A study on window size selection for threshold and bootstrap Value-at-Risk models

Abstract

This paper investigates the effects of window size selection on various models for Value-at-Risk (VaR) forecasting using high performance computing. Subsequently, automated procedures using change-point analysis for optimal window size selection are proposed. In particular, stationary bootstrapping and the peaks-over-threshold methods are utilized for the rolling daily VaR estimation and are contrasted with the classical conditional Gaussian model. It is evidenced that change-point procedures can, on average, result in more adequate risk predictions than a predetermined fixed window size. The data sets analyzed include indices across 5 continents, i.e., the Dow Jones Industrial Average Index (DJI), the Financial Times Stock Exchange 100 Index (UKX), the NIKKEI Top 225 Index (NKY), the Johannesburg Stock Exchange Top 40 Index (JSE Top40), the Ibovespa Brazil Sao Paulo Stock Exchange All Index (IBOV), and the Bombay Stock Exchange Top 500 Index (BSE 500).

Keywords: Value-at-Risk; window size; threshold; bootstrap; change-point.

Key Messages:

- Analysis of the effects of window size selection in VaR forecasting.
- Conditional Gaussian, POT and stationary bootstrap models are considered.
- VaR forecasting is shown to be sensitive to window size selection.
- Change-point procedures are proposed as remedies for this issue.
- Six market indices, from five different continents, are analyzed.
1. Introduction

Value-at-Risk (VaR) is a statistic used to measure the volatility, and therefore the risk, of a given financial portfolio. This statistic is essentially a lower quantile, which quantifies the likelihood that a security return falls below a certain value, at a given confidence level, over a predetermined time period. Typically, historical observations are used to calculate an estimate of VaR. This estimation is however sensitive to the size of the historical window, or period, used. Too large a window will result in little variation in the statistic but will expose the statistic to model bias. Conversely, too small a window will result in a statistic which is sensitive to model changes but is very volatile. This presents portfolio managers with the challenge of selecting a window size to use in order to most accurately describe the risk of a given portfolio in the next period.

The Basel Accord II (Basel Committee on Banking Supervision 2006) laid out various quantitative standards to aid financial institutes in the evaluation of their internal risk models. For example, banks may use VaR to determine the amount of market risk capital to hold as a cushion against adverse market movements (McNeil and Frey 2000). One of the Basel II quantitative standards requires that the historical observation period, for the calculation of VaR, must be constrained to at least one year (i.e., approximately 250 trading days). However, the limited amount of study on the appropriate window size shows mixed results. For example, Hendricks (1996) observed that a long window (i.e., 1250 days) produced the best result in his analysis, relative to four other shorter windows. In contrast, Hoppe (1998) argued for the use of shorter windows (in some cases as short as 30 days) to produce better coverage results and to better align with the non-stationarity of the data.

Čížek, Härdle and Spokoiny (2009), Härdle, Hautsch and Mihoci (2015) and Schröder (2016) suggested the use of adaptive pointwise selection procedures for studying volatility, parameter variations and local trend estimation in financial time series. Our current work explores these ideas for the forecasting of VaR, as a remedy to the problem of window size selection.

The issue above is further complicated by the existence of various VaR models, with varying degrees of success under different scenarios. The cohort of such models is expanded by Basel III which relaxed some of the requirements on the effective window size (Basel Committee on Banking Supervision 2011). This paved way for the use of exponential weighted moving average models and the generalized autoregressive conditional heteroscedastic (GARCH) class of methods. In particular, GARCH-type filtered approaches, with non-parametric (e.g., historical simulation), parametric (e.g., Gaussian distribution), or semi-parametric (e.g., threshold models) procedures for describing the innovations, have become popular for estimating VaR (McNeil and Frey 2000; Orhan and Köksal 2012; Brandolini and Colucci 2012; Laker, Huang and Clark 2017).

The current paper contributes to the literature as follows. Firstly, a comprehensive analysis is performed for three popular VaR models, for fixed window sizes over the range from 500 to 1500 days (with increments of 50 days). This provides an overview of the effects of selecting different window sizes for the process of VaR forecasting. This analysis also identifies the optimal fixed window size. However, it requires high performance computing and may not be practical in real applications. Hence, as a further contribution, we explore the use of change-point analysis for window size selection. In particular, the at-most-one-change (AMOC) and the binary segmentation (BinSeg) algorithms are used to produce superior VaR backtesting results than the average performance of the fixed window size approach.
2. Methodology

Given a series of daily closing stock prices $p_1, p_2, p_3, \ldots$, we define the negative stock, or security, return for day $t$ as

$$R_t = -100 \times \ln \left( \frac{p_t}{p_{t-1}} \right).$$

The negative return of a stock is then decomposed into two parts: an expected return which is predicted by an autoregressive (AR) model (typically an AR(1) model) and an error term which captures deviations from the predicted value. This is presented as follows

$$R_t = E(R_t|\theta_{t-1}) + \varepsilon_t,$$

where $\theta_{t-1}$ denotes all the information about the return process up to day $t - 1$. We will assume a GARCH(1,1) process for the error term by defining

$$\varepsilon_t = \sigma_t Z_t \quad \text{and} \quad \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2,$$

where $\alpha_0 > 0$, $\alpha_1 > 0$ and $\beta > 0$. This is often referred to as a GARCH(1,1) filter and various assumptions may be conjectured about the innovations (or residuals) $Z_t$. Such assumptions can be utilised to estimate the GARCH parameters and produce forecasts for $E(R_{t+1}|\theta_t)$ and $\sigma_{t+1}$.

The VaR statistic is defined as an extreme tail quantile of a financial time series. VaR is said to be violated when the security return is in excess of the VaR quantile. The probability of getting such an extreme return, right of the lower limit of the upper tail of the assumed distribution of negative returns, is denoted by $\tau$, i.e.,

$$\Pr(R_t > \text{VaR}(1 - \tau)) = \tau.$$

Hence, using the above GARCH filter, we can deduce VaR for the next period as

$$\text{VaR}(1 - \tau) = E(R_{t+1}|\theta_t) + \sigma_{t+1} z_{\tau},$$

where $z_{\tau}$ is the 100$(1 - \tau)$-th quantile of the marginal distribution of $Z_t$. If $z_{\tau}$ is estimated by the Gaussian distribution, then we have the popular mean-variance formulation. When its estimate is obtained by bootstrapping, then the procedure is referred to as filtered historical simulation (FHS). The former is a parametric approach and the latter is non-parametric, both being commonly used in practice due to their simplicity in implementation and interpretation. However, we shall extend the standard FHS to cater for dependencies in the residuals by using stationary bootstrapping (Laker, Huang and Clark 2017). A third approach that is semi-parametric estimates $z_{\tau}$ indirectly through the modelling of the corresponding threshold exceedances of $Z_t$. This is termed the conditional peaks-over-threshold (POT) method, which has also gained popularity in recent years (McNeil and Frey 2000). Stationary bootstrapping and the POT method are briefly described below. These are followed by discussions on change-point algorithms (and how they may be used for window size selection) and backtests for VaR.

Stationary bootstrapping

Stationary bootstrapping is a generalization of the ordinary bootstrapping that subdivides the observations into blocks prior to resampling. It also assumes that the block length is a geometric
random variable. Firstly, a value \( p \in (0,1) \) is preselected, which can be optimally taken at \( c^{-1} n^{-1/3} \) for some constant \( c \) (Politis and Romano 1994). Subsequently, each observation is put through a decision rule in the order which they appear. For each observation, a number \( U \) is randomly drawn from the UNIF(0,1) distribution. This observation is then included in the present block if \( U \) is less than \( 1 - p \). Alternatively, if \( U \) is greater than \( 1 - p \), then a new block is started. This decision rule is continued until all the observations have been selected into blocks. This procedure caters for dependency in the observations and preserves stationarity. For our purposes, we shall apply the above to the realized GARCH innovations \( Z_t \) and calculate a bootstrapped value for \( z_t \).

**POT**

The POT method is known to be very useful in the tail estimation of data as this method is based on sound statistical theory and provides an asymptotic parametric form for the tail of the distribution (McNeil and Frey 2000). Furthermore, it is specifically useful for approximating the distribution of financial data due to the leptokurtic nature of the conditional distribution of the errors arising from the GARCH model (Orhan and Köksal 2012). More precisely, given a threshold \( \mu \), we can write the corresponding excess distribution of \( Z_t \) as

\[
F_\mu(y) = \Pr(Z_t - \mu \leq y | Z_t > \mu) = \frac{F(y + \mu) - F(\mu)}{1 - F(\mu)}
\]

where \( F \) is the distribution function of \( Z_t \) and \( 0 \leq y < z_0 - \mu \), for some \( z_0 \) defined as the right endpoint of \( F \). For a large class of distributions \( F \) (i.e., \( F \) in the domain of attraction of the extreme value distribution), there exists a measurable function \( \beta(\mu) \) such that

\[
\lim_{\mu \to z_0} \sup_{0 \leq y < z_0 - \mu} |F_\mu(y) - G_{\epsilon, \beta(\mu)}(y)| = 0
\]

where

\[
G_{\epsilon, \beta}(y) = \begin{cases} 
1 - (1 + ey/\beta)^{-1/\epsilon}, & \text{if } \epsilon = 0 \\
1 - \exp\left(-\frac{y}{\beta}\right), & \text{if } \epsilon \neq 0
\end{cases}
\]

is the generalized Pareto distribution (GPD). In other words, the distribution of threshold exceedances is asymptotically identified with the GPD (Embrechts, Klüppelberg and Mikosch 1997). This allows us, by reverting the arguments above, to calculate

\[
\hat{\epsilon}_t = \mu + \hat{\beta} \left( \left( \frac{1 - \tau k/n}{k/n} \right)^{\hat{\epsilon}} - 1 \right)
\]

where \( \hat{\beta} \) and \( \hat{\epsilon} \) are maximum likelihood estimates for \( \beta \) and \( \epsilon \), respectively, \( k \) is the number of exceedances above \( \mu \), and \( n \) is the number of observations.

**AMOC and BinSeg**

The GARCH model, used in this paper to model the volatility of financial data over time, is known to assume time homogeneity. This may not be appropriate when modelling financial data, as market and institutional changes result in changing variability of the data over time. When these changes are not accounted for, the accuracy of the model fit is compromised. As such, one can employ the technique of change-point analysis to achieve a more flexible model which is able to forecast data more
accurately over longer periods of time (Čížek, Härdle and Spokoiny 2009). This technique essentially aims to identify stationary intervals (with a fixed right endpoint) inside every estimation window. This is achieved by detecting the points at which the variability of the data changes most significantly (Čížek, Härdle and Spokoiny 2009).

The points at which the data’s distribution changes significantly can be found in numerous ways. A significant change in the mean or variance of the series can be monitored and the splits made accordingly. Two such approaches are implemented through the AMOC algorithm (Silva and Texeira 2008) and BinSeg algorithm (Scott and Knott 1974).

Initially, one assumes that the data series captured within the window is time homogeneous (i.e., the null hypothesis). This notion is then tested against there being a change-point required due to the non-homogeneity of the series within the window (i.e., the alternative hypothesis). The AMOC algorithm makes use of the likelihood ratio method, where the maximum likelihood is calculated under both hypotheses across different change-point positions within the selected window. The BinSeg method is a generalized version of the AMOC algorithm, where once a change-point is established the sequence is split and each segment is then again tested for a change point until the prescribed threshold is reached.

Automated Window Size Selection

An important facet of the current work is to identify ways to select an appropriate historic window for VaR forecasting, at each given point in time. This is to avoid the problem of having to artificially specifying a fixed window size in advance and the sensitivity of models to window size (as we shall explore later). This paper proposes the use of AMOC and BinSeg to identify change-points in terms of volatility (points for which the largest change in variance is identified). For each day to be forecasted, the procedure is implemented as follows: Let \( x \) be the day for which VaR is to be estimated and \( U \) and \( L \) are the maximum and minimum sizes (in days) allowed for the estimation window to be used. \( x - 1 \) is a fixed right endpoint, as the closest point to the day of interest. Then, AMOC or BinSeg is applied to identify a change-point in the interval \([x - U; x - L]\), say at \( y \). This results in the estimation window \([y; x - 1]\), which is then used for model estimation.

VaR Backtests

We employ two common backtests to evaluate the adequacy of VaR estimates, namely the Kupiec likelihood ratio test (Kupiec 1995) and the Christoffersen test (Christoffersen, Hahn and Inoue 2001). The former is an unconditional coverage test that compares the realized VaR violations against the specified VaR level (i.e., value of \( \tau \)). The second test is a conditional coverage test that examines the dependency amongst VaR violations (i.e., a robust VaR model should react quick enough so as to minimise the chance of consecutive VaR violations). The resulting p-values from both backtests are examined as a method to compare across different models.

3. Data Analysis

Our data consists of security returns calculated from daily closing prices of the past 16 years (January 2000 to November 2016), sourced from the Bloomberg database. Collecting information from international stock markets results in minor discrepancies in the number of available data points. To overcome this problem, the database of returns has been truncated to 4147 data points for each stock, ensuring that an equal amount of data points is recorded for each stock. Specifically, 6 securities have been chosen, namely the Dow Jones Industrial Average Index (DJI), the Financial Times Stock Exchange 100 Index (UKX), the NIKKEI Top 225 Index (NKY), the Johannesburg Stock Exchange Top 40 Index (JSE
Top40), the Ibovespa Brazil Sao Paulo Stock Exchange All Index (IBOV), and the Bombay Stock Exchange Top 500 Index (BSE 500). A brief description of these indices is given below.

- **DJI Index**: This index is a price-weighted average of 30 significant stocks traded on the NYSE and NASDAQ Stock Exchange.
- **UKX Index**: This index, informally known as the "Footsie", is a price-weighted index of the top 100 companies trading on the London Stock Exchange (LSE) where the companies are ranked according to market capital.
- **NKY Index**: This index is a price-weighted average of the top 225 stocks traded on the Japanese Nikkei Stock Exchange.
- **JSE Top 40 Index**: This index is made up of the 40 largest companies trading on the Johannesburg Stock Exchange (JSE), ranked by market capitalization.
- **IBOV Index**: This index is a gross total return index weighted by market value to the free float. It is comprised of the most liquid stocks on the Sao Paulo Stock Exchange.
- **BSE 500 Index**: This index is a broad representation of the overall performance of the Indian Bombay Stock Exchange (BSE), with the index representing 93% of the total market capitalization and covering 20 major industries in the market. The top 500 stocks listed in the market are represented in the index, with each stock being weighted according to free-float market capitalization weight.

We will apply three VaR models to all 6 return series, namely GARCH with Gaussian innovations (GARCH-Gaussian), GARCH with stationary bootstrapping (GARCH-Boot), and GARCH with POT (GARCH-POT). The last two are implemented by using pseudo maximum likelihood estimation for the GARCH parameters. This is to minimize distributional assumptions of the innovations (McNeil and Frey 2000). As a preliminary analysis, the GARCH model is fitted to the complete horizon of data points in each index and descriptive statistics are drawn from both the original returns and the innovations. These are displayed in Table 1.

| Data               | Mean  | Std. Dev. | Min.  | Max.  | Excess Kurtosis | Skewness |
|--------------------|-------|-----------|-------|-------|-----------------|----------|
| DJI                | 0.005 | 0.503     | -3.561| 4.564 | 8.309           | -0.059   |
| DJI Innovations    | 0.004 | 0.999     | -4.313| 4.428 | 0.907           | 0.136    |
| UKX                | 6.45e-4 | 0.529     | -4.024| 4.076 | 6.209           | -0.149   |
| UKX Innovations    | 0.038 | 0.999     | -3.522| 4.418 | 0.705           | 0.211    |
| NKY                | -4.80e-4 | 0.674     | -5.260| 5.748 | 6.010           | -0.395   |
| NKY Innovations    | 0.029 | 0.999     | -5.570| 4.964 | 0.908           | -0.043   |
| JSE Top 40         | 0.016 | 0.582     | -3.645| 3.347 | 3.423           | -0.131   |
| JSE Top 40 Innovations | 0.006 | 1.000     | -5.366| 5.593 | 1.003           | 0.101    |
| IBOV               | 0.011 | 0.790     | -5.253| 5.940 | 3.702           | -0.072   |
| IBOV Innovations   | 0.006 | 0.999     | -3.684| 4.746 | 0.456           | 0.080    |
| BSE 500            | 0.014 | 0.665     | -5.403| 6.348 | 7.794           | -0.643   |
| BSE 500 Innovations| 0.011 | 0.999     | -5.071| 8.419 | 1.810           | 0.004    |

From Table 1 it can be seen that all return series, irrespective of market maturity, have a near-zero mean and a negatively skewed distribution. This is fairly typical in financial data and is explained by either the return distribution peaking at values larger than the mean with a long tale to the negative end of the return scale, or merely that the higher returns out-weigh the lower returns to the left of the scale.

The excess kurtosis is positive for all markets, with particularly high values being recorded for the BSE 500 Index and the DJI Index. These kurtosis values indicate sharply peaked distributions, with fatter
tails than that of the Gaussian distribution. The extent of this deviation from the Gaussian distribution seems to be larger in the developed market. On the other hand, the innovation excess kurtosis values are much lower than that of the original returns. As expected, this is one effect of the GARCH model catering for the changing volatility. However, when applying the rolling-window mechanism there is the possibility for the excess kurtosis to vary, depending on which window of data has been selected. This phenomenon is explained in Figure 1 below for a rolling window of size 1500 days. It is evidenced that the rolling innovation excess kurtosis can often surge significantly above zero. This justifies the need to cater for conditional heavy tails. Interestingly, the common spike at around the 2000th observation corresponds to the 2008 global credit crunch.

![Figure 1: Rolling excess kurtosis of GARCH innovations using a window size of 1500 days.](image)

We now perform daily 95% VaR forecasts (using all three models described earlier) for each return series using a fixed number of observations in each widow (i.e., a fixed window size). With the aid of high performance computing, this is done for all window sizes from 500 to 1500 (at increments of size 50) for each index. The parameter $p$ for stationary bootstrapping is determined by setting $c$ to 3.15, which is obtained by Monte Carlo simulation using the Gaussian AR(1) process. Each window is bootstrapped 1000 times. The threshold $\mu$ for the POT approach is set to be the 90% quantile of the innovations in the corresponding window. The resulting $p$-values for the Kupiec and Christoffersen tests, for each window size and each index are displayed by the figures in Appendix A. It is interesting to note that the commonly used window sizes, e.g., 500 and 1000, does not always produce the optimal results. The window size for the highest $p$-value also varies across different models, indices and backtests. The results are also summarized in Tables 2 to 7. Column 3 records the average of all $p$-values obtained across different fixed window sizes, while column 4 gives the standard deviation of these $p$-values. Columns 5 and 6 presents the smallest and biggest $p$-value obtained across all window

---

1 It should be noted (as mentioned by a reviewer) that one needs to be careful here not to fall into the trap of the $p$-value fallacy. For the $p$-value to be interpreted in a more direct manner, they need to be calibrated (Sellke, Bayarri and Berger 2001). It is not our aim to interpret them in this way. Having that said, under some conditions, the comparison of $p$-values for ranking model performances is still viable. Indeed it should be noted here that for each test, our sample size (not the same as window size) is identical and we are primarily interested in identifying the best model within each financial index (rather than comparing directly across indices). Such comparisons are still widely used in the literature (see, for example, Zoia, Biffi and Nicolussi 2018; Yu et al. 2018; Sampil, Hasim and Dai 2018).
sizes, with the window size that produced the biggest p-value given in the last column. It is clear that results produced by the two commonly used backtests can drastically change according to the choice of window size.

Interestingly, the models (on average) seem to perform the best for JSE Top 40. This may be attributed to the fact that this index produced the least excess kurtosis. This also resulted in a relatively smaller coefficient of variation (C.V.) for the different p-values across different window sizes. Hence, JSE Top 40 appears to be the least sensitive (relative to other indices) to window size selection. In contrast, UKX and NKY produced significantly higher C.V. values, indicating a high sensitivity to window size selection.

| Innovation       | VaR Test | Mean   | Std. Dev. | Min. | Max.   | C.V. | Optimal Window Size |
|------------------|----------|--------|-----------|------|--------|------|---------------------|
| Gaussian         | Kupiec   | 0.442  | 0.124     | 0.305| 0.746  | 0.281| 500                 |
|                  | Christ.  | 0.370  | 0.104     | 0.235| 0.693  | 0.281| 500                 |
| POT              | Kupiec   | 0.390  | 0.123     | 0.199| 0.617  | 0.314| 550/1500            |
|                  | Christ.  | 0.562  | 0.150     | 0.281| 0.800  | 0.267| 550/1500            |
| Stationary Bootstrap | Kupiec | 0.409  | 0.079     | 0.231| 0.556  | 0.192| 1100/1200           |
|                  | Christ.  | 0.349  | 0.093     | 0.235| 0.556  | 0.265| 1100               |

| Innovation       | VaR Test | Mean   | Std. Dev. | Min. | Max.   | C.V. | Optimal Window Size |
|------------------|----------|--------|-----------|------|--------|------|---------------------|
| Gaussian         | Kupiec   | 0.159  | 0.182     | 0.026| 0.680  | 1.143| 600                 |
|                  | Christ.  | 0.122  | 0.117     | 0.022| 0.404  | 0.955| 500                 |
| POT              | Kupiec   | 0.347  | 0.259     | 0.123| 0.954  | 0.748| 600                 |
|                  | Christ.  | 0.400  | 0.219     | 0.196| 0.882  | 0.547| 500                 |
| Stationary Bootstrap | Kupiec | 0.225  | 0.159     | 0.049| 0.617  | 0.704| 550                 |
|                  | Christ.  | 0.227  | 0.161     | 0.044| 0.594  | 0.711| 550                 |

| Innovation       | VaR Test | Mean   | Std. Dev. | Min. | Max.   | C.V. | Optimal Window Size |
|------------------|----------|--------|-----------|------|--------|------|---------------------|
| Gaussian         | Kupiec   | 0.062  | 0.059     | 0.021| 0.231  | 0.957| 550                 |
|                  | Christ.  | 0.028  | 0.033     | 0.007| 0.123  | 1.160| 550                 |
| POT              | Kupiec   | 0.119  | 0.105     | 0.040| 0.395  | 0.881| 600                 |
|                  | Christ.  | 0.071  | 0.086     | 0.016| 0.351  | 1.217| 500                 |
| Stationary Bootstrap | Kupiec | 0.197  | 0.104     | 0.060| 0.445  | 0.528| 550                 |
|                  | Christ.  | 0.087  | 0.057     | 0.009| 0.252  | 0.656| 550                 |

| Innovation       | VaR Test | Mean   | Std. Dev. | Min. | Max.   | C.V. | Optimal Window Size |
|------------------|----------|--------|-----------|------|--------|------|---------------------|
| Gaussian         | Kupiec   | 0.815  | 0.103     | 0.631| 0.975  | 0.127| 850/1350/1450       |
|                  | Christ.  | 0.728  | 0.181     | 0.536| 0.975  | 0.249| 850                 |
| POT              | Kupiec   | 0.920  | 0.063     | 0.764| 0.975  | 0.068| 600/700/750/1150/1400/1450 |
|                  | Christ.  | 0.713  | 0.197     | 0.486| 0.987  | 0.276| 800/850             |
| Stationary Bootstrap | Kupiec | 0.843  | 0.105     | 0.631| 0.975  | 0.124| 500/600/900/1050/1300 |
|                  | Christ.  | 0.844  | 0.149     | 0.569| 0.987  | 0.176| 650                 |

Table 2: Summary statistics of backtest p-values for DJI across different window sizes.

Table 3: Summary statistics of backtest p-values for UKX across different window sizes.

Table 4: Summary statistics of backtest p-values for NKY across different window sizes.

Table 5: Summary statistics of backtest p-values for JSE Top 40 across different window sizes.
Table 6: Summary statistics of backtest p-values for IBOV across different window sizes.

| Innovation       | VaR Test | Mean  | Std. Dev. | Min.  | Max.  | C.V.  | Optimal Window Size |
|------------------|----------|-------|-----------|-------|-------|-------|---------------------|
| Gaussian         | Kupiec   | 0.241 | 0.100     | 0.123 | 0.445 | 0.414 | 600                 |
|                  | Christ.  | 0.341 | 0.100     | 0.193 | 0.567 | 0.294 | 650                 |
| POT              | Kupiec   | 0.465 | 0.233     | 0.145 | 0.814 | 0.502 | 500/700             |
|                  | Christ.  | 0.306 | 0.162     | 0.104 | 0.710 | 0.530 | 500                 |
| Stationary Bootstrap | Kupiec | 0.610 | 0.166     | 0.348 | 0.975 | 0.272 | 900                 |
|                  | Christ.  | 0.548 | 0.131     | 0.349 | 0.795 | 0.239 | 900                 |

Table 7: Summary statistics of backtest p-values for BSE 500 across different window sizes.

| Innovation       | VaR Test | Mean  | Std. Dev. | Min.  | Max.  | C.V.  | Optimal Window Size |
|------------------|----------|-------|-----------|-------|-------|-------|---------------------|
| Gaussian         | Kupiec   | 0.027 | 0.009     | 0.011 | 0.040 | 0.336 | 500/900/1400/1450    |
|                  | Christ.  | 0.060 | 0.018     | 0.036 | 0.111 | 0.300 | 900                 |
| POT              | Kupiec   | 0.145 | 0.047     | 0.072 | 0.199 | 0.326 | 1050/1200/1250/1300/1400/1450/1500 |
|                  | Christ.  | 0.166 | 0.053     | 0.066 | 0.260 | 0.318 | 600                 |
| Stationary Bootstrap | Kupiec | 0.180 | 0.106     | 0.032 | 0.348 | 0.591 | 1100/1500           |
|                  | Christ.  | 0.243 | 0.121     | 0.099 | 0.511 | 0.499 | 1100                |

On average (considering the mean p-value again), POT and stationary bootstrap approaches seem to produce more robust VaR estimates than the conditional Gaussian model. This is with the exception of DJI, where the three models were very similar in performance. In terms of the maximum p-value achieved at the optimal window size, the results again varied across different indices. For UKX, the best performance is dominated by POT, whereas the difference between POT and the Gaussian approaches is indecisive for DJI (likewise, POT and stationary bootstrap produced similar results for NKY). For IBOV and BSE 500, the best model under optimal window size is clearly achieved by stationary bootstrapping. All three models gave almost identical VaR backtesting results for JSE Top 40, at the optimal window sizes.

At this point, it is imperative to note several shortcomings. Firstly, the VaR backtesting results varied across different indices, models and window sizes. This create a dilemma for practitioners in deciding on an appropriate approach. Secondly, the models seem to be largely sensitive to window size selection. Although the optimal window sizes are determined as part of the process in this analysis, high performance computing is required to record such results. There is also no universal choice for the best window size. These make the approach impractical in some applications. At the same time, these optimal window sizes may be restricted only to the time horizon in the current sets of data.

To overcome the issues above, we propose and implement change-point procedures in each estimation window, with a fixed right endpoint (as described in the Methodology Section). AMOC and BinSeg are both utilized, where the number of splits for BinSeg is capped at three. The minimum window size is set to be 500 and the maximum is 1500 (i.e., default value if no change-point is identified). This procedure is automated without the need to run the models through all window sizes and hence requires significantly less computing power. Table 8 summarizes the results after implementing the change-point procedures. The highest p-values for each index and backtest are highlighted in bold. Columns 3 and 4 records the actual number of VaR exceedances, as produced by implementing AMOC and BinSeg, respectively. The results of the two backtests are then recorded in columns 5 to 8.
Table 8: Backtest results at 95% VaR using varying window sizes via change-point analysis.

| Index | Innovation | Actual Exceedances (Expected = 132) | Kupiec Test | Christ. Test |
|-------|------------|--------------------------------------|-------------|-------------|
|       |            | AMOC | BinSeg | AMOC | BinSeg | AMOC | BinSeg |
| DJI   | Gaussian   | 141  | 137    | 0.445 | 0.680    | 0.450 | 0.639    |
|       | POT        | 144  | 143    | 0.305 | 0.348    | 0.315 | 0.358    |
|       | Stat. Boot.| 147  | 144    | 0.199 | 0.305    | 0.204 | 0.315    |
| UKX   | Gaussian   | 139  | 138    | 0.556 | 0.617    | 0.316 | 0.349    |
|       | POT        | 136  | 133    | 0.746 | **0.954**| 0.671 | **0.773**|
|       | Stat. Boot.| 138  | 138    | 0.617 | 0.617    | 0.583 | 0.583    |
| NKY   | Gaussian   | 149  | 148    | 0.145 | 0.170    | 0.072 | 0.087    |
|       | POT        | 144  | 144    | **0.305**| **0.305**| **0.309**| **0.309**|
|       | Stat. Boot.| 146  | 144    | 0.231 | 0.305    | 0.123 | 0.168    |
| JSE Top 40 | Gaussian | 130  | 131    | 0.834 | **0.904**| 0.965 | 0.971    |
|       | POT        | 134  | 137    | 0.883 | 0.680    | **0.984**| 0.857    |
|       | Stat. Boot.| 131  | 131    | **0.904**| **0.904**| 0.976 | 0.976    |
| IBOV  | Gaussian   | 136  | 136    | 0.746 | 0.746    | 0.671 | 0.671    |
|       | POT        | 133  | 135    | **0.954**| 0.814    | 0.501 | 0.710    |
|       | Stat. Boot.| 128  | 131    | 0.697 | 0.904    | 0.571 | **0.810**|
| BSE 500 | Gaussian | 155  | 156    | 0.049 | 0.040    | 0.074 | 0.060    |
|       | POT        | 151  | 147    | 0.103 | **0.199**| 0.162 | 0.199    |
|       | Stat. Boot.| 151  | 150    | 0.103 | 0.123    | 0.162 | **0.260**|

First, one can clearly observe that, in almost all cases, the result produced by the change-point procedures are more robust than the average p-value obtained by fixed window sizes in Tables 2 to 7. This is strong evidence that implementing a change-point procedure is more useful than a randomly chosen (or selected-by-convenience) fixed window size. More importantly, given the appropriate innovation assumption or treatment, the corresponding model can produce similar results as those arising from having previously identified the optimal fixed window sizes. These make the proposed procedure practically more attractive.

As expected, the Gaussian innovation assumption again produced the least robust model in most indices, except for DJI and JSE Top 40. For JSE Top 40, the results are mixed across the two different backtests, with comparable results between models. More interestingly, the conditional Gaussian model with BinSeg seems to be the best approach for DJI.

### 4. Conclusion

In this paper, we studied the effects of window size selection on three popular VaR models, namely the conditional Gaussian model, the conditional POT model and a generalized FHS method. This was done by utilizing high performance computing to run the models through a range of window sizes for six different indices. The level of sensitivity across different markets is varied. However, without doubt, the choice of the window size plays a vital role in the performance of these models.

Although the optimal window sizes were determined from the above process, the approach is in general not practical. This is due to the large disparity in terms of model choice, the reliance on computing power and the non-existence of a universal optimal window size. At the same time, the optimal window size determined here may be restricted to the current time horizon analyzed. To overcome these problems, we have proposed and implemented change-point procedures to identify
significant changes in volatility within the estimation windows. This then allows us to specify appropriate windows with a fixed right endpoint. The results show that the combination of the different models with change-point procedures are capable of producing adequate VaR forecasts. In particular, they produce (on average) more robust results than randomly selected, or selected-by-convenience, window sizes (such as the traditional 500 or 1000 day windows). More importantly, given the appropriate treatment of the innovations, one can achieve similar backtesting results as in the case where optimal window size is known.

Further investigations could possibly include extending the approach to utilize other adaptive pointwise estimation methods (Čížek, Härdle and Spokoiny 2009) and performing similar analysis on portfolios of public companies. The combination of a switching model strategy (Chiu and Chuang 2016) with varying window sizes may be another interesting exploration. It is also the aim of the authors to extend the study to other risk measures, such as conditional VaR and entropic VaR.
Figure A1: Kupiec (bottom) and Christoffersen (Top) p-values for DJI over different fixed window sizes.
Figure A2: Kupiec (bottom) and Christoffersen (Top) p-values for UKX over different fixed window sizes.
Figure A3: Kupiec (bottom) and Christoffersen (Top) p-values for NKY over different fixed window sizes.
Figure A4: Kupiec (bottom) and Christoffersen (Top) p-values for JSE Top 40 over different fixed window sizes.
Figure A5: Kupiec (bottom) and Christoffersen (Top) p-values for IBOV over different fixed window sizes.
Figure A6: Kupiec (bottom) and Christoffersen (Top) p-values for BSE 500 over different fixed window sizes.
Declaration of Interest

The authors report no conflicts of interest.

Acknowledgments

The authors would like to thank the two anonymous reviewers for their critical review and suggestions that helped improve this work. Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: http://hpc.uct.ac.za

References

Basel Committee on Banking Supervision. (2006). International convergence of capital measurement and capital standards. Bank for International Settlements. http://www.bis.org/publ/bcbs128latest.pdf.

Basel Committee on Banking Supervision. (2011). Revisions to the Basel II market risk framework. Bank for International Settlements. http://www.bis.org/publ/bcbs193.pdf.

Brandolini, D. & Colucci, S. (2012). Backtesting value-at-risk: a comparison between filtered bootstrap and historical simulation. Journal of Risk Model Validation, 6(4), 3-16. doi:10.21314/JRMV.2012.094

Chiu, Y. & Chuang, I. (2016). The performance of the switching forecast model of value-at-risk in the Asian stock markets. Finance Research Letters, 18, 43-51. doi:10.1016/j.frl.2016.03.019

Čížek, P., Härdle, W. & Spokoiny, V. (2009). Adaptive pointwise estimation in time-inhomogeneous conditional heteroscedasticity models. The Econometrics Journal, 12(2), 248-271. doi:10.1111/j.1368-423X.2009.00292.x

Christoffersen, P., Hahn, J. & Inoue, A. (2001). Testing and comparing value-at-risk measures. Journal of Empirical Finance, 8(3), 325-342. doi:10.1016/S0927-5398(01)00025-1

Embretich, P., Klüppelberg, C. & Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance. Springer: Berlin.

Härdle, W. K., Hautsch, N. & Mihoci, A. (2015). Local adaptive multiplicative error models for high-frequency forecasts. Journal of Applied Econometrics, 30(4), 529-550. doi:10.1002/jae.2376

Hendricks, D. (1996). Evaluation of value-at-risk models using historical data. FRBNY Economic Policy Review, 2(1), 39-69. doi:10.2139/ssrn.1028807

Orhan, M. & Köksal, B. (2012). A comparison of GARCH models for VaR estimation. Expert Systems with Applications, 39(3), 3582-3592. doi:10.1016/j.eswa.2011.09.048

Politis, D. N. & Romano, J. P. (1994). The stationary bootstrap. Journal of the American Statistical Association, 89(428), 1303-1313. doi:10.1080/01621459.1994.10476870

Sampid, M. G., Hasim, H. M. & Dai, H. (2018). Refining value-at-risk estimates using a Bayesian Markov-switching GJR-GARCH copula EVT model. PLoS, 13(6), e0198753. doi:10.1371/journal.pone.0198753

Schroeder, A. L. (2016). Methods for change-point detection with additional interpretability. PhD thesis, The London School of Economics and Political Science (LSE). http://etheses.lse.ac.uk/id/eprint/3421
Scott, A. J. & Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3), 507-512. doi:10.2307/2529204

Sellke, T., Bayarri, M. J. & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1), 62-71. doi:10.1198/000313001300339950

Silva, E. G. & Texeira, A. A. (2008). Surveying structural change: Seminal contributions and a bibliometric account. *Structural Change and Economic Dynamics*, 19(4), 273-300. doi:10.1016/j.strueco.2008.02.001

Yu, W., Yang, K., Wei, Y. & Lei, L. (2018). Measuring value-at-risk and expected shortfall of crude oil portfolio using extreme value theory and vine copula. *Physica A: Statistical Methods and its Applications*, 490, 1423-1433. doi:10.1016/j.physa.2017.08.064

Zoia, M. G., Biffi, P. & Nicolussi, F. (2018). Value at risk and expected shortfall based on Gram-Charlier-like expansions. *Journal of Banking & Finance*, 93, 92-104. doi:10.1016/j.jbankfin.2018.06.001