# Backtest Criteria for the Quantile Correction under Model Risk

Mr. Siridej Putsorn

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Finance
Department of Banking and Finance
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2015

# เกณฑ์การทดสอบย้อนหลังสำหรับการแก้ไขค่าควอนไทล์ภายใต้ความเสี่ยงของแบบจำลอง

นายสิริเดช พุฒซ้อน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการเงิน ภาควิชาการธนาคารและการเงิน
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

| | |
|---|---|
| Thesis Title | Backtest Criteria for the Quantile Correction under Model Risk |
| By | Mr. Siridej Putsorn |
| Field of Study | Finance |
| Thesis Advisor | Assistant Professor Sira Suchintabandid, Ph.D. |

Accepted by the Faculty of Commerce and Accountancy, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

____________________________Dean of the Faculty of Commerce and Accountancy
(Associate Professor Pasu Decharin, Ph.D.)

THESIS COMMITTEE

____________________________________________Chairman
(Associate Professor Sunti Tirapat, Ph.D.)

____________________________________________Thesis Advisor
(Assistant Professor Sira Suchintabandid, Ph.D.)

____________________________________________Examiner
(Associate Professor Seksan Kiatsupaibul, Ph.D.)

____________________________________________External Examiner
(Kridsda Nimmanunta, Ph.D.)

สิริเดช พุฒซ้อน : เกณฑ์การทดสอบย้อนหลังสำหรับการแก้ไขค่าควอนไทล์ภายใต้ความเสี่ยงของแบบจำลอง (Backtest Criteria for the Quantile Correction under Model Risk) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.สิระ สุจินตะบัณฑิต, 4 หน้า.

Value-at-Risk (VaR)) เป็นเครื่องมือทางสถิติชนิดหนึ่งที่ใช้ในการพยากรณ์ความเสี่ยงทางการเงิน แต่วิกฤตการณ์ทางการเงินที่ผ่านมาได้บ่งบอกถึงจุดอ่อนที่สำคัญอย่างหนึ่งของ VaR นั่นคือ การวิเคราะห์ความเสี่ยงด้วย VaR ตกอยู่ภายใต้ความเสี่ยงของแบบจำลอง (Model RIsk) ในปัจจุบัน การแก้ปัญหาดังกล่าวได้ถูกแบ่งออกเป็นสองวิธีหลัก วิธีแรก คือ การทดสอบย้อนหลัง (Backtest) เป็นการนำค่า VaR ที่คำนวณในระยะเวลาหนึ่งๆที่กำหนดในอดีตมาเปรียบเทียบกับผลตอบแทนของหลักทรัพย์อ้างอิงที่เกิดขึ้นจริง ณ ช่วงเวลาเดียวกัน เพื่อพิจารณาว่า VaR สามารถครอบคลุมความเสี่ยงที่เกิดขึ้นได้ดีเพียงใด วิธีที่สอง คือการแก้ไขค่า VaR เพื่อลดความเสี่ยงของแบบจำลอง (Model Risk ซึ่งวิธีหนึ่งที่เพิ่งถูกคิดค้นมาได้ไม่นาน คือ การแก้ไขค่าความเสี่ยงแบบจำลองโดยการใช้ผลลัพธ์จากการทดสอบย้อนหลัง (Backtest) โดยแบบทดสอบย้อนหลัง จะถูกนำมาเป็นเกณฑ์ (Criteria) ในการปรับค่า VaR ให้มีความเหมาะสมมากขึ้น วิทยานิพนธ์ฉบับนี้ ได้พัฒนาวิธีแก้ไข VaR ดังกล่าว โดยการรวบรวมวิธีการทดสอบย้อนหลัง (Backtest) ต่างๆจากการวิจัยในอดีตเพื่อนำมาสร้างเป็นเกณฑ์ใหม่ๆในการปรับค่า VaR ซึ่งเลือกจากการพิจารณาความสามารถทางสถิติในการทดสอบแบบจำลอง VaR สำหรับผลการวิจัย พบว่า VaR ที่ถูกปรับค่าด้วยความเสี่ยงแบบจำลองแล้วมีศักยภาพในการวัดความเสี่ยงมากกว่า ค่า VaR เดิมในหลายกรณี สำหรับการเลือกเกณฑ์ (Criteria) ในการปรับค่า VaR นั้น เกณฑ์ที่ประกอบด้วยแบบทดสอบย้อนหลังที่มีประสิทธิภาพมากกว่าจะให้ผลลัพธ์ที่ดีกว่าเมื่อใช้แบบจำลองที่ไม่ซับซ้อน (Static VaR Models)

| | | | |
|---|---|---|---|
| ภาควิชา | การธนาคารและการเงิน | ลายมือชื่อนิสิต | |
| สาขาวิชา | การเงิน | ลายมือชื่อ อ.ที่ปรึกษาหลัก | |
| ปีการศึกษา | 2558 | | |

# # 5783047226 : MAJOR FINANCE
KEYWORDS: BACKTEST / QUANTILE FORECASTING / MODEL RISK

SIRIDEJ PUTSORN: Backtest Criteria for the Quantile Correction under Model Risk. ADVISOR: ASST. PROF. SIRA SUCHINTABANDID, Ph.D., 4 pp.

The fact that financial risks cannot be exactly determined but have to be estimated make Value-at-Risk (VaR) models less reliable. Thus far, VaR-model risk have gained increasing concerns and have been addressed in two general ways. The first way is to evaluate risk models using statistical tests, called backtests. In particular, backtests employ a comparison of VaR series and realized returns in the specified period to examine whether risk estimates are appropriate or not. The second way is adjusting VaR for model risk, which one of the recently proposed frameworks is the quantile correction method via the outcome of backtesting. Set of backtest methods are chosen for being adjustment criteria by considering three desirable properties of VaR models, namely, unconditional coverage, independence, and magnitude of violations (losses that exceed VaR). This thesis extend the general quantile correction framework by applying various backtest methods focusing on their statistical power of backtests shown by authors. Five standard data generating models (DGMs) are used to compute VaR of Stock Exchange of Thailand (SET) index daily returns. The results from ex post validation show that model-risk-adjusted series provide better results than original VaR in many cases. With regards to criteria sets, higher-statistical-power backtest criteria sets outperform their counterparts when static VaR models are used.

| | | | |
|---|---|---|---|
| Department: | Banking and Finance | Student's Signature | |
| Field of Study: | Finance | Advisor's Signature | |
| Academic Year: | 2015 | | |

# ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Sira Suchintabundid, for the continuous support of my study, for his time, patience, and immense knowledge in his quantitative fields. His guidance not only helped me in all the time of research and writing, but also for my presentation which is of utmost importance for my working life. I could not have imagined having a better advisor and mentor for my thesis

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Sunti Tirapat and Prof. Seksan Kiatsupaibul, for their insightful comments and encouragement, and especially for their time provided. Also, I am really grateful to Prof. Kridsda Nimmanunta for being the outside examiner and gave me various perspectives.

Last but not the least, I would like to thank my parents, my brother and sister for supporting me spiritually throughout writing this thesis. I also thank my master student’s fellow for the discussions.

# CONTENTS

Page

THAI ABSTRACT ........ iv

ENGLISH ABSTRACT ........ v

ACKNOWLEDGEMENTS ........ vi

CONTENTS ........ vii

LIST OF TABLES ........ ix

LIST OF FIGURES ........ x

CHAPTER 1: Introduction ........ 1

1.1 Background and Significance of the Problem ........ 1

1.2 Research Questions ........ 5

1.3 Objective and Contributions ........ 5

1.4 Research Hypotheses ........ 7

1.5 Organization of the Paper ........ 9

CHAPTER 2: Literature Review ........ 10

2.1 Value-at-Risk ........ 10

2.1.1 Definition ........ 10

2.1.2 Model Classes ........ 10

2.1.3 Critique ........ 13

2.2 Quantile Model Risk ........ 13

2.2.1 Definition ........ 14

2.2.2 Relevant Researches ........ 15

2.3 Model risk correction via Backtesting Framework ........ 16

2.3.1 Fundamental of Backtests ........ 17

2.3.2 The Original Approach of Boucher et al. (2014) ........ 20

2.4 Other Backtesting Methods for the Criteria ........ 25

2.4.1 Unconditional Coverage Test (UC) ........ 25

2.4.2 Independence Test (IND) ........ 27

2.4.3 Conditional Coverage Test (CC) ........ 29

2.4.4 Magnitude Test (MG) ........ 31

Page
2.5 Finite Sample Inference ........ 32
CHAPTER 3: Data and Methodology ........ 34
3.1 Data ........ 34
3.2. Methodology ........ 34
3.2.1 Criteria sets Construction ........ 34
3.2.2 Ex Post Validation ........ 40
CHAPTER 4: Empirical Result ........ 43
4.1 Model risk Calibrated by Individual (Backtest) Criterion ........ 43
4.2 Model risk Calibrated by Criteria Sets ........ 46
4.3 Ex Post Validation ........ 52
4.3.1 Out-of-Sample Backtest ........ 52
4.3.1.1 Overlapping 250-day Backtest Period ........ 52
4.3.1.2 Non-Overlapping 100-day Backtest Period ........ 55
4.3.2 Risk Ratio Analysis ........ 57
CHAPTER 5: Conclusion ........ 59
REFERENCES ........ 61
VITA ........ 76

# LIST OF TABLES

Page

Table 1. Set of New Criteria Sets Constructed For Hypothesis 1-2. ..................... 38

Table 2. Statistical Power of the Proposed Backtests Claimed by Authors........... 39

Table 3. Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Individual Criterion........................................................................................ 43

Table 4. Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set ............................................................................................... 47

Table 5. Maximum of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set.............................................................................................. 48

Table 6. Ex Post Validation Results for Hypothesis 1 (Overlapping) ................... 54

Table 7. Ex Post Validation Results for Hypothesis 2 (Overlapping) ................... 54

Table 8. Ex Post Validation Results for Hypothesis 1 (Non-overlapping)............ 56

Table 9. Ex Post Validation Results for Hypothesis 2 (Non-overlapping)............ 56

# LIST OF FIGURES

Page

Figure 1. SET Daily Negative Returns and Daily 95% VaR Computed by EWMA Model.......................................................................................... 177

Figure 2. SET Daily Negative Returns and Daily 95% VaR Computed by EWMA Model with Highlighted Violation Clustering.............................................. 18

Figure 3. Boucher et al. (2014)’s Conceptual Framework.................................... 22

Figure 4. Methods to Find $q_t^*$ ................................................................... 23

Figure 5. Backtest Methods for Model Risk Correction framework. .................... 35

Figure 6. $q_t^*$ for Criteria Set a1 Relative to 95% EVaR ........................................ 49

Figure 7. $q_t^*$ for Criteria Set a2 Relative to 95% EVaR ........................................ 49

Figure 8. $q_t^*$ for Criteria Set a3 Relative to 95% EVaR ........................................ 50

Figure 9. $q_t^*$ for Criteria Set a4 Relative to 95% EVaR ........................................ 50

Figure 10. $q_t^*$ for Criteria Set a5 Relative to 95% EVaR ...................................... 50

Figure 11. $q_t^*$ for Criteria Set a1 Relative to 95% EVaR with SET index ............. 51

Figure 12. Acceptance Frequency Ratio for all criteria sets and EVaR (Overlapping) ........................................................................................... 53

Figure 13. Acceptance Frequency Ratio for all criteria sets and EVaR (Non-overlapping)........................................................................................... 55

Figure 14. Risk Ratio of all criteria sets (RaVaR) and EVaR ............................... 57

## LIST OF APPENDIX, FIGURES, SCHEMES

Page

Table A1. Maximum of Positive and Negative $q_t^*$ Relative to 95% VaR using Individual Criterion 65

Table A2. Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set (sub-sample 1: the 1997's crisis) 66

Table A3 Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set (sub-sample 2: the 2008's crisis) 67

Table A4. Ex post validation of RaVaR/EVaR series for 95% VaR on SET index (full sample) 68

Table A5. Ex post validation of RaVaR/EVaR series for 95% VaR on SET index (sub-sample 1: the 1997's crisis) 69

Table A6. Ex post validation of RaVaR/EVaR series for 95% VaR on SET index (sub-sample 2: the 2008's crisis) 70

Figure B1. $q_t^*$ for Criteria Set b1 Relative to 95% EVaR 71

Figure B2. $q_t^*$ for Criteria Set b2 Relative to 95% EVaR 71

Figure B3. $q_t^*$ for Criteria Set b3 Relative to 95% EVaR 71

Figure B4. $q_t^*$ for Criteria Set b4 Relative to 95% EVaR 72

Figure B5. $q_t^*$ for Criteria Set b5 Relative to 95% EVaR 72

Figure B6. $q_t^*$ for Criteria Set c1 Relative to 95% EVaR 72

Figure B7. $q_t^*$ for Criteria Set c2 Relative to 95% EVaR 73

Figure B8. $q_t^*$ for Criteria Set c3 Relative to 95% EVaR 73

Figure B9. $q_t^*$ for Criteria Set c4 Relative to 95% EVaR 73

Figure B10. $q_t^*$ for Criteria Set c5 Relative to 95% EVaR 74

# CHAPTER 1: Introduction

## 1.1 Background and Significance of the Problem

Quantile-based risk assessment, with particular reference to "Value-at-Risk (VaR)", is of paramount importance in market risk management. Internal-model approach for regulatory capital reserve have been based on VaR ever since the 1996 Market Risk Amendment to the Basel Accord. In general, VaR tells us the potential maximum loss expected to incur on a portfolio given a degree of statistical confidence and forecast horizon. For instance, 1-day 99% VaR of $1M means that over a day we are 99% confident that portfolio loss will not higher than $1M. Due to its prevalence, substantial literature of VaR modeling and its financial applications have been documented overtime.

However, the experience from 2008 financial crisis have questioned the VaR accuracy as it played a negative role by distorting the true level of risk. A key reason is that risk models are subject to model risks, e.g. estimation risk and misspecification risk. While there is no unique definition of model risk, it typically occurred when we do not know true data generating process (DGP). As a consequence, if model risk is highly present, VaR tend to be incapable of predicting change in risk dynamics, i.e., underestimating risk in period before the crisis and overestimating risk post-crisis.

Regarding this concern, a number of literature that deals explicitly with VaR model risk have been proposed (see e.g., Bao and Ullah, 2004; Christoffersen and Goncalves, 2005; Kerkhof et al., 2010; Alexander and Sarabia, 2012; Boucher and Mailet, 2013; Boucher et al., 2014). Of particular interest, Boucher et al. (2014) provide a remedy for VaR model risk by proposing a general framework for model risk by allowing models to learn from its past errors dynamically through backtesting process.

In general, backtest is a statistical procedure to evaluate accuracy of VaR models by simulating rolling-window VaR forecasts on past data, (250 days is usually applied in regulatory framework). Then, the process goes through comparing VaR forecasts and realized return over a given backtest horizon. For any day (i.e. in daily basis), if negative returns exceed VaR, in other words, VaR cannot cover the losses, it is called “violation”. Next, sequence of violations are to be examined whether or not the VaR model represents risk appropriately using statistical inference. Roughly speaking, accurate VaR models should meet three desirable (violation) properties: (i) Unconditional coverage (UC) - the probability of realizing violations should be precisely α x 100%; (ii) Independence (IND) - violations should be independent; and (iii) Magnitude (MG) - magnitude of VaR violations should be appropriate, not too small or too large. In addition, when both unconditional coverage and independence properties are combined to be a joint hypothesis, it is known as Conditional coverage (CC) property (See e.g., Campbell, 2005; Haas, 2001; for a review of backtests).

More precisely, the aim of the Boucher et al. (2014)’s framework is to approximately quantify VaR model risk and add that amount into VaR forecasts to obtain risk measures that are robust to model risk, called “Model-risk-adjusted VaR (RaVaR)”. In particular, if VaR model doesn’t pass (accept the null hypothesis) all given backtests criteria, the correction amount is the minimal constant that, when added to all recent 250 daily-VaR observations, make it passes all chosen backtests. That correction amount indicate the proxy of model risk’s magnitude. But if the model already passes all chosen backtests, no correction would be required. For the sake of simplicity, Boucher et al. (2014) used three basic backtests as the adjustment criteria, one accounts for each violation property, namely Kupiec (1995) likelihood ratio test for unconditional coverage property (UC), Christoffersen (1998) first-order Markov test for independence property (IND), and Berkowitz (2001) test for magnitude property (MG).

In my point of view, the general framework of Boucher et al. (2014) provides an effective guideline for model risk quantification and correction calibrated using general backtest procedure. However, it should be noted that the estimated magnitude of model risk in Boucher's framework could be varied with different sets of (backtest) criteria. For instance, to test null hypothesis of a particular VaR-violation property (UC, IND, CC or MG), there are several backtest methods that can be used, especially for IND and CC tests which have been given increased significance over time.

Although Kupiec (1995) UC test and Christoffersen (1998) IND test (used in Boucher's framework) are commonly applied in practical applications, they're well known of its drawback. The most obvious example is that Christoffersen (1998) (first-order) Markov test cannot detect more than one-day (two-consecutive-day violations) independence. Indeed, in financial data, it could be that probability of having violations today is dependent on whether or not there is violation in one week ago, one month ago, and so on. Hence, many backtest methods have been proposed to overcome this limitation, including duration-based test (Christoffersen and Pelletier, 2004), GMM duration based-test (Candelon et al., 2011) dynamic quantile test (Engle and Manganelli, 2004), and Monte Carlo based-test (Ziggel et al., 2014).

For Kupiec (1995) UC test, it has been shown in some papers that its statistical power in detecting inaccurate models is quite low[1] and is outperformed by some UC tests, i.e., GMM duration based-test (Candelon et al., 2011) and Monte Carlo based-test (Ziggel et al., 2014). In addition, since violation clustering (dependence) seems to be more severe in financial data, CC tests tend to be superior to UC tests and are main focuses of recent literature (e.g., Engle and Manganelli, 2004; Candelon et al., 2011; Berkowitz et al., 2011; Ziggel et al., 2014).

These drawbacks on backtesting framework provide a motivation of my thesis. To fill this gap, I apply other backtest methods in literature into Boucher's model risk correction framework aiming to make the outcome of risk forecasts more accurate. In order to evaluate performance of different (backtest) criteria sets, my study also contribute to the literature by providing a two simple ex post validation methods, namely, out-of-sample backtest which based on general backtest procedures, and risk ratio analysis (Danielsson et al., 2014) which is another way to gauge model risk. In ex post validation process, using SET index data, series of RaVaR generated from each criteria set will be examined

[1] In backtesting literature, simulation experiment is generally used to show that when testing an exactly inaccurate model, backtests that have "more statistical power" in detecting inaccurate model are backtests that have "greater chance to reject the null hypothesis (infer that the model is inaccurate). See, for instance, Campbell (2005), Christoffersen and Pelletier (2004), Candelon et al. (2011), etc.

## 1.2 Research Questions

First of all, three backtest criteria (namely Kupiec (1995), Christoffersen (1998), and Berkowitz (2001)) used in Boucher et al. (2014)'s leaves an extension on trying other backtest methods. By focusing mainly on the drawbacks of Kupiec (1995), Christoffersen (1998) tests, the main research question is: *Do criteria sets that contain higher-statistical-power backtests provide better performance in ex post analysis?* Also, to adjust model risk explicitly for model risk, it is worth considering whether to adjust or not the model-risk-adjusted VaR (RaVaR) series outperform the original VaR series (estimated VaR or EVaR) in expost validation analysis? Lastly, following Boucher (2014) that model risk consists of estimation and misspecification risk, which the second part is generally much larger. By accepting that we never know the true data generating process (DGP) of market indexes, an additional question is: *Does amount of model risk generated from the correction framework depend on which methods (models) used to compute VaR*?

## 1.3 Objective and Contributions

The primary objective of this thesis is to strengthen importance of model risk on quantile modeling. According to Kerkhof et al. (2010), Boucher and Maillet (2013), and Boucher et al. (2014), appropriated risk measures that are in particular robust to statistical model risks is of key importance for enhancing market risk management. Although in Basel 3.5 discussion (BCBS, 2013) the regulators indicate that banks assess model risk, there is no explicit guideline. In this regard, I will put forward the Boucher et al. (2014) model risk correction framework by applying a range of other VaR backtest methods in literature, for each particular violation property, that is, Unconditional Coverage (UC) test, Independence (IND) test, and Conditional Coverage (CC) tests.

As an extension of Boucher et al. (2014)'s framework, this thesis contributes to a way to make model risk quantification more precise and obtain safer buffer on VaR forecasts explicitly account for model risk. The experience from global financial crisis as well as other meltdowns have questioned to academics and practitioners on VaR's capability of indicating risks. Although dealing explicitly with model risk will not overcome all VaR drawbacks, it could help solve the problem. By adjusting VaR estimates, hopefully they would be better to absorb changes in market conditions, especially at the turning point of change in market conditions i.e., calm-to-turbulent period, and vice versa.

This study also contributes to the literature by providing two simple ex-post validation methods. The first method called "out-of-sample backtest" use general backtest procedures to evaluate series of model-risk-adjusted VaR (RaVaR). The second method applies the idea of Risk-ratio analysis proposed by Danielsson et al. (2014) as an alternative approach to measure VaR-model risk. Lastly, regarding the amount of model risk ($q_t^*$), with more accurate model risk quantification, ones can be benefited from comparing various data generating models more effectively using the idea that the lower the model risk, the better the model.

## 1.4 Research Hypotheses

To answer the main research question, explaining how higher-statistical-power backtests can affect the model risk quantification and correction may be needed. First of all, backtests that have higher-statistical power means that it has greater chance to reject an inaccurate VaR-model, thus having low probability of incurring Type II error – falsely rejecting the model when the model is inaccurate (or rejecting null hypothesis when the alternative is true). Although backtesting also involve probability of rejecting an accurate VaR-model or Type I error, Gaglianone et al., (2011) and Jorion, (2001) said that in risk management, it is much more costly if the tests have low power to reject inaccurate model. Recall that model risk correction framework is based on idea that, for any time t, if VaR-model already pass all given backtests in a criteria set, there is no model risk and hence no correction will be added, and the opposite is true when there is at least one backtest rejecting the model. Let us consider the case that all chosen backtests in a criteria set are of very low-statistical power. In the presence of inaccurate VaR-model and thus some adjustment (explicitly account for model risk) is needed, but all backtests in the criteria set is so ineffective that they’re falsely accepting the inaccurate model (and no correction is added). That is a sup-optimal outcome that we should intensely be avoided. Because, for example, if the (optimal) correction amount is negative value, meaning that VaR should be more negative (or more conservative) to address model risk, but we potentially ignore, then the resulting VaR will be too low, and vice versa. Hence, the first hypothesis can be written as

**Hypothesis 1**: "Criteria sets that contain higher statistical-power backtests will outperform in ex post validation". (Details are given in methodology section.)

Apart from three violation-property tests, unconditional coverage (UC), independence (IND), and magnitude (MG) tests which were focused in Boucher's (three-backtest) criteria set, there is another violation-property type that can be backtested called conditional coverage (CC) test. It is a test that jointly detect both UC property and IND property at the same time. In this case ones may wonder how CC tests are different from two separate tests (UC and IND tests) in statistical power aspect. The answer is it depends.

According to (Røynstrand et al., 2012) and Campbell (2005), four possible outcomes can appear in financial data regarding UC and IND properties:

1. At a particular time, if VaR-model violate both UC and IND property[2], then a joint or conditional coverage (CC) test have more statistical power (less chance of incurring type II error) to detect this inaccurate model than either UC or IND test alone.
2. At a particular time, if VaR-model violate UC property but not violate IND property, then IND test alone have more statistical power to detect this inaccurate model than conditional coverage (CC) test.
3. At a particular time, if VaR-model violate IND property but not violate UC property, then UC test alone have more statistical power to detect this inaccurate model than conditional coverage (CC) test.
4. At a particular time, if both UC and IND properties are not violated, then the problem will be about type I error instead, which is of much less concerns.

[2] "Violate" here means that VaR model is inaccurate in term of particular property (UC, IND, or both UC and IND) which appropriate backtest methods should reject the model. Please do not be confused about the word "violations", which is the days that VaR cannot cover the loss).

Although we do not know for sure whether one of these situations are going to happen, in practice, violations clustering seems to appear overtime, and so does the “first outcome”. Thus, to strengthen the buffer on VaR forecast in model risk correction framework, I also hypothesize that adding a CC test in a criteria set may improve the performance of risk forecasts in ex post validation.

**Hypothesis 2:** "Adding CC test in a criteria set will improve performance in ex post validation". (Details are given in methodology section.)

### 1.5 Organization of the Paper

In the following: Chapter 2 gives an review of VaR model risk, relevant sesearch, the original quantile correction framework, and other backtest methods proposed in the literature. Chapter 3 illustrates the data used in this study as well as the methodology how I apply other backtest criteria to the original framework. Chapter 4 shows the results of model risk correction and ex post validation. Then, Chapter 5 finally concludes.

# CHAPTER 2: Literature Review

## 2.1 Value-at-Risk

Value at Risk (VaR) has become a standard risk measure since 1996, when financial institutions are allowed to use it as internal-model approach for capital reserve computation under the first Basel Accord (BCBS, 1996). People are well-known of its main advantages that it is easily to understand. Although it has some criticisms, people are still working with VaR models until now.

### 2.1.1 Definition

VaR's mathematical definition can be expressed as α-level quantile of profit and loss (P&L)'s distribution:

$$\mathrm{VaR}_t(\alpha) = \inf\{x \in R : \Pr(X \leq x) > \alpha\} \qquad (1)$$

For instance, 99%VaR is an amount of capital required to cover the loss in 99% of the time. In other words, it is 99% sure that returns won't be less than the over a given horizon. Apart from model specifications, there components are to be assumed for $\mathrm{VaR}_\alpha$ computation, which are confidence level, forecast horizon (also called liquidity horizon), and estimation period (generally regulators require at least 1-year trading days data).

### 2.1.2 Model Classes

Basically there are three main methods for VaR calculation (Jorion, 2001; Alexander, 2008; Hull, 2012), namely Gaussian parametric (or variance-covariance) approach, historical simulation, and Monte Carlo simulation. Over the century, there have been a number of VaR modelling methodologies proposed. However, in my study, model choice is restricted on ones that commonly used in practice, including: 1) Gaussian parametric approach, 2) Historical simulation (HS), 3) RiskMetrics, 4) GARCH(1,1)-N and 5) GARCH(1,1)-t. Descriptions for each model are given below:

Gaussian parametric approach (or Normal VaR method): Given daily time-series data, to forecast $\mathrm{VaR}_{t+1}$ at time t, the first and simplest way is by assuming the logarithmic returns are normally distributed. Then VaR is calculated by:

$$\mathrm{VaR}_{t+1}(\alpha) = \mu_t - z_\alpha \cdot \sigma, \qquad (2)$$

where, $z_\alpha$ is one-sided critical value for standard normal distribution, σ is historical volatility. Note that generally we assume $\mu_t$= 0 for daily returns.

Historical simulation: Another simplest way is using non-parametric empirical distribution as usually applied in financial institutions (Berkowitz, et al. 2011). Having sort historical return in specified period i.e., 250 observations, the unconditional quantile is:

$$\mathrm{VaR}_{t+1}(\alpha) = \mathrm{percentile}(\{r_\tau\}_{\tau=t-250+1}^{t}, \alpha) \qquad (3)$$

For RiskMetrics and GARCH(1,1)-N approaches general mean equation and error term specification are:

$$r_t = \mu_t + u_t, \qquad u_t = \sqrt{h_t}\varepsilon_t, \;\; \varepsilon_t \sim \mathrm{D}(0,1), \qquad (4)$$

where $r_t$ is return, $\varepsilon_t$ is error term that is identically independently distributed (i.i.d.) with specified distribution D, which is usually normalized to have zero mean and unit variance.

RiskMetrics: Instead of applying equal weight for whole historical sample which may not effectively reflect current conditions, the method of exponentially weighted moving average (EWMA) which places more weight on recent observations can be another choice. Its variance equation is defined as:

$$\mathrm{h_t} = \gamma \mathrm{h_{t-1}} + (1-\gamma)u_{\mathrm{t-1}}^2, \tag{5}$$

The Lambda ($\gamma$) is decay factor describing how fast the weights on recent observations are reduced when moving back through earlier observations. I follow a well-known EWMA methodology of RiskMetrics$^{\mathrm{TM}}$, launched by J.P. Morgan which suggested $\gamma$ = 0.94 for daily volatility model. The error term distribution is specified as $\mathrm{D}(0,1) = \mathrm{N}(0,1)$.

GARCH(1,1)-N: Standard or symmetric GARCH model is the generalization of Engle's ARCH model introduced by Bollerslev (1986). The variance equation is:

$$\mathrm{h_t} = \alpha_0 + \sum_{\mathrm{i}=1}^{\mathrm{p}} \alpha_{\mathrm{i}}\, u_{\mathrm{t-i}}^2 + \sum_{\mathrm{j}=1}^{\mathrm{q}} \beta_{\mathrm{j}} \mathrm{h_{t-j}} \tag{6}$$

where p = q = 1 and $\mathrm{D}(0,1) = \mathrm{N}(0,1)$.

GARCH(1,1)-t: Using the same specification as previous model but this time, the probability density function of student-t distribution; or D(0,1) = t(0,1,$\upsilon$), is applied:

$$f(\varepsilon_{\mathrm{t}}) = \frac{\Gamma\frac{(\upsilon+1)}{2}}{\sqrt{\pi(\upsilon-2)}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1+\frac{{\varepsilon_{\mathrm{t}}}^2}{\upsilon-2}\right)^{-\frac{\upsilon+1}{2}}, 2 < \upsilon < \infty \tag{7}$$

Thus, one-day-ahead VaR are given by:

$$\mathrm{VaR_{t+1}}(\alpha) = \mu_{\mathrm{t}} + \mathrm{z}_\alpha \cdot \mathrm{h_t}$$

$$\mathrm{VaR_{t+1}}(\alpha) = \mu_{\mathrm{t}} + \mathrm{t}_\alpha(\upsilon)\sqrt{\frac{\upsilon-2}{\upsilon}} \cdot \mathrm{h_t} \tag{8}$$

For normality and student-t assumption, respectively.

### 2.1.3 Critique

The widespread of using VaR bring forth heavy discussions of its accuracy (see, e.g. Berkowitz, 2002; Escanciano and Pei, 2012; Danielsson et al., 2014). For instance, VaR has been criticized for capability of forecasting risk when severe outcome occurs. Also, it does not say anything about the potential size of loss when exceeds VaR. Moreover, it is widely known that VaR are not meeting one of the requirements of a proper risk measures, that is, sub-additivity (Jorion, 2001; BCBS, 2013; Ziggel, 2014).

By acknowledging these concerns, new statistical modelling approach as well as its evaluation technique have been increasingly improved. Focusing on the latter improvement, literature have also considered model risk as another problem. Generally speaking, there are two main validation processes dealing with model uncertainty. The first and most common way is to backtest, which dealing with model risk indirectly using statistical inference to detect inaccurate model. Another way is model risk quantification dealing straightly with VaR model risk. Details on these model evaluation will be described next.

## 2.2 Quantile Model Risk

Model risks, in general, refer to an imprecision of model’s estimation compared to a true value of interested variable. Similarly, quantile model risk is occurred when estimating quantile with statistical models. Generally, there is no unique definition of model risk but it most happens when we do not know the true data generating process (DGP), resulting in making assumptions (Alexander and Sarabia, 2012).

### 2.2.1 Definition

For mathematical definition of quantile model risk, following Alexander and Sarabia (2012), the $\alpha$ quantile of a continuous distribution F of random variable X where $X \in R$ is denoted by:

$$q_{\alpha}^{F} = F^{-1}(\alpha), \tag{9}$$

where $\alpha$ is a predetermined coverage rate, i.e. 1% for 99%VaR models, and F is assumed true distribution. With statistical models, estimated quantile is based on estimated distribution $\hat{F}$:

$$q_{\alpha}^{\hat{F}} = \hat{F}^{-1}(\alpha) \tag{10}$$

Here, quantile model risk originated from the fact that $\hat{F} \neq F$, and hence $q_{\alpha}^{\hat{F}} \neq q_{\alpha}^{F}$. As a consequence, the model's $\alpha$ quantile $q_{\alpha}^{\hat{F}}$ is at a different quantile of $\alpha$ under $F^{-1}(\cdot)$ and we can use $\widehat{\alpha}$ for this new quantile, that is, $q_{\alpha}^{\hat{F}} = q_{\widehat{\alpha}}^{F}$ . Then, reversing the quantile function, we can obtain the $\widehat{\alpha}$ as

$$\widehat{\alpha} = F[\hat{F}^{-1}(\alpha)] \tag{11}$$

The process when $\alpha$ shifted to $\hat{\alpha}$ under the true distribution $F^{-1}(\cdot)$ is called "probability shifting" (Boucher and Mailet, 2013; Boucher et al., 2014). To amplify model risk, one way is to measure the deviation of $\widehat{\alpha}$ from $\alpha$, or **"quantile probability errors"**:

$$e\big(\alpha \big| F, \hat{F}\big) = \widehat{\alpha} - \alpha \tag{12}$$

Another way to compute a magnitude of model risk is via the difference of quantile value correspond to the true distribution, or **"quantile errors"**:

$$e\big(q_{\alpha}^{F} \big| F, \hat{F}\big) = q_{\alpha}^{F} - q_{\widehat{\alpha}}^{F} \tag{13}$$

In the same analogy, Boucher et al. (2014) define model risk as a bias function bias($\theta_0$, $\hat{\theta}$, $\alpha$,) (or economic value of model risk) that's make the Theoretical VaR (ThVaR) and estimated VaR (EVaR) equal:

$$\mathrm{ThVaR}(\theta_0, \alpha) \ = \ \mathrm{EVaR}(\hat{\theta}, \alpha) \ + \ \mathrm{bias}(\theta_0, \hat{\theta}, \alpha) \qquad (14)$$

### 2.2.2 Relevant Researches

Of course, quantifying amount of true model risk as above can be done only in simulation experiment where true DGP is known. In real situation, however, risk model risks have been approaches in two different ways. The first way is done using a benchmark model. For example, Alexander (2012) proposed to quantify and adjust model risk to regulatory capital using maximum entropy distribution as a benchmark. The second way is examining all feasible models, then evaluating discrepancy of the results. For instance, Bao and llah (2004) studied the bias occur when forecasting VaR via ARCH(1) specification. Christoffersen and Goncalves (2005) quantified ex ante model estimation risk by constructing confidence interval and suggested a resampling technique. Kerkhof et al. (2010) is who first proposed to adjust regulatory market capital charge explicitly for model risk using backtest. Boucher et al. (2014) and its initial version (Boucher and Mailet, 2013) complement Kerkhof et al. (2010)'s approach and generalizing the backtests used for cushion on estimated VaR series.

Generally, there is no unique definition of model risk and various sources could be involved, including misspecification error (or wrong model choice), estimation error (including sampling and parameterization error), and identification error. The first two sources are most heavily documented in literature. For example, Christoffersen and

Goncalves (2005) focus only on estimation risk. Similarly, Kerkhof et al. (2010) calculated estimation error as the difference between upper confidence interval and point estimate. Bao and llah (2004), Alexander and Sarabia (2012), Boucher and Maillet (2013), and Boucher et al. (2014) are those who defined model risk as a combination of misspecification and estimation error. Kerkhof et al. (2010) added identification risk as another source accounting for uncertainty on analyzing subjective approach. Other sources can be liquidity risk, granularity risk, and data contamination (see, Boucher et al., 2014).

## 2.3 Model risk correction via Backtesting Framework

Boucher et al. (2014) recently proposed to approximate model risk ultimately for the adjustment that make VaR forecasts more robust to model risk. Their approach complements the Kerkhof et al. (2010) who first proposed a procedure to incorporate model risk into the calculation of regulatory capital reserves using backtests. Actually, its simple version has been initially documented by Boucher and Maillet (2013) where there is one backtest criterion for model risk adjustment - Kupiec (1995) unconditional coverage test. Thus far, in regulatory framework there is no explicit guideline for model risk, but as Kerkhof et al. (2010) pointed out, the multiplication factor for capital reserve computation is deemed partly to account for model risk. Nevertheless, regulatory framework regards only unconditional coverage property which may not be enough. Other violation properties that should be also considered in model risk evaluation will be described next.

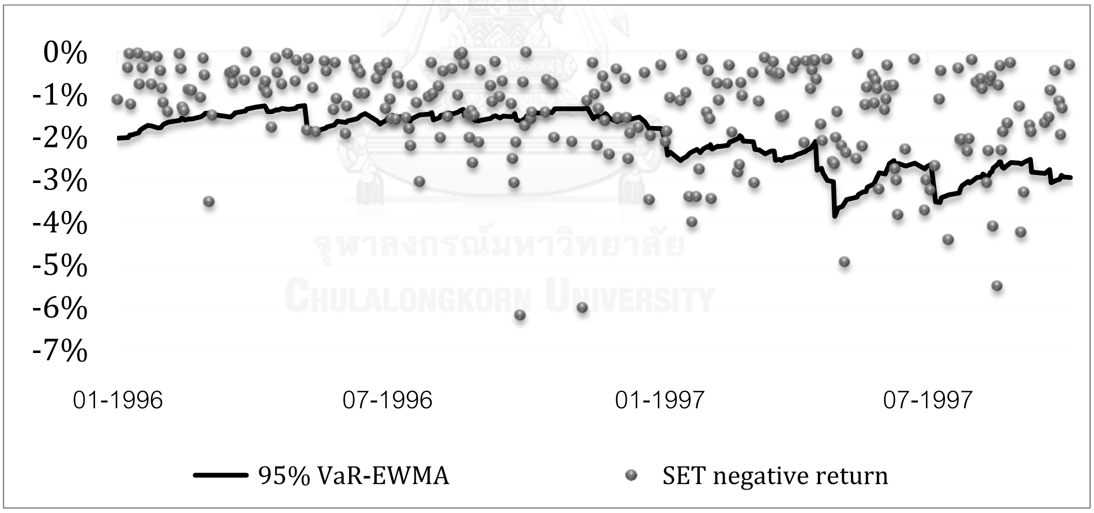
### 2.3.1 Fundamental of Backtests

Backtest is an ex post process in the sense that it compare VaR forecasts to realized returns to examine whether the model represent amount of risks properly or not. Specifically, given a notation of "hit" function as follows:

$$I_t(\alpha) = \begin{cases} 1 & \text{if } r_t \leq -\mathrm{VaR}_t(\alpha) \\ 0 & \text{if } r_t > -\mathrm{VaR}_t(\alpha) \end{cases} \tag{15}$$

Where $r_t$ is the profit and loss return at time t, {t = -1, -2, -3, …, -250}, i.e., backtest period is over the past one-year trading days). If realized losses exceed $\mathrm{VaR}_t(\alpha)$ the function value is 1, this is called VaR violations, meaning that VaR cannot cover the loss on that day. Otherwise the function value is zero.

**Figure 1. SET Daily Negative Returns and Daily 95% VaR Computed by EWMA Model**



To backtest a sequence of VaR violations Campbell (2005) and Haas (2001) suggest three properties to be concerned. The first property is **unconditional coverage** (hereafter UC). Sometimes it's also called frequency or hit test. Consider **Figure 1**, shows SET negative return (dots) and daily 95% VaR computed by EMWA (line). The negative return that exceeds VaR forecast on that day (dots that located below the line**)** are called "violations". The idea of UC test is that proportion of violations

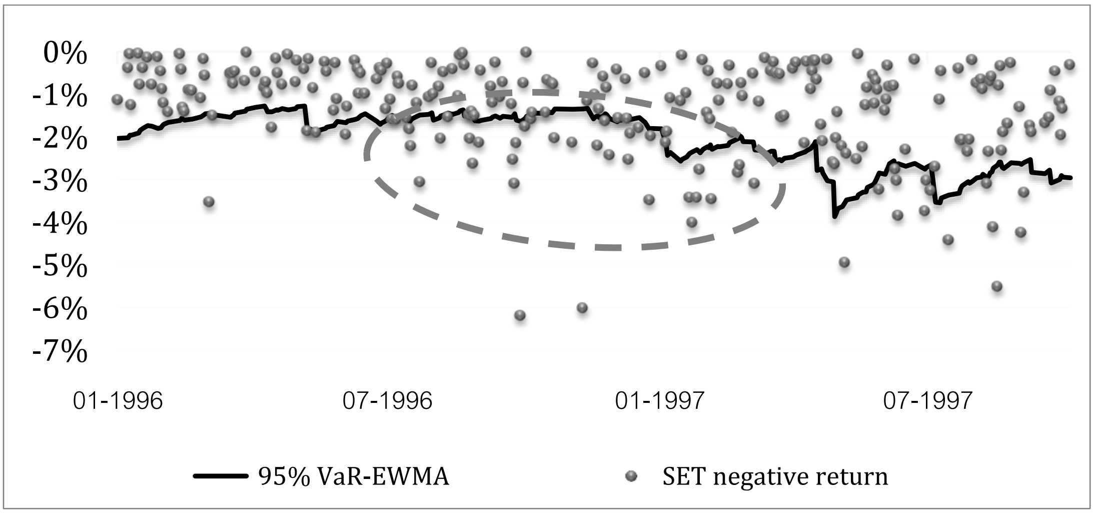
$\left(\frac{\text{No. Violation}}{\text{No. Observations (250)}}\right)$ should be close to Coverage rate (α) or 5% for 95% confidence level of VaR. If the proportion is too high (low), the model underestimate (overestimate) true level of risk. Thus, “in general UC term”, the model is said to possess UC property when:

$$E[I_t(\alpha)] = \alpha, \ \forall t \in \{0,1,2, \ldots T\} \tag{16}$$

For instance, 1% daily VaR with 250 backtkest periods should have 250*0.01 = 2.5 days of violation. In other words, the VaR expected to violate one percent of the time. UC is the most basic violation property and it has been applied with the regulatory traffic light test.

The second property is **independence** (hereafter IND), which places more restriction on violations behavior. Specifically, violation sequence should not be dependent or appeared clusterings unless the model is said to be incapable of predicting change in risk dynamics. For example, in **Figure 2**, around the 1997’s crisis, violations frequently failed to capture downside risk as highlighted by dashed circle.

**Figure 2. SET Daily Negative Returns and Daily 95% VaR Computed by EWMA Model with Highlighted Violation Clustering**



We can expressed the definition of independence property:

$$\mathrm{H_{0,cc}}: \mathrm{I_t}(\alpha) \sim \mathrm{i.i.d.\,Bernoulli}(\hat{\alpha}) \tag{17}$$

hit function have i.i.d. Bernoulli distribution with unspecified parameter denoted by $\hat{\alpha}$, which is probability of observing violations.

If unconditional coverage and independence properties are combined to be a joint test, it is called **conditional coverage** (hereafter CC) property. Null hypothesis can be expressed as:

$$\mathrm{H_{0,cc}}: \mathrm{I_t}(\alpha) \sim \mathrm{i.i.d.\,Bernoulli}(\alpha) \tag{18}$$

Normally the null of CC hypothesis is similar to IND hypothesis except it require another condition, that is, parameter of Bernoulli for example is equal to α (e.g., 5% for 95% VaR). As Campbell (2005) elucidated, if the previous-day violation (positively) is correlated with current violation, then given that violation occurred yesterday, the probability of observing violation today will be higher than α, which violate CC property.

The last property is **magnitude** of exceptions (hereafter MG) which is less popular than the former two properties until recent decades. It is worth noting that severity of loss is the real concern not exception per se. especially in capital reserve computation, and/or margin buffer (Colletaz et al., 2013). For example, given a sequence of violations, Lopez (1998) measure the magnitude of violation using a quadratic score function as:

$$\mathrm{L}(\mathrm{VaR_t}(\alpha), \mathrm{r_t}) = \begin{cases} 1 + \left(\mathrm{r_t} - \mathrm{VaR_t}(\alpha)\right)^2 & \text{if } \mathrm{r_t} \leq -\mathrm{VaR_t}(\alpha) \\ 0 & \text{if } \mathrm{r_t} > -\mathrm{VaR_t}(\alpha) \end{cases} \tag{19}$$

which tells how much predicted losses when negative P&L exceeds VaR will be.

### 2.3.2 The Original Approach of Boucher et al. (2014)

Model risk correction framework via backtests proposed by Boucher et al. (2014) is based on three-backtesting criteria set. Each test separately accounts for each property, including Kupiec (1995) unconditional coverage test, Christoffersen (1998) Markov (independence) test, and Berkowitz (2001) magnitude test. Basically, for any backtest, whether the null hypothesis (or the model) is accepted is based on test statistic. Kupiec (1995) test: One of the earliest backtests is Kupiec (1995) which focus exclusively on unconditional coverage property. Proportion of failures (POF) is used to examine how many violations occur over a given timeframe. Using the sample of observations of T, Kupiec's test statistic is a likelihood ratio (LR) which is asymptotically chi-square distributed with one degree of freedom takes the form of:

$$\mathrm{LR}_{\mathrm{uc}}^{\mathrm{k}} = -2\ln\left(\left(\frac{1-\hat{\alpha}}{1-\alpha}\right)^{\mathrm{T}_0}\left(\frac{\hat{\alpha}}{\alpha}\right)^{\mathrm{T}_1}\right) \quad \xrightarrow{d} \chi^2(1)\ , \qquad (20)$$

where, $\hat{\alpha} = \frac{1}{\mathrm{T}}\mathrm{T}_1\ ,\quad \mathrm{T}_1 = \sum_{t=1}^{T} \mathrm{I}_t(\alpha),\quad \mathrm{T}_0 = \mathrm{T} - \mathrm{T}_1$

Christoffersen (1998) Markov test: This test account for independence property defined as a first-order Markov chain. It is also one of the most frequently referred tests dominated in literature. This means that if the risk model is accurate, any VaR violation should not depend on whether or not there is violation the day before. The null of $\mathrm{H}_{0,\mathrm{ind}}$ is $\Pr(\mathrm{I}_t = 1|\Omega_{t-1}) = \Pr(\mathrm{I}_{t+1} = 1|\Omega_t)$ where $\Omega_{t-1}$ is past information,

The test statistic is a LR which is asymptotically chi-square distributed with one degree of freedom:

$$\mathrm{LR}_{\mathrm{ind}}^{\mathrm{c}} = -2\ln\left[\frac{(1-\pi)^{\mathrm{T}_{00}+\mathrm{T}_{10}}\ \pi^{\mathrm{T}_{01}+\mathrm{T}_{11}}}{(1-\pi_0)^{\mathrm{T}_{00}}\pi_0{}^{\mathrm{T}_{01}}(1-\pi_1)^{\mathrm{T}_{10}}\pi_1{}^{\mathrm{T}_{11}}}\right] \quad \overset{d}{\rightarrow}\chi^2(1)\ , \qquad (21)$$

where, $\pi_0 = \frac{T_{01}}{T_{00}+T_{01}}$ , $\pi_1 = \frac{\mathrm{T}_{11}}{\mathrm{T}_{10}+\mathrm{T}_{11}}$ , $\pi = \frac{\mathrm{T}_{01}+\mathrm{T}_{11}}{\mathrm{T}_{00}+\mathrm{T}_{01}+\mathrm{T}_{10}+\mathrm{T}_{11}}$ , $\mathrm{T}_{\mathrm{ij}}$ is defined as the number of days when i occur today conditional on j occurred the day before, with i and j can be 0 (no violation) or 1 (violation occurred). For example, $\mathrm{T}_{01}$is the number of days of no violation given that there is violation on the previous day, and $\mathrm{T}_{11}$ is the number of days of violation given that there is also violation on the previous day. Thus, $\pi_{\mathrm{i}}$ is a binary Markov chain reflecting one-day dependence by the probability of realizing violation today given the condition i on the previous day. $\pi$ is unconditional probability of realizing violation today.

<u>Berkowitz (2001) Magnitude Test</u>: The last test in original criteria set focuses large losses incurred when realized P&L exceeds VaR. Berkowitz (2001) proposed LR test based on a censored normal likelihood. Consider the left tail of the distribution that is predefined by users, i.e., loss larger than 5% VaR in this case. So it focus only observation (violations) in the tailed part, others are truncated. Then a new variable for large violation is redefined as:

$$\mathrm{r}_{\mathrm{t}}^{*} = \begin{cases} \mathrm{r}_{\mathrm{t}} & \text{if } \mathrm{r}_{\mathrm{t}} < -\mathrm{VaR} \\ \mathrm{VaR} & \text{if } \mathrm{r}_{\mathrm{t}} \geq -\mathrm{VaR} \end{cases} \qquad (22)$$

VaR is the tailed threshold, $\mathrm{r}_{\mathrm{t}}$ is ex post return at time t. Then the LR test statistic is based on the constrained (null hypothesis: $\mu = 0,\ \sigma^2 = 1$) and unconstrained (alternative hypothesis) conditions:

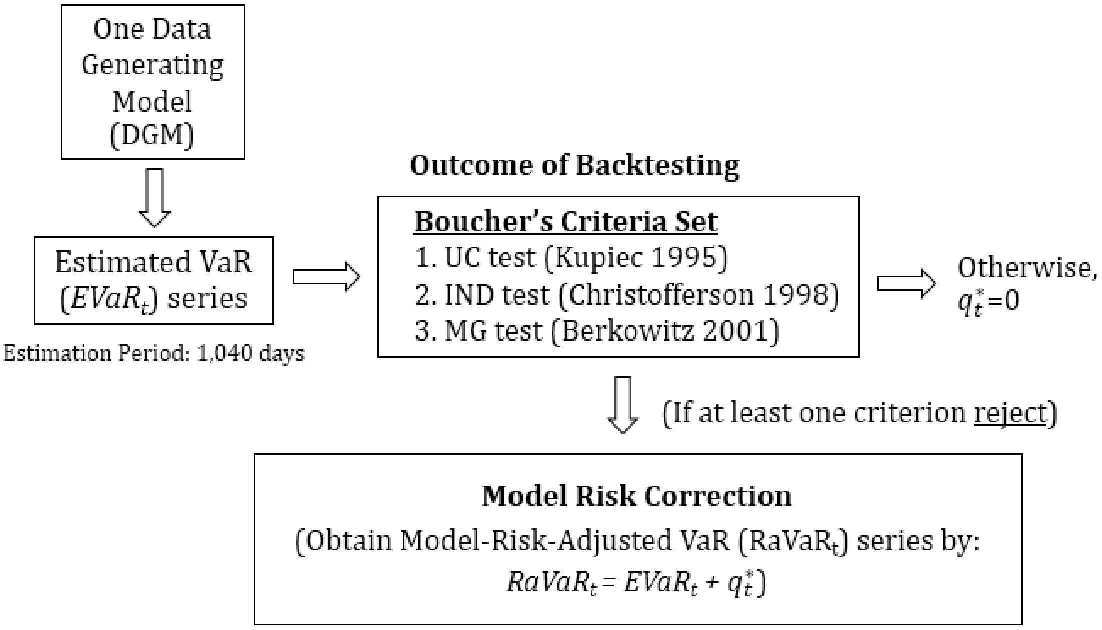
$$\mathrm{LR}_{\mathrm{mg}}^{\mathrm{b}} = -2[\mathrm{L}(0,1) - \mathrm{L}(\hat{\mu}, \hat{\sigma}^2)] \qquad \overset{d}{\rightarrow}\chi^2(2), \qquad (23)$$

where the log-likelihood function for estimation of $\hat{\mu}$ and $\hat{\sigma}$ is:

$$L(\mu, \sigma \,|r^{*}) = \sum\nolimits_{r_t^{*}<\mathrm{VaR}} \left(-\frac{1}{2}\ln(2\pi\sigma^{2}) - \frac{1}{2\sigma^{2}}(r_t^{*} - \mu^{2})\right) + \sum\nolimits_{r_t^{*}=\mathrm{VaR}} \left(1 - \phi\left(\frac{\mathrm{VaR}-\mu}{\sigma}\right)\right) \quad (24)$$

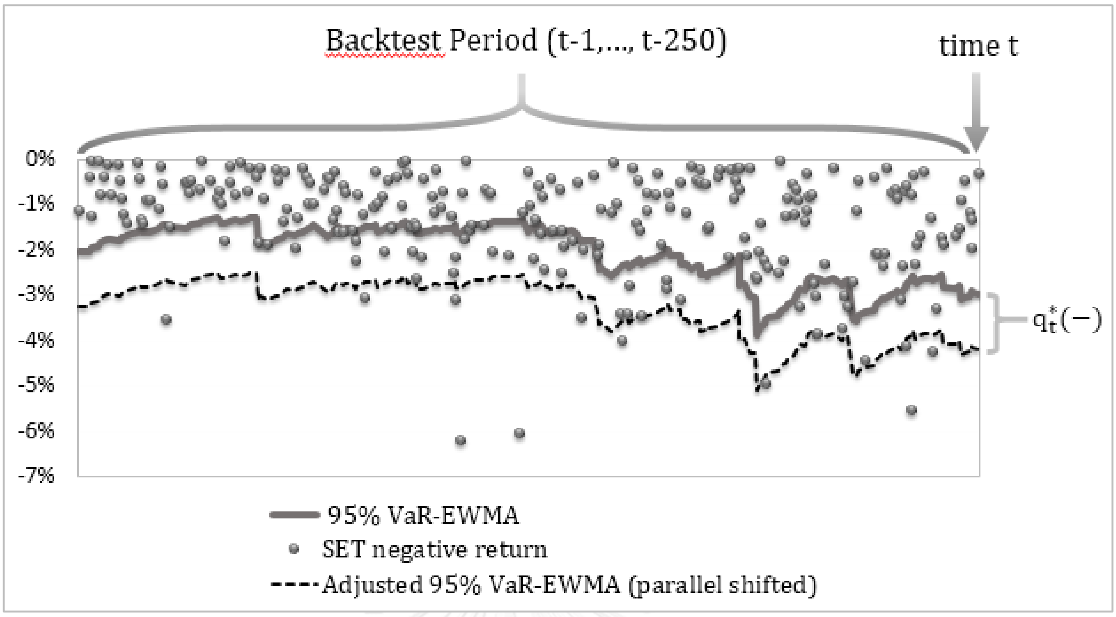
The null will be rejected if large losses are significantly higher or lower than expected losses.

**Figure 3. Boucher et al. (2014)'s Conceptual Framework**


One Data Generating Model (DGM)
Estimated VaR ($EVaR_t$) series
Estimation Period: 1,040 days
Outcome of Backtesting
Boucher's Criteria Set
1. UC test (Kupiec 1995)
2. IND test (Christofferson 1998)
3. MG test (Berkowitz 2001)
Otherwise, $q_t^*$=0
(If at least one criterion reject)
Model Risk Correction
(Obtain Model-Risk-Adjusted VaR ($RaVaR_t$) series by:
$RaVaR_t = EVaR_t + q_t^*$)

Then, three criteria in *Boucher's Criteria Set* are used in the backtesting procedure as shown by **Figure 3**. Starting with one data generating model (DGM) to fit the historical data (i.e., DJIA daily index returns), then they estimated daily VaR series denoted by $\mathrm{EVaR_t}$ using rolling estimation window of 1,040 daily returns. Next, in the backtesting procedure, if at least one criterion in the criteria set reject the model, modek risk coorection is done by finding the new risk forecast series called model-risk-adjusted VaR: $\mathrm{RaVaR_t} = \mathrm{EVaR_t} + \mathrm{q_t^*}$, where $\mathrm{q_t^*}$ is regarded as a proxy of model risk's magnitude. But if all criteria accept the model, no correction will be required ($\mathrm{q_t^*} = 0$).

**Figure 4. Methods to Find $q_t^*$**



Specifically, $q_t^*$ it is the minimal amount that when added to all observations in a given backtest period, i.e. recent 250 daily Estimated VaR (EVaR) series $\{\mathrm{VaR}_{t-1}, \dots, \mathrm{VaR}_{t-250}\}$, make it passes to all criteria (when at first at least one criterion reject the model). For an illustration, **Figure 4** show how to find $q_t^*$ in daily basis using rolling window technique. Suppose we are at time t, given then series of daily 95% VaR-EWMA (line) does not pass all criteria in the criteria set, then Boucher et al. (2014) do a parallel shift from the original series incrementally to obtain a new series that all criteria accept the model. Note that this is the situation when model risk correction $q_t^*$ is negative (downward parallel shift), however, it can be a case of positive $q_t^*$ (upward parallel shift) as well, i.e., when the initial VaR series is too much negative (number of violations is too low).

Boucher et al. (2014)'s quantile correction framework provides very effective guideline for model risk correction using backtests. In their research, 9 famous models (including Gaussian parametric approach, student-t, historical simulation, RiskMetrics, CAViaR, GARCH(1,1), Cournish-Fisher, generalized extreme value, and generalized Pareto distribution) were applied and found to be significantly improved on capturing market conditions after empirically adjusting model uncertainty.

However, one should notice that the approximated magnitude of model risk could be varied with different criteria sets. Further, currently apart from three backtests used in the original framework, there have been a number of backtests proposed which many of them have been appropriately validated to have more statistical power than those in the original set. The most obvious example is the extension of Christoffersen (1998) IND test as the first-order Markov chain is known of the limitation to detect more than one-day dependence (two-consecutive-day clustering) (Campbell, 2005; Christoffersen and Pelletier, 2004; Engle and Manganelli, etc.). Those outperformed tests that are capable of detecting more than one-day dependence include duration-based test (Christoffersen and Pelletier, 2004), dynamic quantile test (Engle and Manganelli, 2004), and Monte Carlo test (Ziggel et al., 2014).

In addition, IND and especially CC tests tend to be superior to UC tests and are main focuses of many literature as violation clustering (dependence) seems to be mostly severe in financial data (e.g., Engle and Manganelli, 2004; Candelon et al., 2011; Berkowitz et al., 2011; Ziggel et al., 2014). This leaves me a way to put forward the Boucher’s framework. By aiming to improve the accuracy of quantile model risk quantification and correction, I will extend the backtesting framework and find optimal criteria sets.

## 2.4 Other Backtesting Methods for the Criteria

Having gathered a range of backtesting methods, they are then categorized into four groups corresponding to their property detection, including (i) UC test, (ii) IND test, (ii) CC test, and (iv) MG test. Generally, backtests proposed by individual papers could be adapted to test more than one property, whether another separate test or joint test. Some methods could even test for three hypotheses (UC, IND, and CC).

### 2.4.1 Unconditional Coverage Test (UC)

With regard to UC property, Kupiec (1995) is among the most common backtests applied heavily in risk management researches and it is also the main backtest for regulatory framework and financial institution. Due to the knowledge that UC property solely is not enough to detect bad models, more researching effort were placed on studying other tests especially IND and CC. However, there are still some improved methodologies on UC hypothesis.

Engle and Manganelli (2004) adopted backtest methods called CaViaR (or Dynamic Quantile (DQ)) using linear regression based on independent variable in information set. Although, their tests focus mainly on IND and CC test, it could be applied to test for UC property. Consider the following equation:

$$\mathrm{I_t} - \alpha = \omega + \sum_{i=1}^{n} \beta_{1,i}\mathrm{I}_{t-i} + \sum_{i=1}^{m} \beta_{2,j}\mathrm{VaR}_{t-j} + \varepsilon_t \tag{25}$$

where $\alpha$ is a quantile level of VaR, $\varepsilon_t$ is an error term, $\beta_{1,i}$ and $\beta_{2,j}$ are regression coefficients of Violations and VaR sequences in the past, respectively. The reasoning is that if the null of UC is true, probability of having violation predicted by the model will be closed to α. In this setting, I use n = m =3 (lagged up to 3 period). For test statistic, they use the following Wald statistic:

$$\mathrm{DQ}_{\mathrm{uc}} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{R}'(\mathbf{R}[\mathbf{X}'\mathbf{X}]\mathbf{R}')^{-1}\mathbf{R}\hat{\boldsymbol{\beta}}}{\alpha(1-\alpha)} \tag{26}$$

where $\hat{\boldsymbol{\beta}} = \left(\hat{\omega}, \hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3,}, \hat{\beta}_{2,1}, \hat{\beta}_{2,2}, \hat{\beta}_{2,3}\right)'$, **R** is the 1 x (3+3+1) matrix (1, 0, …,0), and ***X*** is a matrix that contain ones in the first column, and lagged hit functions in the next three column, and lagged VaR in the last three column. And the null hypothesis is $\mathrm{H}_{0,\mathrm{uc}}\colon \omega = 0$.

In addition, Candelon et al. (2011) framework applied the concept of GMM duration using orthonormal polynomials. Although their main focus are IND and CC tests, they also conduct UC test. To illustrate, first, denoted by $\mathrm{d}_\mathrm{i}$, the duration or number of days between two consecutive violations which $\mathrm{d}_\mathrm{i} = \mathrm{t}_\mathrm{i} - \mathrm{t}_{\mathrm{i}-1}$, where $\mathrm{t}_\mathrm{i}$ is the day of i$^{\mathrm{th}}$ violation (Christoffersen and Pelletier, 2004). Under CC hypothesis the duration $\{\mathrm{d}_\mathrm{i}\}$ have a geometric distribution with parameter α. So Candelon et al. (2011) initially introduced the recursive equation as follows:

$$\mathrm{M}_{\mathrm{j}+1}(\mathrm{d};\beta) = \frac{(1-\beta)(2\mathrm{j}+1)+\beta(\mathrm{j}-\mathrm{d}+1)}{(\mathrm{j}+1)\sqrt{1-\beta}}\mathrm{M}_\mathrm{j}(\mathrm{d};\beta) - \left(\frac{\mathrm{j}}{\mathrm{j}+1}\right)\mathrm{M}_{\mathrm{j}-1}(\mathrm{d};\beta) \tag{27}$$

which $\mathrm{M}_{-1}(\mathrm{d};\beta) = 0$ and $\mathrm{M}_0(\mathrm{d};\beta) = 1$. Hence the null hypothesis of CC is:

$$\mathrm{H}_{0,\mathrm{cc}} : \mathrm{E}\left[\mathrm{M}_\mathrm{j}(\mathrm{d}_\mathrm{i};\alpha)\right] = 0,\ \ \mathrm{j} = \{1,\dots,\mathrm{p}\}, \tag{28}$$

where p is number of moment conditions. And the null hypothesis of UC is when the average of duration is equal to $1/\alpha$, which can be derived as:

$$\mathrm{H}_{0,\mathrm{uc}} : \mathrm{E}[\mathrm{M}_1(\mathrm{d}_\mathrm{i};\alpha)] = 0, \tag{29}$$

The null of IND test, when $\mathrm{d}_\mathrm{i}$ is geometric distribution, is also shown as:

$$\mathrm{H}_{0,\mathrm{ind}} : \mathrm{E}\left[\mathrm{M}_\mathrm{j}(\mathrm{d}_\mathrm{i};\beta)\right] = 0,\ \ \mathrm{j} = \{1,\dots,\mathrm{p}\}, \tag{30}$$

They used Monte Carlo study and showed that the $\mathrm{GMM}_{\mathrm{cc}}$ dominates some existing $\mathrm{LR}_{\mathrm{cc}}$ methods, but didn't compare the performance of $\mathrm{GMM}_{\mathrm{uc}}$ with any other UC test.

However, Ziggel et al. (2014) recently noted that the UC tests of GMM and Kupiec (1995) have drawbacks on realized small sample and inappropriate definition of UC hypothesis. To deal with, they first claim that the UC hypothesis: $\mathrm{E}[\mathrm{I}_{\mathrm{t}}(\alpha)] = \alpha$, which require to test the expected coverage equal to α for all t, is imprecise. Even though $\mathrm{I}_{\mathrm{t}}(\alpha)$ sequence has expected coverage varies overtime it still possess UC if its average over backtest horizon is α. Thus, they redefined it to $\mathrm{E}\left[\frac{1}{\mathrm{n}}\sum_{\mathrm{t}=1}^{\mathrm{n}}\mathrm{I}_{\mathrm{t}}(\alpha)\right] = \alpha$ and proposed a Monte Carlo simulation (MCS) based test, $\mathrm{MCS}_{\mathrm{uc}}$, which rely on a new way to compute critical value via MCS rather than using asymptotic distribution as:

$$\mathrm{MCS}_{\mathrm{uc}} = \sum_{\mathrm{t}=1}^{\mathrm{n}}\mathrm{I}_{\mathrm{t}}(\alpha) + \epsilon\,, \tag{31}$$

where $\epsilon \sim 0.001 \cdot \mathrm{N}(0,1)$ is a continuous random variable for ensuring that two test statistics from sample and from MCS could not be the same (tiebreaking procedure).

### 2.4.2 Independence Test (IND)

IND test is most documented in literature as well as CC. According to the failure of Christoffersen (1998)'s first-order Markov test mentioned before, many tests has been proposed. One of the early introduced ones was the duration-based test of Christoffersen and Pelletier (2004). It is more flexible and effective in that it can detect exception clustering more than two-day-consecutive violations (one-day dependence). By introducing the concept of (no-hit) duration, $\mathrm{d}_{\mathrm{i}}$ as the number of days between two violations (mentioned before), IND hypothesis is detected when there is excessive number of short and long durations. More specifically, in the absence of violation dependence, durations should have exponential distribution, to represent memoryless property (independence) of duration variable. It could be expressed as:

$$f_{\mathrm{exp}}(\mathrm{d};\alpha) = \alpha\exp(-\alpha\mathrm{d}) \qquad (32)$$

In order to find test statistic, Christoffersen and Pelletier (2004) use the Weibull distribution as its hazard function has close-from representation, which become the exponential distribution when b = 1, or a flat hazard case:

$$f_{\mathrm{W}}(\mathrm{d};\mathrm{a};\mathrm{b}) = \mathrm{a^b b d^{b-1}}\exp(-(\mathrm{ad})^{b}) \qquad (33)$$

Thus, IND null hypothesis is $\mathrm{H_{0,ind}}\colon \mathrm{b} = 1$ and corresponding likelihood ratio (LR) test statistic could be calculated. In addition, the CC version of this is added simply as $\mathrm{H_{0,cc}}\colon \mathrm{b} = 1,\ \mathrm{a} = \alpha$. The second duration-based methodology is the GMM test of Candelon et al. (2011) claiming to overcome the LR Markov-chain of Christoffersen (1998) and Christoffersen and Pelletier (2004)'s duration-based framework in realized small backtesting size. Its IND hypothesis was stated in *equation (30).*

Another test is the conditional autoregressive VaR (CaViaR) suggested by Engle and Manganelli (2004). As said before in UC test the CaViaR is mainly for IND/CC test that account for nth-order auto-regression. Using the same linear regression as in *equation (25*): $\mathrm{I_t} - \alpha = \omega + \sum_{\mathrm{i=1}}^{\mathrm{n}} \beta_{1,\mathrm{i}}\mathrm{I_{t-i}} + \sum_{\mathrm{i=1}}^{\mathrm{m}} \beta_{2,\mathrm{j}}\mathrm{VaR_{t-j}} + \varepsilon_{\mathrm{t}}$. the Wald statistic for independent test can be explained by:

$$\mathrm{DQ_{ind}} = \frac{\boldsymbol{\hat{\beta}'P'(P[X'X]P')^{-1}RP}}{\alpha(1-\alpha)} \qquad (34)$$

where $\boldsymbol{\hat{\beta}} = \left(\hat{\omega}, \hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3,}, \hat{\beta}_{2,1}, \hat{\beta}_{2,2}, \hat{\beta}_{2,3}\right)'$, $\boldsymbol{P}$ is the 1 x (3+3+1) matrix (0, 1, …,1) , and ***X*** is a matrix that contain ones in the first columns, and lagged hit functions in the next three column, and lagged VaR in the last three columns. The null is $\mathrm{H_{0,ind}}$: $\beta_{1,i} = 0,$ and $\beta_{2,i} = 0.$

As well as for the UC hypothesis test, Ziggel et al. (2014) argued that traditional IND tests which supposed to detect non-i.i.d VaR violations were focus around auto-regression and thus are not be able to classify inaccurate VaR models during the joint of calm and volatile market period. Consequently, they propose a new IND ($\mathrm{MCS_{ind}}$) test for new i.i.d. property as:

$$\mathrm{MCS_{ind}} = t_1^2 + (n - t_m)^2 + \sum_{i=2}^{m} (t_i - t_{i-1})^2 + \epsilon, \qquad (35)$$

where the sum represent squared duration between two violations. In presence of violations clustering is, the test statistic will be large, so they conduct one-sided test using MCS technique.

### 2.4.3 Conditional Coverage Test (CC)

Many independence tests have their joint version, for the earliest case, Chrisoffersen (1998) extend his IND test to CC test by using the null hypothesis of $\mathrm{H_{0,ind}}$: $\Pr(I_t = 1|\Omega_{t-1}) = \alpha$, for all t. Then the test statistic is computed by:

$$\mathrm{LR_{cc}} = \mathrm{LR_{uc}} + \mathrm{LR_{ind}} \qquad \overset{d}{\to} \chi^2(2) \qquad (36)$$

Another CC test is the conditional autoregressive VaR (CaViaR) or Dynamic Quantile tests (Engle and Manganelli 2004). Recall the *equation (25)*: $I_t - \alpha = \omega + \sum_{i=1}^{n} \beta_{1,i} I_{t-i} + \sum_{i=1}^{m} \beta_{2,j} \mathrm{VaR}_{t-j} + \varepsilon_t$ the Wald test statistic of this regression-based test will be:

$$\mathrm{DQ_{cc}} = \frac{\hat{\beta}'\mathrm{X}'\mathrm{X}\hat{\beta}}{\alpha(1-\alpha)} \tag{37}$$

All parameters definition are still the same as in $\mathrm{DQ_{uc}}$ and $\mathrm{DQ_{ind}}$ tests. The null of CC hypothesis is $\mathrm{H_{0,cc}}: \omega = 0, \beta_{1,i} = 0,$ and $\beta_{2,i} = 0$.

Ziggel et al. (2014) also conduct CC test ($\mathrm{MCS_{cc}}$), as a combination of $\mathrm{MCS_{uc}}$ and $\mathrm{MCS_{iid}}$ and showed that their proposed backtests outperform existing IND and CC test including Markov-based, duration-based and GMM-based. Specifically, they introduced the combined test of CC in that users are allowed to weight UC and IND (or i.i.d. in this case) components which make it more flexible to different risk perspectives. The test statistic was described in the form:

$$\mathrm{MCS_{cc}} = \mathrm{a} \cdot f(\mathrm{MCS_{uc}}) + (1-\mathrm{a}) \cdot \mathrm{g}(\mathrm{MCS_{iid}}), \quad 0 \le \mathrm{a} \le 1, \tag{38}$$

where a is weight of the combined tests, and I set a = 0.5,

$$f(\mathrm{MCS_{uc}}) = \left| \frac{\frac{(\mathrm{MCS_{uc}})}{\mathrm{n}} - \alpha}{\alpha} \right| = \left| \frac{\frac{(\epsilon + \sum_{\mathrm{t=1}}^{\mathrm{n}} \mathrm{I_t})}{\mathrm{n}} - \alpha}{\alpha} \right|, \text{and}$$

$$\mathrm{g}(\mathrm{MCS_{iid}}) = \frac{\mathrm{MCS_{iid}} - \hat{\mathrm{r}}}{\hat{\mathrm{r}}} \cdot l_{\{\mathrm{MCS_{iid}} \ge \hat{\mathrm{r}}\}} \tag{39}$$

The first component is percentage difference between the coverage rate (α) and observed proportion of violations. It makes UC and IND tests comparable via standardization. The second component measures the percentage deviation between the expected sum of square durations under the null hypothesis ($\hat{\mathrm{r}}$ *or* an estimator for $\mathrm{E}(\mathrm{MCS_{iid}}|\mathrm{H_0})$)[3] and corresponding observed value. As their

[3] $\hat{\mathrm{r}}$ is calculated in simulation process, few steps before finding critical value.

UC and IND tests introduced before, they obtain the test statistic by using Monte Carlo simulation.

### 2.4.4 Magnitude Test (MG)

This property had gained less interest relative to other hypothesis, perhaps because traditional regulatory framework focused only on unconditional property (Berkowitz, 2001). Currently, the supervisor have concerned much about magnitude of large losses and introduced stressed VaR in the Basel 3 and point out in the Basel 3.5 proposal to take account for size of backtesting exception or change to use alternative risk measure, expected shortfall (BCBS, 2013). However, the clear-cut process is still ongoing (Embrechts et al., 2014).

In MG hypothesis framework, apart from the Berkowitz's magnitude test, loss function-based procedure has been early introduced by Lopez (1998). Recall quadratic loss function:

$$\mathrm{L}(\mathrm{VaR}_t(\alpha), \mathrm{r}_t) \;=\; \begin{cases} 1 + \left(\mathrm{r}_t - \mathrm{VaR}_t(\alpha)\right)^2 & \text{if } \mathrm{r}_t \;\leq\; -\mathrm{VaR}_t(\alpha) \\ 0 & \text{if } \mathrm{r}_t \;\leq\; -\mathrm{VaR}_t(\alpha) \end{cases} \tag{40}$$

One way to compare VaR models is simply examining which models minimize loss function. Consider the average score function ($\hat{\mathrm{L}}$) for the whole period (T):

$$\hat{\mathrm{L}} = \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \mathrm{L}(\mathrm{VaR}_t(\alpha), \mathrm{r}_t) \tag{41}$$

Then, we can rank the candidate models using the idea that model that provide smaller $\hat{\mathrm{L}}$ is better. Note two ways above are not on test-statistic basis, but comparative analysis. It could be adapted to such a test assuming the specified model is correct under

$H_0$, then computing the associated critical value. However, Piontek (2013) found that its power is quite unstable and choosing the benchmark is very subjective.

Another recent methodology was proposed by Colletaz et al. (2013) called a double-threshold or risk map approach. In their validation framework, in addition to exception they conducted a new variable called super exception $\mathrm{I_t}(\alpha')$ defined as the same analogy as hit function $\mathrm{I_t}(\alpha)$ in *equation (15)* as

$$\mathrm{I_t}(\alpha') = \begin{cases} 1 & \text{if } \mathrm{r_t} \le -\mathrm{VaR_t}(\alpha') \\ 0 & \text{if } \mathrm{r_t} > -\mathrm{VaR_t}(\alpha') \end{cases} \tag{42}$$

Then the joint null hypothesis is $\mathrm{H_{0,MG}}$: $\mathrm{E}[\mathrm{I_t}(\alpha)] = \alpha$ and $\mathrm{E}[\mathrm{I_t}(\alpha')] = \alpha'$. The only difference here is the coverage rage $(\alpha')$ that is predetermined and usually very small to represent extreme tails. This test could be also viewed as a multivariate unconditional coverage (MUC) with 2 coverage levels. Thus likelihood ratio statistic of MUC could be used. I will refer this test as $\mathrm{LR_{MG}}$.

## 2.5 Finite Sample Inference

It is worth mentioning that while the look-back period sample is large, i.e. 250 days, the number of violations to be found can be very small, particularly for 5% or even 1% coverage rate. This scarcity of violations make inference that due to a well-known asymptotic distribution of critical value (i.e., Chi-square distribution) ineffective. Thus, in backtest literature, it is common to implement the Monte Carlo technique proposed by Dufour (2006) in order to make rejection or acceptance more correct in order to gauge empirical size and statistical power of any backtest method (see, e.g.; Christoffersen and Pelletier, 2004; Candelon et al., 2011; and Ziggel et al., 2014). Basically, Dufour (2006)'s technique use Monte Carlo simulation to approximate unknown distribution under the null hypothesis. For this reason, in my study I will apply of Dufour (2006)'s testing technique to make an inference from all backtests methods except the three backtests from MCS-based tests (Ziggel et al., 2014), which already account for small sample bias. Details on computation are described in the following.

Denote by S, the test statistic calculated from any backtest. I first simulate N sequence of violations that have i.i.d. Bernoulli distribution with the parameter equal to coverage rate (α), which all the sequence are according with the null hypothesis that the model is correct. Then, for each violation sequence, compute test statistic and get $S_i, \dots, S_N$. Then, generally we can calculate $S_0$ from empirical data and compare it to the simulated test statistic. However, when working with binary sequence, there is a probability of observing ties between test statistics calculated from empirical data ($S_0$) and simulated data ($S_i$). Thus, tiebreaking procedure is needed. Following Christoffersen and Pelletier (2004), when calculating each simulated test statistic, $S_i$, I draw independent realization of a uniform distribution [0,1], denoted by $U_i, i = 1, \dots, N$. And the Monte-Carlo p-value is calculated by:

$$\widehat{p_N}(S_0) = \frac{N\widehat{G}(S_0) + 1}{N + 1} \tag{43}$$

Where $\widehat{G_N}(LR_0) = 1 - \frac{1}{N}\sum_{i=1}^{N} I(S_i \leq S_0) + \frac{1}{N}\sum_{i=1}^{N} I(S_i = S_0) \cdot I(U_i \geq U_0)$, and indicator $I(\cdot)$ means that if the condition hold, $I(\cdot) = 1$. Otherwise, $I(\cdot) = 0$.

# CHAPTER 3: Data and Methodology

## 3.1 Data

In empirical implementation, SET index daily return will be used as actual P&L in this study. Sample will be collected from the 2nd Jan, 1990 to the 30th December, 2014 for total 6,126 observations. This include the period of 1997 Asian financial crisis and 2008 Global financial crisis. To judge between range of criteria sets in ex post validation, the outcome period between 3rd April, 1995 to the 30th December, 2014 (full-sample period), for total 4,836 observations will be used for RaVaR backtests.

## 3.2. Methodology

### 3.2.1 Criteria sets Construction

Individual backtests criteria that will be used in this are assigned into four groups characterized by VaR-violation property shown in **Figure 5**. There are totally 15 (individual) backtests, including two backtests that were used in Boucher et al. (2014), namely, Kupiec (1995) - Unconditional Coverage (UC) test **[1]**, Christoffersen, (1998) - First-order Markov (ind) test **[5]**[4].

[4] For magnitude (MG) test, I choose Double-threshold test instead of the Berkowitz (2001) magnitude test used in Boucher (2014) due to the convergent problem. The idea of these two tests are the same.

**Figure 5. Backtest Methods for Model Risk Correction framework.**

*Backtest methods for model risk correction framework*

*(i) Unconditional Coverage Test [UC]*

*[1] An Unconditional Coverage Test (Kupiec, 1995) - (hereafter $LR_{UC}$)*

*[2] A GMM Duration-based (UC) Test (Candelon et al., 2011) - (hereafter $GMM_{UC}$)*

*[3] A Dynamic Quantile (UC) Test (Engle and Manganelli, 2004) - (hereafter $DQ_{CC}$)*

*[4] A MCS (UC) Test (Ziggel et al., 2014) - (hereafter $MCS_{UC}$)*

*(ii) Independence Test [IND]*

*[5] A First-order Markov (IND) Test (Christoffersen, 1998) – (hereafter $LR_{IND}$)*

*[6] A Duration-based (IND) Test (Christoffersen and Pelletier, 2004) - ($DUR_{IND}$)*

*[7] A GMM Duration-based (IND) Test (Candelon et al., 2011) - ($GMM_{IND}$)*

*[8] A Dynamic Quantile (IND) Test (Engle and Manganelli, 2004) - ($DQ_{IND}$)*

*[9] A MCS (IND) Test (Ziggel et al., 2014) - (hereafter $MCS_{IND}$)*

*(iii) Conditional Coverage Test [CC]*

*[10] A First-order Markov (CC) Test (Christoffersen, 1998) - ($LR_{CC}$)*

*[11] A Duration-based (CC) Test (Christoffersen and Pelletier, 2004) - ($DUR_{CC}$)*

*[12] A GMM Duration-based (CC) Test (Candelon et al., 2011) - ($GMM_{CC}$)*

*[13] A Dynamic Quantile (CC) Test (Engle and Manganelli, 2004) - ($DQ_{CC}$)*

*[14] A MCS (CC) Test (Ziggel et al., 2014) - ($MCS_{CC}$)*

*(iv) Magnitude Test [MG]*

*[15] A Double-threshold Test (Colletaz et al., 2013) – ($LR_{MG}$)*

All IND tests have a joint version (CC tests) proposed by the same authors. Most authors proposed many (individual) backtests, for example Christoffersen (1998) proposed two Markov tests, which are IND test and CC test. Candelon et al. (2011) and Ziggel et al. (2014) even proposed three (individual) backtests. This can strengthen the fact that both IND and CC tests are the main focus of literature. For magnitude (MG) property test, however, is less popular than others. Therefore, I restrict the extension of backtesting framework mainly on UC, IND, and CC tests.

Regarding the two hypotheses, the constructed set of new criteria sets are shown in **Table 1**. All criteria sets are divided into three groups, **A**, **B**, and **C**. For each criteria set in Group **A** (a1, a2, a3, a4, a5), there are three individual backtests, each backtest account for one type of violation-property (UC, IND, and MG). To test hypothesis 2 ("*Adding CC test in a criteria set will improve performance in ex post validation*"), for each criteria set in a1, a2, a3, a4, a5, I include another backtest (a CC test of the same methods) with MCScc, GMMcc, DQcc, DURcc, and LRcc, respectively and get resulting five criteria set in Group **B**, which are b1, b2, b3, b4, and b5, respectively. Thus, each criteria set in Group **B** include four individual backtests, one accounts for particular type of violation-property (UC, IND, CC, and MG). To compare between the two separate tests (UC and IND), with the joint test (CC), another five criteria sets are also constructed in Group **C**, by replacing UC and IND tests in each criteria set in Group **A** with a CC test.

Regarding the "Class (Methods of test statistic computation)" of criteria set, I have divided all backtests into 5 classes, namely Monte Carlo (MCS) based-test (for criteria set a1/b1), GMM duration-based test (for criteria set a2/b2), dynamic quantile (DQ)-based test (for criteria set a3/b3), duration-based test (for criteria set a4/b4), and simple likelihood ratio (LR)-based test (for criteria set a5/b). Note that class represent specific characteristics for all type of backtest methods except MG test. For MG test, I restrict to use Colletaz et al. (2013) for all criteria set. Note also that duration-based test (for criteria set a4/b4), the authors only proposed "two" backtests (namely IND and CC). Hence, UC test for criteria sets a4/b4 will be the $LR_{UC}$ of Kupiec (1995) instead.

Specifically, to test hypothesis 1: “criteria sets that contain higher statistical-power backtests will outperform in ex post validation”, based on literature, statistical power of the proposed backtests claimed by authors and the following sub-hypothesis are shown in Table 2. Generally, MCS-based test which is the most recent methodology (compared with others) are claim to outperform existing methods especially in small finite sample case, i.e. backtest period of 250 days. And Simple LR tests (criteria set a5/b5), are among the worst in term of statistical performance. Finally, the following sub-hypothesis are provided in **Table 2.**

For hypothesis 2: "Adding CC test in a criteria set will improve performance in ex post validation"), the sub-hypothesis are as follows:

- Criteria set b1 ≻ a1
- Criteria set b2 ≻ a2
- Criteria set b3 ≻ a3
- Criteria set b4 ≻ a4
- Criteria set b5 ≻ a5

Where preference sign “≻” means “*outperform*”

**Table 1. Set of New Criteria Sets Constructed For Hypothesis 1-2.**

| **Group A: UC+IND+MG** | | | | | |
|---|---|---|---|---|---|
| **Criteria set's name** | **set a1** | **set a2** | **set a3** | **set a4** | **set a5** |
| **Criteria for each set** | 1.$MCS_{UC}$ | 1.$GMM_{UC}$ | 1. $DQ_{UC}$ | 1. $LR_{UC}$ | 1.$LR_{UC}$ |
| | 2.$MCS_{IND}$ | 2.$GMM_{IND}$ | 2.$DQ_{IND}$ | 2.$DUR_{IND}$ | 2.$LR_{IND}$ |
| | 3.$LR_{MG}$ | 3.$LR_{MG}$ | 3.$LR_{MG}$ | 3.$LR_{MG}$ | 3.$LR_{MG}$ |

| **Group B: UC+IND+CC+MG** | | | | | |
|---|---|---|---|---|---|
| **Criteria set's name** | **set b1** | **set b2** | **set b3** | **set b4** | **set b5** |
| **Criteria for each set** | 1. $MCS_{UC}$ | 1.$GMM_{UC}$ | 1. $DQ_{UC}$ | 1.$LR_{UC}$ | 1. $LR_{UC}$ |
| | 2.$MCS_{IND}$ | 2.$GMM_{IND}$ | 2.$DQ_{IND}$ | 2.$DUR_{IND}$ | 2.$LR_{IND}$ |
| | 3.$MCS_{CC}$ | 3.$GMM_{CC}$ | 3.$DQ_{CC}$ | 3.$DUR_{CC}$ | 3.$LR_{CC}$ |
| | 4.$LR_{MG}$ | 4.$LR_{MG}$ | 4.$LR_{MG}$ | 4.$LR_{MG}$ | 4.$LR_{MG}$ |

| **Group C: CC+MG** | | | | | |
|---|---|---|---|---|---|
| **Criteria set's name** | **set c1** | **set c2** | **set c3** | **set c4** | **set c5** |
| **Criteria for each set** | 1.$MCS_{CC}$ | 1.$GMM_{CC}$ | 1. $DQ_{CC}$ | 1. $LR_{CC}$ | 1.$LR_{CC}$ |
| | 2.$LR_{MG}$ | 2.$LR_{MG}$ | 2.$LR_{MG}$ | 2.$LR_{MG}$ | 2.$LR_{MG}$ |

Criteria sets in Group **A** include three violation-property tests, namely UC, IND, and MG tests. To test *Hypothesis 2*, each criteria sets in Group **B** I include another violation-property component, CC test. Lastly, to compare between the two separate tests (UC and IND), with the joint test (CC), another five criteria sets are also constructed in Group **C**. For each set in Group **A**, **B** and **C**, there are five criteria sets, chosen for specified class of backtests. For instance, Monte Carlo (MCS)-based tests for set a1 and b1, GMM duration based-test for set a2 and b2, Dynamic quantile (DQ) based-test for set a3 and b3, Duration-based tests in set b4 and b4, and (simple) Likelihood Ratio based-tests in set a5 and b5. For each set, (individual) backtests are shown by their condensed name. For set a4 and b4, there is no Duration (DUR) method for UC test, thus the UC test of $LR_{UC}$ is used.

**Table 2. Statistical Power of the Proposed Backtests Claimed by Authors**

| Authors | Claim | Sub-Hypothesis 1 |
|---|---|---|
| Monte Carlo based-tests (Ziggel et al. 2014): | $MCS_{UC} > GMM_{UC}, LR_{UC}$ | 3.1: “a1 ≻ a2, a4, a5”<br>“b1≻ b2, b4, b5”<br>and “c1≻ c2, c4, c5” |
| | $MCS_{IND} > GMM_{IND}, DUR_{IND}, LR_{IND}$ | |
| | $MCS_{CC} > GMM_{CC}, DUR_{CC}, LR_{CC}$ | |
| GMM duration based-tests (Candelon et al. 2011): | $GMM_{IND} > DUR_{IND}, LR_{IND}$ | 3.2: “a2 ≻ a4, a5”<br>“b2≻ b4, b5” and<br>“c2≻ c4, c5” |
| | $GMM_{CC} > DUR_{CC}, LR_{CC}$ | |
| Duration based-tests (Christoffersen and Pelletier 2004): | $DUR_{IND} > LR_{IND}$ | 3.3: “a4 ≻ a5”<br>“b4 ≻ b5” and<br>“c4 ≻ c5” |
| | $DUR_{CC} > LR_{CC}$ | |
| Dynamic quantile based-tests (Engle and Manganelli 2004): | $DQ_{IND} > LR_{IND}$ | 3.4: “a3 ≻ a5”<br>“b3 ≻ b5” and<br>“c3 ≻ c5” |
| | $DQ_{CC} > LR_{CC}$ | |

The sign “greater than” OR “ > ” means “higher statistical power” in detecting inaccurate model. For example “$MCS_{CC} > GMM_{CC}, DUR_{CC}, LR_{CC}$” means that $MCS_{CC}$ test have greater chance to reject the inaccurate model (with respective to CC hypothesis) than $GMM_{CC}$ test, $DUR_{CC}$ test, and $LR_{CC}$ test. The last column shown the corresponding sub-hypothesis of *Hypothesis 1*. The sign “preference” OR “≻” means outperform in ex post validation. Note that all papers above applied Dufour (2006)’s Monte Carlo method to indicate statistical power of their backtests.

### 3.2.2 Ex Post Validation

Ex post validation will be based on two discretions. The first discretion is out-of-sample backtest using general backtest procedures to evaluate series of RaVaR. Performance of criteria sets will be based on acceptance frequency ratio (AFR) based on 5% significance level. The second ex post validation method is risk ratio analysis (Danielsson et al., 2014). Roughly speaking, given a set of risk model's forecasts, risk ratio is calculated by the highest risk forecasts divided by the lowest risk forecasts. In the absence of model risk, the ratio should be very close to 1. Other numbers from 1 represent the degree of dispersion (and hence degree of model risk) among the models. To judge, criteria sets that provide less model risk on RaVaR series (lower risk ratio) are considered more robust. The steps are as follows:

**3.2.2.1 Forecast EVaR$_t$:** Using SET index daily return, for each model (including Normal, historical simulation, RiskMetrics, GARCH(1,1)-N and GARCH(1,1)-t), calculate daily 95% estimated VaR (EVaR) with rolling (estimation) window of 1,040 observations starting from the 28th March, 1994 to the 30th December, 2014 and obtain 5,086 daily forecasts.

**3.2.2.2 Find $\mathbf{q_t^*}$:** For each of the 15 criteria sets and each of the 5 models, find the optimal adjustment $q_t^*$, for any time t, with rolling backtest window (daily) until the last observation. The method for finding $q_t^*$ is the numerical search algorithm with step size of 0.1% multiplied by $\mathrm{EVaR_t}$.

**3.2.2.3 Compute RaVaR$_t$:** add $q_t^*$ to $\mathrm{EVaR_t}$ to get "Model-Risk-Adjusted VaR" ( $\mathrm{RaVaR_t}$) series from the 3rd April, 1995 to the 30th December, 2014 (4,836 observations). For 15 criteria sets and 5 data generating models, I therefore obtain 75 (15x5) $\mathrm{RaVaR_t}$ series.

**3.2.2.4 RaVaR backtesting:** For each $\mathrm{RaVaR_t}$ series, backtest them for full-sample 4,586 (i.e. 4,836 - 250) observations. This means that for each $\mathrm{RaVaR_t}$ series, I backtest it "4,586" times with rolling backtest window (daily) until the last observation. Eight backtests are used in this process, namely[5]

1. GMM Duration-based (UC) Test (Candelon et al., 2011)
2. GMM Duration-based (IND) Test (Candelon et al., 2011)
3. GMM Duration-based (CC) Test (Candelon et al., 2011)
4. MCS (UC) Test (Ziggel et al., 2014)
5. MCS (IND) Test (Ziggel et al., 2014)
6. MCS (CC) Test (Ziggel et al., 2014)
7. Double-threshold Test (Colletaz et al., 2013)

The performance of criteria sets will be based on **"acceptance frequency ratio ($\mathbf{AFR_i}$)"**. For each of 7 backtests above, acceptance frequency ratio (AFR) is computed by:

$$\mathrm{AFR_i} = \frac{\text{No. of acceptance}}{\text{No. of observations}} \,, \qquad (44)$$

where i = 1, 2, 3, …, 7, No. of observations is 4,586, and No. of acceptance is times that a particular backtest (at any time t) "accept" the model (infer that the model is acceptably accurate). Thus, I'll get 7 acceptance frequency ratios for each $RaVaR_t$ series. Then average it and get:

[5] The choice is based on the backtest methods that were recently proposed and shown to outperform many existing ones in realistic small sample size, i.e., 250 backtest period.

$$\overline{\mathrm{AFR}} = \frac{1}{7}\sum_{i=1}^{7} \mathrm{AFR_i} \qquad (45)$$

Which criteria sets provide RaVaR backtests that have higher $\overline{\mathrm{AFR}}$ are considered outperforming. In addition, two sub sample cases will be applied. The first sub-sample includes 500 days around the 1997's crisis (27th June, 1996 – 13th July, 1998), whereas the second sub-sample includes around the 2008 crisis (11th September, 2007 – 25th September, 2009).

**3.2.2.5 Risk ratio analysis**[6] (Danielsson et al. 2014): For each criteria sets, given 5-model-RaVaR forecasts, the estimated risk ratio is given by:

$$\hat{\mathrm{R}} = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \left(\frac{\mathrm{LowerboundRaVaR_t}}{\mathrm{UpperboundRaVaR_t}}\right) \qquad (46)$$

Where $\mathrm{LowerboundRaVaR_t}$ is the highest risk forecast (in absolute value), $\mathrm{UpperboundRaVaR_t}$ is the lowest risk forecast (in absolute value), and T is 4,836. In the absence of model risk the ratio should be close to 1. Criteria sets that provide less model risk (the ratio is closer to 1) are considered better. Finally, two sub-sample analyses of the 1997's crisis and 2008's crisis will also be applied.

[6] Note that risk ratio analysis entails employing all models (once for each criteria set) to see inconsistency among risk forecasts, whereas RaVaR backtesting employs once for individual model (and once for each criteria set). Thus, they are of different dimensions.

# CHAPTER 4: Empirical Result

This chapter describes the results from model risk quantification and correction, which are divided into three sections. The first section is the implementation of the model risk correction framework quantified by each individual (backtest) criterion, i.e. require each EVaR to pass each baktest alone for adjusting risk. The second section describe the main results from each criteria set, i.e. require each EVaR to pass all backtest in criteria set. The last section will be ex post validation.

## 4.1 Model risk Calibrated by Individual (Backtest) Criterion

Generally, the adjustment amount of model risk are of equal sign, meaning that at particular time t, if initial VaR models are strongly overestimating risk (more negative), $q_t^*$ are generally positive to make the risk forecasts acceptable with regards to particular criteria. Similarly, if our models are strongly underestimating risk (less negative), $q_t^*$ are generally negative. In few situations where our model are neither heavily overestimating nor underestimating, that is, the number of violations stays in acceptable range (around 8-17 times from the mean of 12.5 times (250 x 5% coverage rate), sign's disagreement of model risk occurred in independence (IND) and (CC) tests due to the fact that violations clustering is quite complicated that criteria for IND and CC property are turned from reject to accept the null in different ways (e.g., one case can be done by adding positive $q_t^*$, and another case require negative $q_t^*$). However, it is of no concern because in main study I use criteria set to adjust EVaR for model risk, making all criteria in a particular set to accept null hypothesis, not individual criteria alone, and even though it sill occurred, those sign’s disagreements are not a problem by themselves.

Consider first the results from adjusting each of the five EVaR series to pass each individual criteria alone. **Table 3** show the mean of model risk adjustment ($q_t^*$), including positive adjustment and negative adjustment. The unit is presented in percentage of EVaR. When looking across the property-type test, including Unconditional Coverage (UC) Independence (IND), Conditional Coverage (CC) and Magnitude test (MG). UC tests for all methods/classes of backtest criteria (LR-based, GMM-based, DUR-based, and MCS-based) are required less amount level than IND and CC tests to pass the 5% significance level. This come from the fact that, normally the problem of unconditional coverage (too high/low violations) is much less severe than dependence in violation sequences (violation clusterings) as it means that our risk models is too naive to predict dynamics in financial risks. In fact, frequently unexpected losses may result in even more serious problem to financial institution, i.e. bankruptcy (See, e.g., Campbell, 2005; Christoffersen and Pelletier, 2004).

## Table 3. Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Individual Criterion

| Individual Criterion | GARCH | | GARCH-t | | EWMA | | HS | | Normal | |
|---|---|---|---|---|---|---|---|---|---|---|
| | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) |
| **LRuc** | -1.15% | 1.37% | -0.30% | 1.50% | -0.47% | 1.26% | -1.89% | 1.78% | -1.49% | 1.92% |
| **LRind** | -1.19% | 1.78% | -0.56% | 1.41% | -1.24% | 1.63% | -3.48% | 2.39% | -3.64% | 2.59% |
| **LRcc** | -1.47% | 1.48% | -0.39% | 1.44% | -0.95% | 1.19% | -1.80% | 1.92% | -1.47% | 2.02% |
| **LRmg** | -1.16% | 1.47% | -0.37% | 1.27% | -0.92% | 0.98% | -1.86% | 1.40% | -1.62% | 1.60% |
| **GMMuc** | -1.14% | 1.37% | NA | 1.50% | -0.45% | 1.26% | -1.88% | 1.84% | -1.48% | 1.98% |
| **GMMind** | -1.65% | 0.89% | NA | 1.35% | -1.37% | 1.34% | -1.95% | 1.50% | -2.26% | 1.74% |
| **GMMcc** | -1.28% | 1.35% | NA | 1.51% | -0.85% | 1.23% | -1.77% | 2.09% | -1.64% | 2.13% |
| **DURind** | -1.10% | 1.13% | NA | 1.31% | NA | 0.60% | -1.78% | 1.20% | -1.65% | 1.63% |
| **DURcc** | -1.20% | 1.34% | NA | 1.38% | -0.45% | 1.08% | -1.90% | 1.69% | -1.55% | 1.88% |
| **DQuc** | -1.12% | 1.48% | NA | 1.60% | -0.49% | 0.91% | -1.83% | 1.54% | -1.51% | 1.56% |
| **DQind** | -1.52% | 1.78% | -2.11% | 1.36% | -1.77% | 1.59% | -2.87% | 2.26% | -2.83% | 2.40% |
| **DQcc** | -1.68% | 1.61% | -0.82% | 1.24% | -1.41% | 1.42% | -2.59% | 1.80% | -2.47% | 1.92% |
| **MCSuc** | -1.14% | 1.08% | -0.23% | 1.33% | -0.47% | 1.07% | -1.85% | 1.66% | -1.54% | 1.82% |
| **MCSind** | -2.06% | 0.98% | -0.93% | 1.21% | -1.85% | 1.61% | -2.38% | 1.42% | -2.35% | 1.55% |
| **MCScc** | -1.30% | 1.46% | -0.30% | 1.29% | -0.51% | 0.61% | -1.57% | 1.34% | -1.25% | 1.38% |

Datasource: Bloomberg. SET index daily data from the 2nd Jan, 1990 to the 30th December, 2014. EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. $q_t^*$ are calculared using 250 backtest period from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The unit is present in percentage of EVaR. NA means non-existent value. For example, if q*(-) does not exist, it means that there is no negative adjustment for that criterion (row) when using particular model (column).

According to one research question that focus on whether amount of model risk depend on which methods (data generating models (DGM)) I used to compute VaR, it can be seen that amount of model risks quantified by all individual criteria are relatively high for static VaR methods (namely, HS, and Normal) and relatively low for dynamic VaR methods (namely EWMA, GARCH-N, and GARCH-t. The result does support for both negative and positive $q_t^*$'s mean. In almost all cases, HS and Normal are the models that provide the largest and the second largest average of model risks, respectively. Compare within dynamic VaR method, GARCH-t gave have higher model risk than GARCH-N and EWMA. The higher positive means for GARCH-t in some cases may indicate that sometimes this model is quite overestimating risk.

### 4.2 Model risk Calibrated by Criteria Sets

I now discuss the main results of quantile model risk correction using (backtest) criteria sets. **Table 4** shows the average of positive and negative model risks relative 95% to $EVaR_t$. Not only the results from individual criterion shown in **Table 3** that amount of model risks are be model-dependent, the same applied here. In case of negative $q_t^*$'s mean, HS and Normal still provide the largest and the second largest amount of model risks, respectively, in all cases except some cases, where GARCH-t become the first rank. Mean of positive $q_t^*$'s is also in line with the model-dependence. The support of hypothesis 1 also verifies, to some degree, robustness of this model risk correction framework as the estimated magnitude of VaR-model risk are harmonious with its definition. In addition, although between criteria sets the average amount of model risk does not vary much, but when it does, outcome could be different.

Next, maximum of negative and positive model risk adjustment are shown in **Table 5**. Most of the negative adjustments were happened in stressed period, such as in the 1997's and the 2008's crises, and other volatile markets. The most negative and maximum adjustment happened in criteria set b3, -5.40% (EWMA), and 4.60% (Normal), respectively. In fact, the maximum adjustment for a3 and b3 is the same. This means that adding CC test (or in particular, $DQ_{CC}$) to criteria set a3 does not matter in extreme model risk. The lowest maximum negative (positive) adjustment is -2.10% in EWMA (1.60% in GARCH-t) for criteria set a1.

**Table 4. Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set**

| Criteria | GARCH | | GARCH-t | | EWMA | | HS | | Normal | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sets | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) |
| **a1** | -1.34% | 1.08% | -0.93% | 1.28% | -1.34% | 1.29% | -2.03% | 1.85% | -1.84% | 1.94% |
| **a2** | -1.44% | 1.29% | -0.82% | 1.49% | -1.23% | 1.30% | -2.13% | 2.09% | -1.82% | 2.17% |
| **a3** | -1.45% | 1.57% | -1.83% | 1.53% | -1.52% | 1.43% | -2.25% | 2.01% | -1.91% | 1.91% |
| **a4** | -1.19% | 1.37% | -0.38% | 1.50% | -0.89% | 1.24% | -1.90% | 1.84% | -1.64% | 1.99% |
| **a5** | -1.40% | 1.53% | -0.56% | 1.51% | -1.14% | 1.52% | -2.18% | 2.17% | -1.83% | 2.21% |
| **b1** | -1.33% | 1.09% | -0.93% | 1.29% | -1.34% | 1.29% | -2.03% | 1.86% | -1.83% | 1.96% |
| **b2** | -1.41% | 1.34% | -0.82% | 1.52% | -1.21% | 1.36% | -2.13% | 2.22% | -1.82% | 2.25% |
| **b3** | -1.55% | 1.69% | -1.61% | 1.46% | -1.64% | 1.54% | -2.22% | 2.06% | -1.99% | 1.95% |
| **b4** | -1.19% | 1.38% | -0.35% | 1.50% | -0.89% | 1.22% | -1.90% | 1.84% | -1.64% | 1.99% |
| **b5** | -1.43% | 1.53% | -0.55% | 1.51% | -1.12% | 1.52% | -2.17% | 2.17% | -1.82% | 2.22% |
| **c1** | -1.18% | 1.48% | -0.34% | 1.30% | -0.92% | 0.98% | -1.86% | 1.45% | -1.59% | 1.65% |
| **c2** | -1.29% | 1.35% | -0.47% | 1.51% | -0.96% | 1.32% | -1.67% | 2.20% | -1.51% | 2.23% |
| **c3** | -1.43% | 1.57% | -0.78% | 1.31% | -1.44% | 1.35% | -1.95% | 1.87% | -1.88% | 1.99% |
| **c4** | -1.16% | 1.41% | -0.38% | 1.45% | -0.92% | 1.12% | -1.86% | 1.73% | -1.60% | 1.95% |
| **c5** | -1.47% | 1.48% | -0.40% | 1.44% | -0.95% | 1.17% | -1.64% | 1.96% | -1.48% | 2.08% |

Datasource: Bloomberg. SET index daily data from the 2$^{nd}$ Jan, 1990 to the 30th December, 2014. EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. $q_t^*$ are calculated using 250 backtest period from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The unit is present in percentage of EVaR

### Table 5. Maximum of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set

| Criteria Set | GARCH | | GARCH-t | | EWMA | | HS | | Normal | |
|---|---|---|---|---|---|---|---|---|---|---|
| | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) |
| **a1** | -2.30% | 3.60% | -3.10% | 1.60% | -2.10% | 3.60% | -3.50% | 4.40% | -3.60% | 3.30% |
| **a2** | -3.52% | 5.15% | -3.00% | 1.80% | -2.10% | 3.60% | -3.50% | 4.20% | -3.60% | 3.30% |
| **a3** | -3.40% | 4.20% | -3.20% | 3.40% | -5.40% | 4.40% | -3.30% | 4.40% | -3.50% | 4.60% |
| **a4** | -2.40% | 2.20% | -3.00% | 0.60% | -2.10% | 2.50% | -3.40% | 3.90% | -3.60% | 3.00% |
| **a5** | -3.20% | 2.90% | -3.00% | 1.10% | -2.90% | 3.50% | -3.30% | 4.10% | -3.50% | 3.30% |
| **b1** | -2.40% | 3.60% | -3.10% | 1.60% | -2.00% | 3.60% | -3.50% | 4.40% | -3.60% | 3.30% |
| **b2** | -2.40% | 3.30% | -3.00% | 1.80% | -2.70% | 3.60% | -3.50% | 4.20% | -3.60% | 3.30% |
| **b3** | -3.40% | 4.40% | -3.20% | 3.40% | -5.40% | 4.40% | -3.30% | 4.40% | -3.50% | 4.60% |
| **b4** | -2.40% | 2.20% | -3.00% | 0.50% | -2.10% | 2.50% | -3.40% | 3.90% | -3.60% | 3.00% |
| **b5** | -3.20% | 2.90% | -3.00% | 1.10% | -2.90% | 3.50% | -3.30% | 4.10% | -3.50% | 3.30% |
| **c1** | -2.20% | 2.20% | -0.60% | 2.90% | -1.60% | 2.00% | -3.20% | 3.22% | -2.60% | 3.20% |
| **c2** | -3.48% | 2.30% | -1.40% | 3.00% | -2.61% | 2.20% | -3.20% | 3.22% | -2.60% | 3.33% |
| **c3** | -2.80% | 3.18% | -2.20% | 3.00% | -4.62% | 2.34% | -4.00% | 3.36% | -3.20% | 3.35% |
| **c4** | -2.20% | 2.30% | -0.60% | 3.00% | -1.60% | 2.00% | -3.20% | 3.30% | -2.60% | 3.30% |
| **c5** | -2.80% | 3.00% | -0.60% | 3.00% | -2.20% | 2.10% | -2.80% | 3.00% | -2.30% | 3.20% |

Datasource: Bloomberg. SET index daily data from the 2$^{nd}$ Jan, 1990 to the 30th December, 2014. EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. $\mathrm{q}_\mathrm{t}^*$ from 3rd April 1995 to the 30th December, 2014 (4,836 observations) are calculated using rolling window (250 backtest period). The unit is present in percentage of EVaR

To illustrate more about $\mathrm{q}_\mathrm{t}^*$ series, **Figure 6-10** show time-varying model risk adjustment through the timeline of the 3$^{rd}$ April, 1995 to the 30$^{th}$ December, 2014 calibrated by criteria sets a1-a5, respectively[7]. The left panel of each figure represent $\mathrm{q}_\mathrm{t}^*$ series from two dynamic GARCH models, while the right panel represent $\mathrm{q}_\mathrm{t}^*$ series

[7] For criteria set in Group **B** (b1, b2, b3, b4, b5) the results are similar to Group **A** (a1, a2, a3, a4, a5), respectively. For criteria set in Group **C** (c1, c2, c3, c4, c5), q* are relatively low compared with others. See **Appendix B**.

from others. As expected GARCH and GARCH-t have less frequency of adjustments than other models, in other words, they have higher non-adjustment (zero $q_t^*$). This zero adjustment means no backtest criteria reject the models at particular day and thus, their risk estimates are effective than others. Nevertheless, two GARCH models still can’t capture dynamics of SET index expressed by series of adjustment. When it comes to model risk adjustment (i.e., if one of criteria reject the model), the magnitude of adjustment ($q_t^*$) in GARCH-N and GARCH-t are lower than other models.

**Figure 6. $q_t^*$ for Criteria Set a1 Relative to 95% EVaR**

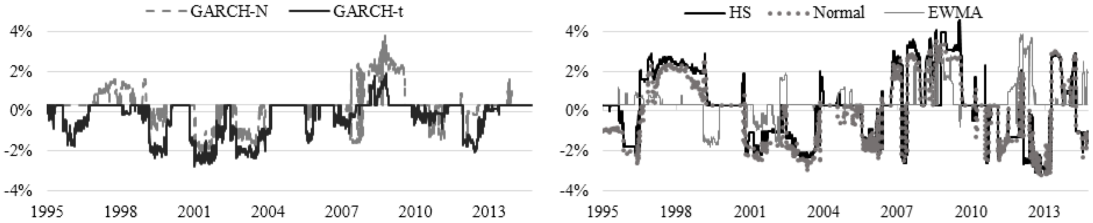


**Figure 7. $q_t^*$ for Criteria Set a2 Relative to 95% EVaR**

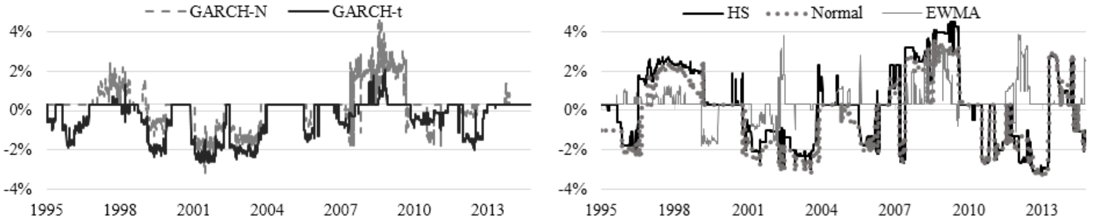


Figure 8. $q_t^*$ for Criteria Set a3 Relative to 95% EVaR

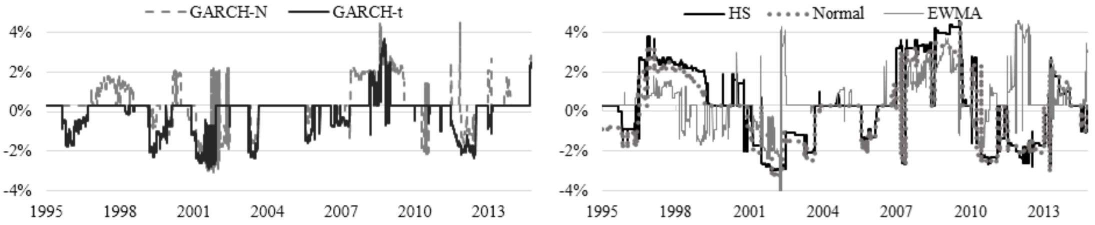


Figure 9. $q_t^*$ for Criteria Set a4 Relative to 95% EVaR

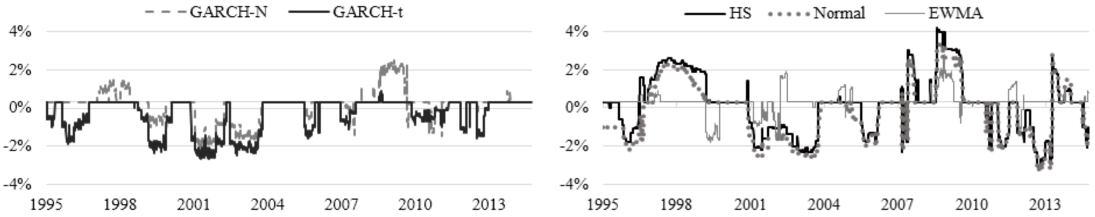
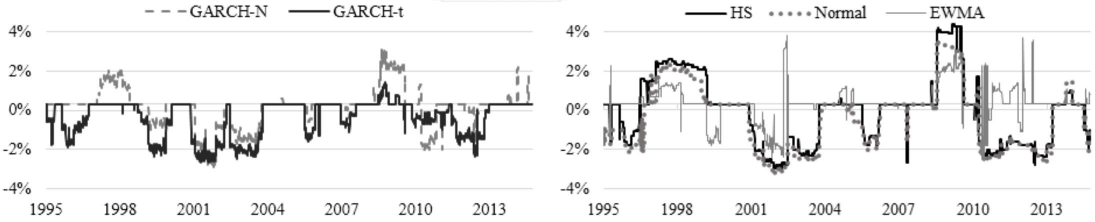
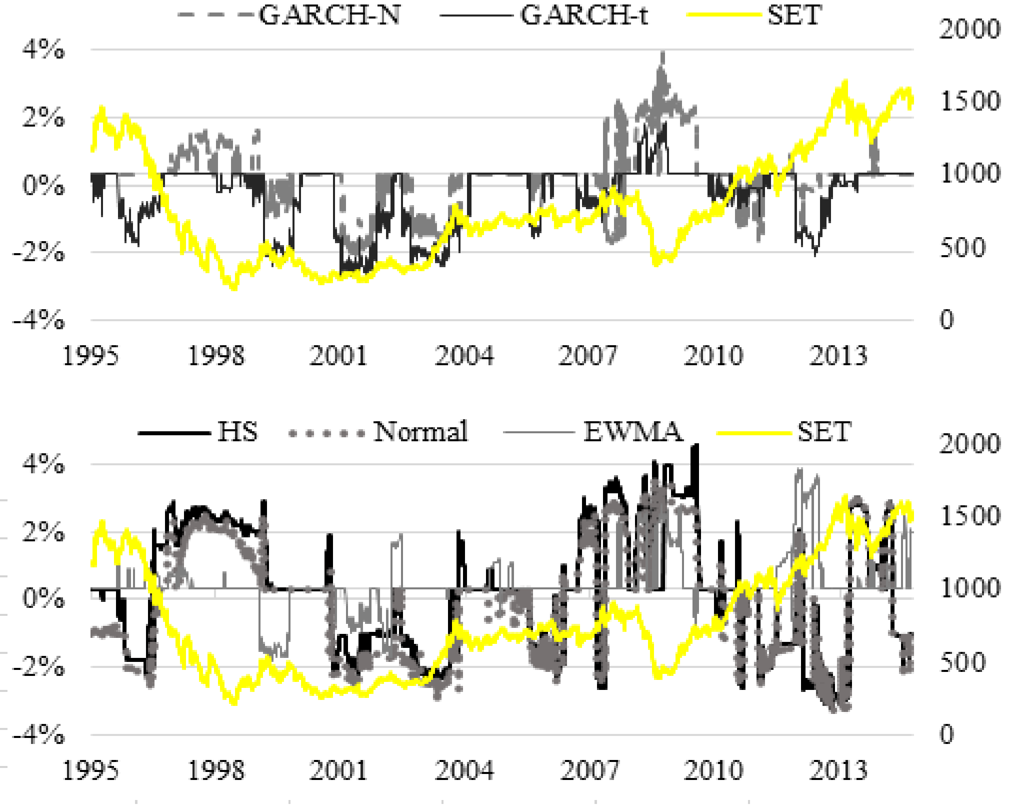


Figure 10. $q_t^*$ for Criteria Set a5 Relative to 95% EVaR



However, in stressed periods, for example, around the Asian financial crisis (July 1997) and the Subprime crisis (August 2008), model risk correction still lags behind the events. For example, **Figure 11** show the result model risk outcome when using criteria set a1 compared with SET index. Specifically, when SET index drastically fell and VaR should be much more negative, the quantile correction framework did not adjust VaR immediately since it required some times to recognize model risks via outcome of backtesting.

**Figure 11. $q_t^*$ for Criteria Set a1 Relative to 95% EVaR with SET index**



Series of SET index, and $q_t^*$ generated from five data generating models include: the 3rd April, 1995 to the 30th December, 2014. The first vertical axis represent amount of model risk in percentage of EVaR, while the secon vertical axis represent point of SET index.

## 4.3 Ex Post Validation

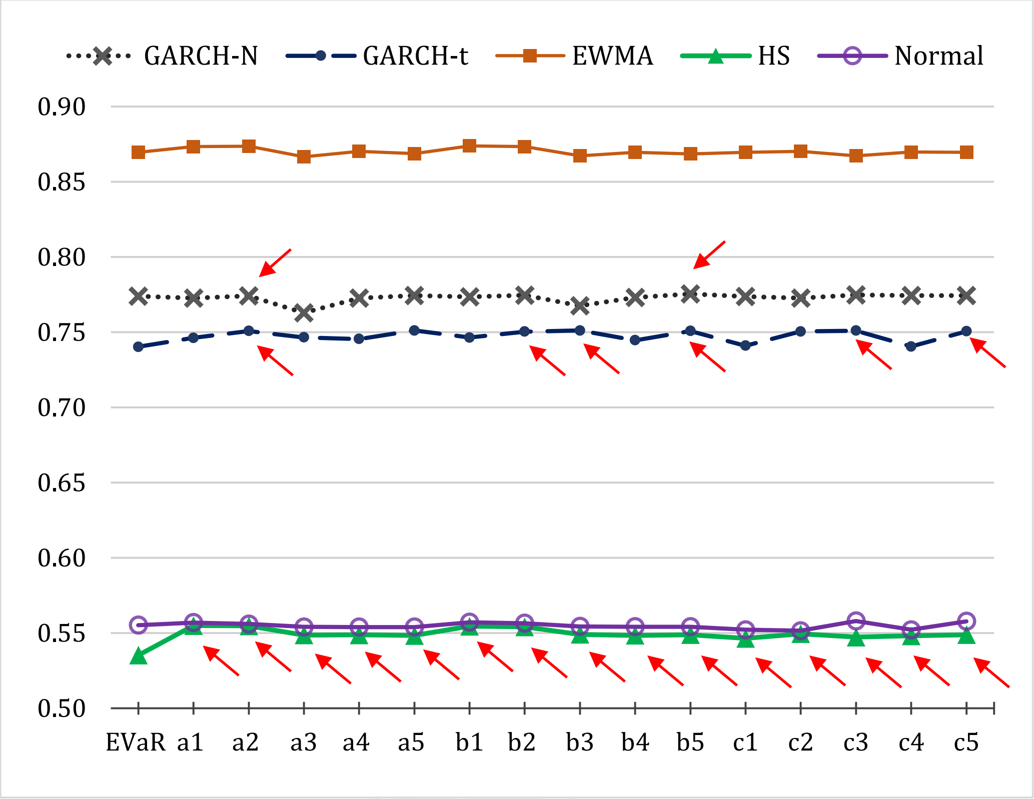
### 4.3.1 Out-of-Sample Backtest

There are two sections for out-of-sample backtest. First, the results generated by 250 days of backtest period will be discussed. However, when using overlapping backtest period the problem could occur on comparing Acceptance Frequency Ratio ($\overline{\text{AFR}}$) between criteria sets. Hence, results when using non-overlapping period of backtest are also performed. But in order to keep amount of sample for doing the t-test, backtest periods's size is reduced to 100 days.

#### 4.3.1.1 Overlapping 250-day Backtest Period

**Figure 12** shows the acceptance frequency ratio ($\overline{\text{AFR}}$) for all RaVaR series generated from all 15 criteria sets as well as for the original VaR series (EVaR). The criteria sets include criteria set a1-a5, b1-b5, and c1-c5. Recall that for the out-of-sample backtest, higher $\overline{\text{AFR}}$ is better. Interestingly, in many cases of criteria sets applied, adjusting EVaR to lessen model risk effects make RaVaR series more reliable by the increased number of acceptance $\overline{\text{AFR}}$ with statistical significance at 5% level, especially in HS. For the risk ratio analysis, all RaVaR series outperform EVaR with statistical significance at 5% level. Also, how large the AFR can be improved depends on methods to calculate VaR. To be more specific, HS and Normal is the first and second rank for having the highest difference between maximum AFR (from RaVaR) and minimum AFR (from EVaR), respectively. However, between each model the ratio does not vary much.

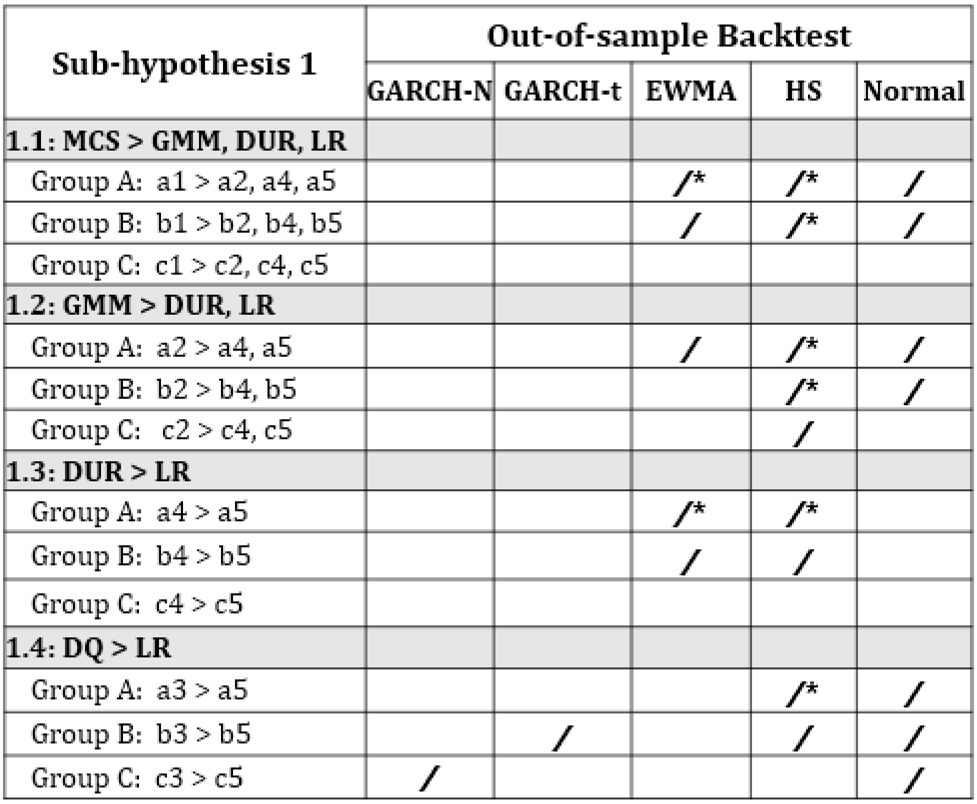
**Figure 12. Acceptance Frequency Ratio for all criteria sets and EVaR (Overlapping)**



"The arrow" means $\overline{\text{AFR}}$ of RaVaR that outperform (higher than) EVaR's at 5% significance level

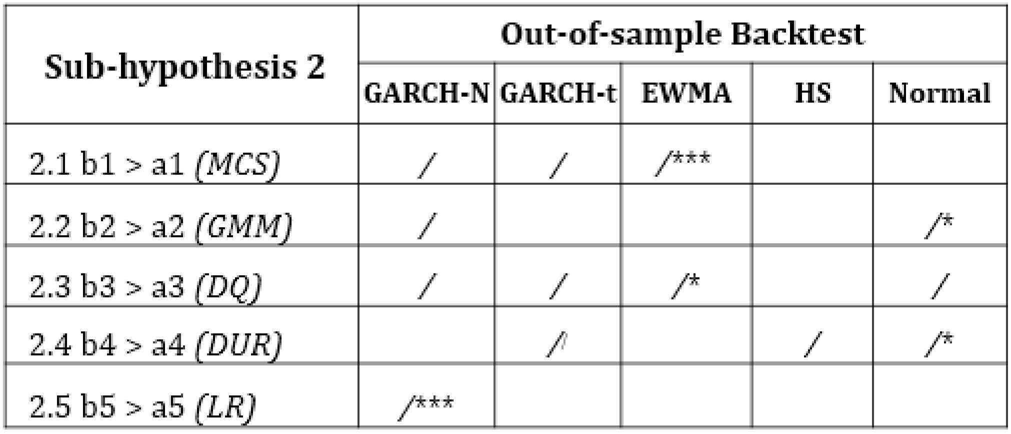
For hypothesis 1, *"Criteria sets that contain higher statistical-power backtests will outperform in ex post validation*", **Table 6** shows that sub-hypothesis is true mostly in static VaR methods and EWMA, especially for HS which hold for all sub-hypothesis, and EWMA. However, most of them are not statistically significance, except in HS VaR method. Again, this also support the idea that data generating models have effects on this quantile correction framework. Turning to analyse hypothesis 2. *"Adding CC test in a criteria set will improve performance in ex post validation"*. **Table 7** shows that although there are many cases that the sub-hypothesis hold, there are not statistically significant. Hence, this hypothesis is inconclusive.

**Table 6. Ex Post Validation Results for Hypothesis 1 (Overlapping)**



| Sub-hypothesis 1 | Out-of-sample Backtest | | | | |
|---|---|---|---|---|---|
| | GARCH-N | GARCH-t | EWMA | HS | Normal |
| **1.1: MCS > GMM, DUR, LR** | | | | | |
| Group A: a1 > a2, a4, a5 | | | /* | /* | / |
| Group B: b1 > b2, b4, b5 | | | / | /* | / |
| Group C: c1 > c2, c4, c5 | | | | | |
| **1.2: GMM > DUR, LR** | | | | | |
| Group A: a2 > a4, a5 | | | / | /* | / |
| Group B: b2 > b4, b5 | | | | /* | / |
| Group C: c2 > c4, c5 | | | | / | |
| **1.3: DUR > LR** | | | | | |
| Group A: a4 > a5 | | | /* | /* | |
| Group B: b4 > b5 | | | / | / | |
| Group C: c4 > c5 | | | | | |
| **1.4: DQ > LR** | | | | | |
| Group A: a3 > a5 | | | | /* | / |
| Group B: b3 > b5 | | / | | / | / |
| Group C: c3 > c5 | / | | | | / |

**Table 7. Ex Post Validation Results for Hypothesis 2 (Overlapping)**

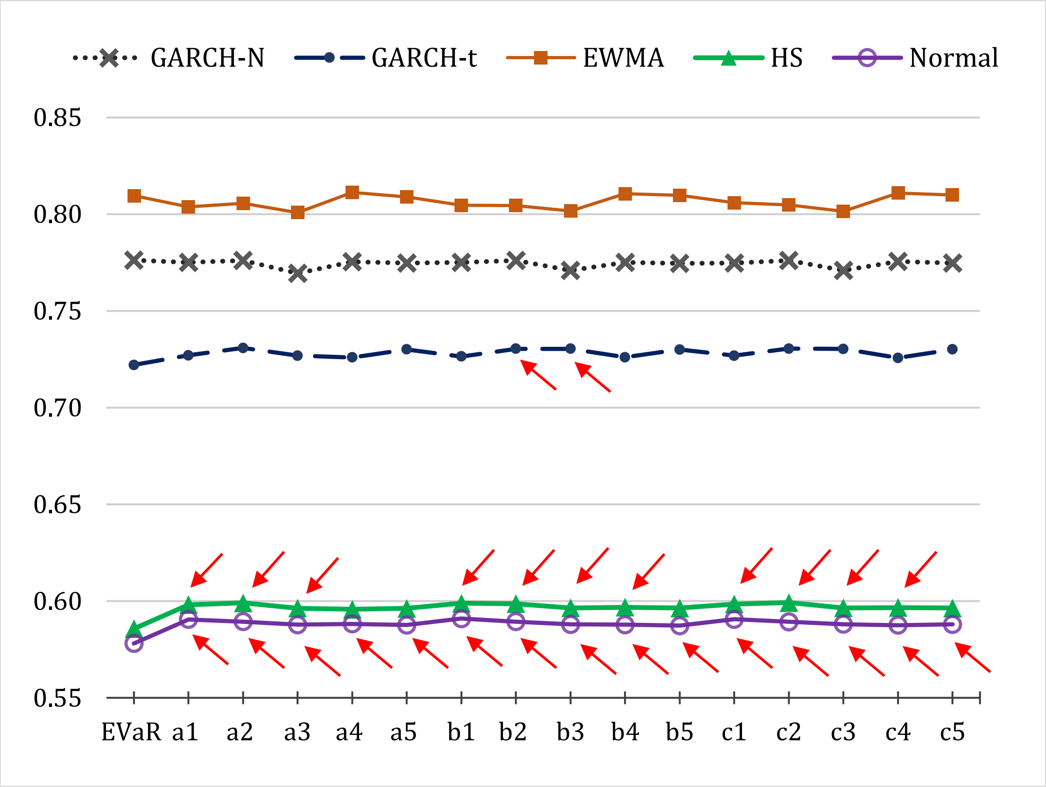


| Sub-hypothesis 2 | Out-of-sample Backtest | | | | |
|---|---|---|---|---|---|
| | GARCH-N | GARCH-t | EWMA | HS | Normal |
| 2.1 b1 > a1 *(MCS)* | / | / | /*** | | |
| 2.2 b2 > a2 *(GMM)* | / | | | | /* |
| 2.3 b3 > a3 *(DQ)* | / | / | /* | | / |
| 2.4 b4 > a4 *(DUR)* | | / | | / | /* |
| 2.5 b5 > a5 *(LR)* | /*** | | | | |

“ / ” means the sub-hypothesis is true
“ > ” means outperform
“ * ” means significance at 10% level
“ *** ” means significance at 1% level

### 4.3.1.2 Non-Overlapping 100-day Backtest Period

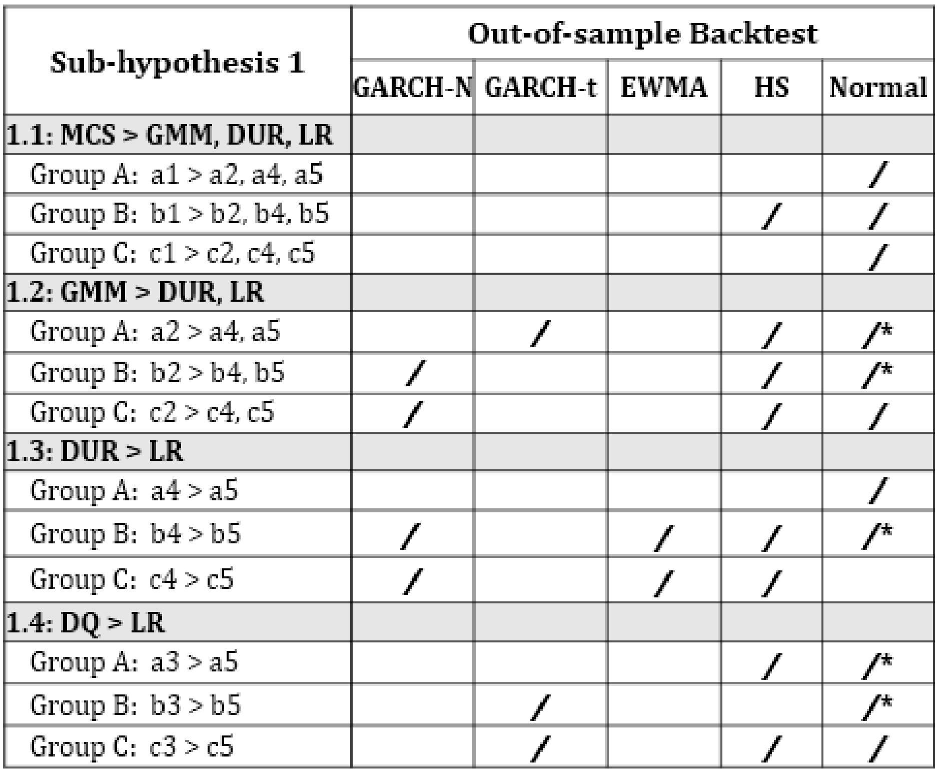
**Figure 13. Acceptance Frequency Ratio for all criteria sets and EVaR (Non-overlapping)**



"The arrow" means $\overline{\text{AFR}}$ of RaVaR that outperform (higher than) EVaR's at 10% significance level

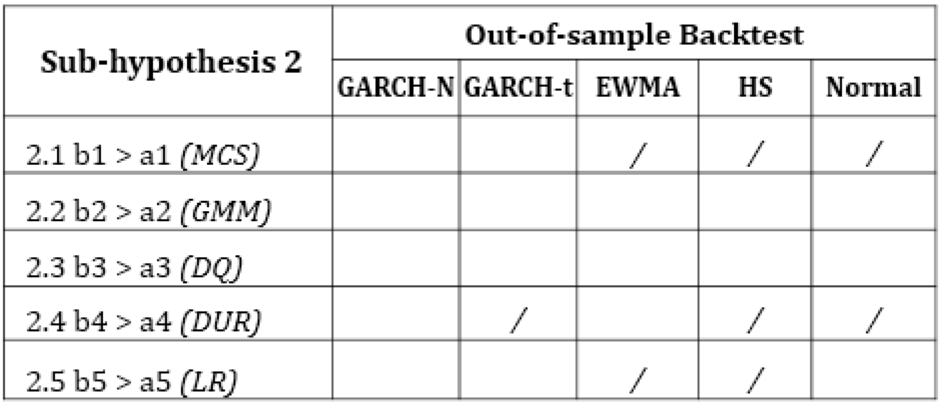
**Figure 13** shows the acceptance frequency ratio ($\overline{\text{AFR}}$) for all RaVaR series generated from all 15 criteria sets as well as for the original VaR series (EVaR) when using 100 days of non-overlapping backtest period. Obviously, RaVaR series are more robust than EVaR series when using static volatility models, HS and Normal distribution with significance at 10% level. But for other DGMs, almost all cases appeared no significant improvement.

**Table 8. Ex Post Validation Results for Hypothesis 1 (Non-overlapping)**



| Sub-hypothesis 1 | Out-of-sample Backtest | | | | |
|---|---|---|---|---|---|
| | GARCH-N | GARCH-t | EWMA | HS | Normal |
| **1.1: MCS > GMM, DUR, LR** | | | | | |
| Group A: a1 > a2, a4, a5 | | | | | / |
| Group B: b1 > b2, b4, b5 | | | | / | / |
| Group C: c1 > c2, c4, c5 | | | | | / |
| **1.2: GMM > DUR, LR** | | | | | |
| Group A: a2 > a4, a5 | | / | | / | /* |
| Group B: b2 > b4, b5 | / | | | / | /* |
| Group C: c2 > c4, c5 | / | | | / | / |
| **1.3: DUR > LR** | | | | | |
| Group A: a4 > a5 | | | | | / |
| Group B: b4 > b5 | / | | / | / | /* |
| Group C: c4 > c5 | / | | / | / | |
| **1.4: DQ > LR** | | | | | |
| Group A: a3 > a5 | | | | / | /* |
| Group B: b3 > b5 | | / | | | /* |
| Group C: c3 > c5 | | / | | / | / |

**Table 9. Ex Post Validation Results for Hypothesis 2 (Non-overlapping)**

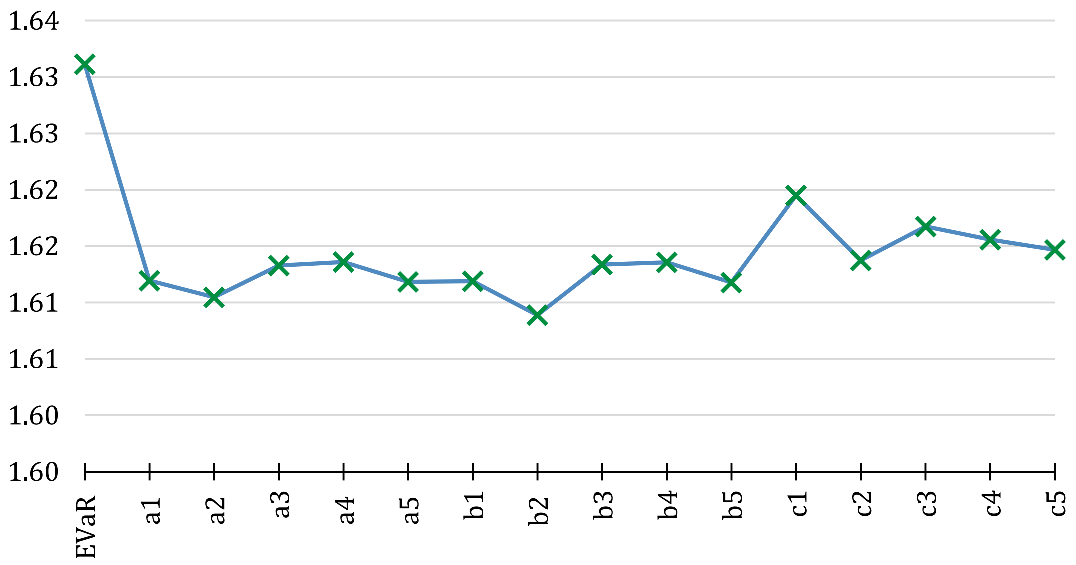


| Sub-hypothesis 2 | Out-of-sample Backtest | | | | |
|---|---|---|---|---|---|
| | GARCH-N | GARCH-t | EWMA | HS | Normal |
| 2.1 b1 > a1 *(MCS)* | | | / | / | / |
| 2.2 b2 > a2 *(GMM)* | | | | | |
| 2.3 b3 > a3 *(DQ)* | | | | | |
| 2.4 b4 > a4 *(DUR)* | | / | | / | / |
| 2.5 b5 > a5 *(LR)* | | | / | / | |

“ / ” means the sub-hypothesis is true
“ > ” means outperform
“ * ” means significance at 10% level

About hypothesis 1 and 2, **Table 8** and **Table 9** show Ex Post Validation Results when using non-overlapping backtest period for hypothesis 1 and 2, respectively. The results for Hypothesis 1 are similar to those when using overlapping backtest period in that, almost no sub-hypothesis is true for dynamic DGMs. But for static DGMs, the sub-hypothesis hold but only for Normal VaR method that significance level appeared. For, hypothesis 2, the results is also conclusive as in the overlapping period case.

**Figure 14. Risk Ratio of all criteria sets (RaVaR) and EVaR**



Risk Ratio of all RaVaR series are lower than the ratio of EVaR with 5% significance level

### 4.3.2 Risk Ratio Analysis

**Figure 14** shows risk ratio for all RaVaR series generated by 15 criteria sets and for original VaR series (EVaR). As can be seen that risk ratio generated from all criteria sets outperform EVaR by giving the lower risk ratio, which the lowest one belongs to set b2. This means that all RaVaR series have lower inconsistency between risk models. However, hypothesis 1 and 2 does not hold with significance level for risk ratio analysis

Overall, results out-of-sample backtest for both overlapping and non-overlapping cases shows that level of statistical power of backtest matter when use static volatility models. For two GARCH models and EWMA, criteria sets seems to be ignorable. Between criteria sets in Group **A**, **B**, and **C**, ex post validation results no obvious difference. Comparing the performance of criteria sets in Group **C** (include only joint test, CC) and criteria sets in Group **A** (include only two separate tests, UC nad IND), although criteria sets in Group **C** provide slightly higher risk ratio, results from AFR are very similar.

# CHAPTER 5: Conclusion

I extend Boucher et al. (2014)'s model risk correction framework by focusing on statistical power to reject inappropriate risk models. I apply other backtest methods in literature to be criteria for model risk adjustment. Regarding the robustness, I use Dufour (2006)'s Monte Carlo testing technique to obtain reliable inference from backtests in realistic sample size. Apart from extension of criteria sets, another contribution of my study is to provide two ex-post validation methods to evaluate series of model-risk-adjusted forecasts (RaVaR). The first validation method is out-of-sample backtest. In risk management, it is vital to backtest EVaR to decide whether the EVaR number and the models used to calculate EVaR is sound or not. Similarly, it is also of great importance to gauge the soundness of RaVaR before using it, as their usefulness are the same, e.g., for indicating risk of a given portfolio and making management decision. Another validation technique is risk ratio analysis which is an alternative measurement of model risk by examining inconsistency between range of risk models.

According to the result, five standard risk models, especially sophisticated ones (GARCH-N and GARCH-t) still cannot capture the structure of SET index as backtest criteria reject the models many times especially in volatile periods. Therefore, amount of correction is required. In some ways, the quantile correction framework does make sense in that estimated magnitude of model risk is harmonious with misspecification risk. But as an enhancement in out-of-sample backtest and risk ratio is quite small, model risk still exists. Nevertheless, static VaR methods especially for HS, are highly recommended to apply the model risk correction approach as room for improvement is strongly high. Also, even though choosing criteria sets does not matter for dynamic models, higher-statistical backtests do make interesting results when static methods are used for both overlapping and non-overlapping backtest periods.

Indeed, statistical power of backtests to reject inappropriate models claimed by authors were done in simulation experiment, where series of dynamic returns (e.g., GARCH(1,1)-t) were simulated and simple VaR models (e.g., HS and Normal) were used to investigate. In the same analogy, this study use series of market index returns which its distribution is unknown but have been suggested to have some complicated forms. That is the reason why level of statistical power does matter when using static models.

## REFERENCES

Alexander, C. (2008). Market risk analysis. Volume IV: Value-at-risk models, Wiley.

Alexander, C. and J. Sarabia (2012). "Quantile uncertainty and value-at-risk model risk." Journal of Risk Analysis **32** (2): 1293–1308.

Bao, Y. and A. Ullah (2004). "Bias of value-at-risk." Finance Research Letters **1**(4): 241-249.

BCBS (1996). Overview of the amendment to the capital accord to incorporate market risks, Basel Committee on Banking Supervision**:** 11.

BCBS (2013). Fundamental Review of the Trading Book: A Revised Market Risk Framework, Basel Committee on Banking Supervision**:** 127.

Berkowitz, J. (2001). "Testing density forecasts with applications to risk management." Journal of Business and Economics Statistics **19**(4): 465-474.

Berkowitz, J., et al. (2011). "Evaluating value-at-risk models with desk-level data." Journal of Management Science **57**(12): 2213-2227.

Berkowitz, J. and J. O'Brien (2002). "How accurate are value-at-risk models at commercial banks?" Journal of Finance **57**(3): 1093-1111.

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." Journal of Econometrics **31**(3): 307-327.

Boucher, C., et al. (2014). "Risk Model-at-Risk." Journal of Banking and Finance **44**: 72-92.

Boucher, C. and B. Maillet (2013). "Learning by Failing: A Simple Buffer for VaR." Financial Markets, Institutions & Instruments Journal **22**: 113-127.

Campbell, S. (2005). A review of backtesting and backtesting procedures. Finance and Economics Discussion Series., Federal Reserve Board, Washington, D.C.

Candelon, B., et al. (2011). "Backtesting value-at-risk: a GMM duration-based test." Journal of Financial Econometrics **9**(2): 314-343.

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

Christoffersen, P. and S. Goncalves (2005). "Estimation Risk in Financial Risk Management." Journal of Risk **7**(3): 1-28.

Christoffersen, P. and D. Pelletier (2004). "Backtesting value-at-risk: a duration-based approach." Journal of Financial Econometrics **2**: 84-108.

Colletaz, G., et al. (2013). "The risk map: a new tool for validating risk models. Journal of Banking and Finance." Journal of Banking and Finance **37** (10): 3843-3854.

Danielsson, J., et al. (2014). Model Risk of Risk Models. SRC Working Paper, London School of Economics.

Dufour, J. (2006). "Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics." Journal of Econometrics **133**(2): 443-477.

Embrechts, P., et al. (2014). "An academic response to Basel 3.5." Journal of Risks **2**: 25-48.

Engle, R. and S. Manganelli (2004). "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles." Journal of Business & Economic Statistics **22**: 367-381.

Escanciano, J. and P. Pei (2012). "Pitfalls in backtesting historical simulation models. Journal of Banking and Finance " Journal of Banking and Finance **36**(8): 2233-2244.

Gaglianone, W., et al. (2011). "Evaluating value-at- risk models via quantile regression." Journal of Business & Economic Statistics **29**(1): 150-160.

Haas, M. (2001). New methods in backtesting. Financial Engineering Research Center Caesar, Bonn.

Hull, J. (2012). Risk management and financial institutions, Wiley.

Jorion, P. (2001). Value at risk: the new benchmark for managing financial risk. McGraw-Hill, New York.

Kerkhof, J., et al. (2010). "Model risk and capital reserves." Journal of Banking and Finance **34**(1): 267-279.

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

Lopez, J. (1998). Methods for Evaluating Value-at-Risk Estimates. Economic Policy Review. Federal Reserve Bank of San Francisco**:** 120-124.

Piontek, K. (2013). Value-at-Risk Backtesting Procedures Based on Loss Functions: Simulation Analysis of the Power of Tests. Data Analysis and Knowledge Organization**:** 273-281.

Røynstrand, T., et al. (2012). Evaluating power of value-at-risk backtests, Norwegian University of Science and Technology. Master thesis.

Ziggel, D., et al. (2013). "New Set of Improved Value-at-Risk Backtests." Journal of Banking and Finance **48**: 29-41.

# APPENDIX

# APPENDIX A: TABLES

**Table A1. Maximum of Positive and Negative $q_t^*$ Relative to 95% VaR using Individual Criterion**

| Individual Criterion | GARCH | | GARCH.t | | EWMA | | HS | | Normal | |
|---|---|---|---|---|---|---|---|---|---|---|
| | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) |
| **LRuc** | -2.20% | 2.40% | -0.30% | 3.00% | -1.00% | 2.10% | -2.90% | 2.60% | -2.30% | 2.90% |
| **LRind** | -2.70% | 3.10% | -1.10% | 3.00% | -4.20% | 2.90% | -7.20% | 3.30% | -6.30% | 3.60% |
| **LRcc** | -2.80% | 3.00% | -0.50% | 3.00% | -2.20% | 2.10% | -2.80% | 3.00% | -2.30% | 3.20% |
| **LRmuc** | -2.20% | 2.20% | -0.60% | 2.90% | -1.60% | 2.00% | -2.80% | 2.40% | -2.30% | 2.80% |
| **GMMuc** | -2.40% | 2.40% | 0.00% | 3.00% | -1.00% | 2.10% | -2.90% | 2.70% | -2.40% | 3.00% |
| **GMMind** | -6.10% | 2.70% | -1.80% | 2.30% | -4.70% | 2.10% | -6.60% | 3.20% | -6.20% | 3.10% |
| **GMMcc** | -2.70% | 2.30% | -1.40% | 3.00% | -2.40% | 2.20% | -3.20% | 3.20% | -2.60% | 3.30% |
| **DURind** | -2.00% | 2.00% | 0.00% | 2.10% | 0.00% | 0.60% | -2.70% | 3.00% | -6.20% | 2.90% |
| **DURcc** | -1.80% | 2.30% | 0.00% | 3.00% | -0.80% | 2.00% | -3.60% | 3.30% | -3.10% | 3.30% |
| **DQuc** | -2.40% | 2.40% | 0.00% | 2.90% | -1.00% | 1.70% | -2.80% | 2.50% | -2.20% | 2.80% |
| **DQind** | -4.40% | 3.60% | -3.90% | 2.70% | -5.50% | 3.30% | -7.10% | 3.30% | -6.40% | 3.40% |
| **DQcc** | -5.70% | 3.00% | -2.30% | 3.00% | -4.50% | 2.40% | -7.20% | 3.10% | -6.40% | 3.30% |
| **MCSuc** | -2.40% | 2.40% | -0.30% | 3.00% | -1.10% | 2.10% | -2.90% | 2.70% | -2.40% | 3.00% |
| **MCSind** | -6.80% | 2.20% | -1.60% | 3.10% | -4.90% | 2.00% | -6.90% | 3.10% | -6.30% | 2.80% |
| **MCScc** | -2.00% | 2.00% | -0.30% | 2.60% | -0.80% | 0.70% | -3.20% | 3.20% | -2.70% | 2.50% |

Datasource: Bloomberg. SET index daily data from the 2$^{nd}$ Jan, 1990 to the 30th December, 2014. EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. $q_t^*$ are calculared using 250 backtest period from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The unit is present in percentage of EVaR.NA means non-existent value. For example, if q*(-) does not exist, it means no negative adjustment for that criterion (row) when using particular model (column)

**Table A2. Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set (sub-sample 1: 1997's crisis)**

| Criteria Set | GARCH-N | | GARCH-t | | EWMA | | HS | | Normal | |
|---|---|---|---|---|---|---|---|---|---|---|
| | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) |
| **a1** | 1.80% | NA | 0.50% | 0.74% | 0.20% | NA | NA | 2.11% | NA | 1.85% |
| **a2** | 1.96% | NA | 0.57% | 0.87% | 0.20% | NA | NA | 1.78% | NA | 2.29% |
| **a3** | NA | NA | 0.81% | 0.82% | NA | 1.38% | NA | 1.33% | NA | 1.39% |
| **a4** | -0.79% | NA | NA | 1.00% | -0.45% | NA | -1.98% | 1.36% | -1.77% | 1.78% |
| **a5** | -0.90% | NA | -0.41% | 0.97% | -0.49% | 1.40% | -1.87% | 1.39% | -1.67% | 1.88% |
| **b1** | -1.03% | NA | NA | 0.73% | -0.67% | NA | -2.27% | 2.10% | -1.91% | 1.90% |
| **b2** | -0.77% | NA | NA | 0.87% | -0.42% | NA | -1.83% | 1.93% | -1.64% | 2.41% |
| **b3** | -1.16% | NA | NA | 0.82% | -0.75% | 1.60% | -1.87% | 1.25% | -1.68% | 1.95% |
| **b4** | -0.79% | NA | NA | 1.07% | -0.45% | NA | -1.97% | 1.36% | -1.76% | 1.78% |
| **b5** | -0.88% | NA | -0.41% | 0.97% | -0.49% | 1.40% | -1.88% | 1.39% | -1.67% | 1.88% |
| **c1** | -0.65% | NA | NA | 0.62% | -0.36% | NA | -1.86% | 0.90% | -1.60% | 1.32% |
| **c2** | -0.69% | NA | NA | 0.80% | -0.38% | NA | -1.78% | 1.71% | -1.60% | 2.32% |
| **c3** | -1.13% | NA | NA | 0.62% | -0.63% | 0.75% | -2.08% | 1.02% | -1.83% | 1.90% |
| **c4** | -0.65% | NA | NA | 1.05% | -0.35% | NA | -1.86% | 1.04% | -1.60% | 1.87% |
| **c5** | -1.13% | NA | NA | 0.81% | -0.64% | 1.00% | -1.92% | 1.17% | -1.69% | 1.67% |

EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. RaVaR series from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The number of out-of-sample backtest is 500 times (from the 27th June, 1996 to the 13th July, 1998) for EVaR and RaVaR to observe Acceptance Frequency Ratio ($\overline{\text{AFR}}$). The underlined numbers are the most outperforming criteria sets. The unit is present in percentage of EVaR. NA means non-existent value. For example, if q*(-) does not exist, it means that there is no negative adjustment for that criteria set (row) when using particular model (column).

**Table A3. Mean of Positive and Negative $q_t^*$ Relative to 95% VaR using Criteria Set (sub-sample 2: 2008's crisis)**

| Criteria Set | GARCH-N | | GARCH-t | | EWMA | | HS | | Normal | |
|---|---|---|---|---|---|---|---|---|---|---|
| | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) | q*(-) | q*(+) |
| **a1** | -1.97% | 1.80% | -0.93% | 0.50% | -1.81% | 0.20% | -2.77% | NA | -2.33% | NA |
| **a2** | -1.87% | 1.96% | -0.85% | 0.57% | -1.93% | 0.20% | -3.07% | NA | -2.44% | NA |
| **a3** | -1.83% | NA | -1.82% | 0.81% | -1.99% | NA | -3.34% | NA | -2.64% | NA |
| **a4** | -1.67% | 0.90% | -0.38% | 0.58% | -1.24% | 0.20% | -2.71% | NA | -2.34% | NA |
| **a5** | -1.91% | NA | -0.56% | 0.58% | -1.72% | NA | -3.68% | NA | -2.93% | 0.31% |
| **b1** | -1.97% | 1.80% | -0.93% | 0.52% | -1.81% | 0.20% | -2.75% | NA | -2.33% | NA |
| **b2** | -1.84% | 2.00% | -0.84% | 0.57% | -1.90% | 0.20% | -3.14% | 1.40% | -2.50% | 1.83% |
| **b3** | -1.89% | NA | -1.70% | 0.82% | -1.99% | NA | -3.34% | NA | -2.64% | NA |
| **b4** | -1.67% | 0.90% | -0.35% | 0.58% | -1.24% | 0.20% | -2.71% | NA | -2.34% | NA |
| **b5** | -1.97% | NA | -0.55% | 0.58% | -1.71% | NA | -3.63% | NA | -2.82% | 0.32% |
| **c1** | -1.67% | NA | -0.34% | NA | -1.12% | 0.20% | -2.51% | NA | -2.15% | NA |
| **c2** | -1.72% | 1.53% | -0.47% | 0.57% | -1.35% | 0.20% | -2.21% | 2.22% | -2.11% | 2.55% |
| **c3** | -1.67% | NA | -0.72% | 0.65% | -1.47% | 1.30% | -2.87% | NA | -2.34% | NA |
| **c4** | -1.66% | NA | -0.38% | 0.59% | -1.12% | 0.20% | -2.64% | NA | -2.14% | NA |
| **c5** | -1.95% | NA | -0.40% | 0.59% | -1.62% | 0.80% | -1.39% | NA | -1.98% | NA |

EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. RaVaR series from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The number of out-of-sample backtest is 500 times (from the 11th September, 2007 to 25th September, 2009) for EVaR and RaVaR to observe Acceptance Frequency Ratio ($\overline{\text{AFR}}$). The unit is present in percentage of EVaR. The unit is present in percentage of EVaR. NA means non-existent value. For example, if q*(-) does not exist, it means that there is no negative adjustment for that criteria set (row) when using particular model (column).

**Table A4. Ex post validation of RaVaR/EVaR series for 95% VaR on SET index (full sample)**

| EVaR / RaVaR (criteria set) | Acceptance Frequency Ratio ($\overline{\text{AFR}}$) | | | | | Risk Ratio |
|---|---|---|---|---|---|---|
| | GARCH | GARCH-t | EWMA | HS** | Normal | |
| **EVaR** | 0.7738 | 0.7402 | 0.8697 | 0.5352 | 0.5552 | 1.6311 |
| **a1 (MCS-based set)** | 0.7727 | 0.7462 | 0.8733 | 0.5549 | 0.5570 | 1.6119 |
| **a2 (GMM-based set)** | 0.7743 | 0.7508 | 0.8737 | 0.5548 | 0.5559 | 1.6105 |
| **a3 (DQ-based set)** | 0.7629 | 0.7465 | 0.8667 | 0.5487 | 0.5541 | 1.6133 |
| **a4 (DUR-based set)** | 0.7727 | 0.7455 | 0.8702 | 0.5488 | 0.5539 | 1.6136 |
| **a5 (LR-based set)** | 0.7745 | 0.7511 | 0.8687 | 0.5483 | 0.5540 | 1.6118 |
| **b1 (MCS-based set)** | 0.7736 | 0.7465 | 0.8739 | 0.5545 | 0.5572 | 1.6119 |
| **b2 (GMM-based set)** | 0.7745 | 0.7503 | 0.8734 | 0.5541 | 0.5565 | 1.6089 |
| **b3 (DQ-based set)** | 0.7676 | 0.7511 | 0.8673 | 0.5491 | 0.5544 | 1.6134 |
| **b4 (DUR-based set)** | 0.7730 | 0.7446 | 0.8695 | 0.5484 | 0.5542 | 1.6136 |
| **b5 (LR-based set)** | 0.7755 | 0.7508 | 0.8686 | 0.5489 | 0.5542 | 1.6118 |
| **c1 (MCS-based set)** | 0.7738 | 0.7410 | 0.8697 | 0.5464 | 0.5521 | 1.6195 |
| **c2 (GMM-based set)** | 0.7727 | 0.7505 | 0.8703 | 0.5495 | 0.5516 | 1.6137 |
| **c3 (DQ-based set)** | 0.7749 | 0.7510 | 0.8673 | 0.5473 | 0.5580 | 1.6167 |
| **c4 (DUR-based set)** | 0.7743 | 0.7405 | 0.8699 | 0.5481 | 0.5521 | 1.6156 |
| **c5 (LR-based set)** | 0.7744 | 0.7505 | 0.8697 | 0.5489 | 0.5578 | 1.6147 |

Datasource: Bloomberg. EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. RaVaR series from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The underlined numbers are ratio of RaVaR that are outperform EVaR (Higher $\overline{\text{AFR}}$ / lower Risk ratio) with 5% significance level.

**Table A5**. **Ex post validation of RaVaR/EVaR series for 95% VaR on SET index (sub-sample 1: the 1997's crisis)**

| **EVaR/ RaVaR (Criteria sets)** | **Acceptance Frequency Ratio ($\overline{\text{AFR}}$)** | | | | | **Risk Ratio** |
|---|---|---|---|---|---|---|
| | **GARCH** | **GARCH-t** | **EWMA** | **HS** | **Normal** | |
| **EVaR (original VaR)** | 0.6517 | 0.8449 | 0.8654 | 0.2954 | 0.3883 | 1.6956 |
| **a1 (MCS-based set)** | 0.6583 | 0.8449 | 0.8651 | 0.3140 | 0.4157 | 1.6659 |
| **a2 (GMM-based set)** | 1.5827 | 1.5745 | 1.5748 | 1.5802 | 1.5767 | 1.6677 |
| **a3 (DQ-based set)** | 0.6580 | 0.8463 | 0.8657 | 0.3137 | 0.4154 | 1.6639 |
| **a4 (DUR-based set)** | 0.6597 | 0.8449 | 0.8663 | 0.3134 | 0.4151 | 1.6694 |
| **a5 (LR-based set)** | 0.6560 | 0.8466 | 0.8660 | 0.3137 | 0.4146 | 1.6690 |
| **b1 (MCS-based set)** | 0.6606 | 0.8466 | 0.8654 | 0.3134 | 0.4146 | 1.6660 |
| **b2 (GMM-based set)** | 0.6583 | 0.8426 | 0.8649 | 0.3134 | 0.4160 | 1.6676 |
| **b3 (DQ-based set)** | 0.6586 | 0.8457 | 0.8654 | 0.3140 | 0.4166 | 1.6639 |
| **b4 (DUR-based set)** | 0.6614 | 0.8437 | 0.8649 | 0.3134 | 0.4151 | 1.6692 |
| **b5 (LR-based set)** | 0.6583 | 0.8449 | 0.8651 | 0.3140 | 0.4157 | 1.6688 |
| **c1 (MCS-based set)** | 0.6577 | 0.8483 | 0.8660 | 0.3131 | 0.4166 | 1.6723 |
| **c2 (GMM-based set)** | 0.6566 | 0.8431 | 0.8657 | 0.3134 | 0.3989 | 1.6730 |
| **c3 (DQ-based set)** | 0.6574 | 0.8471 | 0.8657 | 0.3137 | 0.4149 | 1.6666 |
| **c4 (DUR-based set)** | 0.6580 | 0.8429 | 0.8657 | 0.3137 | 0.4151 | 1.6726 |
| **c5 (LR-based set)** | 0.6589 | 0.8429 | 0.8657 | 0.3134 | 0.4166 | 1.6706 |

EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. RaVaR series from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The number of out-of-sample backtest is 500 times from the 27th June, 1996 to the 13th July, 1998 for EVaR and RaVaR to observe Acceptance Frequency Ratio ($\overline{\text{AFR}}$). The unit is present in percentage of EVaR

**Table A6. Ex post validation of RaVaR/EVaR series for 95% VaR on SET index (sub-sample 2: 2008's crisis**

| EVaR/ RaVaR (Criteria Set) | Acceptance Frequency Ratio ($\overline{\text{AFR}}$) | | | | | Risk Ratio |
|---|---|---|---|---|---|---|
| | GARCH | GARCH-t | EWMA | HS | Normal | |
| **EVaR (original VaR)** | 0.6057 | 0.7903 | 0.8323 | 0.5091 | 0.5497 | 1.5952 |
| **a1 (MCS-based set)** | 0.6009 | 0.7923 | 0.8317 | 0.5226 | 0.4977 | 1.5827 |
| **a2 (GMM-based set)** | 0.6051 | 0.8071 | 0.8311 | 0.5231 | 0.4971 | 1.5745 |
| **a3 (DQ-based set)** | 0.6051 | 0.7929 | 0.8017 | 0.5226 | 0.4969 | 1.5748 |
| **a4 (DUR-based set)** | 0.6006 | 0.8063 | 0.8337 | 0.5229 | 0.4977 | 1.5802 |
| **a5 (LR-based set)** | 0.6054 | 0.8057 | 0.8329 | 0.5226 | 0.4969 | 1.5767 |
| **b1 (MCS-based set)** | 0.6014 | 0.7937 | 0.8320 | 0.5234 | 0.4969 | 1.5821 |
| **b2 (GMM-based set)** | 0.6054 | 0.8063 | 0.8317 | 0.5226 | 0.4974 | 1.5710 |
| **b3 (DQ-based set)** | 0.6051 | 0.8060 | 0.8020 | 0.5237 | 0.4980 | 1.5748 |
| **b4 (DUR-based set)** | 0.6011 | 0.8060 | 0.8320 | 0.5229 | 0.4971 | 1.5801 |
| **b5 (LR-based set)** | 0.6051 | 0.8049 | 0.8320 | 0.5229 | 0.4977 | 1.5762 |
| **c1 (MCS-based set)** | 0.6057 | 0.7940 | 0.8346 | 0.5234 | 0.4974 | 1.5805 |
| **c2 (GMM-based set)** | 0.6006 | 0.8043 | 0.8331 | 0.5231 | 0.4920 | 1.5844 |
| **c3 (DQ-based set)** | 0.6051 | 0.8034 | 0.8029 | 0.5097 | 0.5497 | 1.5912 |
| **c4 (DUR-based set)** | 0.6051 | 0.7923 | 0.8309 | 0.5229 | 0.4974 | 1.5823 |
| **c5 (LR-based set)** | 0.6057 | 0.8037 | 0.8320 | 0.5234 | 0.5500 | 1.5890 |

EVaR series from the 28th March, 1994 to the 30th December, 2014 (5086 daily returns) are calculated by rolling window of 1040 daily returns. RaVaR series from 3rd April 1995 to the 30th December, 2014 (4,836 observations). The number of out-of-sample backtest is 500 times (from the 11th September, 2007 to 25th September, 2009) for EVaR and RaVaR to observe Acceptance Frequency Ratio ($\overline{\text{AFR}}$).

## APPENDIX B: FIGURES

Figure B1. $q_t^*$ for Criteria Set b1 Relative to 95% EVaR

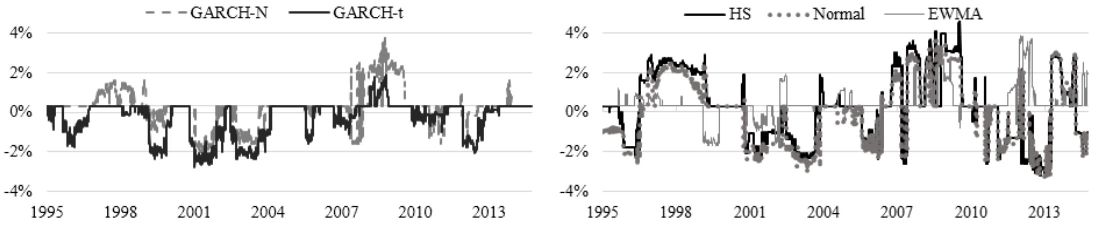


Figure B2. $q_t^*$ for Criteria Set b2 Relative to 95% EVaR

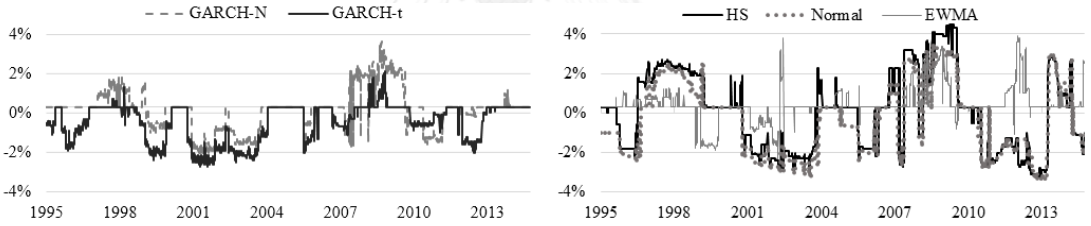


Figure B3. $q_t^*$ for Criteria Set b3 Relative to 95% EVaR

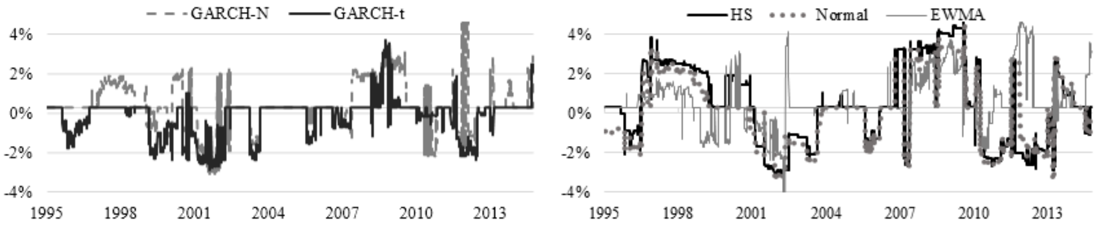


Figure B4. $q_t^*$ for Criteria Set b4 Relative to 95% EVaR

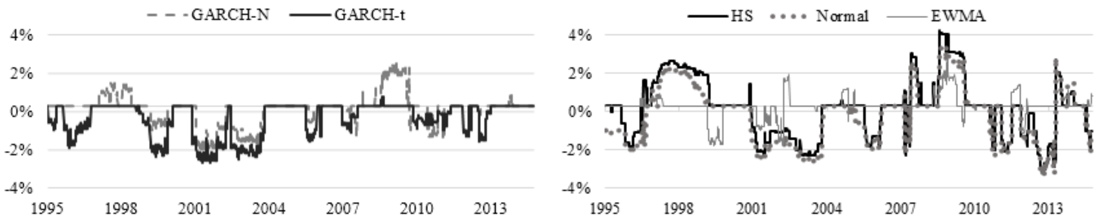


Figure B5. $q_t^*$ for Criteria Set b5 Relative to 95% EVaR

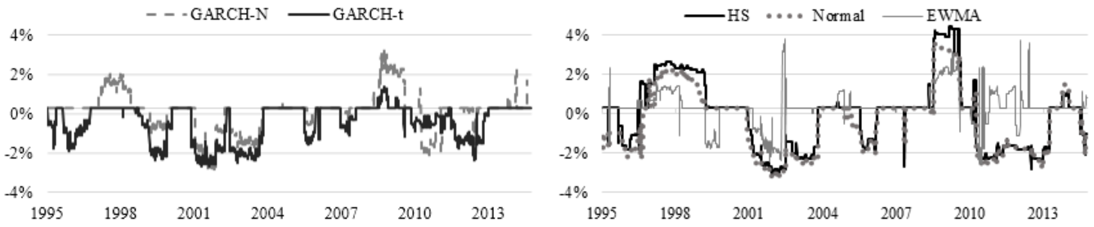
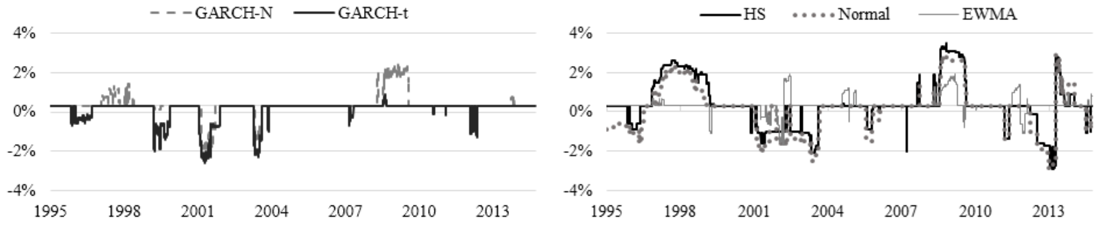
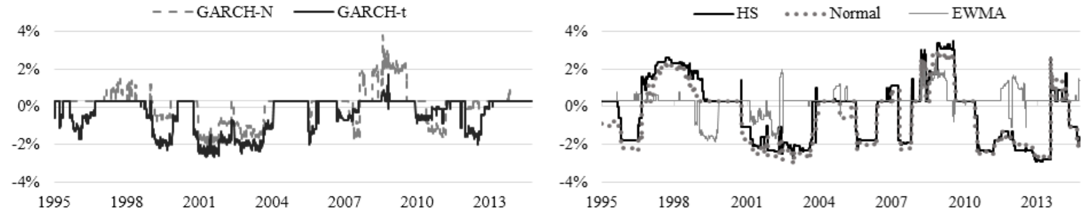


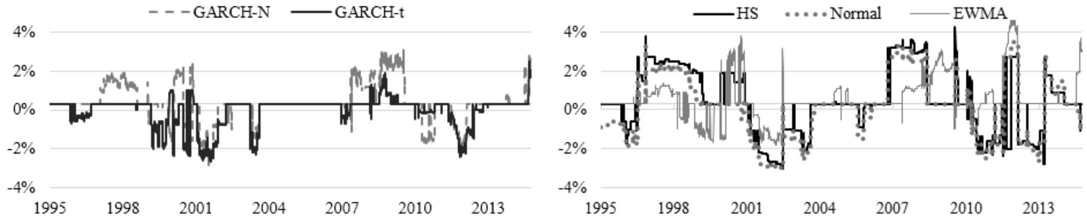
Figure B6. $q_t^*$ for Criteria Set c1 Relative to 95% EVaR



### Figure B7. $q_t^*$ for Criteria Set c2 Relative to 95% EVaR



### Figure B8. $q_t^*$ for Criteria Set c3 Relative to 95% EVaR



### Figure B9. $q_t^*$ for Criteria Set c4 Relative to 95% EVaR

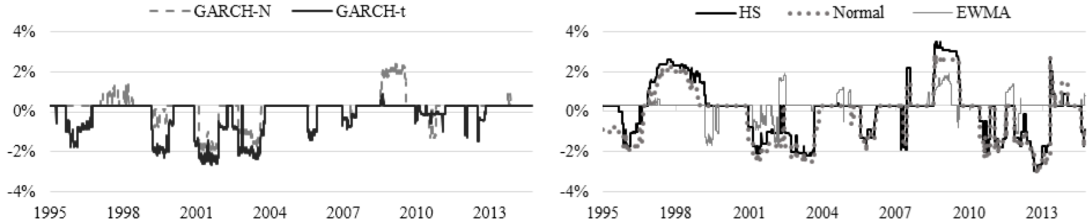
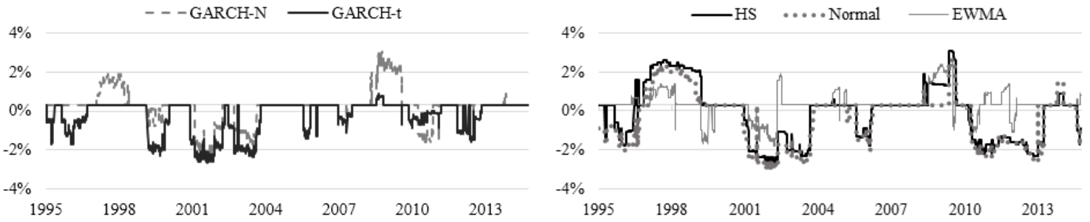


**Figure B10. $q_t^*$ for Criteria Set c5 Relative to 95% EVaR**



# VITA

Name: Siridej Putsorn

Birth date: 27 October 1992

Education:

| | |
|---|---|
| Master of Science in Finance, Chulalonkorn University | 2014-Present |
| Bachelor of Economics, Chulalongkorn University | 2010-2014 |