Three Essays on Improving Financial Risk Estimation, Forecasting and Backtesting

Dissertation

zur Erlangung des akademischen Grades eines Doktors der
Wirtschaftswissenschaften (Dr. rer. pol.)

vorgelegt von
Sebastian Bayer

Universität
Konstanz

Sektion Politik – Recht – Wirtschaft
Fachbereich Wirtschaftswissenschaften
Konstanz, 2017

Tag der mündlichen Prüfung: 19. März 2018

1. Referent: Prof. Dr. Winfried Pohlmeier
2. Referent: Prof. Dr. Ralf Brüggemann
3. Referent: Prof. Dr. Francesco Audrino

Für Marina

Risk is like fire: If controlled it will help you;
if uncontrolled it will rise up and destroy you.

Theodore Roosevelt

Danksagung

An dieser Stelle möchte ich einigen Menschen danken, die mich bei der Entstehung dieser Arbeit begleitet und unterstützt haben.

Mein Dank gilt zunächst meinem Doktorvater Herrn Prof. Dr. Winfried Pohlmeier, der mich seit meiner Anstellung als studentische Hilfskraft unterstützt und gefördert hat. Durch seine Initiative, sein Vertrauen und seine Hilfe habe ich diese Promotion begonnen und erfolgreich abgeschlossen. Herrn Prof. Dr. Ralf Brüggemann danke ich für die Übernahme der Zweitbetreuung dieser Arbeit und hilfreiches Feedback in Seminaren und Herrn Prof. Dr. Francesco Audrino danke ich für die Übernahme der Drittbetreuung.

Meinen Kollegen und Freunden am Lehrstuhl für Ökonometrie und der Graduiertenschule danke ich für private und fachliche Diskussionen, Denkanstöße, Feedback, gegenseitige Hilfe und gemeinsames “Durchstehen”: Timo Dimitriadis, Christoph Frey, Roxana Halbleib, Phillip Heiler, Ekaterina Kazak, Verena Kretz, Jana Marecková, Marco Menner, Anastasia Morozova, Christian Neumeier, Audronė Virbickaitė und Aygul Zagidullina.

Ein ganz besonderer Dank gilt meiner Frau Marina und meiner Familie, die immer für mich da waren und die mich in allen Lebens- und Gemütslagen unterstützt und ertragen haben.

Table of Contents

Summary

This dissertation addresses the risk measures Value-at-Risk (VaR) and Expected Shortfall (ES) that have been used in finance for several years to assess the market risk of investments. The VaR is the maximum loss that will not be exceeded with a given probability (usually between 1% and 5%) over a target period (usually 1 to 10 days) and the ES is the expected (average) loss in case of an exceedance of the VaR. Statistically these two quantities are the quantile of a distribution and the expected value over all observations smaller than the VaR. Financial institutions use these two risk measures to manage bank-internal processes; but above all, the Basel Committee requires banks to stockpile capital reserves determined by these measures. The VaR used for this purpose is to be replaced by the ES at the end of 2019, as it solves some of the VaR's problems. However, the ES is clearly inferior to the VaR in other areas, which are among the topics addressed in this dissertation.

Two fundamental questions in this area of research are how to estimate and forecast these risk measures as precisely as possible, and how to validate ("backtest") our predictions. The three chapters of this thesis are therefore concerned with the estimation, forecasting and backtesting of the VaR and the ES. The methodological link between each article is the use of regression techniques for the corresponding functional of the underlying distribution function. Therefore, we model the conditional VaR or the pair consisting of VaR and ES as a function of covariates in the articles and use the regression models in different contexts.

The essays are independent research papers that I wrote during my doctoral studies at the University of Konstanz. I wrote the first article completely myself, the second and third articles were written together with Timo Dimitriadis. As usual in the related literature, the author or authors are always referred to as "we".

The first chapter, *Combining Value-at-Risk Forecasting Using Penalized Quantile Regression*, currently in press at *Econometrics and Statistics*, is concerned with the combination of VaR forecasts. In particular, here we consider the combination of a large number of predictions, as no risk model has been found so far, which consistently makes good

predictions. Rather, the optimal model depends on the data and can change at any time. Combining many predictions with a data driven choice of combination weights provides a good way to improve and robustify VaR forecasts. We estimate the weights of the standalone forecasts with quantile regressions, which are regularized with the Elastic Net. The main reason for this regularization is an almost perfect multicollinearity between the individual predictions, which leads to instability and overfitting of the weights, if estimated without penalty term. In the empirical application, we find that our method combines the individual VaR predictions more reliably than established methods from the literature. In particular, the hypothesis of the correctness of our combined forecasts is less frequently rejected by backtests, and our forecasts lead to lower values of the tick loss function, with which quantile predictions are often evaluated.

One of the main problems of the functional ES is that it is not *elicitable*, i.e. there is no loss function that is minimized by correct ES forecasts. Among other things, this property implies that it is not possible to directly estimate the parameters of an ES regression model. In a recent paper, however, Fissler and Ziegel (2016) show that the pair consisting of quantile and ES is jointly elicitable and introduce an associated class of loss functions. We use this class in the second chapter, *A Joint Quantile and Expected Shortfall Regression Framework*, to extend the linear quantile regression to the simultaneous modeling of the conditional quantile and the conditional ES. For the estimation of the regression parameters we propose an M- and a Z-estimator, prove the consistency for both and show the asymptotic distribution. Furthermore, we introduce several estimators of the covariance of the parameters, since this contains some nuisance quantities that are difficult to estimate. In an extensive simulation study, we compare several members of the underlying class of loss functions and conclude that, in particular, homogeneous variants are promising, which has since been confirmed in other papers. We also show that the M-estimator is clearly preferable to the Z-estimator, since the latter is unstable. To illustrate the many uses of our method, we simultaneously forecast the VaR and the ES based on realized variances and compare the predictions to those of a parametric and a non-parametric risk model.

In the third chapter, *Regression Based Expected Shortfall Backtesting*, we use the method we introduce in the second Chapter to propose new ES backtests. These are analogous to the well-known method of Mincer and Zarnowitz (1969), which is often used to evaluate predictions of the conditional mean. For our first test, we regress the returns on the ES predictions and use a Wald test to test whether the intercept and slope parameter of the ES regression equation are 0 and 1. In a second proposal we set the slope parameter to 1 and only test the estimated intercept with a t-test, which allows the definition of a one-sided hypothesis that is of interest to the financial authorities. Of particular note is that our backtests are the first to require only ES predictions as input parameters. All previous tests require at least VaR predictions, but often also other quantities such as predictions of the volatility or even the whole distribution function, which are not available for regulators. In extensive simulation studies, we show that our tests have an empirical size close to the chosen level of significance, especially if the tests are applied using the bootstrap procedure. In addition, our tests have good power, so they reliably detect wrong predictions, especially when compared to existing literature proposals. The existing tests fail several times to detect the misspecified predictions, whereas our tests can discriminate between correct and incorrect predictions in every situation examined. Furthermore, our one-sided test performs well in detecting too large ES forecasts and is therefore particularly relevant to financial regulators who are interested in having a sufficiently large capitalization of financial institutions.

References

Fissler, T. and J. F. Ziegel (2016). “Higher order elicitability and Osband’s principle”. *Annals of Statistics* 44 (4), 1680–1707 (see pp. 9, 12, 53–56, 58, 70, 71, 76, 95, 100).

Mincer, J. and V. Zarnowitz (1969). “The Evaluation of Economic Forecasts”. In: *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*. National Bureau of Economic Research, Inc, 3–46 (see pp. 9, 13, 95, 96, 99, 118).

Zusammenfassung

Diese Dissertation befasst sich mit den beiden Risikomaßen Value-at-Risk (VaR) und Expected Shortfall (ES), die seit einigen Jahren im Finanzwesen verwendet werden, um das Marktrisiko von Investitionen zu bewerten. Der VaR ist der maximale Verlust, der mit einer gewissen Wahrscheinlichkeit (meist zwischen 1% und 5%), über einen Zielzeitraum (üblicherweise 1 bis 10 Tage), nicht überschritten wird und der ES ist der erwartete (durchschnittliche) Verlust im Falle einer Überschreitung des VaR. Statistisch gesehen sind diese beiden Größen das Quantil einer Verteilung und der Erwartungswert über alle Beobachtungen kleiner dem VaR. Finanzinstitutionen verwenden diese beiden Risikomaße, um bankinterne Abläufe zu steuern; vor allem aber schreibt der Basler Ausschuss vor, dass Banken Kapitalreserven vorrätig halten müssen, die durch diese Maße bestimmt werden. Der hierfür bisher verwendete VaR soll ab Ende des Jahres 2019 durch den ES abgelöst werden, da dieser einige Probleme des VaR löst. Jedoch ist der ES dem VaR in anderen Bereichen klar unterlegen, die in dieser Dissertation unter anderem thematisiert werden.

Zwei grundlegende Fragen in diesem Forschungsbereich sind, wie sich diese Risikomaße möglichst genau berechnen bzw. vorhersagen lassen und wie wir unsere Vorhersagen validieren ("backtesten") können. Die drei Kapitel dieser Dissertation befassen sich daher mit der Messung, der Vorhersage und dem Backtesting des VaR und des ES. Das methodische Verbindungsglied zwischen den einzelnen Artikeln ist die Verwendung von Regressionstechniken für die entsprechenden Funktionale der zugrundeliegenden Verteilungsfunktion. Wir modellieren deshalb in den Artikeln jeweils den konditionalen VaR bzw. das Paar bestehend aus VaR und ES als eine Funktion von Kovariaten und verwenden die Regressionsmodelle in verschiedenen Kontexten.

Die Aufsätze sind eigenständige Forschungspapiere, die ich während meines Promotionsstudiums an der Universität Konstanz verfasst habe. Den ersten Artikel habe ich vollständig selbst geschrieben, der zweite und dritte Artikel sind gemeinsam mit Timo Dimitriadis

entstanden. Wie in der verwandten Literatur üblich, werden der Autor bzw. die Autoren stets als “wir” (“we”) bezeichnet.

Das erste Kapitel, *Combining Value-at-Risk Forecasts Using Penalized Quantile Regressions*, das derzeit bei *Econometrics and Statistics* im Druck ist, beschäftigt sich mit der Kombination von VaR Vorhersagen. Insbesondere betrachten wir hier die Kombination einer großen Anzahl an Vorhersagen, da bisher kein Risikomodell gefunden wurde, das durchweg gute Vorhersagen trifft. Vielmehr hängt das optimale Modell von den Daten ab und kann sich jederzeit ändern. Die Kombination vieler Vorhersagen durch eine datengetriebene Wahl der Kombinationsgewichte bietet daher eine gute Möglichkeit, die VaR Vorhersagen zu verbessern und zu robustifizieren. Die Gewichtung der einzelnen Vorhersagen schätzen wir über Quantilsregressionen, die mit dem Elastic Net regularisiert werden. Der Hauptgrund für diese Regularisierung ist eine fast perfekte Multikollinearität zwischen den einzelnen Vorhersagen, die zu einer Instabilität und Überschätzung (“overfitting”) der Gewichte führt, wenn diese ohne Bestrafungsterm geschätzt werden. In der empirischen Anwendung kommen wir zu dem Ergebnis, dass unsere Methode die einzelnen VaR Vorhersagen zuverlässiger kombiniert als etablierte Methoden aus der Literatur. Insbesondere wird die Hypothese der Korrektheit unserer kombinierten Vorhersagen seltener durch Backtests verworfen und unsere Vorhersagen führen zu geringeren Werten der Tick Loss Funktion, mit der Quantilsvohersagen oft bewertet werden.

Eines der Hauptprobleme des Funktionals ES ist dass es nicht *elicitable* ist, es also keine Verlustfunktion gibt, die durch korrekte ES Vorhersagen minimiert wird. Diese Eigenschaft impliziert unter anderem, dass es nicht möglich ist, die Parameter eines ES Regressionsmodells direkt zu schätzen. In einem aktuellen Paper zeigen Fissler und Ziegel (2016) jedoch, dass das Paar bestehend aus Quantil und ES gemeinsam elicitable ist und führen eine dazugehörige Klasse von Verlustfunktionen ein. Wir benutzen diese Klasse im zweiten Kapitel, *A Joint Quantile and Expected Shortfall Regression Framework*, um die lineare Quantilsregression auf die gleichzeitige Modellierung des konditionalen Quantils und des konditionalen ES zu erweitern. Für die Schätzung der Regressionsparameter schlagen wir einen M- und einen Z-Schätzer vor, beweisen für beide die Konsistenz und zeigen die asymptotische Verteilung. Weiterhin führen wir mehrere Schätzer der Kovarianz der Parameter ein, da diese einige Störgrößen beinhaltet, die schwer zu schätzen sind. In einer extensiven Simulationsstudie vergleichen wir mehrere Mitglieder der zugrundeliegenden Klasse von Verlustfunktionen und kommen zu dem Schluss, dass insbesondere homogene Varianten vielversprechend sind, was inzwischen auch in anderen Artikeln bestätigt wurde. Wir zeigen weiterhin, dass der M-Schätzer dem Z-Schätzer klar vorzuziehen ist, da letzterer instabil ist. Um die zahlreichen Einsatzmöglichkeiten unserer Methode zu illustrieren, prognostizieren wir gleichzeitig den VaR und den ES basierend auf realisierten Varianzen und vergleichen die Vorhersagen mit denen eines parametrischen und eines nicht-parametrischen Risikomodells.

Im dritten Kapitel, *Regression Based Expected Shortfall Backtesting*, verwenden wir die Methode, die wir im zweiten Kapitel einführen, um neue ES Backtests vorzuschlagen. Diese sind analog zu dem bekannten Verfahren von Mincer und Zarnowitz (1969), das häufig verwendet wird, um Vorhersagen des konditionalen Mittelwerts zu bewerten. Für unseren ersten Test regressieren wir die Renditen auf die ES Vorhersagen und testen mit einem Wald Test, ob der Interzept und Steigungsparameter der ES Regressionsgleichung 0 und 1 sind. In einem zweiten Vorschlag setzen wir den Steigungsparameter auf 1 und testen nur den geschätzten Interzept mit einem t-Test, was die Definition einer einseitigen Hypothese erlaubt die für die Finanzbehören interessant ist. Besonders hervorzuheben ist, dass unsere Backtests die ersten sind, die nur ES Vorhersagen als Inputgröße benötigen. Alle bisherigen Tests benötigen zumindest VaR Vorhersagen, oft aber auch weitere Größen wie Vorhersagen der Volatilität oder sogar die ganze Verteilungsfunktion, die für Regulatoren nicht verfügbar sind. In umfangreichen Simulationsstudien zeigen wir, dass unsere Tests eine empirische Size nahe am gewählten Signifikanzniveau haben, vor allem wenn die Tests mithilfe des Bootstrapverfahrens durchgeführt werden. Zudem haben unsere Tests eine gute Power, erkennen also zuverlässig falsche Vorhersagen, vor allem im Vergleich zu bestehenden Vorschlägen aus der Literatur. Den existierenden Tests gelingt es mehrmals nicht, die fehlspezifizierten Vorhersagen zu erkennen, wohingegen unsere Tests in jeder untersuchten Situation zwischen korrekten und falschen Vorhersagen unterscheiden können. Weiterhin zeigt unser einseitiger Test eine gute Leistung bei der Erkennung von zu großen ES Vorhersagen und ist daher besonders für Finanzregulatoren relevant, die an einer ausreichend großen Kapitalausstattung der Finanzinstitute interessiert sind.

Literatur

Fissler, T. and J. F. Ziegel (2016). “Higher order elicitability and Osband’s principle”. *Annals of Statistics* 44 (4), 1680–1707 (see pp. 9, 12, 53–56, 58, 70, 71, 76, 95, 100).

Mincer, J. and V. Zarnowitz (1969). “The Evaluation of Economic Forecasts”. In: *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*. National Bureau of Economic Research, Inc, 3–46 (see pp. 9, 13, 95, 96, 99, 118).

Chapter 1

Combining Value-at-Risk Forecasts Using Penalized Quantile Regressions

1.1. Introduction

Although difficult, it is important to decide between alternative Value-at-Risk (VaR) modeling and forecasting strategies. A poorly selected risk model may have drastic effects on banks and the economy as a whole, as evidenced during the previous financial crisis when many standard approaches predicted inadequately low levels of risk. Einhorn (2008) compares the VaR to “an airbag that works all the time, except when you have a car accident”. The VaR is defined as the worst possible loss over a target horizon that will not be exceeded with a given probability (Jorion, 2006). Therefore, VaR is a quantile of the distribution of returns over a horizon (usually one or ten days) for a given probability level (typically 1%). A major reason for its popularity is that the Basel Committee (1996, 2006, 2011) utilizes the VaR for calculation of the minimum capital requirements which banks need to keep as reserves to cover the market risk of their investments.

Extensive literature exists on how to estimate and predict VaR (see Kuester et al. (2006), Komunjer (2013) and Nieto and Ruiz (2016) for overviews). The primary issue with VaR forecasting, however, is that the models’ performance and reliability in accurately predicting the risk depends heavily on the data. While a parsimonious model might perform well in economically stable periods, it can fail tremendously during a volatile period. Likewise, highly parameterized models might be adequate during periods of high volatility, but can be easily outperformed by simpler approaches in less turbulent times. To date, no unique model or approach dominates throughout the existing VaR forecasting comparisons (see Kuester et al. (2006), Marinelli et al. (2007), Halbleib and Pohlmeier (2012), Abad and Benito (2013), Boucher et al. (2014), Louzis et al. (2014), Ergen (2015), Nieto and Ruiz (2016) and Bernardi and Catania (2016)). The key reasons for this finding are that the applied models are prone to suffer from model misspecification (e.g. through the application of an overly simplistic model) and estimation uncertainty (e.g. they imply a complicated estimation procedure). For a more detailed discussion of the risks and uncertainties involved in VaR forecasting, see Boucher et al. (2014).

If the best model is unknown or likely to change over time, a promising alternative to deciding on a specific risk model is to combine the predictions stemming from several approaches. In an overview on forecasting combinations, Timmermann (2006) provides three arguments in favor of combining forecasts for the stabilization and improvement of predictive performance upon standalone models: First, there are diversification gains stemming from the combination of forecasts computed from different assumptions, specifications or information sets. Second, combined forecasts tend to be robust against structural breaks. Third, the influence of potential misspecification of the individual models is reduced due to averaging over a set of forecasts stemming from several models.

Halbleib and Pohlmeier (2012) propose the combination of VaR forecasts using quantile regression (QR), introduced by Koenker and Bassett (1978), as the QR estimator minimizes the tick loss function. This asymmetric and piecewise linear loss function is consistent for quantiles, which implies that the true quantile prediction minimizes expected tick loss (Gneiting, 2011b). Therefore, it is reasonable to incorporate the tick loss for the estimation and evaluation of VaR forecasts. If someone aims at combining a large number of VaR forecasts, he or she likely faces the issue of multicollinearity among them, since they stem from the same data and similar mathematical approaches. This is the case in our empirical application: we observe high pairwise correlations among the forecasts (sometimes greater than 99%), which indicates the presence of severe multicollinearity. In this situation, the standard QR estimator is unstable: small variations in the data can lead to large changes in the estimated parameters. Moreover, it can overfit the data such that, for two highly correlated forecasts, we obtain a large positive weight for one and a large negative weight for the other. From an in-sample perspective, this is not problematic because the estimated coefficients still minimize the tick loss function. For out-of-sample purposes, however, such imprecisely estimated parameters can be harmful because the model fails to properly generalize to new data (Hastie, Tibshirani, and Friedman, 2011, p. 38). An obvious solution is to only combine forecasts with small to moderate cross-correlations. However, we aim to avoid manually selecting models over whose forecasts we average, and instead consider combination techniques that can withstand high correlations among the predictions.

In this paper, we propose penalized QR as a novel VaR combination technique. In particular, we consider regularization with the elastic net penalty of Zou and Hastie (2005), which represents a convex combination of the well-known least shrinkage and selection operator (lasso) of Tibshirani (1996) and the ridge penalization of Hoerl and Kennard (1970a,b). Due to the geometry of the penalty function, the elastic net simultaneously induces coefficient shrinkage and variable selection. These two properties allow for the combination of a large number of potentially highly correlated VaR forecasts, since the parameter estimates are stable, overfitting is reduced and variables are automatically selected.

Penalized QR feature a number of advantages over alternative combination techniques: (1) They perform a data-driven selection of forecasts: the coefficients of uninformative standalone models can be set to zero. This has the potential to improve the predictability, as only a subset of the available forecasts enters the combination. We explore this by comparing lasso and elastic net QR with ridge QR, which shrinks the coefficients but selects no variables. (2) They can cope with nearly collinear forecasts through the regularization of the estimated weights. (3) They include an intercept term for the purpose of bias correction (Halbleib and Pohlmeier, 2012). This is important when all standalone predictions systematically over- or underestimate the VaR: an intercept can shift the combined forecast outside the range of the standalone predictions. Simple averaging techniques, for instance the mean over all forecasts, do not include such an intercept term and are furthermore bound between the minimum and maximum of the standalone forecasts. (4) They minimizes the (penalized) tick loss function.

A range of other quantile combination techniques is proposed in the literature. Giacomini and Komunjer (2005) introduce a generalized method of moments (GMM) estimator aimed at the minimization of the tick loss function for the purpose of forecasting combination and encompassing tests. Halbleib and Pohlmeier (2012), in addition to QR, introduce a GMM estimator to determine the optimal combination weights by minimizing the conditional coverage test (Christoffersen, 1998). QR is further applied by Fuertes and Olmo (2013), who utilize it for combining VaR forecasts from intra- and inter-day models and for a conditional QR forecast encompassing test. QR forecast combination under a variety of restrictions is also explored by Jeon and Taylor (2013), who combine predictions stemming from the conditional autoregressive VaR class of models (Engle and Manganelli, 2004) with the predictions from an option implied volatility model. The mean and median over all standalone forecasts is considered by Huang and Lee (2013), who combine VaR predictions from models using high frequency information. McAleer et al. (2013a,b) combine VaR forecasts by the percentiles of their predictions with the goal of minimizing capital charges imposed by the Basel Accord. Shan and Yang (2009) and Casarin et al. (2013) introduce sequential combination approaches wherein the weights of the previously well performing models are increased and vice versa. While Shan and Yang (2009) assess the performance of the standalone models via tick loss, Casarin et al. (2013) evaluate the models with respect to the capital requirements imposed by the Basel Accords. An alternative route is described by Hamidi et al. (2015), who average VaR forecasts stemming from conditional autoregressive expectile models (Taylor, 2008a) and estimate the combination weights by optimizing the squared difference between the nominal probability and the hit rate, i.e. the share of times a VaR prediction is smaller than the realized return. Recently, Bernardi, Catania, and Petrella (2017) suggest filtering the standalone models with the model confidence set of Hansen et al. (2011) prior to averaging the forecasts.

In the empirical portion of this paper, we assess the performance of the proposed penalized QR combination method for a data set comprising 30 constituents of the Dow Jones Industrial Average Index (DJIA) over a horizon of 8 years. We compare the performance of the elastic net, lasso and ridge QR combined forecasts to a large variety of competing approaches. For forecast evaluation, we use backtesting via the dynamic quantile backtest of Engle and

Manganelli (2004) and the unconditional coverage backtest of Kupiec (1995). Furthermore, we compare the forecasts with the model confidence set of Hansen et al. (2011) to determine the approach that produces the most precise predictions. The results indicate that the penalized QR combined forecasts exhibit the lowest number of backtest rejections and the tick losses are comparably low. By splitting the evaluation sample into two subperiods, we additionally determine that during volatile periods, lasso and elastic QR net perform slightly better than ridge QR. This relationship reverses during calm periods. Additionally, we do not face the "forecast combination puzzle" (Stock and Watson, 2004), which states that simple approaches are difficult to outperform. In the combination of VaR forecasts, it appears as though complexity pays off.

The remainder of this paper is organized as follows. Section 1.2 introduces the methodology and provides details on penalized QR. Section 1.3 introduces the data set, evaluation horizons, the forecast evaluation method and the standalone models. Section 1.4 presents the results of the empirical application. Section 3.6 consists of a conclusion and an outlook on potential future research areas.

1.2. Methodology

Let the price of a financial asset or a portfolio at time t be P_t such that the logarithmic return from time t to $t + h$ is $r_{t+h} = \log(P_{t+h}/P_t)$. We denote the VaR forecast from t to $t + h$, conditional on all available information $\mathcal{F}_t$, as $q_{t+h|t}(\alpha)$. The VaR is defined as,

$$\alpha = \Pr\left(r_{t+h} \leq q_{t+h|t}(\alpha)|\mathcal{F}_t\right), \tag{1.1}$$

where $\alpha \in (0,\ 1)$ is the probability level. Throughout the paper, we focus on the probability level $\alpha = 1\%$ and the forecast horizon $h = 1$ day.

In the following, $q_{m,t+1|t}(\alpha)$ is the VaR forecast for day $t + 1$ of model $m = 1, \dots, M$ based on the information available at t and $\boldsymbol{q}_{t+1|t}(\alpha) = (q_{1,t+1|t}(\alpha), \dots, q_{M,t+1|t}(\alpha))'$ is the vector of all forecasts. The linear combination of the M forecasts, including an intercept term, is given by,

$$\begin{aligned} q^c_{t+1|t}(\alpha) &= \beta_{0,t}(\alpha) + \beta_{1,t}(\alpha)q_{1,t+1|t}(\alpha) + \ldots + \beta_{M,t}(\alpha)q_{M,t+1|t}(\alpha) \\ &= \beta_{0,t}(\alpha) + \boldsymbol{q}'_{t+1|t}(\alpha)\boldsymbol{\beta}_t(\alpha), \end{aligned} \tag{1.2}$$

where $\boldsymbol{\beta}_t(\alpha)$ is the quantile-specific vector of slope coefficients, which we loosely call the combination weights, even though the sum of the coefficients is not necessarily one. We explicitly incorporate an intercept term $\beta_{0,t}(\alpha)$ to correct a potential bias of misspecified standalone forecasts. If all standalone predictions systematically over- or underestimate the VaR, an intercept can shift the combined forecast outside the range of the standalone predictions (Halbleib and Pohlmeier, 2012). The time index of the coefficients indicates that the weights are time-varying. In order to incorporate the most recent information into the model parameters, we re-estimate the combination weights every day.

In order to determine the optimal combination weights, we assume that the loss of the forecaster only depends on the forecast error $u_{t+1}(\alpha) = r_{t+1} - \beta_{0,t}(\alpha) - \boldsymbol{q}'_{t+1|t}(\alpha)\,\boldsymbol{\beta}_t(\alpha)$. For quantiles, a consistent loss function is the asymmetric and piecewise linear tick loss (Giacomini and Komunjer, 2005; Gneiting, 2011b) given by,

$$\rho_\alpha(u) = \left(\alpha - \mathbb{1}_{\{u\leq 0\}}\right) u. \tag{1.3}$$

Consistency of the tick loss implies that the true quantile prediction minimizes the expected tick loss, a concept directly linked to the fact that the VaR is elicitable (Gneiting, 2011b). Equipped with a consistent loss function, the optimal forecast combination weights consequently minimize the expected loss of the forecast error,

$$\left(\beta^*_{0,t}(\alpha),\, \boldsymbol{\beta}^*_t(\alpha)\right) = \underset{\beta_{0,t}(\alpha),\,\boldsymbol{\beta}_t(\alpha)}{\arg\min}\ \mathrm{E}\left[\rho_\alpha\left(r_{t+1} - \beta_{0,t}(\alpha) - \boldsymbol{q}'_{t+1|t}(\alpha)\,\boldsymbol{\beta}_t(\alpha)\right) \mid \mathcal{F}_t\right], \tag{1.4}$$

which can be estimated by performing a linear QR of the standalone forecasts on the returns, as the tick loss lies at the heart of QR (Koenker and Bassett, 1978). We therefore obtain a consistent and asymptotically normal estimator of the combination weights by minimizing the average tick loss,

$$\left(\widehat{\beta}_{0,t}(\alpha),\, \widehat{\boldsymbol{\beta}}_t(\alpha)\right) = \underset{\beta_{0,t}(\alpha),\,\boldsymbol{\beta}_t(\alpha)}{\arg\min}\ \frac{1}{t}\sum_{\tau=0}^{t-1}\rho_\alpha\left(r_{\tau+1} - \beta_{0,t}(\alpha) - \boldsymbol{q}'_{\tau+1|\tau}(\alpha)\boldsymbol{\beta}_t(\alpha)\right), \tag{1.5}$$

which can then be used to form the combined forecast for the next day via $\widehat{q}^c_{t+1|t}(\alpha) = \widehat{\beta}_{0,t}(\alpha) + \boldsymbol{q}'_{t+1|t}(\alpha)\widehat{\boldsymbol{\beta}}_t(\alpha)$.

1.2.1. The Effect of Multicollinearity on Forecast Combinations

Although combination weights estimated via QR are optimal for an in-sample combination of the standalone forecasts (they minimize the tick loss), they might not be optimal for out-of-sample purposes. This is due to the almost perfect multicollinearity of the standalone forecasts (here denoted by X), which implies that the XX' matrix is close to singular. As the asymptotic distribution of the QR estimator depends on the inverse of XX' (Koenker and

Bassett, 1978), the variance of the QR estimator increases with the degree of correlation among the standalone forecasts.

In order to understand why a high variance of the combination weights can be an issue when combining forecasts, assume for the moment that we forecast the mean instead of the quantiles. Suppose we have a model of the form $Y = f(X) + \varepsilon$ with $\mathrm{E}[\varepsilon] = 0$ and $\mathrm{V}[\varepsilon] = \sigma^2$, where X are the covariates, Y is the dependent variable, f is a function of the data and $\widehat{f}$ is an estimate of f. Then, the expected prediction error under squared error loss is,

$$\mathrm{E}\left[(Y - \widehat{f}(X))^2\right] = \mathrm{E}\left[f(X) - \widehat{f}(X)\right]^2 + \mathrm{E}\left[(f(X) - \widehat{f}(X))^2\right] + \sigma^2, \tag{1.6}$$

which is the usual bias-variance tradeoff (e.g. Hastie, Tibshirani, and Friedman, 2011, p. 223). Thus, we see that an increase in the variance of $\widehat{f}$ (e.g. through multicollinearity of the covariates) increases the expected squared prediction error.

Such straightforward and general calculations in terms of mean and variance are available only for the mean squared error, but not for the tick loss function. James (2003) generalizes the bias-variance tradeoff to general symmetric loss functions, but the case of asymmetric loss functions, such as the tick loss, is to the best of our knowledge still unexplored. Nevertheless, this logic intuitively carries over to the tick loss: an increase of the variance of the estimated combination weights increases the expected tick loss of the prediction error, although the exact relation is unknown and is likely not as simple as in eq. (1.6).

We thus conclude that forecast combination is mainly beneficial if we can estimate the combination weights with a reasonable precision. The precision, however, correlates negatively with the degree of multicollinearity among the forecasts.

1.2.2. Elastic Net Penalized Quantile Regression

The elastic net penalty of Zou and Hastie (2005) represents a linear combination of the ridge penalty of Hoerl and Kennard (1970a,b) and the lasso of Tibshirani (1996). While the ridge term shrinks the coefficients towards zero, the lasso shrinks the coefficients and additionally selects variables. The automatic variable selection through the lasso is attractive as the weights of uninformative models can be set to zero. With highly correlated variables, however, the lasso tends to select one of the coefficients of the correlated variables randomly, whereas the ridge shrinks them towards each other (Zou and Hastie, 2005). In this case, the elastic net offers a compromise: similar to ridge, the elastic net shrinks the variables in groups and similar to lasso, it sets some coefficients to zero. Thus, the elastic net, combines the strengths of both approaches so that Zou and Hastie (2005) interpret the elastic net as a stabilized version of the lasso penalization. The QR estimator under elastic net penalization is given by

$$\left(\widehat{\beta}_{0,t}(\alpha,\,\lambda,\,\delta),\,\widehat{\boldsymbol{\beta}}_t(\alpha,\,\lambda,\,\delta)\right) = \underset{\beta_{0,t}(\alpha),\,\boldsymbol{\beta}_t(\alpha)}{\arg\min}\ \frac{1}{t}\sum_{\tau=0}^{t-1}\rho_\alpha\left(r_{\tau+1}-\beta_{0,t}(\alpha)-\boldsymbol{q}'_{\tau+1|\tau}(\alpha)\boldsymbol{\beta}_t(\alpha)\right)+ \\ +\lambda\left(\delta||\boldsymbol{\beta}_t(\alpha)||_1+(1-\delta)||\boldsymbol{\beta}_t(\alpha)||_2^2/2\right), \tag{1.7}$$

where λ is the regularization parameter and $\delta \in [0,\,1]$ balances the ridge and the lasso term, given by the sum of the absolute, respectively the sum of the squared parameters. We estimate (1.7) with the semismooth Newton coordinate descent algorithm proposed by Yi and Huang (2017), which is available through the R (R Core Team, 2016) implementation of Yi (2017) in the hqreg library.

If $\lambda \longrightarrow \infty$, eq. (1.7) simplifies to the intercept as it remains unpenalized. In this case, we simply estimate the empirical quantile of the returns. For $\lambda = 0$, eq. (1.7) reduces to unpenalized QR. Therefore, the value of λ controls the influence of the standalone predictions on the combined forecast. Considering the parameter δ, we obtain lasso QR for $\delta = 1$ and a pure ridge penalization for $\delta = 0$. As suggested by Hastie, Tibshirani, and Wainwright (2015, p. 57), we only estimate the λ parameter and consider preselected values of δ. In particular, we consider the three cases of $\delta = 0$ (ridge), $\delta = 1$ (lasso) and $\delta = 0.5$ (elastic net) in the empirical application.

Relation to Convex Weights

In the forecast combination literature (see e.g. Hansen, 2008; Timmermann, 2006), convexity is frequently imposed on the combination weights and this restriction typically improves the predictive performance upon the non-constrained estimator. Convex weights are non-negative and they sum to one, i.e. $0 \le \beta_m(\alpha) \le 1$, for $m = 1,\ldots,M$ and $\sum_{m=1}^{M}\beta_m(\alpha) = 1$. This particular restriction bears an interesting relation to the elastic net penalty, which we can see by rewriting the Lagrangian form of the elastic net QR given in eq. (1.7) in its restricted variant,

$$\left(\widetilde{\beta}_{0,t}(\alpha,\,\xi,\,\delta),\,\widetilde{\boldsymbol{\beta}}_t(\alpha,\,\xi,\,\delta)\right) = \underset{\beta_{0,t}(\alpha),\,\boldsymbol{\beta}_t(\alpha)}{\arg\min}\ \frac{1}{t}\sum_{\tau=0}^{t-1}\rho_\alpha\left(r_{\tau+1}-\beta_{0,t}(\alpha)-\boldsymbol{q}'_{\tau+1|\tau}(\alpha)\boldsymbol{\beta}_t(\alpha)\right) \\ \text{s.t.}\ \left(\delta||\boldsymbol{\beta}_t(\alpha)||_1+(1-\delta)||\boldsymbol{\beta}_t(\alpha)||_2^2/2\right)\le\xi, \tag{1.8}$$

where ξ is the regularization parameter for the restricted estimator. As usual, there is a one-to-one mapping between λ in eq. (1.7) and ξ in eq. (1.8).

If we now consider the case of lasso QR ($\delta = 1$) and we furthermore assume all slope coefficients to be non-negative, then eq. (1.8) collapses to convex QR if $\xi = 1$. Therefore, we can interpret the frequently imposed convexity constraint as a special case of the elastic net penalty, which is more general due to three reasons: (1) the combination weights are allowed to be negative; (2) the weights must not sum to one, as one can choose the value of the regularization parameter and (3) one can select the degree of sparsity that the model is enforcing by varying the balance between the ridge and the lasso terms.

1.2.3. Selection of the Regularization Parameter

The optimal shrinkage parameter for forecasting purposes is the value that minimizes the expected prediction error over the out-of-sample data. The in-sample tick loss, $\frac{1}{t}\sum_{\tau=0}^{t-1}\rho_\alpha(r_{\tau+1} - \widehat{\beta}_{0,t}(\alpha, \lambda, \delta) - \boldsymbol{q}'_{\tau+1|\tau}(\alpha)\widehat{\boldsymbol{\beta}}_t(\alpha, \lambda, \delta))$, can not be used as this loss decreases in λ. The standard approaches for estimating the parameter λ include information criteria and cross validation, which we discuss below. Additionally, we propose a computationally convenient heuristic rule based on the sum of the absolute combination weights.

Bayesian Information Criterion

The simplest approach for estimating the regularization parameter λ is via the Bayesian Information Criterion (BIC), which penalizes the in-sample loss. For the application of the BIC, we require a measure of the effective degrees of freedom of the model. In the case of lasso QR, Li and Zhu (2008) show that the effective degrees of freedom can be estimated by the number of non-zero coefficients, i.e. by $\text{df} = \sum_{m=1}^{M}\mathbb{1}_{\left\{\widehat{\beta}_{m,t}(\alpha,\lambda,\delta=1)\neq 0\right\}}$. Consequently, the BIC for lasso QR regression is given by,

$$\text{BIC}_t(\alpha, \lambda) = \ln\left(\frac{1}{t}\sum_{\tau=0}^{t-1}\rho_\alpha\left(r_{\tau+1} - \widehat{\beta}_{0,t}(\alpha, \lambda, \delta = 1) - \boldsymbol{q}'_{\tau+1|\tau}(\alpha)\widehat{\boldsymbol{\beta}}_t(\alpha, \lambda, \delta = 1)\right)\right) + \frac{\ln t}{2t}\text{df}, \tag{1.9}$$

and we determine the estimate of λ as the value that minimizes the BIC. This approach is implemented only for lasso QR, given that there is no similar approach available for elastic net and ridge QR.

Time Series Cross Validation

More appropriate for out-of-sample purpose is cross validation (CV) as it aims at minimizing the out-of-sample prediction error by evaluating a model on data that was not part of the estimation process. However, the two most common approaches, leave-v-out and K-fold CV, are not applicable in the present context. The reason is a violation of the fundamental assumption of CV that the estimation and evaluation samples are independent (Arlot and Celisse, 2010). Financial returns may be assumed to be at least uncorrelated, but they are neither independent nor identically distributed. Furthermore, VaR predictions exhibit high positive autocorrelation. In our application, the autocorrelations decrease only slowly, even after 250 days the autocorrelations of several forecasts are well above 50%.

In order to account for this dependence in the data, we employ the time series CV method of Hart (1994) which takes the form,

$$\mathrm{CV}_t(\alpha,\,\lambda,\,\delta) = \frac{1}{t - n_{\min}} \sum_{\tau=n_{\min}}^{t-1} \rho_\alpha \left(r_{\tau+1} - \widehat{q}^{\,c}_{\tau+1|\tau}(\alpha,\,\lambda,\,\delta) \right), \tag{1.10}$$

where $\widehat{q}^{\,c}_{\tau+1|\tau}(\alpha,\,\lambda,\,\delta) = \widehat{\beta}_{0,\tau}(\alpha,\,\lambda,\,\delta) + \boldsymbol{q}'_{\tau+1|\tau}(\alpha)\widehat{\boldsymbol{\beta}}_\tau(\alpha,\,\lambda,\,\delta)$ is the combined VaR prediction for $\tau + 1$ based on the information available up to τ and $n_{\min}$ is the minimum number of observations required to initially estimate the combination weights (we set $n_{\min}$ to 4 years in our application). In contrast to leave-v-out or K-fold CV, this approach only employs past information to predict future values and is robust to autocorrelation in the data (Hart and Lee, 2005). Eventually, for a given value of δ, we select λ by the value that minimizes the CV loss.

Heuristic Rule

Apart from the BIC for lasso QR and time series CV for lasso, ridge and elastic net QR, we propose a computationally convenient heuristic rule for selecting the regularization parameter λ. Our suggestion is to choose the most restricted model such that for a given value of δ, the sum of the absolute estimated weights (i.e. the L_1-norm) is smaller than some value s,

$$\widehat{\lambda} = \max \lambda, \qquad \text{s.t.} \sum_{m=1}^{M} |\widehat{\beta}_{m,t}(\alpha,\,\lambda,\,\delta)| \leq s. \tag{1.11}$$

The intuition for this rule stems from the fact that the elastic net penalty generalizes the convexity restriction, for which we find $s = 1$, given that $\beta_{m,t}(\alpha) \geq 0$ for $m = 1, \ldots, M$.

The suggestion given in eq. (1.11) therefore connects a generalized variant of the convexity constraint (the L_1-norm) with the regularization parameter λ.

To find reasonable values of s, we compute $\sum_{m=1}^{M} |\widehat{\beta}_{m,t}(\alpha, \lambda, \delta)|$ when λ is estimated with the time series CV procedure. It turns out that while the estimates of λ vary greatly depending on δ and the data, the values of $\sum_{m=1}^{M} |\widehat{\beta}_{m,t}(\alpha, \lambda, \delta)|$ are remarkably stable over the time, the asset space and even across the different values of δ. We find that most of the L_1-norms of the weights are in the range between (0.7, 1.1) with the majority of values at 0.8. In the empirical application, we therefore include predictions formed with the above rule and set $s = 0.8$. Furthermore, we provide a robustness check on the choice of s in which we show that actually a wide range of values of s yields precise predictions. This rule for selecting λ might be not optimal from a theoretical point of view, but on the practical side, it performs well empirically, it is robust, easy to implement and requires no computationally expensive CV procedure.

1.3. Empirical Application: Setup

In the empirical application, we compare the predictions of the penalized QR with forecasts of the standalone models and several competing combination approaches. This section outlines the data, the models to be combined, some competing combination techniques and the forecast evaluation methodology. The results are presented in Section 1.4.

1.3.1. Data and Evaluation Horizon

The dataset under consideration are the daily closing (dividend and split adjusted) prices of 30 constituents of the DJIA for a time horizon from January 2, 1996 to December 31, 2014, a total of 4784 days, which we obtained from Thomson Reuters Eikon. Note that the DJIA composition as of March 19, 2015 includes Goldman Sachs (GS) and Visa (V), which were only listed after 1996. Consequently, we replace these two stocks with two immediate predecessors, AT&T (T) and Hewlett Packard (HPQ). The symbols of the assets we analyze are thus: AAPL, AXP, BA, CAT, CSCO, CVX, DD, DIS, GE, HD, HPQ, IBM, INTC, JNJ, JPM, KO, MCD, MMM, MRK, MSFT, NKE, PFE, PG, T, TRV, UNH, UTX, VZ, WMT and XOM.

Figure 1.5 in the Appendix shows the log return series of the stocks and Table 1.2 presents the corresponding summary statistics, together with the ticker symbols and company names. The return series show volatility clustering, especially in the time around the dot-com bubble and in the time of the previous global financial crisis. Moreover, the returns exhibit excess kurtosis and non-zero skewness, the Jarque-Bera test strongly rejects normality of the log returns.

As we require data to estimate the standalone models, to estimate the regularization parameter and to combine the forecasts, our evaluation horizon spans the time from January 3, 2007 to December 31, 2014 (2014 days). Besides the full 8 years of data, we split the sample into two equally sized windows of 4 years each, as the first half of the overall sample is mainly driven by the financial crisis and is much more volatile compared to the second subperiod. The goal of this split is to evaluate the models under different volatility regimes, which we term the crisis and the calm period. For an illustration of the data and sample split, consider Figure 1.1, which shows the log returns of the equally weighted portfolio of the 30 return series. The light and dark gray areas depict the crisis period from January 3, 2007 to December 31, 2010 (1008 days), respectively the calm period from January 3, 2011 to December 31, 2014 (1006 days). Both areas taken together represent the overall period.

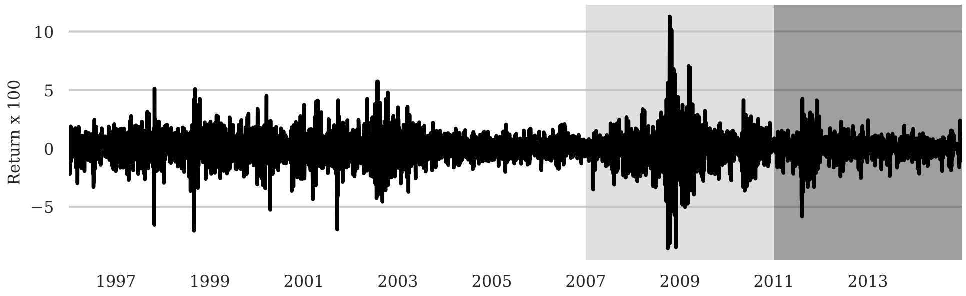


Figure 1.1: Equally weighted portfolio of the 30 assets included in the empirical application. The gray shaded areas indicate the forecast evaluation horizons January 3, 2007 to December 31, 2010 (1008 days, light gray), January 3, 2011 to December 31, 2014 (1006 days, dark gray) and January 3, 2007 to December 31, 2014 (2014 days, both areas).

1.3.2. Standalone Models

Our pool of models which we utilize to form the standalone VaR forecasts consists of 17 approaches. The selected models cover a wide range of frequently used parametric, semi-parametric and non-parametric techniques. While some of them are parsimonious, others are highly parametrized and can account for rich dynamics in the data.

Static Normal Distribution

This approach assumes that the returns are normally distributed with mean μ and variance σ^2. The quantile prediction for the next day is $q_{t+1|t}(\alpha) = \widehat{\mu} + \widehat{\sigma}\Phi(\alpha)^{-1}$, where $\Phi(\cdot)^{-1}$ is the inverse of the standard normal distribution and we estimate μ and σ^2 based on a rolling window of 250 observations.

Weighted Historical Simulation

The historical simulation (HS) approach predicts the next day's VaR by the empirical α-quantile of the past returns. While the standard HS weights all past days equally, the weighted HS technique of Boudoukh et al. (1998) uses a geometrically declining weighting scheme: more recent data points are more important for the prediction. The weight of day $\tau = t - w + 1, \ldots, t$ is $\eta_\tau = \eta^{\tau-1}(1-\eta)/(1-\eta^w)$, where w is the window length and we set $\eta = 0.99$. We estimate the empirical quantile of the HS, respectively the weighted HS approach using a rolling window of 250 observations.

RiskMetrics

The exponential smoothing RiskMetrics method (RiskMetrics Group, 1996) assumes the VaR forecast for day $t+1$ to be $q_{t+1|t}(\alpha) = \sigma_{t+1}\Phi^{-1}(\alpha)$, where $\sigma_t^2 = 0.06r_{t-1}^2 + 0.94\sigma_{t-1}^2$.

CAViaR Models

The conditional autoregressive VaR (CAViaR) class of models introduced by Engle and Manganelli (2004) assumes the VaR forecast to be a function of lagged VaR predictions and other explanatory variables. They propose the following four specifications,

Symmetric absolute value	(SAV)	$q_{t+1\|t}(\alpha) = \beta_0 + \beta_1 q_{t\|t-1}(\alpha) + \beta_2\|r_t\|,$
Asymmetric slope	(AS)	$q_{t+1\|t}(\alpha) = \beta_0 + \beta_1 q_{t\|t-1}(\alpha) + \beta_2(r_t)^+ + \beta_3(r_t)^-,$
Indirect GARCH(1, 1)	(IG)	$q_{t+1\|t}(\alpha) = (\beta_0 + \beta_1 q_{t\|t-1}^2(\alpha) + \beta_2 r_t^2)^{1/2},$
Adaptive	(AD)	$q_{t+1\|t}(\alpha) = q_{t\|t-1}(\alpha) + \beta_1\{[1+\exp(G[r_t - q_{t\|t-1}(\alpha)])]^{-1} - \alpha\},$

where $(x)^+ = \max(x, 0)$, $(x)^- = -\min(x, 0)$ and we set $G = 10$ as in Engle and Manganelli (2004). The estimation of the CAViaR models follows the procedure described in Engle and Manganelli (2004) using a rolling window of 1000 days.

GARCH Models

The remaining 9 models are all of the GARCH-type, i.e. we assume that returns can be decomposed into $r_t = \mu_t + \sigma_t z_t$. The component μ_t is the mean of the conditional distribution of r_t, σ_t is a volatility process and the innovation term z_t is independent and identically distributed with mean zero and unit variance. The VaR forecasts are $q_{t+1|t}(\alpha) = \mu_{t+1|t} + \sigma_{t+1|t}Q_\alpha(z_t)$, where $\mu_{t+1|t}$ and $\sigma_{t+1|t}$ are one-step-ahead forecasts of the mean, respectively the volatility and $Q_\alpha(z_t)$ is the unconditional α quantile of the innovations.

We assume that returns are not predictable and set the conditional mean to zero. For the volatility process, we assume either the standard GARCH(1, 1) of Bollerslev (1986), the exponential GARCH(1, 1) of Nelson (1991) or the asymmetric power ARCH(1, 1) of Ding et al. (1993), subsequently denoted by GARCH, EGARCH and APARCH. They are given by:

$$\begin{aligned}
&\text{GARCH}(1,1) && \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2,\\
&\text{EGARCH}(1,1) && \log\left(\sigma_t^2\right) = \omega + \alpha z_{t-1} + \gamma\left(|z_{t-1}| - \mathrm{E}\left[|z_{t-1}|\right]\right) + \beta\log\left(\sigma_{t-1}^2\right),\\
&\text{APARCH}(1,1) && \sigma_t^\delta = \omega + \alpha\left(|\varepsilon_{t-1}| - \gamma\varepsilon_{t-1}\right)^\delta + \beta\sigma_{t-1}^\delta.
\end{aligned}$$

In contrast to the standard GARCH, the EGARCH and APARCH specifications can respond asymmetrically with respect to positive and negative returns. The APARCH additionally accounts for the Taylor effect, the finding that the autocorrelation of absolute returns is typically larger than that of squared returns (Taylor, 1986).

For the innovations z_t we assume the normal distribution (abbreviated by N in the following), the Student-t distribution (t) and the filtered historical simulation (FHS) method of Barone-Adesi et al. (1999), which estimates $Q_\alpha(z_t)$ by the empirical quantile of the standardized returns. Combining the three variance processes with the three assumptions on the innovations, we obtain a total of 9 models. For estimation of the GARCH models we employ the rugarch library for R by Ghalanos (2015) and a rolling window of 1000 days.

1.3.3. Competing Combination Approaches

This section introduces a range of competing quantile combination approaches. Note that only the first two, the unpenalized and the convex QR, estimate an intercept term. For the seven others $\beta_{0,t} = 0$ and furthermore, the combined forecast of these approaches is bounded between the minimum and the maximum of the standalone predictions since their weights are non-negative and sum to one.

Unpenalized Quantile Regression

For a comparison with its penalized variants, we include the unpenalized QR estimator (which we estimate with the quantreg library by Koenker (2016)),

$$\widehat{\boldsymbol{\beta}}_t(\alpha) = \underset{\beta_{0,t}(\alpha),\,\boldsymbol{\beta}_t(\alpha)}{\arg\min}\ \frac{1}{t}\sum_{\tau=0}^{t-1}\rho_\alpha\left(r_{\tau+1} - \beta_{0,t}(\alpha) - \boldsymbol{q}_{\tau+1|\tau}'(\alpha)\boldsymbol{\beta}_t(\alpha)\right). \tag{1.12}$$

Convex Quantile Regression

Since we interpret the convexity constraint as a restricted variant of the elastic net, we include convex QR to evaluate whether the more general penalty is necessary for good forecast performance,

$$\widehat{\boldsymbol{\beta}}_t(\alpha) = \underset{\beta_{0,t}(\alpha),\boldsymbol{\beta}_t(\alpha)}{\arg\min}\ \frac{1}{t}\sum_{\tau=0}^{t-1}\rho_\alpha\left(r_{\tau+1} - \beta_{0,t}(\alpha) - \boldsymbol{q}'_{\tau+1|\tau}(\alpha)\boldsymbol{\beta}_t(\alpha)\right), \tag{1.13}$$
$$\text{s.t. } \beta_{m,t}(\alpha) \geq 0 \text{ for } m = 1,\dots,M \text{ and } \sum_{m=1}^{M}\beta_{m,t}(\alpha) = 1.$$

Simple Mean

Due to the simplicity and empirical success of this approach in the mean forecasting literature (Timmermann, 2006), we consider the simple average over all forecasts,

$$\widehat{\beta}_{m,t}(\alpha) = \frac{1}{M}, \quad \text{for all } m = 1,\dots,M. \tag{1.14}$$

Trimmed Mean

A trimmed variant of the simple mean combination is proposed by Timmermann (2006), which uses the relative rankings of the models to set the weight of certain models to zero. This method is supposed to be more robust than the simple mean as only the forecasts of the best performing models enter the combination. The weights are given by

$$\widehat{\beta}_{m,t}(\alpha) = \begin{cases} \frac{1}{\lfloor \eta M \rfloor}, & \text{if } R_t^m(\alpha) \leq \eta M \\ 0, & \text{else,} \end{cases} \quad \text{for all } m = 1,\dots,M, \tag{1.15}$$

where $R_t^m(\alpha)$ is the rank of model m at time t with respect to the sum of tick losses up to time t, given by $L_t^m(\alpha) = \sum_{\tau=0}^{t-1}\rho_\alpha(r_{\tau+1} - q^m_{\tau+1|\tau}(\alpha))$. We set $\eta = 0.25$ such that we average over the forecasts of the four previous best models.

Single Best

This approach assigns all weight to the previously best performing model,

$$\widehat{\beta}_{m,t}(\alpha) = \begin{cases} 1, & \text{if } R_t^m(\alpha) = 1 \\ 0, & \text{else.} \end{cases} \quad \text{for all } m = 1,\dots,M. \tag{1.16}$$

Inverse Loss

A further approach from the mean forecasting literature is to weight the forecasts inversely proportional with respect to their historical performance measured by the losses of the standalone models (Timmermann, 2006),

$$\widehat{\beta}_{m,t}(\alpha) = \frac{L_t^m(\alpha)^{-1}}{\sum_{n=1}^{M} L_t^n(\alpha)^{-1}} \quad \text{for all } m = 1, \ldots, M. \tag{1.17}$$

Inverse Rank

Timmermann (2006) suggests to weight the forecasts inversely proportional to their rank instead of the losses directly, as ranks are less sensitive to outliers than losses,

$$\widehat{\beta}_{m,t}(\alpha) = \frac{R_t^m(\alpha)^{-1}}{\sum_{n=1}^{M} R_t^n(\alpha)^{-1}} \quad \text{for all } m = 1, \ldots, M. \tag{1.18}$$

Optimizing the Hit Rate

Hamidi et al. (2015) propose to estimate the combination weights by minimizing Mallows's C_p (Mallows, 1973) on the squared difference between the nominal and empirical hit rates (the share of times the VaR is smaller than the return) subject to the convexity restriction on the weights, i.e.

$$\widehat{\boldsymbol{\beta}}_t(\alpha) = \arg\min C(\boldsymbol{\beta}_t(\alpha)), \quad \text{s.t. } \beta_{m,t}(\alpha) \geq 0 \text{ for } m = 1, \ldots, M \text{ and } \sum_{m=1}^{M} \beta_{m,t}(\alpha) = 1, \tag{1.19}$$

where $C(\boldsymbol{\beta}_t(\alpha)) = (\alpha - \widehat{\alpha}_t(\boldsymbol{\beta}_t(\alpha)))^2 + 2M \sum_{m=1}^{M} \beta_{m,t}(\alpha) s_m^2$. The term $\widehat{\alpha}_t(\boldsymbol{\beta}_t(\alpha))$ is the in-sample hit rate of the combination when using the weight $\boldsymbol{\beta}_t(\alpha)$, $\widehat{\alpha}_{m,t}$ is the hit rate of the mth model and $s_m^2 = (\alpha - \widehat{\alpha}_{m,t})^2/(t - M)$.

Sequential Relative Performance Approach

Shan and Yang (2009) propose a sequential method that is based on the relative historical performance of the standalone forecasts. Their approach takes the form,

$$\widehat{\beta}_{m,t}(\alpha) = \frac{\widehat{\beta}_{m,t-1}(\alpha) \exp\left(-\phi \rho_\alpha \left(r_t - q_{t|t-1}^m(\alpha)\right)\right)}{\sum\limits_{n=1}^{M} \widehat{\beta}_{n,t-1}(\alpha) \exp\left(-\phi \rho_\alpha \left(r_t - q_{t|t-1}^n(\alpha)\right)\right)} \quad \text{for } m = 1, \ldots, M, \tag{1.20}$$

where the initial weights are $\beta_{m,0}(\alpha) = 1/M$ for all $m = 1, \ldots, M$. In each period, this technique increases the weight of the models with low losses in the past and vice versa. We set the tuning parameter to $\phi = 1$ as this value performs best in the empirical application of Shan and Yang (2009).

1.3.4. Forecast Evaluation

We evaluate the VaR forecasts by two approaches. First, we test whether the VaR forecasts are by themselves valid or not, i.e. if the risk prediction is correct. For this purpose, we consider the unconditional coverage backtest by Kupiec (1995) and the conditional coverage backtest by Engle and Manganelli (2004). Second, we evaluate the precision of the forecasts by comparing the tick losses with the Model Confidence Set (MCS) of Hansen et al. (2011). This allows us to find the statistically most precise prediction, that is the model that produces the lowest tick losses.

Christoffersen (1998) terms a VaR forecast efficient with respect to the available information $\mathcal{F}_t$ if the hit variable $H_{t+1}(\alpha) = \mathbb{1}_{\{r_{t+1} \leq q_{t+1|t}(\alpha)\}}$ satisfies the property of correct conditional coverage given by $\mathrm{E}\left[H_{t+1}(\alpha)|\mathcal{F}_t\right] = \alpha$. If it is not possible to reject this hypothesis, we call the VaR forecast to be conditionally correct.

As the original likelihood ratio test of Christoffersen (1998) has inferior size and power properties compared to more recent alternatives (see Berkowitz et al., 2011), we test the hypothesis of correct conditional coverage of a VaR forecast with the dynamic quantile (DQ) backtest of Engle and Manganelli (2004). For the DQ test we estimate the equation

$$H_{t+1}(\alpha) - \alpha = \gamma_0 + \gamma_1 H_t(\alpha) + \gamma_2 q_{t+1|t}(\alpha) + u_{t+1}, \tag{1.21}$$

with least squares. The choice of the regressors is as in Berkowitz et al. (2011), who assess the size and power properties of a wide variety of backtests. The actual backtest is then the Wald test for $\gamma_0 = \gamma_1 = \gamma_2 = 0$, which is asymptotically χ^2_3 distributed.

In addition to the DQ test, we also test the unconditional coverage hypothesis given by $\mathrm{E}\left[H_{t+1}(\alpha)\right] = \alpha$. Tests for this hypothesis thus examine whether the average of the hit variable coincides with the nominal quantile level, without taking the possibility of clustered violations into account. Here, we utilize the frequently used likelihood ratio test of Kupiec (1995), which tests whether $H_{t+1}(\alpha)$ is Bernoulli distributed with success probability α.

For the relative evaluation of the forecasts, we compare the tick losses over the evaluation period by the MCS, similar to McAleer et al. (2013a) and Bernardi and Catania (2016). The MCS procedure repeatedly evaluates the hypothesis $\mathrm{E}\left[d_{ij}\right] = 0$ for all $i, j = 1, \ldots, M$, where d_{ij} is the loss differential between the predictions of model i and model j. Whenever it is possible to reject the hypothesis of equal predictive ability among all forecasts, the worst performing model (with respect to the losses) is eliminated and the procedure starts anew. This approach terminates with a set of models that statistically can not be further distinguished at a certain significance level.

For computation of the MCS we use the ARCH package for Python by Sheppard (2017). We report results for the T_R statistic (Hansen et al., 2011), based on 100,000 repetitions of the moving block bootstrap with a block size of 10 days to account for the possibility of clustered VaR hits. We also check the results for block sizes of 5 and 20 days and find the results to be robust with respect to the choice of the block length. Note that Hansen et al. (2011) express concerns about the validity of the assumption of stationarity loss differentials d_{ij} when the model parameters are recursively estimated. In order to account for this concern, we perform unit-root tests on the loss differentials and do not find evidence against stationarity.

1.4. Empirical Results

1.4.1. Estimation Window of the Penalized Quantile Regressions

Apart from the value of the regularization parameter and the balance between lasso and ridge, we need to decide on the length of the estimation window for the penalized QR estimators. In order to determine the optimal estimation window for lasso, elastic net and ridge QR, we compare the out-of-sample predictive performance depending on the window length used for the estimation of the parameters when we hold the regularization parameter λ fixed. For each stock $i = 1, \ldots, N$, rolling window sizes $w = 250, 500, 1000, 1500$ and a recursively extending window starting in January 3, 2000, we compute the average tick loss over the out-of-sample window with size R,

$$\mathrm{TL}_{i,w}(\alpha, \lambda, \delta) = 1/R \sum_{t=T}^{T+R-1} \rho_\alpha \left(r^i_{t+1} - \hat{q}^{c,i}_{t+1|w}(\alpha, \lambda, \delta) \right), \tag{1.22}$$

where r^i_{t+1} is the return of stock i at time $t+1$ and $\hat{q}^{c,i}_{t+1|w}(\alpha, \lambda, \delta)$ is the penalized QR combined VaR forecast of stock i for day $t+1$ based on a window of data w. Here, the out-of-sample period spans the overall evaluation period from January 3, 2007 to December 31, 2014 ($R = 2014$).

For a simple interpretation of the predictive performance, we average over the assets to get a single number per shrinkage value and window length,

$$\overline{\mathrm{TL}}_w(\alpha, \lambda, \delta) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{TL}_{i,w}(\alpha, \lambda, \delta), \tag{1.23}$$

and thereby obtain a measure of the average precision of the penalized QR estimators. Figure 1.2 shows the average tick loss for ridge ($\delta = 0$), elastic net ($\delta = 0.5$) and lasso ($\delta = 1$) QR for regularization parameters λ between 10^{-5} and 10^{2}. We can see that all loss curves reach their minimum within the considered grid of shrinkage values, which implies that neither the empirical quantile of the data ($\lambda \rightarrow \infty$) nor the unpenalized QR estimator ($\lambda \rightarrow 0$) is optimal. Considering these two extreme cases, we see that the empirical quantile is best estimated with short windows, while less penalized models profit from longer estimation samples. When we consider the minimum of the five loss curves per panel, we see that the minimum loss is decreasing in the length of the estimation window. Therefore, it is reasonable to use all available information for the estimation of the combination weights and in what follows, we estimate the penalized QR with the recursively extending window approach. For a fair comparison with the competing combination approaches, we apply them with the same window of data.

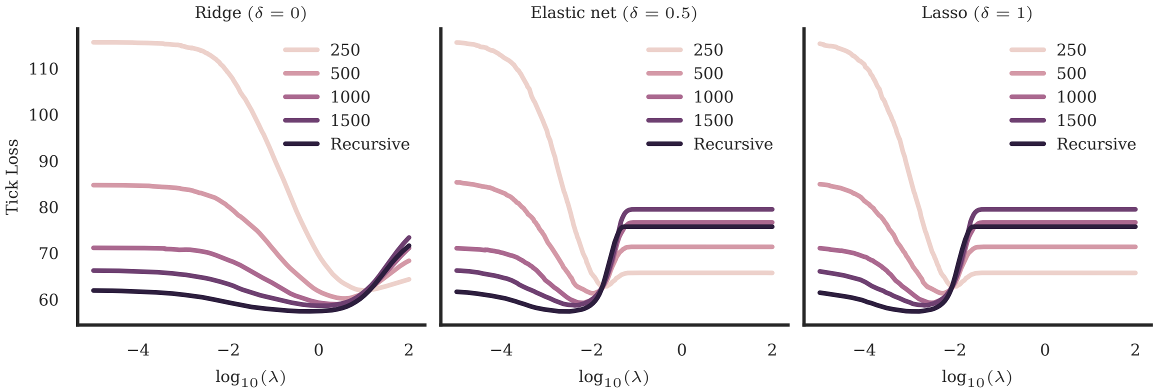


Figure 1.2: Average tick loss x 10^5 over all 30 assets for ridge, elastic net and lasso QR for a grid of values for the regularization parameter λ between 10^{-5} and 10^{2}. Each of the three panels shows the tick losses for a variety of rolling window sizes and a recursively extending window.

1.4.2. Conditional Coverage Backtesting

After having decided on an estimation strategy for the penalized QR, we start the discussion of the forecast comparison results by evaluating the standalone and the combined predictions by conditional coverage backtesting. Since presenting detailed tables with p-values of the backtest for all 30 assets is not feasible, we condense the results by presenting the number of times the forecasts are rejected at certain significance levels. We consider two significance levels. First, we record whether the hypothesis of correct conditional coverage is rejected at the 1% significance level, which indicates severe evidence against the validity of the forecast. We call this a severe rejection. Second, we check if the p-value of the backtest is between 1% and 10%, which could indicate either a valid or a non-valid prediction and we call this a mild rejection. Thus, a good risk model should produce as few severe and mild rejections as possible.

The results for all evaluation periods are presented in Figure 1.3; one panel for each of the three out-of-sample horizons. Each of these panels shows stacked bar plots with the number of backtest rejections at the two significance levels. The red and orange bars denote the number of severe and mild rejections of the hypothesis of a correct VaR forecast. Since we are aggregating the test decisions over all considered assets, the number of rejections can be at most 30 for each risk model and evaluation period.

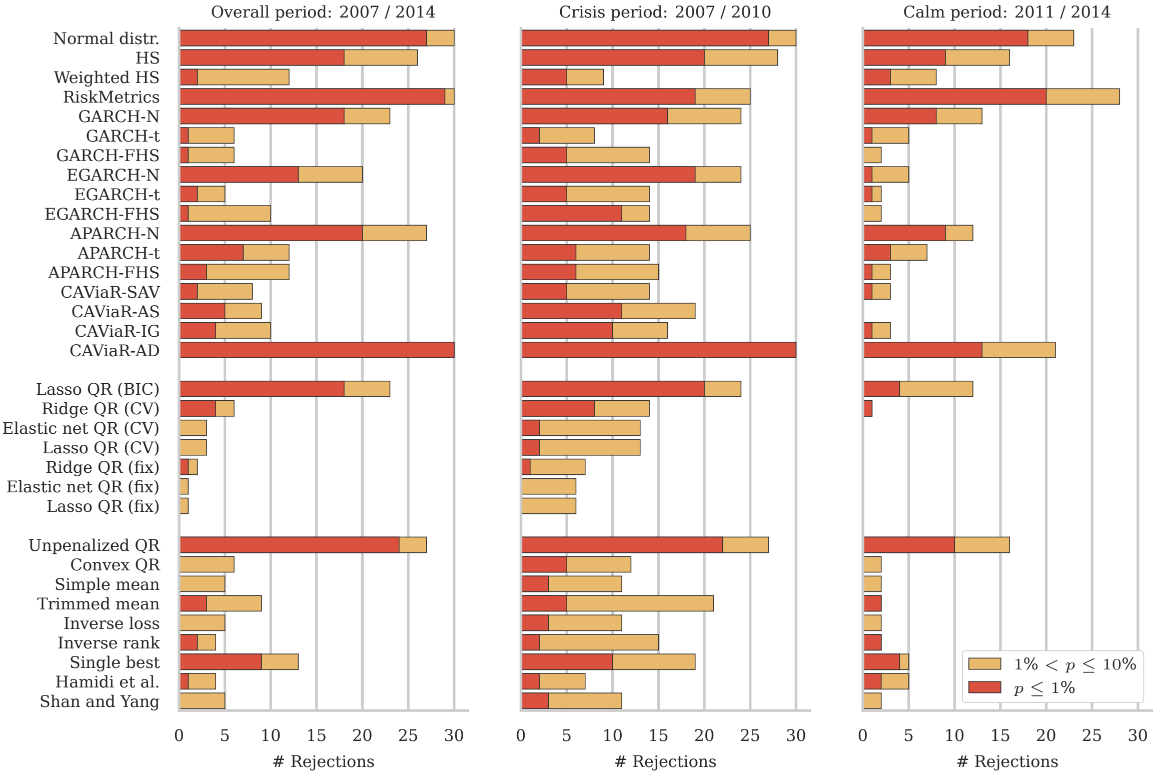


Figure 1.3: Number of dynamic quantile backtest rejections for all approaches at two different significant levels indicated by the colored bars. The three panels show the results for the overall / crisis / calm period, respectively. The empty lines separate the standalone models, the penalized QR and the competing combination techniques.

The first panel shows the number of rejections during the overall evaluation period from January 2007 to December 2014. We can see that the standalone models are often and highly rejected. Especially the HS and RiskMetrics, which are particularly popular in practice, are among the models that are most frequently rejected. The best standalone models are the GARCH models using the t-distribution and the FHS method with one severe and five mild rejections. Considering the combination approaches, we find that with the exception of the unpenalized QR and lasso QR with BIC estimated regularization parameters, the combined forecasts are less often rejected than the standalone predictions. Thus, implementing forecast combination is generally beneficial for VaR prediction and it is a major improvement upon the standalone models.

In evaluating the different combination approaches, we find that the unpenalized QR often fails to produce valid VaR forecasts. The reason for its poor performance is the previously discussed multicollinearity among the forecasts: the unpenalized QR overfits the data and the weights are unstable. Imposing the convexity restriction on the QR estimator improves the predictions, which already indicates that regularizing the QR estimator is beneficial. From the other competing combination approaches, we find that trimming the models prior to averaging them leads to more backtest rejections than the simple mean, even though the trimmed mean is supposed to improve the predictions upon its simpler variant. Also averaging over the inverse of the ranks of the models instead of the inverse of the tick losses does not improve the predictive performance. Thus, two conclusions from the mean forecasting literature (Timmermann, 2006), namely that trimming and averaging based on ranks improve upon the simpler variants, do not apply in our comparison. Moreover, selecting a single model on a day-by-day basis performs worse than the averaging techniques and even worse than many of the standalone models, which is in line with the findings in Aiolfi and Timmermann (2006). The quantile-specific combination approaches of Shan and Yang (2009) and Hamidi et al. (2015) exhibit roughly the same number of rejections as the other competing combination approaches and thus do not improve upon simpler combination techniques.

Considering the penalized QR in more detail, we find that the forecasts of lasso with BIC selected shrinkage values are nearly as often rejected as the unpenalized QR. The reason is that the BIC induces an insufficient amount of shrinkage such that the predictions are too similar to those of the unpenalized QR, which is in line with the findings in Koenker (2011). We furthermore find that lasso and elastic net QR are less often rejected than ridge QR, independent of the approach of selecting the shrinkage parameter. Thus, the sparsity enforcing property of the lasso operator is crucial for good predictions in this period. When we compare the proposed heuristic rule (denoted by fix) and the time series CV (denoted by CV) approaches for selecting the regularization parameter, we see that the heuristic rule leads to slightly less rejections than the CV, which demonstrates the robustness of the proposed way of estimating the regularization parameter. In fact, lasso and elastic net QR with the heuristic rule are just once mildly rejected in the overall period, much less than all other approaches.

Next, we evaluate the forecasts during the first half of the overall sample from, i.e. January 3, 2007 to December 31, 2014. This sample represents a period of high volatility around the 2007 – 2008 global financial crisis. As predicting the risk in volatile periods is difficult (Halbleib and Pohlmeier, 2012), we find in the second panel of Figure 1.3 that the number of backtest rejections generally increases in comparison to the overall period. The number of rejections rises especially for the standalone models, which are now at least 5 times severely rejected. The only exception is the GARCH model with t-distributed innovations, which is 2 times severely and 6 times mildly rejected. Likewise, the forecasts of the combination approaches are more often rejected, which is a direct consequence of the poor performance of the standalone models during this time. Nevertheless, we find that certain approaches still perform well. In particular, the only approaches that are never severely rejected are lasso and elastic net QR with the regularization amount estimated by the heuristic rule. In contrast to that, the competing combination approaches are at least twice severely rejected. Using the penalized QR it is thus possible to obtain VaR forecasts that exhibit a good performance even during this crisis period.

The last panel shows the results for the relatively calm period from January 3, 2011 to December 31, 2014. During this time, the volatility is lower and, thus, the VaR is easier to predict than during the crisis or the overall period. Therefore, we see that several standalone models are now hardly rejected at all (e.g. GARCH-FHS, EGARCH-FHS or CAViaR-AS) and exhibit a good performance. Nevertheless, the techniques involving the Normal distribution are still often rejected, raising doubt against the validity of this assumption, even in relatively calm times. Since most standalone forecasts are quite good, we furthermore find that most competing combination approaches perform well: even very simple techniques (e.g. the simple mean) are just twice rejected. In contrast to that, the unpenalized QR is still rejected frequently. When we evaluate the penalized QR in more detail, we find that with the exceptions of lasso QR (with shrinkage values selected by the BIC) and ridge QR (with CV selected shrinkage values), there is not a single severe or mild rejection of the conditional coverage hypothesis. Thus, the penalized QR once again exhibit less rejections of the conditional coverage hypothesis than the competing standalone and combination approaches. The proposed heuristic in particular performs well since none of the forecasts of lasso, ridge or elastic net QR are rejected with the amount of shrinkage selected using this rule.

1.4.3. Hit Ratios: Unconditional Coverage

We proceed by showing details on the hit rates of the VaR forecasts and the unconditional coverage backtest results of Kupiec (1995). Figure 1.4 displays the empirical hit rates, i.e. the share of times the predicted VaR is smaller than the return. If the VaR forecast is correct, the empirical hit rate should be close to 1%. In each of the panels for the three time periods, the 30 gray dots represent the empirical hit rates in percentage points for the different risk models. The gray line is the 1% value and the two black lines depict the 99% confidence interval of the unconditional coverage backtest. Occasionally, the CAViaR-AD model produces hit rates larger than 4%. For the sake of clarity of Figure 1.4, we do not show these outliers.

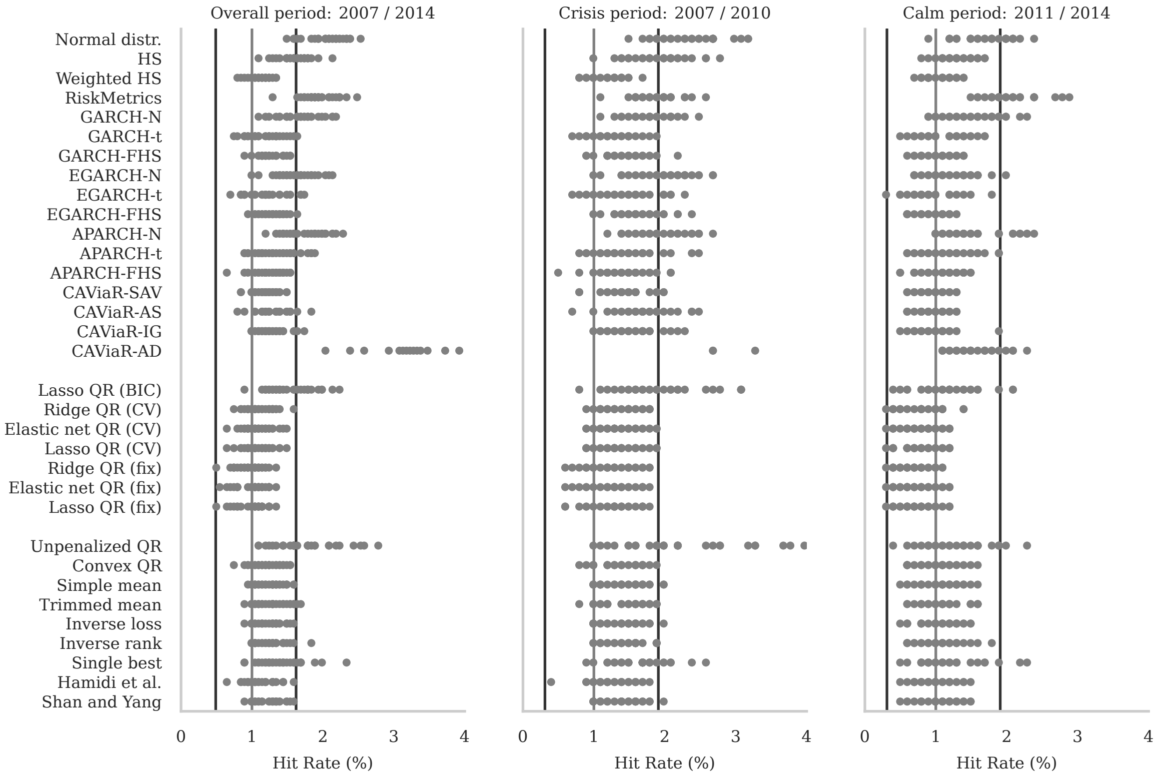


Figure 1.4: Empirical VaR hit rates. The figure is split into panels for the overall / crisis / calm period. In each of them, every dot represents the empirical hit rate in percentage points for one of the assets. The gray line denotes the nominal value of 1% and the two black lines indicate the 99% confidence interval for the unconditional coverage test of Kupiec (1995). This figure only shows empirical hit rates smaller than 4% in order to improve the presentation of the results.

In this figure, we can see that most standalone models tend to produce hit rates larger than 1%, i.e. they underestimate the true risks. Many of the hit rates are moreover outside the confidence bands and are thus rejected by the unconditional coverage backtest at the 1% level. A notably exception is the weighted HS approach which is the only standalone model whose hit rates are always within the 99% confidence interval.

Several combination approaches perform well with respect to the unconditional coverage criterion. For instance, the approach Hamidi et al. (2015) is always within the confidence interval, as this technique estimates the combination weights by optimizing the hit rate. Likewise, the penalized QR perform very well as almost always, the hit rates are within the 99% confidence interval for all three evaluation periods. The hit rates of the penalized QR are moreover centered around the 1% value, i.e. there is no tendency to underestimate the risks. This is in contrast to the combination approaches not involving a bias correcting intercept term which exhibit hit rates usually larger than 1% as the hit rates of the standalone forecasts are mostly larger than 1%. Thus, the bias correcting intercept term of the QR in combination with the regularization helps to obtain precise VaR forecasts.

1.4.4. Relative Evaluation of all Forecast Approaches

Besides the evaluation of the forecasts via backtesting, we next present the results of the MCS. The application of the MCS on the tick losses of the forecasts yields a p-value for each of the models and forecast horizons, which can be used to decide whether some model is in or out of the superior set of models (SSM). For the evaluation, we count for how many assets a model is included in the 90% and the 75% SSM, so that it is statistically not possible to distinguish between the models at the 10%, respectively the 25% significance level. Thus, the more often a model is within the SSM, the better is its predictive power. Like Grigoryeva et al. (2017), we additionally provide the average of the 30 MCS p-values. The larger this average p-value, the higher a model is ranked by the MCS. Table 1.1 contains the results and consists of one panel per evaluation horizon. In each of them, the first and the second column are the number of times a model is included in the 90%, respectively the 75% SSM. The third column contains the average MCS p-value.

Overall, the results of the MCS are less decisive than the backtesting, as typically a large number of models is included in the 90% and 75% SSM. This is similar to the findings of Bernardi and Catania (2016), who can only eliminate a small number of models using the MCS. Nevertheless, we can find some differences between the different risk models.

Regarding the standalone models, we find that some approaches (the Normal Distribution, HS and CAViaR-AD) perform much better during the calm period than during the crisis and the overall period. For instance, the CAViaR-AD is just 4 times within the 75% SSM during the overall and crisis period, but 29 times during the calm period. We furthermore find that the RiskMetrics model is frequently included in the SSM, although both backtests regularly reject its VaR forecasts. In contrast to that, the Weighted HS approach is not often in the SSM but is much less often rejected than the forecasts of RiskMetrics. Thus, for some standalone models there seems to be a tradeoff between absolute and relative accuracy of the forecasts. However, for the forecast combination approaches, we do not face a similar tradeoff since most of them exhibit a good absolute and relative performance.

If we inspect the competing combination forecasts more closely, we find that they perform relatively well throughout all three horizons. They are often included in the SSM and their average MCS p-values are high. Two exceptions are the unpenalized QR and the approach of Hamidi et al. (2015), which are less often in the SSM and exhibit lower average

Table 1.1: Relative Comparison of all forecasting approaches

	Overall period			Crisis period			Calm period		
Approach	$\#_{90\%}$	$\#_{75\%}$	$\overline{p}_{\text{MCS}}$	$\#_{90\%}$	$\#_{75\%}$	$\overline{p}_{\text{MCS}}$	$\#_{90\%}$	$\#_{75\%}$	$\overline{p}_{\text{MCS}}$
Normal Distr	25	17	0.32	24	19	0.34	30	30	0.77
HS	20	9	0.27	24	12	0.31	29	27	0.63
Weighted HS	16	13	0.31	18	15	0.42	17	13	0.30
RiskMetrics	30	28	0.72	30	30	0.82	30	29	0.64
GARCH-N	30	27	0.68	30	26	0.70	30	29	0.75
GARCH-t	28	26	0.74	30	26	0.75	27	24	0.66
GARCH-FHS	28	25	0.62	26	25	0.63	27	23	0.61
EGARCH-N	30	30	0.83	30	29	0.76	30	30	0.92
EGARCH-t	30	30	0.90	30	30	0.87	30	29	0.88
EGARCH-FHS	29	28	0.77	29	28	0.73	29	29	0.81
APARCH-N	30	29	0.72	30	29	0.76	30	30	0.78
APARCH-t	30	29	0.85	29	29	0.84	29	29	0.79
APARCH-FHS	26	23	0.60	25	23	0.57	29	27	0.69
CAViaR-SAV	29	27	0.64	29	25	0.67	28	26	0.69
CAViaR-AS	30	29	0.67	30	27	0.62	29	29	0.84
CAViaR-IG	28	26	0.65	29	25	0.67	27	23	0.66
CAViaR-AD	9	4	0.10	11	4	0.12	30	29	0.71
Lasso QR (BIC)	27	24	0.54	27	24	0.60	28	23	0.59
Ridge QR (CV)	29	29	0.76	29	28	0.76	30	27	0.67
Elastic net QR (CV)	30	30	0.77	30	30	0.80	29	23	0.62
Lasso QR (CV)	30	30	0.79	30	30	0.80	27	23	0.63
Ridge QR (fix)	30	30	0.84	30	29	0.82	29	28	0.70
Elastic net QR (fix)	30	30	0.85	30	30	0.85	29	26	0.72
Lasso QR (fix)	30	30	0.86	30	30	0.88	27	25	0.68
Unpenalized QR	24	19	0.37	28	22	0.46	26	24	0.55
Convex QR	30	29	0.78	30	28	0.77	27	24	0.65
Simple Mean	30	29	0.84	29	27	0.75	30	29	0.84
Trimmed Mean	30	29	0.79	29	28	0.75	29	27	0.75
Inverse Loss	29	29	0.77	27	25	0.63	30	29	0.81
Inverse Rank	30	30	0.83	30	28	0.76	30	30	0.82
Single Best	30	27	0.71	30	29	0.69	29	28	0.76
Hamidi et al	24	22	0.58	25	22	0.65	26	25	0.59
Shan and Yang	30	30	0.84	29	27	0.74	30	29	0.83

This table presents the results of the model confidence set over all 30 assets. $\#_{90\%}$ and $\#_{75\%}$ are the number of times a model is included in the 90%, respectively 75% SSM and $\overline{p}_{\text{MCS}}$ is the average over the 30 individual MCS p-values based on the T_R statistics using 100,000 iterations of the moving block bootstrap with a block length of ten days.

MCS p-values in comparison to the other competitors. In line with the existing literature on forecasting combination, we furthermore find that the simple mean over all forecasts performs quite well and often performs as good as more sophisticated approaches.

When we evaluate the penalized QR combinations in more detail, we find that most of them are almost always included in the 75% and 90% SSM and achieve high average MCS p-values. Therefore, the penalized QR exhibit a good relative accuracy in addition to the excellent backtest results. The only approach that performs not as good as the others is the lasso QR when the shrinkage values are estimated via the BIC, as the BIC induces insufficient shrinkage. Comparing the time series CV with the proposed heuristic rule, we find that the number of times the models are within the SSM are comparable for both ways of estimating the shrinkage parameter. However, the average MCS p-values are larger for the heuristic rule and, therefore these models are higher ranked by the MCS. We furthermore find that during the crisis time, the lasso and elastic net QR is more often included in the 75% and 90% SSM than the ridge QR and their average MCS p-values are larger. That once more indicates that penalized QR performs better when the model is allowed to set certain weights to zero in times when many standalone models perform poorly. During the calm time, this relation reverses: the forecasts of ridge QR are more often included in the SSM. We can thus conclude that the variable selection property of the elastic net and lasso penalties is especially important in volatile times when many of the standalone models fail. When all models perform well (for instance during the calm time), shrinkage without a variable selection suffices to obtain precise predictions.

Summing up the results from backtesting and the relative comparison via the MCS, we find that: (1) The conditional coverage hypothesis is less often rejected for the penalized QR than for the standalone models and the competing combination forecasts. (2) The hit rates of the penalized QR combined forecasts are close to the nominal value of 1% and are hardly rejected by the unconditional coverage test. (3) The regularized QR are often included in SSM and the average MCS p-values are high, which indicates a good relative performance. (4) The proposed heuristic rule performs well and even better than the time series CV. (5) The differences in the performance of lasso, elastic net and ridge QR are rather small, yet the former two perform slightly better during the crisis time and the overall period. (6) Contradictory to the fact that elastic net QR should be superior to ridge and lasso QR as it combines the strengths of both, we do not find that elastic net QR performs better than lasso QR on its own.

1.4.5. Robustness Check: The Heuristic Rule

The empirical comparison reveals that the proposed heuristic rule performs very well with respect to backtesting and tick losses. In order to demonstrate that the results are not simply due to a fortunate choice of the parameter of the heuristic rule, we verify the robustness of the proposition by examining the results when we vary the maximum allowed L_1-norm of the weights, i.e. the parameter s in eq. (1.11).

Figure 1.6a shows the number of 1% conditional coverage backtest rejections for the VaR forecasts of ridge, lasso and elastic net QR for values of s between 0.5 and 1.5 and all three evaluation periods. For a value of s in the range from 0.75 to 1, we find just up to two severe rejections of the conditional coverage hypothesis for all models and periods. Furthermore, for many values there is not even a single rejection. Figure 1.6b presents the tick loss averaged over all 30 assets. In this Figure, can see that the tick loss curves are relatively flat in the considered region of values for s. Thus, the relative performance of the penalized QR is hardly influenced by the choice of the parameter s.

These two findings confirm the robustness and the performance of the proposed way of selecting the regularization parameter of the penalized QR. Therefore, a time consuming cross validation is not essential for good forecast performance of the penalized QR.

1.4.6. Combination Weights and Relative Importance of the Predictors

In order to get some intuition into how the penalized QR estimates the combination weights and selects the standalone models, Figure 1.7 shows the weights for the VaR forecasts of the AT&T stock as an illustration. The three panels show the estimated weights and intercepts for lasso, elastic net and ridge QR over the period from January 3, 2007 to December 31, 2014. For that particular stock, the models RiskMetrics and APARCH-N dominate the estimated weights of lasso and elastic net QR. The weights of the other standalone models are comparably small, so that the lasso and the elastic net penalty set the coefficients of 15 out of 17 variables to zero or almost zero, see Figures 1.7a and 1.7b. When we look at the estimated weights of ridge QR (Figure 1.7c), we find that the weights are very similar for the different standalone models. This reveals the grouping effect of the ridge penalty: the coefficients of highly correlated variables are shrunk towards each other. For all three penalized QR, the combination weights are furthermore relatively stable over time, there is not much variation in the choice of the standalone models or in the estimated weights.

For a more complete picture of the estimated weights, Figure 1.8 displays the median of the estimated weights and intercepts over the out-of-sample period, as the estimated weights are relatively stable over time. In this Figure, we find that for lasso and elastic net QR, the most important predictors are the GARCH models and RiskMetrics. Even though RiskMetrics is individually not a well performing model (it is often rejected by the backtests), the lasso and elastic net strongly opt for its inclusion in the combinations. A potential reason is that RiskMetrics' estimation error is zero as it is a calibrated model, i.e. it may serve as a stabilizing component. From the median weights of ridge QR, we can again observe the strong grouping effect of this penalty as almost all weights are between 0 and 0.1.

Finally, we evaluate the number of active predictors, i.e. the number of non zero coefficients for lasso and elastic net QR as they enforce sparsity. In Figure 1.8, we see that the lasso and elastic net QR combine the predictions of up to 6 models. However, for many stocks there is just one standalone model that dominates the combination weights. For instance, the VaR forecasts of the Wal-Mart stock (WMT) are mainly driven by the forecasts of RiskMetrics. However, the standalone models that dominate the forecasts change throughout the assets, so that a data-driven selection of the standalone models is required.

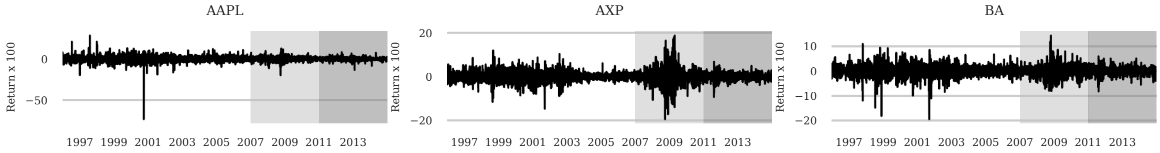
1.5. Conclusion

In this paper, we propose the combination of VaR forecasts with penalized QR. In particular, we consider regularization with the ridge, the lasso and the elastic net penalties. The primary advantage of the regularization over the unpenalized estimator is that it reduces overfitting due to the high multicollinearity of the standalone forecasts. Through the shrinkage and variable selection properties of the penalties, regularized QR stabilize the estimates of the combination weights and thereby improve the predictions.

In the empirical application, we combine the VaR forecasts of 17 standalone models for 30 assets of DJIA and consider three evaluation horizons. We compare the penalized QR combined predictions with the standalone forecasts and a large variety of competing combination approaches. The penalized QR combined forecasts are less often rejected by two backtests than the alternative approaches and are frequently included in the superior set of models. We find that in volatile periods, the lasso and elastic net QR perform slightly better than the ridge QR, i.e. in periods when many standalone models fail, so that the variable selection property is highly relevant. We also observe that the elastic net QR does not perform better than the lasso QR, even though the elastic net is supposed to stabilize the lasso in case of highly correlated covariates (Zou and Hastie, 2005). Furthermore, we find that the proposed heuristic rule performs well for all out-of-sample horizons and all three penalized QR estimators.

For future research, a comparison of penalized QR to QR boosting (Zheng, 2012) would be interesting. One could furthermore consider nonlinear forecast combination via quantile random forests introduced by Meinshausen (2006) or the post-lasso QR estimator by Belloni and Chernozhukov (2011).

Appendix 1.A Plots and Summary Statistics of the Return Series



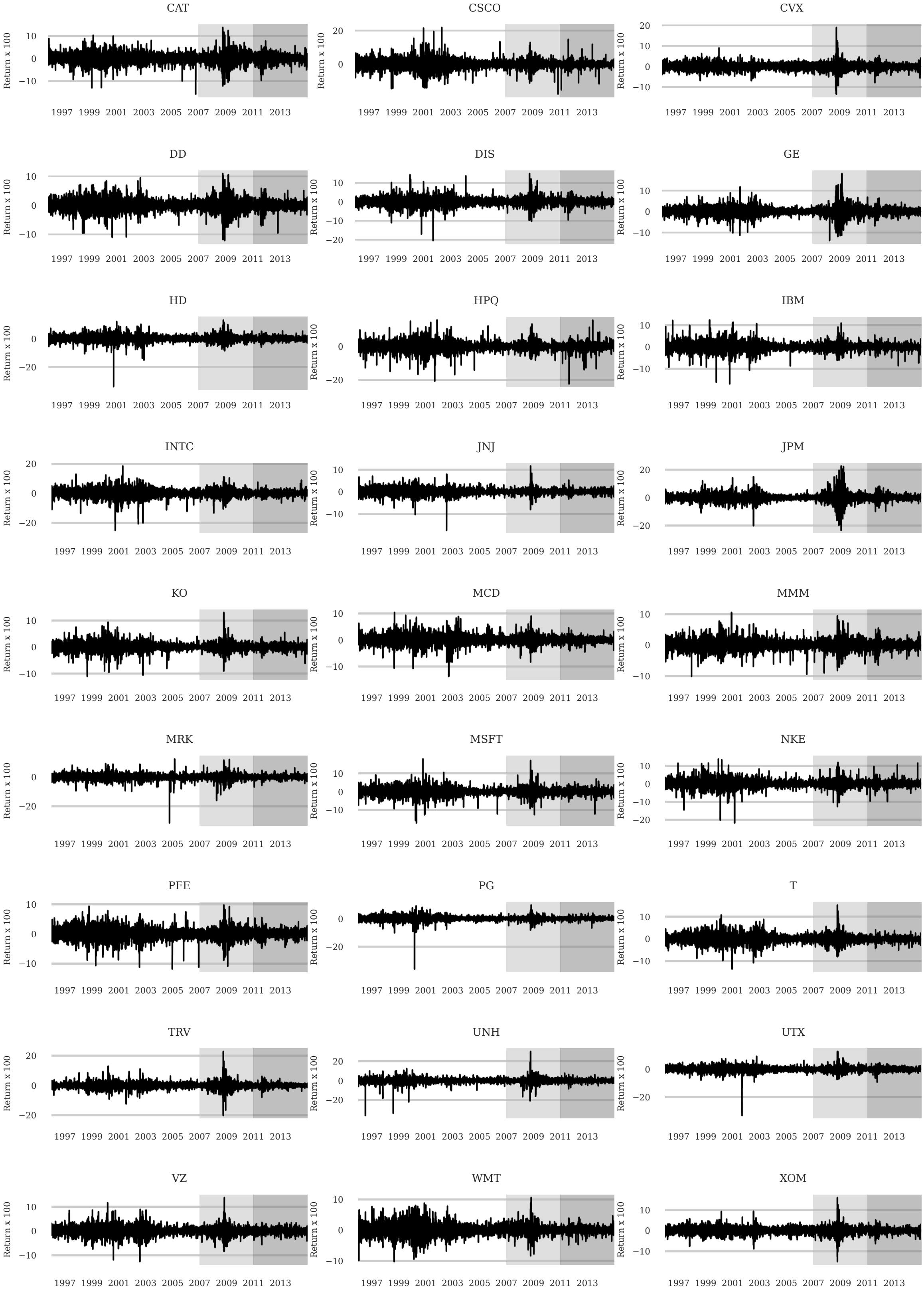
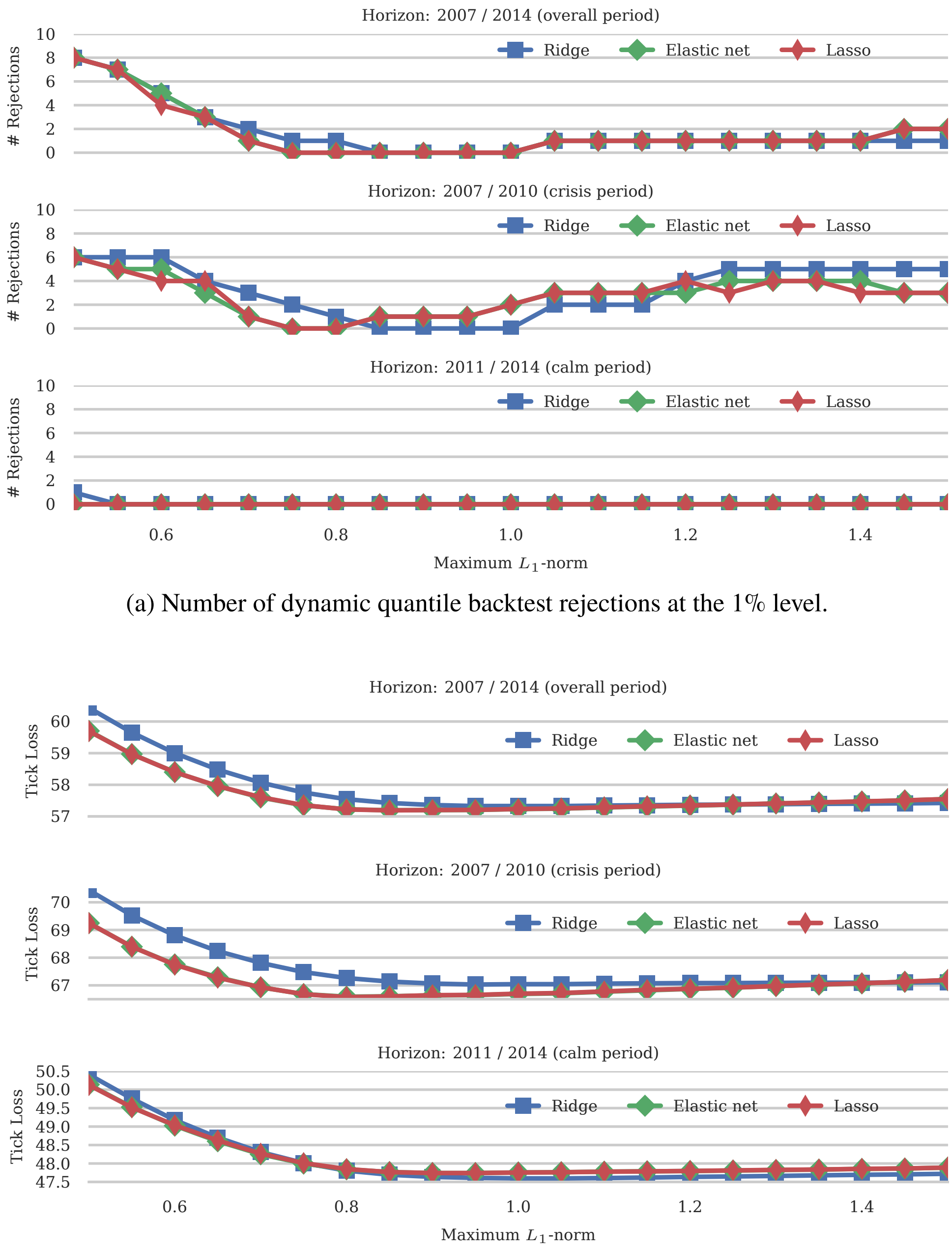


Figure 1.5: Log return series from January 2, 1996 to December 31, 2014. The gray shaded areas indicate the forecast evaluation horizons January 3, 2007 to December 31, 2010 (1008 days, light gray), January 3, 2011 to December 31, 2014 (1006 days, dark gray) and January 3, 2007 to December 31, 2014 (2014 days, both areas).

Table 1.2: Ticker symbols, company names and summary statistics of the log returns x 100.

Symbol	Name	Min	Max	Mean	Var.	Kurt.	Skew.	JB	JB-p
AAPL	Apple	−73.12	28.69	0.10	9.54	73.62	−2.56	1.0e6	0.00
AXP	American Express Company	−19.35	18.77	0.04	5.66	10.97	0.01	1.3e4	0.00
BA	The Boeing Company	−19.39	14.38	0.02	4.07	9.78	−0.37	9.3e3	0.00
CAT	Caterpillar Inc.	−15.69	13.73	0.04	4.55	7.17	−0.09	3.5e3	0.00
CSCO	Cisco Systems, Inc.	−17.69	21.82	0.04	7.21	9.38	0.05	8.1e3	0.00
CVX	Chevron Corporation	−13.34	18.94	0.03	2.71	12.40	0.08	1.8e4	0.00
DD	E. I. du Pont de Nemours and Company	−12.03	10.86	0.02	3.51	7.17	−0.15	3.5e3	0.00
DIS	The Walt Disney Company	−20.29	14.82	0.03	4.07	10.69	−0.07	1.2e4	0.00
GE	General Electric Company	−13.68	17.98	0.02	3.79	10.52	0.01	1.1e4	0.00
HD	The Home Depot, Inc.	−33.88	13.16	0.05	4.47	20.33	−0.79	6.0e4	0.00
HPQ	Hewlett-Packard Company	−22.35	15.95	0.01	6.53	10.08	−0.30	1.0e4	0.00
IBM	International Business Machines	−16.89	12.37	0.04	3.36	10.50	−0.03	1.1e4	0.00
INTC	Intel Corporation	−24.89	18.33	0.03	6.41	9.79	−0.37	9.3e3	0.00
JNJ	Johnson & Johnson	−17.25	11.54	0.03	1.82	13.00	−0.22	2.0e4	0.00
JPM	JPMorgan Chase & Co.	−23.23	22.39	0.02	6.74	14.26	0.23	2.5e4	0.00
KO	The Coca-Cola Company	−11.07	13.00	0.02	2.20	9.57	0.00	8.6e3	0.00
MCD	McDonald’s Corporation	−13.72	10.31	0.03	2.52	8.47	−0.04	6.0e3	0.00
MMM	3M Co	−10.08	10.50	0.03	2.41	7.30	−0.02	3.7e3	0.00
MRK	Merck & Co., Inc.	−31.17	12.25	0.01	3.37	25.90	−1.26	1.1e5	0.00
MSFT	Microsoft Corporation	−16.96	17.87	0.04	4.25	10.11	−0.07	1.0e4	0.00
NKE	Nike	−21.65	13.78	0.05	4.45	11.84	−0.15	1.6e4	0.00
PFE	Pfizer, Inc.	−11.82	9.69	0.02	3.17	6.84	−0.19	3.0e3	0.00
PG	Procter & Gamble	−36.01	9.73	0.03	2.31	73.03	−2.96	9.8e5	0.00
T	AT&T Inc.	−13.54	15.08	0.00	3.14	8.26	0.06	5.5e3	0.00
TRV	The Travelers Companies, Inc.	−20.07	22.76	0.03	3.70	16.55	0.35	3.7e4	0.00
UNH	UnitedHealth Group	−35.59	29.83	0.05	5.61	34.18	−1.36	2.0e5	0.00
UTX	United Technologies Corporation	−33.20	12.79	0.05	3.21	30.93	−1.30	1.6e5	0.00
VZ	Verizon Communications Inc.	−12.61	13.66	0.01	2.91	7.97	0.14	4.9e3	0.00
WMT	Wal-Mart Stores, Inc.	−10.26	10.50	0.04	2.87	7.14	0.08	3.4e3	0.00
XOM	Exxon Mobil Corporation	−15.03	15.86	0.03	2.53	11.61	0.02	1.5e4	0.00

Appendix 1.B Robustness Check

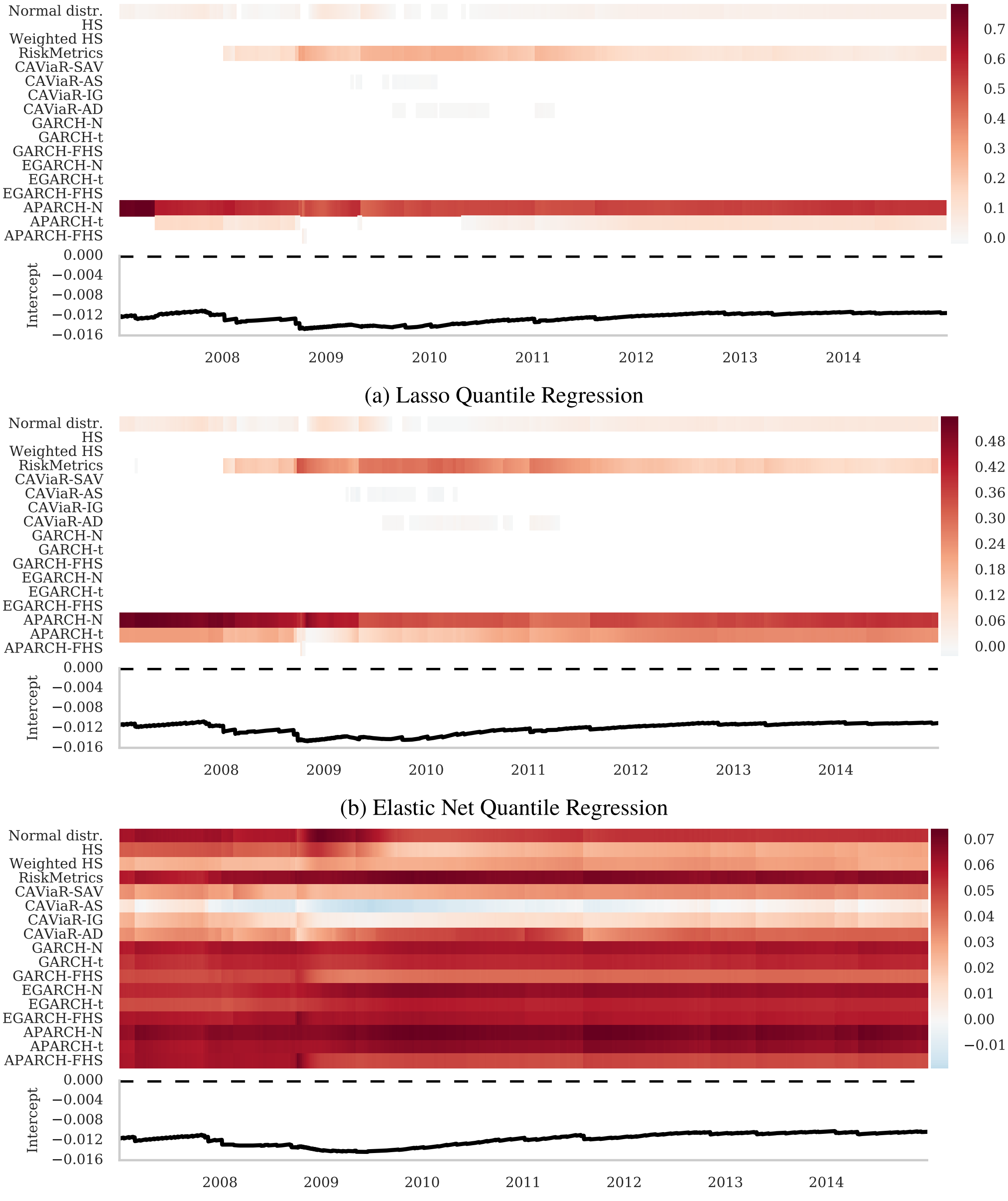


(a) Number of dynamic quantile backtest rejections at the 1% level.

(b) Average tick loss over all 30 assets scaled by 10^5.

Figure 1.6: This figure shows a robustness check for the value of s in the heuristic rule in eq. (1.11). The upper figure shows the number of backtest rejections at the 1% significance level and the lower shows the average tick loss. In each of them, the numbers are depicted for lasso, ridge and elastic net QR and for all three the three evaluation horizons.

Appendix 1.C Estimated Combination Weights for AT&T



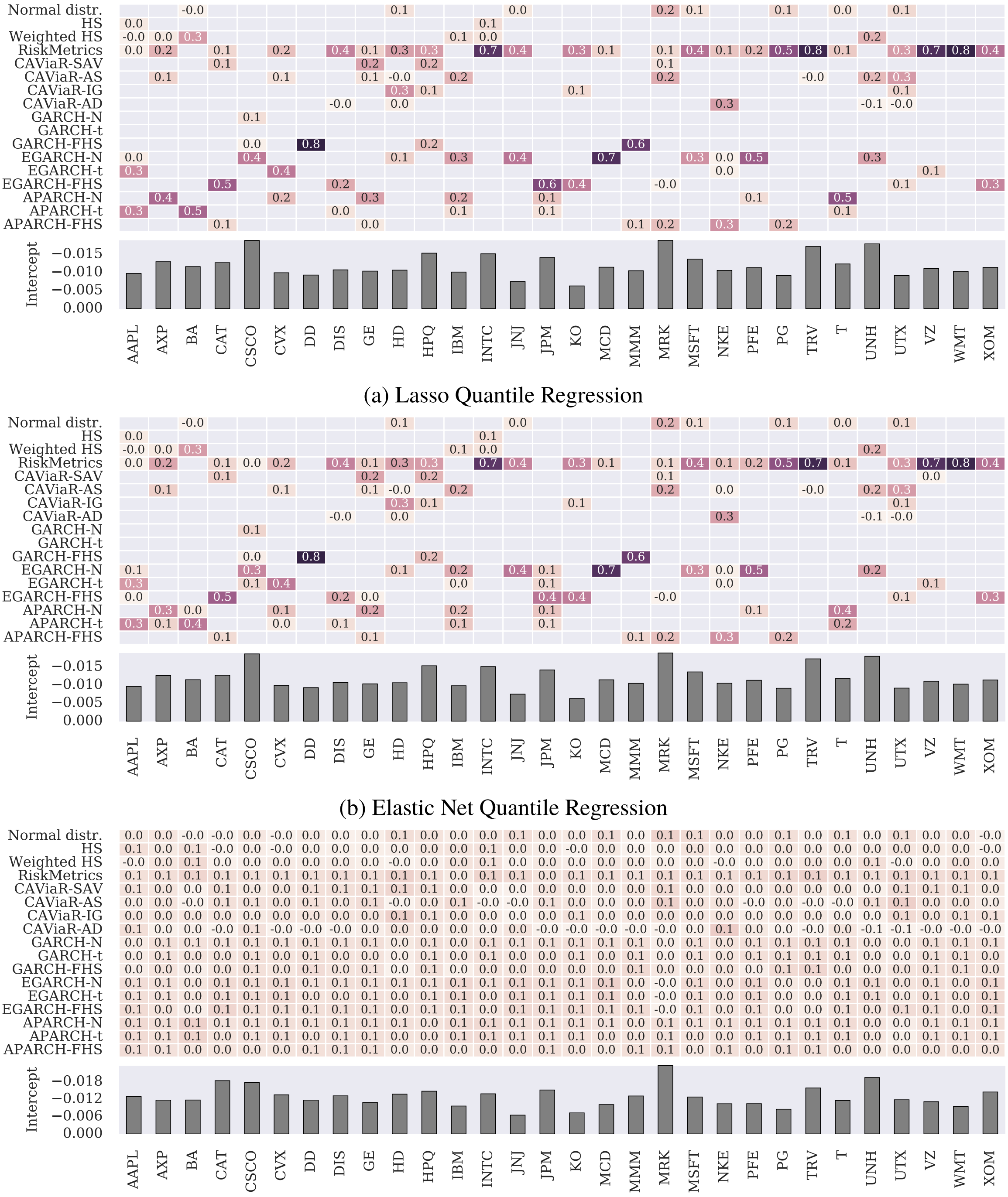
(a) Lasso Quantile Regression

(b) Elastic Net Quantile Regression

(c) Ridge Quantile Regression

Figure 1.7: Estimated combination weights and intercepts for the AT&T stock over the time from January 2007 to December 2014 for lasso, elastic net and ridge QR.

Appendix 1.D Median Estimated Combination Weights



(a) Lasso Quantile Regression

(b) Elastic Net Quantile Regression

(c) Ridge Quantile Regression

Figure 1.8: Median estimated combination weights and intercepts (over the time from January 2007 to December 2014) for lasso, elastic net and ridge QR. The values of the median weights are given in the cells, a blank entry indicates that a weight is on average zero.

References

Abad, P. and S. Benito (2013). “A detailed comparison of value at risk estimates”. *Mathematics and Computers in Simulation* 94, 258–276 (see p. 15).

Aiolfi, M. and A. Timmermann (2006). “Persistence in forecasting performance and conditional combination strategies”. *Journal of Econometrics* 135 (1â€“2), 31–53 (see p. 34).

Arlot, S. and A. Celisse (2010). “A survey of cross-validation procedures for model selection”. *Statistics Surveys* 4, 40–79 (see p. 23).

Barone-Adesi, G., K. Giannopoulos, and L. Vosper (1999). “VaR without correlations for portfolios of derivative securities”. *Journal of Futures Markets* 19 (5), 583–602 (see pp. 27, 117).

Basel Committee (1996). *Overview of the Amendment to the Capital Accord to Incorporate Market Risks*. Tech. rep. Available at `http://www.bis.org/publ/bcbs23.pdf`. Bank for International Settlements (see pp. 15, 96, 102).

— (2006). *International Convergence of Capital Measurement and Capital Standards*. Tech. rep. Available at `http://www.bis.org/publ/bcbs107.pdf`. Bank for International Settlements (see p. 15).

— (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems*. Tech. rep. Available at `http://www.bis.org/publ/bcbs189.pdf`. Bank for International Settlements (see p. 15).

Belloni, A. and V. Chernozhukov (2011). “ℓ_1-penalized quantile regression in high-dimensional sparse models”. *The Annals of Statistics* 39 (1), 82–130 (see p. 41).

Berkowitz, J., P. Christoffersen, and D. Pelletier (2011). “Evaluating value-at-risk models with desk-level data”. *Management Science* 57 (12), 2213–2227 (see p. 30).

Bernardi, M. and L. Catania (2016). “Comparison of Value-at-Risk models using the MCS approach”. *Computational Statistics* 31 (2), 579–608 (see pp. 15, 30, 37).

Bernardi, M., L. Catania, and L. Petrella (2017). “Are news important to predict the Value-at-Risk?” *The European Journal of Finance* 23 (6), 535–572 (see p. 17).

Bollerslev, T. (1986). “Generalized autoregressive conditional heteroskedasticity”. *Journal of Econometrics* 31 (3), 307–327 (see pp. 26, 70, 107).

Boucher, C. M., J. Danielsson, P. S. Kouontchou, and B. B. Maillet (2014). “Risk models-at-risk”. *Journal of Banking & Finance* 44, 72–92 (see p. 15).

Boudoukh, J., M. Richardson, and R. F. Whitelaw (1998). “The Best of Both Worlds: A Hybrid Approach to Calculating Value at Risk”. *Risk* 11 (5), 64–67 (see p. 26).

Casarin, R., C.-L. Chang, J.-A. Jimenez-Martin, M. McAleer, and T. Perez-Amaral (2013). “Risk management of risk under the Basel Accord: A Bayesian approach to forecasting Value-at-Risk of VIX futures”. *Mathematics and Computers in Simulation* 94, 183–204 (see p. 17).

Christoffersen, P. (1998). “Evaluating Interval Forecasts”. *International Economic Review* 39 (4), 841–862 (see pp. 17, 30, 98).

Ding, Z., C. W. J. G. Granger, and R. F. Engle (1993). “A long memory property of stock market returns and a new model”. *Journal of Empirical Finance* 1 (1), 83–106 (see p. 27).

Einhorn, D. (2008). “Private Profits and Socialized Risk”. In: *Global Association of Risk Professionals Risk Review (June/July 2008)*. Ed. by Einhorn, D. and Brown, A. Vol. 42, 10–26 (see p. 15).

Engle, R. F. and S. Manganelli (2004). “CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles”. *Journal of Business & Economic Statistics* 22 (4), 367–381 (see pp. 17, 26, 30, 98).

Ergen, I. (2015). “Two-step methods in VaR prediction and the importance of fat tails”. *Quantitative Finance* 15 (6), 1013–1030 (see p. 15).

Fuertes, A.-M. and J. Olmo (2013). “Optimally harnessing inter-day and intra-day information for daily value-at-risk prediction”. *International Journal of Forecasting* 29 (1), 28–42 (see p. 17).

Ghalanos, A. (2015). *rugarch: Univariate GARCH models.* R package version 1.3-6. (see p. 27).

Giacomini, R. and I. Komunjer (2005). “Evaluation and Combination of Conditional Quantile Forecasts”. *Journal of Business & Economic Statistics* 23 (4), 416–431 (see pp. 17, 19, 119).

Gneiting, T. (2011b). “Quantiles as optimal point forecasts”. *International Journal of Forecasting* 27 (2), 197–207 (see pp. 16, 19).

Grigoryeva, L., J.-P. Ortega, and A. Peresetsky (2017). “Volatility forecasting using global stochastic financial trends extracted from non-synchronous data”. *Forthcoming in Econometrics and Statistics*. DOI: 10.1016/j.ecosta.2017.01.003 (see p. 37).

Halbleib, R. and W. Pohlmeier (2012). “Improving the Value at Risk Forecasts: Theory and Evidence from the Financial Crisis”. *Journal of Economic Dynamics and Control* 36 (8), 1212–1228 (see pp. 15–17, 19, 35).

Hamidi, B., C. Hurlin, P. Kouontchou, and B. Maillet (2015). “A DARE for VaR”. *Finance* 36 (1), 7–38 (see pp. 17, 29, 34, 36, 37).

Hansen, B. (2008). “Least-squares forecast averaging”. *Journal of Econometrics* 146 (2), 342–350 (see p. 21).

Hansen, P. R., A. Lunde, and J. M. Nason (2011). “The Model Confidence Set”. *Econometrica* 79 (2), 453–497 (see pp. 17, 18, 30, 31, 117, 118).

Hart, J. D. (1994). “Automated Kernel Smoothing of Dependent Data by Using Time Series Cross- Validation”. *Journal of the Royal Statistical Society. Series B (Methodological)* 56 (3), 529–542 (see p. 23).

Hart, J. D. and C.-L. Lee (2005). “Robustness of one-sided cross-validation to autocorrelation”. *Journal of Multivariate Analysis* 92 (1), 77–96 (see p. 23).

Hastie, T., R. Tibshirani, and J. Friedman (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer (see pp. 16, 20).

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC (see p. 21).

Hoerl, A. E. and R. W. Kennard (1970a). “Ridge Regression: Applications to Nonorthogonal Problems”. *Technometrics* 12 (1), 69–82 (see pp. 16, 20).

— (1970b). “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. *Technometrics* 12 (1), 55–67 (see pp. 16, 20).

Huang, H. and T.-H. Lee (2013). “Forecasting Value-at-Risk Using High-Frequency Information”. *Econometrics* 1 (1), 127–140 (see p. 17).

James, G. M. (2003). “Variance and Bias for General Loss Functions”. *Machine Learning* 51 (2), 115–135 (see p. 20).

Jeon, J. and J. W. Taylor (2013). “Using CAViaR Models with Implied Volatility for Value-at-Risk Estimation”. *Journal of Forecasting* 32 (1), 62–74 (see p. 17).

Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk*. 3rd ed. McGraw-Hill (see p. 15).

Koenker, R. (2011). “Additive models for quantile regression: Model selection and confidence bandaids”. *Brazilian Journal of Probability and Statistics* 25 (3), 239–262 (see p. 34).

— (2016). *quantreg: Quantile Regression*. R package version 5.29 (see p. 27).

Koenker, R. and G. Bassett (1978). “Regression Quantiles”. *Econometrica* 46 (1), 33–50 (see pp. 16, 19).

Komunjer, I. (2013). “Quantile Prediction”. In: *Handbook of Economic Forecasting*. Ed. by Elliott, G. and Timmermann, A. Vol. 2. Elsevier. Chap. 17, 961–994 (see pp. 15, 98).

Kuester, K., S. Mittnik, and M. Paolella (2006). “Value-at-Risk Prediction: A Comparison of Alternative Strategies”. *Journal of Financial Econometrics* 4 (1), 53–89 (see p. 15).

Kupiec, P. H. (1995). “Techniques for Verifying the Accuracy of Risk Measurement Models”. *The Journal of Derivatives* 3 (2), 73–84 (see pp. 18, 30, 35, 36, 98).

Li, Y. and J. Zhu (2008). “L1-Norm Quantile Regression”. *Journal of Computational and Graphical Statistics* 17 (1), 163–185 (see p. 22).

Louzis, D. P., S. Xanthopoulos-Sisinis, and A. P. Refenes (2014). “Realized volatility models and alternative Value-at-Risk prediction strategies”. *Economic Modelling* 40, 101–116 (see p. 15).

Mallows, C. L. (1973). “Some Comments on Cp”. *Technometrics* 15 (4), 661–675 (see p. 29).

Marinelli, C., S. D’addona, and S. T. Rachev (2007). “A Comparison Of Some Univariate Models For Value-at-risk And Expected Shortfall”. *International Journal of Theoretical and Applied Finance* 10 (06), 1043–1075 (see p. 15).

McAleer, M., J.-A. Jimenez-Martin, and P.-A. Teodosio (2013a). “GFC-robust risk management strategies under the Basel Accord”. *International Review of Economics & Finance* 27, 97–111 (see pp. 17, 30).

— (2013b). “International Evidence on GFC-Robust Forecasts for Risk Management under the Basel Accord”. *Journal of Forecasting* 32 (3), 267–288 (see p. 17).

Meinshausen, N. (2006). “Quantile Regression Forests”. *Journal of Machine Learning Research* 7, 983–999 (see p. 41).

Nelson, D. B. (1991). “Conditional Heteroskedasticity in Asset Returns: A New Approach”. *Econometrica* 59 (2), 347–370 (see p. 27).

Nieto, M. R. and E. Ruiz (2016). “Frontiers in VaR forecasting and backtesting”. *International Journal of Forecasting* 32 (2), 475–501 (see p. 15).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. `http://www.R-project.org`. R Foundation for Statistical Computing. Vienna, Austria (see p. 21).

RiskMetrics Group (1996). *RiskMetrics – Technical Document*. J. P. Morgan and Reuters. New York (see p. 26).

Shan, K. and Y. Yang (2009). “Combining Regression Quantile Estimators”. *Statistica Sinica* 19 (3), 1171–1191 (see pp. 17, 29, 30, 34).

Sheppard, K. (2017). *ARCH*. Python package version 4.0 (see p. 31).

Stock, J. H. and M. W. Watson (2004). “Combination forecasts of output growth in a seven-country data set”. *Journal of Forecasting* 23 (6), 405–430 (see p. 18).

Taylor, J. W. (2008a). “Estimating Value at Risk and Expected Shortfall Using Expectiles”. *Journal of Financial Econometrics* 6 (2), 231–252 (see pp. 17, 52).

Taylor, S. J. (1986). *Modelling Financial Time Series*. World Scientific Publishing (see p. 27).

Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso”. English. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1), 267–288 (see pp. 16, 20).

Timmermann, A. (2006). “Forecast Combinations”. In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, C. W., and Timmermann, A. Vol. 1. Elsevier. Chap. 4, 135–196 (see pp. 15, 21, 28, 29, 34).

Yi, C. (2017). *hqreg: Regularization Paths for Lasso or Elastic-Net Penalized Huber Loss Regression and Quantile Regression*. R package version 1.4 (see p. 21).

Yi, C. and J. Huang (2017). “Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression”. *Journal of Computational and Graphical Statistics* 26 (3), 547–557 (see p. 21).

Zheng, S. (2012). “QBoost: Predicting quantiles with boosting for regression and binary classification”. *Expert Systems with Applications* 39 (2), 1687–1697 (see p. 41).

Zou, H. and T. Hastie (2005). “Regularization and variable selection via the Elastic Net”. *Journal of the Royal Statistical Society. Series B (Methodological)* 67 (2), 301–320 (see pp. 16, 20, 41).

Chapter 2

A Joint Quantile and Expected Shortfall Regression Framework

2.1. Introduction

Measuring and forecasting risks is essential for a variety of academic disciplines. For this purpose, risk measures which are formally defined as a map (with certain properties) from a space of random variables to a real number, are applied to condense the complex nature of the involved risks to a single number (Artzner et al., 1999). In the context of financial risk measurement, to date the most commonly used risk measure is the Value-at-Risk (VaR), which is the α-quantile of the return distribution. Its popularity is mainly due to its simple nature and the fact that up to now, the Basel Accords stipulate its use for the calculation of capital requirements for banks. Besides being not coherent (Artzner et al., 1999), the main drawback of the VaR is its inability to capture tail risks beyond itself. This deficiency is overcome by the risk measure Expected Shortfall (ES) at level α, which is defined as the mean of the returns which are smaller than the α-quantile of the return distribution. The ES has the desired ability to capture information from the whole left tail of the return distribution, which is particularly important for measuring extreme financial risks. Over the past few years, ES has increasingly become the object of interest for practitioners, academics, and regulators, especially since its recent introduction into the Basel Accords (Basel Committee, 2016).

A major drawback of the ES (regarded as a statistical functional) is that it is not elicitable, which means that there exists no loss function (scoring function, scoring rule) which the ES uniquely minimizes in expectation (Gneiting, 2011a; Weber, 2006). This result has two main consequences. First, consistent ranking of competing forecasts for the ES based on such a loss function is infeasible. Second, and more substantial for this paper, modeling the conditional ES given a set of covariates through a regression model without specifying the full conditional distribution is infeasible since estimation of the regression parameters through M-estimation requires such a loss function. Consequently, and in contrast to quantile regression (which can be used to model the VaR), to date, there exists no such regression framework which models the ES based on a set of covariates.

Nadarajah et al. (2014) provide an overview of estimation methods for the ES. However, the reviewed approaches are only applicable for univariate data and not suitable for estimating the conditional ES based on covariates such as in mean and quantile regression. Nevertheless, there are some approaches for the ES which incorporate explanatory variables through indirect estimation procedures. Taylor (2008b) proposes an implicit approach for forecasting ES using exponentially weighted quantile regression and Taylor (2008a) introduces a procedure based on expectile regression and a relationship between the ES and expectiles. Taylor (2017) suggests a joint modeling technique for the quantile and the ES based on maximum likelihood estimation of the asymmetric Laplace distribution. Barendse (2017) proposes generalized method of moments (GMM) estimation for a regression framework for the interquantile expectation.

Even though the ES is not elicitable stand-alone, Fissler and Ziegel (2016) show in their seminal paper that the quantile (the VaR) and the ES are jointly elicitable by introducing a class of joint loss functions, whose expectation is minimized by these two functionals. This joint elicitability result and the associated class of loss functions gives rise to a growing literature in both, joint estimation (Zwingmann and Holzmann, 2016) and in joint forecast evaluation (Acerbi and Szekely, 2014; Fissler, Ziegel, and Gneiting, 2016; Nolde and Ziegel, 2017; Ziegel et al., 2017) for the risk measures VaR and ES.

In this paper, we utilize the class of loss functions of Fissler and Ziegel (2016) for the introduction of a novel simultaneous regression framework for the quantile and the ES and propose both, an M- and a Z-estimator for the joint regression parameters. These strictly consistent loss functions facilitate the opportunity to introduce M- and Z-estimation of the regression parameters without specifying the full conditional distribution of the model, as opposed to maximum likelihood estimation. We show consistency and asymptotic normality for both estimators under weak regularity conditions which are typical for such a regression framework. To the best of our knowledge, we are the first to propose such a joint regression framework for the quantile and the ES together with the joint M- and Z-estimation and the associated results of consistency and asymptotic normality. Furthermore, we are the first to propose a joint semiparametric regression framework for two different functionals based on joint M-estimation without specifying the full conditional distribution.

The employed joint loss function, the estimating equations (for the Z-estimator) and the resulting parameter estimates depend on two specification functions, which can be chosen from some class of functions. Even though consistency and asymptotic normality hold for all applicable choices of these specification functions, they affect the necessary moment conditions, the resulting asymptotic covariance matrices of the estimators, the numerical stability of the optimization algorithm, and the computation times. We discuss the choice of these functions in a theoretical context with respect to asymptotic efficiency and necessary regularity conditions, and with respect to the numerical properties of the optimization algorithm.

The estimation of the asymptotic covariance matrix imposes some difficulties. The first occurs in the estimation of the density quantile function, analogous to quantile regression (cf. Koenker, 2005) and thus, we utilize estimation procedures stemming from this literature. The second issue is the estimation of the variance of the negative quantile residuals conditional on the covariates, a nuisance quantity which is new to the literature. We introduce several estimators for this quantity which are able to cope with limited sample sizes and which can model the dependency of the negative quantile residuals on the covariates. Furthermore, we estimate the covariance matrix using the bootstrap. For ease of application, we provide an R package (Bayer and Dimitriadis, 2017b) which contains the implementation of the M- and Z-estimator. The user can choose the specification functions, the numerical optimization procedure and the estimation method for the covariance matrix of the parameter estimates.

We conduct a Monte-Carlo simulation study where we consider three data generating processes with different properties. We numerically verify consistency and asymptotic normality of the M-estimator for a range of different choices of the specification functions. Furthermore, we find that the Z-estimator is numerically unstable due to the redescending nature of the utilized estimating equations and consequently, we rely on M-estimation of the regression parameters. Moreover, we find that the performance of the M-estimator strongly depends on the specification functions, where choices resulting in positively homogeneous loss functions (Efron, 1991; Nolde and Ziegel, 2017) lead to a superior performance in terms of asymptotic efficiency, computation times, and mean squared error of the estimator.

This joint regression technique for the quantile and ES has a wide range of potential applications as it generalizes quantile regression to the pair consisting of the quantile and the ES. Such estimation, forecasting, and backtesting methods for the ES are particularly sought-after in light of the recent shift from VaR to ES in the Basel Accords. As an illustration, we present an empirical application where we use our regression framework to jointly forecast VaR and ES based on the realized volatility.

The rest of the paper is organized as follows. In Section 2.2, we introduce the joint regression framework, the underlying regularity conditions together with the asymptotic properties of our estimators and discuss the choice of the specification functions. Section 2.3 provides details on the numerical implementation of the estimators and on the estimation of the asymptotic covariance matrix. Section 2.4 presents an extensive simulation study and Section 2.5 contains an empirical application. Section 2.6 provides concluding remarks. The proofs are deferred to Appendices 2.B and 2.C.

2.2. Methodology

2.2.1. The Joint Regression Framework

Following Lambert, Pennock, et al. (2008), Gneiting (2011a) and Fissler and Ziegel (2016), we introduce the concept of (multivariate) p-elicitability. We consider a random variable $Z : \Omega \to \mathbb{R}^d$, defined on some complete probability space $\big(\Omega, \mathcal{F}, P\big)$, a class of distributions $\mathcal{P}$ on $\mathbb{R}^d$, equipped with the Borel σ-field and a functional $T : \mathcal{P} \to D$ with its domain of action $D \subseteq \mathbb{R}^p$, $p \in \mathbb{N}$. We call an integrable loss function $\rho : \mathbb{R}^d \times D \to \mathbb{R}$ *strictly consistent* for the functional T relative to the class of distributions $\mathcal{P}$, if T is the unique minimizer of $\mathbb{E}\big[\rho(Z, \cdot)\big]$ for all distributions $F \in \mathcal{P}$, where F is the distribution of Z. Furthermore, we call a p-dimensional functional T *p-elicitable* relative to the class $\mathcal{P}$, if there exists a loss function ρ which is strictly consistent for T relative to $\mathcal{P}$. If the dimension p is clear from the context, we simply call the functional elicitable instead of p-elicitable.

Given the generalized α-quantile $Q_\alpha(Z) = F^{-1}(\alpha) = \inf\big\{z \in \mathbb{R} : F(z) \geq \alpha\big\}$ for some $\alpha \in (0, 1)$, the ES of the random variable Z at level α is defined as $\text{ES}_\alpha(Z) = \frac{1}{\alpha}\int_0^\alpha Q_u(Z)\,\mathrm{d}u$. If the distribution function of Z is continuous at its α-quantile, this definition can be simplified to the conditional tail expectation $\text{ES}_\alpha(Z) = \mathbb{E}\big[Z \,\big|\, Z \leq Q_\alpha(Z)\big]$. Gneiting (2011a) shows that the ES is not 1-elicitable with respect to any class $\mathcal{P}$ of probability distributions on intervals $I \subseteq \mathbb{R}$, which contain measures with finite support or finite mixtures of absolutely continuous distributions with compact support (see also Weber (2006)). This result has several consequences for the risk measure ES. First, consistent and meaningful ranking of competing forecasts for the functional ES is infeasible. Second, and more consequential for this work, estimating the parameters of a stand-alone regression model for the functional ES in the sense that $\text{ES}_\alpha(Y|X) = X'\theta^e$ by means of M-estimation, i.e. by minimizing some strictly consistent loss function, is infeasible. Even though the ES is not 1-elicitable, Fissler and Ziegel (2016) show that the pair consisting of the ES and the quantile at common probability level α is 2-elicitable relative to the class of distributions with finite first moments and unique α-quantiles and they characterize the full class of strictly consistent loss functions for this pair subject to some regularity conditions. Since the definition of the ES already depends on the respective quantile, the fact that the ES is only elicitable jointly with the quantile is not surprising.

We utilize this joint elicitability result for the introduction of a new joint regression framework for the quantile and the ES where the aforementioned class of strictly consistent loss functions serves as the basis for the M-estimation of the joint regression parameters. For this, let $Y : \Omega \to \mathbb{R}$ and $X : \Omega \to \mathbb{R}^k$ be random variables defined on the same probability space $\big(\Omega, \mathcal{F}, P\big)$ as above. Henceforth, the transpose of X will be denoted by X', the cumulative distribution function of Y given X by $F_{Y|X}$ and the conditional density function by $f_{Y|X}$. For a k-times differentiable real-valued function $G : \mathbb{R} \to \mathbb{R}$, we denote the k-th derivative by $G^{(k)}(\cdot)$.

Assumption 2.2.1 (The joint regression model). The regression framework which jointly models the conditional quantile and ES of Y given X for some fixed level $\alpha \in (0,1)$ is given by

$$Y = X'\theta_0^q + u^q \qquad \text{and} \qquad Y = X'\theta_0^e + u^e, \tag{2.1}$$

where $Q_\alpha(u^q|X) = 0$ and $\mathrm{ES}_\alpha(u^e|X) = 0$. The model is parametrized by $\theta_0 = (\theta_0^{q\prime}, \theta_0^{e\prime})' \in \Theta \subset \mathbb{R}^{2k}$, where the parameter space Θ is compact with nonempty interior, $\mathrm{int}(\Theta) \neq \emptyset$.

We propose both, an M-estimation and a Z-estimation procedure for the compound regression parameter vector θ_0. For the M-estimation, we adapt the class of strictly consistent joint loss functions[1] for the quantile and ES as given in Fissler and Ziegel (2016) such that it can be used in a regression framework,

$$\begin{aligned} \rho(Y,X,\theta) = &\left(\mathbb{1}_{\{Y \le X'\theta^q\}} - \alpha\right) G_1(X'\theta^q) - \mathbb{1}_{\{Y \le X'\theta^q\}} G_1(Y) \\ &+ G_2(X'\theta^e)\left(X'\theta^e - X'\theta^q + \frac{(X'\theta^q - Y)\mathbb{1}_{\{Y \le X'\theta^q\}}}{\alpha}\right) - \mathcal{G}_2(X'\theta^e) + a(Y), \end{aligned} \tag{2.2}$$

where the function G_1 is twice continuously differentiable, $\mathcal{G}_2$ is three times continuously differentiable, $\mathcal{G}_2^{(1)} = G_2$, G_2 and $G_2^{(1)}$ are strictly positive, G_1 is increasing and a and G_1 are integrable. We discuss the choice of the *specification functions* G_1 and $\mathcal{G}_2$ in a theoretical context in Section 2.2.3 and by their numerical performance in Section 2.4.2. The corresponding (ρ-type) M-estimator is defined by a sequence $\hat\theta_{\rho,n}$, such that $\hat\theta_{\rho,n} = \mathrm{argmin}_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \theta)$.

Instead of minimizing some objective function $\rho(Y,X,\theta)$ such as in (2.2), we can also define the corresponding Z-estimator (or ψ-type M-estimator), which sets a vector of estimating equations (moment conditions), denoted by $\psi(Y,X,\theta)$, to zero. More generally, it suffices that these estimating equations converge to zero almost surely. Formally, the Z-estimator is a sequence $\hat\theta_{\psi,n}$, such that $\frac{1}{n}\sum_{i=1}^n \psi(Y_i, X_i, \hat\theta_{\psi,n}) \to 0$ almost surely, where

$$\psi(Y,X,\theta) = \begin{pmatrix} \psi_1(Y,X,\theta) \\ \psi_2(Y,X,\theta) \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha}(\mathbb{1}_{\{Y \le X'\theta^q\}} - \alpha)\left(\alpha X G_1^{(1)}(X'\theta^q) + X G_2(X'\theta^e)\right) \\ X G_2^{(1)}(X'\theta^e)\left(X'\theta^e - X'\theta^q + \frac{1}{\alpha}(X'\theta^q - Y)\mathbb{1}_{\{Y \le X'\theta^q\}}\right) \end{pmatrix}, \tag{2.3}$$

[1] One can interpret the structure of this loss function as follows (Fissler, Ziegel, and Gneiting, 2016): The first summand in (2.2) is a strictly consistent loss function for the quantile (Gneiting, 2011a) and hence only depends on the quantile, whereas the second summand cannot be split into a part depending only on the quantile and one depending only on the ES. This illustrates the fact that the ES itself is not 1-elicitable, but 2-elicitable together with the respective quantile.

which is obtained by differentiating[2] (2.2) and where the functions G_1 and G_2 are given as above. When the loss function $\rho(Y, X, \theta)$ is continuously differentiable in θ, it is obvious that the M- and Z-estimation approaches are equivalent. However, in this case the loss function $\rho(Y, X, \theta)$ is not differentiable and $\psi(Y, X, \theta)$ is discontinuous at the points where $Y = X'\theta^q$. Thus, we treat these two estimation approaches as different estimators and show their asymptotic behavior separately.

2.2.2. Asymptotic Properties

In this section, we present the asymptotic properties of the M- and Z-estimator of the regression parameters. Consistency and asymptotic normality hold under the following set of weak regularity conditions, which are natural for this regression framework.

Assumption 2.2.2 (Regularity Conditions).

($\mathcal{A}$-1) The data (Y_i, X_i) for $i = 1, \ldots, n$ is an iid series of random variables, distributed such as (Y, X) given above. Furthermore, the conditional distribution $F_{Y|X}$ has finite second moments and is absolutely continuous with probability density function $f_{Y|X}$, which is strictly positive, continuous and bounded in a neighbourhood of the true conditional quantile, $X'\theta_0^q$.

($\mathcal{A}$-2) The matrix $\mathbb{E}\left[XX'\right]$ is positive definite.

($\mathcal{A}$-3) The functions $\rho(Y, X, \theta)$ and $\psi(Y, X, \theta)$ are given as in (2.2) and (2.3), where the function G_1 is twice continuously differentiable, $\mathcal{G}_2$ is three times continuously differentiable, $\mathcal{G}_2^{(1)} = G_2$, G_2 and $G_2^{(1)}$ are strictly positive, G_1 is increasing and a and G_1 are integrable.

Remark 2.2.3 (Finite Moment Conditions). We further have to assume that certain moments of X are finite. For the sake of space, we specify the Finite Moment Conditions ($\mathcal{M}$-1) - ($\mathcal{M}$-4) in Appendix 2.A. Note that these general moment conditions simplify substantially for sensible choices of the specification functions G_1 and $\mathcal{G}_2$ as further outlined in Section 2.2.3.

Assumption ($\mathcal{A}$-1) is a combination of typical regularity conditions of mean and quantile regression. Absolute continuity of $F_{Y|X}$ with a strictly positive, bounded and continuous density function in a neighborhood of the true conditional quantile is also imposed for the

[2] Note that the function $\rho(Y, X, \theta)$, given in (2.2) is only differentiable for $Y \neq X'\theta^q$. However, the points of non-differentiability, $Y = X'\theta^q$ form a nullset with respect to the absolutely continuous distribution of Y given X.

asymptotic theory of quantile regression. Existence of the conditional moments of Y given X is subject to the conditions of mean regression and is included in our regularity conditions since ES is a truncated mean. The positive definiteness (full rank condition) in ($\mathcal{A}$-2) is common for any regression design with stochastic regressors in order to exclude perfect multicollinearity of the regressors. The conditions for the specification functions G_1 and $\mathcal{G}_2$ in ($\mathcal{A}$-3) mainly originate from the conditions for the joint elicitability of the quantile and ES in Fissler and Ziegel (2016). Differentiability of these functions is required in this setup for obtaining the estimating equations and for the differentiations in the computation of the asymptotic covariance in Theorem 2.2.6 and Theorem 2.2.7. The existence of certain moments of the explanatory variables as in conditions ($\mathcal{M}$-1) - ($\mathcal{M}$-4) in Appendix 2.A is also standard in any regression design relying on stochastic regressors. Even though compactness of the parameter space Θ in Assumption 2.2.1 generally simplifies the proofs, in this setup it is crucial for consistency of the Z-estimator as the estimating equations ψ_2 are redescending to zero for many reasonable choices of the G_2 function such as e.g. the choices resulting in positively homogeneous loss functions. For details on this, we refer to Section 2.3.1.

Theorem 2.2.4. Assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions ($\mathcal{M}$-1) in Appendix 2.A hold true. Then, for every sequence $\hat{\theta}_{\psi,n} \in \Theta$ satisfying $\frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, X_i, \hat{\theta}_{\psi,n}) \xrightarrow{a.s.} 0$, it holds that $\hat{\theta}_{\psi,n} \xrightarrow{a.s.} \theta_0$.

Theorem 2.2.5. Assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions ($\mathcal{M}$-2) in Appendix 2.A hold true. Then, for every sequence $\hat{\theta}_{\rho,n} \in \Theta$ such that $\frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i, \hat{\theta}_{\rho,n}) \leq \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i, \theta_0) + o_P(1)$, it holds that $\hat{\theta}_{\rho,n} \xrightarrow{\mathbb{P}} \theta_0$.

Theorem 2.2.6. Assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions ($\mathcal{M}$-3) in Appendix 2.A hold true. Then, for every sequence $\hat{\theta}_{\psi,n} \in \Theta$ satisfying $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi(Y_i, X_i, \hat{\theta}_{\psi,n}) \xrightarrow{\mathbb{P}} 0$, it holds that

$$\sqrt{n}\big(\hat{\theta}_{\psi,n} - \theta_0\big) \xrightarrow{d} \mathcal{N}\left(0, \Lambda^{-1} C \Lambda^{-1}\right), \tag{2.4}$$

with

$$\Lambda = \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix} \qquad \text{and} \qquad C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{2.5}$$

where

$$\Lambda_{11} = \frac{1}{\alpha}\mathbb{E}\left[(XX')f_{Y|X}(X'\theta_0^q)\big(\alpha G_1^{(1)}(X'\theta_0^q) + G_2(X'\theta_0^e)\big)\right], \tag{2.6}$$

$$\Lambda_{22} = \mathbb{E}\big[(XX')G_2^{(1)}(X'\theta_0^e)\big], \tag{2.7}$$

$$C_{11} = \frac{1-\alpha}{\alpha}\mathbb{E}\left[(XX')\big(\alpha G_1^{(1)}(X'\theta_0^q) + G_2(X'\theta_0^e)\big)^2\right], \tag{2.8}$$

$$C_{12} = C_{21} = \frac{1-\alpha}{\alpha}\mathbb{E}\left[(XX')\big(X'\theta_0^q - X'\theta_0^e\big)\big(\alpha G_1^{(1)}(X'\theta_0^q) + G_2(X'\theta_0^e)\big)G_2^{(1)}(X'\theta_0^e)\right], \tag{2.9}$$

$$C_{22} = \mathbb{E}\left[(XX')(G_2^{(1)}(X'\theta_0^e))^2\left(\frac{1}{\alpha}\operatorname{Var}\big(Y - X'\theta_0^q \big| Y \leq X'\theta_0^q, X\big) + \frac{1-\alpha}{\alpha}\big(X'\theta_0^q - X'\theta_0^e\big)^2\right)\right]. \tag{2.10}$$

Theorem 2.2.7. Assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions $(\mathcal{M}\text{-}4)$ in Appendix 2.A hold true. Then, for every sequence $\hat{\theta}_{\rho,n} \in \Theta$ such that $\frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \hat{\theta}_{\rho,n}) \leq \inf_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n \rho(Y_i, X_i, \theta) + o_P(n^{-1})$, it holds that

$$\sqrt{n}\big(\hat{\theta}_{\rho,n} - \theta_0\big) \xrightarrow{d} \mathcal{N}\big(0, \Lambda^{-1}C\Lambda^{-1}\big), \tag{2.11}$$

where the matrices Λ and C are given as in Theorem 2.2.6.

Remark 2.2.8 (Quantile Regression). Notice that the asymptotic covariance matrix of the quantile-specific parameter estimates $\hat{\theta}^q$ is given by $\alpha(1-\alpha)D_1^{-1}D_0D_1^{-1}$, where

$$D_1 = \mathbb{E}\left[(XX')f_{Y|X}(X'\theta_0^q)\big(\alpha G_1^{(1)}(X'\theta_0^q) + G_2(X'\theta_0^e)\big)\right] \qquad \text{and} \tag{2.12}$$

$$D_0 = \mathbb{E}\left[(XX')\big(\alpha G_1^{(1)}(X'\theta_0^q) + G_2(X'\theta_0^e)\big)^2\right]. \tag{2.13}$$

This simplifies to the covariance matrix of quantile regression parameter estimates by setting $G_1(z) = z$ and $G_2(z) = 0$, which means ignoring the ES-specific part of our loss function and estimating equations. This demonstrates that the quantile regression method is nested in our regression procedure, also in terms of its asymptotic distribution.

Remark 2.2.9 (Asymptotic Covariance of the ES and the Oracle Estimator). The ES-specific part of the asymptotic covariance is mainly governed by the term C_{22}, which depends on the quantity

$$\frac{1}{\alpha}\operatorname{Var}\big(Y - X'\theta_0^q \big| Y \leq X'\theta_0^q, X\big) + \frac{1-\alpha}{\alpha}\big(X'\theta_0^q - X'\theta_0^e\big)^2 = \frac{1}{\alpha^2}\operatorname{Var}\left((Y - X'\theta_0^q)\mathbb{1}_{\{Y \leq X'\theta_0^q\}}\middle| X\right). \tag{2.14}$$

It is reasonable that the asymptotic covariance of ES regression parameters depends on the truncated variance of Y given X as the asymptomatic covariance of mean regression parameters is driven by the conditional (non-truncated) variance of Y given X. The second term $\left(X'\theta_0^q - X'\theta_0^e\right)^2$ in (2.14) is included since the ES represents a truncated mean where the truncation point itself is a statistical functional (the quantile). In comparison, we consider an oracle M-estimator for the ES-specific regression parameters θ^e, given by the loss function

$$\rho_{\text{Oracle}}(Y, X, \theta^e) = (Y - X'\theta^e)^2 \mathbb{1}_{\{Y \le X'\theta_0^q\}}, \tag{2.15}$$

where we assume that the true quantile regression parameters θ_0^q are known. The resulting asymptotic covariance is given by

$$\text{AVar}\left(\widehat{\theta}^e_{\text{Oracle}}\right) = \frac{1}{\alpha}\mathbb{E}\left[XX'\right]^{-1} \cdot \mathbb{E}\left[(XX')\,\text{Var}\left(Y - X'\theta_0^e \middle| Y \le X'\theta_0^q, X\right)\right] \cdot \mathbb{E}\left[XX'\right]^{-1}, \tag{2.16}$$

which shows that the additional term $\left(X'\theta_0^q - X'\theta_0^e\right)^2$ is not included for this estimator with fixed truncation point $X'\theta_0^q$.

Remark 2.2.10 (Joint Estimation of the Sample Quantile and ES). We can use this regression framework to jointly estimate the quantile and ES of an identically distributed sample $Y_1, \dots, Y_n$ by regressing on a constant only. The asymptotic covariance matrix given in Theorem 2.2.6 and Theorem 2.2.7 then simplifies to Σ with components

$$\Sigma_{11} = \frac{\alpha(1-\alpha)}{f_Y^2(\theta_0^q)}, \tag{2.17}$$

$$\Sigma_{12} = \Sigma_{21} = (1-\alpha)\frac{\theta_0^q - \theta_0^e}{f_Y(\theta_0^q)}, \tag{2.18}$$

$$\Sigma_{22} = \frac{1}{\alpha}\,\text{Var}(Y - \theta_0^q | Y \le \theta_0^q) + \frac{1-\alpha}{\alpha}(\theta_0^q - \theta_0^e)^2, \tag{2.19}$$

where θ_0^q and θ_0^e are the true quantile and ES of Y. The same result is obtained by Zwingmann and Holzmann (2016), who further allow for a distribution function for Y which is not differentiable at the quantile with strictly positive derivative. Notice that in this simplified case without covariates, the asymptotic covariance matrix is independent of the specification functions G_1 and $\mathcal{G}_2$ used in the loss function and in the estimating equations. Furthermore, (2.17) implies that quantile estimates stemming from our joint estimation procedure have the same asymptotic efficiency as quantile estimates stemming from minimizing the generalized piecewise linear loss (Gneiting, 2011a) and as sample quantiles (cf. Koenker (2005)). The same holds true for the efficiency of the sample ES estimators (based on the sample quantile) of Brazauskas et al. (2008) and Chen (2008).

Remark 2.2.11 (Pseudo-R^2 and the choice of $a(Y)$). By choosing $a(Y) = \alpha G_1(Y) + \mathcal{G}_2(Y)$ in (2.2), we can guarantee non-negative losses $\rho(Y, X, \theta) \geq 0$. This choice enables us to define a pseudo-R^2 for our joint regression framework in the sense of Koenker and Machado (1999),

$$R^{QE} = 1 - \frac{\rho(Y, X, \hat{\theta})}{\rho(Y, X, \tilde{\theta})}, \tag{2.20}$$

where $\hat{\theta}$ denotes the parameter estimates of the full regression model and $\tilde{\theta}$ denotes the parameter estimates of a regression model restricted to an intercept term only. However, this choice of $a(Y)$ comes at the cost of more restrictive moment conditions, since we need to impose that $\mathbb{E}\big[G_1(Y) + \mathcal{G}_2(Y)\big] < \infty$.

2.2.3. Choice of the Specification Functions

The loss functions and the estimating equations given in (2.2) and (2.3) depend on two specification functions, G_1 and $\mathcal{G}_2$ (with derivative G_2), which have to fulfill the regularity conditions ($\mathcal{A}$-3) in Assumption 2.2.2. Fissler, Ziegel, and Gneiting (2016) already mention the feasible choices $G_1(z) = 0$, $G_1(z) = z$, $G_2(z) = \exp(z)$ and $G_2(z) = \exp(z)/\big(1 + \exp(z)\big)$ in order to show that this class is non-empty. In contrast to the loss functions of mean, quantile and expectile regression, there is no natural choice for these specification functions for the quantile and ES yet (Nolde and Ziegel, 2017). However, as the choice of these functions strongly influences the performance of our regression procedure in terms of its asymptotic efficiency, the necessary moment conditions of the regressors and the numerical performance of the optimization algorithm, we discuss sensible selection criteria in the following.

Efron (1991) and Nolde and Ziegel (2017) argue that for M-estimation of regression parameters it is crucial that the utilized loss function is positively homogeneous of some order $b \in \mathbb{R}$ in the sense that

$$\rho(cY, X, c\theta) = c^b \rho(Y, X, \theta) \tag{2.21}$$

for all $c > 0$. This is an important property for loss functions since the ordering of the losses should be independent of the unit of measurement, e.g. the currency we measure the prices and risk forecasts with. Loss functions following this property guarantee that we can change the scaling and still obtain the same optima and consequently the same parameter estimates.

For the pair consisting of the quantile and the ES, Nolde and Ziegel (2017) characterize the full class of positively homogeneous[3] loss functions of order b for the case where we restrict the domain of $\mathcal{G}_2$, i.e. the conditional ES to the negative real line[4],

$$
\begin{aligned}
&b < 0: && G_1(z) = -c_0, && \mathcal{G}_2(z) = c_1(-z)^b + c_0, && (2.22)\\
&b = 0: && G_1(z) = d_0 \mathbb{1}_{\{z \le 0\}} + d_0' \mathbb{1}_{\{z > 0\}}, && \mathcal{G}_2(z) = -c_1 \log(-z) + c_0, && (2.23)\\
&b \in (0,1): && G_1(z) = \big(d_1 \mathbb{1}_{\{z \le 0\}} + d_1' \mathbb{1}_{\{z > 0\}}\big)|z|^b - c_0, && \mathcal{G}_2(z) = -c_1(-z)^b + c_0, && (2.24)
\end{aligned}
$$

for some constants $c_0, d_0, d_0' \in \mathbb{R}$ with $d_0 \le d_0'$, $d_1, d_1' \ge 0$ and $c_1 > 0$. There are no positively homogeneous loss functions for the cases $b \ge 1$. Our numerical simulations show that there is no gain in efficiency or numerical accuracy by deviating from the choice $G_1(z) = 0$ (see also Fissler, Ziegel, and Gneiting (2016), Nolde and Ziegel (2017), and Ziegel et al. (2017)), which is also consistent with the homogeneity result. Consequently, we use $G_1(z) = 0$ in the following.

A different natural guiding principle for selecting the specification functions is induced by choosing $\mathcal{G}_2$ (and G_1) such that the moment conditions ($\mathcal{M}$-1) - ($\mathcal{M}$-4) in Appendix 2.A are as least restrictive and as parsimonious as possible. For instance, choosing $\mathcal{G}_2$ such that G_2 and its first and second derivatives are bounded functions (and $G_1(z) = 0$) results in the moment condition $\mathbb{E}\left[||X||^5 + ||X||^4 \mathbb{E}\big[|Y|\big|X\big] + ||X||^3 \mathbb{E}\big[Y^2\big|X\big] + |a(Y)|\right] < \infty$. This motivates the usage of bounded functions[5] for G_2 such as e.g. the second example of Fissler, Ziegel, and Gneiting (2016), $G_2(z) = \exp(z)/\big(1 + \exp(z)\big)$, which is the distribution function of the standard logistic distribution. Further examples of bounded G_2 functions include the distribution functions of absolutely continuous distributions on the real line. In the simulation study in Section 2.4.2, we compare the performance of different specification functions in terms of mean squared error, asymptotic efficiency of the estimator and computation times.

2.3. Numerical Estimation of the Model

In this section, we discuss the difficulties one encounters and the solutions we propose for estimating the joint regression model. Section 2.3.1 illustrates the numerical optimization

[3] For $b = 0$, only the loss differences are positively homogeneous. However, the ordering of the losses is still unaffected under this slightly weaker property.

[4] Since the conditional ES of financial assets for small probability levels is always negative, this is no critical restriction. However, for the numerical parameter estimation, we have to restrict the parameter space Θ such that $X_i'\theta^e < 0$ for all $\theta \in \Theta$ and for all X_i in the underlying sample. For details on this, we refer to Section 2.3.1.

[5] Note that the positively homogeneous loss functions exhibit unbounded $\mathcal{G}_2$ functions. However, as the function $\mathcal{G}_2(z)$ does not grow faster than linear as z tends to infinity, the resulting finite moment conditions are not too restrictive.

procedure we employ for estimating the regression parameters and Section 2.3.2 discusses different estimation methods for the covariance matrix of the estimator.

2.3.1. Optimization

Theorem 2.2.6 and Theorem 2.2.7 imply that both, M-estimation and Z-estimation of the regression parameters θ have the same asymptotic efficiency and consequently, we discuss these estimation approaches in terms of their numerical performance in the following. The numerical implementation of the Z-estimator relies on root-finding of the estimating equations given in (2.3), which we implement as in GMM-estimation by minimizing the inner product $\sum_i \psi(Y_i, X_i, \theta)' \cdot \sum_i \psi(Y_i, X_i, \theta)$. However, the estimating equations are redescending to zero for many attractive choices of $\mathcal{G}_2$ in the sense that $\psi_2(Y, X, \theta) \to 0$ for $X'\theta^e \to -\infty$. Consequently, for θ such that $\theta^q = \theta_0^q$ and $X'\theta^e \to -\infty$, we get the same minimal value of the Z-estimation objective function $\sum_i \psi(Y_i, X_i, \theta)' \cdot \sum_i \psi(Y_i, X_i, \theta)$ as for the true regression parameters θ_0. Thus, the Z-estimator is numerically unstable and diverges in many setups.

Consequently, we rely on M-estimation of the regression parameters in the following. As the loss functions given in (2.2) are not differentiable and non-convex for all applicable choices of the specification functions (Fissler, 2017), we apply a derivative-free global optimization technique. More specifically, we use the Iterated Local Search (ILS) meta-heuristic of Lourenço et al. (2003), which successively refines the parameter estimates by repeated optimizations with iteratively perturbed starting values. Our exact implementation consists of the following steps. First, we obtain starting values for θ^q and θ^e from two quantile regressions of Y on X for the probability levels α and $\tilde{\alpha}$, where we choose $\tilde{\alpha}$ such that the $\tilde{\alpha}$-quantile and the α-ES coincide under normality. Second, using these starting values we minimize the loss function with the derivative-free and robust Nelder-Mead Simplex algorithm (Nelder and Mead, 1965). Third, we perturb the resulting parameter estimates by adding normally distributed noise with zero mean and standard deviation equal to the estimated asymptotic standard errors of the initial quantile regression estimates. Fourth, we re-optimize the model with the perturbed parameter estimates as new starting values. If the loss is further decreased by this re-optimization, we update the estimates and otherwise, we retain the previous ones. Fifth, we iterate over the previous two steps until the loss does not decrease in $m = 10$ consecutive iterations. Our numerical experiments indicate that this repeated optimization procedure yields estimates very close to the ones stemming from other global optimization techniques such as e.g. simulated annealing, whereas the major advantage of ILS is the considerably lower computation time.

For the choices of the specification functions which result in positively homogeneous loss functions, we have to restrict the domain of $\mathcal{G}_2$ to the negative real line as already discussed in Section 2.2.3. Thus, we have to restrict Θ such that $X_i'\theta^e < 0$ for all $\theta \in \Theta$ and for all $i = 1, \dots, n$ during the optimization process. Even though in financial risk management the response variable Y is usually given by financial returns where the true (conditional) ES is strictly negative, there might still be some outliers X_i such that $X_i'\theta_0^e \geq 0$. In such a case, imposing the restriction $X_i'\theta^e < 0$ for all $i = 1, \dots, n$ during the optimization process generates substantially biased estimates for θ^e. In order to avoid this, we estimate the regression model for the transformed dependent variables $Y - \max(Y)$ for the positively homogeneous loss functions and add $\max(Y)$ to the estimated intercept parameters to undo the transformation[6].

We provide an R package for the estimation of the regression parameters (see Bayer and Dimitriadis, 2017b). This package contains an implementation of both, the M- and the Z-estimator, where different optimization algorithms can be chosen (ILS, simulated annealing). The package allows for choosing the specification functions G_1 and $\mathcal{G}_2$ and it includes an option to estimate the model either with or without the translation of the dependent variable. Furthermore, the covariance matrix of the parameter estimates can be estimated either by using the asymptotic theory and the resulting techniques we discuss in the next section, or by using the nonparametric iid bootstrap (Efron, 1979). We recommend applying the M-estimator with the ILS algorithm as this procedure exhibits the best performance in our numerical experiments with respect to accuracy, stability and computation times.

2.3.2. Asymptotic Covariance Estimation

While most parts of the asymptotic covariance matrix given in Theorem 2.2.6 and Theorem 2.2.7 are straightforward to estimate, two nuisance quantities impose some difficulties. The first is the density quantile function $f_{Y|X}(X'\theta_0^q)$, which is already well investigated in the quantile regression literature. In particular, we consider the estimators proposed by Koenker (1994), henceforth denoted by *iid* and by Hendricks and Koenker (1992), henceforth denoted by *nid*. The main difference between these is that the first is based on the assumption that the quantile residuals are independent of the covariates, whereas the second allows for a linear dependence structure. Both approaches depend on a bandwidth parameter which we choose according to Hall and Sheather (1988).

[6] Note that this data transformation changes the average loss function as the applied loss functions are in general not translation invariant. Thus, optimizing the translated loss function can lead to different parameter estimates. However, we do not face the risk of obtaining substantially biased estimates in cases where $X_i'\theta_0^e \geq 0$ for some $i \in \{1, \dots n\}$. Our numerical experiments indicate that the difference between estimating the model for Y and for $Y - \max(Y)$ is small when $X_i'\theta_0^e < 0$ for all $i \in \{1, \dots n\}$, but can be quite substantial if there is an outlier for X_i such that $X_i'\theta_0^e \geq 0$.

The second nuisance quantity is the variance of the quantile residuals, conditional on the covariates and given that these residuals are negative,

$$\operatorname{Var}\left(Y - X'\theta_0^q \middle| Y \leq X'\theta_0^q, X\right) = \operatorname{Var}\left(u^q \middle| u^q \leq 0, X\right). \tag{2.25}$$

Estimation of this quantity is demanding for two reasons. First, for very small probability levels which are typical in financial risk management such as e.g. $\alpha = 2.5\%$, the truncation $u^q \leq 0$ cuts off all but very few (about $\alpha \cdot n$) observations. Second, modeling this truncated variance conditional on the covariates X is challenging, especially considering the very small sample sizes. Under the assumption of homoscedasticity, i.e. that the distribution of u^q is independent of the covariates X, we can simply estimate (2.25) by the sample variance of the negative quantile residuals and we refer to this estimator as *ind* in the following.

We propose two further estimators which allow for a dependence of the quantile residuals on the covariates. For this purpose, we assume a location-scale process with linear[7] specifications of the conditional mean and standard deviation in order to explicitly model the conditional relationship of u^q on X,

$$u^q = X'\zeta + X'\phi \cdot \varepsilon, \tag{2.26}$$

for some parameter vectors $\zeta, \phi \in \mathbb{R}^k$ and where $\varepsilon \sim G(0,1)$ follows a zero mean, unit variance distribution, such that $u^q | X \sim G\big(X'\zeta, (X'\phi)^2\big)$ with distribution function F_G and density f_G. As we need to estimate the truncated variance of u^q given $u^q \leq 0$, i.e. a truncated variant of $(X'\phi)^2$, one possibility is to estimate (2.26) only for those observations where $u^q \leq 0$. However, this approach particularly suffers from the very few negative quantile residuals as we need to estimate additional parameters compared to the *ind* approach.

We present a feasible alternative by estimating the parameters ζ and ϕ using all available observations of u^q and X by quasi generalized pseudo maximum likelihood (Gourieroux and Monfort, 1995, Section 8.4.4) and we obtain the truncated conditional variance by the scaling formula $\operatorname{Var}(u^q | u^q \leq 0, X) = \int_{-\infty}^{0} z^2 h(z)\, dz - \left(\int_{-\infty}^{0} z h(z)\, dz\right)^2$, where $h(z) = f_G(z)/F_G(0)$ is the truncated conditional density of u^q given X and $u^q \leq 0$. We propose one parametric estimator, henceforth denoted by *scl-N*, where we assume that the distribution G is the normal distribution and apply a closed-form solution to the scaling formula. We further propose a semiparametric estimator, henceforth denoted by *scl-sp*, where we estimate the

[7] This approach can further be generalized by considering more general specifications for the conditional mean and standard deviation. However, our numerical experiments indicate that the estimation accuracy for the asymptotic covariance matrix does not increase by deviating from these linear specifications.

distribution G nonparametrically and then apply the scaling formula for this estimated density by numerical integration.

2.4. Simulation Study

In this section, we investigate the finite sample behavior of the M-estimator and verify the asymptotic properties derived in Section 2.2.2 through simulations. Furthermore, we compare the performance of different choices for the specification functions and evaluate the precision of the different covariance matrix estimators described in Section 2.3.2.

2.4.1. Data Generating Process

In order to assess the numerical properties of estimating the joint regression model, we simulate data from a linear location-scale data generating process (DGP),

$$Y = X'\gamma + (X'\eta) \cdot v, \tag{2.27}$$

where $v \sim F(0, 1)$ has zero mean and unit variance, $X = \big(1, X_2, \ldots, X_k\big)'$ and $\gamma, \eta \in \mathbb{R}^k$. For this process, the true conditional quantile and ES are linear functions in X, given by

$$Q_\alpha\left(Y|X\right) = X'(\gamma + z_\alpha \eta) \qquad \text{and} \qquad \text{ES}_\alpha\left(Y|X\right) = X'(\gamma + \xi_\alpha \eta), \tag{2.28}$$

where z_α and ξ_α are the quantile and ES of the distribution $F(0, 1)$, which implies that $\theta_0^q = \gamma + z_\alpha \eta$ and $\theta_0^e = \gamma + \xi_\alpha \eta$. Furthermore, the conditional distributions of the quantile- and ES-residuals are given by

$$u^q|X \sim F\left(-z_\alpha(X'\eta),\, (X'\eta)^2\right) \qquad \text{and} \qquad u^e|X \sim F\left(-\xi_\alpha(X'\eta),\, (X'\eta)^2\right). \tag{2.29}$$

For the simulation study, we want to assess the performance of our regression procedure in various setups. Thus, we specify γ, η and F in the following such that we get data which is homoscedastic (DGP-(1)) and heteroskedastic (DGP-(2)). Furthermore, we include a regression setup with multiple, correlated regressors and a leptocurtic conditional distribution (DGP-(3)),

$$\begin{aligned}
&\text{DGP-(1):} \quad X = (1,\, X_2), \qquad X_2 \sim \chi_1^2 \quad \text{and} \quad Y|X \sim \mathcal{N}\big(-X_2,\, 1\big)\\
&\text{DGP-(2):} \quad X = (1,\, X_2), \qquad X_2 \sim \chi_1^2 \quad \text{and} \quad Y|X \sim \mathcal{N}\big(-X_2,\, (1 + 0.5X_2)^2\big)\\
&\text{DGP-(3):} \quad X = (1,\, X_2,\, X_3) \quad X_2,\, X_3 \sim U[0, 1] \quad \text{with} \quad \text{corr}(X_2, X_3) = 0.5 \quad \text{and}\\
&\qquad\qquad\quad Y|X \sim t_5\left(X_2 - X_3,\, (1 + X_2 + X_3)^2\right).
\end{aligned}$$

We simulate all three processes 25,000 times with varying sample sizes of n = 250, 500, 1000, 2000 and 5000 observations. For each replication and for each of the sample sizes we regress the simulated Y's on the covariates X using our joint regression method for the probability level $\alpha = 2.5\%$.

2.4.2. Comparing the Specification Functions

We start the discussion of the simulation results by investigating the numerical performance of the M-estimator based on different choices of the specification function[8] $\mathcal{G}_2$ used in the loss function in (2.2). We use three natural examples resulting in positively homogeneous loss functions of order $b = -1$, $b = 0$ and $b = 0.5$ respectively[9], a bounded G_2 function and the (unbounded) exponential function:

$$\begin{aligned} &\mathcal{G}_2(z) = -1/z, \qquad \mathcal{G}_2(z) = -\log(-z), \qquad \mathcal{G}_2(z) = -\sqrt{-z}, \\ &\mathcal{G}_2(z) = \log\big(1 + \exp(z)\big), \qquad \text{and} \qquad \mathcal{G}_2(z) = \exp(z). \end{aligned} \tag{2.30}$$

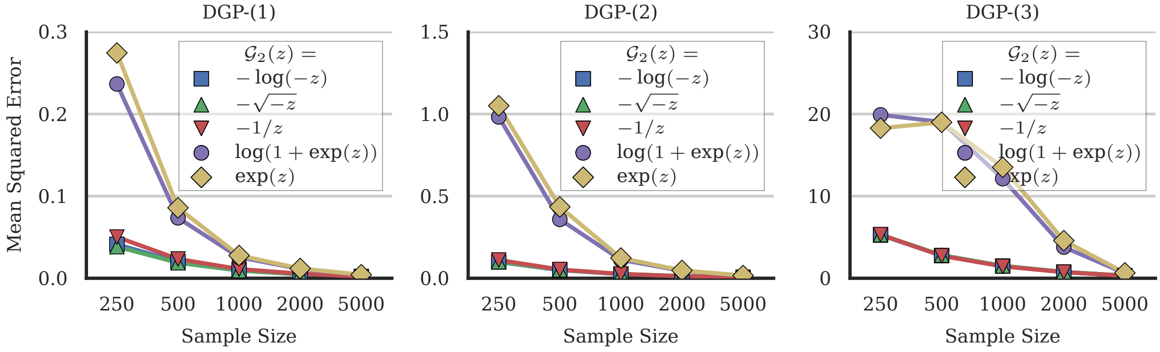


Figure 2.1: Sum of the mean squared errors of the parameter estimates for all three DGPs. The results are shown for the five choices of the specification functions given in (2.30) and a range of sample sizes.

Figure 2.1 presents the sum (over the $2k$ regression parameters) of the mean squared errors (MSE) of the regression parameters for the three DGPs described above, different sample sizes and for the five choices of the specification functions given in (2.30). As implied by the asymptotic theory, we obtain consistent parameter estimates for all five choices of the specification functions as the MSEs converge to zero for all three DGPs. However, they differ substantially with respect to their small sample properties. The three positively homogeneous specifications result in the most accurate estimates, whereas the choices $\mathcal{G}_2(z) = -\sqrt{-z}$ and $\mathcal{G}_2(z) = -\log(-z)$ tend to perform slightly better than the choice

[8] Following the reasoning of Section 2.2.3 and Nolde and Ziegel (2017) and Ziegel et al. (2017), we fix $G_1(z) = 0$ throughout the simulation study.

[9] Our numerical simulations show that the numerical results are unaffected by different choices of the associated constants in (2.22) - (2.24).

$\mathcal{G}_2(z) = -1/z$. Furthermore, the bounded choice $\mathcal{G}_2(z) = \log\big(1 + \exp(z)\big)$ still performs better than the unbounded exponential function.

Table 2.1 reports the Frobenius norms of the lower triangular parts of the true asymptotic covariance matrices and of the respective (lower triangular) quantile-specific and the ES-specific sub-matrices for the three DGPs and for the five choices of the specification functions given in (2.30). For comparison, we also report the Frobenius norm of the lower triangular part of the asymptotic covariance of the quantile regression estimator. We approximate the true asymptotic covariance matrix through Monte-Carlo integration with a sample size of 10^9 using the formulas in Theorem 2.2.6 and by using the true density and conditional truncated variance. On average, the specification functions $\mathcal{G}_2(z) = -\log(-z)$ and $\mathcal{G}_2(z) = -\sqrt{-z}$ exhibit the smallest asymptotic covariances, closely followed by the third choice for a positively homogeneous loss function, $\mathcal{G}_2(z) = -1/z$. The non-homogeneous choices lead to considerably larger asymptotic variances for all considered DGPs and sub-matrices. Furthermore, by comparing the quantile-specific parameters of the joint estimation approach (from the positively homogeneous loss functions) to quantile regression estimates, we roughly obtain the same asymptotic efficiency.

Table 2.1: This table reports the Frobenius norms of the lower triangular parts of the asymptotic covariance matrices and the respective quantile-specific and the ES-specific sub-matrices for the three DGPs and for the five choices of the specification functions given in (2.30). For comparison, we report the same quantity for the asymptotic covariance of the quantile regression estimator.

	DGP-(1)			DGP-(2)			DGP-(3)		
	Q	ES	Full	Q	ES	Full	Q	ES	Full
$\mathcal{G}_2(z) = -\log(-z)$	7.5	13.1	9.2	17.9	26.9	20.0	581.1	1739.1	1053.0
$\mathcal{G}_2(z) = -\sqrt{-z}$	7.0	11.8	8.4	18.0	25.4	19.3	584.5	1740.1	1054.4
$\mathcal{G}_2(z) = -1/z$	9.1	16.9	11.8	24.1	39.4	28.5	613.7	1851.9	1119.8
$\mathcal{G}_2(z) = \log(1 + \exp(z))$	15.4	21.5	16.6	72.4	80.1	67.1	987.9	2393.0	1496.4
$\mathcal{G}_2(z) = \exp(z)$	15.8	22.6	17.2	74.6	84.5	70.0	1001.9	2440.4	1524.6
Quantile Regression	6.8	–	–	21.4	–	–	600.5	–	–

2.4.3. Comparing the Variance-Covariance Estimators

In this section, we compare the empirical performance of the asymptotic covariance estimators discussed in Section 2.3.2. For the comparison of their precision, Figure 2.2 reports the average of the Frobenius norm of the lower triangular part of the differences between the estimated covariances and the empirical covariance of the estimated parameters. We report results for the three homogeneous loss functions and the three DGPs, where each of the plots presents the average norm differences for the four covariance estimators (*iid/nid*, *nid/scl-N*, *nid/scl-sp* and the iid bootstrap) depending on the sample size.

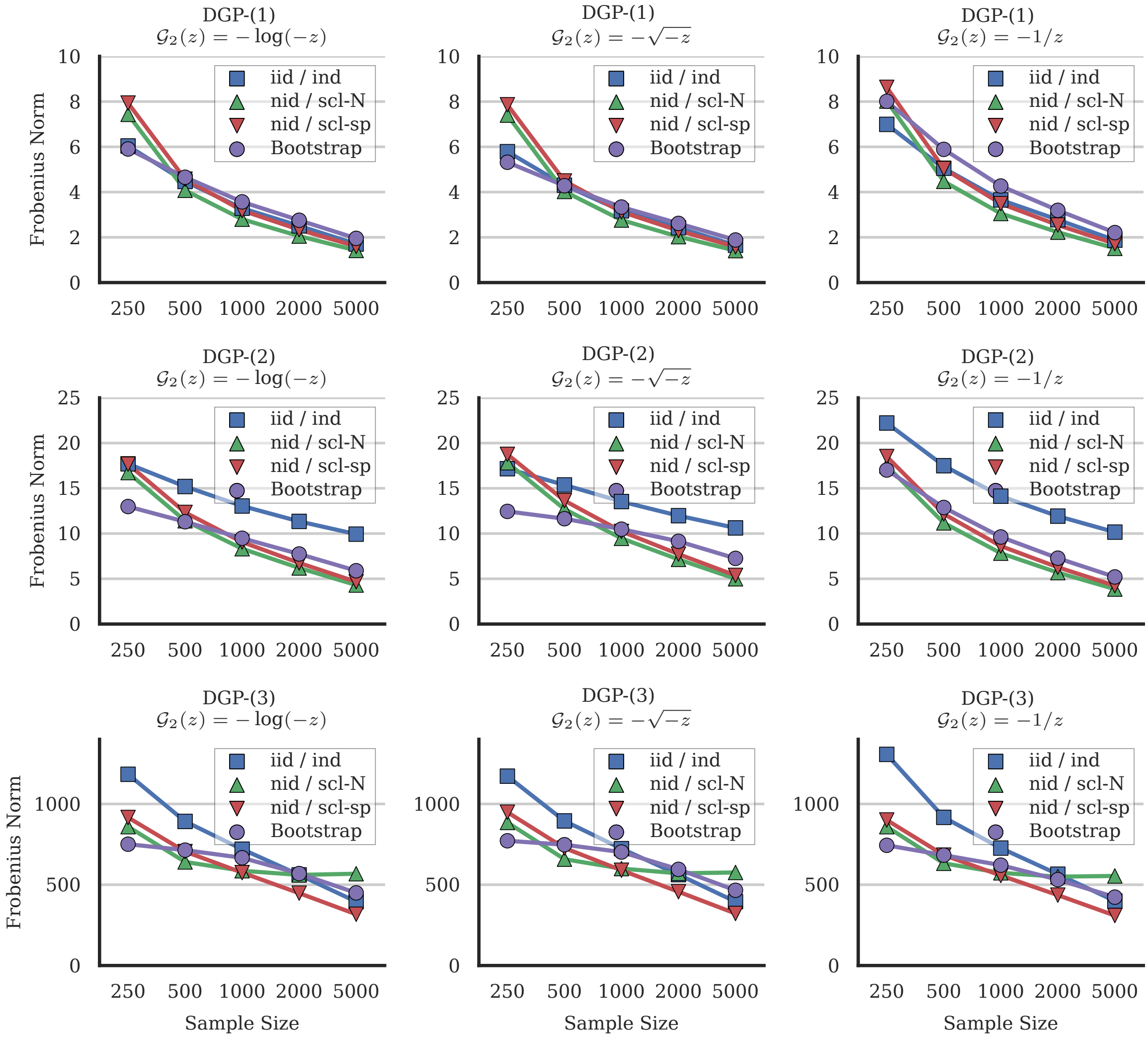


Figure 2.2: This figure compares four covariance estimation approaches described in Section 2.3.2 for the three data generating processes, a range of sample sizes and the three positively homogeneous choices of the $\mathcal{G}_2$-functions. We report the average of the Frobenius norm of the lower triangular part of the differences between the estimated asymptotic covariances and the empirical covariance of the M-estimator.

We find that the *iid/nid* estimator performs well for the first, homoscedastic DGP whereas for the other two DGPs, it fails to capture the underlying more complicated dynamics of the data. The *nid/scl-N* estimator outperforms the other estimation approaches in the first two DGPs, where the underlying conditional distribution follows a normal distribution whereas its performance drops for the third DGP, which follows a Student-*t* distribution. The performance of the flexible *nid/scl-sp* estimator is the most stable throughout all three DGPs. Eventually, the bootstrap estimator accurately estimates the covariance for all three DGPs, whereas in comparison to the other estimators, it is particularly good in small samples. The provided R package contains all four covariance estimators.

2.5. Empirical Application

In this empirical application, we use our joint regression framework for forecasting the VaR and ES of the close-to-close log returns of the IBM stock.

$$Q_\alpha(r_t|\mathrm{RV}_{t-1}) = \theta_1^q + \theta_2^q \mathrm{RV}_{t-1} \quad \text{and} \quad \mathrm{ES}_\alpha(r_t|\mathrm{RV}_{t-1}) = \theta_1^e + \theta_2^e \mathrm{RV}_{t-1}, \tag{2.31}$$

where $\mathrm{RV}_t = (\sum_i r_{t,i}^2)^{1/2}$ denotes the realized volatility estimator (Andersen and Bollerslev, 1998) for day t, where $r_{t,i}$ denotes the i-th high-frequency return of day t. Our dataset consists of the five minute returns of the IBM stock from January 3, 2001 to July 18, 2017 with total of 4120 days, which we obtain from the TAQ database. We estimate the model parameters using a rolling window of 1000 days and evaluate the forecasts on the remaining 3120 days.

We compare the predictive power of this model against three standard models from the literature. The first is the historical simulation (HS) approach, which forecasts the VaR and ES for day t as the sample quantile and ES of the daily returns of the past 250 trading days. The second is an AR(1)-GARCH(1,1)-t model (Bollerslev, 1986), and the third is the Heterogeneous Auto-Regressive (HAR) model of Corsi (2009), based on the realized volatility estimates given above. Forecasts of the VaR and ES for the HAR model are obtained from the volatility forecasts and by assuming a Gaussian return distribution. While the first two of these approaches rely on daily data only, the third one incorporates the same high frequency information as our approach.

We evaluate the forecasting power of the VaR and ES of these models by the class of strictly consistent loss (scoring) functions for the VaR and ES of Fissler and Ziegel (2016). We use *Murphy diagrams* introduced by Ehm et al. (2016) and Ziegel et al. (2017), which provide a parsimonious way to evaluate competing forecasts simultaneously for a full class of strictly consistent loss functions. In fact, one forecasting model significantly dominates another one with respect to the full class of strictly consistent loss functions if and only if the elementary score differences plotted in the Murphy diagrams are strictly negative (positive). For further details on the theory and the implementation of Murphy diagrams, we refer to Ehm et al. (2016) and Ziegel et al. (2017).

Figure 2.3 displays the average of the elementary score differences of the joint VaR and ES regression model against the three alternative models together with the respective 95% pointwise confidence bands for the elementary scores provided in Ziegel et al. (2017) for the pair VaR and ES. Using this graphical method, we can see that the elementary score differences for the joint regression forecasting model against the historical simulation and AR(1)-GARCH(1,1)-t model are significantly negative for the vast majority of threshold values. This implies that the joint regression forecasting model significantly dominates these

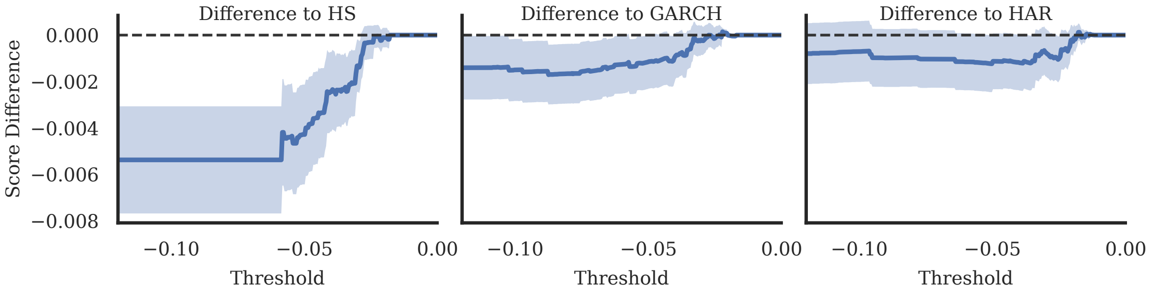


Figure 2.3: Elementary Score Differences of the VaR/ES Regression and the respective comparison models

other two forecasting approaches. Even though we also observe strictly negative elementary score differences in comparison against the HAR model, these differences are not significant and consequently, we cannot significantly outperform this model.

2.6. Conclusion

In this paper, we introduce a joint regression technique for the quantile (the VaR) and the ES. This regression approach relies on the class of strictly consistent joint loss functions introduced by Fissler and Ziegel (2016), which permits the joint elicitation of the quantile and the ES. We introduce an M- and a Z-estimator for the parameters of the joint regression model. Given a set of standard regularity conditions, we show consistency and asymptotic normality for both estimators, which we also verify numerically through extensive simulations. The underlying loss functions, the estimating equations and the asymptotic covariance matrices of the estimators depend on the choice of two specification functions, which we investigate in terms of the resulting moment conditions, asymptotic efficiency, numerical performance and computation times. In our numerical simulations, we find that choices resulting in positively homogeneous loss functions dominate other choices with respect to the aforementioned criteria. Furthermore, we propose several estimation methods for the asymptotic covariance matrix, which are able to cope with different properties of the underlying data. We provide an R package (see Bayer and Dimitriadis, 2017b), which implements the M- and Z-estimation procedures where one can choose the underlying specification functions, the numerical optimization approach and the estimation method for the asymptotic covariance matrix.

Our new joint regression technique allows for a wide range of applications for the risk measures VaR and ES. As an illustration, we present an empirical application in this paper where we use this regression framework to jointly forecast VaR and ES based on realized volatility estimates. Furthermore, Bayer and Dimitriadis (2017c) use this regression to develop an ES backtest which is particularly relevant in light of the recent introduction of ES into the Basel regulatory framework and the present lack of accurate backtesting methods for the ES.

Appendix 2.A Finite Moment Conditions

For convenience of the supremum notation, for all $\theta \in \text{int}(\Theta)$ and for $d > 0$, we define the open neighborhood $U_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| < d\}$ and its closure $\bar{U}_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| \leq d\}$.

($\mathcal{M}$-1) For Theorem 2.2.4, we assume that the following moments are finite for some $d_0 > 0$:

- $\mathbb{E}[||X||^2 \sup_{\theta \in U_{d_0}(\theta_0)} |G_1^{(1)}(X'\theta^q)|]$
- $\mathbb{E}[||X||^2 \sup_{\theta \in U_{d_0}(\theta_0)} |G_1^{(2)}(X'\theta^q)|]$
- $\mathbb{E}[||X||^2 \sup_{\theta \in U_{d_0}(\theta_0)} |G_2(X'\theta^e)|]$
- $\mathbb{E}[||X||^3 \sup_{\theta \in U_{d_0}(\theta_0)} |G_2^{(1)}(X'\theta^e)|]$
- $\mathbb{E}[||X||^3 \sup_{\theta \in U_{d_0}(\theta_0)} |G_2^{(2)}(X'\theta^e)|]$
- $\mathbb{E}[||X||^2 \sup_{\theta \in U_{d_0}(\theta_0)} |G_2^{(1)}(X'\theta^e)| \; \mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^2 \sup_{\theta \in U_{d_0}(\theta_0)} |G_2^{(2)}(X'\theta^e)| \; \mathbb{E}[|Y||X]]$

($\mathcal{M}$-2) For Theorem 2.2.5, we assume that the following moments are finite:

- $\mathbb{E}[||X||^2]$
- $\mathbb{E}[\sup_{\theta \in \Theta} |G_1(X'\theta^q)|]$
- $\mathbb{E}[|G_1(Y)|]$
- $\mathbb{E}[|a(Y)|]$
- $\mathbb{E}[||X|| \sup_{\theta \in \Theta} |G_2(X'\theta^e)|]$
- $\mathbb{E}[\sup_{\theta \in \Theta} |G_2(X'\theta^e)| \; \mathbb{E}[|Y||X]]$
- $\mathbb{E}[\sup_{\theta \in \Theta} |\mathcal{G}_2(X'\theta^e)|]$

($\mathcal{M}$-3) For Theorem 2.2.6, we assume that the following moments are finite for some constant $d_0 > 0$ and for all $\theta \in \bar{U}_{d_0}(\theta_0)$:

- $\mathbb{E}[||X||^3(\sup_{\tau \in \bar{U}_{d_0}(\theta_0)} G_1^{(1)}(X'\tau^q))(\sup_{\tilde{\tau} \in \bar{U}_{d_0}(\theta_0)} G_1^{(2)}(X'\tilde{\tau}^q))]$
- $\mathbb{E}[||X||^3(\sup_{\tau \in \bar{U}_{d_0}(\theta_0)} G_1^{(1)}(X'\tau^q))(\sup_{\tilde{\tau} \in \bar{U}_{d_0}(\theta_0)} G_2^{(1)}(X'\tilde{\tau}^e))]$
- $\mathbb{E}[||X||^3(\sup_{\tau \in \bar{U}_{d_0}(\theta_0)} G_2(X'\tau^e))(\sup_{\tilde{\tau} \in \bar{U}_{d_0}(\theta_0)} G_1^{(2)}(X'\tilde{\tau}^q))]$
- $\mathbb{E}[||X||^3(\sup_{\tau \in \bar{U}_{d_0}(\theta_0)} G_2(X'\tau^e))(\sup_{\tilde{\tau} \in \bar{U}_{d_0}(\theta_0)} G_2^{(1)}(X'\tilde{\tau}^e))]$
- $\mathbb{E}[||X||^3 \sup_{\tau \in \bar{U}_{d_0}(\theta_0)}(G_1^{(1)}(X'\tau^q))^2]$
- $\mathbb{E}[||X||^3 \sup_{\tau \in \bar{U}_{d_0}(\theta_0)}(G_2(X'\tau^e))^2]$
- $\mathbb{E}[||X||^3 \sup_{\tau \in \bar{U}_{d_0}(\theta_0)} G_1^{(1)}(X'\tau^q)G_2(X'\tau^e)]$
- $\mathbb{E}[||X||^5(\sup_{\tau \in \bar{U}_{d_0}(\theta_0)} G_2^{(1)}(X'\tau^e))(\sup_{\tilde{\tau} \in \bar{U}_{d_0}(\theta_0)} G_2^{(2)}(X'\tilde{\tau}^e))]$
- $\mathbb{E}[||X||^5(\sup_{\tau \in \bar{U}_{d_0}(\theta_0)} G_2^{(1)}(X'\tau^e))^2]$

- $\mathbb{E}[||X||^4(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2^{(1)}(X'\tau^e))(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_2^{(2)}(X'\tilde{\tau}^e))\mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^3 G_2^{(1)}(X'\theta^e)(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2^{(1)}(X'\tau^e))\mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^3 G_2^{(1)}(X'\theta^e)(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2^{(2)}(X'\tau^e))\mathbb{E}[Y^2|X]]$
- $\mathbb{E}[||X||^3(\sup_{\tau\in\bar{U}_{d_0}(\theta_0)} G_2^{(1)}(X'\tau^e))(\sup_{\tilde{\tau}\in\bar{U}_{d_0}(\theta_0)} G_2^{(2)}(X'\tilde{\tau}^e))\mathbb{E}[Y^2|X]]$

($\mathcal{M}$-4) For Theorem 2.2.7, we assume that the following moments are finite for some constant $d_0 > 0$:

- $\mathbb{E}[|G_1(Y)|]$
- $\mathbb{E}[|a(Y)|]$
- $\mathbb{E}[||X|| \sup_{\theta\in\bar{U}_{d_0}(\theta_0)} |G_1^{(1)}(X'\theta^q)|]$
- $\mathbb{E}[||X||^2 \sup_{\theta\in\bar{U}_{d_0}(\theta_0)}(G_1^{(1)}(X'\theta^q))^2]$
- $\mathbb{E}[||X||^2 \sup_{\theta\in\bar{U}_{d_0}(\theta_0)} |G_1^{(1)}(X'\theta^q)G_2(X'\theta^e)$
- $\mathbb{E}[||X|| \sup_{\theta\in\bar{U}_{d_0}(\theta_0)} |G_2(X'\theta^e)|]$
- $\mathbb{E}[||X||^2 \sup_{\theta\in\bar{U}_{d_0}(\theta_0)} |G_2^{(1)}(X'\theta^e)|]$
- $\mathbb{E}[||X||^2 \sup_{\theta\in\bar{U}_{d_0}(\theta_0)}(G_2(X'\theta^e))^2]$
- $\mathbb{E}[||X||^4 \sup_{\theta\in\bar{U}_{d_0}(\theta_0)}(G_2^{(1)}(X'\theta^e))^2]$
- $\mathbb{E}[||X|| \sup_{\theta\in\bar{U}_{d_0}(\theta_0)} |G_2^{(1)}(X'\theta^e)|\ \mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^3 \sup_{\theta\in\bar{U}_{d_0}(\theta_0)}(G_2^{(1)}(X'\theta^e))^2\ \mathbb{E}[|Y||X]]$
- $\mathbb{E}[||X||^2 \sup_{\theta\in\bar{U}_{d_0}(\theta_0)}(G_2^{(1)}(X'\theta^e))^2\ \mathbb{E}[Y^2|X]]$

Appendix 2.B Proofs

Henceforth, $||v||$ denotes the maximum norm for a vector $v \in \mathbb{R}^k$ and for a matrix A, $||A||$ denotes the row-sum matrix norm which is induced by the maximum norm for vectors. For convenience of the supremum notation, for all $\theta \in \text{int}(\Theta)$ and for some $d > 0$, we define the open neighborhood $U_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| < d\}$ and its closure $\bar{U}_d(\theta) = \{\tau \in \Theta : ||\tau - \theta|| \le d\}$.

Proof of Theorem 2.2.4. We apply Theorem 2 from Huber (1967) and show that the function $\psi(Y, X, \theta)$ as given in (2.3) satisfies the respective assumptions of this theorem. Note that the parameter space Θ is assumed to be compact and thus, we do not have to show condition (B-4) in the notation of Huber (1967). As the product of continuous functions and the indicator function $\mathbb{1}_{\{Y\le X'\theta^q\}}$, the function ψ is measurable and regarded as a stochastic process in θ, ψ is separable in the sense of Doob as it is almost surely continuous in θ (Gikhman and Skorokhod (2004), p.164). This condition assures measurability of the suprema[10] given below and in Lemma 2.C.1.

[10] Many other authors such as e.g. Andrews (1994), Newey and McFadden (1994), and van der Vaart (1998) rely on outer probability in order to avoid these measurability issues.

In oder to show that ψ has a unique root at θ_0, let us first define the sets

$$U_\theta = \big\{\omega \in \Omega \big| X(\omega)'\theta^q \neq X(\omega)'\theta_0^q\big\}, \qquad \text{and} \qquad W_\theta = \big\{\omega \in \Omega \big| X(\omega)'\theta^q = X(\omega)'\theta_0^q\big\}, \tag{2.32}$$

for all $\theta \in \Theta$ such that $\Omega = W_\theta \cup U_\theta$ and $W_\theta \cap U_\theta = \emptyset$. We first show that $\mathbb{P}(U_\theta) > 0$ for all $\theta \neq \theta_0$. In order to see this, we assume the converse, i.e. let us assume that for a fixed $\theta \neq \theta_0$, it holds that $\mathbb{P}(W_\theta) = \mathbb{P}\big(X'\theta^q = X'\theta_0^q\big) = 1$, which implies that

$$(\theta^q - \theta_0^q)'\, \mathbb{E}[XX']\, (\theta^q - \theta_0^q) = \mathbb{E}\big[\big(X'\theta^q - X'\theta_0^q\big)^2\big] = 0. \tag{2.33}$$

However, since $\theta^q \neq \theta_0^q$, this contradicts the assumption that the matrix $\mathbb{E}[XX']$ is positive definite and we can conclude that $\mathbb{P}(U_\theta) > 0$.

The quantity

$$\lambda_1(\theta) = \mathbb{E}\big[\psi_1(Y, X, \theta)\big] = 1/\alpha\, \mathbb{E}\left[X\big(\alpha G_1^{(1)}(X'\theta^q) + G_2(X'\theta^e)\big)\big(F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)\big)\right]$$

exists under the moment conditions ($\mathcal{M}$-1) in Appendix 2.A and if $\theta^q = \theta_0^q$, it holds that $\lambda_1(\theta) = 0$. Now, we assume that $\theta \in \Theta$ such that $\theta^q \neq \theta_0^q$. By splitting the expectation, we get that

$$\begin{aligned}
&\lambda_1(\theta)'(\theta^q - \theta_0^q)\\
&= 1/\alpha\, \mathbb{E}\left[\big(\alpha G_1^{(1)}(X'\theta^q) + G_2(X'\theta^e)\big)\big(X'\theta^q - X'\theta_0^q\big)\big(F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)\big)\mathbb{1}_{\{\omega \in W_\theta\}}\right]\\
&+ 1/\alpha\, \mathbb{E}\left[\big(\alpha G_1^{(1)}(X'\theta^q) + G_2(X'\theta^e)\big)\big(X'\theta^q - X'\theta_0^q\big)\big(F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)\big)\mathbb{1}_{\{\omega \in U_\theta\}}\right].
\end{aligned}$$

The first summand is obviously zero since for all $\omega \in W_\theta$, $F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q) = 0$. Since the distribution of Y given X has strictly positive density in a neighbourhood of $X'\theta_0^q$, we get that $F_{Y|X}$ is strictly increasing in a neighbourhood of $X'\theta_0^q$ and thus

$$\big(X'\theta^q - X'\theta_0^q\big)\big(F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)\big) > 0 \tag{2.34}$$

for all $\omega \in U_\theta$. Furthermore, since $\alpha G_1^{(1)}(X'\theta^q) + G_2(X'\theta^e) > 0$ for all $\theta \in \Theta$ and $\mathbb{P}(U_\theta) > 0$, we get that

$$\begin{aligned}
&\lambda_1(\theta)'(\theta^q - \theta_0^q)\\
&= 1/\alpha\, \mathbb{E}\left[\big(\alpha G_1^{(1)}(X'\theta^q) + G_2(X'\theta^e)\big)\big(X'\theta^q - X'\theta_0^q\big)\big(F_{Y|X}(X'\theta^q) - F_{Y|X}(X'\theta_0^q)\big)\mathbb{1}_{\{\omega \in U_\theta\}}\right] > 0,
\end{aligned}$$

and consequently $\lambda_1(\theta) \neq 0$. This implies that $\lambda_1(\theta) = 0$ if and only if $\theta^q = \theta_0^q$. Furthermore,

$$\lambda_2(\theta) = \mathbb{E}\left[X G_2^{(1)}(X'\theta^e) \left(X'\theta^q \left(F_{Y|X}(X'\theta^q) - \alpha \right)/\alpha + X'\theta^e - 1/\alpha \, \mathbb{E}\big[Y \mathbb{1}_{\{Y \leq X'\theta^q\}} \big| X \big] \right) \right]. \tag{2.35}$$

Assuming that $\theta^q = \theta_0^q$, which results from $\lambda_1(\theta) = 0$, we get that $F_{Y|X}(X'\theta^q) = F_{Y|X}(X'\theta_0^q) = \alpha$ and $1/\alpha \, \mathbb{E}\big[Y \mathbb{1}_{\{Y \leq X'\theta_0^q\}} \big| X \big] = X'\theta_0^e$. Thus, (2.35) simplifies to $\mathbb{E}\big[(XX') G_2^{(1)}(X'\theta^e) \big] \big(\theta^e - \theta_0^e \big)$ and by applying Lemma 2.C.2, we get that the matrix $\mathbb{E}\big[(XX') G_2^{(1)}(X'\theta^e) \big]$ is positive definite for all $\theta \in \Theta$. Consequently, $\lambda_2(\theta) = 0$ if and only if $\theta^e = \theta_0^e$ and together with the arguments for λ_1, we get that $\lambda(\theta) = 0$ if and only if $\theta = \theta_0$. Eventually, assumption (B-2)' from Theorem 2 of Huber (1967) follows directly from Lemma 2.C.1, which concludes this proof. □

Proof of Theorem 2.2.5. For this proof, we apply Theorem 5.7 from van der Vaart (1998) and show that the respective assumptions of this theorem hold. As in the proof of Theorem 2.2.6, we can conclude measurability of the suprema since the process ρ is continuous and consequently separable in the sense of Doob. Thus, we do not have to rely on outer probability measures such as in van der Vaart (1998). We start by showing uniform convergence in probability of the empirical mean of the objective function by the help of Lemma 2.4 of Newey and McFadden (1994). Since we have iid data, a compact parameter space Θ and $\rho(Y, X, \theta)$ is continuous for all $\theta \in \Theta$, it remains to show that there exists a dominating function $d(Y, X) \geq \big| \rho(Y, X, \theta) \big|$ for all $\theta \in \Theta$ with $\mathbb{E}\big[d(Y, X) \big] < \infty$. We define

$$\begin{aligned} d(Y, X) = &\sup_{\theta \in \Theta} \left| G_1(X'\theta^q) + 1/\alpha \, G_2(X'\theta^e)(X'\theta^q - Y) \right| + \big| G_1(Y) \big| \\ &+ \sup_{\theta \in \Theta} \big| G_2(X'\theta^e) \big(X'\theta^e - X'\theta^q \big) \big| + \sup_{\theta \in \Theta} \left| \mathcal{G}_2(X'\theta^e) \right| + \big| \alpha G_1(Y) + a(Y) \big| \end{aligned} \tag{2.36}$$

and it holds that $d(Y, X) \geq \big| \rho(Y, X, \theta) \big|$ for all $\theta \in \Theta$ and consequently, we can conclude uniform convergence in probability.

We now show that $\mathbb{E}\big[\rho(Y, X, \theta) \big]$ has a unique and global minimum at $\theta = \theta_0$. For this, we assume that $\theta \in \Theta$ such that $\theta \neq \theta_0$ and we define the sets

$$U_\theta = \big\{ \omega \in \Omega \big| X(\omega)'\theta^q \neq X(\omega)'\theta_0^q \quad \text{or} \quad X(\omega)'\theta^e \neq X(\omega)'\theta_0^e \big\} \quad \text{and} \tag{2.37}$$

$$W_\theta = \big\{ \omega \in \Omega \big| X(\omega)'\theta^q = X(\omega)'\theta_0^q \quad \text{and} \quad X(\omega)'\theta^e = X(\omega)'\theta_0^e \big\}, \tag{2.38}$$

such that $\Omega = U_\theta \cup W_\theta$ and $U_\theta \cap W_\theta = \emptyset$. We first show that $\mathbb{P}(U_\theta) > 0$ for all $\theta \neq \theta_0$. In order to see this, we assume the converse, i.e. we assume that $\mathbb{P}(W_\theta) = 1$, which implies that $(\theta^q - \theta_0^q)' \, \mathbb{E}[XX'] \, (\theta^q - \theta_0^q) = \mathbb{E}\left[\left(X'\theta^q - X'\theta_0^q\right)^2\right] = 0$, since $\mathbb{P}\big(X'\theta^q = X\theta_0^q\big) = 1$ and equivalently $(\theta^e - \theta_0^e)'\mathbb{E}[XX'](\theta^e - \theta_0^e) = 0$. However, since $\theta \neq \theta_0$ and consequently either $\theta^q \neq \theta_0^q$ or $\theta^e \neq \theta_0^e$, this contradicts the assumption that the matrix $\mathbb{E}[XX']$ is positive definite and it follows that $\mathbb{P}(U_\theta) > 0$.

From the joint elicitability property of the quantile and ES of Fissler and Ziegel (2016), Corollary 5.5 we get that for all $x \in \mathbb{R}^k$ such that $x'\theta^q \neq x'\theta_0^q$ or $x'\theta^e \neq x'\theta_0^e$, it holds that

$$\mathbb{E}\big[\rho(Y, X, \theta_0)\big|X = x\big] < \mathbb{E}\big[\rho(Y, X, \theta)\big|X = x\big], \tag{2.39}$$

since the distribution of Y given X has a finite first moment and a unique α-quantile. Thus, for all $\omega \in U_\theta$,

$$\mathbb{E}\big[\rho(Y, X, \theta_0)\big|X\big](\omega) < \mathbb{E}\big[\rho(Y, X, \theta)\big|X\big](\omega). \tag{2.40}$$

We now define the random variable

$$h(X, \theta, \theta_0)(\omega) = \mathbb{E}\big[\rho(Y, X, \theta_0)\big|X\big](\omega) - \mathbb{E}\big[\rho(Y, X, \theta)\big|X\big](\omega), \tag{2.41}$$

and (2.40) implies that $h\big(X, \theta, \theta_0\big)(\omega) < 0$ for all $\omega \in U_\theta$. Since $\mathbb{P}(U_\theta) > 0$, this implies that $\mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in U_\theta\}}\right] < 0$. Furthermore, for all $\omega \in W_\theta$, it obviously holds that $h(X, \theta, \theta_0)(\omega) = 0$ and consequently $\mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in W_\theta\}}\right] = 0$. Thus, we get that

$$\mathbb{E}\big[h(X, \theta, \theta_0)\big] = \mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in U_\theta\}}\right] + \mathbb{E}\left[h(X, \theta, \theta_0)\mathbb{1}_{\{\omega \in W_\theta\}}\right] < 0 \tag{2.42}$$

for all $\theta \in \Theta$ such that $\theta \neq \theta_0$, which shows that $\mathbb{E}\big[\rho(Y, X, \theta)\big]$ has a unique minimum at $\theta = \theta_0$. □

Proof of Theorem 2.2.6. We apply Theorem 3 of Huber (1967) for the ψ-function as given in (2.3) and show the respective assumptions of this theorem. Consistency of the Z-estimator is shown in Theorem 2.2.4. For the measureability and separability of the ψ function, we refer to the proof of Theorem 2.2.4. It is already shown in the proof of Theorem 2.2.4 that there exists a $\theta_0 \in \Theta$ such that $\lambda(\theta_0) = 0$. For the technical conditions (N-3), we apply Lemma 2.C.3, Lemma 2.C.1 and Lemma 2.C.4. It remains to show that $\mathbb{E}\big[||\psi(Y, X, \theta_0)||^2\big] < \infty$, which follows from the subsequent computation of C and the

Moment Conditions ($\mathcal{M}$-3) in Appendix 2.A. The asymptotic covariance matrix is given by $\Lambda^{-1}C\Lambda^{-1}$, where $C = \mathbb{E}\big[\psi(Y,X,\theta_0)\,\psi(Y,X,\theta_0)'\big]$ and

$$\Lambda = \frac{\partial\lambda(\theta)}{\partial\theta}\bigg|_{\theta=\theta_0} = \begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{pmatrix} = \begin{pmatrix}\frac{\partial\lambda_1(\theta)}{\partial\theta^q}\Big|_{\theta_0} & \frac{\partial\lambda_1(\theta)}{\partial\theta^e}\Big|_{\theta_0}\\ \frac{\partial\lambda_2(\theta)}{\partial\theta^q}\Big|_{\theta_0} & \frac{\partial\lambda_2(\theta)}{\partial\theta^e}\Big|_{\theta_0}\end{pmatrix}. \tag{2.43}$$

Straightforward calculations yield the matrix C as given in (3.17) - (3.19). For the computation of Λ, we first notice that the function

$$\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big] = \begin{pmatrix}\frac{1}{\alpha}\big(F_{Y|X}(X'\theta^q)-\alpha\big)\big(\alpha G_1^{(1)}(X'\theta^q)+G_2(X'\theta^e)\big)\\ XG_2^{(1)}(X'\theta^e)\Big(X'\theta^e - X'\theta^q + \frac{1}{\alpha}\mathbb{E}\big[(X'\theta^q-Y)\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big]\Big)\end{pmatrix} \tag{2.44}$$

is continuously differentiable for all θ in some neighborhood $U_d(\theta_0)$ around θ_0, since the distribution $F_{Y|X}$ has a density which is strictly positive, continuous and bounded in this area. Let us choose a value $\tilde\theta\in U_d(\theta_0)$ such that $X'\tilde\theta\le X'\theta$. Then,

$$\begin{aligned}\frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big] &= \frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X'\tilde\theta^q\}}\big|X\big] + \frac{\partial}{\partial\theta^q}\mathbb{E}\big[Y\mathbb{1}_{\{X'\tilde\theta^q<Y\le X'\theta^q\}}\big|X\big]\\ &= \frac{\partial}{\partial\theta^q}\int_{X'\tilde\theta^q}^{X'\theta^q} y f_{Y|X}(y)\mathrm{d}y = X(X'\theta^q)f_{Y|X}(X'\theta^q).\end{aligned} \tag{2.45}$$

We consequently get that for all $\theta\in U_d(\theta_0)$,

$$\begin{aligned}\frac{\partial}{\partial\theta^q}\mathbb{E}\big[\psi_1(Y,X,\theta)\big|X\big] &= 1/\alpha\,(XX')\Big[\big(\alpha G_1^{(1)}(X'\theta^q)+G_2(X'\theta^e)\big)f_{Y|X}(X'\theta^q)\\ &\qquad + G_1^{(2)}(X'\theta^q)\big(F_{Y|X}(X'\theta^q)-\alpha\big)\Big],\\ \frac{\partial}{\partial\theta^e}\mathbb{E}\big[\psi_1(Y,X,\theta)\big|X\big] &= \frac{\partial}{\partial\theta^q}\mathbb{E}\big[\psi_2(Y,X,\theta)\big|X\big] = 1/\alpha\,(XX')G_2^{(1)}(X'\theta^e)\big(F_{Y|X}(X'\theta^q)-\alpha\big),\\ \frac{\partial}{\partial\theta^e}\mathbb{E}\big[\psi_2(Y,X,\theta)\big|X\big] &= 1/\alpha\,(XX')G_2^{(2)}(X'\theta^e)\,\Big[X'\theta^q\big(F_{Y|X}(X'\theta^q)-\alpha\big)+\alpha(X'\theta^e)-\mathbb{E}\big[Y\mathbb{1}_{\{Y\le X'\theta^q\}}\big|X\big]\Big]\\ &\qquad + (XX')G_2^{(1)}(X'\theta^e).\end{aligned}$$

In order to conclude that $\frac{\partial}{\partial\theta}\mathbb{E}\big[\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]\big] = \mathbb{E}\big[\frac{\partial}{\partial\theta}\mathbb{E}\big[\psi(Y,X,\theta)\big|X\big]\big]$, we apply a measure-theoretical version of the Leibniz integration rule, which requires that the derivative of the integrand exists and is absolutely bounded by some integrable function $d(Y,X)$, independent of θ. For the first term, this can easily be obtained by defining

$$d(Y,X) = \sup_{\theta\in U_d(\theta_0)}\bigg\|1/\alpha\,(XX')\Big[\big(\alpha G_1^{(1)}(X'\theta^q)+G_2(X'\theta^e)\big)f_{Y|X}(X'\theta^q)+G_1^{(2)}(X'\theta^q)\big(F_{Y|X}(X'\theta^q)-\alpha\big)\Big]\bigg\|,$$

which has finite expectation by the Moment Conditions ($\mathcal{M}$-3). The other two terms follow the same reasoning. Inserting $\theta = \theta_0$ eventually shows (3.15) and (3.16). □

Proof of Theorem 2.2.7. For this proof, we apply Theorem 5.23 from van der Vaart (1998) and show that the respective assumptions of this theorem hold. Theorem 2.2.5 shows consistency of the M-estimator. The map $(Y, X) \mapsto \rho(Y, X, \theta)$ is obviously measurable as the sum of measurable functions. Furthermore, the map $\theta \mapsto \rho(Y, X, \theta)$ is almost surely differentiable since the only point of non-differentiability occurs where $Y = X'\theta^q$, which is a nullset with respect to the joint distribution of Y and X and for all $\theta \in \Theta$ such that $Y \neq X'\theta^q$, its derivative is given by $\psi(Y, X, \theta)$. Local Lipschitz continuity with square-integrable Lipschitz-constant follows from Lemma 2.C.5. We have already seen in the proof of Theorem 2.2.5 that the function $\mathbb{E}\big[\rho(Y, X, \theta)\big]$ is uniquely minimized at the point θ_0 and is twice continuously differentiable and consequently admits a second-order Taylor expansion at θ_0. Thus, we have shown the necessary assumptions of Theorem 5.23 from van der Vaart (1998).

For the computation of the covariance matrix, we notice that the distribution of Y given X has a density $f_{Y|X}$ in a neighborhood of $X'\theta_0$, which is strictly positive, continuous and bounded. Therefore, by the same arguments as in (2.45), we get that $\frac{\partial}{\partial\theta^q}\mathbb{E}\big[G_1(Y)\mathbb{1}_{\{Y \le X'\theta^q\}}\big|X\big] = XG_1(X'\theta^q)f_{Y|X}(X'\theta^q)$. Thus, straight-forward calculations yield that for all $\theta \in U_d(\theta_0)$, it holds that $\frac{\partial}{\partial\theta}\mathbb{E}\big[\rho(Y, X, \theta)\big|X\big] = \mathbb{E}\big[\psi(Y, X, \theta)\big|X\big]$ and by applying the Leibniz integration rule such as in the proof of Theorem 2.2.6, we finally get that

$$\frac{\partial}{\partial\theta}\mathbb{E}\big[\rho(Y, X, \theta)\big] = \mathbb{E}\big[\psi(Y, X, \theta)\big]. \tag{2.46}$$

Consequently, the asymptotic covariance matrix equals the one given in Theorem 2.2.6. □

Appendix 2.C Technical Results

Lemma 2.C.1. Let

$$u(Y, X, \theta, d) = \sup_{\tau \in \bar{U}_d(\theta)} \big\|\psi(Y, X, \tau) - \psi(Y, X, \theta)\big\| \tag{2.47}$$

and assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions ($\mathcal{M}$-1) in Appendix 2.A hold. Then, there are strictly positive real numbers b and d_0, such that

$$\mathbb{E}\big[u(Y, X, \theta, d)\big] \le b \cdot d \qquad \text{for} \quad ||\theta - \theta_0|| + d \le d_0, \tag{2.48}$$

and for all $d \geq 0$.

Proof of Lemma 2.C.1. For measurability of the suprema, we refer to the proof of Theorem 2.2.4. Let in the following $d > 0$ and $\theta \in \Theta$ such that $||\theta - \theta_0|| + d \leq d_0$. We first notice that for some fixed $X \in \mathbb{R}^k$ and for all $\tau \in \bar{U}_d(\theta)$, it holds that

$$\left|\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right| \leq \mathbb{1}_{\{X'\theta^q_- \leq Y \leq X'\theta^q_+\}} \tag{2.49}$$

for all $Y \in \mathbb{R}$ and for some $\theta^q_-, \theta^q_+ \in \bar{U}_d(\theta)$. Since $\bar{U}_d(\theta)$ is compact, we get that

$$\sup_{\tau \in \bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right| \leq \mathbb{1}_{\{X'\theta^q_- \leq Y \leq X'\theta^q_+\}} \tag{2.50}$$

for all $Y \in \mathbb{R}$ and for some values $\theta^q_-, \theta^q_+ \in \bar{U}_d(\theta)$. Note that the values θ^q_- and θ^q_+ depend on X and θ, however they are independent of Y. Consequently, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right| \middle| X\right] \leq \mathbb{E}\left[\mathbb{1}_{\{X'\theta^q_- \leq Y \leq X'\theta^q_+\}} \middle| X\right] \\
&= F_{Y|X}\big(X'\theta^q_+\big) - F_{Y|X}\big(X'\theta^q_-\big) = f_{Y|X}(X'\tilde{\theta}^q)\big(X'\theta^q_+ - X'\theta^q_-\big) \\
&\leq 2||X|| \cdot \sup_{\tau \in \bar{U}_d(\theta)} f_{Y|X}(X'\tau^q) \cdot d,
\end{aligned} \tag{2.51}$$

where we apply the mean value theorem for some $\tilde{\theta}^q$ on the line between θ^q_- and θ^q_+, i.e. $\tilde{\theta}^q \in \bar{U}_d(\theta)$.

For the first component of ψ, we get that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|\left|\psi_1(Y, X, \theta) - \psi_1(Y, X, \tau)\right|\right|\right] \\
&\leq \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|\left| X \left(G_1^{(1)}(X'\theta^q) - G_1^{(1)}(X'\tau^q) + \frac{G_2(X'\theta^e) - G_2(X'\tau^e)}{\alpha}\right)\right|\right|\right] \\
&\quad + \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|\left| X \left(G_1^{(1)}(X'\tau^q) + \frac{G_2(X'\tau^e)}{\alpha}\right)\right|\right| \cdot \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right| \middle| X\right]\right].
\end{aligned} \tag{2.52}$$

The first term in (2.52) is $\mathcal{O}(d)$ since $G_1^{(1)}(X'\theta^q)$ and $G_2(X'\theta^e)$ are continuously differentiable functions w.r.t θ and thus, by the mean value theorem we get that

$$\begin{aligned} \sup_{\tau \in \bar{U}_d(\theta)} \left| G_1^{(1)}(X'\theta^q) - G_1^{(1)}(X'\tau^q) \right| &\leq \sup_{\tilde{\tau} \in \bar{U}_d(\theta)} \left\| X G_1^{(2)}(X'\tilde{\tau}^q) \right\| \cdot \sup_{\tau \in \bar{U}_d(\theta)} \left\| \theta^q - \tau^q \right\| \\ &\leq \sup_{\tilde{\tau} \in \bar{U}_d(\theta)} \left\| X G_1^{(2)}(X'\tilde{\tau}^q) \right\| \cdot d, \end{aligned} \tag{2.53}$$

and the respective moments are finite by assumption. The same arguments hold for the function G_2. For the second term in (2.52), we apply (2.51) and thus get that

$$\begin{aligned} &\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| X \left(G_1^{(1)}(X'\tau^q) + \frac{G_2(X'\tau^e)}{\alpha} \right) \right\| \cdot \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left| \mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}} \right| \middle| X \right] \right] \\ \leq\ &\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| X \left(G_1^{(1)}(X'\tau^q) + \frac{G_2(X'\tau^e)}{\alpha} \right) \right\| ||X|| \cdot \sup_{\tau \in \bar{U}_d(\theta)} f_{Y|X}(X'\tau^q) \right] \cdot d. \end{aligned} \tag{2.54}$$

Since the density $f_{Y|X}$ is bounded in a neighborhood of $X'\theta_0^q$ and the respective moments are finite by assumption, we get that this term is also $\mathcal{O}(d)$.

For the second component of ψ, we get that

$$\begin{aligned} &\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| \psi_2(Y, X, \theta) - \psi_2(Y, X, \tau) \right\| \right] \\ \leq\ &\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| X(X'\theta^e - X'\theta^q) G_2^{(1)}(X'\theta^e) - X(X'\tau^e - X'\tau^q) G_2^{(1)}(X'\tau^e) \right\| \right] \\ &\quad + \mathbb{E}\left[\left\| \frac{X G_2^{(1)}(X'\theta^e) X'\theta^q}{\alpha} \right\| \cdot \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left| \left(\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}} \right) \right| \middle| X \right] \right] \\ &\quad + \mathbb{E}\left[\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| \mathbb{1}_{\{Y \leq X'\tau^q\}} \left(\frac{X G_2^{(1)}(X'\theta^e) X'\theta^q}{\alpha} - \frac{X G_2^{(1)}(X'\tau^e) X'\tau^q}{\alpha} \right) \right\| \middle| X \right] \right] \\ &\quad + \mathbb{E}\left[\left\| \frac{X G_2^{(1)}(X'\theta^e)}{\alpha} \right\| \cdot \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left| Y \left(\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}} \right) \right| \middle| X \right] \right] \\ &\quad + \mathbb{E}\left[\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| \frac{Y \mathbb{1}_{\{Y \leq X'\tau^q\}}}{\alpha} \left(X G_2^{(1)}(X'\theta^e) - X G_2^{(1)}(X'\tau^e) \right) \right\| \middle| X \right] \right] \\ =\ &\text{(i)} + \text{(ii)} + \text{(iii)} + \text{(iv)} + \text{(v)}. \end{aligned}$$

The first, third and fifth term are linearly bounded by (2.53) since the functions $(X'\theta^e - X'\theta^q) G_2^{(1)}(X'\theta^e)$ and $(X'\theta^q) G_2^{(1)}(X'\theta^e)$ and $G_2^{(1)}(X'\theta^e)$ are continuously differentiable.

For the second term, we use the arguments from (2.51). For the fourth term, we use similar arguments as in (2.51), and get that there exist some $\theta^q_-, \theta^q_+ \in \bar{U}_d(\theta)$ and a value $\tilde{\theta}^q$ on the line between θ^q_- and θ^q_+, such that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\| \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|Y\left(\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right)\right| \,\middle|\, X\right]\right] \\
\leq\ &\mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\| \mathbb{E}\left[|Y|\ \mathbb{1}_{\{X'\theta^q_- \leq Y \leq X'\theta^q_+\}} \,\middle|\, X\right]\right] \\
=\ &\mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\| \int_{X'\theta^q_-}^{X'\theta^q_+} |y| f_{Y|X}(y)\mathrm{d}y\right] \qquad (2.55) \\
\leq\ &\mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\| |X'\tilde{\theta}^q| f_{Y|X}(X'\tilde{\theta}^q)\big(X'\theta^q_+ - X'\theta^q_-\big)\right] \\
\leq\ &\frac{2}{\alpha}\mathbb{E}\left[G_2^{(1)}(X'\theta^e)\|X\|^2 \sup_{\tau \in \bar{U}_d(\theta)} |X'\tau^q| f_{Y|X}(X'\tau^q)\right] \cdot d = \mathcal{O}(d)
\end{aligned}$$

since $f_{Y|X}$ is bounded in a neighborhood of $X'\theta_0$ and the respective moments exist by assumption. This concludes the proof of the lemma. □

Lemma 2.C.2. Let the random variable $X \in \mathbb{R}^k$ with distribution $\mathbb{P}$ be such that its second moments exist and the matrix $\mathbb{E}[XX']$ is positive definite. Furthermore, let $\tilde{\Theta} \subset \mathbb{R}^k$ be a compact subspace with nonempty interior and let $g : \mathbb{R}^k \times \tilde{\Theta} \to \mathbb{R}$ be a strictly positive function. Then, the matrix

$$\mathbb{E}\big[(XX')g(X,\theta)\big] \qquad (2.56)$$

is also positive definite.

Proof of Lemma 2.C.2. Since $\mathbb{E}[XX']$ is positive definite, we know that for all $z \in \mathbb{R}^k$ with $z \neq 0$, it holds that $0 < z'\mathbb{E}[XX']z = \mathbb{E}[z'(XX')z] = \mathbb{E}[(X'z)^2]$ and consequently $\mathbb{P}\big(X'z \neq 0\big) > 0$. Since $\sqrt{g(X,\theta)}$ is a strictly positive scalar for all $\theta \in \tilde{\Theta}$, it also holds that $\mathbb{P}\big((X'z)\sqrt{g(X,\theta)} \neq 0\big) > 0$ and thus, for all $z \neq 0$,

$$z'\mathbb{E}\big[(XX')g(X,\theta)\big]z = \mathbb{E}\left[\left(X'z\sqrt{g(X,\theta)}\right)^2\right] > 0. \qquad (2.57)$$

This positivity statement holds since $\big(X'z\sqrt{g(X,\theta)}\big)^2$ is a non-negative random variable and $\mathbb{P}\big((X'z)\sqrt{g(X,\theta)} \neq 0\big) > 0$. This shows that the matrix $\mathbb{E}\big[(XX')g(X,\theta)\big]$ is positive definite. □

Lemma 2.C.3. Assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions ($\mathcal{M}$-3) in Appendix 2.A hold. Then, for

$$\lambda(\theta) = \mathbb{E}\big[\psi(Y, X, \theta)\big], \tag{2.58}$$

there are strictly positive numbers a, d_0, such that

$$||\lambda(\theta)|| \geq a \cdot ||\theta - \theta_0|| \qquad \text{for} \quad ||\theta - \theta_0|| \leq d_0. \tag{2.59}$$

Proof of Lemma 2.C.3. Let $d_0 > 0$ and let $||\theta - \theta_0|| \leq d_0$. Then, applying the mean value theorem, we get that

$$\lambda_1(\theta) = \frac{1}{\alpha}\mathbb{E}\left[(XX')\big(\alpha G_1^{(1)}(X'\theta^q) + G_2(X'\theta^e)\big) f_{Y|X}(X'\tilde{\theta}^q)\right](\theta^q - \theta_0^q) \tag{2.60}$$

for some $\tilde{\theta}^q$ on the line between θ^q and θ_0^q. Similarly, for the second component we get that

$$\begin{aligned} \lambda_2(\theta) = \mathbb{E}&\left[X\frac{G_2^{(1)}(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha}\big[X'(\theta^q - \theta_0^q)\big]\big[X'(\tilde{\theta}^q - \theta^q)\big]\right] \\ &+ \mathbb{E}\big[(XX')G_2^{(1)}(X'\theta^e)\big](\theta^e - \theta_0^e), \end{aligned} \tag{2.61}$$

where $\tilde{\theta}^q$ lies on the line between θ^q and θ_0^q.

We first assume that $||\theta - \theta_0|| = ||\theta^q - \theta_0^q||$, i.e. $||\theta^q - \theta_0^q|| \geq ||\theta^e - \theta_0^e||$. Since the matrix

$$A(\theta) := \mathbb{E}\left[(XX')\frac{\big(\alpha G_1^{(1)}(X'\theta^q) + G_2(X'\theta^e)\big)}{\alpha} f_{Y|X}(X'\tilde{\theta}^q)\right] \tag{2.62}$$

exists and has full rank for all $\theta \in \Theta$ by Lemma 2.C.2 and is obviously symmetric, A has strictly positive real Eigenvalues $\gamma_1(\theta), \dots, \gamma_k(\theta)$ with minimum $\gamma_{(1)}(\theta)$ and we thus get that[11]

$$||\lambda(\theta)|| \geq ||\lambda_1(\theta)|| = ||A(\theta)(\theta^q - \theta_0^q)|| \geq \gamma_{(1)}(\theta) \cdot ||\theta^q - \theta_0^q|| \tag{2.63}$$

$$\geq \left(\inf_{||\theta - \theta_0|| \leq d_0} \gamma_{(1)}(\theta)\right) \cdot ||\theta^q - \theta_0^q|| = c_1\, ||\theta - \theta_0||. \tag{2.64}$$

[11] For a symmetric matrix A with full rank, we can find an orthogonal basis of Eigenvectors $\{v_1, \dots, v_k\}$ with corresponding nonzero Eigenvalues $\{\gamma_1(\theta), \dots, \gamma_k(\theta)\}$ such that $x = \sum b_j v_j$ with $b_j \in \mathbb{R}$. Then, $||Ax|| = ||A \sum b_j v_j|| = ||\sum b_j A v_j|| = ||\sum b_j \gamma_j v_j|| \geq \min|\gamma_j| \cdot ||\sum b_j v_j|| = \min|\gamma_j| \cdot ||x||$.

Since $||\theta - \theta_0|| \leq d_0$ is a compact set and the function $\theta \mapsto \inf_{||\theta-\theta_0||\leq d_0} \gamma_{(1)}(\theta)$, where $\gamma_{(1)}(\theta)$ is the smallest Eigenvalue of the matrix $A(\theta)$, is continuous[12], we get that the infimum coincides with the minimum and thus, the constant $c_1 := \inf_{||\theta-\theta_0||\leq d_0} \gamma_{(1)}(\theta)$ is strictly positive and does not depend on θ.

Now, we assume that $||\theta - \theta_0|| = ||\theta^e - \theta_0^e|| \leq d_0$, i.e. $||\theta^e - \theta_0^e|| \geq ||\theta^q - \theta_0^q||$. For the first term of $\lambda_2(\theta)$, given in (2.61), we define the vector

$$b(\theta) := \mathbb{E}\left[X \frac{G_2^{(1)}(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha} \left[X'(\theta^q - \theta_0^q)\right] \left[X'\tilde{\theta}^q - X'\theta^q\right] \right], \tag{2.65}$$

and for its l-th component, we get that

$$\begin{aligned} |b_l(\theta)| &= \left| \sum_{i,j} (\theta_i^q - \theta_{0i}^q)(\tilde{\theta}_j^q - \theta_j^q) \mathbb{E}\left[X_i X_j X_l \frac{G_2^{(1)}(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha} \right] \right| \\ &\leq \sum_{i,j} \mathbb{E}\left[\left| X_i X_j X_l \frac{G_2^{(1)}(X'\theta^e) f_{Y|X}(X'\tilde{\theta}^q)}{\alpha} \right| \right] \cdot |\theta_i^q - \theta_{0i}^q| \cdot |\tilde{\theta}_j^q - \theta_j^q| \\ &\leq c_2 \sum_{i,j} |\theta_i^q - \theta_{0i}^q| \cdot |\tilde{\theta}_j^q - \theta_j^q| \\ &\leq c_2 k^2 ||\theta - \theta_0||^2, \end{aligned} \tag{2.66}$$

for all $l = 1, \dots, k$, which implies that

$$||b(\theta)|| \leq c_3 ||\theta - \theta_0||^2, \tag{2.67}$$

for some $c_3 > 0$. For $D(\theta) := \mathbb{E}\left[(XX')G_2^{(1)}(X'\theta^e)\right]$, it holds that $||D(\theta)(\theta^e - \theta_0^e)|| \geq c_4||\theta^e - \theta_0^e|| = c_4||\theta - \theta_0||$ for $c_4 > 0$ by the same arguments as in (2.63). From (2.66), we can choose d_0 small enough such that

$$2||b(\theta)|| \leq 2c_3||\theta - \theta_0||^2 \leq c_4||\theta - \theta_0|| \leq ||D(\theta)(\theta^e - \theta_0^e)||. \tag{2.68}$$

[12] This follows since the entries of the matrix $A(\theta)$ are continuous in θ as the expectation of a continuous function which is dominated by an integrable function is again continuous by the dominated convergence theorem. Furthermore, the Eigenvalues of a matrix are the solution of the characteristic polynomial, which has continuous coefficients since our matrix entries are continuous in θ. Eventually, since the roots of any polynomial with continuous coefficients are again continuous, we can conclude that the Eigenvalues of $A(\theta)$ are continuous in θ.

Furthermore, by the submultiplicativity of the matrix norm, we also get that $||D(\theta)(\theta^e - \theta_0^e)|| \leq ||D(\theta)|| \cdot ||\theta^e - \theta_0^e|| = c_5||\theta^e - \theta_0^e||$ and by the inverse triangle inequality, we get that

$$||\lambda(\theta)|| \geq ||\lambda_2(\theta)|| = \left\|D(\theta)(\theta^e - \theta_0^e) + b(\theta)\right\| \geq \Big| ||D(\theta)(\theta^e - \theta_0^e)|| - ||b(\theta)|| \Big|. \tag{2.69}$$

From (2.68), we can choose d_0 small enough such that $||D(\theta^e - \theta_0^e)|| > 2||b||$ and thus

$$\Big| ||D(\theta^e - \theta_0^e)|| - ||b|| \Big| = ||D(\theta^e - \theta_0^e)|| - ||b|| \geq \frac{1}{2}||D(\theta^e - \theta_0^e)|| \tag{2.70}$$

$$\geq \frac{c_4}{2}||\theta^e - \theta_0^e|| = \frac{c_4}{2}||\theta - \theta_0||. \tag{2.71}$$

□

Lemma 2.C.4. Let

$$u(Y, X, \theta, d) = \sup_{\tau \in \bar{U}_d(\theta)} \left\|\psi(Y, X, \tau) - \psi(Y, X, \theta)\right\|. \tag{2.72}$$

and assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions ($\mathcal{M}$-3) in Appendix 2.A hold. Then, there are strictly positive numbers c and d_0, such that

$$\mathbb{E}\left[u(Y, X, \theta, d)^2\right] \leq c \cdot d \qquad \text{for} \quad ||\theta - \theta_0|| + d \leq d_0, \tag{2.73}$$

and for all $d \geq 0$.

Proof of Lemma 2.C.4. Let in the following $d > 0$ and $\theta \in \Theta$ such that $||\theta - \theta_0|| + d \leq d_0$. It holds that

$$\left(\sup_{\tau \in \bar{U}_d(\theta)} \left\|\psi(Y, X, \tau) - \psi(Y, X, \theta)\right\|\right)^2 = \sup_{\tau \in \bar{U}_d(\theta)} \left\|\psi(Y, X, \tau) - \psi(Y, X, \theta)\right\|^2 \tag{2.74}$$

and consequently, we show that

$$\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\|\psi_j(Y, X, \tau) - \psi_j(Y, X, \theta)\right\|^2\right] = \mathcal{O}(d) \tag{2.75}$$

for both components $j = 1, 2$ and for some $d > 0$ small enough.

For the first squared component, we get that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)} \left\|\psi_1(Y,X,\tau)-\psi_1(Y,X,\theta)\right\|^2\right]\\
\leq\ & \max\left(\left|\frac{1-\alpha}{\alpha}\right|^2, 1\right)\cdot\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X\left(\alpha G_1^{(1)}(X'\theta^q)+G_2(X'\theta^e)-\alpha G_1^{(1)}(X'\tau^q)-G_2(X'\tau^e)\right)\right\|^2\right]\\
&+\frac{2}{\alpha^2}\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X\left(\alpha G_1^{(1)}(X'\tau^q)+G_2(X'\tau^e)\right)\right\|^2 ||X|| \sup_{\tau\in\bar{U}_d(\theta)} f_{Y|X}(X'\tau^q)\right]\cdot d\\
&+\frac{2}{\alpha^2}\max\left(1-\alpha,\alpha\right)\mathbb{E}\left[\sup_{\tau\in\bar{U}_d(\theta)}\left\|X\left(\alpha G_1^{(1)}(X'\theta^q)+G_2(X'\theta^e)-\alpha G_1^{(1)}(X'\tau^q)-G_2(X'\tau^e)\right)\right\|\right.\\
&\qquad\qquad\left.\cdot\left\|X\left(\alpha G_1^{(1)}(X'\tau^q)+G_2(X'\tau^e)\right)\right\|\right],
\end{aligned}
$$

where we apply (2.51) for the second summand. The remaining two summands can be bounded linearly by the arguments given in (2.53) since $G_1^{(1)}$ and G_2 are continuously differentiable functions and the respective moments are finite.

For the second component of ψ, we get that

$$
\begin{aligned}
&\left\|\psi_2(Y,X,\tau)-\psi_2(Y,X,\theta)\right\|\\
&\leq\left\|X(X'\theta^e-X'\theta^q)G_2^{(1)}(X'\theta^e)-X(X'\tau^e-X'\tau^q)G_2^{(1)}(X'\tau^e)\right\|\\
&\quad+\left\|\frac{XG_2^{(1)}(X'\theta^e)X'\theta^q}{\alpha}\left(\mathbb{1}_{\{Y\leq X'\theta^q\}}-\mathbb{1}_{\{Y\leq X'\tau^q\}}\right)\right\|\\
&\quad+\left\|\mathbb{1}_{\{Y\leq X'\tau^q\}}\left(\frac{XG_2^{(1)}(X'\theta^e)X'\theta^q}{\alpha}-\frac{XG_2^{(1)}(X'\tau^e)X'\tau^q}{\alpha}\right)\right\|\\
&\quad+\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}Y\left(\mathbb{1}_{\{Y\leq X'\theta^q\}}-\mathbb{1}_{\{Y\leq X'\tau^q\}}\right)\right\|\\
&\quad+\left\|\frac{Y\mathbb{1}_{\{Y\leq X'\tau^q\}}}{\alpha}\left(XG_2^{(1)}(X'\theta^e)-XG_2^{(1)}(X'\tau^e)\right)\right\|\\
&=\text{(i)}+\text{(ii)}+\text{(iii)}+\text{(iv)}+\text{(v)}.
\end{aligned}
\tag{2.76}
$$

Thus, in order to evaluate $\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| \psi_2(Y, X, \tau) - \psi_2(Y, X, \theta) \right\|^2\right]$, we have to consider all the cross products out of the five summands in (2.76). Since the techniques applied are very similar, we only show details for two of the cross products.

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} (\text{ii}) \cdot (\text{v}) \right] \\
&= \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left\| \frac{X G_2^{(1)}(X'\theta^e) X'\theta^q}{\alpha} \left(\mathbb{1}_{\{Y \le X'\theta^q\}} - \mathbb{1}_{\{Y \le X'\tau^q\}} \right) \right\| \right. \\
&\qquad \left. \cdot \left\| \frac{Y \mathbb{1}_{\{Y \le X'\tau^q\}}}{\alpha} \left(X G_2^{(1)}(X'\theta^e) - X G_2^{(1)}(X'\tau^e) \right) \right\| \right] \\
&\le \frac{1}{\alpha^2} \mathbb{E}\left[\left\| X G_2^{(1)}(X'\theta^e) X'\theta^q \right\| \cdot \mathbb{E}\big[|Y| \big| X \big] \cdot ||X|| \cdot \sup_{\tau \in \bar{U}_d(\theta)} \left\| G_2^{(1)}(X'\theta^e) - G_2^{(1)}(X'\tau^e) \right\| \right] \\
&\le \frac{1}{\alpha^2} \mathbb{E}\left[\left\| X G_2^{(1)}(X'\theta^e) X'\theta^q \right\| \cdot \mathbb{E}\big[|Y| \big| X \big] \cdot ||X|| \cdot \sup_{\tau \in \bar{U}_d(\theta)} \left\| X G_2^{(2)}(X'\tau^e) \right\| \right] \cdot d \\
&= \mathcal{O}(d),
\end{aligned}
$$

by (2.53) since $G_2^{(1)}$ is continuously differentiable.

The following crossproducts can be bounded analogously by bounding the indicator functions and by applying the mean value theorem as in (2.53): (i)^2, (iii)^2, (v)^2, (i) · (iii), (i) · (iv), (i) · (v), (ii) · (iv), (ii) · (v), (iii) · (iv), (iii) · (v) and (iv) · (v).

A second type of technique, similar to the arguments in (2.55) arises in the cases (ii)^2, (iv)^2 and (ii) $\cdot$ (iv). We get that there exists $\theta^q_-, \theta^q_+ \in \bar{U}_d(\theta)$ and a value $\tilde{\theta}^q$ on the line between θ^q_- and θ^q_+, such that

$$\begin{aligned}
\mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} (\text{iv})^2\right] &\leq \mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\|^2 \mathbb{E}\left[\sup_{\tau \in \bar{U}_d(\theta)} \left|Y\left(\mathbb{1}_{\{Y \leq X'\theta^q\}} - \mathbb{1}_{\{Y \leq X'\tau^q\}}\right)\right|^2 \middle| X\right]\right] \\
&\leq \mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\|^2 \mathbb{E}\left[Y^2\, \mathbb{1}_{\{X'\theta^q_- \leq Y \leq X'\theta^q_+\}} \middle| X\right]\right] \\
&= \mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\|^2 \int_{X'\theta^q_-}^{X'\theta^q_+} y^2 f_{Y|X}(y)\mathrm{d}y\right] \\
&\leq \mathbb{E}\left[\left\|\frac{XG_2^{(1)}(X'\theta^e)}{\alpha}\right\|^2 (X'\tilde{\theta}^q)^2 f_{Y|X}(X'\tilde{\theta}^q)\left(X'\theta^q_+ - X'\theta^q_-\right)\right] \\
&\leq \frac{2}{\alpha}\mathbb{E}\left[\left\|X\right\|^3 \left(G_2^{(1)}(X'\theta^e)\right)^2 \cdot \sup_{\tau \in \bar{U}_d(\theta)} (X'\tau^q)^2 f_{Y|X}(X'\tau^q)\right] \cdot d \\
&= \mathcal{O}(d),
\end{aligned}$$

where we apply a multivariate version of the mean value theorem and notice that $f_{Y|X}$ is bounded. □

Lemma 2.C.5. Assume that Assumption 2.2.1, Assumption 2.2.2 and the Moment Conditions ($\mathcal{M}$-4) in Appendix 2.A hold. Then, the function $\rho(Y, X, \theta)$, given in (2.2) is locally Lipschitz continuous in θ in the sense that for all $\theta_1, \theta_2 \in U_d(\theta_0)$ in some neighborhood of θ_0, it holds that

$$\left|\rho(Y, X, \theta_1) - \rho(Y, X, \theta_2)\right| \leq K(Y, X) \cdot \left\|\theta_1 - \theta_2\right\|, \tag{2.77}$$

where $\mathbb{E}\left[K(Y, X)^2\right] < \infty$.

Proof. We start the proof by splitting the ρ function into two parts,

$$\rho(Y, X, \theta) = \rho_1(Y, X, \theta) + \rho_2(Y, X, \theta), \tag{2.78}$$

where

$$\rho_1(Y,X,\theta) = \mathbb{1}_{\{Y\leq X'\theta^q\}}\left(G_1(X'\theta^q) - G_1(Y) + \frac{1}{\alpha}G_2(X'\theta^e)(X'\theta^q - Y)\right), \tag{2.79}$$

$$\rho_2(Y,X,\theta) = G_2(X'\theta^e)\big(X'\theta^e - X'\theta^q\big) - \mathcal{G}_2(X'\theta^e) - \alpha G_1(X'\theta^q) + a(Y). \tag{2.80}$$

Local Lipschitz continuity of ρ_2 follows since it is a continuously differentiable function and thus locally Lipschitz. We consequently get that for some $d > 0$ and for all $\theta_1, \theta_2 \in U_d(\theta_0)$, it holds that

$$\big|\rho_2(Y,X,\theta_1) - \rho_2(Y,X,\theta_2)\big| \leq \big\|\theta_1 - \theta_2\big\| \cdot \sup_{\theta\in U_d(\theta_0)} \left\| \begin{pmatrix} -XG_2(X'\theta^e) - \alpha XG_1^{(1)}(X'\theta^q) \\ XG_2^{(1)}(X'\theta^e)\big(X'\theta^e - X'\theta^q\big) \end{pmatrix} \right\|, \tag{2.81}$$

with Lipschitz-constant

$$K(Y,X) = \sup_{\theta\in U_d(\theta_0)} \left\| \begin{pmatrix} -XG_2(X'\theta^e) - \alpha XG_1^{(1)}(X'\theta^q) \\ XG_2^{(1)}(X'\theta^e)\big(X'\theta^e - X'\theta^q\big) \end{pmatrix} \right\|, \tag{2.82}$$

which is square-integrable by the moment conditions ($\mathcal{M}$-4).

For the function ρ_1, we consider three cases. First, let $\theta_1, \theta_2 \in \Theta$ such that $X'\theta_1^q \leq X'\theta_2^q < Y$. Then it holds that,

$$\rho_1(Y,X,\theta_1) = \rho_1(Y,X,\theta_2) = 0, \tag{2.83}$$

since $\mathbb{1}_{\{Y\leq X'\theta_1^q\}} = \mathbb{1}_{\{Y\leq X'\theta_2^q\}} = 0$, which is obviously a Lipschitz continuous function.

Second, let $\theta_1, \theta_2 \in \Theta$ such that $Y \leq X'\theta_1^q \leq X'\theta_2^q$. Then, for $\theta = \theta_1, \theta_2$,

$$\rho_1(Y,X,\theta) = G_1(X'\theta^q) - G_1(Y) + \frac{1}{\alpha}G_2(X'\theta^e)(X'\theta^q - Y), \tag{2.84}$$

which is a continuously differentiable function and thus

$$\big|\rho_1(Y,X,\theta_1) - \rho_1(Y,X,\theta_2)\big| \leq \big\|\theta_1 - \theta_2\big\| \cdot \sup_{\theta\in U_d(\theta_0)} \left\| \begin{pmatrix} XG_1^{(1)}(X'\theta^q) + \frac{1}{\alpha}XG_2(X'\theta^e) \\ \frac{1}{\alpha}XG_2^{(1)}(X'\theta^e)(X'\theta^q - Y) \end{pmatrix} \right\|. \tag{2.85}$$

Finally, let $\theta_1, \theta_2 \in \Theta$ such that $X'\theta_1^q < Y \leq X'\theta_2^q$. Then, since G_1 is increasing, we get that

$$\begin{aligned}\left|\rho_1(Y, X, \theta_1) - \rho_1(Y, X, \theta_2)\right| &= \left|G_1(X'\theta_2^q) - G_1(Y) + \frac{1}{\alpha} G_2(X'\theta_2^e)(X'\theta_2^q - Y)\right| \\ &\leq \left|G_1(X'\theta_2^q) - G_1(X'\theta_1^q)\right| + \left|\frac{1}{\alpha} G_2(X'\theta_2^e)(X'\theta_2^q - X'\theta_1^q)\right| \\ &\leq \left\|\theta_1^q - \theta_2^q\right\| \cdot \sup_{\theta \in U_d(\theta_0)} \left(\left\|X G_1^{(1)}(X'\theta^q)\right\| + \frac{1}{\alpha}\left\|X G_2(X'\theta^e)\right\| \right).\end{aligned}$$

Thus, the function $\rho(Y, X, \theta)$ is locally Lipschitz continuous in θ with square-integrable Lipschitz constants, $\mathbb{E}\left[K(Y, X)^2\right] < \infty$ by the Moment Conditions ($\mathcal{M}$-4) in Appendix 2.A. □

Proposition 2.C.6. Let Y be a real-valued random variable with distribution function F, finite first and second moments and a unique α-quantile $q_\alpha = F^{-1}(\alpha)$. Then,

$$\frac{1}{\alpha^2} \int_{-\infty}^{q_\alpha} \int_{-\infty}^{q_\alpha} F(x \wedge y) - F(x)F(y) \mathrm{d}x\mathrm{d}y = \frac{1}{\alpha} \operatorname{Var}(Y | Y \leq q_\alpha) + \frac{1-\alpha}{\alpha} \left(q_\alpha - \xi_\alpha\right)^2, \quad (2.86)$$

where $\xi_\alpha = \mathbb{E}\left[Y \middle| Y \leq q_\alpha\right]$ denotes the α-ES of Y.

Proof. We first notice that for a distribution F with finite second moment und unique α-quantile, it holds that

$$\mathbb{E}\left[Y \middle| Y \leq q_\alpha\right] = -\frac{1}{\alpha} \int_{-\infty}^{q_\alpha} F(x)\mathrm{d}x + q_\alpha \qquad \text{and} \qquad (2.87)$$

$$\mathbb{E}\left[Y^2 \middle| Y \leq q_\alpha\right] = -\frac{2}{\alpha} \int_{-\infty}^{q_\alpha} xF(x)\mathrm{d}x + q_\alpha^2, \qquad (2.88)$$

which can be obtained by using the identity

$$Y \mathbb{1}_{\{Y \leq q_\alpha\}} = \mathbb{1}_{\{Y \leq q_\alpha\}} \left(\int_0^\infty \mathbb{1}_{\{Y > t\}} \, \mathrm{d}t - \int_{-\infty}^0 \mathbb{1}_{\{Y \leq t\}} \, \mathrm{d}t \right) \qquad (2.89)$$

and by taking expectations on both sides. By applying (2.87), we get that

$$\int_{-\infty}^{q_\alpha} \int_{-\infty}^{q_\alpha} F(x)F(y)\mathrm{d}x\mathrm{d}y = \left(\int_{-\infty}^{q_\alpha} F(x)\mathrm{d}x \right)^2 = \left(\alpha q_\alpha - \alpha \mathbb{E}\left[Y \middle| Y \leq q_\alpha\right]\right)^2 = \alpha^2 \left(q_\alpha - \xi_\alpha\right)^2. \qquad (2.90)$$

Furthermore, notice that

$$\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x \wedge y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{q_\alpha}\int_{-\infty}^{y} F(x)\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{q_\alpha}\int_{y}^{q_\alpha} F(y)\mathrm{d}x\mathrm{d}y, \tag{2.91}$$

and by rearranging the order of integration for the first term in (2.91), we get that

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{y} F(x)\,\mathrm{d}x\mathrm{d}y &= \iint\limits_{\{(x,y):\, y\le q_\alpha,\, x\le y\}} F(x)\,\mathrm{d}x\mathrm{d}y = \iint\limits_{\{(x,y):\, x\le q_\alpha,\, y\ge x\}} F(x)\,\mathrm{d}y\mathrm{d}x \\ &= \int_{-\infty}^{q_\alpha}\int_{x}^{q_\alpha} F(x)\,\mathrm{d}y\mathrm{d}x = \int_{-\infty}^{q_\alpha} F(x)(q_\alpha - x)\,\mathrm{d}x.\end{aligned} \tag{2.92}$$

Thus, by first using (2.91) and (2.92) and by plugging in (2.87) and (2.90), we obtain

$$\begin{aligned}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x \wedge y)\mathrm{d}x\mathrm{d}y &= 2\int_{-\infty}^{q_\alpha}\int_{y}^{q_\alpha} F(y)\,\mathrm{d}x\mathrm{d}y \\ &= 2\int_{-\infty}^{q_\alpha} F(y)(q_\alpha - y)\,\mathrm{d}y \\ &= 2q_\alpha\int_{-\infty}^{q_\alpha} F(y)\,\mathrm{d}y - 2\int_{-\infty}^{q_\alpha} yF(y)\,\mathrm{d}y \\ &= 2q_\alpha\big(\alpha q_\alpha - \alpha\xi_\alpha\big) + \alpha\mathbb{E}\big[Y^2\big|Y \le q_\alpha\big] - \alpha q_\alpha^2 \\ &= \alpha\mathbb{E}\big[Y^2\big|Y \le q_\alpha\big] + \alpha q_\alpha^2 - 2\alpha q_\alpha\xi_\alpha.\end{aligned} \tag{2.93}$$

Eventually, using (2.90) and (2.93), straight-forward calculations yield that

$$\frac{1}{\alpha^2}\int_{-\infty}^{q_\alpha}\int_{-\infty}^{q_\alpha} F(x \wedge y) - F(x)F(y)\mathrm{d}x\mathrm{d}y = \frac{1}{\alpha}\operatorname{Var}(Y|Y \le q_\alpha) + \frac{1-\alpha}{\alpha}\big(q_\alpha - \xi_\alpha\big)^2, \tag{2.94}$$

which concludes the proof. □

Appendix 2.D Separability of almost surely continuous functions

Definition 2.D.1 (Separability of a Stochastic Process). A stochastic process $\psi(x,\theta) : \Omega \times \Theta \to \mathcal{Y}$ is called separable in the sense of Doob, if there exists in Ω an everywhere dense countable set I, and in Ω a nullset N such that for any arbitrary open set $G \subset \Theta$ and every closed set $F \subset \mathcal{Y}$, the two sets

$$\{x|\psi(x,\theta) \in F,\ \forall\theta \in G\} \qquad \text{and} \tag{2.95}$$

$$\{x|\psi(x,\theta) \in F,\ \forall\theta \in G \cap I\} \tag{2.96}$$

differ from each other at most by a subset of N.

Proposition 2.D.2 (Gikhman and Skorokhod (2004)). Let Θ and $\mathcal{Y}$ be metric spaces, Θ be a separable space. The sets (2.95) and (2.96) coincide for all $x \in \Omega$ for which the stochastic process $\psi(x, \theta)$ is continuous in θ.

Proof. It is clear that $\{x | \psi(x, \theta) \in F, \; \forall \theta \in G\} \subseteq \{x | \psi(x, \theta) \in F, \; \forall \theta \in G \cap I\}$. We thus only show the reverse.

Let $G \subset \Theta$ be an arbitrary open set and $F \subset \mathcal{Y}$ an arbitrary closed set. Let furthermore $x \in \Omega$ such that $\psi(x, \theta) \in F$ for all $\theta \in G \cap I$. We have to show that $\psi(x, \tilde{\theta}) \in F$ for all $\tilde{\theta} \in G$ but $\tilde{\theta} \notin I$.

Thus, let $\tilde{\theta} \in G \setminus I$. Since I is a dense set in Θ, there exists a sequence $(\theta_n)_{n \in \mathbb{N}} \in \Theta \cap I$, such that $\theta_n \longrightarrow \tilde{\theta}$ and since G is an open set in Θ and $\tilde{\theta} \in G$, we can conclude that for $m \in \mathbb{N}$ large enough, $\theta_n \in G$ for all $n \geq m$. Furthermore, by continuity at θ, it holds that $\psi(x, \theta_n) \to \psi(x, \tilde{\theta})$ and since $\theta_n \in G \cap I$ for all n large enough, $\psi(x, \theta_n) \in F$ by assumption. Eventually, since F is a closed set, $\psi(x, \tilde{\theta}) \in F$ which proves the proposition. □

Corollary 2.D.3 (Separability of continuous functions). Let Θ and $\mathcal{Y}$ be metric spaces, Θ be a separable space, and let the stochastic process $\psi(x, \theta)$ be almost surely continuous. Then, ψ is separable.

Proof. Since $\psi(x, \theta)$ is continuous for all $x \in \Omega \setminus N$ for some $N \subset \Omega$ with $\mathbb{P}(N) = 0$. We get from Proposition 2.D.2 that the sets (2.95) and (2.96) coincide for all $x \in \Omega \setminus N$, i.e. they differ only by a subset of N. □

References

Acerbi, C. and B. Szekely (2014). “Back-testing Expected Shortfall”. *Risk* December, 76–81 (see pp. 53, 103).

Andersen, T. and T. Bollerslev (1998). “Answering the skeptics: Yes, standard volatility models do provide accurate forecasts”. *International Economic Review* 39 (4), 885–905 (see p. 70).

Andrews, D. (1994). “Empirical Process Methods in Econometrics”. In: *Handbook of Econometrics*. Ed. by Engle, R. and McFadden, D. Vol. 4. Elsevier. Chap. 37, 2247–2294 (see p. 73).

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). “Coherent Measures of Risk”. *Mathematical Finance* 9 (3), 203–228 (see pp. 52, 95).

Barendse, S. (2017). “Interquantile Expectation Regression”. Available at https://ssrn.com/abstract=2937665 (see pp. 53, 95, 100).

Basel Committee (2016). *Minimum capital requirements for Market Risk*. Tech. rep. Available at http://www.bis.org/bcbs/publ/d352.pdf. Bank for International Settlements (see pp. 52, 95).

Bayer, S. and T. Dimitriadis (2017b). *esreg: Joint Quantile and Expected Shortfall Regression*. R package version 0.3.1, available at `https://CRAN.R-project.org/package=esreg` (see pp. 54, 64, 71).

— (2017c). “Regression-based Expected Shortfall Backtesting”. Working Paper (see p. 71).

Bollerslev, T. (1986). “Generalized autoregressive conditional heteroskedasticity”. *Journal of Econometrics* 31 (3), 307–327 (see pp. 26, 70, 107).

Brazauskas, V., B. L. Jones, M. L. Puri, and R. Zitikis (2008). “Estimating conditional tail expectation with actuarial applications in view”. *Journal of Statistical Planning and Inference* 138 (11), 3590–3604 (see p. 61).

Chen, S. X. (2008). “Nonparametric Estimation of Expected Shortfall”. *Journal of Financial Econometrics* 6 (1), 87–107 (see p. 61).

Corsi, F. (2009). “A simple approximate long-memory model of realized volatility”. *Journal of Financial Econometrics* 7 (2), 174–196 (see p. 70).

Efron, B. (Jan. 1979). “Bootstrap Methods: Another Look at the Jackknife”. *The Annals of Statistics* 7 (1), 1–26 (see pp. 64, 102).

— (1991). “Regression percentiles using asymmetric squared error loss”. *Statistica Sinica* 1 (1), 93–125 (see pp. 54, 61).

Ehm, W., T. Gneiting, A. Jordan, and F. Krüger (2016). “Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78 (3), 505–562 (see p. 70).

Fissler, T. (2017). “On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space”. PhD thesis. Universität Bern (see p. 63).

Fissler, T. and J. F. Ziegel (2016). “Higher order elicitability and Osband’s principle”. *Annals of Statistics* 44 (4), 1680–1707 (see pp. 9, 12, 53–56, 58, 70, 71, 76, 95, 100).

Fissler, T., J. F. Ziegel, and T. Gneiting (2016). “Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting”. *Risk* January, 58–61 (see pp. 53, 56, 61, 62, 95).

Gikhman, I. and A. Skorokhod (2004). *The Theory of Stochastic Processes I*. Vol. 210. Classics in Mathematics. Springer Berlin Heidelberg (see pp. 73, 91).

Gneiting, T. (2011a). “Making and Evaluating Point Forecasts”. *Journal of the American Statistical Association* 106 (494), 746–762 (see pp. 52, 54–56, 60).

Gourieroux, C. and A. Monfort (1995). *Statistics and Econometric Models: Volume 1, General Concepts, Estimation, Prediction and Algorithms*. Cambridge University Press (see p. 65).

Hall, P. and S. J. Sheather (1988). “On the Distribution of a Studentized Quantile”. *Journal of the Royal Statistical Society. Series B (Methodological)* 50 (3), 381–391 (see p. 64).

Hendricks, W. and R. Koenker (1992). “Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity”. *Journal of the American Statistical Association* 87 (417), 58–68 (see p. 64).

Huber, P. (1967). “The behavior of maximum likelihood estimates under nonstandard conditions”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 221–233 (see pp. 73, 75, 76).

Koenker, R. (1994). “Confidence Intervals for Regression Quantiles. Proceedings of the Fifth Prague Symposium, held from September 4–9, 1993”. In: *Asymptotic Statistics*. Ed. by Mandl, P. and Hušková, M. Heidelberg: Physica-Verlag HD, 349–359 (see p. 64).

— (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press (see pp. 53, 60).

Koenker, R. and J. A. F. Machado (1999). “Goodness of Fit and Related Inference Processes for Quantile Regression”. *Journal of the American Statistical Association* 94 (448), 1296–1310 (see p. 61).

Lambert, N. S., D. M. Pennock, and Y. Shoham (2008). “Eliciting Properties of Probability Distributions”. In: *Proceedings of the 9th ACM Conference on Electronic Commerce*. ACM, 129–138 (see p. 54).

Lourenço, H. R., O. C. Martin, and T. Stützle (2003). “Iterated Local Search”. In: *Handbook of Metaheuristics*. Ed. by Glover, F. and Kochenberger, G. A. Boston, MA: Springer US, 320–353 (see p. 63).

Nadarajah, S., B. Zhang, and S. Chan (2014). “Estimation methods for expected shortfall”. *Quantitative Finance* 14 (2), 271–291 (see pp. 52, 98).

Nelder, J. A. and R. Mead (1965). “A Simplex Method for Function Minimization”. *The Computer Journal* 7 (4), 308–313 (see p. 63).

Newey, W. and D. McFadden (1994). “Large sample estimation and hypothesis testing”. In: *Handbook of Econometrics*. Ed. by Engle, R. and McFadden, D. Vol. 4. Elsevier. Chap. 36, 2111–2245 (see pp. 73, 75).

Nolde, N. and J. F. Ziegel (2017). “Elicitability and backtesting: Perspectives for banking regulation”. arXiv:1608.05498 [q-fin.RM] (see pp. 53, 54, 61, 62, 67, 95–97, 100, 103–108, 110, 111, 114, 116–122).

Taylor, J. W. (2008a). “Estimating Value at Risk and Expected Shortfall Using Expectiles”. *Journal of Financial Econometrics* 6 (2), 231–252 (see pp. 17, 52).

— (2008b). “Using Exponentially Weighted Quantile Regression to Estimate Value at Risk and Expected Shortfall”. *Journal of Financial Econometrics* 6 (3), 382–406 (see p. 52).

— (2017). “Forecasting Value at Risk and Expected Shortfall Using a Semiparametric Approach Based on the Asymmetric Laplace Distribution”. *Forthcoming in Journal of Business & Economic Statistics*. DOI: 10.1080/07350015.2017.1281815 (see pp. 52, 100).

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (see pp. 73, 75, 78).

Weber, S. (2006). “Distribution Invariant Risk Measures, Information, and Dynamic Consistency”. *Mathematical Finance* 16 (2), 419–441 (see pp. 52, 55).

Ziegel, J. F., F. Krüger, A. Jordan, and F. Fasciati (2017). “Murphy Diagrams: Forecast Evaluation of Expected Shortfall”. arXiv:1705.04537 [q-fin.RM] (see pp. 53, 62, 67, 70).

Zwingmann, T. and H. Holzmann (2016). “Asymptotics for the expected shortfall”. arXiv:1611.07222 [math.ST] (see pp. 53, 60).

Chapter 3

Regression Based Expected Shortfall Backtesting

3.1. Introduction

There is a great demand for reliable backtests for the Expected Shortfall (ES) stemming from the transition from the Value-at-Risk (VaR) to the ES as the primary market risk measure in the Basel Accord (Basel Committee, 2016). In general, backtesting is the process of testing whether forecasts of risk measures are correct, which is done by comparing the history of risk forecasts with the corresponding realized returns. Formally, the ES at level $\tau \in (0, 1)$ is defined as the mean of the returns smaller than the respective τ-quantile (the VaR), and τ is usually chosen to be 2.5% as stipulated by the Basel Accord. The ES is introduced into the banking regulation because it overcomes several shortcomings of the VaR, such as not being coherent and its inability to capture tail risks beyond the τ-quantile (Artzner et al., 1999; Basel Committee, 2013). However, the ES is more difficult to backtest than the VaR since the functional ES is not elicitable (Fissler, Ziegel, and Gneiting, 2016; Nolde and Ziegel, 2017).

Nevertheless, there is now a large amount of literature on ES backtesting, but unfortunately, all the proposed approaches are either joint backtests for a vector of risk measures (such as the triple containing the VaR, the ES, and the volatility) or even for the whole tail distribution (Nolde and Ziegel, 2017). As the proposed backtests require further input parameters, such as forecasts for the volatility, the tail distribution beyond some quantile, or even the entire distribution, they are further not applicable for the regulatory authorities because this additional private information is not reported by the financial institutions (Aramonte et al., 2011; Basel Committee, 2016, 2017). In contrast, this paper is the first to develop ES backtesting procedures that solely rely on ES forecasts (and the observable realized returns) as input parameters, which makes these tests applicable for the regulatory authorities.

Triggered by the seminal paper of Fissler and Ziegel (2016) who show that the ES is jointly elicitable with the VaR by proposing a class of strictly consistent joint loss functions for these functionals, there is a growing amount of literature that utilizes these loss functions to establish a regression framework for the functional ES. Dimitriadis and Bayer (2017) propose a linear regression for the pair VaR and ES and show consistency and asymptotic normality for the M- and the Z-estimators based on this class of strictly consistent loss functions under standard regularity conditions. Patton et al. (2017) and Barendse (2017) generalize these asymptotic results for more general dependence conditions of the underlying stochastic process.

In this paper, we utilize this regression technique to propose a novel backtest for ES forecasts which is based on the classical Mincer and Zarnowitz (1969) forecast evaluation approach. The backtest uses the previously described joint regression framework in which we use financial returns as the response variable and the ES forecasts as the explanatory variable including an intercept term. For correct ES forecasts, the intercept and slope parameters should be equal to 0 and 1, respectively. We use a Wald statistic to test for these parameter values, where we apply both, an asymptotic test using the covariance estimator introduced in Dimitriadis and Bayer (2017), and a bootstrap hypothesis test. Such regression-based forecast evaluation approaches are already used for testing mean forecasts (Mincer and Zarnowitz, 1969), quantile forecasts (Gaglianone et al., 2011; Guler et al., 2017), and expectile forecasts (Guler et al., 2017).

We also introduce a second Mincer-Zarnowitz regression-based ES backtest by fixing the slope parameter in the regression to 1, and by only estimating and testing the intercept term. This second backtest allows for both, one-sided and two-sided hypotheses which contrasts with the classical Mincer-Zarnowitz backtest that only allows a two-sided hypothesis because it is generally unclear how underestimated and overestimated ES forecasts influence the slope and intercept parameters. Because the capital requirements that the financial institutions must keep as a reserve depend on the reported risk forecasts, the market participants have an incentive to overestimate[1] the risk forecasts to minimize these expensive capital requirements. In contrast, underestimation of the risk forecasts results in too conservative risk forecasts and larger capital reserves, which does not have to be punished by the regulatory authorities. Thus, the regulatory authorities only have to prevent and penalize the overestimation of risk forecasts, which demonstrates the necessity of one-sided testing procedures. For example, the currently implemented traffic light system (Basel Committee, 1996) is in fact a one-sided VaR backtest. Both backtesting procedures we introduce in this paper have the desired property to only require ES forecasts as input parameters and consequently can be considered as the first procedures that solely backtest the ES.

We introduce several simulation setups to evaluate the empirical properties of our novel ES backtests and compare them to the existing backtests of McNeil and Frey (2000) and Nolde and Ziegel (2017). In the first setup, we implement the classical size and power analysis for backtesting risk measures, where we simulate data stemming from a realistic data generating process and evaluate the empirical rejection frequencies of the backtests for forecasts stemming from the true and from some misspecified forecasting models. In the second setup, we introduce a novel technique for evaluating the power of backtests for financial risk measures, where we continuously misspecify certain model parameters of the data generating process to obtain a continuum of alternative models with a gradually increasing degree of misspecification. Misspecifying the different model parameters separately allows

[1] Throughout the paper, we use the sign convention that losses are denoted by negative numbers and *overestimation* of risk measures is meant in the mathematical sense, i.e. as reporting too large real numbers.

us to misspecify certain model characteristics (such as the reaction to shocks) in isolation, which permits a closer examination of the proposed backtesting procedures.

From these simulations, we find that our proposed backtests are reasonable sized, especially when the tests are applied using the bootstrap. Moreover, they are more powerful than the existing backtests in most considered simulation designs. This is the case for two-sided hypothesis and for the one-sided version, which is of particular relevance for the financial authorities. Notably, our backtests detect the misspecified forecasts in all considered designs. In comparison to that, the backtests of McNeil and Frey (2000) and Nolde and Ziegel (2017) fail several times to discriminate between the true and the misspecified forecasts, for instance when the forecaster reports risk predictions for a wrong probability level.

The rest of this paper is organized as follows. Section 3.2 introduces the theory of our new backtests, and Section 3.3 reviews the existing ES backtesting techniques. Section 3.4 contains two simulation studies, and Section 3.5 applies the backtests to the risk forecasts of the S&P500 index. Section 3.6 concludes this paper and provides an outlook on potential future research.

3.2. Theory

3.2.1. Setup and Notation

Let us consider a stochastic process

$$Z = \big\{Z_t : \Omega \longrightarrow \mathbb{R}^{k+1}, k \in \mathbb{N}, t = 1, \dots, T\big\}, \tag{3.1}$$

defined on some complete probability space $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$, with the filtration $\mathcal{F} = \big\{\mathcal{F}_t,\ t = 1, \dots T\big\}$ and $\mathcal{F}_t = \sigma\{Z_s,\ s \leq t\}$. We partition the stochastic process $Z_t = (Y_t, X_t)$, where Y_t is an absolutely continuous random variable of interest and X_t is a vector of explanatory variables. We denote the conditional cumulative distribution function of Y_t given the past information $\mathcal{F}_{t-1}$ by $F_t(y) = \mathbb{P}(Y_t \leq y \mid \mathcal{F}_{t-1})$ and the corresponding probability density function by f_t. The mean and variance of the distribution F_t are denoted by $\mathbb{E}_t[\cdot]$ and $\text{Var}_t(\cdot)$, whenever they exist.

In the context of this paper, Y_t can be regarded as the daily log returns of a financial asset (for instance, a stock or portfolio), i.e. $Y_t = \log P_t - \log P_{t-1}$, where P_t denotes the price of the asset at day $t = 1, \dots, T$. This means that throughout this paper, we use the sign convention that positive returns denote profits, and negative returns denote losses. The vector X_t contains further variables that are used to produce forecasts for certain functionals (usually risk measures) of the random variable Y_t.

We are interested in testing whether forecasts of a certain (d-dimensional, $d \in \mathbb{N}$) functional $\rho : \big(\Omega, \mathcal{F}, \mathbb{P}\big) \to \mathbb{R}^d$ of the conditional distribution F_t are correctly specified. For that, we define the most frequently used functionals for financial risk management in the following. The conditional quantile of Y_t given the information set $\mathcal{F}_{t-1}$ at level $\tau \in (0, 1)$ is defined as

$$Q_\tau\big(Y_t \mid \mathcal{F}_{t-1}\big) = F_t^{-1}(\tau) = \inf\big\{y \in \mathbb{R} : F_t(y) \geq \tau\big\}. \tag{3.2}$$

Furthermore, we define the functional ES at level τ of Y_t given $\mathcal{F}_{t-1}$ as

$$\mathrm{ES}_\tau\big(Y_t \mid \mathcal{F}_{t-1}\big) = \frac{1}{\tau}\int_0^\tau F_t^{-1}(s)\,\mathrm{d}s. \tag{3.3}$$

If the distribution function F_t is continuous at its τ-quantile, this definition can be simplified to the truncated tail mean of Y_t,

$$\mathrm{ES}_\tau\big(Y_t \mid \mathcal{F}_{t-1}\big) = \mathbb{E}_t\left[Y_t \mid Y_t \leq Q_\tau\big(Y_t \mid \mathcal{F}_{t-1}\big)\right]. \tag{3.4}$$

We denote an $\mathcal{F}_{t-1}$-measurable one-step-ahead forecast for day t for the risk measure (the functional) ρ of the distribution F_t, stemming from an arbitrary external forecaster or model[2] by $\hat{\rho}_t = \hat{\rho}_t(\mathcal{F}_{t-1})$. Following this notation, we denote forecasts for the τ-quantile (in this context also known as the VaR) by $\hat{v}_t$ and for the τ-ES by $\hat{e}_t$ for some fixed level $\tau \in (0, 1)$. For simplicity, we drop the dependence on τ in the notation as it is a fixed quantity.

Testing correctness for a given series of forecasts $\big(\hat{\rho}_t,\ t = 1, \dots, T\big)$ for the functional ρ relative to the realized (and observed) return series $\big(y_t,\ t = 1, \dots, T\big)$ is called *backtesting*, which we formally define in the following.

Definition 3.2.1. A *backtest* for the series of forecasts $\big(\hat{\rho}_t,\ t = 1, \dots, T\big)$ for the d-dimensional risk measure (functional) ρ relative to the realized return series $\big(y_t,\ t = 1, \dots, T\big)$ is a function

$$f : \mathbb{R}^T \times \mathbb{R}^{T \times d} \to (0, 1), \tag{3.5}$$

which maps the return and forecast series onto the respective p-value of the test.

Almost all of the VaR backtests in the literature satisfy Definition 3.2.1 (see e.g. Christoffersen, 1998; Engle and Manganelli, 2004; Kupiec, 1995). However, as we discuss in Section 3.3, this definition becomes relevant when considering backtesting the risk measure

[2] For recent overviews on VaR and ES forecasting approaches, see Komunjer (2013) and Nadarajah et al. (2014).

ES, because many of the recently proposed ES backtests are based on the knowledge of forecasts for other quantities such as the volatility, the tail distribution or even the entire distribution of the returns.

3.2.2. A Mincer-Zarnowitz regression based ES backtest

We now propose a new backtest for the risk measure ES, that tests whether the ES forecasts $\hat{e}_t$ (stemming from some risk model) coincide with the conditional ES of the returns by regressing the returns Y_t on the forecasts $\hat{e}_t$ and an intercept term, similar to the Mincer-Zarnowitz test for mean forecasts (Mincer and Zarnowitz, 1969). For that, we use a regression equation designed specifically for the functional ES,

$$Y_t = \alpha + \beta \hat{e}_t + u_t^e, \tag{3.6}$$

where $\mathrm{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$. Given the structure in (3.6) and since $\hat{e}_t$ is generated by using the information set $\mathcal{F}_{t-1}$, this assumption on the error term is equivalent to

$$\mathrm{ES}_\tau \left(Y_t \mid \mathcal{F}_{t-1}\right) = \alpha + \beta \hat{e}_t. \tag{3.7}$$

We then test the hypothesis

$$\mathbb{H}_0 : \big(\alpha, \beta\big) = (0, 1) \qquad \text{against} \qquad \mathbb{H}_1 : \big(\alpha, \beta\big) \neq (0, 1), \tag{3.8}$$

and under the $\mathbb{H}_0$ the ES forecasts are correctly specified since $\hat{e}_t = \mathrm{ES}_\tau \left(Y_t \mid \mathcal{F}_{t-1}\right)$.[3]

As outlined in Dimitriadis and Bayer (2017), estimating the parameters (α, β) in (3.6) by M- or Z- (GMM-) estimation stand-alone using a semiparametric method without specifying the full conditional distribution of the error term u_t^e is not possible since the functional ES is not elicitable. However, these parameters can be estimated through a joint regression technique for the quantile and ES that we briefly review in the following. For a response variable Y_t and a k-dimensional vector of covariates X_t following the definition of the

[3] Given that the ES forecasts are correctly specified, i.e. $\hat{e}_t = \mathrm{ES}_\tau \left(Y_t \mid \mathcal{F}_{t-1}\right)$, the condition (3.7) is equivalent to $\alpha = (1-\beta)\hat{e}_t$. This results in the remark of Holden and Peel (1990), who claim that (3.7) is only a sufficient, but not a necessary condition for correctly specified forecasts as $\alpha = (1-\beta)\hat{e}_t$ is the required necessary condition. However, the more general condition requires that the forecasts $\hat{e}_t$ are constant for all $t = 1, \ldots, T$, which is unrealistic given the dynamic nature of financial time series.

stochastic process in (3.1), they model the quantile and the ES at the joint level $\tau \in (0,1)$ through the linear regression equations

$$Y_t = X_t'\theta^q + u_t^q \qquad \text{and} \tag{3.9}$$

$$Y_t = X_t'\theta^e + u_t^e, \tag{3.10}$$

where $Q_\tau(u_t^q \mid \mathcal{F}_{t-1}) = 0$ and $\mathrm{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$. Here, $\theta = \left(\theta^q, \theta^e\right)$ denotes the $2k$-dimensional vector of regression parameters of the joint model. The M-estimator of the regression parameters θ is obtained by

$$\widehat{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} \rho(Y_t, X_t, \theta), \tag{3.11}$$

where the loss function is given by

$$\rho(Y_t, X_t, \theta) = \frac{1}{-X_t'\theta^e}\left(X_t'\theta^e - X_t'\theta^q + \frac{(X_t'\theta^q - Y_t)\mathbb{1}_{\{Y_t \le X_t'\theta^q\}}}{\alpha}\right) + \log(-X_t'\theta^e). \tag{3.12}$$

As shown by Dimitriadis and Bayer (2017), consistent and asymptotically normal M-estimation of these regression parameters can be obtained by employing loss functions from a whole class of functions, originally introduced by Fissler and Ziegel (2016) in the context of forecast evaluation. However, consensus seems to emerge on the 0-homogeneous loss function presented in (3.12), see Barendse (2017), Dimitriadis and Bayer (2017), Patton et al. (2017), and Taylor (2017) and Nolde and Ziegel (2017).

Consistency and the asymptotic normality of the M-estimator of θ is shown by Patton et al. (2017) for a stationary and α-mixing stochastic process $Z_t = (Y_t, X_t)$. Under the further technical conditions in Assumption 1 and 2 in Patton et al. (2017), it holds that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \Lambda^{-1} C \Lambda^{-1}\right), \tag{3.13}$$

where θ_0 denotes the unknown true parameter value where

$$\Lambda = \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix} \qquad \text{and} \qquad C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{3.14}$$

with

$$\Lambda_{11} = -\frac{1}{\alpha T}\mathbb{E}\left[\sum_{t=1}^{T}(X_t X_t') f_t(X_t'\theta_0^q)/(X_t'\theta_0^e)\right], \tag{3.15}$$

$$\Lambda_{22} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(X_t X_t')/(X_t'\theta_0^e)^2\right], \tag{3.16}$$

$$C_{11} = \frac{1-\alpha}{\alpha}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(X_t X_t')/(X_t'\theta_0^e)^2\right], \tag{3.17}$$

$$C_{12} = C_{21} = -\frac{1-\alpha}{\alpha}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(X_t X_t')\big(X_t'\theta_0^q - X_t'\theta_0^e\big)/(X_t'\theta_0^e)^3\right], \tag{3.18}$$

$$C_{22} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(X_t X_t')/(X_t'\theta_0^e)^4\left(\frac{1}{\alpha}\operatorname{Var}_t\big(Y_t - X_t'\theta_0^q \mid Y_t \le X_t'\theta_0^q\big) + \frac{1-\alpha}{\alpha}\big(X_t'\theta_0^q - X_t'\theta_0^e\big)^2\right)\right]. \tag{3.19}$$

We use this joint regression framework and the asymptotic estimation theory for the semiparametric estimation of (3.6) by estimating the joint system,

$$Y_t = \gamma + \delta\hat{e}_t + u_t^q, \tag{3.20}$$

$$Y_t = \alpha + \beta\hat{e}_t + u_t^e. \tag{3.21}$$

Because we only want to test the correct specification in the regression equation for the ES given in (3.21), we only test for the associated parameters (α, β) using a Wald statistic,

$$T_{\text{ESR}} = \left((\hat{\alpha}, \hat{\beta})' - (0,1)'\right)' \widehat{\Sigma}_{\text{ES}}^{-1} \left((\hat{\alpha}, \hat{\beta})' - (0,1)'\right)', \tag{3.22}$$

where $\widehat{\Sigma}_{\text{ES}}$ is an estimator for the submatrix $\Sigma_{\text{ES}} = \Lambda_{22}^{-1} C_{22} \Lambda_{22}^{-1}$. By the asymptotic normality of the parameter estimates in (3.13) and given that $\hat{\Sigma}_{\text{ES}} \xrightarrow{\mathbb{P}} \Sigma_{\text{ES}}$, the test statistic asymptotically follows a χ^2 distribution with two degrees of freedom,

$$T_{\text{ESR}} \xrightarrow{d} \chi_2^2. \tag{3.23}$$

For the estimation of the asymptotic covariance matrix of the parameter estimates, $\hat{\Sigma}_{\text{ES}}$, we employ the methods discussed in Dimitriadis and Bayer (2017). The main difficulty is the estimation of the nuisance quantity $\operatorname{Var}_t\big(Y_t - X_t'\theta_0^q \mid Y_t \le X_t'\theta_0^q\big)$, for which we employ the *scl-sp* method introduced by them.

We also use a bootstrap hypothesis test[4] for testing whether $(\alpha, \beta) = (0, 1)$. For that, we draw $B = 1000$ bootstrap samples from the data with replacement, i.e. we apply the iid bootstrap of Efron (1979), since neither the M-estimator of the parameters nor the covariance estimator depend on the temporal ordering of the data. In each bootstrap sample, we estimate the model parameters and the covariance matrix to compute a total of B bootstrap Wald statistics as in (3.22), where the bootstrap estimates are centered around the estimate for the original sample. Finally, the bootstrap p-value is the share of the B bootstrap test statistics that are larger or equal than the test statistic for the original sample.

3.2.3. A One-sided Mincer-Zarnowitz Intercept Test

The backtesting procedure introduced in the previous section only allows for testing two-sided hypotheses as specified in (3.8) because it is generally unclear how too small or too large risk forecasts influence the parameters α and β. Because the capital requirements the financial institutions have to keep as a reserve depend on the reported risk forecasts, the market participants have an incentive to overestimate the risk forecasts in order to keep as little capital requirements as possible. In contrast, underestimation of the risk measures results in too conservative risk forecasts and higher capital requirements that does not have to be punished by the regulatory authorities. Thus, the regulatory authorities only have to prevent and consequently penalize the overestimation of risk measures, which can be done by using one-sided testing procedures. For example, the traffic light system (Basel Committee, 1996), currently implemented in the Basel Accords, is in fact a one-sided backtest for the hit ratios of VaR forecasts.

Consequently, in the following we introduce a Mincer-Zarnowitz backtesting procedure for the ES that allows for both, a one-sided and a two-sided hypothesis. This backtest is based on regressing the forecast error, $Y_t - \hat{e}_t$, on an intercept term,

$$Y_t - \hat{e}_t = \alpha + u_t^e, \tag{3.24}$$

where $\text{ES}_\tau(u_t^e \mid \mathcal{F}_{t-1}) = 0$ and testing whether the parameter α is zero. Note that this is equivalent to setting the slope parameter of the bivariate ESR test given in (3.6) to one

[4] This approach provides an asymptotic refinement, i.e. the error in the rejection probability decreases faster compared to the asymptotic distribution or to using bootstrapped covariance matrices for the test, see e.g. MacKinnon (2009). In the construction of confidence intervals this is also known as the percentile-t method.

and only estimating and testing the intercept term. By using this restriction we can define one-sided and the two-sided hypotheses,

$$\begin{aligned} &\mathbb{H}_0^{2s} : \alpha = 0 \quad &&\text{against} \quad &&\mathbb{H}_1^{2s} : \alpha \neq 0, \quad \text{and} \\ &\mathbb{H}_0^{1s} : \alpha \geq 0 \quad &&\text{against} \quad &&\mathbb{H}_1^{1s} : \alpha < 0, \end{aligned} \tag{3.25}$$

which we test by using a t-test based on the asymptotic covariance and based on the bootstrap procedure described above.

3.3. Existing Backtests

Over the past two decades and especially driven by the recent transition from VaR to ES in the Basel regulatory framework, a large literature on backtesting the ES has emerged. These backtests are usually introduced with financial regulators in mind who need to verify the risk forecasts they receive from the financial institutions. To be applicable for the regulatory authorities, a proper backtest for the risk measure ES thus follows Definition 3.2.1 and only requires the observed return series and the ES forecasts as input variables. However, many of the proposed backtests for the ES fail to have this property. In particular, several tests require the whole return distribution (Acerbi and Szekely, 2014; Berkowitz, 2001; Graham and Pál, 2014; Kerkhof and Melenberg, 2004; Wong, 2008), the cumulative violation process $\int_0^\tau \mathbb{1}_{\{Y_t \leq \hat{v}_t(p)\}} \, \mathrm{d}p$ (Costanzino and Curran, 2015; Du and Escanciano, 2017; Emmer et al., 2015; Kratz et al., 2017), the volatility (McNeil and Frey, 2000; Nolde and Ziegel, 2017; Righi and Ceretta, 2013, 2015), or the VaR (McNeil and Frey, 2000; Nolde and Ziegel, 2017) in addition to the ES forecasts. However, none of this private information (except the VaR) will be reported by the financial institutions and therefore, most of these tests can not be used by the regulators (Aramonte et al., 2011; Basel Committee, 2017).

Furthermore, when more information than solely the ES forecasts is used for backtesting, then a rejection of the null hypothesis does not necessarily imply that the ES forecasts are wrong. More precisely, a rejection of the null implies that *some* component of the input parameters is wrong (cf. Nolde and Ziegel, 2017). A related concern is raised by Aramonte et al. (2011), who note that financial institutions could be tempted to submit forecasts of this additional information chosen such that the tests have particularly low power, so that correctness of their internal model is not doubted.

Strictly following Definition 3.2.1, we would furthermore have to distinguish between backtests for the ES and joint backtests for the pair VaR and ES. However, as the ES is strongly intertwined with the VaR as the definition of the ES already depends on the VaR, sensible forecasts for the ES are based on correctly specified VaR forecasts. Consequently, it is reasonable to backtest both quantities jointly and thus, we do not distinguish between ES backtests and joint VaR and ES backtests. In the following, we describe the exceedance residual test of McNeil and Frey (2000) and the conditional calibration tests of Nolde and Ziegel (2017) in more detail, since both have versions that only require VaR forecasts in addition to the ES.

3.3.1. Testing the Exceedance Residuals

One of the first and still most frequently used tests for the ES is the exceedance residual (ER) backtest of McNeil and Frey (2000). This approach is based on the ES residuals that exceed the VaR, $er_t = \big(Y_t - \hat{e}_t\big)\mathbb{1}_{\{Y_t \le \hat{v}_t\}}$, that form a martingale difference sequence given that $\hat{v}_t$ and $\hat{e}_t$ are the true $\mathcal{F}_{t-1}$-measurable quantile and ES respectively. McNeil and Frey (2000) furthermore consider a second version that uses ER standardized by the volatility, i.e. $er_t / \hat{\sigma}_t$, instead of the raw values.

For the actual backtest, we need to test whether the expected value of the (raw or standardized) ER, μ, is zero using $\hat{\mu} = 1/(\sum_{t=1}^T \mathbb{1}_{\{Y_t \le \hat{v}_t\}}) \sum_{t=1}^T er_t$ in conjunction with a bootstrap hypothesis test (see Efron and Tibshirani, 1993, p. 224). In the original paper, McNeil and Frey (2000) propose to test μ against the one-sided hypothesis that μ is negative, i.e. that the ES is overestimated. However, in this paper we discuss both, tests based on the one-sided and two-sided hypothesis, so that we in addition to the original proposal also include a two-sided test,

$$\begin{aligned} &\mathbb{H}_0^{2s} : \mu = 0 \qquad \text{against} \qquad \mathbb{H}_1^{2s} : \mu \neq 0, \quad \text{and} \\ &\mathbb{H}_0^{1s} : \mu \geq 0 \qquad \text{against} \qquad \mathbb{H}_1^{1s} : \mu < 0. \end{aligned} \tag{3.26}$$

By Definition 3.2.1, the test using the standardized ER is in fact a joint backtest for the triple VaR, ES and volatility, whereas the test using the raw ER is a joint backtest for the pair VaR and ES. In light of the discussion above, the test using the raw ER is therefore preferred. Nevertheless, in the simulation studies and the empirical application we apply both approaches and find that they perform similar.

Although the intercept ESR test introduced in Section 3.2.3 and the ER backtest appear to be similar, there is a subtle difference between the two test statistics. For the intercept ESR test, we compute the empirical ES of $Y_t - \hat{e}_t$, i.e. the average of $Y_t - \hat{e}_t$ given that $Y_t - \hat{e}_t$ is smaller than its empirical τ-quantile. In contrast, the ER backtest computes the average of $Y_t - \hat{e}_t$, given that Y_t is smaller than the respective forecast for its τ-quantile, $\hat{v}_t$. This difference seems marginal, but it has severe consequences for the theoretical and empirical properties of the tests. In particular, the ER backtest cannot distinguish between correct forecasts of the VaR and ES at level τ and forecasts for a misspecified probability level $\tilde{\tau} \neq \tau$, as the given level τ does not influence the ER test statistic at all. In contrast, by computing the empirical τ-quantile of $Y_t - \hat{e}_t$, the intercept ESR test does not suffer from this shortcoming.

3.3.2. Conditional Calibration Backtests

Nolde and Ziegel (2017) introduce the concept of conditional calibration (CC) based on strict identification functions (also known as moment conditions or estimating equations) of the respective functional and show that many classical backtests for risk measures can be unified using this concept. For the pair VaR and ES at level $\tau \in (0, 1)$, they choose the strict identification function

$$V(Y,\, v,\, e) = \begin{pmatrix} \tau - \mathbb{1}_{\{Y \leq v\}} \\ e - v + \mathbb{1}_{\{Y \leq v\}}(v - Y)/\tau \end{pmatrix}, \tag{3.27}$$

whose expectation is zero if and only if v and e equal the true VaR and ES of Y respectively. The CC test is based on the hypotheses

$$\begin{aligned} &\mathbb{H}_0^{2s} : \mathbb{E}\big[V(Y_t,\, \hat{v}_t,\, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] = 0 && \text{against} && \mathbb{E}\big[V(Y_t,\, \hat{v}_t,\, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] \neq 0, \quad \text{and} \\ &\mathbb{H}_0^{1s} : \mathbb{E}\big[V(Y_t,\, \hat{v}_t,\, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] \geq 0 && \text{against} && \mathbb{E}\big[V(Y_t,\, \hat{v}_t,\, \hat{e}_t) \mid \mathcal{F}_{t-1}\big] < 0, \end{aligned} \tag{3.28}$$

component-wise and almost surely for all $t = 1, \ldots, T$. This is equivalent to testing $\mathbb{E}\big[h_t' V(Y_t, \hat{v}_t, \hat{e}_t)\big] = 0$ for all $\mathcal{F}_{t-1}$ measurable $\mathbb{R}^2$-valued functions h_t. As this is infeasible, Nolde and Ziegel (2017) propose to use an $\mathcal{F}_{t-1}$-measurable sequence of $q \times 2$-matrices of test functions $\boldsymbol{h}_t$ for some $q \in \mathbb{N}$ and to use the Wald-type test statistic

$$T_{\text{CC}} = T \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{h}_t V\left(Y_t, \hat{v}_t, \hat{e}_t\right) \right)' \widehat{\Omega}^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{h}_t V\left(Y_t, \hat{v}_t, \hat{e}_t\right) \right), \tag{3.29}$$

where $\widehat{\Omega} = \frac{1}{T} \sum_{t=1}^{T} \left(\boldsymbol{h}_t V\left(Y_t, \hat{v}_t, \hat{e}_t\right)\right) \left(\boldsymbol{h}_t V\left(Y_t, \hat{v}_t, \hat{e}_t\right)\right)'$ is a consistent estimator of the covariance of the q-dimensional vector $\boldsymbol{h}_t V\left(Y_t, \hat{v}_t, \hat{e}_t\right)$. Under $\mathbb{H}_0$, the test statistic asymptotically follows a χ^2_q distribution with q degrees of freedom.

Nolde and Ziegel (2017) propose two versions of this test, where the first uses no information beside the risk forecasts (termed *simple CC test*), and where the second additionally requires volatility forecasts (termed *general CC* test). For the simple CC test, the test function is the identify matrix, $\boldsymbol{h}_t = I_2$, for both, the one- and two-sided hypotheses. For the general CC test, they propose to choose

$$\boldsymbol{h}_t = \hat{\sigma}_t\big((\hat{e}_t - \hat{v}_t)/\tau,\ 1\big) \quad \text{and} \quad \boldsymbol{h}_t = \begin{pmatrix} 1 & |\hat{v}_t| & 0 & 0 \\ 0 & 0 & 1 & \hat{\sigma}_t^{-1} \end{pmatrix}', \tag{3.30}$$

for the two-sided and for the one-sided test, respectively, where $\hat{\sigma}_t$ is a forecast of the volatility. As with the standardized ER test, the general CC test is strictly speaking a backtest for the triple VaR, ES, and volatility, but we nevertheless include both versions in our empirical comparisons.

We provide implementations of the two ESR backtests proposed in this paper, the ER test of McNeil and Frey (2000) and both CC backtests of Nolde and Ziegel (2017) in our R package `esback` (Bayer and Dimitriadis, 2017a).

3.4. Monte-Carlo simulations

In this section, we evaluate the empirical performance of our proposed ES backtests and compare them to the tests of McNeil and Frey (2000) and Nolde and Ziegel (2017). For that, we first assess the empirical size of the tests, defined as the rejection frequency of the test under the null hypothesis, that should equal the nominal significance level. Then, we analyze the rejection frequency of the null hypothesis for misspecified forecasts, i.e. the empirical power of the tests, that should be as close to one as possible.

This comparison is conducted using two different approaches. The first, presented in Section 3.4.1, follows the typical strategy in the related literature of first assessing the size of the backtests with some realistic data generating process (DGP), followed by an evaluation of the power by backtesting forecasts of misspecified models, in our case the Historical Simulation and the RiskMetrics model.

In the second setup, presented in Section 3.4.2, we misspecify the parameters of the true model and thereby obtain alternative models with a continuously increasing degree of misspecification. This approach of evaluating backtests has two main advantages. First, we can obtain power curves which can be used to draw conclusions how an increasing model misspecification influences the test decision. Second, misspecifying the different model parameters separately allows us to misspecify certain model characteristics while leaving the remaining model unchanged. Thus, we can evaluate which model misspecification the backtests are able to identify, which allows for a closer examination of the backtesting procedures.

3.4.1. Traditional size and power comparisons

For the first simulation study, we simulate asset returns from the same model as used by Nolde and Ziegel (2017), which is an AR(1) - GARCH(1,1) process (Bollerslev, 1986) with skewed Student-t distributed innovations. This model is realistic and highly flexible because it replicates the stylized facts typically found in financial return series such as non-normality, volatility clustering, asymmetries, and fat tails. The model is given by,

$$\begin{aligned} r_t &= \mu_t + \varepsilon_t, \\ \varepsilon_t &= \sigma_t z_t, \\ \mu_t &= -0.05 + 0.3 r_{t-1}, \\ \sigma_t^2 &= 0.01 + 0.1\varepsilon_{t-1}^2 + 0.85\sigma_{t-1}^2, \\ z_t &\overset{\text{iid}}{\sim} \text{skew-}t(5,\ 1.5), \end{aligned} \tag{3.31}$$

where z_t are innovations stemming from the standardized skewed Student-t distribution of Fernandez and Steel (1998) with five degrees of freedom, and a skewness parameter of 1.5. Since (3.31) is of the location-scale form, the conditional VaR and ES forecasts at level τ are given by

$$\hat{v}_t = \mu_t + \sigma_t q_z(\tau) \quad \text{and} \quad \hat{e}_t = \mu_t + \sigma_t \xi_z(\tau), \tag{3.32}$$

where $q_z(\tau)$ and $\xi_z(\tau)$ are the τ-quantile, respectively the τ-ES of the innovations z_t (see Lambert and Laurent (2002) and Trottier and Ardia (2016) for the technical details). For the following size and power analysis of the backtests, we simulate the process (3.31) 10,000 times with sample sizes of 250, 500, 1000, 2500, and 5000 observations and 250 additional pre-sample values required for the power analysis. As stipulated by the Basel Accords, we forecast the two risk measures for the probability level $\tau = 2.5\%$.

We evaluate the empirical sizes of the tests by backtesting the VaR and ES forecasts of the true model and the respective simulated return series by computing the share of simulation replications where we reject the null hypothesis at the significance levels 1%, 5%, and 10%. In this part of the study, we focus on two-sided hypotheses and defer the one-sided case to Sections 3.4.2 and 3.4.3.

Table 3.1 presents the rejection rates for forecasts of the true model for all backtests, sample sizes, and nominal test sizes. We find that in large samples, all backtests display rejection rates close to the respective nominal sizes. However, in small samples all backtests are oversized and they differ with respect to their speed of convergence. Looking at the individual tests in greater detail, we find that especially the tests relying on asymptotic quantities (i.e. the ESR and CC tests) are substantially oversized in small samples and converge to the nominal size comparably slow. However, by using the bootstrap for the intercept and bivariate ESR tests (indicated by (b) in the table), the empirical sizes are much closer to the nominal sizes in small samples than for the asymptotic version. Comparing the intercept and the bivariate ESR test, we find that the former has better size properties in small samples, presumably because less parameters need be estimated and the covariance is simpler. Furthermore, also the two ER tests (which also rely on bootstrapping) exhibit good empirical sizes and there are hardly any differences between the raw and the standardized version.

Table 3.1: Empirical sizes of the backtests

Nominal Size	Sample Size	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
1%	250	0.03	0.14	0.02	0.11	0.01	0.21	0.04	0.04
	500	0.03	0.09	0.02	0.07	0.03	0.12	0.01	0.01
	1000	0.02	0.06	0.02	0.04	0.04	0.08	0.01	0.01
	2500	0.01	0.02	0.01	0.02	0.03	0.04	0.01	0.01
	5000	0.01	0.02	0.01	0.01	0.02	0.03	0.01	0.01
5%	250	0.10	0.23	0.07	0.17	0.09	0.27	0.07	0.08
	500	0.08	0.16	0.07	0.12	0.11	0.19	0.04	0.05
	1000	0.07	0.11	0.06	0.09	0.10	0.14	0.05	0.06
	2500	0.06	0.07	0.05	0.06	0.08	0.09	0.06	0.06
	5000	0.06	0.06	0.04	0.05	0.06	0.07	0.05	0.05
10%	250	0.16	0.29	0.13	0.22	0.18	0.31	0.12	0.13
	500	0.14	0.22	0.12	0.17	0.17	0.23	0.10	0.11
	1000	0.13	0.16	0.11	0.14	0.15	0.19	0.11	0.12
	2500	0.12	0.12	0.10	0.11	0.13	0.15	0.11	0.11
	5000	0.11	0.10	0.09	0.10	0.11	0.12	0.10	0.11

Notes: The table reports the empirical sizes of the backtests for an AR(1)-GARCH(1,1)-skewed-t process. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

For a comparison of the power of the backtests, we evaluate their ability to reject the null hypothesis for risk models producing incorrect ES forecasts. We utilize two models that are popular in practice, the Historical Simulation approach and the RiskMetrics model (Zangari, 1996).

The Historical Simulation approach forecasts the VaR and ES by,

$$\hat{v}_t = \widehat{Q}_\tau \left(r_{t-1}, r_{t-2}, \cdots, r_{t-w}\right) \quad \text{and} \quad \hat{e}_t = \frac{1}{\sum_{i=1}^{w} \mathbb{1}_{\{r_{t-i} \le \hat{v}_t\}}} \sum_{i=1}^{w} r_{t-i} \cdot \mathbb{1}_{\{r_{t-i} \le \hat{v}_t\}}, \tag{3.33}$$

where $\widehat{Q}_\tau$ is the empirical τ-quantile and w is the length of a rolling window, that we set to 250, i.e. one year of data. Since the standardized ER and the general CC backtest both require forecasts of the volatility, we estimate this quantity with the sample standard deviation of the returns in the same rolling window.

The RiskMetrics model can be expressed as a location-scale model with zero mean, conditional variance $\sigma_t^2 = 0.06 r_{t-1}^2 + 0.94 \sigma_{t-1}^2$ (an integrated GARCH model), and normally distributed innovations, so that forecasts of the VaR and ES are given by,

$$\hat{v}_t = \hat{\sigma}_t \Phi^{-1}(\tau) \quad \text{and} \quad \hat{e}_t = -\hat{\sigma}_t \phi(\Phi^{-1}(\tau))/\tau, \tag{3.34}$$

where $\Phi^{-1}(\tau)$ and $-\phi(\Phi^{-1}(\tau))/\tau$ are the τ-quantile and τ-ES for the standard normal distribution.

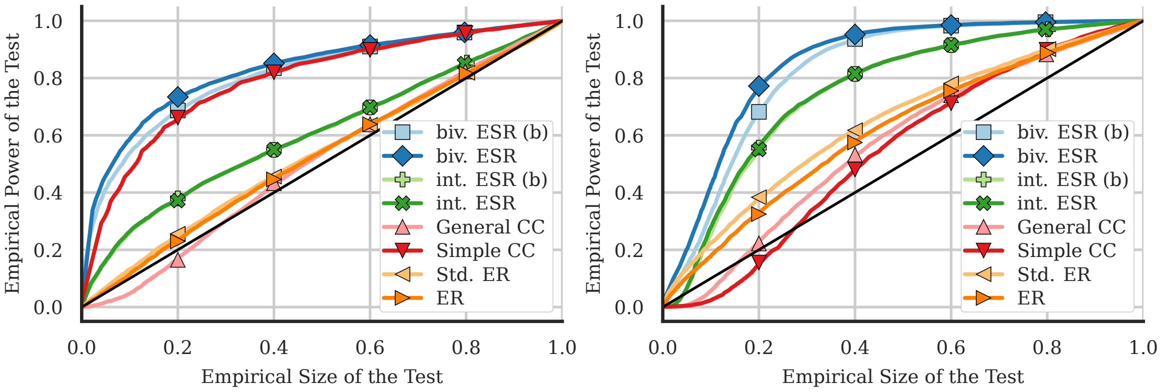
For a meaningful and fair comparison of the power of the backtests to reject the null that the forecasts of these two models are correct, we compare the *size-adjusted power*[5] of the backtests (Lloyd, 2005). For this, the original critical values of the tests are either increased or decreased such that the rejection frequency for the true model equals the nominal test size. Then, we obtain the size-adjusted power by the rejection frequencies for the alternative models using these new critical values.

Figure 3.1 contains the size-adjusted power of the backtests for all empirical sizes in the unit interval against RiskMetrics and the Historical Simulation for the sample size 1000.[6] The black line depicts the case of equal empirical size and power, which can be seen as a lower bound for any reasonable test: whenever the power is below this line, randomly guessing the test decision is more accurate than performing the test. In this figure, we see that the bivariate ESR backtest clearly dominates the others against both alternatives at all empirical sizes, including the most relevant region of test sizes between 1% and 10%. The bivariate ESR test using asymptotic quantities is slightly more powerful than the bootstrap version (indicated by (b)), but the loss in power is negligible compared to the improvements in the sizes we find in Table 3.1.

A common drawback of the other backtests is that they either perform well against RiskMetrics or against the Historical Simulation, but not against both. For instance, the simple CC test performs almost as good as the bivariate ESR test against RiskMetrics, but

[5] A comparison of the *raw power* (i.e. the rejection rate for the null hypothesis that these forecasts are correct, analog to the empirical sizes) could be misleading due to the differences in the empirical sizes of the backtests. In particular, an oversized test would exhibit unrealistically large rejection rates. For completeness, Tables 3.B.5 and 3.B.6 report the raw powers of the tests.

[6] These type of plots are known as the receiver operating characteristic (ROC) curve and origin from the psychometrics literature (Lloyd, 2005). However, they can be used for general binary classification tasks such as hypothesis testing.

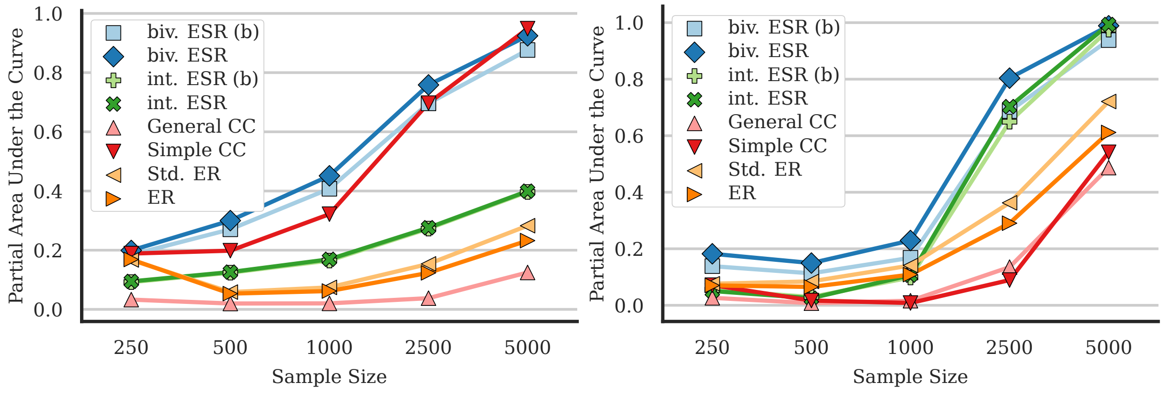


(a) Alternative model: RiskMetrics (b) Alternative model: Historical Simulation

Figure 3.1: Size-adjusted power for both alternative models for a sample size of 1000 days. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

the power against Historical Simulation is below the critical black line in the relevant range of sizes. Analogously, the ER backtest and its standardized version perform well against the Historical Simulation (especially for empirical sizes below 10%), but have hardly any power against RiskMetrics.

In order to present results in condensed form for all sample sizes, we summarize the size-adjusted power by the partial area under the curve (PAUC), as proposed by Lloyd (2005). For the PAUC, we numerically compute the area under each power curve for the empirical sizes between 1% and 10% which is thus the average power to reject a false model for the considered empirical test sizes.



(a) Alternative model: RiskMetrics (b) Alternative model: Historical Simulation

Figure 3.2: Partial area under the curve for empirical sizes between 1% and 10%. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

In Figure 3.2, we present the PAUC for all backtests and sample sizes. As expected, the average power increases in the sample size, so that using more information leads to more reliable decisions about the quality of a forecast. Furthermore, also when evaluating all considered sample sizes we find that the bivariate ESR backtest is the only approach that exhibits good power properties against both alternative models.

As a robustness check for these findings, we repeat the experiment with the DGP used by Gaglianone et al. (2011) which is less flexible than (3.31) due to its parsimonious specification. The results are presented in Section 3.A of the Appendix and we find them to be insensitive towards the choice of the DGP.

3.4.2. Continuous Model Misspecification

In the second simulation study, we use a GARCH(1,1) model with standardized Student-t distributed innovations as the true model,

$$\begin{aligned} r_t &= \sigma_t z_t, \\ \sigma_t^2 &= \gamma_0 + \gamma_1 r_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \\ z_t &\sim t_\nu, \end{aligned} \tag{3.35}$$

with the parameter values $\gamma_0 = 0.01$, $\gamma_1 = 0.1$, $\gamma_2 = 0.85$, and $\nu = 5$ for the true model. For the analysis of the backtests, we simulate 10,000 times from the true model with a sample size of 2500 observations and consider the probability level $\tau = 2.5\%$ for the risk forecasts.

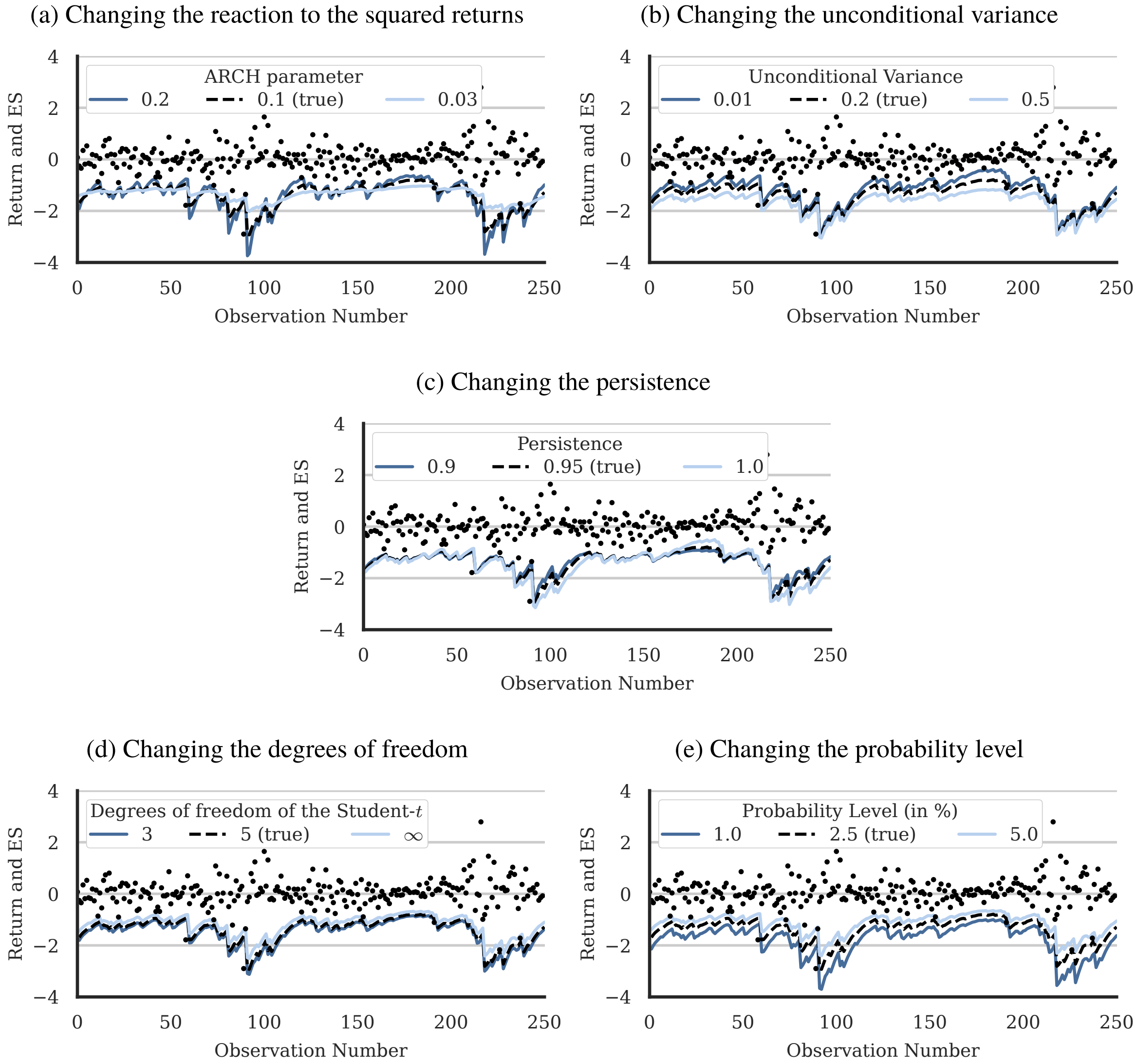
Table 3.2: Empirical sizes (nominal size: 5%) for the second simulation study.

Null Hypothesis	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
Two-Sided	0.06	0.07	0.05	0.06	0.08	0.09	0.05	0.06
One-Sided	–	–	0.07	0.03	0.02	0.02	0.06	0.06

Notes: This table shows the empirical sizes of the backtests for a GARCH(1,1)-t model. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000). Note that the bivariate ESR test does not permit a one-sided hypothesis, and therefore we only present sizes for the two-sided hypothesis.

Table 3.2 presents the empirical sizes of the backtests for a nominal size of 5% for the two- and one-sided hypotheses. As in the first simulation study, we find that most of the backtests are reasonably sized with rejection frequencies close to the nominal value. However, the two

CC tests reject the true model slightly too often in the two-sided, respectively too rarely in the one-sided case.


(a) Changing the reaction to the squared returns

(b) Changing the unconditional variance

(c) Changing the persistence

(d) Changing the degrees of freedom

(e) Changing the probability level

Figure 3.3: Illustration of the consequences for the ES forecasts of changing various aspects of the DGP. In each of the subfigures, the black dashed line corresponds to the true model.

We next describe five misspecifications of the true model, alongside with the effects on the ES forecasts and present the size-adjusted rejection rates for these modifications. As an illustrative example, Figures 3.3a to 3.3e show 250 realizations of the returns of DGP (3.35), together with the corresponding ES forecasts for the true (black dashed line) and two misspecified models. The first misspecification concerns how strongly the conditional variance reacts to the squared returns, i.e. the ARCH parameter γ_1. For this, we vary γ_1 and γ_2 such that the persistence of the GARCH process remains constant, i.e. we choose $\tilde{\gamma}_2 = 0.95 - \tilde{\gamma}_1$. When $\tilde{\gamma}_1$ is below its true value, there is little variation in the ES forecasts due to the reduced response to shocks since the GARCH process approaches a constant volatility model. For the second misspecification, we alter the unconditional variance of the GARCH process $\mathrm{E}\left[\sigma_t^2\right] = \gamma_0/(1-\gamma_1-\gamma_2)$ by changing γ_0 while holding γ_1 and γ_2 constant. Since the conditional variance is a weighted combination of the unconditional variance, the past squared returns and the past conditional variance, this change implies that the ES is always underestimated when the unconditional variance is larger than the true value, and vice versa. In the third design, we alter the persistence of shocks by setting $\tilde{\gamma}_1 = c \cdot \gamma_1$ and $\tilde{\gamma}_2 = c \cdot \gamma_2$ for a constant c that we vary, and $\tilde{\gamma}_0 = \mathrm{E}\left[\sigma_t^2\right](1-\tilde{\gamma}_1-\tilde{\gamma}_2)$ to keep the unconditional variance constant. In the exemplary series, we see that with a persistence larger than true value, the ES forecasts react stronger and longer to shocks. Fourth, we vary the degrees of freedom of the underlying Student-t distribution between 3 and ∞. Since the conditional variance is unaffected, this modification implies a relative horizontal shift of the ES forecasts. The last modification concerns the probability level $\tilde{\tau}$ that the forecaster uses for making ES predictions. This represent the scenario that a forecaster submits (accidentally or on purpose) predictions for some level $\tilde{\tau} \neq \tau$. Similar to changing the degrees of freedom, this modification implies a relative horizontal shift of the ES forecasts.

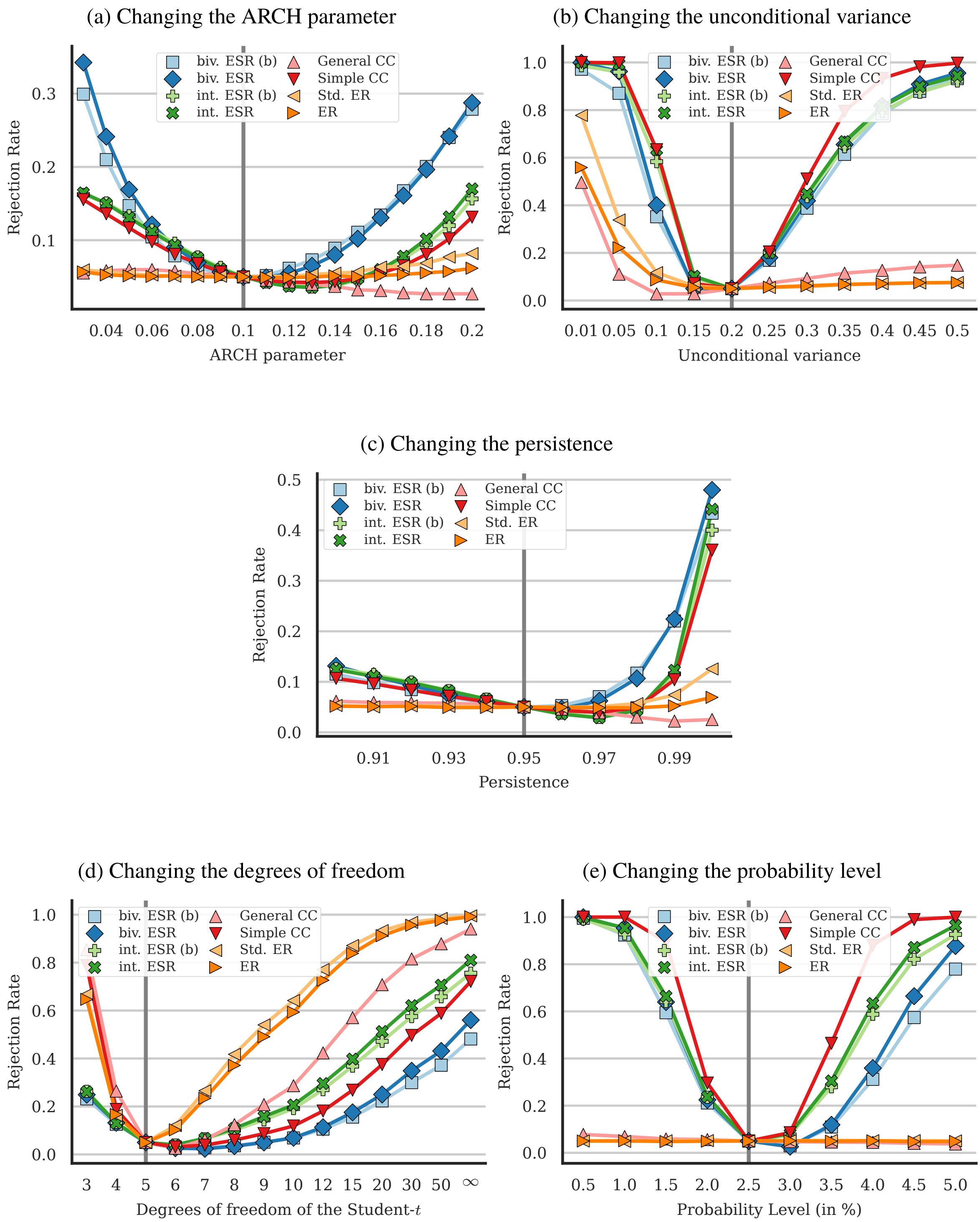
We proceed with presenting the size-adjusted rejection rates for these four designs in Figures 3.4a to 3.4e, in which the true model is indicated by the gray vertical line. Several conclusions can be drawn from this figure.

(1) Unlike in the first simulation study, there is no backtest that dominates the others throughout all considered designs. However, we can identify certain patterns about the relative performance of the tests depending on the type of misspecification.

(2) Our bivariate ESR test performs well when we change the dynamics of the ES forecasts, i.e. in the cases of changing the ARCH parameter, the persistence or the unconditional variance of the GARCH process, see Figures 3.4a to 3.4c. There, the power of the other backtests is comparably low, especially the general CC and the two ER tests are not able to detect these misspecification. However, the bivariate ESR test is not as powerful as our intercept ESR or the simple CC test when we horizontally shift the ES forecasts, i.e. when changing the degrees of freedom of the Student-t distribution or when the forecaster uses the wrong probability level as can be seen in Figures 3.4d and 3.4e.

(3) The application of the bootstrap for our ESR tests mainly affects the empirical sizes, the empirical power of the asymptotic and the bootstrap ESR tests is similar throughout all designs.

(4) The general CC and the two ER tests perform similar across all designs. However, only when altering the degrees of freedom they exhibit good power properties, in the other scenarios, they can hardly distinguish between forecasts of the true and the alternative models.



(a) Changing the ARCH parameter

(b) Changing the unconditional variance

(c) Changing the persistence

(d) Changing the degrees of freedom

(e) Changing the probability level

Figure 3.4: Size-adjusted rejection rates for various types of misspecification. The gray vertical line depicts the true model. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

In particular, they can not discriminate between ES forecasts for the level τ and some $\tilde{\tau} \neq \tau$ (cf. Section 3.3.1), although changing the degrees of freedom and the probability level have a similar effect on the ES forecasts (see Figures 3.3d and 3.3e). Thus, if these backtests would be used by the regulatory authorities, the banks could submit ES forecasts for some level $\tilde{\tau} > \tau$ to minimize the capital requirements without risk of being detected by these backtests.

(5) In this experiment, the simple CC backtest is very powerful in two of the five designs, see Figures 3.4b and 3.4e, and generally exhibits a relatively good performance that is often similar to our proposed tests. However, our two ESR backtests exhibit much better size properties (see Tables 3.1 and 3.2) and do not fail (in contrast to the simple CC test) in rejecting the Historical Simulation forecasts in the first simulation study (see Figure 3.1).

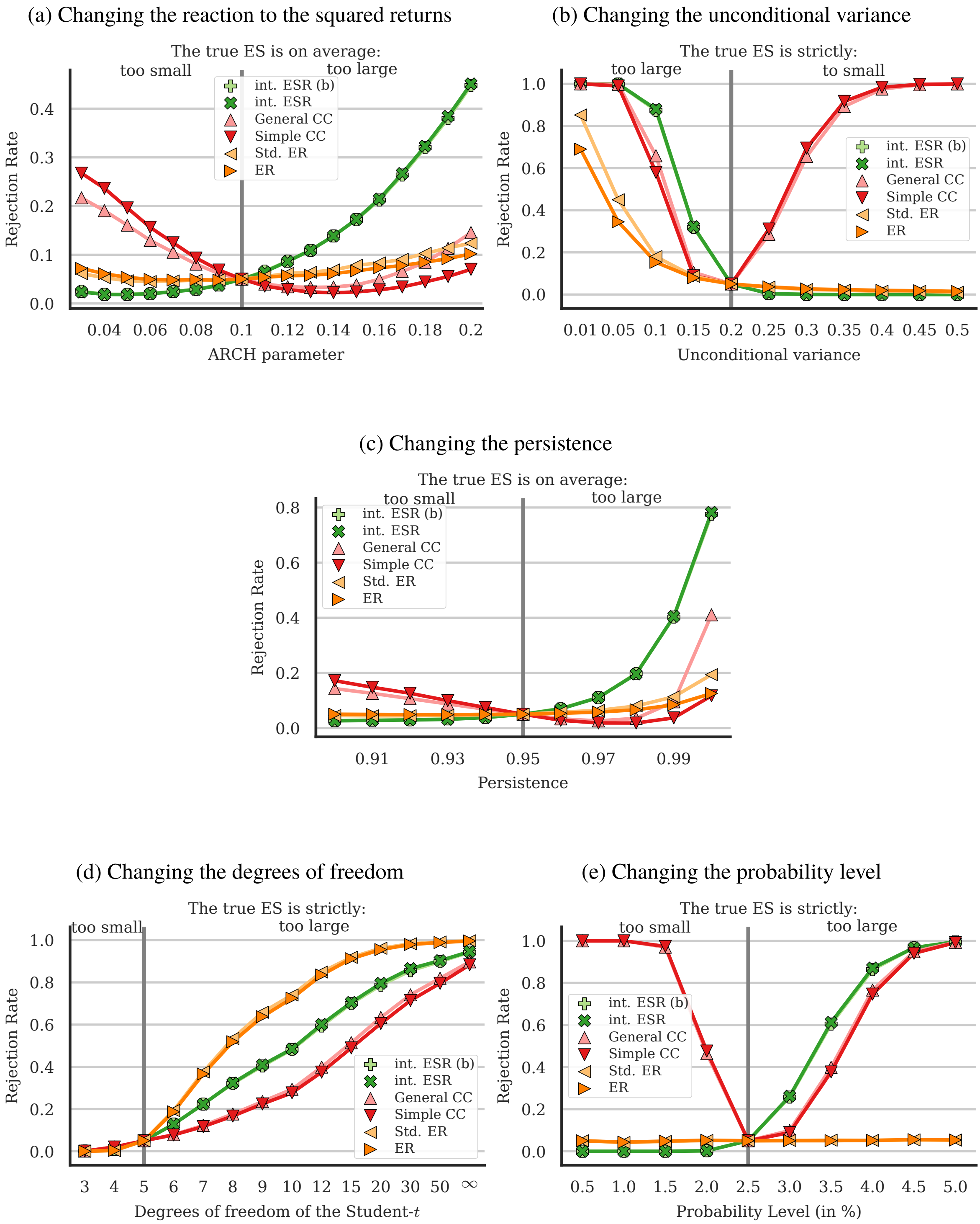
(6) Although the ER and our intercept ESR test are conceptually similar (see the discussion in Section 3.3.1), the latter is in four of the five scenarios clearly more powerful, which shows that jointly backtesting the VaR and ES without explicitly incorporating the probability level τ can be problematic.

Taken as a whole, these findings, together with the results from the first simulation study, show that our proposed ESR backtests are a powerful choice for backtesting ES forecasts. They exhibits good power properties against a variety of misspecifications and are reasonably sized. Notably, in contrast to the existing backtests, there is no single situation where our ESR tests are unable to discriminate between forecasts of the true and the alternative models.

3.4.3. Testing one-sided hypotheses

For the regulatory authorities, a one-sided hypothesis might be more meaningful than the two-sided version we considered so far. Holding more money than stipulated in the Basel accords is no concern for regulators, as it is only important that banks keep enough monetary reserves to cover the risk from their market activities. As all backtests (with exception of the bivariate ESR) allow for testing one-sided hypotheses, we assess their ability to reject the null hypothesis that the misspecified ES forecasts overestimate the true ES.

In Figures 3.5a to 3.5e, we present the size-adjusted rejection rates for one-sided hypothesis tests and the structure of these figures is analog to the two-sided rejection rates we considered in the previous section. However, the backtests should now only reject the null hypothesis for ES predictions that overestimate the ES. The five modifications of the true model in Section 3.4.2 exhibit clear patterns when they are over-, respectively underestimating the true ES. In three of the five designs, the ES is either strictly overestimated or underestimated in every period, whereas in the remaining two designs this is at least on average the case. For details on this, see the upper part of Figures 3.5a to 3.5e which


(a) Changing the reaction to the squared returns

(b) Changing the unconditional variance

(c) Changing the persistence

(d) Changing the degrees of freedom

(e) Changing the probability level

Figure 3.5: Size-adjusted rejection rates for various types of misspecification with a one-sided hypothesis. The gray vertical line depicts the true model. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

indicates the regions of parameter values where the ES forecasts are too large, respectively too small.

In these five figures, we find that our intercept ESR (in the asymptotic and the bootstrap version) backtest clearly dominates the ER and the CC tests in four out of five designs. Only when changing the degrees of freedom, the ER test is slightly more powerful than the intercept ESR. Surprisingly, we see that the one-sided CC tests (both, the simple and the general version) reject too *small* ES forecasts in four out of the five cases, i.e. they make a wrong decision.[7]

Summarizing these results, the proposed intercept ESR backtest is a powerful backtest with good size properties for testing one-sided hypotheses. The existing backtests either fail to detect rather obvious misspecifications or make false decisions altogether.

3.5. Empirical application

In the empirical application, we predict the market risk for the daily close-to-close log-returns of the S&P500 index for the time period from January 3, 2000 to October 18, 2017, totaling up to 4478 days. We predict the ES (and the VaR for application of the existing tests) for this return series using 10 different risk models. The first two are the Historical Simulation estimated with a rolling window of 250 days and RiskMetrics. The other 8 models are combinations of the GARCH(1,1) and the asymmetric GJR-GARCH(1,1) of Glosten et al. (1993) with with four assumptions on the conditional distribution of the innovations. These are the standard normal distribution (abbreviated by N), the standardized Student-t (t), the standardized skewed Student-t (skew-t) and the semi-parametric filtered historical simulation approach (FHS) of Barone-Adesi et al. (1999), where the quantile, respectively the ES of the innovations is estimated from the standardized returns. We estimate these 8 models on a rolling window of 1000 days.

Table 3.3 presents the p-values of the backtests (for the two-sided hypothesis), the average losses of the 0-homogeneous loss function (3.12), and the p-value of the Model Confidence Set (MCS) of Hansen et al. (2011) applied to this loss function. With the MCS p-values, we can determine a set of models having equal predictive ability at a certain significance level with respect to the losses. The models are sorted according to the average loss.

From this table we can draw several conclusions. First, the MCS rejects 7 out of 10 models at the 5% significance level, i.e. only 3 models have equal predictive power with respect to the joint loss function. These three (GJR-GARCH-skew-t, -FHS, -t) share the same assumption on the volatility process and only differ with respect to the assumption on

[7] We verified our implementation of the CC tests with the codes provided by Nolde and Ziegel (2017) at https://github.com/nnolde/Elicitability-and-Backtesting/.

Table 3.3: Results of the empirical application.

Model	biv.	biv	int.	int.	General	Simple	Std.	ER	Mean	MCS
	ESR (b)	ESR	ESR (b)	ESR	CC	CC	ER		Loss	p-value
Historical Simulation	**0.01**	**0.00**	**0.01**	**0.00**	0.11	**0.01**	0.06	0.06	1.132	**0.01**
RiskMetrics	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.075	**0.00**
GARCH-N	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.030	**0.01**
GARCH-t	**0.03**	**0.05**	**0.05**	**0.03**	0.57	**0.02**	0.58	0.60	1.000	**0.03**
GARCH-skew-t	0.12	0.19	0.92	0.92	0.33	**0.05**	0.38	0.10	0.986	**0.05**
GARCH-FHS	0.09	0.14	0.19	0.17	0.67	0.31	0.68	0.69	0.993	**0.03**
GJR-GARCH-N	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.982	**0.02**
GJR-GARCH-t	0.06	0.10	0.06	**0.04**	0.28	0.11	0.23	0.90	0.963	0.28
GJR-GARCH-skew-t	0.07	0.12	0.78	0.77	0.85	0.08	0.87	0.14	0.951	1.00
GJR-GARCH-FHS	0.13	0.20	0.30	0.30	0.39	0.69	0.34	0.55	0.953	0.70

Notes: In this table, p-values smaller than 5% are printed bold-faced and the models are sorted by the average loss. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000). We compute the MCS p-values using the R-statistic of Hansen et al. (2011) and 100,000 bootstrap iterations of the stationary bootstrap (Politis and Romano, 1994) with an average block length of 10 days.

the innovations. Moreover, for these three models the null hypothesis of correct forecasts is not rejected by almost all backtests at the 5% significance level. Thus, the backtests and the MCS agree on which models predict the ES (and the VaR) well. Second, incorporating leverage into the volatility dynamics appears to be important, since mainly the models using the GJR-GARCH are not rejected by the backtests. Additionally, it is crucial to consider models with flexible tails, e.g. by using the skewed Student-t or the FHS approach, since the models based on conditionally normally distributed returns are collectively rejected by the backtests and the MCS. Third, the CC and ER tests reject less forecasts at the 5% significance level than the two ESR backtests, which reflects the findings of the simulation studies where these backtests are often less powerful than our ESR tests. In particular, the null hypothesis is not rejected for the Historical Simulation model, although this approach yields large losses.

3.6. Conclusion

In this paper, we introduce two novel regression-based backtests for forecasts of the risk measure ES. These are based on the idea of Mincer and Zarnowitz (1969) to regress the response variable on the forecasts and to test the resulting parameter estimates. We introduce a bivariate version where we test the intercept and the slope parameter for 0 and 1, and an intercept version that only incorporates an intercept term being estimated and tested for 0. The motivation for the latter test is the possibility to specify a one-sided hypothesis that is especially relevant for the regulatory authorities, whereas the bivariate test only permits a two-sided hypothesis.

A unique feature of the backtests proposed in this paper is that they solely require and consequently test forecasts of the ES. In contrast to that, a common drawback of the existing backtests is that they need forecasts of further input parameters, such as the VaR, the volatility, the tail distribution or even the whole return distribution. Using more information than the ES forecasts is problematic for two reasons. First, these tests are not applicable for the regulatory authorities, who receive forecasts of the risk measures, but not of the additional information required by many tests. Second, rejecting the null hypothesis does not necessarily imply that the ES forecasts are wrong since the rejection could be a result of a false prediction of any of the input parameters.

In several simulation studies, we assess the empirical size and power properties of the proposed tests and compare them to the approaches of McNeil and Frey (2000) and Nolde and Ziegel (2017). We find that our regression-based tests are reasonably sized, especially when they are applied using the bootstrap. Moreover, in most simulation designs our two proposed backtests are more powerful than the existing tests. The backtests from the literature are often not able to distinguish between forecasts of the true model and the misspecified forecasts, for instance when the forecaster reports predictions for a wrong probability level. In contrast to that, our two backtests detect the misspecification in all considered simulation experiments. We provide an implementation of our backtests and several approaches from the literature in the `esback` package for R (Bayer and Dimitriadis, 2017a).

For future research, it could be interesting to disentangle the VaR and ES forecast performance of frequently used risk models to determine whether some models are better suited for predicting the ES (or the VaR) than others. It could also be interesting to introduce an ES encompassing test analog to the quantile encompassing test of Giacomini and Komunjer (2005).

Appendix 3.A Robustness Check

The DGP used by Gaglianone et al. (2011) is a GARCH(1,1) model with standard normally distributed innovations,

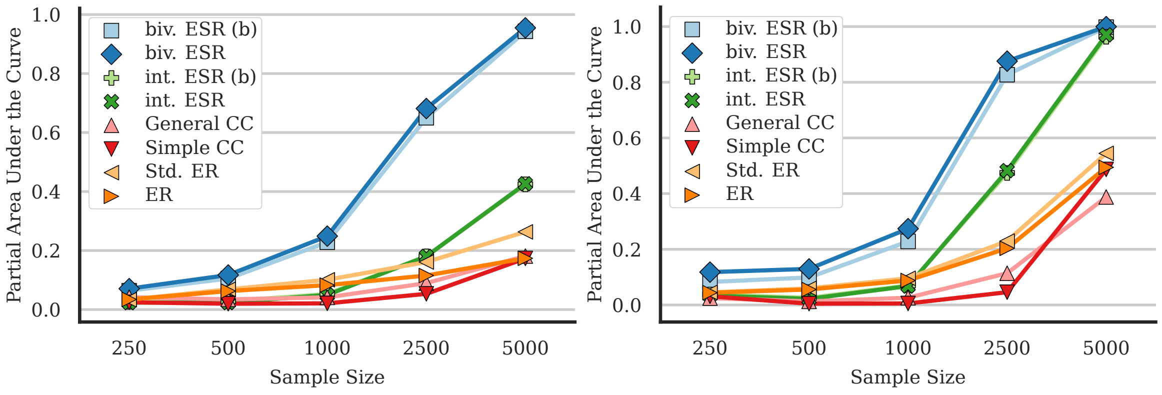
$$\begin{aligned} r_t &= \sigma_t z_t, \\ \sigma_t^2 &= 0.05 + 0.05 r_{t-1}^2 + 0.90 \sigma_{t-1}^2, \\ z_t &\sim \mathcal{N}(0,1)\,. \end{aligned} \tag{3.36}$$

For this DGP, Table 3.A.4 and Figure 3.A.6 present the empirical sizes and the PAUC analog to the results provided in Section 3.4.1.

Table 3.A.4: Empirical sizes of the backtests

Nominal Size	Sample Size	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
1%	250	0.02	0.10	0.01	0.07	0.01	0.17	0.04	0.04
	500	0.02	0.05	0.01	0.05	0.02	0.08	0.00	0.00
	1000	0.01	0.04	0.01	0.03	0.02	0.05	0.00	0.01
	2500	0.01	0.02	0.01	0.02	0.02	0.03	0.01	0.01
	5000	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01
5%	250	0.08	0.18	0.05	0.13	0.06	0.22	0.06	0.07
	500	0.06	0.12	0.05	0.10	0.07	0.14	0.04	0.04
	1000	0.06	0.09	0.05	0.07	0.07	0.10	0.04	0.04
	2500	0.06	0.07	0.05	0.06	0.06	0.07	0.05	0.05
	5000	0.05	0.06	0.05	0.06	0.05	0.06	0.05	0.05
10%	250	0.14	0.24	0.10	0.18	0.13	0.26	0.11	0.11
	500	0.12	0.18	0.10	0.14	0.13	0.19	0.08	0.08
	1000	0.11	0.14	0.10	0.12	0.12	0.15	0.09	0.09
	2500	0.10	0.12	0.10	0.11	0.11	0.12	0.10	0.10
	5000	0.10	0.11	0.10	0.11	0.11	0.11	0.10	0.10

Notes: The table reports the empirical sizes of the backtests for a GARCH(1,1)-N process. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).



(a) Alternative model: RiskMetrics

(b) Alternative model: Historical Simulation

Figure 3.A.6: Partial area under the curve for empirical sizes between 1% and 10%

Appendix 3.B Raw Power

Table 3.B.5: Empirical power of the backtests against RiskMetrics

Nominal Size	Sample Size	bivariate ESR (b)	bivariate ESR	intercept ESR (b)	intercept ESR	General CC	Simple CC	Std. ER	ER
1%	250	0.13	0.40	0.04	0.18	0.02	0.40	0.16	0.16
	500	0.19	0.41	0.06	0.16	0.01	0.37	0.01	0.01
	1000	0.27	0.48	0.08	0.15	0.01	0.42	0.01	0.01
	2500	0.53	0.68	0.11	0.17	0.01	0.68	0.04	0.04
	5000	0.78	0.87	0.19	0.25	0.02	0.93	0.11	0.08
5%	250	0.27	0.50	0.12	0.25	0.05	0.48	0.17	0.18
	500	0.35	0.53	0.15	0.24	0.06	0.49	0.04	0.05
	1000	0.48	0.61	0.19	0.25	0.05	0.57	0.07	0.07
	2500	0.73	0.81	0.27	0.31	0.06	0.83	0.15	0.13
	5000	0.89	0.94	0.38	0.41	0.15	0.97	0.28	0.24
10%	250	0.37	0.56	0.19	0.31	0.11	0.52	0.20	0.21
	500	0.46	0.60	0.23	0.31	0.11	0.55	0.10	0.11
	1000	0.59	0.69	0.27	0.32	0.11	0.65	0.14	0.14
	2500	0.81	0.86	0.36	0.39	0.15	0.88	0.25	0.23
	5000	0.93	0.96	0.49	0.51	0.28	0.99	0.39	0.35

Notes: The table reports the empirical power of the backtests against RiskMetrics for an AR(1)-GARCH(1,1)-skewed-t process. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

Table 3.B.6: Empirical power of the backtests against Historical Simulation

Nominal	Sample	bivariate	bivariate	intercept	intercept	General	Simple	Std.	ER
Size	Size	ESR (b)	ESR	ESR (b)	ESR	CC	CC	ER	
1%	250	0.10	0.34	0.02	0.10	0.01	0.16	0.05	0.05
	500	0.07	0.21	0.00	0.03	0.00	0.04	0.01	0.01
	1000	0.06	0.23	0.01	0.05	0.00	0.01	0.03	0.03
	2500	0.29	0.63	0.14	0.40	0.04	0.04	0.14	0.11
	5000	0.79	0.96	0.79	0.96	0.20	0.28	0.41	0.33
5%	250	0.20	0.46	0.07	0.20	0.04	0.22	0.09	0.09
	500	0.15	0.37	0.04	0.13	0.03	0.08	0.07	0.06
	1000	0.22	0.49	0.11	0.24	0.06	0.06	0.14	0.12
	2500	0.77	0.91	0.66	0.81	0.22	0.21	0.37	0.32
	5000	0.97	1.00	0.98	1.00	0.57	0.74	0.73	0.63
10%	250	0.29	0.54	0.14	0.28	0.11	0.28	0.15	0.15
	500	0.24	0.50	0.12	0.22	0.09	0.14	0.14	0.13
	1000	0.45	0.66	0.29	0.41	0.15	0.14	0.24	0.21
	2500	0.93	0.97	0.88	0.94	0.39	0.39	0.53	0.47
	5000	0.99	1.00	1.00	1.00	0.77	0.90	0.85	0.78

Notes: The table reports the empirical power of the backtests against the Historical Simulation for an AR(1)-GARCH(1,1)-skewed-t process. The number of Monte-Carlo repetitions is 10,000 and the probability level for the risk measures is $\tau = 2.5\%$. ESR refers to the backtests introduced in this paper with (b) indicating the bootstrap version, CC to the conditional calibration tests of Nolde and Ziegel (2017), and ER to the exceedance residuals tests of McNeil and Frey (2000).

References

Acerbi, C. and B. Szekely (2014). “Back-testing Expected Shortfall”. *Risk* December, 76–81 (see pp. 53, 103).

Aramonte, S., P. Durand, S. Kobayashi, M. Kwast, J. A. Lopez, G. Mazzoni, P. Raupach, M. Summer, and J. Wu (2011). *Messages from the academic literature on risk measurement for the trading book*. Tech. rep. Working Paper No. 19, available at http://www.bis.org/publ/bcbs_wp19.pdf. Bank for International Settlements (see pp. 95, 103).

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). “Coherent Measures of Risk”. *Mathematical Finance* 9 (3), 203–228 (see pp. 52, 95).

Barendse, S. (2017). “Interquantile Expectation Regression”. Available at https://ssrn.com/abstract=2937665 (see pp. 53, 95, 100).

Barone-Adesi, G., K. Giannopoulos, and L. Vosper (1999). “VaR without correlations for portfolios of derivative securities”. *Journal of Futures Markets* 19 (5), 583–602 (see pp. 27, 117).

Basel Committee (1996). *Overview of the Amendment to the Capital Accord to Incorporate Market Risks*. Tech. rep. Available at http://www.bis.org/publ/bcbs23.pdf. Bank for International Settlements (see pp. 15, 96, 102).

Basel Committee (2013). *Fundamental review of the trading book: A revised market risk framework*. Tech. rep. Available at http://www.bis.org/publ/bcbs265.pdf. Bank for International Settlements (see p. 95).

— (2016). *Minimum capital requirements for Market Risk*. Tech. rep. Available at http://www.bis.org/bcbs/publ/d352.pdf. Bank for International Settlements (see pp. 52, 95).

— (2017). *Pillar 3 disclosure requirements – consolidated and enhanced framework*. Tech. rep. Available at http://www.bis.org/bcbs/publ/d400.pdf. Basel Committee on Banking Supervision (see pp. 95, 103).

Bayer, S. and T. Dimitriadis (2017a). *esback: Expected Shortfall Backtesting*. R package version 0.1.1, available at https://github.com/BayerSe/esback (see pp. 106, 119).

Berkowitz, J. (2001). "Testing Density Forecasts, With Applications to Risk Management". *Journal of Business & Economic Statistics* 19 (4), 465–474 (see p. 103).

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics* 31 (3), 307–327 (see pp. 26, 70, 107).

Christoffersen, P. (1998). "Evaluating Interval Forecasts". *International Economic Review* 39 (4), 841–862 (see pp. 17, 30, 98).

Costanzino, N. and M. Curran (2015). "Backtesting general spectral risk measures with application to expected shortfall". *Journal of Risk Model Validation* 9 (1), 21–31 (see p. 103).

Dimitriadis, T. and S. Bayer (2017). "A Joint Quantile and Expected Shortfall Regression Framework". arXiv:1704.02213 [math.ST] (see pp. 95, 96, 99–101).

Du, Z. and J. C. Escanciano (2017). "Backtesting Expected Shortfall: Accounting for Tail Risk". *Management Science* 63 (4), 940–958 (see p. 103).

Efron, B. (Jan. 1979). "Bootstrap Methods: Another Look at the Jackknife". *The Annals of Statistics* 7 (1), 1–26 (see pp. 64, 102).

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall (see p. 104).

Emmer, S., M. Kratz, and D. Tasche (2015). "What Is the Best Risk Measure in Practice? A Comparison of Standard Measures". *Journal of Risk* 18 (2), 31–60 (see p. 103).

Engle, R. F. and S. Manganelli (2004). "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles". *Journal of Business & Economic Statistics* 22 (4), 367–381 (see pp. 17, 26, 30, 98).

Fernandez, C. and M. F. J. Steel (1998). "On Bayesian Modeling of Fat Tails and Skewness". *Journal of the American Statistical Association* 93 (441), 359–371 (see p. 107).

Fissler, T. and J. F. Ziegel (2016). "Higher order elicitability and Osband's principle". *Annals of Statistics* 44 (4), 1680–1707 (see pp. 9, 12, 53–56, 58, 70, 71, 76, 95, 100).

Fissler, T., J. F. Ziegel, and T. Gneiting (2016). "Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting". *Risk* January, 58–61 (see pp. 53, 56, 61, 62, 95).

Gaglianone, W. P., L. R. Lima, O. Linton, and D. R. Smith (2011). "Evaluating Value-at-Risk Models via Quantile Regression". *Journal of Business & Economic Statistics* 29 (1), 150–160 (see pp. 96, 111, 119).

Giacomini, R. and I. Komunjer (2005). “Evaluation and Combination of Conditional Quantile Forecasts”. *Journal of Business & Economic Statistics* 23 (4), 416–431 (see pp. 17, 19, 119).

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). “On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks”. *The Journal of Finance* 48 (5), 1779–1801 (see p. 117).

Graham, A. and J. Pál (2014). “Backtesting value-at-risk tail losses on a dynamic portfolio”. *The Journal of Risk Model Validation* 8 (2), 59 (see p. 103).

Guler, K., P. T. Ng, and Z. Xiao (2017). “Mincer–Zarnowitz quantile and expectile regressions for forecast evaluations under aysmmetric loss functions”. *Journal of Forecasting* 36 (6), 651–679 (see p. 96).

Hansen, P. R., A. Lunde, and J. M. Nason (2011). “The Model Confidence Set”. *Econometrica* 79 (2), 453–497 (see pp. 17, 18, 30, 31, 117, 118).

Holden, K. and D. A. Peel (1990). “On Testing For Unbiasedness And Efficiency Of Forecasts”. *The Manchester School* 58 (2), 120–127 (see p. 99).

Kerkhof, J. and B. Melenberg (2004). “Backtesting for risk-based regulatory capital”. *Journal of Banking & Finance* 28 (8), 1845–1865 (see p. 103).

Komunjer, I. (2013). “Quantile Prediction”. In: *Handbook of Economic Forecasting*. Ed. by Elliott, G. and Timmermann, A. Vol. 2. Elsevier. Chap. 17, 961–994 (see pp. 15, 98).

Kratz, M., Y. H. Lok, and A. J. McNeil (2017). “Multinomial VaR Backtests: A simple implicit approach to backtesting expected shortfall”. arXiv:1611.04851 [q-fin.RM] (see p. 103).

Kupiec, P. H. (1995). “Techniques for Verifying the Accuracy of Risk Measurement Models”. *The Journal of Derivatives* 3 (2), 73–84 (see pp. 18, 30, 35, 36, 98).

Lambert, P. and S. Laurent (2002). “Modelling skewness dynamics in series of financial data”. Université Catholique de Louvain and Université de Liège, available at `http://hdl.handle.net/2078.1/91035` (see p. 107).

Lloyd, C. J. (2005). “Estimating test power adjusted for size”. *Journal of Statistical Computation and Simulation* 75 (11), 921–933 (see pp. 109, 110).

MacKinnon, J. G. (2009). “Bootstrap Hypothesis Testing”. In: *Handbook of Computational Econometrics*. Ed. by Belsley, D. A. and Kontoghiorghes, E. J. John Wiley & Sons, Ltd. Chap. 6, 183–213 (see p. 102).

McNeil, A. J. and R. Frey (2000). “Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach”. *Journal of Empirical Finance* 7 (3–4), 271–300 (see pp. 96, 97, 103, 104, 106, 108, 110, 111, 114, 116, 118–122).

Mincer, J. and V. Zarnowitz (1969). “The Evaluation of Economic Forecasts”. In: *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*. National Bureau of Economic Research, Inc, 3–46 (see pp. 9, 13, 95, 96, 99, 118).

Nadarajah, S., B. Zhang, and S. Chan (2014). “Estimation methods for expected shortfall”. *Quantitative Finance* 14 (2), 271–291 (see pp. 52, 98).

Nolde, N. and J. F. Ziegel (2017). “Elicitability and backtesting: Perspectives for banking regulation”. arXiv:1608.05498 [q-fin.RM] (see pp. 53, 54, 61, 62, 67, 95–97, 100, 103–108, 110, 111, 114, 116–122).

Patton, A. J., J. F. Ziegel, and R. Chen (2017). “Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)”. arXiv:1707.05108 [q-fin.EC] (see pp. 95, 100).

Politis, D. N. and J. P. Romano (1994). “The Stationary Bootstrap”. *Journal of the American Statistical Association* 89 (428), 1303–1313 (see p. 118).

Righi, M. B. and P. S. Ceretta (2013). “Individual and flexible expected shortfall backtesting”. *Journal of Risk Model Validation* 7 (3) (see p. 103).

— (2015). “A comparison of Expected Shortfall estimation models”. *Journal of Economics and Business* 78, 14–47 (see p. 103).

Taylor, J. W. (2017). “Forecasting Value at Risk and Expected Shortfall Using a Semiparametric Approach Based on the Asymmetric Laplace Distribution”. *Forthcoming in Journal of Business & Economic Statistics*. DOI: 10.1080/07350015.2017.1281815 (see pp. 52, 100).

Trottier, D.-A. and D. Ardia (2016). “Moments of standardized Fernandez-Steel skewed distributions: Applications to the estimation of GARCH-type models”. *Finance Research Letters* 18, 311–316 (see p. 107).

Wong, W. (2008). “Backtesting trading risk of commercial banks using expected shortfall”. *Journal of Banking & Finance* 32 (7), 1404–1415 (see p. 103).

Zangari, P. (1996). *RiskMetrics – Technical Document*. Tech. rep. New York: Morgan Guaranty Trust Company (see p. 108).

Complete References

Abad, P. and S. Benito (2013). “A detailed comparison of value at risk estimates”. *Mathematics and Computers in Simulation* 94, 258–276 (see p. 15).

Acerbi, C. and B. Szekely (2014). “Back-testing Expected Shortfall”. *Risk* December, 76–81 (see pp. 53, 103).

Aiolfi, M. and A. Timmermann (2006). “Persistence in forecasting performance and conditional combination strategies”. *Journal of Econometrics* 135 (1â€“2), 31–53 (see p. 34).

Andersen, T. and T. Bollerslev (1998). “Answering the skeptics: Yes, standard volatility models do provide accurate forecasts”. *International Economic Review* 39 (4), 885–905 (see p. 70).

Andrews, D. (1994). “Empirical Process Methods in Econometrics”. In: *Handbook of Econometrics*. Ed. by Engle, R. and McFadden, D. Vol. 4. Elsevier. Chap. 37, 2247–2294 (see p. 73).

Aramonte, S., P. Durand, S. Kobayashi, M. Kwast, J. A. Lopez, G. Mazzoni, P. Raupach, M. Summer, and J. Wu (2011). *Messages from the academic literature on risk measurement for the trading book*. Tech. rep. Working Paper No. 19, available at http://www.bis.org/publ/bcbs_wp19.pdf. Bank for International Settlements (see pp. 95, 103).

Arlot, S. and A. Celisse (2010). “A survey of cross-validation procedures for model selection”. *Statistics Surveys* 4, 40–79 (see p. 23).

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). “Coherent Measures of Risk”. *Mathematical Finance* 9 (3), 203–228 (see pp. 52, 95).

Barendse, S. (2017). “Interquantile Expectation Regression”. Available at https://ssrn.com/abstract=2937665 (see pp. 53, 95, 100).

Barone-Adesi, G., K. Giannopoulos, and L. Vosper (1999). “VaR without correlations for portfolios of derivative securities”. *Journal of Futures Markets* 19 (5), 583–602 (see pp. 27, 117).

Basel Committee (1996). *Overview of the Amendment to the Capital Accord to Incorporate Market Risks*. Tech. rep. Available at http://www.bis.org/publ/bcbs23.pdf. Bank for International Settlements (see pp. 15, 96, 102).

— (2006). *International Convergence of Capital Measurement and Capital Standards*. Tech. rep. Available at http://www.bis.org/publ/bcbs107.pdf. Bank for International Settlements (see p. 15).

— (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems*. Tech. rep. Available at http://www.bis.org/publ/bcbs189.pdf. Bank for International Settlements (see p. 15).

Basel Committee (2013). *Fundamental review of the trading book: A revised market risk framework*. Tech. rep. Available at http://www.bis.org/publ/bcbs265.pdf. Bank for International Settlements (see p. 95).

— (2016). *Minimum capital requirements for Market Risk*. Tech. rep. Available at http://www.bis.org/bcbs/publ/d352.pdf. Bank for International Settlements (see pp. 52, 95).

— (2017). *Pillar 3 disclosure requirements – consolidated and enhanced framework*. Tech. rep. Available at http://www.bis.org/bcbs/publ/d400.pdf. Basel Committee on Banking Supervision (see pp. 95, 103).

Bayer, S. and T. Dimitriadis (2017a). *esback: Expected Shortfall Backtesting*. R package version 0.1.1, available at https://github.com/BayerSe/esback (see pp. 106, 119).

— (2017b). *esreg: Joint Quantile and Expected Shortfall Regression*. R package version 0.3.1, available at https://CRAN.R-project.org/package=esreg (see pp. 54, 64, 71).

— (2017c). “Regression-based Expected Shortfall Backtesting”. Working Paper (see p. 71).

Belloni, A. and V. Chernozhukov (2011). “ℓ_1-penalized quantile regression in high-dimensional sparse models”. *The Annals of Statistics* 39 (1), 82–130 (see p. 41).

Berkowitz, J. (2001). “Testing Density Forecasts, With Applications to Risk Management”. *Journal of Business & Economic Statistics* 19 (4), 465–474 (see p. 103).

Berkowitz, J., P. Christoffersen, and D. Pelletier (2011). “Evaluating value-at-risk models with desk-level data”. *Management Science* 57 (12), 2213–2227 (see p. 30).

Bernardi, M. and L. Catania (2016). “Comparison of Value-at-Risk models using the MCS approach”. *Computational Statistics* 31 (2), 579–608 (see pp. 15, 30, 37).

Bernardi, M., L. Catania, and L. Petrella (2017). “Are news important to predict the Value-at-Risk?” *The European Journal of Finance* 23 (6), 535–572 (see p. 17).

Bollerslev, T. (1986). “Generalized autoregressive conditional heteroskedasticity”. *Journal of Econometrics* 31 (3), 307–327 (see pp. 26, 70, 107).

Boucher, C. M., J. Danielsson, P. S. Kouontchou, and B. B. Maillet (2014). “Risk models-at-risk”. *Journal of Banking & Finance* 44, 72–92 (see p. 15).

Boudoukh, J., M. Richardson, and R. F. Whitelaw (1998). “The Best of Both Worlds: A Hybrid Approach to Calculating Value at Risk”. *Risk* 11 (5), 64–67 (see p. 26).

Brazauskas, V., B. L. Jones, M. L. Puri, and R. Zitikis (2008). “Estimating conditional tail expectation with actuarial applications in view”. *Journal of Statistical Planning and Inference* 138 (11), 3590–3604 (see p. 61).

Casarin, R., C.-L. Chang, J.-A. Jimenez-Martin, M. McAleer, and T. Perez-Amaral (2013). “Risk management of risk under the Basel Accord: A Bayesian approach to forecasting Value-at-Risk of VIX futures”. *Mathematics and Computers in Simulation* 94, 183–204 (see p. 17).

Chen, S. X. (2008). “Nonparametric Estimation of Expected Shortfall”. *Journal of Financial Econometrics* 6 (1), 87–107 (see p. 61).

Christoffersen, P. (1998). “Evaluating Interval Forecasts”. *International Economic Review* 39 (4), 841–862 (see pp. 17, 30, 98).

Corsi, F. (2009). “A simple approximate long-memory model of realized volatility”. *Journal of Financial Econometrics* 7 (2), 174–196 (see p. 70).

Costanzino, N. and M. Curran (2015). “Backtesting general spectral risk measures with application to expected shortfall”. *Journal of Risk Model Validation* 9 (1), 21–31 (see p. 103).

Dimitriadis, T. and S. Bayer (2017). “A Joint Quantile and Expected Shortfall Regression Framework”. arXiv:1704.02213 [math.ST] (see pp. 95, 96, 99–101).

Ding, Z., C. W. J. G. Granger, and R. F. Engle (1993). “A long memory property of stock market returns and a new model”. *Journal of Empirical Finance* 1 (1), 83–106 (see p. 27).

Du, Z. and J. C. Escanciano (2017). “Backtesting Expected Shortfall: Accounting for Tail Risk”. *Management Science* 63 (4), 940–958 (see p. 103).

Efron, B. (Jan. 1979). “Bootstrap Methods: Another Look at the Jackknife”. *The Annals of Statistics* 7 (1), 1–26 (see pp. 64, 102).

— (1991). “Regression percentiles using asymmetric squared error loss”. *Statistica Sinica* 1 (1), 93–125 (see pp. 54, 61).

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall (see p. 104).

Ehm, W., T. Gneiting, A. Jordan, and F. Krüger (2016). “Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78 (3), 505–562 (see p. 70).

Einhorn, D. (2008). “Private Profits and Socialized Risk”. In: *Global Association of Risk Professionals Risk Review (June/July 2008)*. Ed. by Einhorn, D. and Brown, A. Vol. 42, 10–26 (see p. 15).

Emmer, S., M. Kratz, and D. Tasche (2015). “What Is the Best Risk Measure in Practice? A Comparison of Standard Measures”. *Journal of Risk* 18 (2), 31–60 (see p. 103).

Engle, R. F. and S. Manganelli (2004). “CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles”. *Journal of Business & Economic Statistics* 22 (4), 367–381 (see pp. 17, 26, 30, 98).

Ergen, I. (2015). “Two-step methods in VaR prediction and the importance of fat tails”. *Quantitative Finance* 15 (6), 1013–1030 (see p. 15).

Fernandez, C. and M. F. J. Steel (1998). “On Bayesian Modeling of Fat Tails and Skewness”. *Journal of the American Statistical Association* 93 (441), 359–371 (see p. 107).

Fissler, T. (2017). “On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space”. PhD thesis. Universität Bern (see p. 63).

Fissler, T. and J. F. Ziegel (2016). “Higher order elicitability and Osband’s principle”. *Annals of Statistics* 44 (4), 1680–1707 (see pp. 9, 12, 53–56, 58, 70, 71, 76, 95, 100).

Fissler, T., J. F. Ziegel, and T. Gneiting (2016). “Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting”. *Risk* January, 58–61 (see pp. 53, 56, 61, 62, 95).

Fuertes, A.-M. and J. Olmo (2013). “Optimally harnessing inter-day and intra-day information for daily value-at-risk prediction”. *International Journal of Forecasting* 29 (1), 28–42 (see p. 17).

Gaglianone, W. P., L. R. Lima, O. Linton, and D. R. Smith (2011). “Evaluating Value-at-Risk Models via Quantile Regression”. *Journal of Business & Economic Statistics* 29 (1), 150–160 (see pp. 96, 111, 119).

Ghalanos, A. (2015). *rugarch: Univariate GARCH models.* R package version 1.3-6. (see p. 27).

Giacomini, R. and I. Komunjer (2005). “Evaluation and Combination of Conditional Quantile Forecasts”. *Journal of Business & Economic Statistics* 23 (4), 416–431 (see pp. 17, 19, 119).

Gikhman, I. and A. Skorokhod (2004). *The Theory of Stochastic Processes I*. Vol. 210. Classics in Mathematics. Springer Berlin Heidelberg (see pp. 73, 91).

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). “On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks”. *The Journal of Finance* 48 (5), 1779–1801 (see p. 117).

Gneiting, T. (2011a). “Making and Evaluating Point Forecasts”. *Journal of the American Statistical Association* 106 (494), 746–762 (see pp. 52, 54–56, 60).

Gneiting, T. (2011b). “Quantiles as optimal point forecasts”. *International Journal of Forecasting* 27 (2), 197–207 (see pp. 16, 19).

Gourieroux, C. and A. Monfort (1995). *Statistics and Econometric Models: Volume 1, General Concepts, Estimation, Prediction and Algorithms*. Cambridge University Press (see p. 65).

Graham, A. and J. Pál (2014). “Backtesting value-at-risk tail losses on a dynamic portfolio”. *The Journal of Risk Model Validation* 8 (2), 59 (see p. 103).

Grigoryeva, L., J.-P. Ortega, and A. Peresetsky (2017). “Volatility forecasting using global stochastic financial trends extracted from non-synchronous data”. *Forthcoming in Econometrics and Statistics*. DOI: 10.1016/j.ecosta.2017.01.003 (see p. 37).

Guler, K., P. T. Ng, and Z. Xiao (2017). “Mincer–Zarnowitz quantile and expectile regressions for forecast evaluations under aysmmetric loss functions”. *Journal of Forecasting* 36 (6), 651–679 (see p. 96).

Halbleib, R. and W. Pohlmeier (2012). “Improving the Value at Risk Forecasts: Theory and Evidence from the Financial Crisis”. *Journal of Economic Dynamics and Control* 36 (8), 1212–1228 (see pp. 15–17, 19, 35).

Hall, P. and S. J. Sheather (1988). “On the Distribution of a Studentized Quantile”. *Journal of the Royal Statistical Society. Series B (Methodological)* 50 (3), 381–391 (see p. 64).

Hamidi, B., C. Hurlin, P. Kouontchou, and B. Maillet (2015). “A DARE for VaR”. *Finance* 36 (1), 7–38 (see pp. 17, 29, 34, 36, 37).

Hansen, B. (2008). “Least-squares forecast averaging”. *Journal of Econometrics* 146 (2), 342–350 (see p. 21).

Hansen, P. R., A. Lunde, and J. M. Nason (2011). “The Model Confidence Set”. *Econometrica* 79 (2), 453–497 (see pp. 17, 18, 30, 31, 117, 118).

Hart, J. D. (1994). “Automated Kernel Smoothing of Dependent Data by Using Time Series Cross- Validation”. *Journal of the Royal Statistical Society. Series B (Methodological)* 56 (3), 529–542 (see p. 23).

Hart, J. D. and C.-L. Lee (2005). “Robustness of one-sided cross-validation to autocorrelation”. *Journal of Multivariate Analysis* 92 (1), 77–96 (see p. 23).

Hastie, T., R. Tibshirani, and J. Friedman (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer (see pp. 16, 20).

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC (see p. 21).

Hendricks, W. and R. Koenker (1992). “Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity”. *Journal of the American Statistical Association* 87 (417), 58–68 (see p. 64).

Hoerl, A. E. and R. W. Kennard (1970a). “Ridge Regression: Applications to Nonorthogonal Problems”. *Technometrics* 12 (1), 69–82 (see pp. 16, 20).

— (1970b). “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. *Technometrics* 12 (1), 55–67 (see pp. 16, 20).

Holden, K. and D. A. Peel (1990). “On Testing For Unbiasedness And Efficiency Of Forecasts”. *The Manchester School* 58 (2), 120–127 (see p. 99).

Huang, H. and T.-H. Lee (2013). “Forecasting Value-at-Risk Using High-Frequency Information”. *Econometrics* 1 (1), 127–140 (see p. 17).

Huber, P. (1967). “The behavior of maximum likelihood estimates under nonstandard conditions”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 221–233 (see pp. 73, 75, 76).

James, G. M. (2003). “Variance and Bias for General Loss Functions”. *Machine Learning* 51 (2), 115–135 (see p. 20).

Jeon, J. and J. W. Taylor (2013). “Using CAViaR Models with Implied Volatility for Value-at-Risk Estimation”. *Journal of Forecasting* 32 (1), 62–74 (see p. 17).

Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk*. 3rd ed. McGraw-Hill (see p. 15).

Kerkhof, J. and B. Melenberg (2004). “Backtesting for risk-based regulatory capital”. *Journal of Banking & Finance* 28 (8), 1845–1865 (see p. 103).

Koenker, R. (1994). “Confidence Intervals for Regression Quantiles. Proceedings of the Fifth Prague Symposium, held from September 4–9, 1993”. In: *Asymptotic Statistics*. Ed. by Mandl, P. and Hušková, M. Heidelberg: Physica-Verlag HD, 349–359 (see p. 64).

— (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press (see pp. 53, 60).

— (2011). “Additive models for quantile regression: Model selection and confidence bandaids”. *Brazilian Journal of Probability and Statistics* 25 (3), 239–262 (see p. 34).

— (2016). *quantreg: Quantile Regression*. R package version 5.29 (see p. 27).

Koenker, R. and G. Bassett (1978). “Regression Quantiles”. *Econometrica* 46 (1), 33–50 (see pp. 16, 19).

Koenker, R. and J. A. F. Machado (1999). “Goodness of Fit and Related Inference Processes for Quantile Regression”. *Journal of the American Statistical Association* 94 (448), 1296–1310 (see p. 61).

Komunjer, I. (2013). “Quantile Prediction”. In: *Handbook of Economic Forecasting*. Ed. by Elliott, G. and Timmermann, A. Vol. 2. Elsevier. Chap. 17, 961–994 (see pp. 15, 98).

Kratz, M., Y. H. Lok, and A. J. McNeil (2017). “Multinomial VaR Backtests: A simple implicit approach to backtesting expected shortfall”. arXiv:1611.04851 [q-fin.RM] (see p. 103).

Kuester, K., S. Mittnik, and M. Paolella (2006). “Value-at-Risk Prediction: A Comparison of Alternative Strategies”. *Journal of Financial Econometrics* 4 (1), 53–89 (see p. 15).

Kupiec, P. H. (1995). “Techniques for Verifying the Accuracy of Risk Measurement Models”. *The Journal of Derivatives* 3 (2), 73–84 (see pp. 18, 30, 35, 36, 98).

Lambert, N. S., D. M. Pennock, and Y. Shoham (2008). “Eliciting Properties of Probability Distributions”. In: *Proceedings of the 9th ACM Conference on Electronic Commerce*. ACM, 129–138 (see p. 54).

Lambert, P. and S. Laurent (2002). “Modelling skewness dynamics in series of financial data”. Université Catholique de Louvain and Université de Liège, available at `http://hdl.handle.net/2078.1/91035` (see p. 107).

Li, Y. and J. Zhu (2008). “L1-Norm Quantile Regression”. *Journal of Computational and Graphical Statistics* 17 (1), 163–185 (see p. 22).

Lloyd, C. J. (2005). “Estimating test power adjusted for size”. *Journal of Statistical Computation and Simulation* 75 (11), 921–933 (see pp. 109, 110).

Lourenço, H. R., O. C. Martin, and T. Stützle (2003). “Iterated Local Search”. In: *Handbook of Metaheuristics*. Ed. by Glover, F. and Kochenberger, G. A. Boston, MA: Springer US, 320–353 (see p. 63).

Louzis, D. P., S. Xanthopoulos-Sisinis, and A. P. Refenes (2014). “Realized volatility models and alternative Value-at-Risk prediction strategies”. *Economic Modelling* 40, 101–116 (see p. 15).

MacKinnon, J. G. (2009). “Bootstrap Hypothesis Testing”. In: *Handbook of Computational Econometrics*. Ed. by Belsley, D. A. and Kontoghiorghes, E. J. John Wiley & Sons, Ltd. Chap. 6, 183–213 (see p. 102).

Mallows, C. L. (1973). “Some Comments on Cp”. *Technometrics* 15 (4), 661–675 (see p. 29).

Marinelli, C., S. D’addona, and S. T. Rachev (2007). “A Comparison Of Some Univariate Models For Value-at-risk And Expected Shortfall”. *International Journal of Theoretical and Applied Finance* 10 (06), 1043–1075 (see p. 15).

McAleer, M., J.-A. Jimenez-Martin, and P.-A. Teodosio (2013a). “GFC-robust risk management strategies under the Basel Accord”. *International Review of Economics & Finance* 27, 97–111 (see pp. 17, 30).

— (2013b). “International Evidence on GFC-Robust Forecasts for Risk Management under the Basel Accord”. *Journal of Forecasting* 32 (3), 267–288 (see p. 17).

McNeil, A. J. and R. Frey (2000). “Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach”. *Journal of Empirical Finance* 7 (3–4), 271–300 (see pp. 96, 97, 103, 104, 106, 108, 110, 111, 114, 116, 118–122).

Meinshausen, N. (2006). “Quantile Regression Forests”. *Journal of Machine Learning Research* 7, 983–999 (see p. 41).

Mincer, J. and V. Zarnowitz (1969). “The Evaluation of Economic Forecasts”. In: *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*. National Bureau of Economic Research, Inc, 3–46 (see pp. 9, 13, 95, 96, 99, 118).

Nadarajah, S., B. Zhang, and S. Chan (2014). “Estimation methods for expected shortfall”. *Quantitative Finance* 14 (2), 271–291 (see pp. 52, 98).

Nelder, J. A. and R. Mead (1965). “A Simplex Method for Function Minimization”. *The Computer Journal* 7 (4), 308–313 (see p. 63).

Nelson, D. B. (1991). “Conditional Heteroskedasticity in Asset Returns: A New Approach”. *Econometrica* 59 (2), 347–370 (see p. 27).

Newey, W. and D. McFadden (1994). “Large sample estimation and hypothesis testing”. In: *Handbook of Econometrics*. Ed. by Engle, R. and McFadden, D. Vol. 4. Elsevier. Chap. 36, 2111–2245 (see pp. 73, 75).

Nieto, M. R. and E. Ruiz (2016). “Frontiers in VaR forecasting and backtesting”. *International Journal of Forecasting* 32 (2), 475–501 (see p. 15).

Nolde, N. and J. F. Ziegel (2017). “Elicitability and backtesting: Perspectives for banking regulation”. arXiv:1608.05498 [q-fin.RM] (see pp. 53, 54, 61, 62, 67, 95–97, 100, 103–108, 110, 111, 114, 116–122).

Patton, A. J., J. F. Ziegel, and R. Chen (2017). “Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)”. arXiv:1707.05108 [q-fin.EC] (see pp. 95, 100).

Politis, D. N. and J. P. Romano (1994). “The Stationary Bootstrap”. *Journal of the American Statistical Association* 89 (428), 1303–1313 (see p. 118).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. `http://www.R-project.org`. R Foundation for Statistical Computing. Vienna, Austria (see p. 21).

Righi, M. B. and P. S. Ceretta (2013). “Individual and flexible expected shortfall backtesting”. *Journal of Risk Model Validation* 7 (3) (see p. 103).

— (2015). “A comparison of Expected Shortfall estimation models”. *Journal of Economics and Business* 78, 14–47 (see p. 103).

RiskMetrics Group (1996). *RiskMetrics – Technical Document*. J. P. Morgan and Reuters. New York (see p. 26).

Shan, K. and Y. Yang (2009). “Combining Regression Quantile Estimators”. *Statistica Sinica* 19 (3), 1171–1191 (see pp. 17, 29, 30, 34).

Sheppard, K. (2017). *ARCH*. Python package version 4.0 (see p. 31).

Stock, J. H. and M. W. Watson (2004). “Combination forecasts of output growth in a seven-country data set”. *Journal of Forecasting* 23 (6), 405–430 (see p. 18).

Taylor, J. W. (2008a). “Estimating Value at Risk and Expected Shortfall Using Expectiles”. *Journal of Financial Econometrics* 6 (2), 231–252 (see pp. 17, 52).

— (2008b). “Using Exponentially Weighted Quantile Regression to Estimate Value at Risk and Expected Shortfall”. *Journal of Financial Econometrics* 6 (3), 382–406 (see p. 52).

— (2017). “Forecasting Value at Risk and Expected Shortfall Using a Semiparametric Approach Based on the Asymmetric Laplace Distribution”. *Forthcoming in Journal of Business & Economic Statistics*. DOI: 10.1080/07350015.2017.1281815 (see pp. 52, 100).

Taylor, S. J. (1986). *Modelling Financial Time Series*. World Scientific Publishing (see p. 27).

Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso”. English. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1), 267–288 (see pp. 16, 20).

Timmermann, A. (2006). “Forecast Combinations”. In: *Handbook of Economic Forecasting*. Ed. by Elliott, G., Granger, C. W., and Timmermann, A. Vol. 1. Elsevier. Chap. 4, 135–196 (see pp. 15, 21, 28, 29, 34).

Trottier, D.-A. and D. Ardia (2016). “Moments of standardized Fernandez-Steel skewed distributions: Applications to the estimation of GARCH-type models”. *Finance Research Letters* 18, 311–316 (see p. 107).

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (see pp. 73, 75, 78).

Weber, S. (2006). “Distribution Invariant Risk Measures, Information, and Dynamic Consistency”. *Mathematical Finance* 16 (2), 419–441 (see pp. 52, 55).

Wong, W. (2008). “Backtesting trading risk of commercial banks using expected shortfall”. *Journal of Banking & Finance* 32 (7), 1404–1415 (see p. 103).

Yi, C. (2017). *hqreg: Regularization Paths for Lasso or Elastic-Net Penalized Huber Loss Regression and Quantile Regression*. R package version 1.4 (see p. 21).

Yi, C. and J. Huang (2017). “Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression”. *Journal of Computational and Graphical Statistics* 26 (3), 547–557 (see p. 21).

Zangari, P. (1996). *RiskMetrics – Technical Document*. Tech. rep. New York: Morgan Guaranty Trust Company (see p. 108).

Zheng, S. (2012). “QBoost: Predicting quantiles with boosting for regression and binary classification”. *Expert Systems with Applications* 39 (2), 1687–1697 (see p. 41).

Ziegel, J. F., F. Krüger, A. Jordan, and F. Fasciati (2017). “Murphy Diagrams: Forecast Evaluation of Expected Shortfall”. arXiv:1705.04537 [q-fin.RM] (see pp. 53, 62, 67, 70).

Zou, H. and T. Hastie (2005). “Regularization and variable selection via the Elastic Net”. *Journal of the Royal Statistical Society. Series B (Methodological)* 67 (2), 301–320 (see pp. 16, 20, 41).

Zwingmann, T. and H. Holzmann (2016). “Asymptotics for the expected shortfall”. arXiv:1611.07222 [math.ST] (see pp. 53, 60).

Eigenabgrenzung

Das erste Kapitel, *Combining Value-at-Risk Forecasts Using Penalized Quantile Regressions*, habe ich selbstständig und nur mit den angegebenen Hilfsmitteln erstellt.

Das zweite und dritte Kapitel, *A Joint Quantile and Expected Shortfall Regression Framework* und *Regression Based Expected Shortfall Backtesting*, sind in Zusammenarbeit mit Timo Dimitriadis entstanden, der ebenfalls Dokotorand an der Graduate School of Decision Sciences der Universität Konstanz ist. Meine individuellen Leistungen bei der Erstellung der Kapitel betragen 40% für das zweite und 60% für das dritte Kapitel.