The London School of Economics and Political Science

Rational Bubble, Short-dated Volatility Forecasting and Extract More from The Volatility Surface

Yiyi Wang

A thesis submitted to the Department of Finance of the London School of Economics for the degree of Doctor of Philosophy, London, September 2011

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

Abstract

The thesis covers three main chapters. The first chapter (which is a joint work) we develop a theoretical model of rational bubble. In equilibrium, a bubble can persist until it bursts following an exogenous shock, even when all the agents are aware of the bubble and that it will burst in finite time. Applying the model in the context of the sub-prime mortgage crisis, we argue that excessive sub-prime lending behaviour may be sensible with the introduction of securitization. We thus provide a rational explanation for the housing bubble and the dramatic increase in sub-prime default rates.

In the second chapter I conduct empirical short-dated volatility forecasting in foreign exchange, and carry out a realistic volatility swap trading strategy based on the forecast. Additional to applying regime-switch technique, I propose a double-step approach to circumvent the disadvantage of employing GARCH-type model in the high frequency data in FX market, so that it can separate the effect of intraday/intraweek seasonality and pre-scheduled macroeconomic data releases from the underlying data process. By keeping a battery of models and rotating among them, the forecast ability gets significantly enhanced and the trading profit is pronounced even after considering transaction cost.

In the third chapter I explore the cross-sectional predictive power of the most important two factors in the implied volatility surface - skew and term structure - at individual firm level. Stocks with lower implied volatility skew and higher implied volatility term structure outperform the comparative peers. In particular, the interaction between these two factors reinforces the predictive power, and the return of a weekly long-short strategy can be improved greatly with the attachment of term structure on skew. By sorting firms based on skew and term structure one may also be able to pick up takeover targets and seize the big positive premium.

Acknowledgements

I am deeply grateful to my doctoral supervisor, Professor Christopher Polk, for his invaluable guidance, great encouragement and help during the past four years. I would also like to thank Professor David Webb, Professor Dimitri Vayanos, Professor Daniel Ferreira, Dr. Xiaoji Lin and my PhD friends Xiaoxia Hao, Zijun Liu for all the discussions, suggestions and tremendous support. My PhD was funded by LSE John C. Phelan Scholarship and I genuinely appreciate his generosity. I want to acknowledge a special debt of gratitude to Caio Natividade, who offered his extraordinary insight of financial industry, and helped me to proofread the thesis. My housemates Hui Zhang, Maria do Mar Pereira, Yiheng Lu have accompanied me through the happy and hard time. Thanks for all the jokes, laughs and touching words. And in the end, without my parents' unconditional support and understanding I would not have walked so far. Thanks for everything.

Table of Contents

A Rational Bubbles Model and Application In Sub-prime Mortgage Crisis

Zijun Liu
London School of Economics

Yiyi Wang
London School of Economics

November 27, 2011

Abstract

We develop a theoretical model of rational bubble. In equilibrium, a bubble can persist until it bursts following an exogenous shock, even when all the agents are aware of the bubble and that it will burst in finite time. Applying the model in the context of the recent sub-prime mortgage crisis, we argue that excessive sub-prime lending behavior may be rational with the introduction of securitization. The process is much like a Ponzi scheme where lending is profitable as long as investors continue to invest, and vice versa. We thus provide a rational explanation for the housing bubble and the dramatic increase in sub-prime default rates.

Keywords: Bubbles, Ponzi Scheme, Sub-prime Mortgage
JEL Classifications: D82, D84, G10, G21

1 Introduction

Bubbles have been the central theme of the financial market in the past decade. Following the burst of the dot-com bubble in 2001, there have been strong deviations of asset prices from fundamentals across many different markets and countries, such as the US housing bubble, the Chinese stock market bubble, the worldwide commodity bubble, and so on. Especially, the collapse of the US housing bubble and the sub-prime mortgage market has eventually lead to the recent global financial crisis.

However, the existence of bubbles still remains theoretically controversial. For example, the backward induction argument and the transversality condition preclude the existence of bubbles in finite-horizon models and infinite-horizon inter-temporal models respectively. Moreover, the efficient market hypothesis implies that the presence of sufficiently many well-informed arbitrageurs will ensure that large deviations from fundamental values are not possible. Nonetheless, Abreu and Brunnermeier (2003) showed that bubbles can persist given dispersion of opinion among rational arbitrageurs and the presence of behavioral traders. Hence the backward induction argument could fail if agents do not have common knowledge about the timing of the burst of the bubble.

We develop a theoretical model of bubbles and argue that speculative bubbles can be sustained even if all agents are rational. Our result relies on the assumption that the bubble will burst following an exogenous shock and agents are only partially informed of the timing of this shock. In our model, the investors are fully aware that the asset price has departed from fundamentals since the start of the bubble, therefore they only have incentives to buy the asset if they expect to sell the asset later at a high price. Ideally each investor would like to sell the asset just before the bubble bursts in order to maximize profits. Because investors have heterogeneous beliefs about the time of the burst, they would have different exit strategies, hence the backward induction argument does not apply. This intuition is very similar to Abreu and Brunnermeier (2003), however our model is different in several ways. Most importantly, there is no behavioral traders in our model, and agents are fully rational in the sense that they are well aware of the bubble since the very beginning rather than being sequentially informed of the mis-pricing.

Our model is a very simplified illustration of the above intuition. In the baseline model, there are two investors who can trade one virtually worthless asset in each period. Each investor needs to pay a fixed premium above the market price when he wants to buy the asset from the other. The process will terminate following an exogenous shock taking place at an unknown date. Each investor will receive a noisy signal of this date but does not know the signal of the other. That is, the investors have heterogeneous beliefs about the date when the bubble will burst. The uncertain timing of the burst of the bubble together with heterogeneous beliefs ensure that the model does not have problems arising from backward induction or transversality condition.

In equilibrium, the investors will ride the bubble until some point according to their own exit strategy which depends on their private beliefs. Although this is a zero-sum game, each investor has the incentives to ride the bubble as far as he could, because the potential gain from winning the game is increasing as the bubble grows. The intuition is that, no matter how large the bubble is, there is always a good chance that my opponent will exit after me in which case I will make a larger profit. Thus the investors would gamble for profits while being perfectly aware of the bubble. Moreover, we extend the model with noisier signals and different strategies of the investors. We show that bubbles tend to last longer in illiquid markets.

We then apply the model in the context of the recent housing bubble and sub-prime crisis in the US in order to derive more economic implications. Although the results of the model can be naturally extended to many speculative bubbles we have experienced in the financial market, our application of the model in the sub-prime mortgage market is not a straightforward extension. In particular, the housing bubble in our model is not supported directly by speculative buyers but through excessive lending and securitization behavior of the mortgage lenders.

The sub-prime mortgage industry has been partly blamed for the emergence and collapse of the recent US housing bubble. Loose lending standards have provided excessive credit to the sub-prime borrowers who have a poor credit history and ability to repay their loans. It has been argued that, as a result of the fact that the sub-prime mortgage lenders securitize their loans and sell the securitized products to the investors, the lenders have shifted the risks associated with sub-prime lending away and do not have enough incentives to maintain the quality of the borrowers, hence we observe the boom and collapse of the sub-prime industry. The dramatic increase in the default rates in the industry after the housing bubble burst might suggest that investors have underestimated the risk inherent in the mortgage-backed assets that they bought. However, this cannot stand if we assume that the investors are rational and anticipate the moral hazard issues ex-ante.

Our paper tries to establish a different perspective by extending our model of rational bubbles. Because sub-prime borrowers frequently incur difficulties to repay, the payoff from lending largely depends on the price of the houses that will be repossessed by the lenders in case of default. If house prices always rise, there would be virtually no risk associated with lending. House prices would keep rising if the lenders can fuel the demand by keeping lending generously to the borrowers and refinancing existing borrowers with financial problems. Thus sub-prime lending would become profitable and rational with rising house prices, or a housing bubble.

House prices cannot rise forever. When the bank stops lending following a liquidity shock[1], the housing bubble will burst and house prices will plummet due to foreclosures on sub-prime loads

[1]Basically we need some exogenous shock to terminate the bubble process, and we did not feel it is crucial to specify the exact nature of the shock. While we have used a liquidity shock in this model, we believe that changes in policy, interest rates and other exogenous factors could all trigger the burst of the bubble.

accumulated over the years. Thus sub-prime lending would become unprofitable before the burst of the housing bubble and the bank would not lend in the very first place by backward induction. However, the introduction of securitization would enable the bank to transfer the risk to investors and gamble in the same way as in the baseline model. That is, banks have incentives to lend as long as investors are willing to buy the mortgage-backed securities (MBS), and investors have incentives to buy the MBS as long as banks are willing to refinance borrowers. Because of the uncertain timing of the shock, a bubble in the house price can persist with the same intuition in the baseline model.

The rest of the paper is organized as follows. Section 2 is a brief literature review. Section 3 introduces the baseline model as well as some extensions and implications. Section 4 describes and discusses the application of the model in the sub-prime mortgage market. Finally Section 5 concludes.

2 Related Literature

Allen et al. (1993) provided necessary conditions for a bubble to occur. They say that a rational expectations equilibrium exhibits a strong bubble if the price is higher than the dividend with probability one. They show that in a finite-period general equilibrium model in which a bubble is possible, each agent must have private information in the period and state in which the bubble occurs, and the agents' trades cannot be common knowledge. Allen and Gorton (1993) show that a bubble can exist because the fund managers without private information will churn at the expense of uninformed investors, who cannot observe the skill of the fund managers. These papers assume that all agents are fully rational.

Abreu and Brunnermeier (2003) developed a model of bubbles in which rational arbitrageurs interact with boundedly rational behavioral traders. They show that the inability of arbitrageurs to temporarily coordinate their selling strategies together with the presence of behavioral traders results in the persistence of bubbles over a substantial period. In Abreu and Brunnermeier (2003), arbitrageurs sequentially become aware of the bubble and hence are uncertain about the timing of the burst of the bubble. The incentives of the arbitrageurs to time their exits as close to the burst of the bubble as possible lead to the persistence of the bubble. Sato (2008) extended the paper and show that the presence of relative ranking tournament among fund managers affects their incentives to attack or instead ride asset bubbles.

Our model is different from the above in the following ways. All agents are fully rational in our model. The bubble will end in finite time with probability one, but the length of the bubble is not bounded above. Unlike Allen et al. (1993), the asymmetric information does not come from the period and state in which the bubble occurs, but the time at which the bubble will burst. That is, all agents are fully aware of the bubble from the beginning, but they do not know each other's beliefs about the timing of the burst of the bubble. Moreover, the agents' trades are common knowledge and there is no sequential awareness.

On the application in the sub-prime mortgage market and securitization, there is a rich literature focusing on the role and implication of securitization. It remains controversial whether lenders exploit asymmetric information to sell riskier loans into the public markets or retain riskier loans in response to regulatory capital incentives (regulatory capital arbitrage). Calem and LaCour-Little (2004) argue that, for most mortgage loans, existing regulatory capital levels are too high, creating an incentive to securitize the least risky loans. In addition to regulatory capital rules favoring securitization, the presence of information asymmetries also encourages securitization. DeMarzo and Duffie (1999) using a liquidity based model of securitization show that if the issuer does not wish to retain any portion of the mortgage backed security, then she should sell only those loans having the lowest degree of asymmetric information into the pool and retain those loans with high degree of asymmetric information.

Our paper innovates in arguing that securitization enabled the bank to gamble with the investors and benefit from lending excessively to sub-prime borrowers and riding the bubble. This result is consistent with Keys et al. (2008) who empirically examine the default probability of portfolios with different securitization levels and find that portfolios with higher securitization volume are like to have high default rates, hence implying that the bank may have incentives to loosen lending standard in presence of securitization. Furthermore, Coval et al. (2007) show that many structured finance instruments can be characterized as economic catastrophe bonds that default only under severe economic conditions. This is consistent with our model in which the MBS would only have a negative return when the bubble bursts with a small probability. The difference is that Coval et al. (2007) blame the rating agencies for the mis-pricing of these instruments whereas we show that a bubble can persist without such intermediation problems.

3 Baseline Model

Using a parsimonious model, we will show that rational agents with heterogeneous beliefs will ride the bubble and gamble on future profits given high enough incentives. The model is very simple but illustrate the intuition why rational investors would speculate on over-priced assets. The basic setup of the model is described below.

3.1 Basic Setup

Consider a discrete-time model with infinite time periods 0,1,2... Suppose there are two investors: investor A and investor B, each with an initial wealth of W_0. They will both receive income in each period[2]. The amount of income they receive in period t is k^t.

There is an asset that has a fundamental value of 1 and can be traded between the two investors. Assume that the asset can only be bought at a multiple k of its current price[3]. So in period t, one investor can decide whether or not to buy this asset from the other at a price of k^t.

The above assumption is important because it provides incentives for the investors to ride the bubble. We can think of it as price impact of trades in the reality.

Suppose that both investors will experience a liquidity shock at a random date T and they will consume all their wealth at this date (the asset itself has a consumption value of 1). Both investors will maximize the expected value of this consumption, which we denote by C. Assume that investors are risk-neutral and have zero discount rate. Hence the expected utility of each investor is the expectation of his final consumption at the date of liquidity shock.

Finally we denote the date of liquidity shock by T and assume that it will be drawn ex-ante from a Geometric distribution[4] with parameter p. Each investor will receive a noisy signal s about T. The signal can be $T-1$ or $T+1$ with probability $\frac{1}{2}$ respectively[5]. This assumption of heterogeneous beliefs was necessary for a bubble equilibrium [6] to exist (see Appendix A.1 for a detailed proof).

3.2 Characterization of Equilibrium

Proposition 3.1. *Given the above assumptions, there exists a Nash equilibrium such that each investor buys the asset until their signaled date, i.e. if an investor received a signal s, he will buy the asset at period t if and only if $t < s$.*

[2]This is to ensure that they have enough funds to sustain the bubble in the long run. That is, we need the investors to have finite capital but not be constrained by capital in the long run. Alternatively, we can assume that they have unlimited access to short-term financing.

[3]Here we show that a particular bubble price path with a constant inflator is possible given certain parametric restrictions. Any other price paths with time-dependent inflator are possible as long as those restrictions are satisfied.

[4]We need distributions with probability densities that do not converge to zero too fast along the tails. For example, Poisson distribution will not work. The intuition is that, when the random variable is Poisson distributed, if one agent received a sufficiently large signal, it will be extremely unlikely that other agents had larger signals. Other distributions that will work in our model include discrete uniform distribution and logarithmic series distribution.

[5]Note that the investors are correct on average. There are good reasons why investors in reality would have dispersed opinions the timing of an exogenous event, such as information cost. We have adopted the simplest possible noise distribution here, which can be extended as we show later.

[6]i.e. an equilibrium where both investors buy the asset from each other for a significant period of time

The following table illustrates the wealth of the investors and the asset price (that one investor needs to pay to buy it from the other) in equilibrium before the bubble bursts:

Date	0	1	2	
Income	N/A	k	k^2	
Wealth of A	$^*W_0$	$W_0 + k + k$	$^*W_0 + k + k^2 + k - k^2$	...
Wealth of B	W_0	$^*W_0 + k - k$	$W_0 + k + k^2 - k + k^2$	...
Asset Price	1	k	k^2	

The asterisk indicates the ownership of the asset. At the date of liquidity shock, the investor will consume all his available wealth and the asset (if he owns it).

The investors are subject to the budget constraint that $W_t > 0$ for all t. We can see from the above that this constraint is always satisfied.

3.3 Proof of Equilibrium

Let's check the incentives of the investors in each period given that the asset is still being traded. Suppose that investor A received a signal s (and that he can buy the asset at period $s, s-2, s-4, \ldots$). Denote his wealth at period t by W_t.

At period s, investor A knows for sure that $T = s + 1$ given a signal s. So he will not buy the asset since he cannot sell it at a higher price in the future.

At period $s-2$, if investor A buys the asset at price k^{s-2}, he will lose if $T = s-1$ with probability $\mathbf{P}(T = s-1|s) = \frac{1}{1+(1-p)^2}$, and he will gain if $T = s+1$ with probability $\mathbf{P}(T = s+1|s) = \frac{(1-p)^2}{1+(1-p)^2}$. Hence his expected final consumption from buying is

$$\mathbf{E}_{s-2}(C|s,\text{buy}) = \frac{(1-p)^2}{1+(1-p)^2}[W_{s-2} + (k-1)k^{s-2} + \sum_{t=s-1}^{s+1} k^t]$$
$$+\frac{1}{1+(1-p)^2}[W_{s-2} + 1 - k^{s-2} + k^{s-1}]$$

If investor A does not buy at $s-2$, his expected final consumption is

$$\mathbf{E}_{s-2}(C|s,\text{don't buy}) = \frac{(1-p)^2}{1+(1-p)^2}[W_{s-2} + \sum_{t=s-1}^{s+1} k^t]$$
$$+\frac{1}{1+(1-p)^2}[W_{s-2} + k^{s-1}]$$

Hence we see that investor A would prefer buying when

$$\begin{aligned}
& \mathbf{E}_{s-2}(C|s,\text{buy}) - \mathbf{E}_{s-2}(C|s,\text{don't buy}) \\
= \; & \frac{(1-p)^2}{1+(1-p)^2}[(k-1)k^{s-2}] + \frac{1}{1+(1-p)^2}[1-k^{s-2}] \\
> \; & 0
\end{aligned}$$

This holds for every s if and only if

$$k > 1 + \frac{1}{(1-p)^2} \tag{3.1}$$

At period $s-4$, investor A knows that investor B will always buy the asset in the next period, so it is optimal for investor A to buy. This is true for all the remaining periods.

Now we go on to check the incentive of investor B given that his counterpart follows the equilibrium strategy. Suppose that investor B received a signal s (and that he can buy the asset at period $s+1, s-1, s-3, \ldots$).

At period $s+1$, the liquidity shock will happen and there will be no trading at this period.

At period $s-1$, since there was no shock taking place at $s-1$, investor B knows for sure $T = s+1$ given the signal s. If he buys the asset at k^{s-1}, he will lose if investor A gets signal s

and quits at s, with probability $\frac{1}{2}$; and he will gain if investor A gets signal $s+2$ and continues to buy at s, with probability $\frac{1}{2}$, too. Thus the expected final consumption from buying is

$$\begin{aligned}\mathbf{E}_{s-1}(C|s,\text{buy}) = \frac{1}{2}[W_{s-1} + (k-1)k^{s-1} + k^{s} + k^{s+1}] \\ +\frac{1}{2}[W_{s-1} + 1 - k^{s-1} + k^{s} + k^{s+1}]\end{aligned}$$

If investor B does not buy at $s-1$, his expected final consumption is

$$\begin{aligned}\mathbf{E}_{s-1}(C|s,\text{don't buy}) = \frac{1}{2}[W_{s-1} + k^{s} + k^{s+1}] \\ +\frac{1}{2}[W_{s-1} + k^{s} + k^{s+1}]\end{aligned}$$

Hence we see that investor B would prefer buying when

$$\begin{aligned} & \mathbf{E}_{s-1}(C|s,\text{buy}) - \mathbf{E}_{s-1}(C|s,\text{don't buy}) \\ = \; & \frac{1}{2}[(k-1)k^{s-1}] + \frac{1}{2}[1-k^{s-1}] \\ > \; & 0\end{aligned}$$

This holds for every s if and only if

$$k > 2 \tag{3.2}$$

At period $s-3$, if investor B buys the asset at the price k^{s-3}, he will lose if investor A gets signal $s-2$ and quits the market then, with probability $\frac{1}{2[1+(1-p)^2]}$; otherwise investor B will gain with probability $1 - \frac{1}{2[1+(1-p)^2]}$. The expected final consumption from buying is

$$
\begin{aligned}
\mathbf{E}_{s-3}(C|s,\text{buy}) = \frac{1}{1+(1-p)^2}\Big(&\frac{1}{2}[W_{s-3}+1-k^{s-3}+k^{s-2}+k^{s-1}] \\
&+\frac{1}{2}[W_{s-3}+(k-1)k^{s-3}+k^{s-2}+k^{s-1}]\Big) \\
&+\frac{(1-p)^2}{1+(1-p)^2}\mathbf{E}_{s-1}(C|s,\text{buy})
\end{aligned}
$$

The expected final consumption if investor B does not buy at $s-3$ is

$$
\begin{aligned}
\mathbf{E}_{s-3}(C|s,\text{don't buy}) = &\frac{1}{1+(1-p)^2}[W_{s-3}+k^{s-2}+k^{s-1}] \\
&+\frac{(1-p)^2}{1+(1-p)^2}[W_{s-3}+\sum_{t=s-2}^{s+1}k^t]
\end{aligned}
$$

which is less than $\mathbf{E}_{s-3}(C|s,\text{buy})$ given that condition (3.2) holds.

At all the remaining periods, investor B will always buy since he knows that investor A will not quit in the next period.

Therefore, the above is a Nash equilibrium if both conditions (3.1) and (3.2) are satisfied. Since $2 < 1+\frac{1}{(1-p)^2}$, the equilibrium holds if and only if

$$
k > 1+\frac{1}{(1-p)^2} \tag{3.3}
$$

Hence we have derived a condition for which the bubble can persist in equilibrium. That is, given a high enough reward, rational agents with heterogeneous beliefs have incentives to ride the bubble, even though they know en-ante that this is a bubble and it will burst in finite time.

3.4 Extension

In the following we will extend the model to show that a bubble equilibrium holds with noisier signals and different strategies of the investors. In particular, there exists a Nash equilibrium in which the investors buy until some period before or after their signals, for different ranges of the price multiplier k.

The extended model will be the same as the baseline model in that the exogenous shock will happen at time T which follows a Geometric distribution with parameter p. The difference is that now each investor will receive ex-ante a signal s which is uniformly distributed over the interval $[\max\{0, T-n\}, T+n]$.

We will show that there exists a set of Nash equilibrium characterized by their equilibrium strategies and conditions on the price multiplier k. In each equilibrium NE_m where m is an integer, an investor with private signal s buys the asset at period t if and only if $t \leq s+m$, for a certain interval of k.

Lemma 3.1. *Suppose that investor B has a signal s' and will buy the asset at t with $t \leq s'+m$. Then if it is optimal for investor A to buy at time t, it is also optimal for him to buy at any time $i \leq t$.*

Proof. See Appendix A.2 □

Lemma 3.1 can be interpreted as that if the value of k is a large enough incentive for the investor to buy at time t, then it is also sufficiently large for the investor to buy at any time before t, given that his counterpart follows the euqilibrium strategy. In other words, the necessary condition on k for the investor to buy at time t is weakly increasing in t.

Lemma 3.2. *Let $g(m,n,p)$ be a function such that*

$$g(m,n,p) = \begin{cases} \dfrac{\sum_{j=0}^{n-m-1}(1-p)^j}{\sum_{j=1}^{n-m-2}\frac{m+n+j+1}{m+n+j+3}(1-p)^j + \frac{2n}{2n+1}(1-p)^{n-m-1}} & \text{if } m \geq -n-1 \\ \dfrac{\sum_{j=0}^{2n}(1-p)^j}{\sum_{j=1}^{2n-1}\frac{j}{j+2}(1-p)^j + \frac{2n}{2n+1}(1-p)^{2n}} & \text{if } m \leq -n-1 \end{cases}$$

where m is an integer, n is a non-negative integer and p is a probability. Then g is an increasing function of m and is greater than 1.

Proof. See Appendix A.3 □

Proposition 3.2. *Given that $g(m,n,p) \leq k \leq g(m+1,n,p)$, there exists a Nash equilibrium NE_m such that each investor with a private signal s will buy the asset at t if and only if $t \leq s+m$ where $m \leq n-2$.*

Proof. See Appendix 3.2 □

We can see that $g(m,n,p)$ is independent of m when $m < -n-1$, and then increasing with m. In other words, when the equilibrium strategy is to exit very early, we have a threshold level of incentive k that must be satisfied to make the investors trade. Otherwise, a larger m leads to a higher lower-bound of k, i.e. in order to encourage the investors to exit the market relatively late and ride the bubble for a longer period of time, a higher incentive is needed. Hence the implication is that bubbles tend to last longer in markets with higher market impacts of trade. On the other hand, a smaller incentive is required with more conservative exit strategies adopted by the investors.

In addition, $g(m,n,p)$ is decreasing with n and increasing with p. The intuition is that the less accurate the signals, the more likely that someone will exit after myself, and hence a small incentive is required to encourage people to trade. Similarly, the less likely that the exogenous shock happens in the next period, the more willingness-to-trade the investors have. An extreme case is when $n=0$ and $g(m,n,p) \to \infty$. This is intuitive since when every investor knows definitely when the shock will happen, backward induction eliminates the bubble ex-ante.

This game is not constrained to be played only between two investors. Since the one who sells at time t and buys at time $t+2$ is not necessarily the same person, it is straightforward to extend the result to more investors, assuming that the investors cannot observe who have exited the market. Otherwise, the investors would update their beliefs upon observation and the model will become more complicated. We leave it to future research.

3.5 Summary

To summarize, we have developed a model of rational bubbles based on several assumptions. Firstly, we have assumed that the asset price needs to be inflated by an exogenous factor every time being traded. Secondly, the bubble will burst following an exogenous shock and the date of the shock follows a Geometric distribution. Thirdly, the investors have heterogeneous beliefs about the date of the shock.

Our model shows that an equilibrium in which a bubble can persistently exists given that the bubble grows fast enough. In the baseline model, the speed at which the bubble grows only depends on k, the exogenous inflation factor, hence the implication is that bubbles are more likely in illiquid markets where trading impact on the market price is relatively high. In the extension we have shown that a higher inflation factor will result in bubbles that last longer. Furthermore, a higher dispersion of signals and a smaller probability of exogenous shock could also extend the life of the bubble.

Our model does not have a realistic background and aims at illustrating the intuition. However, it can be easily extended to many cases of asset price bubbles in the reality. In the following section we will apply the model in the context of the recent US housing bubble and sub-prime mortgage crisis. Note that this is not a straightforward extension. In particular, the housing bubble in our model is not supported directly by speculative buyers but through excessive lending and securitization behavior of the mortgage lenders. One of our objectives in doing so is to show that the intuition can be applied in more sophisticated frameworks.

4 Application in Sub-prime Mortgage Market

The sub-prime mortgage industry has been partly blamed for the emergence and collapse of the recent US housing bubble. Loose lending standards have provided excessive credit to the sub-prime borrowers who have a poor credit history and ability to repay their loans. It has been argued that, as a result of the fact that the sub-prime mortgage lenders securitize their loans and sell the securitized products to the investors, the lenders have shifted the risks associated with sub-prime lending away and do not have enough incentives to maintain the quality of the borrowers, hence we observe the boom and collapse of the sub-prime industry. The dramatic increase in the default rates in the industry after the housing bubble burst might suggest that investors have underestimated the risk inherent in the mortgage-backed assets that they bought. However, this cannot stand if we assume that the investors are rational and anticipate the moral hazard issues ex-ante.

We try to establish a different perspective by extending our model of rational bubbles. We argue that the introduction of securitization would enable the bank to transfer the risk to investors and gamble in the same way as in the baseline model. That is, banks have incentives to lend as long as investors are willing to buy the mortgage-backed securities (MBS), and investors have incentives to buy the MBS as long as banks are willing to refinance borrowers. Because of the uncertain timing of the shock, a bubble in the house price can persist with the same intuition in the baseline model.

The sub-prime mortgage market is a complex mechanism involving the processes of mortgage lending, securitization and derivatives trading with many individual investors and financial intermediates and a variety of practices and contractual terms. In order to illustrate our ideas with as few unnecessary complications as possible, our model is a much simplified version of the real world. In particular, the mortgages have a simple structure and will last for one period only. Moreover, house prices will stay constant at a "bubble" level rather than continuously increasing. These are simplifications to better deliver the intuition rather than necessary consequences or constraints of the model.

4.1 Basic Setup

Our model is set in infinite discrete time periods with $t = 1, 2, 3, \ldots$. There are three classes of agents: borrowers, a bank and an investor. All agents are risk neutral.

As we will see later, the setup of the model is in essence the same as the rational bubbles model we introduced in the previous section. In particular, the bubble process will terminate following an exogenous shock, and the bank and the investor maximize their terminal wealth (or consumption). Therefore, we will need to assume an initial wealth and a deterministic income process for the bank and the investor so that the capital constraints are satisfied[7] For the purpose of simplicity, we omit the details of these assumptions here (See Appendix A.5 for full specification). Similarly, we assume that all agents apart from the borrowers are subject to a liquidity shock and they will behave strategically to maximize consumption at the date of the shock as in the baseline model[8].

4.1.1 Borrowers

In each period, there will be N new sub-prime borrowers who would like to gain access to the housing market. Each new borrower has zero initial wealth and expects to receive an income I in each period. The borrowers are called "sub-prime" because their income is very unstable, that is, their income is subject to a shock with probability λ. After an income shock, the borrower will not receive any income for the current and all remaining periods. Assume that the borrowers realize a large enough utility from living in a house so that they always want to borrow.

4.1.2 Mortgage Lender

There is a bank (or mortgage lender) who has the ability to identify the borrowers and lend them money through issuing mortgages. We assume that the length of the mortgage is one period and the mortgage interest rate is exogenously given by r. That is, after one period, the mortgage borrowers must repay $1 + r$ times the amount of the loan. Assume that the borrowers will be able to repay the loan fully with an income of I.

The mortgage contract also stipulates that the bank has the right to repossess the borrowers' homes should they default. The bank will immediately liquidate the repossessed houses for cash.

The bank can refinance mortgages, i.e. it can lend to the borrowers with existing mortgages.

[7]This is for the purpose of simplicity only. When the bank is constrained by capital, there also exists a bubble equilibrium where house prices will increase.

[8]Again, the shock can be interpreted in many ways as we will discuss in the end of this section.

The contract terms are the same for new and refinancing mortgages. If an existing borrower borrowed P in period t and applies for refinancing in period $t+1$, the amount of the refinancing mortgage will be $(1+r)P$.

In the same way as in the baseline model, we assume that the bank maximizes consumption following the liquidity shock and has an income process ensuring that he is not constrained by wealth (see Appendix A.5 for details).

4.1.3 Housing Market

Now we introduce the housing market. We assume that the house price P_t in each period is the equilibrium price such that total demand is equal to total supply. The fundamental supply of houses is fixed at S, and the fundamental demand[9] of houses is determined by the exogenous function $D(P_t)$, which is assumed to be decreasing and convex in P_t. The total demand will include both the fundamental demand and the demand from sub-prime lending (if any), and the total supply will include the fundamental supply and the liquidated houses of borrowers who have defaulted (if any). Hence, without any lending to the sup-prime borrowers, the house price will be at the fundamental level P^F such that $D(P^F) = S$.

Before introducing the investor, let's first look at the basic case where no securitization is allowed.

4.2 Concept of Bubble

We argue that a bubble in the housing market can exist as a result of excessive lending and securitization. In order to show this, we need to define our concept of bubble first. We define a house price bubble to be the level at which house prices cannot be sustained without securitization.

Proposition 4.1. *Let P_1 and P_2 be such that $D(P_1) + N = S$ and $D(P_2) = S + N$. Assume $(1-\lambda)(1+r) + \lambda\frac{P_2}{P_1} < 1$. Then without securitization, the bank does not have incentives to lend to all sub-prime borrowers and keep the price at P_1.*

Proof. See Appendix A.6 □

We can see from Figure 1 that if the bank lends to all N borrowers at period 0, the house price will firstly rise as a result of excessive lending, and then fall below the fundamental price due to

[9]Because we only have sub-prime borrowers in our model, we can think of the fundamental demand as consisting of the house buyers with no mortgages or prime mortgages.

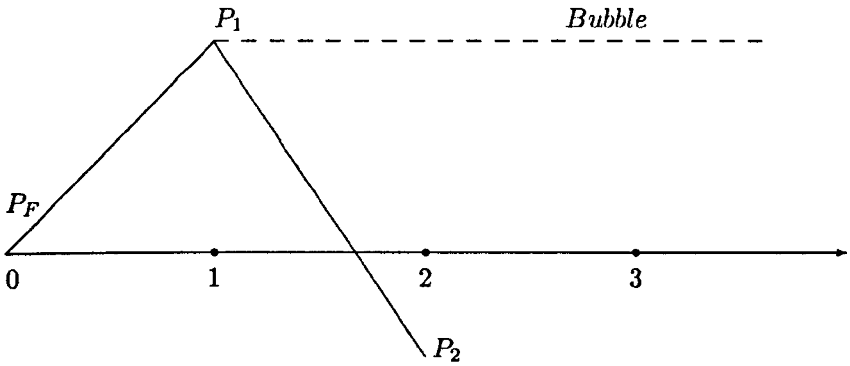


Figure 1: House price movement

foreclosures, which means a negative return to the bank on repossessed houses. The bank would not be willing to refinance the borrowers either, since their incomes would remain zero in the future. Because the fundamental return from lending to all the sub-prime borrowers is negative, the bank does not have incentives to sustain the bubble alone. However, given that the bank can transfer the risk to the investor through securitization, it is possible that a house price bubble is sustained.

4.3 Securitization

We argue that securitization can act as a means for the bank and the investor to take risk and gamble for profit on overvalued assets. As a result, the house price will be kept at artificially high levels before eventually bursting.

There is one investor in the market who maximizes consumption following the liquidity shock and has an income process ensuring that he is not constrained by wealth (see Appendix A.5 for details). In each period, the bank can choose to securitize the mortgages it issued and sell them to the investor. We call those securitized mortgages MBS. Let L_t be the total value of securitized mortgages at period t. We denote the price of the MBS by M_t and assume that $\frac{M_t}{L_t}$ is determined through negotiation between the bank and the investor, i.e. the bank captures a premium/securitization fee on those MBS that is a fixed fraction of the total loan amount. Thus we have $1 < \frac{M_t}{L_t} < 1 + r$ for all t.

The division of profits between the bank and the investor, $\frac{M_t}{L_t}$, is the key parameter in this model and is analogous to the parameter k in the baseline model. The housing bubble can only be sustained given certain restrictions on this parameter ensuring large enough incentives for both parties. While we have exogenously assumed that $\frac{M_t}{L_t}$ is a fixed outcome through negotiation, we acknowledge that there are other possible explanations such as switching cost and competition.

Finally, we assume that the date of liquidity shock T is randomly drawn from a Geometric distribution with parameter p. At the start of the bubble, the bank and the investor each receives a noisy signal s about T, which is binomially distributed on $(T-1, T+1)$ with probability $\frac{1}{2}$.

Proposition 4.2. *Given that* $(2-\delta) < \frac{M_t}{L_t} < \frac{2(1+r)(1-p)^2+\delta}{2(1-p)^2+1}$ *and* $\delta > 1-\frac{r(1-p)^2}{1+(1-p)^2}$ *where* $\delta = \frac{D^{-1}(\infty)}{P^*}$, *there exists a Nash equilibrium such that*[10]

- *Given a signal s, the bank lends and securitizes as much as possible until the liquidity shock*
- *Given a signal s, the investor buys the MBS at period t if and only if $t < s$*

so that the house price will stay at P^* before the bubble bursts, where $D(P^*) + N = S$.

Proof. See Appendix A.7 □

We see that the bubble equilibrium holds when the negotiation outcome $\frac{M_t}{L_t}$ is within a certain interval. The intuition is that the division of profits between the bank and the investor must be "fair" enough to ensure that both parties have incentives to ride the bubble. Moreover, such an interval exists only if δ is small enough relative to r. In particular, if we assume $D^{-1}(\infty) = 0$, we need $r > 1 + \frac{1}{(1-p)^2}$. This may seem unrealistically large, but this is only a result of simplification. As we have seen in the extension of the baseline model, a lower k is required with more conservative strategies etc. Similarly, as we introduce such a natural extension, the magnitude of r will become more sensible, and a higher r will result in longer persistence of bubbles.

4.4 Implications

To summarize, the housing bubble can persist if $(2-\delta) < \frac{M_t}{L_t} < \frac{2(1+r)(1-p)^2+\delta}{2(1-p)^2+1}$ where $\delta > 1-\frac{r(1-p)^2}{1+(1-p)^2}$, i.e. the negotiation outcome is such that both the bank and the investor receive a large enough share of the return on the mortgages. The equilibrium would fail to hold if r is too small, so that the profits are not enough for the bank and the investor to ride the bubble; or if p is too large, so that the risk of losing is too large that riding the bubble is not optimal.

Hence we argue that securitization has enabled the mortgage lenders to transfer the risk to and gamble with the investors on sub-prime mortgages which are otherwise unprofitable. As a result, there is continuous excessive demand in the housing market and a housing bubble can persist.

[10]Here δ is the lower bound of total return of mortgages securitized one period before the bubble bursts.

Before the burst of the bubble, the bank is willing to refinance sub-prime borrowers with difficulties to repay, therefore borrowers with financial problems could avoid foreclosure by refinancing and default rates are kept at artificially low levels. For example, nearly 60% of sub-prime loans originated in 2003 were for refinancing according to Chomsisengphet and Pennington-Cross (2006). As we showed in the model, once the bank stops lending and the housing bubble bursts, the borrowers are no longer able to refinance their mortgages and observed default rates will increase dramatically. Thus we argue that such an increase is not necessarily due to irrational expectation of the investors but the change in the refinancing policy of the mortgage lenders.

As the bubble persists, the amount of total mortgages outstanding will increase. Hence the longer the bubble persists, the more defaults and foreclosures are expected when the bubble bursts, and the negative impact on the housing market will be more drastic. Therefore, as the bank raised lending standards and the number of buyers in the market decreased, we observe a sharp rise in home inventories and fall in house prices in 2006-7.

There were many other factors that have contributed to this process. The tranching practices and lax behavior of rating agencies have resulted in plenty of A-grade MBS and enabled the bank to sell the securitized products to more institutional investors subject to regulations. Speculative home buyers, falling interests and loose regulatory practices have all contributed to the bubble. Hence the exogenous shock can be interpreted in many ways. A sudden change in any of those exogenous factors mentioned above could potentially lead to the collapse of the bubble.

5 Conclusion

We have developed a model of rational bubbles. Assuming an exogenous shock and heterogeneous beliefs, we show that rational agents have incentives to invest in a virtually worthless asset and gamble with each other for profits, even if it is common knowledge ex-ante that it is a bubble and will burst in finite time. As a result, over-valuation of the asset can persist for a significant period before eventually bursting. We have also shown that the higher the price inflator (the rate of increase of the asset price after each trade), the longer the bubble can persist. Thus we have shown that given some exogenous price path, a bubble can exist in a financial market in which the fundamental value of the asset becomes almost irrelevant.

We then apply the model in the sub-prime mortgage market. Despite the poor quality of the sub-prime borrowers, the bank has incentives to lend to and refinance all the borrowers and sustain a housing bubble, given that it is able to securitize the mortgages and sell them to the investors. Hence banks have incentives to lend as long as investors are willing to buy the mortgage-backed securities (MBS), and rational investors have incentives to buy the MBS as long as banks are willing to refinance the borrowers. Because of the uncertain timing of the exogenous shock, a bubble in the house price can persist with the same intuition in the baseline model. Once the bank stops lending and the housing bubble bursts, the borrowers are no longer able to refinance their mortgages and observed default rates will increase dramatically. Thus we argue that such an increase is not necessarily due to irrational expectation of the investors but the change in the refinancing policy of the mortgage lenders.

References

Abreu, D. and Brunnermeier, M. K. (2003), 'Bubbles and crashes', *Econometrica* **71**, 173–204.

Allen, F. and Gorton, G. (1993), 'Churning bubbles', *Review of Economic Studies* **60**, 813–836.

Allen, F., Morris, S. and Postlewaite, A. (1993), 'Finite bubbles with short sale constraints', *Journal of Economic Theory* **61**, 206–229.

Calem, P. S. and LaCour-Little, M. (2004), 'Risk-based capital requirements for mortgage loans', *The Journal of Banking and Finance* **28**, 647–672.

Chomsisengphet, S. and Pennington-Cross, A. (2006), 'The evolution of the subprime mortgage market', *Federal Reserve Bank of St. Louis Review* **88(1)**, 31–56.

Coval, J., Jurek, J. W. and Stafford, E. (2007), Economic catastrophe bonds. HBS Finance Working Paper No. 07-102.

DeMarzo, P. and Duffie, D. (1999), 'A liquidity-based model of security design', *Econometrica* **67:1**, 65–99.

Keys, B. J., Mukherjee, T., Seru, A. and Vig, V. (2008), 'Did securitization lead to lax screening? evidence from subprime loans', *EFA 2008 Athens Meetings Paper* .

Sato, Y. (2008), Fund manager tournaments and the persistence. Working Paper.

A Technical Appendix

A.1 Proof of the necessity of exogenous shock in baseline model

Proof. In order to prove that heterogeneous beliefs is necessary for a bubble equilibrium (i.e. an equilibrium where both investors buy the asset from each other for a significant period of time) to exist, we display below a model that is the same as the baseline model except that investors do not receive signals about the date of the exogenous shock. That is, the probability that the exogenous shock happens in the next period is constant for all periods. We show that a bubble equilibrium does not exist in this case.

Consider a discrete-time model with infinite time periods 0,1,2... Suppose there are two investors: investor A and investor B, each with an initial wealth of W_0. They will both receive income in each period. The amount of income they receive in period t is k^t.

There is an asset that has a fundamental value of 1 and can be traded between the two investors. Assume that the asset can only be bought at a multiple k of its current price. So in period t, one investor can decide whether or not to buy this asset from the other at a price of k^t.

Suppose that the investors do not consume normally, but at each period in time, there is a probability λ that both investors have a liquidity shock and will consume all their wealth (the asset itself has a consumption value of 1). Both investors will maximize the expected value of this consumption, which we denote by C. Assume that investors are risk-neutral and have zero discount rate.

We will show that the Nash Equilibrium where both investors always buy does not hold.

Let's check the incentives of the investors in each period given that the asset is still being traded.

In period t, suppose that investor A can buy the asset from investor B. Let the current wealth of investor A be W_t. Then investor A will buy if

$$\mathrm{E}_t(C|\text{buy}) > \mathrm{E}_t(C|\text{don't buy}) \tag{A.1}$$

The expected consumption if investor A buys is (assuming that both investors will follow the equilibrium strategy in the following periods)

$$\begin{aligned}&\lambda(W_t + k^{t+1} - k^t + 1) + \lambda(1-\lambda)[W_t + k^{t+1} + k^{t+2} + (-1+k)k^t] \\ &+\lambda(1-\lambda)^2[W^t + k^{t+1} + k^{t+2} + k^{t+3} + (-1+k-k^2)k^t + 1] + ...\end{aligned} \tag{A.2}$$

The expected consumption if investor A does not buy is

$$\lambda(W_t + k_{t+1}) + \lambda(1-\lambda)[W_t + k_{t+1} + k_{t+2}] + \lambda(1-\lambda)^2[W_t + k_{t+1} + k_{t+2} + k_{t+3}] + ... \tag{A.3}$$

Hence we can see that $\mathrm{E}_t(C|\text{buy}) > \mathrm{E}_t(C|\text{don't buy})$ if

$$k^t\lambda\sum_{i=1}^{\infty}(1-\lambda)^{i-1}[\sum_{j=1}^{i}(-1)^j k^{j-1}] + \lambda\sum_{i=1}^{\infty}(1-\lambda)^{i-1}\mathbf{I}(i) > 0 \tag{A.4}$$

where $\mathbf{I}(i) = 0$ if i is even, and $\mathbf{I}(i) = 1$ if i is odd. The second term $\lambda\sum_{i=1}^{\infty}(1-\lambda)^{i-1}\mathbf{I}(i) = \lambda[1 + (1-\lambda)^2 + (1-\lambda)^4 + \ldots]$ is equal to $\frac{1}{2-\lambda}$.

Therefore the condition for buying becomes

$$\begin{aligned}
& k^t\lambda\sum_{i=1}^{\infty}(1-\lambda)^{i-1}[\sum_{j=1}^{i}(-1)^j k^{j-1}] + \frac{1}{2-\lambda} \\
= \; & k^t\lambda\sum_{i=1}^{\infty}(1-\lambda)^{i-1}\frac{-1-(-1)^{i+1}k^i}{1+k} + \frac{1}{2-\lambda} \\
= \; & k^t\lambda[-1 + (1-\lambda)(-1+k) + (1-\lambda)^2(-1+k+k^2) + \ldots] + \frac{1}{2-\lambda} \\
= \; & k^t\lambda[-1\sum_{i=0}^{\infty}(1-\lambda)^i + k\sum_{i=1}^{\infty}(1-\lambda)^i + k^2\sum_{i=2}^{\infty}(1-\lambda)^i] + \frac{1}{2-\lambda} \\
= \; & k^t\lambda[-\frac{1}{\lambda} + k\frac{1-\lambda}{\lambda} - k^2\frac{(1-\lambda)^2}{\lambda} + \ldots] + \frac{1}{2-\lambda} \\
= \; & k^t\sum_{i=0}^{\infty}(-1)^{i+1}[(1-\lambda)k]^i + \frac{1}{2-\lambda} \geq 0
\end{aligned}$$

If $(1-\lambda)k \geq 1$, this alternating series does not converge. If $(1-\lambda)k < 1$, the series converges to $\frac{-k^t}{1+(1-\lambda)k}$. Hence the condition can be rewritten as $k^t \leq \frac{1+(1-\lambda)k}{2-\lambda} < \frac{2}{2-\lambda} < 2$. This clearly does not hold for all t given $k > 1$. Hence the proposed equilibrium does not hold. $\square$

A.2 Proof of Lemma 3.1

Proof. First of all, it is easy to show that $t \leq s+n-2$ because the investor will never buy at period $s+n-1$ since he knows for sure that the shock will happen next period. In addition, if $m > 0$, the lemma trivially holds for $t \leq m$ because investor B will trade for at least m periods (say, when $s = 0$). Therefore, in the following we assume $t \in [\max\{0, m+1\}, s+n-2]$.

We will check investor A's incentives before period t given that he will buy the asset at t.

At period t, given that investor B will buy the asset at $t \leq s' + m$, the only possible signals of investor B that would make him stop buying at $t+1$ are $t-m$ or $t-m-1$, since investor B still bought the asset at time $t-1$. So we can find a lower bound for investor A's expected consumption if he buys. We have

$$
\begin{aligned}
&\mathbf{E}_t(C|s, buy)\\
\geq &\mathbf{P}(T = t+1|T \geq \tau_t)\left[W_t + 1 + k^{t+1}\right]\\
&+\mathbf{P}(T = t+2|T \geq \tau_t)\Big[W_t + \mathbf{P}(s' \leq t-m|s' \geq t-m-1, T = t+2)\\
&\qquad + \mathbf{P}(s' > t-m|s' \geq t-m-1, T = t+2)k^{t+1} + k^{t+1} + k^{t+2}\Big]\\
&+\sum_{i=t+3}^{s+n} \mathbf{P}(T = i|T \geq \tau_t)\Big[W_t + \mathbf{P}(s' \leq t-m|s' \geq t-m-1, T = i)\\
&\qquad + \mathbf{P}(s' > t-m|s' \geq t-m-1, T = i)k^{t+1} + k^{t+1} + k^{t+2}\\
&\qquad + \max\left\{\mathbf{E}_{t+2}(C|s, \text{buy}), \mathbf{E}_{t+2}(C|s, \text{don't buy})\right\}\Big] - k^t
\end{aligned}
$$

where $\tau_t \equiv \max\{t+1, s-n, t-m-n-1\}$ and $\mathbf{P}(T = i|T \geq \tau_t)$ $(i = t+1, t+2, \ldots, s+n)$ is the conditional probability of $T = i$ given investor A's information set at time t. This information set contains three parts. First, since the shock has not happened yet, investor A knows that $T \geq t+1$. Second, investor A knows from his private signal s that $T \geq \max\{0, s-n\}$. Third, investor B must have received a signal $s' \geq t-m-1$ since he bought at time $t-1$, hence $T \geq \max\{0, t-m-n-1\}$. Therefore τ_t can be interpreted as the earliest possible shock time given current information set.

In addition, $\mathbf{P}(s'|s' \geq t-m-1, T = i)$ is the conditional probability that investor B received a signal s' given that $T = i$ and $s' \geq t-m-1$. Using Bayesian probability and the property of Geometric distribution, it is easy to show that

$$
\mathbf{P}(T = i|T \geq \tau_t) = \begin{cases} 0 & \text{if} \quad i < \tau_t \\ \dfrac{(1-p)^{i-\tau_t}}{\sum_{i=\tau_t}^{s+n}(1-p)^{i-\tau_t}} & \text{if} \quad i \geq \tau_t \end{cases}
$$

and

$$
\begin{aligned}
&\mathbf{P}(s' > t-m|s' \geq t-m-1, T=i)\\
&=\begin{cases} 0 & \text{if } \ i \leq t-n-m \\ \dfrac{m+n+i-t}{m+n+i-t+2} & \text{if } \ t-n-m < i < t+n-m \\ \dfrac{2n}{2n+1} & \text{if } \ i = t+n-m \\ 1 & \text{if } \ i > t+n-m \end{cases}
\end{aligned}
$$

The expected consumption of investor A if he does not buy at period t is

$$
\mathbf{E}_t(C|s, \text{don't buy}) = \sum_{i=t+1}^{s+n} \mathbf{P}(T=i|T \geq \tau_t)[W_t + \sum_{j=t+1}^{i} k^j]
$$

Therefore, investor A is willing to buy at period t if

$$
\begin{aligned}
&\mathbf{E}_t(C|s, \text{buy}) - \mathbf{E}_t(C|s, \text{don't buy})\\
\geq &\mathbf{P}(T=t+1|T \geq \tau_t)\\
&+ \mathbf{P}(T=t+2|T \geq \tau_t)\Big[\mathbf{P}(s' \leq t-m|s' \geq t-m-1, T=t+2)\\
&\quad + \mathbf{P}(s' > t-m|s' \geq t-m-1, T=t+2)k^{t+1}\Big]\\
&+ \sum_{i=t+3}^{s+n} \mathbf{P}(T=i|T \geq \tau_t)\Big[\mathbf{P}(s' \leq t-m|s' \geq t-m-1, T=i)\\
&\quad + \mathbf{P}(s' > t-m|s' \geq t-m-1, T=i)k^{t+1}\\
&\quad + \max\{\mathbf{E}_{t+2}(C|s, \text{buy}) - \mathbf{E}_{t+2}(C|s, \text{don't buy}), 0\}\Big] - k^t > 0
\end{aligned}
$$

This holds if and only if

$$
k > \frac{\sum_{i=\tau_t}^{s+n}(1-p)^{i-\tau_t}}{X} \tag{A.5}
$$

where

$$
X \equiv \sum_{i=\max\{t+2,\tau_t\}}^{s+n} (1-p)^{i-\tau_t}\mathbf{P}(s' > t-m|s' \geq t-m-1, T=i)
$$

Similarly, investor A will choose to buy at period $t-2$ if

$$
\begin{aligned}
&\mathbf{E}_{t-2}(C|s,\text{buy}) - \mathbf{E}_{t-2}(C|s,\text{don't buy}) \\
\geq &\mathbf{P}(T=t-1|T\geq \tau_{t-2}) \\
&+\mathbf{P}(T=t|T\geq \tau_{t-2})\Big[\mathbf{P}(s'\leq t-m-2|s'\geq t-m-3, T=t) \\
&\quad +\mathbf{P}(s'>t-m-2|s'\geq t-m-3, T=t)k^{t-1}\Big] \\
&+\sum_{i=t+1}^{s+n}\mathbf{P}(T=i|T\geq \tau_{t-2})\Big[\mathbf{P}(s'\leq t-m-2|s'\geq t-m-3, \\
&\quad T=i)+\mathbf{P}(s'>t-m-2|s'\geq t-m-3, T=i)k^{t-1} \\
&\quad +\mathbf{E}_t(C|s,\text{buy})-\mathbf{E}_t(C|s,\text{don't buy})\Big]-k^{t-2}>0
\end{aligned}
$$

This holds if and only if

$$
k > \frac{\sum_{i=\tau_{t-2}}^{s+n}(1-p)^{i-\tau_{t-2}}}{Y} \tag{A.6}
$$

where

$$
Y \equiv \sum_{i=\max\{t,\tau_{t-2}\}}^{s+n} (1-p)^{i-\tau_{t-2}}\mathbf{P}(s'>t-m-2|s'\geq t-m-3, T=i)
$$

Now we need to prove (A.5) is a sufficient condition that (A.6) holds. Since $\tau_t = \max\{s-n, t+1, t-m-1\}$, we need to consider the following three cases:

(I) If $\tau_t = s-n$, i.e. the preconditions satisfy:

$$
\begin{cases} s-n\geq t+1 \\ s-n\geq t-m-n-1 \end{cases} \Longrightarrow \begin{cases} t\leq s-n-1 \\ t\leq s+m+1 \end{cases}
$$

In this case, condition (A.5) becomes

$$
\frac{\sum_{i=s-n}^{s+n}(1-p)^{i-(s-n)}}{X}
$$

where

$$X = \sum_{i=\max\{t+2,s-n\}}^{s+n} (1-p)^{i-(s-n)} \mathbf{P}(s' > t-m | s' \geq t-m-1, T=i)$$

It is straightforward to show that $\tau_{t-2} = s-n$, thus condition (A.6) becomes

$$\frac{\sum_{i=s-n}^{s+n} (1-p)^{i-(s-n)}}{Y}$$

where

$$Y = \sum_{i=s-n}^{s+n} (1-p)^{i-(s-n)} \mathbf{P}(s' > t-m-2 | s' \geq t-m-3, T=i)$$

Because

$$\mathbf{P}(s' > t-m | s' \geq t-m-1, T=i)$$
$$= \begin{cases} 0 & \text{if } \ i \leq t-n-m \\ \dfrac{m+n+i-t}{m+n+i-t+2} & \text{if } \ t-n-m < i < t+n-m \\ \dfrac{2n}{2n+1} & \text{if } \ i = t+n-m \\ 1 & \text{if } \ i > t+n-m \end{cases}$$

and

$$\mathbf{P}(s' > t-m-2 | s' \geq t-m-3, T=i)$$
$$= \begin{cases} 0 & \text{if } \ i \leq t-n-m-2 \\ \dfrac{m+n+i-t+2}{m+n+i-t+4} & \text{if } \ t-n-m-2 < i < t+n-m-2 \\ \dfrac{2n}{2n+1} & \text{if } \ i = t+n-m-2 \\ 1 & \text{if } \ i > t+n-m-2 \end{cases}$$

Comparing the above two expressions, we can show that

$$\mathbf{P}(s' > t-m-2 | s' \geq t-m-3, T=i) \geq \mathbf{P}(s' > t-m | s' \geq t-m-1, T=i)$$

i.e. $Y \geq X$. Therefore

$$\frac{\sum_{i=s-n}^{s+n} (1-p)^{i-(s-n)}}{Y} \leq \frac{\sum_{i=s-n}^{s+n} (1-p)^{i-(s-n)}}{X}$$

Hence condition (A.6) is weaker than condition (A.5) when $\tau_t = s - n$

(II) If $\tau_t = t+1$, then τ_{t-2} will be either $t-1$ or $s-n$.

In this case, the lower bound of k in condition (A.5) is:

$$\frac{\sum_{i=t+1}^{s+n}(1-p)^{i-(t+1)}}{X}$$

where

$$X = \sum_{i=t+2}^{s+n}(1-p)^{i-(t+1)}\mathrm{P}(s' > t-m | s' \geq t-m-1, T=i)$$

$$= \begin{cases} \frac{2n}{2n+1}(1-p) + \sum_{j=3}^{s+n-t}(1-p)^{j-1} & \text{if } m = n-2 \\ \sum_{j=2}^{n-m-1}\frac{m+n+j}{m+n+j+2}(1-p)^{j-1} + \frac{2n}{2n+1}(1-p)^{n-m-1} + \sum_{j=n-m+1}^{s+n-t}(1-p)^{j-1} & \\ & \text{if } m \in (t-s, n-2) \\ \sum_{j=2}^{s+n-t-1}\frac{m+n+j}{m+n+j+2}(1-p)^{j-1} + \frac{2n}{2n+1}(1-p)^{s+n-t-1} & \text{if } m = t-s \\ \sum_{j=2}^{s+n-t}\frac{m+n+j}{m+n+j+2}(1-p)^{j-1} & \text{if } m < t-s \end{cases}$$

(II. 1) If $\tau_{t-2} = t-1$, the preconditions satisfy:

$$\begin{cases} t-1 \geq s-n \\ t-1 \geq t-m-n-3 \end{cases} \Longrightarrow \begin{cases} t \geq s-n+1 \\ m+n+2 \geq 0 \end{cases}$$

Now the lower bound in condition (A.6) is:

$$\frac{\sum_{i=t-1}^{s+n}(1-p)^{i-(t-1)}}{Y}$$

where

$$Y = \sum_{i=t}^{s+n}(1-p)^{i-(t-1)}\mathrm{P}(s' > t-m-2 | s' \geq t-m-3, T=i)$$

$$= \begin{cases} \dfrac{2n}{2n+1}(1-p) + \displaystyle\sum_{j=1}^{s+n-t}(1-p)^{j+1} & \text{if } m = n-2 \\ \displaystyle\sum_{j=0}^{n-m-3}\frac{m+n+j+2}{m+n+j+4}(1-p)^{j+1} + \frac{2n}{2n+1}(1-p)^{n-m-1} + \sum_{j=n-m-1}^{s+n-t}(1-p)^{j+1} & \\ & \text{if } m \in (t-s-2, n-2) \\ \displaystyle\sum_{j=0}^{s+n-t-1}\frac{m+n+j+2}{m+n+j+4}(1-p)^{j+1} + \frac{2n}{2n+1}(1-p)^{s+n-t+1} & \text{if } m = t-s-2 \\ \displaystyle\sum_{j=0}^{s+n-t}\frac{m+n+j+2}{m+n+j+4}(1-p)^{j+1} & \text{if } m < t-s-2 \end{cases}$$

Note that polynomial Y has degree $s+n-t+1$, while X has degree $s+n-t-1$, so we can write Y as $Y = X + f(t)$, where

$$f(t) \equiv \begin{cases} (1-p)^{s+n-t} + (1-p)^{s+n-t+1} & \text{when} \quad t \leq s+m \\ \dfrac{2n}{2n+1}(1-p)^{n-m-1} + (1-p)^{n-m} & \text{when} \quad t = s+m+1 \\ \dfrac{2n-1}{2n+1}(1-p)^{n-m-2} + \dfrac{2n}{2n+1}(1-p)^{n-m-1} & \text{when} \quad t = s+m+2 \\ \displaystyle\sum_{j=s+n-t-1}^{s+n-t}\frac{m+n+j+2}{m+n+j+4}(1-p)^{j+1} & \text{when} \quad t > s+m+2 \end{cases}$$

Therefore

$$\begin{aligned} \frac{\sum_{i=t-1}^{s+n}(1-p)^{i-(t-1)}}{Y} &= \frac{\sum_{i=t+1}^{s+n}(1-p)^{i-(t+1)} + (1-p)^{s+n-t} + (1-p)^{s+n-t+1}}{X+f(t)} \\ &< \frac{\sum_{i=t+1}^{s+n}(1-p)^{i-(t+1)}}{X} \end{aligned}$$

(II. 2) Suppose $\tau_t = t+1$ and $\tau_{t-2} = s-n$. The preconditions are

$$\begin{cases} t+1 \geq s-n \\ t+1 \geq t-m-n-1 \\ s-n \geq t-1 \end{cases} \implies \begin{cases} s-n-1 \leq t \leq s-n+1 \\ m+n+2 \geq 0 \end{cases}$$

Because the previous analysis for $\tau_t = \tau_{t-2} = s-n$ has already included the case of $t = s-n-1$, and the analysis for $\tau_t = t+1$ and $\tau_{t-2} = t-1$ has covered the case of $t = s-n+1$, we only need to look at the case where $t = s-n$.

Then the lower bound of condition (A.6) becomes

$$\frac{\sum_{i=s-n}^{s+n}(1-p)^{i-(s-n)}}{Y}$$

where

$$Y = \sum_{i=s-n}^{s+n}(1-p)^{i-(s-n)}\mathrm{P}(s' > t-m-2|s' \geq t-m-3, T=i)$$

$$= \begin{cases} \frac{2n}{2n+1} + \sum_{j=1}^{2n}(1-p)^j & \text{if } m = n-2 \\ \sum_{j=0}^{n-m-3}\frac{m+n+j+2}{m+n+j+4}(1-p)^j + \frac{2n}{2n+1}(1-p)^{n-m-2} + \sum_{j=n-m-1}^{2n}(1-p)^j & \\ & \text{if } m \in (-n-2, n-2) \\ \sum_{j=0}^{2n-1}\frac{m+n+j+2}{m+n+j+4}(1-p)^j + \frac{2n}{2n+1}(1-p)^{2n} & \text{if } m = -n-2 \end{cases}$$

Y can be written as $Y = \frac{X}{1-p} + g(t)$ where $t = s-n$ and

$$g(t) \equiv \begin{cases} (1-p)^{2n-1} + (1-p)^{2n} & \text{if} \quad m \in [-n, n-2] \\ \frac{2n}{2n+1}(1-p)^{2n-1} + (1-p)^{2n} & \text{if} \quad m = -n-1 \\ \frac{2n-1}{2n+1}(1-p)^{2n-1} + \frac{2n}{2n+1}(1-p)^{2n} & \text{if} \quad m = -n-2 \end{cases}$$

Therefore

$$\begin{aligned} \frac{\sum_{i=s-n}^{s+n}(1-p)^{i-(s-n)}}{Y} &= \frac{\sum_{i=s-n+1}^{s+n}(1-p)^{i-(s-n+1)} + (1-p)^{2n}}{\frac{X}{1-p} + g(t)} \\ &< \frac{\sum_{i=s-n+1}^{s+n}(1-p)^{i-(s-n+1)}}{X} \end{aligned}$$

Hence condition (A.5) is also stronger than condition (A.6) when $\tau_t = t+1$ and $\tau_{t-2} = s-n$.

(III) If $\tau_t = t - m - n - 1$, the preconditions are

$$\begin{cases} t - m - n - 1 \geq s - n \\ t - m - n - 1 \geq t + 1 \end{cases} \Longrightarrow \begin{cases} t \geq s + m + 1 \\ m + n + 2 \leq 0 \end{cases}$$

The case of $m + n + 2 = 0$ has been discussed before, so in this section we only focus on the case of $m + n + 2 < 0$. The lower bound in condition (A.5) is:

$$\frac{\sum_{i=t-m-n-1}^{s+n} (1-p)^{i-(t-m-n-1)}}{X}$$

where

$$\begin{aligned} X &= \sum_{i=t-m-n-1}^{s+n} (1-p)^{i-(t-m-n-1)} \mathrm{P}(s' > t - m | s' \geq t - m - 1, T = i) \\ &= \sum_{j=-m-n+1}^{s+n-t} \frac{m+n+j}{m+n+j+2} (1-p)^{j+m+n+1} \end{aligned}$$

Similar as (II), τ_{t-2} is either $t - m - n - 3$ or $s - n$:

(III. 1) If $\tau_{t-2} = t - m - n - 3$, the preconditions satisfy

$$\begin{cases} t - m - n - 3 \geq s - n \\ m + n + 2 < 0 \end{cases} \Longrightarrow \begin{cases} t \geq s + m + 3 \\ m + n + 2 < 0 \end{cases}$$

Now the lower bound in condition (A.6) is

$$\frac{\sum_{i=t-m-n-3}^{s+n} (1-p)^{i-(t-m-n-3)}}{Y}$$

where

$$\begin{aligned} Y &= \sum_{i=t-m-n-3}^{s+n} (1-p)^{i-(t-m-n-3)} \mathrm{P}(s' > t - m - 2 | s' \geq t - m - 3, T = i) \\ &= \sum_{i=t-m-n-1}^{s+n} (1-p)^{i-(t-m-n-3)} \mathrm{P}(s' > t - m - 2 | s' \geq t - m - 3, T = i) \\ &= \sum_{j=-m-n-1}^{s+n-t} \frac{m+n+j+2}{m+n+j+4} (1-p)^{j+m+n+3} \end{aligned}$$

Hence $Y = X + \sum_{j=s+n-t-1}^{s+n-t} \frac{m+n+j+2}{m+n+j+4}(1-p)^{j+m+n+3}$.

Therefore

$$
\begin{aligned}
&\frac{\sum_{i=t-m-n-3}^{s+n}(1-p)^{i-(t-m-n-3)}}{Y} \\
=&\frac{\sum_{i=t-m-n-1}^{s+n}(1-p)^{i-(t-m-n-1)} + (1-p)^{m+2n+s-t+2} + (1-p)^{m+2n+s-t+3}}{X + \sum_{j=s+n-t-1}^{s+n-t} \frac{m+n+j+2}{m+n+j+4}(1-p)^{j+m+n+3}} \\
<&\frac{\sum_{i=t-m-n-1}^{s+n}(1-p)^{i-(t-m-n-1)}}{X}
\end{aligned}
$$

(III. 2) If $\tau_{t-2} = s - n$, we have

$$
\begin{cases} t-m-n-1 \geq s-n \\ t-m-n-3 \leq s-n \\ m+n+2<0 \end{cases} \Longrightarrow \begin{cases} s+m+1 \leq t \leq s+m+3 \\ m+n+2<0 \end{cases}
$$

The case $t = s+m+1$ has been covered in section (I), and the case $t = s+m+3$ has been included in (III. 1), so in this section we only work with the case $t = s+m+2$.

When $t = s+m+2$, the lower bound in (A.5) becomes

$$
\frac{\sum_{i=s-n+1}^{s+n}(1-p)^{i-(s-n+1)}}{X}
$$

where

$$
\begin{aligned}
X &= \sum_{i=s-n+1}^{s+n}(1-p)^{i-(s-n+1)}\mathrm{P}(s' > t-m | s' \geq t-m-1, T=i) \\
&= \sum_{j=1}^{2n-2} \frac{j}{j+2}(1-p)^{j+1}
\end{aligned}
$$

Now the lower bound for condition (A.6) is:

$$\frac{\sum_{i=s-n}^{s+n}(1-p)^{i-(s-n-1)}}{Y}$$

where

$$\begin{aligned} Y &= \sum_{i=s-n}^{s+n} (1-p)^{i-(s-n-1)} \mathrm{P}(s' > t-m-2 | s' \geq t-m-3, T=i) \\ &= \sum_{j=1}^{2n-1} \frac{j}{j+2}(1-p)^{j+1} + \frac{2n}{2n+1}(1-p)^{2n+1} \end{aligned}$$

Hence

$$Y = X + \frac{2n-1}{2n+1}(1-p)^{2n} + \frac{2n}{2n+1}(1-p)^{2n+1}$$

Thus

$$\begin{aligned} \frac{\sum_{i=s-n}^{s+n}(1-p)^{i-(s-n-1)}}{Y} &= \frac{\sum_{i=s-n+1}^{s+n}(1-p)^{i-(s-n+1)} + (1-p)^{2n} + (1-p)^{2n+1}}{X + \frac{2n-1}{2n+1}(1-p)^{2n} + \frac{2n}{2n+1}(1-p)^{2n+1}} \\ &< \frac{\sum_{i=s-n+1}^{s+n}(1-p)^{i-(s-n+1)}}{X} \end{aligned}$$

Therefore when $\tau_t = t-m-n-1$ and $\tau_{t-2} = s-n$, condition (A.5) is also stronger than condition (A.6).

To conclude, we have proved that condition (A.5) is stronger than condition (A.6). That is, when investor A is willing to buy at period t, it is also optimal for him to buy at period $t-2$. □

A.3 Proof of Lemma 3.2

Proof. First of all, it is easy to see that $g(m,n,p) > 1 + \frac{1}{n}$ from the expression of g.

When $m \leq -n-1$, it is obvious that $g(m-1) \leq g(m)$.

Now suppose $m \geq -n-1$. Let a and b be such that $g(m) = \frac{a}{b}$. Then it can be shown that $g(m-1) = \frac{a+c}{b+d}$ where $c = (1-p)^{n-m}$ and $d = \frac{2n}{2n+2}(1-p)^{n-m-1} + \frac{2n}{2n+1}(1-p)^{n-m} - \frac{2n}{2n+1}(1-p)^{n-m-1}$. We know that, given $\frac{a}{b} > 1 + \frac{1}{n}$, $\frac{a+c}{b+d} < \frac{a}{b}$ if $\frac{c}{d} \leq 1 + \frac{1}{n}$. But

$$\begin{aligned}\frac{c}{d} &= \frac{1}{\frac{2n}{2n+2}(1-p) + \frac{2n}{2n+1} - \frac{2n}{2n+1}(1-p)} \\ &= \frac{1}{p\frac{2n}{2n+1} + (1-p)\frac{n}{n+1}} \\ &\leq 1 + \frac{1}{n}\end{aligned}$$

Therefore it follows that $g(m-1) < g(m)$, i.e. g is an increasing function of m. $\square$

A.4 Proof of Proposition 3.2

Proof. WLOG we only check the incentives of investor A. Recall from the proof of Lemma 3.1 that the condition for investor A to buy the asset at time t is:

$$k > \frac{\sum_{i=\tau_t}^{s+n}(1-p)^{i-\tau_t}}{\sum_{i=\max\{t+2,\tau_t\}}^{s+n}(1-p)^{i-\tau_t}\mathrm{P}(s'>t-m|s'\geq t-m-1,T=i)} \tag{A.7}$$

where

$$\mathrm{P}(s'>t-m|s'\geq t-m-1,T=i)$$
$$=\begin{cases} 0 & \text{if } \ i\leq t-n-m \\ \dfrac{m+n+i-t}{m+n+i-t+2} & \text{if } \ i\in(t-n-m,t+n-m) \\ \dfrac{2n}{2n+1} & \text{if } \ i=t+n-m \\ 1 & \text{if } \ i>t+n-m \end{cases}$$

Replacing t by $s+m$, we can find that the condition for which investor A will buy the asset at $s+m$ is

$$k\geq\begin{cases} \dfrac{\sum_{j=0}^{n-m-1}(1-p)^j}{\sum_{j=1}^{n-m-2}\frac{m+n+j+1}{m+n+j+3}(1-p)^j+\frac{2n}{2n+1}(1-p)^{n-m-1}} & \text{if } m\geq -n-1 \\ \dfrac{\sum_{j=0}^{2n}(1-p)^j}{\sum_{j=1}^{2n-1}\frac{j}{j+2}(1-p)^j+\frac{2n}{2n+1}(1-p)^{2n}} & \text{if } m\leq -n-1 \end{cases}$$

Hence investor A will buy the asset at $t=s+m$ given that $k\geq g(m,n,p)$ where $g(m,n,p)$ is as we defined in Lemma 3.2. Moreover, Lemma 3.1 implies that investor A will be willing to buy the asset at all periods before $s+m$.

Finally, we need to show that investor A does not want to buy the asset at $s+m+1$. Since g is an increasing function of m, we can find an interval $g(m)\leq k\leq g(m+1)$ for k, which implies that investor A is not willing to buy at $s+m+1$ given that investor B's strategy is to buy until $t>s+m+1$. Hence it is sufficient to show that investor A is not willing to buy at $s+m+1$ in NE_m because it is less likely that investor B will buy in the future. Therefore investor A does not want to deviate from the equilibrium strategy.

□

A.5 Specification of wealth and income in the model of sub-prime mortgages

In order to satisfy the wealth constraints, we need the following assumptions:

- The initial wealth of the bank is greater than NP^* where $D(P^*) + N = S$.
- The initial wealth of the investor is greater than $(1 + r)NP^*$.
- The bank has an income of $(1 + r)^{t-1}\lambda NP^*$ in period t.
- The investor has an income of $(1 + r)^t\lambda NP^*$ in period t.

Proof. In the bubble equilibrium, the total amount of capital that the bank needs to lend (outstanding loan) at period t will be $\sum_{i=1}^{t-1}(1 + r)^i\lambda NP^* + NP^*$ where $D(P^*)+N = S$. The amount of capital that the investor needs to buy the MBS will be less than the outstanding loan multiplied by $1 + r$. Hence it is easy to show that the above specification of initial wealth and income garantees that the bank and the investor are not constrained by capital in equilibrium. □

A.6 Proof of Proposition 4.1

Proof. The total payoff to the bank after lending to N borrowers at $t = 0$ and not lending at $t = 1$ will be $\Phi = (1 - \lambda)(1 + r) + \lambda \frac{P_2}{P_1}$ where $P_2 < P_1$. If the bank lends to and refinances everyone at $t = 1$ and keep the price at P_1, the delinquent borrowers will never be able to repay the loan, and the liquidation price of repossessed houses will be lower since the expected number of foreclosures will be higher, hence the payoff would be less than Φ. Therefore the assumption that $\Phi < 1$ ensures that the bank does not have incentives to sustain the housing bubble. $\square$

A.7 Proof of Proposition 4.2

Proof. First of all, in equilibrium, the total amount of mortgages outstanding in period t will be $\sum_{i=1}^{t-1}(1+r)^i\lambda NP^* + NP^*$. Suppose that the liquidity shock will happen in period $t+1$, the total return on mortgages securitized at period t will be $\frac{1-\lambda}{\sum_{i=0}^{t-1}(1+r)^i\lambda+(1-\lambda)}(1+r) + \frac{\sum_{i=0}^{t-1}(1+r)^i\lambda}{\sum_{i=0}^{t-1}(1+r)^i\lambda+(1-\lambda)}\frac{P^S}{P^*}$ where $D(P^S) = S + t\lambda N$. It is clear that this return is bounded below by some $\frac{D^{-1}(\infty)}{P^*}$ for all t. As the equilibrium needs to hold in the infinite horizon, we will use δ in the rest of the proof.

Suppose that the investor received a signal s. We check whether the investor will deviate from equilibrium:

At period s, the investor is certain that there will be a liquidity shock at $s+1$, so he will not buy at s.

At period $s-1$, given that the shock has not happened, the investor knows for sure $T = s+1$. Since the bank always lends until the shock, the investor will buy at $s-1$, and the expected payoff is $(1+r)L_{s-1} - M_{s-1}$.

At period $s-2$, suppose that the investor buys MBS, he will either get δL_{s-2}, if the shock strikes at $s-1$, with conditional probability $\mathbf{P}(T = s-1|s)$ given signal s; or $(1+r)L_{s-2} + [(1+r)L_{s-1} - M_{s-1}]$, if the shock happens at $s+1$, with conditional probability $\mathbf{P}(T = s+1|s)$. In the latter situation the payoff is combined with the expected income both from buying at $s-2$ as well as that from $s-1$ (the item in square brackets), since the investor will go on buying as long as the shock takes place at $s+1$. The expected payoff for the investor is:

$$\begin{aligned}\Phi_{investor,s-2} &= \mathbf{P}(T = s-1|s)\delta L_{s-2} + \mathbf{P}(T = s+1|s) \\ &\cdot\{(1+r)L_{s-2} + [(1+r)L_{s-1} - M_{s-1}]\} - M_{s-2}\end{aligned} \tag{A.8}$$

where $\mathbf{P}(T = s-1|s) = \frac{\mathbf{P}(s|T=s-1)\mathbf{P}(T=s-1)}{\mathbf{P}(s|T=s-1)\mathbf{P}(T=s-1)+\mathbf{P}(s|T=s+1)\mathbf{P}(T=s+1)}$

$$= \frac{\mathbf{P}(T=s-1)}{\mathbf{P}(T=s-1)+\mathbf{P}(T=s+1)} = \frac{1}{1+(1-p)^2}$$

We have $\Phi_{investor,s-2} > 0\ \forall s$ if and only if

$$\frac{M_t}{L_t} < \frac{2(1+r)(1-p)^2 + \delta}{2(1-p)^2 + 1} \tag{A.9}$$

At $s-3$, there is no risk of losing and the investor will buy. This is true for all remaining periods.

The investor has no incentive to deviate from his optimal strategy if (A.9) holds and $L < M < (1+r)L$.

Now suppose that the bank received a signal s.

At period s, the supply shock will surely happen at $s+1$. If the bank continues to lend and securitize, he will get M_s if the investor got signal $s+2$, which happens with probability $\frac{1}{2}$; the bank will get δL_s if the investor got signal s, with probability $\frac{1}{2}$, too. Thus the expected payoff of the bank at s is:

$$\Phi_{bank,s} = \frac{1}{2}(M_s + \delta L_s) - L_s \tag{A.10}$$

We have $\Phi_{bank,s} > 0$ $\forall s$ if and only if

$$\frac{M_t}{L_t} > (2-\delta) \tag{A.11}$$

At period $s-1$, since the shock has not happened yet, the bank knows for sure that it will happen at $s+1$, which implies that the investor will certainly buy at period $s-1$. Hence the bank will not deviate.

At period $s-2$, the shock will be at either $s-1$ or $s+1$, thus the bank's payoff depends on the signal of the investor. If the investor's signal is $s-2$, he will not buy the MBS and the bank will get δL_{s-2} with conditional probability $\mathbf{P}(s-2|s)$; if the investor's signal is s or $s+2$, he will continue to buy and the bank will get $M_{s-2} + \mathbf{E}_{s-2}[\Phi_{bank,s-1\&s}]$ with probability $1 - \mathbf{P}(s-2|s)$.

The expected payoff of the bank to lend at $s-2$ is:

$$\begin{aligned}\Phi_{bank,s-2} = \mathbf{P}(s-2|s)(\delta L_{s-2}) + [1 - \mathbf{P}(s-2|s)]\\ \cdot(M_{s-2} + \mathbf{E}_{s-2}[\Phi_{bank,s-1\&s}]) - L_{s-2}\end{aligned} \tag{A.12}$$

where $\mathbf{P}(s-2|s)$ is the conditional probability for the investor got signal $s-2$ given that the the bank got signal s:

$$\mathbf{P}(s-2|s) = \mathbf{P}(s-2|T=s-1)\mathbf{P}(T=s-1|s) = \frac{1}{2[1+(1-p)^2]} < \frac{1}{2}$$

Therefore, $\Phi_{bank,s-2} > \Phi_{bank,s}$, so that (A.11) can guarantee that (A.12) is positive.

At period $s-3$, the bank knows the investor will definitely buy and it is profitable to lend and securitize. The same is true for the remaining periods. Therefore the bank has no incentive to deviate if (A.11) holds.

Combining (A.9) and (A.11), we see that the equilibrium holds as long as

$$(2-\delta) < \frac{M_t}{L_t} < \frac{2(1+r)(1-p)^2+\delta}{2(1-p)^2+1} \tag{A.13}$$

Finally we check consistency, i.e. we need $1 < \frac{M_t}{L_t} < 1+r$. It can be shown that (A.13) is consistent if and only if $\delta > 1 - \frac{r(1-p)^2}{1+(1-p)^2}$. □

Short-dated Volatility Forecasting

Yiyi Wang

London School of Economics

November 27, 2011

Abstract

This paper studies a wide range of empirical volatility models and the implied volatility to forecast short-dated volatility in FX market using high frequency data. Instead of promoting a single method, this paper argues that each model has its own expertise in the specific environment, and overall forecast can be more robust to keep a battery of models and take a collective signal from them. To mimic the real trading environment and exclude look-ahead bias, estimation and forecasting is conducted on a rolling basis. Moreover, a realistic 2-week volatility swap trading strategy is tested according to the forecast. The profit is pronounced even after considering transaction costs. Especially, the model combination constantly outperforms any single model, and is able to seize the upside from each of them. On the technical side, this paper proposes a double-step approach to circumvent the disadvantage of employing GARCH-type model in high frequency data in FX market, so that it can separate effect of intraday / intraweek seasonality and pre-scheduled macroeconomic data releases from underlying data process. The double-step approach is the first time being applied in the context of volatility forecasting and proved to be significantly effective in model improvement. Furthermore, regime-switch technique is applied to enhance the model adaptability.

Keywords: shorted-dated volatility, volatility forecasting, volatility swap.

1 Introduction

Either as an indicator of the contemporaneous spot price movement, or as an investable asset alone, volatility has been increasingly researched especially after the recent financial crisis. Since 2007 the market has witnessed the ebb and flow of volatility with each strong wave mirroring a turbulence of market sentiment. This experience, nevertheless, provides us with a great opportunity to form a better understanding of volatility evolution.

There have been controversial results in the literature as to the best volatility forecast mechanism (Jorion 1995; Anderson 1998; Jiang and Tian 2005; Andersen and Bondarenko 2007). The traditional empirical models have well documented the stylized features of volatility, including leptokurtosis, clustering and persistence; however, none of these econometric forecast methods escape from the shortcoming of their backward-looking nature. Empirical models use the past to predict the future, which is a setback when forecasting into near-term events that have been known to be more volatile than normal, and which are not part of the estimation window. The obvious alternative – market-implied volatilities – are typically backed out from the Black & Scholes model and thus suffer from model-dependency (and reliability) issues.

The essential problem lies in the fact that the data generating process for volatility is unobservable, or in other words, the "true" realized volatility is a latent variable that can only be "represented" by observable proxy. In this sense, any argument on the best unbiased forecast can be in vain if the proxy choice for realized volatility is arbitrary. Instead, a more relevant and interesting question becomes whether it is possible to construct a strategy to exploit the divergence of forecasts and the proxy given all the observable market conditions. For example, in a volatility swap contract the realized volatility is defined as the square root of annualized daily squared returns. Therefore, the forecasts built around a volatility swap will use this specific proxy as comparison benchmark.

This paper takes the agnostic view that no single model (including implied volatility) can consistently outperform others. Rather than sticking to one method from the beginning to the end, it keeps a battery of models and takes a collective signal from them. To guarantee the adaptability of each model within the selection pool, it further applies a Markov regime switching technique. Interestingly I find that different characteristics embedded in different models tend to highlight themselves in distinct volatility regimes, so rotating across models along with the changes in market turns out to be the best approach.

This research uses intraday high frequency data from the foreign exchange market – the most liquid of all markets - which brings both advantages and challenges. A major improvement is the growth in information, thus reducing output noise relative to the equivalent with daily data. However, there are difficulties associated with FX intraday data. Firstly, currency markets exhibit seasonal volatility patterns linked to relative intra-day and intra-week liquidity. Although the most liquid currencies are traded around the clock, trading volumes can be concentrated around time periods where major time zones (US, Europe, Asia) overlap. Intra-week volume patterns are also notable, leading to a rise in volatility over the course of the week and culminating with an abrupt fall at the end of the week. A blind adoption of an empirical model like GARCH on such a dataset distorts the recursive relationship that the model assumes.

Secondly, since an exchange rate is the relative valuation between two currencies, its price and volatility can be highly vulnerable to events that are idiosyncratic to each currency. Therefore, FX volatility is also subject to regular macroeconomic data releases from each side. Key examples include central bank rates decisions, inflation, manufacturing and unemployment data. The price movement tends to be more volatile around the short period when there is a data release. In certain circumstances the volatility in a 1-hour interval can be 10x that of non-event hourly intervals. This paper proposes a two-step approach to "cleanse" the original data to circumvent the above-mentioned two problems, so that normal empirical models can be operated on high frequency data. The double-step approach is the first time being applied in the context of volatility forecasting and proved to be significantly effective.

Daily data was proper when the research was centred in long-dated (monthly or longer) volatility forecast. With the engagement of high frequency data, it allows us to look into a more flexible forecast length, in particular, short-dated forecast window altering from overnight to several weeks. This provides indication for gamma trading since a delta neutral strategy involves dynamic hedging and profit taking which is essentially long gamma[1]. In the short end, such a strategy is determined

[1]For example, a delta-hedged portfolio can be constructed by a long position in a call option and short position in delta shares of the underlying asset so that the return on the portfolio is risk-free rate. Given theta, gamma of a long position in option is always positive, or in other words delta is an increasing function of stock price. The delta-hedged gain, which is in excess of risk-free rate, could be realized by selling more stocks when share price increases (thus delta increases as well) and buying stocks back when share prices drops (thus delta reduces). This strategy dynamically hedges the underlying asset movement, and it is only exposed to volatility risk in short term.

by how fast the spot prices moves, or namely the volatility. In this paper I choose 2-week window as forecast length due to the tenor of the volatility swap contracts used in the trading strategy tests. Currency options with 2-week maturity have high liquidity and a great amount of transaction volume, which have a similar market size as single-week options. The process presented in this paper can be adapted to any other forecast windows.

Another key difference from previous literature is this paper does not stop at the stage that comparing in-sample fit and out-of-sample squared error. Instead, to better address the dynamics and mimic the real trading, I run a rolling in-and-out-of-sample from 2007 to 2010, resulting in more than 1,000 forecast numbers. Each calibration and forecast only encompasses the information available at that time to avoid hindsight bias. Moreover, a trading strategy is constructed based on the deviation of the forecast from volatility swap strike. With the cumulative P&L (Profit and Loss) and Sharpe Ratio calculated at the end of 2010, this paper exhibits a more comprehensive picture of forecast performance than what delivered by MSE or the Mincer-Zarnowitz R^2. The profit following such a strategy is significant even after accounting for transaction cost, additionally making it practical and realistic.

The rest of the paper is organized as follows: section 2 reviews the literature related to volatility forecasting, high frequency seasonality, and volatility swap. Data sources, general descriptions and methodology are detailed in Section 3. Section 4 demonstrates in-sample fit, out-of-sample predictability as well as volatility swap trading payoffs. Robustness check and comments are laid out in Section 5, while Section 6 concludes the paper.

2 Related Literature

There is a massive amount of research regarding volatility models. Generalized autoregressive conditional heteroskedasticity (GARCH, Bollerslev, 1986) and its derivatives such as GJR (Glosten, Jagannathan and Runkle, 1993), EGARCH (Nelson, 1991), IGARCH (see Bollerslev and Wooldridge 1992 for a survey of GARCH literature) are the most widely applied group. Some papers (Diebold, 1986, and Lamoureux and Lastrapes, 1990) argued that high estimated persistence may originate from structural switches in the variance process, which was not addressed by traditional GARCH models. Their argument is clearly related to the findings of Perron (1989), Cai (1994), Wong and Li (2001), and Lanne and Saikkonen (2003).

Following this argument, Hamilton (1989) raised the idea of applying Markov process on the estimation of autoregressive parameters. Enlightened by it, Lam (1990) and Kim (1994) improved the algorithm; moreover, Hamilton and Susmel (1994), Gray (1996), Klassen (2002), and Haas et al (2004) proposed different theoretical approaches to handle Markov Regime-switch GARCH, and provided practical ways to circumvent the path-dependence problem. More papers (Marcucci, 2005, Cheung and Miu, 2009) showed empirically that Regime-switch GARCH does show a significant outperformance than more traditional single-regime GARCH.

As an alternative, some studies such as Day and Lewis (1993), Jorion (2005) and Christensen and Prabhala (1998) argued implied volatility from traded options data does a better job in predicting future volatility, since it takes market consensus and have forward-looking advantage. The traditional employed implied volatility was a single number backed out from B-S formula, represented by delta-neutral or ATM option. Recent literature improved this calculation and turned to model-free implied volatility. Following the model-free implied volatility derived by Britten-Jones and Neuberger (2000), Jiang and Tian (2005) implemented it using a practical method.

The difficulty surrounding high-frequency GARCH model concentrates in seasonality, autocorrelation, and pre-scheduled macroeconomic data release. Various papers (Anderson and Bollerslev, 1997b, 1998b; Taylor and Xu, 1997; Beltratti and Morana, 1999; Gencay, Selcuk and Whitcher 2001a) suggested different approaches to handle seasonality problem. The typical methods include Flexible Fourier Form (Anderson and Bollerslev, 1997b; Cai et al, 2001, Dominquez and Panthaki, 2006, Laakkonen, 2007); Locally Weighted Scatterplot Smoothing method (Cleveland, 1979) and Intraday Average Observations Model (Omrane and Giot, 2005). Bauwens et al. (2005) also applied the Intraday Average Observations Model on macro news, which is close to the filtering procedure employed in this paper.

Apart from the above literature which either has its technique directly applied in this paper, or aims at tackling similar question, a number of other papers are closely related to this research, especially the discussion regarding to realized volatility measure. Fundamental econometric analysis of realized variance or volatility is studied by Andersen, Bollerslev, Diebold and Labys (2003) and Barndorff-Nielsen and Shephard (2002); Barndorff-Nielsen, Hansen, Lude and shepherd (2008a) introduced realized-kernel estimator. For a more concrete survey, see Andersen, Bollerslev and Diebold (2009).

Papers related to variance/volatility trading and continuous delta-hedging, which is partly employed in this paper at the strategy section, include Neuberger (1990), Carr and Madan (1998), and Carr and Lee (2009).

3 Data and Methodology

3.1 Data

Data in this paper stems from several sources: the full sample of EURUSD spot price is from January 2004 to December 2010. I obtain 1-minute BID and ASK prices from Reuters, and interpolate missing minutes by repeating the last observations – thus mimicking the live data feed and in line with market practice. Since EURUSD is the mostly traded and liquid currency pair, the quality of the raw data is very robust. In 7 years, only less than 1% of the data has been interpolated. Later I select hourly data from the complete minute set because according to previous intraday research (Andersen and Bollerslev), the hourly frequency has already got rid of the bid-ask bounce and become nearly un-correlated, while still retaining the intense intraday information. Original trading data's time tick is GMT; while event calendar and seasonality are both consistent with local market (recorded in London time) trading hours, which have one hour lag to GMT in the summer. To be compatible, all the time tick is modified to London time according to daylight savings. Furthermore, following the global FX market open and close hours, I define one week trading period as from 9pm Sunday (London time, when the Australian market opens) to 9pm Friday (London time, when the North American market closes). Any data which falls out of this range is deleted.

Following Andersen and Bollerslev return calculation in high frequency data, I estimate the MID price as the geometric mean of BID and ASK[2]. Hourly return is calculated as the difference of log MID price and its one hour lag, and hourly variance as return square.

$$
\begin{aligned}
P_t &= \sqrt{A_t \cdot B_t} \\
r_t &= (\log P_t - \log P_{t-1}) * 100 \\
\sigma_t^2 &= r_t^2
\end{aligned}
$$

The first out-of-sample forecast window starts from 01 Jan 2007, so I list general statistics of the

[2] I also test on the more popular definition of MID price as 0.5*(ASK+BID). The results look very similar.

whole sample period and those after 2007 separately in Table 1. The whole picture of hourly return and variance are shown in Figure 1. Even in a glance we are able to catch the wild fluctuation around the credit crunch and European sovereign debt crisis. The nature of the volatility process is ever-changing and it is necessary to base the forecast on a rolling calibration window.

The second data source contains the Macroeconomic data release calendar. From Bloomberg I collect the influential pre-scheduled news announcements in the US and Europe (which mainly consists German and EU news) for the period 2004-1010. Table 2 exhibits a summary of events.

The third part encloses data from the options market. EURUSD 2-week options are considered because this tenor is typically viewed as the shortest length for volatility swap contracts. To be in line with the trading strategy built in the later section of the paper, here I focus on the options with this short-dated window. The option data is recorded daily at 5pm New York time, and the options expire at 3pm New York time 2 weeks ahead.

I take two parallel measurements of implied volatility: the first one is commonly used delta-neutral implied volatility (hereafter DNIV) for options with 2-week tenor, backed out from Black-Scholes formula. The second measurement is so-called "model-free" implied volatility (hereafter MFIV) and takes the entire volatility skew or smile under consideration. I follow the method proposed by Jiang and Tian (2005), including all available options with the same 2-week maturity but at various strike prices. Slightly different from Jiang and Tian (2005), I relax their assumption that interest rate is zero, and download interest curves for both USD and EUR in calculating forward exchange rate for a given strike price. The formula to compute model-free implied variance is as follows:

$$E^Q\left(\int_0^T (\frac{dS_t}{S_t})^2\right) = 2\int_0^\infty \frac{C(T,F(K)) - \max(S_0 - K, 0)}{K^2} dK$$

where $F(K)$ is forward exchange rate for strike price level K, and it can be attained by interest rate parity.

The above formula is the risk-neutral expectation of implied variance. However, due to Jensen's Inequality, as pointed out by Britten-Jones and Neuberger's (2000), the square root of right side would be upward-biased estimator of implied volatility.

$$E^Q\left(\sqrt{\int_0^T (\frac{dS_t}{S_t})^2}\right) \leq \sqrt{2\int_0^\infty \frac{C(T,F(K)) - \max(S_0 - K, 0)}{K^2} dK}$$

Meanwhile, the MFIV by nature is expected to be higher than DNIV owning to the fact that OTM options are generally more expensive than ATM ones (volatility smile). Hence we can anticipate the true implied volatility should lie somewhere between DNIV and the number on the right side.

The data for DNIV stretches throughout the whole forecast period (2007 to 2010). Unfortunately the more complete option data on spectral strikes is only available after Oct 2008, so the MFIV is only able to race half way. As seen from Figure 2 the disagreement between these two time-series is mostly trivial except for the period around credit crisis where the volatility skew is more evident. The insignificant difference is also confirmed by the t-statistics in Table 3. Rendering to these facts, for the rest of this paper I only use DNIV for its longer history. Nonetheless, we shall keep in mind that neither of the two concepts is free from risk premium and thus both will exhibit the upward tilt while forecasting future realized volatility.

3.2 Methodology

3.2.1 Seasonality and Event

After the preliminary filter using a week-frame from 9pm Sunday to 9pm Friday, I apply a two-step approach to clean out the effects of seasonality and of data releases by assuming that the influence from these two are independent.

Thanks to high efficiency in the FX market, prices usually catch up very rapidly following events and the fluctuation is highly concentrated around the exact release time. So I simply define event time as the hour range within which there is a macroeconomic data release. When dealing with seasonality alone, I only select the non-event hours so that the distortion from event effect is controlled.

To remove the stylized intraweek autocorrelation, I follow the intuition from the Intraday Average Observations Model which was first introduced by Bauwens, Ben Omrane and Giot (2005), and modify it to serve my purpose. The following process is repeated in every rolling in-sample window, which covers the most recent 20-week data (2400 observations) immediately before the forecast window. Rolling samples ensure no hindsight.

One week is divided into 120 time-buckets by hour, with each hour indexed by "hour-of-the-week" h=1, 2, ..., 120. After picking the non-event hours, I calculate the average hourly variance for each h across the in-sample period:

$$V_h^{ne} = \frac{1}{W} \sum_{w=1}^{W} V_{w,h}^{ne}$$

where w=1, 2,..., 20 is week index, W=20 and the superscript ne denotes non-event time. V is hourly variance as stated in last session (σ_t^2).

This gives us a picture of how on average volatility changes from hour to hour in a week. Seasonality adjustment coefficient can be computed as the ratio between each hour-of-the-week's variance and its weekly average:

$$S_h = \frac{V_h^{ne}}{\frac{1}{120} \sum_{h=1}^{120} V_h^{ne}}$$

Here $\frac{1}{120} \sum_{h=1}^{120} S_h = 1$, which indicates if normalizing the original variance data by dividing S_h respectively, it will not change the overall weekly variance, but only peel off the part attached with seasonality. In other words, we expect the adjusted time series variance $\widetilde{V}_{w,h}$ is clean of intraday or intraweek pattern, where

$$\widetilde{V}_{w,h} = \frac{V_{w,h}}{S_h}$$

Notice that in the numerator it is $V_{w,h}$ instead of $V_{w,h}^{ne}$ even though I only take the latter to calculate adjustment ratio at the first place. It comes from the assumption that seasonality and event impacts are independent, hence the resulted variance $\widetilde{V}_{w,h}$ will be only exposed to data release for event hours (additional to other irregular factors that affect volatility).

The rest of the task is to get rid of the part associated with events. I do not treat all types of data releases equally since some of them have a much more significant influence in FX market than others. For instance, currency markets are often more sensitive to central bank interest rate decisions and inflation data (due to carry trade), and growth and unemployment data (due to risk appetite), compared to other releases. On the other hand, some other events do not perform as big drivers as these. However, if looking at single event type separately, there are not enough observations in 20 weeks to reach a solid statistics for its effect, as most of the data releases are monthly.

To circumvent this problem I take an anchored window starting from 2004 and first implement the above method to strip out seasonality. The coefficient of influence of Event i is calculated as:

$$E_i = \frac{\frac{1}{N_i}\sum_{j=1}^{N_i} \widetilde{V}_j^i}{\frac{1}{N_{ne}}\sum_{k=1}^{N_{ne}} \widetilde{V}_k^{ne}}$$

where N_i is observation of event i; and N_{ne} is the number of non-event time.

Finally, the clean variance and return can be obtained by

$$\begin{aligned} \widetilde{\widetilde{V}}_{w,h} &= \frac{\widetilde{V}_{w,h}}{E_i I(i) + 1 \cdot (1 - I(i))} \\ \widetilde{\widetilde{r}}_{w,h} &= \sqrt{\widetilde{\widetilde{V}}_{w,h}} \cdot I(r_{w,h} \geq 0) + (-1)\sqrt{\widetilde{\widetilde{V}}_{w,h}} \cdot I(r_{w,h} < 0) \end{aligned}$$

where *I(i)* is the indicator function which is equal to 1 if event i happens during the hour *(w,h)*, or 0 otherwise. Likewise $I(r_{w,h} \geq 0)$ is the indicator that is 1 if the original return is positive.

The above practice can be viewed as "cleansing" before feeding the data into forecast engine. After the double-step filtering, $\widetilde{\widetilde{r}}_{w,h}$ is regarded as return free from seasonality and event influence, and therefore qualifies for recursive calibration of the fitting models (such as GARCH). While after forecast results from empirical models come out, the reversed procedure is attached to "re-pollute" the data by injecting seasonality and event effects back into the forecasted time series.

In the end, the 2-week volatility forecast will be

$$V_{2w}^{fcst} = \sqrt{\frac{252 \times 24}{H}\sum_{h=1}^{H}(V_h^{fcst})^2}$$

where V_h^{fcst} is each single hour's forecasted volatility after "re-pollution"; H is number of hours till the option expires; and V_{2w}^{fcst} is the final forecast for annualized volatility.

3.2.2 Forecasting Models

Among numerous volatility models, I take in the most commonly exploited ones: GARCH, GJR, EWMA (exponentially-weighted moving average), and MA (moving average).

GARCH: $\sigma_t^2 = K + \alpha \cdot \varepsilon_{t-1}^2 + \beta \cdot \sigma_{t-1}^2$

GARCH is the most traditional empirical volatility model, which captures long-term variance (K), consequence from instantaneous innovation (α) and memory from past shocks (β).

GJR: $\sigma_t^2 = K + \alpha \cdot \varepsilon_{t-1}^2 + \beta \cdot \sigma_{t-1}^2 + \theta \cdot \varepsilon_{t-1}^2 \cdot I(\varepsilon_{t-1} < 0)$

In the GJR model the volatility is treated asymmetrically depending on the instantaneous return: negative return has a stronger impact (θ>0).

EWMA: $\sigma_t^2 = (1-\lambda) \cdot \varepsilon_{t-1}^2 + \lambda \cdot \sigma_{t-1}^2$

EWMA is an alteration of MA, setting more weights to the more recent observations than history. In the literature, Riskmetrics EWMA has been widely used in which λ for monthly data is fixed to 0.97 and daily 0.94. In this paper I do not stick to a constant λ, but calibrate it via a rolling sample as GARCH family models.

MA: $\sigma_t^2 = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_{t-i}^2$

MA is an easy, but powerful method as we will see later. One item that needs mentioning is that the in-sample size for the MA estimate is different from others. For models that ask for calibration, a sufficiently large dataset is needed and the recursive nature will automatically set less weight to older samples. However, simple moving average does not discriminate on data history, and I have to exogenously confine the sample N to a shorter period so that the stale information would not be included. Here N is set to 240 (2-week data, similar as the forecast window length).

Except for the MA case, each model goes through a Markov regime-switch technique in order to enhance its adaptability to changing market conditions. I assume that there are 3 volatility regimes (low, medium, and high) and each can be represented by a set of coefficients. Ideally, if states can be explicitly identified, the problem will be reduced to single-regime model that may estimate and forecast individually. However, the real state is unobservable and different regimes can transmit between each other across time. As a result we are only able to make an ex-ante expectation of the future regime or an ex-post expectation of the current regime by Bayes updating, but even these expectations are hard to attain if at all possible. The path-dependence characteristic of regime-switching recursive model makes the full identification unfeasible, and in the existing literature there are different means to approximate the current conditional variance (Gray 1996, Klaassen 2002). In this paper I follow the approach proposed by Haas et al (2004) and only allow the updating conditional variance occur independently within regime. Since this paper is not focused on specific technique, I will not dig deep into the argument on selecting the best approximation. The following example of 3-regime GARCH model illustrates the main algorithm. Similar idea is applied on other empirical models.

A 3-regime GARCH model can be written as:

$$h_{t+1}^i = K^i + \alpha^i \cdot (\varepsilon_t^i)^2 + \beta^i \cdot h_t^i$$

where i=1,2,3 stands for different regimes and h_{t+1}^i is the conditional variance at time $t+1$ for regime i. The conditional expectation of variance at $t+1$ is determined by both h_t^i and $P(s_{t+1} = i|F_t)$, which is the conditional probability that the future state will be i given all information until t:

$$h_{t+1} = \sum_{i=1}^{3} h_{t+1}^i P(s_{t+1} = i|F_t)$$

Assuming a first-order Markov chain, the second part of the right side can be written as

$$P(s_{t+1} = i|F_t) = \sum_{j=1}^{3} P(s_{t+1} = i, s_t = j|F_t) = \sum_{j=1}^{3} P(s_{t+1} = i|s_t = j)P(s_t = j|F_t)$$

On the right side, the first item $P(s_{t+1} = i|s_t = j)$ is decided by a 3×3 transaction probability matrix which is fixed across the time (for a given state j, the probability that it will change into state i next period remains at P_{ij}). The second item $P(s_t = j|F_t)$ denotes the ex-post expectation of state after the observation at time t arrives:

$$P(s_t = j|F_t) = P(s_t = j|F_{t-1}, r_t) = \frac{f(s_t = j, r_t|F_{t-1})}{f(r_t|F_{t-1})} = \frac{f(r_t|s_t = j)P(s_t = j|F_{t-1})}{\sum_{j=1}^{3} f(r_t|s_t = j)P(s_t = j|F_{t-1})}$$

which in turn depends on the previous time's ex-ante expected probability. The recursive equations above demonstrate how the conditional expectations are formed before and after a certain data point arrives. In order to get the whole successive system working, an initial regime probability, $P(s_0 = j|F_0)$ needs to be fed in, and conventionally it is chosen as an unconditional or stationary probability. Please refer to the appendix for details.

In the end, the innovation is assumed to follow a normal distribution, i.e.,

$$f(r_t|s_t = j) = \frac{1}{\sqrt{2\pi h_t^j}} \exp\left(-\frac{(r_t - c)^2}{2h_t^j}\right)$$

Then all the parameters can be optimized through MLE (Maximum Likelihood Estimation) method by maximizing the log-likelihood function

$$\ln L = \sum_{t=1}^{T} f(r_t|\digamma_{t-1}) = \sum_{t=1}^{T} \ln \left[\sum_{i=1}^{3} f(r_t|s_t = i) P(s_t = i|\digamma_{t-1}) \right]$$

4 Emiprical Results

4.1 Seasonality and Event Adjustment

Figure 3 is the autocorrelogram of hourly variance within one week before and after seasonality/event adjustment. The raw variance exhibits a clear pattern of seasonality, and the peak autocorrelation occurs around every 24 hours. This corresponds to the previous research. The hourly variance after cleansing appears to be more random. Even though not all the autocorrelation coefficients strictly collapse to a narrow range, the regular form has vanished.

If we plot the average seasonality adjustment ratio S_h (Figure 4), we are able to see a quite similar shape as the first graph of Figure 3. The adjustment ratio is very high at the very beginning of the week, and is followed by an immediate sharp drop, which is simply because the first hour's return is calculated as the difference between the log price of the first hour on Sunday and that of the last hour on previous Friday. Since the first open hour will reflect the information released during the weekend, it is reasonable to expect a larger jump between these two numbers at this time bucket than during continuous hours. Throughout the week, double peaks can be spotted every day, which signify the two important overlapping periods across time zone: the first happens when the Asian market and the European market both open; and the second is when the European market intersect with the North American market. The analogous outline of the adjustment ratio S_h and raw variance autocorrelation enlightens the fact that when dividing the raw data by S_h, it will result a time series that bears no autocorrelation.

The US dollar, often treated as a safe-haven currency, has exposed itself to news via heterogeneous aspects. When a US data release shows positive surprise, it works both in favour and against the USD: a brighter US economic outlook favours the greenback, while a rise in risk appetite often works against it – since the USD is the world's most used funding currency. Under different market environments, these contradictory interpretations and forces lead to varied price reactions to similar macroeconomic surprises. However, while the direction of the spot move may be time-variant around these events, the impact on variance is generally stable: it usually surges.

Table 4 summarizes the impact from data releases. Several interesting discoveries are worth addressing. First, the influences from different events diverge a lot. The most influential is the US Fed Rate decision, which is announced at monthly FOMC Meeting. Since short-term exchange rate moves are largely determined by interest rate differentials, investors weight heavily on the shocks brought by the Fed. This became more evident around the period of Quantitative Easing I and II. Second, the US NFP (Non-Farm Payroll) and Unemployment Rate are always announced in the same report, so altogether these two share a large weight. On the other side, some event like EU Unemployment Rate does not stir the market, either because investors do not regard it as important, or because European data often produces less surprises to the market.

4.2 Forecast Results

As discussed in the introduction, the true realized volatility is unobservable and subject to definition, or more concretely, the data frequency in use. This becomes a potential problem of all the research on forecasting methods as long as the realized volatility is used as a benchmark or criteria to assess models. In my opinion, a definition of realized volatility is ought to correspond to the strategy that is built around it. For instance, in the context of a volatility swap, realized volatility should be compatible with what is stated in the contract, which is conventionally the square root of annualized daily return square; while in the case of a dynamic delta-hedging strategy, the realized volatility calculation should take the same data frequency as that of rebalance or hedging.

Later in this paper I will construct a volatility swap strategy based on the deviation of forecast and strike price. In this context, when evaluating the models' out-of-sample predictability I calculate realized volatility as defined in a volatility swap, which is based on daily return. In line with forecast length, the window is also set to 2 weeks. The daily price is used and the result is annualized.

$$RV_d = \sqrt{\frac{252}{N}\sum_{t=1}^{N}(\log P_t - \log P_{t-1})^2}$$

The realized volatility throughout 4 years is displayed in Figure 5. If we compare it with Figure 2 and 8, we can see that the realized volatility from a daily calculation is much spikier than the hourly peer, mainly due to noise involved in daily data.

Two conventional criteria, SE (Squared Error) and AE (Absolute Error) from each model are listed in Table 5. Both of them evaluate how the forecasts deviate from the real values, but SE penalizes outliers more. Surprisingly, MA is the best performer in terms of either standard; while what is even more unexpected is the forecasting power of implied volatility is the weakest among all the candidates. The first finding matches what was argued by Andersen, Bollerslev, Diebold and Meddahi (2002) that past integrated volatility had a superior performance than more complicated GARCH models, though I will show in Section 4.4 that if simply take recent realized volatility without seasonality or effect adjustment, the result will be inferior. The second discovery, however, is different from Jorion (1995) who claimed that implied volatility forecasted better than empirical models.

As I will discuss in Section 4.4, DNIV actually outperforms if calculating realized volatility using hourly data, which indicates a more appropriate application of DNIV in a delta-neutral strategy with hourly rebalancing.

Table 5 also presents the Mincer-Zarnowitz R^2, another popular measurement to assess predictability. The idea is to regress out-of-sample realized volatility on the various forecasts and calculate the explanatory power (R^2) of each. All the models show reasonably good fits with MA marginally outperforming others.

So far with the full hindsight, we may argue that the moving average on high frequency data has the best predictability of future (lower frequency) realized volatility; however, it does not rule out the possibility that different models have complementary effect such that a "portfolio" of methods might outperform any single one. Hence, for the next step I carry out an experiment on three different ways to combine the signals, and two other ways to select models on the way. To avoid data mining, the choices are set as simple as possible:

Signal Combination:

[C1] Weighted by the accuracy (inverse of SE) of the latest forecast;

[C2] Weighted by the average accuracy (inverse of the MSE – Mean Squared Error) since the back-test began;

[C3] Equally weighted;

Signal Selection:

[S1] Only apply the model which has the smallest SE in the latest forecast;

[S2] Only apply the model which has the smallest MSE so far.

The difference between signal combination and selection is that the former always takes the full information from every model; the methods that perform better in the most recent history ([C1]) or on average ([C2]) are allocated heavier weights. [C3] is the most straightforward way to the get a collective signal without any discrimination of any model. The signal selection, though also keeping tracking of all the records, only relies on *one* of the forecasts each time. Corresponding to a similar construction idea, [S1] and selection [S2] respectively choose the best latest model and best model on average until the forecast date. The results are presented in Table 6.

These signal combination approaches, especially [C2], have outperformed the best method alone in terms of squared errors and Mincer-Zarnowitz R^2. MSEs have been improved, and the forecasts turn out to be more stable (the standard deviations of the errors are smaller). The MAE of the moving average method is still the smallest even including full battery, which shows the improvement from signal combination is larger on the outliers.

Signal selection, with its emphasis still on the single method, does not improve the result. However, [S2] actually provides us with a more realistic picture that excludes forward-look bias. Imagine an investor selects the model on the way by monitoring each one's performance. Every day she gets a snapshot of Table 5, and applies which is shown to be the best model. She will end up with the last row of Table 6 at the end of 2010.

4.3 Trading Strategy

A trading strategy based on the prediction can be a more practical test to gauge model performance, which also corresponds to the motivation of seeking a superior forecast model. In the case of volatility forecast, one direct and straightforward application is the volatility swap, a contract to exchange a fixed amount (strike) against a varying number that depends on the volatility realized within certain period. If a trader believes the future volatility should be significantly higher than the strike price, she will be the swap buyer; and vice versa. A long position in volatility swap is also widely used as a hedge to the underlying asset since the negative correlation between an asset's spot price and its volatility is commonly observed.

Unfortunately, like with most other swap contracts, volatility swaps are traded over the counter and historical price data is hardly available. There are substitutes, though, to approximate it. One way is to estimate strike through adjusting DNIV by skew. Skew captures the relative cheapness of OTM calls against OTM puts, and since volatility swap can be (imprecisely) replicated by options with a spectrum of strikes, the whole skew is under consideration in pricing volatility swap strike. This approximation by intuition is effectively the same as MFIV, which has been discussed in Section 3. There are disadvantages of MFIV, however, including partial reliance of the accuracy on Black-Scholes formula (despite the fact that it is called "model free") and its static nature. In reality, volatility traders usually apply a more computational intensive model SLV (stochastic-local volatility model) to dynamically set weights to stochastic volatility and local volatility, and blend them into the final volatility swap price.

In this paper, I employ a similar way as SLV to estimate historical volatility swap prices. The result is regarded as the volatility swap mid strike. To better mimic the real trading, a conventional spread (0.5 volatility points) is added (subtracted) to represent the Ask (Bid) price. Therefore, the fact that the forecast is different from strike Mid is not necessary to trigger a position; only when the deviation is big enough, a long or short position will be entered. I set a pre-screening threshold 10% to ensure that only significant mispricing will be exploited and therefore, the profit is more likely to overcome the transaction cost.

I start with the simplest strategy that always enters 1 unit of capital per vega point (i.e. vega notional = 1) once the enter signal is triggered. Suppose an investor joins the market with 10 units of vega notional at the beginning of 2007[3]. Every day after making forecast, she is allowed to buy (sell) one unit of vega notional if the forecast is 10% higher (lower) of the strike MID. Each swap has two-week tenor and the investor always holds the contract until its maturity. Therefore, the maximum absolution position one investor can have is 10 vega notional. The realized P&L (Profit and Loss) at the end of two weeks is calculated as:

$$P\&L = \pm 1 \times (RV_d^{2w} - strike) \times 100$$

The sign before the position is positive (negative) if she is long (short) the volatility swap. Correspondingly, strike price is Ask (Bid) if it is a long (short) position.

I test the strategy performance according to every single model as well as the combined and selected signals. In addition, since the volatility swap is regarded attractive to sellers thanks to the associated risk premium, I include a naïve short strategy as a benchmark, which enters a short volatility swap position (with 1 unit vega notional) every day without any signal filtering.

[3]Note the initial vega notional is not investment cost, as in a volatility swap the notional is not really exchanged.

Table 7 displays the annualized P&L, standard deviation and Sharpe ratio of each alternative. I use P&L instead of investment return is because the volatility swap is virtually a forward contract and the notional is simply a scalar. Or in other words, the investment cost for the trading strategy is merely zero. Therefore, it would be misleading to use returns.

It is interesting to see the discrepancy of real-life strategy performance against the forecast error analysis in Table 5. GARCH, which showed the worst MSE as well as MAE, turns out to be the biggest winner when applied to trading in terms of annualized P&L, Sharpe ratio and the ratio between average win and loss. We can understand this inconsistence by noting the standard deviation of P&L from GARCH model is also far the highest, which is in line with its relative poor predictability: it misses the goal very often (in fact most of the time, see the hit ratio in Table 8); however, it shines once it gets the forecast right, which, is often the case when volatility shoots high. The results from the MA method are somewhat opposite: the forecasts are very accurate (lowest MSE and MAE), the hit ratios are high (see Table 8), the returns and risk are low but the returns in an average losing trade outweigh that of the average winning trade. No single model universally outperforms others. This outcome also confirms the necessity of investigating model forecasts in the context of trading environment.

However, the tradeoffs between different styles start to dissolve when I take signals from a battery of models. The group of model combination delivers significantly improved performance than individual models by both enhancing the average P&L and reducing the risk (slightly higher than MA, though). The annual Sharpe ratios have increased by 40%-100% than the best single GARCH model.

The Naïve Short strategy results in a negative average payoff and far more volatile return. In fact, had an investor blindly gone short volatility during financial crisis, she would have incurred a massive loss, as shown in Figure 6 and Table 9.

I summarize more information about the long/short positions and other return metrics in Table 8. As we expected, empirical model predictions are more prone to enter a short position since the volatility swap strike is often higher than the "fair" value, and average hit ratio is also much higher for short positions. The outcome from the naïve short shows that 47.44% of the time the volatility swap strike has overestimated subsequent realized volatility even after adjusting transaction cost. While each individual model's short signal is trustworthier than the naïve short, as we can see from higher hit ratios for short positions, the long signal often results in a loss. However, this does not necessarily imply it is meaningless to enter a long position, since the upside from buying volatility can be much higher than the downside, due to the limit of lower bound of volatility. What we have experienced during financial crisis is the best proof, and the models do register their best performance in that period as a sharp contrast against the naïve short strategy (Figure 6).

Table 8 also shows Investors would trade less frequently in model combinations (smaller position percentage), which can be understood as "average effect" of taking collective signals from all the models. Hence extreme forecast figures will be alleviated, leaving a more stable and reliable prediction, and long or short signals may only be pronounced when individual methods confirm with each other.

Figure 6 plots the cumulative P&L for single models and the naïve short strategy. We are able to see that MA works the best in quiet period as prior to 2008; while in Q4 2008 GARCH quickly takes over. In fact, the majority of the profits are realized in high volatility regime such as financial crisis, since this is right the time that the models pick up the correct buy signals and benefit from the large volatility space in upside. While in normal/calmer period, risk premium and transaction costs make it hard to beat the benchmark. On the other side, the profit captured in the naïve short is allied with risk premium. During the financial crisis, the risk premium did not catch up quickly enough with the sudden and massive surge of the realized volatility; however, in the aftermath of the crisis, with little risk appetite remaining, risk premium did not recede with the realized volatility and hence the naïve short strategy profited from the volatility swap overpricing. Therefore, the profit of naïve short is closely related to contemporaneous discrepancy of market sentiment and realized volatility. Another strong hit in mid-2010 can be noticed, caused by the European sovereign risk and another wave of risk aversion.

Figure 7 shows the cumulative P&L for different model combinations, which gives a clearer picture of how the collective signal takes effect. The rationale behind it is that different models register outperformance in different regimes. As we see from Figure 7, before the financial crisis the combined model escapes from the losses as most of single models have, but performs as well as the best single model in that period, MA; while once the crisis kicks off, the combined model quickly takes the full power from GARCH and GJR, by entering full capacity of position and consistently making profit.

In building a trading strategy, the maximum drawdown is an important concern from a risk management perspective. The naïve short strategy is very vulnerable to drawdowns and can be easily stopped out once the sharp drop in cumulative return happens. As seen in Table 9, all other models are much steadier in terms of maximum drawdowns and recovery time. More impressively none of them witnessed significant drawdowns during the volatility spike of Q4 2008. Model combinations again show a superior performance.

One alternative is to replace DNIV by volatility swap strike when combining or selecting signals, since there are sufficient reasons to believe the strike contains more comprehensive information than DNIV (the former contains message from the whole volatility surface rather than a single point), even though it can be argued that swap strike is not public information. Indeed this improves the risk-adjusted performance even further by both increasing the return and reducing the risk. To conserve space I will only present the results from signal combinations in Table 10.

So far, we always enter with one unit of vega notional per position regardless of how different the forecast is from the swap strike, as long as it passes the 10% pre-screening criterion. It is possible to outperform even more by scaling the position according to the distance between forecast and strike. I set the position as follows:

$$pos_i = \pm 1 \times [1 + abs(fcst_i - strike)/strike]$$

So the idea that the more positive the spread between model forecast and current strike is, the more the trader buys; the more negative the spread is, the more the trader sells. Note that since now the lower bound of the position is 1.1 vega units (because the pre-condition to enter a position is that the forecast needs to be at least 10% away from the strike) and most of the time it would be larger, and therefore we should not compare the absolute levels of average P&L and standard deviation with previous un-scaled cases. Nevertheless, the annual Sharpe ratio still conveys the useful information in Table 11:

From Table 11, the risk-adjusted payoffs have enhanced substantially in all cases by the extra information acquired from the relative strength of the forecast from the strike. Once again corresponding to what was found in Table 10, using swap strike instead of DNIV also improves the results.

5 Robustness Check

In this section I carry out some alterations of the main procedure, in order to check (1) if the model presented above outperforms simple methods; and (2) whether changing some circumstances will deteriorate the model's robustness.

5.1 Daily Data

The very first thing is to verify the impact that changing data frequency has on results; specifically, whether the results deteriorate if we replace intra-day data with daily data as usually done in models that forecast volatility over longer windows. If the results are similar, there is no gain from the computational hassle associated with hourly data. In Table 12 I present the model performance in daily version. As expected, high frequency data does bring in additional information that is not captured in daily sample. The squared errors and absolute errors are much larger as well as more volatile than in the intraday version, implying the employment of intraday data does lead to a higher accuracy and stability. The annual Sharpe ratios following these forecasts also exhibit an obvious drop, and the MA model's Sharpe ratio even drops down to -1.68. As seen before, moving average in high frequency data gives a very appealing outcome; but in the daily version it is much inferior to other more sophisticated econometric models. I also list the model combinations in Table 12. Similar as above, collective signals improve model prediction and trading payoff even when individual models perform poorly.

5.2 No Seasonality or Event Adjustments

Having confirmed the merits of using intra-day data, the next task is to check whether the seasonality and event adjustment is proper. I rerun the whole process by simply inputting the raw hourly data without taking care of seasonality or macroeconomic data releases. The results are as in Table 13.

Compared to Table 5, the model predictability is much poorer if seasonality and event impact are not removed, as the statistics for squared errors and absolute errors are significantly higher. The difference is more pronounced in GARCH and GJR, which makes sense since this type of models is famous for failing to deal with high frequency data due to seasonality. A similar conclusion can be drawn from the comparison of Sharpe ratios when applying signals generated from forecasts with or without seasonality amendments. In the case of no adjustment, GARCH, GJR, and MA all end up with a negative payoff, but unexpectedly EWMA leads to a positive Sharpe ratio, which is even higher than its counterpart with adjustment. Nevertheless this improvement is much minor compared to other models' deterioration. All in all, the procedure regarding seasonality and event boosts the predictability, and in fact, it is one of the critical steps to make the strategy work.

5.3 Lagged Realized Volatility

Motivated by the MA's superior performanc, here I test if other easy ways such as lagged realized volatility or its various versions of moving average can also reach a decent forecast. Notice in this way only daily data is involved and no intraday information gets employed.

From Table 14, none of the easy ways have a comparable performance as the intraday model in terms of the accuracy of the forecast. The Sharpe ratios are mostly negative.

5.4 Fixed-parameter EWMA

A more common way to apply EWMA is as Riskmetrics does, which fixes λ at 0.97 for monthly data or 0.94 for daily data. In my calibration I keep λ varying so that the parameter may adapt to the most recent information. As a comparison, I also test a static λ on EWMA. The results in Table 15 confirm the rolling calibration does lead to a better forecast.

5.5 Realized Volatility Calculated in Hourly Data

As mentioned before, true realized volatility is an unobservable data process and subject to the definition, which in turn should correspond to the frequency that the trading process is trying to exploit. In the previous sessions I choose to calculate the realized volatility in terms of the square root of annualized squared daily returns, because the payoff of a volatility swap contract is set in that fashion. However, it is not the only way to look at model application, especially when my forecast window is more flexible intraday.

For instance, if I assume that the purpose of the forecast is to serve delta-neutral (or gamma) trading strategy with the hedging/taking profit frequency on hourly basis. In this case, since the profit in the short end exclusively depends on the volatility realized each hour, hourly volatility is more relevant.

$$RV_h = \sqrt{\frac{252 \times 24}{N} \sum_{t=1}^{N} (\log P_t - \log P_{t-1})^2}$$

Compared to Figure 5, the realized volatility in Figure 8 is much smoother, which makes sense since the noise embedded in daily data gets alleviated. I also re-construct the forecast performance table for single models, as shown in Table 16.

The most noticeable change as a comparison to Table 6 is that model predictability is significantly higher when using intraday realized volatility, which shows a even more profitable potential application could be reached in high frequency trading strategy other than volatility swap. Among single methods, MA is still the best approach in terms of all the measurements. That said, delta-neutral implied volatility, which used to be the worst performer to forecast realized volatility calculated from daily data, now becomes one of the top candidates. This indeed should be the case since option prices do reflect intraday information, and the best approach to exploit the information should also extend into higher frequency.

When forecasting in intra-day horizons, the seasonality and event adjustments become even more critical, since this is the case when the individual time bucket really presents its idiosyncratic effect. In Table 17, the result shows model forecast performance without data cleansing, compared to the realized volatility on an hourly basis. Once again it confirms that the seasonality and event adjustments do improve the overall predictability.

6 Conclusion

In order to forecast short-dated volatility in a highly efficient market, this paper explores a wide range of empirical models as well as the implied volatility, which is often perceived as the market's own volatility forecast. A Markov regime switch technique and rolling in-and-out-of sample separations have been applied in order to improve the adaptability. In contrast with previous literature that usually promotes one given method, this paper argues that either no single model can consistently capture the best available information or that, if there is one, it is unlikely to be determined without look-ahead bias. Since each methodology has its own expertise in the specific "golden environment", the overall forecast can be more robust to keep a battery of models and take a collective signal from them.

The availability of intraday data allows the research to stride into a more flexible forecast window. Hourly EURUSD data is employed in this study. To handle the inability of recursive econometric models in dealing with high frequency data, this paper proposes a double-step approach to separate the underlying volatility process from intraday/intraweek seasonality in FX market as well as the short-term exposure to the pre-scheduled macroeconomic announcement. By applying this procedure, the predictability of all empirical models was substantially enhanced.

To access model performance further than evaluating forecast errors, this paper constructs a 2-week volatility swap based on the deviation between forecast and swap strike. The profit is significant even after considering transaction costs. Especially, the model combination constantly outperforms any single model, and is able to seize the upside from each of them.

However, using volatility swaps might not be the best strategy to exploit the full power of the model since the forecast into intraday volatility shows an even more precise outcome. In fact, predictive power alone varies with frequency and definition of the realized volatility, thus there is no given answer whether one model does or does not work; it is the final application and trading instrument altogether that decide on the best applicable model. More sophisticated and higher frequency trading strategies can be designed for the results from this paper, which is left to further research.

A Appendix: Steady-state Probabilities for Markov Chain

This appendix demonstrates the approach to get the unconditional, or steady-state probabilities for a first-order Markov Chain, which is usually used as the initial state probabilities in the recursive process. It is listed here for the purpose of completeness. For more details please refer to Marcucci (2001).

Denote the transition matrix for N states as follows:

$$P = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{12} & p_{22} & \cdots & p_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1N} & p_{2N} & \cdots & p_{NN} \end{pmatrix}$$

where p_{ij} is the probability that the state j will occur conditional on the current state i. Since states are mutually exclusive and collectively exhaustive, every column of the transition matrix sums up to 1, i.e.,

$$\sum_{j=1}^{N} p_{ij} = 1$$

Assume π_t is the $N \times 1$ vector of steady-state probabilities

$$\pi_t = \begin{bmatrix} \Pr(s_t = 1) \\ \Pr(s_t = 2) \\ \vdots \\ \Pr(s_t = N) \end{bmatrix} = \begin{bmatrix} \pi_{1t} \\ \pi_{2t} \\ \vdots \\ \pi_{Nt} \end{bmatrix}$$

The first condition on the steady-state probabilities is they add up to 1. Let i_N be a $1 \times N$ unit vector, hence

$$i_N \cdot \pi_t = 1$$

The second condition comes from its steady feature, which means the state probabilities will remain the same once it is reached:

$$P \cdot \pi_t = \pi_{t+1} = \pi_t$$

which is equivalent to

$$(I_N - P) \cdot \pi_t = 0_N$$

From above two conditions we can get

$$A \cdot \pi_t = \begin{bmatrix} 0_N \\ 1 \end{bmatrix}$$

where

$$A = \begin{bmatrix} I_N - P \\ i_N \end{bmatrix}$$

Therefore

$$\pi_t = (A'A)^{-1} A' \begin{bmatrix} 0_N \\ 1 \end{bmatrix}$$

References

[1] Andersen, T., and T. Bollerslev, 1997a, Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-run in High Frequency Returns, Journal of Finance, 52, 975-1005.

[2] Andersen, T., and T. Bollerslev, 1997b, Intraday Periodicity and Volatility Persistence in Financial Markets, Journal of Empirical Finance, 4, 115-158.

[3] Andersen, T., and T. Bollerslev, 1998a, Answering the Skeptics: Yes, Standard Volatility Models do provide accurate forecasts, International Economic Review, 39, 885-905.

[4] Andersen, T., and T. Bollerslev, 1998b, Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies, Journal of Finance, 53, 219-265.

[5] Andersen, T. G., T. Bollerslev, and F. X. Diebold, 2009, Parametric and Nonparametric Measurement of Volatility, In Y. Ait-Sahalia and L. P. Hansen (Eds.), Handbook of Financial Econometrics, Amsterdam: North Holland. Forthcoming.

[6] Andersen, T., and T. Bollerslev, F. X. Diebold and H. Ebens, 2001, The distribution of Realized Stock Return Volatility, Journal of Financial Economics, 61, 43-76.

[7] Andersen, T., and T. Bollerslev, F. X. Diebold and P. Labys, 2003, Modeling and Forecasting Realized Volatility, Econometrica, 71, 579-625.

[8] Baillie, R. T., and T. Bollerslev, 1990, Intraday and Intermarket Volatility in Foreign Exchange Rates, Review of Economic Studies, 58, 565-585.

[9] Bauwens, L, B. Omrane and P. Giot, 2005, News Announcements, Market Activity and Volatility in the Euro/Dollar Foreign Exchange Market, Journal of International Money and Finance, 24, 1108—1125.

[10] Barndorff-Nielsen, O. E. and N. Shephard, 2002, Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models, Journal of Royal Statistics, Series B (Statistical Methodology), 64, 253-280.

[11] Barndorff-Nielsen, O. E. and N. Shephard, 2007, Variation, Jumps and High Frequency Data in Financial Econometrics. In R. Blundell, T. Persson and W. K. Newey (Eds.), Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress. Econometric Society Monographs, 328–372. Cambridge: Cambridge University Press.

[12] Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde and N. Shephard, 2006. Subsampling Realized Kernels, Working paper, Nuffield College, University of Oxford.

[13] Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde and N. Shephard, 2008a, Designing Realized Kernels to Measure the Ex-post Variation of Equity Prices in the Presence of noise, Econometrica, 76, 1481–536.

[14] Beltratti, A., and C. Morana, 1998, Estimating Variance in the Foreign Exchange Market with High Frequency Data, Bocconi University, manuscript.

[15] Bollerslev, T., 1986, Generalized Autoregressive Conditional Heteroskedasticity, Journal of Economics, 31, 307-327.

[16] Bollerslev, T., 1987, A Conditionally Heteroskedastic Time Series Model for Speculative Prices and rates of Return, Review of Economics and Statistics, 69, 542-547.

[17] Bollerslev, T., and J. Wooldridge, 1992, Qusi-maximum Likelihood Estimation and Inference in Dynamic Models with Time-varing Covariances, Econometric Reviews, 11, 143-172.

[18] Britten-Jones, M., A. Neuberger, 2000, Option Prices, Implied Price Processes, and Stochastic Volatility, Journal of Finance, 55, 839-866.

[19] Cai, J., 1994, A Markov Model of Unconditional Variance in ARCH, Journal of Business and Economic Statistics, 12, 309-316.

[20] Cai, J., Y. Cheung, S. K. L. Raymond, and M. Melvin, 2001, 'Once-in-a-generation' Yen Volatility in 1998: Fundamentals, Intervention, and Order Flow, 20, 327-347.

[21] Carr. P. and R. Lee, 2009, Volatility Derivatives, Annual Review of Financial Economics, 1, 319-339.

[22] Carr, P. and D. Madan, 1998, Towards a Theory of Volatility Trading, Chapter 29 of Volatility: New Estimation Techniques for Pricing Derivatives, R. Jarrow (ed.), Risk Book.

[23] Cheung, C. S., P. Miu, Currency Instability: Regime Switching versus Volatility Clustering. 11, 143-172.

[24] Christensen, B. J., and N. R. Prabhala, 1998, The Relation between Implied and Realized Volatility, Journal of Financial Economics, 50, 125–150.

[25] Cleveland, W., 1979, Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, 74, 829—836.

[26] Dacorogna, M. M., R. Gencay, U.A., Muller, R. B., Olsen and O. V., Pictet, 2001, An Introduction to High-frequency Finance. San Diego: Academic Press.

[27] Day, T., and C. Lewis, 1992, Stock Market Volatility and the Information Content of Stock Index Options, Journal of Econometrics, 52, 267–287.

[28] Day, T., and C. Lewis, 1993, Forecasting Futures Market Volatility, Journal of Derivatives, 1, 33–50.

[29] Diebold, F. X., 1986, Modeling the Persistence of Conditional Variances: A Comment, Econometric Reviews, 5, 51-56.

[30] Dominguez, K.M.E. and F. Panthaki, 2006, What Defines 'News' in Foreign Exchange Markets? Journal of International Money and Finance, 25, 168-198.

[31] Gencay, R., F., Selcuk, and B. Whitcher, 2001a, Differentiating Intraday Seasonalities through Wavelet multiscaling, Physica A, 289, 543-556.

[32] Glosten, L. R., R. Jagannathan, and D. Runkel, 1993, Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, Journal of Finance, 48, 1779-1801.

[33] Haas, M., S. Mittnik and M.S. Paloella, 2004, A New Approach to Markov-Switching GARCH Models, Journal of Financial Economietrics, 2(4), 493-550.

[34] Hamilton, J. D., 1990, Analysis of Time Series Subject to Changes in Regime, Journal of Econometrics, 45, 39-70.

[35] Hamilton, J. D., and R. Susmel, 1994, Autoregressive Conditional Heteroskedasticity and Changes in Regime, Journal of Econometrics, 64, 307-333.

[36] Jiang, G. J., Y. S. Tian, 2005, The Model-free Implied Volatility and Its Information Content, Review of Financial Studies, 18, 1305-1342.

[37] Jorion, P., 1995, Predicting Volatility in the Foreign Exchange Market, Journal of Finance, 50,507-528.

[38] Laakkonen, H., 2007, Exchange Rate Volatility, Macro Announcements and the Choice of Intraday Seasonality Filtering Method, Bank of Finland Research, Discuss papers 23/3007.

[39] Lamoureux, C., and W. Lastrapes, 1990, Heteroskedasticity in Stock Return Data: Volume versus GARCH Effects, Journal of Finance, 45, 221-229.

[40] Lanne, M., and P. Saikkonen, 2003, Modeling the U.S. Short Term Interest Rate by Mixture of Autoregressive Processes, Journal of Financial Econometrics, 1, 96-125.

[41] Nelson, D.B., 1991, Conditional Heteroskedasticity in Asset Returns: A New Approach, Econometrica, 59, 347-370.

[42] Neuberger, A., 1990, Volatility Trading, London Business School Working Paper.

[43] Omrane, W, E., Bodt 2005, Using Self-organizing Maps to Adjust Intra-day Seasonality, Louvain School of Management Working paper 149/05.

[44] Perron, P., 1989, The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis, Econometrica, 57, 1361-1401.

[45] Taylor, S. J. and X. Xu, 1997, The Incremental Volatility Information in One Million Foreign Exchange Quotations, Journal of Empirical Finance, 4, 317-340.

[46] Wong, C., and W. Li, 2001, On a Mixture Autoregressive Conditional Heteroscedastic Model, Journal of the American Statistical Association, 96, 982-995.

Table 1 General Statistics for Hourly Return and Variance

This table shows the general statistics (mean, median, standard deviation, skewness, kurtosis, minimum and maximum) of hourly return and variance of EURUSD for the whole period (2004-2010) and the sub-sample period (2007-2010) when forecasts are carried out. Here $r_t = (\log(P_t) - \log(P_{t-1})) \times 100$; $\sigma_t^2 = r_t^2$

	Whole Sample (2004-2010)		Forecast Sample (2007-2010)	
# of Obs.	42,903		24,138	
	r_t	σ_t^2	r_t	σ_t^2
Mean	2.0817e-004	0.0184	1.7412e-004	0.0225
Median	0	0.0031	0	0.0037
Std. Dev.	0.1358	0.0735	0.1500	0.0876
Skewness	0.1570	26.4741	0.1775	24.9191
Kurtosis	16.9084	1.3305e+003	16.1789	1.1092e+003
Min	-1.9172	0	-1.9172	0
Max	2.3061	5.3179	2.3061	5.3179

Table 2 Macroeconomic Data Releases

This table shows pre-scheduled macroeconomic data releases that are related to USD and EUR market. Bloomberg tickers, event names and release frequency are listed.

Bloomberg Ticker	Corresponding Event	Release Schedule
CPUPXCHG	US Core CPI	Monthly
USURTOT	US Unemployment Rate	Monthly
NFP TCH	US NFP	Monthly
FDTR	FOMC Meeting (Fed Rate)	Monthly
NAPMPMI	US ISM Manufacturing	Monthly
RSTAMOM	US Retail Sales (with Autos)[1]	Monthly
USGDP CQOQ	US GDP Quarterly	Quarterly/Monthly
GRZEWI	German ZEW Survey	Monthly
GRIFPBUS	IFO Germany Business Climate	Monthly
ECCPEMUY	Euro HICP (inflation indicator)	Monthly
PMITMEZ	Euro PMI Manufacturing	Monthly
UMRTEMU	EU Unemployment Rate	Monthly

[1] Retail Sales without Auto is also released at the same time.

Table 3 Statistics of Delta-neutral IV and Model-free IV from Oct 2008-Dec 2010

This table compares the delta-neutral implied volatility (DNIV) and model-free implied volatility (MFIV) for 2-week tenor EURUSD options between Oct 2008 and Dec 2010. MFIV is calculated by following the method proposed by Jiang and Tian (2005).

MFIV= $\sqrt{2\int_0^{\infty}\frac{C(T,F(K))-\max(S_0-K,0)}{K^2}dK}$

	Mean	Std. Dev.	Sample size
DNIV	0.1401	0.0441	543
MFIV	0.1432	0.0474	543
DNIV-MFIV: t stats	1.1323		

Table 4 Event Effect (E_i)

This table shows the event multiplier of each macroeconomic data release on hourly variance, E_i. which is calculated as the ratio of the average hourly variance (net of seasonality) when event i happens, and the average hourly variance (net of seasonality) when there is no event. The estimation period is anchored from the beginning of 2004.

US Event	Effect	EU Event	Effect
US Core CPI	1.43	German ZEW Survey	1.65
US Unemployment Rate US NFP	7.00	IFO Germany Business Climate Euro HICP (inflation indicator)	2.86 1.63
FOMC Meeting (Fed Rate)	15.25	Euro PMI Manufacturing	1.90
US ISM Manufacturing	1.50	EU Unemployment Rate	0.98
US Retail Sales (with Autos)	1.43		
US GDP Quarterly	1.97		

Table 5 Squared Error and Absolute Error (RV from Daily Data)

This table shows the statistics of the squared error and absolute error for each volatility forecast model (GARCH, GJR, EWMA, MA and Delta-neutral implied volatility) from the 1st, January 2007 to the 17th, December 2010. Forecast is conducted every day into the next 2 weeks, and on a rolling basis. GARCH, GJR and EWMA are regime-switch models that are estimated by the most recent 20-week data. Moving Average (MA) is based on the latest 2-week data. Delta-neutral implied volatility (DNIV) is backed out from 2-week EURUSD options. The realized volatility is calculated as the square root of annualized daily average return square. The last column Mincer-Zarnowits R^2 reports the results of regressing realized volatility on each forecasted volatility.

Total # of out-of-sample forecast windows: 1021 (01.01/2007-17.12.2010)					
Model	SE_mean *10000	SE_std *10000	AE_mean *100	AE_std *100	Mincer-Zarnowitz R^2
GARCH	10.871	24.020	2.296	2.368	92.28%
GJR	10.631	22.635	2.274	2.338	92.39%
EWMA	10.655	26.021	2.240	2.376	92.18%
MA	10.250	24.440	2.167	2.358	92.51%
DNIV	12.204	26.896	2.526	2.415	92.44%

Table 6 Model Combination and Model Selection

This table shows the statistics of the squared error and absolute error for "signal combination" and "signal selection" from the 1st, January 2007 to the 17th, December 2010. Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. In each forecast, a set of models: GARCH, GJR, EWMA, MA and DNIV are kept on the battery. [C1] sets weight to each model as the inversed squared error of the last period; [C2] sets weight as the inversed average squared error; [C3] sets equal weights; [S1] only takes the best model for the last period; [S2] only takes the most accurate model on average so far. The realized volatility is calculated as the square root of annualized daily return square average. The last column Mincer-Zarnowits R^2 reports the results of regressing realized volatility on each forecasted volatility.

Total # of out-of-sample forecast windows: 1021 (01.01/2007-17.12.2010)					
	SE_mean *10000	SE_std *10000	AE_mean *100	AE_std *100	Mincer-Zarnowitz R^2
[C1] w=1/SE(t-1)	10.441	23.394	2.252	2.318	92.59%
[C2] w=1/MSE	10.092	22.996	2.201	2.292	92.83%
[C3] Equally weighted	10.086	23.061	2.202	2.290	92.86%
[S1] Model=min SE(t-1)	11.012	25.335	2.303	2.390	92.09%
[S2] Model=min MSE	11.131	25.062	2.268	2.448	92.24%

Table 7 Annual Performance, Average Win and Loss

This table shows performance (annual P&L, standard deviation of P&L and the Sharpe ratio) of a volatility swap trading strategy for each forecast model. Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. Every day, if the volatility forecast is 10% higher (lower) than the MID of 2-week volatility strike, a long (short) position (each with 1 vega notional) will be entered in the 2-week volatility swap, and the trade is unwound after 2 weeks. Trading performance based on each single model (regime-switch GARCH, GJR and EWMA, MA and DNIV), as well as model combinations and selections are reported. For each model combination, [C1] sets weight to each model as the inversed squared error of the last period; [C2] sets weight as the inversed average squared error; [C3] sets equal weights. For each model selection, [S1] only takes the best model for the last period; [S2] only takes the most accurate model on average so far. The last three columns display average win, average loss and the ratio between win and loss for the trading strategy. Performance based on the naïve short strategy (entering a short position every day regardless of model prediction) is listed in the bottom.

Models	Ann. P&L	Ann. Std	Ann. Sharpe	Average Win	Average Loss	Win/Loss
Individual models						
GARCH	22.943	31.511	0.728	2.315	1.830	1.265
GJR	6.063	29.081	0.208	2.036	1.933	1.053
EWMA	15.556	28.379	0.548	2.014	1.926	1.046
MA	16.562	23.004	0.720	1.678	1.759	0.954
DNIV	0	0	n/a	n/a	n/a	n/a
Model Combinations						
[C1] w=1/SE(t-1)	24.311	23.407	1.039	2.128	1.747	1.218
[C2] w=1/MSE	31.585	24.707	1.278	2.294	1.970	1.165
[C3] Equally weighted	35.213	24.156	1.458	2.341	1.931	1.212
Model Selections						
[S1] Model=min SE(t-1)	-1.460	23.001	-0.063	1.886	1.792	1.053
[S2] Model=min MSE	7.096	28.421	0.250	1.998	2.001	0.998
Benchmark						
Naïve Short	-42.050	42.708	-0.985	1.754	1.901	0.923

Table 8 Long/Short Position and Hit Ratio of a Volatility Swap

This table shows the performance of a volatility swap trading strategy: the percentage of the time with positions on, the probability of long and short position, and the hit ratio (possibility of achieving a profit) under long or short position for each forecast model. Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. Every day, if the volatility forecast is 10% higher (lower) than the MID of 2-week volatility strike, a long (short) position (each with 1 vega notional) will be entered in the 2-week volatility swap, and the trade is unwound after 2 weeks. Trading performance based on each single model (regime-switch GARCH, GJR and EWMA, MA and DNIV), as well as model combinations and selections are reported. For each model combination, [C1] sets weight to each model as the inversed squared error of the last period; [C2] sets weight as the inversed average squared error; [C3] sets equal weights. For each model selection, [S1] only takes the best model for the last period; [S2] only takes the most accurate model on average so far. Hit ratio is calculated as the ratio of the number of trades that make profit and the total number of trades. Performance based on the naïve short strategy (entering a short position every day regardless of model prediction) is listed in the bottom.

Models	Position %	Long %	Short %	Hit Ratio		
				Long	Short	Overall
Individual models						
GARCH	46.52%	36.16%	63.84%	32.81%	57.96%	48.87%
GJR	46.78%	36.52%	63.48%	31.54%	60.62%	50.00%
EWMA	45.99%	29.71%	70.29%	30.77%	61.38%	52.29%
MA	38.63%	24.49%	75.51%	37.50%	62.16%	56.12%
DNIV	0	0	0	0	0	0
Model Combinations						
[C1] w=1/SE(t-1)	30.62%	26.18%	73.82%	36.07%	59.30%	53.22%
[C2] w=1/MSE	28.91%	30.00%	70.00%	33.33%	66.23%	56.36%
[C3] Equally weighted	27.07%	32.04%	67.96%	33.33%	68.57%	57.28%
Model selections						
[S1] Model=min SE(t-1)	33.77%	24.12%	75.88%	29.03%	54.36%	48.25%
[S2] Model=min MSE	43.50%	33.84%	66.16%	32.14%	61.64%	51.66%
Benchmark						
Naïve Short	100.00%	0.00%	100.00%	n/a	47.44%	47.44%

Table 9 Maximum Drawdowns

This table shows the performance of a volatility swap trading strategy: maximum drawdown magnitude and drawdown period for each forecast model. Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. Every day, if the volatility forecast is 10% higher (lower) than the MID of 2-week volatility strike, a long (short) position (each with 1 vega notional) will be entered in the 2-week volatility swap, and the trade is unwound after 2 weeks. Trading performance based on each single model (regime-switch GARCH, GJR and EWMA, MA and DNIV), as well as model combinations and selections are reported. For each model combination, [C1] sets weight to each model as the inversed squared error of the last period; [C2] sets weight as the inversed average squared error; [C3] sets equal weights. For each model selection, [S1] only takes the best model for the last period; [S2] only takes the most accurate model on average so far. Maximum drawdown is measured as the largest peak-to-trough decline of the P&L. The peak is the starting time of drawdown, and the drawdown recovery time is when the P&L returns to the previous peak level. Performance based on the naïve short strategy (entering a short position every day regardless of model prediction) is listed in the bottom.

Models	Max. Drawdown	Drawdown starts	Max drawdown time	Drawdown recovers	Duration (days)
Individual models					
GARCH	72.4542	02-Jan-07	10-Mar-08	10-Oct-08	109
GJR	70.2949	02-Jan-07	08-Nov-07	07-Oct-08	191
EWMA	64.3326	02-Jan-07	10-Mar-08	10-Oct-08	109
MA	42.6238	04-Jan-08	26-Mar-08	06-Oct-08	104
DNIV	n/a	n/a	n/a	n/a	n/a
Model Combinations					
[C1] w=1/SE(t-1)	30.7636	20-Apr-09	22-Jul-09	18-Feb-10	122
[C2] w=1/MSE	35.7158	20-Apr-09	13-Aug-09	11-Feb-10	101
[C3] Equally weighted	37.845	20-Apr-09	13-Aug-09	17-Feb-10	105
Model selections					
[S1] Model=min SE(t-1)	56.8758	02-Jan-07	12-Sep-08	02-Mar-10	246
[S2] Model=min MSE	63.3989	22-Jan-07	24-Jul-08	10-Oct-08	43
Benchmark					
Naïve Short	262.4915	11-Jan-07	22-Jan-09	n/a	n/a

Table 10 Annual Performance of Volatility Swap of Model Combinations

This table shows annual P&L, standard deviation, and the Sharpe ratio of the trading performance if replacing DNIV (delta-neutral implied volatility) with the volatility swap strike in model combination. Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. Every day, if the volatility forecast is 10% higher (lower) than the MID of 2-week volatility strike, a long (short) position (each with 1 vega notional) will be entered in the 2-week volatility swap, and the trade is unwound after 2 weeks. The models that come into combined signals include: regime-switch GARCH, GJR and EWMA, MA and the volatility swap strike. [C1] sets weight to each model as the inversed squared error of the last period; [C2] sets weight as the inversed average squared error; [C3] sets equal weights.

Models	Ann. P&L	Ann. Std	Ann. Sharpe
Model Combinations			
[C1] w=1/SE(t-1)	31.889	22.168	1.439
[C2] w=1/MSE	38.966	24.442	1.594
[C3] Equally weighted	36.886	23.492	1.570

Table 11 Sharpe Ratio Comparison

This table compares annual Sharpe Ratios between unit position and adjusted position; and between using DNIV (delta-neutral implied volatility) and the volatility swap strike in model combination. Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. Every day, if the volatility forecast is 10% higher (lower) than the MID of 2-week volatility strike, a long (short) position will be entered in the 2-week volatility swap, and the trade is unwound after 2 weeks. In the case of unit position, each volatility swap is assigned with 1 vega notional; while in the case of adjusted position, the more deviated the forecast moves from volatility swap, the larger vega notional is allocated in the position. i.e., $Pos_i = \pm 1 * (1 + abs(For_i - Strike)/Strike)$. Besides DNIV or swap strike, other models in model combinations include regime-switch GARCH, GJR and EWMA, and MA.

Models	Adjusted Position	Unit Position	Adjusted Position	Unit Position
Individual Models				
GARCH	0.838	0.728		
GJR	0.356	0.208		
EWMA	0.649	0.548		
MA	0.765	0.720		
Model Combinations				
	with DNIV		with Swap Strike	
[C1] w=1/SE(t-1)	1.088	1.039	1.504	1.439
[C2] w=1/MSE	1.334	1.278	1.648	1.594
[C3] Equally weighted	1.505	1.458	1.626	1.570
Model Selections				
	with DNIV		with Swap Strike	
[S1] Model=min SE(t-1)	-0.002	-0.063	0.621	0.539
[S2] Model=min MSE	0.366	0.250	0.648	0.532

Table 12 Model Performance from Daily Data

This table shows the robustness check result when applying daily data to forecast 2-week volatility and trading based on it. Squared errors and absolute errors for each single model and model combinations are reported, where the realized volatility is calculated as the square root of annualized daily average return square. Annual Sharpe ratios for trading volatility swap accordingly are listed in the last column.

Models	SE_mean *10000	SE_std *10000	AE_mean *100	AE_std *100	Sharpe Ratio
Individual models					
GARCH	12.253	28.967	2.414	2.536	0.137
GJR	11.558	29.716	2.335	2.473	0.116
EWMA	12.297	31.477	2.375	2.581	0.062
MA	16.317	37.909	2.764	2.947	-1.680
Model Combinations					
[C1] w=1/SE(t-1)	11.563	27.169	2.354	2.455	-0.515
[C2] w=1/MSE	10.553	24.677	2.254	2.341	0.852
[C3] Equally weighted	10.920	26.625	2.269	2.404	0.813

Table 13 Model Performance without Seasonality or Event Adjustments

This table shows the robustness check result when applying high frequency data to forecast 2-week volatility and trading based on it without the seasonality or event adjustments. Squared errors and absolute errors for each single model are reported, where the realized volatility is calculated as the square root of annualized daily average return square. Annual Sharpe ratios for trading volatility swap accordingly are listed in the last column.

Model	SE_mean *10000	SE_std *10000	AE_mean *100	AE_std *100	Sharpe Ratio
GARCH	14.577	33.437	2.642	2.758	-2.455
GJR	14.053	32.459	2.563	2.737	-2.378
EWMA	11.375	26.613	2.314	2.455	0.702
MA	11.060	26.717	2.269	2.433	-1.670

Table 14 Performance of Lagged Realized Volatility and Its Moving Averages

This table shows the robustness check result when applying a simple lagged moving average of realized volatility to forecast 2-week volatility and trading based on it. Squared errors and absolute errors for each simple moving average are reported, where the realized volatility is calculated as the square root of annualized daily average return square. Annual Sharpe ratios for trading volatility swap accordingly are listed in the last column.

Model	SE_mean *10000	SE_std *10000	AE_mean *100	AE_std *100	Sharpe Ratio
Lagged RV	17.425	45.174	2.795	3.102	-1.980
Lag RV MA=3 (3d)	16.453	42.387	2.682	3.045	-1.449
Lag RV MA=5 (1w)	15.443	39.052	2.598	2.950	-0.956
Lag RV MA=10 (2w)	13.070	31.403	2.443	2.667	-0.463
Lag RV MA=20 (4w)	11.400	27.567	2.283	2.489	-0.186

Table 15 Performance of fixed-parameter EWMA

This table shows the robustness check result when applying RiskMetrics fixed EWMA (λ=0.97) to forecast 2-week volatility and trading based on it. Squared errors and absolute errors are reported, where the realized volatility is calculated as the square root of annualized daily average return square. The annual Sharpe ratio for trading volatility swap accordingly is listed in the last column.

Model	SE_mean *10000	SE_std *10000	AE_mean *100	AE_std *100	Sharpe Ratio
Fixed EWMA	14.310	36.141	2.607	2.742	-3.041

Table 16 Squared Error and Absolute Error (RV from Hourly Data)

This table shows the robustness check result when calculating realized volatility as the square root of annualized hourly average return square. Statistics of squared errors and absolute errors are reported.

Model	SE_mean*10000	SE_std *10000	AE_mean *100	AE_std *100
GARCH	6.555	13.712	1.781	1.840
GJR	6.375	13.115	1.778	1.794
EWMA	5.798	12.286	1.703	1.703
MA	4.536	8.547	1.556	1.455
DNIV	4.850	9.601	1.634	1.477

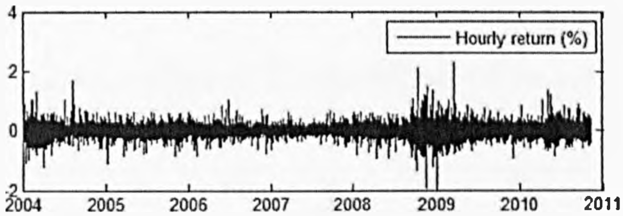
Table 17 Squared Error and Absolute Error (RV from Hourly Data) without Seasonality or Event Adjustments

This table shows the robustness check result when calculating realized volatility as the square root of annualized hourly average return square, and applying high frequency data directly without the seasonality or event adjustments. Squared errors and absolute errors are reported.

Model	SE_mean*10000	SE_std *10000	AE_mean *100	AE_std *100
GARCH	9.185	20.061	2.129	2.158
GJR	8.339	17.839	2.013	2.072
EWMA	6.816	15.415	1.809	1.883
MA	4.997	10.227	1.619	1.542
DNIV	4.850	9.601	1.634	1.477

Figure 1 Hourly Return and Variance

This figure shows historical hourly returns and hourly variance of EURUSD between 2004 and 2011. Hourly returns are calculated as log difference of mid price, which, is the geometric average of bid and ask price. Hourly variance is calculated as the hourly return square.



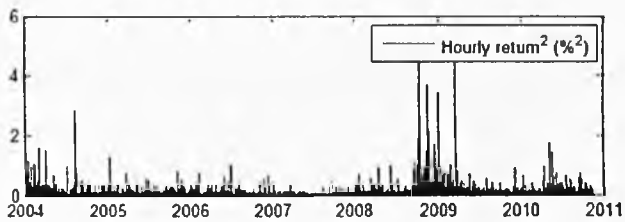


Figure 2 Model-free and Delta-neutral Implied Volatility (2w window, annualized)

This figure shows the delta-neutral implied volatility (DNIV) and model-free implied volatility (MFIV) for 2-week tenor EURUSD options. DNIV is catching a single point (delta-neutral strike, which is approximately ATM) on the implied volatility surface, while MFIV is calculated by following the method proposed by Jiang and Tian (2005).

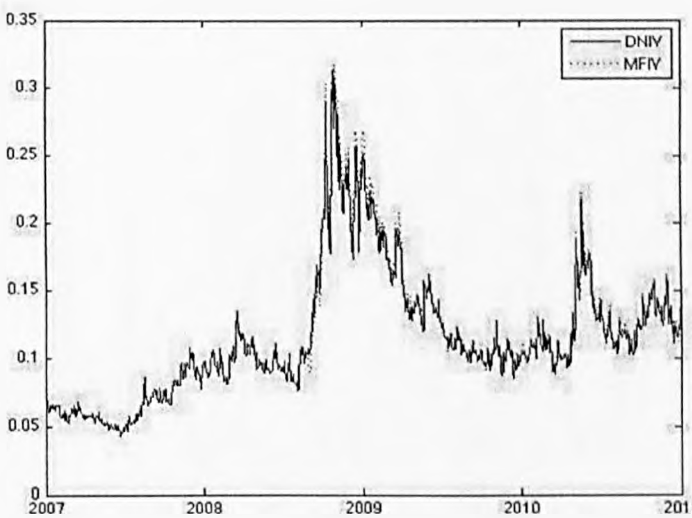


Figure 3 Comparison of Variance Autocorrelation before and after Adjustment

This figure shows autocorrelation of hourly variance before and after the seasonality adjustment.

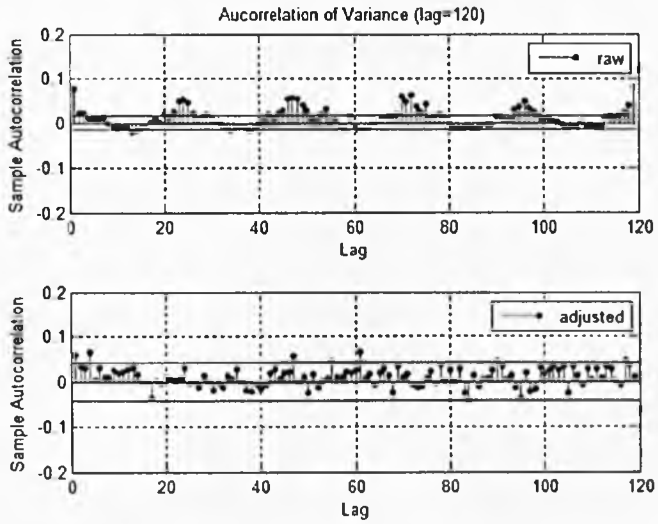


Figure 4 Intraweek Seasonality (S_h)

This figure shows the seasonality adjustment multiplier, S_h, which is calculated as the average hourly variance of hour-of-the-week h, divided by overall hourly variance of a week.

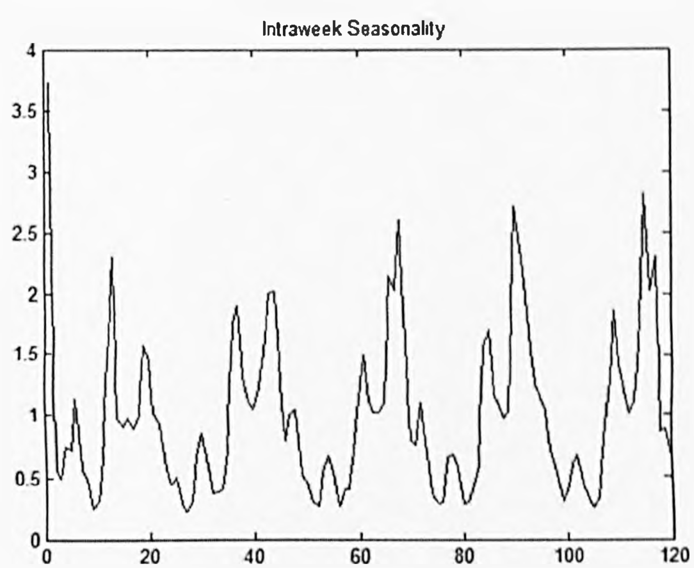


Figure 5 Annualized 2-Week Realized Volatility (Daily Frequency)

This figure shows EURUSD realized volatility of the 2-week tenor, calculated as the square root of the annualized average daily return square.

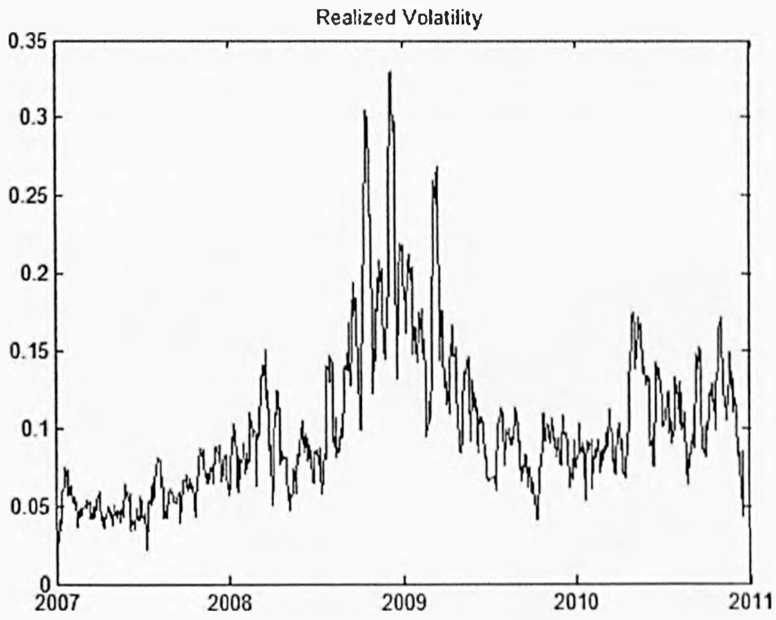


Figure 6 Cumulative P&L of Single Models and Naïve Short

This figure shows the cumulative P&L of trading volatility swap based on single models (regime-switching GARCH, GJR and EWMA; and MA). Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. Every day, if the volatility forecast is 10% higher (lower) than the MID of 2-week volatility strike, a long (short) position (each with 1 vega notional) will be entered in the 2-week volatility swap, and the trade is unwound after 2 weeks. The single trade's P&L is calculated as the difference of annualized realized volatility and volatility swap strike (ask price if enter buy, and bid price if enter sell), multiply the position indication (1 if buy, -1 if sell) and 100. The cumulative P&L is added up from the beginning of 2007. As a contrast, a naïve short strategy enters a sell position every day regardless what the forecasting model predicts.

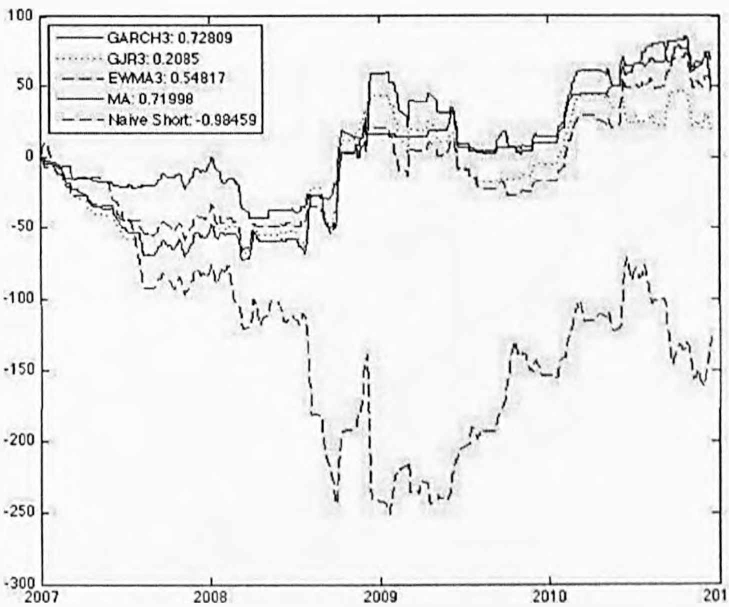


Figure 7 Cumulative P&L of Model Combinations

This figure shows the cumulative P&L of trading volatility swap based on model combinations. Forecasts are conducted every day into the next 2 weeks, and on a rolling basis. The candidate forecasting models include: regime-switch GARCH, GJR, and EWMA; MA; DNIV. After obtaining individual forecast figure, the numbers are aggregated in 3 alternative ways: [C1] sets weight to each model as the inversed squared error of the last period; [C2] sets weight as the inversed average squared error; [C3] sets equal weights. Every day, after the collective signal is generated, if the number is 10% higher (lower) than the MID of 2-week volatility strike, a long (short) position (each with 1 vega notional) will be entered in the 2-week volatility swap, and the trade is unwound after 2 weeks. The single trade's P&L is calculated as the difference of the annualized realized volatility and volatility swap strike (ask price if enter buy, and bid price if enter sell), multiply the position indication (1 if buy, -1 if sell) and 100. The cumulative P&L is added up from the beginning of 2007.

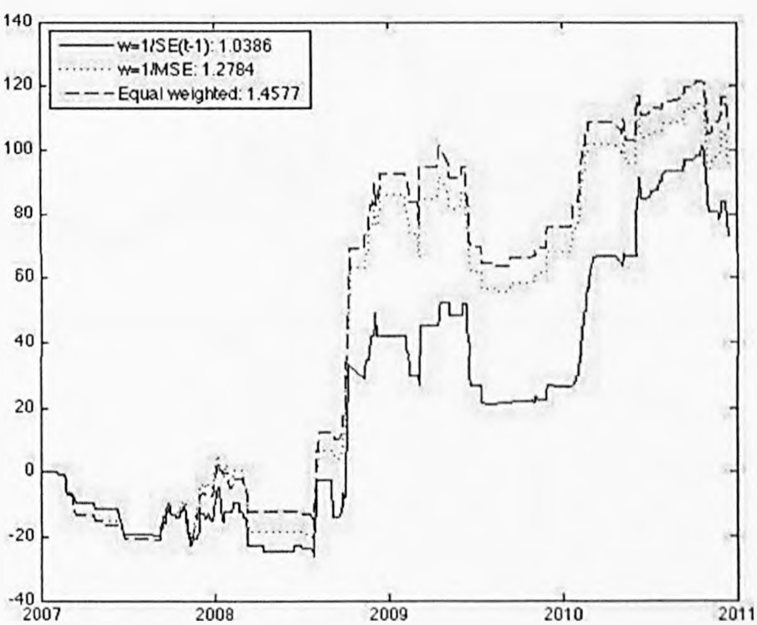
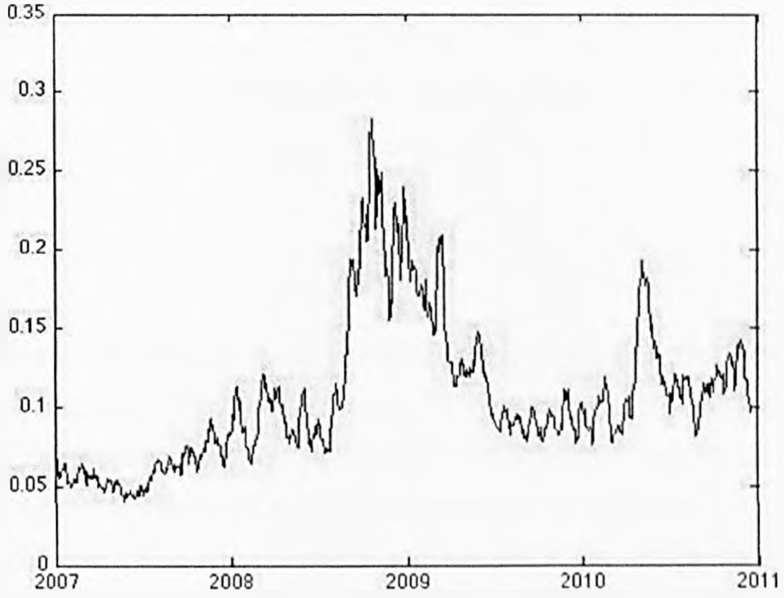


Figure 8 Annualized 2-Week Realized Volatility (Hourly Frequency)

This figure shows EURUSD realized volatility of the 2-week tenor, calculated as the square root of the annualized average hourly return square.



Extract More from The Volatility Surface

Yiyi Wang

London School of Economics

November 27, 2011

Abstract

This paper explores the cross-sectional predictive power of the most important two factors in the implied volatility surface - skew and term structure - at individual firm level. Stocks with lower implied volatility skew and higher implied volatility term structure outperform the comparative peers. In particular, skew represents a directional signal about what has not been captured by the stock price, while term structure reflects upcoming events without indicating the sign (positive or negative) of it. Therefore, the interaction between these two factors reinforces their predictive power, and the annual return of a weekly long-short trading strategy can be enhanced from 14.87% to 20.90% with the attachment of term structure on skew. By linking the factors to information asymmetry proxies, this paper highlights the internal drivers of such forecasting ability. In the end, this paper looks at a subsample of companies that witnessed takeover announcements during the period. The result confirms the options market's leading position in price discovery. By sorting firms based on skew and term structure one may also be able to pick up takeover targets and seize the big positive premium.

Keywords: implied volatility surface, skew, term structure, return prediction, information-based trading.

1 Introduction

The options market is well known for its leading role in price discovery. A large amount of existing research (Easley, O'Hara and Srinivas 1998, Chan, Chung and Fong, 2002) look into the lead-lag relationship between option prices and the underlying stock market, which sheds light on how the information transfers between the two markets and how the trading behaviour links to each other. As some of the literature documented (Cao, Chen and Griffin, 2003), options, especially short-term OTM (out-of-the-money) options are the optimal instruments for the informed investors due to the high leverage. If the cash market is quick enough to capture the information inferred from options market, the additional message will soon get reflected in stock price and the opportunity for profits will vanish. However, as pointed out by Xing, Zhang and Zhao (2010), the underlying market is slow in incorporating what is embedded in option prices, which makes it possible to construct a stock portfolio to exploit the hidden profit.

Instead of looking at option prices directly, this paper chooses to study alternative signals from the implied volatility surface. Implied volatility is the parameter σ that inserts into the Black-Sholes model so that the plain vanilla option price from the formula matches the market price of the option. According to the Black-Sholes assumption, asset returns follow a log normal distribution and the standard deviation (σ) is a constant across various strike prices and time to maturity. However, the implied volatility backed out from traded option prices clearly contradicts this assumption. In fact, it is widely known that the implied volatility demonstrates non-flat instantaneous "volatility smile" (often observed in FX and equity index) or "skew/smirk" (often observed in single equity) for a given tenor; and on the other dimension, implied volatility for different maturities also exhibits a similar pattern. Short-term volatility is usually more sensitive to the contemporaneous market condition change and the upcoming events, thus present a higher "volatility of volatility". Overall, while mapping the implied volatility to different strike levels and tenors, the 3-D plot (so-called implied volatility surface) shows the whole picture of the market expectation for the future volatility. If the information is absorbed quicker in the options market than the underlying stock market, what is concealed in the surface can be applied to predict stock market returns.

According to some principal component analysis (PCA) studies (Cont and Fonseca 2002, Roux 2006), among all the factors, the most important three components of the total variation in volatility surfaces include level, term structure and skew. Among these three, it is most straightforward to understand the relationship between return and level mode. Early studies (Giot 2003, Dennis, Mayhew and Stivers, 2006) showed implied volatility level has strong negative contemporaneous relationship with spot return. Due to the nature of equity market and the sensitivity of supply-demand forces to market sentiment, large jumps in spot tend to be downwards rather than upwards. This asymmetric pattern is also one of the conventional explanations for the existence of implied volatility skew.

Volatility skew, or smirk, presents the phenomenon that OTM put is more expensive than OTM call. Therefore, if taking a snapshot of a volatility surface for a given time to maturity, one can observe a downward-sloping implied volatility curve. To understand the predictability of implied volatility skew, it is helpful to review the other reasons for its presence. Firstly, the underlying asset distribution presents a less heavy right tail and a heavier left tail. In other words, the possibility and magnitude of negative extreme value are both much higher than the positive comparatives. Hence, a steeper skew usually reflects a deteriorate expectation for the future performance. Secondly, the price tends to be more volatile at lower level, which has been widely spotted during each financial crisis. Thirdly, the OTM puts are usually purchased as a hedge against the long position in underlying stock. Since the normal market is net long stock, the demand for puts also pushes the price higher. In this case, if the investors become more confident in the future stock performance, they will either decrease the hedging position in put, or simply turn to long in call, either of which will make the implied volatility skew decrease (become flatter). Therefore, the information inferred from skew is "directional".

Xing, Zhang and Zhao (2010) did find the shape of volatility smirk exhibit pronounced predictability for future stock returns, since the flatter smirk is associated with a better earnings surprise, and vice versa. Their research enlightened an important way to extract useful indications from implied volatility surface, though they confined the study at the level of volatility skew, while other important aspects of the surface remained untouched. This paper makes the contribution in filling the gap of exploring the cross-sectional relationship between term structure and subsequent returns – moreover, the significant information that can be obtained from the interaction between implied volatility skew and term structure, or namely, the whole surface. It shows that when incorporating the term structure and its interaction, a more sophisticated portfolio can be constructed, which leads to a better reward-to-risk profile.

The relationship between term structure and return, however, is not that clear-cut. As the difference of implied volatility between short and long tenors, the term structure tends to be less correlated with changes in fundamental risk in the long run, which leads to a similar shift in options with both short- and long-maturities. Therefore, the volatility term structure is likely to reflect more temporary information, either due to the transitory short-term volatility swing, or because of pending information. The first reason, short-term volatility swing can be viewed as an instantaneous reaction of option market towards market sentiment or stock idiosyncratic risk. Because implied volatility is mean reverting, the implied volatility of shorter maturity options move more volatile than that of longer maturity options (Stein, 1989). If this is the case, the contemporaneous relationship between term structure and market return is expected to display similar pattern as that between volatility and return. On the other hand, if the change in term structure is caused largely by pending information such as earnings announcements or merger and acquisition decisions, it will remain high or increase consistently until the final news comes out and the uncertainty gets resolved. In this case, term structure is more possible to be positive correlated to subsequent return, as a sign of information premium. Overall, it remains an empirical question whether the term structure will reflect more the transitory short-term volatility swing, or pending information.

Even if the volatility term structure does imply pending information, unlike skew, it is "non-directional". However, it serves as an accelerator or confirmer when it interacts with evident skew change. For instance, if the increasing term structure is attached with a decreasing skew, it emphasizes the good information ahead and the chance that it can be realized soon, therefore strengthens the predictive power of skew. Likewise, comparable impact is also expected if the increasing term structure coexists with increasing skew.

This paper starts with a similar approach as Xing, Zhang and Zhao (2010), studying the forecast ability of the information from implied volatility at the single stock level. But different from them, this paper focuses on the change (deviation) of each stock's skew/term structure from its usual level to control the idiosyncratic risk (such as options for some stocks have persistent lower skew, simply because those firms have lower fundamental risk). More importantly, the prediction effect from term structure is analysed in details, especially its interaction with skew. The results show the correlation between skew and subsequent return remains substantially negative and its prediction power lasts as long as 6 month time, as documented by Xing, Zhang and Zhao (2010). On the other side, the term structure displays a positive relationship with future return, thus addressing the statement that the information factor associated with term structures takes on a dominant role in its relationship with returns. However, since the term structure reflects more temporary or undecided/unreleased news which resolves soon, its predictive power does not extend into the future as far as the skew does. Nevertheless, its effect is more enduring when combined with skew. The cross product of these two important elements turns out to be significantly negatively correlated with future returns. Considering the sign of each effect on its own, this result indicates that the signals extracted from skew and term structure do reinforce each other. While applying this idea to portfolio construction, the paper shows that by introducing the information from term structure, the weekly return from a long-short strategy can improve from 28.6 basis points (14.87% annually) to 40.2 basis points (20.90% annually).

By inutition, information acts as the key contributor of the predictive power for both skew and term structure. Then the question is whether this can be shown empirically. If the hypothesis holds, we should be able to observe the stocks with lower skew and higher term structure meanwhile presenting a higher concentration of information. This paper tests on two proxies to measure cross-sectional degree of information asymmetry: probability of informed trading (PIN) and bid-ask spread. The results confirm the strong connection between information and skew/term/cross product. Furthermore, implied volatlity surface factors remain significant after including information proxies in predicting return, hence they are not shadow factors of information proxies that have been studied before.

Following the same idea, this paper carries out an event study on the subsample of takeover targets to investigate how the implied volatility surface evolves before concentrated information release. Merger and acquisitions (M&A) are ideal events to study price discovery in a highly volatile environment driven by information. Because of the considerable takeover premium paid to the target firms during the M&A process, their prices usually demonstrate massive positive jumps on announcement dates. A vast body of literature do research on this topic (Palepu, 1986; Powell, 1997, 2001; Brar, Giamouridis and Liodakis, 2009), aiming at picking up the companies with higher possibility of being taken over by looking at firms' accounting measures. Cao, Chen and Griffin (2005) linked trading behaviour in the options market with takeover targets and showed the higher buy-sell imbalance in options can predict next day return prior to takeover announcements. The analysis of implied surface in this paper is in line with the argument of previous literature that the options market seizes information prior to the event happening. A clear pattern of increasing term structure and decreasing skew can be caught as early as 9-10 months ahead of takeover announcement, which is earlier than the reaction of stock market. Moreover, if following the "long" portfolio constructed as the baseline result, about 9.28% of overall takeover cases can be captured. Even though this paper does not specifically target predicting takeover probability, it provides a potential signal to be employed with the more conventional accounting-based models.

The remainder of the paper is organized as follows. Section 2 describes the data set, key variables and testing hypotheses. Section 3 presents the empirical results from implied volatility surface prediction and portfolio construction. Section 4 investigates the relationship between skew and term structure with information proxies. Section 5 extends the idea to M&A subsample and reports the event study results. Section 6 concludes.

2 Data, Key Variables and Hypotheses

2.1 Data Source

The data used in the paper is from OptionMetrix, which covers the historical option prices, volumes, implied volatility, sensitivity information from January 1996 to December 2010. The data base includes all US listed equities and market indices and all US listed index and equity options. This paper keeps all the valid firm sample as long as the data history is longer than half a year. Besides the main data set, equity returns, trading volume and bid-ask prices are from Centre for Research in Security Prices (CRSP). Fama-French 3 factors together with momentum and long-term reversal factors are downloaded from Professor Kenneth French's website. M&A data is provided by Securities Data Company (SDC) Platinum, and filtered by standard procedure (exclude the bids classified as acquisitions of partial stakes, minority squeeze-outs, buybacks, recapitalizations, and exchange offers; exclude the bids where the acquirer's previous stake exceeding 50%, or a consequent stake less than 50%). In the end, the quarterly PIN measure from 1996 to 2005 is generously delivered by Professor Stephen Brown.

Panel A of Table 1 reports sample size of each year. The sample contains the stocks that have option trading records more than 6 months and have valid link with CRSP return data.

2.2 Key Variables

Skew in intuition is the implied volatility curve for a given expiration date, however, there is no fixed skew measure in past literature. For example, Bates (1991) set skew as arithmetic differences between OTM put volatility and OTM call volatility, based on the percentage moneyness (strike price over spot price); Hull, Nelken and White (2004) defined skew in a similar arithmetic differences, but based on delta. The difference of the above two definitions lies in the convention whether the option price is quoted and traded according to sticky strike, or sticky delta. Alternative skew measures, among others, include Carr and Wu (2007) and Mixon (2009), where skews are normalized versions of the above. In the research closest to this paper, Xing, Zhang and Zhao (2010) adopted the definition similar to Bates (1991), which is implied volatility of OTM put minus that of ATM call, with the moneyness depending on the ratio of strike price to the stock price. Mixon (2010) surveyed these measures and argued that after controlling for volatility and kurtosis, the most robust definition for skew is normalized arithmetic differences of put and call volatilities based on delta. Therefore, this paper applies this measure:

$$SKEW_{i,d} = \frac{VOL_{i,d}^{25\Delta PUT} - VOL_{i,d}^{25\Delta CALL}}{VOL_{i,d}^{50\Delta}} \tag{1}$$

where i is the index for firm and d is the index for date. Since the short-tenor options are the most sensitive to reflect skew shift, when calculating skew only options with maturity=30 days are used. Furthermore, in order to control for the case such as some firms always have lower skew merely because of lower fundamental risk (instead of positive information ahead), for each firm the $SKEW$ is standardized by its mean and standard deviation in the past 6 months[1]. This procedure is also employed by Cao, Chen and Griffini (2003).

The term structure definition follows the same idea as skew. Since now the focus is on differences in maturity, I choose delta-neutral implied volatility for each tenor. The measure is also normalized by past half-year mean and standard deviation.

$$TERM_{i,d} = \frac{VOL_{i,d}^{50\Delta,1M} - VOL_{i,d}^{50\Delta,1Y}}{VOL_{i,d}^{50\Delta,1M}} \tag{2}$$

[1]An alternative way of standardization is only to substract the past 6 months mean from the current value. The main result remains the same.

After computing daily standardized skew and term structure, each measure is averaged to a weekly equivalent (Tuesday close to Tuesday close). The interaction variable of skew and term structure is simply the cross product of both.

$$SKEWTERM_{i,w} = SKEW_{i,w} \cdot TERM_{i,w} \tag{3}$$

Besides the main measures from the surface, this paper also encloses control variables proposed by Xing, Zhang and Zhao (2010). The important control variables for each firm include: *SIZE*, which is the market capitalization; *TURNOVER*, which is the trading volume divided by shares outstanding; last month *VOL* and *SKEWNESS* stand for empirical volatility and skewness of return distribution; *PVOL*, which is difference between short-term delta-neutral implied volatility and last month empirical volatility, represents the risk premium embedded in options market.

In order to link the implied volatility skew, term structure and their cross product with the informativeness of the underlying asset, this paper employs two alternative information proxies. The first one is probability of informed trading (PIN) based on the EKO market microstructure model of information asymmetry (Easley Kiefer and O'Hara, 1997). The EKO model sets out a structure to infer existing informed trading out of the observed order flows. According to the EKO model, PIN can be expressed as a function of number of buy and sell, which is identified by Lee and Ready Algorithm (1991). Previous studies (Easley, Hvidkjaer, and O'Hara, 2002; Vega, 2006; Chen, Goldsteian, and Jiang, 2007; Ferreria and Laux, 2007) have proved that PIN is an effective measure to estimate the information density. This paper uses the quarterly PIN calculated by Professor Stephen Brown, which intersects the main data set from 1996 to 2005.

The other alternative proxy for information asymmetry employed in this paper is bid-ask spread. Theoretically there are two main sources contributing to the spread: liquidity and adverse selection. But while only the latter is driven by information, this paper takes a relatively loose assumption, and argues that if information is the common force of both bid-ask spread and implied volatility surface, at least we should be able find the significant relationship between them.

The statistics of variables are summarized in Table 1.

2.3 Hypotheses

Based on the discussion in the introduction, this paper derives four hypotheses. The first hypothesis tests predictability of implied volatility skew and term structure independently. Since skew is caused by investors' expectation about future negative shock, then the decreasing (flattening) skew maps into a positive future prospect. Term structure, on the other side, predicts subsequent return in terms of information premium.

H1: The implied volatility surface has significant power in forecasting subsequent returns. Returns are likely to rise with decreasing SKEW, and increasing TERM.

The first half of above hypothesis has been tested in Xing, Zhang and Zhao (2010). If SKEW is regarded as a directional bet of future news, then TERM plays more like a timing signal, which indicates an upcoming event without acknowledging its direction. However, under the circumstances when these two factors interact with each other, the effects are expected to be enlarged, such as Hypothesis 2:

H2: The prediction power will be strengthened when SKEW and TERM cooperate.

Considering TERM's timing characteristics, I also test the following hypothesis:

H3: TERM only has predictability that works in the short-term, while SKEW can forecast into future much further.

The last hypothesis looks into the fundamental cause of the prediction power, and explores the interconnection between SKEW, TERM and information proxy.

H4: For firms with lower SKEW and higher TERM, there will be more intensive informed trading of their stocks.

The empirical outcome of these four hypotheses is presented in the next section.

3 Empirical Results

The results from empirical analysis will be represented as follows. In section 3.1, Fama-MacBeth (1973) regression exhibits the main forecasting ability of implied volatility surface, either on its own or after controlling firm variables. Section 3.2 constructs a portfolio based on the rankings of SKEW and TERM, confirming the economic significance of each factor.

3.1 Fama-MacBeth Regression

The baseline Fama-MacBeth regression begins with the following equation, which corresponds to Hypothesis 1.

$$r_{i,w} = \beta_{0,w} + \beta_{1,w} \cdot r_{i,w-1} + \beta_{2,w} \cdot r_{i,m-1} + \beta_{3,w} \cdot r_{i,y-1} + \beta_{4,w} \cdot SKEW_{i,w-1} + \beta_{5,w} \cdot TERM_{i,w-1} \quad (4)$$

where $r_{i,w}$ is the return for firm i in week w. The lagged weekly return for the same firm $r_{i,w-1}$ is included to control for momentum; the previous month's return (excluding the last week) $r_{i,m-1}$ and the previous year's return (excluding the last month) $r_{i,y-1}$ are also enclosed to control for reversal effect. $SKEW_{i,w-1}$ and $TERM_{i,w-1}$ are standardized skew and term structure respectively for week *w-1*. The cross-sectional regression is conducted for each week, and the reported estimates of each parameter are after the Newey-West (1987) adjustment. The calibration and t-statistics for this regression and its univariate version are presented in Table 2 (Column A-C).

As analysed in the introduction, short-term volatility deviations from the long-term mean - prompting a TERM shift - are due to two main reasons: (i) changes in market circumstances (such as instantaneous data release, market sentiment shifts, etc.), or (ii) information foresight. In the market, the first reason can be often detected as a negative contemporaneous correlation between return and TERM; while the second reason takes a form of lead-lag relationship. Therefore, to control for returns at the right side of the regression also separates these two effects, so that TERM performs as a less noisy information carrier.

From Table 2, firms with decreasing SKEW tend to gain higher subsequent return. This phenomenon has been well analysed by Xing, Zhang and Zhao (2010). The coefficient of term structure shows a significantly positive relationship with next week return, though the magnitude and significance is not as large as that of skew. As discussed before, this is because the skew signifies a clear directional surprise, while term structure in itself is non-directional, but only confirms with the presence of undisclosed information.

The next step is to introduce the cross product of two variables. Because the cross product may capture a non-linear relationship between each single variable and the dependent variable, it might easily lead to a spurious significance. In order to control for this, the second moment of SKEW and TERM are taken in the regression[2]:

[2] I also tried controlling for higher moments of each variable, and the results remain unchanged.

$$
\begin{aligned}
r_{i,w} &= \beta_{0,w} + \beta_{1,w} \cdot r_{i,w-1} + \beta_{2,w} \cdot r_{i,m-1} + \beta_{3,w} \cdot r_{i,y-1} + \beta_{4,w} \cdot SKEW_{i,w-1} + \beta_{5,w} \cdot TERM_{i,w-1} \\
&\quad + \beta_{6,w} \cdot SKEW^2_{i,w-1} + \beta_{7,w} \cdot TERM^2_{i,w-1} + \beta_{8,w} \cdot SKEW_{i,w-1} \cdot TERM_{i,w-1} \qquad (5)
\end{aligned}
$$

where the term $SKEW_{i,w-1} \cdot TERM_{i,w-1}$ captures the interaction between them. Since term structure in itself reflects how fast the information (regardless of the sign) can be released and absorbed into the underlying asset price, when we look at the predictive power of the near-future return, high TERM intensifies the effect of SKEW. The significantly negative coefficient of the cross product in Table 2 (Column D) rejects the null hypothesis of H2. The intuition is, when an increasing TERM is attached on a decreasing SKEW, it senses that the hidden positive information can be disclosed soon, hence strengthens the negative relation between SKEW and return; while when a decreasing TERM coexists with a decreasing SKEW, it implies the positive information might be realized later, hence alleviates SKEW's impact in the short term.

To control for other variables at firm level that might also influence return, a more comprehensive regression is carried after adding independent variables. The corresponding results are shown in Table 2 (Column E).

$$
\begin{aligned}
r_{i,w} &= \beta_{0,w} + \beta_{1,w} \cdot r_{i,w-1} + \beta_{2,w} \cdot r_{i,m-1} + \beta_{3,w} \cdot r_{i,y-1} + \beta_{4,w} \cdot SKEW_{i,w-1} \\
&\quad + \beta_{5,w} \cdot TERM_{i,w-1} + \beta_{6,w} \cdot SKEW^2_{i,w-1} + \beta_{7,w} \cdot TERM^2_{i,w-1} \qquad (6) \\
&\quad + \beta_{8,w} \cdot SKEW_{i,w-1} \cdot TERM_{i,w-1} + \beta_{9,w} \cdot CONTROL_{i,w-1}
\end{aligned}
$$

The control variables contain those conventional factors such as SIZE and TURNOVER, and also some other variables that are known to have forecast ability. One is empirical volatility, which is computed as standard deviation of the past month's daily return. Ang, Hodrick, Xiang and Zhang (2006a) found firms with higher idiosyncratic volatility are inclined to have lower future return; Barberis and Huang (2005) and Mitton and Vorkink (2006a) raised the argument that this anomaly was caused by investors' preference for positive skewness. Therefore, skewness of the past month's daily return is also introduced in the regression as a control variable. Furthermore, following Xiang, Zhang and Zhao (2010), I include a volatility premium proxy PVOL, calculated as the difference of 30-day delta-neutral implied volatility and last month's empirical volatility. The idea is that, since all the implied measures are backed out from the B-S model, there is an embedded risk premium in its risk-neutral implied volatility. In order to isolate the prediction power of SKEW/TERM from the effect of market risk preference variation, PVOL is enclosed.

It is noticeable that SKEW becomes even more pronounced after controlling for other factors. However, TERM's prediction in itself becomes weaker and insignificant. This shows the information premium attached on TERM has been overshadowed by other firm-specific variables. Interestingly, the cross product term remains impressively significant. It again confirms the conjecture that non-directional information associated with TERM works through the interaction with directional information embedded in SKEW.

Empirical volatility is negatively correlated with subsequent returns, though not significantly, which matches the previous literature. Empirical skewness, on the other hand, is substantially positively correlated. This result is in line with Xing, Zhang and Zhao (2010), but contradicts earlier studies. Xing et al. (2010) found this positive relationship vanished when extending the return horizon, which can be confirmed with the sample in this paper.

An alternative regression to control other conventional factors is to replace the dependent variable by alpha from Fama-French-Momentum-Reversal five-factor (similar as Fama-French-Carhart (1997) four-factor) model. The results are reported in Table 3 and similar to the baseline regression[3].

In order to test the efficacy of the volatility surface on a longer holding horizon, Table 4 tries out different future returns as dependent variables. Column A lists the regression result for predicting average weekly return of subsequent 4 weeks. Column B to E decompose the next 4 weeks' return into the 1st to the 4th week separately, hence providing a more straightforward way to show how each factor's effect decays with time, and how the information embedded in these factors gets realized in the stock market. As we see from Table 4, skew captures the information which is more long-lasting, slower to be reflected in the underlying stock (even though it does wane along the time); in other words, skew can react much earlier than the information gets completely captured by stock price. Imagine an investor obtains private positive information about a firm without knowing when exactly it will turn out to be public; she would enter into call positions across different expiries. Only when the investor also acknowledges that the good information will be disclosed soon, she will concentrate in buying short-term calls. Therefore, both the term structure and its

[3]Since alpha is computed from the factor model that has controlled for momentum and reversal factors, here the right side of the regression only includes the lagged week's alpha.

interaction with the skew take effect in a shorter horizon, and this information becomes realized in the underlying market much faster. Indeed, although overall the cross product has significant forecast ability on the next 4 week's return, this effect is mostly absorbed in the week coming right after, and reduces to be unremarkable later on. Column G shows that none of the factors calculated 6 weeks ago still sustain a predictive power on the return. Note that this discovery does not necessarily lead to the conclusion that only short horizon strategies are profitable; instead, it is more that a majority of the profit will concentrate in the early weeks.

3.2 Portfolio Trading Strategy

Enlightened by the results from Fama-MacBeth regression, in this subsection I use a portfolio trading strategy to demonstrate the economic benefit of applying information from implied volatility surfaces.

3.2.1 Single Sort

I start by sorting on SKEW. Each week after computing the normalized SKEW factor, the whole universe is ranked and equally divided into 5 groups. Portfolio 1 contains the bottom 20% SKEW while portfolio 5 contains the top 20% SKEW. Portfolio is constructed equally-weighted by its constituents, and rebalanced on a weekly basis. Since firms need half-a-year history as normalized benchmarks, the earliest portfolio construction time is from July 1996, and it lasts till the end of 2010.

Table 5 reports the characteristics of each quintile (Panel A)[4] and the average subsequent

[4]An interesting fact in Panel A is that empirical skewness seems to have a monotonic increasing pattern with implied volatility skew. This is counterintuitive since the skewness measures how the underlying return distribution tilted to the right tail; while implied volatility skew is caused by fat tail in the extreme negative value. Thus in intuition empirical skewness and implied skew should be negative correlated, and have opposite relationship with future returns, which, has been proved by Fama-MacBeth regression (Table 2). One possible explanation is that implied volatility is a forward-looking measure; hence it performs a stronger link with future empirical skewness, rather than past levels. If we regress future empirical skewness on its lagged value and lagged implied skew, we can see that the implied skew is significantly negatively correlated with future skewness (t-stat around -10.3). In the unreported table, I double sort the universe based on skewness and implied skew. After controlling for skewness, portfolios with lower implied skew have higher subsequent returns; after controlling implied skew, portfolios with lower past-month skewness have higher subsequent returns. Both factors are significant but implied skew has a stronger effect.

weekly return for each portfolio (Panel B). The excess return above the risk-free rate (3-month Treasury-Bill rate), alpha after the Fama-French-Momentum-Reversal model and annual Sharpe ratios are also presented. The main finding in Panel B matches that of Xing et al. (2010). If taking a long-short trading strategy (long the lowest SKEW quintile and short the highest), the weekly profit can be 28.6 basis points, which is 14.87% annually. The difference can be even more evident if controlling for the common factors and calculating alpha, which is 30.4 basis points weekly or 15.81% annually. When prolonging the holding horizon, the average weekly return from the long-short strategy becomes smaller, but remains highly significant.

A similar single-sorting based on TERM has been conducted and reported in Table 6. From the lowest quintile to the highest comparative, the average subsequent return has increased monotonically from 17.6 basis points (9.15% annually) to 29.1 basis points (15.13% annually). Alpha also behaves in a similar fashion. As a contrast to SKEW, when increasing the predicting window, the difference between top and bottom portfolios drops significantly and dies out fast. This corresponds to the analysis in Table 4.

3.2.2 Double Sort

It is more interesting to explore the interaction between SKEW and TERM. As we know from the above discussion, SKEW denotes directional information, while TERM carries the signal of timing. Since TERM itself is non-directional, its effect on return prediction is asymmetric when it attaches with high SKEW or low SKEW.

The asymmetry can be detected in the following way: I first sort on SKEW and divide the whole universe into 5 groups. Then I take out the top and bottom 20%, and rank each group based on TERM respectively. The idea is: the lowest SKEW quintile carries the most positive prospective information. Within this sample, the higher the TERM, the sooner the good news will be realized, so the higher the short-term return (Table 7, Column A). On the contrary, the highest SKEW quintile signifies the most adverse future information. Within this sample, the higher TERM subgroup should have lower near-end return because it indicates that the bad information will be soon absorbed into price. Indeed, Column B in Table 7 shows the return in the highest TERM subsample is worse than the lowest TERM group, even though the difference is not significant; it forms a clear contrast to Column A.

An alternative way to look at the asymmetry is to sort by TERM first and then SKEW, which is exhibited in Table 8. Column A shows the influence of SKEW within the lowest TERM quintile; and Column B is the comparative in the highest TERM quintile. Now that the high TERM acts as a barometer for the pending news to happen, SKEW's effect is more remarkable in the highest TERM quintile, either in terms of the difference between the top and bottom sub-quintiles, or in terms of t-statistics.

Both of the above double sorting approaches highlight the fact that it is possible to improve the portfolio performance by taking into account the TERM factor and its interaction with SKEW. As we see in Table 5, the long-short strategy based on single sorting on SKEW can bring about weekly return 28.6 basis points (14.87% annually), now while after double sorting the weekly return can increase to 40.2 basis points (20.90% annually).

4 Information Proxies

The previous sections have assessed the efficacy of the measurements from implied volatility surface SKEW and TERM. In this section, I am going on to explore the driver of their predictive power. The reason that implied volatility can forecast return lies in the fact that informed traders are more inclined to trade in options market due to the high leverage (Esley, O'Hara and Srinivas, 1998; Cao, Chen, Griffin, 2003). Hence information is first captured by the options market, and then spills to the stock market. Other research (Mayhew, Sarin and Shastri, 1995; Kumar, Sarin and Shastri, 1998) also found that the stocks that have corresponding options traded generally show greater price efficiency.

Two information proxies: probability of informed trading (PIN) and bid-ask spread are employed here to test their relationship with implied volatility surface. As shown in the EKO model, PIN is a measure estimates cross-sectional degree of information asymmetry[5]:

$$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} \tag{7}$$

where α is the probability that traders aquire private inforamtion about the firm's fundamentals at the beginning of a trading day; μ is the average arrival rate of buy(sell) orders from informed traders based on good(bad) news; ε_b and ε_s are arrival rate of buy and sell orders from uninformed traders. Hence the denominator $\alpha\mu + \varepsilon_b + \varepsilon_s$ denotes total amount of orders and the numerator

[5] Please check Easley, Kiefer and O'Hara (1997) for details.

is the portion from informed traders. So PIN calibrates the fraction of informed trading among overall order flows.

In microstructure literature, the bid-ask spreads mainly associate with two causes: inventory costs (Amihud and Mendelson, 1980, 1982) and adverse selection (Copeland and Galai, 1983, Glosten and Milgrom, 1985). The rationale of the latter is a market-maker will optimize his position by setting a bid-ask spread to maximize the difference between the expected profit from uninformed investors and expected loss occurred during trading with informed investors. In this sense, the bid-ask spread can be regarded as a "polluted" measure of asymmetric information (due to the inventory cost component) and only employed as a robustness check and supplement to PIN. Two alternative definitions of bid-ask spread are tested:

$$SPREAD = ASK - BID \tag{8}$$

$$SPREAD\% = \frac{ASK - BID}{Traded\ price} \tag{9}$$

If the information is the common force that moves SKEW and TERM, a strong connection between these factors and information proxies is expected. Therefore, the first step is to check the relationship between PIN/Bid-ask spread and SKEW/TERM.

The PIN measure is computed quarterly, while the SKEW and TERM, as used in the first section, are aggregated weekly. Therefore, I try two slightly different ways: the first is to match quarterly PIN with corresponding weekly SKEW/TERM directly; and the second is to first aggregate SKEW/TERM into quarterly and then link them with the same quarter PIN. For the bid-ask spread (as well as spread%) measure, it is calculated from daily close ask, bid and traded prices. Then the weekly average is matched with the weekly SKEW/TERM.

Table 9 illustrates the results from the panel regression when running information proxy (PIN or bid-ask spread) on SKEW, TERM and their cross product:

$$INF_{i,t} = SKEW_{i,t} + TERM_{i,t} + SKEW \cdot TERM_{i,t} \tag{10}$$

where *INF* refers to one of the measures that were just discussed: PIN or bid-ask spread. The errors are clustered by firm. Just as expected, firms with higher TERM and lower SKEW are engaged with higher information density. Furthermore, the interaction term also behaves as in the return prediction, such that the coexistence of higher TERM and lower SKEW intensify one another.

Table 10 carries more detailed double sorting on SKEW and TERM. The numbers in the table stand for the average PIN of each grid. Since the regression might only reflect an overall linear relationship, the sorting method can uncover how the information proxy varies across different quantiles of each measure. Generally the observation confirms what can be seen from panel regression. Especially for the relationship between TERM and PIN: across different SKEW quintiles, PIN almost always increases with TERM. However, the interesting bit is that across different TERM quintiles, PIN in the largest SKEW quintile appears consistently larger than that in the second largest SKEW quintile.

The non-monotonicity of the relationship between information content and SKEW is not surprising, though. On the contrary, it once again confirms what was analysed before - that SKEW designates directional signals - hence the information density is actually a U-shape function of SKEW. That is, both highest and lowest quintiles represent a higher possibility of private information presence, with opposite signs of prospective information. The lowest end, however, has a much stronger effect and dominates the correlation on average.

The above results have proved implied volatility surface indeed takes a root in private information, and has a direct link with the price informativeness. Now the question is whether the employed information proxy, PIN or Bid-ask spread, has already captured the explanatory power of future return, and makes SKEW/TERM redundant. In order to test this question, I include the information proxy and its second moment[6] in the baseline regression, such as

$$\begin{aligned} r_{i,w} &= \beta_{0,w} + \beta_{1,w} \cdot r_{i,w-1} + \beta_{2,w} \cdot r_{i,m-1} + \beta_{3,w} \cdot r_{i,y-1} + \beta_{4,w} \cdot SKEW_{i,w-1} \\ &\quad + \beta_{5,w} \cdot TERM_{i,w-1} + \beta_{6,w} \cdot SKEW^2_{i,w-1} + \beta_{7,w} \cdot TERM^2_{i,w-1} \\ &\quad + \beta_{8,w} \cdot SKEW_{i,w-1} \cdot TERM_{i,w-1} + \beta_{9,w} \cdot CONTROL_{i,w-1} \\ &\quad + \beta_{10,w} \cdot INF_{i,w-1} + \beta_{11,w} \cdot INF^2_{i,w-1} \end{aligned} \tag{11}$$

Because the PIN data is from 1996 to 2005, here I re-run the baseline regression without adding

[6]The second moment is added to capture the possible non-linear relationship. The positive relationship between PIN and return was initially found by Easley et al. (2002). For a critique of the results, check Duarte and Young (2008), and Mohanram and Rajgopal (2009).

information measures as a comparison in Column A, Table 11. The main effects are very close to what I have found in Section 1[7]. Column B-D list the results from the controlled regression after enclosing PIN, bid-ask spread and normalized spread respectively. The magnitude and significance of SKEW/TERM/cross product remain at a similar level as the un-controlled regression. Therefore, even though the volatility surface exhibits a very close relationship with information density, its prediction for future return is not overshadowed by stock market information proxies.

5 Example: Merger and Acquisitions

As discussed in previous sections, since the options market can be a platform for informed traders to execute front-running, implied volatility surface will display a strong and clear pattern before the information eventually gets disclosed. This section will go on to shed light on the implied volatility change during a typical example of concentrated information release, merger and acquisition.

Takeover activity has been widely studied in financial literature because it represents an ideal event that trading activity and information exchange both reach an enormous level. In particular, it is usually followed by an immediate corporate control shift and coupled with a large amount of price premium for takeover targets. Therefore, there is a great incentive for investors to hunt for the information regarding merger and acquisition and trade accordingly.

First, this paper conducts an event study on the target companies that have takeover announcements occurred between 1996 and 2010, and have options trading on the stocks, which end up with 1411 company takeovers[8]. With full backward-looking bias, I calculate the weekly average SKEW and TERM up to 53 weeks (about a year) prior to takeover announcement. Chart 1 illustrates the average pattern of SKEW and TERM, with 0 on the horizontal axis stands for the week of event. As we expected, SKEW exhibits a persistent downtrend, which is possibly caused by large purchase in call options or unwinding positions in put – both send out a positive signal about the future return; TERM, on the other hand, has a consistent uptrend that reinforces the probability that pending information is ahead.

[7]I also tried across different sub-periods as a robustness check. The prediction power of SKEW, TERM and the cross product is generally time-homogeneous.

[8]Apart from the standard filters as mentioned before, I also exclude the cases that the target companies have another takeover within previous one year, and only keep the first announcement. Takeovers which have the time gap longer than one year are remained.

Chart 1 also plots cumulative abnormal return (CAR) as a comparison to the trend of SKEW and TERM. The reason is while informed traders act on the private information in the options market (buy calls and sell puts), they will also trade the underlying stocks. To check if the options market indeed has a leading position in price discovery, weekly CAR is plotted with the two implied volatility surface measurements. The way to calculate CAR is similar to computing alpha in Section 1. Firm-specific factor loadings are calibrated based on Fama-French-Momentum-Reversal five-factor model, and then alpha for each firm each day is calculated as the difference between realized return and expected return, which applies the factor loadings and four factor levels on that day. Daily abnormal return is averaged into weekly, and aggregated from 53 weeks prior to the announcement week, which reaches a single target's CAR. In the end, the cross-sectional average CAR is reported. It is clear to see that SKEW and TERM (the options market) react earlier than CAR (the stock market), with a more remarkable leading trend spotted in SKEW. This once again confirms how SKEW and TERM differ in how far they predict future return. Although both of them start to show a trend as early as about 40 weeks (9-10 months) before the announcement, SKEW's pattern is more pronounced; while the significant upward jump in TERM happens much later.

So far, the result needs perfect hindsight to see how implied volatility surface evolves ahead of a big event. The next step is trying to extract ex-ante signals. Motivated by the portfolio sorting approach in Section 1, I summarize each target's portfolio ranking (both SKEW and TERM) at the last week prior to takeover. The idea is to explore the cross-sectional distribution of target firms' SKEW and TERM rightly before the event; and if following the portfolio purchase proposed by Section 1 (buy the lowest SKEW and highest TERM), how many of the overall takeover cases can be captured in the portfolio.

Table 12 lists number of merger and acquisitions that belong to each portfolio. Out of 1411 announcements, 131 (9.28%) have located within the intersection of lowest SKEW quintile and highest TERM quintile. This number is much higher than all the other occurrences. Summing across each quintile for each measure also indicates generally the targets have lower ranking in SKEW and higher ranking in TERM before the final price jump at the announcement. Without involving return comparison, this table itself states sorting portfolio according to implied volatility surface displays good ability to pick up takeover targets.

Past literature exerts large effort to forecast the possibility that a company will become a takeover target. The common methodology is to employ accounting data (ROE, P/E, growth, etc.) and construct a Logit or Probit model (Palepu, 1986; Powell, 1997, 2001). This paper does not specifically aim at predicting M&A probability; however, it provides an alternative way to tackle that type of problem. It would be interesting to see if combining the signals from relatively higher frequency market data (such as implied volatility surface in this paper) with those from lower frequency accounting data will improve the prediction performance.

6 Conclusion

Options are one of the best instruments to trade on private information; hence the implied volatility surface contains the information that has not been quick enough for the stock market to reflect. This paper explores the predictive power of two most important components on implied volatility surface: skew and term structure, as well as their interaction with each other. Using all available options data in the U.S. market from 1996 to 2010, it shows firms with lower skew and higher term structure tend to have higher subsequent returns. Implied volatility skew denotes a directional bet on firms' future outlook, while the term structure serves more as a non-directional signal about how fast the pending information will be released. Therefore, when high term structure occurs together with low skew, the predictive power for a positive jump in stock price will be even greater. Both Fama-MacBeth regression and portfolio construction support this argument, and significant economic interest can be realized by trading on implied volatility surface.

By connecting implied volatility surface measures to information proxies, this paper demonstrates it is private information that drives the changes in skew and term structure. Alligning to previous research on lead-lag relationships in option and stock prices, this paper reaches a similar conclusion (but from a new angle) that firms with more extreme implied volatility surface change are more informative. Controlled regression indicates the forecastability of skew, term structure and their cross product is robust even after introducing the information proxies and other firm-specific variables, showing that the information contained in implied volatility surface is not redundant.

This paper also applies the idea on M&A sample, which represents as an event of condensed information release. Implied volatility skew and term structure react accordingly much earlier before the announcements, and earlier than the reaction from the stock market. It shows that if following the weekly portfolio construction based on the lowest skew quintile and the highest term structure quintile, about 9.2% of overall target companies can be picked in the portfolio. This provides one promising application of the study: that is to combine it with other conventional M&A prediction models.

References

[1] Amihud, Y., and Mendelson, H., 1980, Dealership Market: Market-making with Inventory, Journal of Financial Economics, 8, 31-53.

[2] Amihud, Y., and Mendelson, H., 19082, Asset Price Behavior in a Dealership Market, *Financial Analysits Journal*, 50.

[3] Ang, A., Hodrick, R., Xing, Y., and Zhang X., 2006, The Cross-section of Volatility and Expected Returns, *The Journal of Finance*, 61, 259-299.

[4] Barbaris, N., and Huang, M., 2008, Stocks as Lotteries: The Implications of Probability Weighting for Security Prices, American Economic Review, 98, 2066, 2100.

[5] Bates, D., 1991, The crash of '87: was it expected? The evidence from options markets, *The Journal of Finance,* 46, 1009–1044.

[6] Brar, G., Giamouridis, D., and Liodakis, M., 2009, Predicting European Takeover Targets, *European Financial Management*, 15, 430-450.

[7] Brown, S., Hillegeist, S., and Lo, K., 2004, Conference calls and information asymmetry, *Journal of Accounting and Economics*, Vol 37, 343-366.

[8] Cao, C., Chen, Z., and Griffin, J., 2003, Informational Content of Option Volume Prior to Takeovers, *Yale SOM Working Paper,* No. ES-31.

[9] Carhart, M., 1997, On Persistence in Mutual Fund Performance, *Journal of Finance*, 52, 57-82.

[10] Carr, P., and Wu, L., 2007, Stochastic Skew in Currency Options, *The Journal of Fiancial Economics*, 86, 213-247.

[11] Chan, K., Chung, Y., and Fong, W., 2002, The Informational Role of Stock and Option Volume, *Review of Financial Studies,* 15, 1049-1975.

[12] Chen, Q., Goldstein, I., and Jiang, W., 2007, Price Informativeness and Investment Sensitivity to Stock Price. *Review of Financial Studies,* 20, 619–650.

[13] Cont, R., and Fonseca, J., 2002, Dynamics of Implied Volatility Surface, *Quantitative Finance,* 2, 45-60.

[14] Copeland, T., and Galai, D., 1983, Information Efffects on the Bid-Ask Spread, *Journal of Finance*, 38, 1457-1469.

[15] Dennis, P., Mayhew, S., and Stivers, C., 2006, Stock Returns, Implied Volatility Innovations, and the Asymmetric Volatility Phenomenon, *Journal of Financial and Quantitative Analysis*, 41, 381-406.

[16] Duarte, J., and Young, L., 2009, Why is PIN Priced? *Journal of Financial Economics*, 91, 119-138.

[17] Easley D., Kiefer, N., O'Hara M., and Paperman J., 1996, Liquidity, Information, and Infrequently Traded Stocks, *Journal of Finance*, 51, 1405-1436.

[18] Easley D., Kiefer, N., and O'Hara M., 1997, The Information Content of the Trading Process, *Journal of Empirical Finance*, 4, 159-186.

[19] Easley D., O'Hara M., and Srinivas, 1998, Option Volume and Stock Prices: Evidence on Where Informed Traders Trade, *Journal of Finance*, 53, 431-466.

[20] Easley, D., Hvidkjaer, S., and O'Hara, M., 2002. Is Information Risk a Determinant of Asset Returns? *Journal of Finance,* 57, 2185–2221.

[21] Fama, E., and French, K., 1993, Common risk Factors in the Returns on Stocks and Bonds, *Journal of Financial Economics*, 33, 3-56.

[22] Fama, E., and French, K., 1996, Multifactor Explanation of Asset Pricing Anomalies, *Journal of Finance*, 51, 55-84.

[23] Fama, E., and MacBeth, J., 1973, Risk, Return and Equilibrium: Empirical Tests, *Journal of Political Economy*, 607-636.

[24] Ferreira, D., Ferreira, M., and Raposo, C., 2011, Board Structure and Price Informativeness, *Journal of Financial Economics*, 99, 523-545.

[25] Ferreira, M., and Laux, P., 2007. Corporate Governance, Idiosyncratic Risk and Information Flow. *Journal of Finance,* 62, 951–989.

[26] Giot, P., 2005, Relationships between Implied Volatility Indexes and Stock Index Returns, *Journal of Portfolio Management*, 31, 92-100.

[27] Glosten, L., and Milgrom, P., 1985, Bid, Ask and Transaction Prices in a Special Market with Heterogenously Informed Traders, *Journal of Financial Economics*, 13, 71-100.

[28] Hull, J., Nelken, I., and White, A., 2004, Merton's Model, Credit Risk and Volatility Skews, *Journal of Credit Risk*, 1, 1-27.

[29] Kumar, R., Sarin, A., and Shastri, K., 1998, The Impact of Options Trading on the Market Quality of the Underlying Security: An Empirical Analysis, *Journal of Finance*, 53, 717-732.

[30] Lee C., and Ready, M., 1991, Inferring Trade Direction from Intraday Data, *Journal of Finance*, 2, 733-746.

[31] Mayhew, S., Sarin, A., and Shastri, K., 1995, The Allocation of Informed Trading across Related Markets: An Analysis of the Impact of Changes in Equity-option Margin Requirements, *Journal of Finance*, 50, 1635-1654.

[32] Mitton, T., and Vorkink, K., 2007, Equilibrium Underdiversification and the Preference for Skewness, *Review of Financial Studies*, 20, 1255-1288.

[33] Mixon, S. 2009, Option Markets and Implied Volatility: Past versus Present, *Journal of Financial Economics*, 94, 171-191.

[34] Mixon, S., 2010, What Does Implied Volatility Skew Measure? SSRN working paper series.

[35] Mohanram, P., and Rajgopal, S., 2009, Is PIN Priced Risk? Journal of Accounting and Economics, 47, 226-243.

[36] Palepu, K., 1986, Predicting takeover targets : A methodological and empirical analysis, *Journal of Accounting and Economics*, 8, 3-35.

[37] Powell, R., 1997, Modelling Takeover Likelihood, *Journal of Business Finance and Accounting*, 24, 1009-1030.

[38] Powell, R., 2001, Takeover Prediction and Portfolio Performance, *Journal of Business Finance and Accounting,* 28, 993-1011.

[39] Roux, M., 2006, A Long-term Model of the Dynamics of the S&P Implied Volatility Surface, *North American Actuarial Journal*, 11, No. 4.

[40] Stein, J., 1989, Overreactions in the Options Market, *Journal of Finance*, 44, 1011-1023.

[41] Vega, C., 2006, Stock Price Reaction to Public and Private Information, *Journal of Financial Economics*, 82, 103–133.

[42] Xing, Y., Zhang, X., and Zhao, R., 2010, What Docs thc Individual Option Volatility Smirk Tell Us about Future Equity Returns?, *Journal of financial and Quantitative Analysis*, 45, 641-662.

Table 1 Summary Statistics

This table shows the summarized statistics of main variables covered in this paper. Panel A lists number of firms exist in each year. Any firm that has valid intersection between CRSP and OptionMetrics data that is longer than 6 months is retained in the sample. Panel B shows mean and a range of quantiles of employed variables. For each firm, SKEW is the weekly average of daily implied volatility skew, which is the difference between 30-day tenor 25 Δ OTM PUT implied volatility and 25 Δ OTM CALL implied volatility, divided by 30-day tenor 50 Δ implied volatility; z-SKEW is the normalized SKEW by the firm-specific mean and standard deviation for the previous 6 months; TERM is the weekly average of daily implied volatility term structure, which is the difference between 30-day tenor 50 Δ implied volatility and 1-year 50 Δ implied volatility, divided by 30-day 50 Δ implied volatility; z-TERM is the normalized TERM by the firm-specific mean and standard deviation; SIZE is the market capitalization; TURNOVER is daily transaction volume divided by shares outstanding; VOL is the standard deviation of daily return for the previous month; SKEWNESS is the empirical skewness of daily return for the previous month; PVOL is difference between 30-day 50 Δ implied volatility and VOL; PIN is quarterly probability of informed trading calculated from microstructure data, as proposed by EKO model; Bid-ask spread/price is the daily close bid-ask spread normalized by close prices.

Panel A: number of firms

Year	# of firms	Year	# of firms
1996	1769	2004	2385
1997	2257	2005	2580
1998	2569	2006	2839
1999	2706	2007	3141
2000	2505	2008	3180
2001	2351	2009	3120
2002	2362	2010	3081
2003	2223		
Overall	6108		

Panel B: statistics of main variables

Variable	Mean	5%	10%	25%	50%	75%	90%	95%
SKEW	0.090	-0.161	-0.070	0.016	0.088	0.163	0.248	0.324
z-SKEW	0.021	-1.432	-1.049	-0.508	0.011	0.540	1.099	1.503
TERM	0.042	-0.133	-0.083	-0.016	0.040	0.100	0.170	0.223
z-TERM	-0.022	-1.752	-1.343	-0.714	-0.040	0.655	1.327	1.770
SIZE (billion $)	5.925	0.135	0.208	0.466	1.269	3.779	23.390	84.713
TURNOVER (%)	1.433	0.156	0.223	0.386	0.714	1.334	2.420	3.600
VOL (%)	2.908	0.905	1.118	1.598	2.386	3.601	5.297	6.675
SKEWNESS	0.183	-1.210	-0.784	-0.291	0.169	0.650	1.205	1.654
PVOL	0.463	0.183	0.218	0.294	0.409	0.579	0.780	0.917
PIN	0.145	0.067	0.081	0.105	0.136	0.175	0.219	0.252
Bid-ask spread/price (%)	0.637	0.036	0.052	0.095	0.232	0.799	1.724	2.525

Table 2 Fama-MacBeth Regression on Weekly Return

This table shows results from the Fama-MacBeth regression (4), (5) and (6). The dependent variable is the leading weekly return. For the independent variables, RET (lag. w) is the past weekly return; RET (lag. m) is the past monthly return excluding the last week, scaled to the weekly level; RET (lag. y) is the past yearly return excluding the last month, scaled to the weekly level; SKEW is the normalized implied volatility skew and $SKEW^2$ is its square; TERM is the normalized implied volatility term structure and $TERM^2$ is its square; SKEW*TERM is the cross product of SKEW and TERM; Log(SIZE) is the nature log of market capitalization; TURNOVER is the transaction volume normalized by shares outstanding; SKEWNESS and VOL are the empirical skewness and standard deviation of the previous month daily return; PVOL is difference between 30-day 50 Δ implied volatility and VOL. All independent variables are one week lagged from the dependent variable. The results are after Newey-West adjustment.

Dependent variable=$r_{i,w}$					
	(A)	(B)	(C)	(D)	(E)
RET (lag. w)	-0.0188	-0.0165	-0.0186	-0.0184	-0.0260
	(-5.06)***	(-5.27)***	(-5.03)***	(-4.97)***	(-7.98)***
RET (lag. m)	-0.00104	-0.00362	-0.00094	-0.00073	-0.01702
	(-0.15)	(-0.52)	(-0.13)	(-0.10)	(-2.91)***
RET (lag. y)	0.0578	0.0544	0.0562	0.0558	0.0544
	(1.51)	(1.42)	(1.47)	(1.46)	(1.80)*
SKEW	-0.000938		-0.000954	-0.00102	-0.00100
	(-10.33)***		(-10.57)***	(-10.11)***	(-11.60)***
$SKEW^2$				0.0000222	-1.06e-6
				(0.54)	(-0.29)
TERM		0.000220	0.000203	0.000225	0.0000857
		(2.26)**	(2.09)**	(2.18)**	(0.73)
$TERM^2$				0.0000212	-5.13e-6
				(0.68)	(-0.13)
SKEW*TERM				-0.000220	-0.000221
				(-3.73)***	(-3.83)***
log(SIZE)					-0.000207
					(-1.62)
TURNOVER					0.0322
					(3.55)***
SKEWNESS					0.00120
					(8.94)***
VOL					-0.0200
					(-1.48)
PVOL					-0.00275
					(-1.28)

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 3 Fama-MacBeth Regression on Weekly Alpha

This table shows the results from the Fama-MacBeth regression (4), (5) and (6) but with weekly alpha as the dependent variable. Alpha is calculated as the residual of the Fama-French-Momentum-Reversal five-factor model. For the independent variables, ALPHA is the lagged weekly alpha; SKEW is the normalized implied volatility skew and $SKEW^2$ is its square; TERM is the normalized implied volatility term structure and $TERM^2$ is its square; SKEW*TERM is the cross product of SKEW and TERM; Log(SIZE) is the nature log of market capitalization; TURNOVER is the transaction volume normalized by shares outstanding; SKEWNESS and VOL are the empirical skewness and standard deviation of the previous month daily return; PVOL is difference between 30-day 50 Δ implied volatility and VOL. All independent variables are one week lagged from the dependent variable. The results are after Newey-West adjustment.

Dependent variable=$\alpha_{i,w}$					
	(A)	(B)	(C)	(D)	(E)
ALPHA	-0.0214	-0.0220	-0.0212	-0.0211	-0.0239
	(-7.17)***	(-7.40)***	(-7.14)***	(-7.10)***	(-8.47)***
SKEW	-0.00103		-0.00104	-0.00110	-0.00113
	(-11.83)***		(-12.03)***	(-11.15)***	(-12.87)***
$SKEW^2$				0.0000387	0.0000146
				(1.01)	(0.40)
TERM		0.000147	0.000127	0.000135	0.0000410
		(1.70)*	(1.46)	(1.45)	(0.43)
$TERM^2$				-0.0000122	-0.0000288
				(-0.32)	(-0.79)
SKEW*TERM				-0.000217	-0.000208
				(-3.67)***	(-3.56)***
log(SIZE)					0.0000581
					(0.68)
TURNOVER					0.0404
					(4.68)***
SKEWNESS					0.000811
					(7.97)***
VOL					-0.0198
					(-1.80)*
PVOL					-0.000525
					(-0.36)

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 4 Fama-MacBeth Regression of Multiple Horizons

This table shows the results from the Fama-MacBeth regression (6) with the dependent variable as returns from different leading periods. In column (A), the dependent variable is the average weekly return for the leading 4 weeks; in column (B)-(G), the dependent variable is the weekly return for the 1st – 6th week in the future respectively. For the independent variables, RET (lag. w) is the past weekly return; RET (lag. m) is the past monthly return excluding the last week, scaled to the weekly level; RET (lag. y) is the past yearly return excluding the last month, scaled to the weekly level; SKEW is the normalized implied volatility skew and $SKEW^2$ is its square; TERM is the normalized implied volatility term structure and $TERM^2$ is its square. SKEW*TERM is the cross product of SKEW and TERM; Log(SIZE) is the nature log of market capitalization; TURNOVER is the transaction volume normalized by shares outstanding; SKEWNESS and VOL are the empirical skewness and standard deviation of the previous month daily return; PVOL is difference between 30-day 50 Δ implied volatility and VOL. All independent variables are one week lagged from the 1st week dependent variable. The results are after Newey-West adjustment.

Dependent variable=$r_{i,k}$							
	(A) Next 4w	(B) 1st week	(C) 2nd week	(D) 3rd week	(E) 4th week	(F) 5th week	(G) 6th week
RET (lag. w)	-0.00976	-0.0260	-0.00914	-0.00384	0.00143	-0.00211	0.000119
	(-7.07)***	(-7.98)***	(-3.16)***	(-1.46)	(0.52)	(-0.81)	(0.05)
RET (lag. m)	-0.00922	-0.01702	-0.00724	-0.00204	-0.00827	-0.00343	-0.00271
	(-3.47)***	(-2.91)***	(-1.26)	(-0.36)	(-1.50)	(-0.63)	(-0.48)
RET (lag. y)	0.0618	0.0544	0.0644	0.0643	0.0615	0.0630	0.0536
	(3.92)***	(1.80)*	(2.19)**	(2.19)**	(2.11)**	(2.19)**	(1.90)*
SKEW	-0.000503	-0.00100	-0.000458	-0.000344	-0.000225	-9.75e-5	-5.13e-5
	(-12.27)***	(-11.60)***	(-5.55)***	(-4.34)***	(-2.73)***	(-1.25)	(-0.61)
$SKEW^2$	-1.75e-5	-1.06e-6	-2.21e-6	-1.62e-5	-3.79e-5	-2.43e-5	-4.23e-5
	(-1.08)	(-0.29)	(-0.07)	(-0.51)	(-1.22)	(-0.77)	(-1.28)
TERM	6.84e-6	8.57e-5	-0.000101	-4.43e-6	-4.89e-5	-0.000134	-0.000162
	(0.12)	(0.73)	(-0.85)	(-0.04)	(-0.43)	(-1.10)	(-1.39)
$TERM^2$	-4.18e-5	-5.13e-6	-8.81e-5	-4.53e-5	-3.80e-5	-6.24e-5	5.02e-6
	(-2.33)**	(-0.13)	(-2.04)**	(-1.15)	(-0.95)	(-1.39)	(0.12)
SKEW*TERM	-8.84e-5	-0.000221	-1.60e-5	-2.59e-5	-7.01e-5	2.97e-5	8.41e-6
	(-2.98)***	(-3.83)***	(-0.28)	(-0.44)	(-1.22)	(0.50)	(0.13)
log(SIZE)	-0.000190	-0.000207	-0.000218	-0.000241	-0.000247	-0.000235	-0.000210
	(-3.24)***	(-1.62)	(-1.69)*	(-1.86)*	(-1.90)*	(-1.83)*	(-1.64)
TURNOVER	0.0132	0.0322	0.00945	0.00364	0.00125	0.00565	-0.00849
	(3.31)***	(3.55)***	(1.08)	(0.45)	(0.17)	(0.74)	(-1.02)
SKEWNESS	0.000489	0.00120	0.000418	0.000220	0.000130	7.17e-5	-2.83e-5
	(8.18)***	(8.94)***	(3.16)***	(1.72)*	(0.55)	(0.62)	(-0.23)
VOL	-0.00859	-0.0200	-0.00366	-0.00868	-0.000931	-0.0135	-0.00842
	(-1.45)	(-1.48)	(-0.28)	(-0.67)	(-0.07)	(-1.04)	(-0.64)
PVOL	-0.00270	-0.00275	-0.00279	-0.00236	-0.00266	-0.00190	-0.00184
	(-2.37)**	(-1.28)	(-1.35)	(-1.16)	(-1.34)	(-0.95)	(-0.94)

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 5 Portfolio Trading Strategy (Single Sort on SKEW)

This table shows the portfolio characteristics and performance when sorting on normalized SKEW and dividing the universe into 5 portfolios. Portfolio is constructed equally weighted by its constituents and rebalanced on a weekly basis. LOW contains the firms with the lowest 20% SKEW and HIGH contains those with the highest 20% SKEW.

Portfolio Criteria: SKEW (normalized)

Panel A: Quintile characteristics

	SKEW (normalized)	SKEW (not normalized)	size(b$)	last month vol	last month skewness	TERM (normalized)
LOW	-1.206	-0.073	6.055	0.0300	0.124	0.021
2	-0.400	0.039	5.766	0.0295	0.174	0.029
3	0.009	0.083	5.539	0.0295	0.194	0.006
4	0.425	0.131	5.783	0.0291	0.212	-0.042
HIGH	1.278	0.257	6.384	0.0282	0.235	-0.110
low-high	-2.483	-0.330				
t-stat	-193.27***	-101.89***				

Panel B: future returns

	return	ex-ret[1]	alpha	ann. Sh	next 4w	next 8w	next 12w	next 16w	next 20w	next 24w
LOW	0.377%	0.318%	0.269%	0.621	0.294%	0.273%	0.269%	0.264%	0.262%	0.260%
2	0.292%	0.234%	0.175%	0.474	0.263%	0.254%	0.254%	0.251%	0.251%	0.253%
3	0.217%	0.158%	0.088%	0.324	0.232%	0.236%	0.238%	0.236%	0.240%	0.240%
4	0.191%	0.132%	0.061%	0.274	0.213%	0.220%	0.222%	0.223%	0.226%	0.228%
HIGH	0.091%	0.032%	-0.034%	0.069	0.159%	0.182%	0.195%	0.199%	0.203%	0.205%
L-H	0.286%	0.290%	0.304%		0.130%	0.091%	0.074%	0.065%	0.059%	0.056%
t-stat	8.11***	8.11***	11.75***		7.90***	7.02***	7.52***	7.90***	8.19***	8.30***

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

[1] Return minus risk-free rate, which is the contemporaneous 3-month treasury rate.

Table 6 Portfolio Trading Strategy (Single Sort on TERM)

This table shows the portfolio characteristics and performance when sorting on normalized TERM and dividing the universe into 5 portfolios. Portfolio is constructed equally weighted by its constituents and rebalanced on a weekly basis. LOW contains the firms with the lowest 20% TERM and HIGH contains those with the highest 20% TERM.

Portfolio Criteria: TERM (normalized)

Panel A: Quintile characteristics

	TERM (normalized)	TERM (not normalized)	size(b$)	last month vol	last month skewness	SKEW (normalized)
LOW	-1.387	-0.058	6.000	0.0278	0.208	0.096
2	-0.518	0.006	6.091	0.0285	0.194	0.034
3	-0.030	0.041	5.908	0.0292	0.187	0.006
4	0.466	0.078	5.943	0.0297	0.177	-0.017
HIGH	1.373	0.148	5.586	0.0310	0.173	-0.013
low-high	-2.760	-0.206				
t-stat	-226.18***	-135.76***				

Panel B: future returns

	return	ex-ret	alpha	ann. Sh	next 4w	next 8w	next 12w	next 16w	next 20w	next 24w
LOW	0.176%	0.117%	0.069%	0.241	0.210%	0.224%	0.234%	0.234%	0.236%	0.241%
2	0.201%	0.142%	0.093%	0.291	0.215%	0.220%	0.226%	0.226%	0.229%	0.232%
3	0.231%	0.172%	0.117%	0.351	0.234%	0.235%	0.235%	0.234%	0.236%	0.238%
4	0.271%	0.212%	0.135%	0.431	0.253%	0.243%	0.240%	0.240%	0.242%	0.238%
HIGH	0.291%	0.232%	0.145%	0.471	0.250%	0.244%	0.242%	0.237%	0.241%	0.238%
L-H	-0.115%	-0.115%	-0.076%		-0.041%	-0.020%	-0.007%	-0.003%	-0.005%	0.004%
t-stat	-3.19***	-3.19***	-2.86***		-2.37**	-1.79*	-0.760	-0.34	-0.62	0.51

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 7 Effect of TERM in the Lowest and Highest SKEW Quintiles

This table shows the effect of TERM in portfolios with the lowest and highest SKEW. Portfolio is first constructed by sorting on SKEW. Column (A) contains the quintile with the lowest SKEW; column (B) contains the highest SKEW quintile. Within each quintile, portfolio is formed again based on TERM. LOW contains those with the lowest 20% TERM; HIGH contains those with the highest 20% TERM.

		(A) The Lowest SKEW Quintile	(B) The Highest SKEW Quintile
TERM	LOW	0.318%	0.064%
	2	0.257%	0.081%
	3	0.387%	0.111%
	4	0.448%	0.145%
	HIGH	0.476%	0.053%
	L-H	-0.158%	0.011%
	t-stat	-2.8***	0.21

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 8 Effect of SKEW in the Lowest and Highest TERM Quintiles

This table shows the effect of SKEW in portfolios with the lowest and highest TERM. Portfolio is first constructed by sorting on TERM. Column (A) contains the quintile with the lowest TERM; column (B) contains the highest TERM quintile. Within each quintile, portfolio is formed again based on TERM. LOW contains those with the lowest 20% SKEW; HIGH contains those with the highest 20% SKEW.

		(A) The Lowest TERM Quintile	(B) The Highest TERM Quintile
SKEW	LOW	0.290%	0.465%
	2	0.246%	0.413%
	3	0.119%	0.287%
	4	0.160%	0.224%
	HIGH	0.063%	0.064%
	L-H	0.227%	0.402%
	t-stat	4.40***	7.27***

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 9 Panel Regression of Information Proxies

This table shows the result of running panel regression of PIN (probability of informed trading) and bid-ask/spread on SKEW, TERM and their cross product. The panel regression is clustered by firm. When using PIN as the independent variable, in column "Weekly" the dependent variables are weekly; in column "Quarterly" the dependent variables are averaged into quarterly. When using bid-ask spread as the independent variable, in column "Spread" the bid-ask spread is calculated as the difference between daily close bid and ask, and then average into weekly; in column "Spread%" the bid-ask spread is calculated as the difference between bid and ask, normalized by daily close price and then average into weekly.

	PIN		Bid-ask spread (weekly)	
	Weekly	Quarterly	Spread	Spread%
SKEW	-0.0239	-0.0412	-0.0562	-0.00580
	(-10.87)***	(-7.10)***	(-15.40)***	(-18.28)***
TERM	0.0189	0.0422	0.134	0.00730
	(5.79)***	(6.60)***	(19.74)***	(27.90)***
SKEW*TERM	-0.0102	-0.0131	-0.0134	-0.00120
	(-4.43)***	(-0.53)	(-7.09)***	(-5.36)***

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 10 PIN in Double Sorting on SKEW and TERM

This table shows the average PIN in different quintile groups of SKEW and TERM. In Panel A, SKEW and TERM are both weekly measures and they match with the corresponding PIN which is quarterly. In Panel B, SKEW and TERM are first averaged into quarterly and then match with the PIN. Results of double-sorted and single-sorted (while controlled for the other measure) are listed.

Panel A: Match quarterly PIN with corresponding weekly SKEW and TERM

		TERM				
		LOW	2	3	4	HIGH
SKEW	LOW	0.1596	0.1602	0.1609	0.1620	0.1676
	2	0.1473	0.1483	0.1487	0.1487	0.1487
	3	0.1373	0.1392	0.1391	0.1388	0.1410
	4	0.1312	0.1337	0.1345	0.1348	0.1373
	HIGH	0.1411	0.1397	0.1392	0.1404	0.1494

	TERM (control SKEW)	SKEW (control TERM)
LOW	0.1434	0.1620
2	0.1442	0.1480
3	0.1445	0.1390
4	0.1448	0.1346
HIGH	0.1489	0.1424
L-H	-0.0055	0.0196
t-stat	8.33***	31.72***

Panel B: Match quarterly PIN with quarterly averaged SKEW and TERM

		TERM				
		LOW	2	3	4	HIGH
SKEW	LOW	0.1649	0.1647	0.1683	0.1681	0.1763
	2	0.1513	0.1522	0.1517	0.1518	0.1539
	3	0.1387	0.1420	0.1418	0.1417	0.1412
	4	0.1301	0.1319	0.1332	0.1353	0.1385
	HIGH	0.1393	0.1349	0.1340	0.1326	0.1495

	TERM (control SKEW)	SKEW (control TERM)
LOW	0.1452	0.1686
2	0.1453	0.1515
3	0.1450	0.1409
4	0.1464	0.1341
HIGH	0.1523	0.1391
L-H	-0.0071	0.0295
t-stat	3.01***	10.87***

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 11 Fama-MacBeth Regression after Including Information Proxy

This table shows the results from the Fama-MacBeth regression (11). The dependent variable is the leading weekly return. Besides the independent variables in Table 2, it also controls for PIN or bid-ask spread (absolute spread or percentage of trading prices). The results are after Newey-West adjustment.

Dependent variable=$r_{i,w}$				
	(A)	(B)	(C)	(D)
RET (lag. w)	-0.0392 (-10.23)***	-0.0392 (-10.32)***	-0.0263 (-8.13)***	-0.0269 (-8.28)***
RET (lag. m)	-0.0248 (-3.59)***	-0.0250 (-3.66)***	-0.0169 (-2.89)***	-0.0177 (-3.00)***
RET (lag. y)	0.0880 (2.60)***	0.0877 (2.59)***	0.0565 (1.89)*	0.0510 (1.71)*
SKEW	-0.00137 (-12.92)***	-0.00137 (-12.87)***	-0.00100 (-11.36)***	-0.00101 (-11.42)***
SKEW^2	1..89e-5 (0.41)	1.93e-5 (0.42)	-8.83e-6 (-0.24)	-1.00e-5 (-0.27)
TERM	0.000145 (1.05)	0.000128 (0.94)	7.45e-5 (0.64)	4.22e-5 (0.36)
TERM^2	-7.55e-5 (-1.52)	-7.76e-5 (-1.56)	2.96e-6 (0.07)	-2.43-6 (-0.06)
SKEW*TERM	-0.000219 (-2.73)***	-0.000214 (-2.65)***	-0.000215 (-3.47)**	-0.000217 (-3.53)**
log(SIZE)	-0.000216 (-1.20)	-0.000218 (-1.25)	-0.000213 (-1.58)	-0.000254 (-2.03)**
TURNOVER	0.0597 (3.80)***	0.0610 (4.12)***	0.0293 (3.25)***	0.0268 (3.12)***
SKEWNESS	0.00141 (8.42)***	0.00141 (8.48)***	0.00121 (8.88)***	0.00122 (8.93)***
VOL	-0.0298 (-1.83)*	-0.0300 (-1.86)*	-0.0209 (-1.55)	-0.0179 (-1.32)
PVOL	-0.00310 (-1.02)	-0.00311 (-1.06)	-0.00320 (-1.51)	-0.00266 (-1.23)
PIN		0.00688 (1.00)		
PIN^2		-0.0224 (-1.69)*		
SPREAD			-0.00585 (-1.95)*	
SPREAD2			0.00633 (0.79)	
SPREAD%				0.0320 (0.37)
SPREAD%2				-9.15 (-1.26)

*** indicates significance at 1%; ** indicates significance at 5%; * indicates significance at 10%.

Table 12 Takeover Distribution in SKEW-TERM Portfolios

This table shows within each portfolio constructed by double sorting on weekly SKEW and TERM (5×5), number of firms that have been taken over in the following week. For example, the number for the intersection of LOW SKEW and HIGH TERM is 131, which means if forming a portfolio by selecting lowest SKEW and highest TERM during 1996 to 2010, there were 131 firms in this portfolio being taken over within the holding horizon (one week).

		TERM					
		LOW	2	3	4	HIGH	SUM
SKEW	LOW	49	51	56	61	131	348
	2	43	52	44	55	94	288
	3	31	46	40	52	81	250
	4	49	54	43	66	62	274
	HIGH	41	38	45	42	85	251
	SUM	213	241	228	276	453	1411

Chart 1 SKEW, TERM and CAR before Taken Over

This chart plots the weekly implied volatility skew and term structure along with cumulative abnormal return (CAR) of the firms that have been taken over, up to one year before the merger and acquisition announcements. SKEW is the weekly average of daily implied volatility skew, which is the difference between 30-day tenor 25 Δ OTM PUT implied volatility and 25 Δ OTM CALL implied volatility, divided by 30-day tenor 50 Δ implied volatility; TERM is the weekly average of daily implied volatility term structure, which is the difference between 30-day tenor 50 Δ implied volatility and 1-year 50 Δ implied volatility, divided by 30-day 50 Δ implied volatility; CAR is calculated as the sum of abnormal return from one year prior to takeover announcement, in which abnormal return is the difference between actual return and the return estimated from Fama-French-Momentum-Reversal five-factor model.