Nonlinear Limits to Arbitrage^{$\stackrel{r}{\approx}$}

Charlie X. Cai¹, Jingzhi Chen², Robert Faff³, Yongcheol Shin⁴

Abstract

Arbitrage costs and funding constraints are two major drivers behind limits to arbitrage. We develop a theoretical framework that can analyze their impacts on arbitrage activities. When the funding constraint is not binding, arbitrage costs are the main concern, and arbitrageurs will be able to adjust investment strategy to achieve the expected risk-adjusted return (the capital allocation effect). The higher the mispricing, the more arbitrage capital will be allocated to the investment opportunity. When the funding constraint is binding, however, we show that the funding constraint effect dominates the capital allocation effect, such that the available capital will dictate the level of arbitrage. Extremely larger mispricing would rather trigger the slow moving capital effect, which further reduces arbitrage activities. In sum, our analysis provide the prediction that overall arbitarge activites tend to display an inverse U-shape against the magnitude and volatility of mispricing errors. We apply a state-dependent, general error-correction model to testing our theoretical predictions in the S&P 500 index spot and future markets. Our empirical results provide stong support for such regime-dependent nonlinear limits to arbitrage. Furthermore, we identify the stress periods when the funding constraint has been binding, the years of 1987, 1998, 2000 and 2007-08, consistent with the slow moving capital period documented in the literature.

Keywords: Limits to Arbitrage, Slow Moving Capital, Mispricing Correction, Noise Momentum,

 $^{^{\}diamond}$ This is a preliminary version created to stimulate discussion and critical comments. Any errors or omissions are the responsibility of the authors.

¹Professor of Finance and Deputy Director of the Institute of Banking and Investment (IBI) in Leeds University Business School, Leeds, UK, E-mail: xc@lubs.leeds.ac.uk

 $^{^2\}mathrm{Ph.d}$ in Economics, Department of Economics and Related Studies, University of York, York YO105dd, UK, E-mail: jzc500@york.ac.uk

 $^{^3}$ Professor of Finance, UQ Business School, University of Queensland, Queensland, 4072, Australia. Phone: +61 7 3346 8055. Fax: +61 7 3346 8188. E-mail: r.faff@business.uq.edu.au

⁴Professor of Economics, Department of Economics and Related Studies, University of York, York YO105dd, UK, E-mail: ys789@york.ac.uk

1. Introduction

Arbitrageurs' active search for and the subsequent elimination of arbitrage opportunities ensure that mispricing is short-lived. However, arbitrage is costly and risky, and the presence of limits to arbitrage deters arbitrageurs from taking full mispricing corrections, leading to the market anomalies and the resource misallocations. It is therefore crucial to understand what limits arbitrage, see Gromb and Vayanos (2010) for an excellent survey.

In a broad context, there have been two main studies focusing on the impediments to arbitrage related to the arbitrage costs and the funding constraints. On the one hand, the earlier studies highlight the presence of arbitrage costs borne by arbitrageurs when exploiting the mispricing opportunity. Arbitrage costs include transaction and holding cost (Pontiff, 2006): commissions, short-selling cost and the risk exposure (Ali, Hwang and Trombley, 2003; Abreu and Brunnermeier, 2002; Doukas, Kim and Pantzalis, 2010). Arbitrage cost tends to deter arbitrage activity if the arbitrage profit cannot cover the costs. The earlier empirical evidence are provided in Dwyer et al. (1996) and Martens, Kofman and Vorst (1998), who apply a threshold error correction model and find that arbitrageurs exploit arbitrage opportunity only if mispricing error exceeds a certain threshold. See also Tse (2001) who extends the approach by applying a smooth transition autoregressive model to heterogeneous arbitrageurs faced with various trading costs. In this setup, arbitrageurs tactically allocate capital across different arbitrage strategies so as to earn the maximized risk-adjusted returns, which we call the capital allocation effect.

Whereas the initial research on limits to arbitrage focuses on the asset side of the balance sheet (e.g., the fundamental value of the arbitrage opportunity), more recent theoretical studies focus on the funding risk of arbitrage, and attempt to explain how the shocks to the arbitrageurs' funding would result in severe and prolonged mispricing (Mitchell and Pulvino, 2012). A seminal paper by Shleifer and Vishny (1997) develops the model in which performance-based investors redeem capital from arbitrageurs following negative performance. Consequently, arbitrageurs are forced to reduce positions, even at the time when expected returns are highest. Eventually, large and persistent mispricing will

be resolved only when new capital can be raised, but it takes several months, especially in the presence of search costs and agency and information asymmetries. This phenomenon is called the slow moving capital hypothesis (Mitchell, Pulvino and Pedersen, 2007). Brunnermeier and Pedersen (2009) further derive the theoretical predictions that the market illiquidity and the funding illiquidity are mutually reinforcing, leading to vicious cycle of liquidity spirals.⁵ See also the growing number of studies (e.g., Acharya, Shin and Yorulmazer, 2010; Duffie, 2010;). We call this chain of mechanisms the funding constraint effect, which tends to deter arbitrage activities and thus destabilize prices especially in the presence of the extremely large mispricing errors.

While most of the theoretical and empirical studies tend to focus on one aspect of limits to arbitrage, we aim to propose a unified approach, which can provide the testable predictions associated with both the capital allocation and the funding constraint effect. To this end we extend the seminal work by Shleifer and Vishny (1997, henceforth SV), and analyze nonlinear impacts of mispricing on overall arbitrage activities. To replicate the situation where arbitrageurs attempt to exploit the price discrepancy while simultaneously facing agency problems, market frictions and funding constraints, we should allow arbitrageurs to face a trade-off in the equilibrium between making investments now and waiting for larger arbitrage opportunities in the subsequent period when mispricing deepens.⁶ To capture such multi-period arbitrage activities, Faff, Cai and Shin (2015, FCS hereafter) develop a generalized error correction model (GECM) and estimate both the initial mispricing correction and the subsequent noise momentum parameters where the latter is designed to measure persistence of the uncorrected pricing errors and affect the ex ante capital allocation strategy by arbitrageurs. Applying it to a wide range of international spot-futures market pairs, FCS document pervasive evidence of noise momentum around the world. In this unified theoretical framework, a higher initial mispricing correction and a lower mispricing persistence induce a faster overall speed of adjustment. Furthermore,

 $^{^{5}}$ A fall in prices reduces the value of the assets, leading to an increase in leverage (the loss spiral). Simultaneously, the subsequent increase in volatility erodes the collateral value of the assets leading to an increase in margins (the margin spiral). This forces the arbitrageur to reduce his positions. In turn it creates a selling pressure on prices, further leading to another round of loss and margin spirals. This vicious cycle causes market prices to deviate significantly from their fundamental values.

 $^{^{6}}$ Acharya, Shin and Yorulmazer (2010) show that "the two choices — whether to invest in profitable activities or to set aside funds for arbitrage in the future — earns the same rate of return when viewed ex ante. As a consequence, limited provision of arbitrage capital and fire sales emerge as robust features of the equilibrium. This result explains why crises that erupt after a long boom are associated with sharper, more severe downturns."

we can decompose arbitrageurs' initial mispricing correction into the capital allocation and the funding constraint effect. We derive the main theoretical prediction that the funding constraint effect dominates the capital allocation effect only if the capital constraint is binding, which may heavily deter overall arbitrage.

In particular, we will investigate the implications of these limits to arbitrage under two alternative paths to equilibrium – characterized by the partial-investment and the full-investment strategy. Under the partial-investment equilibrium, the funding constraint is not binding and thus the larger mispricing would induce more and faster arbitrage activities. Meanwhile we observe that the subsequent mispricing persistence is potentially high, which captures the higher risk of deepening error in the next period. Hence, in this state, arbitrageurs' main concern is to evaluate whether the maximum risk-adjusted return can cover arbitrage costs. By contrast, under the full-investment equilibrium, the funding constraint is binding such that the (extremely) larger mispricing may result in severe and persistent funding constraints, which tends to deter arbitrageurs' initial mispricing correction and slow the overall speed of adjustment, even though mispricing persistence will be significantly lower (the subsequent price recovery is less uncertain). Therefore, in this state, arbitrageurs are likely to suffer from the limit to arbitrage mainly due to the slow moving capital. Thus, our theoretical analysis reveals new insights about the impacts of the limits to arbitrage on arbitrageurs' reactions to a deepening mispricing error.

Despite that rich insights have been offered in the existing theoretical literature on the limits to arbitrage, an analysis of short-term dynamics has been more or less under-researched, at least empirically (FCS). To the best of our knowledge, this paper is the first attempt to rigorously develop both theoretical and empirical frameworks to investigate whether and when the binding funding constraint becomes a major source of limits to arbitrage.⁷ Our theory suggests that overall arbitrage activities do not rise linearly with mispricing errors. Rather, their relationship tends to be the regime-dependent, and overall arbitrage activities display an inverse U-shape against the magnitude and volatility of

⁷?? Sofianos (1993) and Tse (2001) consider the capital constraint as one of the arbitrage cost, though the capital constraint is not binding in their nonlinear price dynamics analysis. However, unlike their studies suggesting that arbitrageurs may skip the small mispricing error and wait for the larger one in the presence of the limited capital, our analysis will focus on the role of the binding capital constraint, which tends to deter arbitrage behavior even when mispricing error is huge.

mispricing errors. Our approach helps us to identify whether it is an arbitrage cost or a funding constraint, which mainly limits arbitrage activities. In particular, when the capital constraint is binding, the capital constraint effect will dominate the capital allocation effect, which tends to render overall arbitrage activities deterred, consistent with the slow moving capital hypothesis. Further, the estimates of the regime-dependent noise momentum parameter also provide valuable information about the future uncertainty as the critically small noise momentum (with the future arbitrage opportunity being less attractive) induces arbitrageurs to adopt the full investment in the initial period instead of waiting, and vice versa. To test the empirical validity of theoretical predictions about the nonlinear state-dependent arbitrage activities, we propose to extend the general error-correction model by FCS to the state-dependent Markov-swiching model. Applying this model with three states to the S&P 500 index spot and future markets over the period 1986 - 2015, we find strong evidence in favor of such regime-dependent nonlinear limits to arbitrage. Furthermore, we identify the stress periods when the funding constraint has been binding as the years of 1987, 1998, 2000 and 2007-08, consistent with the slow moving capital period documented in the literature.

In particular, we are able to identify three distinct regimes: a normal market state with the low mispricing error and the low volatility, a transition market state with the medium error and the medium volatility and an extreme market state with high mispricing error and high volatility. The normal state (capturing 53 percent of the sample) shows that the price-fundamental relationship tends to be more efficient, in line with the cost of carry model than is suggested by a linear model. In this state we observe a relatively small mispricing correction together with a relatively large mispricing persistence, suggesting that arbitrageurs refuse to take a large position initially due to the fact that the smaller arbitrage return is unable to cover arbitrage costs. The transition state covers 44 percent of the sample, and tends to stay between the bull and the bear markets with mispricing error and volatility being 10 and 12 basis point higher than in the normal state. Mispricing correction is the highest among three regimes since the larger mispricing errors generally induce arbitrageurs to engage in more corrections. Mispricing persistence is also relatively high and close to that in normal state, suggesting that future uncertainty is still the main concern for arbitrageurs, who tend to maintain the partial investment strategy. Finally, the extreme regime covers only 3 percent of the sample, which correspond to the stress periods including the market crash in 1987, the Asian and Russian financial crisis in 1997-1998, and the latest financial crisis from the mid-2007. Mispricing errors and volatilities are extremely higher than those in normal and transition regimes. This state is characterized by the lowest mispricing correction and the lowest noise momentum, both of which clearly indicate that the funding constraint is binding, consistent with our main theoretical predictions, suggesting that some arbitrageurs are unable to raise external funds or have to withdraw internal funds, exactly when the arbitrage opportunity is at best.

Regarding the question of speed of adjustment, our dynamic multiplier analysis documents that during normal (transition) market states, mispricing is small (large) but longer lived (shorter lived). Specifically, it takes approximately 25 and 17 days for the price change to converge to its long-run value in the normal and transition regimes, respectively. This finding is more or less consistent with the capital allocation effect hypothesis that the larger the mispricing, the stronger is the incentive of each individual arbitrageur to trade based on that information, and the sooner the price will revert towards the fundamental value. More importantly, during extreme market state with substantial mispricing error, the speed of adjustment is much slower (with an overshooting) and it takes approximately 22 days for the price convergence. This finding is also consistent with our main prediction that the capital constraint is binding during extreme state, which renders the capital constraint effect dominate the capital allocation effect. As a result, overall arbitrage activities slow down, consistent with the slow moving capital.

The advantage of using a Markov switching model is the ability to estimate the likelihood of being in a given latent state, which can then be examined for potential linkages to various observable economic factors. Our analysis linking observables to the extreme state suggests that it is a state characterized by larger mispricing, higher cost of capital, higher transaction costs, larger trading volume and a market downturn with higher spot volatility. Overall, it captures a period of considerable market stress. From the perspective of arbitrageurs the extreme state presents a 'cocktail' of good and bad phenomena. On the good side: it is the large mispricing and higher valuation uncertainty – thus presenting arbitrageurs with more profitable opportunities to exploit. However, on the bad side: they will tend to face a higher cost of capital and higher transaction costs that lower net profit.

Our study significantly contributes to the research on limits of arbitrage. First, we extend a theoretical model that consider the limits of arbitrage from both the left and the right hand sides of the arbitrageurs' balanced sheet. Work on limited arbitrage focuses mainly on the left-hand side, showing that market frictions and risk exposure prevent arbitrageurs from forcing immediate price convergence while work on the right-hand side demonstrate the detrimental effect of the funding constraint (Mitchell and Pulvino, 2012). We unify both explanations in one single theoretical framework. An important benefit of such unified framework allows us to develop a directly matching empirical method to identify which channel of limits to arbitrage is at work in a flexible manner. Second, we demonstrate that our proposed empirical strategy can help to enhance our understanding of the time-varying patterns of limited arbitrage. Financial markets have been increasingly represented by financial intermediaries such as mutual funds and hedge funds. This poses a significant challenge to a proper analysis of investment decisions due to principle-agent and information asymmetries, especially when arbitrage opportunities present themselves. In the face of mispricing, fund managers may like to maintain or increase their position while their clients may decide to withdraw their funds. There is a large growing literature that documents the significant impact of fund-flow in creating mispricing such as short-term momentum and long-term reversal (Lamont, 2008, Luo, 2012; Akbas et al., 2015). Moreover, the binding funding constraint is the cause of prolonged mispricing (Mitchell, Pulvino and Pedersen, 2007). In this regard, as we have demonstrated in our empirical application, identifying the conditions and periods under which the funding constraint is binding is, therefore, pivotal to enhance our understanding of limits to arbitrage.

Our paper is also related to the growing literature on the slow moving capital. Mitchell, Pulvino and Pedersen (2007) and Mitchell and Pulvino (2012) conduct case studies during the extreme market circumstance, and provide evidence that the funding constraint is the main reason behind prolonged mispricing. Mancini-Griffoli and Ranaldo (2012) show that the binding funding constraint, not the arbitrage risk, is the main cause of mispricing in covered interest parity during crisis. Dick-Nielson and Rossi (2013) also provide the empirical support for the slow moving capital by examining investors' ownership of convertible bond during the 2005 bond market crash. Acharya et al. (2013) study the impact of liquidity shock on bond returns using the two-regime Markov-switching and document evidence on the flight to liquidity, i.e., when arbitrageurs face the binding capital constraint, they tend to provide liquidity to assets that do not use too much capital (with low volatility or higher reputation). Schuster and Uhrig-Homburg (2015) examine the validity of the nonlinear relationship predicted by Brunnermeier and Pedersen (2009) that the funding liquidity is strongly (marginally) related to the market liquidity when the capital constraint does (not) bind. Employing a two-regime Markov-switching model to the German bond market, they find that the relationship between funding liquidity and market liquidity is significant and strong only during the stressed and low liquidity periods. Both studies provide an empirical support for the slow moving capital hypothesis. There are two fundamental differences between our approach and the above studies: First, we examine the aggregate arbitrage activities through the nonlinear relationship between mispricing and overall speed of adjustment while these studies investigate the impacts of liquidity shocks on the return spreads and the relationship between market and funding liquidities, using the different proxies (??). Next and importantly, our approach is more coherent as our unified can show how asset prices are affected by the slow moving capital both theoretically and empirically whilst the above papers are mainly empirical studies.

The remainder of our paper is organized as follows: In Section 2 we present the theoretical framework which builds on an important extension of Shleifer and Vishny (1997), and develop the main propositions and hypotheses. In Section 3 we develop a general empirical framework designed to best capture the number of predictions derived from our theoretical framework and we link it explicitly to the key hypotheses. In Section 4 we outline a specific empirical application based on the linkage between S&P 500 index futures and spot markets, and we present and discuss our empirical results. Section 5 contains concluding remarks. Mathematical proofs are collected in Appendix.

2. Theoretical Framework

2.1. The model

We begin with an introduction of a range of basic concepts that accords with the SV setup. There are three types of market participants: noise traders, arbitrageurs and investors. Let V be the fundamental value of the asset, which is known to the arbitrageurs but not to the fund investors and noise traders. Denote the noise traders' shocks at period 1 and 2 by S_1 and S_2 , respectively. These shocks represent the extent to which noise traders in aggregate under-value the asset price relative to its fundamental value, V. S_1 is assumed to be known by arbitrageurs, but S_2 is allowed to be stochastic, taking a value of 0 with probability 1 - q or $S_2 = S_2^* > S_1$ with probability q. Noise traders realize the fundamental value, V at period 3 such that the price goes back to fundamental, *i.e.* $P_3 = V$ for certainty. As a result, arbitrageurs do not invest at period 2 when $S_2 = 0$, but they are fully invested when $S_2 = S_2^*$. Similarly, we let F_1 and F_2 be cumulative funding resources under management by arbitrageurs at period 1 and 2. Notice that the funding constraint is always imposed in the SV model, such that $F_1 < S_1$. Thus, arbitrageurs are unable to fully correct the mispricing error in period 1, and they might face a loss when mispricing deepens in period 2.

SV assume that F_1 is exogenously given while F_2 is determined endogenously. Arbitrageurs invest $D_1 = \beta_1 F_1$ at period 1 where the parameter, $0 \le \beta_1 \le 1$ measures their trading strategies. Arbitrageurs' funding is performance-based such that investors may withdraw their funding if arbitrageurs lose. Parameter α captures the sensitivity of arbitrage funds under management at time 2 to its initial performance, suggesting that they would withdraw or increase funds according to performance-based arbitrage. Thus, the supply of funding to arbitrageurs at period 2 becomes:

$$F_2^0 = F_1 \left[1 + a\beta_1 \left(\frac{V}{P_1} - 1 \right) \right] \text{ and } F_2^* = F_1 \left[1 + a\beta_1 \left(\frac{P_2^*}{P_1} - 1 \right) \right]$$
(1)

where F_2^0 and P_2^0 (F_2^* and P_2^*) are the funding and asset price at period 2 if $S_2 = 0$ ($S_2 = S_2^*$). Notice that arbitrageurs lose funds in the latter case, i.e. $F_2^* < F_1$, since $P_2^* < P_1$, and the loss is more than the negative return if α is greater than 1.

We now introduce the stability condition that guarantee $F_2^* > 0$ even when mispricing error intensifies in the second period:

$$a < (V - S_1 + F_1) / (S_2^* - S_1 + F_1).$$
(2)

The stability condition in (2) provide a upper bound to the sensitivity parameter, a, implying that fund investors are not overly sensitive to poor performance such that arbitrageurs won't lose all their fund at period 2.

2.1 The model

Asset price is then determined by the market clearing condition:

$$P_t = V - S_t + \beta_t F_t. \tag{3}$$

Under this model setup arbitrageurs choose their investment strategy so as to maximize their final wealth at t = 3:

$$EW = (1-q)F_2^0 + q\frac{V}{V - S_2^* + F_2^*}F_2^*.$$
(4)

We obtain the following first order condition:

$$\frac{V}{V - S_1 + D_1} - 1 \ge q \left(\frac{V}{V - S_2^* + F_2^*} - 1\right)$$
(5)

from which arbitrageurs can determine an equilibrium trading strategy, $\hat{\beta}_1$.Notice that an inequality holds for $\hat{\beta}_1 = 1$ (the full investment strategy) and the equality for $0 < \hat{\beta}_1 < 1$ (the partial investment strategy). The strategy $\hat{\beta}_1$ reflects the two choices arbitrageurs face in the initial period: to correct the mispricing error right now or to set aside funds for the possible better opportunity in the near future. The left hand side term in (5), $V/(V - S_1 + \beta_1 F_1) - 1$ represents the maximum return of the arbitrageur's investment strategy to correct the mispricing right now, denoted R_{13} . The right hand side, $q(V/(V - S_2^* + F_2^*) - 1)$ is the expected future return from period 2 to 3 if the arbitrageur sets aside funds at period 1 and fully invests at period 2, denoted R_{23} . It is clear from (5) that the full investment strategy is optimal only if $R_{13} > R_{23}$. Suppose that arbitrageurs maintain the full investment strategy but observe that $R_{13} < R_{23}$. In this case it is optimal to reduce investment until $R_{13} = R_{23}^{-8}$. This implies that the two returns should be equalized under the partial-strategy equilibrium.

SV consider an example of the extreme circumstance under which arbitrageurs tend to adopt the full investment strategy due to the huge mispricing error at an initial period. However, when mispricing deepens in period 2, SV highlight that arbitrageurs have to liquidate their holdings, causing the further downward pressure on asset price and resulting in a loss spiral (Brunnermeier and Pedersen, 2009).

We provide a formally proposition for binding funding constraint based on the consequence that

⁸As $\hat{\beta}_1$ falls, R_{13} rises and R_{23} falls, see appendix for proof.

asset fire sale is triggered.

Proposition 1. Consider the model with (1) - (5). Suppose that mispricing error deepens in period 2; namely $S_2^* > S_1$. Then, arbitrageurs will lose their funds in period 2 as $F_2^* < F_1$. Furthermore, in the extreme case where the funding losses are so huge that $F_2^* < D_1$, then arbitrageurs must liquidate their holdings in period 2. We call this the binding funding constraint, which occurs only if the investment strategy $(\hat{\beta}_1)$ exceeds the liquidation bound, $P_1/(a(S_2^* - S_1) + P_1)$.

The liquidation bound is not necessarily equal to 1, but still close to 1, since the difference in noise trader shocks, $S_2^* - S_1$ is much smaller than P_1 . Thus, Proposition 1 provides a more general liquidation bound and completes the SV's conjecture that the full investment is a sufficient but not necessary condition for liquidation. The funding constraint is always imposed under the SV setup, but it is still useful to derive such condition. Proposition 1 is also consistent with the existing studies, documenting that the funding constraint tends to be binding when speculators have large holdings, e.g. Comerton-Forde et al. (2008), Acharya, Shin and Yorulmazer (2009) and Brunnermeier and Pedersen (2009). We will investgate the two equilibrium startegies: the partial investment equilibrium in which arbitrageurs is not likely to face with the binding funding constraint, and the full investment equilibrium under which the funding constraint is binding.

2.2. Mispricing correction and noise momentum

Recently, FCS introduce the concept of the initial mispricing correction and the subsequent mispricing persistence (called "noise momentum") in the framework of the two-period generalized ECM (GECM), in order to provide the better framework for analyzing the asset pricing dynamics and overall arbitrage process. They document the empirical evidence in S&P 500 index-future relationship that the traditional one-period ECM is misleading in the presence of noise momentum. Following FCS, we define the initial mispricing correction by

$$K = \frac{D_1}{S_1} = \frac{\beta_1 F_1}{S_1},\tag{6}$$

where S_1 and F_1 are the mispricing error and arbitrageurs' fund in period 1. K is designed to capture the proportion of mispricing correction achieved by arbitrageurs in period 1. At one extreme, K = 0, implying that there is no error correction. At the other extreme, $\kappa = 1$, indicating that the full and immediate adjustment occurs. FCS also demonstrate that the subsequent mispricing error persistence, called the noise momentum, is another important parameter to understanding mispricing dynamics. Define the noise momentum by

$$\Lambda = \frac{V - P_2}{V - P_1} = \frac{V - P_2}{S_1 - D_1},\tag{7}$$

where V is the fundamental value, P_1 and P_2 are the price in period 1 and 2. Λ is designed to capture the degree of mispricing error persistence. At one extreme, when $V = P_2$ such that none of the error persists or $S_2 = 0$, we have $\Lambda = 0$. Conversely, if all of period 1 pricing error persists, then $P_2 = P_1$ in which case there is 100% error persistence, *i.e.*, $\Lambda = 1$.

To develop further testable hypotheses for K and Λ and match them with empirical analysis, we express the two parameters as the expectation with respect to q in period 1, such that

$$\kappa = E_q(K) = \frac{\hat{\beta}_1 F_1}{S_1}, \ \lambda = E_q(\Lambda) = q \frac{V - P_2^*}{V - P_1} = q \frac{V - P_2^*}{S_1 - D_1},$$
(8)

where $\hat{\beta}_1$ is the equilibrium investment strategy informed by the FOC (Eq. (5)); q the probability that $S_2 = S_2^* > S_1$, $P_2^* < P_1$ is is the price observed in period 2 when $S_2^* > S_1$. The expectation form implies that both κ and λ are below unity⁹ when rational arbitrageurs choose the equilibrium investment strategy to engage in the arbitrage opportunity.

Notice that the initial arbitrage effect, κ is the product of $\hat{\beta}_1$ and the F_1/S_1 , which we call the capital allocation effect and the funding constraint effect, respectively. The capital allocation effect captures arbitrageurs' strategical response to the risky arbitrage opportunity as they are faced with the trade-off between correcting mispricing now and waiting for future arbitrage opportunities. On the other hand, the funding constraint effect measures the arbitrage capital availability relative to the

$$\frac{V-P_1}{P_1} \ge q\left(\frac{V-P_2^*}{P_2^*}\right),$$

$$\lambda = q \frac{V - P_2^*}{V - P_1} \le \frac{P_2^*}{P_1} < 1.$$

⁹To derive $\lambda < 1$, we rewrite the first order condition, Eq. (5) as

where the inequality is likely to hold when the arbitrageurs tend to adopt the full investment strategy. Then it is easily seen that $V = P^* = P^*$

mispricing shocks. The former represents arbitrageurs' willingness to engage in arbitrage activities whereas the latter the capability to conduct arbitrage. When arbitrageurs are partially invested, κ is determined by the interplay between the capital allocation and the funding constraint effects. Under the full investment, however, κ is determined mainly by the funding constraint effect. Therefore, the initial arbitrage effect reflects the two important impediment to arbitrage activities: arbitrage cost and funding constraint.

While the mispricing correction parameter measures the immediate arbitrage effect, the noise momentum will capture the subsequent price recovery. It is highlighted in Duffie (2010) that the pattern of subsequent price recovery also implies the limits to arbitrage borne by arbitrageurs. In the SV model, the probability q of deepening error in the subsequent period reflects the future arbitrage risk. There exist a threshold point q^* , such that when $q < q^*$, the probability that mispricing error deepens is relatively low, arbitrageurs will be more likely to fully invest at time 1. Alternatively, when $q > q^*$ (i.e. the probability of deepening misperceptions is 'critically' high), arbitrageurs will defer some of their investment. However, both q and q^* are unobservable in practice. Eq. (8) shows that λ captures the important information of the probability q, such that higher noise momentum indicates high probability of deepening misperceptions, thus results in higher uncertainty in the pattern of price recovery.

FCS suggest that mispricing correction together with noise momentum provide a better description of price dynamics and the overall speed of adjustment. We now analyse the impacts of the initial mispricing error, S_1 on the arbitrage activities: k and λ , and develop the testable hypotheses for further empirical studies.

First, it is clear that the capital allocation effect, $d\hat{\beta}_1/dS_1 > 0$ is always positive under the partial investment strategy, but becomes nil under the full investment strategy.

Next, we provide the impacts of the mispricing errors on the pricing efficiency in Proposition 2.

Proposition 2. Consider the model with (1) - (5). Then, we derive the impacts of mispricing errors

2.2 Mispricing correction and noise momentum

on pricing efficiency as follows:

$$\frac{\partial P_1}{\partial S_1} < 0, \quad \frac{\partial P_2^*}{\partial S_1} \left\{ \begin{array}{ll} = 0 & \mbox{for } \hat{\beta}_1 = 0 \\ < 0 & \mbox{for } 0 < \hat{\beta}_1 < 1 \\ > 0 & \mbox{for } \hat{\beta}_1 = 1 \end{array} \right\},$$

and $\left| \frac{\partial P_1}{\partial S_1} \right|_{0 < \hat{\beta}_1 < 1} < \left| \frac{\partial P_1}{\partial S_1} \right|_{\hat{\beta}_1 = 1}$.

Proposition 2 shows that a larger initial pricing deviation is caused by a larger current shocks S_1 . It captures the simple intuition that financial constraint arbitrageurs are limited to bear against the mispricing. What's worth noticing is that the increment in mispricing is different with respect to trading strategies. It is easily seen that in partial investment strategy, since arbitrageurs keep investing more of their funds in period 1, it will slow down the speed of increment in mispricing error. But they fail to do so in full investment strategy, where mispricing error grows faster.

We are also interested in investigating the pricing impact of the period 1 mispricing (S_1) on the second pricing efficiency. Proposition 2 shows that the pricing impact of S_1 at t = 2 depends crucially on the investment strategy. In particular, under partial investment strategy, arbitrageurs tend to augment their investment in response to greater mispricing error, thus they might lose more funding when mispricing deepens, which deteriorates the pricing efficiency in the subsequent period. When arbitrageurs are fully invested, however, these rational arbitrageurs ensure that to take the arbitrage opportunity initially must earn at least the same the rate of return, in most case even higher returns (See the inequality relation in Eq. (5)), as to wait for future opportunity. It means that the second period arbitrage opportunity is relatively smaller, as well as the second period loss when mispricing error deepens. As a result, under the full investment strategy, the larger shocks S_1 can lead to more efficient pricing for P_2^* . This additional information will become a crucial input to an analysis of the overall mispricing persistence or the duration of mispricing errors as we shall explain next.

Finally, we provide the main results on the impacts of the mispricing errors on arbitrage activities in Proposition 3.

Proposition 3. Consider the model with (1) - (5). Then, the impacts of mispricing error on arbitrage

2.2 Mispricing correction and noise momentum

activities are derived as follows:

$$\frac{\partial \kappa}{\partial S_1} \left\{ \begin{array}{ll} = 0 & \text{for } \hat{\beta}_1 = 0 \\ > 0 & \text{for } 0 < \hat{\beta}_1 < 1 \\ < 0 & \text{for } \hat{\beta}_1 = 1 \end{array} \right\}, \quad \frac{\partial \lambda}{\partial S_1} < 0.$$

Furthermore, we have:

$$|\frac{\partial\lambda}{\partial S_1}|_{0<\hat{\beta}_1<1}<|\frac{\partial\lambda}{\partial S_1}|_{\hat{\beta}_1=1}$$

Proposition 3 shows arbitrageurs' response to initial noise trader shocks. As mispricing error, S_1 becomes higher, arbitrageurs are willing to invest more funds (higher $\hat{\beta}_1$) to correct under the partial investment strategy, which is consistent with the earlier studies documenting that the larger mispricing error induces the greater mispricing correction¹⁰. However, arbitrageurs' funds also shrink (lower F_1/S_1). Proposition 3 shows that the positive capital allocation effect dominates the negative funding constraint effect under the partial investment equilibrium, i.e. $d\kappa/dS_1 > 0$. On the contrary, under the full investment equilibrium when the funding constraint becomes binding, the initial mispricing correction is determined mainly by the funding availability (i.e. $d\beta_1/dS_1 = 0$.). As S_1 grows, arbitrageurs is forced to face with a deteriorating funding condition and disengage in arbitrage activities, i.e. $d\kappa/dS_1 < 0$.

Proposition 3 clearly demonstrates that the mispricing correction, κ does not always have a positive association with the magnitude of the mispricing errors, as suggested by the earlier studies focusing only on the arbitrage costs, but it also highlights an importance of taking into account the limit to arbitrage related to funding illiquidity. In particular, we find an inverse U-shaped relationship between mispricing error and arbitrage activities, which provides a number of important implications in relation to the slow moving capital. First, when the funding constraint does not bind under the partial investment strategy, conventional arbitrage costs such as transaction cost and arbitrage risk, tend to be the dominant factors of arbitrage, suggesting that arbitrageurs' correction rises mostly with

 $^{^{10}}$ Under the presence of transaction costs and heterogeneous traders, the price dynamics are nonlinear, and will positively adjust according to the size of deviation. Empirical evidence are documented under threshold ECM model (Dwyer et al. (1996) and Martens, Kofman and Vorst (1998)) and smooth transition model (Tse (2001)).

mispricing error during normal periods. Next, in the extreme periods, however, the funding constraint is likely to be binding due to margin/haircut increases and the industry-wide fire sales. Suppose that the magnitude of initial mispricing error is sufficiently large, prompting arbitrageurs to take relatively large positions. The subsequent severe arbitrage funding constraints rather deter or prevent arbitrage activities as our proposition also derives the prediction that an extremely large mispricing error may result in the lower mispricing correction.

Furthermore, Proposition 3 provides the additional important implications about the mispricing impacts on the noise momentum parameter. First, λ is always negatively related to S_1 , rendering an overall speed faster as S_1 rises. However, the negative relationship between λ and S_1 is not monotonic between the partial and the full investment strategy. Consider the partial investment strategy under which arbitrageurs tend to take relatively small positions in period 1 and invest more funds to exploit the more profit opportunities in period 2, the magnitude of λ is large, capturing the high probability (q) of deepening errors. As S_1 rises, the naegative impact of S_1 on λ is relatively small since the increase in S_1 renders both P_1 and P_2 less inefficient (see Proposition 2). On the contrary, under the full investment, λ declines more sharply with S_1 , because the larger mispricing error can lead to a more efficient pricing for P_2^* in this case (see Proposition 2). The smaller noise momentum reflects the relatively lower probability (q) of deepening errors in the second period, suggesting that the future uncertainty is relatively small and the funding liquidity would be the priority concern.

From Proposition 4 we also derive the predictions on overall speed of price adjustment as follows: When mispricing error is relatively small, κ is relatively small and λ relatively large. As a result, the speed of price adjustment becomes relatively slow. As S_1 increases, κ rises and λ marginally falls simultaneously until β_1 reaches the liquidation bound (see Proposition 1). Thus, in this region, the overall speed of adjustment improves with S_1 . However, with extremely large mispricing error and substantially low λ , arbitrageurs will be more likely to fully invest and face with the binding funding constraint. In this situation κ starts to drop due to the slow moving capital while λ keeps falling significantly. In this case, the impact of S_1 on the overall speed of adjustment is uncertain and empirically determined.

Figure 1 illustrates how arbitrageurs' initial mispricing correction and subsequent noise persistence



Figure 1: Strategic response effect on mispricing correction and noise persistence Figure 1 shows the nonlinear impact of mispricing errors on mispricing correction (Left) and noise momenutm (Right). The left figure also plots the capital allocation effect $(\hat{\beta}_1)$ and the funding constraint effect (F_1/S_1) , while the right figure provides additional plots of prices in period 1 and 2. Both figures highlights the nonlinearity of the impact under different investment strategy.

innovates with respect to mispricing error. The impact of mispricing error is nonlinear based on the partial and full investment equilibrium. In particular, the initial mispricing correction shows an inverse U-shape relation with respect to mispricing error. Under the partial investment equilibrium, both the capital allocation and funding constraint effect contribute to the mispricing correction, while only the negative funding constraint effect remains under full investment equilibrium. Noise momentum remains stable under partial investment equilibrium since prices in period 1 and 2 drops simultaneously, but sharply drops under full investment equilibrium due the the pricing efficiency in P_2 .

2.3. Hypotheses Developments

We consider the three regimes according to magnitude of mispricing error (small, medium and large), and denote κ^r and λ^r with $r \in (s, m, l)$. In the SV model there is a key parameter, a threshold point for q, denoted q^* . If $q > q^*$ (i.e. the probability of mispricing deepening is relatively high), arbitrageurs will defer some of their investment, expecting that the time 2 price will be further away from fundamentals. Alternatively, when $q < q^*$ (i.e., the probability that mispricing deepens is 'critically' low), arbitrageurs will be more likely to fully invest in period 1. Specifically, the concept of noise momentum This logic can be captured exactly by the concept of noise momentum parameter, λ under our setup, which delivers an extra, rich dimension into understanding the price discovery process. Thus, the full (partial) investment strategy is characterized by the critically low (high) value of λ .

We summrise three main predictions derived from our theoretical framework as follows:

Hypothesis 1: The U-shaped initial arbitrage activities. (i) $\kappa^m > \kappa^s$; an initial mispricing correction rises with the size of mispricing error. (ii) $\kappa^l < \kappa^m$; In the presence of the binding funding constraint, further rise in mispricing error will induce the slower mispricing correction.¹¹

Hypothesis 2: The regime-dependent mispricing persistence. (i) λ^s is relatively large, suggesting that mispricing tends to persist in period 2 as the initial mipricing error is too small. (ii) $\lambda^m \approx \lambda^s$; the difference between λ^s and λ^m is relatively negligible under the partial investment strategy. (iii) If β_1 is greater than the liquidation bound, then λ^l will be significantly smaller than λ^s and λ^m .

Hypothesis 3: The overall speed of adjustment (SOA) tends to be faster with the larger mispricing error under the partial investment strategy when the funding constraint is not binding. On the contrary the impact of mispricing error on SOA will be uncertain when the funding constraint is binding.

We now discuss the mispricing dynamics in terms of the interplay between κ and λ with respect to mispricing error, S_1 across three regimes. State l is characterized by relatively lower mispricing error and relatively high λ . As we move from state s to m with the larger S_1 , arbitrageurs tend to maintain the partial investment strategy in which case we expect that κ rises significantly while λ falls very slightly. As a result, the overall speed of adjustment tends to be faster. Next, consider the case when we move from state m to l, where state l is characterized by extremely large mispricing error and substantially low λ . In this situation, arbitrageurs are more likely to not only adopt the full investment strategy but they are also faced with the binding funding constraint. We then expect that both κ and λ fall with S_1 . Hence, the overall speed of adjustment may increase or decrease, though it is more likely to be slower especially when the funding constraint.

¹¹For completeness we may also consider the case when $\kappa^l > \kappa^m$ in which we have: $(\kappa^l - \kappa^m) / (S_1^l - S_1^m) < (\kappa^m - \kappa^s) / (S_1^m - S_1^s)$, implying that the slope is flatter as we move from state m to l.

3. Nonlinear approach: a Markov-Switching Generalized Error Correction Model

3.1. Basic GECM set-up

Building upon the basic analytic foundation provided in in the previous section, we develop a two-period regime-switching generalized error correction model to explicitly test the validity of the main hypotheses regarding the limits of arbitrage and its impact on the asset pricing dynamics. We first follow FCS for the two-period GECM, capturing the initial mispricing correction κ and noise momentum λ in an arbitrage process:

$$\Delta f_t = \kappa z_{t-1} + \lambda \left(1 + \kappa\right) z_{t-2} + \delta \Delta f_{it}^* + \gamma \Delta f_{it-1} + \mu_t, \ \mu_t \sim iid(0, \ \sigma_\mu^2) \tag{9}$$

where where f_t is the (observed) market price; f_t^* is the fundamental value for the asset; $z_t = f_t - f_t^*$, the pricing error; Δ is the difference operator and $\kappa, \lambda, \delta, \gamma$ are the parameters of interest. Equation (9) is called a generalized error correction model (GECM).

The GECM simultaneously captures the (complex) dynamics of the two-period interaction between arbitrageurs and noise traders. The distinguishing feature of the GECM is that we can accommodate 'noise momentum' effects through the term $\lambda (1 + \kappa) z_{t-2}$, with the parameter λ measuring the strength of noise momentum and $(1 + \kappa) z_{t-2}$ represents the unarbitraged component of the pricing errors from the previous period¹².

 δ measures the degree of the over- or under-reaction with respect to the contemporaneous fundamental changes, while γ presents the short-run momentum effect. Generally, $\omega = \delta - 1$ is likely to be non-zero unless the market is perfectly efficient. The sign of ω determines the direction of the price reaction to fundamental impact. If ω is positive (i.e. the pricing error innovation, ε_t is positively correlated with the fundamental valuation innovation, e_t), then the price overreacts on impact to the fundamentals irrespective of the signs of the innovation. On the other hand, a negative ω implies that the price underreacts on impact. The sign of γ ($\gamma = -\omega\pi$) is generally ambiguous since it depends on the product of the correlation coefficient, ω , and the feedback trading coefficient, π . As such this is an empirical issue. The last component, u_t is an idiosyncratic error term with zero mean and finite

 $^{^{12}}$ See Appendix B for a derivation

variance, σ_u^2 . Notice that the total variance of the mispricing innovation, ε_t , is obtained simply as the sum of the variance of fundamental innovation, e_t , and the variance of idiosyncratic error, u_t .

Finally, in the context of equation (9) we see that the magnitude and amplitude of the initial pricing errors are determined mainly by parameters ω , γ and σ_u^2 , while the overall speed of convergence to equilibrium is determined jointly by k and λ , namely $(k + \lambda(1 + \kappa))$. Importantly, positive noise momentum would make the pricing errors more persistent.

3.2. Regime Switching Analysis

One of our key theoretical predictions is that arbitrage activities depends on the level of mispricing error. By construction the linear model cannot test our hypotheses because it imposes (potentially invalid) symmetry restrictions and, thus, is likely to yield misleading results. Accordingly, in our empirical application we choose to embed the GECM within the popular Markov switching model. In particular, we consider a three-state (regime) setup which is compatible with our hypotheses of three different magnitude of mispricing error (small, medium and large):

$$\Delta f_{it} = \kappa_{R_j} z_{it-1} + \lambda_{R_j}^* z_{it-2} + \delta_{R_j} \Delta f_{it}^* + \gamma_{R_j} \Delta f_{it-1} + \mu_{itR_j}, \tag{10}$$

where R_j is a scalar geometric ergodic Markov chain with a 3-dimensional state space, having associated transition probabilities $\rho_{ij} = Pr(R_i \mid R_j)$, for i, j = 1, 2, 3. The model is designed to provide further insights into the asymmetric price discovery process in a flexible manner.

3.3. Price Discovery Analysis with Dynamic Multipliers

While the estimation results obtained from equation (10) will provide useful insights, we are also interested in uncovering the dynamic price discovery process in a more intuitive manner. To this end, we present such information through a dynamic multiplier analysis that evaluates the impact of a one unit change in the fundamental value on the price change. In the long-run, the dynamic multiplier ought to converge to the equilibrium value of unity, implying that one percent change in the fundamentals will result in one percent change in the market price. As discussed above, however, we expect to observe over- or under-reaction of the market price to fundamental changes in the short run due to market frictions. The combined impact on the dynamic price discovery process will be completely represented by a dynamic multiplier analysis¹³. The ability of the dynamic multipliers to illuminate the trajectory between initial equilibrium, short-run disequilibrium following a shock to the fundamentals, and a long-run equilibrium makes them a powerful tool for the analysis of the price-fundamental dynamics.

4. Empirical Application to Index Future

4.1. Empirical Model

To apply our model, a long-run price-fundamentals relationship is required to be well-defined, although the long-run coefficient need not take a value of unity. In what follows we focus on examining the short-term dynamics in the index futures market. The cost of carry model is based on the exclusion of arbitrage and assumes that the risk-free rate and dividend yield are given. Specifically, we expect the following relationship to hold in equilibrium:

$$F_{t,T}^* = S_t \times exp\left[\left(r_t - q_t\right)\tau_t\right] \tag{11}$$

where $F_{t,T}^*$ is "fair value" of a futures contract maturing at time T; S_t is the current value of the spot index; r_t is the risk-free interest rate, and q_t is the dividend yield on the index; τ_t is the the time to maturity.

Assuming that the risk-free rate and dividend yield are deterministic, and will share the same stochastic trend. The futures and spot prices are cointegrated under general conditions (Ghosh (1993), Wahab and Lashgari (1993), Brenner and Kroner (1995)). Significant deviations from the prediction of cost of carry can reflect violations of the model's assumptions. The key assumption underlying the cost of carry model is that market participants take advantage of arbitrage opportunities as soon as they occur (Hull (2008, Ch. 3)). However, empirically, only partial adjustment is found (e.g., Stoll and Whaley (1986), MacKinlay and Ramaswamy (1988)). The MS-GECM developed above provides an ideal tool for helping us understand the rich dynamics behind the pricing error generated and the

 $^{^{13}}$ see Appendix C for a derivation

associated convergence processes. Its empirical counterpart is given by:

$$\Delta f_t = \alpha_{R_j} + \kappa_{R_j} \hat{z}_{t-1} + \lambda_{R_j}^* \hat{z}_{t-2} + \delta_{R_j} \Delta f_t^* + \gamma_{R_j} \Delta f_{t-1} + \mu_{tR_j}, \ \mu_{tR_j} \sim iid\left(0, \ \Sigma_{R_j}\right), \tag{12}$$

where f_t is the natural log of the futures contract price; f_t^* is the natural log of the fundamental value implied by the cost of carry model $S_t + [(r_t - q_t) \tau_t]$; S_t is the natural log of the spot index price; r_t is the risk-free rate; q_t is the dividend yield on the index; and $\{\alpha_{R_j}, \kappa_{R_j}, \lambda_{R_j}^*, \delta_{R_j}, \gamma_{R_j}\}$ are state-dependent coefficients with the covariance of the residuals (\sum_{R_j}) taking different values across the two states. The pricing error, \hat{z}_t , is estimated from the long-run equation:

$$f_t = \mu + \theta f_t^* + z_t \tag{13}$$

Comparing equation Eq. (12) with (10), we allow for both an intercept and a non-unity long-run coefficient for general purposes. According to the cost of carry model, the theoretical value of θ equals 1 which (as will be seen below) is strongly supported by our empirical analysis.

The regime-specific noise momentum coefficients, λ_{R_j} can then be obtained from: $\lambda_{R_j}^* = \lambda_{R_j} (1 + \kappa_{R_j})$ for j = 1, 2, 3. We estimate a three-regime model. The transition matrix is given by:

$$\begin{bmatrix} P_{11} & P_{21} & P_{31} \\ P_{12} & P_{22} & P_{32} \\ P_{13} & P_{23} & P_{33} \end{bmatrix}$$

The parameter P_{ij} is the probability that State *i* is followed by State *j*, i.e., the transition probability from State *i* to State *j*.

4.2. Data

Our empirical approach is different from previous studies in two primary respects. First, we use daily data instead of intraday transactions data. Most previous studies on the linkage between spot and futures indices have focused on very short-term dynamics using high-frequency intraday data (e.g. Dwyer et al. (1996) – though a notable exception is Sarno and Valente (2000) who use daily data). While higher frequency data contain more information, they are also more prone to excessive noise. Over a very short window as is typically used in high-frequency datasets, the only input in the cost of carry model expected to change is the spot price. For the interest rate and dividend yield, one would expect negligible changes over a narrow intraday window. Our study complements previous works by examining a longer window in which meaningful changes in interest rates and dividend yields would be captured.

Second, one of the constraints in using high-frequency data is the length of the sample period: a short sample period of 3 to 6 months is commonly used. Using longer periods of high-frequency data would be computationally expensive. Using lower frequency data allows us to study the dynamics over a longer sample period. Hence, the current study uses a time period covering the complete lifespan of the daily S&P 500 index spot and futures contract between June 1986 and December 2015. ¹⁴

Our proxies for the risk-free interest rate is the US three-month T-bill rate. Divided yields on the indices are also collected. All data are sourced from DataStream. A continuous series of the nearest term futures contracts is constructed by DataStream. The series switch to the next nearest contract on the first day of the expiry month for the nearest term contract. We use a full set of daily price information of every contract to ensure correct matching of the date to maturity with the continued futures price series. Table 1 reports the descriptive statistics for all variables (measured in percentage terms).

As expected, the movements of the spot and futures prices closely mimic each other. The average price changes are of the same magnitude while the volatilities are slightly higher in the futures contracts. The average basis (the log difference between futures and spot prices) is 0.24 percent. After applying the cost of carry model, the difference between the futures price and the fair estimate $(f - f^*)$ is about a quarter of the basis (0.05 percent).

¹⁴We do not use the early data from 1982 to 1986, since the estimated mispricing errors are more than doubled on average during this early period comparing to the period of 1986-2015. The index future contracts are first introduced in 1982, where the market has lack of index arbitrage and higher transaction costs. Thus larger mispricing errors occur. Errors then become more stable after 1986, and fluctuates with major market events. See also Figure ?? in Appendix for a plot of the moving-average mispricing error through time.

	Mean	Median	Minimum	Maximum	Std Dev
Δs	0.028	0.058	-22.833	10.957	1.169
Δf	0.028	0.062	-33.700	17.749	1.262
f-s	0.243	0.196	-11.027	2.958	0.541
$f - f^*$	0.055	0.056	-11.451	2.767	0.294
r	3.380	3.790	0.000	9.100	2.493
q	2.275	2.080	1.070	4.100	0.718

Table 1: Basic Descriptive Statistics

Table 1 reports the the descriptive statistics for all variables. The sample used is the daily series of the S&P 500 index and its futures contract covering the period June 4, 1986 to December 3, 2015. There are a total of 7,442 observations. $\Delta s \ (\Delta f \)$ is the first difference of log spot (futures) price. $f_{t,T}^* = s_t + (r_t - q_t) \tau_t$ where $s_t = \ln (S_t)$ and r_t is the annualized risk-free (3 month T-bill) interest rate on an investment for the period , and q_t is the annualized dividend yield on the index. All numbers are recorded in percentage point terms.

5. Empirical results

5.1. Characterizing the Three Regimes

The estimation results for the MS-GECM in equation (12) are reported in Table 2 and the smoothed regime probability is plotted in Figure 2. We see that the three estimated regimes contrast each other in three distinctive ways. First, they differentiate periods of low volatility from those of medium and (extremely) high volatility. The variance (Σ) of the medium volatility state (State 2) is twice as large as that in the low volatility state (State 1). The extremely high volatility state (State 3) is around 5 times greater than that in State 2. Second, we find that the mispricing error ($|z_{t-1}|$) is the smallest in State 1 with a magnitude of 0.103, whereas they slight grow to 0.207 in State 2. Notice that the average mispricing error estimated from equation (4) throughout the sample period is 0.161. However, they are still relatively small comparing to that in State 3, which is substantially large (0.774) in absolute term.

Third, these three regimes do not have a balanced incidence over time. Panel C of Table 2 shows that State 1 is one of the dominants with an ergodic probability of 50%. Moreover, the transition probabilities confirm that State 1 is one of the dominants with an ergodic probability of 53%. Moreover, the transition probabilities confirm that State 1 is very persistent with a transition probability of 98%. According to Figure 2, the state coincides with the three major bull markets during the sample period, 1991-1996, 2003-2006 and 2012-2015. We classify State 1 as the normal regime, while State 2 is another

Panel A: Estimation Results										
	State 1		State 2		State 3		State 2-1		State 3-2	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
α	-0.005***	-2.51	0.016^{***}	3.34	0.058	0.66	0.022***	4.15	0.042	0.64
δ	0.991^{***}	332.	1.024^{***}	259.	1.142^{***}	48.3	0.031***	6.36	0.118^{***}	4.93
γ	-0.004	-1.38	0.015^{***}	4.18	0.092***	3.58	0.018***	4.11	0.078^{***}	2.98
κ	-0.699***	-36.7	-0.819***	-41.5	-0.596***	-6.56	-0.120***	-4.38	0.224^{**}	2.41
λ^*	0.113^{***}	12.8	0.167^{***}	9.24	0.113	1.29	-0.079***	-3.00	-0.053	-0.59
Σ	0.113^{***}	59.1	0.239^{***}	43.2	1.077^{***}	13.2	0.126***	21.4	0.838***	10.2
SOA	0.453^{***}	19.1	0.652^{***}	27.6	0.482^{***}	8.73	-0.199***	-5.98	0.171^{***}	2.84
$\mid z_{t-1} \mid$	0.103		0.207		0.774					
Panel B. Recovered Coefficients										
ω	-0.008**	-2.67	0.024^{***}	5.94	0.142^{***}	5.99	0.032***	6.35	0.118^{***}	4.93
π	-0.474	-1.24	-0.630***	-3.42	-0.652***	-3.22	-0.156	-0.36	-0.022	-0.09
λ	0.817^{***}	8.96	0.922***	5.86	0.279	1.03	0.107	0.58	-0.647^{***}	-4.92
Panel C. Matrix of Markovian Transition Probabilities										
	State 1_{t-1}		State 2_{t-1}		State 3_{t-1}					
$State1_t$	0.980		0.024		0.000					

	relation r	b care $= l = 1$	otatto ot=1
$State1_t$	0.980	0.024	0.000
$State2_t$	0.019	0.968	0.143
$State3_t$	0.000	0.008	0.857
Ergodic	0.531	0.443	0.024

Table 2: Estimation of the Markov Switching Generalized Error Correction Model

Table 2 reports the estimation of the Markov Switching Generalized Error Correction Model. The sample used is the daily series of the S&P 500 index and its futures contract covering the period June 4, 1986 to December 3, 2015. There are a total of 7,442 observations, of which 4,070, 3,222 and 148 fall into State 1, State 2 and State 3. Specifically, Panel A reports the estimation results for:

$$\Delta f_t = \alpha_{R_i} + \kappa_{R_i} \hat{z}_{t-1} + \lambda_{R_i}^* \hat{z}_{t-2} + \delta_{R_i} \Delta f_t^* + \gamma_{R_i} \Delta f_{t-1} + \mu_{tR_i}$$

where \hat{z}_t is estimated from equation 13, $\{\alpha_{R_j}, \delta_{R_j}, \gamma_{R_j}, \kappa_{R_j}, \lambda_{R_j}^*\}$ are state-dependent coefficients with the covariance of the residuals (Σ_{R_j}) taking different values across the two states. Panel B reports the recovered coefficients. Specifically, ω_{R_j} is recovered by $\delta_{R_j} - 1$; $\pi_{R_j} = -\gamma_{R_j}/\omega_{R_j}$ and $\lambda_{R_j} = \lambda_{R_j}^*/(1 + \kappa_{R_j})$. The final two columns in Panels A and B report the difference in estimated coefficients and associated t-statistic between States 1 and 2. For non-linear combinations of the coefficients, a delta method is applied to obtain the variance of the recovered coefficients and their differences. Panel C reports the transition and ergodic probabilities. ***, ** and * indicate significance at 1%, 5% and 10% levels, respectively.



Figure 2: The smoothed regime probability

Figure 2 plots the smoothed regime probability of being in State 1, 2 and 3. State 1 consists of the years 1991-1996, 2003-2007 and 2012-2015, State 2 consists of the years 1986-1991 and 1996-2002, while State 3 are found in the years of 1987-1991, 1997, 1998 and also 2007-2008.

dominant, labeled as the transition regime and characterized by slightly higher pricing errors and high volatility. Difference in pricing error indicates that arbitrageurs' trading strategy might change in between the two regimes. This State captures the transition period from bull to bear market state, especially the period around 1986-1991 and 1996-2002. State 3 is characterized as the extreme regime, with an ergodic probability of only 2.6% but also extremely high pricing error and volatility. During this state, arbitrageurs might face binding capital constraint due the existence of extremely high mispricing error. The regime tends to reside in the early part of the sample, with specific concentrations around the years of 1987-1988, 1997-1998, and also 2007-2008. Given that these periods coincide with the 1987 stock market crash, the Asian and Russian financial crises in 1997 and 1998, and the global financial crisis in 2007, we are encouraged that the Markov regimes have meaningfully distinguished between major stressful (large mispricing error), the transition period (medium mispricing error) and the calm periods (small mispricing) witnessed over our sample period.

5.2. Main Results

In Table 2, various findings are worthy of special focus – primarily, those linking to our key hypotheses. Due to the various mispricing erro across the three regimes, we expect that the interplay of κ and λ should follow Hypothesis 1 and 2. We see in the table, that the estimated error correction

parameters, κ_{R_j} , are negative and significantly less than unity in both regimes, supporting the hypothesis in the work of FCS. Based on the point estimates, more than half of the prior period pricing error is corrected 69.9% in State 1, which is the medium among the regimes. The relatively small correction in State 1 is in line with its smallest pricing error (0.103), where arbitrageurs are less tempting to exploit the arbitrage profit. Arbitrageurs in State 2 face a relatively higher mispricing error (0.207), and are now able to correct nearly 82% of the pricing error. The difference between State 2 and 1 is a significant 12% of correction with the mispricing error doubled. It indicates the positive reaction of arbitrageurs in response to the pricing error, when funding constraint is not binding. Although the mispricing error is massively higher in the State 3, it appears that arbitrageurs' activities are limited with a significant smaller mispricing correction. Only 59.5% of the pricing errors are corrected initially by arbitrageurs, which is a strong implication of binding capital constraint and limits of arbitrage. The initial correction drops around 22.4%, comparing to the transition state. Therefore, under the extreme market circumstance, arbitrageurs fail to bring the price back to its fundamental value due to their funding condition. Overall, these results support our prediction of κ_{R_j} in Hypothesis 1.

Our primary focus is on the estimated noise momentum coefficients (λ_{R_j}). Notably, we find that the effect of noise momentum is significant among the normal and transition regimes – the estimated coefficients are positive as predicted (Panel B). This finding provides strong support for FCS noise momentum hypothesis which highlights that λ is an important parameter in characterizing the overall speed of adjustment process. Consistent with FCS, λ_{R_j} in all regimes are less than unity, indicating the dominating position of rationael arbitrageurs in the market. Panel B of Table 2 shows that there is approximately 82%, 92% and 28 %, in normal, transition and extreme state respectively, of the unarbitraged error coming from the previous period being reflected in the current trading period. This differential is consistent with Hypothesis 2 – indeed, the state-dependent difference for parameter λ shown in Panel B of Table 2 indicates formal support of this hypothesis i.e. the State 1 estimate is indifferent from that in Sate 2, while, at the 1% level of significance, the State 3 estimate substantially exceeds its State 2 counterpart.

Our findings qualitatively suggest a combination of nonlinear price dynamics, and gives statistical support for the interplay of κ and λ at conventional levels. Moreover, combining both the error

correction and noise momentum coefficients, we find that overall speeds of adjustment (given by the quantity $\kappa + \lambda(1+\kappa)$) are 45.3%, 65.2% and 48.2%, respectively, for the normal, transition and extreme regimes. We find that the overall speed of adjustment in the transition state is much faster that in normal state. Notably, the results suggest that the κ asymmetry, i.e. the increment of mispricing correction, is the main driver behind the increased adjustment speed between these two states, while the λ asymmetry has only a marginal effect. Comparing between transition and extreme states, the slower speed of adjustment is mainly derived by the interplay between κ and λ . The reduction of mispricing correction overpowers the reduction in noise persistence, and results in a significant slower speed of adjustment. The outcomes again suggest the limits of arbitrage during the extreme market state, and support our Hypothesis 3.

The empirical results suggest that the aggregate arbitrage behavior we captured in the Markovswitching Error Correction model provide a strong support of slow moving capital during the stress period. Consistent with the literature on slow moving capital, the regimes with binding capital constraint coincide with the 1987 market crash and 1998 collaspe of Long Term Capital Management, which are found in Mitchell, Pulvino and Pedersen (2007), as well as the 2007 global crisis, which is found in Schuster and Uhrig-Homburg (2015).

We also observe a range of notable findings for the other parameters in our model. First, the intercepts (α_{R_j}) in state 1 and 2 are statistically significant. A large positive intercept is found in State 2 and 3, while it is negative but negligible in State 1 – the positive sign indicating a regime, in which the futures price is more bullish than the spot price, other things equal. As such, this suggests that the transition and extreme regime coincides with periods in which the futures market is more bullish relative to the spot market. On average, during these two regimes, there are more than 16 and 58 basis point returns daily in the futures market, respectively, regardless of the spot market movement. However, such returns are accompanied by larger risks as reflected by the variance of the futures return. Such a pricing difference between the two markets under different market conditions is not directly considered in previous theories, and thus presents a new result to the literature. $w_{R_j} = \delta_{R_j} - 1$

Second, Table 2 shows that the contemporaneous market reaction coefficient, $w_{R_j} = \delta_{R_j} - 1$, is statistically different from zero, but small and negative in the normal regime, while it is highly



Figure 3: Dynamic Multiplier Analysis Figure 3 plots the accumulated dynamic multipliers for the Markov Switching GECM (as reported in Table 2). The dynamic multipliers are constructed as outlined in Appendix C.

significant and positive in the transition and extreme regime. In the case of the latter, it suggests that for a one percentage point change (up/down) in the fundamental value there will be 0.24 and 1.42 percentage point price movement in the futures market in the same direction (i.e. a 2.4 and 14 basis point overreaction respectively). Third, Table 2 also shows that while in the normal market regime there is no feedback trading (i.e. we cannot reject $\pi = 0$), there is large negative (and significant) feedback trading in extreme market conditions. Such negative feedback trading is consistent with lower noise momentum in the market under this regime.

5.3. Dynamic Multiplier Analysis

Dynamic multipliers evaluate the impact of a unit change in the fundamental value (the fair value estimated by the cost of carry model) on the price of the futures contract. Our main interest lies in discovering the dynamic reaction of futures prices with respect to such changes, before converging to the long-run equilibrium level. Figure (3) plots the accumulated dynamic multipliers. Panels A, B and C, respectively, show the multipliers relating to the normal, transition and extreme states of the nonlinear model in Table 2.

Under the normal market state, although the initial underreaction is the lowest, the duration of

the error convergence process is the longest. This is mainly caused by a relatively low level of initial arbitrage price pressure and a far higher level of noise momentum trades. These two effects outweigh the increased positive feedback trading effect in this regime. Thus in normal market conditions the divergence from the cost of carry model is very low but long lived. On the other hand, under transition state, although the initial overreaction is higher, the duration of the error convergence process is the shortest. Our findings provide direct support for the proposition of Abreu and Brunnermeier (2002) that very large mispricing cannot be sustained for very long, while small mispricing can persist for some time due to the differential incentives of arbitrageurs when facing different levels of mispricing. However, that is only half of the story. In the extreme market circumstance with the highest overreaction, the duration of error convergence process is rather similar to that in normal states, due to a difference reason. Although the level of noise momentum is the lowest, the level of initial price pressure also reaches the lowest since funding constraint is binding.

Furthermore, our evidence documents new insights into the cause(s) of a prolonged error correction process which is often explained by the presence of transaction costs in the previous literature (Sercu et al. (1995), Panos et al. (1997) and Roll et al. (2007)). We show that arbitrageurs' anticipation of noise momentum (arbitrage risk) can play an important role in delaying price convergence. We find that in extreme market conditions, arbitrageurs expect mispricing is more likely to be reduced than enlarged which is supported by the empirical evidence of significant negative feedback trading during the period. In other words, noise momentum is lower in extreme market conditions.

5.4. Linking the Hidden States to Obervables (rewrite when update)

TBC

6. Conclusions

We develop a unified approach, which can provide the main theoretical predictions on the effects of both the capital allocation and the funding constraint on limits to arbitarge. Building on the seminal work by Shleifer and Vishny (1997), we analyze the nonlinear impacts of mispricing on arbitrage activities. To replicate the situation where arbitrageurs attempt to exploit the price discrepancy while simultaneously facing agency problems, market frictions and funding constraints, we allow arbitrageurs to face a trade-off between making investments now and waiting for larger arbitrage opportunities in the subsequent period. To capture such multi-period arbitrage activities, we follow the framework advanced by FCS and investigate the impacts of mispricing on both the initial mispricing correction and the subsequent mispricing persistence, where the latter is designed to measure persistence of the uncorrected pricing errors and affect the ex ante capital allocation strategy of arbitrageurs.

We investigate these issues under two alternative paths to equilibrium. Under the partial-investment strategy characterized by the substantially high future mispricing persistence, the funding constraint is not binding and thus the larger mispricing would induce faster arbitrage activities. Next, consider the extreme market state characterized by extremely large mispricing error and critically low future mispricing persistence. Arbitrageurs are then more likely to adopt the full investment strategy but also face with the severe and persistent funding constraints. This tends to deter arbitrageurs' initial mispricing correction and slow the overall speed of adjustment. In sum, our theory suggests that overall arbitrage activities do not rise linearly with mispricing errors. Rather, their relationship tends to be the regime-dependent, and overall arbitrage activities display an inverse U-shape against the magnitude and volatility of mispricing errors. Our approach helps us to identify whether it is an arbitrage cost or a funding constraint, which mainly limits arbitrage activities. In particular, when the funding constraint is binding, the funding constraint effect dominates the capital allocation effect, rendering overall arbitrage activities deterred, consistent with the slow moving capital hypothesis (Mitchell, Pulvino and Pedersen, 2007; Brunnermeier and Pedersen, 2009). Further, the estimates of the regime-dependent noise momentum parameter also provide valuable information about the future uncertainty as the critically small noise momentum (with the future arbitrage opportunity being less attractive) induces arbitrageurs to adopt the full investment in the initial period instead of waiting, and vice versa.

To test the empirical validity of theoretical predictions, we extend the general error-correction model by FCS to the state-dependent Markov-swiching model. Applying this model with three states to the S&P 500 index spot and future markets over the period 1986 - 2015, we find strong evidence in favor of regime-dependent nonlinear limits to arbitrage. Furthermore, we are able to identify the stress periods when the funding constraint has been binding as the years of 1987, 1998, 2000 and 2007-08, consistent with the slow moving capital period documented in the literature.

Our paper is closely related and significantly contributes to the growing literature on limits of arbitrage and the slow moving capital theory (Mitchell, Pulvino and Pedersen, 2007; Mitchell and Pulvino, 2012; Acharya et al., 2013; Schuster and Uhrig-Homburg, 2015). First, we extend a theoretical model that consider the limits of arbitrage from both the left and the right hand sides of the arbitrageurs' balanced sheet. Our unified framework allows us to develop a directly matching empirical method to identify which channel of limits to arbitrage is at work in a flexible manner. Next, we demonstrate that our empirical strategy can help to enhance our understanding of how the capital allocation and the funding constraint interact to derive regime-dependent nonlinear limits to arbitrage. Therefore, our approach is more coherent and direct as we are able to analyse an implication of how asset prices are affected by the slow moving capital both theoretically and empirically.

The potential applications of this approach go well beyond that developed in the current paper. For example, our approach can be applied to explore the short-term dynamics associated with fundamental long-run cointegrating relationships (e.g., price-dividend relationship) and the pricing dynamics between segmented markets for single assets (for example, cross-listing and commodity contracts in different markets). Another important extension is to develop the unified framework for analysing the liquidity spiral (i.e. the market illiquidity and the funding illiquidity can be reinforcing during the stressed period) predicted by Brunnermeier and Pedersen (2009), theoretically and empirically. We commend these and other meaningful extensions to future research agenda.

AppendixA. Proofs

This Appendix provides additional details about the our discussion of the SV model in Section 2. We first provide three implications from the SV model which will be used in developing the proof of the main propositions.

Fact 4. Consider the SV model with (1) - (5). For fixed investment strategy β_1 , then F_2^* increases as S_1 rises.

Proof. From (1) we have

$$F_{2}^{*} = F_{1} \left[1 + \frac{a\beta_{1}S_{1} + a\beta_{1}F_{1}(1 - \beta_{1}) - a\beta_{1}S_{2}^{*}}{V - S_{1} - (a - 1)\beta_{1}F_{1}} \right]$$

$$= F_{1} + F_{1} \left[\frac{a\beta_{1}F_{1}(1 - \beta_{1}) - a\beta_{1}(S_{2}^{*} - S_{1})}{V - S_{1} - (a - 1)\beta_{1}F_{1}} \right]$$

Taking the first differentiation with respect to S_1 , we have

$$\frac{\partial F_2^*}{\partial S_1} = a\beta_1 F_1 \frac{V - S_2^* + (1 - a\beta_1) F_1}{\left(V - S_1 - (a - 1)\beta_1 F_1\right)^2},$$

Under the stability condition, $V > aS_2^* + (a-1)(F_1 - S_1)$, it is easily seen that

$$V - S_2^* > (a - 1) \left(S_2^* + F_1 - S_1 \right) > (a - 1) F_1 \ge (a\beta_1 - 1) F_1$$

QED.

Fact 5. Consider the SV model with (1) - (5). Suppose that mispricing error deepens at period 2; namely $S_2^* > S_1$. In this case arbitrageurs lose their funds as $P_2^* < P_1$ and $F_2^* < F_1$.

Proof. Consider the partially invested case. Then, we obtain the equality of the first order condition (5) at $0 < \beta_1^* < 1$, which can be expressed as

$$\frac{V}{P_1} - 1 = q\left(\frac{V}{P_2^*} - 1\right)$$

For 0 < q < 1, we must have $\frac{V}{P_1} < \frac{V}{P_2^*}$, which proves that $P_2^* < P_1$. Using the definition of F_2^* in (1), it is easily seen that $F_2^* < F_1$.

Next, consider the full investment case where we have $F_2^* = F_1 \left(1 - \frac{a(S_2^* - S_1)}{P_1 - aF_1}\right)$ with $\beta_1 = 1$. When $P_2^* < P_1$, then $F_2^* < F_1$ from (1). Suppose that $P_2^* = P_1$. Then, we must have: $F_2^* = F_1$ from (1). Using F_2^* in full investment case, this implies that $S_2^* - S_1 = 0$, but this violates the mispricing deepening assumption, $S_2^* > S_1$. Next, suppose that $P_2^* > P_1$, and thus $F_2^* > F_1$ from (1). Again using F_2^* , this implies that $P_1 - aF_1 < 0$, but this violates the original SV condition, $P_1 - aF_1 > 0$. Hence, under the full investment strategy, we can only have: $P_2^* < P_1$ and $F_2^* < F_1$.

Fact 6. Consider the SV model with (1) - (5). Suppose that arbitrageurs adopt the partial investment strategy $\left(0 < \hat{\beta}_1 < 1\right)$. Then, the sign of $\frac{\partial P_1}{\partial S_1}$ is the same as that of $\frac{\partial P_2}{\partial S_1}$.

Proof. We rewrite (5) as

$$P_2^* = \frac{qVP_1}{V - (1 - q)P_1}$$

Taking the first differentiation with respect to S_1 , then

$$\frac{\partial P_2^*}{\partial S_1} = qV \left[\frac{\theta \left(V - (1-q)P_1 \right) + \theta \left(1 - q \right) P_1}{\left(V - (1-q)P_1 \right)^2} \right] = \frac{\theta q V^2}{\left(V - (1-q)P_1 \right)^2}$$

where $\theta = \frac{\partial P_1}{\partial S_1}$. Hence, the proposition holds. QED.

Proof. Proposition 2:

SV show that $\frac{\partial P_1}{\partial S_1} < 0$, $\frac{\partial P_2^*}{\partial S_2^*} < 0$ and $\frac{\partial P_1}{\partial S_2^*}$ under the assumption of fixed F_1 . For $\beta_1 = 0$, $F_2 = F_1$, and thus by definition of asset price, $\frac{\partial P_2^*}{\partial S_1} = 0$. Using Fact 6, we have $\frac{\partial P_2^*}{\partial S_1} < 0$ for $0 < \beta_1 < 1$. Finally, for $\beta_1 = 1$, it is seen that F_2^* rises as S_1 rises (See Fact 4). Thus, P_2^* rises with S_1 . QED.

Proof. Proposition 3:

We first consider the partial investment strategy with $0 < \hat{\beta}_1 < 1$. We rewrite (5) in equality as:

$$P_1 = \frac{P_2^* V}{qV + (1-q) P_2^*}$$

then we can express $\frac{P_2^*}{P_1}$ as:

$$\frac{P_2^*}{P_1} = q + P_2^* \frac{1-q}{V} \tag{A.1}$$

Solving asset price P_t and (1) for P_2^* , we have:

$$\frac{P_2^*}{P_1} = 1 - \frac{S_2^* - S_1}{P_1 - a\hat{\beta}_1 F_1}$$

QED

or

$$P_1 - P_2^* = \frac{P_1(S_2^* - S_1)}{P_1 - a\hat{\beta}_1 F_1}$$
(A.2)

Using (5), we can express (A.2) as

$$\frac{S_2^* - S_1}{P_1 - a\hat{\beta}_1 F_1} = \frac{1 - q}{V} (V - P_2)$$
(A.3)

Since $\frac{\partial P_2^*}{\partial S_1} < 0$, the left hand side of (A.3) is an increasing function of S_1 . Taking the first differentiation we have the following result:

$$\frac{\partial \hat{\beta}_1}{\partial S_1} > \frac{V - S_2^* - (a-1)\hat{\beta}_1 F_1}{(a-1) F_1 \left(S_2^* - S_1\right)} \tag{A.4}$$

The first differentiation of $\kappa = \frac{\hat{\beta_1} F_1}{S_1}$ with respect to S_1 is given by

$$\frac{\partial \kappa}{\partial S_1} = F_1 \left(\frac{\frac{\partial \hat{\beta}_1}{\partial S_1} S_1 - \hat{\beta}_1}{S_1^2} \right)$$

Using (A.4), it is straightforward to show that the numerator is positive:

$$\begin{split} \frac{\partial \hat{\beta}_1}{\partial S_1} S_1 - \hat{\beta}_1 &> \quad \frac{V - S_2^* - (a-1)\,\hat{\beta}_1 F_1}{(a-1)\,F_1\,(S_2^* - S_1)} S_1 - \hat{\beta}_1 \\ &> \quad \frac{S_1\,(V - S_2^*) - (a-1)\,S_2^*\hat{\beta}_1 F_1}{(a-1)\,F_1\,(S_2^* - S_1)} \\ &> \quad \frac{S_1\,(V - S_2^*) - (a-1)\,S_2^*F_1}{(a-1)\,F_1\,(S_2^* - S_1)} \\ &> \quad 0 \end{split}$$

The inequality holds under the stability condition, $V - S_2^* > (a - 1) S_2^* + (a - 1) (F_1 - S_1)$.

Further, from (5) under partial investment case, we can express λ as

$$\lambda = q \frac{V - P_2^*}{V - P_1} \tag{A.5}$$

$$= \frac{P_2^*}{P_1}$$
(A.6)

$$= q + P_2^* \frac{1-q}{V}$$
 (A.7)

The last equality comes from the expression of the first order condition in Eq. (A.1). Using Proposition 2, it is easily seen that as S_1 rises, λ falls. This proves $\frac{\partial \kappa}{\partial S_1} > 0$ and $\frac{\partial \Lambda}{\partial S_1} < 0$ for $0 < \beta_1 < 1$.

Next, we consider the full investment case $\hat{\beta}_1 = 1$, in which case

$$\kappa = \frac{\beta_1 F_1}{S_1} = \frac{F_1}{S_1}.$$

Then, it is easily seen that $\frac{\partial \kappa}{\partial S_1} < 0$. Furthermore, from the definition of λ and Fact 5, it is straightforward to show $\frac{\partial \lambda}{\partial S_1} < 0$.

Finally, when $\hat{\beta}_1 = 0$, the results follow trivially. **QED**

AppendixB. The Generalized Error Correction Model (GECM)

In this appendix, we show detailed steps in developing the GECM, given by equation (1) introduced in the main text. Consider the long-run price-fundamentals relationship given by:

$$f_t = f_t^* + z_t \tag{B.1}$$

where f_t is the observed market price, f_t^* is the fundamental value of the asset, and z_t is the short-term deviation of observed price from its fundamental value. Notice that f_t^* is the martingale difference sequence such that z_t is stationary but serially correlated. For simplicity we represent z_t as an AR process:

$$z_t = \phi z_{t-1} + \varepsilon_t, \ \varepsilon_t \sim iid \ (0, \ \sigma_{\varepsilon}^2)$$

where ε_t is regarded as the mispricing innovations. Taking the first difference of (B.1), and using $\Delta z_t = \kappa z_{t-1} + \varepsilon_t$ with $\kappa = \phi - 1$, we obtain the standard ECM:

$$\Delta f_t = \Delta f_t^* + \kappa z_{t-1} + \varepsilon_t \tag{B.2}$$

The parameter, κ , measures the impact of arbitrage trading activity in correcting the pricing error towards the long-run equilibrium relationship, and lies between -1 and 0. Next, we suppose that the

reduced form data generating process for f_t^* is given by:

$$\Delta f_t^* = \pi \Delta f_{t-1} + e_t, \ e_t \sim iid\left(0, \ \sigma_e^2\right)$$

which allows for the (possible) feedback trading pattern, where positive (negative) π implies positive (negative) feedback trading, and e_t captures the innovations from the fundamental value of the asset, after controlling for feedback trading. This setup is motivated by both empirical and theoretical evidence that market price might potentially induce fundamental changes. It has been documented that the futures market can influence pricing of the underlying index (see, e.g., Chen, 1992).

One important issue is whether or not the pricing error innovation (ε_t) and the innovation of fundamentals (e_t) are independent of each other. If they are not correlated, then the pricing error innovation is random noise. In general, the pricing error is likely to be linked to fundamental news, in which case these two error innovations are correlated. If their contemporaneous correlation is significantly different from zero, Δf_t^* is weakly endogenous with respect to ε_t in (B.2). To deal with this issue, we consider the following regression:

$$\varepsilon_t = \omega e_t + \mu_t = \omega \left(\Delta f_t^* - \pi \Delta f_{t-1}\right) + \mu_t, \ \mu_t \sim iid\left(0, \sigma_\mu^2\right) \tag{B.3}$$

where u_t is uncorrelated with e_t by construction. Then, replacing ε_t in (B.2) by (B.3) and rearranging, we obtain the more efficient ECM as follows:

$$\Delta f_t = \kappa z_{t-1} + \gamma \Delta f_{t-1} + \delta \Delta f_t^* + \mu_t \tag{B.4}$$

where $\gamma = -\omega\pi$ and $\delta = 1 + \omega$. Notice that the model (B.4) accommodates the dynamics of price overreaction or underreaction with respect to fundamental changes through the contemporaneous reaction coefficient, δ , as well as the short-run momentum effects through the coefficient, γ . Only if the market is efficient (i.e. ε_t is iid, in which case $\omega = 0$ trivially), then we expect that one unit (permanent) change in fundamentals should cause one unit change in the market price, instantaneously.

The model developed so far, called the standard (one-period) ECM, is a natural starting point

for an analysis of hypotheses relating to the limits of arbitrage. However, the ECM suffers from a fundamental drawback as only limited dynamics are covered; namely, the speed of adjustment (or duration of mispricing) in (B.2) is measured solely by the error correction coefficient, κ . SV's model suggests that extending the analysis of arbitrage into a two-period model is important. Our extended theoretical analysis shows that the error persistence in the second period is significantly different under alternative investment strategies. Specifically, we examine the second period error correction by measuring the percentage of uncorrected pricing error persisting after time 2 trading.

To accommodate this important issue, we now suppose that the pricing errors, z_t follow an AR(2) process of the form:

$$z_{t} = \phi z_{t-1} + \lambda \left(\phi z_{t-2}\right) + \varepsilon_{t}, \ \varepsilon_{t} \sim iid\left(0, \ \sigma_{\varepsilon}^{2}\right)$$

where $\phi z_{t-2} = (1 + \kappa)z_{t-2}$ is the unarbitraged error carried over from the previous period and the parameter, λ , measures the further pricing impact of these (initial) unarbitraged pricing error components, which we call 'noise momentum' effects. The higher is the λ , the higher is the noise momentum in the price. Thus we finally obtain the two-period GECM given by:

$$\Delta f_t = \kappa z_{t-1} + \lambda^* z_{t-2} + \gamma \Delta f_{t-1} + \delta \Delta f_t^* + \mu_t \tag{B.5}$$

where $\lambda^* = \lambda \phi$.

AppendixC. Dynamic Multipliers

To conduct the dynamic multiplier analysis, it is more convenient to transform the GECM, (B.5), into the following autoregressive distributed lag (ARDL) specification:

$$f_t = \phi_1 f_{t-1} + \phi_2 f_{t-2} + \gamma_0 f_t^* + \gamma_1 f_{t-1}^* + \gamma_2 f_{t-2}^* + \mu_t$$

where

$$\phi_1 = 1 + \kappa + \gamma, \ \phi_2 = \lambda \left(1 + \kappa\right) - \gamma, \ \gamma_0 = \delta, \ \gamma_1 = -\left(\delta + \kappa\right), \ \gamma_2 = -\lambda \left(1 + \kappa\right)$$

It is clear that the overall speed of adjustment associated with (B.5) is evaluated by $\phi_1 + \phi_2 - 1 = \kappa + \lambda(1 + \kappa)$.

Dynamic multipliers are defined as the effect of a unit change in f_t^* at time t on the future trajectory of $f_t(t+h)$ for h = 0, ..., H. Rewriting (B.1) as:

$$\Phi(L) f_t = a + \Gamma(L) f_t^* + \mu_t$$

where $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2$, and $\Gamma(L) = 1 - \gamma_1 L - \gamma_2 L^2$, then the dynamic multipliers can be evaluated by:

$$m_h = \sum_{j=0}^h \frac{\partial f_{t+j}}{\partial f_t^*} = \sum_{j=0}^h \eta_j, \ h = 0, ..., H,$$

where η_j for $j = 0, 1, 2, \dots, H$ can be evaluated easily using the following recursive relationships:

$$\eta_j = \phi_1 \eta_{j-1} + \phi_2 \eta_{j-2} + \dots + \phi_{j-1} \eta_1 + \phi_j \eta_0, \ j = 3, 4, \dots$$

with $\eta_0 = \gamma_0$, $\eta_1 = \phi_1 \eta_0 + \gamma_1$, and $\eta_2 = \phi_1 \eta_1 + \phi_2 \eta_0 + \gamma_2$. Notice that by construction, as $h \to \infty$, $m_h \to \theta$, where θ is the long-run coefficient.

- Abreu, D., and M. K. Brunnermeier. "Synchronization Risk and Delayed Arbitrage." Journal of Financial Economics, 66 (2002), 341-360. "Bubbles and crashes." Econometrica, 71 (2003), 173-204.
- [2] Acharya, V.V., Shin, H.S., Yorulmazer, T., "A theory of slow-moving capital and contagion." (2009) Available at SSRN 1331610
- [3] Acharya, V.V., Amihud, Y., Bharath, S. "Liquidity risk of corporate bond returns: conditional approach." Journal of Financial Economics 110, (2013) 358–386.
- [4] Ali, A.; L. S. Hwang; and M. A. Trombley. "Arbitrage Risk and the Book-to-market Anomaly." Journal of Financial Economics, 69 (2003), 355-373.
- [5] Brenner, R. J., and K. F. Kroner. "Arbitrage, Cointegration, and Testing the Unbiasedness Hypothesis in Financial Markets." Journal of Financial and Quantitative Analysis, 30 (1995), 23-42.
- Brunnermeier, M.K., 2009. Deciphering the liquidity and credit crunch 2007–2008. Journal of Economic Perspectives 23, 77–100.
- Brunnermeier, M.K., Pedersen, L.H., 2009. Market liquidity and funding liquidity. The Review of Financial Studies 22, 2201–2238.
- [8] Cai, C. X.; P. B. McGuinness; and Q. Zhang. "The Pricing Dynamics of Cross-listed Securities: The Case of Chinese A- and H-shares." Journal of Banking and Finance, 35 (2011), 2123-2136.
- [9] Campbell, J. Y. and A. S. Kyle. "Smart Money, Noise Trading, and Stock Price Behavior." Review of Economics Studies, 60 (1993), 1-34.
- [10] Chan, K. "A Further Analysis of the Lead-lag Relationship between the Cash Market and Stock Index Futures Markets." Review of Financial Studies, 5 (1992), 123-152.
- [11] Cochrane, J. "Presidential Address: Discount Rates." Journal of Finance, 66 (2011), 1047-1108.
- [12] DeLong, J. B.; A. Shleifer; L. H. Summers; and R. J. Waldmann. "Noise Trader Risk in Financial Markets." Journal of Political Economy, 98 (1990a), 703-738. "Positive Feedback Investment Strategies and Destabilizing Rational Speculation." Journal of Finance, 45 (1990b), 379-395.

- [13] Dick-Nielsen, J., Feldhütter, P., Lando, D., 2012. Corporate bond liquidity before and after the onset of the subprime crisis. Journal of Financial Economics 103, 471–492.
- [14] Dick-Nielsen, J., & Rossi, M., 2013. Arbitrage crashes: Slow-moving capital or market segmentation?. Available at SSRN 2364362.
- [15] Dwyer, G. P.; Jr, P. R. Locke; and W. Yu. "Index Arbitrage and Nonlinear Dynamics between the S&P 500 Futures and Cash." Review of Financial Studies, 9 (1996), 301-32.
- [16] Duffie, D., 2010. Presidential address: asset price dynamics with slow-moving capital. The Journal of Finance 65, 1237–1267.
- [17] Duffie, D., & Strulovici, B. 2012. Capital mobility and asset pricing. Econometrica, 80(6), 2469-2509.
- [18] Ericsson J. and O. Renault. "Liquidity and Credit Risk." Journal of Finance, 61(2006), 2219-2250.
- [19] Faff, R., X. Cai, and Y. Shin. "Noise Momentum around the World." Abacus (2015).
- [20] Fiess, N. and R. Shankar. "Determinants of Exchange Rate Regime Switching." Journal of International Money and Finance, 28 (2009), 68-98.
- [21] Garbade, K. D., and W. L. Silber. "Price Movements and Price Discovery in Futures and Cash Markets." Review of Economics and Statistics, 65 (1983), 289-97.
- [22] Ghosh, A. "Cointegration and Error Correction Models: Intertemporal Causality between Index and Futures Prices." Journal of Futures Markets, 13 (1993), 193-198.
- [23] Goldstein, I. and A. Pauzner. "Demand-Deposit Contracts and the Probability of Bank Runs." Journal of Finance, 60 (2005), 1293-1327.
- [24] Gromb, D., and D. Vayanos. "Limits of Arbitrage: The State of the Theory" (March 8, 2010).
 Available at SSRN: http://ssrn.com/abstract=1567243.
- [25] Hasbrouck, J. "Measuring the Information Content of Stock Trades." Journal of Finance, 46 (1991), 179-207.

- [26] Hombert, J., and D. Thesmar. "Limits of Limits of Arbitrage: Theory and Evidence" (March 5, 2009). Available at SSRN: http://ssrn.com/abstract=1352285.
- [27] Hull, J. C. Fundamentals of Futures and Options Markets, 6th edition, Prentice Hall (2008).
- [28] Kondor, P. "Rational Trader Risk." Discussion paper, 533, Financial Markets Group, London School of Economics and Political Science, London, UK (2004). http://eprints.lse.ac.uk/24646/.
 "Risk in Dynamic Arbitrage: Price Effects of Convergence Trading." Journal of Finance, 64 (2009), 638-658.
- [29] Kozhan, R. and W. Tham. "Arbitrage Opportunity: A Blessing or a Curse." Warwick Business School, Working Paper (2009).
- [30] Kurov, A., and D., Lasser. "Price Dynamics in the Regular and E-Mini Futures Market." Journal of Financial and Quantitative Analysis, 39 (2004), 365-384.
- [31] Liu, J., and F. A. Longstaff. "Losing Money on Arbitrages: Optimal Dynamic Portfolio Choice in Markets with Arbitrage Opportunities." Review of Financial Studies, 17 (2004), 611–41.
- [32] MacKinlay, A. C., and K. Ramaswamy. "Index-futures Arbitrage and the Behavior of Stock Index Futures Prices." Review of Financial Studies, 1 (1988), 137-158.
- [33] Mancini-Griffoli, Tommaso, and Angelo Ranaldo. "Limits to arbitrage during the crisis: funding liquidity constraints and covered interest parity." Available at SSRN 1569504 (2011).
- [34] Martens, M.; P. Kofman; and T. C. F. Vorst. "A Threshold Error Correction Model for Intraday Futures and Index Returns." Journal of Applied Econometrics, 13(1995), 245-263.
- [35] Mitchell, M., Pedersen, L.H., Pulvino, T., 2007. Slow moving capital. American Economic Review 97, 215–220.
- [36] Oehmke, M. "Gradual Arbitrage" (December 16, 2009). Available at SSRN: http://ssrn.com/abstract=1364126.
- [37] Panos, A. M.; R. Nobay; and D. A. Peel. "Transactions Costs and Nonlinear Adjustment in Real Exchange Rates: An Empirical Investigation." Journal of Political Economy, 105 (1997), 862-879.

- [38] Pontiff, Jeffrey. "Costly arbitrage and the myth of idiosyncratic risk." Journal of Accounting and Economics 42.1 (2006): 35-52.
- [39] Roll, R.; E. Schwartz; and A. Subrahmanyam. "Liquidity and the Law of One Price: The Case of the Futures/Cash Basis." Journal of Finance, 62 (2007), 2201–2234.
- [40] Schuster, P., and M. Uhrig-Homburg. "Limits to arbitrage and the term structure of bond illiquidity premiums." Journal of Banking and Finance, 57 (2015) 143-159
- [41] Sarno, L., and G. Valente. "The Cost of Carry Model and Regime Shifts in Stock Index Futures Markets: An Empirical Investigation." Journal of Futures Markets, 20 (2000), 603-624.
- [42] Schuster, P., & Uhrig-Homburg, M. 2015. Limits to arbitrage and the term structure of bond illiquidity premiums. Journal of Banking & Finance, 57, 143-159.
- [43] Sercu, P.; R. Uppal; and C. V. Hulle. "The Exchange Rate in the Presence of Transaction Costs: Implications for Tests of Purchasing Power Parity." Journal of Finance, 50 (1995), 1309-1319.
- [44] Shleifer, A., and R. W. Vishny. "The Limits of Arbitrage." Journal of Finance, 52 (1997), 35-55.
- [45] Stein, J. C. "Presidential Address: Sophisticated Investors and Market Efficiency." Journal of Finance, 64 (2009), 1517-1548.
- [46] Stoll, H. R., and R. E. Whaley. "Expiration Day Effects of Index Options and Futures." Monograph Series in Finance and Economics, Monograph 1986-3. "The Dynamics of Stock Index and Stock Index Futures Returns." Journal of Financial and Quantitative Analysis, 25 (1990), 441-468.
- [47] Tse, Y. 2001, Index arbitrage with heterogeneous investors: A smooth transition error correction analysis. Journal of banking & finance, 25(10), 1829-1855.
- [48] Wahab, M., and M. Lashgari. "Price Dynamics and Error Correction in Stock Index and Stock Index Futures Markets: A Cointegration Approach.", Journal of Futures Markets. 13 (1993), 711-742.
- [49] Yadav, P. K.; P. F. Pope; and K. Paudyal. "Threshold Autoregressive Modeling in Finance: The Price Differences of Equivalent Assets." Mathematical Finance, 4 (1994), 205-221.

[50] Zhang, X. F. "Information Uncertainty and Stock Returns." Journal of Finance, 61 (2006), 105-137.