Measures of Fragility for Tail Risk Models*

Anne-Florence Allard †

Claudia Chmielowska ‡

Massimo Guidolin §

Manuela Pedio¶

This draft: February 3, 2023

Abstract

A fragile tail risk model is one that can only forecast tail risks accurately when backtested under a specific selection of the possible choices required by the backtest. In the literature, it is not yet clear how model fragility should be measured and quantified. We develop several novel measures of model fragility. Firstly, we propose indices – the success rate (SR), local fragility (LF), relative local fragility (RLF), and the relative area (RA) – that exploit comparative algorithms based on a statistical test that can either reject or not the null that a candidate model is no worse than an alternative according to some loss function. However, because of their limitations, we also propose measures that have an absolute nature and capture intrinsic features of forecasting models in terms of the stability of the loss function over the backtesting parameter space: ruggedness (RG), mean semi-elasticity (ME), and integrated semi-elastic radius (IE). The use and relevance of all of these fragility measures are shown with reference to an application of VaR and ES estimation on daily S&P 500 index returns.

Keywords: Fragility, Tail risk, Superior predictive accuracy test, backtesting.

JEL codes: C52, G32, C53

*The Authors are grateful to Paolo Pepe for his excellent work as a research assistant. Do not circulate or cite without the consent of the Authors.

†University of Bristol, UK. E-mail: anneflorence.allard@bristol.ac.uk (corresponding author).

‡Baffi CAREFIN Centre (Bocconi University), Milan, Italy.

§University of Liverpool Management School, UK and Baffi CAREFIN Centre (Bocconi University), Milan, Italy.

¶University of Bristol, UK and Baffi CAREFIN Centre (Bocconi University), Milan, Italy.

1 Introduction

The measurement and forecasting of tail market risks play an increasingly crucial role in the practice (e.g., see Bhansali (2008)) and in the academic research in finance (e.g., see Gao et al. (2019); Kelly and Jiang (2014)). For instance, the entire literature on the econometrics of risk management (see, e.g., Christoffersen (2011a)) and its practical applications (see, e.g., Hull (2012); Power (2004)) showcases the key role played by methodologies and perceptions of risks in the management of financial institutions of all kinds. The measurement of market risk consists of two main tasks: specifying and estimating models to forecast risk and evaluating the success of the models.[1]

The evaluation of the performance of all types (parametric, semiparametric, and nonparametric) of market risk models is usually performed through (pseudo) out-of-sample recursive backtesting exercises (see (Campbell, 2006); Harvey and Liu (2015)). In a single backtesting exercise, ex-post realized measures of tail risk that depend on actual, realized (asset or portfolio) returns are systematically compared to ex-ante tail risk forecasts in order to assess whether one or more candidate models might "adequately" forecast tail risk or their induced losses from the perspective of a model's user. However, when implementing a specific backtest, the user/modeller typically needs to select a range of parameters that affect the scope and nature of the exercise. For instance, the backtester needs to specify (i) the nature of the application (i.e., on the returns of which assets or portfolios the model ought to be estimated and used in forecasting); (ii) the length of the data sample over which the backtest is performed; (iii) the length of the sample over which the model is estimated vs. the length of the (pseudo) out-of-sample (OOS) period over which the predictive performance is to be assessed; (iv) the quantile index for which the tail risk measurement and forecasting ought to occur (e.g., the confidence level of a VaR or Expected Shortfall measure); (v) the loss functions used to quantify any deviations of the tail risk forecasts from the ex-post tail events or measures. We say that the *backtesting parameter space* consists of all the possible combinations of these parameters within their natural or sensible (as implied by a backtesting framework) ranges.

Yet, a model which is classified as "adequately" performing at measuring and forecasting tail risk by a given backtesting exercise is exposed to the risk of being *fragile*. Fragility means that regardless of the measured realized loss function within a given backtest, a model may turn out to lack

[1] Risk is not an observable quantity, and thus to quantify the risk, a measure must be defined. The unobservable nature of risk also creates a layer of complexity when assessing a risk model's performance, as standard comparisons of forecasts to realised values are not possible.

robustness if it returns a rather different performance when backtested under different parametric selections concerning the nature and execution of the backtest. For instance, a quantitative analyst may report some last-generation GARCH-type model as guaranteeing correct frequency coverage in the appropriate VaR tail and displaying approximate independence of violations of the VaR thresholds over the backtesting sample. Nonetheless, such a model may turn deeply disappointing when applied to a different time series of portfolio returns or asset class, when estimated on a shorter or longer sample, when a different split between estimation and OOS periods is used, when a different VaR confidence level is selected, or when an alternative loss function not based on coverage and independence is employed.[2] A candidate model can be considered to be not fragile (or robust) if it performs adequately–to be later defined–across a range of backtesting parameter values.

In other words, a fragile tail market risk model is one that forecasts tail risks accurately only under some specific choices of the backtesting parameters. However, it is not clear in the literature how model fragility–at least from the perspective of tail risk forecasting–might be measured and quantified as an essential characteristic additional to standard notions of forecasting accuracy. Our key objective in this paper is to develop, discuss the pros and cons of, and deploy to a concrete, relevant application a few measures of market risk model fragility. While in the literature backtesting procedures that consider the VaR across multiple confidence levels have been documented (see, e.g., the backtesting method by Crnkovic and Drachman (1996), in general, standard approaches fail to assess the outcomes relative to the set of possible parameters within the backtesting space. To the best of our knowledge, there exists no generally accepted unified approach to evaluate the variation of results across the backtesting parameter space. Therefore our paper remedies this lack of measures of model fragility with a view to forecasting tail risk.

In fact, summarising the robustness of a model into a single value that increases with fragility can provide additional, easily interpretable information that is useful in model selection problems. We start out our quest by attempting to measure the fragility of a model when it has to be chosen over an alternative in a pairwise comparison of predictive accuracy based on some loss function. For instance, using the superior predictive ability (SPA) test of Hansen (2005), in Section 2, we propose four indicators: the success rate (SR), the local fragility (LF), the relative local fragility (RLF), and the relative area (RA). These measures capture various aspects of the fragility of models when

[2] As we shall see in Sections 2 and 3, fragility can manifest itself both in a comparative sense when the candidate model is paired with some alternative, or in an absolute sense. For the time being, this distinction is inessential to our main point.

considered in pairs, i.e., when the backtester has to choose one model over another. As they establish the preferability of one model over the other, but they cannot be used to produce a ranking, we call these measures *ordinal*.

The limitation of working with fragility measures that can only have a comparative meaning is self-evident. Therefore in Section 3, we also propose three additional measures that are not based on comparison but rather on the value assumed by the chosen loss function and therefore can be used to rank models irrespective of the pairs candidate-alternative model. Accordingly, we call these measures *cardinal*. In particular, we propose three cardinal measures: ruggedness (RG), mean semi-elasticity (ME), and integrated semi-elastic radius (IE). Their intuition is that for a model to be robust, its backtested performance should be uniform throughout the backtesting parameter space. In fact, a lack of uniformity points to the existence of traces of model misspecification compared to the true risk-generating process. These indicators do not depend on the binary outcomes (reject/not reject) of some test of hypothesis but posit that a user can single out a loss function to capture her attitudes towards the performance of a model, following Lopez (1999).

In particular, RG looks at the variation of the loss function over the backtesting parameter grid on which the market risk models are to be evaluated. The ME and IE implement a more generalised approach, designed to give consistent results independently of the backtesting parameters and the grid of values used. The problem with the otherwise intuitive notion of ruggedness (RG) is that the scales and meaning behind the parameters are not generally comparable (for example, the units of a quantile level and the length of the sample on which the model is to be evaluated are not comparable). The key concept in establishing the ME and IE measures of fragility is semi-elasticity. Section 3 introduces semi-elasticity as a way of making changes in the loss function over different parameters directly comparable: by considering the change in loss function relative to the fractional change in the underlying parameter, the semi-elasticity constructs vectors with dimensions of the parameter space that give pseudo gradients of the loss function.

Finally, to illustrate the properties of the proposed methods, all the new fragility measures are estimated with reference to three typical, yet simple models, that are routinely applied in risk management and a classical data set of 20 years of daily returns on the S&P 500 index. In our application, we consider three alternative models: one non-parametric (historical simulations) and two parametric (GARCH and IGARCH). We find that historical simulations are very robust in terms of their pairwise

comparison with both GARCH and IGARCH. In fact, when tested using the SPA test proposed by Hansen (2005), historical simulations are generally considered adequate (at least no worse than the alternative models in terms of their predictive performance) for most of the points in the backtesting parameter space. In contrast, the comparisons between GARCH and IGARCH prove to be much less informative from the perspective of the ordinal fragility measures, irrespective of which of the two models is the candidate or the alternative. Interestingly, the results are reversed when the cardinal fragility measures are considered. In fact, for historical simulations, the realized loss in case of violations of the tail quantile forecasts is heavily affected by small perturbations to the backtest parameters. In contrast, in the case of both IGARCH and GARCH models, the realized loss tends to be stable, especially for intermediate values of the backtest parameters. While the results may seem contradictory, they clearly illustrate the different information conveyed by ordinal vs. cardinal fragility measures. In fact, while the former only considers violations of tail quantile forecasts, the latter considers the loss conditional on violations.

As already commented earlier, the literature on the fragility of risk models in backtesting exercises is scant, to say the least. For instance, while Kupiec (1995) introduced a "proportion of failures" test of unconditional coverage, where the proportion of days where the loss exceeds the VaR is compared to the proportion expected given the confidence level, and statistical significance is established, tests such as that in Crnkovic and Drachman (1996) apply unconditional coverage evaluation at various VaR confidence levels, postulating that, if appropriately modelled, the $\alpha \times 100\%$-VaR should be violated in $\alpha \times 100\%$ of the sample. By evaluating the forecast across the confidence levels, the test has more power in assessing if the model is adequate.[3] With specific reference to backtesting, Boucher et al. (2014), Cont et al. (2008), and Kou et al. (2013), amongst others, investigated the sensitivity of risk measures, and, more specifically, their robustness to model misspecification and changes in the data. In particular, Boucher et al. (2014) have proposed a method for adjusting imperfect risk forecasts with a buffer using results obtained from backtesting.[4] Their approach aims to directly correct VaR estimates for any errors caused by estimation uncertainty or model misspecification, accounting to

[3]However, as noted by Campbell (2006), it may be the case that the distribution of the losses is modelled well, yet moderate returns are not well captured. A test over the confidence levels could then incorrectly classify such a model as inaccurate, even though from a risk management perspective the forecasts are adequate. Instead, Campbell suggested that Pearson's Q test for goodness of fit can be applied, in which a manually selected partition of confidence levels is used.

[4]They defined three desirable criteria of the market risk model: that the unconditional coverage test of Kupiec (1995), the independence test of Christoffersen (1998), and an additional test on the magnitude of the VaR violations are passed. Then, through numerical optimisation, they find the set of minimum VaR adjustments required, such that all the tests are passed on the preceding year of VaR predictions.

some extent for the fragility of the model. Yet, even in their approach, only a single set of parameters is used to implement the backtests. In our paper, we instead seek to introduce a unified way to summarise the fragility of the backtesting outputs applied to market tail risk models with respect to the plausible range of parameters in the backtesting space. These measures can be used in conjunction with the average degree of "adequacy" of a model to judge its overall appropriateness.

The rest of this paper is structured as follows. In Section 2, we introduce the ordinal measures that take steps from some binary outcomes of a test of superior accuracy (more generally, adequacy) of a candidate model within a given pair. Section 3 presents our cardinal fragility measures. Section 4 shows through an extensive application to daily S&P 500 index returns what the different fragility measures entail for selecting among standard market risk models in a typical VaR/ES backtesting framework. Section 5 concludes and offers a few directions for additional inquiry.

2 Ordinal Measures of Model Fragility

We start by defining an array of measures of fragility that can be applied when choosing between a candidate model and an alternative one. The objective is to show that the candidate is no worse than the alternative model according to some criterion or loss function that provides a 0-1 binary response from a backtesting (pseudo) OOS exercise. For example, the candidate model can be a parametric time series model used to forecast tail risk (e.g., a GARCH), the alternative model a nonparametric historical simulation algorithm, and the binary outcome is provided by the rejection of the null hypothesis that the candidate is no worse than the alternative in an SPA test. This realistic example is implemented in our application in Section 4.

Consider a candidate model of interest and any statistically valid test procedure to establish whether this candidate model is at least as good as some alternative model in a specific backtesting design. The backtesting space is defined by the choices of the backtester along k dimensions. For instance, when $k = 2$, the backtesting framework can be characterized by the choice of the total sample size (e.g., $T = 2{,}520$ observations) and of the fraction of data τ that is used to perform model estimation and the remaining is used to perform a recursive, rolling-window pseudo-OOS exercise. Regarding the test procedure, we use Hansen's (2005) superior predictive ability (SPA) test methodology. However, our ordinal measures can be generalized to any algorithm being able to code the fact that a given candidate model is not inferior to a given alternative model or not. More specifically, the measures

could be computed for any testing algorithm such that failure to reject

H_0: *the candidate model is no less accurate than the alternative*

is coded as one, while rejection of H0 in favor of H_1: *the candidate model is less accurate than the alternative* is coded as a zero. This creates a binary grid test output.[5]

2.1 Success Rate (SR)

SR is the fraction of the points in the backtesting parameter space for which the null hypothesis of the test is not rejected (i.e., the candidate model is not inferior to the alternative):

$$SR = \frac{1}{IJ...K} \sum_{i,j,...,k=1}^{I,J,...,K} \mathbb{1}_{\{b(i,j,...,k)=1\}}, \tag{1}$$

where I, J, ..., K are the number of possible values taken by each parameter of backtesting parameter space and $b(i, j, ..., k)$ represents the value of the binary grid described above at the position $(i, j, ..., k)$ in the backtesting parameter space.

SR can be used to establish whether the candidate model adequately performs when compared to the alternative model across the backtesting parameter space, which is the case when SR is equal to or greater than 0.5. However, SR does not indicate how fragile the performance of such a candidate model is. To study such fragility, indicators must also take into account the spatial distribution of the binary output of the test (i.e., how the output of the candidate model and of alternative model are distributed over the backtesting parameter space relative to each other.

One simplistic approach to do so consists in evaluating the candidate model relative to the alternative in a two-stage process. In the first stage, if SR is smaller than 0.5 then the candidate is said to perform worse than the alternative. If SR is larger than 0.5, the candidate is said to perform at least as well as the alternative, and additional information about performance is then acquired through the fragility indicators in the second stage.[6]

[5] This coding algorithm may need to be supplemented by a rule that arbitrarily sets as 0s or 1s the cases in which the test of superior model performance cannot be performed or defined. For instance, in what follows, we set that the null hypothesis is not rejected at points in the parameter space where there are no VaR violations for either the candidate or the alternative model, as in this case the models are considered to be equally accurate.

[6] Model selection and model fragility can be thought of as separate dimensions, which is best illustrated through an example. A candidate model which performs as good as or better than an alternative model on shorter backtesting periods, but is outperformed by the alternative on longer backtesting periods (yielding an SR of less than 0.5) can still be a better model to use. This could occur in the case of structural changes in the underlying processes resulting in the candidate being more relevant to the more recent data, while the alternative captures the average behaviour across the

2.2 Local Fragility (LF)

In the backtesting parameter space, every point (i.e., a combination of backtesting parameters) for which the candidate model is not inferior to the alternative is surrounded by other points. The LF measure explores the behaviour of these adjacent points.

Every point for which the candidate model is not inferior to the alternative is classified by an indicator of instability g_i defined as,

$$g_i = \frac{1}{F} \sum_{f=1}^{F} \mathbb{1}_{\{b_f=0\}}, \tag{2}$$

where the sum over f is over the adjacent points on the backtesting parameter grid. In the three-dimensional case, $F = 6$ for all the non-border points on the parameter grid. Non-border points are the points on the grid that are not at the extreme values of any of the ranges of the backtesting parameters.[7] Under LF, fragility is summarised by averaging over all the points where the candidate model is not inferior to the alternative,

$$LF = \frac{1}{P} \sum_{p=1}^{P} g_p, \tag{3}$$

where P denotes the number of points for which the performance of the candidate model is no worse than the alternative (e.g., when the null hypothesis in an SPA test is not rejected).

LF is an indicator of local fragility as it only takes into account the behaviour (i.e., rejecting the null hypothesis of the SPA test or not) of adjacent points on the backtesting parameter grid. Instead, looking at the behaviour of more distant points would give an indicator of wide-span fragility. For example, all the points located five steps away (instead of one) from a point for which the candidate model is not inferior to the alternative could be taken into account. Local fragility will generally also imply a high level of wide-span fragility (in particular, in the case where every adjacent point on the parameter grid has different outcomes of the binary test). However, the opposite need not be true – there may exist adjacent points with the same outcome of the binary test, while outcomes become increasingly heterogeneous as one examines more distant backtesting configurations. Therefore, from

different regimes better.

[7] On the border, the number of points F can take the values of 3, 4, and 5, depending on the point's location on the parameter grid.

a practical perspective, indicators of local fragility are to be preferred, as small changes to the choice of backtesting parameters (regardless of whether these depend on a user's choice or occur endogenously, for instance, because more data become available) should ideally not affect the assessment of the candidate model. However, LF is affected by the number of points on the backtesting parameter grid for which the candidate model is not inferior to the alternative model (i.e., the higher this number of points, the smaller the contribution of each g_i to LF).

2.3 Relative Local Fragility (RLF)

RLF is an indicator of the fragility of the candidate model relative to the maximum fragility of a model with the same SR,

$$RLF = \frac{LF}{LF_{\max}}, \tag{4}$$

where LF is the local fragility indicator of the candidate model relative to the alternative, and LF_{max} is the maximum local fragility that can be achieved with the same SR for the candidate model. LF_{max} can be estimated under a variety of methods, but in what follows and in our application in 4, we suggest using Monte Carlo random simulation methods. Constructed in this way, RLF is not affected by the specific SR of the candidate model, while, as argued above, LF is, by construction.

Using Monte Carlo simulation methods, the distribution of the backtesting parameters for which the null hypothesis of the binary test is rejected in the case of a maximum local fragility configuration can be constructed in two ways–random simulation or quasi-random simulation. In a random simulation, at every point in the backtesting parameter space, the null hypothesis is not rejected with a probability given by the SR of the candidate model. This is achieved by sampling a value u from the standard uniform distribution at every point on the backtesting parameter grid, and setting the null hypothesis to be rejected if $u > SR$ and not rejected otherwise.

Random simulations will result in a grid with the desired SR that is locally fragile, but it will not generally achieve the maximum fragility possible. Quasi-random simulation can instead be used to achieve the maximum fragility possible and is therefore employed in our empirical application. The first step consists in constructing a grid of maximum fragility–every combination of backtesting parameters for which the null hypothesis is not rejected is completely surrounded by points where the null hypothesis is rejected. This configuration will result in an LF of one, with a success rate SR'. Then, the second step consists in achieving the desired success rate SR. If $SR > SR'$, every

combination of parameters for which the null hypothesis is rejected is turned into a point where the null hypothesis is not rejected with probability $p = SR - SR'$. If $SR < SR'$, any of the excess points where the null hypothesis is not rejected are turned into points where the null hypothesis is rejected, again subject to the constraint that $SR = SR'$.[8]

2.4 Relative Area (RA)

The RLF indicator above is calculated as an average across the whole backtesting parameter space. However, it can also be thought as a function of $w \times 100\%$ of the least fragile backtesting parameter combinations, $LF(w)$. That is, consider sorting the values of g_i from Equation 2 into the set $\{g_i\}_{i=1}^{P}$ and compute LF from 3 as a function of the first $w \times 100\%$ of the values in $\{g_i\}_{i=1}^{P}$. By construction, $LF(1)$ corresponds to the LF in Equation 3. Constructed in such a way, LF now captures the fragility given the position of a combination in the backtesting parameter space.

RA is then the ratio between the areas under the curve of $LF(w)$ and that of $LF_{max}(w)$:

$$RA = \frac{\int_0^1 LF(w)dw}{\int_0^1 LF_{max}(w)dw}. \tag{5}$$

A discrete approximation of RA can be obtained by using the trapezoidal rule to compute the two integrals as sums over a discrete range of values of w. RA gives more complete information about fragility when compared to LF and RLF. Specifically, RA takes into account how fragility changes when moving from focusing on the most stable regions of the backtesting parameter space to the whole space. In particular, a convex $LF(w)$ indicates concentrated regions of high fragility in the binary test outcomes, while a concave $LF(w)$ is associated with the fact that fragility in performance is more dispersed through the backtesting space. Because $LF(w)$ is bounded between 0 and 1 (as $LF(0) = 0$ and $LF(1) = 1$), from basic math it follows that a convex function $LF^v(w)$ has a smaller area beneath than a concave function $LF^c(w)$, provided that $LF^c(0) = LF^v(0)$ and $LF^c(1) = LF^v(1)$. This is taken into account by RA: the larger the RA, the stronger the fragility of the candidate model

[8] A noteworthy feature of quasi-random simulation is that the dimensionality k of the backtesting parameter space determines whether the resulting grid in the first step is unique. When the backtesting parameter grid is such that at least two of the dimensions are of even cardinality (as measured by the number of possible parameter values), the number of points where the null hypothesis is not rejected will be equal to the total number of points on the grid divided by four. However, if two or three of the cardinalities are odd, there are different arrangements corresponding to the maximum instability that have different values of SR'. Since in the second step of a quasi-random simulation the probability that the parameter combinations for which the null hypothesis is rejected are replaced by combinations for which the null is not rejected depends only on the probability given by $SR - SR'$, the higher the SR' the lower the number of backtesting grid points that are replaced, and thus the smaller the decrease in fragility.

with respect to the alternative.

3 Cardinal Measures of Model Fragility

The indicators presented in Section 2 are comparative indicators based on the binary outcomes of some valid statistical test (e.g., the outcomes of an SPA test in terms of rejecting or not the null that a given candidate model is no less accurate than some alternative). However, using the binary outcomes of some test of predictive accuracy for tail risk fails to optimise the extraction of information contained in the loss function used in the underlying test. Additionally, the comparative measures of fragility developed in Section 2 are defined over the entire backtesting parameter space used to evaluate the performance of a candidate model. They can therefore be affected by redundant and/or misleading information about the relative fragility of the models, for example when the stability of a model performance turns out to be influenced by areas of the backtesting space that are unlikely to be used in practice. Last but not least, all the indicators introduced in Section 2 can only be used to compare two models at a time: the candidate and one alternative.

In this section, we propose a number of *absolute* measures of fragility across the backtesting parameter space for individual models. These indicators are based on the behaviour of a loss function over the backtesting space. Several loss functions can be used. In our application in Section 4, we resort to a standard square loss that accounts for violations of tail quantile forecasts, as in the SPA test.

Firstly, the ruggedness index (RG) is proposed. RG is dependent on the choice of the grid of backtesting parameter values over which the loss function is computed. To create an index that is more general and not tied to the specific combinations of parameter values used in the backtesting, we also introduce the concept of semi-elasticity. Based on this quantity, the mean semi-elasticity (ME) along with the standard deviation of semi-elasticities (SDE), and the integrated semi-elastic radius (IE) are proposed as useful indices to capture various aspects of the behaviour of semi-elasticity.

3.1 Ruggedness Index (RG)

The concept of ruggedness originates in topography, where it is used to quantify the average deviation of elevation between adjacent points on a terrain map (Riley et al., 1999). In our case, ruggedness is used as a measure of the change in a loss function of interest when moving across a grid of parameter

values defining a range of plausible backtesting exercises. More specifically, the ruggedness index is defined as an average change in the loss between vertices of the parameter space:

$$RG = \frac{1}{S}\sum_{s=1}^{S}\sqrt{\sum_{w=1}^{W}(Loss_{m,w} - Loss_{m,s})^2}, \tag{6}$$

where $Loss_{m,x}$ is the loss of model m evaluated at some point x on the parameter space, S is the total number of non-border points on the parameter space, and the sum over W sums over all the adjacent points within the set that implies s. Adjacent points are those that can be reached by one step in one or more of the parameters. For instance, when a $k = 3$-dimensional parameter space is considered, $W = 26$; each non-border point on the parameter grid is surrounded by 26 points relative to which the difference in the loss of the model $Loss_{m,w} - Loss_{m,s}$ is calculated. RG can be extended to a more general N-dimensional parameter space by adjusting the number of adjacent points W accordingly.

RG is a measure of the average variation of the loss across the backtesting parameter space–it resembles a spatial variance measure. The larger the RG, the more the loss varies between adjacent cells on the parameter grid, and so the "rougher" and thus more fragile the model across the backtesting space is. To evaluate a model's tail risk forecasting performance, the RG could be used in combination with an average measure of the loss function across the parameter space. A low mean loss combined with a low RG across the backtesting space implies that not only the model performs well (as shown by the low mean loss), but also that the performance does not deviate much when the backtesting parameters are changed.

There are two main drawbacks to this approach. Firstly, RG relies on the grid of parameters in κ being fixed to compare levels of fragility across models. Secondly, RG measures the changes in all directions as being of the same importance (i.e., the changes in the loss between adjacent cells fail to carry weights when they are summed to compute RG) regardless of the size of the change in the underlying parameters across points on the grid. The backtesting parameters each have their own (problem-dependent) natural scale, which ideally should decide the weights of their contributions to RG. In order for RG to effectively capture the fragility of the model, the user should use carefully selected ranges and increments of the parameters within the κ vector as a function of their individual goals, in an attempt to specify the natural scale of the parameters.

The RG index captures the average behaviour of the change in the loss. However, as with the LF

and RLF measures, by averaging over the backtesting space, spatial information about the contribution to the ruggedness by each backtest on the parameter grid is lost. This can be problematic for at least two reasons. Firstly, the use of average values across the parameter space results in a loss of information. A model with a very low RG across most of the backtests in the grid but characterized by a small fraction of combinations near the edge of the parameter space where the RG is very high may be preferred to a model for which its RG is at a moderate level throughout the entire space of backtests performed. The first model would be seen as better because the results of the backtest across the values of parameters not at the edge of the grid are more stable than under the second model where the RG is higher over these areas. Secondly, this feature may allow for results to be manipulated by the user. If values of the backtest parameters that do not represent realistic scenarios when used as inputs in the backtest are introduced to the parameter space (for example, excessively long subsample lengths and small fractions of the subsample used to estimate a model), the RG may imply low fragility, but the fragility would no longer be relevant to the application for which the backtest was to be used.[9]

3.2 Semi-Elasticity

A natural way to quantify model fragility is to look at the steepness of the loss function as the backtestesting parameters change, i.e., at the gradient of the loss function with respect to the backtesting parameters. However, this approach is made difficult due to the parameters being non-comparable across the k alternative dimensions of the space of backtests. The main obstacle is defining a correspondence between the parameters (e.g., how does an increase of the OOS subsample by 100 observatios correspond to a 10 percentage point increase in the fraction of the subsample used to estimate the model?). This creates a difficulty when comparing or combining the quantities–for example the gradient of the loss function with respect to $\mathbf{k}$ alternative types of parameters–to compute a single measure. The magnitude of a vector with components expressed in heterogeneous units cannot be found, unless some weighting scheme of the components is selected. Additionally, not only do different backtesting parameters are expressed in different units, they also do not naturally span the same ranges.[10]

[9]The sensitivity of the results to the selected backtesting parameter range is ubiquitous among the other indicators and indices as well. Henceforth, we assume that the user selects a space of backtests that is reasonable, which will convey useful information when used as inputs for backtesting. These ranges should be chosen a priori, based on expectations or requirements of uniformity of model performance over the ranges of backtesting parameter values.

[10]For instance, anticipating a few aspects of our application in Section 4, the length of the OOS sub-sample may range from two observations (one to estimate the model and one OOS forecasting observation) to infinity (which is not possible

Semi-elasticity, defined as the ratio of the change in a target variable and the percentage change in the parameters, remedies to these difficulties. By looking at percentage change in the parameters, the semi-elasticity has the same units as the loss function (which in our intended applications will be unitless as it is a function of returns). Additionally, when semi-elasticity is used, there is no need to make adjustments for the fact that some of the parameters are bounded from above and below (for example the confidence level and the fraction of the subsample used to estimate the model) and others are, in principle, only bounded from below (for example the subsample size). Therefore, in contrast to the gradient of the loss function, the semi-elasticities with respect to each of the backtesting parameters can be compared and combined. The semi-elasticity is a flexible quantity–it can be calculated using any backtesting parameter, and this parameter can then be covered in the fragility analysis.

If we call the backtesting parameters κ_i, $i \in 1, 2, ..., k$, and the average loss for model m, $Loss_m(x_1, x_2, x_3)$, the semi-elasticity (E) is defined as:

$$E_m = (\frac{dLoss_m}{d\kappa_1/\kappa_1}, \frac{dLoss_m}{d\kappa_2/\kappa_2}, ..., \frac{dLoss_m}{d\kappa_k/\kappa_k})'. \tag{7}$$

This vector can be summarised into a single value through a norm. For instance, as in our application, the L^2 norm (the square root of the sum of the squared components) is used, but other norms such as the L^1 norm (the sum of the absolute value of the components), and the infinity norm (the maximum of the absolute value of the components, also denoted by L^∞) may be used as well.

The semi-elasticity measures how much the loss function changes per unit of percentage change in the backtesting parameters. A semi-elasticity of zero corresponds to the case of a constant loss function across the parameter space. Models with a semi-elasticity of zero throughout the backtesting parameter space are completely stable: their performance, as quantified by the loss function, is the same regardless of the parameters used to backtest the model.

in practice due to finite data). However, the confidence level of, say, VaR and ES forecasts lies in (0,1). The fraction of the subsample used to estimate the model ranges also falls in the range (0,1). Instead of considering the fraction of the subsample used to estimate the model, this parameter could be seen as the number of days on which the model is estimated, for which the values of the parameter would be given by the product of the fraction of the subsample used to estimate the model and the length of the subsample. The values of such a parameter would then fall in the range [1,T), where T is the length of the subsample. The ruggedness index attempts to avoid these issues by defining the natural scale of the parameters as the grid of parameters implemented by the user.

3.3 Mean Semi-Elasticity (ME)

A further measure potentially useful to summarise the fragility of a model across the backtesting space that uses semi-elasticity is the mean semi-elasticity (ME). It is defined as

$$ME = \frac{1}{IJ...K} \sum_{i,j,...,k=1}^{I,J,...,K} E_m(\kappa_i, \kappa_j, ..., \kappa_k), \tag{8}$$

where the sum over i, j, ..., k is the sum over the number of elements spanned by the parameter space. The larger the ME, the more fragile the model is on average. It can be used in conjunction with the standard deviation of the semi-elasticities (SDE),

$$SDE = \sqrt{\frac{1}{IJ...K} \sum_{i,j,...,k=1}^{I,J,...,K} (E_m(\kappa_i, \kappa_j, ..., \kappa_k) - ME)^2}, \tag{9}$$

from which the (symmetric, squared) spread of semi-elasticities across the parameter space can be measured.

While ME provides useful information on the measurable level of the semi-elasticity across the backtesting parameters, it does not take into account the spatial distribution of the semi-elasticities. Hence the disadvantage of ME is equivalent to that of the ruggedness. This can be seen when considering a simple example of two models, one with moderately large semi-elasticities throughout the parameter space, and another one with two distinct regions of κ characterized by small and large semi-elasticities. As with ruggedness, the second situation ought to be preferred when selecting a model from a practical perspective, since a small change in the backtesting parameters is less likely to result in a large change in the semi-elasticity, and thus the loss function will propose less deviations across this area of the parameter space. The standard deviation of the semi-elasticities captures the extent to which they vary over the whole κ space, but it fails to take into account their spatial distribution (i.e., distinguishing whether values of semi-elasticity change monotonically as the parameter space is traversed or vary between large and small values on nearby points of the parameter grid).

3.4 Integrated Semi-Elastic Radius (IE)

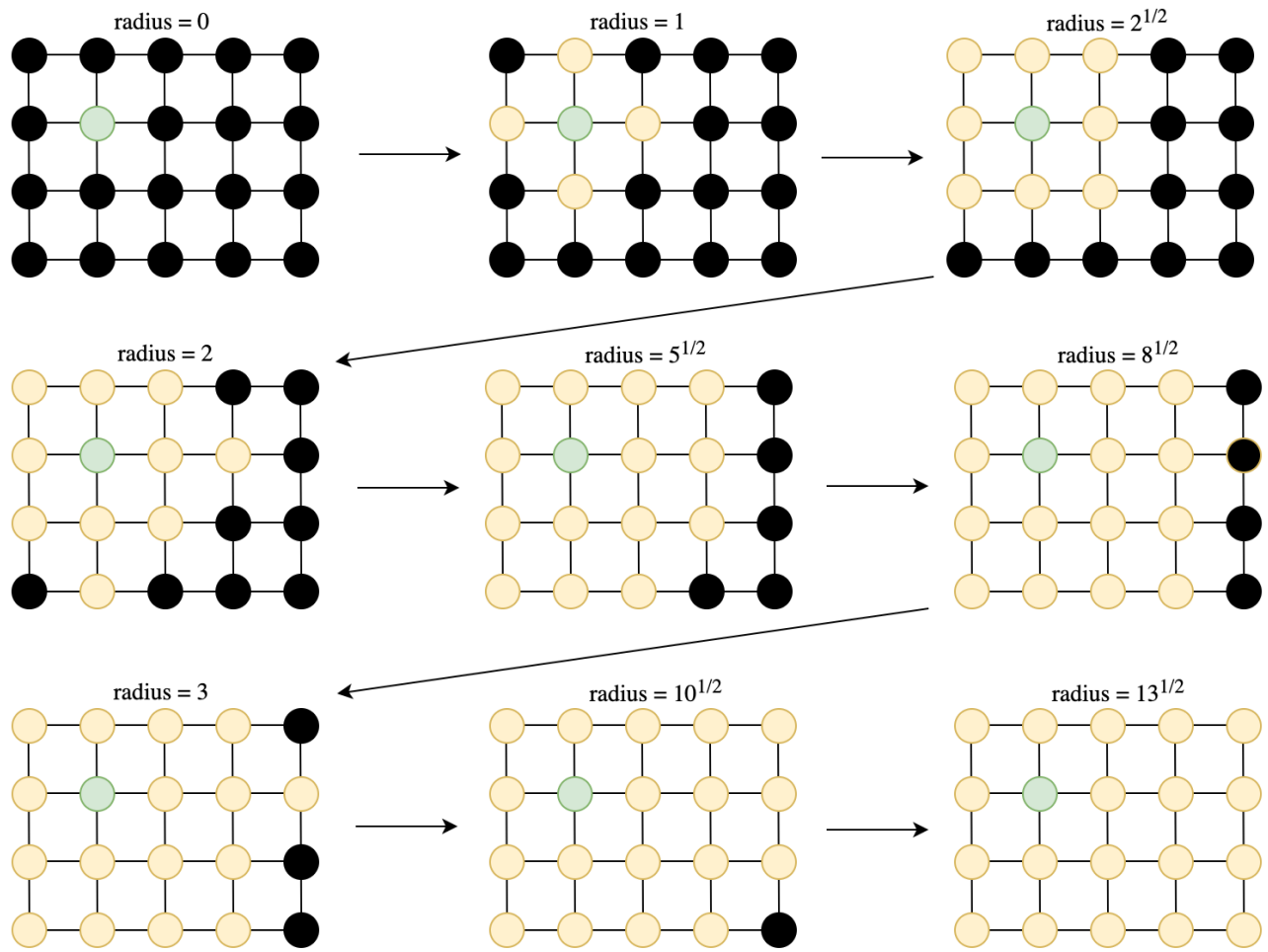
To address the issues described above, we propose the integrated semi-elastic radius (IE) as an index of absolute fragility. The construction of the measure is specified through a step-wise algorithm, before providing a general definition that extends to the limiting case where all the parameters are continuous and the denseness of κ diverges.

The IE can be estimated through the following algorithmic steps:

1. The location of the point of minimum loss on the parameter space is found. This is the point in the backtesting parameter space where the model appears to perform best given the user's loss function.
2. For each point on the parameter grid, the semi-elasticity vector is calculated using the discrete approximation to Equation 7. These vectors are summarised into single values of the semi-elasticity by means of a norm (say, the L^2 norm).
3. From the point of minimum loss, the distance to every point on the backtesting parameter grid is calculated. A distance of one corresponds to traversing one edge on the grid. By considering semi-elasticities on the grid, the grid is now unitless. Therefore one edge of the parameter grid is given a value of one, before being normalised later. When more than one parameter is varied, the distance is given by the value of some norm applied to the distances in each parameter direction. These values are called *radii* because they represent the radial distance from the point of minimum loss on this unitless grid. From these distances, an array of unique sorted values $u = [0, ..., U]$, where U is the maximum possible minimum distance to get from the most stable point to an edge-point of the grid, is computed. An illustration of this process using a two-dimensional parameter space can be found in Figure 1.
4. For each $u = [0, ..., U]$, the value of the mean of the semi-elasticities within that distance from the point of minimum loss is calculated, forming an array of means m of length $U + 1$ (where the first value of m is the semi-elasticity at the point of the minimum loss).
5. The fractional radii, r, are obtained by normalising the radii from $[0, U]$ to $[0, 1]$.

Having constructed an array of normalised radii, r, and the corresponding mean semi-elasticities

Figure 1: **Diagram illustrating integrated semi-elastic radius (IE) calculation.** The vertices represent semi-elasticities calculated on a two-dimensional parameter space. In the first stage, the point of minimum loss (green) is found. At each iteration, the mean of the points within the radius of the most stable point (indicated in yellow) is calculated. During calculation, a radius of 1 corresponds to the traversal of one edge of the grid. The radii are later normalised to fall in the range [0,1].



within each of the radii, m, the integrated semi-elastic radius IE is then:

$$IE = \sum_{u=1}^{U} \frac{(r_{u+1} - r_u)(m_{u+1} + m_u)}{2}. \tag{10}$$

In the continuous limit of the denseness of the parameters in κ, IE converges to the integral of the mean semi-elasticity as a function of the fractional distance from the point of minimum loss,

$$IE = \int_0^1 m(r)dr, \tag{11}$$

where m is the mean semi-elasticity at normalised radius r from the backtest combination of minimum loss.

Constructed in this way, IE measures the spatial distribution of the semi-elasticities, and, more precisely, the average change in the loss function when moving away from the point of minimum loss.[11] The point of minimum loss is where the model can be considered as most performing. A model which is not fragile will have consistently low values of semi-elasticity when moving away from the point of minimum loss, indicating a uniform performance throughout the parameter space. Meanwhile, large increases in IE at small normalised radii indicate that the changes in the loss function near the point of minimum loss are dramatic and that the results of a backtesting exercise may vary significantly with small changes in the corresponding parameters. A gradual increase in the IE as the normalised radius increases represents a middle ground, where the backtest will not classify uniformly the performance across the backtesting parameter space, but the changes are gradual as parameters are changed. IE combines both important aspects of a fragility index: the distribution of the changes in model performance across the backtesting parameter space and the average of the changes in model performance as the parameters in κ are varied.

[11] Normalising the radii to the interval $[0, 1]$ is done to minimise the effect of performing this analysis on a finite backtesting parameter space, as otherwise the case of the point of minimum loss lying near the border of the parameter grid would yield a different range of radii compared to a point of minimum loss near the centre of the grid. It is assumed that the backtesting parameter ranges are chosen from a practical perspective, and so the potential behaviour of the fragility outside the parameter space should not contribute to the fragility of the model as the parameters are not relevant in practice.

4 An Application to the S&P 500 Tail Risk Prediction

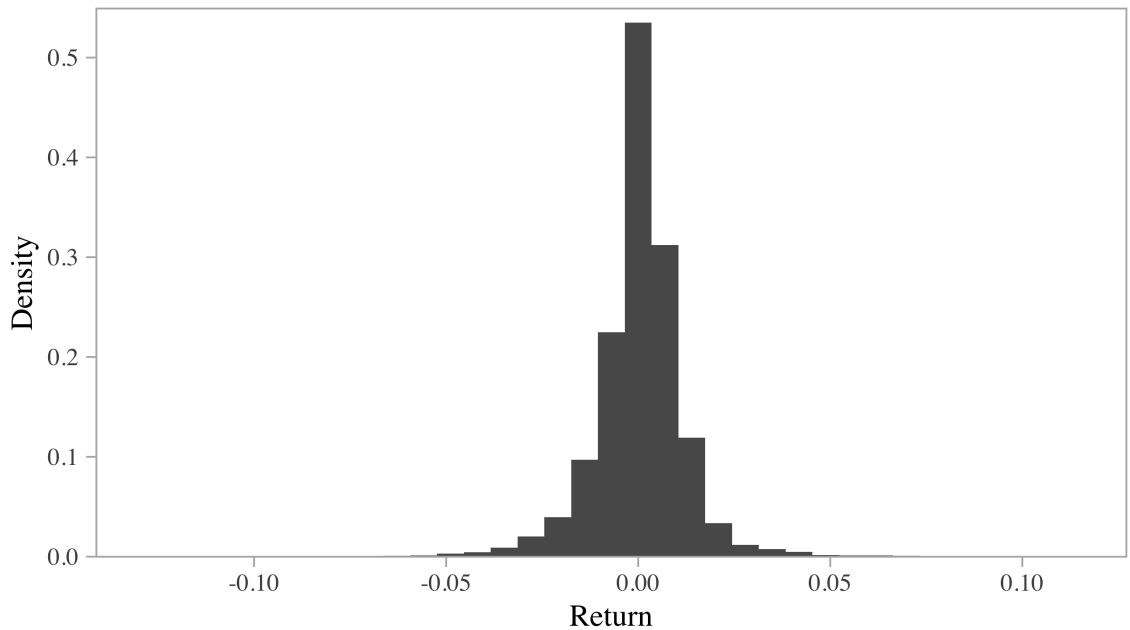
4.1 Data and Some Common Market Risk Models

4.1.1 Data and Some Common Market Risk Models. In our analysis, we use the logarithmic returns on the S&P 500 between May 4, 2001 and February 19, 2021 (5,000 daily return observations in total). S&P 500 data are the most widely used in applied risk management analysis because of their key role in option pricing and portfolio management and hence represent the ideal framework to develop an application illustrative of the methodological notions of Sections 2 and 3. The empirical distribution histogram for S&P 500 returns is shown in Figure 2. Summary statistics concerning returns can be found in Table 1, which shows evidence of non-normal returns, a key stylized fact that needs to be addressed by the tail market risk models reviewed below.

Table 1: **Summary statistics of S&P 500 returns.** 5000 daily returns are considered between 04/05/2001 and 19/02/2021. σ denotes the standard deviation. The SW p-value is the p-value of the Shapiro-Wilk test for normality.

Mean	Median	Min.	Max.	σ	Skewness	Kurtosis	SW p-value
0.00025	0.00067	-0.12765	0.10957	0.01241	-0.42768	14.95924	0.00000

Figure 2: **Empirical distribution of S&P 500 returns.** The distribution is shown for daily returns between 04/05/2001 and 19/02/2021 (5000 returns in total).



4.1.2 Market Risk Models. Three methods are used to model and forecast tail risk: time-weighed historical simulations (BRW), GARCH(1,1), and IGARCH(1,1). Of course, these just represent examples used to illustrate the earlier points. (Unweighted, simple) historical simulation is a non-parametric approach in which it is assumed that the one-day-ahead forecasts for the returns follow the empirical distribution of the past T returns.[12] Whilst simple to implement, choosing the value of T (and so the "smoothness" of the resulting tail forecasts) poses a challenge. Using a time-weighted historical simulation approach, as was first introduced by Boudoukh et al. (1997), BRW reduces the importance of choosing T. At time t an empirical distribution of past returns is generated by considering the set of returns, $\{R_{t-\tau+1}\}_{\tau=1}^{T}$, with probability weights given by $\eta_t = \{\eta^{\tau-1}(1-\eta)/(1-\eta^T)\}_{\tau=1}^{T}$.[13]

While the model-free approach of BRW can be seen as advantageous since there is no need to rely on a (potentially misspecified) parametric distribution, modelling the distribution of the returns non-parametrically also has some disadvantages. In general, the BWR method only captures the extremal aspect of the leverage effect between volatility (and therefore tail risk) and returns.[14] On the contrary, parametric approaches to modelling tail risk aim at capturing stylised facts about the distribution of returns and may generate richer dependence between the mean and the tails:

$$R_t = \mu_t + \sigma_t z_t, \quad \text{with } z_t \sim \text{i.i.d. } D(0,1), \tag{12}$$

where σ_t is the conditional volatility, μ_t is the conditional mean, and $D(0,1)$ is a standardised distribution. For concreteness, in our analysis the shocks are assumed to follow a Gaussian distribution.[15]

The GARCH family of models can be used to forecast volatility. In our application GARCH(1,1)

[12] Owing to its ease of implementation, non-parametric historical simulation tends to be the most common method employed by banks internationally (Perignon and Smith, 2010).

[13] When tail risk is measured by a specific quantile and its estimate falls between two observations, interpolation can be used to approximate the risk measure. Linear interpolation was used in our application. The value of η used is 0.99. Boudoukh et al. evaluated the method using $\eta = 0.97$ and $\eta = 0.99$. However, in general, to obtain the best performance possible, this value should be optimised to the data at hand. Žiković and Aktan (2011) suggested an optimisation approach in which the quadratic loss function is minimised. Calculating the one-day-ahead 1%-VaR forecasts for the S&P 500 between 04/01/2000 and 02/01/2009, they found the optimal value of η to be between 0.979 and 0.995 for various out-of-sample periods.

[14] Since only the left tail of the empirical distribution of returns is considered, estimates of tail risk only increase when returns become more negative. When large positive returns occur the tail risk estimate does not increase, even though such changes also suggest an increase in market risk (Pritsker, 2006).

[15] The use of Gaussian distributions for the shocks is common. However, it is well known that parametric distributions other than the normal could be used to better approximate the conditional distribution of returns, including Student's t-distribution and asymmetrical t-distributions (Christoffersen, 2011b).

and IGARCH(1,1) models are compared.[16]

4.1.3 Tail Risk Measures and their Backtesting. For both the parametric and the nonparametric risk models, once the one-day ahead variance forecasts are obtained, the $\alpha \times 100\%$-VaR and $\alpha \times 100\%$-ES values (under the assumption of normality of conditional returns) were obtained by multiplying the square root of predicted variance by the α-quantile of the standard normal distribution and the mean below the α-quantile of the standard normal distribution respectively.[17]

Three backtesting parameters were considered in what follows ($k = 3$): the VaR and ES confidence level (α, taken to be same for the two), the sub-sample length (l), and the fraction of the sub-sample length on which the models were estimated (q). The subsample was taken to always finish on the final day of the total sample (i.e., 19/02/2021). The length $l \times q$ defines the period on which the models were estimated, with $(l - l \times q)$ days used for out-of-sample forecast evaluation. Rolling windows are used to estimate the models and produce one-day-ahead forecasts, at each iteration moving one day forward in time until forecasts are produced for the whole out-of-sample period. An illustration of this process can be found in Figure 3.

This was performed for every combination of the parameters, with values in the ranges: 0.005 to 0.05 (in increments of 0.005) for the confidence level, 100 to 5000 (in increments of 100) for the subsample length, and 0.1 to 0.9 (in increments of 0.1) for the subsample fraction on which the models were estimated. Such combinations represent on a cubic grid the three-dimensional vector κ in our application.[18]

[16] Under GARCH(1,1), the variance is modelled as

$$\sigma_{t+1}^2 = \omega + \alpha R_t^2 + \beta \sigma_t^2, \tag{13}$$

with parameters ω, α and β. The parameters are constrained such that $\alpha + \beta < 1$, ensuring covariance stationarity. IGARCH(1,1) imposes the restriction $\alpha + \beta = 1$ so that the variance process is not stationary. The parameters of both the GARCH(1,1) and the IGARCH(1,1) models are estimated via maximum likelihood.

[17] A risk measure is a mapping from the random profit and loss of a financial position to the set of real numbers. In this paper, we focus on two commonly used risk measures: Value-at-Risk (VaR) and Expected Shortfall (ES). VaR is the most popular method of measuring market risk. It quantifies an asset's or a portfolio's maximum expected loss for a given confidence level and time horizon. For a confidence level α and returns R_t over time t, the one-step-ahead VaR is defined as the α-quantile of the expected loss:

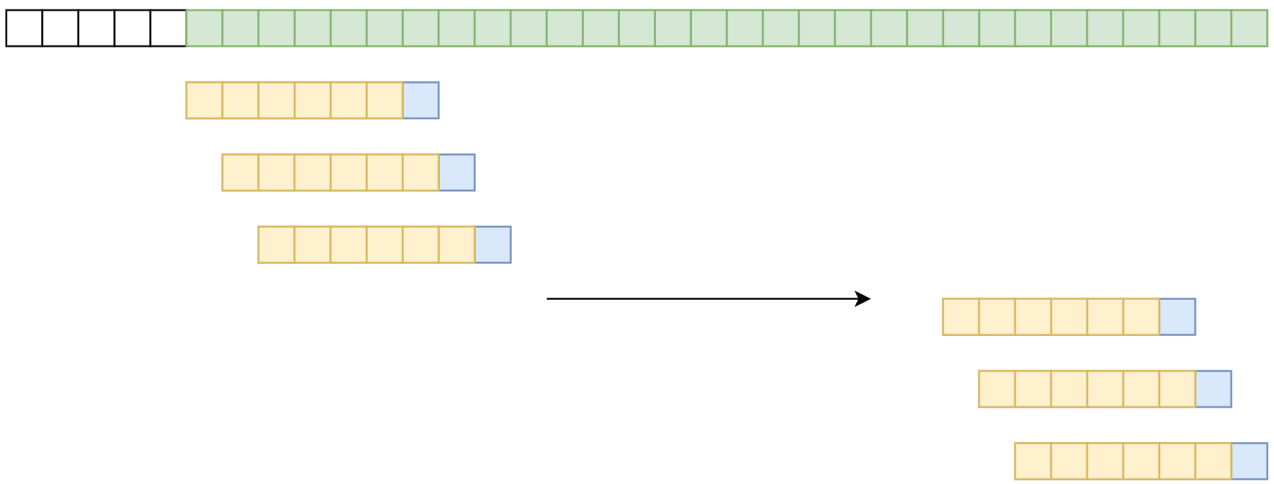
$$\Pr(R_{t+1} < -\mathrm{VaR}_{t+1}^{\alpha}) = \alpha. \tag{14}$$

However, the VaR fails to reflect information about the magnitude of the loss in the case of a violation. The ES is a risk measure that takes into account the confidence level and the distribution of the losses in the tail, given that the VaR is violated. The ES is defined as the expected loss conditional on the return being worse than the VaR:

$$ES_{t+1}^{\alpha} = -E_t(R_{t+1} | R_{t+1} < -\mathrm{VaR}_{t+1}^{\alpha}). \tag{15}$$

[18] These three parameters are fundamental inputs into the market risk model back-tests, but the robustness of models

Figure 3: **Illustration of the backtesting parameters and forecasting structure**. This example shows a total sample length of 35, a subsample length of $l = 30$ (green), and the fraction of the sample length on which the models were estimated of $q = 0.2$ (yellow). The subsample finishes on the last day of the sample. The length of $l \times q = 6$ defines the period on which the models were estimated. A rolling window approach is used to produce $l - l \times q = 24$ one-step-ahead forecasts (blue).



4.1.4 Superior Predictive Ability (SPA) Tests.

Hansen's SPA test [2005] is used to create a binary classification of the market risk forecasting performance of two models. Under the SPA test, one or more candidate models are compared to an alternative, testing the null hypothesis of the alternative is not inferior to any of the alternative models in terms of their expected loss. In this application, only one candidate model is evaluated at any point. The loss used here is similar to the loss proposed by Lopez (1999). The loss is zero when no violation occurs, while at each point in time when the return violates the VaR the loss is the quadratic deviation from the ES:

$$C_{m,t} = \begin{cases} (R_t - ES_{m,t})^2 & \text{if } R_t < -VaR_{m,t} \\ 0 & \text{if } R_t \geq -VaR_{m,t}. \end{cases} \tag{16}$$

The test proceeds as follows. The difference in the loss between the tested model and the alternative model is defined as d_t at every point in the (pseudo-) OOS $t \in [1, T]$. The null hypothesis is then $E(d_t) \leq 0$. The time series $\{d_t\}_{t=1}^{T}$ is assumed to be strictly stationarity and to satisfy a strong mixing condition (such that the use of the stationary bootstrap, described below, to implement the

could be considered across other parameters as well. This includes, for example, the horizon at which VaR and ES are estimated or the way in which forecasts are evaluated through a specific loss function to be selected by the user.

test is justified). Consequently, under the central limit theorem,

$$\sqrt{T}(T^{-1}\sum_{t=1}^{T} d_t - E(d_t)) \xrightarrow{D} N(0, \Omega), \tag{17}$$

where $\Omega = \lim_{T\to\infty} var(\sqrt{T}(T^{-1}\sum_{t=1}^{T} d_t - E(d_t))$.

The test is implemented using the stationary bootstrap by Politis and Romano (1994). B pseudo time series are generated from the data, using combinations of blocks from the original sample, where the length of the blocks is geometrically distributed. Here, 25,000 bootstrap replications are used. The optimal mean block length for the stationary bootstrap depends on the purpose of the bootstrapping. It should increase with the length of the time series T so that the asymptotic p-values are correct (Härdle et al., 2003). Modifying the approach of Pepe (2018), the mean block length used here is $T^{1/4}$, which is of the optimal order for determining the critical values of a one-sided hypothesis test (Hall et al., 1995), as the difference between the true and obtained probabilities of the test (known as the error in rejection probability) is minimized.

As an estimator for the bootstrap population variance, Hansen (2005) recommends using weighted empirical covariances,

$$\hat{\omega}^2 = \hat{\gamma}_0 + 2\sum_{i=1}^{T-1} \kappa(i)\hat{\gamma}_i, \tag{18}$$

where

$$\hat{\gamma}_i = \frac{1}{T}\sum_{j=1}^{T-i}(d_j - \bar{d})(d_{j+i} - \bar{d}), \tag{19}$$

using $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$, and

$$\kappa(i) = \frac{T-i}{T}(1-q)^i + \frac{i}{T}(1-q)^{T-i}. \tag{20}$$

For each bootstrapped time series $\{d^*_{b,t}\}$, the test statistic under the null hypothesis is constructed by defining: $Z^*_{b,t} = d^*_{b,t} - \bar{d}\cdot \mathbb{1}_{\{\bar{d}\geq\sqrt{-(\hat{\omega}^2/T)2\log(\log(T))}\}}$, $\bar{Z}^*_b = \frac{1}{T}\sum_{t=1}^{T} Z^*_{b,t}$, and

$$\mathcal{T}^{SPA*}_b = \max\{0, T^{1/2}\bar{Z}^*_b/\hat{\omega}\} \quad \text{for } b = 1, ..., B. \tag{21}$$

Given the test statistic for the original sample, $\mathcal{T}^{SPA}$, which is calculated in the same way as for the

bootstrapped samples, the p-value of the SPA test is then approximated as

$$\hat{p} = \sum_{b=1}^{B} \frac{\mathbb{1}_{\{\mathcal{T}_b^{SPA*} > \mathcal{T}^{SPA}\}}}{B}. \tag{22}$$

4.2 Ordinal Fragility Measures

The pairwise, relative performance indicators allow us to establish the fragility of each candidate market risk model based on its statistical comparison to some alternative model. Because we are considering three market risk models as potential candidates, i.e., BRW, GARCH, and IGARCH, there are nine possible comparisons to be carried out using the SPA test. In other words, the pairwise, relative measures assess the likelihood that a researcher comparing a candidate model against an alternative one will find different results when she changes the backtest parameters.

Table 2 reports the SR, i.e., the percentage of points in the backtesting parameter space where the null hypothesis of no worse performance than the candidate model against the alternative model is not rejected according to the SPA test when a standard 5 percent size level is adopted. Each row contains a pair defined by the candidate and alternative model. A model is adequate and resilient (i.e., not fragile) when its success rate is high. Very little fragility can be detected for most of the nine combinations, as the candidate model is classified as adequate over most of the points in our three-dimensional parameter space. For all the combinations but two, SR varies from 0.9833 (when BRW is the candidate model and IGARCH is the alternative) to 0.9860 (when BRW is the candidate model and GARCH is the alternative). The comparison between GARCH and IGARCH represents a notable exception. When the GARCH is the candidate model and the IGARCH is the alternative, the SR is 0.55; when the candidate model is the IGARCH and the GARCH is the alternative the SR is 0.64. This implies that the choice between IGARCH and GARCH will heavily depend on the backtest parameters.

As an example, Figure 4 shows the results of the SPA test comparing GARCH vs. IGARCH when the backtest parameters are let to vary over the three dimensions (test size, subsample size, fraction of subsample used to estimate the parameters). Blue points indicate that the null hypothesis of the SPA test was not rejected, while green points indicate that the null hypothesis was rejected. Visibly, the results of the SPA test seem to depend heavily on the size of the subsample and the proportion of the subsample that is used to estimate the model. As the fraction of the subsample that we use

Table 2: **Outputs of the SPA test for all model comparisons.** The success rate (SR) is given by the fraction of points where the null hypothesis (H_0) of superior predictive accuracy is not rejected.

	$\mathbf{H_0}$ **not rejected**	$\mathbf{H_0}$ **rejected**
BRW vs. GARCH	0.9860	0.0140
BRW vs. IGARCH	0.9833	0.0167
GARCH vs. BRW	0.9858	0.0142
GARCH vs. IGARCH	0.5476	0.4524
IGARCH vs. BRW	0.9838	0.0162
IGARCH vs. GARCH	0.6378	0.3622

to estimate the model increases, the null hypothesis is rejected more often, independently from the size of the subsample. Instead, when the fraction of the subsample that we use to estimate the model is low, we reject the null hypothesis for smaller subsamples but not for larger ones. As a matter of comparison, Figure 5 shows the same tridimensional plot for a SPA test comparing GARCH vs. BRW. Visibly, in this case, the null hypothesis is rejected for most of the points in the backtest parameter space.

Table 3 reports the three remaining binary, comparative fragility measures related to the comparison of BRW vs. GARCH, BRW vs. IGARCH, GARCH vs. BWR, GARCH vs. IGARCH, IGARCH vs. BWR, and IGARCH vs. GARCH.

Table 3: **Comparative fragility indicators.** LF denotes the local fragility indicator. RLF denotes the relative, local fragility measure. RA denotes the relative area. RLF and RA are computed relative to the simulated maximum instability given the success rate of the model pair.

	LF	**RLF**	**RA**
BRW vs. GARCH	0.0071	0.5086	0.2019
BRW vs. IGARCH	0.0081	0.4850	0.1831
GARCH vs. BRW	0.0076	0.5327	0.2207
GARCH vs. IGARCH	0.2040	0.4508	0.2007
IGARCH vs. BRW	0.0084	0.5194	0.1940
IGARCH vs. GARCH	0.1781	0.4915	0.2342

When a model is resilient in its backtesting performance, we expect that if the null hypothesis of superior predictive accuracy of the candidate model would not be rejected for a given point in the backtest parameter grid, it should also not be rejected for small perturbations of the parameters in κ. As the LF indicator measures the proportion of rejections close to a point where the candidate model is assessed as adequate by the SPA test, the smaller the indicator, the less fragile the model is. Table 3 shows that the historical simulation method displays strong resilience when the alternative

Figure 4: **Result of the SPA test with GARCH as the candidate model and IGARCH as the alternative model.** Blue points indicate the null hypothesis of the SPA test was not rejected, i.e., the candidate model is adequate (it performs at least as well as the alternative). Green points indicate the null hypothesis of the SPA test was rejected.

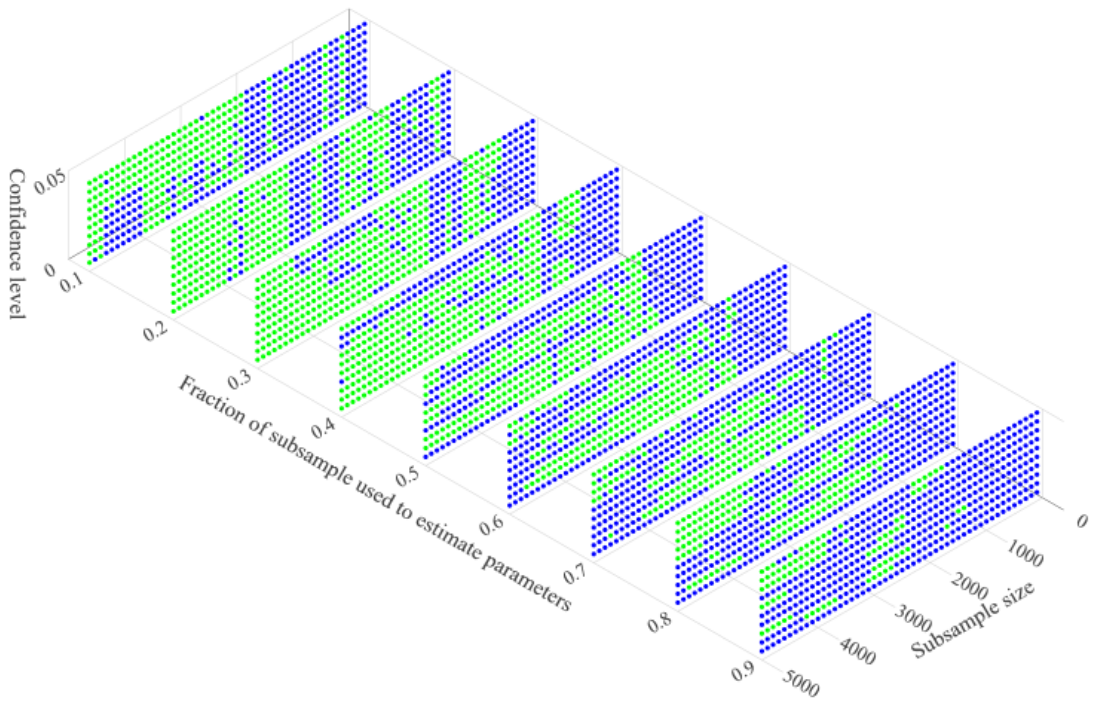
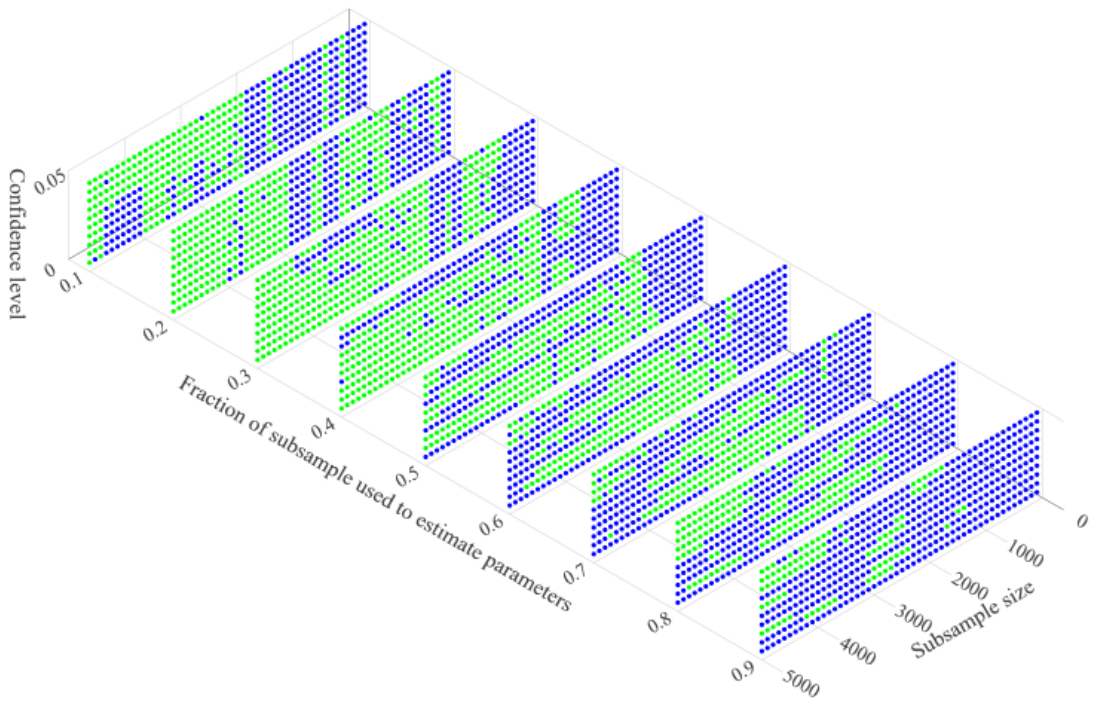


Figure 5: **Result of the SPA test with GARCH as the candidate model and BRW as the alternative model.** Blue points indicate the null hypothesis of the SPA test was not rejected, i.e., the candidate model is adequate (it performs at least as well as the alternative). Green points indicate the null hypothesis of the SPA test was rejected.



model is either GARCH or IGARCH (with an LF indicator of 0.0071 and 0.0081, respectively). In contrast, a researcher comparing IGARCH with GARCH will face an LF of approximately 0.20 and a researcher comparing IGARCH to GARCH will face an LF of approximately 0.20. The RLF measures the average fragility across the backtesting parameter combinations for which the candidate model is not worse than the alternative according to the SPA test, relative to the maximum possible fragility given the SR of the candidate. Notably, all the model pairs display an RLF of around 50%. However, RLF is mostly useful to discriminate among models with a similar level of local instability. In our case, the comparison between IGARCH and GARCH appears more fragile than the comparison between GARCH and IGARCH as the RLF is equal to 0.49 in the former case and to 0.45 in the latter. Finally, the RA indicator contains additional information about the spatial distribution of fragility. The larger the RA, the more fragile the SPA-driven performance comparison of the candidate model against the alternative. Again, the lowest RA is achieved by the BWR model, which seems to be adequate and robust when compared to both GARCH and IGARCH. Conversely, the comparison between IGARCH and GARCH reveals much more fragility, especially when IGARCH is taken as a candidate (the RA indicator is 0.23).

RA considers how the local model pseudo-OOS fragility may change when moving from focusing on the most stable regions of the parameter space to the whole backtesting space. A visual representation of how local fragility changes when we increase the fraction of the least fragile parameter combinations based on which it is computed is given in Figures 6, 7, 8, and 9.

Figures 4.2 and 4.2 show that both LF functions are steeply convex and in fact that they turn up and depart from the horizontal axis starting at levels of w around 0.9. Yet this occurs before the same effect shows up in the case of the maximum LF, which returns a modest but positive RA throughout.

Figures 4.2 and 4.2 show that LF can be close to piece-wise linear or even non-convex (in the latter case, when the LF function is quasi-concave for w in excess of 0.2) and massively lie on top and far away from the maximum LF function, thus returning much larger values of RA and hence evidence of widespread model fragility–of GARCH vs. IGARCH and especially when IGARCH is candidate model and it is compared with GARCH–over the backtesting parameter space, which should warn a user from resorting to either GARCH or IGARCH when there is substantial uncertainty as to their relative adequacy in tail risk prediction.

Figure 6: **Local Instability (LF) as a function of an increasing fraction of the least fragile points when BRW is the candidate model and GARCH is the alternative.** LF is shown in black and maximum LF is shown in red. The ratio between the areas under the LF and the maximum LF curve gives the RA of the candidate model.

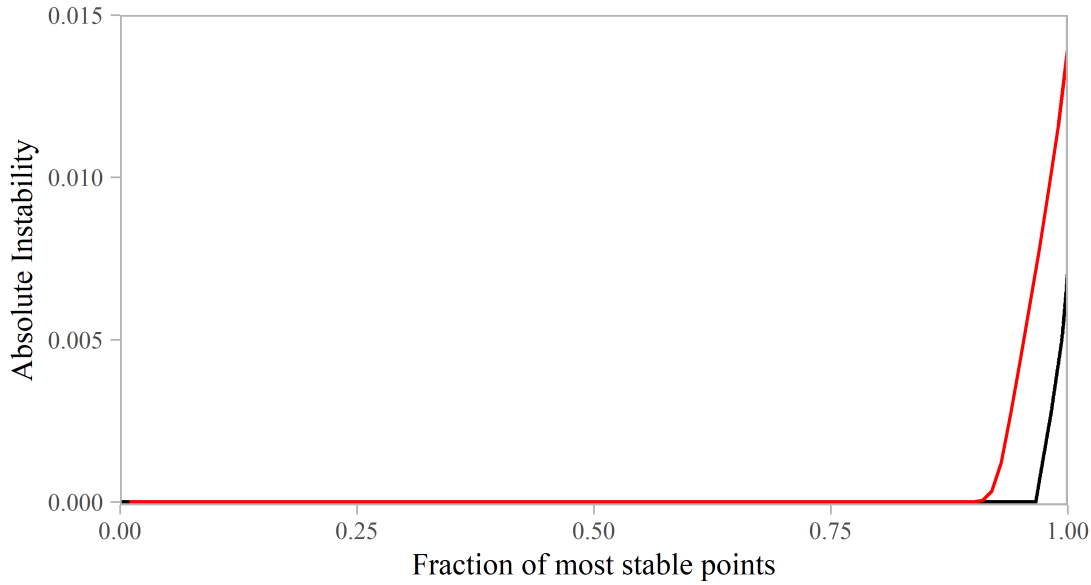


Figure 7: **Local Instability (LF) as a function of an increasing fraction of the least fragile points when BRW is the candidate model and IGARCH is the alternative.** The LF is shown in black. The maximum LF is shown in red and the ratio between the areas under the LF and the maximum LF curve gives the RA of the candidate model.

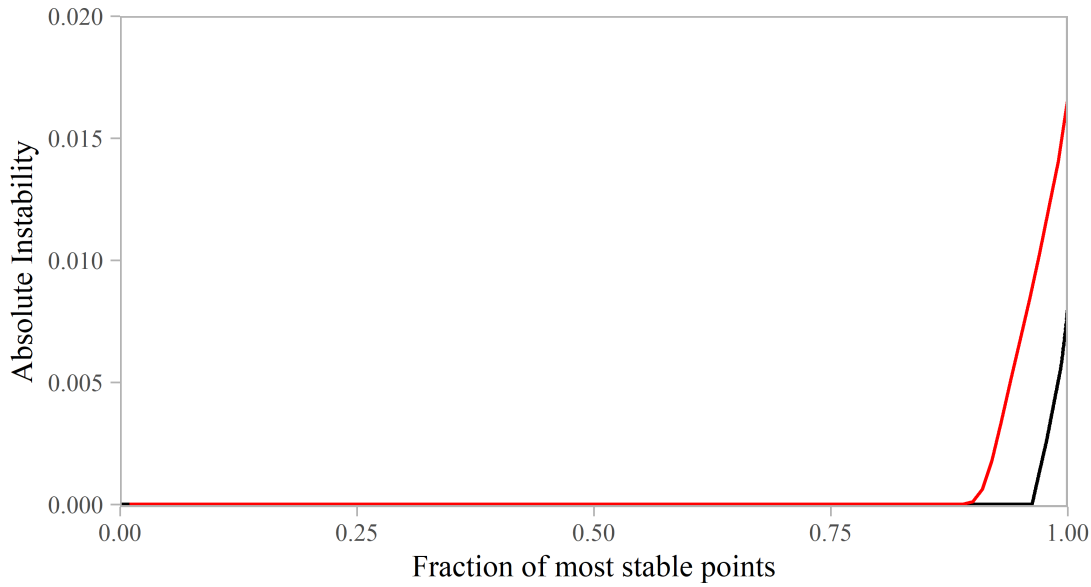


Figure 8: **Local Instability (LF) as a function of an increasing fraction of the least fragile points when GARCH is the candidate model and IGARCH is the alternative.** The LF is shown in black and the maximum LF is shown in red. The ratio between the areas under the LF and the maximum LF curve gives the RA of the candidate model.

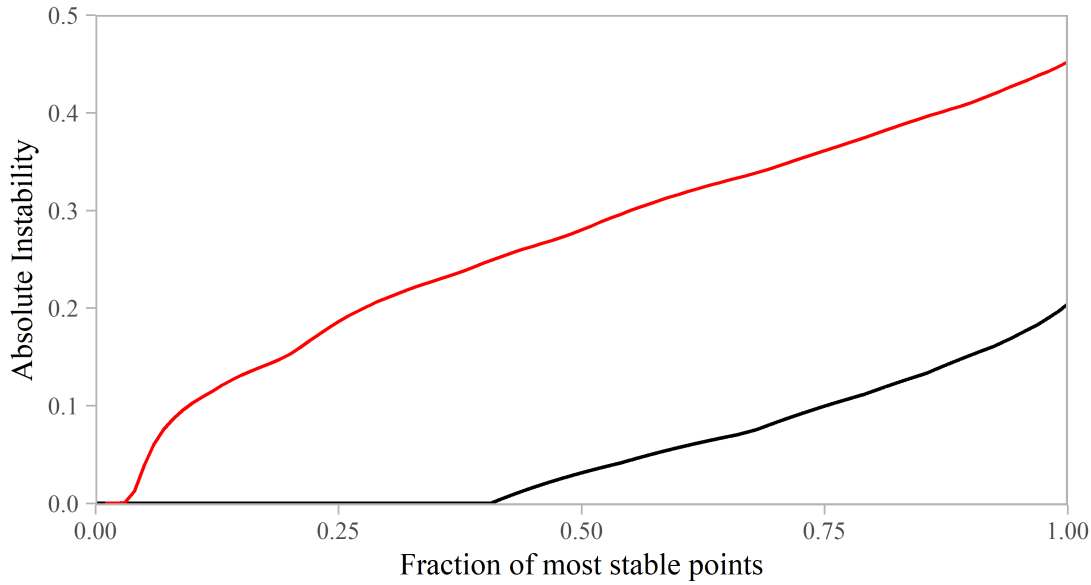
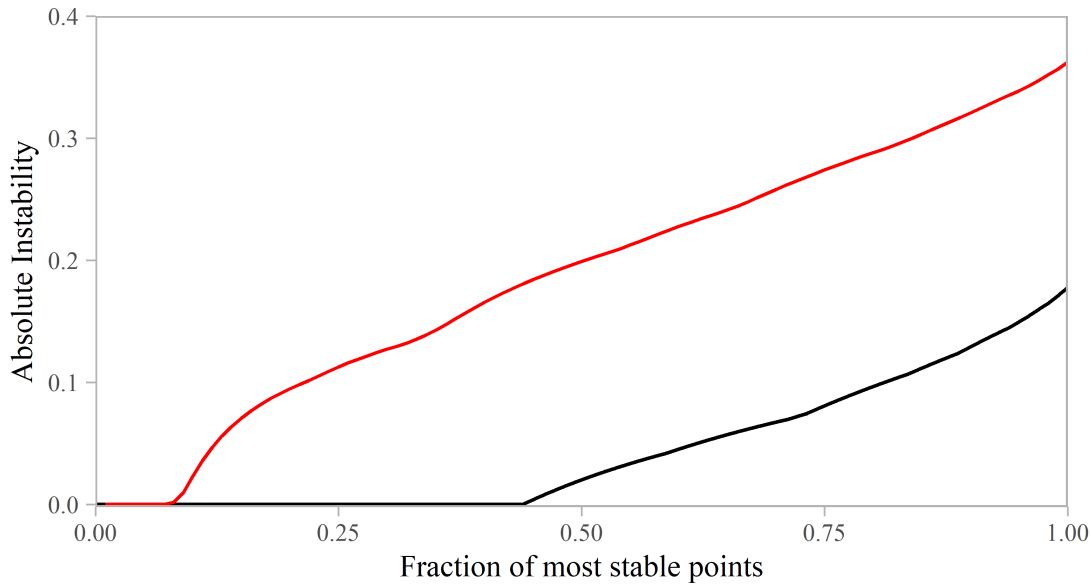


Figure 9: **Local Instability (LF) as a function of an increasing fraction of the least fragile points when IGARCH is the candidate model and GARCH is the alternative.** The LF is shown in black and the maximum LF is shown in red. The ratio between the areas under the LF and the maximum AI curve gives the RA of the candidate model.



4.3 Cardinal Fragility Measures

In the previous subsection, we discussed the fragility of a model from the point of view of a researcher that has to choose between two models, a candidate and an alternative, and wants to assess whether her choice will be robust to perturbations to the backtest parameters. However, the resulting comparative measures of fragility do not allow a researcher to produce a ranking of the models under analysis. In addition, they are based on a binary outcome rather than exploiting the information conveyed by the loss function that should be minimized. In this subsection, we compare historical simulations, GARCH, and IGARCH based on four absolute indices of fragility, namely ruggedness, mean semi-elasticity, the standard deviation of semi-elasticities, and the integrated semi-elastic radius. Table 4 summarizes the results.

Table 4: **Indices of fragility.** The table reports the values of ruggedness (RG), mean semi-elasticity (ME), the standard deviation of semi-elasticities (SDE), and of integrated semi-elastic radius (IE) for each of the three market models investigated.

	RG ($\times 10^{-5}$)	**ME** ($\times 10^{-5}$)	**SDE** ($\times 10^{-5}$)	**IE** ($\times 10^{-5}$)
BRW	1.9214	24.8912	32.2644	29.1790
GARCH	0.4914	7.7692	6.5815	9.1650
IGARCH	0.4334	7.2211	6.7891	8.7596

All the measures seem to consistently point toward the historical simulations as the most fragile model among the three, followed by GARCH and IGARCH. In particular, BRW has a RG index of $1.92x10^{-}5$, almost twice as much as the RG index of the GARCH and IGARCH models. This implies that the historical simulations experience large deviations in the loss function when we move from one point to another in the backtesting parameter space. This is clearly illustrated in Figure 10, which depicts the realized loss function over the parameter space for the BRW model. Visually, the presence of areas with different colors, especially when we increase the proportion of the data that is used to estimate the model, displays the enormous variability in the realized loss function across the backtest parameter space. For the sake of comparison, Figure 11 proposes a similar plot for the GARCH model. While the color changes from purple to dark blue as we move along the axes, the variations appear to be much more modest than in the case of the BRW. In addition, there are large areas of stability for intermediate values of all the backtest parameters.

The difference in the fragility of BRW compared to IGARCH and GARCH models is even starker

Figure 10: **Loss function across the parameter space for BRW.** The loss function is the quadratic deviation of the return from the Expected Shortfall if the Value-at-Risk is violated, and zero otherwise. The lower the value of the loss function, the better the predictions of the market risk model on average.

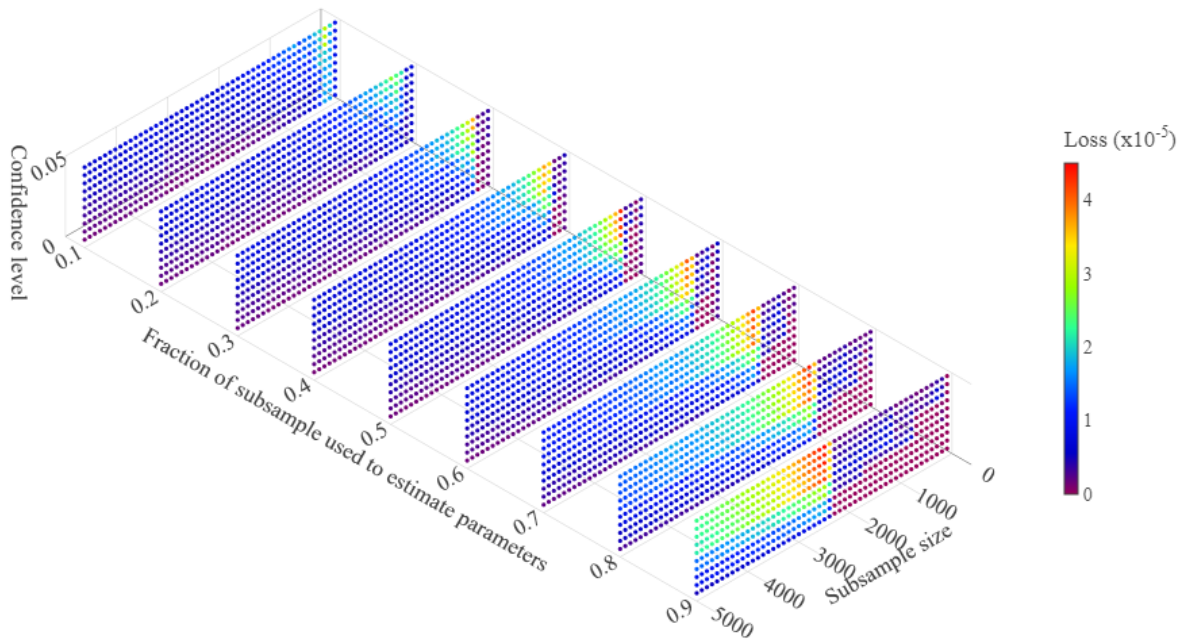
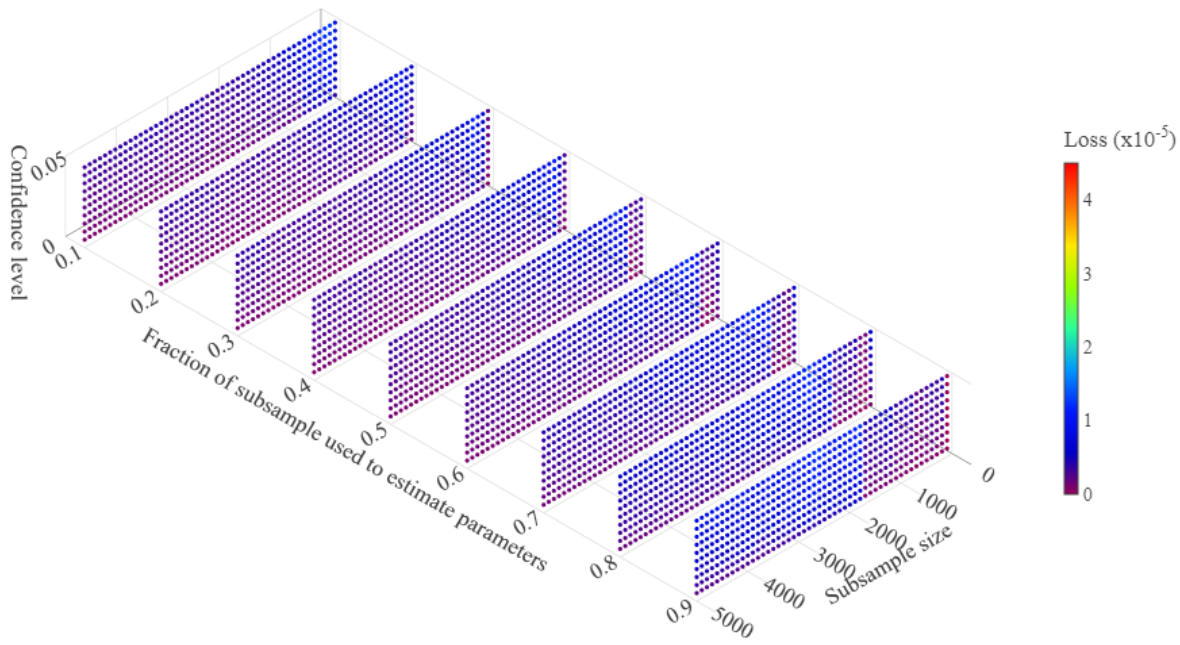


Figure 11: **Loss function across the parameter space for GARCH.** The loss function is the quadratic deviation of the return from the Expected Shortfall if the Value-at-Risk is violated, and zero otherwise. The lower the value of the loss function, the better the predictions of the market risk model on average.



when we consider the mean and standard deviation of semi-elasticities, which allow us to make a fair comparison across the different dimensions of the parameter space, which are otherwise expressed in different scales. While both GARCH and IGARCH have a ME of about $7x10^{-}5$, BRW displays a ME of $24.89x10^{-}5$, more than 3 times as much.

Finally, as outlined in Section 3, IE carries crucial information as it also takes into account the distribution of the changes in model performance across the backtesting parameter space. In fact, from a practical perspective, a model with areas of instability at the centre of the grid is more problematic than one that only displays instability for extreme values in the parameter space. Also in this case, the BRW is the most unstable model, with an IE that is three times as much as that of the other two models. This is unsurprising, given the patterns highlighted in Figure 10, with high areas of instability at the centre of the grid.

4.3.1 Robustness. To ensure the robustness of our metrics, we check that they deliver consistent rankings when we change our choices regarding the level of sparsity in the grid of parameters and the norm used to compute the semi-elasticities (as explained in Section 3). Table 5 reports the values of the RG, ME, SDE, and IE when computed on a parameter grid that is twice as sparse as the original one (i.e., the distance across adjacent points is twice as large as in the initial grid).

Table 5: **Indices of fragility calculated on the sparse parameter grid.** The table reports the values of the ruggedness (RG), the mean semi-elasticity (ME), the standard deviation of semi-elasticities (SDE), and the integrated semi-elastic radius (IE) for each of the three market models.

	RG ($\times 10^{-5}$)	**ME** ($\times 10^{-5}$)	**SDE** ($\times 10^{-5}$)	**IE** ($\times 10^{-5}$)
BRW	3.5996	21.5486	20.2468	24.7745
GARCH	0.8204	6.9390	4.4471	8.0550
IGARCH	0.7119	6.5730	4.3324	7.7169

Comparing Table 5 with 4, we notice that the magnitude of the values of the RG index increases significantly for all the models, when the sparser grid is used. This is expected, as the deviations in the loss between adjacent grid points reasonably become larger when the grid is more sparse. In contrast, the other indices generally decrease when the grid is sparse, but those changes are relatively smaller compared to changes in ruggedness. More importantly, the ranking of the models is always preserved across all the metrics when we use a sparser parameter space, demonstrating the robustness of our measures to the choice of the grid. This is an important advantage. To obtain the most precise

approximations of the ME, SDE, and IE, one should, in principle, have a parameter grid that is as finer as possible. However, increasing the number of parameter values over which the loss function is evaluated increases the computational burden. Therefore, robustness to the use of sparser grids is a convenient feature in practical terms. Finally, Table 6 reports the values of ME and IE when various norms (namely, L^1, L^2, and L^∞ norms) are adopted to compute the semi-elasticity. Notably, the values of ME and IE are very similar when different norms are employed and the ranking of the models remains unchanged across both measures.

Table 6: **Indices of fragility calculated using various norms to compute the semi-elasticity.** Values for the mean semi-elasticity (ME) and integrated semi-elastic radius (IE) are presented for each of the models, using the L^2 norm, the L^1 norm, and L^∞ norm.

	ME ($\times 10^{-5}$)			**IE** ($\times 10^{-5}$)		
Norm	L^2	L^1	L^∞	L^2	L^1	L^∞
BRW	24.8912	27.1955	24.4997	29.1790	31.6185	28.7876
GARCH	7.7692	8.6087	7.6733	9.1650	10.1208	9.0519
IGARCH	7.2211	7.9409	7.1394	8.7596	9.6108	8.6620

5 Conclusions

Backtesting is a powerful method to elicit from the data crucial insights as to whether a given candidate model is able to forecast accurately or to assist a user's decision adequately. However, a backtest requires several choices in terms of the type of data analyzed, the length of the sample and of the OOS period, the size of the statistical tests applied, etc. As no model can perfectly capture the data-generating process, the results of a backtesting exercise may heavily depend on such choices. In this paper, we develop a range of measures that capture the potential fragility of the forecasting performance of different models to alternative choices concerning the parameters of the backtesting exercises. For concreteness, our methodological development and application refer to tail risk estimation. Our novel measures are either of a binary type as based on comparative tests of predictive accuracy or absolute, when based on the stability of a loss function on vector of parameters of the backtesting space.

To illustrate the properties of these measures, we deploy them to assess the fragility of three models (BRW, GARCH, and IGARCH) that are routinely applied in risk management. When pairwise comparisons are performed, we find that BRW models display rather strong robustness in compari-

son to both GARCH and IGARCH. In fact, tail risk estimators based on historical simulations are generally not rejected against either the GARCH and IGARCH models when tested using the SPA approach in Hansen (2005). This result holds for most of the points in the backtest parameter space, as captured by our relative fragility measures. In contrast, the comparison between GARCH and IGARCH proves to be much less decisive as its outcomes tend to heavily depend on which of the two models is considered as the candidate vs. the alternative model.

Notably, the results are reversed when the absolute fragility measures are considered. In fact, the BRW model performs considerably worse than GARCH and IGARCH according to all the absolute instability measures, with values that are more than three times those of the other two models. While this result may seem contradictory, it illustrates well the different characteristics of relative vs. absolute fragility measures. In fact, while the former only consider violations of tail quantile forecasts, the latter measures consider the loss conditional on violations.

To the best of our knowledge, our paper represents the first attempt to formally measure the fragility of the predictive performance of risk models. Our work can be extended in several directions. First, in this paper, for the sake of demonstration, we only focus on three simple and widely applied market risk models. However, more sophisticated models should be considered, including more flexible models from the GARCH family (including semiparametric variants) as well as machine learning techniques, such as artificial neural networks. Second, the degree of complexity of the models (as measured by the number of parameters to be estimated) could also be considered as a parameter of choice in the backtesting space. Finally, while our application has been entirely focused on typical risk management estimates, other tail risk estimators and applications–or more generally, forecasting widely defined–could be studies as well.

References

Bhansali, V. (2008). Tail risk management. *The Journal of Portfolio Management*, 34(4):68–75.

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

Boudoukh, J., Richardson, M., and Whitelaw, R. F. (1997). The best of both worlds: a hybrid approach to calculating value at risk. *Risk*, 11(4):64–67.

Campbell, S. D. (2006). A review of backtesting and backtesting procedures. *Journal of Risk*, 9(2):1–17.

Christoffersen, P. (2011a). *Elements of financial risk management*. Academic press.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.

Christoffersen, P. F. (2011b). *Elements of Financial Risk Management*. Elsevier, Oxford, 2nd edition.

Cont, R., Deguest, R., and Scandolo, G. (2008). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10:593–606.

Crnkovic, C. and Drachman, J. (1996). 'Quality control' in VaR: understanding and applying Value-at-Risk. *Risk*, 9:139–143.

Gao, G. P., Lu, X., and Song, Z. (2019). Tail risk concerns everywhere. *Management Science*, 65(7):3111–3130.

Hall, P., Horowitz, J., and Jing, B. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Harvey, C. R. and Liu, Y. (2015). Backtesting. *The Journal of Portfolio Management*, 42(1):13–28.

Hull, J. (2012). *Risk management and financial institutions,+ Web Site*, volume 733. John Wiley & Sons.

Härdle, W., Horowitz, J., and Kreiss, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review*, 71(2):435–459.

Kelly, B. and Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, 27(10):2841–2871.

Kou, S., Peng, X., and Heyde, C. C. (2013). External risk measures and Basel accords. *Mathematics of Operations Research*, 38(3):393–417.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84.

Lopez, J. (1999). Methods for evaluating value-at-risk estimates. *Federal Reserve Bank of San Francisco Economic Review*, 2:3–15.

Pepe, P. (2018). Measuring model robustness and fragility in the back-testing of risk management indicators. Bocconi University, unpublished master's thesis.

Perignon, C. and Smith, D. (2010). The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance*, 34:362–377.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Power, M. (2004). The risk management of everything. *The Journal of Risk Finance*.

Pritsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking & Finance*, 30(2):561–582.

Riley, S., DeGloria, S., and Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, 5(1–4):23–37.

Žiković, S. and Aktan, B. (2011). Decay factor optimisation in time weighted simulation - evaluating var performance. *International Journal of Forecasting*, 27:1147–1159.