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

February 2015

Abstract

This paper considers the robustness of the standard techniques in risk analysis, with a special emphasis on the Basel III. The robustness analysis covers the relationship 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.

Keywords: Value–at–Risk, expected shortfall, Basel III

[*] Corresponding author Chen Zhou. We thank 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 Bank of The Netherlands.

1 Introduction

Financial risk is usually forecasted with sophisticated statistical methods. However, in spite of the prevalence of such methods in industry and financial regulations, the performance of statistical risk forecast methods is poorly understood. Addressing this deficiency is the main objective of our work, where compare and contrast the most common risk measures, investigate their performance characteristics in typical usage scenarios and study recent methodological proposals.

Minor variations in model assumptions can lead to vastly different risk forecasts. Consequently, one can get vastly different forecasts for the same portfolio, forecasts that are all equally plausible ex–ante. A detailed understanding of how that can arise constitutes a key motivation for this paper. We study the two most commonly used risk measures, Value–at–Risk (VaR) and expected shortfall (ES), comparing and contrasting them, as well as investigating their theoretic and practical properties. The ultimate aim is to see how these popular risk measures perform, 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 and statistical method. Ever since its introduction by J.P. Morgan (1993) and especially the incorporation into financial regulations in Basel Committee (1996), VaR is the most commonly used market risk measure. While it has come under considerable criticism, VaR has generally been preferred, partly because of the work of Yamai and Yoshiba (2002, 2005).

Data is challenging. While any statistical method benefits from as much data as possible, in practice financial data sets may be short or the world may be changing rapidly, rendering old data irrelevant. That puts strict upper limits on how long data sets can be. This practical limitation is brought into an especially sharp contrast when coupled with a desire for longer holding periods, for example 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. The other two alternatives are based on aggregating daily data across time. Supposing we have 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 regulators, as expressed in the Basel III Proposals, are keen on both long holding periods and the overlapping approach.

A large number of statistical methods for forecasting risk has 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 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 five US bank sample of O'Brien and Szerszen (2014), three 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. That means that it is not possible to find a parametric model that performs consistently well across all data generating processes. Although HS is the simplest estimation method, it has the advantage of not being dependent on a particular parametric data generating process. The main deficiency of HS, its poor performance in the presence of structural breaks, will not affect our 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 is becoming increasingly preferred both by practitioners and regulators, most significantly expressed in the Basel III Proposal (Basel Committee on Banking Supervision, 2013). While the Proposal is scant on motivation, the little that is stated only refers to theoretic 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 independent 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 holds that view. 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 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 forecast.

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 data generating process of the observed data and neither do we have any assurance that data across time and assets maintains consistent statistical properties. Therefore, we focus our attention on theoretic and simulation analysis, augmenting it with real data analysis assuming a time invariant data generating process.

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 this 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 incorporates the vast majority of asset returns, VaR and ES are related by a 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 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 the 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 get close to achieving 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 were the cost of type II error 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, when we compare the square–root–of–time approach to the overlapping approach and find that the overlapping approach results in more uncertainty. This result holds theoretically, and in the simulation study, and is robust to both i.i.d. and dependent data generating processes.

2 Risk measure theory

Many authors have consider the various theoretical properties of statistical risk measures and the underlying estimation methods. Here we are interested in particular aspects of risk measures, one that sees little coverage in the extant literature. In particular, we are interested in three questions: 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%) vs. 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 the 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, ${}^{e_F(0.99)}\!/_{q_F(0.99)} = 1.146$. It is also possible to consider different p levels in the two risk measures, such as in comparing the Basel II and III proposals. We get that ${}^{e_F(0.975)}\!/_{q_F(0.99)} = 1.005$. Hence the two risk measures are roughly identical under normality.

Since financial returns exhibit heavy tails, a more realistic distribution is heavy tailed. 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 level. 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 to 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, for the same probability both, 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%), one should take c and s in Proposition 1, such that $1 - s = 0.99$ and $1 - cs = 0.975$, i.e. taking $s = 1\%$ and $c = 2.5$. Then, we get that

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

That is, when comparing ES(97.5%) to II 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 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 further 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 $U = ({}^1/_{1-F})^{\leftarrow}$. 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. 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, when considering the same probability level, the relative estimation uncertainty in the ES measure is higher than that in the VaR measure. The difference is larger for lower α, heavier distributions.

Next, we compare the estimation uncertainty between 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 with respect to α. By solving $g(\alpha) = 1$, we get the break–even point at $\alpha^{\mathrm{be}} = 6$. 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 aggregation

Consider the problem of forecasting risk over holding periods longer than one day, denoting the holding period by H. In this case, we have three main choices, use H day returns (the non–overlapping approach), timescale daily risk forecasts, typically by $\sqrt{H}$ (the square–root–of–time approach) or use H day overlapping data, (the overlapping approach).

Each approach has its own pros and cons, the first is strictly 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 dependence. If one has the data, the first should always be used, in the absence of that one has to choose between the latter two, and that is where we focus our attention.

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 on 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 is 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 the 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 a parallel result as that in Proposition 2.

Proposition 3 *Suppose that $Z_1, \cdots, Z_N$ 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*

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

A similar results holds for ES. To conclude, for both risk measures, 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.

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, in what follows, we focus on the Basel II and III proposals, i.e. the VaR(99%) and ES(97.5%), and so can omit the probability in the discussion.[1]

In each simulated sample, the sample sizes, N, ranges from 100 to 250,000, in intervals of 100, or more. For presentation purposes, we convert sample size 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.

We forecast risk by HS, that is, we apply (1). This means that the HS estimator is slightly biased, (see e.g. Blom, 1958), but the uncertainty of the estimator is not. This means that for our purpose we don't need to adjust the bias by the methods proposed by Hyndman and Fan (1996). A potentially bigger concern is that the mean of HS estimates across multiple draws is a biased estimator of the relevant quantile as noted by Danielsson et al. (2015), and we could use the methods discussed in the paper to adjust the estimate. However, since our main interest is in the uncertainty and not the bias, that 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.

Why such a large simulation necessary? Taking the VaR measure as

[1] In this paper, we focus on the Basel II and III combination in the interest of simplicity. Nevertheless we report the full set of results in the the web appendix at `www.ModelsandRisk.org/VaR-and-ES`.

an example, 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

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

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

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

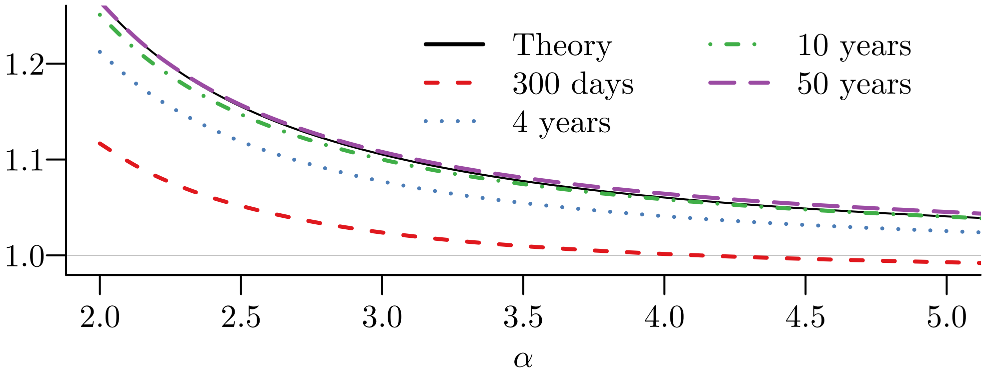
A minimum of $S = 10^7$ samples is necessary for our simulation study and that is the number of simulated samples we use throughout this section.

3.2 Level comparison of the risk measures

The theoretic results in Section 2.1.1 indicate that the relative difference between VaR(99%) and ES(97.5%) is small for distributions that don't have very heavy tails, where the difference is inversely related to the tail thickness. We explore this relation by Monte Carlo simulations with finite sample size. For each given α, we simulate observations from standard Student–t distribution with degree of freedom $\nu = \alpha$. 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 averaged 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 to VaR 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 virtually indistinguishable from the theoretic results.

As sample size decreases, the relative difference between ES(97.5%) and VaR(99%) decreases sharply, especially for the heavier tails. For example, while asymptotic theory suggests that for $\alpha = 3$, ES(97.5%) is 11% larger than VaR(99%) 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 for falls below the VaR.

3.3 Estimation accuracy

The asymptotic results in Section 2.1.2 show that ES(97.5%) is estimated more precisely than VaR(99%)for relatively thin distributions, and less precisely for the fatter and more typical, 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 ratios between the estimates and their corresponding true values, e_F and q_F. Such a procedure is repeated S times for each given sample sizes. 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.[2]

Table 1: Comparison on 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(97.5%) and VaR(99%) 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

[2] The more extensive simulation schedule includes more cases and results in more statistics. Those results are available on the web Appendix at `www.ModelsandRisk.org/VaR-and-ES`.

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(99%) 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 square–root–of–time approach and the overlapping 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. Below 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.[3]

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 VaR(99%) of H–day holding period using both the time scaling and overlapping date approaches. Similar to what we did above, we divide the estimates by the true values, except since we intend to estimate the VaR(99%) of 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

[3] The full results are available on the web Appendix at `www.ModelsandRisk.org/VaR-and-ES`.

Table 2: Impact of overlapping data: 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(99%) 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.

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 that time scaling results in much 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.

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: the GARCH case

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 captured 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

Table 3: Impact of overlapping data, 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(99%) 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.

model, our theoretical result in Proposition 3 may not hold. Therefore, we have to rely on the simulation result for comparing the two strategies.

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 of using overlapping approach for dependent data, even though the time scaling approach still performs better in terms of estimation accuracy. This benefit is contracted by the second observation, where from the 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.

4 Empirical results

Ideally, one would want to validate the theoretic and simulation results above by employing actual financial returns. This would have much more realistic dependence structure and therefore be a much better test case for the overlapping approach vs. the square–root–of–time approach. The downside is that since we do not know the true value of the risk measures, we cannot validate the empirical results. However, we can go some ways towards that by using block bootstrap approaches.

In order to do this, we need a long data set, and we opted to use continuously compounded daily returns of the S&P 500 index from Jan 1, 1928 to Aug 15, 2014 (22,840 daily observations). We consider sample sizes $N = 600$, 1000 and 5000.[4] For each given sample size, we divide the dataset into non–overlapping samples, and perform the analysis in each. For example, for $N = 5000$, only 4 samples are analyzed. The block size needs to be large enough to capture the inherent dependence in the data, and we opted for 200 days.

4.1 Comparing VaR(99%) with ES(97.5%): level and estimation accuracy

We start with the theoretic results in Section 2.1.1 and 2.1.2, where we showed that while ES(97.5%) is higher than VaR(99%), the relative difference between the two is small, and that the relative estimation standard error of ES(97.5%) is higher than that of VaR(99%).

We estimate VaR(99%) and ES(97.5%) for each estimation window and then estimate the standard errors of these estimates by the block bootstrapping method. For example, to estimate the standard error of the estimate on VaR(99%), we bootstrap a random sample with the same sample size N by randomly select a few blocks of 200 consequential observations from the estimation window. Then we re–estimate the VaR(99%) from the random sample, by repeating this procedure for $K = 100$ times, thus obtaining K

[4] The reason to choose $N = 600$ instead of 500 is due to the block bootstrapping procedure. The sample size is required to be a multiple of the block size, 200.

estimates of VaR(99%). The standard error of the estimation of VaR(99%)is calculated as the sample standard error among these K estimates. The same bootstrap procedure is applied to the estimate of ES(97.5%). Then we calculate the ratios ${}^{ES(97.5\%)}/_{VaR(99\%)}$ and ${}^{s.e.(ES(97.5\%))/ES(97.5\%)}/_{s.e.(VaR(99\%))/VaR(99\%)}$ for each estimation window. We report the mean, maximum and minimum of these ratios across all estimation windows in Table 4.

Table 4: Comparison of VaR(99%) and ES(97.5%): empirical evidence

	N	$\frac{ES(97.5\%)}{VaR(99\%)}$	$\frac{\text{Relative s.e. of } ES(97.5\%)}{\text{Relative s.e. of } VaR(99\%)}$
mean	600	1.03	0.93
max	600	1.23	1.77
min	600	0.92	0.49
mean	1,000	1.04	0.94
max	1,000	1.41	1.35
min	1,000	0.95	0.62
mean	5,000	1.08	1.03
max	5,000	1.14	1.48
min	5,000	1.02	0.84

Note: The results are based on the continuously compounded daily returns of the S&P 500 index from Jan 1, 1928 to Aug 15, 2014 (22,840 daily observations). For each sample size N, the dataset is divided into non–overlapping estimation windows with N observations each. In each estimation window, the ES(97.5%) and VaR(99%) are estimated with their standard errors estimated by the block bootstrapping method. The block size is $B = 200$, and the number of bootstrapped samples is $K = 100$ times. The relative s.e. of ES(97.5%) is then calculated by ${}^{s.e.(ES(97.5\%))}/_{point(ES(97.5\%))}$. The relative s.e. of VaR(99%) is estimated in a similar way. Then the ratios $\frac{ES(97.5\%)}{VaR(99\%)}$ and $\frac{\text{Relative s.e. of } ES(97.5\%)}{\text{Relative s.e. of } VaR(99\%)}$ are constructed for each estimation window. The table reports the mean, maximum and minimum of these ratios across all estimation windows.

The table shows that the theoretic and simulation result on point estimates holds up in practice, as the ratios of ${}^{ES(97.5\%)}/_{VaR(99\%)}$ are 1.03, 1.04 and 1.08 for the three selected sample sizes respectively. In addition, for the sample size $N = 5,000$, the ratio ${}^{ES(97.5\%)}/_{VaR(99\%)}$ ranges from 1.02 to 1.14. For the other two smaller sample sizes, the range attain some values that are below, but nevertheless close to, one. These observations are in line with the simulation results that as sample size increases the ratio ${}^{ES(97.5\%)}/_{VaR(99\%)}$ is more stable around its theoretical value given by $f(\alpha)$ in Section 2.1.1, which is above one. In addition, to support the validity of the theory, we invert the relation $f(\alpha) = 1.08$ and obtain a solution $\alpha^{SP} = 3.44$. This is in line with the estimated tail index of the returns of the S&P500 index, see, e.g. Jansen and de Vries (1991).

The second theoretic result on estimation uncertainty is also supported by the empirical results with sample size $N = 5000$. In that case, the mean of the ratio ${}^{s.e.(ES(97.5\%))/ES(97.5\%)}\!/_{s.e.(VaR(99\%))/VaR(99\%)}$ is at 1.03 with a range from 0.84 to 1.48. However, for smaller sample sizes $N = 600$ and $N = 1000$, the validity of the statement that ES(97.5%)has a smaller relative estimation error is challenged. Especially for $N = 1000$, where we observe a rather symmetric range of $\frac{s.e.(ES(97.5\%))/ES(97.5\%)}{s.e.(VaR(99\%))/VaR(99\%)}$ from 0.62 to 1.35 with a mean below one.

Next, we consider the theoretic result in Section 2.2 which concludes that using aggregated data with overlapping aggregation windows will lead to higher estimation uncertainty. We estimate 10–day and 50–day VaR(99%) and the corresponding standard errors by a block bootstrapping method, using both the overlapping approach and the square–root–of–time approach. Then we calculate the ratio between relative standard errors normalized by the point estimates in each estimation window, reporting the mean, maximum and minimum of these ratios across all estimation windows in Table 5.

We note from the table that for the 10–day VaR(99%), the relative estimation error using the overlapping approach is in general higher, with the mean ratio 1.37, 1.42 and 1.31 for the three chosen sample sizes respectively. The ranges of the ratios vary according to different sample sizes. For small samples size such as $N = 600$, it can vary from 0.2 to 4.72, while for $N = 5,000$, the range covers a small region from 0.87 to 1.98. In general, it is highly asymmetric and skewed towards above one. Similar results are obtained for the 50–day VaR(99%)with a more pronounced difference.

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 does not extend to applications employing typical sample sizes of a few hundred to a few thousand observations. In those cases, the uncertainty in

Table 5: Comparison of the overlapping approach and the square–root–of–time approach: empirical evidence

	N	$\frac{\text{Relative s.e. of the overlapping approach}}{\text{Relative s.e. of the square–root–of–time approach}}$	
		$H = 10$	$H = 50$
mean	600	1.37	2.19
max	600	4.72	5.28
min	600	0.20	0.78
mean	1,000	1.42	2.00
max	1,000	3.28	5.11
min	1,000	0.51	0.88
mean	5,000	1.31	2.16
max	5,000	1.98	2.96
min	5,000	0.87	0.97

Note: The results are based on the continuously compounded daily returns of the S&P 500 index from Jan 1, 1928 to Aug 15, 2014 (22,840 daily observations). For each sample size N, the dataset is divided into non–overlapping estimation windows with N observations each. In each estimation window, for a given H, the H–day VaR(99%) is estimated with its corresponding standard errors estimated by the block bootstrapping method, using both the overlapping approach and the square–root–of–time approach. The block size is $B = 200$, and the number of bootstrapped samples is $K = 100$ times. The relative s.e. of each approch is then calculated by ${}^{s.e.(VaR(99\%))}/_{point(VaR(99\%))}$. Then the ratio $\frac{\text{Relative s.e. of the overlapping approach}}{\text{Relative s.e. of the square–root–of–time approach}}$ is constructed for each estimation window. The table reports the mean, maximum and minimum of these ratios across all estimation windows.

the estimation of ES is much higher than the uncertainty in VaR.

The second result is that the overlapping approach is much less accurate than the square–root–of–time approach. There seems to be little reason to use the overlapping approach when forecasting risk. If one is interested in longer holding periods, the Basel II square–root–of–time approach is more accurate than the Basel III overlapping approach.

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 are half a century or more. For the smaller sample sizes, below 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%).

Ultimately, our conclusion is that it would be better for end–users to use VaR, scaling as needed to capture the ES and/or longer holding periods.

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 extremely difficult. It is notable that the estimation uncertainty issue is above and beyond the well-known small sample problem in the Bernoulli 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, with the overlapping approach for analyzing risks in longer holding period. Our results suggest that this would 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.

We also observed estimation biases using the HS method. 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.

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 estimators or picking trades known to be on the lower edge. The former is likely to be difficult, since models can't be changed easily and there is no guarantee that a model that was at the lower bound yesterday will be in the same place tomorrow. It is much easier to manipulate the risk forecasts by picking positions that are known to have favorable risk distribution. On the other hand, the fact that the confidence bound is highly asymmetric, with a much higher upside, means that 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.

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, with focusing on their practical properties. We conclude that overall, VaR is superior to ES, yielding more accurate risk forecasts. When it comes to longer holding periods, the square–root–of–time approach approach has the advantage over the overlapping approach. Last but not least, it is necessary to use a sample with at least 10-year of data to make any meaningful risk forecasts.

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.

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.