EMPIRICAL ASSESSMENT AND MODEL SELECTION IN OPTION PRICING

by
Berk Orbay
B.S., Industrial Engineering, Middle East Technical University, 2008
M.S., Operations Research, Middle East Technical University, 2011

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Industrial Engineering
Boğaziçi University
2016

EMPIRICAL ASSESSMENT AND MODEL SELECTION IN OPTION PRICING

APPROVED BY:

Prof. Refik Güllü (Thesis Supervisor)	....
Assist. Prof. İsmail Başoğlu	....
Assoc. Prof. Wolfgang Hörmann	....
Assist. Prof. Cenk Karahan	....
Assist. Prof. Emrah Şener	....

DATE OF APPROVAL: 18.05.2016

ACKNOWLEDGEMENTS

I don't think there are enough words of appreciation for professors who encourage, guide and support their students to the full extent, but I will try to do my best. I owe much to my supervisor Prof. Refik Güllü, not just for the thesis but also for providing me all the opportunities and resources for my research. Many thanks to my "co-supervisor" Assoc. Prof. Wolfgang Hörmann for his valuable insights and pointing out to the right and critical points that always improved the quality of the research. I would like to thank Asst. Prof. Cenk Karahan for being on my thesis committee for all those years and his contributions. I would also like to thank Asst. Prof. İsmail Başoğlu and Asst. Prof. Emrah Şener for their feedback in my thesis defense.

Family is the greatest supporter for endeavors like this and mine did a fantastic job; way more than I deserve. My deepest gratitude to all my family, especially to my father Prof. Atilla Orbay and my mother Prof. Yeşim Kamile Aktuğlu for their endless love and unconditional support. I wouldn't be where I am right now without them.

These five years would be unbearable without the help of great friends. Many thanks to Jehanzeb, Müge, Orkun, Debbi, Cansu, Atilla, Nezih, Dilber, Erkmen, Bora, Ece, Kadir, Alperen, Oğuzcan, Berkem, Doğan, Eda, Senem, Utku, Emre, Özge, Ferda Şahmalı and all other friends. Special thanks to Nihan for constantly encouraging (nagging) me about finishing my thesis. Special thanks to fellow PhDs Mine, Pınar, Nisa and Şirin. Special thanks to Asst. Prof. Mustafa Baydoğan, especially for the coffee supply. Music is one of the primary elements that kept me going. Special thanks to Queen, Deep Purple, David Bowie, ACDC, The Who and many others.

I would also like to thank Hakan Yamaçoğuz. Brief but helpful discussions with him changed the course of my research. Finally, I gracefully appreciate the support of my university. Part of this thesis is supported by Bogazici University Scientific Research Projects (no: 8101).

ABSTRACT

EMPIRICAL ASSESSMENT AND MODEL SELECTION IN OPTION PRICING

The majority of empirical option pricing studies consider the distance from the market option prices as the performance metric. Though, this kind of assessment is limited to the objectives of proper hedging of options and fair pricing of OTC contracts. Options can also be used in market efficiency tests. Efficiency tests require positions in options and other assets (e.g. underlying security) with the objective to yield risk-adjusted profits. If a model can generate excess profit consistently, then it might claim that markets are not efficient.

This study consists of four main parts. Different empirical option pricing studies are investigated in terms of assessment methodology. It is shown that error aggregation with data mining algorithms is a more robust way to assess model performance of a model. New model error metrics based on efficiency tests are introduced. Finally, all the new findings are used in the introduction of a model selection framework. Model selection uses the price and hedge estimates of individual pricing models (e.g. Black-Scholes) to come up with better price and hedge structure according to the performance metric. Main motivation is to avoid overfitting of complex models, by switching between simpler but effective models according to performance shifts.

Experiments with SPX and NDX contracts between 2009 and 2013 indicate that a model selection method with only a data mining algorithm to group errors and a simple selection rule, yields consistently better results than the individual models it uses.

ÖZET

OPSİYON FİYATLAMADA AMPİRİK DEĞERLENDİRME VE MODEL SEÇİMİ

Ampirik opsiyon fiyatlama çalışmalarının pek çoğu model tahminlerinin piyasa opsiyon fiyatlarından uzaklığı performans ölçütü olarak kabul ederler. Yalnız, bu tür bir değerlendirmenin faydası doğru hedging değerini bulma ve tezgah üstü kontratları fiyatlama ile sınırlıdır. Opsiyonlar piyasa etkinlik testlerinde de kullanılmaktadır. Etkinik testleri opsiyonlarda ve diğer varlıklarda (ör. dayanak varlık) pozisyonlar alarak risk ayarlı karlılığı hedeflemektedir. Eğer bir model yeterince yüksek seviyede ve tutarlı olarak risk ayarlı karlılık elde edebiliyorsa piyasaların etkin olmadığı iddia edilebilir.

Bu çalışma dört ana kısımdan oluşmaktadır. Farklı ampirik opsiyon fiyatlama çalışmalarının değerlendirme metodolojileri incelenmiştir. Veri madenciliği algoritmaları ile hata gruplaması yapmanın bir modelin performansını değerlendirmenin daha gürbüz bir yolu olduğu gösterilmiştir. Etkinlik testlerinden yola çıkarak oluşturulan yeni performans ölçütleri önerilmiştir. Son olarak, bütün bu bulguların kullanıldığı model seçim yöntemi oluşturulmuştur. Model seçimi, opsiyon fiyatlama modellerinin (ör. Black-Scholes) fiyat ve hedge tahminlerini kullanarak farklı amaç fonksiyonları için daha iyi fiyatlama ve hedge oluşturmayı hedeflemektedir. Model seçimindeki ana motivasyon, daha basit ama etkili modeller arasında performans değişimlerine bağlı olarak geçiş yaparak, çok parametreli karmaşık modellerin yarattığı aşırı uygunluk probleminden kaçınmaktır.

2009-2013 yıllarının SPX ve NDX kontratlarıyla yapılan ampirik testler, basit bir seçim kuralı ve hata gruplayan bir veri madenciliği algoritması ile bile, bir model seçim yönteminin kullandığı modellerden daha iyi sonuçlar verdiğini ortaya koymaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\mathcal{F}_t$	Filtration
K	Strike price of the option
$\mathbb{P}$	Physical probability measure
$\hat{P}$	Price estimate of a model on an option contract
P	Market price of a contract
q	Continuous dividend rate
$\mathbb{Q}$	Risk-neutral probability measure
r	Risk-free rate
S_0	Initial spot price of the asset
T	Time to maturity
Δ	Delta value of an option contract
μ	Drift
σ	Volatility

LIST OF ACRONYMS/ABBREVIATIONS

APE	Absolute Pricing Error
ARPE	Absolute Relative Pricing Error
ATM	At-the-Money
BS	Black-Scholes option pricing model
CAC 40	Cotation Assistée en Continu 40 Index (French Market)
DITM	Deep-in-the-Money
DHE	Dynamic Hedging Error
DOTM	Deep-out-of-the-Money
EC	European Call Option
EMH	Efficient Market Hypothesis
EP	European Put Option
EWMA	Exponentially Weighted Moving Average
GARCH	Generalized Autoregression Conditional Heterosketasticity
HPL	Profit and Loss with Static Hedging
HN	Heston-Nandi GARCH option pricing model
ITM	In-the-Money
MAPE	Mean Absolute Percentage Error
MCMM	Mean Correcting Martingale Measure
MS	Model Selection
MSE	Mean Squared Error
NDX	NASDAQ 100 Index Option
NPE	Naked Positioning Error
NPL	Naked Postioning Profit and Loss
OTC	Over the Counter
OTM	Out-of-the-Money
P&L	Profit and Loss
RMSE	Root Mean Squared Error
RPE	Relative Pricing Error

SHE	Static Hedging Error
SPX	Standard and Poor's 500 Index Option
YoY	Year-over-Year

1. INTRODUCTION

Financial derivative products are an integral part of the market. Option contracts, the right but not the obligation to buy or sell an asset in a predetermined payout plan and schedule, have been distinguished instruments for speculation, hedging and arbitrage.[1] The challenge of options is to determine the fair price of such contracts. An extensive and still expanding literature provides an arsenal of models and methodologies to put price estimates on any type of option ever issued.

The main problem with the option pricing models, concerning empirical experiments, is about performance assessment. Model performance changes according to the assets, time periods and performance metrics they are assessed with. For instance, would model A and B perform relatively the same (e.g. A better than B) if we changed the asset from S&P 500 to CAC 40, changed time period from 2006 to 2016, or measure with root mean squared error (RMSE) instead of absolute relative pricing error (ARPE)? If so, how do we choose the best model that is fit to our needs?

In this study, a model selection framework is proposed to pick the 'right' option pricing model from a set of models for any contract at any time with the assistance of data mining algorithms to guide the model selection process for any given performance metric as the main contribution. Thus, in overall, the out-of-sample performance of the the model selection becomes more robust against time, asset and performance metrics than any individual model in the model set.

Model selection framework is the outcome of a series of research in the empirical option pricing and data mining literature. This study consists of four main chapters, including the model selection chapter, and each of them is an individual research on a different topic that is essential to model selection framework.

[1] Even though there are different types of options, this study is interested only in European type options.

The rest of the chapters are organized as follows. Chapter 2 introduces the pricing models used in this study. In chapter 3, a number of empirical pricing studies are examined and compared with each other in terms of their pricing and assessment methodology. These studies, albeit not large in number, are a representative group of different empirical option pricing methodologies. Even though there are some common points in the process; their methods vary, even sometimes contradict at almost every stage from data filtering to the presentation of results.

One of the major weaknesses of empirical option pricing studies is how little information their results tables present. Only very few studies check for the stability of their model performance over different time periods. For instance, the error representation table as a result of an experiment with data from 2010 can present a largely different result than the error representation table of the same experiment with data from 2011. The performance difference between different moneyness maturity regions could also invoke different comments had the experiment been done in a different time period. This kind of instability weakens the findings of experiments since their results are not the indicator of future performance or relative performance.

In addition, arbitrary grouping of errors to comment about model performance in local moneyness maturity regions might cause hindrance of true performance regions of the models. Not all models adhere to the same rigid, unchanging set of boundaries for all the time. In Chapter 4, data mining models are tested on different models, assets and time periods. The results show that algorithms perform better than arbitrary grouping methods in terms of detecting true performance regions and maintaining prediction stability.

Pricing errors (e.g. ARPE) are not, and should not be, the only performance metrics that can be used in option pricing. Pricing errors measure just the distance between the model estimate and the market price but provide no insight about the outcome (payoff) of the contract. In addition, because their optimality is the market, they cannot suggest any action regarding the market. They can be used to calculate hedges and OTC contracts, though. Pricing error metric can be thought as in accordance with Efficient Market Hypothesis. Existence of market efficiency means that no strategy or model can make risk adjusted excess profits in the market. Efficient Market Hypothesis was first coined by Samuelson (1965) and Fama (1970).

The natural way of testing market efficiency is to provide a model that “beats” the market by taking positions and realizing the profit and loss. Efficiency tests were quite popular and EMH is still a heavily debated topic in the literature. Even Black-Scholes model is first used in a market efficiency test by Black and Scholes (1972). Though, joint hypothesis problem (or the bad model problem) dictates it is difficult, if not impossible, to “disprove EMH”.

Chapter 5 discusses differences between pricing errors and efficiency tests. In addition, new model error metrics based on efficiency tests and bad model problem are introduced. These metrics measure the distance from the equilibrium as the model’s inability to cover for risks and assess the model performance with its outcome.

Chapter 6 introduces the specifics of the model selection framework. It starts with the assumption that no option pricing model is constantly dominant to other models for even a single contract parameter combination (e.g. OTM, short-term, Call options). In addition, as model complexity increases so does the risk of overfitting after a degree of complexity. It means, option pricing models can be improved only so far. On the other hand, each model in a carefully designed model set can provide the best price (and hedge) estimate for at least one contract. So, a set of comparably simpler models can be more effective than a single large and complex model. Though, the problem of picking the best option model for each contract remains a challenge.

A model selection framework is introduced to pick price and delta estimates from individual models (e.g. Black-Scholes, Heston-Nandi GARCH) with the motivation that model selection method would yield sustainably better results than all individual models in the model set. The chapter benefits from the outcomes of the previous chapters to determine the design of numerical experiments, choice of data mining algorithm and performance metrics. Model selection framework is empirically tested with different objectives, different assets and in different time periods.

Chapter 7 summarizes the findings of the model selection framework and the rest of the study. All extra material and fundamental information about options and pricing models can be found in the Appendix.

2. PRELIMINARIES FOR OPTION PRICING

This chapter presents the details about option pricing models used in this study. There are three main models and their different parametrizations are used in the computational experiments. Even though results of Lévy process based models are not reported, the model is extensively tested to reach important conclusions about model selection in Chapter 6.

2.1. Black Scholes Model

The simple yet powerful formula of Black-Scholes allows the calculation of a self replicating portfolio consisting of the price and delta of a given European Call or Put Option under complete market assumptions. The mathematical representation of the European Call is as follows.

$$C(S_0, K, r, T, \sigma, q) := S_0 N(d_+) - e^{-rT} K N(d_-) \tag{2.1}$$

where

$$d_- = \frac{log\left(\frac{S_0}{K}\right) + \left(r - q - \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \tag{2.2}$$

$$d_+ = d_- + \sigma\sqrt{T} \tag{2.3}$$

$$N(.) := \text{c.d.f of standard normal variate } N(0,1) \tag{2.4}$$

The setting of a European Call (EC) is as follows. S_0 is the price of the underlying at the time of the inception of the contract, K is the strike price, r is the risk free rate, σ is the volatility and T is the time to maturity. The price of the European Put option can be found by the put-call parity.

$$C(S_0, K, r, T, \sigma) + e^{-rT} = P(S_0, K, r, T, \sigma) + S_0 \tag{2.5}$$

$$P(S_0, K, r, T, \sigma) = e^{-rT} KN(-d_-) - S_0 N(-d_+) \tag{2.6}$$

Finally, delta (Δ) of the option is calculated as follows.

$$\Delta(Call) = \frac{\partial V}{\partial S} = N(d_+) \tag{2.7}$$

$$\Delta(Put) = \frac{\partial V}{\partial S} = N(d_+) - 1 \tag{2.8}$$

2.1.1. Assumptions

Coming up with such a formula requires several assumptions. First of these assumptions is the underlying asset's price process follows Geometric Brownian Motion (GBM).

$$\frac{dS}{S} = \mu dt + \sigma dW(t) \tag{2.9}$$

This yields

$$S_T = S_0 exp\left(\left(\mu - \frac{1}{2}\sigma^2\right) T + \sigma dW(t)\right) \tag{2.10}$$

where μ is the drift parameter of the underlying and $W(t)$ is a Wiener process. So the log-return of the underlying asset is normally distributed.

$$Y_T := log\left(\frac{S_T}{S_0}\right) \sim N\left(\left(\mu - \frac{1}{2}\sigma^2\right) T, \sigma^2 T\right) \tag{2.11}$$

The second assumption is the volatility term is constant. This assumption is one of the strongest assumptions of the Black-Scholes model. It is constantly criticized in the literature.

Other assumptions are short selling is allowed, perfectly liquid and continuous market, no friction (i.e. transaction costs), securities are infinitely divisible, borrowing and lending rates (i.e. riskless rate) are the same and constant and there is no arbitrage opportunity.

Some of these assumptions are quite strong and not applicable to every security. Ever since, none of these assumptions are exempt from criticism and many papers were published offering a remedy to one or more of those assumptions. Even though, at the inception of the formula in the 70s, only few of these assumptions were valid (e.g. short selling was heavily restricted and even banned for some periods), today's market conditions are closer to these assumptions.

It is widely known that BS formula has problems with pricing out-of-money (large difference between strike price and the initial stock price) options due to the low tails of the underlying's assumed price process. The underlying asset log-return distribution is known to be leptokurtic (i.e. fat-tailed and higher kurtosis). Failure to recognize this property, leads to severe losses to the option issuers as the potential loss for the issuer is in theory unbounded.

2.2. Lévy Processes

The assumption of Geometric Brownian Motion governing the asset log-return movements is often challenged by the argument of log-return distributions often exhibiting leptokurtic behavior, in other words the distributions have higher peaks and fatter tails. Lévy processes are proposed as a remedy. Therefore, option pricing with Lévy processes is quite popular in the academic literature. For some examples see Eberlein (2001), Schoutens (2003) and Tankov and Cont (2003). Schoutens' work is generally followed for definitions and explanations in this sub-section.

An infinitely divisible stochastic process $X = X_t, t \geq 0$ can be identified as a Lévy process if it starts at zero $X(0) = 0$, has independent and stationary increments such as $X_{s+t} - X_s; s, t \geq 0$ is equivalent to X_t, and sample paths of X_t are almost surely continuous from the right and have limits from the left.

Some prominent examples of Lévy processes are Poisson process, Compound Poisson process, Gamma process, Inverse Gaussian (IG) process, Generalized Inverse Gaussian process (GIG), Carr-Geman-Madan-Yor (CGMY) process, Generalized Hyperbolic process (GHYP), Hyperbolic process (HYP), Normal Inverse Gaussian process (NIG), Variance Gamma process (VG) and Meixner process. Brownian Motion is also a special case of Lévy processes.

The Lévy processes of interest here are Generalized Hyperbolic process (GHYP) and its special cases Hyperbolic process (HYP), Normal Inverse Gaussian process (NIG) and Variance Gamma process (VG).

2.2.1. Generalized Hyperbolic Distribution (GHYP)

Generalized Hyperbolic Distribution is an important continuous probability distribution which is popular in financial modeling due to its semi-heavy tails. Its density function is formulated as follows.

$$d_{GHYP}(x;\alpha,\beta,\delta,\mu,\lambda) = \frac{(\sqrt{\alpha^2-\beta^2}/\delta)^{\lambda}}{\sqrt{2\pi}K_{\lambda}(\delta\sqrt{\alpha^2-\beta^2})} e^{\beta(x-\mu)} \frac{K_{\lambda-1/2}(\alpha\sqrt{\delta^2+(x-\mu)^2})}{(\sqrt{\delta^2+(x-\mu)^2}/\alpha)^{1/2-\lambda}} \tag{2.12}$$

Moment generating function is frequently used in option pricing, therefore given below.

$$M(x;\alpha,\beta,\delta,\mu,\lambda) = e^{x\mu}(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+x)^2})^{\lambda/2}\frac{K_{\lambda}(\delta\sqrt{\alpha^2-(\beta+x)^2})}{K_{\lambda}(\delta\sqrt{\alpha^2-\beta^2})} \tag{2.13}$$

where λ, α, δ are parameters regulating the shape, β the asymmetry and μ the location. K_{λ} is a modified Bessel function of the third kind. Generalized Hyperbolic distribution can also be represented as a Normal mean-variance mixture model where the mixture is a Generalized Inverse Gaussian distribution.

If λ is fixed to 1, the distribution becomes Hyperbolic distribution (HYP), if λ is fixed to $-1/2$ it becomes Normal Inverse Gaussian distribution (NIG) and if δ is set to 0 it becomes Variance Gamma distribution (VG).

An important weakness of the GHYP process is, it is not closed under convolution. In other words, sum of GHYP random variates is not GHYP distributed. When meddling with financial data if log-returns are modeled as daily under GHYP distribution, weekly and monthly returns are not GHYP distributed.

To illustrate the fit of GHYP distribution, we examined how S&P 500 behaves in short term and long term. Figures 2.1, 2.2, 2.3 and 2.4 belong to S&P 500 levels and logreturn density comparisons between December 21, 2009-2014 and December 21, 2013-2014 respectively.

When looked through a span of 5 years, the empirical density is close to fitted GHYP distributions. Asymmetric and symmetric GHYP distributions have also similar density shapes. Fitted normal distribution shape is quite unsuccessful to represent the empirical log-returns.

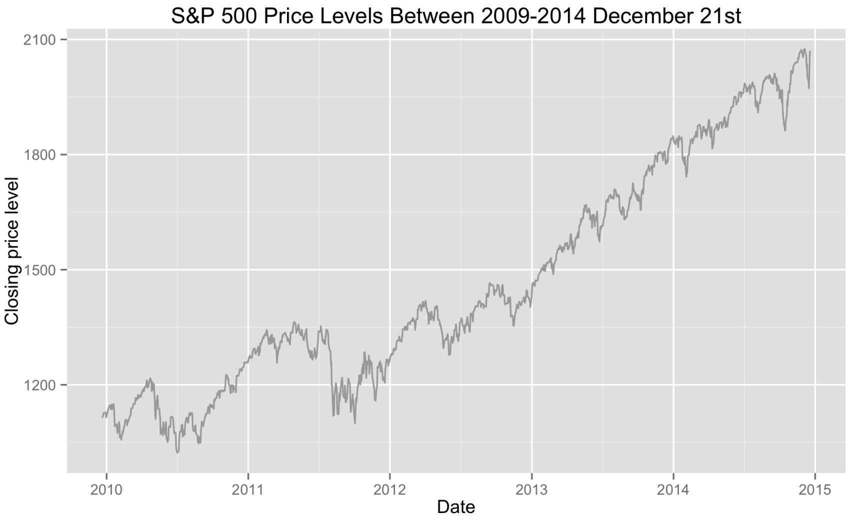

Figure 2.1. S&P 500 Levels Between 2009-2014

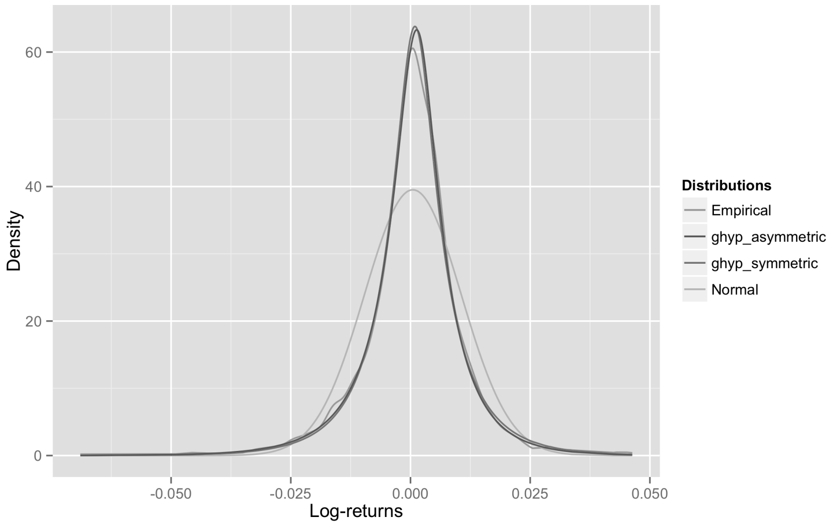

Figure 2.2. S&P 500 Log-return Density Between 2009-2014.

But when we focus only on a single year, the empirical log-return density become irregular in shape, especially on the lower side of the negative log-return tail and on the upper side of the positive log-return tail. But GHYP still provides a better fit than the others.

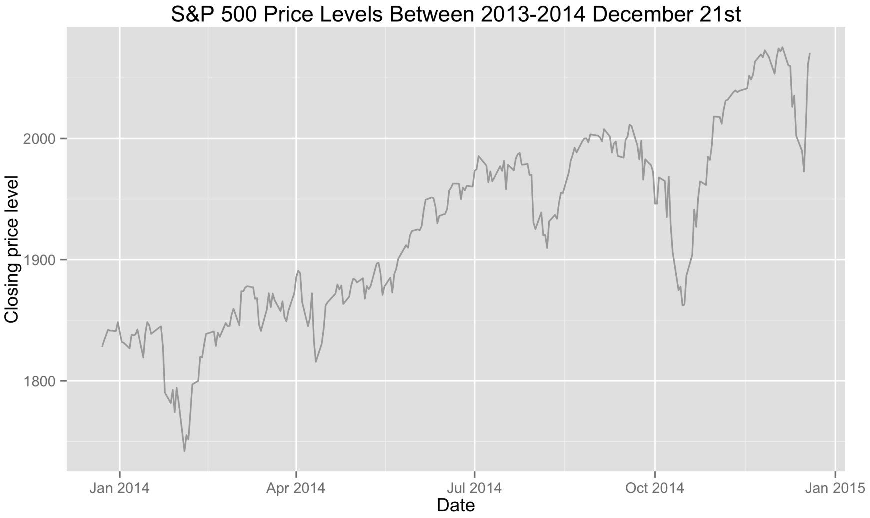

Figure 2.3. S&P 500 Levels Between 2013-2014.

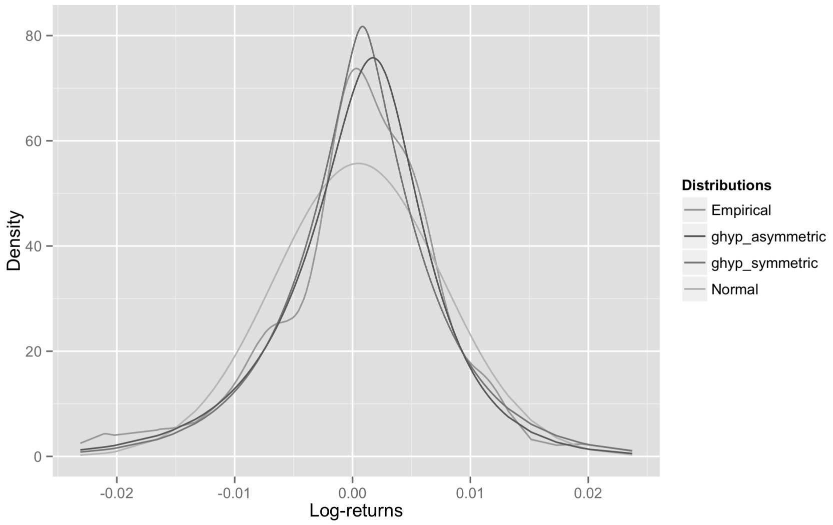

Figure 2.4. S&P 500 Log-return Density Between 2013-2014.

2.3. GARCH

GARCH(p, q) models, first suggested by Bollerslev (1986), based on ARCH models of Engle (1982). The literature is quite rich with GARCH-type option pricing models. Yet, first a simple GARCH(1,1) model with Normal distribution as the error distribution is given below.

$$Y_t = \mu + \sigma_t z_t,\ z_t \sim N(0,1) \tag{2.14}$$

$$\sigma_t^2 = \omega + \alpha(z_{t-1})^2\sigma_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{2.15}$$

$$\omega, \alpha, \beta \geq 0,\ \alpha + \beta < 1 \tag{2.16}$$

Y_t is the log-return of the underlying assets, μ is the drift parameter of the returns, σ_t is the spot volatility and z_t is the error term regulated by the standard Normal distribution. The variance is a function of the long-term variance term ω, error term of the previous time period with coefficient α and the variance of the previous time period with coefficient β. Next time period's volatility is deterministically affected from the previous estimation error And the previous volatility term.

There are serious differences from the GBM framework when using GARCH models. GARCH(1,1) with Normal distribution is leptokurtic, meaning it has fatter tails and higher peak. GARCH offers a discrete framework for modeling time series which can be considered as a disadvantage. Volatility is not constant and also volatility innovations are not independent, which in turn helps modeling the volatility clustering. Variance of the next time period is always deterministic given the previous time periods' information.

GARCH is also quite modular as a framework model. Features can easily be added and changing the distribution governing the error term is almost effortless. The error term ϵ_{t-1}^2 associated with coefficient α is considered as a function called "news impact function" ($\varphi(\epsilon_t)$) and can be stylized differently (e.g. NGARCH and TGARCH). In fact, some small changes in those two features are reported to improve the model significantly.

Following Duan (1995), the simple GARCH(1,1) framework goes under a small change under probability measure $\mathbb{P}$. The drift parameter μ is modeled as $r + \lambda\sigma_t - 1/2\sigma_t^2$. λ can be considered as unit "risk premium".

$$Y_t = r + \lambda\sigma_t - \sigma_t^2/2 + \sigma_t z_t,\ z_t \sim N(0,1) \tag{2.17}$$

$$\sigma_t^2 = \omega + \alpha(z_{t-1})^2\sigma_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{2.18}$$

$$\omega, \alpha, \beta \geq 0,\ \alpha + \beta < 1 \tag{2.19}$$

2.3.1. Heston-Nandi GARCH

Heston-Nandi GARCH model by Heston and Nandi (2000) introduces a semi-closed form solution of the GARCH option pricing model using a variation of the GARCH price process. Authors indicate that their model is more similar to NGARCH and VGARCH of Engle and Ng (1993) than the classic GARCH representation of Bollerslev (1986) and Duan (1995). The pricing process for one period GARCH under physical measure is as follows.

$$\begin{aligned} Y_t &= r + \lambda\sigma_t - \sigma_t^2/2 + \sigma_t z_t,\ z_t \sim N(0,1) && (2.20)\\ \sigma_t^2 &= \omega + \alpha(z_{t-1} - \gamma\sigma_{t-1})^2 + \beta\sigma_{t-1}^2 && (2.21)\\ &\omega, \alpha, \beta \geq 0,\ \alpha + \beta < 1 && (2.22) \end{aligned}$$

where γ is the asymmetry parameter.

Under risk neutral measure and under the assumption of "the value of a call option with one period to expiration obeys the Black-Scholes-Rubinstein formula" the process becomes as follows

$$\begin{aligned} Y_t &= r - \sigma_t^2/2 + \sigma_t(z_t + (\lambda + 1/2)\sigma_t),\ z_t \sim N(0,1) && (2.23)\\ \sigma_t^2 &= \omega + \alpha(z_{t-1} - (\gamma + \lambda + 1/2)\sigma_{t-1})^2 + \beta\sigma_{t-1}^2 && (2.24)\\ &\omega, \alpha, \beta \geq 0,\ \alpha + \beta < 1 && (2.25) \end{aligned}$$

Finally, the value of the call option under risk neutral measure can be found using the below formula.

$$\begin{aligned} C &= e^{-rT}E^*[Max(S(T) - K, 0)] && (2.26)\\ &= \frac{S(0)}{2} + \frac{e^{-rT}}{\pi}\int_0^\infty Re\left[\frac{K^{-iu}f^*(iu+1)}{iu}du\right] && (2.27)\\ &- Ke^{-rT}\left(\frac{1}{2} + \frac{1}{\pi}\int_0^\infty Re\left[\frac{K^{-iu}f^*(iu)}{iu}du\right]\right) && (2.28) \end{aligned}$$

2.4. Martingale Measures Under Incomplete Markets

Under complete market assumption the martingale measure is known to be unique. For the incomplete markets, martingale measure is also no longer unique. The existence of multiple martingale measures also results in multiple risk neutral prices, and not all martingale measures are valid for all models and parameter specifications. For instance, Esscher transform martingale measure under Lévy processes with GHYP distribution is only defined for specific combinations of the parameters and the risk free rate.

Three of those martingale measures will be used in the following chapters: (Generalized) Local Risk Neutral Valuation Relationship ((G)LRNVR), Mean Correcting Martingale Measure (MCMM) and the Esscher Transform. All martingale measures covered in this study are commonly used in the option pricing models.

2.4.1. (Generalized) Local Risk Neutral Valuation Relationship

Local Risk Neutral Valuatıon Relationship (LRNVR) is first proposed by Duan (1995) and later generalized (GLRNVR) by Duan (1999) again. We are going to follow Duan (1995) for the definition below.

“A pricing measure $\mathbb{Q}$ is said to satisfy the locally risk-neutral valuation relationship (LRNVR) if measure $\mathbb{Q}$ is mutually absolutely continuous with respect to measure $\mathbb{P}$, and $S_t/S_{t-1}|\phi_{t-1}$ is distribute lognormally (under $\mathbb{Q}$),

$$E^{\mathbb{Q}}\left[\frac{S_t}{S_{t-1}}\middle|\phi_{t-1}\right] = e^r \tag{2.29}$$

and

$$Var^{\mathbb{Q}}\left(log\left(\frac{S_t}{S_{t-1}}|\phi_{t-1}\right)\right) = Var^{\mathbb{P}}\left(log\left(\frac{S_t}{S_{t-1}}|\phi_{t-1}\right)\right) \tag{2.30}$$

almost surely with respect to measure $\mathbb{P}$.”

The LRNVR is claimed to hold under any of the three utility function types; constant relative risk aversion (CRRA) or constant absolute relative risk aversion (CARRA) if the changes in the aggregate consumption are normal distributed with constant mean and variance, or the utility function is linear.

LRNVR works for GARCH option pricing models with Normal distribution as the error distribution. For other distribution functions the generalized version in Duan (1999) can be used.

2.4.2. Mean Correcting Martingale Measure (MCMM)

Another way to obtain a martingale measure under incomplete markets is to modify the mean parameter of the model. We denote the old mean parameter with m_{old} (i.e. $\mu - 1/2\sigma_t^2$ in BS setting) and the mean-corrected version as m_{new}. We follow Schoutens (2003) definition.

$$m_{new} = m_{old} + r - log(\phi(-i)) \tag{2.31}$$

where r is the risk-free rate and $\phi(x)$ is the characteristic function of the distribution. It is also possible to use moment generating function $M(u)$ of the distribution, then the equation becomes.

$$m_{new} = m_{old} + r - log(M(1)) \tag{2.32}$$

For the GBM case $log(M(1)) = \mu$ and m_{new} becomes

$$\mu - \frac{1}{2}\sigma_t^2 + r - \mu = r - \frac{1}{2}\sigma_t^2 \tag{2.33}$$

Under Lévy processes with GHYP distribution m_{new} becomes

$$m_{new} = r - log\left(\left(\frac{\alpha^2 - \beta^2}{\alpha^2 - (\beta+1)^2}\right)^{\lambda/2} \frac{K_\lambda\left(\delta\sqrt{\alpha^2 - (\beta+1)^2}\right)}{K_\lambda\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}\right) \tag{2.34}$$

2.4.3. Esscher Transform

Esscher transform is actually a method from actuarial science, initially proposed by Esscher (1932). Gerber and Shiu (1994) is the first study which uses Esscher transform as a martingale measure to price options. Siu et al. (2004) used Esscher transform with GARCH framework. Chorro et al. (2012) use Esscher transforms with GARCH framework under GHYP distribution. The general approach is to find a parameter θ that satisfies the following equation.

$$r = log\left(\frac{M(\theta+1)}{M(\theta)}\right) \tag{2.35}$$

Risk neutral density with the Esscher transform becomes

$$f(x;\theta) = \frac{e^{x\theta}f(x)}{M(\theta)} = e^{x\theta - log(M(\theta))}f(x) \tag{2.36}$$

3. EMPIRICAL ASSESSMENT OF OPTION PRICING MODEL PERFORMANCES

Empirical tests, using real life options data, are the common way to assess and benchmark performance of option pricing models. Though, there is no single methodology. The way models are assessed and benchmarked affect the outcome of the assessment. In this chapter, a layout of the empirical option pricing process is presented. Then, a set of studies is used to show how different studies embrace different ways to handle each stage of the empirical pricing process.

3.1. Introduction

Even though options were traded on the exchanges before, seminal works of Black and Scholes (1973) and Merton (1973) made them quite popular with the simplicity and accuracy of Black-Scholes model[2] . Though, in time, it is understood that neither markets are so simple nor the model is so accurate.

There is wide criticism about Black-Scholes model, even by its authors. For instance, Black (1975) makes some remarks about the use of the model starting with the constant volatility assumption. This wide criticism is followed by many other models claiming to deliver where Black-Scholes falls short. Usually, these models lack the simplicity of the Black-Scholes model. Their intricate nature might hazard other risks such as overfitting.

The main objective of this chapter is to lay out the differences in how experiments are conducted with empirical option pricing data. These differences include, but are not limited to, handling of option contracts, parameter inference methods and results presentation. To show differences we use a set of studies which perform empirical experiments with options[3] . It must also be noted that this study is limited to European

[2] See Mackenzie (2008) for a detailed history about the development of Black-Scholes model

[3] This chapter is not a thorough literature survey, so the number of studies are limited.

options.

The rest of the chapter is organized as follows. Literature used in this study is provided in Section 3.2 in detail. Section 3.3 is about the data and preparation used in the experiments. Those include, asset types, handling of risk free rate, handling of splits and dividends. Section 3.4 is about parameter optimization of the models. Especially the difference between parameter optimization with options and historical asset returns are discussed. Results representation and benchmarking methods are discussed in Section 3.5. Non-standard procedures such as robustness checks and correlations between error terms and contract parameters are discussed in Section 3.6. Finally, findings are summarized and overall observations are shared in Section 3.7.

3.2. Studies

Criteria for chosen studies from the literature is in line with our objectives. Each study has an empirical option pricing aspect (i.e. model out-of-sample estimates are assessed with market option prices). They describe how the data handled in their experiments and how parameters are optimized for estimation. The list of studies here are not considered as extensive but as representative.

Pricing error models are as follows. Bakshi et al. (1997) propose an important improvement on Black-Scholes model with the addition of stochastic volatility, stochastic interest rates and jumps. Dumas et al. (1998) propose BS models with implied volatility values inferred from market prices and contract parameters. Heston and Nandi (2000) propose a semi-closed form for pricing options with GARCH processes. Christoffersen and Jacobs (2004b) benchmark different GARCH settings from the perspective of option pricing. They come up with the conclusion that prudent models work better than models with many parameters. Lehar et al. (2002) benchmark BS, Hull-White stochastic volatility model and GARCH using both pricing and risk metrics. Barone-Adesi et al. (2008) propose a nonparametric approach for GARCH error distributions. Fan and Mancini (2009) implement a model guided non-parametric guidance method for better pricing results than parametric methods. Chorro et al. (2012) provide different

ARCH type models with Generalized Hyperbolic distribution for the distribution of error terms.

It can easily be observed that most of the studies mentioned above use ARCH-type processes. This is no coincidence. ARCH (or GARCH) is extensively supported as a good option pricing base method. In addition, it is informative to examine pricing studies of evolving on a single branch to eliminate additional confusion from different model structures.

3.3. Data

3.3.1. Underlying Assets

Majority of the studies use S&P 500 and its options (SPX) as the single underlying asset. There are several exceptions though. Lehar et al. (2002) use FTSE 100. Chorro et al. (2012) both S&P 500 and CAC 40 indices. This is good practice to see whether model performances are dependent on the asset. In other words, if the same experiment was repeated with another asset, would the outcome be the same?

It is possible to use different prices for each asset. Main quote types are either closing prices or bid-ask mid-prices. There is also a synchronization issue for S&P 500 (perhaps other assets too) closing prices, since closing times of options exchange and stock exchange are different. Some studies use mid-day prices to match the prices. A list can be found in Table 3.1

3.3.2. Contract Range and Filtering

Not all contracts in the data set are used for experiments. Studies generally put moneyness and maturity restrictions on the contracts since pricing contracts with too high and low moneyness/maturity values does not yield consistent results and skews the overall results. There are also fewer contracts and there is low trading volume at the edges of moneyness maturity region which makes market prices scarcer and unstable.

Table 3.1. Quote types of different studies.

Quote Type	Matching Studies
Closing prices	Fan and Mancini (2009) Chorro et al. (2012)
Intraday bid-ask	Bakshi et al. (1997) (3 PM CT) Dumas et al. (1998) (2:45-3:15 PM) Heston and Nandi (2000) (2:30-3:15 PM CT)
Real-time trades	Lehar et al. (2002)

Bakshi et al. (1997) use data between 1988-1991. Although same calculations are also done for put options, they report only results on the call options. There are 38,749 call options in their data set. Contracts with less than 6 days to maturity, price lower than \$3/8, not satifying no-arbitrage conditions are removed.

Dumas et al. (1998) use data between 1988-1993. They use only Wednesdays' data. Contracts with moneyness outside $[0.9, 1.1]$ and with maturity less than 6 or more than 100 days are removed.

Heston and Nandi (2000) use data between 1992-1994. They use only Wednesdays' data. They report to follow Dumas et al. (1998) in their contract filtering. Also, if a contract is represented more than once on each day, only the earlier recorded contract is considered. A total of 10,100 contracts are reported.

Lehar et al. (2002) use data between 1993-1997, 1210 trading days. They report to remove contracts with maturity less than 2 weeks, price less than 0.05, moneyness outside $[0.9, 1.1]$ and annualized implied volatility outside the interval of 5% and 50%. The net number of contracts they use in their experiments are 65,549.

Christoffersen and Jacobs (2004b) use the same data set as Bakshi et al. (1997). In addition, only Wednesday contracts are used.

Fan and Mancini (2009) use data between 2002-2004. They report only call options results. Contracts with implied volatility higher than 70%, price less than 0.125, maturity less than 20 days or more than 240 days are removed from the data set. A total of 101,036 observations are reported.

Chorro et al. (2012) use data between 2006-2007. They report to use only quarterly contracts and discard monthly contracts. Contracts outside moneyness $[0.8, 1.2]$ are removed. The number of contracts is 82,657 for CAC 40 and 39,885 for S&P 500.

3.3.3. Risk Free Rate

There are generally three ways to handle risk free rates. It is possible to fix the risk-free rate to a value, fix to a risk-free yield or interpolate between available maturities. The list of different risk-free rate handling methods can be found in Table 3.2.

Table 3.2. Risk free rate handling methods.

Risk Free Rate	Matching Studies
Fixed	Christoffersen and Jacobs (2004b) (5%)
T-Bill bid-ask	Bakshi et al. (1997) Dumas et al. (1998) Heston and Nandi (2000)
Treasury rates	Fan and Mancini (2009)
GBP LIBOR	Lehar et al. (2002)
Overnight Indexed Swap	Chorro et al. (2012)

3.3.4. Dividends and Splits

Dividends can be adjusted in two ways. First method is to convert discrete values to a continuous yield. It is a plausible method for indices which contain many dividend yielding assets and European options. Another method is; discounted[4] discrete dividends are subtracted from the initial spot price of the underlying asset (S_0).

[4] Discount rate is the risk-free rate.

Table 3.3. Dividend handling methods.

Dividend Method	Matching Studies
Discount from S_0	Bakshi et al. (1997) Dumas et al. (1998) Heston and Nandi (2000)
Continuous yield	Lehar et al. (2002)
Implied from put-call parity	Fan and Mancini (2009) Chorro et al. (2012)
No mention	Christoffersen and Jacobs (2004b)

3.4. Inference

Both parametric and nonparametric models need training data to optimize their parameters or nonparametric distributions. In this section, studies show greater differences in the methodology than in the other sections. Proper parameter inference is crucial for a model's out-of-sample success. See Christoffersen and Jacobs (2004a) and Bams et al. (2009) for further discussion.

Bakshi et al. (1997) use sum of squared differences between the model estimate and the market option price as the objective to train their models. They also compare results with maximum-likelihood estimation of historical asset returns and see that while other parameters show no significant difference, volatility estimate of implied method is four times larger than the MLE estimate.

$$SSE = \sum_{i=1}^{N} (\hat{P}_i - P_i)^2 \tag{3.1}$$

Dumas et al. (1998) fit a regression using volatilities as the response variable and transformations of strike price and time to maturity as the covariates. Both local and implied volatilities are used.

$$\sigma = max(0.01, a_0 + a_1K + a_2K^2 + a_3T + a_4T^2 + a_5KT) \tag{3.2}$$

Heston and Nandi (2000) use a mixed approach. They use a non-linear least squares method to estimate GARCH parameters $(\omega, \alpha, \beta, \gamma)$ from option prices but, since they use only Wednesdays as option data they optimize next day variance σ^2_{t+1} using underlying asset returns.

Lehar et al. (2002) use "sliding time windows" consisting of groups of option contracts for every 10 days. During the 1210 trading days time period they report to have 120 time windows. They optimize the model parameters by minimizing (Average) Squared Relative Pricing Error of each time window.

$$SRPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{P}_i - P_i}{P_i}\right)^2 \tag{3.3}$$

Christoffersen and Jacobs (2004b) test with both historical assets maximum likelihood and NLS minimization of option prices.

Fan and Mancini (2009) fit their automated correction method using a time-weighted non-parametric regression function.

Chorro et al. (2012) embrace a two stage estimation using historical log-returns of the underlying asset. Since their models consist of GARCH processes with GH error distributions, there are two sets of parameters. At the first stage, GARCH parameters are estimated. Then, by fixing the GARCH parameters GH parameters are estimated from the residuals.

As it can be clearly seen, parameter inference methods are quite dissimilar due to both model requirements and inference data sources.

3.5. Results

Studies covered in this chapter have one thing in common. They represent their out-of-sample estimation errors using result tables grouped by moneyness, maturity and, for some studies, years. They usually differ in terms of performance metric and it is not uncommon to use multiple performance metrics.

3.5.1. Grouping

Pricing errors are generally grouped by call/put, moneyness and maturity. Some studies also include yearly performances to demonstrate model performance stability.

Bakshi et al. (1997) use 6 moneyness groups $\{< 0.94$, $[0.94, 0.97)$, $[0.97, 1.00)$, $[1.00, 1.03)$, $[1.03, 1.06)$, $\geq 1.06\}$ and 3 maturity groups $\{< 60$, $[60, 180)$, ≥ 180 days$\}$.

Dumas et al. (1998) use 4 moneyness groups $\{[0.9, 0.95)$, $[0.95, 1.00)$, $[1.00, 1.05)$, $[1.05, 1.11)\}$ and 3 maturity groups $\{< 40$, $[40, 70)$, ≥ 70 days$\}$.

Heston and Nandi (2000) use 4 moneyness groups $\{< 0.95$, $[0.95, 0.99)$, $[0.99, 1.01)$, $[1.01, 1.05]$, $> 1.05\}$ and 3 maturity groups $\{< 40$, $[40, 70)$, ≥ 70 days$\}$. They also include yearly changes in aggregate results.

Lehar et al. (2002) use 7 moneyness groups $\{[0.9, 0.92)$, $[0.92, 0.95)$, $[0.95, 0.98)$, $[0.98, 1.02)$, $[1.02, 1.05)$, $[1.05, 1.08)$, $[1.08, 1.1]\}$ and 5 maturity groups $\{< 73$, $[73, 146)$, $[146, 219)$, $[219, 292)$, ≥ 292 days$\}$[5] .

Christoffersen and Jacobs (2004b) show grouping results only in the appendix. They use the same moneyness and maturity grouping as Bakshi et al. (2012).

Fan and Mancini (2009) use 5 moneyness groups $\{< 0.8$, $[0.8, 0.94)$, $[0.94, 1.04)$, $[1.04, 1.2)$, $\geq 1.2\}$ and 3 maturity groups $\{< 60$, $[60, 160)$, ≥ 160 days$\}$. Different from

[5] Maturity grouping in the study is originally fraction of years (i.e. 0.2, 0.4, 0.6 and 0.8).

other considered studies, their ATM region is not symmetric around 1 (i.e. $S_0 = K$).

Chorro et al. (2012) use 6 moneyness groups $\{< 0.8$, $[0.8, 0.9)$, $[0.9, 1.0)$, $[1.0, 1.1)$, $[1.1, 1.2)$, $\geq 1.2\}$ and 3 maturity groups $\{< 182$, $[182, 274)$, ≥ 274 days$\}$[6] .

Even though studies follow similar moneyness and maturity ranges, their partitioning styles are different. Those differences might cause performance changes within the groups and even for the selection of the best model in the benchmarks.

3.5.2. Performance Metrics

There are a number of pricing error types used by these considered studies, but they are usually of two common types. Pricing error (PE) is the difference of price estimate of the model ($\hat{P}$) and the market option price (P), calculated simply by $PE = \hat{P} - P$. Relative pricing error (RPE) is the division of price difference with the market price, calculated as $RPE = (\hat{P} - P)/P$. Root mean squared error ($RMSE = \sqrt{1/N \sum_i^N (PE_i)^2}$) and Absolute Relative Pricing Error ($ARPE = |RPE|$) are frequently used derivations of PE and RPE.

The main difference, in terms of option pricing errors, is relative errors are more biased towards short term and out-of-the-money (OTM) options because they are usually cheap and pricing errors are more biased towards expensive options such as long term and in-the-money (ITM) contracts. This difference is explicitly noted in several studies such as Bakshi et al. (1997), Dumas et al. (1998) and Heston and Nandi (2000).

Bakshi et al. (1997) use RPE and absolute pricing error ($APE = |PE|$) as the performance metrics.

Dumas et al. (1998) use RMSVE[7] , Mean Outside Error (MOE) and Mean Absolute Error ($MAE = 1/N \sum_i^N APE_i$). MOE measures the PE if the model estimate

[6] Maturity grouping in the study is originally fraction of years (i.e. 0.25, 0.5, 0.75 and 1).

[7] Root mean squared valuation error. Essentially same as RMSE.

is outside bid-ask spread. If the price estimate is below bid price or above ask price, then the price difference is taken as error.

$$MOE = 1/N \sum_{i}^{N} \min\{\hat{P}_i - Bid_i; 0\} + \max\{\hat{P}_i - Ask_i; 0\} \tag{3.4}$$

They also use Akaike Information Criterion (AIC) as a goodness of fit test in their performance assessment metric.

Heston and Nandi (2000) use RMSE, MOE, MAE and %Error in their assessments. Their version of MAE consider absolute pricing errors only if model estimate is outside the bid-ask spread. %Error is calculated as the RMSE divided by the average option price in the given moneyness-maturity group.

Lehar et al. (2002) measure only RPE and ARPE in their pricing error assessments. Christoffersen and Jacobs (2004b) use MSE and RMSE in their result tables.

Fan and Mancini (2009) use Bias (i.e. PE) RMSE, MAE and their min/max and percentage versions such as Bias(%) (RPE), MAE(%) (ARPE) and RMSE(%) ($1/N \sum_i^N SRPE_i$). The last metric Err>0% measures the number of contracts with positive pricing errors relative to all contracts.

Chorro et al. (2012) use AARPE ($1/N \sum_i^N ARPE_i$) as their only performance metric.

3.6. Other

Aside from pricing assessment and benchmarking, many studies include extra analysis about their models. Most common analysis is about (especially Delta) hedging performance of the models. Other analyses mainly include risk (e.g. VaR) and robustness checks.

Bakshi et al. (1997) look into explaining pricing errors with regression. Regression covariates include moneyness, maturity, risk-free interest rate, bid-ask spread and implied volatility. They also examine delta hedging with only the underlying asset and delta hedging for all risks except the jump risk.

Dumas et al. (1998) check for hedging performance and robustness. Their robustness checks mention three issues. Quadratic formulation's response to distinct values are tested. They also check for in-sample estimation stability by increasing lookback period. To check for prediction stability, they divide the contract data into subsamples and compare subsample performances.

Heston and Nandi (2000) try S&P 500 futures to filter volatility, in addition to index levels and conclude there is no significant difference in out-of-sample performance of their model.

Part of the study of Lehar et al. (2002) is dedicated to risk management. Their calculations are based on Value-at-Risk and Monte Carlo simulations. They first check for the proportion of values exceeding threshold value. Then they also check for the distribution of errors with Kupiec test. They make an additional test in pricing, by fitting a regression on ARPE pricing errors. Covariates are time to maturity, moneyness and call/put categorization. The methodology and findings are similar to Bakshi et al. (1997).

Christoffersen and Jacobs (2004b) experiment further with periodic parameter updates and different parameter inference methods to check the robustness of their methodology. They also compare risk neutral densities of different models and inference methods.

Fan and Mancini (2009) visually investigate mispricing of models belonging to different moneyness and maturity regions.

3.7. Conclusion

In this chapter we investigated the methodologies embraced by empirical option pricing studies. All studies are thorough with their analyses and their option pricing process stages are similar, but there is no unified methodology.

We showed that they usually disagree in how to measure model performance and how to present them. Different maturity and moneyness intervals can be deemed short/long term and ITM/OTM. Data handling processes are also different from the data filtering to the choice of risk free rates and dividends.

We made no conclusive arguments about which methodology is better, because all of them have their own merits and pitfalls. There should be just few well established rules such as proper consideration of dividends. Also, it should be kept in mind that, as reported by Christoffersen and Jacobs (2004a) and Bams et al. (2009), having the parameter estimation optimization metric and performance assessment metric the same improves the performance of the models significantly. Other issues, such as measuring performance with absolute (e.g. RMSE) or relative (e.g. ARPE) metrics is entirely up to the experimentation. Nevertheless, Table 3.4 summarizes our methodology used in the numerical experiments in other chapters and compares it with the literature covered in this chapter.

In addition, RPE, ARPE and APE pricing error metrics are commonly used in all the main chapters for model assessment. Grouping as in the literature is only done in Chapter 5. There are 4 moneyness groups $\{[0.5, 0.9)$, $[0.9, 1.00)$, $[1.00, 1.1)$, $[1.1, 1.5)\}$ and 5 maturity groups $\{$ $[7, 30)$,$[30, 60)$,$[60, 90)$,$[90, 180)$ and $[180, 365]$ days$\}$.

Table 3.4. Empirical experiments methodology comparison.

Part of the methodology	Literature	Our approach
Assets	Single Asset	Multiple Assets (SPX, NDX)
	Multiple Assets	
Quote type	Closing prices	Closing prices
	Intraday bid-ask mid	
	Real time trades	
Contract Filtering	Moneyness limits	Moneyness limits [0.5,1.5]
	Maturity limits	Maturity limits [7,365] calendar days
	Arbitrage conditions	Arbitrage conditions
		Trading volume limit ($\geq$ 100)
	Price limit	Price limit ($\geq$ \$0.05)
	Wednesdays-only	
	IV limit	IV limit ($> 0.01, \leq 0.7$)
Risk-free rate	Fixed	US Treasury yields (approximated)
	Fixed to an asset (i.e. 90 days T-Bills)	
	Yield approximation	
Dividends	No dividend	Continuous dividend yield
	Discrete dividends	
	Continuous dividend yield	

4. CLUSTER STABILITY OF ERROR REPRESENTATION IN OPTION PRICING STUDIES

Model performance is generally not sustained over a long period of time in option pricing studies. Therefore, aggregate result tables showing model errors should be approached with suspicion about posterior performance. By extension, prediction stability over time is problematic. Aggregating over arbitrary moneyness-maturity partitions aggravates it further, because each partition might contain a mixture of strong and weak regions of the model. Assuming that all models have their own moneyness and maturity boundaries for weak and strong regions, we examine several unsupervised learning algorithms. We test cluster stabilities and show that using algorithm predictions is better in error representation than merely aggregating results over arbitrary boundaries.

4.1. Introduction

Empirical option pricing studies aim to measure the difference of their option model estimates from the market prices. They assess models' performances and benchmark the models with the error metrics such as (Absolute) Relative Pricing Error (A-RPE) and (Absolute) Pricing Error (A-PE). Then, errors are aggregated into arbitrarily defined groups based on contract parameters. Criteria used in those studies are generally based on contract type (call/put), moneyness and maturity. The resulting tables display the performance of the option model in different moneyness and maturity regions and for call and put options respectively.

There are many examples in the literature following the above procedure. For instance, Bakshi et al. (1997) use APE and RPE in their out-of-sample assessments. A summary table of call option errors are reported. Errors are grouped into moneyness and maturity intervals, then group averages are displayed and commented on. Grouping is based on some loose criteria. Moneyness intervals are symmetric around 1 (i.e.

spot asset price and strike price are equal) and maturity intervals are labeled as short-term (less than 60 days), mid-term (60-180 days) and long-term. Similar grouping with different interval values can be observed in majority of similar option pricing studies. Some prominent examples are Dumas et al. (1998), Heston and Nandi (2000), Lehar et al. (2002), Christoffersen and Jacobs (2004b) and Chorro et al. (2012).

There are several weaknesses to this kind of representation, especially for benchmark studies. First, the summary table is a snapshot of the results based on the data set of that time period. There is no indication of how these results might change over time. For instance, model A can seem better than model B for a time period, say T_1. But for another, perhaps much longer, time period T_2 model B could be the better model.[8] Error evolution in different time periods is not the concern of these studies, neither is cross-validation. One of the very few exceptions is Heston and Nandi (2000). Their summary table consists of benchmark of 4 models' aggregate errors (no moneyness/maturity grouping) for the contracts between 1992-1994. They provide information for each year's errors and the overall values of all 3 years. Year over year values show that the model at the third place (*Ad hoc* BS) was actually in the second place in 1992. It implicitly indicates that if the data period were chosen from contracts between 1991-1993, relative model performance could have been different.

The second problem we observe is that, aggregate tables display the performance of models in given moneyness maturity groups. But from a different perspective, this also hinders proper representation of their weak and strong parts. Clustering results over arbitrary boundaries bears the risk of mixing strong and weak regions and report them as 'average' regions.

Total elimination of these problems is not possible due to the mercurial nature of the financial markets. Relative model performances and strong/weak regions will shift in time. Yet, it is possible to increase cluster stability and provide consistent summary results for longer periods of time. In order to achieve this objective, we propose clus-

[8] Even though the term "better" is dependent on the performance metric, better can be defined here as the lowest error on an agreed single performance metric (e.g. RPE, APE). See Chapter 5 for further information.

tering and regression algorithms to assess model pricing errors. Figure 4.1 illustrates

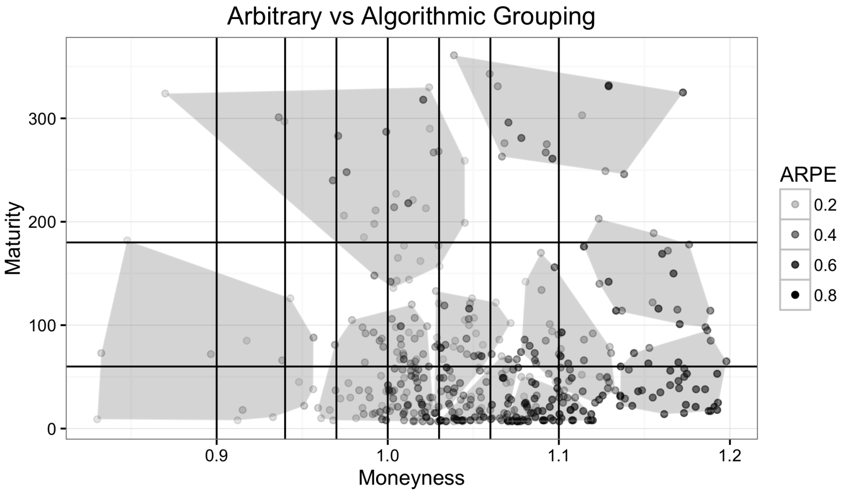

Figure 4.1. Algorithmic vs arbitrary grouping example

an example of the difference between arbitrary grouping vs algorithmic grouping of ARPE pricing errors for a sample set of 2011 SPX put contracts priced by Heston-Nandi GARCH model. Different shading of the points denotes the ARPE magnitude on the contract. A K-Means clustering algorithm is fitted on the sample set and clusters are illustrated with convex polygons[9] . The main advantage of the algorithmic method is its freedom in determining group boundaries according to the error magnitudes.

Data mining algorithms are a good way to dynamically and automatically partition moneyness and maturity regions via unsupervised learning. K-means is one of the widely used clustering algorithms for almost any clustering purpose. It will also be a good base benchmark. In other words, if even K-Means can yield better performance than arbitrary grouping, then the current practice can definitely be improved. We call the arbitrary grouping 'static clustering'.

Other models in consideration are more subtle and powerful in comparison. Support Vector Machine (SVM) introduced by Cortes and Vapnik (1995) is a classification and regression algorithm. In this study, only their regression abilities will be used. Their success in numerous other applications makes the algorithm a good candidate

[9] Convex polygons are used just for illustrative purposes. Actual K-Means regions are naturally different.

method and a benchmark against actual clustering methods are the reasons to implement SVM in this study.

Decision Trees (DT) are introduced by Breiman et al. (1984). What the algorithm basically does is splitting the covariate (i.e. moneyness and maturity) values into regions and fitting a regression model in each region. Splits are binary (hence the tree structure), so what we actually get is a partitioning close to incumbent static clustering method. Conditional Inference Trees (CIT) of Hothorn et al. (2006) claim to improve on overfitting and variable selection bias issues, compared to classical decision trees.

All algorithms as well as static clustering methods will be tested in different time periods. Only out-of-sample option pricing errors will be used in experiments. Cluster stability will be measured using difference of realized pricing errors from cluster estimates. In order to avoid confusion between model pricing error and clustering stability error, different error terms will be used.

This study makes the following contributions. To the best of our knowledge, result representation via time series cross-validation is very rare in option pricing. Thus, our approach enables better error representation stability over long periods. Training data mining algorithms to group pricing errors is totally a new approach to model assessment and benchmarking. The first advantage of such an approach is that comparisons can be better justified and are more robust. The second advantage is to assess model advantages and disadvantages at a more granular level. This way, more accurate decisions can be made to choose the specific model for each contract to obtain better results.

This chapter is organized as follows. Section 4.2 elaborates on data mining models used in this study and their parametrizations. Section 4.3 describes the conducted experiment and numerical results. Finally, findings will be summarized in Section 4.2 and further uses will be discussed.

4.2. Clustering and Regression Algorithms

Clustering models are part of unsupervised learning algorithms. Given a set of training data, a clustering model aims to construct the best cluster arrangement according to its objective function. Some algorithms, such as K-Means, might require predefined information such as number of clusters. Others, such as Decision Trees and Conditional Inference Trees, also fit regressions to improve their prediction capabilities.

The main purpose of using clustering and regression algorithms is to improve prediction stability. K-Means, Support Vector Machine (SVM), Decision Tree (DT) and Conditional Inference Tree (CIT) algorithms are tested to observe their relative performance to each other and arbitrary grouping methods. All clustering algorithms used here are powerful tools that are proven to be effective in other clustering, classification and regression problems. They are also on the top 10 list of Wu et al. (2007) for data mining algorithms.

4.2.1. K-Means Clustering

K-means method used in this study, is an euclidean distance minimizer clustering algorithm. The objective is, simply, to find clusters where minimal internal variation occurs in the clusters. Mathematical representation of the algorithm is as follows.

$$min \sum_{k=1}^{K} \sum_{d=1}^{D} \sum_{i=1}^{N_k} (x_{i,d} - \bar{x}_{k,d})^2 \tag{4.1}$$

where K is the total number of clusters, k denotes a particular cluster, D is the number of attributes (i.e. Moneyness, Maturity), d denotes a particular dimension, N_k is the number of items in cluster k, $x_{i,d}$ is the value of attribute d of item i and $\bar{x}_{k,d}$ is the value of attribute d of the center of cluster k.

To train K-Means clusters, pricing error, moneyness and maturity attributes are used to generate clusters. Because if pricing error is not included in the cluster creation phase, then the only criteria will be the "location" of option contracts. While other data mining algorithms have the advantage of calibrating their parameters using error values, denying it from K-Means would mean unequal terms in benchmarking. On the other hand, both in-sample and out-of-sample estimations will be made with only moneyness and maturity, not pricing error.

4.2.2. Support Vector Machine

Support Vector Machine (SVM) is a classification and regression algorithm. SVM algorithms perform class separation by finding the optimal separating hyperplane, which maximizes the margin between two classes; and, if such linear separator cannot be found, data is transformed into a higher dimension space via kernel techniques. If there are overlapping classes with data points on the unintended regions, their weights are reduced. Then, a quadratic optimization model is constructed; which can be solved easily with known methods. The ϵ-regression version deviation from the fitted values more than the predefined ϵ constant is aimed to be minimized.

Dual of the mathematical model is reported to be generally easier to solve. The dual representation for one covariate and one response variable set (x_i, y_i) is as follows.

$$max \quad -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) \tag{4.2}$$

$$-\epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + y_i\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) \tag{4.3}$$

subject to

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \tag{4.4}$$

$$0 \leq \alpha_i \leq C, \forall_{i=1,\ldots,l} \tag{4.5}$$

$$0 \leq \alpha_i^* \leq C, \forall_{i=1,\ldots,l} \tag{4.6}$$

where α_i and α_i^* are Lagrange multipliers of the primal problem, C is a constant which regulates the sensitivity of tolerance in terms of deviation from ϵ, y_i is the response variable (i.e. pricing errors) and x_i is the explanatory variables (covariates such as moneyness and maturity) while $i = 1, .., l$ is the index of the training data set and $k(., .)$ is the specified kernel function.

Kernel function used in this study is a radial basis function. Mathematical representation is as follows.

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \tag{4.7}$$

Further information about the SVM model can be found in the tutorials of Smola and Schölkopf (2004) and Hsu et al. (2003).

4.2.3. Decision Trees

Decision Trees are statistical methods which use recursive partitioning, with roots in Automated Interaction Detection (AID) by Morgan and Sonquist (1963). Although there are numerous variations, mainly, there are two fundamental steps to tree based algorithms. First step is to create regions in the data set via univariate splits. Second step is, unless it is not a classification problem, the second step is to fit predictive models to each of those regions. Unlike K-Means and SVM, output tree structure is similar to static clustering methods in terms of data partitioning. Union of the partitions cover the whole covariate (parameter) space. An illustration over a basic example is given in Figure 4.2.

Classification and Regression Trees (CART) of Breiman et al. (1984) is chosen as the decision tree algorithm. CART tries to minimize the difference between the cluster estimate and the true value of the response variable.

$$\varepsilon = \sum_{i=1}^{N} (y_i - \sum_{k=1}^{K} c_k I_{i,k})^2 \tag{4.8}$$

where y_i is the value of the response variable of observation i, c_k is the estimate of the cluster[10] k on the response variable and $I_{i,k}$ is whether observation i is classified under cluster k.

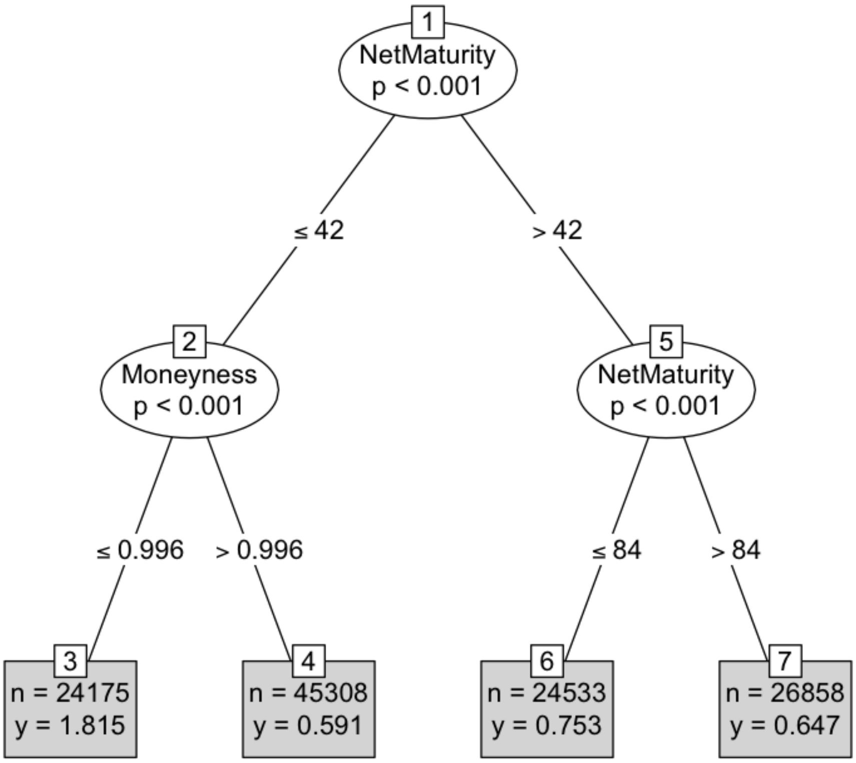

Figure 4.2. Decision Tree example.

Now, the covariate space is supposed to be split into two clusters; based on the binary split such as the example given in Figure 4.2 (Maturity $\leq$ 42, Maturity $>$ 42). To find the best split, a greedy algorithm is used to minimize the sum of the error of the first split cluster ($C_1(p,m) = i\epsilon\{x_{i,p} \leq m\}$) and the second split cluster ($C_2(p,m) = i\epsilon\{x_{i,p} > m\}$) among different alternatives of covariates (p) and covariate splitting points (m).

$$\min_{p,m} \varepsilon = \sum_{C_1(p,m)} (y_i - c_1)^2 + \sum_{C_2(p,m)} (y_i - c_2)^2 \tag{4.9}$$

[10] For the decision tree method used in this study, $c_k = 1/N_k \sum_{i=1}^{N} y_i I_{i,k}$ where N_k is the number of observations in cluster k.

This procedure is repeated until stopping conditions are satisfied. Some example stopping conditions are the minimum number of items in a node to be split and the minimum number of items in a terminal node.

It is highly probable that the resulting tree is too large and overfits the data (i.e. high out-of-sample error). So, a *pruning* procedure should be adapted. A cost-complexity parameter (α) is used to determine where the pruning should end if the pruned tree's error plus $\alpha|T_u|$ is minimized. T_u is a sub-tree (or pruned) of the resulting large tree denoted by T_0 and $|T_u|$ denotes the number of terminal nodes in sub-tree u. The optimal tree can be found using the following minimization formula.

$$T^*_u = \min_u \ \varepsilon_u + \alpha|T_u| \tag{4.10}$$

Pruning starts with the terminal node which has the lowest improvement at each iteration. The trees are cross-validated with different training and prediction subsets sampled from the data set. Further information on decision trees can be found in Hastie et al. (2009).

4.2.4. Conditional Inference Trees

Hothorn et al. (2006) propose Conditional Inference Trees with the argument that decision trees have two weaknesses; overfitting and biased favor towards estimators (covariates) with many split possibilities. They claim to overcome these problems via applying statistical tests to both splitting and termination conditions. They also claim to have a more efficient algorithm. Because, thanks to the statistical tests, their algorithm does need not to do an exhaustive search on splits of covariates for variable selection.

The main contribution of CIT to the decision tree structure is the selection of the split covariate. Instead of doing an exhaustive search on all covariates and covariate values, each covariate's statistical significance to the response variable is measured with a significance level α^{11} . Among the statistically significant covariates, the one with the minimal p-value is chosen. Split point can be determined with the same procedure as the decision tree. In addition to the already available stopping conditions, if no covariate has a p-value less than α then splitting procedure is terminated.

For pruning, another significance level α' is proposed. α' should be significantly smaller than α so that the tree shall be pruned until all terminal nodes reach to the significance level of α'.

4.2.5. Static Clusters

Static clustering uses the (almost) arbitrarily defined moneyness and maturity boundaries to partition the contract data. They are usually symmetric around the at-the-money (ATM) moneyness and have equivalent intervals. It is the common practice to show model estimation results in almost all empirical option pricing studies. In this study, two representative boundary sets from the literature are considered to be tested.

First boundary set is taken from Bakshi et al. (1997). The moneyness boundaries they use are 0.94, 0.97, 1.00 (ATM), 1.03 and 1.06. In addition, their maturity boundaries are 60 and 180 calendar days. It is approximately 42 and 126 trading days, respectively.

Second boundary set is taken from Lehar et al. (2002). The moneyness boundaries they use are 0.92, 0.95, 0.98, 1.02, 1.05 and 1.08. Here, at-the-money (ATM) position is within the interval of [0.98, 1.02]. In addition, their maturity boundaries are approximately 50, 100, 150 and 200 trading days.

Static clustering estimation is quite straightforward. Training data will be partitioned adhering to the predefined boundaries. Average pricing error in each cluster will be taken as that cluster's error prediction.

[11] α parameter is different than the cost-complexity parameter mentioned above.

4.3. Numerical Experiments

4 clustering models (K-Means, SVM, DT and CIT) and 2 different arbitrary groupings will be tested with 3 different pricing error types (RPE, ARPE and APE) with 2 option pricing models (HN and BS). 10 different time periods will be used and both in-sample and out-of-sample cluster stabilities will be measured. Option pricing models and pricing error types are given in Section 4.3.1 and Section 4.3.2. Data and how the model parameters are inferred are explained in Section 4.3.3. Elaboration of clustering algorithm testing and cross validation method used in this study are given in Section 4.3.4. Finally, results and observations based on the benchmarks are shared in Section 4.3.5.

4.3.1. Option Pricing Models

Experiments will include two different option pricing models, namely Black-Scholes model of Black and Scholes (1973) and Heston-Nandi GARCH model of Heston and Nandi (2000). Both of them are widely used as benchmarks in previous empirical option pricing studies. Black-Scholes model has a closed form solution. The formula of the call price under risk neutral measure, $C(.)$, can be calculated as follows.

$$C(S_0, K, r, q, T, \sigma) = S_0 N(d_+) - e^{-rT} K N(d_-) \tag{4.11}$$

$$d_{\pm} = \frac{log\left(\frac{S_0}{K}\right) + \left(r - q \pm \frac{\sigma^2}{2}\right) T}{\sigma\sqrt{T}} \tag{4.12}$$

where S_0 is the spot price of the asset, K is the strike price, r is the risk free rate, q is the dividend yield, T is time to maturity (in years) and σ is the annual volatility estimate.

GARCH-type models is widely in use to price option contracts, starting with Duan (1995). Even though GARCH-type models lack closed form solutions and need other methods such as Monte Carlo simulation, their performance reportedly make it up to the computational costs. Heston-Nandi GARCH (HN) model of Heston and

Nandi (2000) consists of five parameters, which means it is harder to perform parameter inference than for Black-Scholes[12] . Good news about HN model is it has a semi-closed form solution, different from other GARCH-type option pricing models, making it easier to compute the pricing estimates. The call price C at time t for a maturity of T is given as follows.

$$\begin{aligned} C &= e^{-rT} E^{*}[Max(S(T) - K, 0)] \qquad (4.13) \\ &= \frac{S(0)}{2} + \frac{e^{-rT}}{\pi} \int_0^{\infty} Re\left[\frac{K^{-iu} f^{*}(iu+1)}{iu} du\right] \qquad (4.14) \\ &- Ke^{-rT}\left(\frac{1}{2} + \frac{1}{\pi}\int_0^{\infty} Re\left[\frac{K^{-iu} f^{*}(iu)}{iu} du\right]\right) \qquad (4.15) \end{aligned}$$

where $E_t^*[.]$ is the risk neutral expectation and Re is the real part of a complex number and $f(iu)$ is the characteristic function of the log spot price.

4.3.2. Option Pricing Error Types

Model option pricing performance is generally measured with the distance of the model estimates from the market option prices. Most popular metrics are Relative Pricing Error (RPE), Absolute Relative Pricing Error (ARPE) and Absolute Pricing Error (APE).

$$RPE = \frac{(\hat{P} - P)}{P} \qquad (4.16)$$

$$ARPE = \left|\frac{\hat{P} - P}{P}\right| \qquad (4.17)$$

$$APE = |\hat{P} - P| \qquad (4.18)$$

where $\hat{P}$ is the model price estimate and P is the market price of the option contract.

RPE can be considered as the 'bias' in the price estimates. Since it includes a sign, it is an indicator of overpricing or underpricing. ARPE measures the magnitude of error. When aggregating results, ARPE yields more accurate results on the aggregate

[12] It has one or two parameters to optimize, depending on the inference method.

error. APE measures the dollar error. APE can be easily used to compute SE and their aggregate counterparts MSE and RMSE.

Error types are known to react differently to different moneyness and maturity values. For instance, RPE and ARPE weigh more on OTM and short-term options but APE weighs more on ITM and long-term options.

One of the objectives of this study is to observe the behavior of cluster stability with different error types. Therefore, the same experiment will be repeated for all error types mentioned above.

4.3.3. Data and Option Pricing Model Calibration

Daily S&P 500 index prices and option contracts (SPX) will be used to test models. All contracts given in the time windows are within the maturity interval of 1 calendar week (7 days) and 1 calendar year (365 days). Allowed moneyness interval is between 0.9 and 1.1. There are several reasons for determining such limits. Any values outside these intervals have already unstable pricing estimates by different models due to being either too cheap or too expensive. Bakshi et al. (1997) explains the effects of contract values on pricing error types. Briefly, for cheap options relative errors are more sensitive and for expensive options absolute errors are more sensitive. There are also fewer contracts outside these regions, so clustering algorithms would have to include the extended regions which in turn would have yielded inefficient results.

All tests are conducted with the out-of-sample pricing estimates of both Black-Scholes and Heston-Nandi GARCH models. Parameter inference is done via using historical log-returns of S&P 500 index of the last 2 and 5 years respectively. For each trading day, the inference window is moved by one trading day and parameter inference is repeated with the updated information. For both models, dividends are assumed to be continuous, and risk-free rates are calculated using interpolations of US Treasuries with different maturity terms.

The reason why these models and their respective parametrizations are chosen is a practical one. Black-Scholes model, however widely used, is known to frequently misprice ITM and OTM options due to its assumption of a low tails risk neutral density. Constant volatility assumption is also another known drawback of the model. On the other hand, Heston-Nandi GARCH model's multi-parameter setting with varying volatility and extended parameter inference period makes it a better alternative for pricing options. It is also shown by Heston and Nandi (2000) that Heston-Nandi GARCH model generally performs better than Black-Scholes in terms of lower pricing errors.

It can also be assumed that if a pricing model has greater errors in pricing, it will affect cluster stability. Its individual results, therefore overall results will be more inconsistent. Therefore both Black-Scholes and Heston Nandi GARCH models are tested to observe the behavior of clustering algorithm performances with models of varying consistency.

4.3.4. Assessment and Cross Validation

As mentioned earlier, pricing error and cluster stability are different kinds of errors. Cluster stability measures the difference between the actual pricing error and cluster estimate for each contract. Cluster stability is measured by Mean Absolute Percentage Error (MAPE) metric. MAPE is suitable metric to observe relative improvement as we are aiming to minimize the divergence from the past error estimate in order to achieve better cluster stability.

$$MAPE = \frac{1}{N}\sum_{i}^{N}\left|\frac{\epsilon_i - \hat{\epsilon}_i}{\epsilon_i}\right| \tag{4.19}$$

where ϵ_i is the realized pricing error (i.e. ARPE, RPE or APE) of the contract i and $\hat{\epsilon}_i$ is the data mining model's error estimate. Since pricing errors can be very small, even zero, due to the pricing model's performance; to avoid large MAPE deviations minimum value of the denominator of the MAPE is fixed to 0.01.

For a given time period, pricing error data used in experiment is divided into two groups; namely Training and Prediction. 6 months data is used for training and the following month is labeled as prediction data.

To calibrate clustering algorithms and group estimates only the training data is used. Then for each contract, pricing error estimates ($\hat{\epsilon}_i$) and resulting MAPE values are calculated. Aggregate training and prediction MAPE values are reported separately as in-sample and out-of-sample estimation. The difference between prediction MAPE and training MAPE is that, no contract in prediction data set is used to calibrate clustering models.

This procedure is repeated for 10 different time periods between 2009 and 2013, 2 time periods for each year. There are some advantages of this methodology to the classical 70%/30% assessment. First, prediction data set is relatively shorter (approximately 85/15), so immediate accuracy is measured. Repeating the experiment over different time periods, in a sense performing time series cross validation, gives the advantage of making the prediction results more robust. One of the main objectives of this study is to see if clustering algorithms' prediction power fluctuates when different time periods were considered. Aggregate results are reported so a clustering algorithm can be chosen based on overall results.

4.3.5. Results

When MAPE results are examined by pricing model, contract type, error type and in-sample/out-of-sample estimations, each analysis bears varying and interesting results in itself. All eight tables in this section provide different results but one conclusion. CIT is the best overall model.

Data sets are enumerated in a chronological order. Each year contains two data sets[13] (DS). For instance, DS 1 and 2 contain data from 2009, and DS 9 and 10 contain

[13] A data set might include some of either training or prediction data from the previous or next year. 'Contains' means the starting date of the prediction data belongs to that year.

data from 2013.

Tables 4.1 and 4.2 display the best algorithms for each data set, option type (call or put) and error type with respect to in-sample MAPE performance.

Table 4.1. Algorithms with the best in-sample MAPE (HN-GARCH).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
DS 1	CIT	CIT	K-Means	CIT	CIT	K-Means
DS 2	SVM	SVM	K-Means	CIT	CIT	SVM
DS 3	SVM	SVM	K-Means	CIT	CIT	CIT
DS 4	SVM	SVM	SVM	CIT	CIT	SVM
DS 5	CIT	CIT	CIT	CIT	SVM	CIT
DS 6	SVM	SVM	K-Means	CIT	SVM	SVM
DS 7	SVM	SVM	K-Means	CIT	DT	K-Means
DS 8	CIT	CIT	CIT	SVM	CIT	CIT
DS 9	CIT	CIT	CIT	SVM	CIT	K-Means
DS 10	CIT	CIT	CIT	CIT	DT	CIT

For HN in-sample results; CIT has the highest frequency in terms of best performing model. SVM comes the second with significant presence in Call ARPE and RPE. K-Means' best performing error types are APE in both Call and Put options. The outcome is similar for BS model in-sample results. CIT has increased frequency especially in Call APE.

For HN out-of-sample results are given in Table 4.3. CIT, again, has the highest frequency in terms of the best performing model in different data sets and error types. K-Means has increased presence in out-of-sample results than in-sample. Interestingly, static clustering methods became the best methods for Put RPE type errors in 6 data sets and Put ARPE type errors in 3 data sets.

Table 4.2. Algorithms with the best in-sample MAPE (BS).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
DS 1	SVM	SVM	K-Means	CIT	CIT	SVM
DS 2	CIT	CIT	CIT	CIT	CIT	CIT
DS 3	CIT	CIT	CIT	CIT	CIT	CIT
DS 4	CIT	CIT	CIT	SVM	SVM	CIT
DS 5	CIT	CIT	K-Means	CIT	SVM	SVM
DS 6	SVM	SVM	SVM	CIT	CIT	SVM
DS 7	SVM	SVM	CIT	CIT	DT	CIT
DS 8	CIT	CIT	CIT	CIT	SVM	K-Means
DS 9	K-Means	K-Means	CIT	SVM	SVM	CIT
DS 10	CIT	CIT	CIT	CIT	CIT	K-Means

Table 4.3. Algorithms with the best out-of-sample MAPE (HN-GARCH).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
DS 1	CIT	CIT	K-Means	CIT	CIT	K-Means
DS 2	Static 1	DT	SVM	K-Means	K-Means	K-Means
DS 3	SVM	DT	SVM	CIT	Static 1	CIT
DS 4	CIT	CIT	SVM	CIT	CIT	SVM
DS 5	K-Means	K-Means	CIT	CIT	Static 2	K-Means
DS 6	SVM	SVM	K-Means	CIT	Static 1	SVM
DS 7	K-Means	K-Means	K-Means	Static 1	Static 2	SVM
DS 8	CIT	CIT	K-Means	Static 1	Static 2	CIT
DS 9	CIT	CIT	CIT	Static 2	Static 2	CIT
DS 10	CIT	CIT	CIT	CIT	DT	CIT

Table 4.4 quantifies the improvement of the best data mining algorithm over the best static clustering method in terms of MAPE difference. The highest improvement is in DS6 Call ARPE by SVM with %72.15 better MAPE compared to the best performing static clustering method. For some data set error type comparisons, the introduction of data mining algorithms is detrimental to the performance (highest decrease is %21.38 in DS8 Put ARPE), but in overall improvements brought by data mining algorithms is significantly larger than the poor performance in some data set error type combinations.

Table 4.4. MAPE improvement of the best performing data mining algorithm over the best performing Static method (HN-GARCH) (%).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
DS1	27.86	27.88	46.36	20.24	20.45	40.96
DS2	-0.92	7.17	19.5	11.25	11.14	43.71
DS3	12.29	19.31	36.87	8	-11.05	2.14
DS4	31.93	44.24	42	6.98	8.2	15.95
DS5	21.41	22.52	28.91	10.57	-5.45	26.32
DS6	72.15	66.42	28.91	0.51	-5.85	18.03
DS7	21.98	18.22	32.87	-3.57	-2.19	25.32
DS8	42.33	42.4	40.96	-21.38	-14.62	24.2
DS9	21.89	23.72	25.02	-6.41	-7.61	58.73
DS10	19.06	18.2	31.78	7.34	2.39	63.54

The results are in favor of data mining algorithms regarding the out-of-sample performance for Black-Scholes model. Table 4.5 shows slightly fewer CIT frequency and presence of some static clusters (albeit few in number). Presence of static clustering methods is lower than HN out-of-sample results, but still they are in Put ARPE and Put RPE. K-Means has significantly decreased presence in out-of-sample results and performs worse in BS than HN.

Table 4.5. Algorithms with the best out-of-sample MAPE (BS).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
DS 1	CIT	Static 1	K-Means	CIT	Static 1	SVM
DS 2	CIT	CIT	CIT	CIT	CIT	CIT
DS 3	CIT	CIT	CIT	CIT	CIT	CIT
DS 4	CIT	CIT	CIT	CIT	CIT	CIT
DS 5	SVM	SVM	CIT	CIT	SVM	SVM
DS 6	SVM	SVM	K-Means	CIT	SVM	DT
DS 7	SVM	CIT	CIT	Static 1	Static 1	SVM
DS 8	CIT	CIT	CIT	CIT	DT	CIT
DS 9	CIT	K-Means	CIT	Static 2	Static 2	CIT
DS 10	CIT	CIT	CIT	SVM	SVM	SVM

Table 4.6 displays the performance improvement of the best data mining algorithm to the best static clustering method for out-of-sample BS results. Compared to HN, gains of data mining algorithms are higher (tops at 76.74% for DS3 Call APE by CIT) and failures are relatively modest (bottoms at -10.47% for DS9 Put ARPE).

Aggregate tables are more suitable for robust observations, while above tables provide more granular insights. With the same order of above tables, Table 4.7 shows CIT has the best overall performance for all error types. SVM follows CIT in Call ARPE and RPE and K-Means follows CIT in Call and Put APE but for Put ARPE and RPE second best methods are Static 1 and DT respectively.

BS in-sample overall results can be seen in Table 4.8. CIT performs significantly well in all error types except Put RPE, where SVM has the best overall performance. Contrary to HN in-sample performance, SVM is worse in Call type errors and better in Put type errors. K-Means on the other hand performs well in all Call type errors and Put APE. But, K-Means performs the worst in Put ARPE and Put RPE.

Table 4.6. MAPE improvement of the best performing data mining algorithm over the best performing Static method (BS) (%).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
DS1	0.39	-0.04	16.09	1.18	-0.29	21.29
DS2	6.05	5.58	55.67	4.4	3.88	64.33
DS3	23.28	23.13	76.74	8.38	8.43	67.51
DS4	42.17	41.89	62.74	20.27	18.61	55.35
DS5	43.17	39.18	21.7	4.13	12.13	32.05
DS6	65.26	53.07	56.73	2.13	1.96	3.52
DS7	24.77	23.37	16.79	-1.88	-7.17	34.8
DS8	48.66	49.26	64.38	10.44	6.4	6.24
DS9	12.38	12.9	39.71	-10.47	-1.38	25.22
DS10	37.44	34.77	42.94	11.62	6.66	29.97

Table 4.7. Overall in-sample MAPE of algorithms (HN).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
K-Means	2.54	2.67	1.75	1.38	1.53	2.35
SVM	2.46	2.41	2.12	1.43	1.33	3.08
CIT	**2.24**	**2.18**	**1.74**	**1.28**	**1.15**	**2.30**
DT	2.96	2.80	2.36	1.39	1.22	2.87
Static 1	2.68	2.57	2.74	1.37	1.24	3.05
Static 2	2.76	2.64	2.76	1.38	1.24	3.09

Overall out-of-sample results are the most important, since they determine the prediction consistency of displayed errors in result tables. HN model results are displayed in Table 4.9. CIT is notably fares the worst in Call ARPE and RPE type errors, where SVM excels by a far margin. CIT is the best overall model in both Call and Put

Table 4.8. Overall in-sample MAPE of algorithms (BS).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
K-Means	1.81	1.87	1.97	1.62	1.76	2.51
SVM	2.61	2.61	2.54	1.45	**1.27**	2.73
CIT	**1.59**	**1.58**	**1.92**	**1.40**	1.32	**2.40**
DT	2.56	2.42	2.67	1.53	1.38	2.75
Static 1	2.11	2.06	3.06	1.55	1.40	2.94
Static 2	2.17	2.12	3.27	1.55	1.39	2.96

APE. It is in tie with Static 1 method for Put ARPE and Static 2 is the best performing model for Put RPE type errors. Results are consistent with in-sample estimations but the magnitude of errors differs.

Table 4.9. Overall out-of-sample MAPE of algorithms (HN).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
K-Means	3.39	3.32	2.26	1.56	1.77	2.90
SVM	**2.97**	**3.03**	2.65	1.85	1.89	3.42
CIT	3.51	3.60	**2.16**	**1.52**	1.48	**2.87**
DT	4.01	3.89	2.78	1.57	1.48	3.45
Static 1	3.67	3.70	3.17	**1.52**	1.42	3.60
Static 2	3.77	3.79	3.08	1.53	**1.41**	3.69

Finally, overall out-of-sample MAPE results of BS model is given in Table 4.10. CIT is the overall best method for Call ARPE and RPE and Put ARPE. K-means is the best overall method in Call APE, SVM is the best overall method for Put APE and Static 2 is the best overall model for Put RPE. Even though CIT is not dominant, it is either the second best model or fairly close to the winner in terms of overall performance.

Table 4.10. Overall out-of-sample MAPE of algorithms (BS).

	Call			Put		
	ARPE	RPE	APE	ARPE	RPE	APE
K-Means	3.54	3.65	**3.69**	1.91	2.21	3.50
SVM	3.66	3.84	4.81	1.87	1.82	**3.30**
CIT	**3.14**	**3.24**	3.80	**1.72**	1.78	3.42
DT	3.86	4.06	4.50	1.81	1.79	3.62
Static 1	3.79	3.88	5.11	1.77	1.77	3.93
Static 2	3.79	3.87	5.14	1.78	**1.76**	3.98

Regarding the overall and individual data set results, several observations can be made.

Observation 1. CIT performs significantly better than other methods. CIT is the overall best model in HN and BS for both in-sample and out-of-sample estimates. For individual data sets and error types, CIT has a general good performance, closely following the best performing algorithm in terms of MAPE when it is not the best model. Its performance is not spotless, especially for Put RPE errors. CIT suffers from MAPE outliers more than other models, it is one of the main reasons why it failed to deliver for some error types.

What makes CIT almost superior to DT is employing statistical significance criteria when choosing the split covariates and clusters. This way errors following a pattern can be clustered more consistently. And it seems to be the case for pricing errors.

Observation 2. When models are compared, BS MAPE values are generally higher than HN MAPE values for all clustering methods used, including . This is not unexpected since HN is reported to be a better model in terms of minimizing pricing errors than BS by Heston and Nandi (2000).

Observation 3. K-Means performs well in APE type errors regardless of the contract type for HN both in-sample and out-of-sample. Though, it is not performing that well in BS model assessments, but again most of the presence is in APE type errors.

Observation 4. SVM is the runner up of the algorithms covered in this study, in terms of highest frequency in different data sets, contract types and error types. Its overall out-of-sample performance in HN displays as the best algorithm for Call ARPE and RPE type errors. However, the same cannot be said about BS results. For BS out-of-sample results, SVM has the best overall performance for Put APE type errors. Its performance in other error types, especially call error types, is poor by comparison.

Observation 5. Static clustering methods perform quite poor in except several error types and data sets. Though as an exception, presence of static clustering methods in out-of-sample Put ARPE and RPE results of HN model is significant. This significance presents itself in overall results too. Though, the marginal MAPE benefit brought by static clustering methods is modest compared to data mining algorithms' in all other contract and error types for both models.

As Static 2 is more granular than Static 1, it yields slightly better results than Static 1. But this might not always be the rule since too many clusters might end up in overfitting for the method.

4.4. Conclusion

Aggregate pricing error tables in empirical option pricing studies are usually a snapshot of the time period which the experiment is conducted. Whether those aggregate results represent the future states of the same partitions is rarely examined. In addition, arbitrary grouping of contracts bears the risk of hindering true performance of models; every model has different strong and weak regions in terms of pricing error.

Four different algorithms are proposed for better clustering and error prediction: K-Means, Support Vector Machine (SVM), Decision Trees (DT) and Conditional Inference Trees (CIT). These algorithms and two static clustering methods previously used in the literature are tested and benchmarked in terms of cluster stability. MAPE metric is used to calculate relative deviance in pricing errors in each cluster.

10 different data sets from S&P 500 option contracts (SPX) between 2009 and 2013 are used in the experiments. Call and put options are evaluated separately. Each data set is priced with two different models: Black-Scholes and Heston-Nandi GARCH. Three different pricing errors are used: RPE, ARPE and APE. Both in-sample and out-of-sample performances are investigated.

Each data set's best performing algorithm or method is reported for each error type, option type combination. Even though results are shown to change with factors such as time, model and error type (and even contract type); CIT is shown to be a good general model for evaluating and clustering pricing errors.

Using data mining algorithms is a good idea in creating robust predictions about future states of the pricing errors as well as it might be a good idea to select models. If a model can predict future error states of each contract, it is also possible to compare strong and weak regions for each pricing model and choose the better pricing model for each contract. Ultimately, it will lead to better pricing performance.

5. BAD MODEL PROBLEM: HOW TO MEASURE MISSPECIFICATION FOR OPTIONS?

Choice of the performance metric is a crucial part of empirical assessment of option pricing model performance and benchmarking, mainly because the outcome is subject to change based on the metric. Usage of performance metrics in the empirical option pricing literature are examined and their agreement is assessed. We also propose new metrics based on market efficiency tests for more robust assessment and to eliminate the confusion of metric choice. We claim efficiency metrics are the better choice if the objective is to evaluate option models with their financial outcome, not just price fitting.

5.1. Introduction

What is the determinant factor for a model to be successful? There are certainly many answers to this question. But, even with the same objective, same experiment and same data; the outcome may vary due to the performance metric.

For instance, consider two models; namely A and B. Also consider two performance metrics; namely X and Y. If model A is deemed better than model B according to performance metric X and the opposite happens according to performance metric Y, then the choice of the performance metric affects the assessment of the model performance. Lehar et al. (2002) can be shown as an example to this problem. They report both relative pricing error (RPE) and absolute-RPE (ARPE) values for different models and there are cases where ARPE of a model is better than the others, but for the same cases RPE of another model is better than that model in the same experiment.

Although there is no identified ‘best’ way of determining the correct performance metric there are some possible solutions. Performance metric, also commonly named as loss function, is also used in model parameter inference. Parameter inference is generally done via minimization of a loss function and a parameter set yielding the minimal loss value is deemed ‘optimal’ or ‘best’. Christoffersen and Jacobs (2004a) claim that keeping parameter inference (in-sample estimation) loss function of the model and performance assessment (out-of-sample assessment) loss function the same would increase the model’s performance on that loss function’s assessment significantly and that model will generally do better in benchmarks compared to other models which do not use the same loss function in their parameter inference process. Their claim is backed with empirical testing of different models and loss functions, with result of the best model changing with the loss function. This seemingly obvious recommendation is largely ignored by many other empirical studies as stated in the study. They also report that dollar mean squared error ($MSE) might be an appropriate metric because it yields the overall best results. Bams et al. (2009) also agree with these results and propose a framework for the selection of the appropriate loss function via the distributions of considered loss functions.

Many methods and models are proposed to price options over the last several decades. For a general overview of the field, see the studies of Bates (2003), Broadie and Detemple (2004) and Hull (2012). Nevertheless, most empirical option pricing studies have the objective of getting their model estimates as close to market prices as possible. The main motivation is to come up with a hedge using the same parameters of the relevant model. Only spot option prices are relevant to the assessment of these kind of studies. So, performance metrics based on this type of studies measure the distance from the market prices. Examples of such studies are Bakshi et al. (1997), Heston and Nandi (2000), Christoffersen and Jacobs (2004b) and Christoffersen et al. (2009).

There is also another branch of studies testing market efficiency with option pricing models. Their concern is to analyze the difference between estimated prices and hedges and actual market prices. The main objective of the assessment is to outperform the market in terms of risk adjusted profitability. Tests include dynamic positions in both options and underlying markets, so both asset and option price evolution data are required. Usually a profit-loss metric is considered and unlike hedging oriented studies, the outcome is affected by the payouts but not by the spot distance from the market. Although efficiency tests are mainly for underlying markets, studies such as Black and Scholes (1972) and Galai (1977) perform efficiency tests on options markets.

The problem with market efficiency tests is due to joint hypothesis problem. Any error (excess profit) found in the market might be due to the model not considering all the risk sources. Studies measuring the distance from the market should have implicitly assumed markets are informationally efficient, as they measure model performance with the 'best' benchmark and associate all error with the model specification. Studies testing market efficiency implicitly assume that their model is well built for the purpose and any 'error' found in their experiments is against the efficiency in the market.

We propose a performance metric to measure option pricing errors by taking positions and calculating the outcome of those positions in terms of profit and loss, similar to Galai (1977) did to test market efficiency with ex-post and ex-ante delta strategies. Differently, our aim is not to test market efficiency but to test the model performance. It can be made plausible by assuming EMH perfectly holds and all error belongs to model misspecification. The objective of such a performance measure would be to keep risk adjusted profit/loss as close to zero as possible since EMH does not allow for excess profits. Though, we also keep in mind that, such performance metric can be used to find 'profitable' models with little change but that is currently beyond our scope since this is not a trading exercise but a model assessment.

Our main contributions are two fold: (1) we demonstrate that efficiency based metrics differ from pricing error metrics, and (2) we show that efficiency based metrics are a better choice if outcome of a contract is significant to the objective of the study. We then introduce efficiency based metrics based on the efficiency tests but with the objective of minimizing model error instead of testing market efficiency. Finally, we compare those metrics and make suggestions about how to use them in model assessment experiments. We also claim that if the financial output of the option pricing models is of more interest to the experiment, efficiency based metrics are better alternatives than measuring the fit of model prices to the market prices.

This paper is organized as follows. Section 5.2 elaborates on efficent markets hypothesis, joint hypothesis problem and their relevance to the performance metric. Section 5.3 lists commonly used pricing error metrics and makes comparisons within those metrics. Section 5.4 elaborates on efficiency based metrics and how they can be used as more robust model error metrics than basic pricing error metrics. Comparisons are made both within efficiency metrics and between efficiency and pricing error metrics. Section 5.5 presents benchmarking of two models to display the differences between pricing and efficiency errors. Finally, Section 5.6 summarizes the findings and provides explanations on how they can be used in model assessment and selection.

5.2. Efficient Markets Hypothesis and Joint Hypothesis Problem

In his seminal paper on capital markets efficiency, Fama (1970) defines market efficiency as prices ‘fully reflect’ available information. In other words, market prices are the best estimates on the fair value of the assets and it is impossible to get consistent excess risk-adjusted returns. This claim is also called Efficient Market Hypothesis (EMH).

Efficiency in capital markets is categorized under three forms. Weak form of efficiency considers only information about the past prices. It is also the most popular one to test among the researchers. The common conclusion is; even though it is possible make small profits by making frequent trades, but not significant enough considering transaction costs.

Semi strong form of efficiency is about market’s adaption to all available public information such as dividends and stock splits. This form is also reported to support efficient markets hypothesis.

Strong form of efficiency is about information asymmetry, assuming one party has exclusive access to some information which can be used in profit making. Specialists (or market makers) and corporate insiders are examples of this kind of exclusivity but any private information can be categorized under strong market efficiency.

One of the rather unnerving results of the efficient markets is, fluctuations become completely random; therefore 'unforecastable'. In other words, there is no model or any other way to make consistent profits in efficient markets. One conjecture of such state of efficiency would also be the loss of incentive to be active in the market due to loss of prospective profits. Grossman and Stiglitz (1980) formalize this state in their study on their market model with informed and uninformed participants.

As in many other great contributions to the financial literature, from its formal definition by Samuelson (1965), efficient markets hypothesis sparked a debate about whether markets are efficient or not, which is still ongoing. Ever since, many studies dealt with EMH in terms of both theoretical and empirical aspects. Jensen (1978), Fama (1991), Fama (1998), Lo (2007) and Alajbeg et al. (2012) provide a good history of the evolution of proponents and opponents of EMH.

One of the most significant developments in EMH discussions is the "joint hypothesis problem". After Fama (1970) the best way to support or reject EMH was thought to be testing it with data. If a model or trading strategy could provide excess risk adjusted returns, then EMH is rejected by that model. Else, it is documented as another supporting evidence of EMH. But later Fama (1991) and Fama (1998) state in order to reject EMH, one should use an asset equilibrium model. Any excess returns found by the model can be due to market inefficiency as well as the model's inability to account for some other risks in the market. Since making the distinction is not a trivial task, EMH became harder to 'falsify'. Joint hypothesis problem is also called "bad model problem". Further information about joint hypothesis problem can be found in the works of Jensen (1978), Fama (1991) and Fama (1998).

On the options side, theoretical developments gained pace after the seminal works of Black and Scholes (1973) and Merton (1973). Black-Scholes formula laid out a self-financing portfolio in a simple yet elegant way. As Mackenzie (2008) report theoretical work was built on Black-Scholes formula, not the other way around. No arbitrage, risk neutral measures, complete markets and Funadamental Theorems of Asset Pricing, all came later. See Shreve (2004) for detailed explanations of these concepts.

Option markets are no exception to market efficiency tests. Both Black and Scholes (1972) and Galai (1977) performed empirical tests using Black-Scholes model to check for market efficiency. Their conclusions are similar. Even though there is profit opportunity, the transaction costs wipe out most of the excess profits. Both Galai (1978) and Bhattacharya (1983) examine market efficiency boundary conditions of Merton (1973) with spread tests and come to similar conclusions that any profits yielded by models are too small compared to transaction costs. Not all studies confirm those results. Coval and Shumway (2001) reject the claim of risk free returns with delta neutral positions by testing with straddle positions. Ahmad and Wilmott (2005) examine profiting opportunities while hedging when volatility estimates of the market and the individual differ. Constantinides et al. (2007) claim post-1987 crash show stochastic dominance and boundary violations in OTM call options.

Although there were empirical studies, for a long time there were no studies explicitly connecting EMH to option pricing theory. Jarrow and Larsson (2012) study this problem and provide theoretical framework for the EMH using no arbitrage and no domination. Jarrow (2012) makes the connection with derivatives markets and propose the Third Fundamental Theorem of Asset Pricing. Jarrow (2013) discusses the implications of EMH in option pricing. The cited studies claim, in order to reject market efficiency without falling to joint hypothesis problem, one must seek arbitrage opportunities or asset dominance. This in turn would affect many option pricing model assumptions such as no arbitrage and risk neutral valuation.

Option pricing model performances are usually measured with their distance from the market prices. It is a reasonable method given the EMH and its connection with risk neutral valuation relation. If a pricing model can represent the market correctly, then it can be used to construct a ‘perfect hedge’, if applicable, and a self replicating portfolio.

From the efficiency perspective, it is not the distance but the consequences of the actions taken by the trading that defines the performance of a model. That is especially important if positions on contracts have asymmetric payoffs. Two models might have the same distance to the market price from opposite signs, but the magnitude of the outcome, most probably, will not be similar.

5.3. Pricing Error Metrics

Empirical option pricing literature is quite rich with various performance metrics. But at least one of the following fundamental metrics, on a relative or absolute scale, are used in almost every study. Those metrics, also used in this study, are Relative Pricing Error (RPE), Absolute Relative Pricing Error (ARPE) and Squared Pricing Error (SE). In the literature, their applications measure only the distance between the market price and the model estimate for a contract. Measuring pricing error in an absolute scale or relative scale, and absolute values or with positive/negative signs have different impacts on the assessment of performance.

The rest of this section is about detailed examination of these performance metrics and their comparative results using the same empirical data set. Our data set consists of daily S&P 500 option contracts (SPX) of 2011. There are a total of 47,715 contracts in our data set, 21,167 calls and 26,548 puts. Maturity values range between 7 calendar days and 1 year, and moneyness[14] values range betweeen 0.5 and 1.5. Risk-free interest rate is calculated from US Treasury yields approximated to the maturity values for each contract. Dividend yield is taken as continuous. Closing prices are used in both option contracts and underlying asset prices. Contract prices range from \$0.05 to \$1239.45.

The option model used to get price and delta estimates in the numerical experiments is the Heston Nandi GARCH model of Heston and Nandi (2000) with continuous dividends[15] . Parameters are esimated using log-returns of the underlying of the past 5 years for each contract.

[14] Moneyness is defined as the ratio between spot price and the strike price (S_0/K).

[15] See Hull (2012) and Stentoft (2011) for implementation.

5.3.1. Relative Scale Pricing Errors

Relative Pricing Error (RPE) measures the underpricing or overpricing of model estimates relative to its market price. $\hat{P}$ denotes the model price estimate and P denotes the market price.

$$RPE = \frac{\hat{P} - P}{P} \tag{5.1}$$

$$ARPE = \frac{|\hat{P} - P|}{P} \tag{5.2}$$

Measuring the bias with RPE is not without problems. It is generally useful information to know whether a model is biased towards a side. Unfortunately, magnitude of error is usually underrepresented when aggregated, depending on how evenly RPE is distributed to negative and positive values. The magnitude of aggregate values can be measured better with ARPE. Another issue is, since price cannot take values less than zero, the most negative an RPE value can take is -1. Depending on the market price and model estimate the positive side can be much larger.

Figures 5.1 and 5.2 display the RPE values of call and put options respectively with axis values of moneyness and maturity in days. Error magnitudes are illustrated with the shades of the data points (darker points are contracts with larger error) and negative and positive errors are given equal shading as if they are on an absolute scale. Due to the large number of contracts, for better visibility, random sample of 500 contracts are displayed for each plot.[16]

The main difference between call and put options is the extreme overpricing in call options. This is mainly due to the price difference in cheap options such as out-of-money options and/or short term options. Due to the low market price, the denominator in RPE is small and even a dollar difference seems huge. For instance, if the contract price is \$0.05 and model estimate is \$1.05, RPE becomes 20 and skews the error metric. In Figure 5.1, majority of severe errors are clustered between moneyness

[16] 11 points are removed from display in Figure 5.1 for being very few and on distant values of moneyness, so they would not squeeze the graph. They are not removed from calculations however.

values of 0.9 and 1.0 (slightly OTM) and they are overpriced by the model.

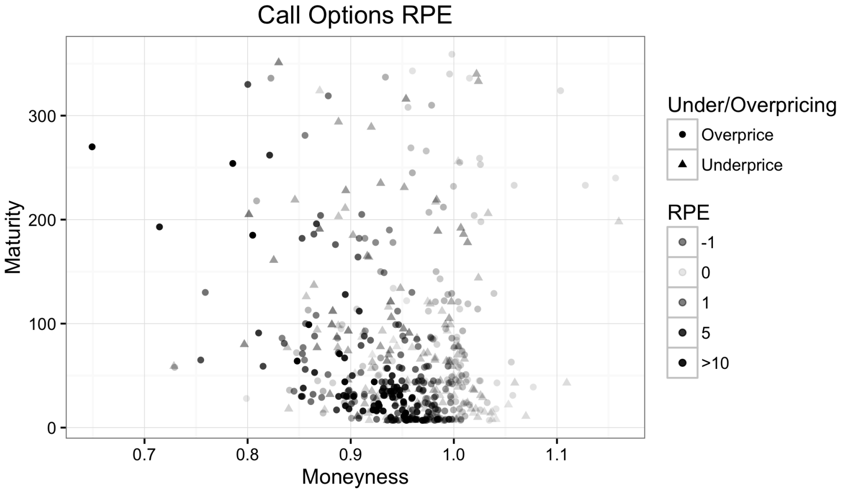

Figure 5.1. RPE values of Call options.

RPE errors of put options shown in Figure 5.2, are restricted in a more reasonable RPE interval. Majority of the 'high' errors start from low maturity and increasing OTM but they are not restricted to low maturity as the moneyness increases.

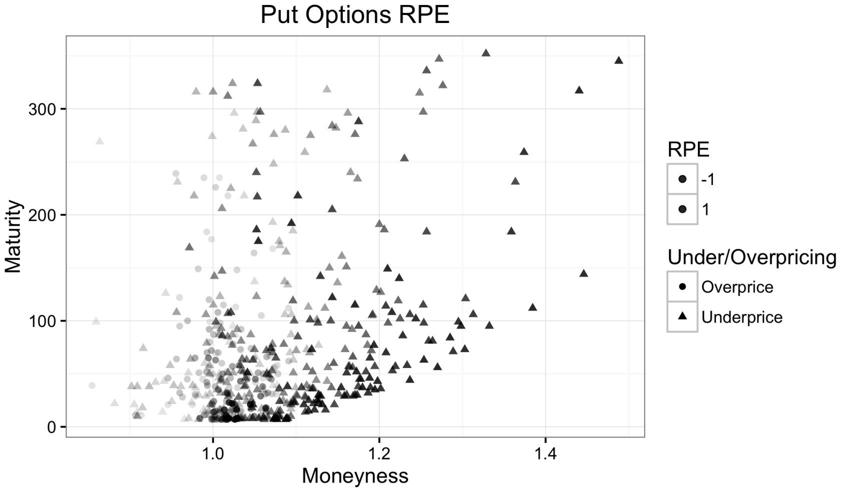

Figure 5.2. RPE values of Put options.

5.3.2. Absolute Scale Pricing Errors

Squared Error (SE), or in aggregate terms Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), measures model estimate distance from market prices in dollar terms. Absolute Pricing Error (APE) is also used but not as popular as the MSE.

Nevertheless, APE will be used as this section's representative error, mainly because plots are represented per contract and it has a better scale perception than SE. Exponential increase in SE would cause the maximum error to be too large and mid error values would be dwarfed in illustrations. Also, APE can easily be interpreted to RMSE through transformation.

$$APE = |\hat{P}_t - P_t| \tag{5.3}$$

$$SE = (\hat{P}_t - P_t)^2 \tag{5.4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{P_{i,t}} - P_{i,t})^2} \tag{5.5}$$

Main disadvantage of using the squared or absolute error is the opposite of relative error's: Expensive options such as in the money and long term options are more prone to large error values than errors in cheap options.

Despite the intuition that relative pricing errors are more informative in explaining option pricing errors, Christoffersen and Jacobs (2004a) claim RMSE is the best candidate metric if error function is not specifically aligned with the parameter inference loss function. Their results are also backed up by Bams et al. (2009).

As mentioned above, absolute scale error values are generally more biased towards expensive options. The more an option is ITM and long term, the higher its price will be and so the absolute difference between market price and model estimate.

Figures 5.3 and 5.4 illustrate absolute pricing errors of call and put prices, respectively, dispersed among moneyness and maturity values.[17] The largest APE values belonging to call options are around the location of ATM and 300 days of maturity.

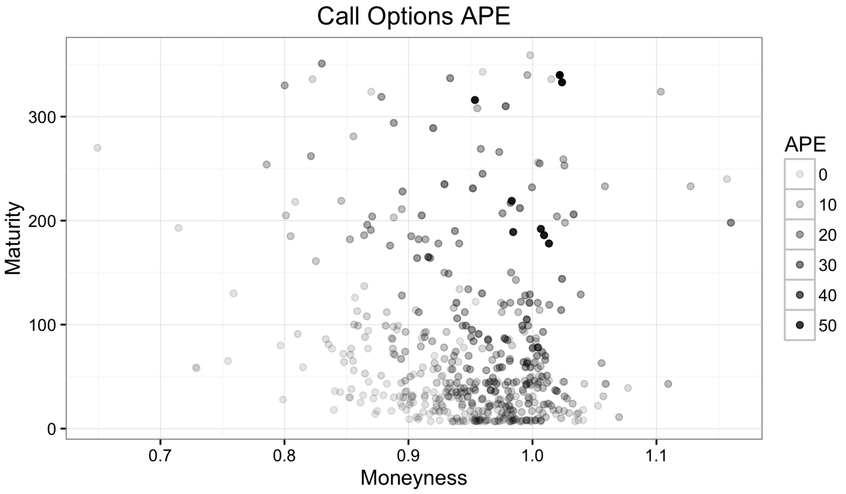

Figure 5.3. APE values of Call options.

For put options, larger errors are dispersed on a wider region of moneyness and maturity but still around the same location of ATM to OTM values and high maturity.

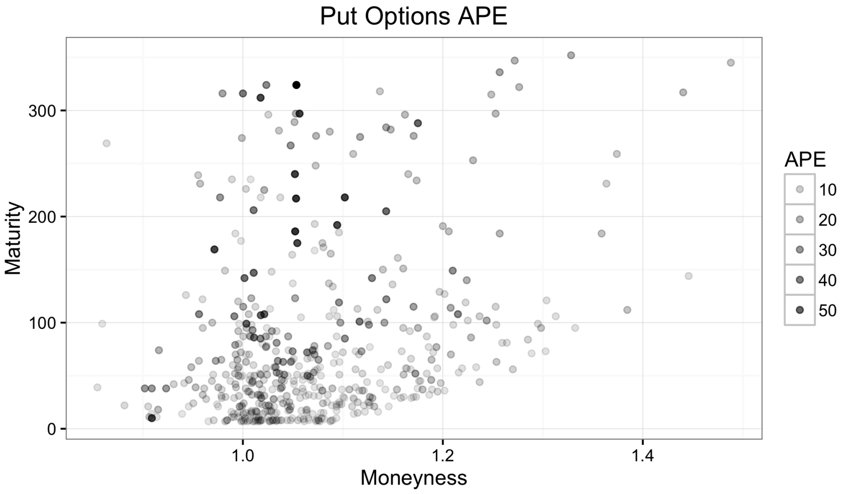

Figure 5.4. APE values of Put options.

[17] Same sample of the data set with the RPE is used for illustration, so side by side visual comparison becomes possible.

5.3.3. Relative vs Absolute Errors

When compared to the relative error with the same set of data, Figures 5.5 and 5.6 indicate important differences.[18] On the left side of the Figure 5.5, a group of contracts have a high APE amount while their ARPE is still low. It also happens, with low APE high ARPE this time, at the bottom right of the graph. Those contracts where APE and ARPE disagree are 'cheap' and 'expensive' contracts mentioned in the previous section. It is commonly observed and reported in other studies also. For instance, Bakshi et al. (1997) prepare different tables for APE and ARPE errors with similar explanation.

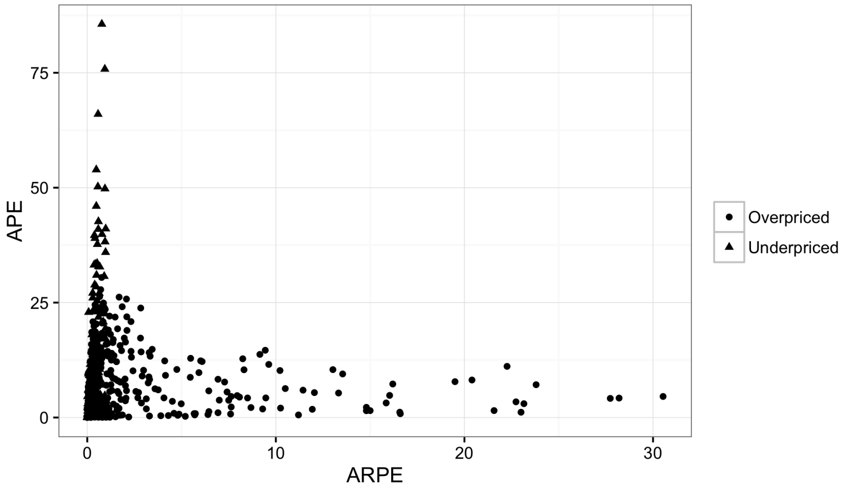

Figure 5.5. Comparison of ARPE and APE type errors with Call options.

For the put options, even though there is increasing APE values with increasing ARPE, the points rather disperse than making a clear pattern.

Aggregate graphs can help us see how error metrics behave and disagree by showing distinctive patterns. Even a single contract can show us how error metrics differ with the difference between model estimate and the market price. Table 5.1 is an example contract of how the pricing errors changed during the life of the contract. It is a put option, with strike $1280 and expiration is within almost a month of the first data

[18] Figures below make use of only APE and ARPE values, despite RPE values are also reported on previous sections. But it is easy to interpret ARPE values to RPE values since shapes denoted as Overpriced/Underpriced imply positive/negative signs of RPE.

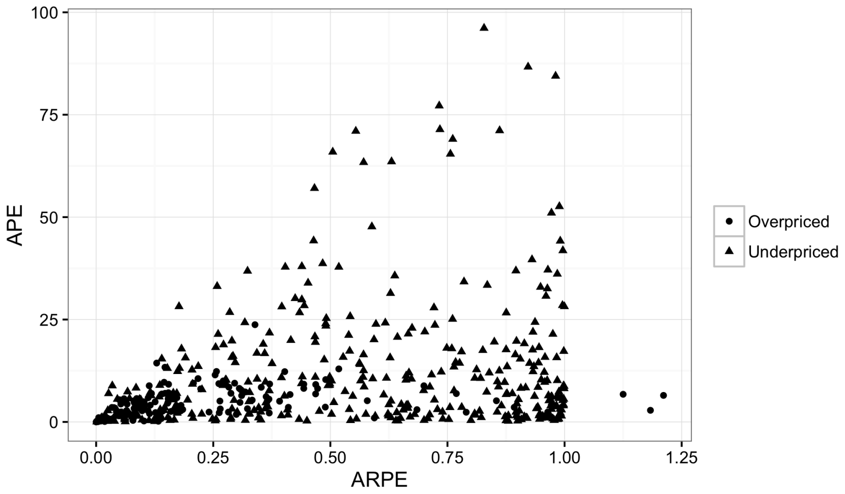

Figure 5.6. Comparison of ARPE and APE type errors with Put options.

date (2011-11-19). The underlying was $1254.19 at its first day.

Table 5.1. Pricing errors on example contract.

	Date (Y/m/d)	Moneyness	Market Price	Model Price	RPE	ARPE	APE
1	2011-10-24	0.98	45.77	45.71	-0.00	0.00	0.06
2	2011-10-27	1.00	26.00	26.27	0.01	0.01	0.27
3	2011-10-28	1.00	25.50	25.10	-0.02	0.02	0.40
4	2011-10-31	0.98	36.90	42.38	0.15	0.15	5.48
5	2011-11-03	0.99	36.15	34.50	-0.05	0.05	1.65
6	2011-11-04	0.98	42.20	38.70	-0.08	0.08	3.50
7	2011-11-11	0.99	26.10	25.54	-0.02	0.02	0.56

Error estimates can vary in their assessments with the moneyness and maturity values as also shown in the above graphs. Tables 5.2 and 5.3 show the average values of each metric within predefined intervals. Five maturity intervals between 7, 30, 60, 180 days and 1 year are defined. Moneyness breaks are defined between 0.5, 0.9, 1, 1.1 and 1.5. Number of contracts in each interval is also given on the third column. Fourth and fifth columns are the average market and model prices.[19]

[19] Average RPE, ARPE and APE might not be calculated from reported average prices, because reported errors are averaged after applied to each contract (not to the averages of each interval).

Table 5.2. Pricing error averages of Call options based on moneyness and maturity intervals.

Maturity	Moneyness	Number of Contracts	Market Price	Model Price	RPE	ARPE	APE
[7,30]	(0.5,0.9]	718	0.34	0.52	2.22	2.61	0.37
[7,30]	(0.9,1]	4501	6.14	8.46	3.38	3.58	4.54
[7,30]	(1,1.1]	1165	38.74	39.56	0.09	0.21	6.21
[7,30]	(1.1,1.5]	19	154.10	147.55	-0.04	0.04	6.59
(30,60]	(0.5,0.9]	754	1.08	2.38	5.03	5.14	1.50
(30,60]	(0.9,1]	3537	12.13	18.22	2.96	3.04	8.41
(30,60]	(1,1.1]	950	50.17	52.45	0.12	0.23	9.34
(30,60]	(1.1,1.5]	23	147.64	139.89	-0.05	0.06	8.14
(60,90]	(0.5,0.9]	473	2.54	5.31	5.74	5.81	3.24
(60,90]	(0.9,1]	1593	21.74	28.47	1.32	1.41	11.00
(60,90]	(1,1.1]	599	58.64	60.78	0.09	0.21	10.92
(60,90]	(1.1,1.5]	18	164.01	160.41	-0.02	0.03	4.12
(90,180]	(0.5,0.9]	783	5.14	10.89	5.68	5.71	6.42
(90,180]	(0.9,1]	1741	30.22	39.25	0.79	0.88	13.37
(90,180]	(1,1.1]	579	74.12	76.27	0.06	0.18	12.09
(90,180]	(1.1,1.5]	34	179.95	173.73	-0.03	0.06	10.33
(180,365]	(0.5,0.9]	708	13.69	26.45	3.46	3.48	13.36
(180,365]	(0.9,1]	1261	54.86	67.41	0.38	0.43	16.86
(180,365]	(1,1.1]	443	101.04	106.67	0.07	0.15	13.88
(180,365]	(1.1,1.5]	35	193.17	192.93	0.00	0.05	9.06

Both call and put aggregate tables show different sections were error terms peak. While RPE and ARPE are expectedly closer, largest errors are with mid-term maturity and OTM regions on the call options and short to mid-term maturity and OTM regions on the put options. APE peaks on long term options for both call and put options, while moneyness is between ATM and ITM.

Table 5.3. Pricing error averages of Put options based on moneyness and maturity intervals.

Maturity	Moneyness	Number of Contracts	Market Price	Model Price	RPE	ARPE	APE
[7,30]	(0.5,0.9]	61	209.27	199.04	-0.04	0.04	10.25
[7,30]	(0.9,1]	1219	45.15	43.91	0.03	0.18	6.51
[7,30]	(1,1.1]	4192	9.99	7.66	-0.16	0.49	4.70
[7,30]	(1.1,1.5]	1874	2.95	0.25	-0.89	0.89	2.70
(30,60]	(0.5,0.9]	61	210.59	202.89	-0.04	0.04	8.11
(30,60]	(0.9,1]	858	54.84	54.79	0.05	0.18	8.86
(30,60]	(1,1.1]	3056	21.30	18.37	-0.07	0.32	7.29
(30,60]	(1.1,1.5]	2635	5.82	1.02	-0.81	0.81	4.80
(60,90]	(0.5,0.9]	51	234.49	227.98	-0.03	0.03	7.51
(60,90]	(0.9,1]	427	67.67	65.10	0.00	0.17	11.36
(60,90]	(1,1.1]	1657	33.19	27.98	-0.11	0.25	9.22
(60,90]	(1.1,1.5]	1615	9.90	2.39	-0.77	0.77	7.52
(90,180]	(0.5,0.9]	83	235.86	232.18	-0.01	0.03	7.18
(90,180]	(0.9,1]	493	83.12	77.94	-0.04	0.15	12.33
(90,180]	(1,1.1]	1670	46.49	38.84	-0.13	0.21	10.83
(90,180]	(1.1,1.5]	2046	17.58	5.74	-0.71	0.71	11.84
(180,365]	(0.5,0.9]	38	315.10	314.80	0.01	0.03	8.74
(180,365]	(0.9,1]	360	113.60	103.76	-0.07	0.12	13.84
(180,365]	(1,1.1]	1072	78.85	64.62	-0.17	0.18	15.06
(180,365]	(1.1,1.5]	1436	35.70	16.30	-0.59	0.59	19.40

5.4. Efficiency Centered Performance Metrics

So far, commonly used pricing performance metrics and their explanatory power are discussed. As explained in the introduction, pricing errors and market efficiency tests have different objectives.

In this section we propose to use efficiency testing methods with the objective of pricing errors. To elaborate, position and hedging payoffs will be used to assess models with the assumption of EMH is correct. Thus, any excess profit (or loss) generated by the model is admitted to be the model's lack of covering for other risks of the market (i.e. model error). So if the same level of hedging is maintained for all models in consideration, benchmarking using efficiency testing metrics becomes plausible.

Another advantage of efficiency centered performance metrics over pricing error metrics is that the consequences of actions are assessed instead of merely measuring the distance from spot prices. Also, since options market is actually a forecast on the underlying asset's market prices, efficiency metrics cover both markets. One disadvantage of measuring with efficiency metrics is it requires increased number of calculations and data.

Three different efficiency centered metrics are considered with increasing hedging activity. First metric is 'naked' positioning. A position on the option is taken and the resulting payoff is also considered with the market option price. Second metric, static delta or 'hedge and forget' strategy also considers taking a delta position with the option position but never change until the expiration of the contract and closing the position only at expiration. Third metric, the most comprehensive and computation intense of all is to go with dynamic positioning and delta hedging throughout the lifetime of the option. The process is very similar to dynamic hedging tests of Galai (1977).

One of the advantages of using efficiency metrics with increasing hedging activity is to measure each added hedging activity's marginal contribution. For instance, it would be interesting to know how much additional effort to hedge actually contributes to model performance.

5.4.1. Naked Positioning Error

Naked positioning error is a simple estimate on the outcome of the option at expiration, dependent on the taken position.

$$NPE_t = \rho(P_t, \hat{P}_t)[\eta(S_T, K, call/put)e^{-r(T-t)} - P_t] \tag{5.6}$$

S_T is the price of the underlying at expiration, K is the strike price and r is the risk free rate. Position function, $\rho(P_t, \hat{P}_t)$ (or simply ρ_t), is an if else condition taking values of 1, if $\hat{P}_t \geq P_t$ or -1, if $\hat{P}_t < P_t$. η is the payoff function of the contract. For European options, it is $max\{S_T - K, 0\}$ for calls and $max\{K - S_T, 0\}$ for puts.

This metric is essentially the bare model error value, given there is no hedging measures in place. Since all equilibrium models are suspected of not covering for all risks according to bad model problem, providing a base for errors is a reasonable first action. Hence, model estimate has no effect on the error except determining the position, therefore the sign of the outcome, of the contract. Also, under risk neutral measure, the expected value of the absolute value of NPE_t is equivalent to APE.

Table 5.4 shows the position changes and NPE of the same contract used in Table 5.1. At expiration, underlying price was at the level of \$1215.65 and moneyness was close to 0.95. The relationship between the relative position of model prices vs market prices and outcomes is quite different.

Figure 5.7 illustrates the NPE histogram for all call and put options, respectively. Both type of options have high kurtosis and heavy tails. Skew is observable in the call options but it is more significant in put options.

Table 5.4. NPE on example contract.

	Date (Y/m/d)	Underlying Price	Market Price	Model Price	ρ_t	NPE
1	2011-10-24	1254.19	45.77	45.71	-1.00	-18.58
2	2011-10-27	1284.59	26.00	26.27	1.00	38.35
3	2011-10-28	1285.09	25.50	25.10	-1.00	-38.85
4	2011-10-31	1253.30	36.90	42.38	1.00	27.45
5	2011-11-03	1261.15	36.15	34.50	-1.00	-28.20
6	2011-11-04	1253.23	42.20	38.70	-1.00	-22.15
7	2011-11-11	1263.85	26.10	25.54	-1.00	-38.25

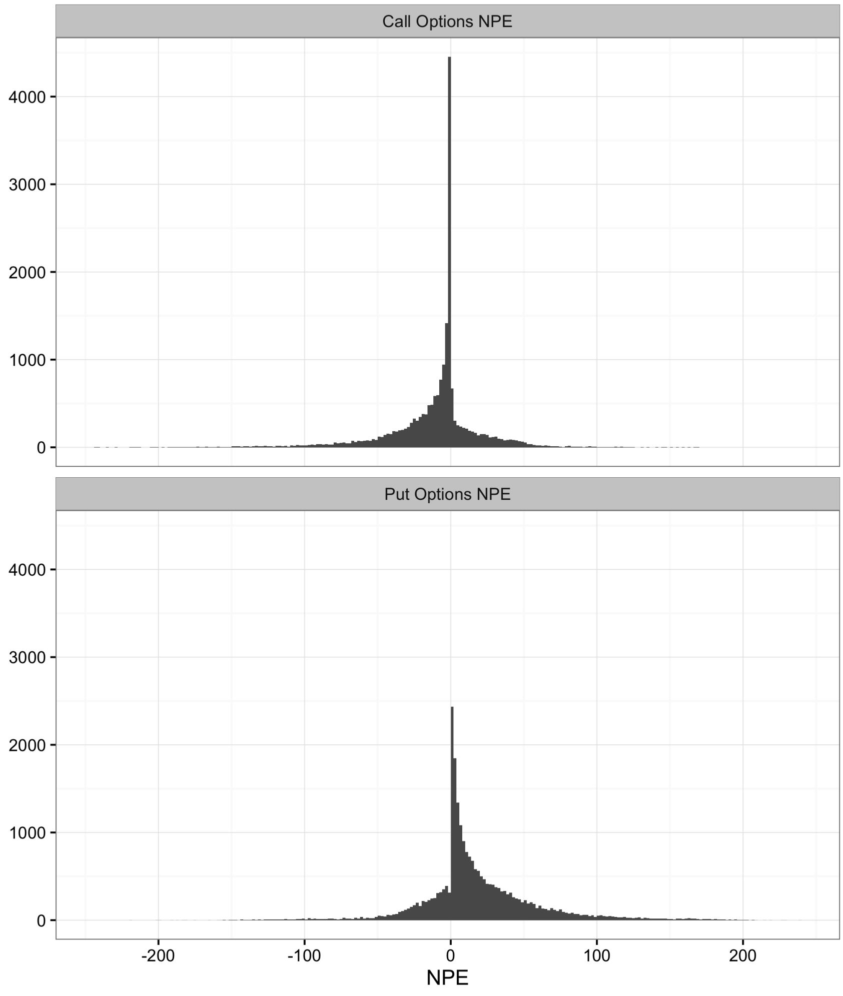

Figure 5.7. Histogram of NPE values for Call and Put options.

5.4.2. Static Hedging Error

Static hedging, also known as a hedge-and-forget positioning, consists of taking a delta position together with the option positioning and then making no adjustments until the expiration of the contract or changing the option position itself. At the expiration, the outcome after the delta position is closed and the payoff is realized gives the Static Hedging Error.

$$SHE = \rho(P_t, \hat{P}_t)[\eta(S_T, K, call/put)e^{-r(T-t)} - P_t] + \Delta(S_T - S_0) \tag{5.7}$$

Δ value shows the amount of the underlying stock to be hedged. Position of Δ is determined by the position taken on the option according to the basic rules of delta hedging. For deep out of the money options where Δ is practically zero, SHE becomes equivalent to NPE.

Table 5.5. SHE on example contract.

	Date	Market	Model			
	(Y/m/d)	Price	Price	$\|\Delta\|$	ρ_t	SHE
1	2011-10-24	45.77	45.71	0.65	-1.00	6.38
2	2011-10-27	26.00	26.27	0.49	1.00	4.63
3	2011-10-28	25.50	25.10	0.49	-1.00	-5.15
4	2011-10-31	36.90	42.38	0.67	1.00	2.21
5	2011-11-03	36.15	34.50	0.64	-1.00	0.91
6	2011-11-04	42.20	38.70	0.70	-1.00	4.00
7	2011-11-11	26.10	25.54	0.67	-1.00	-6.09

Figure 5.8 illustrates the distribution of SHE for call and put options, respectively. The most notable improvement is that its tails are much shorter than of NPE[20] . When compared, SHE provides lower values than NPE. That is the natural effect of delta hedging. Though, even with a single hedge position the outcome error value is reduced

[20] A single contract with SHE greater than 200 is removed from the histogram, but not from calculations.

by up to 90%. Also, due to the hedging error, outcome sign is changed for some contracts.

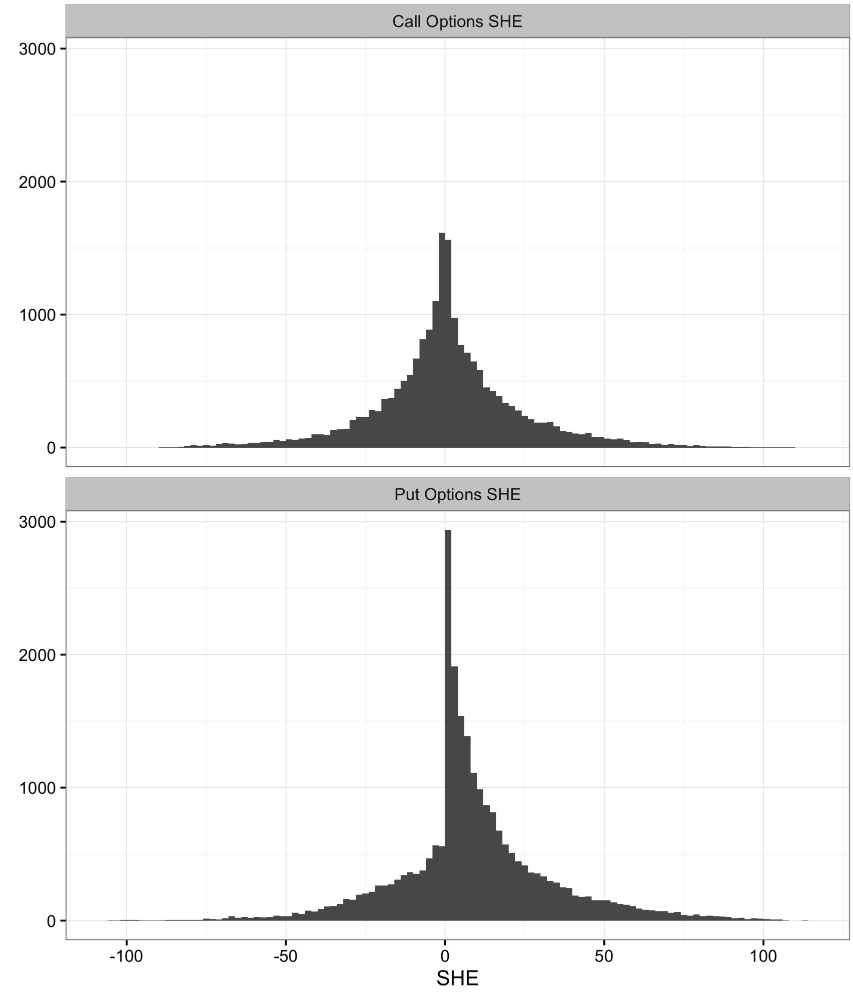

Figure 5.8. Histogram of SHE values for Call and Put options.

Tables 5.6 and 5.7 represent the aggregate efficiency metric errors of call and put contracts, respectively. APE is also appended for easier side by side comparison with pricing errors.

First point of interest is; even if SHE is larger than NPE for some moneyness maturity combinations, the error reduction of SHE is much more significant than NPE in other combinations. Improvement brought by SHE with added delta hedging is most visible for ATM options on mid to long range of maturities in call options.

Table 5.6. Efficiency error averages of Call options based on moneyness and maturity intervals.

Maturity	Moneyness	Number of Contracts	Market Price	Model Price	NPE	SHE	APE
[7,30]	(0.5,0.9]	718	0.34	0.52	0.02	0.16	0.37
[7,30]	(0.9,1]	4501	6.14	8.46	-0.51	2.61	4.54
[7,30]	(1,1.1]	1165	38.74	39.56	-0.30	4.63	6.21
[7,30]	(1.1,1.5]	19	154.10	147.55	16.10	-2.82	6.59
(30,60]	(0.5,0.9]	754	1.08	2.38	-0.37	-0.94	1.50
(30,60]	(0.9,1]	3537	12.13	18.22	-2.47	6.70	8.41
(30,60]	(1,1.1]	950	50.17	52.45	-8.27	10.68	9.34
(30,60]	(1.1,1.5]	23	147.64	139.89	-6.00	9.28	8.14
(60,90]	(0.5,0.9]	473	2.54	5.31	-0.37	-1.83	3.24
(60,90]	(0.9,1]	1593	21.74	28.47	-7.52	6.61	11.00
(60,90]	(1,1.1]	599	58.64	60.78	-22.00	3.95	10.92
(60,90]	(1.1,1.5]	18	164.01	160.41	7.28	4.11	4.12
(90,180]	(0.5,0.9]	783	5.14	10.89	-4.04	-3.91	6.42
(90,180]	(0.9,1]	1741	30.22	39.25	-21.70	1.37	13.37
(90,180]	(1,1.1]	579	74.12	76.27	-53.62	-6.28	12.09
(90,180]	(1.1,1.5]	34	179.95	173.73	-80.70	-3.85	10.33
(180,365]	(0.5,0.9]	708	13.69	26.45	-8.14	-13.49	13.36
(180,365]	(0.9,1]	1261	54.86	67.41	-36.92	-16.83	16.86
(180,365]	(1,1.1]	443	101.04	106.67	-61.69	-20.53	13.88
(180,365]	(1.1,1.5]	35	193.17	192.93	-116.79	-28.50	9.06

For put options displayed in Table 5.7, there is a more of an overall improvement by SHE. Except for several moneyness maturity regions where difference is quite small (e.g. Maturity [7,30], Moneyness (1.1,1.5]) SHE provides much smaller values.

Table 5.7. Efficiency error averages of Put options based on moneyness and maturity intervals.

Maturity	Moneyness	Number of Contracts	Market Price	Model Price	NPE	SHE	APE
[7,30]	(0.5,0.9]	61	209.27	199.04	17.26	10.88	10.25
[7,30]	(0.9,1]	1219	45.15	43.91	12.60	3.70	6.51
[7,30]	(1,1.1]	4192	9.99	7.66	2.18	0.94	4.70
[7,30]	(1.1,1.5]	1874	2.95	0.25	0.91	0.95	2.70
(30,60]	(0.5,0.9]	61	210.59	202.89	16.22	9.81	8.11
(30,60]	(0.9,1]	858	54.84	54.79	31.48	9.88	8.86
(30,60]	(1,1.1]	3056	21.30	18.37	14.45	6.22	7.29
(30,60]	(1.1,1.5]	2635	5.82	1.02	3.83	4.14	4.80
(60,90]	(0.5,0.9]	51	234.49	227.98	39.50	10.75	7.51
(60,90]	(0.9,1]	427	67.67	65.10	48.80	11.35	11.36
(60,90]	(1,1.1]	1657	33.19	27.98	22.63	5.51	9.22
(60,90]	(1.1,1.5]	1615	9.90	2.39	9.34	9.03	7.52
(90,180]	(0.5,0.9]	83	235.86	232.18	46.40	12.45	7.18
(90,180]	(0.9,1]	493	83.12	77.94	55.43	4.54	12.33
(90,180]	(1,1.1]	1670	46.49	38.84	32.76	10.48	10.83
(90,180]	(1.1,1.5]	2046	17.58	5.74	16.71	16.14	11.84
(180,365]	(0.5,0.9]	38	315.10	314.80	35.15	1.61	8.74
(180,365]	(0.9,1]	360	113.60	103.76	65.58	9.51	13.84
(180,365]	(1,1.1]	1072	78.85	64.62	58.99	34.75	15.06
(180,365]	(1.1,1.5]	1436	35.70	16.30	35.69	31.98	19.40

5.4.3. Dynamic Hedging Error

Merton (1973) proves that it is possible to achieve perfect hedging given delta of the option is updated continuously, under Black-Scholes model assumptions. Though, in reality, delta portfolio is rebalanced periodically due to practical reasons and transaction costs.

Our dynamic positioning and hedging error calculations are similar to Galai (1977). Option positioning and rebalancing is performed daily. The differences are due to missing data and closing contract positions earlier than expiration. Since our options data set is on a wider moneyness scale than most studies, not every contract is traded in volume each day. So option position is re-evaluated each day and if no market value emerges due to lack of sufficient volume, no change is made to the contract. Also, due to our data filtering rules, contracts with maturity less than a week are discarded.

By extension, it is also assumed that there is prior knowledge on which days a contract can be traded. Any open position is closed on the last trade possible before the expiration. Even though this assumption, spotting where the trading ceases 100%, seems slightly unrealistic, it is entirely possible to set some strict rules with trading volumes and prices to determine early exit opportunities. But it would further reduce the data set due to false positives, contracts which can still be traded but discarded due to filtering rules. It is considered as a good trade-off between adhering to reality and keeping the data set as large as possible.

$$DHE = \sum_{t=1}^{T'} \rho^*(P_t, \hat{P_t}) * (-P_t) + (\Delta_t - \Delta_{t-1}) * (-S_t) \tag{5.8}$$

ρ^* denotes the net position change in the option contract. For the first position $\rho^*(P_t, \hat{P_t}) = \rho(P_t, \hat{P_t})$ and takes the value of either 1 or -1. But if, at some time point t, the position on the contract changes then the position should be taken twice. First position is taken to close the existing position and the second is to take the opposite position. Also, if $\rho(P_t, \hat{P_t}) = \rho(P_{t-1}, \hat{P}_{t-1})$ no position change should be made. Finally, when totally closing the position on the contract, the opposite position of the net position (say ρ_t^{net}) of the previous period. The terminal position of the contract is determined by $\rho_{T'}^*(P_{T'}, \hat{P_{T'}}) = -\rho_{T'-1}^{net}$. So ρ_t^* can take values of $-2, -1, 0, 1$ and 2 different from ρ_t.

ρ_t^{net} is used for proper calculation of ρ_t^*. $\rho_t^* = \rho_t(P_t, \hat{P}_t) - \rho_{t-1}^{net}\ \forall_t, 1 < t < T'$ and $\rho_t^{net} = \rho_t(P_t, \hat{P}_t)$ if P_t exists. Otherwise $\rho_t^{net} = \rho_{t-1}^{net}$ and $\rho_t^* = 0$.

The procedure is explained in detail in Figure 5.9. The resulting DHE is the error of not a single contract traded at a single date, but the cumulative value of that contract from the time it is first traded till the closing of the position.

```
Take option position according to $\rho(P_1, \hat{P}_1)$ and take $\Delta$ position accordingly.
$DHE = \rho(P_1, \hat{P}_1) * (-\hat{P}_1) + \Delta_1 * (-S_1)$
$\rho_1^{net} = \rho(P_1, \hat{P}_1)$
for $t = 2$ to $T$ (expiration) do
  if $\hat{P}_t$ exists then
    if $\hat{P}_t$ is not the last tradeable contract. then
      Re-evaluate option position.
      Rebalance $\Delta$.
      $DHE = DHE + \rho^*(P_t, \hat{P}_t) * (-\hat{P}_t) + (\Delta_t - \Delta_{t-1}) * (-S_t)$
      $\rho_t^{net} = \rho_{t-1}^{net} - \rho(P_t, \hat{P}_t)$
    else
      Close both option and $\Delta$ positions.
      $DHE = DHE + \rho^-(P_t, \hat{P}_t) * (-\hat{P}_t)$
      break (Interrupt the for loop.)
    end if
  else
    Rebalance $\Delta$.
    $DHE = DHE + (\Delta_t - \Delta_{t-1}) * (-S_t)$
    $\rho^*(P_t, \hat{P}_t) * (-\hat{P}_t) = 0$
  end if
end for
return DHE
```

Figure 5.9. Dynamic Hedging Error calculation algorithm.

Table 5.8 displays DHE process for a single contract. Missing values in Market Price column means either the contract is not always traded in sufficiently high volume or there is missing data. Underlying price is seamless for all the trading days and since the model uses only past underlying asset innovations for price estimation process the model price and model delta evolution can also be calculated seamlessly. ρ_t^* column contains the action information on option positions. The values with -2 or 2 mean the model changed its position on the contract from long to short or short to long, respectively. One position is taken to close the current position and another position to be net one contract on the opposite position. DHE error evolution is represented in the last column, but only the last value, 5.82 is taken into account. Because it is the final outcome value where all option and delta positions are netted to zero.

Table 5.8. DHE on example contract.

Date	Underlying	Market	Model			
(Y/m/d)	Price	Price	Price	$\|\Delta\|$	ρ_t^*	DHE
2011-10-24	1254.19	45.77	45.71	0.65	-1.00	857.73
2011-10-25	1229.05		62.59	0.76	0.00	999.69
2011-10-26	1242.00		52.41	0.71	0.00	937.59
2011-10-27	1284.59	26.00	26.27	0.49	2.00	-658.36
2011-10-28	1285.09	25.50	25.10	0.49	-2.00	644.45
2011-10-31	1253.30	36.90	42.38	0.67	2.00	-877.04
2011-11-01	1218.28		68.00	0.83	0.00	-1075.13
2011-11-02	1237.90		51.65	0.76	0.00	-988.73
2011-11-03	1261.15	36.15	34.50	0.64	-2.00	852.33
2011-11-04	1253.23	42.20	38.70	0.70	0.00	922.39
2011-11-07	1261.12		32.34	0.65	0.00	867.03
2011-11-08	1275.92		22.36	0.55	0.00	731.27
2011-11-09	1229.10		54.71	0.86	0.00	1124.09
2011-11-10	1239.70		44.91	0.83	0.00	1075.86
2011-11-11	1263.85	26.10	25.54	0.67	1.00	**5.82**

Finally, DHE histograms of call and put options are illustrated in Figure 5.10. Tails are significantly lowered and shortened. Hence, the amount of data is also significantly reduced due to the requirement of using the lifetime of a contract as a single contract. For instance, Table 5.8 displays 8 different market prices which are all data points in all previous error terms (including pricing errors), but only the final outcome is used as a data point in the histogram. DHE might be a better method in assessing option model performances. But, due to data requirement, parameter inference might better be made with SHE instead.

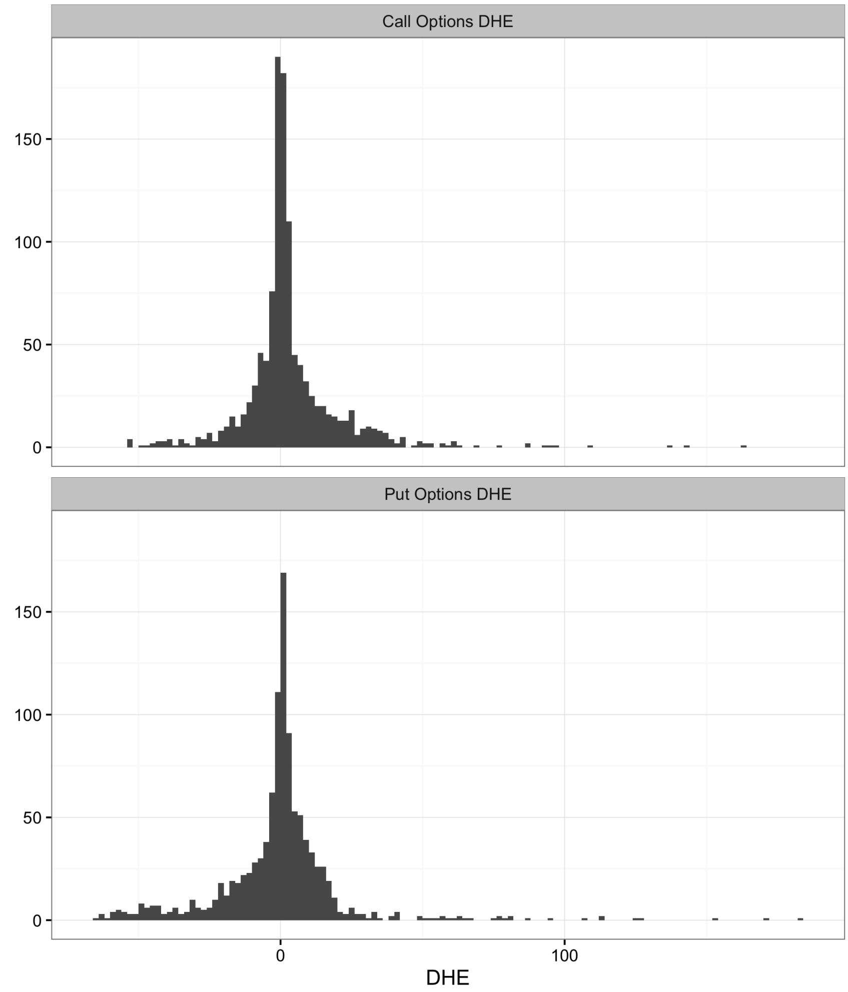

Figure 5.10. Histogram of DHE values for Call and Put options.

It is already shown that pricing errors and efficiency errors are not in accordance in terms of the outcome if contract and/or delta positions are taken on the options. If efficiency metrics are to be used in assessing and benchmarking options DHE will give the theoretically most accurate information. Because both contract and delta positions are taken periodically as described in Galai (1977). But contract-wise SHE is useful, since it uses much fewer data points. SHE can be used in model selection or more granular analysis of option performances. NPE, on the other hand, is useful in auxiliary analysis and trading purposes. It can be thought as a 'weighted positioning error', since position is the only element of the profit or loss. For trading purposes, if a model consistently proposes correct positions more often than not in 'lucrative' (i.e. market price and payoff difference is substantial) contracts, then hedging requirements can be lowered according to the decision maker's taste of risk and trust of the model.

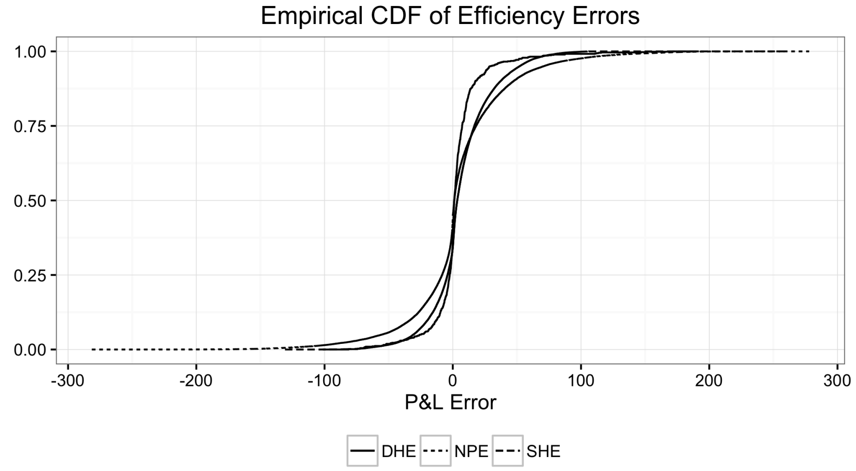

Figure 5.11. Empirical CDF graph of NPE, SHE and DHE.

Figure 5.11 illustrates the comparison of tails and kurtosis of efficiency metrics using empirical cumulative distribution functions. DHE (denoted with the solid line) presents the shortest tails and steepest ascent of all as desired. SHE (dashed line) has asymmetric but moderate tails with again a moderate ascent. NPE has expectedly the longest tails and angular ascent.

5.5. Benchmarking with Efficiency Metrics

Efficiency metrics can be used as a benchmark measure just like pricing errors. Their difference lies on the information they use. Efficiency metrics use the outcome of the position (or positions) taken on the contract. Positions should be closed or contracts should expire to have a conclusion on the benchmark. Pricing errors, on the other hand, use only the spot price of the contract. Therefore, benchmarking can be done real time with pricing errors.

To display the differences of those two metrics in terms of benchmarking, we benchmark the results of the Heston-Nandi GARCH model with the Black-Scholes model. Tables 5.9 and 5.10 show NPE, SHE and APE type errors of different moneyness and maturity intervals for call and put options, respectively.[21] Our main objective is not to compare these two models' performances, but rather to show how performance interpretation of models can change when efficiency metric is used instead of pricing errors.

In Table 5.9, there are notable disagreements between pricing and efficiency errors which can be found both between models or within models. For instance for BS for the same maturity group of (30,60], at the money intervals (0.9,1] and (1,1.1] have similar APE values (11.86 and 11.66 respectively). On the other hand, for the same moneyness maturity intervals, NPE and SHE values are significantly different. For maturity interval (60,90] and moneyness interval (0.9,1] BS model has a higher APE value but a lower SHE than HN model. It means either of the models can be deemed better depending on the performance measure. For maturity interval (60,90] and moneyness interval (1.1,1.5], both HN and BS have similar APE values but this time the SHE and NPE values differ significantly.

Similar disagreements can be found in put options as well. In Table 5.10, HN model has similar APE values for moneyness-maturity intervals (30,60]-(1,1.1] and (90,180]-(0.5,0.9] but SHE value in the latter interval is almost the double of the former. For moneyness-maturity interval (60,90]-(0.5,0.9] APE values of HN and BS are quite close but SHE value of HN is twice the BS SHE value.

We repeated the same experiment with different time periods, assets and models.

[21] DHE is not reported since its calculation requires multiple contracts which are not necessarily in the same moneyness and maturity region.

Even though error value tables and relative performances changed, disagreement between efficiency errors and pricing errors continued to occur for all experiments. The difference between these two types of errors is especially striking for those who calibrate their model for price fitting, but still care about the outcome of the positions on contracts based on information from their models.

Table 5.9. BS and HN benchmarks for Call options.

			HN			BS	
Maturity	Moneyness	NPE	SHE	APE	NPE	SHE	APE
[7,30]	(0.5,0.9]	0.02	0.16	0.37	-0.09	-0.14	0.53
[7,30]	(0.9,1]	-0.51	2.61	4.54	-1.74	2.33	6.17
[7,30]	(1,1.1]	-0.30	4.63	6.21	-5.02	3.64	7.78
[7,30]	(1.1,1.5]	16.10	-2.82	6.59	22.61	-1.86	7.35
(30,60]	(0.5,0.9]	-0.37	-0.94	1.50	-0.71	-2.12	2.35
(30,60]	(0.9,1]	-2.47	6.70	8.41	-4.54	4.41	11.86
(30,60]	(1,1.1]	-8.27	10.68	9.34	-11.43	7.15	11.66
(30,60]	(1.1,1.5]	-6.00	9.28	8.14	7.52	9.77	8.06
(60,90]	(0.5,0.9]	-0.37	-1.83	3.24	-1.32	-4.86	4.92
(60,90]	(0.9,1]	-7.52	6.61	11.00	-7.45	3.44	15.14
(60,90]	(1,1.1]	-22.00	3.95	10.92	-17.79	2.04	13.66
(60,90]	(1.1,1.5]	7.28	4.11	4.12	29.74	5.49	4.62
(90,180]	(0.5,0.9]	-4.04	-3.91	6.42	-2.25	-6.49	9.84
(90,180]	(0.9,1]	-21.70	1.37	13.37	-17.62	-0.80	18.54
(90,180]	(1,1.1]	-53.62	-6.28	12.09	-45.07	-7.43	14.85
(90,180]	(1.1,1.5]	-80.70	-3.85	10.33	-72.87	-6.56	11.44
(180,365]	(0.5,0.9]	-8.14	-13.49	13.36	-5.84	-17.66	20.04
(180,365]	(0.9,1]	-36.92	-16.83	16.86	-28.45	-19.99	23.75
(180,365]	(1,1.1]	-61.69	-20.53	13.88	-54.52	-22.03	17.84
(180,365]	(1.1,1.5]	-116.79	-28.50	9.06	-118.75	-33.89	11.16

Table 5.10. BS and HN benchmarks for Put options.

			HN			BS	
Maturity	Moneyness	NPE	SHE	APE	NPE	SHE	APE
[7,30]	(0.5,0.9]	17.26	10.88	10.25	7.53	10.05	9.07
[7,30]	(0.9,1]	12.60	3.70	6.51	16.50	4.18	7.68
[7,30]	(1,1.1]	2.18	0.94	4.70	5.83	2.76	5.45
[7,30]	(1.1,1.5]	0.91	0.95	2.70	1.95	1.75	2.11
(30,60]	(0.5,0.9]	16.22	9.81	8.11	28.91	8.82	7.02
(30,60]	(0.9,1]	31.48	9.88	8.86	15.91	2.98	11.67
(30,60]	(1,1.1]	14.45	6.22	7.29	16.39	5.35	9.49
(30,60]	(1.1,1.5]	3.83	4.14	4.80	4.73	3.91	3.65
(60,90]	(0.5,0.9]	39.50	10.75	7.51	26.51	5.37	7.88
(60,90]	(0.9,1]	48.80	11.35	11.36	23.64	5.79	14.61
(60,90]	(1,1.1]	22.63	5.51	9.22	14.29	-1.78	11.95
(60,90]	(1.1,1.5]	9.34	9.03	7.52	7.01	4.13	5.44
(90,180]	(0.5,0.9]	46.40	12.45	7.18	-3.48	0.08	11.72
(90,180]	(0.9,1]	55.43	4.54	12.33	15.55	-5.44	15.58
(90,180]	(1,1.1]	32.76	10.48	10.83	14.29	-14.23	13.89
(90,180]	(1.1,1.5]	16.71	16.14	11.84	10.20	2.98	8.20
(180,365]	(0.5,0.9]	35.15	1.61	8.74	-12.00	-10.63	24.65
(180,365]	(0.9,1]	65.58	9.51	13.84	-6.89	-23.81	19.05
(180,365]	(1,1.1]	58.99	34.75	15.06	-3.85	-31.75	16.07
(180,365]	(1.1,1.5]	35.69	31.98	19.40	12.48	0.31	10.28

5.6. Conclusion and Directions for Future Research

In this study we have shown that current model performance (error) metrics widely used in measuring the distance such as RPE, ARPE and APE are not strongly relevant with the consequences of the contracts.

From the perspective of the efficient markets hypothesis and bad model problem, these metrics are not wrong. Bad model problem dictates that to disprove market efficiency an equilibrium model must be used. But, any disagreement between the model and the market might also be due to model's inability to address some risks included in the prices. Pricing error functions implicitly assume markets are informationally efficient and disagreement with the market is attributed to the 'model error'.

We, on the other hand, combined metrics previously used to measure market efficiency with the assumption of pricing error metrics. In other words, any 'inefficiency' displayed by these metrics are attributed to model's inability to take risks into account. Naked Pricing Error, Static Hedging Error and Dynamic Hedging Error metrics are used to show the model's performance throughout the lifetime of the option and the incremental improvement in model performance when hedging activity is increased.

Relation between efficiency metrics and pricing error metrics are examined and disparities are examined. It is shown that contracts with similar pricing errors might yield significantly different results when the payoffs of the contracts are calculated. A clear advantage of using efficiency metrics is that consequences of actions taken on options (such as positions) becomes quantifiable. So the model misspecification can be measured more consistently. In exchange, more data is required to calculate the error of the contract and the calculations become more cumbersome.

The difference between SHE and DHE are comparably small to the difference between SHE and NPE, so SHE can also be used to benchmark model performances and model selection. Since calculating DHE requires a time series of the same contract and different models might give better results at different points of the lifetime of that contract, SHE is a better choice for model selection despite static hedging is a less effective model performance metric.

New efficiency metrics can also be introduced in the future. For instance, spread based tests of Galai (1977) can also be considered as model errors. Relative versions of such metrics can also be useful.

6. A MODEL SELECTION FRAMEWORK FOR PRICING OPTIONS

Empirical studies show that even the best performing option pricing models cannot sustain their performance for all contracts. It can also be added that each model can give the best price estimate for at least a set of contracts. Our aim is to detect which model (and parametrization) is the best price estimate for each individual contract and delta hedging. A model selection framework is proposed to achieve this aim. Both model selection and individual models are benchmarked with different error metrics and underlying assets. Results indicate that model selection is a good and consistent way of pricing option contracts.

6.1. Introduction

Currently, there are hundreds, if not thousands, of option pricing models out there; some of which are used extensively for a broad range of assets and contracts. Those models are often iterations on previous models, claiming to improve the weaknesses of their predecessors. Empirical tests are employed to prove and measure their effectiveness over benchmark models, or merely to benchmark different models to see which model prevails under that experiment assumptions. Although there is usually an overall “best” model, detailed examination shows significant room for improvement.

Results may differ based on the choice of performance metric, asset and time period. There is also the usual practice of partitioning of the results in moneyness and maturity regions. The “best” model is expected to be the best in all the regions, though occasionally another model is shown to be better (even by a small margin) in some regions. It shows that a model is rarely, if ever, a dominant model for even a contract region for all time periods. What if there was a way to detect which model would perform better than all others in a given model set for a given objective (e.g. minimize ARPE)?

There is enough evidence in the literature to arouse concern about trusting a model for prolonged periods of time or use the same model for all contract parameter regions. Heston and Nandi (2000) experiments show that year-over-year results have different ranking of models in terms of performance. Even though the best model didn't change with their calculations, the second place is occupied by different models in different years. Lehar et al. (2002) benchmark Black-Scholes (BS), GARCH and Hull-White Stochastic Volatility (HW) models in terms of option pricing and risk. Their result tables accommodate different pricing error types and moneyness maturity regions. The outcome indicates that even the base benchmark model, BS, can be the model with the least error for a given moneyness-maturity region. Chorro et al. (2012) assess the performances of their proposed models using two different assets; S&P 500 and CAC 40 contracts. The results and the best models differed based on the underlying asset.

There are three ways to improve model performance. First way is to improve parameter fit without changing the model itself. Second way is parametric improvement, such as adding new parameters to the model or making constant parameters time varying. Third way is structural improvement, by introducing new processes or underlying distributions. Parameter fitting improvement is stressed highly in the literature, such as implied parameters (e.g. implied volatility) from market option prices is better than using historical returns, because market option prices are 'forward looking'. The actual reason is captured by Christoffersen and Jacobs (2004a) by the simple but effective thumb rule "optimize the parameters of the model with the same objective function that model performance is assessed with". Parametric improvements can increase model performance only so far. After a degree of complexity, models overfit (i.e. provide worse predictions). For instance, Christoffersen and Jacobs (2004b) examine different streams of GARCH models to come up with conclusion that high complexity models with many parameters do not fare better than more parsimonious models. In addition, it is harder to optimize parameter values. Structural improvements such as using a GARCH model, instead of a simple BS or transforming underlying asset log-return innovation distribution assumption from Brownian Motion to Generalized Hyperbolic (GHYP) distribution can also improve the performance of the model up to a point. Even then, the fortified model provides no guarantee that it will dominate all other models for all contracts and all time periods.

We propose a model selection framework that will work with a set of option pricing models. The method will not produce a price estimate but predict which individual model provides the best estimate and copy that model's price estimate as its own estimate. Model selection is expected to perform better than all individual models for any given objective (not just pricing errors, but also market efficiency tests). Model selection is easily scalable (i.e. it is possible to add any number of models) and is less prone to overfitting. Also, it can pick different models' estimates for different moneyness maturity regions, therefore eliminating individual models' local weaknesses.

Rest of this chapter is organized as follows. Section 6.2 describes the model selection framework and the generalized pricing model. Section 6.3 describes the data and methodology of the emprical analysis as well as numerical results for SPX and NDX, and observations about model selection. Finally, Section 6.4 summarizes the findings and discusses further directions for model selection framework.

6.2. Model Selection Framework

The idea behind model selection framework is simple: A method to pick models to price options which gives better results than the individual models it picks from. The required execution for better performance is not necessarily as simple. The term better can be defined as better performance in the given objective (e.g. minimizing pricing error).

Any model selection method consists of two parts: individual models and selection methodology. There are two assumptions required to make a set of individual models for model selection to operate on.

Assumption 1. Market assumptions of model selection should be able to accommodate the market assumptions of all individual models.

Assumption 2. Each model in the model set should yield the best result for at least a contract in the given set of contracts. In other terms, no model is dominated by other models.

For every $k' \in \mathcal{K}$ there exists $i \in I$ such that $\epsilon_i^{k'} < \epsilon_i^k$ for all $k \in \mathcal{K}, k \neq k'$ where ϵ_i^k is the error[22] of the individual model k, k' in the model selection set $\mathcal{K}$ for contract i in the contract set I.

Assumption 1 is required so each model can perform in a market where its assumptions are satisfied, therefore its estimates are valid. Assumption 2 is to ensure no inferior model is considered in the model set. Inferior models have no added advantage to model selection and can even be detrimental to the performance if model selection picks their estimates. This assumption is also in line with the claim that no model is dominant.

Highest similarity for model selection in the literature can be found in hyper-heuristics studies. Like model selection, hyper-heuristics employ a heuristic to "pick" the best "heuristic". Though, similarity ends here. There are few, perhaps even zero, hyper-heuristics studies regarding option pricing. Selected entities are not heuristics, they are pricing models. Hyper-heuristic studies usually pick from simple heuristic methods using a more sophisticated heuristic method, although it can be done either way. Current proposition of model selection employs a rather simple and greedy rule to pick from fairly sophisticated models. Detailed coverage of hyper-heuristics can be found in Burke et al. (2010).

6.2.1. Model Selection Process

The proposed model selection process is straightforward. The criteria (ξ) consists of two parts; an objective and cluster error estimate. For each model a Conditional Inference Tree (CIT) is trained in the clustering part. Out-of-sample predictions of

[22] Any error measure including measures with maximization objective (i.e. efficiency tests) can be used as long as negative value of that measure is taken.

these CIT models are calculated for each contract ($\hat{\epsilon}_i^k$). Decision is made upon the error term of individual models. Model selection picks the best model according to its objective function for each contract. Let $\xi_i \in \mathcal{K}$ be the index of the best model, so

$$\xi_i := argmin\{\hat{\epsilon}_i^k,\ k \in \mathcal{K}\} \tag{6.1}$$

The term $\hat{\epsilon}_i^k$ should not be confused with ϵ_i^k. While $\hat{\epsilon}_i^k$ is the model selection method's error prediction for the contract i and model k and model selection method's decision is based on $\hat{\epsilon}_i^k$ values; ϵ_i^k is the realized value of the error on contract i by model k and both individual models and model selection methods are evaluated on ϵ_i^k.

The objective can be the minimization of model error (pricing or efficiency) as well as efficiency tests themselves (i.e. maximizing profit from trading strategies with models). Both naked positioning error (NPE) and static hedging error (SHE) can be rearranged for maximization of naked P&L (NPL) and hedge and forget P&L (HPL) respectively. Model selection process is illustrated in Figure 6.1 as a pseudo algorithm.

```
for $k = 1$ to $K$ (each model) do
  Train CIT with the training data
  for $i = 1$ to $I$ (each out-of-sample contract) do
    Predict CIT errors
  end for
end for
for $i = 1$ to $I$ (each out-of-sample contract) do
  Select a model for each contract using the objective function $argmin\{\hat{\epsilon}_i^k, k \in \mathcal{K}\}$
  Use that model's price and delta estimates for that contract as model selection's
  estimates.
end for
```

Figure 6.1. Model selection framework process.

There are two immediate disadvantages of using this model selection framework: It requires the computation of all individual models' price estimates and model selection performance is still limited if individual models agree. For instance, with naked position profit maximization objective case, if all models decide to be on the wrong (money losing) side of the contract then model selection's only choice is also taking that side of the contract. The second disadvantage might be alleviated by increasing the number of models, therefore decreasing the probability of model consensus, at the expense of aggravating computational burden. The trade-off should be adjusted properly.

Training data lookback period and update frequency are important matters for the performance of the model selection. Preliminary tests indicated shorter lookback periods and frequent updates are better in catching shifts in the model performances. However, shorter lookback periods would mean fewer data in calibration of especially OTM/ITM and long term contracts and frequent updates would mean increased computational burden. Lookback period of one year and monthly update is a good compromise according to initial runs. Especially switching from yearly update to monthly update boosted algorithm performance significantly. Even though tests indicated two years training lookback and one year training lookback are less significant, one year lookback displayed more promising results to the favor of model selection.

6.3. Numerical Experiments

Model selection performance is benchmarked empirically with S&P 500 and NASDAQ 100 indices between 2009 and 2013. 5 option pricing models are used to calculate contract prices and delta values for both assets. Then, 7 model selection methods with different objectives draw information from all individual models by using the model selection framework.

Both model sets are benchmarked with each other according to different error types. It is expected for the model selection methods to provide more sustainable if not better results than individual models in their respective objectives. It is also interesting to see a model selection method showing better performance when using not its own selection methods. For instance, the model selection method aimed to minimize ARPE yields also better results in RPE than the model selection method with the objective to minimize RPE.

This section includes preparation of models and data as well as the experiment results. First, individual models and model selection methods are introduced. Then, details of the contract data for both assets are provided. Finally, for both assets, overall numerical results are given for each model and error type. Year over year detailed tables are given in the Appendix.

6.3.1. Models and Model Selection Methods

Performance of model selection framework is tested with 5 different parametrization of 2 option pricing models. These are Heston-Nandi GARCH model of Heston and Nandi (2000) and Black-Scholes model of Black and Scholes (1973).

Model parameters are inferred using historical log-returns of the underlying assets, and parameter inference is updated on each trading day. Although it is widely acknowledged that using option prices for inference is better than using historical log-returns mainly because log-returns are "backward looking" and option prices are "forward looking", there are several reasons why log-returns are used. Even though Dumas et al. (1998) show parameter inference using option prices is better for Black-Scholes model, tests of Christoffersen and Jacobs (2004b) display no significant difference in GARCH-type models[23] . Also recent studies such as Chorro et al. (2012) and Guégan et al. (2013) frequently use historical asset prices, since market option prices are not always consistent predictors of the future (especially after the 2008 crisis[24]).

All seven of the model selection methods draw information from these five individual models. Model selection methods differ only on the objective function, inference method is the same. Minimization of ARPE, RPE[25] and APE are the pricing error objectives. Minimization of Naked Pricing Error (NPE) and Dynamic Hedging Error (DHE) are also included as efficiency metrics. Finally, market efficiency test metrics

[23] They also start with option prices, then test with historical prices for additional insight.

[24] See Guégan et al. (2013) for a similar explanation.

[25] Proximity to zero is the objective.

are included with the objective of maximization of trading profit and loss (P&L). These metrics can be considered as the opposite of the efficiency model error metrics. They are named as naked positions P&L (NPL) and hedge and forget P&L (HPL). The list of model selection methods is given in Table 6.1. All model selection methods are tested with the same data sets. Their performances are assessed and benchmarked according to their respective objectives.

Table 6.1. Model Selection methods.

Model Selection Method	Type	Error
MS ARPE	Pricing Error	Absolute Relative Pricing Error
MS RPE	Pricing Error	Relative Pricing Error
MS APE	Pricing Error	Absolute Pricing Error
MS NPE	Efficiency Model Error	Naked Pricing Error
MS SHE	Efficiency Model Error	Static Hedging Error
MS NPL	Efficiency Test	Naked P&L
MS HPL	Efficiency Test	Hedge and Forget P&L

6.3.2. Data

Both model selection methods and individual models are assessed and benchmarked using closing prices of two different assets; S&P 500 (SPX) and NASDAQ 100 (NDX) indices. Very few empirical studies embrace testing with two assets and when such course of action is taken results may vary significantly. For instance, Chorro et al. (2012) use both S&P 500 and CAC 40[26] and best models are different in at least half of the moneyness maturity regions in the out-of-sample error summary tables.

Option data used in this study is subject to some filtering as common in other empirical studies. Contracts which have moneyness less than 0.5 and more than 1.5

[26] French stock market index.

are removed. Also, maturity values below a week[27] and above a year[28] are removed. Finally, contracts with volume less than 100 or contracts which violate no-arbitrage rules are removed. Both SPX and NDX option data sets are prepared in the same way.

Table 6.2 displays the number of contracts and average price, moneyness and maturity values of SPX and NDX contracts for each year. Each contract on each day counts as a data point if it is within the filtration described above. SPX has about four times more contracts than NDX, longer average maturity. For all years, there are more put options than call options. Average moneyness values indicate that far OTM put options are in high demand[29] .

Table 6.2. YoY descriptive statistics and contract parameter averages of SPX and NDX contracts.

		SPX				NDX			
		Contracts	Avg. Price	Avg. Maturity	Avg. Moneyness	Contracts	Avg. Price	Avg. Maturity	Avg. Moneyness
2009	Call	18084	24.50	51	0.94	4670	29.91	31	0.94
	Put	22159	27.28	50	1.13	5571	25.03	30	1.13
	Total	40243	26.03	51	1.04	10241	27.26	30	1.05
2010	Call	19045	21.36	54	0.95	4483	23.76	30	0.95
	Put	26724	22.48	51	1.13	5795	24.14	29	1.12
	Total	45769	22.01	52	1.05	10278	23.97	29	1.05
2011	Call	20030	24.00	55	0.95	4390	32.23	27	0.96
	Put	27959	26.21	51	1.12	5538	30.05	27	1.10
	Total	47989	25.29	53	1.05	9928	31.01	27	1.04
2012	Call	17961	22.77	59	0.96	3435	25.85	26	0.96
	Put	27255	20.86	55	1.12	4692	21.51	25	1.09
	Total	45216	21.62	57	1.06	8127	23.34	25	1.04
2013	Call	20502	25.33	60	0.97	3110	20.03	23	0.97
	Put	30016	19.95	55	1.11	4618	13.44	23	1.08
	Total	50518	22.14	57	1.05	7728	16.09	23	1.04

[27] 7 calendar days or approximately 5 trading days.

[28] 365 calendar days or approximately 252 trading days

[29] Although not reported, median values are also examined. Especially put options are significantly closer to ATM.

6.3.3. Model Selection Preferences

Each model selection method, according to their objectives, picks price estimates from individual models for each contract. Table 6.3 presents frequency of each individual model for each objective for SPX. Black Scholes models are in high frequency in error metrics but HN models are more favored in efficiency test metrics.

Table 6.3. Frequency of individual models for each objective (SPX).

	APE	ARPE	RPE	SHE	NPE	HPL	NPL
BS2	61656	54955	67524	55002	60292	24642	38577
BS5	39503	38864	36556	43320	50142	34415	46965
HNS2	34948	33763	31621	36648	41795	22333	33630
HNA2	36142	41033	39989	36704	38604	42507	38842
HNA5	50293	53927	46852	50868	31709	98645	64528

Figure 6.2 shows a sample of individual model preferences of the model selection with the objective of minimize ARPE. Calls are predominantly HN models and a significant portion of OTM puts are BS5 and mid-to-long term ATM puts are BS2.

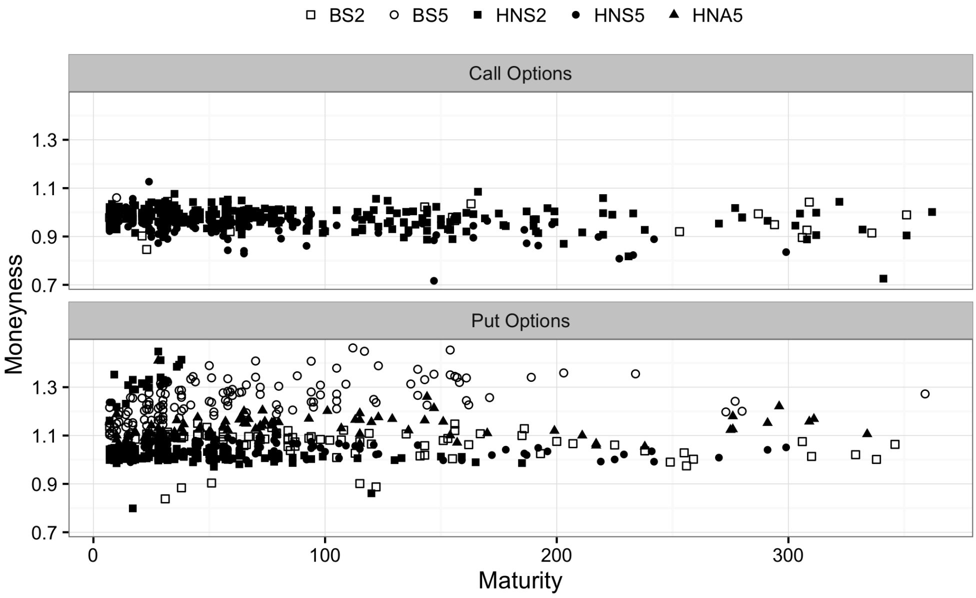

Figure 6.2. Model selection results for ARPE objective on sample data (SPX).

Table 6.4 presents frequency of each individual models for each objective for NDX. HNA5 model is considered as the best individual model by the model selection methods for all objectives except APE and RPE. In those two objectives BS2 model has the highest frequency of preference. Though, overall, all individual models are considered useful by the model selection algorithms.

Table 6.4. Frequency of individual models for each objective (NDX).

	APE	ARPE	RPE	SHE	NPE	HPL	NPL
BS2	13846	11245	16128	12397	11315	7011	8194
BS5	6858	6556	6190	6919	9870	6530	11568
HNS2	7536	6477	6611	8283	8298	4310	7266
HNA2	6818	8969	7253	6744	7459	7616	7040
HNA5	13708	15519	12584	14423	11824	23299	14698

Figure 6.3 shows a subsample of individual model preferences of the model selection method with the objective of minimizing ARPE for NDX contracts. Almost all options are priced by HN models, but BS models are more frequent in OTM and mid-to-long term options.

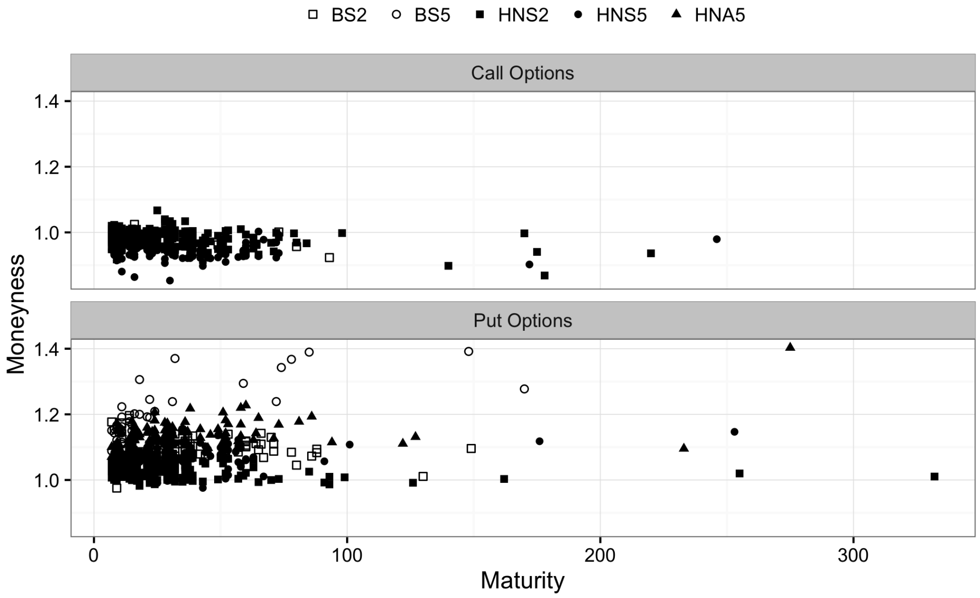

Figure 6.3. Model selection results for ARPE objective on sample data (NDX).

In the following sections, the outcome of out-of-sample estimates will be examined. Model selection methods' performance will be assessed based on predictive power.

6.3.4. S&P 500 (SPX) Results

Result tables, both SPX and NDX, are split into three groups. The first group consists of individual option pricing models. The second group consists of MS methods with efficiency related metrics. Finally, the third group consists of MS methods with pricing error metrics. Total NPE and Total SHE rows can also be interpreted as Total NPL and Total HPL, respectively. Since individual models have no objectives to minimize or maximize the output, the results will be the same for Total NPE and Total NPL. It will also be the same for Total SHE and Total HPL values. For model selection methods, since each method has its own objective, the Total NPE/NPL results are given in the same row as well ass Total SHE/HPL results.

For total results (i.e. Total NPE, Total SHE and Total DHE), overall results are the sum of yearly outcomes. For other error types, the overall results are averages of yearly results.

Each model selection method should be benchmarked with the individual models on its own objective. For instance, ARPE performance is important for MS ARPE against the individual models to assess the success of model selection. However, there might be cases where other model selection methods surpass the performance of model selection method in its own performance metric.

Table 6.5 presents the results of individual models. Best Total NPE, Total SHE and Total DHE values are provided by BS5. HNA2 brings the highest Total NPL, closely followed by HNS2. Yet, highest Total HPL is provided by HNA5. Regarding pricing error metrics, BS2 model yields the best results in relative errors (i.e. ARPE and RPE) but HNS2 and HNA2 models are better in absolute error metric (APE). Significant differences between BS2 and BS5 models imply the effect of lookback period in errors.

Table 6.5. Aggregate outputs of individual models for SPX contracts.

	BS2	BS5	HNS2	HNA2	HNA5
Total NPE/NPL	270379.76	54586.48	1205714.81	1287788.54	1138496.25
Total SHE/HPL	-122310.49	-420028.59	99058.49	174927.34	396374.78
Total DHE	-2945.79	9315.46	13933.55	19606.82	20639.16
Median DHE	0.10	0.28	0.53	0.75	1.31
Mean DHE	-0.26	0.85	1.15	1.61	1.85
SD DHE	13.95	14.52	13.49	13.58	12.69
ARPE	1.71	2.20	1.34	1.52	1.59
RPE	1.26	1.79	0.80	0.90	0.97
APE	9.13	11.19	7.88	8.18	9.11

Results of the model selection methods with efficiency metrics are given in Table 6.6. Respective performance metric results in alignment with the model selection objectives are given in italics. For instance, MS NPL aims to maximize naked profit and yields the highest Total NPE. On the other hand MS NPE aims to minimize Naked P&L metric and yields significantly lower Total NPE value.

Overall results indicate model selection performance is in alignment with their respective objectives. MS NPL yields the highest Total NPE, MS HPL yields the highest Total SHE and MS SHE yields the lowest MS SHE. Only MS NPE is the second model in its objective (MS SHE is lower), but it is low enough. DHE results are also in line with the model selection objectives. As expected, MS HPL and MS SHE have the highest and lowest Total DHE values. Median values also support the objectives. As an interesting observation standard deviation of DHE values are fairly close to each other. So, the effect of model objective in SD DHE seems limited.

Compared to individual models, model selection methods show promising results. MS NPL yields better Total NPE than all individual models. In terms of naked profit generation, MS NPL achieves its aim to come up with better decisions. MS HPL is second only to HNA5 in Total SHE metric. For minimization of model error objectives

BS5 model is the best model in all efficiency model error metric.

Table 6.6. Aggregate outputs of MS efficiency metric models for SPX contracts.

	MS NPE	MS SHE	MS NPL	MS HPL
Total NPE/NPL	*373748.80*	285641.30	*1663650.90*	1218328.77
Total SHE/HPL	-69237.43	*-98016.93*	209426.41	*251420.65*
Total DHE	15952.22	8401.57	23749.45	28035.58
Median DHE	0.48	0.37	0.90	0.97
Mean DHE	1.36	0.78	2.01	2.40
SD DHE	14.10	12.93	13.42	13.32
ARPE	1.57	1.45	1.42	1.23
RPE	1.12	1.01	0.82	0.61
APE	8.62	8.72	8.84	8.08

Results of the model selection methods with pricing errors are given in Table 6.7. MS ARPE yields the best results in all pricing error metrics. MS ARPE also yields better results than both individual models and other MS methods. In addition all pricing error metric MS methods are better than other models and MS methods in RPE and APE. For SPX contracts for the given period of 5 years, MS methods with pricing error metrics perform parallel to expectations. In terms of efficiency metrics, performance of MS methods with pricing error objectives is not even as good as individual models.

Table 6.8 presents the relative performance changes brought by the model selection method in its own performance metric.[30] For instance, MS NPL improves the NPL value of the best individual model (HNA2) by 29.19%.[31] All model selection methods except MS NPE and MS HPL performed better than all other individual models in their respective performance metrics. Even so, MS NPE failed to improve only on BS2 and BS5, and MS HPL failed to improve only on HNA5. MS SHE, even by a small margin, succeeded to surpass all individual models. Even though MS ARPE surpasses

[30] Signed performance metrics (NPE, SHE and RPE) are evaluated with absolute values.

[31] Missing values in MS HPL row indicate the individual models BS2 and BS5 actually incurred net losses, so the relative performance improvement cannot be measured simply because of the change is from loss to profit.

Table 6.7. Aggregate outputs of MS with pricing error metrics for SPX contracts.

	MS ARPE	MS RPE	MS APE
Total NPE/NPL	421149.81	292899.04	294598.17
Total SHE/HPL	-12931.22	-77642.85	-2313.18
Total DHE	20753.89	11349.10	16958.76
Median DHE	0.52	0.34	0.42
Mean DHE	1.80	1.02	1.46
SD DHE	13.15	13.33	13.26
ARPE	*0.94*	1.02	1.06
RPE	0.54	*0.64*	0.67
APE	6.17	6.75	*6.60*

the performances of MS RPE and MS APE in their respective performance metrics, those model selection methods still performed better than all individual models in their respective performance metrics.

Table 6.8. Respective performance changes for each model selection method against individual models for SPX (%).

	BS2	BS5	HNS2	HNA2	HNA5
MS NPE	-38.23	-584.69	69.00	70.98	67.17
MS SHE	19.86	76.66	1.05	43.97	75.27
MS NPL	515.30	2947.73	37.98	29.19	46.13
MS HPL	-	-	153.81	43.73	-36.57
MS ARPE	45.03	57.27	29.85	38.16	40.88
MS RPE	49.21	64.25	20.00	28.89	34.02
MS APE	27.71	41.02	16.24	19.32	27.55

6.3.5. NASDAQ 100 (NDX) Results

Same experiment is repeated with NDX contracts. Overall results are similar to SPX results, although not without subtle differences. Table 6.9 displays the individual models' benchmark results. Once again, HNA2 has the greatest value in Total NPE and all HN models yield greater Total NPE values than BS models. HNA5 yields the greatest Total SHE and all HN models yield greater Total SHE than BS models.

Interestingly, BS models yield net losses in Total DHE and BS2 has the greater loss. Also in terms of efficiency metric model error objective HNS2 yields the best value in Total DHE. Differences in Mean DHE and Median DHE imply skew in terms of error magnitude.

Regarding pricing error metrics, results are quite similar to SPX. BS2 model yields the best results in ARPE and RPE, and all HN models yield lower APE than BS models. HNS2 and HNA2 provide the best result in APE.

Table 6.9. Aggregate outputs of individual models for NDX contracts.

	BS2	BS5	HNS2	HNA2	HNA5
Total NPE/NPL	103887.69	97793.50	167356.67	198505.87	192383.62
Total SHE/HPL	23806.38	18085.35	54032.99	69326.63	98325.65
Total DHE	-9262.31	-1509.32	-2105.08	-163.37	4168.42
Median DHE	-0.12	0.12	0.57	1.08	1.26
Mean DHE	-2.23	-0.25	-0.44	0.07	1.31
SD DHE	19.55	19.32	18.77	18.65	17.45
ARPE	1.70	2.12	1.36	1.45	1.57
RPE	1.29	1.76	0.91	0.90	1.07
APE	11.28	14.28	9.44	9.48	10.91

Overall results of MS methods with efficiency metrics are given in Table 6.10. They contain both increased success and objective-outcome conflict.

MS NPL has the highest Total NPE value with a wide margin. On the other hand, MS HPL model yields a lower Total NPE value, than MS NPE and MS SHE. Outcome is the opposite of the MS method's objective. In addition, MS SHE performs better (i.e. yields less) in Total NPE but worse in Total SHE than MS NPE. In terms of Total DHE, MS SHE performs the best, in line with the expectations. Median DHE and Mean DHE results also support MS SHE. SD DHE results do not differ significantly.

Compared to individual models, MS NPL yields 58% higher Total NPE than HNA2 (highest Total NPE value in individual models). It strikes as a remarkable improvement. Only HNA5 model yields more Total SHE than MS methods. BS5 also yields the best Total SHE, meaning no MS method performs the best in any objective regarding SHE. But, in terms of Total DHE, MS SHE method is better than HNS2. For mean and median DHE values BS5 and HNS2 provide better results, respectively.

Table 6.10. Aggregate outputs of MS efficiency metrics for NDX contracts.

	MS NPE	MS SHE	MS NPL	MS HPL
Total NPE/NPL	*101306.50*	111733.75	*315217.61*	131937.86
Total SHE/HPL	26058.90	*64423.27*	79982.22	*64816.33*
Total DHE	-3167.71	-2410.62	-96.72	2734.53
Median DHE	0.42	0.31	1.09	1.06
Mean DHE	-0.62	-0.37	0.22	0.83
SD DHE	18.91	17.70	19.00	17.95
ARPE	1.69	1.47	1.51	1.39
RPE	1.27	1.06	1.00	0.90
APE	10.99	10.90	10.66	9.75

Overall results of MS methods with pricing error metrics are given in Table 6.11. The outcome is parallel to SPX results. MS ARPE, again, yields the best results for all pricing error metrics compared to both other MS methods and individual models.

Table 6.11. Aggregate outputs of MS with pricing error models for NDX contracts.

	MS ARPE	MS RPE	MS APE
Total NPE/NPL	169863.60	139126.60	144199.96
Total SHE/HPL	42029.90	32595.91	44632.14
Total DHE	61.37	-3784.25	-2202.36
Median DHE	0.42	-0.09	0.08
Mean DHE	0.11	-0.86	-0.45
SD DHE	17.48	17.63	17.64
ARPE	*0.99*	1.08	1.13
RPE	0.60	*0.72*	0.77
APE	7.88	8.62	*8.30*

Table 6.12 presents the relative performance changes brought by the model selection method in its own performance metric.[32] Model selection methods except MS NPE, MS SHE and MS HPL performed better than all other individual models in their respective performance metrics. Even so, MS NPE failed to improve only on BS5, and MS HPL failed to improve only on HNA2 and HNA5. Even though MS ARPE surpasses the performances of MS RPE and MS APE in their respective performance metrics, those model selection methods still performed better than all individual models in their respective performance metrics.

Benchmarking individual models with model selection methods over 5 years and with 2 different assets helps us to come up with some interesting observations about both option pricing models and model selection methods.

Observation 1. Model Selection shows promising results beyond initial expectations for both underlying assets, given the simple structure of the selection methodology. Best overall MS methods are MS NPL and MS ARPE in their respective objectives (i.e. MS NPL has the highest naked P&L and MS ARPE has the lowest ARPE).

[32] Signed performance metrics (NPE, SHE and RPE) are evaluated with absolute values.

Table 6.12. Respective performance changes for each model selection method against individual models for NDX (%).

	BS2	BS5	HNS2	HNA2	HNA5
MS NPE	2.48	-3.59	39.47	48.97	47.34
MS SHE	-170.61	-256.22	-19.23	7.07	34.48
MS NPL	203.42	222.33	88.35	58.80	63.85
MS HPL	172.26	258.39	19.96	-6.51	-34.08
MS ARPE	41.76	53.30	27.21	31.72	36.94
MS RPE	44.19	59.09	20.88	20.00	32.71
MS APE	26.42	41.88	12.08	12.45	23.92

Observation 2. Regarding efficiency error metrics (i.e. NPE, SHE and DHE), some individual models perform better than MS methods. BS5 is the best model in all three metrics for SPX and it is the best in all but Total DHE for NDX. HNS2 yields the best Total DHE for NDX. Though, model selection objective for efficiency errors can easily be confused with poor model performance. Better efficiency should aim for long-term stable minimal divergence from “perfect hedge” returns (i.e. risk-free rate of return). High, but symmetric deviations from efficient returns might yield better efficiency error. Even so, there is still room for improvement for both MS NPE and MS SHE methods.

Observation 3. About pricing errors; MS ARPE performance is the best not only in ARPE, but also in RPE and APE. Year over year results provided in the Appendix indicate stable performance. Considered with the findings of Chapter 4, CIT performed the best in clustering ARPE values. MS RPE also performs well, but not just as good as MS ARPE. An improvement can be to change the clustering method dependent on the model selection objective. For instance, K-Means performed better in APE type errors. Had contracts been clustered with K-Means, performance of MS APE can be improved for APE metric.

Observation 4. DHE is the hardest metric to interpret. Since it includes daily rebalancing, the models are assessed over a longer term. For individual models, best DHE changes between different models and parametrizations. For MS methods, MS SHE provides the best overall results as expected.

Observation 5. Model selection framework is not without problems. Model selection decisions are naturally dependent on the success of the individual models. Extreme shifts in performance by individual models are detrimental to model selection success. Also, especially for metrics requiring taking short and long positions on option contracts model selection performance can be affected without choice. In other words, if all models agree on a position, model selection must also agree even if it is the wrong decision.

These problems can be overcome by increasing the number of individual models in theory. Unfortunately, our initial tests show decrease in performance especially if low performing models are added to the individual models set. So either individual models should be carefully selected a priori or framework should be improved to address these problems.

6.3.6. Significance of Model Selection Methods

Numerical results show that model selection methods perform in parallel to expectations for most of the objectives. But are they statistically so different from the individual models and how do they rank in terms of having the least error? Friedman and post-hoc Nemenyi tests are employed to check for significance using complete 5 years' out-of-sample estimations. All individual models are compared with their respective model selection methods for each objective and with each other.

Demsar (2006) is followed for the implementation of the statistical tests. Friedman test is the nonparametric equivalent of ANOVA with repeated measures. ANOVA test has assumptions like normality and equal variances and those assumptions are likely to be violated by the data mining algorithms and the underlying data sets.

Friedman test's relaxation of those assumption costs statistical power for two-model comparisons, but given the sample size and the number of models there should be less concern about the power. The null hypothesis of Friedman test is all average ranks of the models are equal.

$$H_0 : M_1 = M_2 = \cdots = M_k = \cdots = M_K \tag{6.2}$$

$$H_1 : \text{At least one } M_k, k = 1 \dots K \text{ is different.} \tag{6.3}$$

In order to measure that, error values[33] are replaced with ranks (for K models, ranked from 1 to K) and average of each ranks are considered. For equal values (tied ranks), mean value of ranks are taken. It gives a slight disadvantage for MS method since its error value should be equal to an individual model's error for all data sets and it will always get a tie with an individual model[34] . Friedman test statistic is as follows.

$$\chi_F^2 = \left[\frac{12}{NK(K+1)} \sum_{k=1}^{K} R_k^2 \right] - 3N(K+1) \tag{6.4}$$

where N is the total number of data points (blocks)[35] . R_k^2 is the square of the sum of the ranks ($r_{n,k}$) of model k for contract n ($R_k = \sum_{n=1}^{N} r_{n,k}$). Test statistic is chi-squared distributed with degrees of freedom $K - 1$.

If Friedman Test null hypothesis is rejected, post-hoc Nemenyi test can be applied for pairwise comparisons of the models. Null hypothesis for any two models is their performance is statistically the same. Nemenyi test introduces a critical distance value. If two models' average ranks differ by more than the critical distance, then those two models are significantly different in terms of performance.

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6N}} \tag{6.5}$$

[33] Error terms with signs (i.e. RPE, NPE and SHE) are considered as absolute values and negative values are taken for criteria with the maximization objective (i.e. NPL and HPL).

[34] For lower ranks it can be thought as an advantage. Though, it means using model selection is pointless if it consistently finds itself in lower ranks among individual models.

[35] Complete block design is followed for this test meaning each block consists of only one data point.

where α is the significance level and q_α is the critical value based on the Studentized range statistic divided by $\sqrt{2}$.

Friedman test results show significant power to reject the null hypothesis for all objectives and both underlying assets. Mean rank values for all individual models and respective model selection methods are given in Table 6.13 for SPX and Table 6.14 for NDX.

For SPX, Model Selection results are in line with the error results. MS ranks are high for the reported objectives in the numerical experiments. It supports our claim of MS can pick good individual models for each contract. The biggest failure of the MS method is in the hedge-and-forget type errors (SHE and HPL).

Table 6.13. Nemenyi test rank averages (SPX).

	MS	BS2	BS5	HNA2	HNA5	HNS2
NPL	**3.31**	3.66	3.79	3.37	3.43	3.44
HPL	3.27	3.77	3.75	**3.24**	3.39	3.59
SHE	3.44	3.63	3.82	3.37	3.49	**3.26**
ARPE	**2.63**	3.70	4.05	3.61	3.67	3.33
RPE	**2.80**	3.67	4.03	3.58	3.63	3.29
APE	**2.81**	3.67	4.02	3.58	3.63	3.28

For NDX contracts, The results are similar only with increased difference between MS and the individual models.

Nemenyi critical distances for SPX and NDX are 0.016 and 0.035 respectively. There are very few model pairs which failed to reject Nemenyi test at 95% significance level. Those are HNA5-HNS2 for NPL criterion for SPX; HNS2-HNA2 for SHE and HNA2-MS for HPL for NDX.

Table 6.14. Nemenyi test rank averages (NDX).

	MS	BS2	BS5	HNA2	HNA5	HNS2
NPL	**3.35**	3.63	3.69	3.39	3.45	3.50
HPL	3.35	3.79	3.56	3.34	**3.28**	3.68
SHE	3.47	3.69	3.84	**3.25**	3.51	**3.25**
ARPE	**2.68**	3.82	4.09	3.51	3.65	3.25
RPE	**2.90**	3.79	4.05	3.46	3.61	3.20
APE	**2.88**	3.79	4.05	3.46	3.61	3.20

6.4. Conclusion

In this study, a model selection framework is proposed, under the assumptions that no model is dominant and every model can perform better than the other for some contracts. Model selection framework criteria consists of two parts; the clustering method and the objective (e.g. minimize ARPE). Our main objective is to come up with better results via model selection methods than all the individual models they use.

Conditional Inference Trees (CIT) algorithm is used for clustering and different objectives are used for model selection methods. Five option pricing model parametrizations stemming from Black-Scholes and Heston Nandi GARCH models are chosen as individual models to be used by MS methods. Seven different objectives are used in MS methods based on either pricing or market efficiency metrics. Two different assets (SPX and NDX) are used in experiments to see if results change with different asset price processes.

Overall results indicate that model selection framework improves the option pricing process for most of the objectives, despite its simplicity. Especially, efficiency tests with maximization of naked positioning objective performs well. In addition, successful model selection methods' overall performances are not only dependent on one time success but they display sustainable performance over long time periods.

Despite its success, model selection is not without cost or problems. First of all, model selection requires pricing and delta estimates by all models and extra clustering and selection processes over it. Therefore model selection consumes more computational resources than any individual model. Second, model performance is extremely dependent on individual model performance and number of models. Selection criteria needs improvement to address scalability and model dependence issues.

7. CONCLUSION AND FUTURE RESEARCH

This study examined some very significant issues in empirical option pricing and model assessment process with the main contribution of the introduction of model selection framework.

First, methodologies of empirical option pricing experiments in the literature are presented in Chapter 3 and they are examined in all stages from data preparation to results representation. Even with the small number of studies explicitly examined, methodological differences became apparent. In addition, remarks about the experimentation itself by the other studies are brought to the attention of the reader that had the effect of a discussion. For instance, different opinions of studies from different authors and time periods about the choice of performance metric and its role in the model performance served to understand the underlying reasons for choosing the proper performance metric. Finally, we also presented our methodology used in the experiments in this study.

Chapter 4 claims current pricing error grouping practices widely seen in empirical option pricing studies is not robust enough to hold for further time periods, so any assessment about the model's future performance is actually in question. In addition, the rigid division of moneyness and maturity regions does not necessarily fit the weak and strong parts of each model. So, real high and low performance regions of the model can be averaged out by these predetermined regions.

We propose the employment of data mining algorithms to learn the true performance regions of each model with the objective of better and more robust pricing error grouping for model assessment. K-Means, Support Vector Machine, Decision Tree and Conditional Inference Tree algorithms with the two benchmark static clustering methods are used to determine the best method to group pricing errors. The experiment is repeated for two pricing models (BS and HN), three performance metrics (RPE, ARPE and APE) and data from ten different time periods for proper cross-validation.

Conditional Inference Tree turned out to be the best overall method to cluster pricing errors, but other algorithms are also proved to be prominent in different pricing error types. Static clustering methods, even though they had some success in some data sets and error types, lag in terms of cluster stability.

Chapter 5 discusses the reasons of pricing errors to assess option model performance. Pricing errors, as in the distance from market prices in absolute (e.g. APE) or relative (e.g. RPE) form, are widely used in option pricing studies to assess model performance. Nevertheless, pricing errors implicitly admit Efficient Market Hypothesis and they provide little information about the outcome (payoff) of the contract.

We propose new performance metrics by converting some of the efficiency testing metrics by admitting Efficient Market Hypothesis with the bad model problem. We admit any excess profit (or loss) generated by taking positions in the market is actually the model's lack of covering the risks associated with the positions. This way, the proposed efficiency metrics can now be an alternative to pricing errors to measure the outcome of the model. Three metrics with increasing risk coverage measures are introduced: NPE, SHE and DHE. These metrics are compared with the pricing errors and shown that, similar pricing error values do not necessarily mean similar payoffs.

Finally, in Chapter 6 a model selection framework is proposed to price option contracts. Objective of this study is to find out whether better option price estimates can be found using a set of pricing models than using any of the individual pricing models.

The proposed option pricing method consist of an objective, selection rule and a data mining algorithm (i.e. Conditional Inference Tree) as a selection support. Past performances of individual models are grouped and out-of-sample estimates for each contract and individual model are made. Model selection then selects the models with the least predicted "error" (according to the objective) for each contract. Then the model selection can be assessed just as any option pricing model based on its picked price estimates. Objective function of the model selection can be the minimization of a pricing metric (e.g. ARPE), efficiency error (e.g. NPE, SHE) or an efficiency test to maximize profit and loss. Numerical experiments with different assets (i.e. SPX and NDX), time periods (2009-2013) and objectives show that model selection is a good and sustainable way of pricing options.

Even though model selection method used in this study fails to deliver the best performance all the time, there are still many possibilities for improvement. For starters, improving the selection procedure and using different data mining algorithms for different objectives are the most promising routes to better performance. Though, it cannot be claimed for the individual models. An option pricing model requires an increase in complexity in order to increase its performance, given parameters fit optimally. Improvement with model complexity eventually hits a performance barrier and suffers from overfitting. In addition, complex models' parameters are harder to optimize. Finally, if another good pricing model emerges, model selection can easily add that model to its model set.

In the future, model selection can be improved to be even more sustainable. First improvement would be to the scalability issue. Model selection should be able to draw information from more individual models without being detrimental to performance. Improvement can also be achieved by improving the clustering method behind the selection process by either refining it or changing the algorithm with the objective. Also, model selection is only applied to European type options and might need further changes to handle American, Barrier or exotic options (i.e. Asian, Bermudan).

REFERENCES

Ahmad, R. and Wilmott, P. (2005). Which free lunch would you like today, sir?: Delta hedging, volatility arbitrage and optimal portfolios. *Wilmott*, pages 64–79.

Alajbeg, D., Bubas, Z., and Sonje, V. (2012). The efficient market hypothesis: problems with interpretations of empirical tests. *Financial Theory and Practice*, 36(1):53–72.

Bakshi, G., Cao, C., and Chen, Z. (1997). Empirical Performance of Alternative Option Pricing Models. *The Journal of Finance*, 52(5):2003–2049.

Bakshi, G., Cao, C., and Zhong, Z. (2012). Assessing Models of Individual Equity Option Prices. SSRN Scholarly Paper ID 2038551, Social Science Research Network, Rochester, NY.

Bams, D., Lehnert, T., and Wolff, C. C. P. (2009). Loss Functions in Option Valuation: A Framework for Selection. *Management Science*, 55(5):853–862.

Barone-Adesi, G., Engle, R. F., and Mancini, L. (2008). A GARCH Option Pricing Model with Filtered Historical Simulation. *Review of Financial Studies*, 21(3):1223–1258.

Bates, D. S. (2003). Empirical option pricing: a retrospection. *Journal of Econometrics*, 116(1–2):387–404.

Bhattacharya, M. (1983). Transactions data tests of efficiency of the Chicago board options exchange. *Journal of Financial Economics*, 12(2):161–185.

Black, F. (1975). Fact and Fantasy in the Use of Options. *Financial Analysts Journal*, 31(4):36–41.

Black, F. (1976). Studies of Stock Price Volatility Changes. In *Proceedings of the 1976 meetings of the business and economic statistics section.*, pages 177–191. American Statistical Association.

Black, F. and Scholes, M. (1972). The valuation of option contracts and a test of market efficiency. *The Journal of Finance*, 27(2):399–417.

Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):637–654.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bouchaud, J.-P., Matacz, A., and Potters, M. (2001). Leverage Effect in Financial Markets: The Retarded Volatility Model. *Physical Review Letters*, 87(22).

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Broadie, M. and Detemple, J. B. (2004). ANNIVERSARY ARTICLE: Option Pricing: Valuation Models and Applications. *Management Science*, 50(9):1145–1177.

Burke, E. K., Hyde, M., Kendall, G., Ochoa, G., Özcan, E., and Woodward, J. R. (2010). A Classification of Hyper-heuristic Approaches. In Gendreau, M. and Potvin, J.-Y., editors, *Handbook of Metaheuristics*, number 146 in International Series in Operations Research & Management Science, pages 449–468. Springer US.

Chorro, C., Guégan, D., and Ielpo, F. (2012). Option pricing for GARCH-type models with generalized hyperbolic innovations. *Quantitative Finance*, 12(7):1079–1094.

Christoffersen, P., Heston, S., and Jacobs, K. (2009). The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work So Well. *Management Science*, 55(12):1914–1932.

Christoffersen, P. and Jacobs, K. (2004a). The importance of the loss function in option valuation. *Journal of Financial Economics*, 72(2):291–318.

Christoffersen, P. and Jacobs, K. (2004b). Which GARCH model for option valuation? *Management Science*, 50(9):1204–1221.

Constantinides, G. M., Jackwerth, J. C., and Perrakis, S. (2007). Mispricing of S&P 500 Index Options. *Review of Financial Studies*, 22(3):1247–1277.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Coval, J. D. and Shumway, T. (2001). Expected option returns. *The Journal of Finance*, 56(3):983–1009.

Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7(Jan):1–30.

Duan, J. C. (1995). The GARCH option pricing model. *Mathematical finance*, 5(1):13–32.

Duan, J. C. (1999). Conditionally fat-tailed distributions and the volatility smile in options. *Rotman School of Management, University of Toronto, Working Paper.*

Dumas, B., Fleming, J., and Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *The Journal of Finance*, 53(6):2059–2106.

Eberlein, E. (2001). Application of Generalized Hyperbolic Lévy Motions to Finance. In Barndorff-Nielsen, O. E., Resnick, S. I., and Mikosch, T., editors, *Lévy Processes*, pages 319–336. Birkhäuser Boston.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4):987.

Engle, R. F. and Ng, V. K. (1993). Measuring and Testing the Impact of News on Volatility. *The Journal of Finance*, 48(5):1749–1778.

Esscher, F. (1932). On the probability function in the collective theory of risk. *Scandinavian Actuarial Journal*, 1932(3):175–195.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2):383–417.

Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance*, 46(5):1575–1617.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, 49(3):283–306.

Fan, J. and Mancini, L. (2009). Option Pricing With Model-Guided Nonparametric Methods. *Journal of the American Statistical Association*, 104(488):1351–1372.

Galai, D. (1977). Tests of market efficiency of the Chicago Board Options Exchange. *The Journal of Business*, 50(2):167–197.

Galai, D. (1978). Empirical tests of boundary conditions for CBOE options. *Journal of Financial Economics*, 6(2–3):187–211.

Gerber, H. U. and Shiu, E. S. W. (1994). Option pricing by Esscher transforms. *Transactions of the Society of Actuaries*, 46.

Grossman, S. J. and Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, 70(3):393–408.

Guégan, D., Ielpo, F., and Lalaharison, H. (2013). Option pricing with discrete time jump processes. *Journal of Economic Dynamics and Control*, 37(12):2417–2445.

Hastie, T., Tibshirani, R., and Jerome, F. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2 edition.

Heston, S. L. and Nandi, S. (2000). A closed-form GARCH option valuation model. *Review of Financial Studies*, 13(3):585–625.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science, National Taiwan University.

Hull, J. (2012). *Options, futures, and other derivatives.* Prentice Hall, Boston, 8th ed edition.

Jarrow, R. (2012). The third fundamental theorem of asset pricing. *Annals of Financial Economics*, 07(02):1250007.

Jarrow, R. (2013). Option Pricing and Market Efficiency. *The Journal of Portfolio Management.*

Jarrow, R. A. and Larsson, M. (2012). The Meaning of Market Efficiency. *Mathematical Finance*, 22(1):1–30.

Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economics*, 6(2–3):95–101.

Lehar, A., Scheicher, M., and Schittenkopf, C. (2002). GARCH vs. stochastic volatility: Option pricing and risk management. *Journal of Banking & Finance*, 26(2–3):323–345.

Lo, A. W. (2007). Efficient Markets Hypothesis. In *The New Palgrave: A Dictionary of Economics, L. Blume, S. Durlauf, eds., 2nd Edition*. Palgrave Macmillan Ltd.

Mackenzie, D. (2008). *An Engine, Not a Camera: How Financial Models Shape Markets*. The MIT Press, 1st edition edition.

Merton, R. C. (1973). Theory of Rational Option Pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58(302):415–434.

Orbay, B. (2016). Model Selection Framework GitHub Code Repository. `https://github.com/berkorbay/msf`. Accessed at May 16, 2016.

Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2):41–49.

Schachermayer, W. (2008). The Notion of Arbitrage and Free Lunch in Mathematical Finance. In Yor, M., editor, *Aspects of Mathematical Finance*, pages 15–22. Springer Berlin Heidelberg.

Schoutens, W. (2003). *Lévy Processes in Finance: Pricing Financial Derivatives*. Wiley Series in Probability and Statistics. Wiley.

Shreve, S. E. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer.

Siu, T. K., Tong, H., and Yang, H. (2004). On pricing derivatives under GARCH models: a dynamic Gerber-Shiu approach. *North American Actuarial Journal*, 8:17–31.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Stentoft, L. (2011). What we can learn from pricing 139,879 Individual Stock Options. *Available at SSRN 1975779*.

Tankov, P. and Cont, R. (2003). *Financial Modelling with Jump Processes*. Chapman and Hall/CRC, Boca Raton, Fla, 1 edition edition.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

APPENDIX A: OPTION PRICING FUNDAMENTALS

This chapter includes brief information about the fundamentals of option pricing. Expanded information can be found in textbooks such as Hull (2012).

A.1. What is an option?

An option is derivative product which is a contract on the future state of the underlying asset. The distinctive feature of an option, from other derivative product types (e.g. futures), is only one side (writer of the contract) is bound by the contract. There are many different kinds of options, some of which are; European, American, Bermudan, Asian, Lookback, Barrier etc.

Options are one of the most actively traded derivative contracts both in exchanges and over the counter. CBOE Market Statistics, in 2014, option trade volume in CBOE was above 570 billion dollars and approximately 380 billion dollars of the that volume were on S&P 500 contracts (SPX).

The common definition of a European option is: "A European Call/Put option is a contract which gives the owner the right but not the obligation to buy/sell the underlying asset at a predetermined price at a predetermined time." This study is almost completely about European type options, though it can be generalized to other contract types.

European Options are only profitable if the spot price at the time of exercise is higher/lower than the exercise price for a call/put option.

There are several parameters of or related to an option contract: Spot price (S_0) is the current price of the underlying asset, strike price (K) is the exercise price determined by the contract, maturity ($T - t_0$) is the time remaining to expiration date (also determined by the contract) and moneyness (S_0/K) is simply the relative position of spot price to moneyness.

Maturity groups are labeled as short term, middle term and long term; but boundary values usually depend on the study.

Moneyness groups are identified by the criteria of yielding profit in case of immediate exercise. They are labeled as deep-in-the money (DITM), in-the-money (ITM), at-the-money (ATM), out-of-the-money (OTM) and deep-out-of-the-money (DOTM). Boundaries on the levels are also dependent on the study but ATM options are always around moneyness level 1.

In addition, same moneyness level is adversely labeled for calls and puts; a DOTM labeled moneyness level for a call (e.g. 0.85) option is DITM for a put.

A.1.1. An example option pricing problem

The European Option pricing problem can be illustrated with a simple example. Suppose a European Call (buy) option contract is to be written on $GARAN (Garanti Bank stock symbol) on April 22, 2016. Spot price of $GARAN is 8.45 TL. The specifications of the contract is as follows. Strike price (predetermined buy price of the contract) is 8.5 TL and expiration date (predetermined date) is June 21, 2016.

Parameters of the problem are

Spot price (S_0): 8.45 TL
Strike price (K): 8.5 TL
Maturity ($T - t_0$): 2 months (time before expiration)
Moneyness (S_0/K): $7/7.5 = 0.99$

Suppose on the exercise date, November 30, 2015, spot price of $GARAN is 10 TL. The contract will give the owner the opportunity to exercise it and buy $GARAN at 8.5 TL and sell it to the market at 10 TL; therefore making a profit of 1.5 TL per share. But, if the spot price is below 8.5 TL, say 7 TL, the contract will be worthless. Because it will always be better to buy the stock from the market for 7 TL, instead of buying from the writer for 8.5 TL. The payoff of the European call option can be formulated as $(S_T - K)^+$ or $max\{S_T - K, 0\}$.

If it were a put option with the same parameters, the outcome would be the reverse. If the exercise spot price (S_T) would be above 8.5, say 10 TL, the contract would be useless. Else, say 7 TL, it would be possible to buy the stock from the market for 7 TL and sell it to the writer for 8.5TL, therefore make a profit of 1.5 TL per share. The payoff of the European put option can be formulated as $(K - S_T)^+$ or $max\{K - S_T, 0\}$.

See Figure A.1 for the payoff behavior of the put and call options.

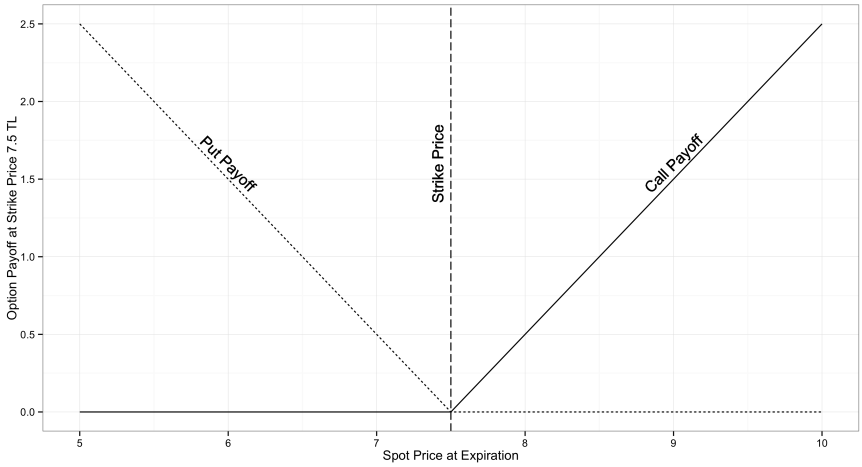

Figure A.1. Payoff of Call and Put options.

The question is simple: “What should be the fair value of the contract?”. Due to the uncertain nature of the future prices of an asset, there are many answers and models.

American options are quite similar to European options with the single difference of the opportunity to exercise the option at ‘or’ before the expiration date. For the others, Hull (2012) covers the formulation of many different option types.

A.1.2. Fair Value

Fair value or fair price (used interchangeably in this study) of an option is a price level which neither party of the contract have a greater probability of making a profit than the other, including time value of the money (i.e. risk-free discounting).

Market price is often quoted as the fair prices. This is especially the claim for underlying stock market by the Efficient Market Hypothesis. Fama (1970) provide details about market efficiency.

However, options, being derivative products on underlying assets and having their own market (or exchange) where prices are not necessarily 'in synchronization' with the underlying market. Although the problem is approached differently in many studies, occasional mismatches are admitted (for examples see REF HERE). It might be due to 'forward looking nature' of options market or some other reason. Therefore since efficiency of options market is at question (if underlying asset's is not), so is options market prices being accepted as fair prices.

A.2. Volatility

Volatility is the main parameter of uncertainty in the asset log-return process which is not observed directly. Therefore it should be estimated. The most basic construction is as the standard deviation of past log-returns of the underlying asset, though there are different formulations of volatility. Every formulation is devised obtain more information from the price process in order to come up with better prices and hedging schemes.

A.2.1. Historical Volatility

Historical volatility basically uses past innovations over a given time period. The sample standard deviation of log-returns is the common calculation of the historical volatility.

$$Y_t = log\left(\frac{S_{t+1}}{S_t}\right) \ , \ \bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t \tag{A.1}$$

$$\sigma_{historical} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \approx \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{A.2}$$

S_t is the price of the security, Y_t is the log-return, n is the time period (number of observations) and $\bar{Y}$ is the average of log-returns. With large enough sample, replacing $n-1$ with n will not make a big difference. Some formulations take $\bar{Y} = 0$ to further simplify the calculation process.

Historical volatility can be measured with almost any time interval (e.g. daily, weekly, hourly) and can be converted to other time intervals by adjusting with a time coefficient assuming underlying distribution is normal or another distribution with similar scalability property. For example, to convert daily historical volatility to yearly (annual) historical volatility it is generally assumed[36] that there are 252 trading days and the daily volatility is multiplied by the square root of that value.

$$\sigma_{historical}^{annual} = \sigma_{historical}\sqrt{T} \tag{A.3}$$

A.2.2. Implied Volatility

While historical volatility uses underlying asset's log-returns, implied volatility is the volatility measure extracted from option prices usually the market quotes. Derivation of implied volatility is predominantly done by reversing Black-Scholes formula (explained in detail in Section 2.1).

[36] Some studies use different trading days per year

A.2.2.1. Volatility Smile (Skew) and Surface. An interesting phenomena observed in options markets especially after 1987 crash (see Mackenzie (2008) for details), is the volatility smile (smirk/skew). Volatility smile is observed on the contracts on the same asset with the same maturity but different moneyness values. Normally, implied volatilities of those contracts are expected to be very similar and the implied volatility line to be flat. Because, strike price is not actually an element which affect the volatility parameter. But, it is commonly observed that implied volatility values display a skewed pattern resembling a “smile/smirk”.

One of the main reasons of observing this smile is about the assumptions about the underlying asset’s log-return distribution. Since the actual distribution is observed to have higher tails and kurtosis than the GBM, volatility smile appears. Hull (2012) provide further details about volatility smile and its reasons.

DOTM puts as well as DITM calls usually yield higher implied volatilities. Implied volatility is expected to be lowest at ATM leevels. See Figure A.2 for a representation.

Volatility Surface is the volatility smile with the added dimension of time to maturity. See Figure A.3 for the graphical representation.

A.2.3. Non-Constant Volatility

Constant volatility assumption (i.e. volatility stays the same at least during the lifetime of the option) is a strong one. It is frequently questioned even by Fischer Black on Black (1975). It is empirically proven volatility tends to change, cluster and there is a negative correlation between volatility and stock price process. Bouchaud et al. (2001) provide tests and elaborations and they conclude negative correlation is more strongly observed in index prices.

Predicting the behavior of volatility in time is no easy task. Though, it can be said that mere inclusion of a varying or stochastic volatility is an important improvement

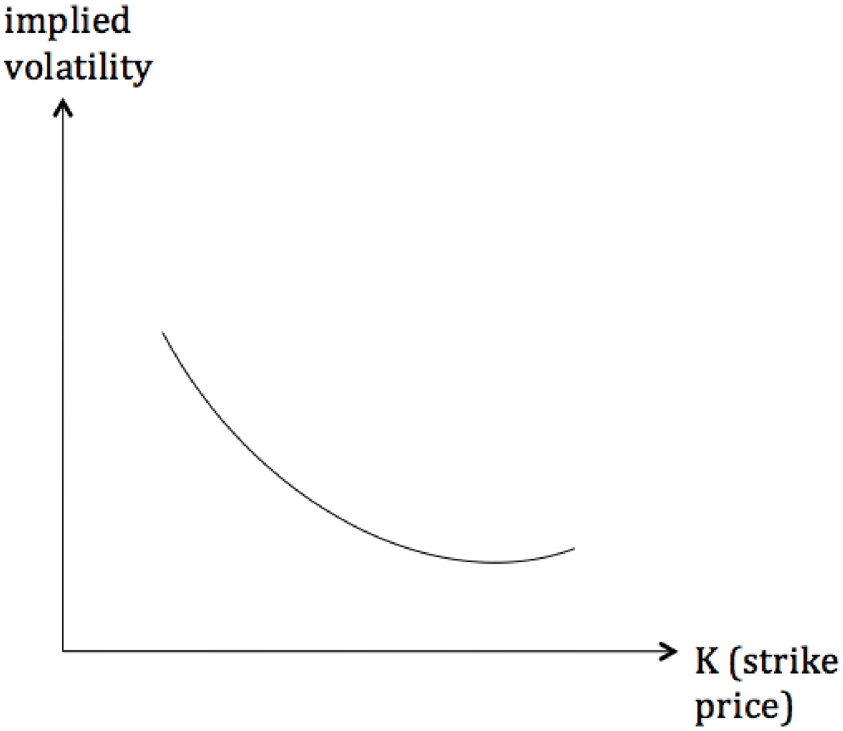

Figure A.2. Volatility smile.

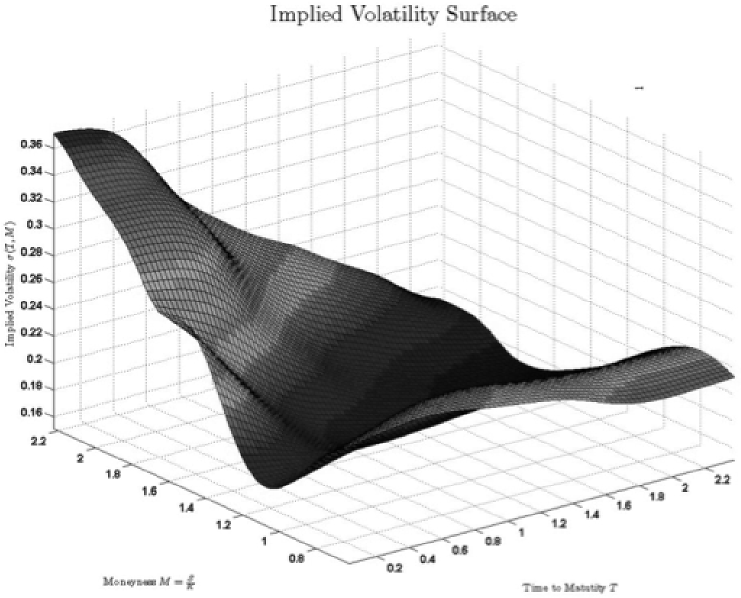

Figure A.3. Volatility surface.

over models with constant volatility assumptions.

Volatility clustering is simply the behavior of high volatility periods following high volatility periods and low volatility periods following low volatility periods. Figure A.4 shows the daily log-returns of Google ($GOOG) stock between 2007 and 2012. The effects of 2008 crisis (around data point 500) result as an increase in volatility and the calm period after the high volatility period can be observed.

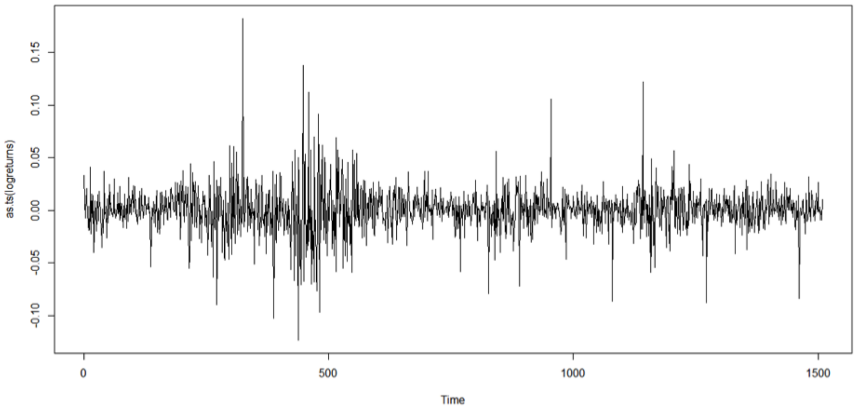

Figure A.4. Log-returns of Google stock price between 2007 and 2012.

Volatility and asset price process being adversely affected, also known as leverage effect, is an observed phenomenon especially in indexes such as SPX and NDX. Although by leverage, it is initially implied by Black (1976) that the company being highly leveraged in terms of debt equity ratio, the relationship is not clear. Bouchaud et al. (2001) report that leverage effect and individual firm leverages have little in common.

A.3. Arbitrage

Arbitrage in the financial markets, simply put, is the opportunity to exploit price differences of the same asset in different markets. To define the concepts in this subsections, Schachermayer (2008) work is followed generally. For instance, if the exchange rate of USD/TRY is 2.10 in market A and 2.1001 in market B the arbitrage opportunity here is to buy in market A and sell in market B at the same time. This way in each exchange a 0.0001 profit is made without any risk of losing money.

Formal definition of arbitrage is "an investment with zero probability of losing money with positive probability of making profits without net investment of capital." This opportunity is actually an anomaly in the financial markets but is reported to appear on occasion. Real arbitrage opportunities is usually more complex in nature consisting of several markets and several assets; but frictions (i.e. trading costs) in real world makes it harder to find and exploit arbitrage opportunities.

For the option prices, an example arbitrage opportunity can be defined as the risk-free rate being higher (lower) than the upper (lower) bound of the option payoff rate. So if the risk free rate is higher than the upper bound of the option payoff rate, one can short the option and invest in the riskless asset (i.e. a bond). Otherwise if the risk fee rate is lower than the lower bound of the option payoff rate one can borrow from the money market at the risk free rate and long the option contract therefore making money without risk or net capital. In the earlier days, there were also arbitrage opportunities by just using options, though those opportunities are much less common (see Mackenzie (2008)).

Though, while modeling the financial systems, markets are assumed that they do not allow arbitrage. Otherwise, the problem would obviously be trivial.

A.4. Risk neutral measure (Martingale)

Prices of financial products usually incorporate risk premiums. But detecting and measuring risk preferences of market players, therefore the magnitude of those premiums is not an easy task.

Risk neutrality is the assumption of investors being unwilling to increase their expected returns in exchange for increased risk. A world where all investors behave in the norms of risk neutrality is called a risk neutral world. Also a fair price found in the risk neutral world is a fair price in all other 'worlds'.

Despite seemingly an incorrect description of the real world, risk neutral pricing is actually plausible with the possibility of perfect and continuous hedging with the underlying asset.

Under the risk neutral world assumption, the expected return rate of an investment is equal to the risk-free rate and the fair price of an option is the expected payoff of that option discounted by the risk free rate.

Detailed information on risk neutral measure and risk neutral world can be found in Hull (2012).

If perfect hedging is possible (i.e. all risks can be hedged with the available instruments) then the market is called a complete market (see Shreve (2004)). Complete market's relevant feature is they contain a unique risk neutral measure. If there are risks dependent on unknown parts (i.e. stochastic volatility) and there are no other securities to hedge the risk, then that market is called incomplete.

Multiple martingale measures exist for given models and also not each martingale is defined for all model parameters. For instance, Esscher Transform martingale measure under Lévy processes with GHYP distribution is only defined for specific combinations of the parameters and the risk free rate. In this study three different martingale measures will be covered: (Generalized) Local Risk Neutral Valuation Relationship ((G)LRNVR), Mean Correcting Martingale Measure (MCMM) and the Esscher Transform. Further details on different martingale measures are given in Chapter 3.

A.5. Dividends and Stock Splits

Dividends are payments from individual firms to stock holders. Although it is generally known to be in cash form, dividends in terms of shares are also known. There are two types of dividends; regular and special. Regular dividends are announced long time before and distributed in a periodic time period within a plan. For instance, Company A declaring they will be distributing dividends of 0.5$ in each quarter in the following year is the announcement of a regular dividend. Special dividends, on the other hand, are usually one off and announced only a short time before the dividend occurs.

The empirical relevance of dividends is although they are discounted from stock prices at the ex-dividend date, the option prices are not adjusted for regular dividends. For European Options, regular dividends are usually discounted from initial asset prices or assumed as continuous. Details of dividends in pricing process are discussed in Chapter 3.

A.6. Risk-Free Rate

Risk free rate is the rate of return from an assumed 'riskless asset'. Examples of riskless asset can be bonds, treasuries and swaps. Details of risk free rate in pricing process are discussed in Chapter 3.

A.7. Delta Hedging

Hedging, in general terms, is the covering of the risks of an investment by taking positions in other assets with the objective to reduce substantial unexpected loss.

Hedging in options is taking positions in different instruments based on the underlying asset. Delta hedging in options include taking a position in the underlying asset itself proportionate to the option position to offset the movements from the underlying asset. Delta calculation of Black-Scholes model is given in Section 2.1.

Delta hedging is not the only hedging method for options. For instance Gamma hedging is used to offset the changes in the underlying asset's delta.

APPENDIX B: SOURCE CODE AND SOFTWARE

R programming language is extensively used in this study. Source code required to replicate the calculations can be found in Orbay (2016). GitHub repository will also be updated to further accommodate experiments and ease replication of results. Following R packages and their depending packages are used in computational experiments, i/o operations and graphics: fOptions, party, e1071, rpart, tidyr, plyr, reshape2, ggplot2, grid, gridExtra, RcolorBrewer, scales, xlsx, readxl, xtable, lazyeval, NbClust, foreach, doMC, devtools and Quandl.