LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Why risk is so hard to measure

Jon Danielsson
Chen Zhou

SRC Discussion Paper No 36
April 2015

SRC Systemic Risk Centre

Discussion Paper Series

ISSN 2054-538X

Abstract
This paper analyzes the robustness of standard risk analysis techniques, with a special emphasis on the specifications in Basel III. We focus on the difference between Value–at–Risk and expected shortfall, the small sample properties of these risk measures and the impact of using an overlapping approach to construct data for longer holding periods. Overall, risk forecasts are extremely uncertain at low sample sizes. By comparing the estimation uncertainty, we find that Value–at–Risk is superior to expected shortfall and the time-scaling approach for risk forecasts with longer holding periods is preferable to using overlapping data.

Keywords: Value–at–Risk, expected shortfall, finite sample properties, Basel III.

JEL classification: C10, C15, G18

This paper is published as part of the Systemic Risk Centre's Discussion Paper Series. The support of the Economic and Social Research Council (ESRC) in funding the SRC is gratefully acknowledged [grant number ES/K002309/1].

Jon Danielsson, Systemic Risk Centre, London School of Economics and Political Science
Chen Zhou, Bank of the Netherlands and Erasmus University Rotterdam

Published by
Systemic Risk Centre
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE

Why risk is so hard to measure*

Jon Danielsson
Systemic Risk Centre
London School of Economics

Chen Zhou
Bank of the Netherlands and Erasmus University Rotterdam

April 2015

Abstract

This paper analyzes the robustness of standard risk analysis techniques, with a special emphasis on the specifications in Basel III. We focus on the difference between Value–at–Risk and expected shortfall, the small sample properties of these risk measures and the impact of using an overlapping approach to construct data for longer holding periods. Overall, risk forecasts are extremely uncertain at low sample sizes. By comparing the estimation uncertainty, we find that Value–at–Risk is superior to expected shortfall and the time-scaling approach for risk forecasts with longer holding periods is preferable to using overlapping data.

Keywords: Value–at–Risk, expected shortfall, finite sample properties, Basel III. JEL codes C10, C15, G18

*Corresponding author Chen Zhou. We thank Dong Lou, Robert Macrae and Lerby Ergun for valuable comments, and the Economic and Social Research Council (UK) [grant number: ES/K002309/1] for supporting the research. All errors are ours. Updated versions of this paper can be found on `www.RiskResearch.org` and the Webappendix for the paper is at `www.ModelsandRisk.org/VaR-and-ES`. Views expressed by Chen Zhou do not reflect the official positions of the Bank of The Netherlands.

1 Introduction

Financial risk is usually forecasted with sophisticated statistical methods. However, in spite of their prevalence in industry applications and financial regulations, the performance of such methods is generally poorly understood. This is a concern since minor variations in model assumptions can lead to vastly different risk forecasts for the same portfolio, forecasts that are all equally plausible ex–ante.

Addressing this deficiency is the main objective of our work. We compare and contrast the most common risk measures, Value–at–Risk (VaR) and expected shortfall (ES), investigate their performance characteristics in typical usage scenarios and study recent methodological proposals. The ultimate aim to analyze how these risk measures perform as well as what works best and what should be avoided in implementation.

Three main challenges arise in the forecasting of financial risk: the choice of risk measure, data choice and usage and which statistical method is used. Ever since its introduction by J.P. Morgan (1993) and especially the incorporation into financial regulations by the Basel Committee (1996), VaR is the most commonly used market risk measure. While it has come under considerable criticism, it has generally been preferred, partly because of the work of Yamai and Yoshiba (2002, 2005).

While any statistical method benefits from as much data as possible, in practice that is usually not possible because of the presence of new instruments, structural breaks, financial innovations or other aspects that reduce the relevance of old data. This practical limitation is brought into an especially sharp contrast when coupled with a desire for longer liquidity horizons as expressed in the Basel regulations.

Suppose data is observed at the daily frequency. There are three ways one can obtain multi–day holding period risk forecasts: First, estimate a daily risk forecast and apply some scaling law to obtain the multi–day forecast, typically square–root–of–time. We call this approach the *time–scaling approach*. Alternatively one can time aggregate daily observations. Focussing on log returns, a 10 day return would be the sum of 10 one–day returns, and the risk forecasts would then be made by these 10 day returns. Here we have two alternatives. We can either use non–overlapping aggregated data, or allow the aggregation periods to overlap. We term the first the *non–overlapping approach* and the second the *overlapping approach*. The Basel Committee on Banking Supervision (BSBC), as expressed in the 2013 version of the Basel III Proposals, was keen on the overlapping approach, but not in the revised

2014 version.

A large number of statistical methods for forecasting risk have been proposed, but as a practical matter, only a handful have found significant traction, as discussed in Danielsson et al. (2014). Of these, all but one depend on a parametric model, while one, historical simulation (HS), is model independent. Our objective in this paper is not to compare and contrast the various risk forecast methods: after all, a large number of high–quality papers exist on this very topic. Instead, we want to see how a representative risk forecast method performs, identifying results that are related technical choices on the other two key issues: risk measure and data.

Considering our objectives, it is appropriate to focus on HS, not only is it a commonly used method, for example, in the US bank sample of O'Brien and Szerszen (2014), 60% use HS. More fundamentally, the good performance of a specific parametric model is usually driven by the fact that the model is close to the data generating process (DGP) and it is not possible to find a parametric model that performs consistently well across all DGPs. Although HS is the simplest estimation method, it has the advantage of not being dependent on a particular parametric DGP. Its main deficiency, the poor performance in the presence of structural breaks, will not affect our main analysis, since we do not impose structural breaks in our simulation setup.

Our first contribution is the practical comparison of VaR to ES. A common view holds that VaR is inherently inferior to ES, a view supported by three convincing arguments. First, VaR is not a coherent measure unlike ES, as noted by Artzner et al. (1999). Second, as a quantile, VaR is unable to capture the risk in the tails beyond the specific probability, while ES accounts for all tail events. Finally, it is easier for financial institutions to manipulate VaR than ES. Perhaps swayed the theoretical advantages, ES appears increasingly preferred both by practitioners and regulators, most significantly expressed in Basel III. While the Proposal is light on motivation, the little that is stated only refers to theoretic tail advantages.

The practical properties of both ES and VaR are less understood, and are likely to provide conflicting signals since implementation introduces additional considerations, some of which work in the opposite direction. The estimation of ES requires more steps and more assumptions than the estimation of VaR, giving rise to more estimation uncertainty. However, ES smooths out the tails and therefore might perform better in practice.

Our second contribution is to investigate how best to use data. Ideally, the non–overlapping approach is preferred, but in practice it would likely result in excessively large data requirements, beyond what would be available

in most cases. This means that any implementation needs to depend on either the time–scaling approach or overlapping approach. From a purely theoretic point of view, the time–scaling approach is not very attractive, the common square–root–of–time approach is only correct for scaling VaR or ES for independently and identically distributed (i.i.d.) normal returns, and in practice is either higher or lower depending on the unknown underlying stochastic process. This suggests that the overlapping approach might be preferred, both because by aggregating high frequency observations we get smoother forecasts due to the smoothing of what might be seen as anomalous extreme outcomes and also when dealing with infrequent trading, high–frequency (daily) observations become unreliable.

Our purely anecdotal observation of practitioners suggests that using the overlapping approach is increasingly preferred to the scaling method. Certainly, the Basel Committee held that view in 2013. However, the overlapping approach gives rise to a particular theoretical challenge, induced dependence in the constructed dataset, and hence the potential to increase the estimation uncertainty. The pros and cons of using the overlapping approach for forecasting risk have until now been mostly conjectured, and not been supported by analytical work. While some theoretical results exist on the properties of square–root–of–time approach compared to overlapping approach, little to none exists on the impact on risk forecasts.

In our third and final contribution we study whether the estimation of risk measures is robust when considering smaller — and typical in practical use — sample sizes. Although the asymptotic properties of risk measures can be established using statistical theories, and such analysis is routinely reported, sample sizes vary substantially. This implies that the known asymptotic properties of the risk forecast estimators might be very different in typical sample sizes.

We address each of these three questions from both theoretic and empirical points of view. Ideally, one would evaluate the robustness of risk analysis with real data, but that is challenging because we do not know the DGP of the observed data and neither do we have any assurance that data across time and assets maintains consistent statistical properties. We therefore do the analysis in three stages starting with theoretic and simulation analysis, which allows us to study properties of the risk measures when we know the DGP. We then augment the analysis by results from the CRSP universe of stock returns.

Our theoretic analysis directly relates to the vast extant literature on risk measures. By contrast, it is surprising that so little Monte Carlo analysis of

the practical statistical properties of risk measures exist. We surmise that an important reason relates to computational difficulties, especially the very large simulation size needed. We are estimating not only the risk measures but also the uncertainty of those estimates, where for example, we need to capture the "quantiles of the quantiles". To achieve robust results, in the sense that they are accurate up to three significant digits, one needs to draw ten million samples, each of various sizes.

We obtain three sets of results. First, we confirm that for Gaussian and heavy tailed return distributions, which encompass the majority of asset returns, VaR and ES are related by a small constant. In the special case of Basel III, the 97.5% ES is essentially the same as the 99% VaR in the Gaussian case, while for heavy–tailed distributions, ES is somewhat larger, but not by much. As the sample size gets smaller, the 97.5% ES gets closer and closer to the the 99% VaR, falling below it at the smallest sample sizes. This suggests that even if ES is theoretically better at capturing the tails, in practice one might just multiply VaR by a small constant to get ES.

Second, ES is estimated with more uncertainty than the VaR. We find this both when we estimate each at the same 99% probability levels and also when using the Basel III combination, ES(97.5%) and VaR(99%). A sample size of half a century of daily observations is needed for the empirical estimators to achieve their asymptotic properties. At the smallest sample sizes, 500 or less, the uncertainty becomes very large, to an extent that it would be difficult to advise using the resulting risk forecasts for important decisions, especially those where the cost of type II errors is not trivial. The confidence bounds around the risk forecasts are very far from being symmetric, the upper 99% confidence bound is a multiple of the forecast, which obviously cannot be the case for the lower confidence bound. This means that if one uses the standard error as a measure of uncertainty, it will be strongly biased downwards.

In our final result, we compare the square–root–of–time approach to the overlapping approach and find that the overlapping approach results in more uncertainty.

The structure of the paper is as follows. We discuss the theoretic properties of the risk measures in Section 2 and the Monte Carlo results in Section 3. Empirical results using stock returns in the Center for Research in Security Prices (CRSP) database are presented in Section 4. The implications of out analysis are discussed in Section 5 and Section 6 concludes. Some of the mathematical derivations have been relegated to the Appendix.

2 Risk measure theory

Many authors have considered the various theoretical properties of statistical risk measures and the underlying estimation methods. Here we are interested in three particular aspects of risk measures that see little coverage in the extant literature: The relationship between ES and VaR, the impact of using the overlapping approach and the small sample properties of the risk measures. We consider both the case where the probability for VaR and ES is the same, and also the Basel III case where the comparison is between ES(97.5%) and VaR(99%).

Denote the profit and loss of a trading portfolio as PL and let $X \equiv -PL$, so we can indicate a loss by a positive number. Suppose we obtain a sample of size N, where, without loss of generality, we assume that the observations are i.i.d., with distribution F. Denote the probability by p.

We refer to VaR and ES by $q_F := q_F(p)$ and $e_F := e_F(p)$, respectively, where p is the tail probability level and estimate them by HS. Rank the N observations as $X_{N,1} \leq X_{N,2} \leq \cdots \leq X_{N,N}$. Then

$$\begin{aligned} \hat{q}_F &= X_{N,[pN]}, \\ \hat{e}_F &= \frac{1}{(1-p)N} \sum_{j=1}^{(1-p)N} X_{N,N-j+1}. \end{aligned} \tag{1}$$

Asymptotically, these estimators are unbiased, but a well–known result, dating at least back to Blom (1958), finds that quantile estimators, like HS, are biased in small samples, so that the $X_{N,[pN]}$ quantile does not correspond exactly to the p probability, instead, the probability is slightly lower. It is straightforward to adjust for this bias, using for example the methods proposed by Hyndman and Fan (1996).

2.1 VaR and ES

2.1.1 The levels of VaR and ES

Consider the ratio of ES to VaR, e_F/q_F, for a range of distribution functions F. Starting with the Gaussian, with mean zero and standard deviation σ, then $q_F = \sigma q_{N(0,1)}$ and $e_F = \sigma e_{N(0,1)}$, where $N(0,1)$ denotes the standard normal distribution. It is obvious that for the same p levels, $e_F(p)/q_F(p) > 1$

and it is straightforward to verify that in this case:

$$\lim_{p\to 1} \frac{e_F(p)}{q_F(p)} = 1.$$

In other words, as we consider increasing, but same for both, extreme probabilities, although the ES is higher than the VaR, the relative difference diminishes. At a finite level, such a ratio can be explicitly calculated, for example, ${e_F(0.99)}/{q_F(0.99)} = 1.146$. When considering different p levels for the two risk measures, such as in comparing the Basel III proposal, we get ${e_F(0.975)}/{q_F(0.99)} = 1.005$. Hence the two risk measures are roughly identical under normality.

Since financial returns exhibit heavy tails, a more realistic distribution takes that into account. Similar to the normal case, it is straightforward to calculate the ratio of ES to VaR for the Student–t distribution with any particular degrees of freedom, and probability levels. Supposing that we consider the Student–t with degrees of freedom three. Then ${e_F(0.99)}/{q_F(0.99)} = 1.54$ and ${e_F(0.975)}/{q_F(0.99)} = 1.11$.

However, we are more interested in a general expression of the relationship between VaR and ES, one that applies to most heavy–tailed distributions. To this end, we make use of Extreme Value Theory (EVT) and define a heavy–tailed distribution by regular variation. That is, we assume that

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha},$$

for some $\alpha > 0$, known as the tail index. For the Student–t distribution, the tail index equals the degrees of freedom. Note that the assumption of regular variation only applies to the right tail of F, and thus does not impose any restriction on the rest of the distribution, allowing this approach to capture a large range of models. Indeed, an assumption of on regular variation is sufficient for inference on tail risk measures.

The following proposition gives the theoretical foundation on comparing the levels of VaR and ES at high probability levels. For Proof see the Appendix.

Proposition 1 *Suppose F is a heavy–tailed distribution with tail index α. Given any constant $c > 0$, we have that*

$$\lim_{s\to 0} \frac{e_F(1 - cs)}{q_F(1 - s)} = \frac{\alpha}{\alpha - 1} c^{-1/\alpha}.$$

To compare the VaR and ES with the same probability level, one can take $c = 1$ and $s = 1 - p$ in Proposition 1, and get that

$$\lim_{p \to 1} \frac{e_F(p)}{q_F(p)} = \frac{\alpha}{\alpha - 1},$$

that is, when the probability is the same for both risk measures, the ES is equivalent to the VaR times the multiplier $\alpha/(\alpha-1)$. This ratio is higher than one, which gives the essential difference between heavy–tailed distributions and thin–tailed distributions such as the normal distribution. Since the multiplier is decreasing in α, the more heavy–tailed the distribution of F, the larger the difference between ES and VaR.

To compare the VaR(99%) with ES(97.5%), set c and s in Proposition 1 such that $s = 1\%$ and $cs = 2.5\%$. Then;

$$\frac{e_F(p_2)}{q_F(p_1)} \approx \frac{\alpha}{\alpha - 1}(2.5)^{-1/\alpha} := f(\alpha).$$

That is, when comparing ES(97.5%) to VaR(99%), the ratio is given by the function $f(\alpha)$. We plot this function in Figure 1 for α ranging from 2 to 5, which is more than wide enough to cover tail thicknesses commonly observed. Note the ratio is decreasing in α, ranging between 1.105 and 1.041 for α ranging from 3 to 5.

2.1.2 The estimation uncertainty of VaR and ES

In what follows, we focus our attention on the best case scenario where the data is i.i.d. and we know it is i.i.d. If we also had to estimate the dynamic structure, the estimation uncertainty would be further increased. We focus our attention on the case where F is heavy–tailed with a tail index α, with the Gaussian as the special case where $\alpha = +\infty$.

We only consider the HS method, and derive the asymptotic properties of two estimators, $\hat{q}_F$ and $\hat{e}_F$, as given in (1). In HS estimation, only the top $(1-p)N$ order statistics are used.

Denote the number of observations used in estimators (1) as $k_q := k_q(N)$ and $k_e := k_e(N)$, such that $k_q, k_e \to \infty$, $k_q/N \to 0$, $k_e/N \to 0$ as $N \to \infty$. We can then generalize (1) by defining the ES and VaR as:

$$\hat{q}_F(1 - k_q/N) = X_{N,N-k_q},$$

$$\hat{e}_F(1 - k_e/N) = \frac{1}{k_e} \sum_{j=1}^{k_e} X_{N,N-j+1}.$$

The following proposition gives the asymptotic properties of these estimators for a general k sequence. For Proof see the Appendix.

Proposition 2 *Suppose that $X_1, \cdots, X_N$ are i.i.d. and drawn from a heavy tailed distribution function F with $\alpha > 2$. Denote $U = \left(1/(1-F)\right)^{\leftarrow}$ as the quantile function. Assume the usual second order condition holds:*

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^{1/\alpha}}{A(t)} = x^{1/\alpha}\frac{x^{\rho}-1}{\rho},$$

for a constant $\rho \leq 0$ and a function $A(t)$ such that $\lim_{t\to\infty} A(t) = 0$. Suppose $k := k(N)$ is an intermediate sequence such that as $N \to \infty$, $k \to \infty$, $k/N \to 0$ and $\sqrt{k}A(N/k) \to \lambda$ with a constant λ. Then, we have that as $N \to \infty$,

$$\sqrt{k}\left(\frac{\hat{q}_F(1-k/N)}{q_F(1-k/N)} - 1\right) \xrightarrow{d} N\left(0, \frac{1}{\alpha^2}\right),$$
$$\sqrt{k}\left(\frac{\hat{e}_F(1-k/N)}{e_F(1-k/N)} - 1\right) \xrightarrow{d} N\left(0, \frac{2(\alpha-1)}{\alpha^2(\alpha-2)}\right).$$

From Proposition 2, both estimators are asymptotically unbiased. We focus on comparing their asymptotic variances.

First, we consider the case in which the ES and VaR probability is the same. In that case, $k_e = k_q = (1-p)N$. Consequently, we get that

$$\frac{\mathrm{Var}\left(\frac{\hat{e}_F(p)}{e_F(p)}\right)}{\mathrm{Var}\left(\frac{\hat{q}_F(p)}{q_F(p)}\right)} \approx \frac{\frac{2(\alpha-1)}{\alpha^2(\alpha-2)}\frac{1}{k_e}}{\frac{1}{\alpha^2}\frac{1}{k_q}} = \frac{2(\alpha-1)}{\alpha-2} = 1 + \frac{\alpha}{\alpha-2}.$$

Which means that when considering the same probability level, the relative estimation uncertainty of the ES measure is higher than that of the VaR measure. The difference is larger for lower α, i.e. heavier distributions.

Next, we compare the estimation uncertainty of VaR(99%) and ES(97.5%). In this case, we need to set $k_e/k_q = (1-p_2)/(1-p_1) = 2.5$, which reflects the relative difference in the two tail probabilities. By applying the Proposition with k_q and k_e such that $k_e/k_q = 2.5$, we get that

$$\frac{\mathrm{Var}\left(\frac{\hat{e}_F(p_2)}{e_F(p_2)}\right)}{\mathrm{Var}\left(\frac{\hat{q}_F(p_1)}{q_F(p_1)}\right)} \approx \frac{\frac{2(\alpha-1)}{\alpha^2(\alpha-2)}\frac{1}{k_e}}{\frac{1}{\alpha^2}\frac{1}{k_q}} = \frac{4(\alpha-1)}{5(\alpha-2)} =: g(\alpha).$$

The function $g(\alpha)$ is decreasing in α. By solving $g(\alpha) = 1$, we get the break–even point at $\alpha^{\text{be}} = 6$. For $\alpha > 6$, $g(\alpha) < 1$; if $\alpha < 6$, then $g(\alpha) > 1$.

That means, if the losses are heavy–tailed with $\alpha < 6$, the estimation uncertainty in ES(97.5%) is higher than that of VaR(99%).

2.2 Overlapping and time–scaling approaches

Consider the problem of forecasting risk over holding periods longer than one day, say a H days. We have three main choices in risk forecasting: using H–day returns from the non–overlapping approach, using the time–scaling approach for daily risk forecasts, typically $\sqrt{H}$, or using the overlapping approach to construct H–day returns.

Each approach has its own pros and cons, the first is the most accurate, but likely to result in unreasonable data requirements. The second is only strictly correct when the returns are i.i.d. normal, while the last induces serial dependence in the constructed H–day returns. If one has sufficient data, the first should always be used. In the absence of that one has to choose between the latter two. Hence, in this Section, we focus on comparing the overlapping approach and time–scaling approach.

Suppose $Y_1, Y_2, \cdots, Y_{N+H-1}$ are i.i.d. daily observations with the common distribution function G. We can then define the two alternatives by;

The overlapping approach Calculate overlapping observations by

$$Z_i = \sum_{j=1}^{H} Y_{i+j-1}$$

and use $Z_1, \cdots, Z_N$ in estimation with (1). Denote the distribution of Z_i by F;

The square–root–of–time approach Use $Y_1, \cdots, Y_N$, to estimate q_G by $\hat{q}_G$ from (1). Then we estimate q_F by $\sqrt{H}\hat{q}_G$.

The number of observations used in these approaches is $N + H - 1$ and N, respectively so the required sample sizes are comparable. The overlapping approach provides a direct estimate of q_F, while the time scaling approach only provides an estimate of $\sqrt{H}q_G$, which is an approximation of q_F. In practice, this approximation turns to be an exact relation if F is i.i.d. normal, and slightly too high for i.i.d. heavy–tailed distributions.

Consider the overlapping approach. In this case, the $H-$day observations $Z_1, \cdots, Z_N$ are not independent but exhibit a moving average. Clearly, if G is Gaussian, so is F. If G is heavy–tailed with tail index α, F will also be heavy–tailed with tail index α; see Feller (1971).

Consider the estimation uncertainty of risk measures based on dependent observations. The following Proposition is parallel to Proposition 2.

Proposition 3 *Suppose $Z_1, \cdots, Z_N$ are dependent observations defined as $Z_i = \sum_{j=1}^{H} Y_{i+j-1}$, where $Y_1, Y_2, \cdots, Y_{N+H-1}$ are i.i.d. observations from a heavy tailed distribution function with $\alpha > 2$. Under the same notations and conditions as in Proposition 2, we have that as $N \to \infty$,*

$$\sqrt{k}\left(\frac{\hat{q}_F(1-k/N)}{q_F(1-k/N)} - 1\right) \xrightarrow{d} N\left(0, H\frac{1}{\alpha^2}\right),$$
$$\sqrt{k}\left(\frac{\hat{e}_F(1-k/N)}{e_F(1-k/N)} - 1\right) \xrightarrow{d} N\left(0, H\frac{2(\alpha-1)}{\alpha^2(\alpha-2)}\right).$$

Proposition 3 shows that using overlapping data enlarges the estimation variance for the estimators on both VaR and ES by a factor proportional to H, leading to the following corollary on the comparison of estimation variance across the strategies.

Corollary 4 *As $N \to \infty$, for a given k sequence satisfying the conditions in Proposition 3, we have that*

$$\operatorname{Var}\left(\frac{\hat{q}_F(k/N)}{q_F(k/N)}\right) \sim H \operatorname{Var}\left(\frac{\sqrt{H}\hat{q}_G(k/N)}{\sqrt{H}q_G(k/N)}\right).$$

A similar results holds for ES.

To conclude, for both risk measures, given our assumptions, the overlapping approach will result in a standard deviation that is $\sqrt{H}$ times higher as the standard deviation using the square–root–of–time approach. For non– i.i.d. data it is necesary to use simulations to compare the two cases, and we do that in the next Section.

3 Simulation study

While the theoretic results in Section 2 provide guidance as to the asymptotic performance of the estimators on VaR and ES, in typical sample sizes the

asymptotics may not yet hold. For that reason, it is of interest to investigate the properties of the estimators for a range of sample sizes that might be encountered in practical applications, and we do that by means of an extensive simulation study.

Though, it would be straightforward to consider other probability levels, for the remainder of this Section, we focus on presenting results from the Basel II and III probabilities, and so we can omit the probability from the notation, unless otherwise indicated. Hence VaR means VaR(99%) and ES ES(97.5%) below. We report the full set of results for both probabilities in the the web appendix at `www.ModelsandRisk.org/VaR-and-ES`.

In each simulated sample, the sample sizes, N, range from 100 to 12,500 or more, in intervals of 100, or more. For presentation purposes, we convert sample sizes above 300 into number of years with a year consisting of 250 observations, so a sample size at 250,000 corresponds to 1,000 years. In the interest of tractability, below we focus on a representative subset of the sample sizes, with full results reported in the web appendix.

We consider two fat tailed distributions, the Student–t and the Pareto. The results from both are qualitatively similar, so in what follows we focus on the Student–t, leaving the Pareto to the web appendix.

We forecast risk by HS as in (1). Although it is well known that the HS estimator is slightly biased, (see e.g. Blom, 1958), since we focus on comparing the estimation uncertainty, we do not need to adjust the bias by the methods proposed by Hyndman and Fan (1996). Another concern is that the mean of HS estimates across multiple draws is also biased (see e.g. Danielsson et al., 2015). Again since our main interest is in the uncertainty and not the bias, a bias adjustment is not necessary.

3.1 The number of simulations

The simulations are used not only to obtain estimates of the risk measures, but more importantly the uncertainty of those estimates. This means that in practice we aim to capture the quantiles of the quantiles. Our somewhat ad hock criteria for the results is that they are accurate for at least three significant digits, and as it turns out it requires at least $S = 10^7$ simulations. For the largest sample sizes, we are then generating $S \times N = 10^7 \times 2.5 \times 10^5 = 2.5 \times 10^{12}$ random numbers, and for each sequence need to find a quantile and a mean.

Why is such a large simulation necessary? Taking the VaR measure as an ex-

ample, from each sample, we obtain one simulated quantity $\hat{q}_F/q_F - 1$. Across S simulated samples, we obtain S such ratios denoted as $r_1, r_2, \cdots, r_S$. They are regarded as i.i.d. observations from the distribution of $\hat{q}_F/q_F$, denoted as F_R. Since we intend to obtain the 99% confidence interval of this ratio, $[F_R^{-1}(0.005), F_R^{-1}(0.995)]$, we take the $[0.005S]$-th and $[0.995S]$-th order statistics among $r_1, \cdots, r_S$, $r_{S,[0.0005S]}$ and $r_{S,[0.995S]}$ to be the estimates of the lower and upper bounds respectively. For the lower bound, following Theorem 2 in Mosteller (1946), we get that as $S \to \infty$,

$$\sqrt{S}\left(\frac{r_{S,[0.0005S]}}{F_R^{-1}(0.005)} - 1\right) \xrightarrow{d} N\left(0, \frac{0.0005 \cdot (1 - 0.0005)}{\left(F_R^{-1}(0.005)\right)^2 f_R^2(F_R^{-1}(0.005))}\right),$$

where f_R is the density function of F_R. Following Proposition 2, the distribution F_R can be approximated by a normal distribution with a given standard deviation σ_N. Using this approximated distribution, we can calculate the asymptotic variance of $r_{S,[0.0005S]}/F_R^{-1}(0.005)$ as

$$\sigma_R^2 = \frac{0.005 \cdot (1 - 0.005)}{\left(\sigma_N \Phi^{-1}(0.005)\right)^2 \left(\frac{1}{\sigma_N}\phi\left(\frac{\sigma_N \Phi^{-1}(0.005)}{\sigma_N}\right)\right)^2} = 3.586.$$

Note that this variance is independent of σ_N. Therefore this result can be applied to any estimator that possesses asymptotic normality.

To ensure that the relative error between our simulated lower bound $r_{S,[0.0005S]}$ and the actual lower bound $F_R^{-1}(0.005)$ is less than 0.001 with a confidence level of 95%, the restriction requires a minimum S such that

$$S \geq \sigma_R^2 * \left(\frac{\Phi^{-1}(0.975)}{0.001}\right)^2 = 1.378 \times 10^7.$$

A minimum of $S = 10^7$ samples is necessary for our simulation study and that is the number of simulated samples we use throughout this section.

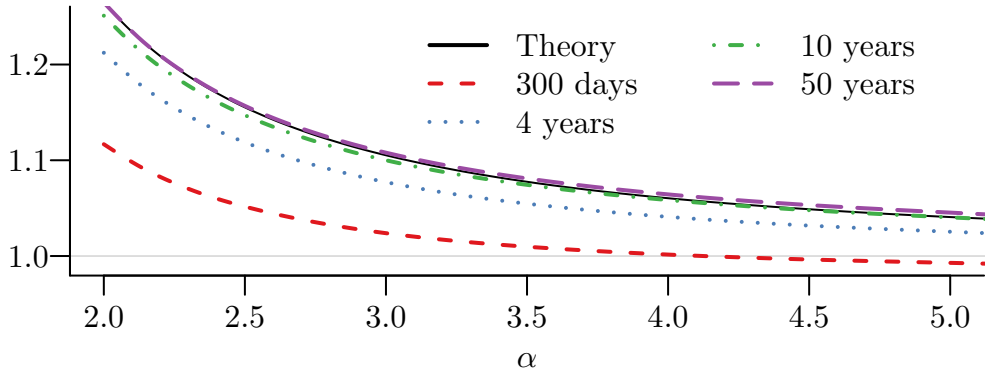
3.2 Level comparison of the risk measures

The theoretic results in Section 2.1.1 indicate that the relative difference between VaR and ES is small for distributions that do not have very heavy tails, where the difference is inversely related to the tail thickness. We explore this relation by Monte Carlo simulations with a finite sample size. For each given α, we simulate observations from standard Student–t distribution with degree of freedom $\nu = \alpha$, and for each simulated sample, we calculate the

ratio between the two estimators $\hat{e}_F$ and $\hat{q}_F$. Such a procedure is repeated S times for each given sample sizes. We plot the average ratio across different simulated samples with respect to the variation of α, reported in Figure 1.

Figure 1: Ratio of ES(97.5%) to VaR(99%)

The number of simulations is $S = 10^7$. The figure shows the ratio of ES(97.5%) to VaR(99%) for a range of sample sizes.



The solid line shows the theoretical level of the ratio, the $f(\alpha)$ function in Section 2.1.1, declining towards one as the tails become progressively thinner. The same decline is observed for every sample size. The results for the larger sample sizes, 10 and 50 years, are increasingly close to the theoretic results, and at 50 years, virtually the same.

As the sample size decreases, the relative difference between ES and VaR decreases sharply, especially for the heavier tails. For example, while asymptotic theory suggests that for $\alpha = 3$, ES is 11% larger than VaR, at 300 days it is only 3% and 8% at 4 years. At the smallest sample sizes, for tails that are slightly thinner than is usual, the ES falls below the VaR.

3.3 Estimation accuracy

The asymptotic results in Section 2.1.2 show that ES is estimated more precisely than VaR for relatively thin distributions, and less precisely for the fatter and more typical distributions, with the break even point at $\alpha = 6$. Below we investigate this results further for finite samples.

For each given α, we simulate N observations from a standard Student–t distribution with degrees of freedom $\nu = \alpha$, where N varies from 100 to 125,000. For each simulated sample, we obtain the two estimates $\hat{e}_F$ and $\hat{q}_F$ and calculate the relative estimation error as the ratio between the

estimates and their corresponding true values, e_F and q_F. Such a procedure is repeated S times for each given sample size. We then report the mean and standard error of the estimation errors, as well as the 99% empirical confidence interval, corresponding to the 0.5% and 99.5% quantiles from the S simulated estimation errors, respectively. Table 1 gives the summary information for various sample sizes and tail thicknesses.

Table 1: Comparison of the estimation accuracy

		VaR(99%)			ES(97.5%)		
α	Sample size	bias	se	99% conf	bias	se	99% conf
2.5	300 days	1.11	(0.33)	[0.61,2.46]	1.01	(0.38)	[0.54,2.64]
2.5	2 years	1.06	(0.22)	[0.67,1.89]	1.01	(0.29)	[0.61,2.20]
2.5	10 years	1.01	(0.09)	[0.82,1.28]	1.00	(0.13)	[0.78,1.48]
2.5	50 years	1.00	(0.04)	[0.91,1.11]	1.00	(0.06)	[0.89,1.19]
3	300 days	1.09	(0.27)	[0.65,2.16]	1.01	(0.27)	[0.60,2.14]
3	2 years	1.05	(0.19)	[0.70,1.73]	1.00	(0.21)	[0.66,1.82]
3	10 years	1.01	(0.08)	[0.84,1.23]	1.00	(0.09)	[0.82,1.31]
3	50 years	1.00	(0.03)	[0.92,1.09]	1.00	(0.04)	[0.91,1.12]
4	300 days	1.07	(0.21)	[0.69,1.85]	1.00	(0.19)	[0.66,1.73]
4	2 years	1.04	(0.15)	[0.74,1.55]	1.00	(0.15)	[0.72,1.52]
4	10 years	1.01	(0.06)	[0.86,1.19]	1.00	(0.06)	[0.86,1.20]
4	50 years	1.00	(0.03)	[0.93,1.08]	1.00	(0.03)	[0.93,1.08]

Note: For each given α and sample size N, $S = 10^7$ observations from a standard Student–t distribution with degree of freedom $\nu = \alpha$ are simulated. For each simulated sample, the ES and VaR are estimated and then divided by their corresponding true values. The resulting ratio is regarded as the relative estimation error. The table reports the bias (mean), standard error and 0.5% and 99.5% quantiles of these ratios across the simulated samples. The two quantiles are reported as the lower and upper bounds of the 99% confidence interval. In comparing across the two risk measures, the red values indicates whose with the higher standard error.

We obtain three main results: First, the Monte Carlo results are consistent with the theoretic result in Proposition 2, i.e. ES is estimated with more uncertainty than VaR. This simulation results show that the only exception occurs at the very small sample size combined with a higher α.

Second, the estimation bias increases as the sample size becomes smaller. This is expected given the HS bias of Blom (1958) and Monte Carlo bias of Danielsson et al. (2015). It also follows that the use of ES will partly offset

the HS bias.

Finally, the empirical confidence bounds indicate that the estimation errors are highly positively skewed, especially for the small sample sizes. For example, at $N = 300$ and $\alpha = 2.5$, the 99% confidence interval for VaR ranges from about 61% to 246% of the true value. Even for an uncommonly large 10–year sample, the confidence bound is [0.82,1.28]. For ES(97.5%), the confidence bounds are wider at [0.54,2.64] and [0.78,1.48], respectively.

3.4 The overlapping approach and the time–scaling approach

The theoretic results in Section 2.2 provided insights into the impact of using overlapping estimation or time–scaling to obtain multi–day holding period risk forecasts. We further extend those results by means of Monte Carlo simulations, both investigating the final example properties but also examining the impact of using dependent data. Below we only report a subset of the results for VaR, as the results for ES were qualitatively similar, with the full results available in the web Appendix.

For each given distribution and H, we simulate N daily observations, $S = 10^7$ times, varying N from 300 to 12,500 (50 years). For each simulated sample, we estimate the H–day holding period VaR using both the time–scaling and overlapping date approaches. Similar to the above, we divide the estimates by the true values. Since we estimate the VaR of the H–day holding period, which is not analytically tractable, we have to rely on simulation results with very large data samples to obtain the true values. We consider $H = 10$ and $H = 50$.

3.4.1 Data generating process: The i.i.d. case

We start with the i.i.d. case and report the results in Table 2, which is similar to Table 1, with the addition of two columns that show the ratios of the se and the width of the confidence interval, for the overlapping approach over the square–root–of–time approach.

The i.i.d. simulation results are consistent with those predicted by Proposition 3 in that time–scaling results in better estimation accuracy than the overlapping destination. Both the standard errors and the width of the confidence intervals for the overlapping approach are much higher than that of the time–scaling approach, ranging from 1.3 to 3.9 times larger.

Table 2: Impact of overlapping data on VaR: Student–t

(a) H=10

N	α	overlapping approach			square–root–of–time approach			Ratios of overlap to scaling	
		mean	se	99% conf	mean	se	99% conf	se	range
300 days	2.5	1.01	(0.81)	[0.43,4.48]	1.13	(0.33)	[0.62,2.49]	2.5	2.2
2 years	2.5	1.12	(0.88)	[0.50,5.31]	1.08	(0.23)	[0.68,1.91]	3.8	3.9
10 years	2.5	1.03	(0.21)	[0.70,1.92]	1.03	(0.089)	[0.83,1.29]	2.4	2.7
50 years	2.5	1.00	(0.080)	[0.84,1.26]	1.02	(0.039)	[0.92,1.12]	2.1	2.1
300 days	3.0	1.00	(0.49)	[0.48,3.25]	1.16	(0.29)	[0.69,2.30]	1.7	1.7
2 years	3.0	1.06	(0.54)	[0.56,3.65]	1.12	(0.20)	[0.75,1.84]	2.7	2.8
10 years	3.0	1.01	(0.15)	[0.75,1.61]	1.08	(0.081)	[0.90,1.32]	1.9	2.0
50 years	3.0	1.00	(0.060)	[0.87,1.19]	1.07	(0.035)	[0.98,1.17]	1.7	1.7
300 days	4.0	0.98	(0.29)	[0.53,2.23]	1.17	(0.23)	[0.75,2.02]	1.3	1.3
2 years	4.0	1.02	(0.28)	[0.61,2.32]	1.13	(0.17)	[0.81,1.70]	1.6	1.9
10 years	4.0	1.00	(0.10)	[0.80,1.35]	1.10	(0.068)	[0.94,1.30]	1.5	1.5
50 years	4.0	1.00	(0.043)	[0.90,1.13]	1.09	(0.030)	[1.02,1.17]	1.4	1.5

(b) H=50

N	α	overlapping approach			square–root–of–time approach			Ratios of overlap to scaling	
		mean	se	99% conf	mean	se	99% conf	se	range
300 days	2.50	0.72	(0.43)	[0.15,2.44]	1.15	(0.34)	[0.63,2.53]	1.3	1.2
2 years	2.50	0.81	(0.46)	[0.27,2.79]	1.09	(0.23)	[0.69,1.94]	2.0	2.0
10 years	2.50	1.08	(0.77)	[0.56,4.55]	1.04	(0.090)	[0.84,1.31]	8.6	8.5
50 years	2.50	1.01	(0.16)	[0.75,1.69]	1.03	(0.039)	[0.94,1.14]	4.1	4.7
300 days	3.00	0.76	(0.32)	[0.17,1.95]	1.20	(0.30)	[0.71,2.37]	1.1	1.1
2 years	3.00	0.84	(0.32)	[0.30,2.12]	1.15	(0.21)	[0.77,1.89]	1.5	1.6
10 years	3.00	1.02	(0.39)	[0.61,2.89]	1.11	(0.083)	[0.92,1.35]	4.7	5.3
50 years	3.00	1.00	(0.11)	[0.80,1.39]	1.10	(0.036)	[1.01,1.20]	3.1	3.1
300 days	4.00	0.78	(0.27)	[0.18,1.65]	1.20	(0.24)	[0.77,2.08]	1.1	1.1
2 years	4.00	0.86	(0.25)	[0.33,1.70]	1.16	(0.17)	[0.83,1.74]	1.5	1.5
10 years	4.00	0.99	(0.19)	[0.65,1.74]	1.13	(0.070)	[0.97,1.33]	2.7	3.0
50 years	4.00	1.00	(0.076)	[0.83,1.23]	1.12	(0.031)	[1.05,1.21]	2.5	2.5

Note: For each given α and holding period H, N daily observations from standard Student–t distribution with degree of freedom $\nu = \alpha$ are simulated with a year consisting of 250 days. For each simulated sample, the VaR of H-day holding period is estimated using the two strategies in Section 2.2 separately, and then divided by the corresponding true value obtained from pre-simulations. The resulting ratio is regarded as the relative estimation error. The table reports the mean, standard error and 0.5% and 99.5% quantiles of these ratios across S simulated samples with $S = 10^7$. The two quantiles are reported as the lower and upper bounds of the 99% confidence interval. The last two columns show the ratios of the se and the width of the confidence interval, for the overlapping approach over the square–root–of–time approach.

The bias and uncertainty for the overlapping approach first increases and then decreases as the sample size increases, something not observed for the time–scaling approach. We surmise that this happens because as N increases from 300 to 1,000, so does the probability of having very large daily losses. These daily losses will persist in H–day losses for H days, which is a significant fraction of the sample of H–day losses. This reduces the estimation accuracy. Eventually, as the N increases further, we move away from the scenario that the persistent large H–day losses can be regarded as a large fraction of the sample. Therefore, the sample size effect starts to perform. This implies that the overlapping approach performs the worst when used for typical sample sizes, such as two years.

3.4.2 Data generating process: Dependent data

Table 3: Impact of overlapping data on VaR, t-GARCH(0.01, 0.04, 0.94, 6.0)

(a) H=10

N	overlapping approach			square–root–of–time approach			Ratios of overlap to scaling	
	mean	se	99% conf	mean	se	99% conf	se	range
300 days	1.01	(0.33)	[0.49,2.38]	1.14	(0.29)	[0.68,2.33]	1.1	1.1
2 years	1.02	(0.29)	[0.57,2.28]	1.11	(0.22)	[0.72,2.00]	1.3	1.3
10 years	1.01	(0.14)	[0.75,1.52]	1.06	(0.100)	[0.86,1.40]	1.4	1.4
50 years	1.00	(0.059)	[0.87,1.18]	1.05	(0.044)	[0.95,1.18]	1.3	1.3

(b) H=50

N	overlapping approach			square–root–of–time approach			Ratios of overlap to scaling	
	mean	se	99% conf	mean	se	99% conf	se	range
300 days	0.78	(0.32)	[0.17,2.00]	1.16	(0.29)	[0.69,2.37]	1.1	1.1
2 years	0.85	(0.31)	[0.30,2.05]	1.12	(0.23)	[0.74,2.02]	1.3	1.4
10 years	0.99	(0.24)	[0.60,2.00]	1.08	(0.10)	[0.87,1.42]	2.4	2.5
50 years	1.00	(0.10)	[0.79,1.33]	1.07	(0.044)	[0.96,1.20]	2.3	2.3

Note: For each holding period H, N daily observations from the GARCH model (2) are simulated with a year consisting of 250 days. For each simulated sample, the VaR of H–day holding period is estimated using the two strategies in Section 2.2 separately, and then divided by the corresponding true value obtained from pre-simulations. The resulting ratio is regarded as the relative estimation error. The table reports the mean standard error and 0.5% and 99.5% quantiles of these ratios across S simulated samples with $S = 10^7$. The two quantiles are reported as the lower and upper bounds of the 99% confidence interval. The last two columns show the ratios of the se and the width of the confidence interval, for the overlapping approach over the square–root–of–time approach.

The overlapping approach induces serial dependence and is therefore likely to be especially sensitive to the inherent dependence of the data. We therefore also explore the impact of simulating from dependent data, using a specification that both captures the fat tails and dependence. There are many different ways one could specify such a model. A commonly used specification would be a normal GARCH model, but such a model would not adequately capture the tails, (see e.g. Sun and Zhou, 2014) and we therefore opted for a GARCH model with Student–t innovations. We parameterized the simulation model by estimating the same specification for a number of stocks and picking a typical set of parameters. In particular:

$$\begin{cases} X_t &= \sigma_t \varepsilon_t; \\ \sigma_t^2 &= 0.01 + 0.94\sigma_{t-1}^2 + 0.04X_{t-1}^2; \end{cases} \tag{2}$$

where ε_t are i.i.d. innovation terms following a standardized Student–t distribution with degree of freedom 6 and unit variance.

Table 3 reports the result based daily observations generated from the GARCH model. Notice that due to the serial dependence in the GARCH model, our theoretical result in Proposition 3 may not hold. Therefore, we have to rely on the simulation result for comparing the two approaches.

Compared to the i.i.d. case, the time–scaling approach results in even lower standard errors than the overlapping approach. In addition, there are two important differences between the i.i.d. and dependent cases for the overlapping approach. First, in the dependent case, the standard errors and biases decrease as N increases. Second, for $H = 50$, and N less than 10 years, there is a downward bias, i.e. the estimates are lower than the true value. The bias can be around 20% for low values of N.

The first difference provides some support for using the overlapping approach for dependent data, even though the time–scaling approach still performs better in terms of estimation accuracy. This benefit is counteracted by the second observation, where, for example, from a viewpoint of a prudential regulator, the lower bound of the confidence interval based on overlapping approach is much lower than that based on time–scaling approach.

4 Empirical results

While the theoretic and simulation results above provide a clear picture of the relative properties of various risk measures, they were obtained under particular distributional assumptions. To augment those results, we also

employ observed returns from CRSP. The downside is that since we do not know the true value of the risk measures, we cannot directly validate the estimation uncertainty, but can approximated by means of a block bootstrap procedure.[1]

Our data sets consists of daily returns on all stock prices traded on NASDAQ, NYSE or AMSE from 1926 to 2014. We removed illiquid stocks[2] as well as those with less than 650 observations.[3] This filtering procedure results in 7,686 stocks. For each stock, we split the time series into non–overlapping samples with sample sizes N and then pool all samples from different stocks together. The sample sizes in this analysis are $N = 600$, 1000 and 5000 resulting in 34,244, 20,097 and 2,503 samples for each of the three sample sizes, respectively.

Similar to the simulation analysis, the probability levels are 99% for VaR and 97.5% for ES. For expositional expediency, we therefore drop a reference to the probabilities from the notation.

In line with the theoretic and simulation analysis, the real data study consists of three parts: a comparison of the levels of VaR and ES, the relative estimation uncertainty between VaR and ES and overlapping and time-scaling approaches.

4.1 Level comparison

To compare the levels of VaR and ES, we calculate the ratio $^{\text{ES}}/_{\text{VaR}}$ for each sample and report the cross-sectional mean, median and standard error. In addition, we report the fraction of samples with a ratio above one in the row ratio. Finally, we do a t–test to test the mean equaling one across all samples, and report the p–value in the row p–value. All results are in Column 4 of Table 4.

We observe that the level of ES is slightly but statistically significantly higher

[1]The block size needs to be large enough to capture the inherent dependence in the data, and we opted for 200 days.

[2]The liquidity criterion is related to the sample splitting procedure. A non–overlapping sample of one stock is included in our analysis if on the first day of the sample the stock has a share price above 5$ and market capitalization is higher than the 10% quantile of the market capitalization of all stocks traded on NYSE on that day.

[3]The minimum sample size is determined at $N + H$, where N is the minimum sample size 600 and H is the longest liquidity horizon in our analysis 50. The reason to choose the minimum sample size at $N = 600$ instead of 500 is due to the block bootstrapping procedure: the sample size is required to be a multiple of the block size, 200.

than that of VaR for each of the sample sizes respectively. However, the economic significance is rather weak, especially for the smaller and more typical sample sizes such as $N = 600$.

Following the simulation results in Figure 1, the ratio for large sample sizes is close to its theoretical value $f(\alpha)$. By taking the mean ratio 1.06 under $N = 5,000$, we invert the relation $f(\alpha) = 1.06$ to get that $\alpha^* = 4.01$. This result is in line with the estimated tail index of stock returns in literature, see, e.g. Jansen and de Vries (1991).

4.2 Estimation accuracy

In order to compare the estimation uncertainty of VaR and ES, we start with calculating the *variation coefficient ratio* (VCR) between ES and VaR, for each sample. The VCR between two risk measures φ_1 and φ_2 is defined as

$$\mathrm{VCR}(\varphi_1, \varphi_2) = \frac{\sigma(\varphi_1)/\hat{\varphi_1}}{\sigma(\varphi_2)/\hat{\varphi_2}},$$

where $\hat{\varphi}_i$ is the point estimate for φ_i and $\sigma(\varphi_i)$ is the standard error of the estimation obtained from block bootstrapping, $i = 1, 2$.

The block bootstrapping procedure is as follows. We randomly draw a number of blocks consisting of consecutive observations with a block size $B = 200$ from the given sample. With N/B blocks, we construct a bootstrapped sample with sample size N. For each bootstrapped sample j, we get the point estimate of φ_i as $\hat{\varphi_i}^{(j)}$, where $j = 1, 2, \cdots, K$. Here the number of replication $K = 5,000$. Then we calculate the standard error $\sigma(\varphi_i)$ as the sample standard deviation among the K bootstrapped estimates.

From the theoretic analysis of i.i.d. data, one would expect that the VCR (ES,VaR) exceeds one. However, the results in Column 5 in Table 4 show that the VCR (ES,VaR) exceeds one only for the largest sample size $N = 5,000$ so the empirical results only partially support the theoretic results. This is comparable with some exceptional results in the simulation study, see Table 1, the last panel with $\alpha = 4$.

To further explore the economic impact of estimation uncertainty, we compare the lower bound of the 99% confidence interval when estimating the two risk measures. For each given sample, using the bootstrapping procedure, we take the $0.005K$–th and $0.995K$–th quantiles among the K bootstrapped estimates of VaR to construct the 99% confidence interval for the VaR estimate. We focus on the lower bound and denote it as l(VaR). Similarly, we obtain

l(ES) Then, we calculate the ratio between the standardized lower bounds as

$$\mathrm{Q(VaR,ES)} = \frac{\iota(\mathrm{VaR})/\widehat{\mathrm{VaR}}}{\iota(\mathrm{ES})/\widehat{\mathrm{ES}}}$$

We report the outcome based on these ratios in the 5th column of Table 4. The lower bound of VaR is significantly higher than that of ES across all sample sizes, in line with our simulation results in Table 1.

4.3 The overlapping approach and the time–scaling approach

Finally, we compare the overlapping approach and the time–scaling approach, in particular the square–root–of–time approach. The notation VCR (VaR $H10$) is the VCR is between the VaR measures using the overlapping approach and square–root–of–time approach, and similarly for ES and $H = 50$. The results are reported in the last four columns of Table 4.

We find strong evidence that all four VCRs are significantly above one. These results are in line with our qualitative conclusion drawn from theoretical analysis and simulations. The average VCR for $H = 10$ is below $\sqrt{10}$ and that for $H = 50$ is below $\sqrt{50}$. Therefore, the empirical VCR is lower than the predicted VCR when assuming independence. Nevertheless, they are close to the simulation results when the DGP is a t–GARCH process, see Table 3. Hence, we conclude both from simulation and real data analysis that the serial dependence leads to VCR that are lower than those derived from the independent case.

5 Analysis

The theoretic, simulation and real data analysis together paint a clear picture of the performance of common risk measures and provide a deep understanding of results often observed in practice.

Three main results emerge.

First, the theoretic superiority of ES over VaR comes at the cost of higher estimation error for ES. For many end–users, using typical sample sizes of a few hundred to a few thousand observations, this may well tip the advantage in favor of VaR.

Table 4: Empirical analysis with the CRSP data

N	Number of Samples		Level	ES and VaR uncertainty		VCR holding period comparisons			
			$^{\text{ES}}/_{\text{VaR}}$	VCR (ES,VaR)	Q(VaR,ES)	VaR $H10$	VaR $H50$	ES $H10$	ES $H50$
600	34,244	mean	1.03	0.91	1.02	1.39	2.28	1.41	2.53
		median	1.01	0.86	1.01	1.16	1.78	1.24	2.04
		se	0.09	0.39	0.11	1.01	2.49	0.73	2.48
		ratio	0.57	0.29	0.55	0.61	0.85	0.70	0.93
		p–value	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	20,097	mean	1.04	0.96	1.01	1.40	1.90	1.33	1.97
		median	1.03	0.92	1.01	1.22	1.63	1.24	1.73
		se	0.08	0.29	0.09	0.80	1.06	0.53	0.96
		ratio	0.71	0.34	0.54	0.67	0.88	0.73	0.93
		p–value	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5000	2,503	mean	1.06	1.01	1.01	1.43	2.03	1.35	1.85
		median	1.06	0.98	1.01	1.35	1.83	1.31	1.75
		se	0.04	0.18	0.04	0.49	0.89	0.33	0.64
		ratio	0.97	0.45	0.58	0.83	0.95	0.87	0.96
		p–value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: The table shows the empirical results using daily returns on all liquid traded stocks on NASDAQ, NYSE or AMSE from 1926 to 2014. Each stock return series is split into non–overlapping samples with sample sizes $N = 600, 1,000, 5,000$. All samples from different stocks are pooled together. A sample is included if on the first day of the sample the stock has a share price above 5$ and market capitalization higher than the 10%th quantile of the market capitalization of all stocks traded on NYSE on that day. The number of samples are reported in column 2. Column 4 reports the summary statistics of the ratios $^{\text{ES}}/_{\text{VaR}}$ across all samples. The row "ratio" reports the fraction of samples with a ratio above one. The row "p–value" reports the p–value of a t-test that the mean equals to one. The empirical results for the variation coefficient ratio (VCR) between ES and VaR are reported in Column 5. For the calculation of the VCR, a block bootstrapping procedure is employed with details in Section 4. Column 5 reports the empirical results based on the ratio Q(VaR, ES), which measures the relative difference in the lower bound of the confidence intervals for VaR and ES with the calculation details in Section 4. The last four columns report the empirical results based on the VCR between the a given risk measure and liquidity horizon using the overlapping approach and square–root–of–time approach.

The second result is that the overlapping approach is much less accurate than the time–scaling approach. This certainly holds for i.i.d. data and with the dependent structure considered here. Indeed, there seems to be little reason to use the overlapping approach when forecasting risk. If one is interested in longer holding periods, the time–scaling approach approach allowed in Basel II and the 2014 version of Basel III, is more accurate than the overlapping approach in the 2013 version of Basel III.

Finally, both ES and VaR are highly sensitive to the sample size. The asymptotic properties of the estimators can only be attained when sample sizes span half a century or more. For the smaller sample sizes, below a few thousand days, the uncertainty becomes considerable, and at 500 or less very little signal remains. Consider the case of typically thick tails ($\alpha = 3$) and a 500 day sample size. In this case, the 99% confidence interval around the true value of one is $[0.70, 1.73]$ for VaR(99%) and [0.66,1.82] for ES(97.5%).

We also observed estimation biases using the HS method, expected given Blom (1958). Nevertheless, sometimes the estimation bias can work in the opposite direction: an upwards bias may lead to a relatively high value on the lower bound. Therefore, a full discussion on estimation uncertainty may take into account both bias and variance. Regardless, it is straightforward to adjust for this bias, using for example the methods proposed by Hyndman and Fan (1996).

From a purely statistical point of view, there is little to recommend ES over VaR. Still, there can be very good reasons for opting for a risk measure like ES that otherwise is less accurate statistically than the alternatives. Because VaR is only one point on the distribution and profit and loss and ES captures the entire tail from that point, ES harder to manipulate.

A financial institution may manipulate the risk forecast by picking a particular estimation method. However, there is no guarantee that an estimation method that delivers a favorable result today will deliver similar favorable results in the future. Instead, manipulation is much likely to happen by picking trades that place assets on the on the lower boundary of the confidence interval, something very easy to do while being virtually undetectable. Because ES captures all tail events, it is be harder to implement such manipulation than if VaR is used.

5.1 Implications for back testing

The high estimation uncertainty for both risk measures provides one explanation for why violation ratios in back testing so often deviate from the expected values by large amounts. Since the lower bound can be regarded as an estimate of the risk metric within a reasonable confidence level, the violation ratio based on the lower bound can also be regarded as acceptable at the same confidence level. However, due to the large difference in absolute value, the actual violation ratio using the lower bound may largely deviate from the expectation. This makes the backtesting procedure on reported risk measures challenging.

The high uncertainty of risk forecasts explains why back–testing procedures, such as violation ratio analysis, often perform so badly, since the maintained Bernoulli distribution for violations does not hold in typical sample sizes. This is above and beyond the well-known small sample problem in the Bernoulli distribution when calculation violation ratios.

5.2 Implications for Basel III

In the latest Basel III market risk proposals, the Basel committee suggests replacing 99% VaR with 97.5% ES. Our results indicate that this will lead to less accurate risk forecasts. If the regulators are concerned by precision, VaR is preferred. However, ES is harder to manipulate than VaR and therefore might be preferred even if it is less accurate.

When looking at the confidence intervals, a particular regulatory focus is on the lower bounds. This does give banks some scope for deliberately underreporting risk, perhaps by cherry picking trades known to be on the lower edge. On the other hand, since the confidence bound is highly asymmetric, with a much higher upside, if banks use the point estimate of the risk measures, they are more likely to hold excessively large trading book capital than they are to hold too little capital.

Finally, in some cases the 99% VaR is larger than the 97.5% ES, meaning that the move from Basel II to Basel III may result in lower risk forecasts. This is especially likely for the smaller sample sizes most likely to be encountered in practice.

5.3 The importance of confidence bounds

These results indicate that it is not advisable for neither financial institutions nor the regulators to rely solely on point estimates of risk forecasts. It is important to also report the confidence bounds. Furthermore, given the highly asymmetric nature of these bounds, the actual bounds should be reported, rather than reporting the standard error only.

6 Conclusion

In this paper we focus on three issues in risk analysis: the choice of risk measures, the aggregation method when considering longer holding period and the number of observations needed for accurate risk forecast. We compare the most commonly used risk measures, the VaR and ES.

We conclude that overall VaR is superior to ES, yielding more accurate risk forecasts. When it comes to longer holding periods, the time–scaling approach has the advantage over the overlapping approach. Finally, we need half a century of daily data for the estimators to reach their asymptotic properties, with the uncertainty increasing rapidly with lower sample sizes. At sample sizes of few hundred, the risk forecast retain very little information content.

Appendix

Proof of Proposition 1.
Recall that F is a heavy-tailed distribution with tail index α. Danielsson et al. (2006) showed that if $\alpha > 1$

$$\lim_{s\to 0} \frac{e_F(1-s)}{q_F(1-s)} = \frac{\alpha}{\alpha-1}. \tag{3}$$

In addition, from the regular variation condition, we get that

$$\lim_{s\to 0} \frac{q_F(1-cs)}{q_F(1-s)} = (2.5)^{-1/\alpha}. \tag{4}$$

The proposition is proved by combining Eq. (3) and (4). ■

Proof of Proposition 2.
Under the conditions in the proposition, Theorem 2.4.8 in de Haan and Ferreira (2006) showed that there exists a proper probability space with Brownian motions $\{W_N(s)\}_{s\geq 0}$ such that as $N \to \infty$,

$$\left| \sqrt{k}\left(\frac{X_{N,N-[ks]}}{U(N/k)} - s^{-1/\alpha} \right) - \frac{1}{\alpha} s^{-\frac{1}{\alpha}-1} W_N(s) - \sqrt{k} A(N/k) s^{-\frac{1}{\alpha}} \frac{s^{-\rho}-1}{\rho} \right| \xrightarrow{P} 0 \tag{5}$$

holds uniformly for all $0 < s \leq 1$.

By taking $s = 1$, the first statement on $\hat{q}_F(1-k/N)$ follows immediately. To prove the second statement on $\hat{e}_F(1-k/N)$, we apply the integral for $s \in (0,1]$ to (5) and obtain that as $N \to \infty$

$$\sqrt{k}\left(\frac{\hat{e}_F(1-k/N)}{U(N/k)} - \frac{1}{1-1/\alpha} \right) - \int_0^1 \frac{1}{\alpha} s^{-\frac{1}{\alpha}-1} W_N(s) ds - \lambda \frac{1}{(1-\rho)(1-1/\alpha-\rho)} \xrightarrow{P} 0.$$

Notice that it is necessary to have $\alpha > 2$ to guarantee the integrability of $\int_0^1 \frac{1}{\alpha} s^{-\frac{1}{\alpha}-1} W_N(s) ds$.

Similarly, from the inequality (2.3.23) in de Haan and Ferreira (2006), we get that for any $\varepsilon > 0$, with sufficiently large N,

$$\left| \sqrt{k}\left(\frac{U(N/ks)}{U(N/k)} - s^{-1/\alpha} \right) - \sqrt{k} A(N/k) s^{-\frac{1}{\alpha}} \frac{s^{-\rho}-1}{\rho} \right| \leq \varepsilon \sqrt{k} A(N/k) s^{-1/\alpha-\rho-\varepsilon},$$

holds for all $0 < s \leq 1$. With a small ε such that $1/\alpha + \rho + \varepsilon < 1$, we can take integral for $s \in (0,1]$ on both sides and obtain that as $N \to \infty$,

$$\sqrt{k}\left(\frac{e_F(1-k/N)}{U(N/k)} - \frac{1}{1-1/\alpha} \right) \to \lambda \frac{1}{(1-\rho)(1-1/\alpha-\rho)}.$$

Therefore, by comparing the asymptotics of $\frac{\hat{e}_F(1-k/N)}{U(N/k)}$ and $\frac{e_F(1-k/N)}{U(N/k)}$, we get that

$$\sqrt{k}\left(\frac{\hat{e}_F(1-k/N)}{e_F(1-k/N)}-1\right) \xrightarrow{d} \frac{\alpha-1}{\alpha^2}\int_0^1 s^{-\frac{1}{\alpha}-1}W(s)ds.$$

The proof is finished by verifying the variance of the limit distribution as follows.

$$\begin{aligned}
\mathrm{Var}\left(\frac{\alpha-1}{\alpha^2}\int_0^1 s^{-\frac{1}{\alpha}-1}W(s)ds\right) &= \frac{(\alpha-1)^2}{\alpha^4}\int_0^1 ds\int_0^1 dt\left(s^{-\frac{1}{\alpha}-1}t^{-\frac{1}{\alpha}-1}\min(s,t)\right)\\
&= \frac{2(\alpha-1)^2}{\alpha^4}\int_0^1 dt\left(t^{-\frac{1}{\alpha}-1}\int_0^t s^{-\frac{1}{\alpha}}ds\right)\\
&= \frac{2(\alpha-1)}{\alpha^3}\int_0^1 t^{-\frac{2}{\alpha}}dt\\
&= \frac{2(\alpha-1)}{\alpha^2(\alpha-2)}.
\end{aligned}$$

■

Proof of Proposition 3.
Proposition 2 was proved based on the limit relation (5). We refer to a similar relation based on dependent data, see Theorem 2.1 in Drees (2003). There exists a proper probability space with Gaussian processes $\{B_N(s)\}_{s\geq 0}$ such that

$$\left|\sqrt{k}\left(\frac{X_{N,N-[ks]}}{U(N/k)}-s^{-1/\alpha}\right)-\frac{1}{\alpha}s^{-\frac{1}{\alpha}-1}B_N(s)-\sqrt{k}A(N/k)s^{-\frac{1}{\alpha}}\frac{s^{-\rho}-1}{\rho}\right| \xrightarrow{P} 0 \tag{6}$$

holds uniformly for all $0<s\leq 1$, as $N\to\infty$. Here the Gaussian processes $\{B_N(s)\}_{s\geq 0}$ has a covariance function $c(x,y) := \mathrm{Cov}(B_N(x), B_N(y))$ determined by the dependence structure as follows. Denote $c_m(x,y)$ as the tail dependence function between X_1 and X_{1+m} as

$$\lim_{t\to\infty} t\Pr(X_1 > U(t/x), X_{1+m} > U(t/y)) = c_m(x,y).$$

Then

$$c(x,y) = \min(x,y) + \sum_{m=1}^{+\infty}(c_m(x,y)+c_m(y,x)).$$

We calculate the specific c function under the moving average structure $X_i = \sum_{j=1}^{H} Y_{i+j-1}$. It is clear that $c_m(x,y)=0$ for $m\geq H$. Next, for $1\leq m<H$,

we have that

$$\begin{aligned} c_m(x,y) &= \lim_{t\to\infty} t \Pr(X_1 > U(t/x), X_{1+m} > U(t/y)) \\ &= \lim_{t\to\infty} t \Pr(\sum_{j=m+1}^{H} Y_j > \max(U(t/x), U(t/y))) \\ &= \lim_{t\to\infty} t(H-m) \Pr(Y_j > U(t/\min(x,y))) \\ &= \frac{H-m}{H} \min(x,y). \end{aligned}$$

Consequently,

$$c(x,y) = \min(x,y) + \min(x,y) \cdot 2 \sum_{m=1}^{H-1} \frac{H-m}{H} = H \min(x,y).$$

The covariance function of $B_N(s)$ indicates that we can write $B_N(s) = \sqrt{H} W_N(s)$, where W_N is a standard Brownian motion. The proposition is thus proved by following similar steps as in the proof of Proposition 2. ■

References

Artzner, P., F. Delbaen, J. Eber, and D. Heath (1999). Coherent measure of risk. *Mathematical Finance 9*(3), 203–228.

Basel Committee (1996). *Amendment to the Capital Accord to Incorporate Market Risks*. Basel Committee on Banking Supervision. `http://www.bis.org/publ/bcbs24.pdf`.

Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: A revised market risk framework. Technical report, Basel Committee on Banking Supervision.

Basel Committee on Banking Supervision (2014). Fundamental review of the trading book: outstanding issues. Technical report, Basel Committee on Banking Supervision.

Blom, G. (1958). *Statistical Estimates and Transformed Beta–Variables*. John Wiley.

Danielsson, J., L. M. Ergun, and C. G. de Vries (2015). Pitfalls in worst case analysis. mimeo, LSE.

Danielsson, J., K. James, M. Valenzuela, and I. Zer (2014). Model risk of risk models. Working paper, Systemic Risk Centre and Federal Reserve Board.

Danielsson, J., B. N. Jorgensen, M. Sarma, and C. G. de Vries (2006). Comparing downside risk measures for heavy tailed distributions. *Economics letters 92*(2), 202–208.

de Haan, L. and A. Ferreira (2006). *Extreme value theory: an introduction*. Springer.

Drees, H. (2003). Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli 9*(4), 617–657.

Feller, W. (1971). *An introduction to probability theory and its applications*, Volume II. New York: Wiley.

Hyndman, R. J. and Y. Fan (1996). Sample quantiles in statistical packages. *The American Statistician 50*(4), 361–365.

Jansen, D. W. and C. G. de Vries (1991). On the frequency of large stock returns: Putting booms and busts into perspective. *The Review of Economics and Statistics 73*(1), 18–24.

J.P. Morgan (1993). *RiskMetrics-technical manual*.

Mosteller, F. (1946). On some useful "inefficient" statistics. *The Annals of Mathematical Statistics 17*(4), 377–408.

O'Brien, J. and P. J. Szerszen (2014). An evaluation of bank var measures for market risk during and before the financial crisis. working paper, Federal Reserve Board.

Sun, P. and C. Zhou (2014). Diagnosing the distribution of GARCH innovations. *Journal of Empirical Finance 29*, 287–303.

Yamai, Y. and T. Yoshiba (2002). Comparative analyses of expected shortfall and VaR: their estimation error, decomposition, and optimization. *Monetary and Economic Studies 20*(1), 87–121. IMES Discussion Paper Series 2001-E-12, `http://www.imes.boj.or.jp/english/publication/edps/2001/01-E-12.pdf`.

Yamai, Y. and T. Yoshiba (2005). Value-at-risk versus expected shortfall: A practical perspective. *Journal of Banking and Finance 29*(4), 997–1015.

The London School of Economics
and Political Science
Houghton Street
London WC2A 2AE
United Kingdom

Tel: +44 (0)20 7405 7686
systemicrisk.ac.uk
src@lse.ac.uk