TOPICS ON THE FUNDAMENTAL REVIEW OF THE TRADING BOOK

Janine Balter

PhD

University of York

Management

July 2021

Abstract

The Basel Committee on Banking Supervision has introduced a revised market risk framework and banks are expected to adopt the new rules by January 2023. This work addresses two of the major amendments to the framework: the capital calculation formula and the model validation standards for the internal model approach. The capital calculation formula has changed fundamentally and now relies on the expected shortfall measure and takes into account liquidity risk. We explain the details and state the mathematical foundations of the new formula for capital calculation and extend it to elliptical distributions. We demonstrate that in this case the formula gives an upper bound which can be explained by the aggregation across time taking place in the formula. The model validation regime for internal models now requires that models perform satisfactorily at trading desk level and for a range of quantiles in the tails of the loss distributions. To this end, we develop multi-desk backtests, which simultaneously backtest all trading desk models and which exploit all the information available in the presence of correlation between desks. We first consider tests based on indicator time series for VaR violations before proposing a multi-desk extension of the mono-desk spectral test of Gordy and McNeil (2020) which allow the evaluation of a model at more than one confidence level. The spectral tests make use of realised probability integral transform values based on the estimated loss distribution and these contain more information than the VaR violation indicators. A further proposed test extends Cochran Q-test to obtain a test that VaR violation rates across desks are equal as well as consistent with the targeted exception rate. The new backtests are easy to implement with reasonable running time. In extensive simulation studies, we compare the performance of the tests in terms of size and power.

Contents

List of Tables

List of Figures

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. This work includes the content of the published paper 'On the Basel Liquidity Formula for Elliptical Distributions' (Balter, J. and McNeil, A., 2018. Risks, 6(3), p.92). The opinions expressed in this paper are mine and do not necessarily reflect views shared by the Deutsche Bundesbank or its staff.

Chapter 1

Introduction

1.1 The Basel Accords

The establishment of consistent rules for banking regulation worldwide and the strengthening of the ability of banks to absorb unexpected shocks due to market stress is an ongoing concern for international regulators. The Basel Committee on Banking Supervision (BCBS) was established in 1974 in order to improve and align banking supervision globally across its member jurisdictions. In 2019 it had 45 members from 28 jurisdictions. The members consist of central banks and other authorities responsible for banking regulation[1]. The BCBS sets high-level supervisory standards and guidelines for the regulation of banks and promotes exchange in banking supervision. The BCBS itself is coordinated on a supranational level by the Financial Stability Board, which in turn is authorised by the 'Group of Twenty'. Developed standards set out by the BCBS are interpreted as minimum requirements and are not binding automatically. However, members are expected to implement them through their own domestic authorities. The most important task of the BCBS is the development of the Basel Accords (Basel I, Basel II, Basel II.5 and Basel III[2]). Note that Basel III has fully been integrated into the consolidated Basel framework in January 2021, BCBS (2021), which is now known as Basel III.5[3]. The Basel Accord sets high-level standards for banking supervision and specifically provides recommendations for the three pillars: (1) minimum capital and liquidity requirements; (2) supervisory review process and firm-wide risk management and capital planning; (3) risk disclosure and market discipline. The rules of the minimum capital requirements concern the preservation of a minimum amount of capital which has to be held by banks in order to absorb unexpected losses.

[1] Bank of International Settlements (2020).
[2] BCBS (1988), BCBS (2006), BCBS (2011a) and BCBS (2011b).
[3] Sometimes referred to as Basel IV.

The Basel III Accord focusses on the three major risks that financial institutions are exposed to: credit risk, operational risk and market risk. Market risk is defined as the risk of losses resulting from movements in market prices. This includes, for example, the subtypes interest rate risk, credit spread risk and foreign exchange risk. On a European level, the Basel standards developed by BCBS are passed by the European Parliament under the label Capital Requirements Regulation (CRR) and Capital Requirements Directive (CRD). The adopted framework also guides the responsibility for the implementation of the standards. The CRR becomes binding European law immediately, whereas the directives (CRD) have to be implemented by domestic authorities.

The BCBS issued its first recommendations on market risk as early as 1996, in the form of the 'Amendment to the Capital Accord to incorporate market risks' also known as the 'Market Risk Amendment'[4]. Banks with trading activities have been required by their national competent authorities[5] to set aside capital to protect the bank against extreme losses resulting from adverse movements in market prices of instruments held in the trading book. However, the financial crisis of 2007 to 2009 revealed that the existing framework of Basel II was insufficient since many banks were materially undercapitalised in their trading books during this period of extreme market stress. As a consequence, the BCBS introduced some enhancements to the existing market risk framework, which are referred to as Basel II.5[6]. Simultaneously, the BCBS started an extensive review of the existing market risk framework, known as the Fundamental Review of the Trading Book (FRTB). This review addresses structural deficiencies and shortcomings in the existing minimum capital requirements of Basel II and II.5 for market risk and aims to ensure a more robust market risk measurement. On 14 January 2016 (revised February 2019), the BCBS published its initial version of the revised minimal capital requirements for market risk (BCBS (2019b)). On 27 June 2019 CRR II, which enacts the new FRTB requirements, became European law. Due to the COVID-19 crisis, the deadline for the implementation of the revised market risk framework was postponed by one year to January 2023.

1.2 The Fundamental Review of the Trading Book

According to BCBS (2019a), p.1, the revised minimum capital requirements for market risk address the following four areas:

[4] BCBS (1996).

[5] For example Bundesanstalt für Finanzdienstleistungsaufsicht and Deutsche Bundesbank for Germany, Prudential Regulation Authority and Financial Conduct Authority for United Kingdom.

[6] In particular the introduction of a stressed value-at-risk.

- specification of stricter criteria for the assignment of instruments to the trading book,
- the overhaul of the internal model approach to better address risks that were observed during the crisis,
- the reinforcement of the supervisory approval processes for the use of internal models,
- the introduction of a new, more risk-sensitive standardised approach.

Stricter criteria for the assignment of instruments to either the banking book or trading book were introduced since the assignment had previously been based on trading intentions only. Positions in the trading book are subject to market risk capital requirements, whereas positions in the banking book are subject to credit risk capital requirements. Therefore, banks could potentially optimize their capital requirements and understate the risk inherent in an instrument (BCBS (2019a), p.3). The distinction between the two books is still based on a bank's trading intentions but this is buttressed by some predefined allocation rules for specific instruments. Furthermore, the movement of positions between books is now restricted. The standardised approach is a non-model-based approach to calculating the market risk capital requirements. It is a simple formulaic approach prescribed by the regulatory framework. It is applied by banks which do not use or are not allowed to use the internal model approach. Its design in Basel II was not risk-sensitive enough, which is why changes have been made in the latest version. In the revised framework, the internal market risk model approach is subject to major changes both in the area of model validation and in the capital calculation formula. Banks that have approval to adopt the internal model approach can use an internally developed model to determine their minimum capital requirements for the market risk inherent in their trading book activities. One of the major changes in the internal model approach is the shift from the value-at-risk (VaR) to the expected shortfall risk measure, which now is part of the equation which determines the capital that banks have to set aside for market risk. The intention of this change is to better capture tail risk, since the VaR only measures the loss that a portfolio will not exceed with a certain probability, whereas the expected shortfall measures the average of the losses beyond the VaR cut-off point. Since it also turned out during the financial crisis that many instruments were not as liquid as thought (in the sense that a bank can quickly exit the position or hedge its risk without affecting market prices). Therefore, the new approach includes the risk of market illiquidity. Whereas the old framework assumed a uniform ten day holding period[7], the revised framework recognises granular, asset specific liquidity buckets of differing lengths.

[7]The time required to liquidate asset positions in the trading book or to execute trades to hedge their risks.

The revised framework also introduces more stringent validation standards for internal market risk models. The mandated validation tests still rely heavily on the VaR measure. However, validation actions now have to be performed at bank-wide level as well as at trading desk level[8]. Desks failing to meet the regulatory requirements fall back to the standardised approach, which is generally more costly in terms of capital requirements[9]. Under the revised framework, it is now also required that the performance of the internal models should be satisfactory beyond the usual mandated 99% confidence level. Increased emphasis on the accuracy of VaR estimates at a wider range of confidence levels in the estimated loss distribution seems natural since assessing a wider range of the model will also inform about the adequateness of risk measures read off the estimated distribution function. This also relates to the expected shortfall measure, which has a new prominent role in the capital calculation formula and is calculated by averaging all of the returns in the distribution that are worse than the VaR cut-off point.

1.3 Aims of this thesis

These two revisions motivated this work, which specifically covers aspects of the FRTB in connection with the liquidity-adjusted capital formula based on the expected shortfall measure and the new validation requirements. Regarding the new capital formula, we were especially interested in thoroughly stating the theoretical foundations of the formula and in extending the underlying assumption to more realistic distributions. For the new validation requirements we are interested in formulating a new methodology for bank-wide and desk-level model validation tests, since this issue is left open by the regulations.

The work starts with the analysis of the new Basel liquidity formula. In a nutshell, the formula works as follows. The base liquidity horizon used for the calculation of the expected shortfall is 10 days. Mandated risk factor categories (like interest rates or commodity prices) are mapped to liquidity horizons of differing lengths, where the mapping rule is given by the regulator. The 10-day expected shortfall measure is calculated for various subportfolios of risk factors. The formula then aggregates the various risk factor categories by an approach that relies on square-root-of-time scaling of the base values for the 10-day expected shortfall. Since the regulation itself gives no detailed motivation or thorough statement of the theoretical background and mathematical assumptions for the formula, our research question is *what justifies the Basel*

[8]Large banks typically organise their trading operations in separate trading desks for different kind of instruments such as equities, currencies or derivatives.

[9]Approval for the use of an internal model is now also given on a trading desk basis, and is thus more granular than the previous process mandated in Basel II.5, which allowed the use of an internal model directly at bank-wide level.

liquidity formula and does it remain valid for non-Gaussian risk factors? We will explain the Basel liquidity formula which is based on the assumption that the 10-day risk factor changes form a multivariate Gaussian white noise process, and that the changes in the portfolio value are a linear function of these risk factor changes. We will also analyse the formula under the more general assumption that the risk factor changes form a multivariate elliptical distribution and will show that the regulatory formula gives an upper bound in that case. We will also derive a new formula for calculating the expected shortfall by Fourier inversion for univariate spherical random variables with known characteristic generator, which can be used when we have no access to the density.

The remaining part of this work concerns the revised validation standards. We thereby focus on the new requirement that internal models have to demonstrate satisfactory performance both at bank-wide and trading desk level, as well as the stipulation that a wider range of the estimated loss distribution should be validated. The proposed statistical tests complement bank's existing validation frameworks and can detect potential problems in their models that the latter are likely to miss.

A key research question is thus *are the trading risks of the bank modelled adequately across all desks?* Since the number of trading desks can be quite large, univariate backtesting, that is backtesting of each trading desk model separately, can be quite time consuming and resource intensive; if enough parallel tests are carried out, some are very likely to fail. Moreover, univariate backtesting does not account for cross-sectional dependencies between desks. Conversely, aggregation of the trading desk data into one portfolio data set will result in loss of information and potential biases in the results (for instance one trading desk model under-estimates the risk which is compensated by the over-estimation of risk in another trading desk model). On the other hand, multi-desk backtesting, where all desks are tested jointly and simultaneously, increases the amount of data and thus, in theory, the power of the test.

We suggest that a comprehensive validation framework for a bank should include backtests which exploit all the information available in the separate internal trading desk models and which also control for correlation across desks.

We will first consider tests based on indicator time series for VaR violations, since most literature concentrates on those. Another aim is to develop a multi-desk extension of the more general framework of mono-desk spectral backtests developed by Gordy and McNeil (2020). Their spectral tests make use of the realised probability integral transform (PIT) values of the estimated loss distribution. More precisely, their tests are built upon transformations of the

realised PIT level exceedance indicator function, which contains more information than simple VaR violation indicators. The tests we are going to propose are also very flexible, since the risk modelling group can emphasize the region of the estimated loss distribution about which it is most concerned with respect to model performance and can choose a weighting scheme for observed violations accordingly.

Both types of multi-desk backtests (multi-desk VaR backtesting and multi-desk backtesting using PIT-values) take the form of Z-tests. In these tests the distribution of the test statistic is asymptotically normal but the rate of convergence may distort the performance of the test. Although we will show that the new proposed tests work satisfactorily for backtesting periods typically considered by banks (the number of observations used in backtesting is usually 250 or 500 days), we will also study the benefits of using bootstrap methods. We will see that in our setting, the bootstrap methods generally do not offer any clear advantages over the Z-tests, particularly when the additional computational complexity and increased time resources are taken into account.

The final research question is also connected to desk-based backtesting but focuses on comparative tests, that is tests comparing the performance of the various internal desk models. The formulated research question is *are the trading desks equally good at modelling risk?* To address this research question, the well-known Cochran Q-test (Cochran (1950)) can be used, which is designed to detect significant differences between outcomes for correlated binary data. We will first provide a more thorough derivation of the test's statistical distribution, which is derived heuristically in Cochran's original paper. However, the standard application of the Cochran Q-test can only provide an indication if all desks perform equally (meaning that the observed VaR violation rates do not differ significantly across desks). Since the regulator is particularly interested in whether all internal models lead to numbers of VaR violations that are no more than expected, our aim is to extend the Cochran Q-test to a test that also tests the unconditional coverage hypothesis at the same time. The unconditional coverage hypothesis tests if the unconditional probability of a VaR violation is equal to the expected α coverage rate[10]. Since the new test's distribution has no simple closed-form expression, we will exploit the quadratic form of the test and use a little-known method called Imhof's method (Imhof (1961)) to approximate the distribution of the quadratic form.

All tests in this work are implemented in R (R Core Team (2020)). All codes related to the simulation studies are availably publicly on the Open Science Framework (OSF), see Balter (2021).

[10]See Definition 15 in Section 4.2 for the definition of the unconditional coverage hypothesis.

Chapter 2

Literature review

This chapter gives an overview of the regulatory and academic literature on the topics studied in this thesis.

2.1 Risk measures and capital calculation

The amount of required market risk capital is related to the portfolio risk both under the old and new framework. Under the old framework of Basel II and Basel II.5, banks using the internal model approach were obliged to measure and report their market risk based on a (stressed) 99% VaR of the profit and loss distribution of the their aggregated trading book on a 10-day horizon (see BCBS (2006), p.195, para. 718(b) and (c)). The application of the square-root-of-time rule to scale a one-day VaR measure to a 10-day horizon was not forbidden (see BCBS (2006), p.195, para. 718(c)). The square-root-of-time rule is an attempt to account for illiquidity risk and is based on the scaling of variances with time in diffusion models for asset prices.

The VaR at level α of a portfolio is the loss (measured in currency units) which is not exceeded with probability α over the given time horizon, where α is some high percentage[11]. The VaR measure came under criticism during the financial crisis (see, for example, Committee on Science and Technology (2009)), as it does not give any indication about the potential loss when violated. Also, it is not a coherent risk measure, generally lacking the property of sub-additivity (see Artzner et al. (1999), Rootzén and Klüppelberg (1999) and Embrechts (2000), for example). Therefore, in Basel III.5, the formula for the calculation of the capital requirements for market risk is based on the ES, which is defined as the average of all losses which are greater than the VaR at some level α. That is, the VaR can be interpreted as an interval estimate for large

[11] See Definition 2.

losses at some cut-off point of the profit and loss distribution, whereas the ES is the average of this interval. Also, in contrast to the VaR, the ES is a coherent risk measure (see, for example, Acerbi and Tasche (2002), Frey and McNeil (2002), Inui and Kijima (2005)).

In total, the new formula for the calculation of the capital requirements for market risk is the sum of three components (see BCBS (2019a), pp.5). The ES is used for 'modellable' risk factors, meaning risk factors for which a sufficient amount of historically observed data exists. A second capital add-on is required for risk factors not deemed as modellable (non-modellable risk factors). The third component covers the default risk for credit and equity positions since credit defaults and/or significant downgrades (the migration of an instrument from one rating class to a worse rating class, leading to depreciation of the market value of the position) are often accompanied by disturbances in market liquidity.

The ES metric has to be calculated according to certain rules set out in BCBS (2019b), pp.89. It determines the potential loss a bank might face under a period of market stress, where the time horizon depends on the liquidity of its trading book positions. The required level of confidence for the calculation of the ES is set to 97.5%[12]. Another important change is that the square-root-of-time rule is no longer permitted and has been replaced by a more complex liquidity calculation.

2.2 Liquidity risk

Under Basel II, the relevant VaR measure had to be estimated on a standard 10-day horizon with the option to obtain the 10-day horizon by scaling a one-day VaR to the required 10-day VaR. This implicitly assumed that all positions in the trading book could be exited or hedged within 10 days. The validity of this standard 10-day horizon for all trading positions and the adequacy of the square-root-of-time rule has long been questioned in the literature. Diebold et al. (1998) emphasized that the square-root-of-time rule applied to the VaR only holds under the assumption of independent identically normal distributed underlying (high-frequency) returns, which is usually violated. Danielsson and Zigrand (2006) showed that VaR is underestimated by the square-root-of-time rule in the case where the underlying returns follow a jump-diffusion process. Wang et al. (2011) considered different stylized facts in returns and analysed how they impact the applicability of the square-root-of-time rule. They found that for certain behaviours in the data, like serial dependence, the application of the square-root-of-time

[12]The decrease in the confidence level is due to the fact that for the normal distribution, the 99% quantile mandated in Basel II and Basel II.5 for the VaR is approximately the average of the 97.5% tail.

rule leads to biased results. Regarding the standard 10-day horizon there has been discussions in the literature about the adequacy of this assumption since it does not incorporate liquidity risk and does not recognize that some positions probably require a longer time horizon to be hedged or exited. Finger (2009) noted that internal risk management in banks itself is realised on positions level. That is, the relevant risk horizon is the frequency with which a position can be traded. This is also acknowledged by Lawrence and Robinson (1997), for example. They argue that liquidity risk (in the VaR) is best incorporated by applying the time horizon the investor believes he needs to exit a position. In other words the risk calculation should not rely on one standard time horizon for all positions. A study by Hung et al. (2020) analysed the effect of liquidity on the VaR measure. Based on their findings they suggested liquidity risk in VaR models to prevent underestimated market risk capital requirements.

The new regulatory framework addresses these shortcomings. Under the new framework, the ES has to be calculated on a base liquidity horizon of 10 days, using overlapping 10-day changes of the risk factors. To include illiquidity risk, the granularity of liquidity horizons is increased. Different liquidity buckets are applied for different risk factors, ranging from 10 days (for example for interest rates of specific currencies), to 120 days (for example for volatility of certain commodity prices)[13]. In total, five liquidity buckets are defined. A bank has to calculate five unscaled ES measures on the base 10-day liquidity horizon according to the five given liquidity buckets. That is, first the ES for all risk factors with a liquidity horizon of 10 days or more has to be calculated. Then the ES for all risk factors with liquidity horizon of 20 days or more and so on. The formula for the total ES is obtained from these five ES measures by scaling. As stated in BCBS (2019a), p.7, the new formula also reduces the potential for risk-reducing diversification effects, since the capital requirement resulting from the ES metric is calculated as the average of (i) the sum of single risk factor ES values, which does not benefit from diversification effects and (ii) an integrated ES calculation, which recognizes diversification effects across risk factors.

Despite the shift from VaR to ES in the calculation of the capital requirements for market risk, the VaR still plays a central role, namely in the validation of internal market risk models since the new framework still anticipates backtesting of the VaR value, see BCBS (2019b), p.18, para. 32.4. Under Basel II and Basel II.5, a 'traffic light approach' was mandated in order to assess the accuracy of a VaR model by counting the number of VaR violations within the last 250 observed days. Under the new framework, model validation standards are more stringent. The validation actions have to be performed at desk level and bank-wide level, with desks failing to

[13]The liquidity buckets are specified in BCBS (2019b), p.92, table 2, which are 10, 20, 40, 60 and 120 days.

meet the requirements obliged to compute their capital using the standardised approach. The framework also explicitly mentions, that at bank-wide level, backtesting may also be conducted at calibration levels other than the 99% one and other statistical tests may be performed (see BCBS (2019b), para. 32.13). At trading desk level, a bank must analyse the performance of the VaR measure not only at the 99% but also at the 97.5% level (see BCBS (2019b), para. 32.18). This indicates that the regulator has become more interested in the performance of internal models not only at one confidence level but also at other quantiles. Since the ES is now the measure which determines the capital charge, an indirect validation by assessing the model performance at a wider range of quantiles in the tail of distribution seems reasonable.

2.3 VaR estimation and validation

Over the last few decades, many VaR estimation methods have been proposed in the literature. Methods for the direct modelling of the quantile have been suggested as well as methods for the estimation of the profit and loss distribution (PnL) from which the VaR can be inferred. The various options include non-parametric methods (like historical simulation), parametric models (like GARCH or Monte Carlo simulation) or semi-parametric ones (like extreme value theory or the CaViaR method). For an extensive review of VaR estimation methods see, for example, Abad et al. (2014)).

In this thesis we assume that the risk modelling group estimates or forecasts the unknown PnL distribution function on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{N}_0}, \mathbb{P})$ using the information available at time $t-1$. The conditional PnL distribution function at each time point is denoted by $F_t = \mathbb{P}(L_t \leq x|\mathcal{F}_{t-1})$, where L_t is an $\mathcal{F}_t$-measurable random variable representing the portfolio loss at time t in currency units (i.e. we consider the negative PnL where losses are positive, profits negative).

Since the VaR itself is not observable ex-post, traditional assessment criteria like the mean squared error are not applicable. A method of testing the predictive models using historical data, known as backtesting[14], has become the industry standard for validating VaR models. Under the regulatory framework of Basel II and Basel II.5, backtesting required the comparison of the one-day-ahead VaR forecast to the ex-post observable realised loss. Backtesting strategies in the literature are therefore often based on VaR violation indicators. Most proposed backtests for the accuracy of VaR forecasts are of *univariate* nature and are based on univariate time

[14] Jorion (2007), for example, defines backtesting as a set of statistical procedures designed to check if the real losses are in line with VaR forecasts.

series of VaR violations for *one* portfolio. Let $I_t(\alpha) = I_{\{L_t \geq \widehat{\text{VaR}}_t(\alpha)\}}$ denote a time series of n consecutive violation indicators based on forecasts $\widehat{\text{VaR}}_t(\alpha)$. Christoffersen (1998) defined criteria which should be fulfilled if these forecasts are adequate estimates of the true underlying VaR values. These are known as the *unconditional coverage* and *independence hypotheses*. Both can be combined to obtain the *conditional coverage hypothesis*. One can easily show that the conditional coverage hypothesis holds true if the sequence $(I_t(\alpha))_{t\in\mathbb{N}}$ is a series of independent identically distributed (i.i.d.) Bernoulli variables with event probability $1-\alpha$[15]. Christoffersen emphasizes the importance of testing for the conditional coverage hypothesis in the presence of dynamics associated with stochastic volatility. These are expected in VaR time series since the underlying financial market data typically show stochastic volatility. Moreover, the portfolio composition changes over time and thus portfolio value changes may be non-stationary. Various tests have been proposed using the unconditional coverage, independence and conditional coverage hypotheses as test criteria. Multivariate extensions have largely been neglected. The extension is indeed not straightforward since the data from, for example, several trading desks are likely to be serially correlated over time and cross-correlated across desks. Multivariate backtesting needs to take this into account.

Testing for the unconditional coverage hypothesis means testing whether the number of observed VaR violations is significantly different from the expected number of VaR violations. As already mentioned the VaR forecast is made quite deep in the tail of the PnL distribution with levels around $\alpha = 0.99$. Considering a historical period of 250 days (or one trading year), which is a standard time frame used in market risk model validation, we expect 2.5 violations. If the number of observed VaR violations is significantly higher than $(1-\alpha)\times n$, the risk is potentially under-estimated. Tests of the unconditional coverage hypothesis are closely connected to the old regulatory framework for the calculation of the minimum capital requirements of market risk. The market risk capital multiplier applied to a bank was determined according to the number of VaR violations observed over the past 250 trading days. One of the earliest approaches for verifying the unconditional coverage hypothesis is given by Kupiec (1995) who proposes a likelihood-ratio statistic based on the proportion of VaR violations observed over a certain period of time. The null hypothesis states that the observed VaR violation rate is equal to the expected violation rate under the applied α level. A rejection of the null hypothesis indicates inaccuracies in the internal market risk model used to estimate the VaR. As Kupiec points out, tests for the unconditional coverage hypothesis usually have low power and therefore struggle to detect misspecified VaR models for sample sizes which are typical in the validation of market

[15] See Christoffersen (1998), p.844, Lemma 1.

risk models (one or two years of data, that is 250 or 500 trading days, respectively). The importance of correct conditional coverage of an interval forecast like the VaR (which can be thought of as a special interval forecast bounded at one side only) is also highlighted in Engle (1982): An interval forecast should be dynamic in the sense that it is narrow during an economic stable environment and wider in volatile times. If not, the forecast can be correct on average, i.e. it can satisfy the requirement of correct unconditional coverage, but it may show undesired clustering of VaR exceptions violating the requirement of independence. This is particularly important in the presence of dynamics caused by stochastic volatility in price changes, which can typically be observed in financial data.

The unconditional coverage hypothesis analyses the frequency of VaR violations, whereas the independence hypothesis says something about the way in which they may occur. Independence of any two observations of the VaR violation sequence means that the previous history of VaR violations does not include any further valuable information for future VaR predictions. For example, the presence of violation clusters (VaR violations appearing jointly within a short period of time) contradicts the independence hypothesis and hence contradicts the validity of the underlying VaR model. A valid VaR model must be able to respond to changing market conditions and to reflect them in the VaR predictions so that VaR violations are spread evenly across time and do not appear in clusters. Tests of the unconditional coverage hypothesis usually have no power or only very weak power to detect unwanted dependencies between the VaR violation indicator variables.

One of the first approaches to test for the independence of the sequence of VaR violation indicators $(I_t(\alpha))_{t\in\mathbb{N}}$ is presented in Christoffersen (1998). The author proposes a likelihood ratio based test of the hypothesis of independence against an explicit first-order Markov chain (Markov test). The sequence of VaR violations is expressed as a first-order Markov chain with two states, where the transition probability matrix is derived under the independence assumption. The test then compares the likelihood function under independence with the likelihood under the more general Markov model. Under the null hypothesis, these transition probability matrices should be equal. As pointed out in Christoffersen (1998), this test only tests for independence and the unconditional coverage hypothesis of correct coverage rate can still be violated.

Note that most tests of the independence hypothesis have in common that they need an assumption about the dependence structure which is tested against in the alternative hypothesis. In the case where the VaR violation sequence exhibits a different dependence structure to the one assumed in the alternative hypothesis, the test may have no power to detect this. Another

drawback is also pointed out in Ziggel et al. (2014) who noted that previously, tests for the occurrence of clustered violations focus on testing the violations for independence and neglect testing whether they have an identical distribution. They argue that VaR violation clusters can occur even when the time series of VaR violations is independent. As an example, they mention the fact that the series may not be identically distributed, showing varying probabilities of VaR violations across time. Therefore, they propose a univariate VaR backtest for the full i.i.d. hypothesis, explicitly testing for clusters in the time series of VaR violations.

Duration-based tests are a variety of tests which can be used for both the verification of the conditional coverage hypothesis, the independence hypothesis as well as the unconditional coverage hypothesis. Under the null hypothesis of a correctly specified model, the time between two VaR violations should be 'memory free' and the average duration between violations should be equal to $1/\alpha$ days. As pointed out in Christoffersen and Pelletier (2004), the only memory free continuous function is the exponential function. Duration-based tests need an alternative which allows for dependent durations and which nests the exponential function. For example, Christoffersen and Pelletier (2004) postulate a duration-based test for independence using the Weibull distribution, amongst others. They show that the proposed test has more power than the Markov test which relies on a more strict assumption on the form of the clustering in the alternative hypothesis. A drawback of these kinds of duration-based tests is that they need a specification for the distribution of the duration under the alternative hypothesis. Hurlin et al. (2010) introduce duration based tests for unconditional coverage, conditional coverage and independence hypothesis, which do not require an explicit distribution under the alternative hypothesis. Their tests use the durations between VaR violation indicator variables as input variable and exploit the definition of orthonormal polynomials associated with the geometric distribution. They then use the fact that the polynomial evaluated at a random variable following a geometric distribution is zero under the null hypothesis.

As pointed out in Campbell (2007), joint tests of the unconditional coverage and independence are not automatically preferable to separate testing. Inadequate VaR models, which violate only one of the two hypotheses, are less likely to be detected by a joint test than by the two separate tests. Campbell (2007) therefore suggests separate testing of the hypotheses when previous knowledge about the potential deficiency (a violation of the unconditional or independence property) is present.

Another strand of testing is based on the fact, that the VaR estimate should be valid for every α level and not just for one specific α level when a PnL distribution is estimated adequately. This is also acknowledged by the revised regulatory requirements for the measurement of market risk,

which emphasize that testing should be carried out not only at the 99% level. Testing for more than one specific α level becomes also more important with the new market risk regulations since the expected shortfall is now one of the components of the formula which determines the required capital for market risk. Defined as the average of all losses which are greater than the VaR at level α, an adequate estimate of the expected shortfall requires a PnL forecast, which is accurate in the respective tail of the distribution.

Tests based on multiple VaR α levels, which test simultaneously for the adequacy of a finite number of VaR estimates at different levels $\alpha_1, \ldots, \alpha_N$ were proposed by Kratz et al. (2018), for example. They propose tests for both the unconditional coverage and the independence hypothesis and show that they are more powerful than single-level binomial VaR violation tests when $N \geq 4$. They suggest that these tests can also be used as an implicit backtest of the expected shortfall. In Pérignon and Smith (2008), a multivariate unconditional coverage test is proposed, which takes the form of a multilevel likelihood-ratio test, where the VaR is backtested at several distinct α-values. It is a multivariate generalization of the unconditional test of Kupiec. Another multilevel test is presented in Campbell (2007). He uses the Pearson's chi-squared test for goodness of fit in order to examine regions of α values.

2.4 Backtesting based on PIT-values

Testing for the accuracy of the VaR at more than one α level implies that we move away from a simple assessment of the validity of value-at-risk to a more thorough assessment of the forecasted PnL distribution from which the VaR forecast is calculated. If the PnL distribution is adequately estimated, the VaR must be adequate for every α level in $[0, 1]$. In the extreme case that we test for *every* α level, we backtest the complete PnL distribution. One advantage of backtesting (a region of) the forecasted PnL distribution is that we exploit much more information to verify the validity of the internal model in comparison to using the series of VaR violation indicators at a specific α level. It is obvious that the VaR violation indicator only exploits a limited amount of data. It takes only the values zero and one at every time point t, which provides information whether a VaR violation has occurred or not. Moreover, in risk management applications, we typically face α levels around 99%, so violations occur very rarely. Consequently, the indicator variable is mostly zero. Some backtests for the forecasted PnL distribution are based on realised probability-integral transform (PIT) values[16]. For this, the series $L_1, \ldots, L_n$ of observed profits

[16]The nomenclature in the Fundamental Review of the Trading Book has also been realised p-values in some versions.

and losses is transformed to a series of i.i.d. variables via the probability integral transform. Rosenblatt (1952) shows[17] that the series $(U_t)_{t\in\mathbb{N}}$ consisting of the variables

$$U_t = \int_{-\infty}^{L_t} f_t(u|L_1,\ldots,L_{t-1})\mathrm{d}u = F_t\left(L_t\right)$$

forms a series of i.i.d. standard uniform distributed variables. In this formula, f_t denotes the conditional density of the PnL distribution function of the variable L_t. In the following U_t is called the *theoretical PIT value*. Note that this transformation holds even when the random variables L_t, $t = 1,\ldots,n$, are non-stationary.

Using this insight, backtests of the forecasted PnL distribution function $\widehat{F}_t$ can be constructed by arguing that one would expect that the sequence of *realised PIT-values* $P_t = \widehat{F}_t\left(L_t\right)$, $t = 1,\ldots,n$, to behave like i.i.d. standard uniform variables when a risk modelling group provides ideal forecasts $\widehat{F}_t$ of F_t at any point in time in the sense of Gneiting et al. (2007). As pointed out in Gordy and McNeil (2020), PIT-values contain information about VaR violations at any α level[18]:

$$U_t \geq \alpha \;\Rightarrow\; L_t \geq \mathrm{VaR}_t(\alpha).$$

So tests based on realised PIT-values are expected to be more powerful than tests based on the VaR violation sequence alone when it comes to the detection of deficiencies in the the forecasts $\widehat{F}_t$, $t = 1,\ldots,n$. Note that the criteria defined by Christoffersen (1998) can be directly applied to the sequence of realised PIT-values $(P_t)_{t\in\mathbb{N}}$:

Uniformity In the case that $\widehat{F}_1,\ldots,\widehat{F}_n$ are ideal forecasts of the PnL distribution function, the sequence $(P_t)_{t\in\mathbb{N}}$ must be uniformly distributed on the unit interval [0,1]. Note that this condition is directly linked to the unconditional coverage hypothesis when using the sequence of VaR violation indicators, since both require that the VaR at level α should be violated $(1-\alpha)$-times:

$$\mathbb{P}(P_t \geq \alpha) = \mathbb{E}(I_{\{P_t\geq\alpha\}}) = 1-\alpha.$$

Independence For ideal forecasts $\widehat{F}_1,\ldots,\widehat{F}_n$, the realised PIT-values $P_1,\ldots,P_n$ must be independent from each other. A VaR violation at level α should not contain any information about the occurrence of a future VaR violation at level $\tilde{\alpha}$. This property is directly linked

[17] Under the assumption that the conditional PnL distribution is continuous at each point in time.
[18] Note that the distribution function has to be strictly increasing to apply the following equivalence.

to the independence hypothesis of the sequence of VaR violations indicators.

Both properties can be summarized in the hypothesis that

$$P_t \overset{i.i.d.}{\sim} U(0,1),\ t = 1, \ldots, n,$$

where $U(0,1)$ denotes the standard uniform distribution on $[0,1]$. Backtests can be constructed by testing the uniformity and/or the independence hypothesis. An external regulator can use realised PIT-values as the basis for further validation actions when the forecasts $\widehat{F}_1, \ldots, \widehat{F}_n$ made by the risk modelling group are reported to them. Note that these properties hold irrespective of the method used to forecast the true underlying PnL distribution.

Berkowitz (2001) proposed a backtest based on realised PIT-values based on the fact that for a sequence of i.i.d. standard uniform variables $(U_t)_{t\in\mathbb{N}}$ it holds that $Z_t = \mathcal{N}^{-1}(U_t)$, $t = 1, \ldots, n$, is a sequence of i.i.d. standard normal distributed variables. Using $(Z_t)_{t\in\mathbb{N}}$ he exploits that this sequence must be independent across observations and that the variables must be serially independent and standard normally distributed. More precisely, he proposes a likelihood-ratio test based on a censored (or truncated) likelihood. As he explains, the test compares only the tail of the estimated density to the observed tail and not the entire density (see Berkowitz (2001), p.469). A standard (non-censored) likelihood test will reject a model no matter in which part of the density the forecast deviates from the observation in the first two conditional moments of the data. The proposed test uses the transformed sequence of realised PIT-values $(Z_t)_{t\in\mathbb{N}}$, which are truncated to a desired cut-off point like VaR at a certain α level and then calculates the log-likelihood function using the insight that the values Z_t should have a normal distribution for all t. Another test which accentuates certain areas of a distribution forecast in the backtesting exercise has been proposed by Amisano and Giacomini (2007). Their weighted likelihood ratio test is a relative test, comparing the performance of two competing density forecasts. Their approach is based on weighted averages of the logarithmic scoring rule, where the weights can be chosen according to the preferences of the risk modelling group. Diks et al. (2011) also proposed tests for the comparison of competing forecasts but based them on conditional likelihood and censored likelihood scoring rules. Their test has the advantage that it circumvents unjustified favouritism of density forecasts which place more probability mass on the accentuated region. Gneiting and Ranjan (2011) adopted the weighting approach in Amisano and Giacomini (2007) but applied the (quantile-weighted) continuous ranked probability score instead of the logarithmic score. All of these tests have in common that they analyse a specific region of interest.

Recently, Gordy and McNeil (2020) proposed so-called *spectral tests* of the unconditional and conditional coverage hypothesis, which covers many of previously proposed approaches to backtesting as special cases. Their tests are absolute tests evaluating the weighted performance of one distribution forecast. The tests are based on transformations of the realised PIT-value level exceedances indicators $Y_t(\alpha) = I_{\{P_t \geq \alpha\}}$ of the form

$$W_t = \int_{[0,1]} I_{\{P_t \geq u\}} \mathrm{d}\nu(u),\ t = 1, \ldots, n \tag{2.1}$$

where ν is a Lebesgue-Stieltjes measure defined on $[0, 1]$. This form is very flexible. The risk modelling group or banking regulator can choose the kernel window, which defines the area of the forecast distribution he or she is interested in assessing, mostly in the area around the 99% quantile. Moreover, by choice of the kernel it can also be decided how to weight the probabilities within the observed kernel window since preferences may not be monotonic. Gordy and McNeil (2020) state that the statistic translates to many existing backtesting approaches by choosing specific kernels. For example, the Dirac measure concentrated at α conforms to the test of Kupiec. The statistic also subsumes the test in Berkowitz (2001). The spectral risk measure test of Costanzino and Curran (2015) and the expected shortfall test of Du and Escanciano (2017) are further special cases of the spectral tests in Gordy and McNeil (2020) obtained by choosing a kernel truncated to tail probabilities. Costanzino and Curran (2015) proposed a backtest for any spectral risk measure that is based on spectrally weighted VaR violations. The test assesses the spectral risk measure's performance on a specific tail of interest. Du and Escanciano (2017) proposed a backtest for the expected shortfall.

To apply the spectral tests, only the observed realised PIT-values are required and no information about the forecast distribution is needed. A test of the unconditional coverage hypothesis means to test that the distribution of the W_t values is the one implied when the realised PIT-values are in fact uniform. One test, which is proposed in Gordy and McNeil (2020) for the unconditional coverage hypothesis, takes the form of a Z-test, based on the asymptotic behaviour of the average of the values $W_1, \ldots, W_n$. For a large class of kernels, this Z-test statistic is analytically tractable and its distribution function is approximated by the standard normal distribution. Besides the univariate transformation W_t where one single kernel measure ν is considered, they also propose a multivariate transformation where a set of different kernel measures $\nu_1, \ldots, \nu_m$ is considered. Hence we have a vector $\boldsymbol{W}_t = (W_t^1, \ldots, W_t^m)$, where each entry is an univariate transformation W_t with a different kernel with the form given in (2.1).

2.5 Multivariate backtesting

So far, all of the above tests make use of univariate time series. Under the revised regulatory framework for the trading book (BCBS (2019b)), multivariate backtesting has acquired a new relevance since backtesting of VaR is now required at a bank-wide level *and* trading desk-level. However, multivariate extensions of univariate VaR backtests have not been widely developed (see, for example, Berkowitz et al. (2011), p.8, where some ideas to approach this topic are mentioned but left for future research).

To the best of our knowledge, only two papers deal with multivariate backtesting of the VaR, where we interpret multivariate backtesting in the sense of backtesting VaR forecasts for several portfolios (or desks) simultaneously.

Both Danciulescu (2016) and Wied et al. (2016) propose simultaneous backtests for a set of VaR forecasts using the multivariate time series of the VaR violation indicators as input variable. In the multivariate setting we assume that the bank builds a PnL model $F_{t,i}$ for the Loss $L_{t,i}$ at time t for each desk (or sub-portfolio/bank) $i = 1, \dots, d$. $\widehat{\text{VaR}}_{t,i}(\alpha)$ is the VaR forcast for time t for desk i. Then the multivariate VaR violation indicator variable is given by

$$I_{t,i}(\alpha) = \begin{cases} 1 & \text{if} \quad L_{t,i} \geq \widehat{\text{VaR}}_{t,i}(\alpha), \\ 0 & \text{if} \quad L_{t,i} < \widehat{\text{VaR}}_{t,i}(\alpha). \end{cases} \tag{2.2}$$

The *VaR violation vector* can be defined as $\boldsymbol{I}_{t,\alpha} = (I^1_{t,\alpha}, \dots, I^d_{t,\alpha})'$, for $t = 1, \dots, n$.

Danciulescu (2016) suggests a test for the unconditional coverage hypothesis and a test for the independence hypothesis. The author proposes a multivariate portmanteau test statistic of Ljung-Box type which is applied to the multivariate time series of VaR violation indicator variables. The test for the independence hypothesis simultaneously tests for the absence of cross- and autocorrelations in the multivariate time series of VaR violation indicator variables up to lag K. The sample covariance matrix of the VaR violation vector is used in the test statistic.

In Wied et al. (2016) multivariate tests are proposed for the detection of clustered VaR violations, which violates the conditional coverage hypothesis. They argue there can be two reasons for the occurrence of clustered violations. One reason why clustered violations occur is that the probability of observing a VaR violation could vary over time. For this, Wied et al. propose a backtest for non-constant expectations, formally testing if the sum of observed violations across

desks is constant across time. Another reason for the occurrence of clustered violations is that the time series of VaR violations may not fulfill the multivariate independence hypothesis that the time series are independent across time and across desks[19]. For this, they propose a χ^2-test for the conditional coverage hypothesis, testing for cross-sectional and serial dependence (with correct coverage rates). The null hypothesis assumes that the time series show no serial dependence or cross-sectional dependence up to a lag K with $K > 0$[20].

It also has to be mentioned that due to limited availability there are very few studies using empirical data on bank-wide PnL data and even less on desk-level PnL data. One study on desk-level PnL data is given by Berkowitz et al. (2011). Their data set includes daily realisations from the PnL distribution and the daily forecast of the VaR (calculated using historical simulation) from each of four separate business lines of a large international commercial bank. In this study, various univariate tests like the Markov test of Christoffersen (1998) for the conditional coverage hypothesis, the CaViaR test for autocorrelation of Engle and Manganelli (2004) and the unconditional coverage hypothesis test of Kupiec (1995) are used to backtest the VaR for each trading desk separately. Moreover, the tests are assessed based on their finite-sample size and power properties in a Monte Carlo study. The authors find that the VaR models for two out of the four business lines are rejected due to volatility clustering and that the model of a third business line is rejected by the unconditional coverage hypothesis test of Kupiec.

2.6 Expected shortfall validation

As described in Section 2.3, backtesting of the estimated VaR can be very simple. In order to prove its adequateness, it suffices to compare the series of realised losses with the VaR estimates. For the expected shortfall, validation is not that simple, since it is calculated as the averages of the losses beyond the VaR cut-off point. In fact, it was shown by Gneiting (2011) that the expected shortfall (as opposed to VaR) lacks the property of elicitability (see Definition 19 in Section 4.2). For elicitable risk measures scoring functions exist (see Definition 18 in Section 4.2) which can be used to rank competing forecasting methods. It has long been argued that the lack of elicitability leads to the impossibility of a rigorous backtesting of the expected shortfall (see discussions in Acerbi and Szekely (2017), Emmer et al. (2013), Pitera and Schmidt (2018) or Ziegel (2016), for example). However, Fissler et al. (2015) showed that the expected shortfall is jointly elicitable with VaR and they provide strictly consistent scoring functions which are used

[19]Note that the authors do not speak about desks but business lines, sub-portfolios or more generally banks.
[20]They also proposed a test for the more restrictive null hypothesis assuming $K \geq 0$.

to construct comparative backtests for model selection. It has already been argued in Acerbi and Szekely (2014) that the lack of elicitability only impacts the feasibility of model selection and not the possibility for model testing. In fact, many implicit and explicit backtesting methods for expected shortfall have been proposed in the recent past, which serve as absolute tests of the adequacy of the expected shortfall measure (see Costanzino and Curran (2015), Costanzino and Curran (2018), Du and Escanciano (2017), Kratz et al. (2018), for example).

Chapter 3

On the Basel liquidity formula

3.1 Introduction

With the introduction of the Basel III.5 standard (see BCBS (2021)), the formula for the calculation of the capital requirements for market risk changed fundamentally. Whereas the market risk capital was based on a stressed 99% value-at-risk (VaR) of the profit and loss (PnL) distribution under the old Basel Accords, the Basel III.5 formula relies on expected shortfall (ES). Moreover, the revised rules acknowledge that the previously applied standard 10-day liquidity horizon did not differentiate sufficiently between liquidity risks associated with different risk factors. In fact, during the last financial crisis it turned out that not all instruments are equally liquid.

The ES metric has to be calculated according to certain rules set out in BCBS (2019b), pp.89. In short, the ES has to be calculated at a 97.5% level on a daily basis both for the bank-wide internal model and for each trading desk that uses the internal model approach. The base liquidity horizon for the ES is 10 days, which has to be computed by using overlapping 10-day changes of the relevant risk factors. The formula then aggregates the risk factor categories based on the new concept of liquidity horizons and square-root-of-time scaling from the 10-day ES. This concept of liquidity horizons defines five different liquidity horizons ranging from 10 days (for example for interest rates of specific currencies), to 120 days (for example for volatility of certain commodity prices). These horizons are (conservative) estimates of the amount of time that would be required to execute trades that would eliminate the effects of these risk factors on the value of the portfolio in a period of market illiquidity. The prescribed mapping of risk factors to one of these five liquidity horizons is given in the Basel III.5 framework. Every risk factor affecting the value of positions in the trading book has to be assigned by the bank to it's

unique liquidity horizon LH_j. In total, a bank has to calculate five unscaled ES measures on the base 10-day liquidity horizon (one for each liquidity bucket). The first ES measure includes all risk factors belonging to the first to the fifth liquidity horizon, the second ES measure includes all risk factors belonging to the second to the fifth liquidity horizon and so on. That is, the expected shortfall measures are calculated with respect to shocks to the risk factors with specified liquidity horizons while the other risk factors are held constant. The precise formula for the regulatory liquidity-adjusted expected shortfall is given by

$$\mathrm{ES} = \sqrt{(\mathrm{ES}_T(P))^2 + \sum_{j\geq 2}\left(\mathrm{ES}_T(P,j)\sqrt{\frac{\mathrm{LH}_j - \mathrm{LH}_{j-1}}{T}}\right)^2} \tag{3.1}$$

(see BCBS (2019b), p.90), where

- $T = \mathrm{LH}_1$ is the the base liquidity horizon of 10 days.
- $\mathrm{ES}_T(P)$ is the expected shortfall at horizon T for a portfolio P with respect to shocks to all risk factors to which the positions in the portfolio are exposed (total ES).
- $\mathrm{ES}_T(P,j)$ is the expected shortfall at horizon T for a portfolio P with respect to shocks to the risk factors which have a liquidity horizon of length LH_j or greater, with all other risk factors held constant (four partial ES).

The expression $\frac{\mathrm{LH}_j - \mathrm{LH}_{j-1}}{T}$ in formula (3.1) is the scaling factor used to scale up $\mathrm{ES}_T(P,j)$ from the base liquidity horizon of $T = 10$ days to the required liquidity horizon. The Basel III.5 framework considers five liquidity buckets of liquidity horizons of 10, 20, 40, 60 and 120 days. Under the assumption that all risk factors belong to the bucket of the longest liquidity horizon of 120 days, the regulatory liquidity-adjusted ES would directly scale from the base expected shortfall $\mathrm{ES}_T(P)$ by $\mathrm{ES} = \mathrm{ES}_T(P)\sqrt{12}$, which is roughly a tripling of the base ES.

In this chapter we study this new formula and formulate the following research question:

Question 1. ***What justifies the Basel liquidity formula and does it remain valid for non-Gaussian risk factors?***

So the aim of this chapter is to provide a theoretical justification for the Basel liquidity formula in (3.1) and to investigate how it can be generalised to allow more realistic distributional assumptions. Precisely, we show that the Basel liquidity formula can be justified under the assumption that the 10-day risk factor changes form a multivariate Gaussian white noise process

with the additional common assumption that changes in the portfolio value are a linear function of these risk factor changes. Following this, we analyse the Basel liquidity formula under the more general assumption that the 10-day risk factor changes form a multivariate elliptical distribution and derive a generalised formula. The behaviour of the Basel liquidity formula for elliptically distributed risk factor changes is interesting because the elliptical family contains more realistic distributions which are able to model heavy tails and tail dependencies. In fact, many results in quantitative risk management apply equally when multivariate normal assumptions are generalised to multivariate elliptical assumptions. In particular, when losses depend linearly on the underlying risk factors, aggregation of risk measures across different business lines, desks or risk factors can generally be based on a common formulaic approach, regardless of the exact choice of elliptical distribution, see McNeil et al. (2015), pp.275. The difference in the regulatory liquidity-adjusted ES formula is that aggregation takes place not only across risk factors, but also across time and a central limit effect takes place. We show in a simulation study that formula (3.1) gives an upper bound when the distribution of risk factor changes belong to the family of heavier-tailed symmetric generalised hyperbolic distributions. That means that in this case the formula leads to a higher capital charge than is actually necessary to achieve the targeted level. The simulation study also analyses how the level of accurateness of the Basel liquidity formula is influenced by correlated risk factors. As a by-product of our analyses we also demonstrate a new (Fourier) approach for the calculation of VaR and ES for general univariate symmetric loss distributions when the characteristic function is known. This approach is particularly useful in cases where we take convolutions of symmetrically distributed univariate random vectors and do not have a simple closed-form expression for their probability densities.

3.2 Mathematical foundations

In this section we will provide some mathematical background which is used in this chapter. In particular, we state the formal definition of value-at-risk and expected shortfall. We also review some definitions and properties in multivariate time series which are used both in this chapter and subsequent chapters of this thesis. Moreover, we define some specific multivariate distributions which are considered in this chapter including elliptical distributions. We also present some prominent examples of normal variance mixture models which are used in the simulation study. The primary reference is McNeil et al. (2015).

3.2.1 Risk measures

We now turn to the definitions of the risk measures *value-at-risk* and *expected shortfall*. For this we have to define the quantile function.

Definition 1 (Quantile Function)**.** *Given some distribution function F, the generalised inverse $F^{\leftarrow}$ is called the quantile function of F. For $\alpha \in (0,1)$ the α-quantile of F is given by*

$$q_\alpha(F) = F^{\leftarrow}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

See McNeil et al. (2015), p.65, Definition 2.9 (ii).

In the following, we consider a portfolio of risky assets. By L we denote the random variable which describes the portfolio loss over some fixed time horizon[21]. The corresponding distribution function is given by $F_L(l) = \mathbb{P}(L \leq l)$. The distribution function of L is also called the Profit and Loss distribution. Note that we consider the *negative* Profit and Loss function.

The value-at-risk (VaR) defines the maximum loss which is not exceeded with a given probability $1-\alpha$, where α describes the desired confidence level (also known as the coverage rate).

Definition 2 (Value-at-risk)**.** *Given some confidence level $\alpha \in (0,1)$, the VaR of the distribution function of L at confidence level α is given by the smallest number l such that the probability that the loss L exceeds l is not larger than $1-\alpha$. Formally,*

$$\mathrm{VaR}(\alpha) = \mathrm{VaR}_L(\alpha) = \inf\{l \in \mathbb{R} : \mathbb{P}(L > l) \leq 1-\alpha\} = \inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\}.$$

See McNeil et al. (2015), p.64. So the VaR is basically a quantile of the profit and loss distribution and can also be interpreted as a one-sided interval. Next, we define the expected shortfall (ES), which is the average of the VaR values over all confidence levels $u > \alpha$.

Definition 3 (Expected Shortfall)**.** *For a loss L with $\mathbb{E}(|L|) < \infty$ and distribution function $F_L(l) = \mathbb{P}(L \leq l)$, the expected shortfall at confidence level $\alpha \in (0,1)$ is defined as*

$$ES(\alpha) = \mathrm{ES}_L(\alpha) = \frac{1}{1-\alpha}\int_\alpha^1 q_u(F_L)\mathrm{d}u,$$

[21] More precisely, it describes the change in the portfolio value, also describing the profits (loss positive, profits negative). However, the general term loss is often used in the literature so we follow this notation.

where $q_u(F_L)$ denotes the quantile function.

See McNeil et al. (2015), p.69. Note that when F_L is continuous we have $\mathrm{ES}_L(\alpha) = \mathbb{E}(L \mid L \geq \mathrm{VaR}_L(\alpha))$.

3.2.2 Multivariate time series

Mostly, market risk modelling is based on historical observations of financial risk factor changes or Monte-Carlo methods to simulate possible future financial risk factor changes which influence the value of a bank's trading portfolio. The set of risk factor changes forms a multivariate time series. In the following we will list some important definitions and properties of multivariate time series which are either used later or are helpful for the understanding of some of the concepts.

A multivariate time-series model for two or more risk factors is a stochastic process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$, i.e. a family of random vectors, indexed in time order and defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. That is, for any time t we have a d-dimensional random vector $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,d})$, which follows some multivariate distribution. The i-th entry of $\boldsymbol{X}_t$ is the random variable $X_{t,i}$.

For a strictly stationary process the joint probability distribution function of any subset of the multivariate times series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is independent of time.

Definition 4 (Strictly Stationary)**.** *A multivariate time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is strictly stationary if*

$$(\boldsymbol{X}_{t_1}, \ldots, \boldsymbol{X}_{t_n}) \overset{d}{=} (\boldsymbol{X}_{t_1+k}, \ldots, \boldsymbol{X}_{t_n+k}),$$

for all $t_1, \ldots, t_n$, $k \in \mathbb{Z}$ and for all $n \in \mathbb{N}$.

A stochastic process is covariance (or weakly or second-order) stationary if the mean of the process is independent from time, the covariance matrix is finite and the covariance between two observations X_t and X_s only depends on the lag (that is the time between two realisations).

Definition 5 (Covariance Stationary)**.** *A multivariate time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is covariance sta-*

tionary (or weakly or second-order stationary) if the first two moments exist and satisfy

$$\boldsymbol{\mu}(t) = \boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_t], \quad t \in \mathbb{Z},$$
$$\Gamma(t,s) = \Gamma(t+k, s+k), \quad t,s,k \in \mathbb{Z}$$

with mean function $\boldsymbol{\mu}(t)$ *and covariance matrix function* $\Gamma(t,s) = \mathbb{E}[(\boldsymbol{X}_t - \boldsymbol{\mu}(t))(\boldsymbol{X}_s - \boldsymbol{\mu}(s))']$ *of* $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$.

For a multivariate time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ which is covariance stationary, it holds that the covariance between two observations X_t and X_s only depends on the lag $h = t - s$. To see this we can choose $k = s$ and calculate $\Gamma(t-s, 0) = \Gamma(t,s)$. Due to this, we can define the covariance matrix function of a weakly stationary process simply by $\Gamma(h) = \Gamma(h, 0)$.

Note that $\Gamma(0) = \mathrm{Var}(\boldsymbol{X}_t) = \mathbb{E}[(\boldsymbol{X}_t - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})']$ is the variance-covariance matrix

$$\mathrm{Var}(\boldsymbol{X}_t) = \begin{pmatrix} \mathrm{var}(X_{t,1}) & \dots & \mathrm{cov}(X_{t,1}, X_{t,d}) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(X_{t,d}, X_{t,1}) & \dots & \mathrm{var}(X_{t,d}) \end{pmatrix}.$$

The covariance matrix function of a weakly stationary process gives in general the cross-covariance between $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t+h}$. The cross-covariance between two random vectors $\boldsymbol{Z}_1 = (Z_{1,1}, \dots, Z_{1,d})'$ and $\boldsymbol{Z}_2 = (Z_{2,1}, \dots, Z_{2,d})'$ is defined as

$$\mathrm{Cov}(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \mathbb{E}[(\boldsymbol{Z}_1 - \mathbb{E}[\boldsymbol{Z}_1])(\boldsymbol{Z}_2 - \mathbb{E}[\boldsymbol{Z}_2])'].$$

A univariate white noise process is a random process whose values are uncorrelated at different points in time, have zero mean and finite variance. The multivariate extension is given below formally. For the definition we first need to introduce the correlation matrix function.

Definition 6 (Correlation Matrix Function)**.** *Let* $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ *be a covariance stationary multivariate time series. Writing* $D = \mathrm{diag}\left(\sqrt{\mathrm{var}(X_{t,1})}, \dots, \sqrt{\mathrm{var}(X_{t,d})}\right)$, *the correlation matrix function* $R(h)$ *of* $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ *is*

$$R(h) := D^{-1}\Gamma(h)D^{-1}, \quad \forall h \in \mathbb{Z}.$$

With this we can formally define a multivariate white noise process.

Definition 7 (Multivariate White Noise)**.** $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ *is multivariate white noise if it is covariance stationary with correlation matrix function given by*

$$R(h) = \begin{cases} R, & h = 0, \\ 0, & h \neq 0 \end{cases}$$

for some positive-definite correlation matrix R*.*

Note that in our notation the correlation matrix is $R = \mathrm{Cor}(\boldsymbol{X}_t) = D^{-1}\mathrm{Var}(\boldsymbol{X}_t)D^{-1}$ where D is defined as in Definition 6:

$$R = \mathrm{Cor}(\boldsymbol{X}_t) = \begin{pmatrix} 1 & \mathrm{cor}(X_{t,1}, X_{t,2}) & \dots & \mathrm{cor}(X_{t,1}, X_{t,d}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cor}(X_{t,d}, X_{t,1}) & \mathrm{cor}(X_{t,d}, X_{t,2}) & \dots & 1 \end{pmatrix}.$$

Note that $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ being a multivariate white noise process implies that the components of $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ show no cross-correlation except for lag zero. In the case that the components are independent (with finite covariance matrix) we have a multivariate strict white noise process.

Definition 8 (Multivariate Strict White Noise)**.** $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ *is a multivariate strict white noise if it is a series of i.i.d. random vectors with finite covariance matrix.*

We now assume that the stochastic process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ which represents the information available up to time t. That is the process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{Z}}, \mathbb{P})$. A process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ for every $t \in \mathbb{Z}$ if X_t is $\mathcal{F}_t$-measurable for every t. We can now define a martingale difference sequence, which is a stochastic process with certain properties.

A *martingale* is a sequence of random vectors $\boldsymbol{Z}_t$ such that the expected value of $\boldsymbol{Z}_t$ given the previous history of the sequence is $\boldsymbol{Z}_{t-1}$. Defining $\boldsymbol{X}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}$ we obtain a martingale difference:

Definition 9 (Multivariate Martingale Difference)**.** $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ *has the multivariate martingale-*

difference property with respect to the filtration $(\mathcal{F}_t)$ *if* $\mathbb{E}|\boldsymbol{X}_t| < \infty$ *and*

$$\mathbb{E}\left(\boldsymbol{X}_t|\mathcal{F}_{t-1}\right) = \mathbf{0}, \quad \forall t \in \mathbb{Z}.$$

3.2.3 Multivariate distributions

Definition 10 (Multivariate Normal Distribution)**.** *A n-dimensional vector of random variables* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *has a multivariate normal or Gaussian distribution if*

$$\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\mu} + A\boldsymbol{Z}, \tag{3.2}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ *is a vector of i.i.d. univariate standard normal random variables,* $A \in \mathbb{R}^{n\times k}$ *and* $\boldsymbol{\mu} \in \mathbb{R}^n$ *are a matrix and a vector of constants, respectively.*

See McNeil et al. (2015), p.178. Since the multivariate normal distribution is characterised by its mean and covariance matrix we write for multivariate normal distributed random variables

$$\boldsymbol{X} \sim \mathcal{N}_n\left(\boldsymbol{\mu}, \Sigma\right).$$

It is easy to calculate that the mean of $\boldsymbol{X}$ is given by $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and the covariance matrix by $\mathrm{Var}(\boldsymbol{X}) = \Sigma = AA'$, where Σ is a positive-semidefinite matrix. It is well-known that the characteristic function of a standard univariate normal random variable Z is given by $\phi_Z(t) = \mathbb{E}[e^{itZ}] = \exp(-t^2/2)$. The characteristic function of a multivariate normal distributed random vector $\boldsymbol{X}$ can then easily be derived

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left(e^{i\boldsymbol{t}'\boldsymbol{X}}\right) = \exp\left(i\boldsymbol{t}'\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}\right), \boldsymbol{t} \in \mathbb{R}^n. \tag{3.3}$$

See McNeil et al. (2015), p.178. The multivariate normal distribution has some drawbacks, when it comes to the modelling of short-term financial returns (like daily returns). It is well-known that financial returns show certain typical characteristics (often summarized as so-called stylized facts, see for example Cont (2001)), like fat-tails and a profit/loss asymmetry.

Multivariate distribution functions which generalise the multivariate normal distribution to capture fat tails are multivariate normal variance mixture distributions which introduce randomness into the covariance matrix.

Definition 11 (Normal Variance Mixture Distribution)**.** *The random vector* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *is said to have a multivariate normal variance mixture distribution if*

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \sqrt{W} A \boldsymbol{Z}, \tag{3.4}$$

where

- $\boldsymbol{Z} \sim \mathcal{N}_k(\boldsymbol{0}, I_k)$*;*
- $W \geq 0$ *is a non-negative, scalar-valued random variable that is independent of* $\boldsymbol{Z}$*, and*
- $A \in \mathbb{R}^{n \times k}$ *and* $\boldsymbol{\mu} \in \mathbb{R}^n$ *are a matrix and a vector of constants, respectively.*

See McNeil et al. (2015), p.183. The vector $\boldsymbol{\mu}$ is referred to as the *location vector* and $Q = AA'$ is the *dispersion matrix* of the distribution. Note that $\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}(W)Q$, that is the dispersion matrix Q only coincides with the covariance matrix of $\boldsymbol{X}$ when $\mathbb{E}(W) = 1$. For the characteristic function we have

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}'\boldsymbol{\mu}\right) \hat{H}\left(\frac{1}{2}\boldsymbol{t}'Q\boldsymbol{t}\right), \tag{3.5}$$

where $\hat{H}(\theta) = \int_0^\infty e^{-\theta\nu} \mathrm{dH}(\nu)$ is the Laplace-Stieltjes transform of the distribution function H of W. In case that the distribution function H has a density h we can write $\hat{H}(\theta) = \int_0^\infty e^{-\theta\nu} h(\nu) \mathrm{d}\nu$. Based on this remark we will write for normal variance mixture distributed random variables

$$\boldsymbol{X} \sim \mathcal{M}_n\left(\boldsymbol{\mu}, Q, \hat{H}\right). \tag{3.6}$$

Next we will introduce elliptical distributions. The members of the family of elliptical distributions build a class of symmetric distributions which are generally able to model fat tails. Multivariate normal variance mixture distributions belong to this class. Spherical distributions are a special case of elliptical distributions and extend the theory of multivariate standard normal distributions. Spherical distributions are elliptical distributions with a zero mean and uncorrelated components. That is for two vectors of the same length, the value of the density at these points is the same.

Definition 12 (Spherical Distribution)**.** *A random vector* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *has a spherical*

distribution if, for every orthogonal map $U \in \mathbb{R}^{n\times n}$ *(i.e. maps satisfying* $UU' = U'U = I_n$*),*

$$U\boldsymbol{X} \overset{d}{=} \boldsymbol{X}. \tag{3.7}$$

See McNeil et al. (2015), p.196. It is shown in McNeil et al. (2015), p.196, Theorem 6.18, that the following is equivalent

(1) $\boldsymbol{X}$ is spherical.

(2) There exists a function $\psi : \mathbb{R} \to \mathbb{R}$ of a scalar variable such that, for all $\boldsymbol{t} \in \mathbb{R}^n$,

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left(e^{i\boldsymbol{t}'\boldsymbol{X}}\right) = \psi(\boldsymbol{t}'\boldsymbol{t}) = \psi\left(t_1^2 + \cdots + t_n^2\right). \tag{3.8}$$

Equation (3.8) shows that the characteristic function of a spherically distributed random variable is fully determined by a function ψ of a scalar variable. The function ψ is also known as the *characteristic generator* of the spherical distribution. We therefore write

$$\boldsymbol{X} \sim S_n(\psi) \tag{3.9}$$

for a spherical distributed random vector. Next we state the characteristic generators for two special cases: the standard multivariate normal and the standardized normal variance mixture distribution which we will need later.

Examples 1.

- *For a standard normal distributed random vector with uncorrelated margins,* $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{0}, I_n)$*, which is also a spherical distributed random vector,* $\boldsymbol{X} \sim S_n(\psi)$*, since the density is* $f(x) = 1/\sqrt{2\pi}\exp(-1/2x^2)$*, the characteristic generator* ψ *is given by*

$$\psi(s) = e^{-s/2} \tag{3.10}$$

because of $\phi(\boldsymbol{t}) = e^{i\boldsymbol{t}'\boldsymbol{0}-1/2\boldsymbol{t}'I_n\boldsymbol{t}} = e^{-\boldsymbol{t}'\boldsymbol{t}/2} = \psi(\boldsymbol{t}'\boldsymbol{t})$*.*

- *For a standardized normal variance mixture distributed random vector with uncorrelated margins,* $\boldsymbol{X} \sim \mathcal{M}_n(\boldsymbol{0}, I_n, \hat{H})$*, which is also a spherical random vector, the characteristic generator is*

$$\psi(s) = \hat{H}\left(\frac{1}{2}s\right) \tag{3.11}$$

because of $\phi(\boldsymbol{t}) = \exp(i\boldsymbol{t}'\boldsymbol{0})\,\hat{H}\left(\frac{1}{2}\boldsymbol{t}'I_n\boldsymbol{t}\right) = \hat{H}\left(\frac{1}{2}\boldsymbol{t}'\boldsymbol{t}\right) = \psi(\boldsymbol{t}'\boldsymbol{t})$*.*

See McNeil et al. (2015), p.185, equation (6.20) and p.197.

Elliptical distributions are a generalisation of the spherical distribution, given by affine transformations of spherical distributions.

Definition 13 (Elliptical Distribution)**.** $\boldsymbol{X}$ *has an elliptical distribution if*

$$\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\mu} + A\boldsymbol{Y}, \tag{3.12}$$

where $\boldsymbol{Y} \sim S_k(\psi)$ *for some characteristic generator and* $A \in \mathbb{R}^{n\times k}$ *and* $\boldsymbol{\mu} \in \mathbb{R}^n$ *are a matrix and vector of constants, respectively.*

See McNeil et al. (2015), p.200.

That is, elliptical distributions are a class of distributions with symmetric margins, where the joint density has an elliptical contour with mean (location vector) $\boldsymbol{\mu}$ and dispersion matrix $Q = AA'$. The characteristic function of an elliptical distribution is given by

$$\begin{aligned}\phi_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathbb{E}[e^{i\boldsymbol{t}'\boldsymbol{X}}] = \mathbb{E}[e^{i\boldsymbol{t}'(\boldsymbol{\mu}+A\boldsymbol{Y})}] = e^{i\boldsymbol{t}'\boldsymbol{\mu}}\mathbb{E}[e^{i(A'\boldsymbol{t})'\boldsymbol{Y}}] \\ &= e^{i\boldsymbol{t}'\boldsymbol{\mu}}\phi(A'\boldsymbol{t}) = e^{i\boldsymbol{t}'\boldsymbol{\mu}}\psi_{\boldsymbol{Y}}\left(\boldsymbol{t}'Q\boldsymbol{t}\right).\end{aligned} \tag{3.13}$$

Consequently, the characteristic function of a centred elliptical distribution is reduced to $\psi\left(\boldsymbol{t}'Q\boldsymbol{t}\right)$. Due to the dependence of the distribution on its characteristic generator we write

$$\boldsymbol{X} \sim E_n\left(\boldsymbol{\mu}, Q, \psi\right) \tag{3.14}$$

for elliptical distributions. The following is important in relation to the characteristic generator.

Remark 1. *The distribution of an elliptical distributed random vector determines only* $\boldsymbol{\mu}$ *uniquely.* Q *and* ψ *are only determined up to a positive constant. That is, when an elliptical distributed random vector* $\boldsymbol{X}$ *can be represented by* $E_n\left(\boldsymbol{\mu}, Q, \psi\right)$ *it is equally represented by* $E_n\left(\boldsymbol{\mu}, cQ, \psi(\cdot/c)\right)$ *with a constant* $c > 0$*.*

The following properties of elliptical distributions will be useful later.

Theorem 1. *Let* $\boldsymbol{X} \sim E_n\left(\boldsymbol{\mu}, Q, \psi\right)$*, where* $\boldsymbol{\mu} \in \mathbb{R}^n$ *and* $Q \in R^{n\times n}$*, be an elliptical distributed*

random vector

- *Every linear combination of an elliptical distributed random vector is elliptical distributed with the same characteristic generator $\psi : \mathbb{R} \to \mathbb{R}$:*

$$B\boldsymbol{X} + \boldsymbol{b} \sim E_k\left(B\boldsymbol{\mu} + \boldsymbol{b}, QB', \psi\right) \tag{3.15}$$

  for all $B \in \mathbb{R}^{k\times n}$ and $\boldsymbol{b} \in \mathbb{R}^k$.

- *The convolution of two independent elliptical distributed random vectors with the same dispersion matrix Q is also elliptical. If $\boldsymbol{X} \sim E_n\left(\boldsymbol{\mu}, Q, \psi\right)$ and $\boldsymbol{Y} \sim E_n\left(\tilde{\boldsymbol{\mu}}, Q, \tilde{\psi}\right)$ then we have*

$$\boldsymbol{X} + \boldsymbol{Y} \sim E_n\left(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, Q, \bar{\psi}\right), \tag{3.16}$$

  where $\bar{\psi}(u) = \psi(u)\tilde{\psi}(u)$.

See McNeil et al. (2015), p.202.

3.2.4 Symmetric generalised hyperbolic distributions

In the previous section, we have introduced multivariate normal variance mixture distributions, which are a generalisation of the multivariate normal distribution. In a later simulation study, we assume risk factor changes to follow some distribution belonging to the family of symmetric generalised hyperbolic distributions, a prominent example of the normal variance mixture distribution and a special case of the more general so-called generalised hyperbolic distributions.

A multivariate symmetric generalised hyperbolic distribution is obtained by choosing a generalised inverse Gaussian distribution for the non-negative, scalar-valued random variable W in Definition 11 of normal variance mixture models. The density of the generalised inverse Gaussian is given by

$$f_W(w) = \frac{\chi^{-\lambda}\left(\sqrt{\chi\kappa}\right)^{\lambda}}{2K_\lambda\left(\sqrt{\chi\kappa}\right)} w^{\lambda-1} e^{-1/2\left(\chi w^{-1} + \kappa w\right)} \quad , \begin{cases} \chi > 0, \kappa \geq 0 & \text{if } \lambda < 0 \\ \chi > 0, \kappa > 0 & \text{if } \lambda = 0 \\ \chi \geq 0, \kappa > 0 & \text{if } \lambda > 0, \end{cases}$$

where K_λ denotes a Bessel function of the third kind.

In the simulation study, we will consider special cases of the symmetric generalised hyperbolic

distribution. The symmetric hyperbolic distribution and the normal inverse Gaussian distribution are special cases of the symmetric generalised hyperbolic distribution. The multivariate Student t distribution and the variance-gamma distribution are boundary cases of the symmetric generalised hyperbolic distribution:

Hyperbolic distribution (HY)

For $\lambda = 1$, we get the symmetric multivariate hyperbolic distribution (HY).

Normal Inverse Gaussian distribution (NIG)

For $\lambda = -0.5$, we get the normal inverse Gaussian distribution (NIG) which is often used in the modelling of univariate financial returns.

Multivariate Student t distribution (t)

$\boldsymbol{X} \sim t_d(\nu, \mathbf{0}, I_d)$. Note that the multivariate Student t distribution is a boundary case of the symmetric generalised hyperbolic distribution (choosing $\lambda = -\nu/2$ and $\chi = \nu$, $\kappa \to 0$).

Variance-Gamma distribution (VG)

Choosing $\lambda > 0$ we get the symmetric variance-gamma distribution (VG) as a boundary case ($\chi \to 0$) of the symmetric generalised hyperbolic distribution.

Note again that if the random vector $\boldsymbol{X}$ has a spherical distribution, then the characteristic function $\phi_{\boldsymbol{X}}$ takes the form $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \psi\left(t_1^2 + \cdots + t_n^2\right)$, $t \in \mathbb{R}^n$, where ψ is the characteristic generator of the spherical distribution (see also Equation (3.8)). Therefore, the characteristic generator is related to the characteristic function of an univariate spherical distribution by $\phi_X(x) = \psi(x^2)$. Thus it is easy to infer the characteristic function of a multivariate spherical distribution from the characteristic function of the univariate spherical distribution. In the following, we give the characteristic function of some special cases of univariate spherical distributions, which are needed in the remainder of this chapter.

From (3.5) we know that the characteristic function of an univariate normal variance mixture random variable is given by

$$\phi(s) = \int_0^\infty \exp(-1/2 s^2 w) f_W(s). \tag{3.17}$$

From this, the characteristic functions of the Student t distribution and the symmetric generalised hyperbolic distribution can be derived (see Appendix for the detailed steps). The characteristic function of all other distributions mentioned in this section can be derived from those cases as boundary cases.

Student t distribution. Let Z have a Student t distribution with ν degrees of freedom. The characteristic function is given in terms of a Bessel function of the third kind K_λ and a gamma function Γ by

$$\phi(s) = \frac{\nu^{\nu/4} s^{\nu/2}}{2^{\nu/2-1}\Gamma(\frac{1}{2}\nu)} K_{\nu/2}(\sqrt{\nu}s), \quad s > 0 \tag{3.18}$$

and $\text{var}(Z) = \nu/(\nu-2)$ for $\nu > 2$.

Symmetric generalised hyperbolic. Let Z have a symmetric generalised hyperbolic (GH) distribution with parameters λ, χ and κ satisfying $\chi > 0, \kappa \geq 0$ if $\lambda < 0$, $\chi > 0, \kappa > 0$ if $\lambda = 0$ and $\chi \geq 0, \kappa > 0$ if $\lambda > 0$. The characteristic function is given by

$$\phi(s) = \left(\frac{\kappa}{s^2+\kappa}\right)^{\lambda/2} \frac{K_\lambda\left(\sqrt{\chi(s^2+\kappa)}\right)}{K_\lambda\left(\sqrt{\chi\kappa}\right)}, \quad s > 0. \tag{3.19}$$

3.3 A justification for the Basel liquidity formula

Let $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ be a d-dimensional time series of risk-factor changes for all relevant risk factors and assume that these are all defined in terms of simple differences or log-differences. We interpret $\boldsymbol{X}_{t+1}$ as the vector of risk factor changes over the time step $[t, t+1]$. For $h \in \mathbb{N}$, the risk factor changes over the time step $[t, t+h]$ are given additively by

$$\boldsymbol{X}_{[t,t+h]} = \sum_{j=1}^{h} \boldsymbol{X}_{t+j}. \tag{3.20}$$

Without loss of generality let the risk calculation be made at time $t = 0$. We make the following assumptions.

Assumption 1. *(i) The risk factor changes $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ form a multivariate white noise process with mean zero and covariance matrix Σ. That is the process is covariance stationary and is serially uncorrelated.*

(ii) Each risk factor is assigned to a liquidity bucket B_k defined by a liquidity horizon $h_k \in \mathbb{N}$,

$k = 1, \dots, n$.

(iii) The loss (or profit) attributable to risk factors in bucket B_k over any time horizon $h \in \mathbb{N}$ is given by $\boldsymbol{b}_k' \boldsymbol{X}_{[0,\min(h,h_k)]}$ where $\boldsymbol{b}_k$ is a weight vector with zeros in any position that corresponds to a risk factor that is not in B_k.

Assumption 1 (iii) ensures that only risk factors assigned to liquidity bucket B_k are taken into account in bucket B_k and the minimum function $\min(h, h_k)$ guarantees that the holding period is only h_k in case the the liquidity horizon h_k assigned to liquidity bucket B_k is shorter than h. Note that Assumption 1 includes a linearity assumption which is that the change in the portfolio value is a linear function of the underlying risk factors. This assumption is quite common when short time horizons are considered.

Under these assumptions we compute the portfolio loss L over the maximum time horizon h_n assigned to liquidity bucket B_n (for example 120 days). This will be the hypothetical loss incurred if for each risk factor in each liquidity bucket the full liquidity horizon of these liquidity buckets were required to hedge/sell the risk positions. We find that

$$\begin{aligned} L &= \sum_{k=1}^{n} \boldsymbol{b}_k' \boldsymbol{X}_{[0,h_k]} \overset{(1)}{=} \sum_{k=1}^{n} \sum_{j=1}^{k} \boldsymbol{b}_k' \boldsymbol{X}_{[h_{j-1},h_j]} \overset{(2)}{=} \sum_{k=1}^{n} \sum_{j=k}^{n} \boldsymbol{b}_j' \boldsymbol{X}_{[h_{k-1},h_k]} \\ &= \sum_{k=1}^{n} \boldsymbol{\beta}_k' \boldsymbol{X}_{[h_{k-1},h_k]} \end{aligned} \tag{3.21}$$

where $\boldsymbol{\beta}_k = \sum_{j=k}^{n} \boldsymbol{b}_j$ and $h_0 = 0$. That is $\boldsymbol{\beta}_k$ describes the considered portfolio subset. For example, $\boldsymbol{\beta}_1 = \sum_{j=1}^{n} \boldsymbol{b}_j$ describes the total portfolio, whereas $\boldsymbol{\beta}_n = \boldsymbol{b}_n$ represents the risk factors belonging to liquidity bucket B_n.

- Step (1) follows from the fact that $\boldsymbol{X}_{[0,h_k]} = \boldsymbol{X}_{[h_0,h_1]} + \cdots + \boldsymbol{X}_{[h_{k-1},h_k]} = \sum_{j=1}^{k} \boldsymbol{X}_{[h_{j-1},h_j]}$ for the mutually exclusive time steps $[h_{j-1}, h_j]$, $j = 1, \dots, k$, see equation (3.20).

- To see step (2) compare

$$\begin{aligned} \sum_{k=1}^{n} \sum_{j=1}^{k} \boldsymbol{b}_k' \boldsymbol{X}_{[h_{j-1},h_j]} = \boldsymbol{b}_1' \left(\boldsymbol{X}_{[h_0,h_1]} \right) + \boldsymbol{b}_2' \left(\boldsymbol{X}_{[h_0,h_1]} + \boldsymbol{X}_{[h_1,h_2]} \right) \\ + \cdots + \boldsymbol{b}_n' \left(\boldsymbol{X}_{[h_0,h_1]} + \cdots + \boldsymbol{X}_{[h_{n-1},h_n]} \right) \end{aligned}$$

with

$$\sum_{k=1}^{n}\sum_{j=k}^{n} \boldsymbol{b}_j' \boldsymbol{X}_{[h_{k-1},h_k]} = \left(\boldsymbol{b}_1' + \cdots + \boldsymbol{b}_n'\right) \boldsymbol{X}_{[h_0,h_1]} + \cdots + \boldsymbol{b}_n' \boldsymbol{X}_{[h_{n-1},h_n]}.$$

Note that we transformed the expression for the Loss L in (3.21) to a sum of vectors of risk factor changes over the time steps $[h_{k-1}, h_k]$, $k = 1, \ldots, n$. To derive a formula for the variance of L let us write

$$L_k = \boldsymbol{\beta}_k' \boldsymbol{X}_{[h_{k-1},h_k]} \quad \text{for} \quad k = 1, \ldots, n \tag{3.22}$$

for the summands in the final expression in (3.21). These are uncorrelated due to Assumption 1 (i) and we may easily calculate that

$$\operatorname{var}(L) = \sum_{k=1}^{n} \boldsymbol{\beta}_k' \operatorname{Var}(\boldsymbol{X}_{[h_{k-1},h_k]}) \boldsymbol{\beta}_k = \sum_{k=1}^{n} (h_k - h_{k-1}) \boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k, \tag{3.23}$$

where the first step follows from the assumption that the risk factor changes are serially uncorrelated. The final step follows from the assumption that the process of risk factor changes $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is assumed to be covariance stationary with covariance matrix Σ and because equation (3.20) implies that

$$\boldsymbol{X}_{[h_{k-1},h_k]} = \sum_{j=1}^{h_k - h_{k-1}} \boldsymbol{X}_{h_{k-1}+j}. \tag{3.24}$$

We now introduce the random variables

$$L^{(k)} = \boldsymbol{\beta}_k' \boldsymbol{X}_{[0,h_1]}$$

with $k = 1, \ldots, n$. These represent the losses attributable to all risk factors in the union of liquidity buckets $B_k \cup \cdots \cup B_n$ over the base liquidity horizon h_1. Note that the random variables L_k and $L^{(k)}$ differ in general (unless $k = 1$ which describes the loss attributed to all risk factors over the time step $[0, h_1]$).

Since $\operatorname{var}(L^{(k)}) = h_1 \boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k$, we have

$$\operatorname{var}(L) = \sum_{k=1}^{n} (h_k - h_{k-1}) \boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k = \sum_{k=1}^{n} \frac{h_k - h_{k-1}}{h_1} \operatorname{var}(L^{(k)}). \tag{3.25}$$

With that we obtain the following formula for the standard deviation of L

$$\mathrm{sd}(L) = \sqrt{\sum_{k=1}^{n} \left(\sqrt{\frac{h_k - h_{k-1}}{h_1}} \, \mathrm{sd}(L^{(k)}) \right)^2}. \tag{3.26}$$

Suppose that $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is a **Gaussian** process. In this case $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is actually a strict white noise process (a process with independent and identically distributed vectors). It follows that

$$L_k \sim N(0, (h_k - h_{k-1})\boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k)$$

due to the well-known properties of multivariate Gaussian distributions and applying (3.24). Moreover, we have that the random variables L_k are independent for all k. Recall that $L = \sum_{k=1}^{n} L_k$. Thus, by the convolution property for independent normals, which says that the sum of two independent normal distributed random vectors is again normal distributed where the mean and the covariances are simply added up (see McNeil et al. (2015), p.180), we have

$$L \sim N\left(0, \sum_{k=1}^{n} (h_k - h_{k-1})\boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k\right). \tag{3.27}$$

Moreover, $L^{(k)} \sim N(0, h_1 \boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k)$.

For any zero mean normal random variable Z it is easy to show that

$$\mathrm{ES}_Z(\alpha) = c_\alpha \, \mathrm{sd}(Z) \quad \text{with} \quad c_\alpha = \phi(\Phi^{-1}(\alpha))/(1-\alpha),$$

where ϕ denotes the density of the standard normal distribution and $\Phi^{-1}(\alpha)$ denotes the α-quantile of the standard normal distribution function Φ, see McNeil et al. (2015), p.70, for example.

It follows from (3.26) that

$$\begin{aligned} \mathrm{ES}_L(\alpha) &= c_\alpha \sqrt{\sum_{k=1}^{n} \left(\sqrt{\frac{h_k - h_{k-1}}{h_1}} \, \mathrm{sd}(L^{(k)}) \right)^2} \\ &= c_\alpha \sqrt{\sum_{k=1}^{n} \left(\sqrt{\frac{h_k - h_{k-1}}{h_1}} \, \frac{\mathrm{ES}_{L^{(k)}}(\alpha)}{c_\alpha} \right)^2} \\ &= \sqrt{\sum_{k=1}^{n} \left(\sqrt{\frac{h_k - h_{k-1}}{h_1}} \, \mathrm{ES}_{L^{(k)}}(\alpha) \right)^2}, \end{aligned} \tag{3.28}$$

which is the proposed standard formula for the trading book.

3.4 The liquidity formula for elliptical distributions

In this section we assume that the process of risk factor changes $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ follows a centred elliptical distribution. This class of distributions includes the multivariate normal distribution as a special case. In addition to Assumption 1 we assume the following:

Assumption 2. *(i) The process $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is a multivariate strict white noise (an i.i.d. process).*

(ii) The distribution of $\boldsymbol{X}_t$ is elliptical with location vector $\mathbf{0}$, positive-definite dispersion matrix Σ and characteristic generator function $\psi = \psi(s)$. We write $\boldsymbol{X}_t \sim E_d(\mathbf{0}, \Sigma, \psi)$.

Assumption 2 (ii) means that it is $\boldsymbol{X}_t = A\boldsymbol{Y}_t$ for some matrix $A \in \mathbb{R}^{d\times d}$ satisfying $\Sigma = AA'$ and some random variable $\boldsymbol{Y}_t$ which is spherically distributed, written $\boldsymbol{Y}_t \sim S_d(\psi)$. The characteristic function of $\boldsymbol{Y}_t$ is then given by $\phi(\boldsymbol{s}) = E(e^{i\boldsymbol{s}'\boldsymbol{Y}_t}) = \psi(\boldsymbol{s}'\boldsymbol{s})$ (see equation (3.8)). See also Fang et al. (1990) and McNeil et al. (2015) for further details of this special class of distributions.

Note that under the Basel III.5 framework, we consider 10-day risk factor changes. The independence assumption is a strong assumption but less problematic in case of 10-day periods compared to daily data. The assumption is required to analyse convolutions of elliptical distributed random vectors with different characteristic generators.

We need three key properties of elliptical distributions, which follow from Remark 1 and Theorem 1. Let $\boldsymbol{X} \sim E_d(\mathbf{0}, \Sigma, \psi_{\boldsymbol{X}})$ and $\boldsymbol{Y} \sim E_d(\mathbf{0}, \Sigma, \psi_{\boldsymbol{Y}})$ be independent elliptically distributed random variables with the same dispersion matrix Σ and possibly different characteristic generators $\psi_{\boldsymbol{X}}$ and $\psi_{\boldsymbol{Y}}$. Then

$$\boldsymbol{\beta}'\boldsymbol{X} \sim E_1(0, \boldsymbol{\beta}'\Sigma\boldsymbol{\beta}, \psi_{\boldsymbol{X}}) \quad \text{for } \boldsymbol{\beta} \in \mathbb{R}^d \text{ and } \boldsymbol{\beta} \neq \mathbf{0}. \tag{3.29}$$

$$\boldsymbol{X} \sim E_d(\mathbf{0}, c\Sigma, \psi(s/c)) \quad \text{for any } c > 0. \tag{3.30}$$

$$\boldsymbol{X} + \boldsymbol{Y} \sim E_d(\mathbf{0}, \Sigma, \tilde{\psi}), \quad \text{where } \tilde{\psi}(s) = \psi_{\boldsymbol{X}}(s)\psi_{\boldsymbol{Y}}(s). \tag{3.31}$$

First note that if the process $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is assumed to be elliptical distributed, $\boldsymbol{X}_t \sim E_d(\mathbf{0}, \Sigma, \psi)$, then it follows from (3.20) and (3.31) that $\boldsymbol{X}_{[h_{k-1},h_k]} \sim E_d(\mathbf{0}, \Sigma, \psi_k)$ with $\psi_k = \psi^{h_k - h_{k-1}}$. Recall that we need to calculate the distributions of

$$L_k = \boldsymbol{\beta}_k'\boldsymbol{X}_{[h_{k-1},h_k]}, \quad L = \sum_{k=1}^{n} L_k \quad \text{and} \quad L^{(k)} = \boldsymbol{\beta}_k'\boldsymbol{X}_{[0,h_1]}. \tag{3.32}$$

Using property (3.29), we have for L_k and $L^{(k)}$ that

$$L_k \sim E_1\left(0, \boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k, \psi_k\right) \quad \text{and} \quad L^{(k)} \sim E_1(0, \boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k, \psi_1)$$

with $\psi_1 = \psi^{h_1}$.

Using property (3.30) which states that the dispersion matrix of an elliptical distribution is only uniquely determined up to a positive constant $c > 0$, we can rewrite $L_k \sim E_1\left(0, 1, \psi_k(s\boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k)\right)$. Now, each L_k has the same dispersion matrix and we can use the convolution property (3.31) to conclude that $L \sim E_1(0, 1, \psi^*)$ with

$$\psi^*(s) = \prod_{k=1}^{n} \psi_k(s\boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k). \tag{3.33}$$

It may be easily verified that when $\psi(s) = \exp(-s/2)$ (the Gaussian case) that the characteristic function $\phi(s) = \psi^*(s^2)$ implied by (3.33) is the characteristic function of the normal distribution in (3.27).

The key distributional statements may also be written in terms of spherical distributions since

$$\boldsymbol{X} \sim E_d\left(\boldsymbol{\mu}, \Sigma, \psi\right) \Longleftrightarrow \Sigma^{-1/2}\left(\boldsymbol{X} - \boldsymbol{\mu}\right) \sim S_d(\psi)$$

in case that Σ is positive definite:

$$L^{(k)} / \sqrt{\boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k} \sim S_1\left(\psi_1\right), \quad L \sim S_1(\psi^*).$$

For any location-scale family (like the family of elliptical distributions), there exist positive constants c_{α,ψ^*} and c_{α,ψ_1}, depending on the characteristic generators ψ^* and ψ_1, such that $\mathrm{ES}_L(\alpha) = c_{\alpha,\psi*}\,\mathrm{sd}(L)$ and $\mathrm{ES}_{L^{(k)}}(\alpha) = c_{\alpha,\psi_1}\,\mathrm{sd}(L^{(k)})$. In general these constants will differ (except for the Gaussian case). We can then easily conclude that the general liquidity formula is

$$\begin{aligned}
\mathrm{ES}_L(\alpha) &= c_{\alpha,\psi*} \sqrt{\sum_{k=1}^{n} \left(\sqrt{\frac{h_k - h_{k-1}}{h_1}} \frac{\mathrm{ES}_\alpha(L^{(k)})}{c_{\alpha,\psi_1}} \right)^2} \\
&= \frac{c_{\alpha,\psi*}}{c_{\alpha,\psi_1}} \sqrt{\sum_{k=1}^{n} \left(\sqrt{\frac{h_k - h_{k-1}}{h_1}}\, \mathrm{ES}_{L^{(k)}}(\alpha) \right)^2}.
\end{aligned} \tag{3.34}$$

In general, we expect that $c_{\alpha,\psi^*} \leq c_{\alpha,\psi_1}$, due to the central limit/aggregation across time effect when the distribution of the risk factor changes is heavier-tailed than the normal distribution.

Due to the aggregation across time, more and more 10-day returns are aggregated, which makes the distribution less heavy tailed and more normal-like. Due to the applied square-root-of-time scaling it is implicitly assumed that the distribution of the 20-day to 120-day risk factor changes follow the same distribution as the 10-day risk factor changes. This additionally overestimates the risk. So we expect that the Basel liquidity formula, which is given when $c_{\alpha,\psi^*} = c_{\alpha,\psi_1}$, gives an upper bound for the ratio of the two constants.

3.5 Calculating expected shortfall by Fourier inversion

To calculate $\mathrm{ES}_\alpha(L)$, we require a method of computing the constants for univariate spherically distributed random variables. We can not rely on always having access to a density since we will take convolutions of spherical distributions. In the following, we derive a formula that can be used to compute the expected shortfall from the characteristic function of a spherical random variable, which will always be accessible to us, when the characteristic generator is known.

A univariate spherical random variable $Z \sim S_1(\psi)$ is symmetric about the origin with a real-valued even characteristic function given by $\phi_Z(s) = \psi(s^2)$. We give a general result that applies to univariate random variables that are symmetric about the origin.

Theorem 2. *Let Z be symmetrically distributed about the origin with an integrable characteristic function $\phi(s)$. Let $-\infty < a < b < \infty$. Then the following formulas hold:*

$$f_Z(z) = \frac{1}{\pi}\int_0^\infty \cos(sz)\phi(s)\mathrm{d}s, \tag{3.35}$$

$$F_Z(z) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\sin(sz)}{s}\phi(s)\mathrm{d}s, \tag{3.36}$$

$$\mathbb{E}(ZI_{\{a\le Z\le b\}}) = \frac{1}{\pi}\int_0^\infty \frac{bs\sin(bs)+\cos(bs)-as\sin(as)-\cos(as)}{s^2}\phi(s)\mathrm{d}s\,. \tag{3.37}$$

Proof. The characteristic function $\phi(s)$ of a random variable that is symmetric about the origin is real-valued and even. If ϕ is integrable then the density exists and the standard Fourier inversion formula for the characteristic formula yields

$$f_Z(z) = \frac{1}{2\pi}\int_{-\infty}^\infty e^{-isz}\phi(s)\mathrm{d}s = \frac{1}{\pi}\int_0^\infty \cos(sz)\phi(s)\mathrm{d}s.$$

The step is a consequence of the symmetry property of the sine and cosine function.

The formula (3.36) for the distribution function is obtained from a well-known representation

of the distribution by Gil-Pelaez (1951).

To derive (3.37) we observe that

$$\begin{aligned}\int_a^b z f_Z(z) &= \frac{1}{\pi}\int_a^b \int_0^\infty z\cos(sz)\phi(s)\mathrm{d}s\mathrm{d}z \\ &= \frac{1}{\pi}\int_0^\infty \left(\int_a^b z\cos(sz)\mathrm{d}z\right)\phi(s)\mathrm{d}s\end{aligned}$$

by Fubini's Theorem since $|z\cos(sz)\phi(s)| \leq |z||\phi(s)|$ and the latter is integrable on $[a,b]\times[0,\infty)$. The inner integral can be solved by parts to obtain

$$\int_a^b z\cos(sz)\mathrm{d}z = \frac{bs\sin(bs)+\cos(bs)-as\sin(as)-\cos(as)}{s^2}$$

and (3.37) follows. □

These formulas permit the accurate evaluation of the value-at-risk and expected shortfall using one-dimensional integration. Calculation of $\mathrm{VaR}_\alpha(Z)$ for $\alpha > 0.5$ is accomplished by numerical root finding using (3.36). If $\mathbb{E}|Z| < \infty$ for the distribution in question, the expected shortfall is defined and can be calculated by setting $a = \mathrm{VaR}_\alpha(Z)$ and computing the limit

$$\mathrm{ES}_Z(\alpha) = \lim_{b\to\infty} \frac{1}{\pi(1-\alpha)}\int_0^\infty \frac{bs\sin(bs)+\cos(bs)-as\sin(as)-\cos(as)}{s^2}\phi(s)\mathrm{d}s\,. \tag{3.38}$$

3.6 Calculation of the scaling ratio

In this section, we describe the steps required to calculate the scaling ratio $c_{\alpha,\psi^*}/c_{\alpha,\psi}$ in (3.34) when the underlying real-valued risk factors have some known symmetric distribution function.

Recall the basic components we require for the calculation. It is $L \sim S_1(\psi^*)$, where ψ^* is given in (3.33). It further is $L^{(k)} = \sqrt{\boldsymbol{\beta}_k'\Sigma\boldsymbol{\beta}_k}Z$ with $Z \sim S_1(\psi_1)$ and $\psi_1 = \psi^{h_1}$. We assume that $X \sim S_1(\psi)$ has some known characteristic function $\phi_X(s) = \psi(s^2)$ (one-to-one correspondence with inversion formula) with known standard deviation. The steps are:

1. Calculate $\mathrm{ES}_Z(\alpha)$ using (3.38) and $\phi_Z(s) = \phi_X^{h_1}(s)$
2. Calculate $\mathrm{sd}(Z) = \sqrt{h_1}\,\mathrm{sd}(X)$
3. Hence calculate $c_{\alpha,\psi_1} = \mathrm{ES}_Z(\alpha)/\,\mathrm{sd}(Z)$
4. Calculate $\mathrm{ES}_L(\alpha)$ using (3.38)

5. Calculate $\mathrm{sd}(L)$ using (3.26) and the fact that $\mathrm{sd}(L^{(k)}) = \sqrt{h_1 \boldsymbol{\beta}_k' \Sigma \boldsymbol{\beta}_k}\,\mathrm{sd}(X)$

6. Calculate $c_{\alpha,\psi*} = \mathrm{ES}_L(\alpha)/\,\mathrm{sd}(L)$

7. Calculate the ratio $c_{\alpha,\psi*}/c_{\alpha,\psi_1}$

Note that $\mathrm{sd}(Z) = \sqrt{h_1}\,\mathrm{sd}(X)$ follows from the fact that $Z \overset{d}{=} \sqrt{h_1}X$ and $\mathrm{var}(\sqrt{h_1}X) = h_1\,\mathrm{var}(X)$ due to the assumption that $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is a strict white noise process.

3.7 Application to market data

We follow the calculation steps described in the previous section to calculate $c_{\alpha,\psi^*}/c_{\alpha,\psi}$ in (3.34) using real market data. We assume that the market data follow some known symmetric distribution function, with parameter values that are estimated from the data.

In order to calibrate the considered distributions, we use 10 years (2,518 observations) of standardised log-returns of the S&P500 index, ranging from 15 May 2006 to 14 May 2016. The daily log-returns are calculated using the adjusted closing prices. The distribution parameters are derived by fitting them to biweekly log-returns, i.e. roughly 10 trading days, which is the liquidity base horizon under the Basel III.5 framework. For the fitting, we use the R package `ghyp` (see Breymann and Lüthi (2013)).

In our experiment, we consider distributions belonging to the family of symmetric generalised hyperbolic distributions (see Section 3.2.4), which also belong to the class of elliptical distributions.

Table 3.1 states the fitted shape parameters.

dists | para	λ/ν	α
Gauss	0	
t	2.93	
NIG	-0.5	0.49
Hyp	1	0.11
VG	0.96	0

Table 3.1: Hyperbolic distributions fitted to market data

Next, we calculate the ratio $c_{\alpha,\psi*}/c_{\alpha,\psi}$ for each distribution, applying the parameters stated in Table 3.1. Note that for normal variance mixture models (like the multivariate Student t and symmetric generalised hyperbolic distribution), the covariance matrix is in general not given by

the dispersion matrix Σ, but by $\mathbb{E}(W)\Sigma$. So, we have to calculate $\mathbb{E}(XI_{\{X\geq a\}})/\mathbb{E}(W)$ in order to obtain $c_{\alpha,\psi*}$ and $c_{\alpha,\psi}$. For all distributions we consider here, $\mathbb{E}(W)$ is known[22].

We conduct four experiments:

- We consider two independent risk factors following one of the distribution functions mentioned above. For the dispersion matrix we assume $\Sigma = I_d$ (i.e. independent risk factors). One risk factor belongs to liquidity bucket B_1 with a liquidity horizon of 10 days, the other risk factor belongs to liquidity bucket B_2 with a liquidity horizon of 20 days.
- The second experiment follows in the same fashion but we assume five independent risk factors, each belonging to one of the five liquidity buckets stipulated in the Basel III.5 framework. Recall that the five buckets have liquidity horizons of 10 ($h_1 = 1$), 20 ($h_2 = 2$), 40 ($h_3 = 4$), 60 ($h_4 = 6$) and 120 days ($h_5 = 12$).
- Both experiments are also considered under an equicorrelation model with correlation $\rho = 0.5$.

Table 3.2 and Table 3.3 state the results for various confidence levels α. We report the values for the constants $c_{\alpha,\psi}$ and c_{α,ψ^*} as well as the ratio of the two.

We have seen in Section 3.3 that the standard formula for the trading book holds true for the Gaussian case meaning that the ratio of c_{α,ψ^*} and $c_{\alpha,\psi}$ must be one. This is verified by our experiments (see row 'Gauss'). It also becomes apparent that the scaling ratios are smaller than one for all other distributions meaning that the Basel liquidity formula is indeed conservative when the risk factors follow one of the four multivariate elliptical distributions from a generalised hyperbolic sub-family.

Introducing correlation (see column $\rho = 0.5$) in general leads to larger values for the constant c_{α,ψ^*}. The reason is that the constant depends on the characteristic generator ψ^* and with that depends on the 'variance factors' $\boldsymbol{\beta}_k'\Sigma\boldsymbol{\beta}_k$ (see equation (3.33)) and larger correlation means bigger variance factors. We conclude that the weaker the correlation, the more inaccurate the liquidity formula is (in the sense that the ratio deviates more from the value one). We attribute this to the central limit effect. Considering formula (3.23) we can think of the variance factors as the relative weights attached to each of the n liquidity buckets. When $\rho = 0$, these weights are $(5, 4, 3, 2, 1)$ but when $\rho = 0.5$, for example, they are $(15, 10, 6, 3, 1)$. The intuition is that, for high correlation, the first few liquidity buckets dominate more in the convolution calculation and the central limit effect is mitigated.

[22] If unknown, the calculation of the standard deviation of X is possible using the density function (3.35).

By comparing Table 3.2 with Table 3.3 we further see that the convolution of five independent random variables produces more conservative estimates than the convolution of two independent random variables. The intuition is that we aggregate over more liquidity buckets which increases the central limit effect.

Analysing the results for the different distributions, we see that the Student t distribution produces the least conservative estimates with ratios close to one. Between the variance-gamma, normal inverse Gaussian and the hyperbolic distribution, only slight differences appear. For $\alpha = 0.975$ the normal inverse Gaussian shows the most conservative estimates in the case of the convolution of five independent random variables.

Higher confidence levels α lead to more conservative estimates which indicates that the central limit effect due to convolution is stronger in the tails. Considering the relevant confidence level for the expected shortfall applied under the Basel III.5 framework ($\alpha = 0.975$), we see that the overestimation amounts up to 19.47% in case of independence (16.96% in case of equicorrelation) when we consider five liquidity buckets.

	α	0.95		0.975		0.99	
dists	res | ρ	0	0.5	0	0.5	0	0.5
Gauss	$c_{\alpha,\psi}$	2.063	2.063	2.338	2.338	2.665	2.665
	c_{α,ψ^*}	2.063	2.063	2.338	2.338	2.665	2.665
	ratio	1	1	1	1	1	1
t	$c_{\alpha,\psi}$	2.225	2.225	2.906	2.906	4.063	4.063
	c_{α,ψ^*}	2.213	2.213	2.831	2.839	3.866	3.892
	ratio	0.995	0.995	0.974	0.977	0.951	0.958
VG	$c_{\alpha,\psi}$	2.344	2.344	2.842	2.842	3.497	3.497
	c_{α,ψ^*}	2.247	2.26	2.67	2.696	3.218	3.267
	ratio	0.959	0.964	0.94	0.949	0.92	0.934
Hyp	$c_{\alpha,\psi}$	2.331	2.331	2.821	2.821	3.463	3.463
	c_{α,ψ^*}	2.237	2.25	2.653	2.679	3.194	3.241
	ratio	0.959	0.965	0.94	0.949	0.922	0.936
NIG	$c_{\alpha,\psi}$	2.373	2.373	2.974	2.974	3.828	3.828
	c_{α,ψ^*}	2.295	2.304	2.799	2.823	3.498	3.551
	ratio	0.967	0.971	0.941	0.949	0.914	0.928

Table 3.2: Constants $c_{\alpha,\psi}$, c_{α,ψ^*} and ratios for two risk factors

	α	0.95		0.975		0.99	
dists	res | ρ	0	0.5	0	0.5	0	0.5
Gauss	$c_{\alpha,\psi}$	2.063	2.063	2.338	2.338	2.665	2.665
	c_{α,ψ^*}	2.063	2.063	2.338	2.338	2.665	2.665
	ratio	1	1	1	1	1	1
t	$c_{\alpha,\psi}$	2.22	2.22	2.898	2.898	4.042	4.042
	c_{α,ψ^*}	2.16	2.169	2.636	2.67	3.398	3.483
	ratio	0.973	0.977	0.91	0.921	0.841	0.862
VG	$c_{\alpha,\psi}$	2.344	2.344	2.842	2.842	3.497	3.497
	c_{α,ψ^*}	2.112	2.132	2.429	2.467	2.823	2.89
	ratio	0.901	0.909	0.855	0.868	0.807	0.826
Hyp	$c_{\alpha,\psi}$	2.33	2.33	2.819	2.819	3.458	3.458
	c_{α,ψ^*}	2.108	2.128	2.423	2.459	2.814	2.877
	ratio	0.905	0.913	0.859	0.872	0.814	0.832
NIG	$c_{\alpha,\psi}$	2.373	2.373	2.975	2.975	3.83	3.83
	c_{α,ψ^*}	2.142	2.167	2.49	2.544	2.939	3.041
	ratio	0.902	0.913	0.837	0.855	0.767	0.794

Table 3.3: Constants $c_{\alpha,\psi}$, c_{α,ψ^*} and ratios for five risk factors

3.8 Summary

The main conclusions of this chapter are:

- We explained the theoretical background of the Basel liquidity formula, which is based on the assumption that the risk factor changes form a Gaussian white noise process when the portfolio loss depends linearly of these risk factor changes.
- We derived a formula for the more general assumption that risk factor changes form an elliptical distribution and showed that the Basel liquidity formula gives an upper bound in that case, which we attributed to the aggregation across time effect.
- We presented a new Fourier approach for the calculation of VaR and ES for univariate symmetric loss distributions with known characteristic function. The approach is used to calculate the VaR and ES of convolutions of spherical distributions when we do not have access to the density function.
- In the simulation study we analysed the impact of correlation in the risk factors on the Basel liquidity formula. The weaker the correlation the more inaccurate is the Basel liquidity formula, which can be attributed to the fact that the buckets with smaller liquidity horizon dominate more for correlated data, which mitigates the aggregation effect.

Chapter 4

Multi-desk VaR backtesting

4.1 Introduction

We begin this chapter by recalling some key facts from Chapter 2 about validation of value-at-risk (VaR) estimates, which are relevant in this chapter.

Accurate backtesting of the VaR forecast plays a crucial role for both bank regulators and bank risk managers due to its high importance in the bank's internal steering of market risk and because of its role in the regulatory requirements for internal market risk models. The internal market risk model produces a series of out-of-sample VaR forecasts for a historical period (typically one or two years, i.e. 250 and 500, respectively, trading days), which can be compared to the ex-post observable changes in the PnL distribution. VaR backtesting is often based on VaR violations, which are defined as a realised portfolio loss which exceeds the VaR forecast[23]. Backtesting of the VaR value is therefore well studied in the literature (see Section 2.3). Following Christoffersen (1998), the series of the VaR violation indicator $(I_t(\alpha))_{t\in\mathbb{N}}$ should satisfy two hypotheses in order to be viewed as being consistent with a valid model used for VaR predictions: the unconditional coverage hypothesis (see Definition 15 in Chapter 4.2) and the independence hypothesis (see Definition 16 in Chapter 4.2). Both properties combined mean that VaR violations follow a martingale difference sequence. Many tests have been proposed testing for these properties.

In the Basel III.5 framework, the revised minimum capital requirements for market risk postulate extended validation actions. Previously, the formal requirements stipulated daily backtesting on the bank-wide portfolio level. This means, the ex-ante forecast for the VaR aggregated for

[23] Both under the former and new capital requirements regulation, the capital charge for market risk is directly linked to VaR violations, see Articles 366(2) and 366(3) of Council of European Union (2013) and Article 325bf(6) of Council of European Union (2019).

all relevant trading book positions had to be compared to the ex-post realised loss. Under the Basel III.5 framework, approval for the usage of an internal market risk model is given at trading desk level. In large banks, financial assets which are exposed to market risk are traded on separate trading desks, like desks for commodities, equities or exchange rates. Under the Basel III.5 requirements, a bank must demonstrate for each trading desk separately that the desk's internal market risk model is adequate to measure the inherent risk. Only then approval is given by the regulator to include (respectively to keep) the trading desk in the scope of the internal model (otherwise the trading desk will fall/remain under the standardised approach).

However, it is still required that an internal market risk model is also accurate at the bank-wide portfolio level (so on the aggregated desk level data). The requirements in BCBS (2019b) do not specify which methods and tests have to be used in order to verify the adequateness of the internal market risk models but gives some orientation. It is stated in BCBS (2019b), paras. 32.5 and 32.13, that backtesting of bank-wide risk models should be based on a 99% VaR and that banks should also consider backtesting on other quantiles than the 99% and other statistical tests not listed in the framework. As a consequence, banks are required to implement meaningful tests in order to assess the adequacy of their internal models both at bank-wide and trading desk level. Since the number of trading desks can be quite large, univariate backtesting, that is separate backtesting of each trading desk model separately can be quite consuming in terms of resources (like IT resources and time). Also, separate univariate testing raises the multiple testing issue. If one carries out a certain number of tests each of nominal level β, β% of the tests will falsely reject the null hypothesis (since the probability of falsely rejecting the null hypothesis is given by the nominal level β for each test). Moreover, testing each individual trading desk model separately will not incorporate any cross-correlations across desks. On the other hand, aggregating all trading book data will result in loss of information and probably biased results (imagine that one desk under-estimates the risk but is compensated by the over-estimation of risk in another desk). However, multi-desk backtesting, where all desks are tested jointly, increases the amount of data and thus, in theory, the power of the test, that is the probability of correctly rejecting the null hypothesis when it is false.

Therefore, a meaningful backtesting framework should also include backtests which exploit all the information available in the separate internal desk models and which controls for any correlation structures present across desks.

One way to achieve this is to consider what we call *multi-desk tests* opposed to the previously described univariate, or *single desk tests*. A multi-desk test is a simultaneous backtest of all trading desk models. That is we do not aggregate the univariate time series of VaR violations

across the entire portfolio but instead use all desk's univariate time series of VaR violations separately. This not only allows us to incorporate methods to control for potential dependencies across desks but also allows us to use more information compared to single desk backtesting (simple bank-wide portfolio backtesting). Considering that more information is utilized, multi-desk backtesting could prove to be more powerful in the detection of any falsely specified desk models.

The formulated research question is as follows:

Question 2. ***Are the trading risks of the bank modelled adequately across desks?***

This research question concerns the overall quality of the internal risk models across desks. The aim is to assess the overall quality of the models used by banks to calculate the market risk they are exposed to in their different trading desks. The key technical challenge in the construction of multi-desk tests is to develop a testing method that can overcome the lack of information about the correlation between the trading desks' data.

The chapter is organized as follows. We first state some useful mathematical background required in this and the following chapters. Then we explain the general methodology. That is we will describe the input data of the tests, the hypothesis linked to our research question and we present the general simulation setup, which holds throughout this thesis.

After presenting the general methodology, we consider a univariate[24] binomial score test. We first show why it is problematic to extend this univariate test to a multivariate test by using the well-known Bonferroni method without any control for dependencies across desks. We will show theoretically and in a simulation study why the control for inter-desk dependence data is crucial in order to produce well-sized tests. Then we develop a multi-desk version of the binomial score test which controls for inter-desk dependence and we analyse the size and power properties of the test in an extensive simulation study. The proposed new multi-desk VaR backtests extends the traditional form of univariate testing used in banks.

[24] Univariate in the sense that one univariate time series of VaR violations is considered.

4.2 Mathematical foundations

4.2.1 VaR violation indicator and (un)conditional coverage hypothesis

Accurate backtesting of the VaR forecast plays a crucial role for both bank regulators and bank risk managers due to its high importance in the financial risk measurement for banks and due to its role in the regulatory requirements for internal market risk models. The internal market risk model produces a series of out-of-sample VaR forecasts for a historical period (typically one or two years, i.e. 250 and 500, respectively, trading days), which can be compared to the ex-post observable realisation of the change in the PnL distribution. VaR backtesting is often based on VaR violations, which are defined as a realised portfolio loss which exceeds the VaR forecast. For backtesting purposes, banks have to analyse a historical time series of VaR violations of a minimum length of 250 observations. For every historical point in time, the internal one day ahead VaR forecast is compared to the realised loss. Therefore, we are now considering the loss L_t at time t to introduce the time component. Formally, we consider a portfolio of risky assets, where L_t is a random variable, defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{N}_0}, \mathbb{P})$, which describes the loss at time t. $\mathcal{F}_t$ represents the information available at time t and $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. At every point in time $t \in \mathbb{N}$, L_t is a $\mathcal{F}_t$-measurable random variable. We now define the VaR violation indicator.

Definition 14 (VaR Violation Indicator)**.** *Let L_t denote the portfolio loss realised at time t. $I_t(\alpha)$ denotes the variable associated to the ex-post observation of a $\widehat{\mathrm{VaR}}_t(\alpha)$ violation*

$$I_t(\alpha) = \begin{cases} 1, & \text{if} \quad L_t \geq \widehat{\mathrm{VaR}}_t(\alpha) \\ 0, & \text{if} \quad L_t < \widehat{\mathrm{VaR}}_t(\alpha), \end{cases} \tag{4.1}$$

where $\widehat{\mathrm{VaR}}_t(\alpha)$ is an estimate of the VaR of L_t (a VaR forecast made at t-1).

In Christoffersen (1998) criteria are defined which have to be fulfilled by the VaR violation indicator sequence $(I_t(\alpha))_{t\in\mathbb{N}}$ to speak of valid VaR forecast $\widehat{\mathrm{VaR}}_t(\alpha)$. VaR forecasts are valid if and only if the violation process $I_t(\alpha)$ satisfies the unconditional coverage hypothesis and the independence hypothesis. Both hypotheses can be combined to the conditional coverage hypothesis. We will define and explain the three hypotheses below.

Definition 15 (Unconditional Coverage Hypothesis)**.** *The unconditional probability of a VaR violation must be equal to* $1-\alpha$*, where* α *is also referred to as the coverage rate:*

$$\mathbb{P}(I_t(\alpha) = 1) = \mathbb{E}(I_t(\alpha)) = 1 - \alpha\,, \forall\, t.$$

In the case that the probability is higher than $1-\alpha$, we have observed more violations than expected, that is we face a (possibly significant) underestimation of risk.

The importance of correct conditional coverage of an interval forecast like the VaR is also highlighted in Engle (1982). An interval forecast should be dynamic in the sense, that it is narrow during an economic stable environment and wider in volatile times. If not, the forecast can be correct on average, i.e. having correct unconditional coverage, but can show unwanted clustering of outliers (observations outside the predicted range). This is particularly important in the presence of higher-order dynamics in the underlying time series, which can typically be observed in financial data.

Definition 16 (Independence Hypothesis)**.** *VaR violations are said to occur independently, if*

$$I_t(\alpha) \text{ and } I_s(\alpha) \text{ are independent for every } t \neq s.$$

The unconditional coverage hypothesis deals with the frequency of VaR violations, whereas the independence hypothesis says something about the way in which they may occur. Independence of the violation sequence means that the previous history of VaR violations does not include any further valuable information for the next VaR predictions. For example, violation clusters (the event of more than expected VaR violations in a short period of time) contradicts the independence hypothesis and thereby throws doubt on the validity of the VaR model. A valid VaR model must be able to respond to changing market conditions and to take account of them in the VaR prediction so that VaR violations are spread evenly and not clustered over the considered time period. An unconditional coverage test only tests whether $\mathbb{E}(I_t(\alpha)) = 1-\alpha$. These tests usually have no power to detect unwanted dependencies between the indicator variables. They only test whether the observed number of VaR violations differs significantly from the expected number.

The conditional coverage hypothesis implies that the series of VaR violations $(I_t(\alpha))_{t\in\mathbb{N}}$ is independent and each violation indicator has correct coverage.

Definition 17 (Conditional Coverage Hypothesis)**.** *The violation process is said to have correct conditional coverage if*

$$\mathbb{E}\left(I_t(\alpha)|I_{t-1}(\alpha),\ldots,I_1(\alpha)\right) = 1-\alpha,\ \forall\, t.$$

Note that any two of these three hypotheses imply the other. Christoffersen (1998) shows that testing for correct conditional coverage is equivalent to testing that the underlying sequence $(I_t(\alpha))_{t\in\mathbb{N}}$ is an i.i.d. Bernoulli process, where each Bernoulli variable $I_t(\alpha)$ has correct coverage probability α

$$I_t(\alpha) \sim \text{Bern}(1-\alpha),\ \forall\, t.$$

See Christoffersen (1998), proof of Lemma 1.

The following remark will be used later.

Remark 2. *The sum of independent, identically Bernoulli distributed variables with some parameter α is Binomial distributed with parameter α:*

$$\sum_{t=1}^{n} I_t(\alpha) \sim \text{B}(n, 1-\alpha).$$

Those insights can be used to construct tests for the accuracy of VaR models by testing for the unconditional coverage, independence or conditional coverage hypothesis.

Especially in the context of backtesting the expected shortfall measure, the concept of elicitability is often mentioned. According to McNeil et al. (2015), elicitable statistical functionals are 'functionals that minimise expected scores where the expected scores are calculated using scoring functions' (see p.356). The scoring functions (or scoring rules) describe the difference between a realised value and its forecast by assigning a score and are used to assess the quality of probabilistic forecasts. The formal definition of scoring functions is as follows.

Definition 18 (Scoring Function)**.** *A scoring function is a function $S : \mathbb{R}\times\mathbb{R} \to [0,\infty)$ satisfying, for any $y, l \in \mathbb{R}$:*

(i) $S(y,l) \geq 0$ *and* $S(y,l) = 0$ *if and only if* $y = l$*;*

(ii) $S(y,l)$ *is increasing for* $y > l$ *and decreasing for* $y < l$*;*

(iii) $S(y,l)$ *is continuous in* y*.*

A scoring function is called proper if the forecaster maximises the expected score when the true distribution is forecasted. Scoring functions are closely connected to the theory of elicitability.

Definition 19 (Elicitability)**.** *A real-valued statistical functional T defined on a space of distribution functions $\mathcal{X}$ is said to be elicitable on $\mathcal{X}_T \subseteq \mathcal{X}$ if there exists a scoring function S such that, for every $F \in \mathcal{X}_T$,*

(1) $\int_{\mathbb{R}} S(y,l)\mathrm{dF(l)} < \infty$, $\forall y \in \mathbb{R}$,

(2) $T(F) = \mathrm{argmin}_{y\in\mathbb{R}} \int_{\mathbb{R}} S(y,l)\mathrm{dF(l)}$.

See McNeil et al. (2015), p.356. The scoring function is then said to be strictly consistent for T. If L is a random variable with loss distribution function F_L, then an elicitable risk measure minimises the expected score

$$\mathbb{E}\left(S(y,L)\right) = \int_{\mathbb{R}} S(y,l)\mathrm{dF_L(l)} \tag{4.2}$$

with respect to y for every $F_L \in \mathcal{X}_T$, where $\mathcal{X}_T$ is the set of distribution functions for which the integral in (4.2) is defined (compare McNeil et al. (2015), p.356).

4.2.2 Properties of estimators

We will later consider estimators, which we prove to be consistent. In this section, we formally introduce consistent estimators and state some well-known theorems which are required to prove consistency. An estimator is said to be consistent if it converges in probability: $(T_n \xrightarrow{P} \theta)$

Definition 20 (Consistent Estimators)**.** *Let T_n be an estimator for θ on a sample $X_1, \ldots, X_n$. T_n is a consistent estimator for θ if for every $\varepsilon > 0$*

$$\mathbb{P}\left(|T_n - \theta| > \varepsilon\right) = 0 \text{ for } n \to \infty.$$

The continuous mapping theorem is useful to obtain consistency for functions of consistent estimators.

Theorem 3 (Continuous Mapping Theorem)**.** *Let T_n be a consistent estimator for θ and g is a real-valued function continuous at point θ, then $g(T_n)$ is consistent for $g(\theta)$.*

A useful theorem to proof consistency of an estimator is the strong law of large numbers, which states that the sample mean converges to the expected value almost surely.

Theorem 4 (Strong Law Of Large Numbers)**.** *Let $X_1, \ldots, X_n$ be a sequence of independent identically distributed (i.i.d.) random variables with finite mean*

$$\mathbb{E}[X_i] = \mu < \infty \, \forall \, i = 1, \ldots, n.$$

Then the sample mean $\bar{X}_n = 1/n \sum_{i=1}^{n} X_i$ converges to μ almost surely: $\bar{X}_n \overset{a.s.}{\to} \mu$.

See Resnick (2005). Note that the assumption of i.i.d. random variables can be weakened. For this we define an ergodic stochastic process.

Definition 21 (Ergodic Stochastic Process)**.** *A stationary stochastic process $(\boldsymbol{X}_t)_{t \in \mathbb{N}}$ is said to be ergodic if, for any two bounded functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$,*

$$\begin{aligned} &\lim_{n \to \infty} \left| \mathbb{E}[f\left(X_t, \ldots, X_{t+k}\right) g\left(X_{t+n}, \ldots X_{t+n+m}\right)] \right| \\ &\quad = \left| \mathbb{E}[f\left(X_t, \ldots, X_{t+k}\right)] \right| \left| \mathbb{E}[g\left(X_t, \ldots, X_{t+k}\right)] \right| \end{aligned} \tag{4.3}$$

That is, a stationary process is ergodic if it is asymptotically independent in the sense that dependencies weaken with increasing separation of the variables. The i.i.d. assumption can be weakened to stationary ergodic processes as it is stated in the next proposition.

Proposition 1. *Let $X_1, \ldots, X_n$ be a stationary and ergodic sequence of random variables with finite mean*

$$\mathbb{E}[X_i] = \mu < \infty \, \forall \, i = 1, \ldots, n.$$

Then the sample mean $\bar{X}_n = 1/n \sum_{i=1}^{n} X_i$ converges to μ almost surely: $\bar{X}_n \overset{a.s.}{\to} \mu$.

4.2.3 Copulas

The concept of a copula can be used to describe the dependence between random variables, isolating the description of the dependence structure between individual random variables. The

marginal distributions of the multivariate distribution are thereby uniform on $[0, 1]$. The formal definition is given in the following (see McNeil et al. (2015), p.221).

Definition 22 (Copula)**.** *A d-dimensional copula is a distribution function on* $[0, 1]^d$ *with standard uniform marginal distributions.*

We will use the notation $C(\boldsymbol{u}) = C(u_1, \ldots, u_d)$ for multivariate distribution functions that are copulas. As stated in McNeil et al. (2015), p.221, a copula C is a mapping $C : [0, 1]^d \to [0, 1]$ and three properties must hold:

1. $C(u_1, \ldots, u_d) = 0$ if $u_i = 0$ for any i.
2. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for all $i \in \{1, \ldots, d\}$, $u_i \in [0, 1]$.
3. For all $(a_1, \ldots, a_d)$, $(b_1, \ldots, b_d) \in [0, 1]^d$ with $a_i \leq b_i$ we have

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} C(u_{1i_1}, \ldots, u_{di_d}) \geq 0,$$

   where $u_{j1} = a_j$ and $u_{j2} = b_j$ for all $j \in \{1, \ldots, d\}$.

In this thesis we use two specific copulas, which are the Gauss copula and a t copula (see McNeil et al. (2015), p.226 and p.228). If a d-dimensional random vector $\boldsymbol{X}$ is multivariate normal, $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$, its copula is called Gauss copula. Note that for the Gauss copula the marginal distributions are independent if $\Sigma = I_d$. For the t copula this does not hold true, since uncorrelated t-distributed random variables are not independent. We use the two copulas in our simulation study to generate data from them (see McNeil et al. (2015), pp.229, for a description of the algorithm).

4.3 Methodology

4.3.1 Data

Let the number of desks approved for the use of an internal model be d. The variable $L_{t,i}$ describes the profit and loss (PnL) of the risky assets attributed to desk i at time t (note that we consider the negative PnL distribution, following the convention of a positive sign for a loss). $F_{t,i}$ denotes the conditional profit and loss distribution function given the information up

to time $t-1$ of the random variable $L_{t,i}$, $F_{t,i}(x) = \mathbb{P}(L_{t,i} \leq x|\mathcal{F}_{t-1})$. We assume throughout that the PnL distribution functions are continuous [25]. At time $t-1$ the risk modelling group builds a one-day ahead forecast of the PnL distribution function for each desk i, represented by the estimate $\widehat{F}_{t,i}$. This is the internal model used by the bank to generate a one-day ahead forecast of the risk measure of interest, like the VaR. Let $\widehat{\text{VaR}}_{t,i}(\alpha)$ (see Definition 2) represent the one-day ahead forecast (that is the estimate) of the VaR at level α (typically 99%) of desk i generated at time $t-1$. A VaR violation at time t in desk i, $I_{t,i}(\alpha) = I_{\{L_{t,i} \geq \widehat{\text{VaR}}_{t,i}(\alpha)\}}$ (compare Definition 14), occurs when the realised loss $L_{t,i}$ exceeds the estimated one-day ahead forecast of the VaR for desk i. Let n be the length of the backtesting period, that is the number of days used for the backtesting exercise (typically 250 or 500 days, which is roughly one, two years, respectively). $(I_{t,i}(\alpha))_{t\in\mathbb{N}}$ denotes the time series of VaR violations (*VaR violation indicator sequence*) of desk i.

We group the data of VaR violation indicators $I_{t,i}(\alpha)$, $t = 1, \ldots, n$, $i = 1, \ldots, d$, in a matrix where the columns contain the desks data, so that the rows represent the days. Table 4.1 shows the structure for some stylized sample data.

t \ i	desk 1	desk 2	desk 3	desk 4	desk 5	...	desk d
day 1	1	1	1	1	0	...	0
day 2	$I_{2,1}$=0	$I_{2,2}$=0	0	0	0	...	1
day 3	$I_{3,1}$=1	0	0	0	0	...	1
day 4	0	1	0	0	1	...	1
day 5	0	0	0	0	0	...	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
day n	1	0	0	0	0	...	0

Table 4.1: Stylized sample of the data structure of the VaR violation indicator sequence

Note that VaR violation clusters can naturally both appear across time or across desks. Violations across time violate the independency hypothesis, see Definition 16. Heuristically, this means that the information available at time t (which also includes the previous history of VaR violations up to time t) is not helpful to predict the future state of the VaR violation indicator. VaR violation indicators should be independent across time. In the multivariate setting considering several desks this translates to

$$I_{t,i}(\alpha) \text{ and } I_{s,j}(\tilde{\alpha}) \text{ are independent for every } t \neq s \tag{4.4}$$

[25] In our simulation study, the distribution functions are continuous as well.

for all $\alpha, \tilde{\alpha} \in [0,1]$. Note that in condition (4.4), VaR violation indicators can be correlated across desks within one trading day. Correlation across desks does not violate the property of independence across time. For fixed α we define

$$Y_{t,i} = I_{t,i}(\alpha) = I_{\{L_{t,i} \geq \widehat{\text{VaR}}_{t,i}(\alpha)\}} \tag{4.5}$$

4.3.2 Hypotheses

In this section, we formulate the hypothesis corresponding to our research question 2. Our research question concerns the overall quality of the internal risk models across desks. So we like to assess the overall quality of the models used by banks to calculate the market risk they are exposed to in their different trading desks. Let $\boldsymbol{Y}_t = (Y_{t,1}, \ldots, Y_{t,d})$ be the vector of VaR violation indicators of the desk's data at time t. An appropriate hypothesis for research question 2 is given by

$$H_0 : \mathbf{Y}_1, \ldots, \mathbf{Y}_n \text{ are i.i.d. Bernoulli with } \mathbb{P}(Y_{t,i} = 1) = 1 - \alpha \quad \forall\, t, i. \tag{4.6}$$

We understand our test as a test of the unconditional coverage hypothesis, meaning that the probability of observing a realised loss larger than the estimated VaR (respectively a realised PIT-value larger than α as considered in chapter 5) must be equal to $1 - \alpha$. Although the tests we propose in this chapter are derived under the independence (across time) assumption under the null hypothesis, it is not specifically designed as a test for the independence hypothesis. The proposed tests will only have limited power to detect independence, as noted in Gordy and McNeil (2020). So the focus of our tests remains on the unconditional coverage hypothesis.

4.3.3 Simulation study

In our simulation studies, we analyse both the size (the probability of falsely rejecting the null hypothesis when it is true) and the power (the probability of correctly rejecting the null hypothesis when it is false) of various tests under different settings. The tests are one-sided right-tailed tests since we follow the regulator's perspective, who is primarily interested in a potential risk underestimation. For tests with an asymptotic χ^2-distribution, the tests are two-sided, since the test statistics look for both under- and overestimation by construction.

We will analyse the performance of the tests by varying the following variables potentially having an influence:

- **sample size**: We look at different sample sizes. The sample size corresponds to the number of days used in the bank's backtesting exercise. The length of the backtesting period is typically small (one or two years of daily data). If we only know the asymptotic distribution of a test, the test statistic will converge to its respective asymptotic distribution with the sample size. Therefore, it is of interest how the tests behave for finite sample sizes.

- **number of desks**: The number of trading desks in a bank can be quite large. It is of interest in which way the number of desks influence the test's performance, in particular if only a small number of desks violates the null hypothesis.

- **dependence across desks**: We will assume different dependence structures across desks via sampling from different copulas (in order to consider different correlation structures between the marginal distributions, that is different dependence structures). Further, we will simulate correlated data varying the levels of correlation across desks.

The test performances are compared with respect to *size* and *power*:

- **size**: in order to analyse the size of the test statistics, we assume that the risk modelling group builds correct forecasting models for all desks throughout time and measure how often the null hypothesis is falsely rejected on average.

- **power**: in order to analyse the power, we assume that the risk modelling group uses a misspecified model for a certain amount of desks (ranging from 25% to 100%) and measure how often the null hypothesis is correctly rejected on average.

The nominal level β of all tests is 0.05. A test is called oversized when its empirical size is larger than the nominal level of $\beta = 0.05$.

In our simulation study, we highlight the size and power results with colours. Table 4.2 shows the logic behind the colouring, which shows when we deem the size and power as good, neutral, poor or bad. Note that the choice is rather arbitrary but aids in the interpretation and comparison of the results.

	good	neutral	poor	bad
size	≤ 6	(6,9)	[9,12)	≥ 12
power	≥ 70	(30,70)	(10,30]	≤ 10

Table 4.2: Colouring of size and power results

The general simulation setup is as follows:

Let $\boldsymbol{P_t} = (P_{t,1}, \ldots, P_{t,d})'$ denote the vector of realised probability integral transform values (PIT-values) at time t for all desks $i = 1, \ldots, d$. To examine the size and power of the test we will generate independent random vectors

$$\boldsymbol{P_t} \sim C\left(G_1(u_1), \ldots, G_d(u_d)\right), \tag{4.7}$$

for different copulas C (see Definition 22) and distribution functions G_i which aim to simulate the effects of correct and incorrect estimation of desk-level models. We consider two copula families: the Gauss copula $C_{\mathcal{R}}^{Ga}$ and the t4 copula $C_{4,\mathcal{R}}^{t}$, where $\mathcal{R}$ is a given correlation matrix. For both copulas we consider two different correlation levels for $\mathcal{R}$: the identity matrix (zero correlation) and an equicorrelation (flat) structure in which all pairwise correlations are 0.5.

To check size and power, we simulate from (4.7) choosing $G_i(u) = \Phi(F_i^{-1}(u))$, where Φ is the standard normal distribution (choice of the risk modeller). To check size, all distribution functions F_i, $i = 1, \ldots, d$, must be standard normal distributed. Then, the marginal distributions are uniformly distributed, $G_i(u) = u$, which means that the null hypothesis is correct. To check power, we choose a distribution with a heavier tail than the normal distribution for a certain proportion of the desks. Concretely, we choose all or a certain proportion of distributions F_i being a standardized t4 distribution (scaled so that the variance is one). Then it holds that $G_i(u) < u$, so that we can check the power of the test. Note that these two distribution choices are just a device to generate data which fulfil, respectively violate the null hypothesis. We do not claim that these distributions are the true distributions. We vary the proportion of misspecified desks between 25% (one in four desks is not correctly specified) to 100% (all desks are not correctly specified).

Note that in the case of the t copula, when parametrised with the identity matrix for the correlation structure, the marginal distributions are not independent since the t copula models tail dependence, meaning that the null hypothesis is not true.

In this setting we basically test if the marginal distributions are correct under the null hypothesis. We do not test for the correlation structure. For our simulation study we have chosen the Gauss copula and the very flexible t copula, which allows for tail dependencies and varying correlation across desks. It could be thought of using further, especially asymmetric copulas (like Gumbel or Clayton). This was dismissed in this thesis for simplicity of exposition and restricted resources.

Throughout this thesis we apply the notations presented in Table 4.3. Further parameters may be introduced in the respective sections. The values in the table are the predominantly used

values. Deviations can occur in specific simulation studies which will be made clear in the context.

Variable	Description	Value
`n.sim`	number of repetitions	1000
`n`	number of days	250 or 500
`F`	True marginal distribution of desk	Normal or t4
`C`	Copula used for simulation	Gauss or t4
`d`	number of desks	25 or 50
`rho`	level of correlation	0 ('Zero') or 0.5 ('Flat')
`fraction`	ratio of misspecified desks	25 to 100 percent
`alpha`	VaR level α	0.99

Table 4.3: Description of variables used in the simulation studies

The simulation studies were performed using the R package `simsalapar`, which is a very flexible tool for conducting large-scale simulation studies (see Hofert and Mächler (2016)). The described simulation approach is used throughout the whole thesis. See, for example, Section 4.4.1, Table 4.4 for the first practical example.

4.4 Multi-desk VaR violation tests

In this section we first present a simple approach to extend a univariate test to a multivariate test problem by applying the well-known Bonferroni method to the binomial score test. We show why this simple approach does not work well within our framework. The second approach also considers the binomial score test but controls for dependencies across desks.

4.4.1 Bonferroni method

Kratz et al. (2018) showed that the best alternative out of several compared variants of the binomial test for VaR violations counts is the binomial score test:

$$Z_{\text{BinScore}} = \frac{\frac{1}{n}\sum_{t=1}^{n} I_{\{L_t \geq \widehat{\text{VaR}}_t(\alpha)\}} - (1-\alpha)}{\sqrt{\frac{1}{n}\alpha(1-\alpha)}}. \tag{4.8}$$

A simple way to extend this (or other) test(s) to a multi-desk VaR violation test is the Bonferroni method, named after Carlo Emilio Bonferroni (see, for example, Dunn (1959)). The method corrects for the cumulation of Type I errors when considering multiple tests simultaneously (i.e. testing multiple hypotheses at the same time).

Let us suppose that we consider d desks. This means that we face d null hypotheses, which allows us to use the Bonferroni method to obtain an overall test of desired nominal size β.

Recall that we defined the binomial variable $Y_{t,i} = I_{\{L_{t,i} \geq \widehat{\text{VaR}}_{t,i}(\alpha)\}}$ in (4.5), which takes the value one in case of an observed VaR violation.

Consider the i-th desk. The observed total desk exception rate is given by $\bar{Y}_i = N_i/n$, where $N_i = \sum_{t=1}^{n} Y_{t,i}$ is the number of observed violations in desk i. We expect $n(1-\alpha)$ VaR violations in a desk in the considered time period of n days.

The binomial score test in (4.8) for desk i is based on the statistic

$$T_i = \sqrt{n}\, \frac{\bar{Y}_i - (1-\alpha)}{\sqrt{\alpha(1-\alpha)}}. \tag{4.9}$$

Using asymptotic theory, this test statistic is approximately normally distributed under the null hypothesis, when $Y_{1,i}, \ldots, Y_{n,i}$ are independent. The p-value is then given by $Q_i = 1 - \Phi(T_i)$ (for a right-tailed test), Φ being the standard normal distribution.

In order to form a multivariate test of size β based on the Bonferroni method for testing multiple hypotheses simultaneously, the d null-hypotheses

$$\begin{aligned} H_0^i &: P(Y_{t,i} = 1) < 1 - \alpha \quad \forall\, t \\ H_1^i &: P(Y_{t,i} = 1) \geq 1 - \alpha \quad \text{for at least one } i \end{aligned}$$

are rejected if $\min\{Q_1, ..., Q_d\} \leq \beta/d$.

This is a rather crude backtest. For example, for $n = 250$ days, $d = 50$ desks, $\alpha = 0.99$ and a desired nominal level of $\beta = 0.05$ we have

$$1 - \Phi(T_i) \leq 0.05/d \iff T_i \geq \Phi^{-1}(0.999) \iff T_i \geq 3.09 \iff N_i \geq 7.36. \tag{4.10}$$

This means that all d single hypotheses are rejected when more than eight violations are observed in one of the desks. This can be attributed to the fact that T_i is of discrete nature since $\bar{Y}_i$ describes the average number of violations in desk i, so the actual distribution is discrete and not normal. So the assumption of uniform distributed p-values Q_i under the null hypothesis is a false assumption. This is problematic in the setting of market risk measurement. We face sample sizes of around $n = 250$ or $n = 500$ in the bank's backtestings, combined with required α levels close to one (like 0.99). Moreover, the Bonferroni method implicitly assumes that all tests are independent of each other. For correlated desk data this is certainly not the case.

The probability of rejecting the null hypothesis using the Bonferroni method under the assumption that the underlying desks data are independent is given in the following proposition.

Proposition 2. *Let $Q_1, \ldots, Q_n$ be independent identically distributed random variables. Then we have*

$$\mathbb{P}\left(\min\left\{Q_1, ..., Q_d\right\} \leq \beta/d\right) = 1 - \left(1 - \mathbb{P}\left(Q_i \leq \beta/d\right)\right)^d.$$

Proof. The expression $\mathbb{P}\left(\min\left\{Q_1, ..., Q_d\right\} \leq \beta/d\right)$ implies that at least one Q_i, $i = 1, \ldots, d$, is smaller than β/d. It follows that

$$\mathbb{P}\left(\min\left\{Q_1, ..., Q_d\right\} \leq \beta/d\right) = 1 - \mathbb{P}\left(Q_1 > \beta/d, \ldots, Q_d > \beta/d\right).$$

Since $Q_1, \ldots, Q_d$ are independent, the last equation is equal to

$$1 - \prod_{i=1}^{d} \mathbb{P}\left(Q_i > \beta/d\right).$$

Since all $Q_1, \ldots, Q_d$ have the same distribution it follows $1 - \left(1 - \mathbb{P}\left(Q_i \leq \beta/d\right)\right)^d$. □

Using (4.9) we can calculate $\mathbb{P}\left(Q_i \leq \beta/d\right)$ for one i:

$$\mathbb{P}\left(Q_i \leq q\right) = \mathbb{P}\left(N_1 \geq n(1-\alpha) + \sqrt{n\alpha(1-\alpha)}\Phi^{-1}(1-q)\right).$$

Using this formula with Proposition 2 it can be calculated that for $n = 250$, $d = 50$, $\alpha = 0.99$ and $\beta = 0.05$ that $\mathbb{P}\left(\min\left\{Q_1, ..., Q_d\right\} \leq \beta/d\right) \approx 0.182$. This greatly exceeds the desired nominal size of $\beta = 0.05$. So the nominal size cannot be controlled using the Bonferroni method. This is also evident in the simulation study. We consider the single desk binomial score test in (4.9), counting VaR violations at level $\alpha = 0.99$ with the Bonferroni method for correction of the accumulated sizes. The desired nominal size is $\beta = 0.05$. Table 4.4 shows the results. It is obvious that the Bonferroni method greatly exceeds the desired nominal level of $\beta = 0.05$ (see rows with F = Normal for the empirical sizes of the test under various settings).

	d	50				100			
	ρ	Zero		Flat		Zero		Flat	
n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
250	Normal	18.4	16.5	14.7	11.3	32.8	27.0	26.5	18.2
	t4	85.9	73.4	67.8	49.5	97.4	86.4	81.6	64.1
500	Normal	20.6	21.6	19.3	14.1	15.5	16.0	13.4	9.2
	t4	97.8	93.2	86.9	76.1	98.1	91.7	83.4	70.5

Table 4.4: Estimated size and power of the binomial score test applying the Bonferroni method with $\alpha = 0.99$

4.4.2 Multi-desk VaR violation tests with adjustment for unknown dependencies across desks

In the previous section we have seen that the Bonferroni method is unable to control the size of the mono-desk binomial score test. To construct better sized tests, we will build multi-desk VaR violation tests, considering all desks simultaneously. That is we are not testing d single hypothesis like in the Bonferroni method but rather the Hypothesis in (4.6). We define the total number of VaR violations on a day t by $N_t = \sum_{i=1}^{d} Y_{t,i}$ and the *desk exception rate on day* t by $Z_t = d^{-1}N_t$. Let the expected value given by $\mu_Z = \mathbb{E}(Z_t) = 1 - \alpha$ and the variance by $\sigma_Z^2 = \text{var}(Z_t)$. Let

$$\bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_t$$

be the total average desk exception rate. Then the test statistic

$$T_{MVaR} = \sqrt{n}\,\frac{\bar{Z} - (1-\alpha)}{\sigma_Z} \dot{\sim} \mathcal{N}(0,1) \tag{4.11}$$

is asymptotically standard normally distributed due to the Central Limit Theorem. The test in (4.11) can be used to test if the observed number of VaR violations in the d desks over the backtesting period of length n is significantly higher than expected.

The variance σ_Z is unknown and has to be estimated. Since we expect intra-day correlation across desks (that is correlation on the same trading day t), we have to find appropriate estimates that can control for this unknown correlation. We propose different estimation methods for σ_Z^2 and will analyse the performance of test statistic (4.11) using the proposed alternatives.

Variance estimation This estimation of σ_Z^2 is particularly tricky due to the data basis. Typical α levels, which are wide in the tail, mean that for an appropriate VaR model, a VaR violation will occur quite rarely. For example, a confidence level of 0.99 means that we would expect 2.5 VaR violation in 250 days on average for an appropriate VaR model[26]. Therefore, the VaR violation indicators of desk i, $Y_{1,i}, \ldots, Y_{n,i}$, will mainly be zeros. This leads to severe problems in the estimation of σ_Z^2. If the individual desks VaR violation indicators are independent across desks at any day t, we have

$$\sigma_Z^2 = \frac{\alpha(1-\alpha)}{d}. \tag{4.12}$$

However, positive dependencies across desks are likely on the same trading day t so it can be expected that the unknown variance is larger than (4.12).

In our simulation study, we will apply

$$\hat{\sigma}_Z^2 = \max(\sigma_{\min}^2, \tilde{\sigma}_Z^2),$$

where $\sigma_{\min}^2 = \alpha(1-\alpha)/d$ is the variance of Z_t when we ignore any dependencies between desks and $\tilde{\sigma}_Z^2$ is one of the variance estimators which we will describe below. By taking the maximum, we correct for estimation errors and ensure that the variance is at least as high as the variance under independence.

In the following, we present four alternatives for the estimation of $\tilde{\sigma}_Z^2$. Their performance will be analysed in a simulation study.

None We set $\widetilde{\sigma}_Z^2 = 0$. That is $\hat{\sigma}_Z^2 = \sigma_{\min}^2$. This is true in the case of independent desks. In the case of dependencies across desks, we expect this measure to be too small (since we ignore the correlation structure).

M1 Under the null Hypothesis in (4.6) (when the random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are independent across time and identical distributed) we infer for any t

$$\begin{aligned}\sigma_Z^2 &= \operatorname{var}\left(\frac{1}{d}\sum_{i=1}^{d} Y_{t,i}\right) = \frac{1}{d^2}\sum_{i=1}^{d}\operatorname{var}(Y_{t,i}) + \frac{2}{d^2}\sum_{i=1}^{d}\sum_{i\neq j}\operatorname{cov}(Y_{t,i}, Y_{t,j}) \\ &= \frac{\alpha(1-\alpha)}{d} + \frac{2}{d^2}\sum_{i=1}^{d}\sum_{i\neq j}\operatorname{cov}(Y_{t,i}, Y_{t,j}),\end{aligned}$$

[26]The regulatory framework requires that the backtest compares if the number of observed VaR violations is consistent with a 99% confidence level with the notation that other α levels should be added to the backtesting exercise, see BCBS (2019b), p.81, para. 32.5 and p.83, para. 32.13.

where we used $\text{var}(Y_{t,i}) = \alpha(1-\alpha)\ \forall i$. For the covariance we exploit the equality

$$\text{cov}(Y_{t,i}, Y_{t,j}) = \mathbb{E}(Y_{t,i}Y_{t,j}) - (1-\alpha)^2$$

for $t = 1, \dots n$ and $i, j = 1, \dots, d, i \neq j$.

Using the consistent estimator for $\mathbb{E}(Y_{t,i}Y_{t,j})$ given by $\widetilde{Y}_{ij} = 1/n \sum_{t=1}^{n} Y_{t,i}Y_{t,j}$, we get

$$\widetilde{\sigma}_Z^2 = \frac{\alpha(1-\alpha)}{d} + \frac{2}{d^2} \sum_{i=1}^{d} \sum_{i \neq j} (\widetilde{Y}_{ij} - (1-\alpha)^2).$$

As we will also see later in the simulation study, the usage of the proposed variance estimate **M1** leads to an undersized test. It seems that the covariance estimate overestimates the true covariance. Since the distribution of variance or covariance estimates tends to be asymmetric, the intuition is that there is a positive bias in the estimate which leads to an overestimation. This in turns leads to undersized tests.

M2 Using the coefficient of variation of Z_t given by $c_\nu = \sigma_Z/\mu_Z$ we exploit

$$\sigma_Z^2 = (1-\alpha)^2 c_\nu^2 = (1-\alpha)^2 \frac{\sigma_Z^2}{\mu_Z^2}.$$

Replacing the squared coefficient of variation with the empirical estimator,

$$\hat{c}_\nu^2 = \frac{\frac{1}{n-1} \sum_{t=1}^{n} (Z_t - \bar{Z})^2}{\left(\frac{1}{n} \sum_{t=1}^{n} Z_t\right)^2}$$

we use

$$\widetilde{\sigma}_Z^2 = (1-\alpha)^2 \hat{c}_\nu^2.$$

Note that by using the coefficient of variation, we 'correct' the sample variance estimator by its sample mean in the denominator. So we would expect better estimates compared to using **M1**.

M3 Using $Z_t = d^{-1} \sum_{i=1}^{d} Y_{t,i} = d^{-1} \mathbf{1}' \boldsymbol{Y}_t$, $\mathbf{1} = (1, \dots, 1)'$, we infer

$$\sigma_Z^2 = \frac{1}{d^2} \mathbf{1}' \text{Var}(\boldsymbol{Y}_t) \mathbf{1} = \frac{\sigma_Y^2}{d^2} \mathbf{1}' \mathcal{R}_{\boldsymbol{Y}} \mathbf{1},$$

where $\sigma_Y^2 = \text{var}(Y_{t,i}) = \alpha(1-\alpha)\ \forall t, i$ and $\mathcal{R}_{\boldsymbol{Y}}$ is the correlation matrix of the random vector $\boldsymbol{Y}_t$. Note that under the null hypothesis we have that $\mathcal{R}_{\mathbf{Y}}$ is independent of t. This

suggests the estimator

$$\widetilde{\sigma}_Z^2 = \frac{\alpha(1-\alpha)}{d^2}\mathbf{1}'\widehat{\mathcal{R}}_{\boldsymbol{Y}}\mathbf{1},$$

using the sample correlation matrix $\widehat{\mathcal{R}}_{\boldsymbol{Y}}$ with entries

$$\widehat{\mathrm{cor}}(Y_{t,i}, Y_{t,j}) = \frac{\sum_{t=1}^n \left(Y_{t,i} - \bar{Y}_i\right)\left(Y_{t,j} - \bar{Y}_j\right)}{\sqrt{\sum_{t=1}^n \left(Y_{t,i} - \bar{Y}_i\right)^2}\sqrt{\sum_{t=1}^n \left(Y_{t,j} - \bar{Y}_j\right)^2}} \tag{4.13}$$

with $\bar{Y}_k = 1/n \sum_{t=1}^n Y_{t,k}$. Compared to **M1** and **M2** we do not use a variance or covariance estimate. Since correlation estimates are less prone to positive biases than variance and covariance estimates we deem this a good candidate too.

Note that when estimating $\widehat{\mathcal{R}}_{\boldsymbol{Y}}$, we have to disregard desks with zero violations (i.e. the daily observed loss for this desk did not exceed the respective estimated VaR on any occasion in the considered backtesting period). If we would not exclude these desks, the sample variance of the desk is zero, resulting in an error in the calculation of the sample correlation in (4.13)[27]. The probability that a desk is excluded depends on the considered α level. If we consider a time series of VaR violation indicators of length 250 for one desk, the probability to observe no violations is $(1-\alpha)^{250}$, when we assume independence.

Consistency of the variance estimates All proposed estimates described under **M1**, **M2**, **M3** are consistent estimates for the unknown variance σ_Z^2 of Z_t under the null hypothesis as a consequence of the well known strong law of large numbers; see Theorem 4 in Section 4.2.

Proposition 3. *Let $\mathbf{Y}_t = (Y_{t,1}, \ldots, Y_{t,d})$, $t = 1, \ldots, n$, denote the vector of desk's data at time t. The vectors $\mathbf{Y}_t$ are assumed to be i.i.d. Bernoulli distributed. Then, the variance estimates $\tilde{\sigma}_Z^2$ described in **M1**, **M2**, **M3** are consistent estimates for the unknown variance σ_Z^2 of Z_t.*

Proof. **M1** We will proof almost sure convergence for the expression $\widetilde{Y}_{ij} = n^{-1}\sum_{t=1}^n Y_{t,i}Y_{t,j}$, from which convergence in probability (see Definition 20) follows. Since the vectors $\mathbf{Y}_1, \ldots \mathbf{Y}_n$ are assumed to be i.i.d. Bernoulli, $\{Y_{t,i}Y_{t,j}\}_{t=1}^n$ is an i.i.d. sequence (with some finite mean). Due to the strong law of large numbers we have

$$\widetilde{Y}_{ij} \xrightarrow{a.s.} \mathbb{E}\left(Y_{t,i}Y_{t,j}\right).$$

[27]The same problem arises in case that we observe 100% exceptions in a desk. However, due to the usually chosen α levels in the area close to one, this case is irrelevant.

Note that $\text{var}(Y_{t,i})$ and with that $\mathbb{E}\left(Y_{t,i}Y_{t,j}\right)$ exists $\forall\, t,i,j$, because $Y_{t,i}$ is bounded. Using the continuous mapping theorem (see Theorem 3 in Section 4.2) and Fatou's Lemma it follows that

$$\widetilde{\sigma}_Z^2 = \frac{\alpha(1-\alpha)}{d} + \frac{2}{d^2}\sum_{i=1}^{d}\sum_{i\neq j}(\widetilde{Y}_{ij} - (1-\alpha)^2) \xrightarrow{P} \sigma_Z^2.$$

M2 By the strong law of large numbers we have $n^{-1}\sum_{t=1}^{n} Z_t \xrightarrow{a.s.} \mu_Z$ since the sequence $Z_1,\ldots,Z_n$ is i.i.d. (with some finite mean). It follows by the continuous mapping theorem that

$$\left(1/n\sum_{t=1}^{n} Z_t\right)^2 \xrightarrow{a.s.} \mu_Z^2.$$

For the expression $\frac{1}{n-1}\sum_{t=1}^{n}(Z_t-\bar{Z})^2$ we have $n^{-1}\sum_{t=1}^{n}(Z_t-\bar{Z})^2 = n^{-1}\sum_{t=1}^{n} Z_t^2 - \bar{Z}^2$. Applying the strong law of large numbers, we have

$$\frac{1}{n}\sum_{t=1}^{n} Z_t^2 \xrightarrow{a.s.} \mathbb{E}\left(Z_t^2\right) \quad \text{and} \quad \bar{Z} \xrightarrow{a.s} \mathbb{E}\left(Z_t\right) = \mu_Z.$$

By the continuous mapping theorem it follows that

$$\frac{1}{n}\sum_{t=1}^{n}(Z_t-\bar{Z})^2 \xrightarrow{a.s} \mathbb{E}\left(Z_t^2\right) - \mathbb{E}\left(Z_t\right)^2 = \sigma_Z^2.$$

We directly get

$$\frac{n}{n-1}\frac{1}{n}\sum_{t=1}^{n}(Z_t-\bar{Z})^2 \xrightarrow{a.s.} \sigma_Z^2 \tag{4.14}$$

since $n/(n-1)\to 1$ as $n\to\infty$. Using again the continuous mapping theorem we conclude

$$\widetilde{\sigma}_Z^2 = (1-\alpha)^2\hat{c}_\nu^2 = (1-\alpha)^2\frac{\frac{1}{n-1}\sum_{t=1}^{n}(Z_t-\bar{Z})^2}{\left(\frac{1}{n}\sum_{t=1}^{n} Z_t\right)^2} \xrightarrow{a.s.} \sigma_Z^2.$$

Note that $\left(\frac{1}{n}\sum_{t=1}^{n} Z_t\right)^2 \neq 0$ because the sample mean of Z_t converges to the expected value μ_Z which is $1-\alpha$. Almost sure convergence implies convergence in probability.

M3 To show the consistency of $\widehat{\mathcal{R}}_{\boldsymbol{Y}}$, we rewrite the entries of $\widehat{\mathcal{R}}_{\boldsymbol{Y}}$ as

$$\frac{\frac{1}{n}\sum_{t=1}^{n}\left(Y_{t,i}-\bar{Y}_i\right)\left(Y_{t,j}-\bar{Y}_j\right)}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_{t,i}-\bar{Y}_i\right)^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_{t,j}-\bar{Y}_j\right)^2}}. \tag{4.15}$$

Following the same arguments as in (4.14) we know that the denominator in (4.15) converges almost surely. We rewrite the numerator in (4.15) as

$$\frac{1}{n}\sum_{t=1}^{n}\left(Y_{t,i}Y_{t,j} - Y_{t,j}\bar{Y}_i - Y_{t,i}\bar{Y}_j + \bar{Y}_i\bar{Y}_j\right). \tag{4.16}$$

- The expression $\frac{1}{n}\sum_{t=1}^{n} Y_{t,i}Y_{t,j}$ is the sample mean of $Y_{t,i}Y_{t,j}$, $\forall i, j$. We have already seen for estimate **M1** that this expression converges to $\mathbb{E}\left(Y_{t,i}Y_{t,j}\right)$ almost surely.
- The term $V_t = \frac{1}{n}\sum_{t=1}^{n} Y_{t,j}\bar{Y}_i = \frac{1}{n}\sum_{t=1}^{n} Y_{t,j}\frac{1}{n}\sum_{t=1}^{n} Y_{t,i}$, converges to $\mathbb{E}\left(Y_{t,j}\right)\mathbb{E}\left(Y_{t,i}\right)$ almost surely for every $i, j, i \neq j$ due to the law of large numbers and the continuous mapping theorem.

Summarizing, we have for the numerator in (4.15) that

$$\frac{1}{n}\sum_{t=1}^{n}\left(Y_{t,i} - \bar{Y}_i\right)\left(Y_{t,j} - \bar{Y}_j\right) \overset{a.s.}{\to} \mathbb{E}\left(Y_{t,i}Y_{t,j}\right) - \mathbb{E}\left(Y_{t,j}\right)\mathbb{E}\left(Y_{t,i}\right).$$

Therefore we have $\widehat{\text{cor}}(Y_{t,i}, Y_{t,j}) \overset{a.s.}{\to} \text{cor}(Y_{t,i}, Y_{t,j})$ for all entries of $\widehat{\mathcal{R}}_{\mathbf{Y}}$ and with that $\widehat{\mathcal{R}}_{\mathbf{Y}} \overset{a.s.}{\to} \mathcal{R}_{\mathbf{Y}}$. From this it follows that

$$\widetilde{\sigma}_Z^2 = \frac{\alpha(1-\alpha)}{d^2}\mathbf{1}'\widehat{\mathcal{R}}_{\boldsymbol{Y}}\mathbf{1} \overset{P}{\to} \sigma_Z^2.$$

□

4.4.3 Results simulation study: multi-desk VaR violation tests

We analyse the ability of the different variance estimation methods described above to control the unknown correlation cross desks in a simulation study. For this we use the multi-desk binomial score test presented in (4.11). We assume the the risk modelling group either correctly specifies or misspecifies all desk models (variable fraction is frozen at 100). The remaining setting is as stated in Table 4.3 in Section 4.3.3. Table 4.5 shows the results. The rows called 'Normal' show the empirical size of the test. The rows called 't4' the power. The following is observed

- Ignoring dependencies across desks by using method **None**, that is $\hat{\sigma}_Z^2 = d^{-1}\alpha(1-\alpha)$, gives strongly oversized tests. The only exception is when simulating data from a Gauss copula with correlation matrix equal to the identity matrix (correlation structure 'Zero'), which corresponds to independent desks.

- Proposal **M1** leads to undersized tests. No clear structural pattern appears when moving from 50 to 100 desks or from sample size 250 to 500.
- Using alternative **M2** (coefficient of variation) gives slightly oversized tests and shows good power properties.
- Method **M3** shows better size properties than **M2** and a comparable level of power.
- In general, the differences between $d = 50$ desks and $d = 100$ desks is rather small and no clear improvement in size or power is recognised. Choosing a larger sample size tends to improve power when comparing tests with comparable level of empirical size.

		d	50				100			
		ρ	Zero		Flat		Zero		Flat	
cor	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
None	250	Normal	3.9	20.4	24.2	27.8	4.5	27.2	29.5	31.7
		t4	99.9	93.2	90.1	78.4	100.0	97.1	92.9	83.0
	500	Normal	4.5	22.0	26.0	32.9	6.1	28.2	31.6	36.2
		t4	100.0	99.6	97.8	92.1	100.0	99.8	98.8	93.9
M1	250	Normal	3.1	1.9	1.9	1.2	3.4	1.5	2.1	1.3
		t4	99.9	67.2	51.5	23.6	100.0	69.8	52.4	24.1
	500	Normal	3.9	2.6	2.6	1.9	4.4	2.4	1.9	1.6
		t4	100.0	94.2	81.9	51.5	100.0	95.5	84.6	52.3
M2	250	Normal	3.9	6.0	7.0	7.8	4.4	6.3	6.6	7.4
		t4	99.9	83.7	71.7	49.6	100.0	83.8	73.1	50.2
	500	Normal	4.5	5.8	6.8	7.2	6.1	5.5	6.7	7.0
		t4	100.0	97.0	91.1	71.7	100.0	98.1	92.3	72.4
M3	250	Normal	3.9	4.1	4.1	4.0	4.4	4.5	4.0	4.2
		t4	99.9	79.8	65.3	39.5	100.0	80.6	67.5	41.5
	500	Normal	4.2	4.3	4.8	4.6	6.0	4.5	4.7	4.5
		t4	100.0	96.2	87.9	64.5	100.0	97.7	89.5	66.1

Table 4.5: Estimated size and power of binomial score test using $\alpha = 0.99$ with different variance estimation methods

4.5 Summary

The main conclusions of this chapter are:

- We considered the well-known binomial score test and applied the Bonferroni method to extend this test to a multi-desk VaR violation test, based on VaR violation indicators as input variable. The Bonferroni method tests the adequateness of several desk models (several null hypotheses) at once. We showed both theoretically and in a simulation study that the Bonferroni method is not able to control the size of the test and leads to heavily oversized tests.

- We proposed a new multi-desk VaR backtest, based on VaR violation indicators as input variables, which consider all desks simultaneously in the test function. In particular, we considered a Z-test where the average of the desk's daily exceptions rates is standardized, so that the test is asymptotically standard normally distributed due to the Central Limit Theorem.

- We suggested and analysed three different variance estimation methods to estimate the unknown variance in the test function. In a simulation study it turned out that the method, where only the correlation part of the covariance is estimated (and the known analytical variance under the null hypothesis is used), is best to control for the unknown correlation between desks. In this case, the test is well-sized with good power properties (mostly well above 70%).

- The simulation study also revealed that an increase of the number of desks from 50 to 100 did not improve the size and power performance of the proposed multi-desk VaR violation test. The performance improved when the number of observations was increased from 250 (roughly one year of observations) to 500 (roughly two years of observations).

Chapter 5

Multi-desk backtesting using PIT-values

5.1 Introduction

We have argued in Chapter 4 that multi-desk backtesting is beneficial and that controlling for the unknown dependency across desks is crucial to develop well-sized tests. While we developed tests based on indicator time series for VaR violations in Chapter 4, we will generalise this theory to multi-desk tests based on realised probability integral transform values (PIT-values). The new tests also answer research question 2 (see Section 4.1).

In Gordy and McNeil (2020), mono-desk spectral backtests for the unconditional (see Definition 15) and conditional coverage (see Definition 17) hypothesis are proposed. Their spectral tests make use of the realised PIT-values of the estimated PnL distribution, which serve as the input variable. The proposed class of spectral backtests includes many well-known tests as special cases including the binomial likelihood ratio test of Kupiec (1995), which simply counts the number of VaR violations in the observed sequence of VaR violations. More precisely, the proposed tests are built upon transformations of the realised PIT-value level exceedance indicator function, which are weighted. By the choice of the weighting function (or kernel function) used to weight the realised PIT-value level exceedances, the risk modelling group can emphasize the region of the PnL distribution about which it is most concerned with respect to model performance. Moreover, the benefit of using realised PIT-values - as already exploited in the USA - becomes apparent since the new tests turn out to be quite powerful.

The new Basel framework also indicates that the regulator is not only interested in the perfor-

mance of the VaR model at the confidence level used to calculate the required capital (99%) but also in the performance at the 97.5% level (see in BCBS (2019b), p.83, para. 32.13). The proposed tests in Gordy and McNeil (2020) permit the risk modelling group to choose a region of the estimated PnL distribution on which the assessment should be focused (e.g. an area around the 99% quantile or just the 99% quantile).

Generally, tests based on realised PIT-values level exceedance indicators are expected to be more powerful than tests based on the VaR violation indicator sequence alone, since they include information about any α level and can be used to assess a wider range of the PnL distribution. However, one drawback is that information about the whole PnL distribution function F is required and not only information on one specific quantile $F^{\leftarrow}(\alpha)$. In practice, more emphasis is given to the modelling of the tail of the PnL distribution function, where the extreme losses occur. It is expected that for moderate α levels, the estimated distribution is likely to be less adequate. Consequently, for the purposes of backtesting one has to find a range of α levels which give a good balance between powerful tests and precision.

In this chapter we draw on the findings in Gordy and McNeil (2020) and develop various tests based on realised PIT-value level exceedance indicators which extend the theory of the proposed mono-desk tests[28] to multi-desk tests, which test several trading desk models simultaneously. Since the realised PIT-value exceedances are likely to be correlated across desks, we apply the proposed variance estimates from Chapter 4 to account for correlation.

The chapter is organized as follows. We first explain the rational behind PIT-values and state the hypothesis linked to the research question. After that we present multi-desk monospectral tests and multi-desk bispectral tests, all being generalisations of the tests proposed in Gordy and McNeil (2020) to the multi-desk framework. The chapter concludes with an extensive simulation study, which follows the principles set out in Section 4.3.3.

5.2 Methodology

5.2.1 Data

The general setup is the same as in Section 4.3.1. The number of desks approved for the use of an internal model is d. The variable $L_{t,i}$ describes the profit and loss of the risky assets attributed to desk i at time t. It is defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{N}_0}, \mathbb{P})$,

[28] In our interpretation, the univariate case corresponds to a mono-desk test or a test at bank-wide portfolio level.

where $\mathcal{F}_t$ represents the information available at time t and $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. $F_{t,i}$ denotes the conditional profit and loss (PnL) distribution function given the information up to time $t-1$

$$F_{t,i}(x) = \mathbb{P}\left(L_{t,i} \leq x | \mathcal{F}_{t-1}\right).$$

We assume throughout that the PnL distribution functions are continuous. As before let $\widehat{\text{VaR}}_{t,i}(\alpha)$ represent the one-day ahead forecast of the VaR at level α of desk i generated at time $t-1$.

The multi-desk tests presented in this chapter are based on realised probability integral transform or PIT-values (sometimes also known as p-values) of the conditional PnL distribution function $F_{t,i}$. Formally, a theoretical PIT-value is defined as

$$U_{t,i} = F_{t,i}\left(L_{t,i}\right),$$

where the probability integral transformation is applied which transforms a random variable with some continuous distribution function into a random variable with standard uniform distribution. More precisely, Rosenblatt (1952) showed that, under the assumption that the conditional distribution function is continuous, the sequence $(U_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d. standard uniform distributed random variables.

The Rosenblatt transformation maps a n-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_n)'$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with some distribution function $F(x_1, \dots, x_n)$, to a uniform distribution on the n-dimensional hypercube. The transformation is defined by $\boldsymbol{z} = (z_1, \dots, z_n)' = T(x_1, \dots, x_n)'$ with

$$\begin{aligned} z_1 &= P\left(X_1 \leq x_1\right) = F_1(x_1) \\ z_2 &= P\left(X_2 \leq x_2 | X_1 = x_1\right) = F_2(x_2|x_1) \\ &\vdots \\ z_n &= P\left(X_n \leq x_n | X_{n-1} = x_{n-1}, \dots, X_1 = x_1\right) = F_n(x_n|x_{n-1}, \dots, x_1). \end{aligned} \tag{5.1}$$

Rosenblatt showed that for $\boldsymbol{Z} = T(\boldsymbol{X})$ with $\boldsymbol{Z} = (Z_1, \dots, Z_n)'$ the variables $Z_1, \dots, Z_n$ are uniformly, independently distributed. Applying this transformation, we have that the sequence $(U_{t,i})_{t\in\mathbb{N}}$ with $U_{t,i} = F_{t,i}(L_{t,i})$ is a sequence of i.i.d. standard uniform variables. Note that this transformation holds true even when the process $(L_{t,i})_{t\in\mathbb{N}}$ is non-stationary.

Under the assumption that the bank constructs an ideal probabilistic forecast $\widehat{F}_t$ of F_t at

any point in time, we would expect that the sequence of realised PIT-values $(P_{t,i})_{t\in\mathbb{N}}$ with $P_{t,i} = \widehat{F}_{t,i}(L_{t,i})$ forms a series of i.i.d. standard uniform variables. Note that we interpret an ideal forecast $\widehat{F}_t$ of F_t at any point in time in the sense of Gneiting et al. (2007), where the risk modelling group always knows the truth.

One advantage of PIT-values is that they give information about the entire forecast distribution since for any $\alpha \in [0,1]$

$$U_{t,i} \geq \alpha \Longrightarrow L_{t,i} \geq F^{\leftarrow}_{t,i}(\alpha) \Longleftrightarrow L_{t,i} \geq VaR_{t,i}(\alpha). \tag{5.2}$$

The distribution function has to be increasing to apply the first inference and the equality holds for continuous distribution functions $F_{t,i}$. For discontinuous functions small issues exists for certain values. That is the realised PIT-value level exceedance contains information about a VaR violation at any level α, whereas the VaR violation indicator just provides information if there was a violation at the pre-specified level α of the VaR. Due to (5.2), the (realised) PIT-value level exceedance indicator function $I_{\{P_{t,i}\geq\alpha\}}$ is equivalent to the VaR violation indicator function $I_{\{L_{t,i}\geq\widehat{\mathrm{VaR}}_{t,i}(\alpha)\}}$.

As in Gordy and McNeil (2020), the tests we consider are based on transformations of indicator variables for realised PIT-value level exceedances. We set

$$W_{t,i} = \int_{[0,1]} I_{\{P_{t,i}\geq u\}} \mathrm{d}\nu(u), \tag{5.3}$$

where ν is a Lebesgue-Stieltjes measure defined on $[0,1]$, which is designed to apply different weights to different quantile levels. Following Gordy and McNeil (2020) we will refer to ν as the *kernel measure* for the transformed indicator variables for realised PIT-value level exceedances. By choosing ν, the risk modelling group has the flexibility to choose the area of the PnL distribution function about which it is most concerned with respect to model performance. The risk modelling group can also assign more weight to specific quantiles or quantile areas within its specified region of interest. We will refer to (5.3) as *W-values*. Let $\boldsymbol{P}_t = (P_{t,1}, \ldots, P_{t,d})$ denote the vector of PIT-values containing the PIT-values of each desk at point in time t. We structure the desk data into the columns, as shown in Table 5.1.

t \ i	desk 1	desk 2	desk 3	desk 4	desk 5	...	desk d
day 1	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$	$P_{1,4}$	$P_{1,5}$	...	$P_{1,d}$
day 2	$P_{2,1}$	$P_{2,2}$	$P_{2,3}$	$P_{2,4}$	$P_{2,5}$	...	$P_{2,d}$
day 3	$P_{3,1}$	$P_{3,2}$	$P_{3,3}$	$P_{3,4}$	$P_{3,5}$	...	$P_{3,d}$
day 4	$P_{4,1}$	$P_{4,2}$	$P_{4,3}$	$P_{4,4}$	$P_{4,5}$	...	$P_{4,d}$
day 5	$P_{5,1}$	$P_{5,2}$	$P_{5,3}$	$P_{5,4}$	$P_{5,5}$	...	$P_{5,d}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
day n	$P_{n,1}$	$P_{n,2}$	$P_{n,3}$	$P_{n,4}$	$P_{n,5}$	...	$P_{n,d}$

Table 5.1: Stylized sample of the data structure of the PIT-values

For example, the risk modelling group could choose a continuous weighting function, placing weights in an area $[\alpha_1, \alpha_2] \subset [0,1]$ according to some continuous function $g : [0,1] \to [0,\infty)$. Then the W-values are

$$W_{t,i} = \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u,\ i = 1, \ldots, d \tag{5.4}$$

(more details on this can be found in Section 5.3). For example, if we choose a uniform weighting function $g(u) = 1$, $W_{t,i}$ is given by the function $P_{t,i}$ shown in Figure 5.1. The maximum value of $W_{t,i}$ is reached at $\alpha_2 - \alpha_1$. We used $\alpha_1 = 0.75$ and $\alpha_2 = 0.90$ in our example. In the region between α_1 and α_2, equal weights are placed on realised PIT-value level exceedances.

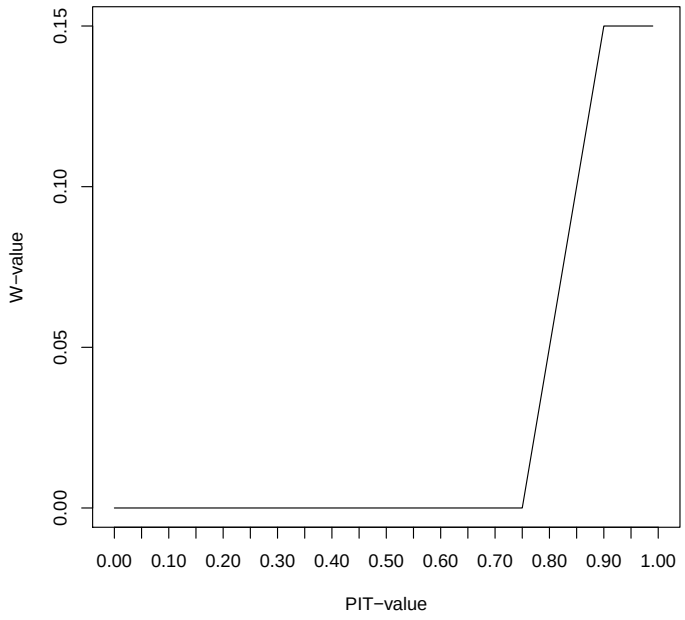

Figure 5.1: W-values for a uniform weighting function

5.2.2 Hypotheses

Research question 2 concerns the overall quality of the internal risk models across desks. So we like to assess the overall quality of the models used by banks to calculate the market risk they are exposed to in their different trading desks.

We know from Section 5.2.1 that under the assumption that the bank builds an ideal forecast $\widehat{F}_t$ of F_t at any point in time, that the sequence of realised PIT-values $(P_{t,i})_{t\in\mathbb{N}}$ with $P_{t,i} = \widehat{F}_{t,i}(L_{t,i})$ forms a series of i.i.d. standard uniform variables. Let $F^0_{\mathbf{W}}$ be the unknown distribution function of the d-dimensional vector of W-values $\boldsymbol{W}_t = (W_{t,1}, \ldots, W_{t,d})$, when the realised PIT-values $P_{t,1}, \ldots, P_{t,d}$ are indeed uniform for any t, $t = 1, \ldots, n$. Then we can formulate the null hypothesis

$$H_0 : \mathbf{W}_1, ..., \mathbf{W}_n \text{ are i.i.d. with the same distribution function } F^0_{\mathbf{W}}. \tag{5.5}$$

Note that the null hypothesis implicitly assumes that two desks are independent for two different point in times, like $W_{t,1}$ to $W_{s,2}$ for $t \neq s$. Note also that we explicitly allow for correlations across desks on the same trading day t under the null hypothesis in our convention. This seems to be a reasonable assumption since desk models are usually built using the same underlying

methodology and the same correlated risk factor data. The challenge is to develop a test that overcomes the lack of information about the correlation between the trading desks data since correlation cannot be observed but has to be estimated.

We understand our tests as tests of the *unconditional coverage hypothesis* (see Definition 15), meaning that the probability of observing a realised PIT-value larger than α must be equal to $1-\alpha$. Although the tests we propose in this chapter are derived under the independence (across time) assumption under the null hypothesis, they are not specifically designed as a test for the independence hypothesis. The proposed tests will only have limited power to detect deviations from the independence hypothesis, as noted in Gordy and McNeil (2020). So the focus of our tests remains on the unconditional coverage hypothesis.

In the case of a bispectral test, we face a two-dimensional vector of W-values using two different kernel measures ν^j, $j = 1, 2$:

$$W_{t,i}^j = \int_{[0,1]} I_{\{P_{t,i} \geq u\}} \mathrm{d}\nu^j(u),\ j = 1, 2$$

such that $\boldsymbol{W}_t^j = (W_{t,1}^j, \ldots, W_{t,d}^j)$ (see Section 5.3.2 for more details on the bispectral test). In this case, the hypothesis reads as

$$H_0 : \widetilde{\mathbf{W}}_\mathbf{1}, \ldots, \widetilde{\mathbf{W}}_\mathbf{n} \text{ are i.i.d. with distribution function } F_\mathbf{W}^0, \tag{5.6}$$

where $F_\mathbf{W}^0$ is the marginal distribution function of the two-dimensional vector $\widetilde{\mathbf{W}}_\mathbf{t} = (\mathbf{W}_t^1, \mathbf{W}_t^2)$, when $P_{t,1}, \ldots, P_{t,d}$ are i.i.d. uniformly distributed.

5.3 Multi-desk tests based on PIT-values

We build on the findings in Gordy and McNeil (2020) and extend their theory of mono-desk spectral tests to multi-desk spectral tests.

From expression (5.3), the closed form expression for the W-values is easily derived:

$$W_{t,i} = \int_{[0,1]} I_{\{P_{t,i} \geq u\}} \mathrm{d}\nu(u) = \int_{[0,P_{t,i}]} \mathrm{d}\nu(u) = \nu\left([0, P_{t,i}]\right). \tag{5.7}$$

Following Gordy and McNeil (2020), we will refer to ν as the *kernel measure* for the transformed indicator variables for realised PIT-value level exceedances. From (5.7) it becomes clear that $W_{t,i}$ is an increasing function in $P_{t,i}$. Recall that $\boldsymbol{W_t} = (W_{t,1}, \ldots W_{t,d})$ denotes the vector of the

W-values of the d desks at time t.

In Gordy and McNeil (2020) both discrete and continuous weighting functions are considered for the weighting scheme implied by the measure ν. For simplicity of exposition, we concentrate on the continuous case in this thesis. For continuous weighting, we can choose a continuous function defining a weighting structure for a subset of $[0,1]$. For this, the measure ν has the form $\mathrm{d}\nu = g(u)\mathrm{d}u$ for some continuous function g defined on $[0,1]$. The measure ν can be thought of as a probability measure with probability density g. However, it is not needed that $\nu([0,1]) = 1$.

Assumption 3. *For the weighting function g it holds that $g(u) = 0$ for $u \notin [\alpha_1, \alpha_2]$, g is continuous and positive on (α_1, α_2).*

With this assumption it is guaranteed that $\nu([\alpha_1, \alpha_2]) = \int_{\alpha_1}^{\alpha_2} g(u)\mathrm{d}u$ is positive, that $[\alpha_1, \alpha_2]$ is the kernel window and that we exclude non-trivial tests, where $\nu([\alpha_1, \alpha_2]) = 0$.

For a continuous measure, the expression for the W-values in (5.7) then becomes

$$W_{t,i} = \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u.$$

Let G be the integral of the function g

$$G(u) = \int_{-\infty}^{u} g(t)dt = \int_{\alpha_1}^{u} g(t)dt,$$

where the latter is due to Assumption 3. Then, $W_{t,i}$ can be written as

$$\begin{aligned} W_{t,i} &= \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u = I_{\{P_{t,i} \geq \alpha_1\}} \int_{\alpha_1}^{\min(P_{t,i}, \alpha_2)} g(u)\mathrm{d}u \\ &= I_{\{P_{t,i} \geq \alpha_1\}} \left(G(P_{t,i} \wedge \alpha_2) - G(\alpha_1)\right) = G(\alpha_1 \vee (P_{t,i} \wedge \alpha_2)) - G(\alpha_1) \\ &= G(P^*_{t,i}) - G(\alpha_1). \end{aligned} \tag{5.8}$$

with $P_{t,i} \wedge \alpha = \min\{P_{t,i}, \alpha\}$ and $P_{t,i} \vee \alpha = \max\{P_{t,i}, \alpha\}$. From (5.8) it is clear that $W_{t,i}$ truncates the realised PIT-values $P_{t,i}$ on $[\alpha_1, \alpha_2]$ and that $W_{t,i}$ is always a strictly increasing and continuous function of the *truncated realised PIT-values* $P^*_{t,i} := \alpha_1 \vee (P_{t,i} \wedge \alpha_2)$ for g (see

Assumption 3). The truncated PIT-values satisfy

$$P^*_{t,i} = \begin{cases} \alpha_2 & \text{for } P_{t,i} \geq \alpha_2 \\ P_{t,i} & \text{for } P_{t,i} \in [\alpha_1, \alpha_2] \\ \alpha_1 & \text{for } P_{t,i} < \alpha_1. \end{cases} \tag{5.9}$$

So for $W_{t,i}$ we obtain

$$W_{t,i} = \begin{cases} G(\alpha_2) - G(\alpha_1) & \text{for } P_{t,i} \geq \alpha_2 \\ G(P^*_{t,i}) - G(\alpha_1) & \text{for } P_{t,i} \in [\alpha_1, \alpha_2] \\ 0 & \text{for } P_{t,i} < \alpha_1. \end{cases} \tag{5.10}$$

This means that observed PIT-values smaller than α_1 count as a zero observation. Realised PIT-values in $[\alpha_1, \alpha_2]$ are violations which are weighted according to the chosen weighting function g. For realised PIT-values larger than α_2, the weighting is fixed at $G(\alpha_2) - G(\alpha_1)$. In our simulation study, we consider different functions g.

The following two expressions can be easily calculated analytically. The expected value of $W_{t,i}$ is

$$\mu_W = \mathbb{E}(W_{t,i}) = \mathbb{E}\left(\int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u\right) = \int_{\alpha_1}^{\alpha_2} g(u) \mathbb{E}\left(I_{\{P_{t,i} \geq u\}}\right) \mathrm{d}u = \int_{\alpha_1}^{\alpha_2} g(u)(1-u)\mathrm{d}u.$$

We can further derive

$$\begin{aligned} \mu_{2,W} &= \mathbb{E}(W^2_{t,i}) = \mathbb{E}\left(\left(\int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u\right)^2\right) \\ &= \mathbb{E}\left(\left(\int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u\right)\left(\int_{\alpha_1}^{\alpha_2} g(v) I_{\{P_{t,i} \geq v\}} \mathrm{d}v\right)\right) \\ &= \int_{\alpha_1}^{\alpha_2} 2g(u)(G(u) - G(\alpha_1))(1-u)\mathrm{d}u. \end{aligned} \tag{5.11}$$

To see the final step of (5.11) we calculate

$$\begin{aligned}
W_{t,i}W_{t,i} &= \left(\int_{\alpha_1}^{\alpha_2} g(u)I_{\{P_{t,i}\geq u\}}\mathrm{d}u\right)\left(\int_{\alpha_1}^{\alpha_2} g(v)I_{\{P_{t,i}\geq v\}}\mathrm{d}v\right) \\
&= \int_{\alpha_1}^{\alpha_2}\int_{\alpha_1}^{\alpha_2} g(u)g(v)I_{\{P_{t,i}\geq u\}}I_{\{P_{t,i}\geq v\}}\mathrm{d}u\mathrm{d}v \\
&= \int_{\alpha_1}^{\alpha_2}\int_{v=\alpha_1}^{u} g(u)g(v)I_{\{P_{t,i}\geq u\}}\mathrm{d}v\mathrm{d}u + \int_{\alpha_1}^{\alpha_2}\int_{v=u}^{\alpha_2} g(u)g(v)I_{\{P_{t,i}\geq v\}}\mathrm{d}v\mathrm{d}u \\
&= \int_{\alpha_1}^{\alpha_2} g(u)\left(\int_{v=\alpha_1}^{u} g(v)dv\right)I_{\{P_{t,i}\geq u\}}\mathrm{d}u + \int_{v=\alpha_1}^{\alpha_2} g(v)\left(\int_{u=\alpha_1}^{v} g(u)\mathrm{d}u\right)I_{\{P_{t,i}\geq v\}}\mathrm{d}v \\
&= \int_{u=\alpha_1}^{\alpha_2} g(u)\left(G(u)-G(\alpha_1)\right)I_{\{P_{t,i}\geq u\}}\mathrm{d}u + \int_{v=\alpha_1}^{\alpha_2} g(v)\left(G(v)-G(\alpha_1)\right)I_{\{P_{t,i}\geq v\}}\mathrm{d}v \\
&= \int_{\alpha_1}^{\alpha_2} g(u)\left(G(u)-G(\alpha_1)\right)I_{\{P_{t,i}\geq u\}}\mathrm{d}u + \int_{\alpha_1}^{\alpha_2} g(v)\left(G(v)-G(\alpha_1)\right)I_{\{P_{t,i}\geq v\}}\mathrm{d}v \\
&= 2\int_{\alpha_1}^{\alpha_2} g(u)\left(G(u)-G(\alpha_1)\right)I_{\{P_{t,i}\geq u\}}\mathrm{d}u.
\end{aligned}$$

For the variance we have $\sigma_W^2 = \mu_{2,W} - \mu_W^2$. Note that the variance must be finite in order to apply the Central Limit Theorem underpinning those test, which take the form of a Z-test statistic. Assumption 3 satisfies the existence of moments. Conditions for the finite expectation of the product $W_{t,i}W_{t,i}$ are given in Gordy and McNeil (2020), Proposition 3.2. All the weighing functions considered in this thesis satisfy the condition.

5.3.1 Multi-desk monospectral tests

Our spectral test takes the form of a Z-test statistic which distribution function can be approximated by the normal distribution under the null hypothesis. We test for the hypothesis stated in (5.5) in Section 5.2.2. The test is based on the W-values $W_{t,i}$ presented above, using a continuous kernel measure. The desk exception rate is given by

$$Z_t = \frac{1}{d}\sum_{i=1}^{d} W_{t,i} = \frac{1}{d}\sum_{i=1}^{d}\int_{\alpha_1}^{\alpha_2} g(u)I_{\{P_{t,i}\geq u\}}\mathrm{d}u.$$

The proposed Z-test is then

$$T_{Mspec} = \sqrt{n}\,\frac{\bar{Z}-\mu_W}{\sigma_Z}$$

with average desk exception rate $\bar{Z} = n^{-1}\sum_{t=1}^{n} Z_t$, variance $\sigma_Z^2 = \mathrm{var}(Z_t)$ and

$$\mu_W = \mathbb{E}(\bar{Z}) = n^{-1}\sum_{t=1}^{n}\mathbb{E}(Z_t) = \mathbb{E}\left(W_{t,i}\right) = \int_{\alpha_1}^{\alpha_2} g(u)(1-u)\mathrm{d}u.$$

Under the null hypothesis and with a consistent estimator of the variance σ_Z^2, we can use the Central Limit Theorem to base a test on the normality of the test statistic:

$$T_{Mspec} = \sqrt{n}\,\frac{\bar{Z} - \mu_W}{\sigma_Z} \dot{\sim} \mathcal{N}(0,1)\,. \tag{5.12}$$

Since σ_Z is unknown, it has to be estimated. We will apply different alternative estimates derived in the following section. In general, these are the same estimators presented in Section 4.4.2 adapted to the setting of our monospectral tests.

Estimation of the variance We use the different variance estimation approaches presented in Section 4.4.2. Note that in our simulation study, we will again truncate the variance estimate to the minimum variance which holds true under the assumption of independent desks:

$$\hat{\sigma}_Z^2 = \max(\sigma_{\min}^2, \tilde{\sigma}_Z^2).$$

The variable $\sigma_{\min}^2$ is the variance of Z_t when we ignore any dependencies between desks, i.e. $\sigma_{\min}^2 = \mathrm{var}(Z_t) = d^{-1}(\mu_{2,W} - \mu_W^2)$, and $\tilde{\sigma}_Z^2$ is one of the variance estimators described below.

None We set $\widetilde{\sigma}_Z^2 = 0$.

M1 We set

$$\widetilde{\sigma}_Z^2 = \frac{1}{d}(\mu_{2,W} - \mu_W^2) + \frac{2}{d^2}\sum_{i=1}^{d}\sum_{i\neq j}(\tilde{\mu}_{ij} - \mu_W^2).$$

M2 We set

$$\widetilde{\sigma}_Z^2 = \mu_W^2 \hat{c}_\nu^2 \quad \text{with} \quad \hat{c}_\nu^2 = \frac{\frac{1}{n-1}\sum_{t=1}^{n}(Z_t - \bar{Z})^2}{\left(\frac{1}{n}\sum_{t=1}^{n} Z_t\right)^2}.$$

M3 We set

$$\widetilde{\sigma}_Z^2 = \frac{\mu_{2,W} - \mu_W^2}{d^2}\mathbf{1}'\widehat{\mathcal{R}}_{\boldsymbol{W}}\mathbf{1},$$

where $\widehat{\mathcal{R}}_{\boldsymbol{W}}$ is the estimated sample correlation matrix of the vector $\boldsymbol{W}_t$. Note that when estimating $\widehat{\mathcal{R}}_{\boldsymbol{W}}$, we have to disregard desks with zero violations, which means that the realised PIT-value $P_{t,i}$ of this desk did not lie once within the kernel window (i.e. the realised PIT-value did not exceed the threshold α_1) in the considered backtesting period. If we do not exclude those desks, the sample variance of the desk will be zero in those cases,

resulting in an error in the calculation of the sample correlation. In our simulation studies, we consider regions in the extreme right tail (for example the region $[0.9805, 0.9995]$). Considering for example $n = 250$ days, this amounts to a probability of 4.88×10^{-431} that one desk has to be excluded (no violations observed) in the simulations under the assumption of independence.

All three estimates described in **M1**, **M2** and **M3** are consistent estimates of σ_Z^2 under the null hypothesis. To show this one can follow the arguments as in the proof of Proposition 3 in Section 4.4.2 and use the fact that $W_{t,i}$ is bounded for all $t, i, t = 1, \ldots, n$, $i = 1, \ldots, d$.

5.3.2 Multi-desk bispectral tests

In the monospectral Z-test, we considered one continuous weighting function g. That is we had a d-dimensional vector $\boldsymbol{W}_t = (W_{t,1}, \ldots, W_{t,d})$ for every point in time t:

$$W_{t,i} = \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u.$$

This setting can be generalised to a multispectral test, where more than one weighting function g is considered, that is we consider a set of W-values. In this thesis we concentrate on the multi-desk bispectral test, i.e. we consider the set $(W_{t,i}^1, W_{t,i}^2)$ where

$$W_{t,i}^1 = \int_{\alpha_1}^{\alpha_2} g^1(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u,$$
$$W_{t,i}^2 = \int_{\alpha_1}^{\alpha_2} g^2(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u,$$

where g^1 and g^2 are functions fulfilling Assumption 3. In this way, the risk modelling group can choose two weighting schemes for the PIT-value level exceedances for each desk, which can give a better idea about the quality of the tail of the model. Note that in this setting each of the d desks receives the same two weighting functions. The advantages of using multispectral tests have also been highlighted in Gordy and McNeil (2020). Whereas the monospectral Z-test tests for the first moment of $W_{t,i} = G(P_{t,i}^*) - G(\alpha_1)$, the bispectral Z-test effectively tests for two first moments which differ in the chosen weighting scheme. This can be beneficial if one chosen weighting scheme fails to efficiently detect deviations of the realised PIT-values distribution function from the PIT-values distribution function expected under the null hypothesis (uniformity). It is generally expected that the power of the spectral test is higher the larger the discrepancy between these two distribution functions is within the considered kernel window. The spectral Z-tests measure the average distance of these two distribution functions within the kernel win-

dow, weighting the deviation according to the chosen weighting scheme. Two weighting schemes are beneficial in the case where the distribution function of realised PIT-values crosses the true distribution function of PIT-values within the kernel window. This becomes apparent in Figure 5.2, which shows the distribution function for realised PIT-values when the true model $F_{t,i}$ is either standard normal, scaled t3 or scaled t5 and the risk modelling group assumes a normal distribution. That is we plot $\mathbb{P}(P_{t,i} \leq u) = \mathbb{P}(\Phi(L_{t,i}) \leq u) = \mathbb{P}(L_{t,i} \leq \Phi^{-1}(u)) = F_{t,i}(\Phi^{-1}(u))$. When the true distribution function is normal, that is when the null hypothesis is true, the cumulative distribution function is the identity line. A crossing of the identity line in the chosen kernel window decreases the average distance to the identity line. This is challenging for the monospectral test, which tests for the first moment of $W_{t,i}$.

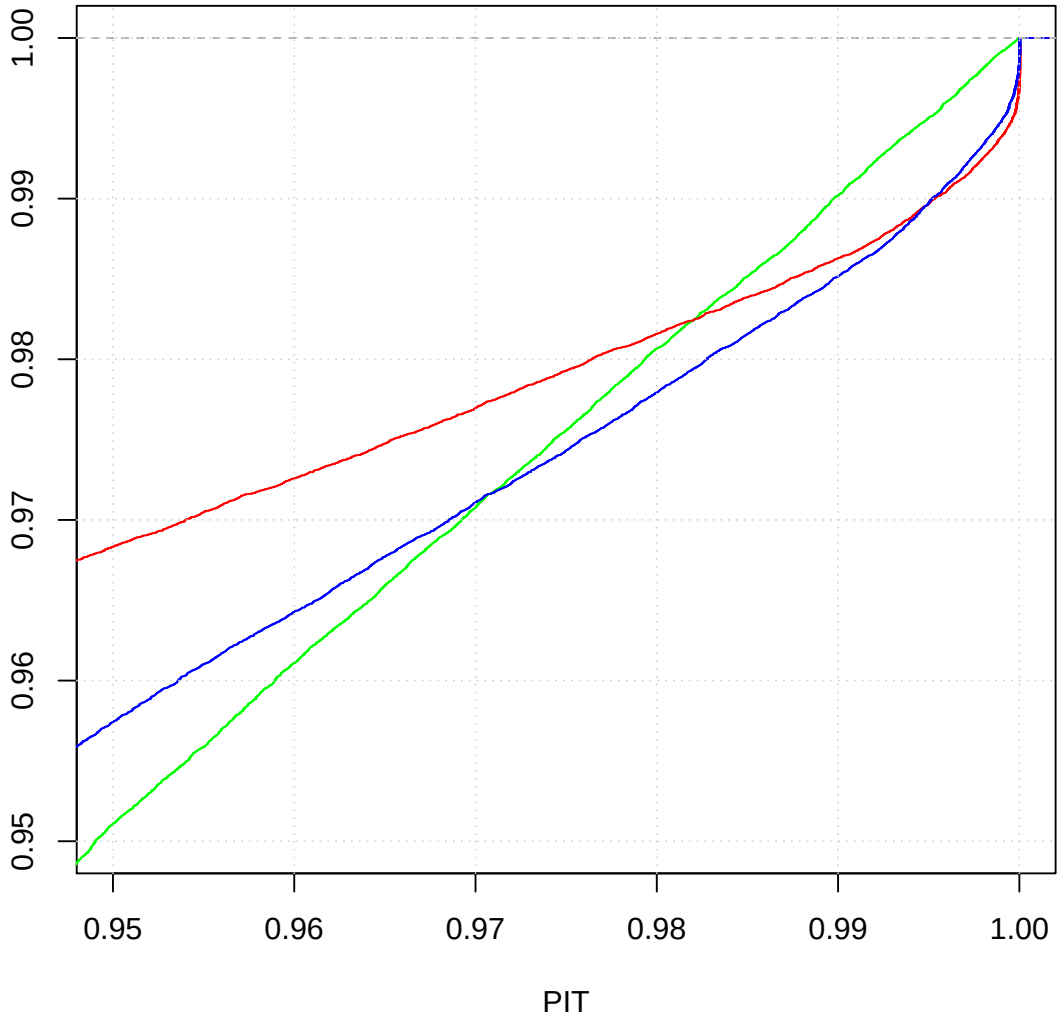

Figure 5.2: Cumulative distribution function of realised PIT-values when true loss model is standard normal (green line), scaled t3 (red line) or scaled t5 (blue line)

For the bispectral test, we define two desk exception rates

$$Z_t^1 = \frac{1}{d}\sum_{i=1}^{d} W_{t,i}^1, \quad Z_t^2 = \frac{1}{d}\sum_{i=1}^{d} W_{t,i}^2,$$

which differ in the choice of the weighting function g. Let $\boldsymbol{W}_t^1 = (W_{t,1}^1, \dots W_{t,d}^1)$ and $\boldsymbol{W}_t^2 = (W_{t,1}^2, \dots W_{t,d}^2)$ denote the vectors of the respective W-values at time t with weighting function g^1, respectively g^2. We further define the 2-dimensional vector $\widetilde{\boldsymbol{W}}_t = (\boldsymbol{W}_t^1, \boldsymbol{W}_t^2)'$.

We consider the bispectral test

$$T_{Bspec} = n\left(\bar{\boldsymbol{Z}} - \boldsymbol{\mu_W}\right)' \Sigma_{\boldsymbol{W}}^{-1} \left(\bar{\boldsymbol{Z}} - \boldsymbol{\mu_W}\right), \tag{5.13}$$

where

$$\bar{\boldsymbol{Z}} = (\bar{Z}^1, \bar{Z}^2) \quad \text{with} \quad \bar{Z}^k = n^{-1}\sum_{t=1}^{n} Z_t^k, \quad k = 1, 2$$

is the average desk exception rate and

$$\begin{aligned}\boldsymbol{\mu_W} &= (\mu_{W^1}, \mu_{W^2}) = \left(\mathbb{E}(\bar{Z}^1), \mathbb{E}(\bar{Z}^2)\right) = \left(\mathbb{E}(W_{t,i}^1), \mathbb{E}(W_{t,i}^2)\right) \\ &= \left(\int_{\alpha_1}^{\alpha_2} g^1(u)(1-u)\mathrm{d}u, \int_{\alpha_1}^{\alpha_2} g^2(u)(1-u)\mathrm{d}u\right)'.\end{aligned}$$

Note that $\mathrm{Var}(\bar{\boldsymbol{Z}}) = n^{-1}\Sigma_{\boldsymbol{W}}$ with

$$\Sigma_{\boldsymbol{W}} = \mathrm{Var}(Z_t^1, Z_t^2) = \begin{pmatrix} \mathrm{var}(Z_t^1) & \mathrm{cov}(Z_t^1, Z_t^2) \\ \mathrm{cov}(Z_t^1, Z_t^2) & \mathrm{var}(Z_t^2) \end{pmatrix}. \tag{5.14}$$

Under the null hypothesis, the test statistic in (5.13) is approximately χ^2-distributed with 2 degrees of freedom. This is a consequence of the multivariate Central Limit Theorem (Van der Vaart (2000)) from which follows that $\bar{\boldsymbol{Z}}$ is approximately bivariate normal distributed under the null hypothesis with finite covariance matrix $\Sigma_{\boldsymbol{W}}$. For the estimation of the variances in (5.14) we use method **M3**, which turned out to produce the best simulation results. That is we split the covariance into a variance part (which is known explicitly under the null hypothesis) and a correlation part (which is estimated from data). We have

$$\mathrm{Var}(Z_t^k) = \frac{1}{d^2}\mathbf{1}'\mathrm{Var}(\boldsymbol{W}_t^k)\mathbf{1} = \frac{\mu_{2,W^k} - \mu_{W^k}^2}{d^2}\mathbf{1}'\mathcal{R}_{\boldsymbol{W}^k}\mathbf{1},$$

with correlation matrix $\mathcal{R}_{\boldsymbol{W}^k} = \mathrm{Cor}(\boldsymbol{W}^k)$ for any t. For the correlations in $\mathcal{R}_{\boldsymbol{W}^k}$ we use the sample correlation estimate.

For the covariances in (5.14) we follow the same idea:

$$\begin{aligned}\mathrm{Cov}(Z_t^1, Z_t^2) &= \frac{1}{d^2}\mathbf{1}'\mathrm{Cov}(\boldsymbol{W}_t^1, \boldsymbol{W}_t^2)\mathbf{1} \\ &= \frac{1}{d^2}\sqrt{\mu_{2,W^1} - \mu_{W^1}^2}\sqrt{\mu_{2,W^2} - \mu_{W^2}^2}\,\mathbf{1}'\mathcal{R}_{\boldsymbol{W^1W^2}}\mathbf{1}.\end{aligned}$$

with $\mathcal{R}_{\boldsymbol{W^1W^2}} = \mathrm{Cor}(\boldsymbol{W}^1, \boldsymbol{W}^2)$ for any t.

Note that we assume that $\mathcal{R}_{\boldsymbol{W^1 W^2}}$ is independent of t. This suggests the estimator

$$\frac{1}{d^2}\sqrt{\mu_{2,W^1} - \mu_{W^1}^2}\sqrt{\mu_{2,W^2} - \mu_{W^2}^2}\boldsymbol{\iota}'\widehat{\mathcal{R}}_{\boldsymbol{W^1 W^2}}\boldsymbol{\iota},$$

where we use the sample correlation matrix as an estimator for $\mathcal{R}_{\boldsymbol{W^1 W^2}}$.

5.4 Results simulation study: PIT-value based tests

In this section, we present the results of our simulation study, which analyses the size and power of our spectral backtests. We illustrate how size and power are influenced by the various simulation parameters presented in Section 4.3.3, like the number of desks or the level of correlation. Additionally, we consider different kernels to see if the choice of the kernel function influences the performance of the tests.

For our Z-tests we decided to look at the following three continuous weighting functions g defined on $[\alpha_1, \alpha_2]$:

(1) The uniform weighting function $g(u) = 1$

(2) The linear weighting $g(u) = u$

(3) The exponential weighting function $g(u) = \exp(k(u - \alpha_1)) - 1$ with $k = 1$.

The values α_1 and α_2 determine the kernel window. We choose $\alpha_1 = 0.9805$ and $\alpha_2 = 0.9995$, which gives a symmetric interval around 0.99. Note that the functions are non-decreasing, placing more weight on more extreme outcomes (in the right tail).

For the multi-desk monospectral Z-tests the kernel functions listed above lead to three different tests which we denote by:

- SP.U: Uniform weighting function
- SP.L: Linear weighting function
- SP.E: Exponential weighting function

For multi-desk bispectral Z-tests, we will look at two different Z-tests, which combine the continuous kernel functions mentioned above:

- SP.UL: Uniform and linear weighting function
- SP.UE: Uniform and exponential weighting function

Comparison of variance estimation methods In Section 5.3.1 we introduced various alternatives for the estimation of the unknown variance in our spectral Z-test. In this section, we analyse the performance of these estimators with respect to the monospectral tests. The fraction parameter is kept at 100%, meaning that the risk modelling group misspecifies all desks. The results are presented in Table 5.2. We observe the following:

- For method 'None', dependencies across desks are ignored. All three multi-desk monospectral Z-tests are strongly oversized except for the case when we draw random variables from the Gauss copula with zero correlation. Note that in the case that we draw from a t4 copula, our test is oversized even in the case of zero correlation. This is due to the fact that t4 copulas exhibit tail dependencies, which are not grasped by the variance estimator.
- Method **M1** leads to strongly undersized tests. Method **M2** (coefficient of variation) is close to the nominal size but slightly oversized for samples with correlation, especially for small samples.
- Method **M3** leads to a well sized test both for $n = 250$ and $n = 500$ and in case of correlated data by simultaneously showing very good power results, especially in the larger sample of $n = 500$.
- Regarding the kernel weighting functions, no function seems to systematically outperform the others.
- Considering more desks in the simulation leads to slightly better power results.

Due to the findings above, we will concentrate on method M3, which turned out to be the best method to control for correlations across desks.

Bispectral tests We analyse the performance of the multi-desk bispectral Z-tests SP.UL and SP.UE. For the estimation of the variance, we apply method **M3**, which appeared to work best for the monospectral test. The fraction parameter is kept at 100%, meaning that the risk modelling group misspecifies all desks. Table 5.3 shows the results.

Interestingly, for both tests the tests' sizes are higher than for the monospectral Z-tests, when using method **M3**. The power is slightly higher, which can also be a consequence of the size. Concerning the size we observe that the results are noticeable better when $n = 500$.

We therefore analysed the performance under increased and decreased sample size, see Table 5.4 for the results. Increasing the sample size to 1000 leads to reasonable (slightly undersized)

tests. Testing for more than one moment seems to require a larger sample size so that the test statistic can be assumed to be χ^2-distributed.

Fraction In the simulation studies before the desk models were either all correctly specified or all were misspecified. This is a rather theoretical assumption. In a realistic scenario it is more likely that only a certain proportion of desk models is misspecified.

Therefore, we analyse the performance of the monospectral Z-tests SP.U, SP.L and SP.E when a certain proportion of the desks models were not correctly classified by the risk modelling group. In the simulation study, the risk modelling group assumes that the desks underlying data follow a standard normal distribution, $G_i(u) = \Phi(F^{-1}(u))$. The simulation parameter 'fraction' gives the proportion of desks, which underlying data follow a t-distribution with 4 degrees of freedom (meaning that F is a t4-distribution). The higher the fraction, the more desks are wrongly classified as being standard normal distributed.

Table 5.5 shows the results. We restrict to variance estimation method **M3**. The following is observed:

- For 25% of misspecified desks, tests are well-sized but the power is rather low (for both considered sample sizes n). The power increases slightly when more desks ($d = 100$) are considered.
- The power is good for larger samples when the data are generated from the Gauss copulas with zero correlation. We attribute this to the intuition that independent data increase the effective sample size, whereas correlated data decrease it.
- The power highly increases when we consider 50% wrongly classified desks and further increases for 75%, where we have acceptable power even for small sample sizes.

			d	50				100			
			ρ	Zero		Flat		Zero		Flat	
test	cor	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
SP.U	None	250	Normal	5.0	23.6	26.2	29.2	5.1	29.6	30.5	32.9
			t4	100.0	93.6	89.8	79.0	100.0	96.4	92.4	82.4
		500	Normal	5.1	24.7	27.5	32.9	5.4	29.4	32.5	36.7
			t4	100.0	99.4	97.7	91.1	100.0	99.7	98.6	93.4
	M1	250	Normal	3.9	1.8	2.0	1.1	3.5	1.6	1.9	0.9
			t4	100.0	64.5	46.7	22.3	100.0	66.6	47.3	22.2
		500	Normal	4.7	1.7	2.4	1.6	4.1	2.2	2.0	1.7
			t4	100.0	91.6	78.0	46.6	100.0	93.5	80.6	47.2
	M2	250	Normal	5.0	5.9	6.7	7.6	5.1	6.0	6.7	7.0
			t4	100.0	79.6	67.7	47.0	100.0	79.6	69.3	46.6
		500	Normal	5.1	5.2	6.8	6.9	5.4	5.0	6.5	6.7
			t4	100.0	95.8	87.7	68.2	100.0	97.0	89.1	68.5
	M3	250	Normal	4.6	4.4	4.2	4.5	4.9	4.9	4.4	4.6
			t4	100.0	80.0	65.7	41.7	100.0	80.0	67.2	42.9
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	100.0	95.9	87.3	65.8	100.0	97.3	88.9	66.3
SP.L	None	250	Normal	4.6	22.7	25.5	28.9	5.1	28.5	29.3	32.1
			t4	100.0	98.7	96.4	90.2	100.0	99.2	97.8	92.2
		500	Normal	5.1	25.5	26.6	34.2	5.5	30.4	32.4	37.7
			t4	100.0	100.0	99.7	97.4	100.0	100.0	99.9	98.4
	M1	250	Normal	3.9	1.5	1.8	1.1	3.6	1.3	1.7	1.1
			t4	100.0	87.7	75.0	34.7	100.0	89.2	77.4	34.9
		500	Normal	4.5	1.7	2.4	1.2	4.3	2.3	1.7	1.3
			t4	100.0	99.5	96.7	71.0	100.0	99.6	97.2	70.8
	M2	250	Normal	4.6	6.2	7.0	7.6	5.1	6.3	6.9	8.1
			t4	100.0	94.2	89.5	65.8	100.0	96.3	90.3	67.0
		500	Normal	5.0	5.2	7.2	7.1	5.5	4.7	7.0	7.1
			t4	100.0	99.9	98.1	86.5	100.0	99.8	98.7	87.0
	M3	250	Normal	4.3	4.6	4.2	5.4	4.8	5.1	4.3	5.2
			t4	100.0	94.3	88.9	62.8	100.0	96.4	90.1	64.5
		500	Normal	5.0	3.3	5.1	5.0	5.5	3.5	4.6	5.2
			t4	100.0	99.9	98.1	86.0	100.0	99.8	98.7	86.3
SP.E	None	250	Normal	5.0	23.6	26.2	29.2	5.1	29.6	30.5	32.9
			t4	100.0	93.6	90.0	79.0	100.0	96.5	92.5	82.5
		500	Normal	5.1	24.7	27.5	32.9	5.4	29.4	32.5	36.8
			t4	100.0	99.4	97.7	91.3	100.0	99.7	98.6	93.5
	M1	250	Normal	3.9	1.8	2.0	1.1	3.5	1.6	1.9	0.9
			t4	100.0	64.7	47.0	22.3	100.0	66.8	47.4	22.3
		500	Normal	4.7	1.7	2.4	1.6	4.1	2.2	2.0	1.7
			t4	100.0	91.8	78.1	46.6	100.0	93.6	80.8	47.5
	M2	250	Normal	5.0	5.9	6.7	7.6	5.1	6.0	6.7	7.0
			t4	100.0	79.7	67.9	47.0	100.0	80.1	69.5	46.7
		500	Normal	5.1	5.2	6.8	6.9	5.4	5.0	6.5	6.7
			t4	100.0	95.8	88.0	68.4	100.0	97.1	89.5	68.6
	M3	250	Normal	4.6	4.4	4.2	4.4	4.9	4.9	4.4	4.6
			t4	100.0	80.2	66.1	42.0	100.0	80.3	67.8	43.0
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	100.0	96.1	87.4	66.1	100.0	97.5	89.1	66.4

Table 5.2: Size and power of monospectral Z-tests using different variance estimation methods

		d	50				100			
		ρ	Zero		Flat		Zero		Flat	
test	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
SP.UL	250	Normal	5.8	9.4	11.6	18.6	4.1	12.3	12.2	21.0
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	Normal	2.7	5.4	5.0	10.2	3.1	7.2	6.6	10.0
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SP.UE	250	Normal	6.6	11.2	12.7	20.2	4.9	14.0	14.5	22.4
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	Normal	3.0	6.2	5.4	10.6	3.6	7.7	6.8	10.7
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 5.3: Size and power of bispectral Z-tests

		d	50				100			
		ρ	Zero		Flat		Zero		Flat	
test	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
SP.UL	100	Normal	17.7	33.9	30.2	49.9	20.2	38.9	35.0	54.6
		t4	100.0	99.7	100.0	98.6	100.0	100.0	100.0	99.4
	1000	Normal	3.0	3.1	3.1	4.8	2.3	4.5	4.6	5.5
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SP.UE	100	Normal	22.6	40.0	35.4	54.8	25.2	45.2	38.0	57.2
		t4	100.0	99.7	100.0	98.7	100.0	100.0	100.0	99.5
	1000	Normal	3.1	3.0	3.3	5.2	2.5	4.7	4.8	6.0
		t4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 5.4: Size and power of bispectral Z-tests with decreased and increased sample size

			d	50				100			
			ρ	Zero		Flat		Zero		Flat	
test	fraction	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
SP.U	25	250	Normal	4.6	4.4	4.2	4.5	4.9	4.9	4.4	4.6
			t4	43.0	15.6	12.8	10.1	69.9	16.8	14.1	10.7
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	66.7	22.2	18.3	13.4	90.4	24.0	19.6	13.1
	50	250	Normal	4.6	4.4	4.2	4.5	4.9	4.9	4.4	4.6
			t4	89.0	36.5	29.6	18.5	99.3	39.6	29.3	19.9
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	99.3	56.2	43.1	27.8	100.0	58.9	43.5	27.7
	75	250	Normal	4.6	4.4	4.2	4.5	4.9	4.9	4.4	4.6
			t4	99.6	59.2	46.7	29.3	100.0	63.1	48.6	29.9
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	100.0	82.5	69.8	46.5	100.0	85.3	72.3	47.7
SP.L	25	250	Normal	4.3	4.6	4.2	5.4	4.8	5.1	4.3	5.2
			t4	67.2	23.1	19.1	13.0	89.7	25.9	20.3	14.5
		500	Normal	5.0	3.3	5.1	5.0	5.5	3.5	4.6	5.2
			t4	88.8	34.7	28.1	18.7	99.3	37.8	30.8	19.7
	50	250	Normal	4.3	4.6	4.2	5.4	4.8	5.1	4.3	5.2
			t4	99.2	56.7	46.5	27.4	100.0	58.7	47.4	27.2
		500	Normal	5.0	3.3	5.1	5.0	5.5	3.5	4.6	5.2
			t4	100.0	80.0	68.7	43.3	100.0	82.1	70.2	44.5
	75	250	Normal	4.3	4.6	4.2	5.4	4.8	5.1	4.3	5.2
			t4	100.0	82.0	71.0	45.1	100.0	84.2	74.8	45.6
		500	Normal	5.0	3.3	5.1	5.0	5.5	3.5	4.6	5.2
			t4	100.0	96.5	91.3	67.7	100.0	98.1	92.8	69.0
SP.E	25	250	Normal	4.6	4.4	4.2	4.4	4.9	4.9	4.4	4.6
			t4	43.1	15.6	12.9	10.3	70.1	16.9	14.1	10.7
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	66.8	22.3	18.3	13.4	90.4	24.2	19.6	13.1
	50	250	Normal	4.6	4.4	4.2	4.4	4.9	4.9	4.4	4.6
			t4	89.1	36.6	29.6	18.5	99.3	39.8	29.5	20.0
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	99.3	56.6	43.4	27.9	100.0	59.1	43.7	28.0
	75	250	Normal	4.6	4.4	4.2	4.4	4.9	4.9	4.4	4.6
			t4	99.6	59.3	46.9	29.6	100.0	63.3	48.9	30.4
		500	Normal	5.0	4.1	5.0	4.5	5.4	4.0	4.5	4.6
			t4	100.0	82.6	69.9	46.9	100.0	85.5	72.5	47.8

Table 5.5: Size and power of monospectral Z-tests when a certain proportion of desks is misspecified

5.5 Summary

The main conclusions of this section are:

- Based on the theory in Gordy and McNeil (2020) we proposed multi-desk spectral tests to answer the research question if the trading risks of the bank are modelled adequately across desks. The tests take the form of a Z-test, averaging the daily desks exception rates, and make use of transformed indicator variables for realised PIT-value level exceedances as input variable. The tests are designed in a way that the risk modelling group can emphasize the region of the estimated loss distribution about which it is most concerned with respect to model performance and can choose a weighting scheme for observed violations.

- In the case of multi-desk monospectral tests, one weighting function is considered. We also proposed multi-desk bispectral tests, which consider more than one weighting function.

- We suggested and analysed three different variance estimation methods to estimate the unknown variance in the test function. In a simulation study it turned out that the method, where only the correlation part of the covariance is estimated (applying the analytically calculable variance), is best to control for the unknown dependencies between desks. In this case, the test is well-sized with good power properties (mostly well above 70% when a certain proportion of desks are misspecified).

- The simulation study revealed that an increase of the number of desks from 50 to 100 did not improve the size and power performance of the proposed multi-desk tests in a significant way. The performance improved when the number of observations was increased from 250 (roughly one year of observations) to 500 (roughly two years of observations). The choice of the weighting function did not have an impact on the performance.

- According to the simulation study more than 25% of the desks must be misspecified to yield acceptable power levels.

- Compared to the monospectral Z-test, the bispectral Z-tests lead to (slightly) oversized tests, especially for $n = 250$ observations. Testing for more than one first moment seems to require larger sample sizes for the Central Limit Theorem.

Chapter 6

Bootstrap Z-tests

6.1 Introduction

The proposed multi-desk spectral tests presented in Chapter 5 take the form of Z-tests. In these tests the distribution of the test statistic is asymptotically normal but the rate of convergence may distort the performance of the test. Although we have shown in the previous chapter that the new proposed tests work satisfactorily for the typical backtesting periods applied in risk management applications (where the number of observations is usually 250 or 500 days), we want to analyse the effect of using bootstrap methods instead of relying on the Central Limit Theorem. Since our multi-desk spectral Z-tests have been analysed using simulated rather than real bank backtesting data[29], insights into the possible gains of using bootstrapping can be valuable to decide whether the asymptotic Z-tests can be improved upon. Using bootstrap methods, we are independent of any distributional assumption. This can be beneficial in comparison to the application of asymptotically justified tests since the latter may sometimes result in too few correct rejections of the null hypothesis, which lead to a reduction in the power of a statistical test. In fact, it has been shown that confidence intervals based on bootstrapped data can be asymptotically more accurate than confidence intervals built on the normality assumption (DiCiccio and Efron (1996)). As explained in Efron and Tibshirani (1993), p.156, every bootstrap hypothesis test is dual to a bootstrap confidence interval. We will use this insight to construct bootstrap confidence intervals equivalent to our null hypotheses for multi-desk monospectral and multi-desk bispectral tests.

In this chapter, we propose a bootstrap confidence interval for the multi-desk monospectral test,

[29]There is currently very little data available to researchers due to the sensitivities surrounding bank risk reporting both in the EU and the USA.

which is more or less straightforward since it builds on the well-known theory of (Studentized-t) bootstrap confidence intervals in the one-dimensional space. For the multi-desk bispectral test, finding a confidence region for the vector-based test statistic is more complex. We propose and analyse two alternatives: an intuitive circular confidence region; a confidence region based on Tukey's concept of half-space depth. In a simulation study, we compare the size and power properties of the proposed bootstrap methods. For the multi-desk bispectral test we also analyse the advantages and disadvantages of the two proposed methods for confidence regions.

6.2 Mathematical foundations

6.2.1 Bootstrapping

Bootstrapping is a computational method which can be used to estimate the standard error of an estimate (for example, an estimate of the mean of a random variable) in order to assess its accuracy. It was popularized in the seminal paper of Efron (1979). The advantage of empirical bootstrapping is that it is a non-parametric method so that it does not require any assumptions about the distribution of the underlying data. Using the bootstrap estimate for the standard error it is possible to construct bootstrap confidence intervals for a point estimate. Since every confidence interval is dual to a statistical hypothesis, this interval can be used for hypothesis testing.

Suppose we have univariate data points $x_1, \ldots, x_n$, which have been randomly drawn from some distribution function F. An empirical bootstrap sample of $x_1, \ldots, x_n$ is generated by sampling with replacement from $x_1, \ldots, x_n$ using the same sample size n. The bootstrap sample $x_1^*, \ldots, x_n^*$ can be thought of as a sample drawn from the empirical distribution function F_n of the data points $x_1, \ldots, x_n$. The standard deviation σ_θ of an estimator $\theta = f(x_1, \ldots, x_n)$ is approximated by the bootstrap estimate of σ_θ, which is the standard deviation of $\theta = f(x_1, \ldots, x_n)$ calculated on the data sets of the same size n which were randomly sampled from the empirical distribution function F_n. Note that this form of bootstrapping procedure is designed for independent and identically distributed data. For dependent data various generalisations of the original bootstrapping procedure, like block-bootstrap methods, have been proposed. For a review, see for example Kreiss and Lahiri (2012).

6.2.2 Tukey depth and multivariate trimmed means

We will later study multivariate tests, where the test statistic is a two-dimensional vector. The test performance will be analysed using both the asymptotic distribution derived from the Central Limit Theorem and using a bootstrap procedure. For the latter, a number of bootstrap samples (say R) are drawn and for each of these samples the particular test statistic is calculated. The R bootstrap test statistic values are used to derive a rejection area for the null hypothesis. In the two-dimensional case, the test statistics form two-dimensional vectors.

In the univariate case we form an empirical confidence interval from the bootstrap sample using empirical quantiles of the sample. In the multivariate case we need an analogous procedure, but there is more flexibility in how to proceed. We need a measure of centrality for each bootstrap data point and one method for achieving this is based on the concept of Tukey depth (Tukey (1975)). Instead of simply taking the sample mean, a (multivariate) trimmed mean is used for the 'most central point' as proposed in Massé (2009). A depth function measures the centrality of a data point within a point cloud or a probability distribution. Tukey depth is one out of a number of depth concepts in the literature and arguably the most popular.

Definition 23 (Tukey Depth)**.** *Let $\boldsymbol{X}$ be a $\mathbb{R}^p$ random variable with probability distribution F and let $\mathcal{H}$ denote the class of closed halfspaces H in $\mathbb{R}^p$. The Tukey depth (or halfspace depth) of a point $x \in \mathbb{R}^p$ (with respect to F) is defined as*

$$\mathcal{D}(F, x) := \inf\{F(H) : H \in \mathcal{H}, x \in H\},$$

where $F(H)$ denotes the probability that the random variable $\boldsymbol{X}$ is in the halfspace H.

See Massé (2009), p.367. When we have a data sample this concept can be analogously applied based on the empirical distribution of the data. Given a finite set S of n data points and a point $s \in \mathbb{R}^p$, the Tukey depth of s is then the minimum number of points of S contained in any closed halfspace with s on its boundary.

One can think of depth as a value which decreases monotonically the farther a data point is away from the centre of the distribution F in any direction. The concept of depth is used in Massé (2009) to define a multivariate analogue of a univariate trimmed mean by calculating the mean of a finite set of points which have a certain depth with respect to a probability function. An α-trimmed region with respect to F is defined as $Q_\alpha = Q_\alpha(F) = \{x : \mathcal{D}(F, x) \geq \alpha\}$. A

depth-based multivariate trimmed mean is then defined by averaging the elements that belong to the trimmed region. The sample trimmed mean based on the Tukey depth for two or more dimensions is implemented in the R package `depth` (Genest et al. (2019)).

6.3 Spectral tests with bootstrap confidence intervals

In this section we consider the case of multi-desk monospectral tests. Note that the method explained here can also be applied to the multi-desk Binomial score test in (4.11). Recall that the multi-desk monospectral Z-test takes the form

$$T_{Mspec} = \sqrt{n}\ \frac{\bar{Z} - \mu_W}{\sigma_Z} \tag{6.1}$$

with average desk exception rate

$$\bar{Z} = n^{-1} \sum_{t=1}^{n} Z_t,$$

and where $\sigma_Z^2 = \text{var}(Z_t)$ and

$$\mu_W = \mathbb{E}(\bar{Z}) = n^{-1} \sum_{t=1}^{n} \mathbb{E}(Z_t) = \mathbb{E}\left(W_{t,i}\right) = \int_{\alpha_1}^{\alpha_2} g(u)(1-u)\mathrm{d}u,$$

see Section 5.3.1. Under the null hypothesis and using a consistent estimator of the variance σ_Z^2, we constructed a test that assumed the asymptotic normality of the test statistic.

Now we are looking for a confidence interval with confidence level $1-\beta$, such that we reject the null hypothesis, if

$$T = \frac{\mu_W^0 - \mu_W}{\sigma_Z^0} \notin \widetilde{\mathcal{R}}_\beta,$$

where μ_W^0 and σ_Z^0 denote estimates of μ_W and σ_Z based on the original sample before bootstrapping and the factor $n^{1/2}$ is absorbed in T compared to expression (6.1). An interval of desired confidence level $1-\beta$ for μ_W is then given by

$$[\mu_W^0 - T_{\beta/2}\sigma_Z^0, \mu_W^0 + T_{1-\beta/2}\sigma_Z^0], \tag{6.2}$$

where T_β is the β quantile of the distribution of T. The null hypothesis is rejected if

$$\mu_W \notin [\mu_W^0 - T_{\beta/2}\sigma_Z^0, \mu_W^0 + T_{1-\beta/2}\sigma_Z^0] = \widetilde{\mathcal{R}}_\beta. \tag{6.3}$$

Since this distribution of T is unknown, the empirical bootstrap distribution is used as an approximation for the unknown distribution. The so-called Studentized-t bootstrap confidence interval (see, for example, Hall (1987)), is based on the distribution of values

$$T^{b_i} = \frac{\mu_W^{b_i} - \mu_W^0}{\sigma_Z^{b_i}}, \tag{6.4}$$

where $\mu_W^{b_i}$ and $\sigma_Z^{b_i}$ denote the bootstrap estimates from bootstrap replication i for $i = 1, \ldots, R$. From this, we can construct a Studentized-t bootstrap confidence interval of the form (6.2) replacing T_β with the respective quantiles from the empirical bootstrap distribution.

6.4 Bispectral tests with bootstrap confidence regions

The idea of using bootstrap methods to find an appropriate rejection region dual to some hypothesis can also be applied to estimates of an unknown *vector* of dimension d. In such cases, we are looking for *confidence regions* instead of confidence intervals. Whereas the construction of bootstrap confidence intervals ($d = 1$) is widely studied, the construction of bootstrap confidence regions ($d \geq 2$) is more complex (as pointed out, for example, in Yeh and Singh (1997)). The construction is not straightforward since it requires a measure which excludes a proportion β of the set of possible values of some parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ in higher dimensions in order to construct a confidence region of confidence level $1 - \beta$ for the parameter vector $\boldsymbol{\theta}$. For the special case $d = 2$, the problem reduces to a point cloud in the two-dimensional space where some contour has to be found which includes $1 - \beta$ percent of the data points. This contour defines the bootstrap confidence region.

Similar to our monospectral tests, for the bispectral test we want to test whether the unknown vector $\boldsymbol{\mu_W} = (\mu_{W^1}, \mu_{W^2})$ lies within a bootstrap confidence region of confidence level $1 - \beta$. If $\boldsymbol{\mu_W}$ does not lie in this confidence region, the null hypothesis must be rejected.

Recall that the bispectral test is based on a random vector $\boldsymbol{S}$ given by

$$\boldsymbol{S} = \Sigma_{\boldsymbol{Z}}^{-1/2} \left(\bar{\boldsymbol{Z}} - \boldsymbol{\mu_W} \right), \tag{6.5}$$

where $\bar{\boldsymbol{Z}} = (\bar{Z}_1, \bar{Z}_2)$ with $\bar{Z}_k = n^{-1} \sum_{t=1}^n Z_t^k$, $Z_t^k = d^{-1} \sum_{i=1}^d W_{t,i}^k$, $\boldsymbol{\mu_W} = \mathbb{E}(\bar{\boldsymbol{Z}})$ and $\Sigma_{\boldsymbol{Z}} =$

$\text{Var}(Z_t^1, Z_t^2)$, which is assumed to be positive definite. The form of the bispectral test in (5.13) is obtained by taking the square of the norm of (6.5) times the sample size n.

We want to find a confidence region $\mathcal{R}_\beta$ such that we reject the null hypothesis in (5.6) if $\boldsymbol{\mu_W} \notin \mathcal{R}_\beta$. A confidence region of desired confidence level $1-\beta$ is given by

$$\mathcal{R}_\beta = \left\{ \bar{\boldsymbol{Z}} - \Sigma_{\boldsymbol{Z}}^{1/2} s : s \in \mathcal{S}_\beta \right\},$$

where $\mathcal{S}_\beta$ is any region such that $\mathbb{P}(\boldsymbol{S} \in \mathcal{S}_\beta) = 1-\beta$ under the null hypothesis. This is easy to see. Let $\Sigma_{\boldsymbol{Z}}^{-1/2}\left(\bar{\boldsymbol{Z}} - \boldsymbol{\mu_W}\right) = s$ with $s \in \mathcal{S}_\beta$. This is equivalent to $\bar{\boldsymbol{Z}} - \boldsymbol{\mu_W} = \Sigma_{\boldsymbol{Z}}^{1/2} s$ which can be written as $\boldsymbol{\mu_W} = \bar{\boldsymbol{Z}} - \Sigma_{\boldsymbol{Z}}^{1/2} s$. Consequently, the shape of the region $\mathcal{R}_\beta$ depends on the shape of $\mathcal{S}_\beta$. Ghosh and Polansky (2014), for example, considered ellipsoidal confidence regions, which would in our case be implied when choosing a circular form for $\mathcal{S}_\beta$.

Using the boostrap idea, we approximate $\bar{\boldsymbol{Z}}$ and $\Sigma_{\boldsymbol{Z}}$ by their bootstrap estimates to find the contour for $\mathcal{S}_\beta$. We consider two alternative forms which are

- an intuitive **circular** form for $\mathcal{S}_\beta$, which implies an ellipsoidal form for $\mathcal{R}_\beta$ (we refer to it as circular confidence region);
- a convex hull, using half-space **depth sets** based on Tukey's depth (we refer to it as depth-based confidence region).

The circular form encloses a proportion of $1-\beta$ of all observations around the center, where the center is calculated as the trimmed mean based on the Tukey's depth. The confidence region based on half-space depth encloses all data with a depth of $1-\beta$.

6.4.1 Circle

Let S_β be a d-dimensional **circle** centred at the origin such that $\mathbb{P}(\boldsymbol{S} \in S_\beta) = 1-\beta$. Then

$$\mathcal{R}_\beta = \{\bar{\boldsymbol{Z}} - \Sigma_{\boldsymbol{Z}}^{1/2} s : s \in S_\beta\} \tag{6.6}$$

is a confidence region for $\boldsymbol{\mu_W}$ with confidence level $1-\beta$ (under the null hypothesis). In the previous section we have seen that the confidence region $\mathcal{R}_\beta$ is designed to contain a possible set of values around $\boldsymbol{\mu_W}$ such that $\mathbb{P}(\boldsymbol{S} \in \mathcal{R}_\beta) = 1-\beta$. A circular bootstrap Studentized-t confidence region based on (6.6) can be constructed by considering

$$T^{b_i} = \left(\Sigma_{\boldsymbol{Z}}^{b_i}\right)^{-1/2} \left(\boldsymbol{\mu}_{\boldsymbol{W}}^{b_i} - \boldsymbol{\mu}_{\boldsymbol{W}}^{0}\right), \tag{6.7}$$

where $\boldsymbol{\mu}_{\boldsymbol{W}}^{0}$ is the value for the expectation of $\bar{\boldsymbol{Z}}$ calculated on the original sample and $\boldsymbol{\mu}_{\boldsymbol{W}}^{b_i}$ the expectation based on the bootstrap sample. We calculate the center of the R bootstrap statistics $T^{b_1}, \ldots, T^{b_R}$ using the multivariate trimmed mean (see Section 6.2.2). Then we determine the radius of a circle with this center which encloses $(1-\beta)\%$ of the data. This is the circular region S_β. If it holds for $\boldsymbol{S}$ that $\boldsymbol{S} = \Sigma_{\boldsymbol{Z}}^{-1/2}\left(\bar{\boldsymbol{Z}} - \boldsymbol{\mu}_{\boldsymbol{W}}\right) \in S_\beta$ the null hypothesis cannot be rejected.

6.4.2 Depth sets

The first steps are the same as in the derivation of the circular bootstrap confidence region. We first calculate the studentized statistic

$$T^{b_i} = \left(\Sigma_{\boldsymbol{Z}}^{b_i}\right)^{-1/2}\left(\boldsymbol{\mu}_{\boldsymbol{W}}^{b_i} - \boldsymbol{\mu}_{\boldsymbol{W}}^{0}\right)$$

using the bootstrap sample to estimate the variance-covariance matrix and the mean $\boldsymbol{\mu}_{\boldsymbol{W}}^{b_i}$. These bootstrap test values are used to construct $\mathcal{S}_\beta$. Instead of using a circle, we use Tukey's concept of depth to delete all points T^{b_i} until the desired confidence level of $1-\beta$ is reached. We do this by calculating the Tukey depth of each data point T^{b_i} and exclude the $\beta\%$ points with smallest depth. So we remove the $\beta\%$ most exterior points from the data cloud. A boostrap confidence region of the form

$$\mathcal{R}_\beta = \left\{\bar{\boldsymbol{Z}} - \Sigma_{\boldsymbol{Z}}^{1/2} s : s \in \mathcal{S}_\beta\right\},$$

is built with $\mathcal{S}_\beta$ being the contour around all data points with depth larger than the $\beta\%$-quantile of the depth values distribution.

6.5 Simulation setup

In the following, we briefly describe the simulation procedure used to construct the bootstrap confidence intervals and regions described in the previous sections.

Note again that we consider functions of the form

$$W_{t,i} = \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_{t,i} \geq u\}} \mathrm{d}u,$$

which are transformed indicators of realised PIT-values $P_{t,i} = \widehat{F}_{t,i}(L_{t,i})$.

6.5.1 Bootstrap confidence intervals

For monospectral tests, the simulation works as follows: Let $\boldsymbol{P}_1^0, \ldots, \boldsymbol{P}_n^0$ be the vectors of daily PIT-values across all desks so that $\boldsymbol{P}_t^0 = (P_{t,1}^0, \ldots, P_{t,d}^0)$, $t = 1, \ldots, n$, where $P_{t,i}$ denotes the PIT-value of desk i at day t. In each bootstrap replication a bootstrap sample $\boldsymbol{P}_1^{b_i}, \ldots, \boldsymbol{P}_n^{b_i}$, $i = 1, \ldots, R$, is generated by sampling directly total vectors of length d from the data set $\boldsymbol{P}_1^0, \ldots, \boldsymbol{P}_n^0$ with replacement. Note that by sampling vectors from the 'rows data' $\boldsymbol{P}_1^0, \ldots, \boldsymbol{P}_n^0$, the correlation structure inherent in the desks on a day t is not destroyed. For each bootstrap sample, we calculate the bootstrap mean $\mu_W^{b_i}$ and the bootstrap standard deviation $\sigma_Z^{b_i}$, $i = 1, \ldots, R$,

$$\left(\mu_W^{b_1}, \sigma_Z^{b_1}\right), \ldots, \left(\mu_W^{b_R}, \sigma_Z^{b_R}\right).$$

These empirical bootstrap estimates are then used to calculate the required β-quantiles in the calculation of the Studentized-t bootstrap confidence interval for the mean μ_W:

$$[\mu_W^0 - T_{\beta/2}^b \sigma_Z^0, \mu_W^0 + T_{1-\beta/2}^b \sigma_Z^0], \tag{6.8}$$

where μ_W^0 and σ_Z^0 are the empirical mean and empirical standard deviation of $\bar{Z}$ calculated from the observed data sample. T_β^b denotes the β-quantile of the empirical bootstrap distribution of the R bootstrapped Student's t values

$$T^{b_i} = \frac{\mu_W^{b_i} - \mu_W^0}{\sigma_Z^{b_i}}, \; i = 1, \ldots, R,$$

that is the sample quantile of level β. In the simulation study, a Studentized-t bootstrap interval is generated based on the R bootstrapped data for each simulation dataset. We check for each simulated dataset if the confidence interval (6.8) covers the value μ_W i.e. we check

$$\mu_W \in [\mu_W^0 - T_{\beta/2}^b \sigma_Z^0, \mu_W^0 + T_{1-\beta/2}^b \sigma_Z^0].$$

If μ_W is not covered by the interval, the null hypothesis in (5.5) must be rejected. If the null hypothesis is true, we expect μ_W to lie within the confidence interval for a proportion of $(1-\beta)$ of the simulation samples. For the calculation of the Studentized-t bootstrap confidence interval we use the R package `boot`, see Angelo and Ripley (2020) and Davison and Hinkley (1997).

6.5.2 Bootstrap confidence regions

For bispectral tests, we proposed confidence regions of two different shapes. The bootstrap test statistics T^{b_i}, $i = 1, \ldots, R$, are simulated in the same way for both cases. The difference is in the decision against the null hypothesis, which incorporates the shape of the region (circle or contour derived from halfspaces).

For each bootstrap sample, we calculate the empirical bootstrap means and the components of the empirical bootstrap variance-covariance matrix $\Sigma_{\boldsymbol{Z}}^{b_i}$

$$\Sigma_{\boldsymbol{Z}}^{b_i} = \begin{pmatrix} \sigma_{Z^1}^{b_i} & \text{cov}_Z^{b_i} \\ \text{cov}_Z^{b_i} & \sigma_{Z^2}^{b_i} \end{pmatrix},$$

where $\sigma_{Z^j}^{b_i}$ are the empirical bootstrap variances of Z_t^j, $j = 1, 2$, and $\text{cov}_Z^{b_i}$ denotes the empirical bootstrap covariance $\text{cov}(Z_t^1, Z_t^2)$. The variances $\sigma_{Z^j}^{b_i}$ are estimated according to variance estimation method **M3** (see Section 5.3.1). All formulas can be found in Section 5.3.2.

The empirical bootstrap estimates are then used to calculate the bootstrap test statistics T^{b_i}. For each simulated dataset, we calculate the test value $\boldsymbol{S}$ in (6.5).

For the circular confidence region we check for each simulation step if the distance of the test value to the center of the bootstrap data T^{b_i}, $i = 1, \ldots, R$, is larger than the distance of the bootstrap value T^{b_i} with the 95% largest distance. If so, the null hypothesis is rejected. The center of the point cloud T^{b_i} is calculated using the multivariate trimmed mean of Tukey (see Section 6.2.2).

For the depth-based confidence region we check in each simulation step, if the depth of the test value is smaller than the 95% quantile of the distribution of depth values of T^{b_i}, $i = 1, \ldots, R$. If so, the null hypothesis is rejected.

6.6 Results simulation study: bootstrapping

The higher the number of bootstrap samples‘R the smaller the potential random sampling error, which can lead to less robust standard error estimates. In Efron et al. (2015) is has been shown that numbers of repetitions as low as 50 can lead to reliable estimates. However, for the construction of bootstrap confidence intervals, higher numbers are recommended (see Efron and Tibshirani (1993), p.161). We compared the performance of the multi-desk Binomial score test for $R = 100$ and $R = 500$ for $d = 50$ desks and $\alpha = 0.99$. Comparing the results for $R = 100$

and $R = 500$ in Table 6.1, we see that an increase of the number of bootstrap repetitions from $R = 100$ to $R = 500$ does not significantly improve the size. Therefore, we consider $R = 100$ bootstrap repetitions in all of the following simulation studies, unless stated otherwise.

Monospectral tests We analyse the performance of the three versions of the monospectral Z-tests, which have been presented and analysed in Section 5.3.1. That is we consider the same weighting functions: the uniform (test SP.U); the linear (test SP.L); and the exponential (test SP.E) one. Due to computational speed, we restrict ourselves to $d = 50$ desks, considering that the previous simulation results in Section 5.4 have already shown that the test performance is quite similar for $d = 50$ and $d = 100$ desks. We assume that the risk modelling group misspecifies a certain proportion of desks ranging from 25% to 75%. Table 6.2 shows the results. The following can be observed:

- SP.U is well-sized. However, the power properties are generally mediocre, especially for 25% misspecified desks. Interestingly, the power is quite high with values above 70% when at least 50% desks are misspecified and the simulated data are independent (Gauss copula with zero correlation). The latter can be explained by the fact that for independent desks, the effective sample size is increased. Power properties improve when considering 500 instead of 250 days.

- The findings for SP.L and SP.E are very similar to the ones of SP.U. The choice of the weighting function does not heavily influence the performance of the spectral tests with respect to size and power, whereby the general results seem to be best for SP.L.

- Comparing the results with Table 5.5, which gives the analogous results of the monospectral Z-test using the Central Limit Theorem instead of bootstrapping, we see that the general results are the same. There are no significant performance differences in the size. However, the power is on a higher level when using the normality approximation. For example, for 25% misspecified desks for SP.U for independent data, the power is 27.8% using the bootstrap method, whereas it is 43.0% using the normality approximation. In both versions (Central Limit Theorem and bootstrapping), we observe that the power is highest for Gaussian independently distributed data.

Bispectral tests For the bispectral Z-test we analyse the performance of the SP.UL test, both using a circular confidence region and using a depth-based confidence region. We concentrate on SP.UL since previous results in Section 5.4 have shown that the performance of the bispectral

Z-tests is not much influenced by the choice of the weighting functions.

First, we compare the size and power of the tests, when 100% of the desks are misspecified (parameter fraction is 100%). The results for the circular confidence region can be found in Table 6.3, the findings for the depth-based confidence region in Table 6.4. For $R = 100$, it can be observed that for the circular confidence region, the considered bispectral test shows adequate sizes, which are however a bit lower than expected, especially for $n = 250$ days. The size stabilises around the desired nominal level for all considered copulas and correlation levels when the number of observations increases from $n = 250$ to $n = 500$. Increasing the number of bootstrap repetitions from $R = 100$ to $R = 500$, we see that the test tends to be more undersized especially for $n = 250$ days. The power ranges up to 100%, indicating that the confidence region could be too small. For the depth-based confidence region, the sizes are too large to be acceptable for $R = 100$. This also does not change when increasing the number of observations to $n = 500$. For the depth-based confidence region, we see that the size generally improves when using $R = 500$ instead of $R = 100$ bootstrap repetitions, leading to the conclusion that for depth-based confidence regions a larger number of bootstrap repetitions is necessary. In nearly all simulations settings, the test detects around 100% of the misspecified desks, especially for $n = 500$ days. However, the high size levels could be the reason for the observed high power when using the depth-based confidence region.

Comparing the results obtained using bootstrap confidence regions with Table 5.3, which shows the size and power properties of the SP.UL test when using the normality assumption, we see that for $n = 250$ days the results are more similar to the findings using the depth-based confidence region, whereas for $n = 500$ the results resemble the ones using a circular confidence region. We have already seen in Section 5.4 that the observed large levels of size for the bispectral test using the normality assumption improves when increasing the sample size to $n = 1000$ (see Table 5.4). We conclude that testing for more than one moment seems to require a larger sample size so that the test statistic can be assumed to be χ^2-distributed. The simulation study in this section indicates that a circular confidence region could be the better option compared to the normality assumption when the observed sample size is less than 1000 days.

Next, we compare the performance when the risk modelling group misspecifies the distribution for a certain ratio of desks. Table 6.5 shows the findings for the SP.UL test when a circular and depth-based confidence region is used. We again observe that using a depth-based confidence region leads to highly oversized test. For the circular confidence region, the size is around the nominal level for $n = 500$ for all fractions. The power is mediocre for 25% misspecified desks but clearly improves for 50% and 75% of misspecified desks, power results being close to 100%

for 75% misspecified desks.

To get a clearer picture why the performances are quite different for the circular confidence region and the depth-based confidence region, we plot the bootstrapped test values as a point cloud and include the test statistics value as well as the circular confidence region and the depth-based confidence region. Doing this 10 times we get an impression why the test using a circular rejection area tends to be undersized whereas the test using a rejection area based on halfspace depths is heavily oversized. We used $R = 500$ bootstrap repetitions[30] and we concentrate on the Gauss copula with flat correlation and $n = 250$ observations.

Figure 6.1 shows the situation under the null hypothesis. This is the situation where the risk modelling group correctly specifies the risk model for each desk. Therefore, we would expect around 5% rejections. However, as we have seen in the previous study, using a circular region leads to slightly undersized tests for $n = 250$ (for both $R = 250$ and $R = 500$ bootstrap repetitions), whereas applying depth-based confidence regions leads to oversized tests for both $n = 250$ and $n = 500$. This can be explained by the pictures as well. In all but one case the test value lies within the circular confidence region[31]. For the depth-based confidence region, the test value lies outside the confidence region in two cases (20%). The radius of the circle seems to be a bit too high. The depth-based confidence region is shaped very closely around the point cloud so that the test is more likely to reject the null hypothesis, especially in cases where the point cloud is less circular shaped.

Figures 6.2-6.5 show the situation under the alternative hypothesis. The setting is generally the same as before (bispectral test SP.UL with $\alpha_1 = 0.9805$ and $\alpha_2 = 0.9995$ and $d = 50$ desks) but a certain proportion (25%, 50%, 75% and 100%) of desks are t-distributed. That is the risk modelling group misspecifies the distribution of a certain proportion of desks. For 25% misspecified desks, three test values lie outside the circular and two outside the depth-based confidence region. The more desks are misspecified, the better the power performance becomes. For 75% and 100% misspecified desks, the test values lie (far) outside of both confidence regions, which explains the observed high power results found in the simulation studies describe above. In comparison, especially for 25% to 50% of misspecified desks, the circular confidence region shows more rejections than the depth-based confidence region, which is supported by the results shown in Table 6.5.

[30] We used $R = 500$ instead of the usually applied $R = 100$ because it turned out in the simulation study that for $R = 100$, the convex hull originated by the halfspace depths just goes through all outer points, so that a comparison to the circular region is difficult.

[31] In fact, for 10 cases we would expect zero to one rejected null hypothesis on average for a well-sized test of nominal level 5%.

		R	100				500			
		ρ	Zero		Flat		Zero		Flat	
fraction	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
25	250	Normal	4.7	4.7	5.6	5.3	4.2	4.6	5.8	5.6
		t4	23.8	7.0	8.3	6.0	26.2	8.0	9.4	5.0
	500	Normal	5.9	3.8	6.1	4.7	5.8	3.9	6.3	5.0
		t4	44.6	12.5	12.0	8.1	47.8	12.2	12.8	7.6
50	250	Normal	4.7	4.7	5.6	5.3	4.2	4.6	5.8	5.6
		t4	75.9	21.8	17.8	9.2	78.0	23.7	17.9	10.1
	500	Normal	5.9	3.8	6.1	4.7	5.8	3.9	6.3	5.0
		t4	95.1	40.9	30.5	17.3	96.5	42.7	32.5	18.2
75	250	Normal	4.7	4.7	5.6	5.3	4.2	4.6	5.8	5.6
		t4	96.4	43.3	31.7	16.2	97.5	43.3	33.2	16.3
	500	Normal	5.9	3.8	6.1	4.7	5.8	3.9	6.3	5.0
		t4	100.0	69.4	54.4	31.3	100.0	72.4	57.9	31.2

Table 6.1: Estimated size and power of the multi-desk Binomial score test using a bootstrap confidence interval, $d = 50$ and $\alpha = 0.99$

		test	SP.U				SP.L				SP.E			
		ρ	Zero		Flat		Zero		Flat		Zero		Flat	
fraction	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
25	250	Normal	5.6	4.6	5.6	4.9	5.0	4.7	6.3	5.1	5.6	4.6	5.6	4.9
		t4	27.8	7.5	8.3	5.5	48.8	11.1	10.5	7.1	27.9	7.5	8.3	5.5
	500	Normal	5.8	4.7	5.8	5.1	5.8	4.4	5.9	5.1	5.8	4.7	5.7	5.1
		t4	51.0	11.7	12.0	8.2	76.9	20.2	18.0	10.9	51.2	11.7	12.0	8.2
50	250	Normal	5.6	4.6	5.6	4.9	5.0	4.7	6.3	5.1	5.6	4.6	5.6	4.9
		t4	77.4	20.8	16.5	9.2	96.5	36.1	29.8	13.3	77.8	21.0	16.5	9.4
	500	Normal	5.8	4.7	5.8	5.1	5.8	4.4	5.9	5.1	5.8	4.7	5.7	5.1
		t4	96.7	39.7	28.7	15.5	99.9	62.8	50.9	27.1	96.7	39.9	28.9	15.5
75	250	Normal	5.6	4.6	5.6	4.9	5.0	4.7	6.3	5.1	5.6	4.6	5.6	4.9
		t4	97.4	40.0	29.8	15.0	99.9	61.7	51.2	24.1	97.4	40.2	29.9	15.2
	500	Normal	5.8	4.7	5.8	5.1	5.8	4.4	5.9	5.1	5.8	4.7	5.7	5.1
		t4	100.0	66.6	50.5	29.0	100.0	90.9	80.0	47.6	100.0	66.8	50.7	29.2

Table 6.2: Estimated size and power of SP.U, SP.L and SP.E using a bootstrap confidence interval, $R = 100$, $d = 50$, $\alpha_1 = 0.9805$ and $\alpha_2 = 0.9995$

	R	100				500			
	ρ	Zero		Flat		Zero		Flat	
n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
250	Normal	4.0	2.8	3.0	1.8	3.3	1.8	1.8	1.1
	t4	100.0	99.8	98.9	63.5	100.0	99.8	98.9	59.9
500	Normal	5.0	5.2	5.4	6.3	3.6	4.5	5.0	5.6
	t4	100.0	100.0	100.0	98.9	100.0	100.0	100.0	99.4

Table 6.3: Estimated size and power of SP.UL with circular confidence region for 100% misspecified desks, $d = 50$, $\alpha_1 = 0.9805$ and $\alpha_2 = 0.9995$

	R	100				500			
	ρ	Zero		Flat		Zero		Flat	
n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
250	Normal	7.7	10.6	14.1	17.6	5.0	8.5	11.9	16.5
	t4	100.0	95.8	91.5	45.1	100.0	97.4	94.2	38.1
500	Normal	8.0	11.0	12.1	13.7	5.4	8.3	9.4	12.5
	t4	100.0	99.9	99.8	91.9	100.0	100.0	100.0	96.0

Table 6.4: Estimated size and power of SP.UL with depth-based confidence region for 100% misspecified desks, $d = 50$, $\alpha_1 = 0.9805$ and $\alpha_2 = 0.9995$

			circular				depth-based			
		ρ	Zero		Flat		Zero		Flat	
fraction	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
25	250	Normal	4.0	2.8	3.0	1.8	7.7	10.6	14.1	17.6
		t4	63.7	15.4	29.0	2.2	55.5	10.6	20.3	7.7
	500	Normal	5.0	5.2	5.4	6.3	8.0	11.0	12.1	13.7
		t4	94.2	51.8	65.7	11.2	91.0	40.3	55.0	12.1
50	250	Normal	4.0	2.8	3.0	1.8	7.7	10.6	14.1	17.6
		t4	99.9	77.9	84.9	18.5	99.2	54.4	65.4	16.5
	500	Normal	5.0	5.2	5.4	6.3	8.0	11.0	12.1	13.7
		t4	100.0	99.5	99.4	59.6	100.0	94.7	95.3	49.1
75	250	Normal	4.0	2.8	3.0	1.8	7.7	10.6	14.1	17.6
		t4	100.0	97.7	96.3	43.3	100.0	85.8	84.7	31.3
	500	Normal	5.0	5.2	5.4	6.3	8.0	11.0	12.1	13.7
		t4	100.0	100.0	100.0	91.2	100.0	99.3	99.0	79.6

Table 6.5: Estimated size and power of SP.UL with circular and depth-based bootstrap confidence region for $R = 100$, 100% misspecified desks, $d = 50$, $\alpha_1 = 0.9805$ and $\alpha_2 = 0.9995$

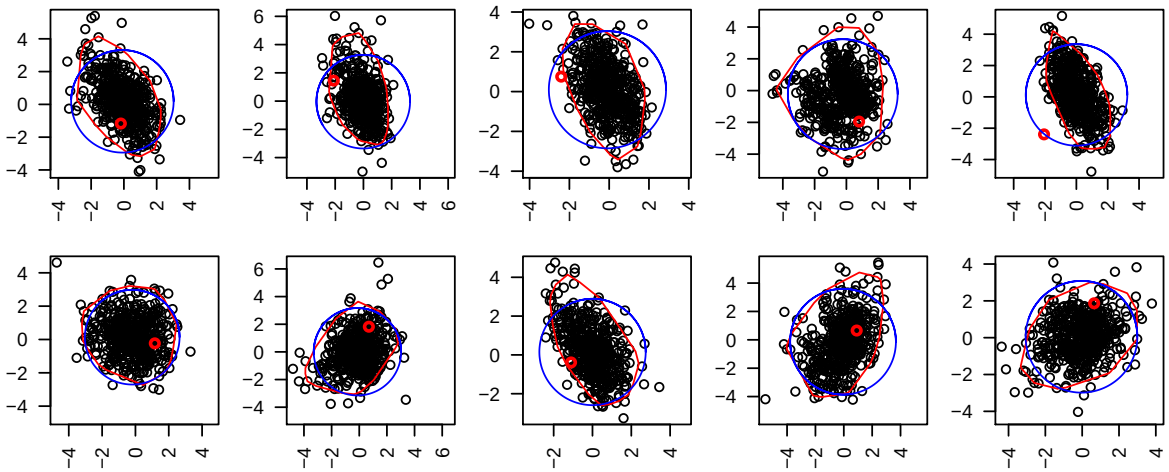

Figure 6.1: Null hypothesis, $R = 500$; circular (blue) and depth-based (red). The test value is highlighted in red

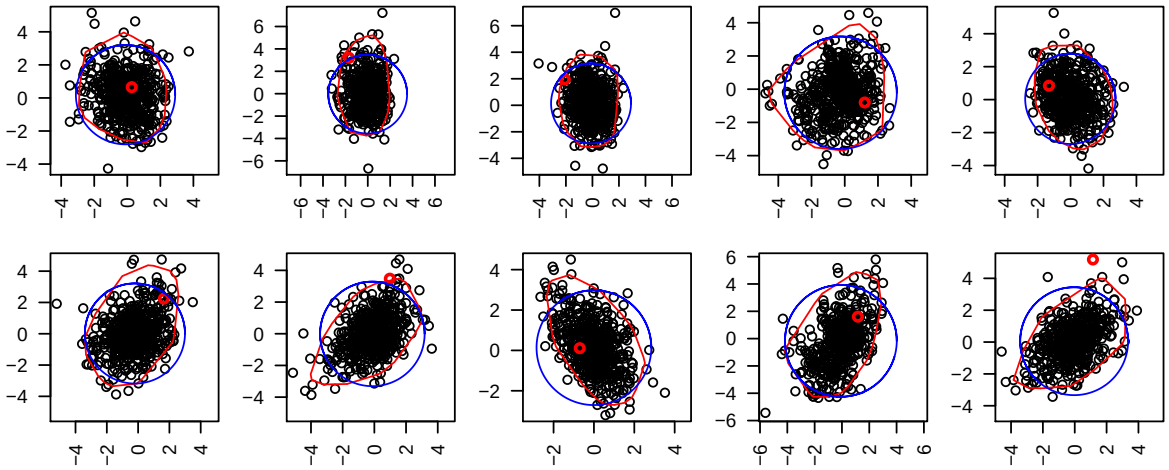

Figure 6.2: Alternative hypothesis with fraction 25%, $R = 500$; circular (blue) and depth-based (red). The test value is highlighted in red

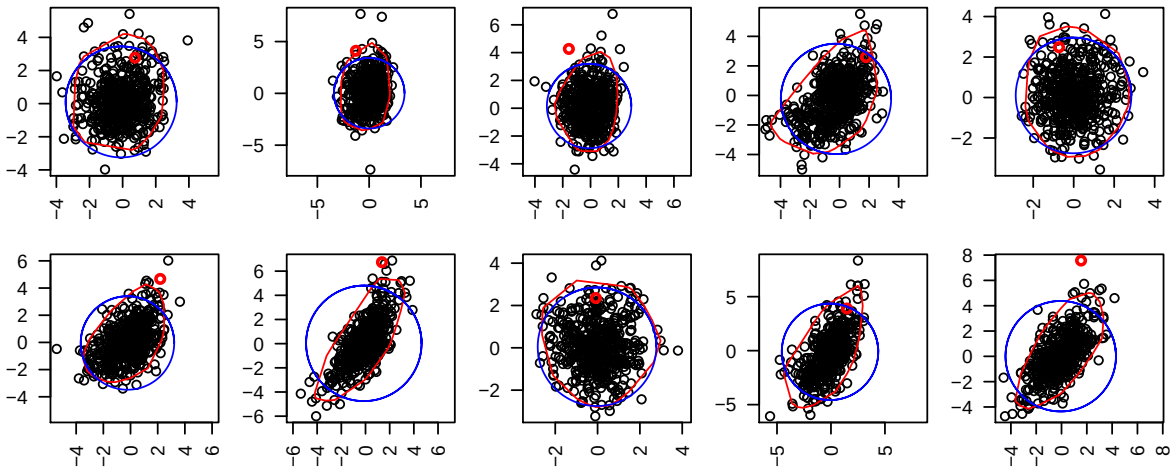

Figure 6.3: Alternative hypothesis with fraction 50%, $R = 500$; circular (blue) and depth-based (red). The test value is highlighted in red

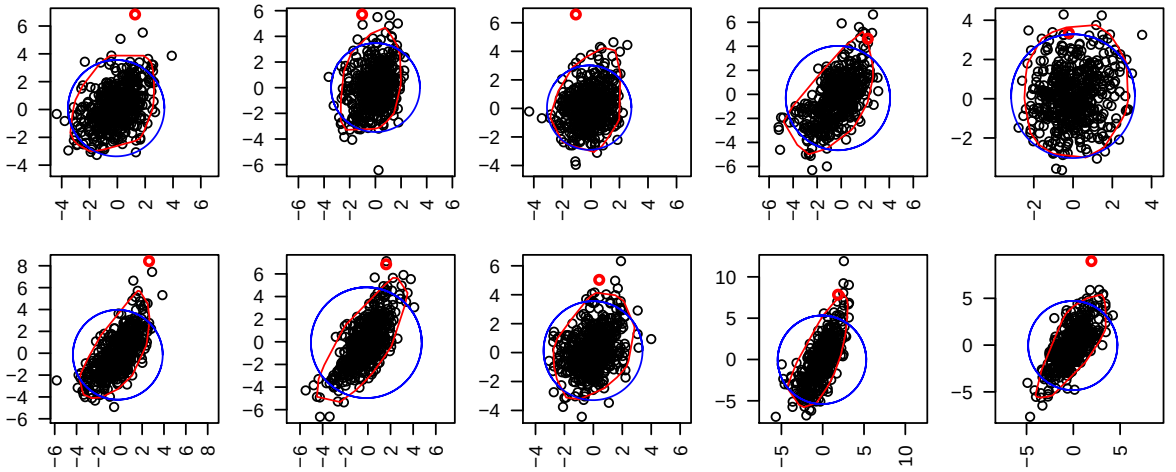

Figure 6.4: Alternative hypothesis with fraction 75%, $R = 500$; circular (blue) and depth-based (red). The test value is highlighted in red

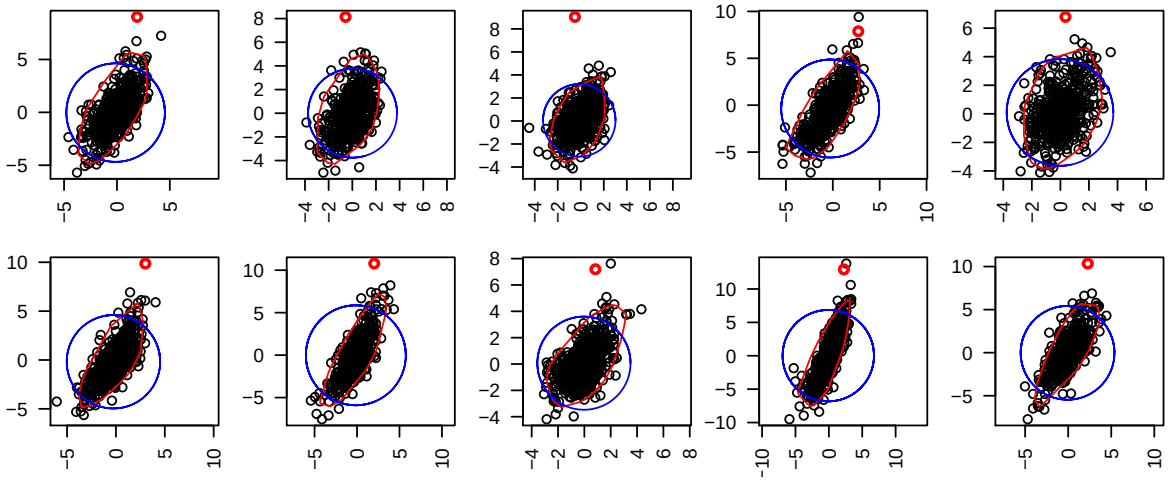

Figure 6.5: Alternative hypothesis with fraction 100%, $R = 500$; circular (blue) and depth-based (red). The test value is highlighted in red

6.7 Summary

The main conclusions of this section are:

- We have explored bootstrap options both for the new multi-desk monospectral test and the multi-desk bispectral test to derive the test statistic's distribution and analysed the possible benefits from not using the asymptotic normality assumption.
- For the multi-desk monospectral desk we considered the Studentized-t bootstrap confidence interval which is dual to our hypothesis. The simulation study did not indicate significant differences in the general performance compared to using the normality assumption. However, the power of the Z-test is higher than that of the bootstrap test. This means that using the normality assumption, we have a better chance to detect wrong models (under the given simulation parameters, especially 100 bootstrap repetitions).
- For the multi-desk bispectral test, we analysed two alternatives to build confidence regions for the mean vector. Firstly, we considered an intuitive circular confidence region (with centre at the multivariate trimmed mean of Tukey). Secondly, we analysed half-space depth sets (based on Tukeys depth) to build a confidence region more adapted to the observed data.
- In an extreme case where all desks have been misspecified or correctly specified by the risk modelling group, the circular confidence region results in power values around 100%

but values for size that tend to be on the low side. The depth-based confidence region is heavily oversized in this situation.

- In a more realistic scenario where a certain proportion of desks is misspecified, power results deteriorate for the circular confidence region and are mediocre for 25% misspecified desks. For the depth-based confidence region power results are good for 75% misspecified desks but then also quickly deteriorate for lower fractions.

- Analysing the situation graphically it seems that the circular confidence region tends to be too large. As can be seen, the test value lies more often than expected within the confidence region when the null hypothesis is true. On the other hand, the depth-based confidence region is too adapted to the data such that the test value lies outside the confidence region too often. When the alternative hypothesis is true and a high ratio of desks is misspecified, the graphical analysis showed that the test value lies far beyond both confidence regions, which explains the observed high power values.

- Overall, using bootstrapping methods instead of the normality assumption does not seem to come with many benefits considering that bootstrapping methods are very computationally intensive. However, for the bispectral test, a circular bootstrap confidence region can be the better choice compared to using the normality assumption when observations are below 1000 days.

Chapter 7

Comparative tests across desks

7.1 Introduction

The previous two chapters have covered research question 2 which examined whether trading risk was adequately modelled in aggregate across all desks. In this chapter, we concentrate on comparative tests, which assess whether the internal models used to measure the market risk are of similar quality for all desks. That is we address our third research question:

Question 3. ***Are the trading desks equally good at modelling risk?***

This research question can be addressed in the framework of the well-known Cochran Q-test (Cochran (1950)). Cochran presented a test for significant differences between outcomes for correlated binary data, violating the assumption of the standard χ^2-test which requires independent samples. He showed that his proposed test statistic follows a χ^2-distribution asymptotically when certain conditions are met. The test is a generalisation of one proposed in McNemar (1969) to more than two samples. Cochran Q-test is a non-parametric statistical test, which means that no distributional assumption for the underlying data is required.

A typical application for Cochran Q-test is testing for significant differences in the proportions between matched samples in a before and after scenario, for example examination results (passed/not passed) before and after a lecture. Other test problems concern clinical studies where a patient is subject to different treatments and the relative effectiveness of treatments is analysed[32]. In this kind of testing problem, each patient is given each of the treatments

[32]This is also one of the examples given in Cochran (1950).

under consideration at different times and the response to the treatments is measured. Each patient represents one correlated (matched) sample. The null hypothesis is that all treatments are equally effective (or equally ineffective).

Research question 3 fits into this kind of testing framework since we expect all trading desks to show the same proportion of violations if they are equally good at modelling risk. The matched sample consists of the results for all desks on one particular day. Since Cochran Q-test requires binary data as input variables, we use the VaR violation indicators $Y_{t,i} = I_{\{L_{t,i} > \widehat{VaR}_{t,i}(\alpha)\}}$. Recall that this variable is equivalent to a PIT-value level exceedance indicator (see the relation in (5.2)). That is for each day we have a vector of length d, $\boldsymbol{Y}_t = (Y_{t,1}, \ldots, Y_{t,d})$ consisting of the observations for the d desks. Each of these vectors represents a matched correlated sample of binary observations. We can then test for significant differences in the proportions of VaR violations between the d desk with Cochran Q-test.

However, this standard application of Cochran Q-test can only provide an indication if all desks perform equally (equally good *or* equally bad). In this chapter, we therefore also develop a new extension of Cochran Q-test which tests for the unconditional coverage hypothesis (see Definition 15) at the same time. This test not only assesses whether all desks show an equal proportion of violations, but also tests whether this proportion is in line with the expected value when all desks are modelling risk effectively.

The remaining of this chapter is organised as follows: We first provide the necessary mathematical background. We then describe the general methodology, including the input data and the formal definition of the null hypotheses of interest. This is followed by the presentation of the standard version of Cochran Q-test. Since this theory requires that the underlying data are exchangeable, we present a generalisation of the theory of Cochran Q-test for non-exchangeable data. After that we present the new test which extends Cochran Q-test to a joint test of equality across desks and unconditional coverage. For this, we exploit the quadratic form representation of the test because there is no simple closed-form expression for it's distribution. The chapter closes with a simulation study, specifically comparing the performance of the two presented tests.

7.2 Mathematical foundations

The only new mathematical theory required is a general expression for sums of squares of Gaussian random vectors and the definition of exchangeable random variables.

7.2.1 Exchangeability

For an exchangeable random vector it holds that it's distribution function does not change under any permutation.

Definition 24 (Exchangeability)**.** *A d-dimensional random vector* $\boldsymbol{X}$ *is exchangeable if*

$$(X_1, \ldots, X_d) \stackrel{d}{=} \left(X_{\Pi(1)}, \ldots, X_{\Pi(d)}\right) \tag{7.1}$$

for any permutation $(\Pi(1), \ldots, \Pi(d))$ *of* $(1, \ldots, d)$*.*

See McNeil et al. (2015), p.234, Definition 7.16. It implies that all variables X_i, $i = 1, \ldots, d$, have the same variance and that the variables are equicorrelated.

7.2.2 Walsh's theorem

In this section we present a well-known theorem of Walsh (1947), see Theorem 1, p.94, which is used by Cochran to derive the limit distribution of his test statistic.

Theorem 5. *A vector of d random variables* $X_1, \ldots, X_d$ *with multivariate normal distribution with common variance* σ_X^2 *and common covariances* $\rho_X \sigma_X^2$ *satisfies*

$$\sum_{i=1}^{d} \left(X_i - \bar{X}\right)^2 \sim \sigma_X^2 (1 - \rho_X) \chi^2 \left(d - 1\right), \tag{7.2}$$

where $\bar{X} = 1/d \sum_{i=1}^{d} X_i$*.*

Note that this Theorem includes the special case of exchangeable random vectors $\boldsymbol{X} = (X_1, \ldots, X_d)$ but is more general since the expected values of $X_1, \ldots, X_d$ can be different.

7.2.3 Imhof's method

The next proposition shows how the sum of squared Gaussian random variables can be expressed as a linear combination of independent non-central χ^2-distributed variables.

Proposition 4. *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be a random vector which follows a multivariate normal*

distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ *and non-singular covariance matrix* $\Sigma \in \mathbb{R}^{n \times n}$. *Let* $\boldsymbol{c} = (c_1, \ldots, c_n)$ *be some constant vector and let* $A \in \mathbb{R}^{n \times n}$ *be some constant matrix. The quadratic form* $Q = (\boldsymbol{X} + \boldsymbol{c})' \boldsymbol{A} (\boldsymbol{X} + \boldsymbol{c})$, *has the representation*

$$Q = \sum_{r=1}^{m} \lambda_r Y_r, \tag{7.3}$$

where m *is the rank of* $A\Sigma$, λ_r *are the non-zero eigenvalues of* $\boldsymbol{A}\Sigma$ *and* Y_r, $r = 1, \ldots, m$, *are independent* χ^2*-distributed variables with* h_r *degrees of freedom and non-centrality parameter* δ_r. *The* h_r *are the respective orders of multiplicity of the eigenvalues of* $\boldsymbol{A}\Sigma$.

For the derivation, see Imhof (1961), p.419 and references therein. The variables Y_r in (7.3) can be represented as $Y_r = (Z_1 + \delta_r)^2 + \sum_{i=2}^{h_r} Z_i^2$, where $Z_1, \ldots, Z_h$ are independent unit normal variables. Note that for $\delta_r = 0$ this represents the usual χ^2- distribution for Y_r.

It is well-known that the distribution function F_Q of a positively weighted sum of i.i.d. χ^2-distributed variables Q has no closed-form (Bodenham and Adams (2016)). However, in applications one is often interested in $\mathbb{P}(Q > x)$ for some given value x. There are some approximations for this distribution available. Bodenham and Adams (2016), p.917, compare the computational speed and accuracy of some of the proposed methods. They point out that the method proposed by Imhof (1961) should be the preferred choice when computational resources are sufficient. Imhof's method provides a numerical solution for $F_Q(x) = \mathbb{P}(Q < x)$ for fixed x with a desired accuracy by numerically inverting the characteristic function of Q. The numerical solution of Imhof's method is implemented in the `R` package `CompQuadForm` (Duchesne and de Micheaux (2010)).

7.3 Methodology

7.3.1 Data

As binary input data we use the VaR violation indicators, $Y_{t,i} = I_{\{L_{t,i} > \widehat{VaR}_{t,i}(\alpha)\}}$. The vector $\boldsymbol{Y}_t = (Y_{t,1}, \ldots, Y_{t,d})$ contains the data of the d desks at time t. We group the indicator variables $Y_{t,i}$ into a $n \times d$ table in which the columns correspond to the d desks and the rows to the n observed days. Each day then represents a matched correlated sample of binary observations. We can then test for significant differences in the proportions of VaR violations between the d desks. Table 7.1 shows the ordering of the data.

t \ i	desk 1	desk 2	desk 3	...	desk d
day 1	$Y_{1,1}$	$Y_{1,2}$	$Y_{1,3}$	...	$Y_{1,d}$
day 2	$Y_{2,1}$	$Y_{2,2}$	$Y_{2,3}$	...	$Y_{2,d}$
⋮	⋮	⋮	⋮	⋮	⋮
day n	$Y_{n,1}$	$Y_{n,2}$	$Y_{n,3}$	...	$Y_{n,d}$
colum totals	Y_1	Y_2	Y_3	...	Y_d

Table 7.1: Stylized sample of the data structure of the VaR violation indicators

In the table, Y_i represents the column totals $Y_i = \sum_{t=1}^{n} Y_{t,i}$. Note that correlation between desks is explicitly allowed.

7.3.2 Hypotheses

In this section, we will formulate the hypotheses corresponding to our research question 3. We want to assess whether all desk models are equally good at modelling market risk. We deem the desk models to perform equally when the numbers of observed VaR violations show no significant differences across desks. To test the assertion of research question 3 we can therefore test

$$\begin{aligned} H_0 &: \mathbb{P}(Y_{t,1} = 1) = ... = \mathbb{P}(Y_{t,d} = 1), \, \forall \, t \\ H_1 &: \mathbb{P}(Y_{t,i} = 1) \neq \mathbb{P}(Y_{t,j} = 1) \text{ for at least one } i \neq j \end{aligned} \tag{7.4}$$

Under the null hypothesis, we expect an equal amount of VaR violations in each desk. If the null hypothesis is rejected, there is at least one desk with average violation rate which differs significantly from the others. This would be an indication for a misspecified desk model. Note that the interpretation is not restricted to (technical) misspecifications in the internal desk models. The reason for a rejection of the null hypothesis can also come from an increased riskiness of one or more trading desks (for example due to economic circumstances) or because traders do not respect their VaR trading desk limits.

With this hypothesis, we do not test for the correct coverage rate in each desk but merely if there is a significant difference in the observed violation rates across desks (regardless of whether there are significantly more or less violations than expected). However, we will also develop a

test which extends the null hypothesis to

$$H_0 : \mathbb{P}(Y_{t,1} = 1) = ... = \mathbb{P}(Y_{t,d} = 1) = 1 - \alpha, \quad \forall\, t$$
$$H_1 : \mathbb{P}(Y_{t,i} = 1) \neq 1 - \alpha \quad \text{for at least one desk } i. \tag{7.5}$$

This hypothesis includes the unconditional coverage hypothesis (see Definition 15), which is tested for all desks at once. Note that in the case of a rejection of the null hypothesis, it is unknown which desk triggered this decision. A violation of the hypothesis should therefore ideally be followed by post hoc analyses to find the root of the significant test result.

7.4 Cochran Q-test

In this section, we present more details on the theory behind the derivation of Cochran Q-test as given in Cochran (1950). This test is used to test the hypothesis stated in (7.4). Cochran Q-test is based on the idea that the random variables forming the rows (the desk data on some day t) are exchangeable (see Definition 24), which is a bit restrictive in our application. It implies that every pair of desks is equicorrelated. Therefore, we also derive a generalisation of Cochran Q-test under non-exchangeable data. Finally, we state an extension of Cochran Q-test to a joint test of equality across desks and the unconditional coverage hypothesis. This test is used to test the hypothesis stated in (7.5).

Note that for deriving the results we assume throughout that the desks data $Y_{t,i}$, $t = 1, ..., n$, are independent across time for each desk d.

7.4.1 Methodology Cochran Q-test

The Cochran Q-test formally applies the test criterion

$$\sum_{i=1}^{d} \left(Y_i - \bar{Y}\right)^2,$$

where $Y_i = \sum_{t=1}^{n} Y_{t,i}$ describes the sum of VaR violations in desk i over the observation period of n days (the column totals) and $\bar{Y} = 1/d \sum_{i=1}^{d} \sum_{t=1}^{n} Y_{t,i}$ is the average number of VaR violations across desks. To derive the limit distribution of the test, Cochran uses Theorem 5 stated in Section 7.2.2.

An exchangeable Gaussian random vector with $\mathbb{E}(X_i) = \mu$ for all $i = 1, \ldots, d$ satisfies (7.2)

where σ_X^2 is the common variance of all variables and ρ_X is the common correlation for all pairs. In our application this would mean that VaR violation indicators across desks on day t, $Y_{t,i}$, $i = 1, \ldots, d$, must be exchangeable in order to apply Theorem 5. Although we assume that the VaR violation indicators across desks have the same variance on day t under the null hypothesis, the equicorrelation assumption for VaR violation indicators across desks seems unlikely in our framework. We will later argue that Cochran Q-test is fairly robust against a relaxation of this assumption (see Section 7.4.2).

Cochran derives his test statistic by analysing the conditional distribution of the data given the row totals $U_t = \sum_{i=1}^{d} Y_{t,i}$ are fixed at the observed values u_t. He then exploits that when the row total of violations is fixed, that the row total has zero variance. For exchangeable data, every desk is then equally likely to show one of the u_t violations on day t. That is the probability to observe a violation in some desk i at day t given $U_t = u_t$ is $p_t = u_t/d$. We define the vector $\boldsymbol{U} = (U_1, \ldots, U_n)$, which consists of the n row totals and the vector $\boldsymbol{u} = (u_1, \ldots, u_n)$.

We have to assume that the following terms converge to some limit value when n goes to infinity.

(i) $\bar{p}(n) = n^{-1} \sum_{t=1}^{n} p_t$,
(ii) $\bar{\sigma}^2(n) = n^{-1} \sum_{t=1}^{n} p_t(1 - p_t)$.

It is for each single row t, $t = 1, \ldots, n$

$$\begin{aligned}\operatorname{var}\left(\sum_{i=1}^{d} Y_{t,i} \middle| U_t = u_t\right) &= \sum_{i=1}^{d} \operatorname{var}\left(Y_{t,i} \middle| U_t = u_t\right) + \sum_{i=1}^{d} \sum_{j \neq i} \operatorname{cov}\left(Y_{t,i}, Y_{t,j} \middle| U_t = u_t\right) \\ &= dp_t(1 - p_t) + d(d-1)p_t(1 - p_t) \operatorname{cor}\left(Y_{t,i}, Y_{t,k} \middle| U_t = u_t\right) = 0,\end{aligned}$$

from which directly follows that $\operatorname{cor}\left(Y_{t,i}, Y_{t,k} \middle| U_t = u_t\right) = -1/(d-1)$.

We define the variable $Z_i = n^{-1} \sum_{t=1}^{n} Y_{t,i}$, which presents the average column total of column i. It is $\mathbb{E}(Z_i \big| \boldsymbol{U} = \boldsymbol{u}) = n^{-1} \sum_{t=1}^{n} p_t = \bar{p}(n)$. For the conditional variance of Z_i we have $\operatorname{var}(Z_i | \boldsymbol{U} = \boldsymbol{u}) = n^{-2} \sum_{t=1}^{n} p_t(1 - p_t) = n^{-1} \bar{\sigma}^2(n)$. For the calculation of both conditional moments, we used the assumption that rows are independent.

It directly follows

$$\operatorname{cov}(Z_i, Z_k | \boldsymbol{U} = \boldsymbol{u}) = \frac{1}{n^2} \sum_{t=1}^{n} \sum_{s=1}^{n} \operatorname{cov}(Y_{t,i}, Y_{s,k} \big| \boldsymbol{U} = \boldsymbol{u}) = -\frac{1}{n^2} \sum_{t=1}^{n} \frac{p_t(1 - p_t)}{d-1} = -\frac{1}{n(d-1)} \bar{\sigma}^2(n),$$

where we exploited the fact that $Y_{t,i}$, $i = 1, \ldots, d$, are exchangeable.

In vector notation, if $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$ then $\mathbb{E}(\boldsymbol{Z} \big| \boldsymbol{U} = \boldsymbol{u}) = \bar{p}(n)\mathbf{1}$ and $\operatorname{cov}(\boldsymbol{Z} \big| \boldsymbol{U} = \boldsymbol{u}) =$

$n^{-1}\bar{\sigma}^2(n)P$ with matrix

$$P = \begin{pmatrix} 1 & -1/(d-1) & \dots & -1/(d-1) \\ -1/(d-1) & 1 & \dots & -1/(d-1) \\ \vdots & \vdots & \vdots & \vdots \\ -1/(d-1) & -1/(d-1) & \dots & 1 \end{pmatrix}.$$

Then we can apply the Central Limit Theorem conditional on $\boldsymbol{U}$ to see that

$$\frac{\sqrt{n}\,(\boldsymbol{Z} - \bar{p}\mathbf{1})}{\bar{\sigma}} \dot{\sim} \mathcal{N}_d\,(\mathbf{0}, P)$$

as $n \to \infty$ using assumption (i) and (ii). We can now apply Theorem 5 with the random variables $X_1, \dots, X_d$ taken to be $X_i = \sqrt{n}(Z_i - \bar{p})$ since they are approximately normally distributed for large n. Then we have that

$$\sum_{i=1}^{d} n\,(Z_i - \bar{p})^2 \dot{\sim} \bar{\sigma}^2 \frac{d}{d-1}\chi^2_{d-1}$$

is asymptotically χ^2-distributed with $d-1$ degrees of freedom if $n \to \infty$. Simple conversion gives

$$\sum_{i=1}^{d} \frac{\left(Y_i - \bar{Y}\right)^2}{n} \dot{\sim} \sum_{t=1}^{n} \frac{p_t(1-p_t)}{n} \frac{d}{d-1}\chi^2_{d-1}.$$

Summarised we have

$$T_{QC} = \frac{d-1}{d} \frac{\sum_{i=1}^{d}\left(Y_i - \bar{Y}\right)^2}{\sum_{t=1}^{n} p_t(1-p_t)} \dot{\sim} \chi^2(d-1), \tag{7.6}$$

which is the Cochran Q-test statistic.

7.4.2 Generalisation of Cochran Q-test to non-exchangeable data

As mentioned above, the asymptotic theory of Cochran Q-test is based on the assumption that the VaR violation indicators across desks on day t, $Y_{t,i}$, $i = 1, \dots, d$, are equicorrelated, which is a strong assumption. In this section, we generalise this theory to non-exchangeable data relaxing the assumption of equicorrelation.

Note that in the standardised test statistic in (7.6), the expression $\sum_{i=1}^{d}\left(Y_i - \bar{Y}\right)^2$ is 'weighted'

with the factor

$$\frac{d-1}{d\sum_{t=1}^{n} p_t(1-p_t)}.$$

This has been derived under the assumption that the vector of random variables $Y_1, \ldots, Y_d$ is (approximately) multivariate normal distributed with common correlation ρ_Y (see Section 7.4.1). We show how this changes for non-exchangeable data making use of Proposition 4 stated in Section 7.2.3. That is we consider the situation where we allow for different pairwise correlations $\rho_{Y_i,Y_k} = \text{cor}(Y_i, Y_k)$:

$$P = \begin{pmatrix} 1 & \rho_{Y_1,Y_2} & \ldots & \rho_{Y_1,Y_d} \\ \rho_{Y_2,Y_1} & 1 & \ldots & \rho_{Y_2,Y_d} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{Y_d,Y_1} & \rho_{Y_d,Y_2} & \ldots & 1 \end{pmatrix},$$

where P is the correlation matrix of the vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$. Under this setting the following can be shown for the test criterion $\sum_{i=1}^{d} \left(Y_i - \bar{Y}\right)^2$ of Cochran Q-test:

Proposition 5. *Let $Y_{1,i}, \ldots, Y_{n,i}$, $i = 1, ..., d$, be independent Bernoulli distributed random variables with parameter $1-\alpha$. Let $Y_i = \sum_{t=1}^{n} Y_{t,i}$, $i = 1, \ldots, d$ and $\bar{Y} = d^{-1}\sum_{i=1}^{d}\sum_{t=1}^{n} Y_{t,i}$. Let $A = I_d - d^{-1}J_d$, where I_d is the identity matrix and J_d the all-ones matrix. Then*

$$\sum_{i=1}^{d} \left(Y_i - \bar{Y}\right)^2 \dot{\sim} \sum_{i=1}^{d} \lambda_i^* Z_i^2,$$

where Z_i are i.i.d. standard normal distributed random variables and λ_i^* are the eigenvalues of $n\alpha(1-\alpha)APA$ with correlation matrix $P \in \mathbb{R}^{d\times d}$ given by $P_{i,k} = \text{cor}(Y_i, Y_k)$ for all i, k.

Proof. We consider the expression

$$V_i = \frac{1}{\sqrt{n}} \left(Y_i - \bar{Y}\right), i = 1, \ldots, d. \tag{7.7}$$

Note that $\left(Y_i - \bar{Y}\right)$ in (7.7) is part of the test criterion of Cochran Q-test, which is multiplied with $n^{-1/2}$ to get V_i. We have $\boldsymbol{V} = (V_1, \ldots, V_d) = n^{-1/2}A\boldsymbol{Y}$ with $A = I_d - d^{-1}J_d$. For the covariance matrix it holds that $\text{Var}(\boldsymbol{V}) = n^{-1}A\text{Var}(\boldsymbol{Y})A$. Since $\text{var}(Y_i) = n\alpha(1-\alpha)$ it follows

that $\text{Var}(\boldsymbol{V}) = \alpha(1-\alpha)APA$. In the limit $n \to \infty$ we have

$$\boldsymbol{V} \to \mathcal{N}\left(\mathbf{0}, \alpha(1-\alpha)APA\right).$$

We now consider $\boldsymbol{V}\boldsymbol{V}' = n^{-1}\sum_{i=1}^{d}\left(Y_i - \bar{Y}\right)^2$. In the limit, as $n \to \infty$, this quadratic form is distributed as

$$\sum_{i=1}^{d} \lambda_i Z_i^2,$$

where Z_i are the i.i.d. standard normal distributed random variables and λ_i are the eigenvalues of $\alpha(1-\alpha)APA$ (see (7.3)). Using again central limit arguments we can deduce that the distribution of $Y_i = \sum_{t=1}^{n} Y_{t,i}$, $i = 1, \ldots, d$, can be approximated by the normal distribution for large n. Equivalently, we have for $n \to \infty$ that

$$n\boldsymbol{V}\boldsymbol{V}' = \sum_{i=1}^{d}\left(Y_i - \bar{Y}\right)^2 \dot{\sim} \sum_{i=1}^{d} \lambda_i^* Z_i^2, \tag{7.8}$$

where λ_i^* are the eigenvalues of $n\alpha(1-\alpha)APA$. □

To calculate the p-value $\mathbb{P}(Q > x)$ with $Q = \sum_{i=1}^{d}\left(Y_i - \bar{Y}\right)^2$ one can use the Imhof's method (see Proposition 4). We performed a small simulation with equicorrelated data (that is $\rho_{Y_i,Y_k} = \rho$ for all i, k) and compared the p-values of the test T_{QC} in (7.6) and the p-values using the Imhof's method. The p-values were quite similar. We then considered a situation with differing pairwise correlations. It turned out that in this situation the p-values were quite similar as well. Thus, the relaxation of the equicorrelation assumption should have a minor effect on the asymptotic distribution.

7.4.3 Extended Cochran Q-test to test for unconditional coverage

As already mentioned, Cochran Q-test has power against the null hypothesis that all desks are modelled equally well or badly. The test has no power to detect 'badly' modelled desks where the sum of violations observed in a desk differs from the expected number of violations. However, regulators are interested in testing whether one or more desk's VaR violation rates exceed the expected ratio, possibly leading to an underestimation of the capital set aside for market risk.

In the following, we consider the test statistic of the Cochran Q-test and extend the approach to a joint test of equality across desks and unconditional coverage, meaning that we test the null

hypothesis that the desk's VaR violation rates are consistent with their theoretical expectation under the null hypothesis (see hypothesis in (7.5)). That is we want to test whether the column totals Y_i differ from the expected value of Y_i, which is $\mu_Y = n(1-\alpha)$ under the null hypothesis. The applied test criterion of Cochran Q-test then changes to

$$T_{QE} = \sum_{i=1}^{d} (Y_i - \mu_Y)^2 . \tag{7.9}$$

The limit distribution of this extended version of the test statistic of Cochran Q-test does not follow a χ^2-distribution. This is easy to see when we decompose the test statistic in the following way

$$\sum_{i=1}^{d} (Y_i - \mu_Y)^2 = \sum_{i=1}^{d} \left(Y_i - \bar{Y}\right)^2 + d\left(\bar{Y} - \mu_Y\right)^2 . \tag{7.10}$$

We know from Walsh's Theorem (see Theorem 5 stated in Section 7.2.2) for the more restricting assumption of exchangeable data that the first summand asymptotically follows a $\chi^2(d-1)$ distribution scaled by the common covariance ρ_Y of the random variables $Y_1, \ldots, Y_d$, which have common variance $\sigma_Y^2 = np(1-p)$ under the hypothesis in (7.5). However, the second term in (7.10) has a scalar multiple of a $\chi^2(1)$ distribution in the limit. Since the scaling factors differ, it is not possible to simply add the distributions of the first and second term together. To find the limiting distribution of (7.9), we will express the test statistic as a linear combination of independent χ^2-distributed variables and then exploit the Imhof's method (see Section 7.2.3) to find the required quantiles of the distribution.

Proposition 6. *Let $Y_{1,i}, \ldots, Y_{n,i}$, $i = 1, \ldots, d$, be independent Bernoulli distributed random variables with parameter $1-\alpha$ and invertible variance-covariance matrix Σ. Let $Y_i = \sum_{t=1}^{n} Y_{t,i}$, $i = 1, \ldots, d$, where $\mu_Y = n(1-\alpha)$ is the mean of Y_i for all i. Furthermore, let $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_d)$ with $\widetilde{Y}_i = Y_i - \mu_Y$. Then as $n \to \infty$*

$$\sum_{i=1}^{d} (Y_i - \mu_Y)^2 \dot{\sim} \sum_{i=1}^{d} \lambda_i U_i^2,$$

where $U_1, \ldots, U_d$ are i.i.d. standard normal variables and λ_i are the eigenvalues of Σ.

Proof. Let $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_d)$ with $\widetilde{Y}_i = Y_i - \mu_Y$. The sum $\sum_{i=1}^{d} \widetilde{Y}_i^2$ can be written in the

quadratic form $Q = \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}$. Defining $\boldsymbol{Z} = \Sigma^{-1/2}\widetilde{\boldsymbol{Y}}$ we have

$$Q = \boldsymbol{Z}'\Sigma^{1/2}\Sigma^{1/2}\boldsymbol{Z} = \boldsymbol{Z}'\Sigma\boldsymbol{Z}. \tag{7.11}$$

It is well-known by the spectral theorem that for a normal matrix like Σ (since Σ is symmetric it holds $\Sigma'\Sigma = \Sigma\Sigma'$), there exists an orthogonal matrix V, i.e. $V'V = VV' = I$, such that $\Sigma = V'DV$, where D is a diagonal matrix with the eigenvalues of matrix Σ as diagonal elements. The matrix V has the eigenvectors of Σ as rows. We can then compute

$$Q = \boldsymbol{Z}'\Sigma\boldsymbol{Z} = \boldsymbol{Z}'V'DV\boldsymbol{Z} = (V\boldsymbol{Z})'\,DV\boldsymbol{Z}. \tag{7.12}$$

It directly follows

$$Q = (V\boldsymbol{Z})'\,D\,(V\boldsymbol{Z}) = \sum_{i=1}^{d}\lambda_i U_i^2, \tag{7.13}$$

where $\boldsymbol{U} = (U_1, \ldots, U_d)$ is a multivariate normal distributed random vector with zero mean and identity covariance matrix and λ_i are the eigenvalues of the variance-covariance matrix Σ. For the last equation we used (7.3). Since $\widetilde{\boldsymbol{Y}}$ follows approximately for large n a multinormal distribution with zero mean and covariance matrix Σ (by central limit arguments), the stochastic presentation (7.13) holds true in the limit as $n \to \infty$. □

To derive the distribution of (7.11) numerically, we use the R package `CompQuadForm` (Duchesne and de Micheaux (2010)), which computes the distribution function of a quadratic form of a multivariate normal random vector with a regular covariance matrix.

More precisely, we use Imhof's method. For a multivariate normal random vector $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)$ with covariance matrix Σ, Imhof's method computes the quantile $\mathbb{P}(Q > q)$ for a quadratic form $Q = \widetilde{\boldsymbol{Y}}'A\widetilde{\boldsymbol{Y}}$, such as the quadratic form in (7.13). In our application, we estimate the unknown variance-covariance matrix Σ by $\widehat{\Sigma} = n\alpha(1-\alpha)\mathcal{R}$, where $\mathcal{R}$ is the empirical correlation matrix. Note that, since we apply a consistent estimate for the correlation matrix, the eigenvalues of Σ converge to the true eigenvalues.

7.5 Results simulation study: comparative tests across desks

In this section, we present the results of our simulation study, which analyses the size and power properties of the two comparative tests we presented:

- Cochran Q-test for equal VaR violation rates across desks as in Cochran (1937) (T_{QC} in (7.6), called QC in the simulation study)
- An extended version of the Cochran test which also tests for unconditional coverage (T_{QE} in (7.9), called QE in the simulation study)

The general simulation setup is presented in Section 4.3.3. We illustrate how size and power are influenced by the different simulation parameters.

7.5.1 Impact of days without VaR violations

When it comes to market risk modelling, typical sample sizes used to validate the model are 250 or 500 trading days (approximately one, two years respectively). In fact, the regulatory backtesting requirements mandate a period of one year when counting the number of VaR violations, which can trigger supervisory actions (BCBS (2019b), p.81, para. 32.4 and p.82, Table 1). Combined with the required high confidence levels around 99% a rather low number of observed violations can be expected, even for misspecified models.

The comparative tests we consider in this chapter could potentially be affected by this since days without violations do not add any additional information. This can influence the test's power. Both versions of Cochran Q-test use binary data as an input variable. As Cochran pointed out (Cochran (1950), p.258), the value of Cochran Q-test is not affected by rows consisting of only zeros (meaning that no violations are observed on a specific day across the desks) or only ones (meaning that all desks show a violation). Since these rows add no information, the effective sample size can be significantly lower than assumed (having a large number of zero-rows effectively decreases the number of days). This can be problematic for the speed of convergence to the asymptotic limit. This is also pointed out by Tate and Brown (1970), who summarised some results from previous studies regarding the necessary sample sizes. According to their analyses, as a rule of thumb, the number of rows times the number of columns *after* deleting rows with only zero's or only one's should be at least 24 when the number of rows is at least four (Tate and Brown (1970), p.159). In this case, the approximation can be deemed satisfactory.

To get an impression how likely it is to observe days without any violations and how likely it is to observe at least a certain number of violations in our simulation study, we consider the following experiment. Given a set of confidence levels around 99%, we state the proportion of days showing no VaR violations at all as well as the proportion of days on which we observe at least five or ten VaR violations across desks.

Table 7.2 shows the results. Let nex_t denote the number of desks showing exceptions on day t. The variable nex expresses the average percentage of days on which nex_t takes certain values. The row $nex = 0$ shows the average proportion of days t showing no violation (that is $L_{t,i} \leq \widehat{VaR}_{t,i}(\alpha)$ for all desks i on day t). For each simulation step we determine the number of days with zero VaR violations. The average is then taken over the total number of simulation steps. Row $nex = 10$ shows the average proportions of days showing at least 10 violations (that is $L_{t,i} \geq \widehat{VaR}_{t,i}(\alpha)$ for at least 10 desks on some day t). We chose the values $nex_t = 5$ and

$nex_t = 10$ since this roughly means that at least 10% of desks showed a violation on some day t in our simulation setting.

Looking at Table 7.2, row $nex = 0$, we see that in the case of independent, Gaussian distributed desks ($\rho =$ Zero, $F =$ Normal and $C =$ Gauss), the proportion of days without any VaR violation violation equates to α^d as expected. These percentages increase up to 99% for $\alpha = 0.999$ when the desks data are correlated ($\rho =$ Flat). In this case it is nearly the case that $L_{t,i} \leq \widehat{\text{VaR}}_{t,i}(\alpha)$ for all t. That is, nearly no violations were simulated at all. Considering the rule of thumb mentioned in Tate and Brown (1970), considering 50 and 100 desks (column data) in our simulation study, the constraint is that at least four rows should remain after deleting all zero-rows. For $n = 500$ days the rule of thumb is violated as soon as more than 99.2% of the days show no violation across desks. For $n = 250$ days, a percentage of more than 98.4% violates the rule of thumb. For better representation, we coloured results below 98.0% green and results greater 99.2% red for $n = 500$ (for $n = 250$: green for results below 97.0% and red for results greater 98.4%). Taking into account that this is a rough rule of thumb, according to our results, convergence speed could be problematic for α values around 0.999, where the average number of violations observed in the sample breaks or are close to the rule of thumb in various cases, especially for $n = 250$. At such high α levels, the number of violations in the sample data may be too low for the asymptotic in T_{QC} and T_{QE} to work.

Looking at row $nex = 5$ and $nex = 10$, we see that the percentage is well below 5% and in most settings lies around zero for the higher α levels. So the proportion of days showing at least 10% desk violations on some day is rather low.

When looking at t4-distributed marginal distributions ($F =$ t4) and comparing it to normal distributed marginal distributions ($F =$ Normal), we see that the number of violations are lower for t4 when looking at the higher α levels. This reverses for the lower α levels considered where the number of violations are higher for t4. This is reasonable, since the variance of the t4 distribution is normalised to one in our simulation so that at one point in the distribution, the t4 distribution lies below the Normal distribution.

Moving from a Gauss copula ($C =$ Gauss) to a t4 copula ($C =$ t4) we observe more days without violations. Showing at least five violations ($nex = 5$) is more likely when sampling from the Gauss copula and not from the t4 copula. This pattern is reversed for $nex = 10$. This could be attributed to the tail-dependence in the t4 copula. When violations are observed, there is a tendency for clustered violations across desks on a day.

The results show that violations are generally a rare event in our simulation study when looking

at confidence levels around 99%. Since this is consistent with what can be observed in reality as well, it is of interest, which tests still can extract enough information from the (simulated) data to show reasonable power.

			d	50				100			
			ρ	Zero		Flat		Zero		Flat	
α	nex	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
0.900	0	250	Normal	0.52	12.32	34.72	43.45	0.00	5.83	26.20	35.82
			t4	2.39	23.73	43.63	54.14	0.06	14.35	34.39	46.82
		500	Normal	0.53	12.21	34.77	43.43	0.00	5.79	26.19	35.81
			t4	2.36	23.63	43.63	54.07	0.06	14.24	34.36	46.71
	5	250	Normal	56.74	44.22	32.02	29.12	97.61	68.07	45.16	39.52
			t4	29.01	30.20	23.69	21.37	85.35	51.22	35.53	29.97
		500	Normal	56.81	44.29	32.02	29.12	97.65	68.07	45.22	39.50
			t4	29.09	30.23	23.74	21.37	85.45	51.18	35.60	30.00
	10	250	Normal	2.42	15.59	17.84	17.72	54.80	42.48	30.86	28.24
			t4	0.25	9.30	12.14	12.41	18.41	28.28	22.67	20.68
		500	Normal	2.44	15.64	17.80	17.74	54.76	42.53	30.92	28.28
			t4	0.26	9.35	12.09	12.44	18.35	28.33	22.67	20.60
0.950	0	250	Normal	7.75	38.07	53.11	64.54	0.60	27.02	43.64	58.01
			t4	12.87	46.19	58.31	69.89	1.64	35.07	48.89	63.90
		500	Normal	7.68	37.86	53.11	64.48	0.59	26.90	43.60	57.88
			t4	12.78	46.00	58.28	69.87	1.62	34.95	48.89	63.87
	5	250	Normal	10.32	19.52	16.66	14.93	56.33	35.72	26.58	21.70
			t4	5.06	15.15	13.44	12.10	37.57	28.70	22.10	17.77
		500	Normal	10.31	19.54	16.64	14.98	56.29	35.84	26.56	21.65
			t4	4.96	15.16	13.36	12.11	37.56	28.76	22.14	17.76
	10	250	Normal	0.02	5.46	7.76	8.38	2.78	17.97	15.71	14.37
			t4	0.00	4.06	5.90	6.65	0.72	13.74	12.58	11.59
		500	Normal	0.01	5.47	7.67	8.40	2.81	17.95	15.70	14.34
			t4	0.00	4.07	5.85	6.65	0.73	13.77	12.52	11.56
0.990	0	250	Normal	60.56	82.33	82.84	90.76	36.58	76.54	76.54	88.34
			t4	46.76	74.76	77.19	86.74	21.78	67.20	69.73	83.40
		500	Normal	60.52	82.24	82.82	90.76	36.63	76.50	76.51	88.37
			t4	46.69	74.70	77.25	86.65	21.86	67.16	69.80	83.35
	5	250	Normal	0.01	3.11	2.88	2.99	0.34	6.66	5.88	4.69
			t4	0.10	4.93	4.62	4.55	1.77	10.26	8.94	7.01
		500	Normal	0.01	3.12	2.87	3.01	0.32	6.68	5.83	4.68
			t4	0.10	4.95	4.62	4.56	1.79	10.30	8.86	6.99
	10	250	Normal	0.00	0.69	0.88	1.54	0.00	2.70	2.55	2.83
			t4	0.00	1.13	1.60	2.34	0.00	4.31	4.16	4.32
		500	Normal	0.00	0.71	0.90	1.54	0.00	2.71	2.53	2.85
			t4	0.00	1.16	1.60	2.35	0.00	4.33	4.14	4.33
0.999	0	250	Normal	95.16	97.96	97.14	98.93	90.53	97.12	95.54	98.63
			t4	74.16	88.83	88.14	94.19	54.93	84.90	83.21	92.60
		500	Normal	95.15	97.96	97.15	98.94	90.56	97.12	95.53	98.64
			t4	74.15	88.81	88.21	94.22	54.92	84.87	83.28	92.68
	5	250	Normal	0.00	0.26	0.16	0.28	0.00	0.61	0.44	0.47
			t4	0.00	1.76	1.55	1.80	0.04	3.89	3.41	2.85
		500	Normal	0.00	0.28	0.17	0.30	0.00	0.61	0.45	0.48
			t4	0.00	1.78	1.55	1.80	0.04	3.89	3.37	2.85
	10	250	Normal	0.00	0.05	0.03	0.14	0.00	0.22	0.13	0.27
			t4	0.00	0.38	0.42	0.88	0.00	1.52	1.33	1.70
		500	Normal	0.00	0.06	0.03	0.15	0.00	0.24	0.14	0.28
			t4	0.00	0.39	0.44	0.90	0.00	1.55	1.33	1.69

Table 7.2: Average percentage of days with zero, at least 5 or at least 10 violations across all desks

7.5.2 Cochran Q-test

In this section, we analyse the performance of the two presented tests. First, we look at the results for the two extreme cases when all desks are specified either correctly (F = Normal) or wrongly (F = t4) by the risk modelling group. Table 7.3 shows the results.

For the assessment of the performance of the original Cochran Q-test (QC) we have to keep in mind that the test is designed as a test to detect significant differences in correlated binary data. Therefore, we expect that both in the case of 100% correctly specified desks and in case of 100% misspecified desks, the null hypothesis is rejected with the nominal level since the null hypothesis is true in both cases. Looking at Table 7.3, this is exactly what we see. The original Cochran Q-test shows rejection rates in the area of the nominal level of 5%. Looking at row QE, which shows the extended version of the Cochran Q-test, we see that the test generally shows satisfactorily sizes around the nominal level. The only exception is in the case of independent data in combination with $n = 250$ days, when the null hypothesis is satisfied ($n = 250$, ρ = Zero, F = Normal, C = Gauss). In this case, the test is oversized with values of 12% and 13%. For $n = 500$ days, the effect does not occur. Looking at Table 7.5 which shows the simulation results for different α levels, we see that the effect is also absent for $n = 250$ days when the level of α is decreased, i.e. when there is a tendency to observe more violations on a day. The asymptotic convergence seems to be too slow in case of 250 days in combination with α levels of 0.99 or higher for Imhof's method. This is supported by the findings in Section 7.5.1. For 250 days, as a rule of thumb, α levels beyond 98.4% become critical. What also has to be noted is that for QE, the correlation between the VaR violation indicators is estimated from the data. This brings additional uncertainty when there are too many zero-rows in the data. The question remains, why this only is a problem for independent data and not, for example, for the t4 copula (C = t4). Judging from the results stated in Table 7.2 for $nex = 5$ and $nex = 10$, the case of independent data, when the null hypothesis is satisfied (F = Normal, C = Gauss, ρ = Zero) produces the fewest number of violations across all considered simulation settings, what could explain the observed behaviour.

The power of the extended Cochran Q-test is quite satisfactory even for 250 days, with levels generally above 60%, when 100% of desks are misspecified.

Next we look at the tests performances when the risk modelling group misspecifies a certain fraction of desks. Table 7.4 shows the results.

For the original Cochran Q-test, QC, it is only relevant if all desks behave the same. Therefore, the power is highest (albeit still on a moderate level) in the case that 50% of desks are misspec-

ified. In this case an equal amount of desks is equally bad, respectively equally good. In case of 25% and 75% misspecified desks, the number of equally good, respectively equally bad desks seems to be too high for the test to be significant. The size of the test is satisfactory.

For the extended version of Cochran Q-test, QE, the power increases significantly compared to QC. This is expected since the extended Cochran Q-test tests a narrower hypothesis. Whereas the null hypothesis of QC encompasses all models for which desks are equally good or bad, the null hypothesis of QE is more detailed testing for equally good desks, which violation rates have to match the mean. For QE, power is highest when 75% of desks are misspecified. This make sense since in case of 75% misspecified desks, the unconditional coverage hypothesis is violated in most desks. In case of 25% misspecified desks, the power is rather poor in most cases but still better compared to the analogous case for the original Cochran Q-test.

Next, we analyse how the tests are influenced by the choice of the α level. We only look at α levels below 0.99, since the results in Section 7.5.1 indicate that levels beyond 0.99 might show too few violations within a day. Table 7.5 shows the results for various levels of α. We see that for α levels of 0.90 and 0.95, the size is around the nominal level for both for QC and QE, the latter being a bit undersized. For QE, the test's power increases when moving from $\alpha = 0.95$ to $\alpha = 0.90$. This is expected since lower α levels create more violations, that is a lower number of zero-rows. This implies more information for both Cochran tests which tendentially increases the power. For $\alpha = 0.95$ and $n = 250$ the power is rather low. This is non-intuitive to the findings in Table 7.3 where the power is better, albeit the α level is higher with $\alpha = 0.99$. In general, the case of $n = 250$ seems to be problematic in combination with higher α levels.

		d	50				100			
		ρ	Zero		Flat		Zero		Flat	
test	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
QC	250	Normal	4.7	4.2	4.0	4.3	5.4	4.9	3.9	4.5
		t4	3.9	4.3	5.6	5.2	4.9	4.4	4.5	4.4
	500	Normal	5.0	3.8	5.7	4.0	4.6	3.6	6.2	4.5
		t4	4.7	3.7	4.4	5.0	4.8	3.7	5.6	4.4
QE	250	Normal	12.0	5.3	4.4	5.0	12.7	4.1	4.3	4.9
		t4	97.0	70.8	60.2	35.9	100.0	75.9	63.3	37.4
	500	Normal	4.8	3.2	4.4	4.6	4.2	2.8	3.7	4.4
		t4	100.0	90.5	82.5	60.1	100.0	95.0	86.5	60.9

Table 7.3: Estimated size and power of Cochran Q-test and extended Cochran Q-test for $\alpha = 0.99$ and 100% misspecified desks

			d	50				100			
			ρ	Zero		Flat		Zero		Flat	
test	fraction	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
QC	25	250	Normal	4.7	4.2	4.0	4.3	5.4	4.9	3.9	4.5
			t4	14.8	14.0	13.0	15.9	19.9	21.0	21.3	24.4
		500	Normal	5.0	3.8	5.7	4.0	4.6	3.6	6.2	4.5
			t4	26.9	28.4	30.4	39.7	42.0	43.3	48.8	58.3
	50	250	Normal	4.7	4.2	4.0	4.3	5.4	4.9	3.9	4.5
			t4	15.2	16.6	17.4	20.2	24.5	24.3	26.4	30.4
		500	Normal	5.0	3.8	5.7	4.0	4.6	3.6	6.2	4.5
			t4	33.8	32.9	41.0	48.4	51.8	56.4	60.9	71.6
	75	250	Normal	4.7	4.2	4.0	4.3	5.4	4.9	3.9	4.5
			t4	11.0	10.5	12.2	12.7	15.2	16.4	16.6	18.2
		500	Normal	5.0	3.8	5.7	4.0	4.6	3.6	6.2	4.5
			t4	21.9	20.4	24.7	32.1	32.2	35.9	40.1	48.7
QE	25	250	Normal	12.0	5.3	4.4	5.0	12.7	4.1	4.3	4.9
			t4	42.6	20.8	18.4	11.2	60.9	24.0	17.7	11.5
		500	Normal	4.8	3.2	4.4	4.6	4.2	2.8	3.7	4.4
			t4	56.2	34.5	27.5	15.9	78.4	40.0	31.2	16.2
	50	250	Normal	12.0	5.3	4.4	5.0	12.7	4.1	4.3	4.9
			t4	72.3	42.6	35.1	20.8	90.0	47.5	37.7	19.8
		500	Normal	4.8	3.2	4.4	4.6	4.2	2.8	3.7	4.4
			t4	89.6	66.1	56.5	31.6	99.2	73.6	61.3	32.7
	75	250	Normal	12.0	5.3	4.4	5.0	12.7	4.1	4.3	4.9
			t4	87.7	58.8	49.3	28.7	98.8	65.0	53.7	29.1
		500	Normal	4.8	3.2	4.4	4.6	4.2	2.8	3.7	4.4
			t4	98.7	81.5	73.1	47.6	100.0	89.6	77.7	50.0

Table 7.4: Estimated size and power of Cochran Q-test and extended Cochran Q-test for various fractions at level $\alpha = 0.99$

			d	50				100			
			ρ	Zero		Flat		Zero		Flat	
test	α	n	$F \mid C$	Gauss	t4	Gauss	t4	Gauss	t4	Gauss	t4
QC	0.90	250	Normal	4.3	4.4	4.6	4.9	4.4	4.0	5.1	3.5
			t4	4.7	4.7	5.4	4.1	4.6	5.2	4.5	4.3
		500	Normal	4.1	5.1	4.9	5.9	5.3	5.0	4.9	5.1
			t4	3.9	5.1	5.5	5.5	4.5	4.1	5.6	4.6
	0.95	250	Normal	4.7	3.4	6.4	4.6	4.5	4.7	4.6	5.6
			t4	5.2	3.7	5.2	5.0	3.6	5.6	4.2	4.7
		500	Normal	5.2	5.4	4.9	3.9	4.7	4.5	6.0	3.8
			t4	5.2	3.9	4.9	4.1	5.0	4.4	4.3	4.6
QE	0.90	250	Normal	3.0	3.1	5.6	4.6	1.6	2.8	5.2	5.1
			t4	100.0	99.0	80.9	71.5	100.0	99.4	81.9	73.6
		500	Normal	3.6	5.0	5.0	5.5	4.0	4.8	5.4	5.7
			t4	100.0	100.0	98.6	95.8	100.0	100.0	99.0	96.3
	0.95	250	Normal	3.8	3.0	4.7	4.0	1.8	3.7	5.4	4.0
			t4	34.7	28.5	22.7	18.2	55.2	32.9	24.3	19.6
		500	Normal	4.7	3.9	4.4	5.9	3.7	4.2	4.4	5.2
			t4	95.4	72.1	50.9	38.3	99.7	78.5	53.0	39.3

Table 7.5: Estimated size and power of Cochran Q-test and extended Cochran Q-test for various levels of α

7.6 Summary

The main conclusions of this chapter are:

- We presented the Cochran Q-test which can be used to address our third research question which concerns the question of whether all trading desk models are equally good at modelling risk. Since this test has only power against the null hypothesis that all desks behave equally (either equally good or equally bad) we developed a new test based on Cochran's test statistic which also tests for the unconditional coverage hypothesis. To derive the asymptotic theory of the new test, we used Imhof's method, which approximates the distribution of a quadratic form.

- We analysed the performance of these comparative tests in a simulation study. We found that the new test has comparably better power. This is explainable, since the new test tests a narrower hypothesis. Whereas power is only on a moderate level for 25% misspecified desks, it improves considerably when 50% and 75% are misspecified.

- The new test shows some non-intuitive behaviour in simulation studies where we consider

250 (backtesting) days in combination with high α levels of 0.95 and 0.99. Looking at some complementary studies we attribute this to the fact that the effective sample size is not high enough when the combination of number of days and the α level leads to too few violations to guarantee a sufficient rate of convergence. This is supported by a conducted simulation study analysing the percentage of days showing no violations. We therefore recommend looking at lower percentile levels such as 0.90 when applying the new test with 250 days in regulatory contexts.

Chapter 8

Summary and Outlook

After more than 10 years in the making, the reforms of the Basel III framework finally have an impending deadline for their full implementation. The reforms encompass a major revision of the market risk framework. Among other new features, the formula for the calculation of the minimum capital requirements for market risk is now based on the expected shortfall measure and includes liquidity risk. Moreover, banks now have to show that their internal market risk models work appropriately not only at bank-wide level but also at trading desk level and at a wider range of quantiles.

In this work we have thoroughly derived the revised capital formula, which is based on the assumption that the 10-day risk factor changes form a multivariate Gaussian white noise process, and that the changes in the portfolio value are a linear function of these risk factor changes. We have analysed the formula under the more general assumption that the risk factor changes form a multivariate elliptical distribution and showed that it gives an upper bound in this case. This means that the formula leads to a higher capital charge than actually necessary to achieve the target. The results provide valuable insights into the functionality of the formula.

The remaining part of this work concerned the revised validation standards, which will likely motivate banks to review their internal validation framework. We thereby focused on the new requirement that internal models have to demonstrate satisfactory performance both at bank-wide and trading desk level, as well as the stipulation that a wider range of the estimated loss distribution should be validated. To this end we proposed various multi-desk backtests, which simultaneously backtest all trading desk models in the presence of intra-desk correlation. Multi-desk backtests have several advantages to univariate (mono-desk) backtests, which test each of the internal trading desk models separately. However, to the best of our knowledge, only two papers have proposed concrete tests which could be used as multi-desk backtests. Both

of them use indicator variables for value-at-risk violations as input variables. We developed a multivariate extension of the spectral tests proposed in Gordy and McNeil (2020), which are based on probability integral transform values, exploiting more data than the usual indicator variables of value-at-risk violations. In fact, the newly developed multi-desk spectral tests turn out to be quite powerful. The tests are also very flexible, since the risk modelling group can emphasize the region of the estimated loss distribution about which it is most concerned with respect to model performance and can choose a weighting scheme for observed violations accordingly. We also analysed the benefits of using bootstrap methods but generally found no clear advantages over the Z-tests, particularly when the additional computational complexity and increased time resources are taken into account.

The final research question was also connected to desk-based backtesting but focused on comparative tests, that is tests comparing the performance of the various internal desk models. To address this research question, the well-known Cochran Q-test can be used, which is designed to detect significant differences between outcomes for correlated binary data. We first provided a more thorough derivation of the test's statistical distribution, which is derived heuristically in Cochran's original paper. The standard application of the Cochran Q-test can only provide an indication if all desks perform equally. Since the regulator is particularly interested in whether all internal models lead to numbers of VaR violations that are no more than expected, we extended the Cochran Q-test to a test that also tests the unconditional coverage hypothesis at the same time. Our simulation studies showed that the new proposed test has good size and power properties. Due to some complementary simulation studies, we recommend using this new test only for percentiles below 0.95 when the length of the used backtesting history is 250 days or less.

Overall, the performance and structural advantages of the new proposed multi-desk tests (such as the possibility to assess a wider range of the estimated loss distribution or the fact that they control for correlation across desks) suggest that they are a valuable addition to a bank's validation framework. Nevertheless, it has to be noted, that we generally found that multi-desk backtests require a certain proportion of desks to be misspecified. Detecting a handful of 'bad' desks in a (realistic) setting of 50 to 100 trading desks is rather unrealistic. The tests are more powerful when at least 25% of internal models are misspecified. However, compared to univariate mono-desk tests, multi-desk tests are able to incorporate existing correlation structures across trading desks and avoid the multiple testing issue and are therefore still a valuable addition to a validation framework. If a multi-desk test fails, it is likely that a larger amount of internal trading desk models do not perform satisfactorily. A post hoc testing should then be used to

identify the internal models, which are inadequate. For future research, the simulation setting could be extended to a fewer number of desks and asymmetric copulas (like Gumble or Clayton) in order to analyse if the observed effects still persist. Moreover, due to the fact that real data are not easily accessible, it could be thought about using indices to approximate desk level data.

It can also pointed out that the multi-desk spectral backtests can be used to assess a wider range of the internal model. This in turn can inform about the adequacy of the expected shortfall measure, which measures the average of the losses beyond the VaR cut-off point. Additionally, since the benefits of using probability integral transform values are already being exploited in the USA, it is likely that other regulatory authorities will probably exploit this kind of information more systematically in the future as well. Moreover, it could be explored in the future if the multi-desk spectral test can be extended in a way that the PIT-value level exceedances of every desk are weighted differently across the desks. This seems to be a useful extension considering that less risky trading desks can receive less attention compared to higher risky desks.

Since the preferred tests have the form of Z-tests, they are easy to implement and the calculation time is not excessive. Moreover, perceived model risk in the Z-tests and comparative tests is no higher than for other tests already widely adopted in validation frameworks. This is an important factor. Especially in on-site investigations, the regulator audits in-depth the underlying model assumptions and challenges the model choices. Banks have to be prepared to show that all model choices made are reasonable and valid[33].

Finally, we can also say that the new multi-desk backtests can probably also serve to backtest a whole banking system instead of several trading desks whilst including spill-over effects across banks. In this way, systemic risk could be detected. Another interesting field of application could be in the area of capital allocation. The new multi-desk backtests could serve to detect traders, which are not respecting their allocated VaR limits, for example. However, these aspects as well as the application of the new tests on real data, are left for future research. The application on real data is complicated in particular since there is currently very little data available to researchers due to the sensitivities surrounding bank risk reporting both in the EU and the USA.

[33]See the requirement as laid down in BCBS (2019b), p.69, para. 30.17(1), which requires banks to show that any assumptions made within internal models are appropriate.

Appendix

Characteristic function of univariate spherical distributions In the following, we derive the characteristic functions for the symmetric generalised hyperbolic distribution and the Student t distribution. In both cases we reduce the characteristic function of the respective distribution to the integral over the support of the density of the generalised inverse Gaussian distribution, which is one.

Symmetric generalised hyperbolic distribution Let Y be a univariate symmetric generalized hyperbolic random variable. That is it can be represented by a normal variance mixture

$$Y \stackrel{d}{=} \sqrt{W}Z, \tag{8.1}$$

where $W \geq 0$ is a non-negative, scalar-valued random variable (see Definition 11) which follows a so-called Generalised Inverse Gaussian (GIG) distribution $W \sim N^{-}(\lambda, \chi, \kappa)$ and Z is a standard normal variable. The density of the GIG is given by

$$f_W(w) = \frac{\chi^{-\lambda}\left(\sqrt{\chi\kappa}\right)^{\lambda}}{2K_\lambda\left(\sqrt{\chi\kappa}\right)} w^{\lambda-1} e^{-1/2\left(\chi w^{-1}+\kappa w\right)} \quad , \begin{cases} \chi > 0, \kappa \geq 0 & \text{if } \lambda < 0 \\ \chi > 0, \kappa > 0 & \text{if } \lambda = 0 \\ \chi \geq 0, \kappa > 0 & \text{if } \lambda > 0, \end{cases}$$

where K_λ denotes a Bessel function of the third kind. We know from (3.5) that the characteristic function of Y is given by

$$\phi(s) = \int_0^\infty e^{-1/2s^2 w} f_W(s) \tag{8.2}$$

and the variance is $\mathbb{E}[W]$.

Replacing the density f_W in (8.2) by the density of the GIG we get

$$\phi(s) = \int_0^\infty e^{-1/2s^2 w} \frac{\chi^{-\lambda}\left(\sqrt{\chi\kappa}\right)^\lambda}{2K_\lambda\left(\sqrt{\chi\kappa}\right)} w^{\lambda-1} e^{-1/2\left(\chi w^{-1}+\kappa w\right)} \mathrm{d}w.$$

This can be transformed to

$$\frac{\chi^{-\lambda}\left(\sqrt{\chi\kappa}\right)^\lambda}{2K_\lambda\left(\sqrt{\chi\kappa}\right)} \int_0^\infty e^{-1/2\left(\chi w^{-1}+(\kappa+s^2)w\right)} A w^{\lambda-1} \frac{1}{A} \mathrm{d}w \tag{8.3}$$

with

$$A = \frac{\chi^{-\lambda}\left(\sqrt{\chi(\kappa+s^2)}\right)^\lambda}{2K_\lambda\left(\sqrt{\chi(\kappa+s^2)}\right)}.$$

Cancelling expressions out in (8.3) leads to

$$\frac{\left(\sqrt{\chi\kappa}\right)^\lambda}{K_\lambda\left(\sqrt{\chi\kappa}\right)} \frac{K_\lambda\left(\sqrt{\chi(\kappa+s^2)}\right)}{\left(\sqrt{\chi(\kappa+s^2)}\right)^\lambda} \int_0^\infty e^{-1/2\left(\chi w^{-1}+(\kappa+s^2)w\right)} A w^{\lambda-1} \mathrm{d}w.$$

The integral in the last expression is the integral over the density of the GIG distribution which is one (note that the support of the Inverse Gaussian is $(0,\infty)$). We therefore have

$$\phi(s) = \left(\frac{\kappa}{s^2+\kappa}\right)^{\lambda/2} \frac{K_\lambda\left(\sqrt{\chi(\kappa+s^2)}\right)}{K_\lambda\left(\sqrt{\chi\kappa}\right)}, s > 0$$

which gives (3.19).

Student t distribution By assuming W in (8.1) to be a Inverse Gamma (IG) distributed random variable, $W \sim IG(1/2\nu, 1/2\nu)$, we obtain the Student t distribution with ν degrees of freedom with $\mathbb{E}(W) = \nu/(\nu-2)$, $\nu > 2$.

That is, in the case of the Student t distribution, f_W is the density of an IG random variable $W \sim \mathrm{IG}(\frac{1}{2}\nu, \frac{1}{2}\nu)$. Replacing f_W in (8.2) with the density of the IG distribution we obtain

$$\phi(s) = \int_0^\infty e^{-\frac{1}{2}s^2 w} \frac{(\frac{1}{2}\nu)^{\nu/2}}{\Gamma(\frac{1}{2}\nu)} w^{-\frac{\nu}{2}-1} e^{-\frac{1}{2}\nu w^{-1}} \mathrm{d}w.$$

We transform the former expression to

$$\frac{(\frac{1}{2}\nu)^{\nu/2}}{\Gamma(\frac{1}{2}\nu)} \int_0^\infty w^{-\frac{\nu}{2}-1} e^{(-\frac{1}{2}(s^2 w+\nu w^{-1}))} B \frac{1}{B} \mathrm{d}w \tag{8.4}$$

with

$$B = \frac{2K_{\nu/2}(\sqrt{\nu s^2})}{\nu^{\nu/2}\left(\sqrt{\nu s^2}\right)^{-\nu/2}}.$$

We then transform (8.4) to

$$\frac{(\frac{1}{2}\nu)^{\nu/2}}{\Gamma(\frac{1}{2}\nu)} \frac{2K_{\nu/2}(\sqrt{\nu s^2})}{\nu^{\nu/2}\left(\sqrt{\nu s^2}\right)^{-\nu/2}} \int_0^\infty w^{-\frac{\nu}{2}-1} e^{(-\frac{1}{2}(s^2 w + \nu w^{-1}))} \frac{1}{B} \mathrm{d}w.$$

In the last expression we have the integral of the density of the GIG distribution, which equals one. We therefore have

$$\phi(s) = K_{\nu/2}(\sqrt{\nu}s) \frac{\left(\sqrt{\nu s^2}\right)^{\nu/2}}{2^{\nu/2-1}\Gamma(\frac{1}{2}\nu)}, s > 0$$

which is (3.18).

References

Abad, P., Benito, S. and López, C. (2014). A comprehensive review of value at risk methodologies. *The Spanish Review of Financial Economics*, 12(1), 15–32.

Acerbi, C. and Szekely, B. (2014). *Backtesting expected shortfall*. [Online]. Available at: `https://www.msci.com/documents/10199/22aa9922-f874-4060-b77a-0f0e267a489b` [Accessed: 21 Oct 2021].

Acerbi, C. and Szekely, B. (2017). *General properties of backtestable statistics*. [Online]. Available at: `https://ssrn.com/abstract=2905109` [Accessed: 21 Oct 2021].

Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7), 1487–1503.

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2), 177–190.

Angelo, C. and Ripley, B. (2020). *boot: bootstrap R (S-Plus) functions*. R package version 1.3-25.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203–228.

Balter, J. (2021). *R codes - topics on the fundamental review of the trading book*. [Online]. Available at: `https://osf.io/hpy5j/` [Accessed: 27 July 2021].

Bank of International Settlements (2020). *History of the Basel committee*. [Online]. Available at: `https://www.bis.org/bcbs/history.htm` [Accessed: 27 Apr 2020].

Basel Committee on Banking Supervision (1988). *Basel Capital Accord*. [Online]. Available at: `https://www.bis.org/publ/bcbs04a.htm` [Accessed: 15 June 2020].

Basel Committee on Banking Supervision (1996). *Amendment to the capital accord to incorporate market risk*. [Online]. Available at: `https://www.bis.org/publ/bcbs24.pdf` [Accessed: 13 Jan 2021].

Basel Committee on Banking Supervision (2006). *International convergence of capital measurement and capital standards*. [Online]. Available at: `https://www.bis.org/publ/bcbs128.pdf` [Accessed: 30 Mar 2020].

Basel Committee on Banking Supervision (2011*a*). *Basel III: a global regulatory framework for more resilient banks and banking systems*. [Online]. Available at: `https://www.bis.org/publ/bcbs189.pdf` [Accessed: 30 Mar 2020].

Basel Committee on Banking Supervision (2011*b*). *Revisions to the Basel II market risk framework*. [Online]. Available at: `https://www.bis.org/publ/bcbs193.pdf` [Accessed: 13 Jan 2021].

Basel Committee on Banking Supervision (2019*a*). *Explanatory note on the minimum capital requirements for market risk*. [Online]. Available at: `https://www.bis.org/bcbs/publ/d457_note.pdf` [Accessed: 13 Jan 2021].

Basel Committee on Banking Supervision (2019*b*). *Minimum capital requirements for market risk*. [Online]. Available at: `https://www.bis.org/bcbs/publ/d457.pdf` [Accessed: 18 Jan 2021].

Basel Committee on Banking Supervision (2021). *The Basel framework*. [Online]. Available at: `https://www.bis.org/basel_framework/` [Accessed: 29 May 2021].

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4), 465–474.

Berkowitz, J., Christoffersen, P. and Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science*, 57(12), 2213–2227.

Bodenham, D. A. and Adams, N. M. (2016). A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, 26(4), 917–928.

Breymann, W. and Lüthi, D. (2013). *ghyp: a package on generalized hyperbolic distributions*. R package version 1.5.6.

Campbell, S. (2007). A review of backtesting and backtesting procedures. *Journal of Risk*, 9(2), 1–17.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.

Christoffersen, P. and Pelletier, D. (2004). Backtesting value-at-risk: a duration-based approach. *Journal of Financial Econometrics*, 2(1), 84–108.

Cochran, W. G. (1937). The efficiencies of the binomial series tests of significance of a mean and of a correlation coefficient. *Journal of the Royal Statistical Society*, 100(1), 69–73.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37(3/4), 256–266.

Committee on Science and Technology (2009). *The risks of financial modeling: Var and the economic meltdown*. [Online]. Available at: `https://www.govinfo.gov/content/pkg/CHRG-111hhrg51925/pdf/CHRG-111hhrg51925.pdf` [Accessed: 15 Oct 2021].

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236.

Costanzino, N. and Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Journal of Risk Model Validation*, 9(1), 21–31.

Costanzino, N. and Curran, M. (2018). A simple traffic light approach to backtesting expected shortfall. *Risks*, 6(1), 2.

Council of European Union (2013). *Council regulation (EU) no 575/2013*. [Online]. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02013R0575-20200627` [Accessed: 15 Jan 2021].

Council of European Union (2019). *Council regulation (EU) no 2019/876*. [Online]. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02019R0876-20200627` [Accessed: 3 June 2021].

Danciulescu, C. (2016). Backtesting aggregate risk. *Journal of Forecasting*, 35(4), 285–307.

Danielsson, J. and Zigrand, J.-P. (2006). On time-scaling of risk and the square-root-of-time rule. *Journal of Banking & Finance*, 30(10), 2701–2713.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge University Press. Cambridge. ISBN 0-521-57391-2.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189–212.

Diebold, F., Hickman, A., Inoue, A. and Schuermann, T. (1998). Scale models. *Risk*, 11, 104–107.

Diks, C., Panchenko, V. and van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2), 215–230.

Du, Z. and Escanciano, J. C. (2017). Backtesting expected shortfall: accounting for tail risk. *Management Science*, 63(4), 940–985.

Duchesne, P. and de Micheaux, P. L. (2010). Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54, 858–862.

Dunn, O. J. (1959). Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics*, 30(1), 192–197.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.

Efron, B., Rogosa, D. and Tibshirani, R. (2015). Resampling methods of estimation. In International Encyclopedia of the Social & Behavioral Sciences. 2nd ed. Oxford: Elsevier. pp. 492–495.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. 1st ed. New York: Chapman & Hall/CRC.

Embrechts, P. (2000). Extreme value theory: Potential and limitations as an integrated risk management tool. *Derivatives Use, Trading & Regulation*, 6(1).

Emmer, S., Kratz, M. and Tasche, D. (2013). What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk*, 18(2), 31–60.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.

Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of business & economic statistics*, 22(4), 367–381.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. 1st ed. New York: Chapman & Hall/CRC.

Finger, C. C. (2009). *IRC comments*. [Online]. Available at: `https://www.msci.com/documents/10199/95a00649-09f2-49c8-8a09-e5f9b96d5a96` [Accessed: 18 Jan 2021].

Fissler, T., Ziegel, J. F. and Gneiting, T. (2015). Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*, .

Frey, R. and McNeil, A. J. (2002). Var and expected shortfall in portfolios of dependent credit risks: conceptual and practical insights. *Journal of Banking & Finance*, 26(7), 1317–1334.

Genest, M., Masse, J.-C. and Plante, J.-F. (2019). *depth: nonparametric depth functions for multivariate analysis*. R package version 2.1-1.1.

Ghosh, S. and Polansky, A. M. (2014). Smoothed and iterated bootstrap confidence regions for parameter vectors. *Journal of Multivariate Analysis*, 132, 171–182.

Gil-Pelaez, J. (1951). Note on the inversion theorem. *Biometrika*, 38(3-4), 481–482.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, 69(2), 243–268.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3), 411–422.

Gordy, M. B. and McNeil, A. J. (2020). Spectral backtests of forecast distributions with application to risk management. *Journal of Banking & Finance*, 116, 105817.

Hall, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika*, 74(3), 481–493.

Hofert, M. and Mächler, M. (2016). Parallel and other simulations in R made easy: an end-to-end study. *Journal of Statistical Software*, 69(4), 1–44.

Hung, J.-C., Su, J.-B., Chang, M. C. and Wang, Y.-H. (2020). The impact of liquidity on portfolio value-at-risk forecasts. *Applied Economics*, 52(3), 242–259.

Hurlin, C., Colletaz, G., Tokpavi, S. and Candelon, B. (2010). Backtesting value-at-risk: a GMM duration-based test. *Journal of Financial Econometrics*, 9(2), 314–343.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4), 419–426.

Inui, K. and Kijima, M. (2005). On the significance of expected shortfall as a coherent risk measure. *Journal of Banking & Finance*, 29(4), 853–864.

Jorion, P. (2007). *Value at risk. The new benchmark for managing financial risk.* 3rd ed. New York: McGraw-Hill.

Kratz, M., Lok, Y. H. and McNeil, A. J. (2018). Multinomial VaR backtests: a simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, 88.

Kreiss, J.-P. and Lahiri, S. N. (2012). *Bootstrap methods for time series.* Vol. 30. 1st ed. Oxford: Elsevier.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2), 73–84.

Lawrence, C. and Robinson, G. (1997). Liquidity, dynamic hedging and value at risk. *Risk Management for Financial Institutions*, 1(9), 63–72.

Massé, J.-C. (2009). Multivariate trimmed means based on the tukey depth. *Journal of Statistical Planning and Inference*, 139(2), 366–384.

McNeil, A., Frey, R. and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools.* 2nd ed. Princeton: Princeton Series in Finance.

McNemar, Q. (1969). *Psychological Statistics.* 4th ed. New York: John Wiley & Sons.

Pérignon, C. and Smith, D. R. (2008). A new approach to comparing VaR estimation methods. *The Journal of Derivatives*, 16(2), 54–66.

Pitera, M. and Schmidt, T. (2018). Unbiased estimation of risk. *Journal of Banking & Finance*, 91, 133–145.

R Core Team (2020). *R: a language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. Available at: `http://www.R-project.org/`.

Resnick, S. I. (2005). *A probability path.* 1st ed. Boston: Birkhäuser.

Rootzén, H. and Klüppelberg, C. (1999). A single number can't hedge against economic catastrophes. *Ambio*, 28(6), 550–555.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3), 470–472.

Tate, M. W. and Brown, S. M. (1970). Note on the Cochran Q test. *Journal of the American Statistical Association*, 65(329), 155–160.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians (Vancouver)*, 2, 523–531.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *The Annals of Mathematical Statistics*, 18(1), 88–96.

Wang, J.-N., Yeh, J.-H. and Cheng, N. Y.-P. (2011). How accurate is the square-root-of-time rule in scaling tail risk: a global study. *Journal of Banking & Finance*, 35(5), 1158–1169.

Wied, D., Weiß, G. and Ziggel, D. (2016). Evaluating value-at-risk forecasts: a new set of multivariate backtests. *Journal of Banking & Finance*, 72, 121–132.

Yeh, A. B. and Singh, K. (1997). Balanced confidence regions based on Tukey's depth and the bootstrap. *Journal of the Royal Statistical Society*, 59(3), 639–652.

Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance*, 26(4), 901–918.

Ziggel, D., Berens, T., Weiß, G. and Wied, D. (2014). A new set of improved value-at-risk backtests. *Journal of Banking & Finance*, 48, 29–41.