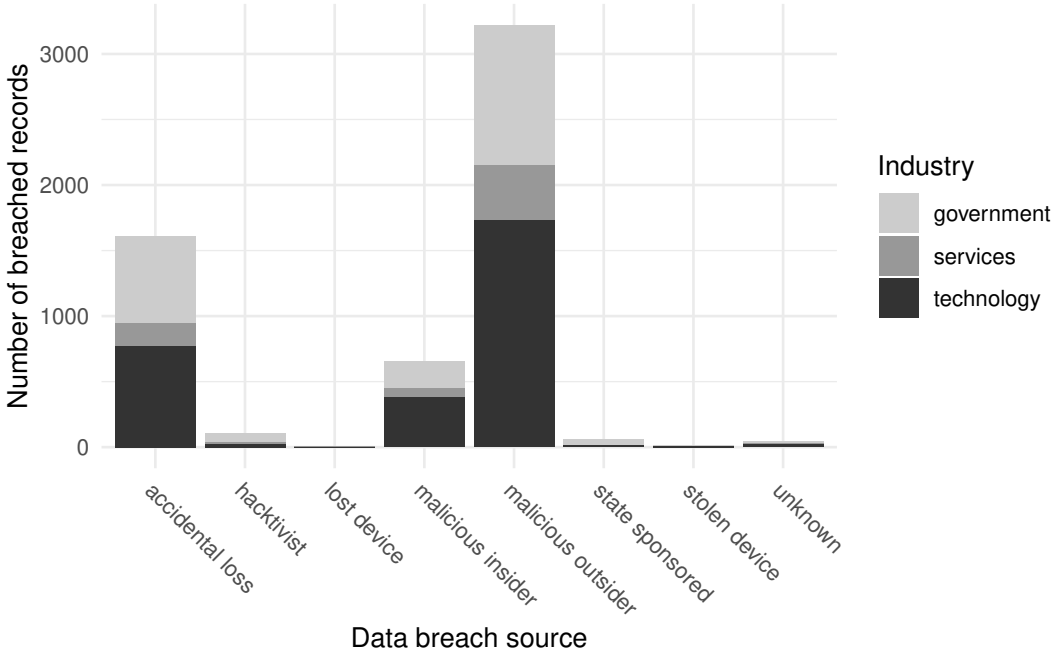
Figure 5.1 and Figure 5.2 present the dataset. The most common source of data breaches is malicious outsider. Accidental loss is the second most frequent source. Most data breaches in the dataset come from North America.

Loss events in the dataset come from various organisations. Loss amount in one organisation should not affect loss amount in a different organisation because

Table 5.1: Summary statistics of numbers of breached records in individual loss events broken down by industries

summary statistic	all industries	technology	services	government
number of observations	5,713	2,946	694	2,073
mean	2,576,163	1,963,863	6,672,738	2,074,866
standard deviation	43,057,592	31,647,109	87,939,504	33,083,023
median	1,661	1,934	1,434	1,250
min	1	1	1	1
max	2,200,000,000	1,200,000,000	2,200,000,000	1,340,000,000

Figure 5.1: Number of breached records between years 2013 and 2018 broken down by source of data breach and industry

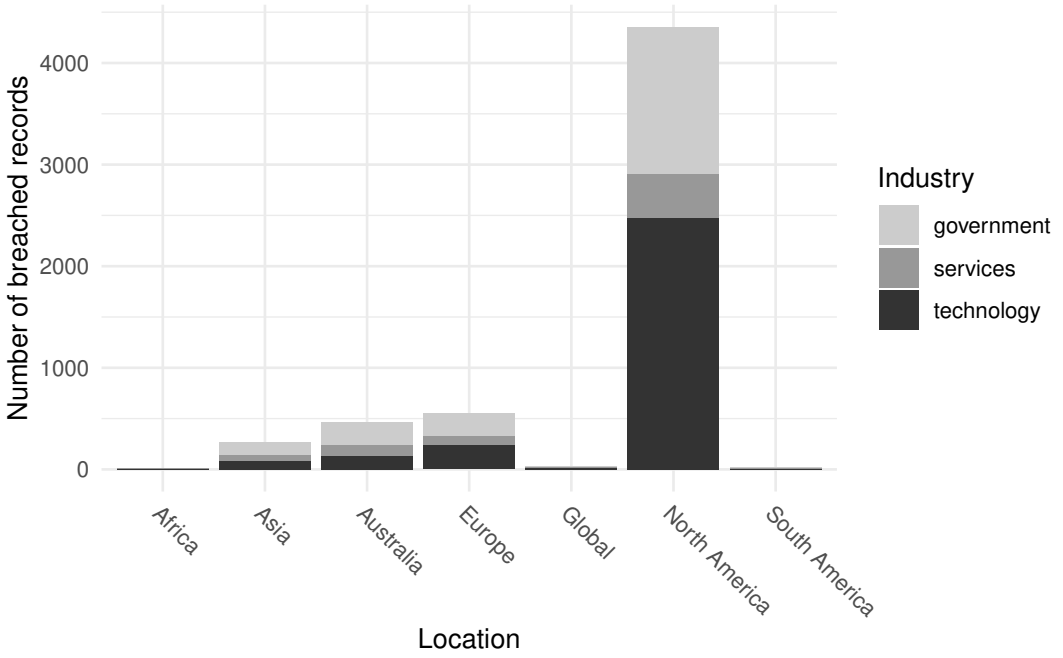


companies employ different security mechanism, store variously sensitive data and attackers do not target all organisation equally. Simultaneous loss events in different industries do not violate assumptions of the model. Other assumptions of the model are trivial. Therefore it is reasonable to expect that the data meet assumptions of the model.

We convert the 99% CVaR measured in number of breached records into a 99% CVaR measured in USD using a cost of USD 150 per one breached record estimated by IBM Security and Ponemon Institute (2019). Because we compare the 99% CVaR with the global GDP and the GDP of the Czech Republic, we have to obtain both these quantities in USD.

According to the Czech Statistical Office (2020) the GDP of the Czech

Figure 5.2: Number of breached records between years 2013 and 2018 broken down by location and industry



Republic in 2019 was CZK 5,751 billion. And according to the Czech National Bank (2020) the average USD/CZK exchange rate in the fourth quarter of 2019 was 23.107. Therefore at this exchange rate the Czech GDP in 2019 was USD 249 billion. According to the World Bank (2020) the global GDP in 2019 was USD 87,752 billion.

Chapter 6

Results and discussion

6.1 Results

First, we report results of goodness of fit methods applied to data. In particular it includes histograms, Q-Q plots and goodness of fit tests applied to data and estimated frequency and severity distributions. Second, we review estimates of parameters of generalized Pareto distribution estimated under extreme value theory. Third, we introduce copula parameter estimates along with copula goodness of fit tests and copula information criterion to simplify copula family selection. Fourth, we present risk measure estimates and compare results between various risk measures and between different dependence modelling approaches. We also report results of VaR backtesting. Finally, we discuss result of the t-test comparing 99% CVaR with the GDP of the Czech Republic. We explain decisions that led to the choices of particular distributions and parameters.

Figure 6.1 and Figure 6.2 illustrate histograms of loss frequency data and probabilities of fitted Poisson and negative binomial distributions, respectively. There are four plots, one for all industries combined and then one for each industry. Especially Poisson distribution seems to overestimate the number of loss events per day. Overall, negative binomial distribution shows a better fit to the data.

Figure 6.3 and Figure 6.4 depict Q-Q plots for loss frequency data against Poisson and negative binomial distributions, respectively. There are four plots, one for all industries combined and then one for each industry. Again, negative binomial distribution shows a better fit. While histograms show discrepancies for lower quantiles, Q-Q plots show differences between theoretical and empirical

Figure 6.1: Histograms of loss frequency data and probabilities of fitted Poisson distribution

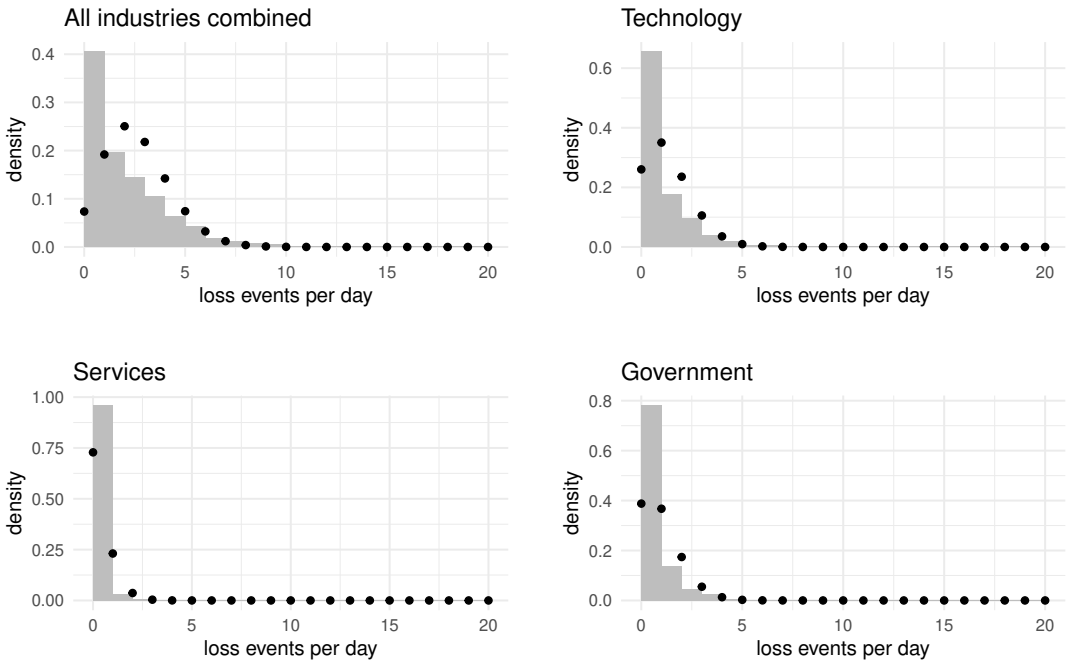
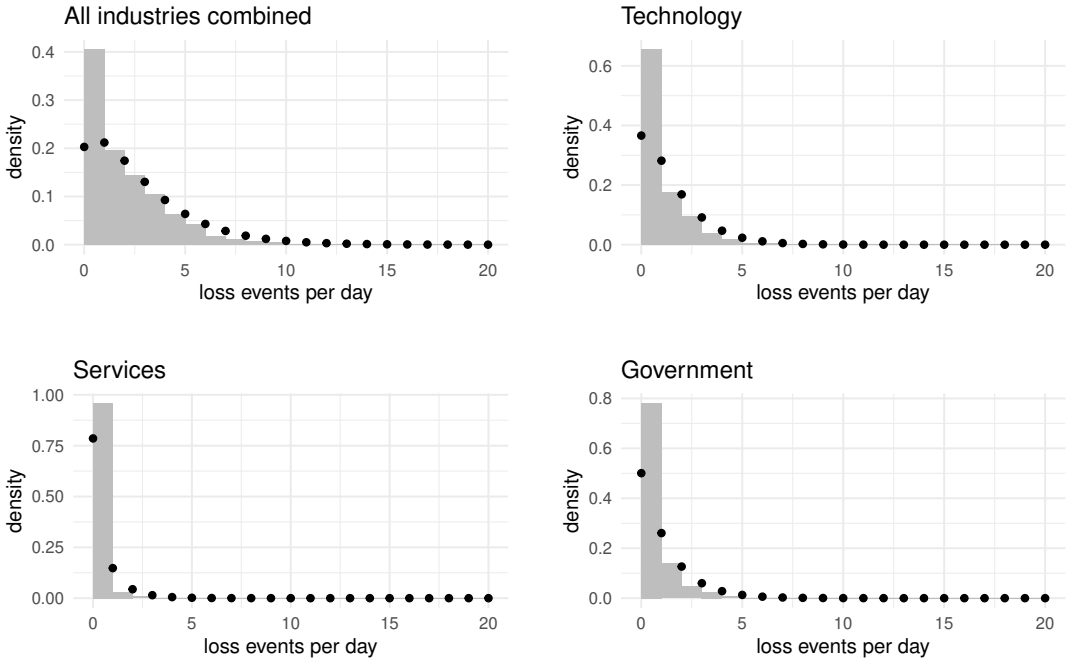


Figure 6.2: Histograms of loss frequency data and probabilities of fitted negative binomial distribution



distributions mainly in the right tails of the distributions. All industries show similar patterns for both distributions.

Table 6.1 gives information about parameter estimates of loss frequency

Figure 6.3: Q-Q plots against Poisson distribution for loss frequency data

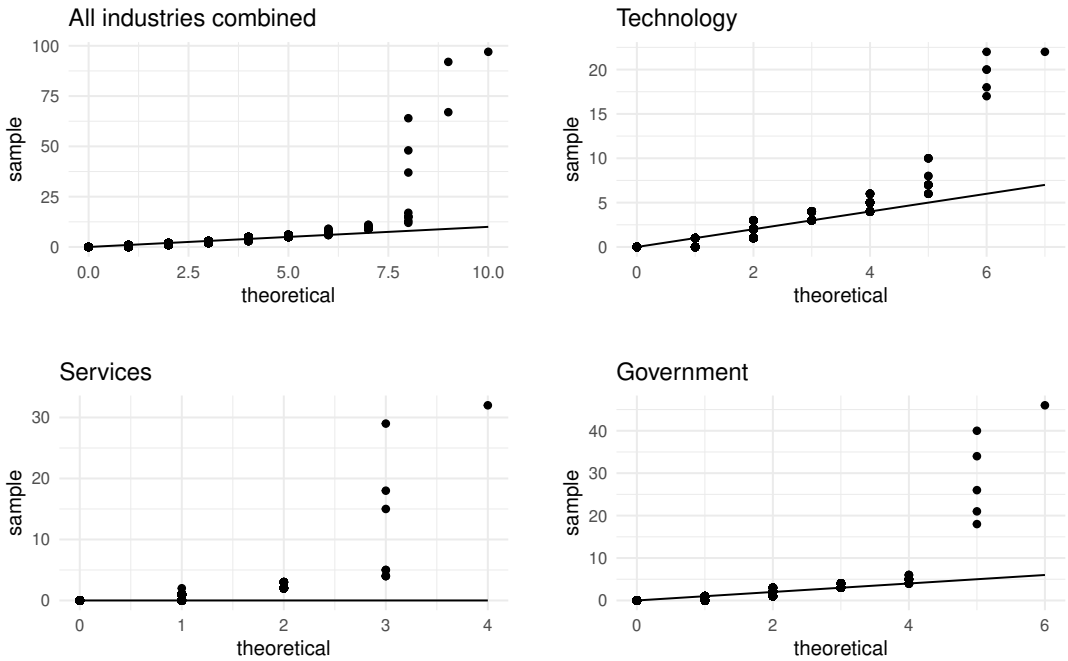
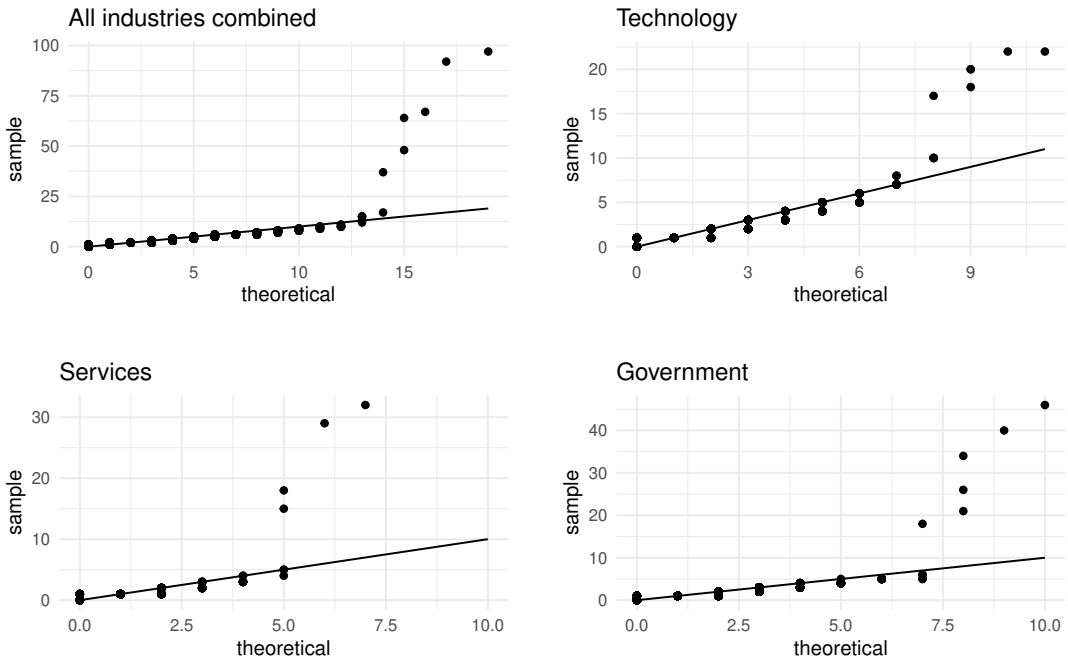


Figure 6.4: Q-Q plots against negative binomial distribution for loss frequency data



distributions. Table 6.1 gives information about p-values of Anderson-Darling and Cramér-von Mises goodness of fit tests for loss frequency distributions. We have to reject all hypothesis regarding Poisson and negative binomial

distributions being the distributions of loss frequency. Since goodness of fit p-values do not constitute a ranking of distribution suitability, we have to rely on histograms and Q-Q plots and use negative binomial distribution for calculation of aggregate losses.

Table 6.1: Estimated parameters of loss frequency distributions

	Poisson	negative binomial	
industry	λ	n	p
all industries	2.6099	1.7426	0.4004
technology	1.3458	1.7991	0.5721
services	0.3170	0.4627	0.5934
government	0.9470	1.1569	0.5499

Table 6.2: Anderson-Darling (AD) and Cramér-von Mises (CvM) goodness of fit tests p-values for loss frequency distributions

	Poisson		negative binomial	
industry	AD	CvM	AD	CvM
all industries	0.00059982	0.04833411	0.00059982	0.00112033
technology	0.00059982	0.00282555	0.00061286	0.00002353
services	0.00059982	0.00000000	0.00059982	0.00000000
government	0.00059982	0.00000000	0.00059982	0.00000000

Figure 6.5 and Figure 6.6 depict Q-Q plots for loss severity data against exponential and log-normal distributions, respectively. Figure A.6, Figure A.7 and Figure A.8 in Appendix show similar information for normal, Weibull and Cauchy distributions, respectively. Log-normal distribution shows a better fit especially in the body of the distribution. Since the tail is modelled with extreme value theory anyway, some discrepancies in the tail are not concerning.

Table 6.3 gives information about parameter estimates of loss severity distributions. Table 6.4 gives information about p-values of Kolmogorov-Smirnov, Anderson-Darling and Cramér-von Mises goodness of fit tests for loss severity distributions. We cannot reject the null hypothesis that loss severity data come from log-normal distribution. We can reject similar hypotheses for all other considered loss severity distributions including those that are not included in Table 6.4. Since only one distribution is not rejected, it makes the choice of severity distribution for aggregate loss model straightforward.

Figure 6.5: Q-Q plots against exponential distribution for loss severity data

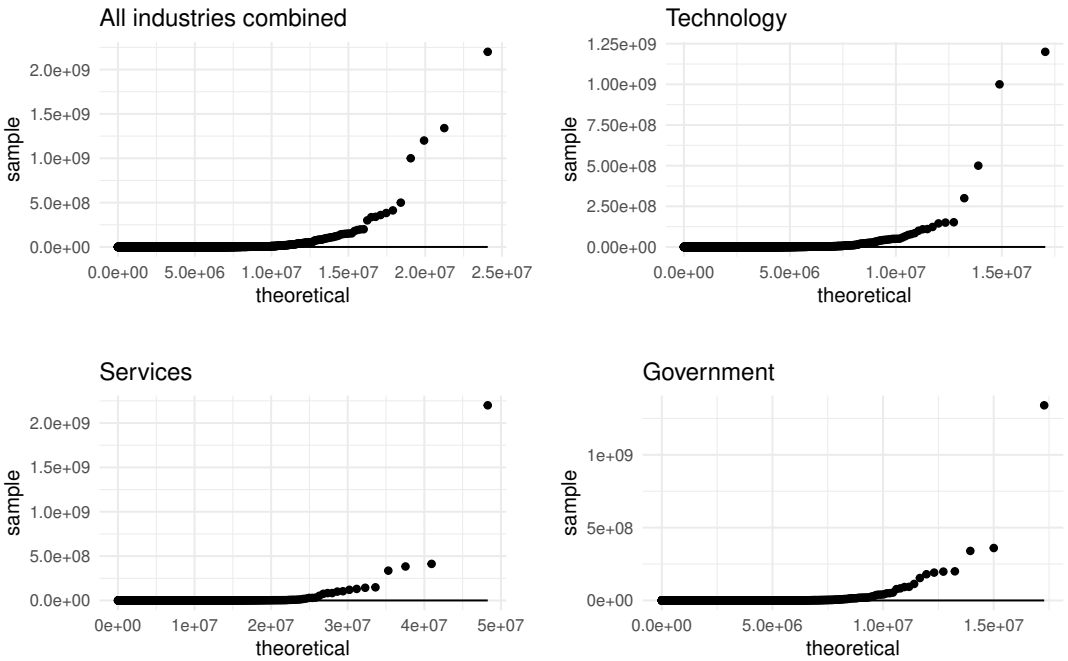


Figure 6.6: Q-Q plots against log-normal distribution for loss severity data

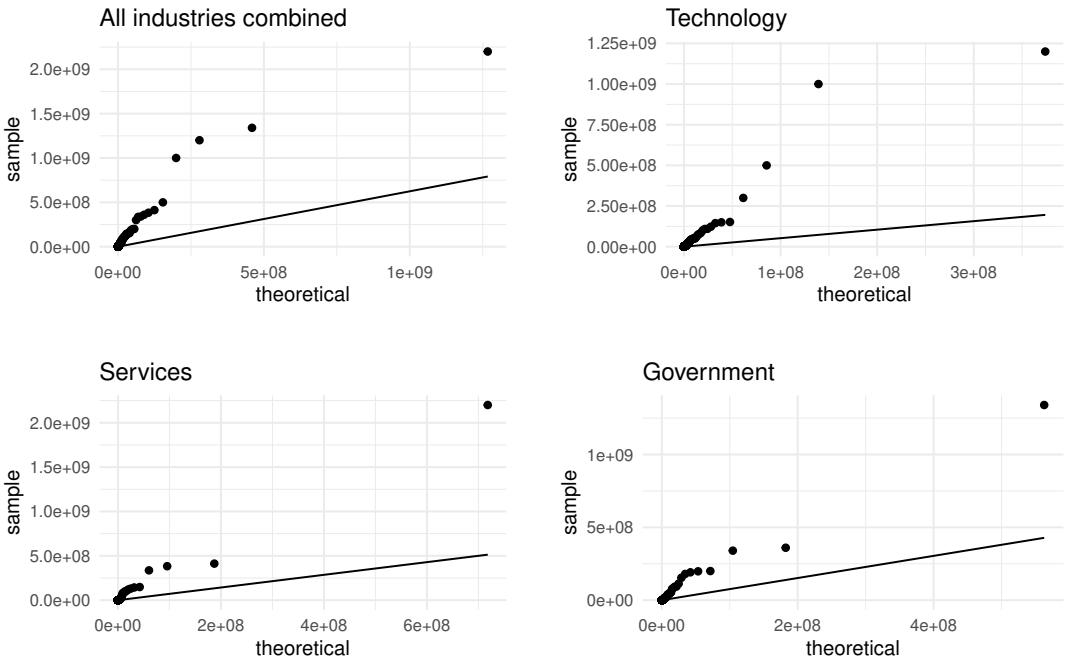


Figure 6.7 shows mean excess plots. We use them to select threshold for application of peaks over threshold method. The plot seems to be linear an upward sloping above 3×10^8 for all industries combineed, above 1.2×10^8 for

Table 6.3: Estimated parameters of loss severity distributions

	normal		exponential	log-normal		Weibull		Cauchy	
industry	μ	σ	λ	μ	σ	a	σ	l	s
all industries	2,576,163	43,053,823	0.00000039	7.60	3.56	0.25	10,872	779	1,427
technology	1,963,863	31,641,737	0.00000051	7.85	3.32	0.27	12,494	1,021	1,588
services	6,672,738	87,876,124	0.00000015	7.51	4.04	0.22	12,570	502	1,178
government	2,074,866	33,075,042	0.00000048	7.26	3.69	0.25	8,290	516	1,145

Table 6.4: Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Cramér-von Mises (CvM) goodness of fit tests p-values for loss severity distributions

	normal			exponential			log-normal		
industry	KS	AD	CvM	KS	AD	CvM	KS	AD	CvM
all industries	0.0000	0.0006	0.0000	0.0000	0.0006	0.0000	0.0000	0.4817	0.3490
technology	0.0000	0.0006	0.0000	0.0000	0.0006	0.0000	0.0000	0.0125	0.5924
services	0.0000	0.0006	0.0000	0.0000	0.0006	0.0000	0.0009	0.5342	0.5651
government	0.0000	0.0006	0.0000	0.0000	0.0006	0.0000	0.0034	0.2427	0.3967

technology, above 1.1×10^8 for services and above 1.8×10^8 for government.

Figure 6.7: Mean excess plots of loss severity data

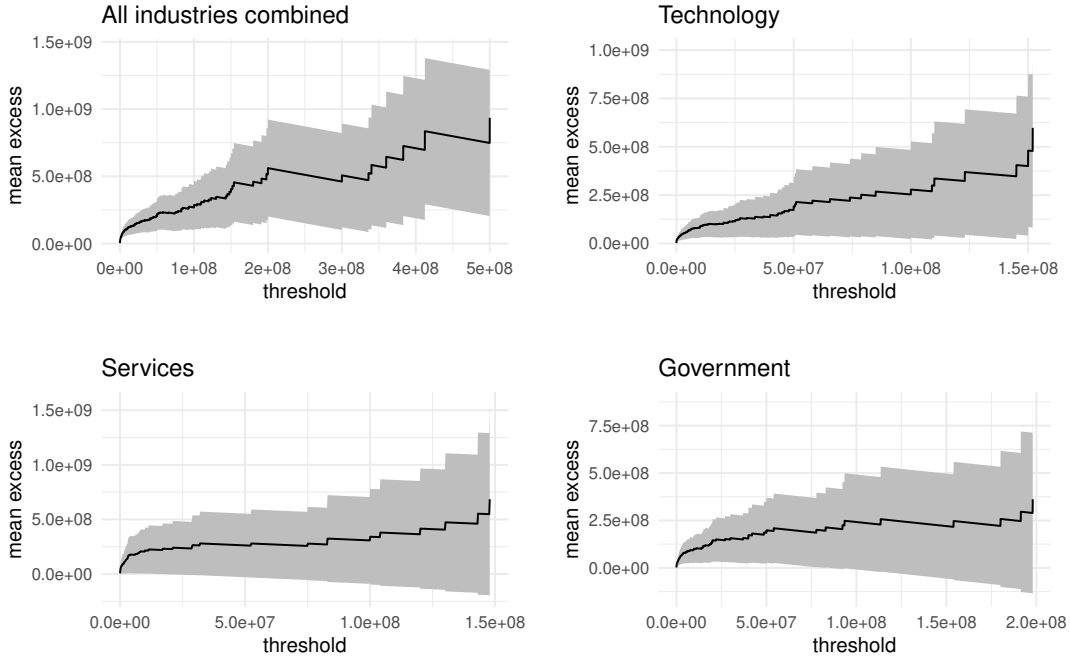


Table 6.5 summarizes parameter estimates of generalized Pareto distribution and two summary statistics related to peaks over threshold method. Namely, number of exceedances and proportion of exceedances. All industries combined have the smallest scale parameter β and the largest shape parameter ξ, while

government has the largest scale parameter β and the smallest shape parameter ξ.

Table 6.5: Estimated scale (β) and shape (ξ) parameters of generalized Pareto distribution along with threshold (u), number of exceedances (N_u) and proportion of exceedances (N_u/n) in the peaks over threshold method

industry	β	ξ	u	N_u	N_u/n
all industries	507,121,430	0.1229	300,000,000	10	0.0018
technology	326,250,000	0.1937	120,000,000	8	0.0027
services	374,014,287	0.3833	110,000,000	8	0.0115
government	221,348,867	0.4531	180,000,000	7	0.0034

Table 6.6 displays sample Kendall's taus between aggregate losses in different industries. We can see especially strong dependence between technology and government and a weak dependence between technology and services.

Table 6.6: Kendall's taus between aggregate weekly losses in three different industries

	technology	services	government
technology	1.0000	0.0201	0.1105
services	0.0201	1.0000	0.0931
government	0.1105	0.0931	1.0000

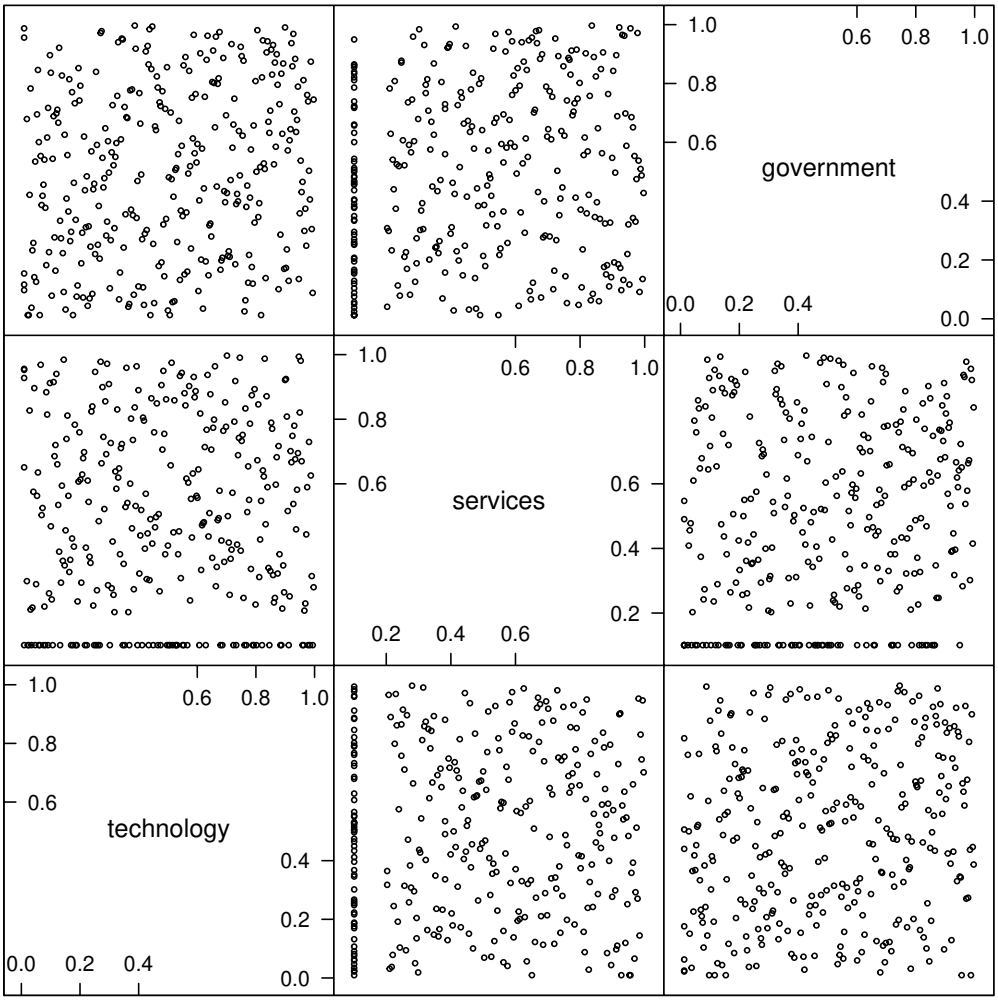
Figure 6.8 displays scatter plots of pseudo-observations of aggregate losses. We provide these scatter plots for all combinations of industries. It is left to the readers discretion where the dependence between aggregate losses confirmed by Kendall's taus can be seen in the scatter plots.

The estimated normal copula has parameter $\rho = 0.11$. The estimated t copula has parameter $\rho = 0.12$ and 19.33 degrees of freedom. The estimated Clayton, Frank, Gumbel-Hougaard and Joe copulas have parameter ψ equal to 0.18, 0.58, 1.05 and 1.02, respectively.

Figure 6.9 compares contour plots of two-dimensional fitted and empirical copulas. There are contour plots for normal and Clayton copulas in Figure 6.9. The fit seems to be relatively strong in both cases.

Table 6.7 presents copula goodness of fit test p-values and leave-one-out cross validation copula information criterion. We do not reject any null hypothesis that a particular copula describes dependence between aggregate losses in

Figure 6.8: Scatter plots of pseudo-observations of aggregate losses in three industries



different industries. In particular, we do not reject the hypothesis that Clayton copula describes such dependence.

Clayton copula also has the highest copula information criterion. Therefore we prefer this copula when reporting risk measures or comparing 99% CVaR using t-test. "It is proposed [...] that Gaussian or Normal-like copulas are not to be used for operational risk modelling. For instance a T-Student copula with few degrees of freedom (e.g. 3 or 4) in most cases appears more appropriate to capture the dependencies between operational risk events" (European Banking Authority 2014). Our results are consistent with this recommendation because we prefer the Clayton copula. Nonetheless we have to mention that normal copula scored second.

Figure 6.9: Contour plots of two-dimensional fitted normal copula and two-dimensional empirical copula, and the same situation with Clayton copula

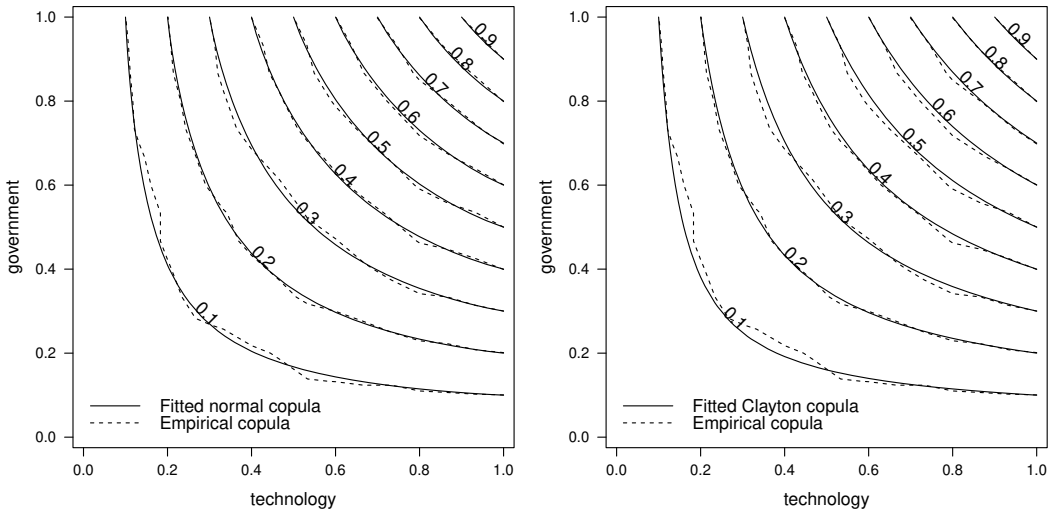


Table 6.7: Copula goodness of fit test p-values and leave-one-out cross validation copula information criterion

copula	goodness of fit p-value	cross-validation criterion
normal	0.7607	3.3643
t	0.7527	2.5206
Clayton	0.9955	6.2261
Frank	0.4481	2.9330
Gumbel-Hougaard	0.1044	0.0833
Joe	0.0085	-1.0251

Table 6.8 shows VaR and CVaR estimates when all industries are modelled together. Table 6.9 shows VaR and CVaR estimates with either full dependence or copula dependence structure. All risk measure estimates in Table 6.9 assume log-normal loss severity distribution. All risk measure are reported in number of breached records. Table A.1 in Appendix shows the same type of information under the assumption of exponential loss severity distribution. When loss severity is modelled with log-normal distribution and dependence between aggregate losses is modelled with Clayton copula then at a cost of USD 150 per one breached record the one-year 99% CVaR due to global data breach risk is USD 2,213 billion. It is 38.7% less compared to a situation with full dependence between aggregate losses in different industries.

Except for 95% VaR all other risk measures are higher when dependence is

modelled with copulas than under full dependence. When dependence structure is taken into account either through full dependence or through copulas, 99% VaR is roughly at the level of 95% CVaR. Despite few exceptions, in general we can say that modelling all industries together slightly underestimates the risk measures.

Table 6.8: VaR and CVaR estimates when all industries are modelled together, unit of measurement is number of breached records

copula	95% VaR	99% VaR	95% CVaR	99% CVaR
normal	414,441,370	1,013,614,641	798,092,160	1,716,477,213
exponential	94,650,336	981,124,425	619,756,629	1,671,977,056
log-normal	42,852,552	962,526,680	592,344,919	1,630,533,583
Weibull	15,705,227	922,392,391	555,684,503	1,600,582,859
Cauchy	180,679	917,960,076	544,172,136	1,574,850,628

Table 6.9: VaR and CVaR estimates with either full dependence or copula dependence structure, log-normal loss severity distribution is assumed in all cases, and unit of measurement is number of breached records

copula	95% VaR	99% VaR	95% CVaR	99% CVaR
full dependence	30,708,450	1,364,493,909	909,416,623	3,340,172,762
independence	246,954,478	1,086,660,479	814,458,638	2,043,386,444
normal	249,717,998	1,075,442,171	838,306,929	2,144,085,524
t	232,013,098	1,020,231,460	817,907,213	2,135,745,465
Clayton	253,625,078	1,065,736,706	819,130,934	2,045,888,804
Frank	250,479,827	1,058,790,644	810,420,925	2,004,924,756
Gumbel-Hougaard	233,003,869	1,099,158,551	823,617,799	2,113,897,090
Joe	232,989,165	1,009,021,492	770,314,689	1,915,477,677

Table 6.10 shows number of violations (exceedances) in terms of Kupiec's proportion of failures test for VaR backtesting under an assumption of copula dependence structure and log-normal loss severity distribution. The confidence intervals for a period of 6 years are $(60, 94)$ and $(8, 24)$ for VaR at confidence levels 95% and 99%, respectively. Results do not vary very much between different copulas. 95% VaR seems to be very conservative while 99% VaR is close to the lower bound of the corresponding confidence interval. Therefore 99% VaR is sightly conservative.

The t-test with a null hypothesis that the mean of one-year 99% CVaR

Table 6.10: Number of violations (exceedances) in terms of Kupiec's proportion of failures test for VaR backtesting under an assumption of copula dependence structure and log-normal loss severity distribution

copula	exceedances of 95% VaR	exceedances of 99% VaR
independence	25	5
normal	25	5
t	28	5
Clayton	25	5
Frank	25	5
Gumbel-Hougaard	27	5
Joe	27	5

is equal or smaller than the GDP of the Czech Republic in 2019 has p-value 1.0993×10^{-139}. We have enough evidence to say that one-year global data breach risk measured with 99% CVaR is greater than the GDP of the Czech Republic in 2019.

Observations for this t-test come from a statistical model which is believed to reasonably well represent the data generating process of 99% CVaR due to global data breach risk. This statistical model assumes that loss frequency distribution is negative binomial, loss severity distribution is log-normal with generalized Pareto distribution in the right tail and Clayton copula captures the dependence structure between aggregate losses in different industries.

It is also possible to directly compare risk measures with other quantities without a test. Under the same assumptions as before one-year 99% CVaR due to global data breach risk amounts to 2.5% of the global GDP in 2019.

Finally, we understand that many distribution families are possible options for frequency and severity distributions. Also the dependence structure can be modelled with many different copulas. There are many goodness of fit tests with various modifications and there are many different risk measures. We believe that we took all reasonable care to ensure that our model depicts reality to the best of our abilities. Nonetheless, as Box (1976) explains, we should not forget that "since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration."

6.2 Summary of results

In this section we compare our results with other researchers and highlight our contribution. We test the following hypotheses with the following results.

Hypothesis #1: Rejected. We reject the hypothesis that Poisson distribution is the distribution of loss frequencies. We use two tests to test this hypothesis, namely Anderson-Darling and Cramér-von Mises goodness of fit tests, with the same results. We prefer the negative binomial distribution for loss frequency distribution based on Q-Q plots. This is consistent with Eling and Jung (2018).

Hypothesis #2: Not rejected. We do not reject the hypothesis that log-normal distribution is the distribution of loss severities. We use Anderson-Darling and Cramér-von Mises goodness of fit tests to test this hypothesis. Both provide the same results. Log-normal distribution clearly provides the best fit to the data. This result is consistent with Abbate, Gourier, and Farkas (2009). Second best option would be the exponential distribution.

Hypothesis #3: Not rejected. We do not reject the hypothesis that the Gaussian copula describes dependencies between aggregate losses in different industries. Nonetheless, the Clayton copula has higher leave-one-out cross validation copula information criterion and other researchers like European Banking Authority (2014) do not suggest using the Gaussian copula in the context of operational risk modelling. Therefore we proceed with the Clayton copula to estimate VaR and CVaR. We use a copula goodness of fit test to test this hypothesis.

Hypothesis #4: Rejected. We reject the hypothesis that the possible total worldwide cost of data breaches per one year is smaller than the nominal GDP of the Czech Republic in 2019. We have enough evidence to say that it is greater than the GDP of the Czech Republic. We use a t-test on a random sample of 99% CVaR observations to test this hypothesis.

Table 6.11 compares our results with other authors studying operational risk and cyber risk in particular.

To the best of our knowledge, Eling and Jung (2018) is so far the most advanced study of cyber risk modelling using extreme value theory and copula. We extend their research with the following four improvements. First, in contrast

Table 6.11: Comparison of methodology and results in this thesis with previous studies

authors	fully disclosed methodology	publicly available data	adjusted goodness of fit test	EVT	copula	frequency distribution	severity distribution
Di Clemente and Romano (2004)	yes	no	no	yes	t	Poisson	log-normal
Valle, Fantazzini, and Giudici (2008)	yes	no	no	no	multiple	multiple	gamma
Abbate, Gourier, and Farkas (2009)	yes	no	no	yes	Frank	Poisson	log-normal
H. Herath and T. Herath (2011)	no	no	no	no	Gumbel	Poisson	Weibull
Lu (2011)	yes	yes	no	yes	t	multiple	multiple
Carrillo-Menéndez and Suárez (2012)	no	no	no	yes	none	Poisson	log-normal
Eling and Jung (2018)	yes	yes	no	yes	vine	negative binomial	log-normal
Eling and Wirfs (2019)	yes	no	no	yes	none	Poisson	none
this thesis	yes	yes	yes	yes	Clayton	negative binomial	log-normal

Note: The data are considered to be publicly available only if the authors provide a public source. Copula, frequency distribution and severity distribution columns represent the authors' recommended distribution.
Source: Author based on individual papers and own results.

to them we use goodness of fit tests adjusted for distribution functions with estimated parameters. Second, unlike them we are backtesting our risk measures using Kupiec's proportion of failures test. Third, they neither compare the estimated cyber risk with other quantities, nor their report the cyber risk as a percentage of other quantity. We do both. Fourth, their dataset is a database of data breaches provided by Privacy Rights Clearinghouse. This database is limited to loss events recorded in the United States. We have access to a much larger dataset containing 5,713 loss events recorded around the whole world.

Neither study presented in Table 6.11 uses goodness of fit tests adjusted for distributions with estimated parameters. Their choices of distribution families for frequency and severity distributions might be heavily biased. We use a method described by Braun (1980) to solve this problem.

H. Herath and T. Herath (2011), Abbate, Gourier, and Farkas (2009), Di Clemente and Romano (2004) and Carrillo-Menéndez and Suárez (2012) do not compare risk measures with quantities known to the general public. We extend their research by comparing the 99% CVaR of data breach risk with the GDP of the Czech Republic. So that even a reader with practically no understanding of the topic is able to understand some of our results.

Han, W. Wang, and J. Wang (2015) and Lu (2011) interpret risk measures in terms of regulatory capital which might not mean much to a reader without any further knowledge about the topic. Especially when they do not disclose the company whose capital requirements they calculate. As a consequence it is difficult for a reader to put their results into a wider concept. We extend their research by giving the 99% CVaR an easy to understand interpretation.

Eling and Jung (2018) calculate risk measures with unit of measurement in number of breached records. They do not go the extra step to convert the risk measures to monetary units. We go even further than calculating risk measures in US dollars. We provide statistically significant comparison of the 99% CVaR with the GDP, i.e. a quantity that is easily comprehensible to a wide audience.

Di Clemente and Romano (2004) use generated data. Their study is focused more on methodology than on application therefore it is acceptable. Abbate, Gourier, and Farkas (2009) and Yao, Wen, and Luan (2013) disclose the source of their data, however the data are not public. On one hand, a concern about privacy is an understandable reason. On the other hand, using proprietary data makes reproducibility impossible. Nonetheless, as our data were public at some point in time, we can disclose all information about the data that we have available. The trustworthiness of our research is therefore larger than theirs.

Eling and Wirfs (2019) do not use copula to model dependence between losses which according to our results exaggerates the risk measures. Moreover they use a small dataset to calibrate their model. We extend their research by applying copulas to aggregate loss distributions and by using a larger dataset.

Biener, Eling, and Wirfs (2014) calculate risk measures which are even more sophisticated than CVaR, but they practically do not disclose their methodology. One of the goals that this work tries to achieve is to fully explain methodology in order to facilitate more research of cyber risk.

Di Clemente and Romano (2004) and Valle, Fantazzini, and Giudici (2008) are among those rare studies which explain their methodology meticulously. We go even further. We extend their research by explaining into a great detail the algorithm for calculation of aggregate loss distribution under extreme value theory and copula.

This work tries to continue in a line of research started by Chalupka and Teplý (2008), Rippel (2009), Rippel and Teplý (2011) and Lebovič (2012). In particular, we improve an algorithm used by Rippel (2009) for estimating aggregate loss distribution under the peaks over threshold method with Monte Carlo simulation. Their version sums losses generated from the body distribution of the loss severity distribution and from the generalized Pareto distribution. Such algorithm does not produce the desired outcome. It is not possible to create a random sample from the loss severity distribution by separately generating some observations from the body distribution and some from the tail distribution because the choice whether a particular observation should come from the tail or from the body is itself random and the value of the observation depends on the result of this random event. In other words, different quantiles are generated from the body distribution than from the tail distribution. We use a piecewise quantile function to generate random numbers from the loss severity distribution under the peaks over threshold method. In order to obtain a random number we first generate a random number from an interval $[0, 1]$. Depending on the size of this number we pass it to either the body or the tail distribution function.

We fill the gaps in the literature by combining improvements which are used in other studies individually, but which to our best knowledge do not appear together in one single study. Eling and Jung (2018) is the most advanced study, yet it lacks some improvements from older studies related to operational risk. As far as we know, there is no single study that implements all improvements presented in older studies. We implement many of these remaining missing features in order to fill some gaps in the literature. Moreover, we believe that we are the first to compare the cyber risk with other quantities using statistical tests, and to report the cyber risk measured with 99% CVaR as a percentage of the global GDP. This fills even more gaps in the literature.

6.3 Policy recommendation

By now we have hopefully provided enough evidence to say that the global cyber risk is often underestimated. The world does not seem to be ready for

this amount of risk. Therefore we would like to provide policy recommendation. On the other hand, we understand that the digital revolution is still ongoing. And it is challenging to design an efficient policy when the subject is constantly changing.

In general, our recommendation is similar to that of Verizon (2020). First, we suggest to apply "better safe than sorry." Prevention is crucial in the fight against cyber crime. And it usually does not require more than applying common sense. Second, we recommend to get prepared for data breaches. The question is not any more whether a data breach happens. The question is when. An emergency plan is a cornerstone of data breach preparedness. Protection against data breaches starts with individuals and ends with critical infrastructure. Finally, in the context of the Czech Republic we suggest to increase the budget of the National Cyber and Information Security Authority.[1] This institution can help to increase the awareness of the public about cyber risk.

Data breaches are an integral part of our society. They cannot be fully eliminated. The best we can aim for is to reduce the losses due to data breaches. This however does not mean that compensations to the victims of data breaches should be reduced. The opposite is true. Organizations processing personal data create large databases which serve as a target for attackers. It is therefore reasonable to assume that these organizations are punished when this risk materialises into losses.

Article 34 of the General Data Protection Regulation (GDPR) of European Parliament and Council of the European Union (2016) gives individuals a right to be informed about data breaches. However the GDPR does not entitle individuals to any kind of compensation when they are affected by a data breach. The only option for individuals in order to receive a compensation is to file a case in court against the organization which suffers a data breach. In many cases the loss for the individual does not manifest instantly after the data breach. It might take several months or years before the individual realizes that they were a victim of an identity theft and that they incurred an actual loss due to the data breach. Some organizations realize that it is often inefficient and impractical for individuals to take legal action in order to claim compensation. This constitutes a moral hazard because these organizations create risk for which they cannot be made responsible.

While the loss incurred by one person might be relatively small, the total loss

[1] For more information, see `https://nukib.cz/en/`.

of thousands of victims of one data breach can be substantial. We believe that organizations should be responsible for data breaches even when the victims are unable to prove direct losses. Therefore in order to compensate for the burden of moral hazard to the society we recommend taxing data breaches. In our opinion the GDPR provides a reasonable starting point for the protection of personal data. Furthermore, we understand that agreeing on the GDPR was accompanied with obvious difficulties and strong opposition of many data processing organizations. Nonetheless, the moral hazard of not being financially responsible for data breaches will not be eliminated unless a policy similar to the one that we introduce is implemented.

Machuletz and Böhme (2020) give another example that GDPR does not always follow its purpose and in some cases it makes individuals worse off despite trying to help them. Under GDPR websites must inform its visitors that they are using cookies. Machuletz and Böhme (2020) discover that if the consent dialog has a highlighted default button with all purposes selected then users are tempted to agree with more purposes than they would without such a button.

GDPR is criticised by advocates of laissez-faire data market as well. For instance Zarsky (2017) argues that "Article 22 is perhaps the most salient example of the GDPR's rejection of the Big Data revolution." They maintain that GDPR has a prejudice against automated systems without a reason. Allen et al. (2019) on the other hand suggest that GDPR creates additional risk for organisations while completely ignoring that GDPR aims to offset the risk for individuals. They predict that a new insurance product will appear which will provide protect for organizations against losses due to GDPR.

6.4 Further research opportunities

Further research opportunities are at least threefold. First, a larger dataset would allow us to model cyber risk separately for each country. On one hand, as data collection gradually becomes cheaper and cyber risk awareness rises, we expect a higher availability of data in the future. On the other hand, cyber risk research is becoming commercialized and even datasets which were public in the past are now becoming private. This situation clearly suits consulting companies which earn money this way. Therefore, it is also possible that data will be less available in the future.

Second, both Erhardt and Czado (2012) and Brechmann, Czado, and Paterlini (2014) propose a method which utilizes copulas to handle zero losses

in empirical loss distribution. We could apply this method to the aggregate loss distribution and reduce aggregation from one week to one day this way. It might also be useful if we decide to model cyber risk in a particular geographic region for which there might be less data.

Third, the idea of taxing data breaches should be developed further. In this thesis we discuss legal and ethical aspects of such policy. Our arguments are justified by the large scale of global cyber risk that we estimate. Nonetheless, the full economic implications should be further investigated in order to properly set rules under which such tax is charged.

Chapter 7

Conclusion

In this thesis we deal with cyber risk modelling using copulas. First, we define operational risk, cyber risk and data breach risk. Second, we analyse current cyber threads. Third, we propose an operational risk model for data breach risk. It uses fitted loss frequency and severity distributions which are combined into an aggregate loss distribution defined by an actuarial model. The aggregate loss distribution is computed using Monte Carlo simulation. We apply peaks over threshold method to the loss severity distribution. We calculate risk measures from the distribution of aggregate losses. As risk measures we use value at risk and conditional value at risk which is a coherent risk measure.

We model distributions of aggregate losses both separately for each of the three industries and together for all industries combined. The three industries are technology, services and government. We estimate two elliptical and four Archimedean copulas on aggregate loss data for these three industries. We use a copula goodness of fit test and a copula information criterion to select the most suitable copula family. Using graphical tools and goodness of fit tests we conclude that negative binomial distribution provides the best fit to the loss frequency data and log-normal distribution provides the best fit to the loss severity data. Clayton copula has the highest copula information criterion out of all considered copulas. We define a distribution of conditional value at risk obtained from an empirical distribution function of aggregate losses. We use a t-test on the random sample of conditional value at risk. Finally, we summarize results which we obtained from applying the model on the loss data. We discuss contribution of this thesis mainly in Section 6.2.

When testing for goodness of fit of frequency and severity distributions we use Anderson-Darling and Cramér-von Mises tests adjusted for distribution

functions with estimated parameters. To achieve this we use a method presented in Braun (1980). It turns out that not all previous studies on cyber risk use some approach for correcting the loss of power of goodness of fit tests when they are used with a distribution function with estimated parameters.

We discover that there is some positive dependence between aggregate losses in technology and government industries. We use an average cost of one breached record calculated by IBM Security and Ponemon Institute (2019) to convert the unit of measurement of 99% CVaR from number of breached records to US dollars. Risk measures are lower when all industries are modelled together than if the dependence structure between aggregate losses is modelled in some way. These results are consistent across different marginal loss severity distributions and copulas. When dependence structure between aggregate losses in different industries is modelled with Clayton copula then one-year 99% CVaR due to global data breach risk is USD 2,213 billion. It is 38.7% less compared to a situation with full dependence between aggregate losses.

We have enough evidence to say that one-year global data breach risk measured with 99% CVaR is greater than the GDP of the Czech Republic in 2019. Furthermore, one-year global cyber risk measured with 99% CVaR amounts to 2.5% of the global GDP in 2019. This comparison of conditional value at risk with the GDP allows even a reader with a very little experience with operational risk modelling to understand the magnitude of the cyber risk. Our main policy recommendation as a result of the substantial size of the cyber risk is to consider taxation of data breaches.

We fill many gaps in the literature by combining improvements of operational risk models which are used in other studies individually, but which as far as we know do not appear together in one single study. We fill further gaps in the literature by thoroughly explaining our methodology with particular attention to the algorithm for calculation of aggregate loss distribution under extreme value theory and copula.

The contribution of this thesis over the so far most advanced study of cyber risk by Eling and Jung (2018) is fourfold:

1. We use a large unique dataset consisting of 5,713 loss events between 2013 and 2018. Our dataset covers the whole world. The dataset used by Eling and Jung (2018) covers only the United States.

2. We use Anderson-Darling and Cramér-von Mises tests adjusted for dis-

tribution functions with estimated parameters. The p-values might be invalid without a similar adjustment.

3. We are backtesting our risk measures using Kupiec's proportion of failures test.

4. We report the 99% CVaR as a percentage of the global GDP and we compare the 99% CVaR with the GDP of the Czech Republic.

Bibliography

Abbate, Donato, Elise Gourier, and Walter Farkas (2009). "Operational Risk Quantification Using Extreme Value Theory and Copulas: From Theory to Practice". In: *Journal of Operational Risk* 3, pp. 1–24.

Aldasoro, Iñaki et al. (2020). "The drivers of cyber risk". In: *BIS Working Papers* 865.

Allen, Darcy W E et al. (2019). "Some Economic Consequences of the GDPR". In: *Economics Bulletin* 39.2, pp. 785–797.

Anderson, T. W. and D. A. Darling (1952). "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes". In: *The Annals of Mathematical Statistics* 23.2, pp. 193–212.

— (1954). "A Test of Goodness of Fit". In: *Journal of the American Statistical Association* 49.268, pp. 765–769.

Artzner, Philippe et al. (1999). "Coherent Measures of Risk". In: *Mathematical Finance* 9.3, pp. 203–228.

Basel Committee on Banking Supervision, Bank for International Settlements (2019). *Basel Framework: Calculation of RWA for operational risk: Definitions and application*. URL: `https://www.bis.org/basel_framework/chapter/OPE/10.htm` (visited on 06/30/2020).

Bhatti, M. Ishaq and Hung Quang Do (2019). "Recent development in copula and its applications to the energy, forestry and environmental sciences". In: *International Journal of Hydrogen Energy* 44.36, pp. 19453–19473.

Biener, Christian, Martin Eling, and Jan Wirfs (2014). "Insurability of Cyber Risk: An Empirical Analysis". In: *Geneva Papers on Risk and Insurance - Issues and Practice* 40, pp. 1–28.

Bouveret, Antoine (2018). *Cyber Risk for the Financial Sector: A Framework for Quantitative Assessment*. URL: `https://www.imf.org/en/Publications/WP/Issues/2018/06/22/Cyber-Risk-for-the-Financial-Sector-A-Framework-for-Quantitative-Assessment-45924` (visited on 06/30/2020).

Box, George E. P. (1976). "Science and Statistics". In: *Journal of the American Statistical Association* 71.356, pp. 791–799.

Braun, Henry (1980). "A Simple Method for Testing Goodness of Fit in the Presence of Nuisance Parameters". In: *Journal of the Royal Statistical Society* 42.1, pp. 53–63.

Brechmann, Eike, Claudia Czado, and Sandra Paterlini (2014). "Flexible dependence modeling of operational risk losses and its impact on total capital requirements". In: *Journal of Banking & Finance* 40, pp. 271–285.

Carrillo-Menéndez, Santiago and Alberto Suárez (2012). "Robust quantification of the exposure to operational risk: Bringing economic sense to economic capital". In: *Computers & Operations Research* 39.4, pp. 792–804.

Cebula, James, Mary Popeck, and Lisa Young (2014). *A Taxonomy of Operational Cyber Security Risks Version 2*. URL: `https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=91013` (visited on 06/30/2020).

Chalupka, Radovan and Petr Teplý (2008). "Operational Risk Management and Implications for Bank's Economic Capital – a Case Study". In: *Working Papers IES*.

Chapelle, Ariane et al. (2008). "Practical methods for measuring and managing operational risk in the financial sector: A clinical study". In: *Journal of Banking & Finance* 32.6, pp. 1049–1061.

Chavez-Demoulin, Valérie, Paul Embrechts, and Marius Hofert (2016). "An Extreme Value Approach for Modeling Operational Risk Losses Depending on Covariates". In: *Journal of Risk and Insurance* 83.3, pp. 735–776.

Chavez-Demoulin, Valérie, Paul Embrechts, and J. Nešlehová (2006). "Quantitative models for operational risk: Extremes, dependence and aggregation". In: *Journal of Banking & Finance* 30.10, pp. 2635–2658.

Chernobai, Anna S., Svetlozar T. Rachev, and Frank J. Fabozzi (2007). *Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis*. John Wiley & Sons, Inc.

Czech National Bank (2020). *Central bank exchange rates fixing - monthly averages*. URL: `https://www.cnb.cz/en/financial-markets/foreign-exchange-market/central-bank-exchange-rate-fixing/central-bank-exchange-rate-fixing/currency_average.html?currency=USD` (visited on 06/30/2020).

Czech Statistical Office (2020). *Gross domestic product - time series*. URL: `https://www.czso.cz/csu/czso/hdp_ts` (visited on 06/30/2020).

D'Agostino, Ralph B. and Michael A. Stephens (1986). *Goodness-of-fit techniques*. MARCEL DEKKER, INC.

Darling, D. A. (1957). "The Kolmogorov-Smirnov, Cramer-von Mises Tests". In: *Ann. Math. Statist.* 28.4, pp. 823–838.

Department for Digital, Culture, Media & Sport, Her Majesty's Government (2019). *Cyber Security Breaches Survey 2019: Main report*. URL: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2019 (visited on 06/30/2020).

Di Clemente, Annalisa and Claudio Romano (2004). "A copula-Extreme Value Theory approach for modelling operational risk". In: *Oper. Risk Model. Anal.* 9, pp. 189–208.

Eling, Martin and Kwangmin Jung (2018). "Copula approaches for modeling cross-sectional dependence of data breach losses". In: *Insurance: Mathematics and Economics* 82, pp. 167–180.

Eling, Martin and Jan Wirfs (2019). "What are the actual costs of cyber risk events?" In: *European Journal of Operational Research* 272.3, pp. 1109–1119.

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch (1997). *Modelling Extremal Events: for Insurance and Finance*. Springer-Verlag Berlin Heidelberg.

Erhardt, Vinzenz and Claudia Czado (2012). "Modeling dependent yearly claim totals including zero claims in private health insurance". In: *Scandinavian Actuarial Journal* 2012.2, pp. 106–129.

European Banking Authority (2014). *Consultation Paper: Draft Regulatory Technical Standards on assessment methodologies for the Advanced Measurement Approaches for operational risk under Article 312 of Regulation (EU) No 575/2013*. URL: https://eba.europa.eu/file/64728/download?token=RzTgmsYB (visited on 06/30/2020).

European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (visited on 06/30/2020).

Gaidosch, Tamas et al. (2019). *Cybersecurity Risk Supervision*. URL: https://www.imf.org/en/Publications/Departmental-Papers-Policy-Papers/Issues/2019/09/23/Cybersecurity-Risk-Supervision-46238 (visited on 06/30/2020).

Gemalto (2019). *Breach Level Index: Data Breach Database*. URL: https://breachlevelindex.com/data-breach-database (visited on 10/31/2019).

Gençay, Ramazan, Faruk Selçuk, and Abdurrahman Ulugülyağci (2003). "High volatility, thick tails and extreme value theory in value-at-risk estimation". In: *Insurance: Mathematics and Economics* 33.2, pp. 337–356.

Genest, Christian, Bruno Rémillard, and David Beaudoin (2009). "Goodness-of-fit tests for copulas: A review and a power study". In: *Insurance: Mathematics and Economics* 44.2, pp. 199–213.

Ghosh, Souvik and Sidney Resnick (2010). "A discussion on mean excess plots". In: *Stochastic Processes and their Applications* 120.8, pp. 1492–1517.

Greenspan, Alan (1996). *Remarks at the Financial Markets Conference of the Federal Reserve Bank of Atlanta, Coral Gables, Florida*. URL: https://fraser.stlouisfed.org/title/452/item/8561 (visited on 06/30/2020).

Han, Jinmian, Wei Wang, and Jiaqi Wang (2015). "POT model for operational risk: Experience with the analysis of the data collected from Chinese commercial banks". In: *China Economic Review* 36, pp. 325–340.

Herath, Hemantha and Tejaswini Herath (2011). "Copula Based Actuarial Model for Pricing Cyber-Insurance Policies". In: *Insurance Markets and Companies* 2.

Hofert, Marius et al. (2018). *Elements of Copula Modeling with R*. Springer International Publishing AG.

IBM Security and Ponemon Institute (2019). *Cost of a Data Breach Report*. URL: https://databreachcalculator.mybluemix.net/ (visited on 06/30/2020).

Jarrow, Robert A. (2008). "Operational risk". In: *Journal of Banking & Finance* 32.5, pp. 0378–4266.

Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmont (2012). *Loss models: from data to decisions*. John Wiley & Sons, Inc.

Lebovič, Michal (2012). *The use of coherent risk measures in operational risk modeling*. URL: https://dspace.cuni.cz/handle/20.500.11956/40523 (visited on 06/30/2020).

Lloyd's and Cyence (2017). *Counting the cost: Cyber exposure decoded*. URL: https://www.lloyds.com/~/media/files/news-and-insight/risk-insight/2017/cyence/emerging-risk-report-2017---counting-the-cost.pdf (visited on 06/30/2020).

Lu, Zhaoyang (2011). "Modeling the yearly Value-at-Risk for operational risk in Chinese commercial banks". In: *Mathematics and Computers in Simulation* 82.4, pp. 604–616.

Machuletz, Dominique and Rainer Böhme (2020). "Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR". In: *Proceedings on Privacy Enhancing Technologies* 2020.2, pp. 481–498.

Mejstřík, Michal, Magda Pečená, and Petr Teplý (2015). *Banking in Theory and Practice*. Karolinum Press, Charles University.

Mukhopadhyay, Arunabha et al. (2013). "Cyber-risk decision models: To insure IT or not?" In: *Decision Support Systems* 56, pp. 11–26.

Nelsen, Roger B. (2006). *An Introduction to Copulas*. Springer Science+Business Media, Inc.

NortonLifeLock (2020). *What is a data breach?* URL: `https://us.norton.com/internetsecurity-privacy-data-breaches-what-you-need-to-know.html` (visited on 06/30/2020).

Pan, Yue et al. (2019). "Modeling risks in dependent systems: A Copula-Bayesian approach". In: *Reliability Engineering & System Safety* 188, pp. 416–431.

Rippel, Milan (2009). *Operational Risk: Scenario analysis*. URL: `https://dspace.cuni.cz/handle/20.500.11956/32873` (visited on 06/30/2020).

Rippel, Milan and Petr Teplý (2011). "Operational Risk - Scenario Analysis". In: *Prague Economic Papers* 2011.1, pp. 23–39.

Rockafellar, R. Tyrrell and Stanislav Uryasev (2000). "Optimization of conditional value-at-risk". In: *Journal of Risk* 2.3.

Shah, Anand (2016). "Pricing and Risk Mitigation Analysis of a Cyber Liability Insurance using Gaussian, t and Gumbel Copulas – A Case for Cyber Risk Index". In: *CANADIAN ECONOMICS ASSOCIATION (CEA) 2016 Ottawa meetings paper*.

Stephens, Michael A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". In: *Journal of the American Statistical Association* 69.347, pp. 730–737.

Thales (2019). *2019 Thales Data Threat Report: Global Edition*. URL: `https://www.thalesesecurity.com/2019/data-threat-report` (visited on 01/31/2020).

Thomson Reuters (2019). *Data breaches infographic*. URL: `https://legal.thomsonreuters.com/content/dam/ewp-m/documents/legal/en/pdf/infographics/databreaches-infographic.pdf` (visited on 06/30/2020).

Uryasev, Stan (2010). *VaR vs CVaR in Risk Management and Optimization*. URL: https://www.ise.ufl.edu/uryasev/files/2011/11/VaR_vs_CVaR_CARISMA_conference_2010.pdf (visited on 06/30/2020).

Valle, Luciana Dalla, Dean Fantazzini, and Paolo Giudici (2008). "Copulae and Operational Risks". In: *International Journal of Risk Assessment and Management* 9.

Verizon (2018). *2018 Data Breach Investigations Report*. URL: https://enterprise.verizon.com/resources/reports/DBIR_2018_Report.pdf (visited on 06/30/2020).

— (2019). *2019 Data Breach Investigations Report*. URL: https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf (visited on 06/30/2020).

— (2020). *2020 Data Breach Investigations Report*. URL: https://enterprise.verizon.com/resources/reports/2020-data-breach-investigations-report.pdf (visited on 06/30/2020).

Wang, Yinghui, Jianping Li, and Xiaoqian Zhu (2017). "A Method of Estimating Operational Risk: Loss Distribution Approach with Piecewise-defined Frequency Dependence". In: *Procedia Computer Science* 122, pp. 261–268.

Watson, G. S. (1958). "On Chi-Square Goodness-Of-Fit Tests for Continuous Distributions". In: *Journal of the Royal Statistical Society* 20.1, pp. 44–72.

World Bank (2020). *GDP (current US$)*. URL: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD (visited on 06/30/2020).

Xu, Chi et al. (2019). "Double correlation model for operational risk: Evidence from Chinese commercial banks". In: *Physica A: Statistical Mechanics and its Applications* 516, pp. 327–339.

Yamai, Yasuhiro and Toshinao Yoshiba (2002a). "Comparative Analyses of Expected Shortfall and Value-at-Risk: Their Estimation Error, Decomposition, and Optimization". In: *Monetary and Economic Studies* 20.1, pp. 87–121.

— (2002b). "On the Validity of Value-at-Risk: Comparative Analyses with Expected Shortfall". In: *Monetary and Economic Studies* 20.1, pp. 57–85.

Yao, Fengge, Hongmei Wen, and Jiaqi Luan (2013). "CVaR measurement and operational risk management in commercial banks according to the peak value method of extreme value theory". In: *Mathematical and Computer Modelling* 58.1, pp. 15–27.

Zarsky, Tal (2017). "Incompatible: The GDPR in the Age of Big Data". In: *Seton Hall Law Review* 47.4.

Appendix A

Appendix

Table A.1: VaR and CVaR estimates with either full dependence or copula dependence structure, exponential loss severity distribution is assumed in all cases, and unit of measurement is number of breached records

copula	95% VaR	99% VaR	95% CVaR	99% CVaR
full dependence	124,914,789	1,446,497,299	996,745,644	3,410,781,688
independence	299,436,350	1,095,817,516	866,197,485	2,107,297,949
normal	306,177,024	1,097,565,628	893,637,571	2,206,545,655
t	290,940,078	1,049,499,954	868,536,482	2,173,006,372
Clayton	309,330,808	1,082,823,381	879,149,749	2,143,293,397
Frank	308,448,714	1,074,110,138	868,519,480	2,090,544,496
Gumbel-Hougaard	287,680,477	1,109,519,073	892,347,037	2,248,720,435
Joe	288,472,099	1,036,749,181	837,752,121	2,051,530,635

Figure A.1: Density function of Weibull distribution with two different choices of parameters

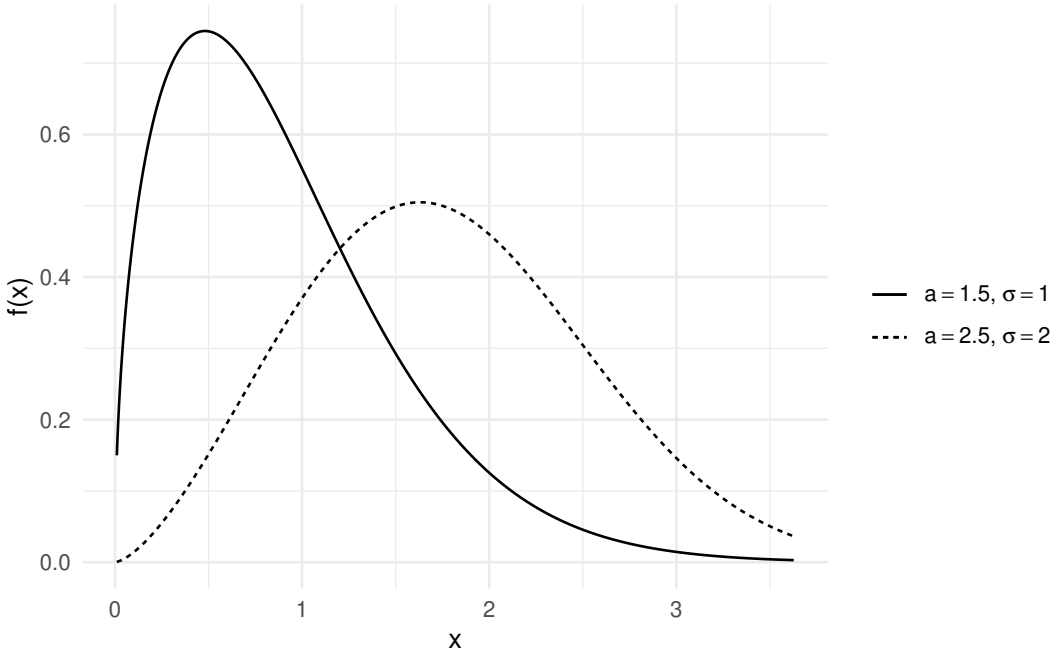


Figure A.2: Density function of Cauchy distribution with two different choices of parameters

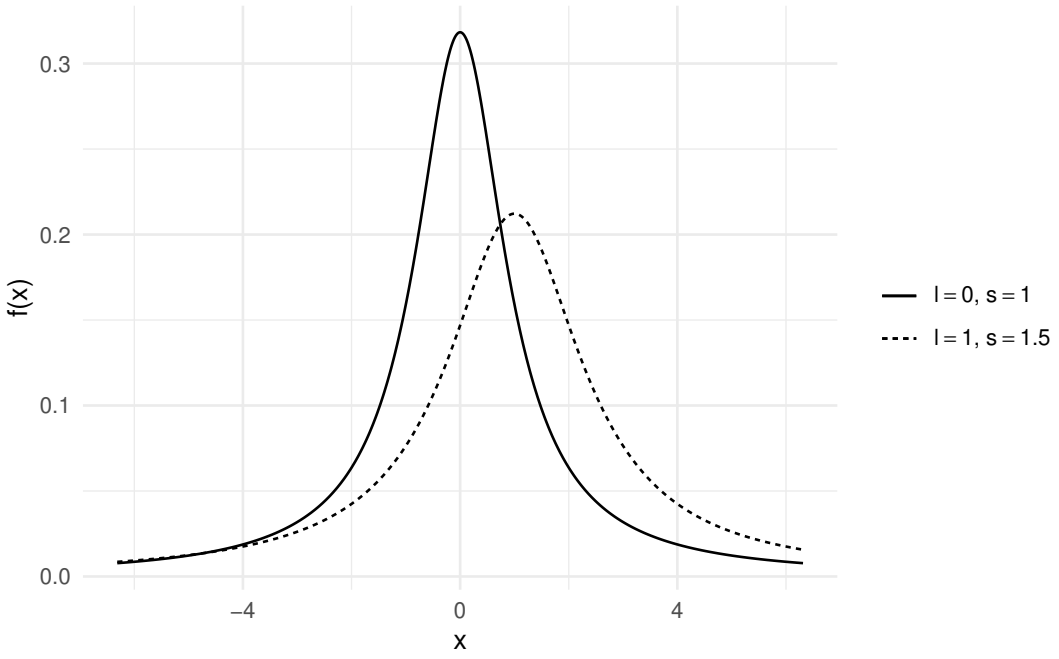


Figure A.3: Wireframe plot of density function of Frank copula with a Kendall's tau $\tau = 0.5$ and a sample of size $n = 1000$ from the same copula

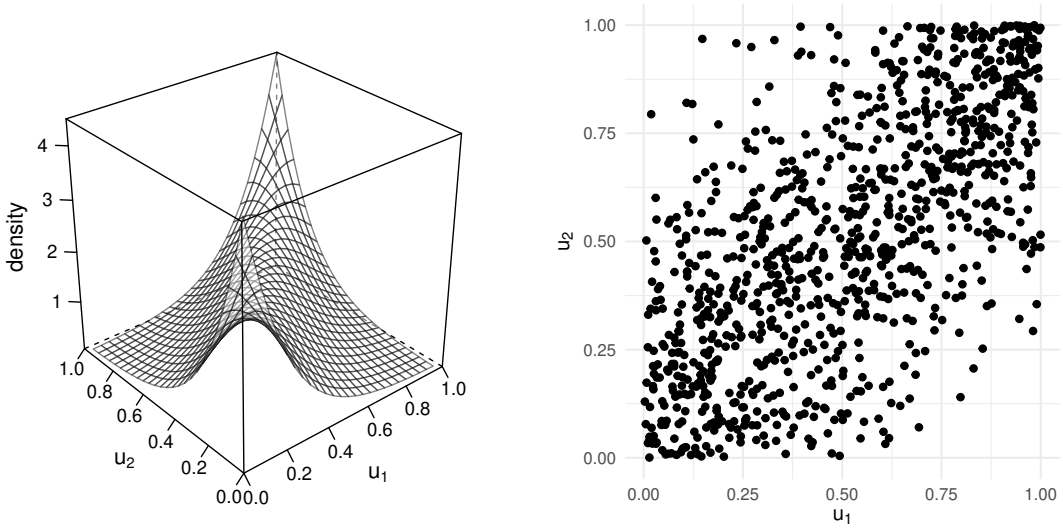


Figure A.4: Wireframe plot of density function of Gumbel-Hougaard copula with a Kendall's tau $\tau = 0.5$ and a sample of size $n = 1000$ from the same copula

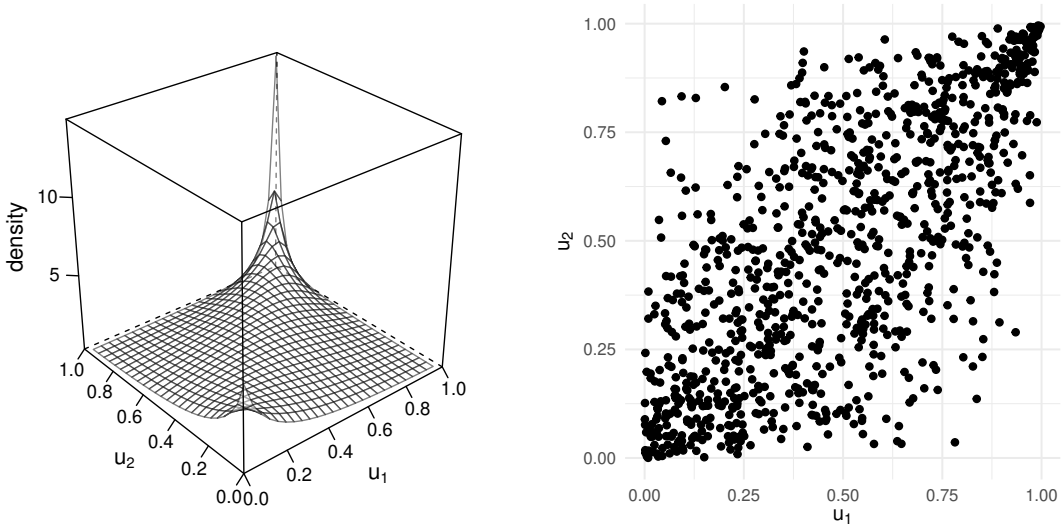


Figure A.5: Wireframe plot of density function of Joe copula with a Kendall's tau $\tau = 0.5$ and a sample of size $n = 1000$ from the same copula

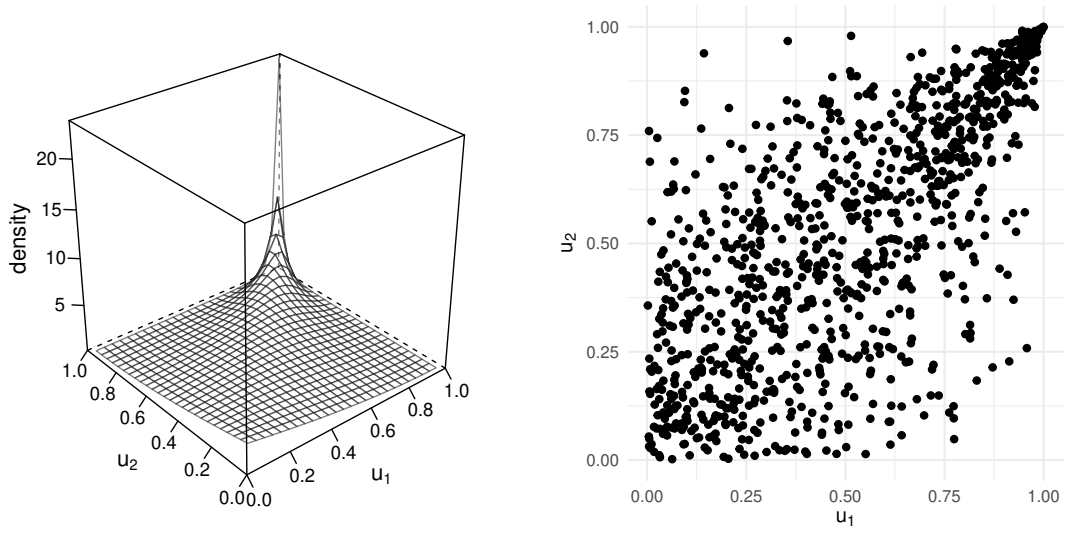


Figure A.6: Q-Q plots against normal distribution for loss severity data

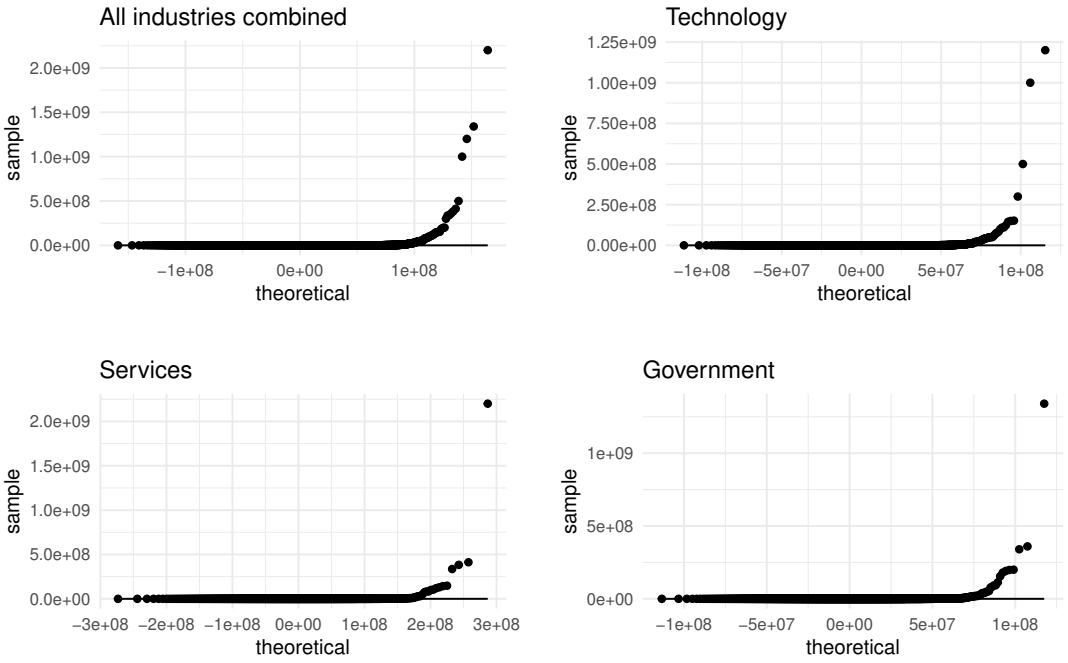


Figure A.7: Q-Q plots against Weibull distribution for loss severity data

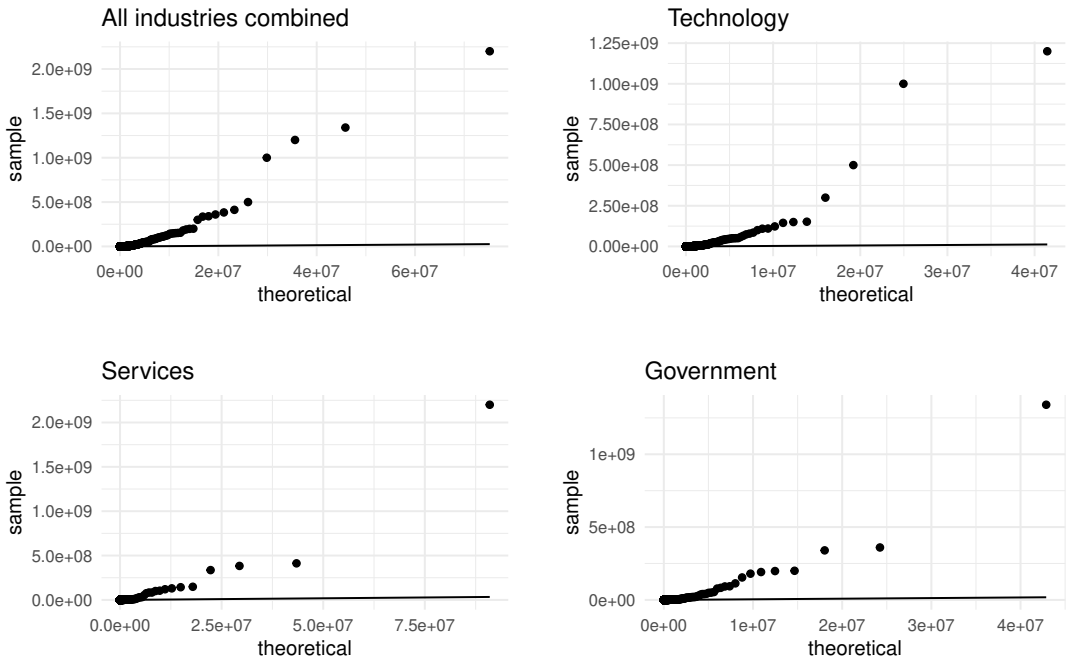


Figure A.8: Q-Q plots against Cauchy distribution for loss severity data